
Generative Search and Brand Visibility Engineering

GEO Advanced Studies Project

Research methods, evidence, and reproducible engineering

Independent working edition · Version 0.6 · Research cutoff: August 24, 2026

Reading Contract	2
1 From Ranked Links to Generated Answers	4
2 Evidence Before Optimization	14
3 Web Acquisition, Representation, and Query Planning	20
4 Retrieval, Fusion, Reranking, and Context Allocation	25
5 Generation, Citation, Attribution, and Source Influence	34
6 Metrics, Labels, Judges, and Benchmark Design	43
7 Experiments, Causal Inference, and Temporal Uncertainty	55
8 Content and Evidence Engineering	66
9 Optimization Algorithms and Agentic GEO	72
10 Multimodal, Platform, Brand, and Evidence Ecosystems	82
11 Manipulation, Detection, Defense, and Mechanism Design	90
12 Governance, Standards, Regulation, and Incident Response	100
References	107
A Versioned Research Map	111
B Reproducibility and Release Protocols	117
C Standards and Official-Guidance Register	119
D Exercises and Reproduction Problems	124
E Notation, Symbols, and Terminology	139
F Prerequisite Bridges: IR, Statistics, and Causal Inference	155

Reading Contract

Purpose and scope

Search interfaces increasingly synthesize an answer before a user visits a source. For organizations, this creates a tempting but underspecified question: “How do we optimize for AI answers?” A useful answer requires more than a list of writing tactics. It requires a model of the information pipeline, a theory of what is being measured, and a discipline for deciding which observations can support which claims.

These English notes accompany, but do not reproduce, the course’s week and grading structure. They are designed for readers in information retrieval, machine learning, marketing science, communications, product, analytics, and governance. Mathematics is used when it removes ambiguity; a lab is specified when a claim should be tested rather than repeated.

Evidence, claims, and versioning

The initial paper library, GEO-42, is derived from the latest user-provided Zotero export. It contains 42 records: 4 with a recorded publication year of 2024, 13 of 2025, and 25 of 2026; 41 are typed as preprints and one as a journal article. Every record contains a DOI and a local PDF attachment, while one lacks a URL and abstract. Five records carry a metadata warning, most commonly a disagreement between the exported year and the year encoded by the arXiv identifier. These warnings are not silently “fixed” because the export is an auditable source object. The normalized catalog, BibTeX file, and anomaly report are separate artifacts.

The paper library is only one evidence layer. Official documentation, standards, source code, benchmark data, blogs, white papers, tutorials, and videos each have a different evidentiary ceiling. A vendor white paper can document a workflow without proving its effect; a preprint can establish a testbed result without revealing a production platform; an official help page can define a policy without describing its ranking model.

Contributions, non-claims, and review state

The manuscript makes four design contributions. First, it decomposes generative-search visibility into versioned conditional events rather than one ranking score. Second, it joins source identity, atomic claims, evidence spans, and provenance in a Claim–Evidence–Source control layer. Third, it specifies measurement and intervention protocols whose units, denominators, uncertainty, failure states, and evidence ceilings remain visible. Fourth, it connects those methods to reversible engineering, safety, and governance gates.

These are working scholarly contributions, not evidence that the framework is optimal, that a named platform implements its stages, or that an intervention improves real-world visibility or business outcomes. The controlled project contains the L^AT_EX source, bibliography, claim and notation registers, 42 evidence cards, figure-provenance records, and synthetic lab specifications. Automated build and rendering checks do not replace human scientific review, independent reproduction, accessibility review, or publication-rights approval.

Source identities, version locks, numerical corrections, and negative findings are retained rather than silently overwritten. A public repository, approved author metadata, citation format, license, and errata channel will be declared only when the responsible authors approve them; their absence in this draft is not filled with placeholder affiliations or invented contact details.

Reading paths

Research path

Chapters 1–7, followed by the paper map and reproducibility appendix. Focus on estimands, protocols, uncertainty, and external validity.

Engineering path

Chapters 1, 3–5, 8, and 9. Follow the source lifecycle from availability to retrieval, generation, attribution, correction, and controlled intervention.

Brand and content path

Chapters 1, 2, 6–10. Build a fact base before optimizing expression, and distinguish mention, citation, absorption, prominence, and business outcomes.

Governance path

Chapters 2, 7, 11, and 12. Use the claim ledger, threat model, release gates, standards register, and incident protocol.

What the notation promises

Symbols in this manuscript name observable units whenever possible. If a stage inside a closed system is not observable, the text does not disguise a proxy as that hidden variable. Platform observations are versioned by date, surface, locale, language, account state, and repetition. A result is never promoted from “observed in this protocol” to “how the platform works” without an additional identification argument.

1 From Ranked Links to Generated Answers

Key idea

Generative engine optimization is not one ranking problem. It is the study and responsible intervention of a sequence in which a source may be discovered, retrieved, admitted to context, used in generation, attributed, interpreted, and acted upon. A measurement is meaningful only after it names which event in that sequence is its target.

1.1 Questions this chapter answers

By the end of the chapter, a reader should be able to answer four questions without collapsing them into one visibility score:

1. What is the observable research object when a search interface returns a generated answer rather than a ranked list?
2. Which events—mention, retrieval, source use, attribution, fidelity, or action—does a reported metric actually count?
3. What can and cannot be inferred about latent pipeline stages from a response, displayed citations, and an interaction trace?
4. How should an intervention objective preserve user utility, evidence, fidelity, and safety while seeking source-side visibility?

1.2 An interface change is not yet a mechanism

Traditional web search makes a ranked list visible. A user can inspect titles, snippets, domains, positions, and links before choosing a source. A generative search surface may instead show a synthesized answer, a set of citations, a follow-up dialogue, or an action. This changes the allocation of attention, but the visible interface does not disclose all of the machinery that produced it. The system may have rewritten the query, issued several searches, combined lexical and neural retrieval, reranked candidates, discarded passages for lack of context budget, generated an answer, and attached citations in a separate step. Some products reveal fragments of this process; others reveal only the final response.

The early GEO formulation made the creator’s visibility inside generated responses an explicit optimization object and introduced a benchmark with fixed source sets [1]. That formulation was foundational, but later work has made a critical distinction: optimizing within a pre-selected context does not determine whether the modified document would survive retrieval and reranking in a large corpus. SAGEO Arena was designed to evaluate the retrieval–reranking–generation pipeline and reports that an edit that helps at one stage can harm another [23]. The two settings answer different questions; neither should be presented as a universal model of a proprietary platform.

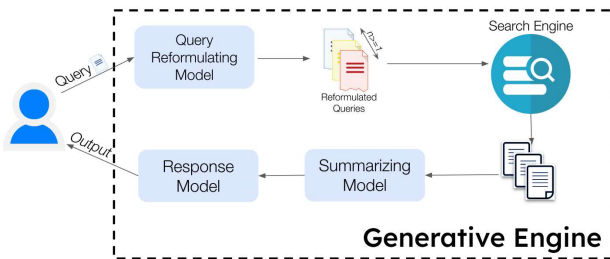


Figure 1: A foundational research abstraction of a generative engine: query reformulation, retrieval, source summarization, and response generation.

Source note. Rectangular vector crop of Aggarwal et al. [1, Fig. 2], physical PDF page 3, CC BY 4.0. The diagram is conceptual, not a disclosure of any named production system.

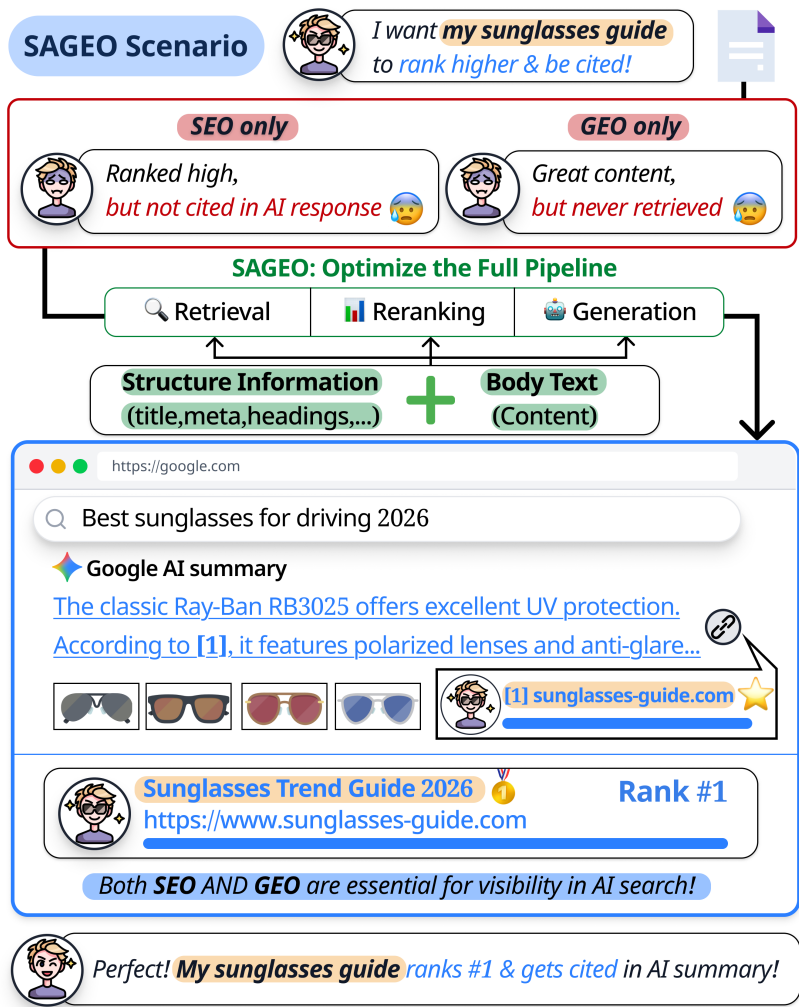


Figure 2: Stage mismatch in a reconstructed search-augmented GEO environment: ranking alone can fail at generation, while generation-only optimization can fail before retrieval.

Source note. Rectangular crop of Kim et al. [23, Fig. 1], physical PDF page 2, CC BY 4.0. It summarizes that benchmark’s staged testbed; it is not a commercial-platform architecture or a universal performance law.

Table 1: Controlled source objects for the synthetic running case. The values are didactic inputs, not claims about a real organization or platform.

Controlled object	Evidence scope	Role in later chapters
Product page	Nominal range 0–50 mg/L; no cold-weather validation.	Candidate generation and claim segmentation.
Calibration report	Independently validated range 5–35 mg/L at 10–25 °C; named protocol and uncertainty.	Evidence support within the tested range.
Current safety notice	A warmed enclosure is required below 5 °C; the notice supersedes the launch manual.	Qualification, conflict, and supersession tests.

Definition 1.1: Generative search system

For a user state u and query q , a generative search system is represented by a stochastic kernel

$$(Y, \mathcal{A}, L) \sim \mathcal{G}_\theta(\cdot \mid q, u, \mathcal{W}_t), \quad (1)$$

where $\mathcal{W}_t$ is the accessible information environment at time t , θ denotes the system state that the protocol can identify or version, Y is a generated response, $\mathcal{A}$ is a set of displayed attributions or citations, and L is any observable interaction trace. A realized observation is written $(y, \neg, \ell)$. The definition does not assume that every system searches the open web, exposes citations, or uses one retrieval call.

The explicit time argument matters. Web documents change, indexes lag, model versions move, and platform policies vary. A response is not a timeless fact about a brand or a source; it is an observation under a system state.

1.3 Running case: Orchid Bay Instruments

The book uses one synthetic organization to keep the abstract pipeline tied to an inspectable task. *Orchid Bay Instruments* is a fictional manufacturer of portable water-quality spectrometers. Its governed source set contains a product specification, an independently issued calibration report, and a safety notice. A procurement analyst asks:

Which portable spectrometer is suitable for field nitrate screening when samples may exceed 40 mg/L and the device must operate below 5 °C?

The case is deliberately inconvenient. The product page advertises a nominal 0–50 mg/L measurement range, the calibration report covers only 5–35 mg/L at 10–25 °C, and the safety notice requires a warmed enclosure below 5 °C. A fluent answer that recommends the device from the product page alone may mention the correct range while omitting the temperature condition and exceeding the independent evidence range. Visibility, factual fidelity, source selection, and decision utility can therefore move in different directions.

Later chapters will version these objects, retrieve their passages, allocate a context budget, score claim–citation support, design a repeated-measure panel, and evaluate a reversible evidence-surfacing intervention. No result from the case is empirical evidence about a production system.

Table 2: Different observables require different units and interpretations.

Observable	Unit	A positive event means	It does not establish
Mention	entity–response	The entity string or a resolved alias occurs	Source use, accuracy, or persuasion
Citation	source–response	A displayed attribution resolves to the source	That the answer used a particular claim faithfully
Absorption	claim–source–response	Source-specific evidence is reflected in the answer	Correct attribution or business value
Fidelity	claim–response	The answer preserves the scoped proposition	Which source or pipeline stage caused it
Prominence	span–response	The entity or source has a declared position or share	Positive sentiment or user attention
Action	user–task	A predeclared behavior occurs	Why it occurred, without an identification design

1.4 The units that are routinely conflated

A precise GEO study separates at least six objects:

Entity

The organization, product, person, place, or concept a user is asking about.

Claim

A minimal proposition about an entity, with time, locale, and other scope conditions.

Source

A governed information object—for example a report, product page, dataset, or regulatory record—that can support or contest a claim.

Passage

A retrievable segment derived from a source. Segmentation can change what evidence is available to a ranker or generator.

Answer

The generated language shown to the user, including uncertainty, ordering, and recommendations.

Outcome

A declared consequence: mention, citation, factual absorption, correct description, visit, qualified lead, purchase, or another action.

An entity can be mentioned without a citation. A source can be cited while none of its distinctive evidence appears in the answer. A claim can be absorbed but attributed to the wrong source. A brand can receive a prominent mention that is factually harmful. A user can act without visiting any cited page. Consequently, “AI visibility” is not a scalar until a protocol defines both the event and its value.

The recent measurement literature reinforces this separation. One study explicitly distinguishes *citation selection* from *citation absorption*: a platform may choose a page at the search layer without allowing that page to shape the final answer to the same degree [65]. Work on citation failures similarly treats non-citation as a family of failures distributed across the pipeline rather than as one content-quality defect [47]. These papers are evidence for useful constructs and protocols; their reported platform comparisons remain tied to their respective datasets, dates, and access methods.



Each arrow is a conditional event, not a guaranteed funnel conversion.

Figure 3: The conditional visibility chain. The diagram is an original conceptual decomposition; individual platforms may merge, repeat, or omit observable stages.

1.5 A conditional visibility chain

Let s be a governed source and q a query sampled from a declared distribution Q . Define binary or graded stage variables:

- Z_D : the source is discoverable and fetchable,
- Z_R : the source or one of its passages is retrieved,
- Z_C : the source enters the effective generation context,
- Z_G : the source contributes to generated content,
- Z_T : the contribution is attributed faithfully,
- Z_A : the response supports the declared downstream outcome.

For an outcome that requires all six stages, the chain rule yields

$$\begin{aligned}
 \Pr(Z_D, Z_R, Z_C, Z_G, Z_T, Z_A \mid q, s) &= \Pr(Z_D \mid q, s) \\
 &\quad \cdot \Pr(Z_R \mid Z_D, q, s) \\
 &\quad \cdot \Pr(Z_C \mid Z_R, Z_D, q, s) \\
 &\quad \cdot \Pr(Z_G \mid Z_C, Z_R, Z_D, q, s) \\
 &\quad \cdot \Pr(Z_T \mid Z_G, Z_C, Z_R, Z_D, q, s) \\
 &\quad \cdot \Pr(Z_A \mid Z_T, Z_G, Z_C, Z_R, Z_D, q, s). \tag{2}
 \end{aligned}$$

No independence assumption is required. Indeed, dependence is the central problem: a rewrite can increase one conditional probability and reduce another.

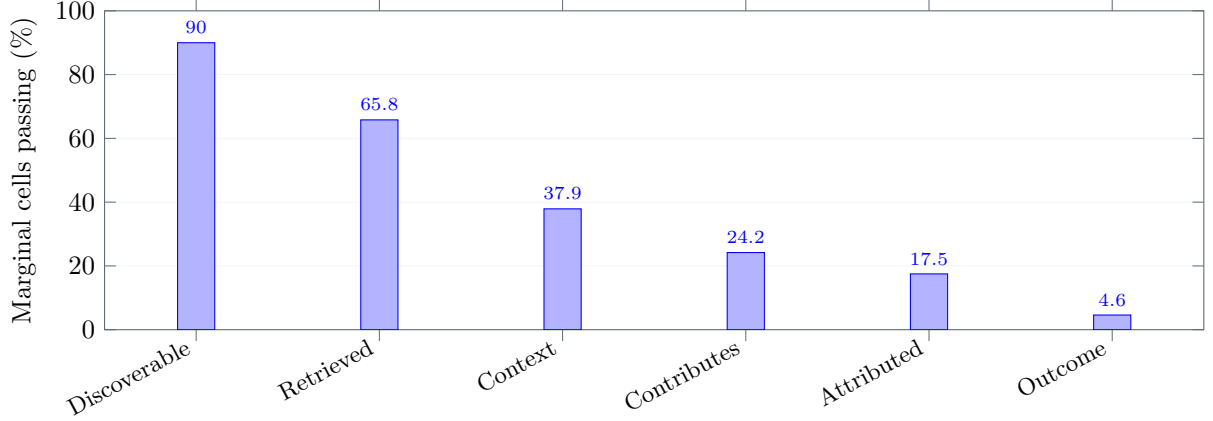
The chain is a diagnostic device, not a claim that a closed system implements six named modules. If retrieval is unobservable, a citation may serve as a proxy for some events, but the report must call it a proxy. If a platform searches several times, Z_R may be a sequence rather than a single event. If the generator produces uncited parametric knowledge, Z_G can occur without a displayed Z_T .

1.5.1 A worked stage-survival audit

Consider a didactic audit with 240 frozen query–system–time cells for the Orchid Bay case. The collection protocol observes acquisition directly in an open index, records retrieved and context passages in a controlled system, and uses double-coded claim and attribution labels. The nested counts in table 3 are synthetic; they exist so that the estimand and arithmetic can be reproduced without implying a result for a named platform.

Table 3: Synthetic nested counts for a stage-resolved audit. Conditional survival uses the preceding observed stage as its denominator; marginal survival always uses all 240 cells.

Stage	Passing cells	Conditional survival	Marginal survival
Discoverable Z_D	216	$216/240 = 0.900$	0.900
Retrieved Z_R	158	$158/216 = 0.731$	0.658
In context Z_C	91	$91/158 = 0.576$	0.379
Contributes Z_G	58	$58/91 = 0.637$	0.242
Attributed Z_T	42	$42/58 = 0.724$	0.175
Outcome Z_A	11	$11/42 = 0.262$	0.046

**Figure 4:** Stage survival in the synthetic audit from table 3. The plot is generated directly from the declared counts; it is a worked teaching artifact, not a production-platform estimate.

The product of the observed conditional rates telescopes:

$$\frac{216}{240} \frac{158}{216} \frac{91}{158} \frac{58}{91} \frac{42}{58} \frac{11}{42} = \frac{11}{240} \approx 0.0458. \quad (3)$$

This equality is bookkeeping, not an independence argument. It also does not identify which internal module caused a loss. The largest count reduction is between retrieval and effective context (158 to 91), but that pattern could reflect relevance, redundancy, context budget, safety filtering, or an error in the stage labels. A diagnosis must return to cell-level traces.

Proposition 1.2: Observational equivalence limits mechanism claims

Let two candidate system states θ_1 and θ_2 induce the same joint distribution of the audit observables $(Y, \mathcal{A}, L)$ for every tested $(q, u, \mathcal{W}_t)$. Any decision rule that is measurable only with respect to those observables has the same distribution under θ_1 and θ_2 . The audit therefore cannot identify which candidate internal mechanism generated the observations.

Proof. Let $d(Y, \mathcal{A}, L)$ be any observable decision rule and let B be a measurable set in its output space. Equality of the joint observable distributions implies

$$\Pr_{\theta_1}(d(Y, \mathcal{A}, L) \in B) = \Pr_{\theta_2}(d(Y, \mathcal{A}, L) \in B).$$

Hence no statistic computed only from the audit observables can distinguish the two candidate mechanisms. Iden-

Table 4: A problem-oriented comparison of adjacent practices.

Practice	Primary observable	Typical intervention surface	GEO-relevant limitation
SEO	index eligibility, ranked links, clicks	crawlability, information architecture, content, links	A high-ranked page may not enter or shape a generated answer
AEO	direct-answer eligibility and extraction	concise answer structures, entities, structured data	Extraction does not by itself establish multi-source attribution or fidelity
RAG engineering	retrieval and grounded generation in a controlled corpus	chunking, embeddings, retrievers, rerankers, prompts, models	Results may not transfer to a web-scale or proprietary ecosystem
Digital PR	third-party publication and reputation	newsworthiness, expert sources, distribution, relationships	Coverage frequency is not equivalent to retrieval, citation, or causal business impact
GEO	declared events inside a generative-search pipeline	source availability, evidence, expression, ecosystem, evaluation	Mechanisms are often partially observed; optimization can create manipulation incentives

tification requires an additional observable, intervention, or assumption whose induced distributions differ. $\square$

1.5.2 From a scalar to a visibility profile

For a source s , query panel Q , system m , and time window T , define a visibility profile

$$\mathbf{V}(s; Q, m, T) = [\hat{p}_{\text{ret}}, \hat{p}_{\text{cite}}, \hat{p}_{\text{abs}}, \hat{a}_{\text{attr}}, \hat{f}_{\text{claim}}, \hat{\sigma}_{\text{time}}], \quad (4)$$

where the hats emphasize that these are protocol-dependent estimates: retrieval incidence when observable, citation incidence, source-specific absorption, attribution accuracy, claim fidelity, and temporal variability, respectively. Their estimators, denominators, missingness rules, and intervals must be declared separately. This diagnostic profile is not a proposal to average incompatible quantities into a proprietary “GEO score.” It is a reminder that a result can improve along one dimension while degrading along another.

Static measurement can be especially misleading when generated responses vary across repetitions and time. Recent work argues for treating visibility as a distribution rather than a one-shot value [40]. We develop the associated sampling and uncertainty protocol in [chapter 6](#).

1.6 GEO, SEO, AEO, and adjacent fields

GEO inherits infrastructure from web search but changes the observable output and the strategic environment. The categories in [table 4](#) should be treated as overlapping problem formulations, not professional territories with fixed borders.

The literature connecting search services and large language models covers query understanding, retrieval, ranking, data annotation, and system evaluation well beyond the marketing label GEO [58]. A rigorous curriculum therefore cannot begin with content tactics. It must begin with information retrieval, experimental design, and source governance.

1.7 Three stakeholders and a misaligned objective

Let U_u denote user utility, U_p platform utility, and U_s source or supplier utility. A creator-side intervention may maximize expected exposure $\mathbb{E}[U_s]$ while reducing answer quality or attribution accuracy. A platform defense may preserve U_p while unfairly suppressing legitimate new sources. A useful objective must therefore be constrained:

$$\max_{\delta \in \Delta} \mathbb{E}[U_s(Y^\delta, \mathcal{A}^\delta, L^\delta)] \quad \text{subject to} \quad \begin{cases} \text{Fidelity}(s \oplus \delta) \geq \tau_f, \\ \text{EvidenceCoverage}(s \oplus \delta) \geq \tau_e, \\ \text{Risk}(s \oplus \delta) \leq \tau_r. \end{cases} \quad (5)$$

Here δ is a permitted source change and $(Y^\delta, \mathcal{A}^\delta, L^\delta) \sim \mathcal{G}_\theta(\cdot \mid q, u, \mathcal{W}_t^{(s, \delta)})$, where $\mathcal{W}_t^{(s, \delta)}$ is the information environment in which source s carries the governed revision. The thresholds are governance decisions, not model parameters discovered automatically.

This multi-stakeholder view is increasingly explicit in the research record. Risk-oriented work frames GEO as a potential route to ranking manipulation and undisclosed influence [53]. More recent mechanism-design work models repeated supplier–platform adaptation and evaluates incentives for verifiable content rather than purely punitive filters [59]. The latter’s empirical results belong to its stated simulations and benchmarks; the broader contribution for these notes is the shift from one-shot tactics to repeated ecosystem dynamics.

Boundary and failure conditions

The existence of a statistically significant effect in a benchmark does not license the claim that the same edit will improve a named production platform. Transfer requires at least a compatible retrieval regime, candidate set, query distribution, answer surface, model state, metric, and time window.

1.8 A stage-resolved audit protocol

[Algorithm 1](#) turns the chapter’s distinctions into a minimum executable protocol: an unavailable stage is recorded as unobserved, never inferred from downstream success or coded as zero. Later chapters expand its acquisition, retrieval, annotation, and repeated-measure steps.

Algorithm 1 Stage-resolved visibility audit

Inputs: Frozen source and query manifests; declared audit cells $\mathcal{C}$; stage-observation, missingness, and stop rules.

Outputs: Cell ledger E ; stage estimates and intervals; deviation and mechanism-uncertainty logs.

- 1: Freeze source hashes, query IDs, system cards, outcomes, and the analysis plan.
- 2: Construct the Cartesian audit cells $\mathcal{C}$ required by the sampling design; retain query and time clusters.
- 3: **for all** $c \in \mathcal{C}$ **do**
- 4: Collect the response, attributions, trace, time, system state, and collection status.
- 5: **for all** $Z \in \{Z_D, Z_R, Z_C, Z_G, Z_T, Z_A\}$ **do**
- 6: **if** Z is directly observed under the frozen rule **then**
- 7: Record value, locator, coder or instrument, and adjudication state.
- 8: **else if** a predeclared proxy for Z is available **then**
- 9: Record the proxy separately and preserve its validation boundary.
- 10: **else**
- 11: Record $Z = \text{UNOBSERVED}$; never infer zero from downstream events.
- 12: **end if**
- 13: **end for**
- 14: Classify refusal, outage, censoring, parse failure, and absent UI.
- 15: **end for**
- 16: Deduplicate sources; adjudicate the double-coded sample; retain disagreements and exclusions.
- 17: Estimate each event with its unit, denominator, weights, clusters, interval, and sensitivity analysis.
- 18: Release E , manifests, tables, and logs; stop on a severe evidence, fidelity, or safety gate.

1.9 A minimal research question template

Every empirical claim in these notes can be rewritten into the following template:

For *which source population* and *which query distribution*, under *which system version and surface*, does intervention δ change outcome Y relative to comparator c , over *which time window*, by how much uncertainty, and under which fidelity and risk constraints?

If a sentence cannot fill these slots, it may still be a useful observation or hypothesis, but it is not yet a decision-grade effect claim.

Reproduction Studio 1 · Decompose a visibility claim

Input. Five public claims about improving “AI visibility,” drawn from at least one paper, one platform document, one vendor white paper, one blog, and one video.

Task. For each claim, identify the source object, intervention, candidate pipeline stage, observable outcome, comparator, query distribution, system state, and evidence level. Draw the conditional chain and mark every unobserved transition.

Deliverable. A one-page claim registry plus a corrected version of each claim that does not exceed its evidence.

Stop rule. Stop when every noun in the outcome is operationalized and every causal verb has either an identification argument or has been replaced by an observational verb.

Chapter summary

1. Generated answers transform visibility into a sequence of conditional events rather than eliminating the classical search pipeline.
2. Mention, citation, absorption, fidelity, prominence, and action are different outcomes with different units.
3. Observable-equivalent systems cannot be separated by mechanism claims without an additional observable, intervention, or identifying assumption.
4. Fixed-context and end-to-end benchmarks answer different questions.
5. A stage-resolved audit keeps proxies and unobserved states distinct and reports a profile rather than an unexplained scalar.
6. A responsible GEO objective constrains exposure by evidence, fidelity, and risk. The next chapter establishes how claims and sources earn their place in that objective.

2 Evidence Before Optimization

Key idea

The first deliverable of a GEO program is not optimized copy. It is a versioned map from claims to evidence and sources, together with a record of what each source can and cannot establish. Optimization performed before this map exists can only make uncertainty more fluent.

2.1 Documents have identities; claims have ceilings

Papers, standards, official documentation, repositories, white papers, blog posts, videos, and screenshots are not interchangeable containers for facts. Each has an issuer, version, method, intended audience, and change process. A standard can define a requirement without proving an intervention’s effect. A source-code repository can expose an implementation without establishing external validity. A vendor case study can document a customer narrative while leaving the denominator, comparator, and failure cases unknown.

This motivates a two-step audit. The **document identity audit** establishes what the item is. The **claim audit** establishes what propositions, if any, the item supports.

Definition 2.1: Source identity record

A source identity record is the tuple

$$I(s) = \langle \text{issuer, type, title, version, date, status, method, artifacts, license, URL} \rangle. \quad (6)$$

Unknown fields remain explicitly unknown. Visual polish, an institutional logo, or a DOI does not fill missing method or artifact fields.

2.1.1 An evidence ladder for use, not prestige

Table 5 organizes evidence by the claims a source can normally sustain. The levels are not a ranking of intellectual worth. A product help page can be the best source for a current product setting; a randomized study can be the best source for a treatment effect; neither substitutes for the other.

The original GEO benchmark, for example, provides controlled evidence about specified content modifications and visibility metrics in its experimental settings [1]. An end-to-end benchmark such as CC-GSEO-Bench adds a content-centric view of source influence under retrieval contexts [6]; SAGEO Arena adds large-corpus retrieval and reranking [23]. None of these sources, by itself, discloses the causal mechanism of every commercial answer engine.

2.2 The Claim–Evidence–Source graph

Marketing asset libraries usually organize files by campaign, channel, or format. A generative search program needs a different primitive: the smallest proposition the organization is willing and able to defend.

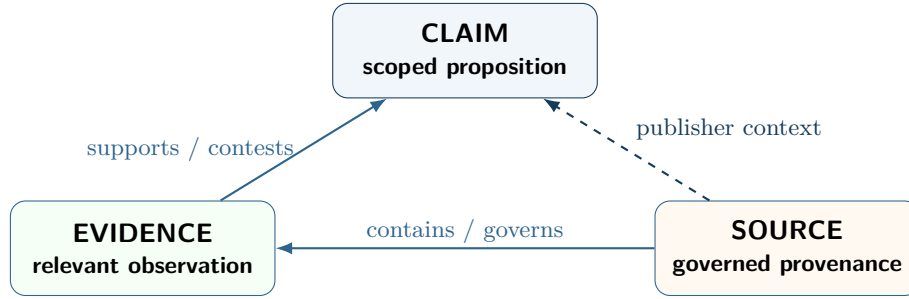
Definition 2.2: Scoped claim

A scoped claim is

$$c = \langle h, r, o, \tau, \lambda, \kappa, q \rangle, \quad (7)$$

Table 5: Evidence layers and their typical claim ceilings.

Layer	Typical material	Can normally support	Common overreach
E0	promotional copy, uncited post, social clip	what an actor publicly claims	treating repetition or reach as validation
E1	official documentation, policy, standard	declared behavior, status, requirements, terminology	inferring hidden ranking weights or real-world efficacy
E2	method-rich white paper, tutorial, code walkthrough	a reproducible procedure or implementation choice	generalizing beyond disclosed data and system
E3	benchmark or observational study	associations and performance under a stated protocol	converting a benchmark delta into platform causality
E4	controlled intervention with comparator	an effect in the experimental environment	ignoring interference, drift, or external validity
E5	independent multi-setting replication or strong field design	a more transportable effect estimate	calling a heterogeneous effect universal

**Figure 5:** The Claim–Evidence–Source graph. Support is typed, scoped, and versioned; a URL alone is not an evidence edge.

where h is a head entity, r a relation or predicate, o an object or value, τ a validity interval, λ a locale, κ other applicability conditions, and q the claim status (for example, proposed, approved, disputed, expired, or withdrawn).

An evidence item e is not just a file. It is a relevant span, table, dataset, measurement, record, or observation with provenance. A source s contains or governs that evidence. We represent support as a typed edge

$$e \xrightarrow[\rho, \iota, \phi]{\text{supports}} c, \quad (8)$$

where ρ is relevance to the exact claim, ι is independence from the claimant, and ϕ is freshness relative to the claim’s time scope. Contradictory, qualifying, and non-informative edges are first-class; they are not deleted to make a dashboard green.

Table 6: Examples of claim-specific minimum evidence.

Claim family	Minimum internal evidence	Stronger external support
Product capability	versioned specification, test condition, owner approval	independent test or current user documentation
Performance comparison	protocol, baseline, dataset, uncertainty, failure cases	independent replication on a relevant task
Market position	defined market, period, geography, denominator	traceable third-party dataset and methodology
Customer result	authorized case record, treatment timeline, metric definition	comparator or design that addresses alternative causes
Certification	certificate identifier, scope, issuer, validity dates	issuer registry or signed verification
Safety or compliance	control description, responsible owner, audit status	applicable standard mapping and independent assessment

2.2.1 Minimal evidence adequacy

Let $E(c)$ be the evidence set attached to claim c . A practical adequacy function can be represented as

$$A(c) = \min \{ \text{Coverage}(E(c)), \text{Relevance}(E(c), c), \text{Freshness}(E(c), c), \text{Authority}(E(c)), \text{Consistency}(E(c)) \}. \quad (9)$$

The minimum makes a governance point: a beautifully sourced claim can still be unusable if every source is expired, and a current source can still be useless if it addresses a broader or different proposition. The function is a release gate, not a scientifically calibrated universal score.

2.3 Reading an empirical GEO paper without flattening it

A paper record should have at least four layers:

1. **Identity:** title, authors, venue or preprint status, version, date, DOI, code/data links, and license.
2. **Design:** research question, unit, sampling frame, systems, query distribution, candidate corpus, intervention, comparator, and metrics.
3. **Result:** estimand, point estimate, uncertainty, heterogeneity, negative results, and failure cases.
4. **Transfer:** which components resemble the deployment of interest and which do not.

The phrase “up to 40%” in the foundational GEO paper is a maximum reported improvement under the paper’s benchmark and metric, not a generic expected uplift for a website [1]. The correct editorial action is not to suppress a striking number; it is to restore its outcome, denominator, comparator, variation across methods and domains, and experimental setting.

Likewise, the observation that a formatting feature does not improve absorption in one multi-platform study is valuable precisely because it is a negative result under a disclosed data pipeline [65]. It does not prove that formatting is irrelevant to retrieval, readability, accessibility, or every other system. Negative results narrow claims; they do not erase adjacent mechanisms.

Evidence note

The first edition’s 42-record library is recent and therefore unusually rich in preprints. Recency is useful for mapping an emerging field, but it increases the need to track version changes, peer-review status, artifact availability,

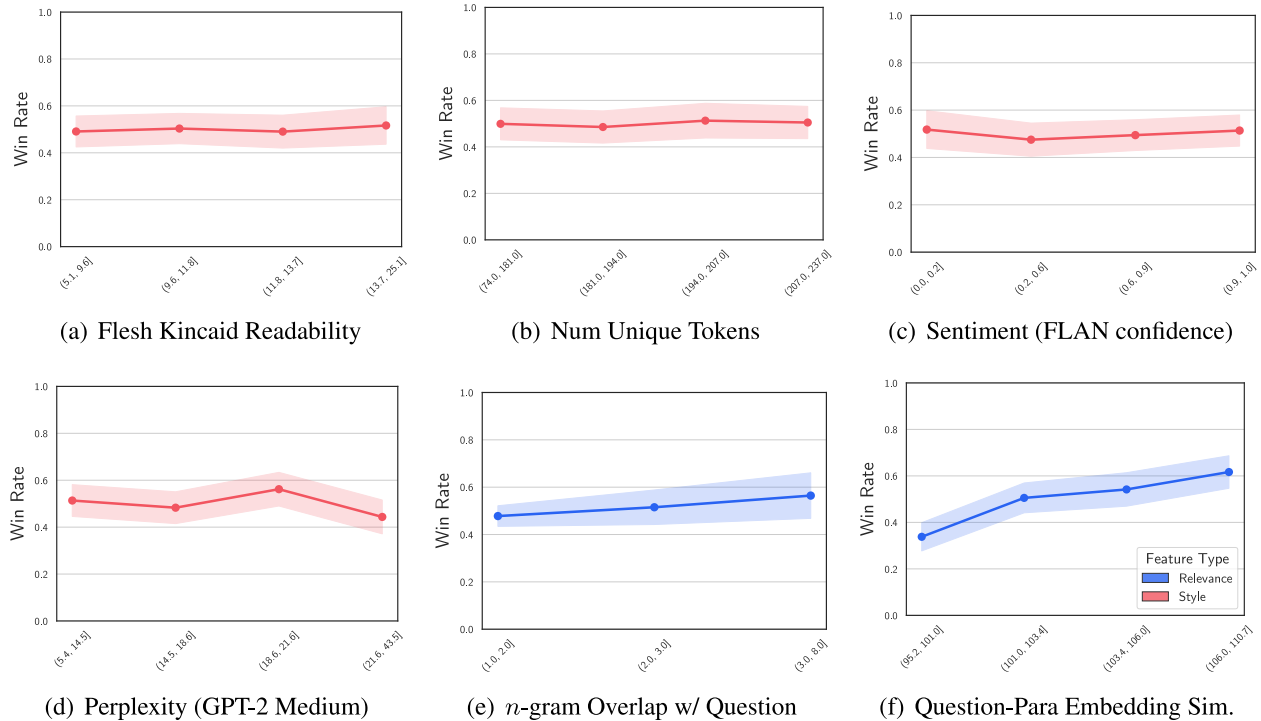


Figure 6: Feature-binned paragraph win rates in a fixed conflicting-evidence setting. Relevance features change more clearly than the measured style features, but the plots are descriptive associations rather than causal rules.

Source note. Vector crop of Wan et al. [50, Fig. 4], official ACL Anthology PDF page 7, CC BY 4.0; framing was cropped and six Type 3 font programs were converted to vector outlines, with marks, values, colors, and geometry unchanged. The panel reports LLaMA-2-Chat-13B results on 242 filtered examples with 95% intervals and does not establish a universal recipe for persuasion, truth, retrieval, or citation.

and overlapping authorship or benchmarks. Every number that enters a table or figure in the Notes must be checked against the attached PDF version, not only against an abstract or a blog summary.

2.4 The GEO-42 snapshot

The latest export has strong basic completeness: all 42 records include a DOI and a local PDF; 41 include both URL and abstract. No duplicate normalized title, DOI, or URL was detected. Five records require editorial attention:

1. four have a recorded publication year that differs from the year encoded by the arXiv identifier;
2. one journal record lacks a URL and abstract in the export.

These conditions lead to three separate fields in the eventual paper index: *source year*, *arXiv identifier year*, and *verified publication status*. Collapsing them into one silently corrected “year” would make the library easier to display but harder to audit.

The automatic first-pass taxonomy also has a known limitation. Keyword rules over-assign papers to “optimization” when titles use the field name GEO but the substantive contribution concerns measurement, system design, cultural variation, or governance. The appendix therefore treats automatic categories as routing hypotheses. Chapter placement is finalized only after abstract, method, results, and limitations are read.

Table 7: How non-paper materials enter an evidence-first manuscript.

Material	Legitimate use	Required audit fields
Official platform documentation	current eligibility, syntax, policy, or supported feature	owner, surface, locale, version/date, archived URL, exact scope
National or international standard	normative terminology, process, controls, or evaluation requirements	issuer, standard number, status, edition, effective date, applicable clauses
GitHub tutorial or repository	reproducible implementation, data schema, test fixture	commit hash, license, environment, data provenance, test result
Industry blog or community tutorial	hypothesis generation, workflow example, vocabulary, emerging practice	author/organization, conflicts, date, evidence links, claim registry
White paper	disclosed methodology, aggregate pattern, operational framework	sponsor, sample, systems, dates, raw unit, analysis, artifacts, missingness
Video or webinar	observable demonstration and time-indexed speaker claims	stable URL, duration, timestamp, speaker role, transcript, claim type, later verification

2.5 Source classes in a mixed research–practice field

Video claims deserve special care because explanation and demonstration are interleaved. A transcript line is tagged as *observation*, *explanation*, *hypothesis*, *forecast*, or *prescription*. Only the first is directly established by what appears on screen, and even then the environment may be staged. Blogs and community tutorials are treated similarly: they are valuable for discovering workflows and vocabulary, but their linked primary sources carry the factual burden.

2.6 Figures are claims with a visual grammar

A figure can overstate evidence more efficiently than a paragraph. Three rules govern the visual program of these Notes.

1. **Mechanism diagrams mark observability.** Solid borders indicate observed components; dashed borders indicate hypothesized or hidden stages. Platform colors and logos are not used to imply a verified architecture.
2. **Data plots preserve the experiment.** Captions name the unit, system, query set, time, intervention, comparator, uncertainty, and source. Axes do not begin at a value chosen only to dramatize a small effect.
3. **External figures require a license record.** If reuse is not clearly permitted, the manuscript cites the result and replaces the image with an original concept diagram or a reproducible plot from licensed data. A screenshot is not reconstructed data.

The TeX source supplied for arXiv:2506.02070v3 demonstrates a strong pedagogical rhythm of definitions, key ideas, examples, remarks, algorithms, figures, and summaries. Because that source is licensed CC BY-NC-ND 4.0, these Notes use it only to study source architecture and teaching roles. The GEO prose, diagrams, and box system are independently written and designed.

Reproduction Studio 2 · Build a claim registry

Input. One organization’s homepage, product documentation, two third-party sources, one white paper, one tutorial, and three papers.

Task. Extract 25 explicit or implied claims. Normalize each into subject, predicate, object, time, locale, conditions,

and status. Link evidence spans and source identities; include contradictory and expired evidence. Assign an evidence ceiling and write the cheapest ethical action that could close each critical gap.

Deliverable. A machine-readable CES table, a conflict report, and a release decision for every public-facing claim.

Stop rule. Stop when each approved claim has a current evidence path and every unapproved claim has an owner, reason, and next review date. Do not stop when every row merely has a URL.

Chapter summary

1. Source identity and claim support are separate audits.
2. Evidence levels constrain what a source can establish; they do not rank sources by prestige.
3. The Claim–Evidence–Source graph gives content engineering a governed factual substrate.
4. The latest 42-paper export is a versioned research library with explicit anomalies, not a bag of conclusions.
5. Visual design must preserve uncertainty, observability, and licensing.

3 Web Acquisition, Representation, and Query Planning

Key idea

Content does not travel directly from a web page into an answer. It passes through a changing information environment, one or more queries, retrieval and selection operations, a finite context, generation, and attribution. The right intervention depends on the first stage at which the intended evidence is lost.

3.1 A reference pipeline, not a platform blueprint

The pipeline in figure 7 is a research abstraction. It collects stages that recur across search, retrieval-augmented generation, and the GEO literature [1, 23, 58]. A named product may fuse stages, execute them repeatedly, retrieve from licensed collections rather than the open web, or generate without a fresh search. The diagram is useful because each stage exposes a different failure mode and a different class of observable evidence.

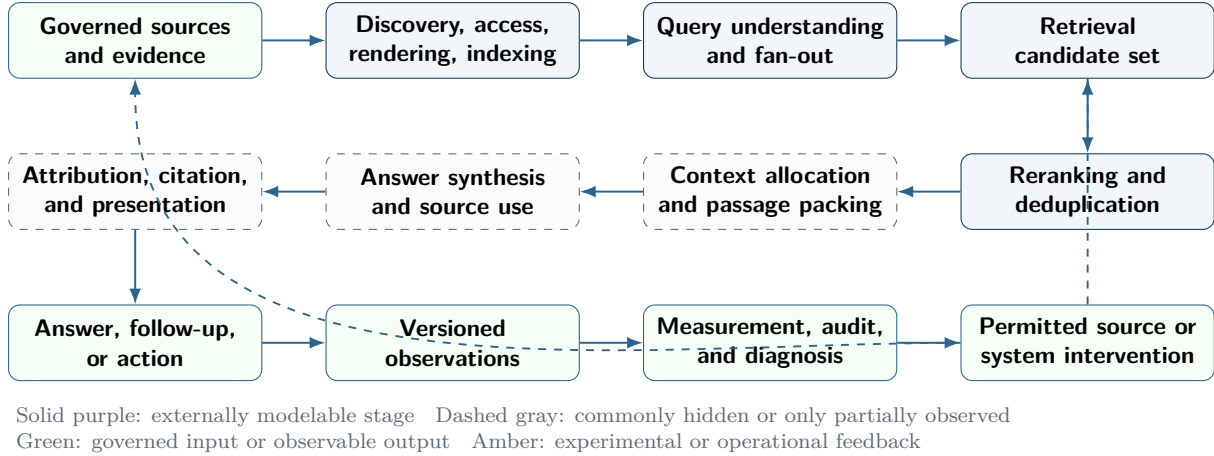


Figure 7: An end-to-end reference pipeline for diagnosis and evaluation. The composition is original. Dashed stages must not be presented as verified internals of a closed platform.

3.2 Stage 0: governed sources and the accessible world

Let $\mathcal{W}_t$ denote the information environment that a system can access at time t . It is not identical to the public web. Access may depend on crawler policy, robots directives, authentication, rendering, licensing, freshness, geography, or a product-specific index. A source s is eligible only if the relevant representation $\pi(s)$ can be obtained and interpreted:

$$\text{Eligible}(s, t) = \text{Discoverable}(s, t) \wedge \text{Fetchable}(s, t) \wedge \text{Renderable}(s, t) \wedge \text{Indexable}(\pi(s), t). \quad (10)$$

This decomposition prevents a common diagnostic error. If a client-side page returns an empty initial document to a non-rendering fetcher, rewriting the visible prose will not fix acquisition. If a canonical rule consolidates the wrong URL, publishing more duplicates can increase ambiguity. Conversely, a successful fetch does not prove indexing, retrieval, or use.

The source layer is also where permissions live. An organization should not simultaneously demand that every

agent refrain from fetching a source and claim that failure to cite that source proves a ranking penalty. Access policy, training policy, search acquisition, and user-triggered retrieval may be different controls and must be audited separately.

3.3 Stage 1: query understanding and fan-out

A natural-language request q may be transformed into a set of retrieval queries

$$\mathcal{Q}(q, u) = \{\hat{q}_1, \dots, \hat{q}_K\}, \quad (11)$$

conditioned on user state u , conversation history, language, locale, and the system’s plan. A procurement question can generate subqueries for requirements, alternatives, reviews, pricing, risks, and recent changes. More fan-out is not automatically better: each query consumes latency and cost and can introduce irrelevant or low-quality evidence.

Because closed products rarely expose the complete set $\mathcal{Q}$, a researcher should maintain two columns: *observed calls* and *reconstructed hypotheses*. Network traces, displayed search terms, or an open agent log can populate the first; an analyst’s plausible query tree belongs in the second. Confusing the two turns an explanatory diagram into fabricated telemetry.

3.3.1 A stopping rule for query expansion

Let E_k be the unique evidence coverage after k subqueries and C_k their cost. One practical stopping condition is

$$\Delta E_{k+1} < \epsilon_E \quad \text{or} \quad \frac{\Delta E_{k+1}}{\Delta C_{k+1}} < \tau, \quad (12)$$

subject to explicit requirements for counterevidence and source diversity. The thresholds are protocol choices. They should not be retrospectively selected to match a preferred answer.

3.3.2 Controlled trajectories and the organic-discovery gap

Two evidence classes answer different questions. In the EcoGEO controlled trajectory study, a synthetic target result was inserted at rank five among nine open-web results, after which a GPT-5.1 search agent could issue and fetch additional pages under a bounded tool budget. Across the three reported task sets, the coordinated TRACE condition produced target recommendation rates of 67.2%, 71.9%, and 73.9% [60]. The design demonstrates that an entry page, linked evidence, and the agent’s subsequent searches can change a controlled browsing trajectory. Because exposure was injected rather than earned, it does not estimate natural crawling, indexing, organic rank, or the probability that a previously unknown commercial page will be discovered.

Live-surface panels offer ecological observation but weaker mechanism access. In one German-language, four-vertical panel, repeated runs in the same time window produced mean source-set Jaccard similarities between 0.321 and 0.434; adjacent-day values were between 0.336 and 0.423 [40]. These figures justify repeated sampling for that panel, but do not isolate decoding randomness from retrieval, routing, caching, or product updates and do not define a universal repeat count.

Boundary and failure conditions

The local evidence base does not identify a causal probability that a newly published, previously unindexed page will be organically discovered by a commercial generative-search surface. Controlled trajectory experiments and live black-box observations bracket that question; neither fills the gap.

3.4 Stage 2: retrieval and candidate construction

For a query $\hat{q}$, a retriever assigns a score to passages or documents $d \in \mathcal{W}_t$:

$$r(d, \hat{q}) = \alpha r_{\text{lex}}(d, \hat{q}) + (1 - \alpha) r_{\text{dense}}(d, \hat{q}) + \gamma r_{\text{prior}}(d, t). \quad (13)$$

The components can represent lexical match, embedding similarity, and allowed priors such as freshness or source quality. The candidate set is $\mathcal{D}_{\hat{q}} = \text{TopK}_d r(d, \hat{q})$. This equation is a generic experimental model, not a reverse-engineered platform score.

Retrieval operates on representations, not author intent. A page may contain the correct fact but lose the relevant passage during segmentation. A visually prominent chart may be inaccessible to a text-only representation. A concise sentence can be retrieved for one query and be too underspecified for another. Therefore, “the fact is on the page” is not an adequate retrieval diagnosis.

Worked example 3.1: The missing version condition

A product page states “The system supports 128 projects” in one passage and “Enterprise limits vary by contract” in a distant footnote. A query about the standard plan retrieves the first passage but not the condition. The page is factually complete at document level and misleading at passage level. The appropriate intervention is to co-locate the number, plan, date, and exception, not to repeat the larger number more often.

3.5 Stage 3: reranking, deduplication, and context allocation

A reranker estimates a task-specific ordering over the candidate pool

$$\rho(d, \hat{q}, \mathcal{D}) = \text{Rerank}(d, \hat{q}, \mathcal{D}; \theta). \quad (14)$$

The score may depend on other candidates, which means a source can change rank without changing when competitors enter or leave the pool. Deduplication can collapse mirrored copies; diversity rules can reserve slots for different domains or viewpoints.

Context assembly then solves a constrained selection problem. Given passage utility v_i , token length l_i , redundancy penalty $g(i, j)$, and budget B , one abstraction is

$$\max_{x_i \in \{0,1\}} \sum_i x_i v_i - \lambda \sum_{i < j} x_i x_j g(i, j), \quad (15)$$

$$\text{subject to } \sum_i x_i l_i \leq B. \quad (16)$$

A high retrieval score does not guarantee context admission. Long passages, duplication, and competition for a limited budget can change the selected evidence. This is one reason end-to-end environments provide information that fixed-candidate experiments cannot [23].

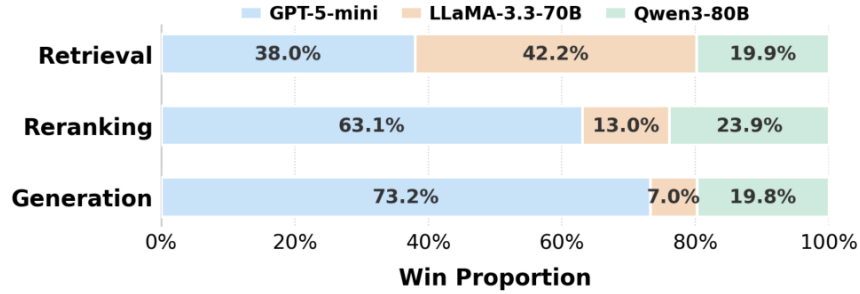


Figure 8: Stage-specific win proportions for three backbones in SAGEO Arena. The ordering changes between retrieval, reranking, and generation, so a single “best model” label would erase the stage being measured.

Source note. Rectangular crop of Kim et al. [23, Fig. 6], physical PDF page 8, CC BY 4.0. A win is highest rank among the three backbones inside that reconstructed pipeline; the panel gives no uncertainty interval and is not a comparison of commercial answer engines.

3.6 Stage 4: generation, attribution, and absorption

Let C denote the effective context and M the generative model. The response is sampled or decoded from

$$y \sim p_M(y \mid q, u, C, \eta), \quad (17)$$

where η captures decoding and system-state variables. The model may compress, combine, omit, qualify, or conflict with the evidence. Attribution may be generated jointly, attached after sentence generation, or resolved from separate passage metadata.

Three evaluations are therefore distinct:

Entailment

Does the cited passage support the answer span?

Completeness

Do material answer claims receive appropriate support?

Absorption

Which source-specific facts, wording, or structure actually contribute to the response?

Citation diagnosis research proposes failure categories across the pipeline and tests repair procedures rather than assuming a generic “quality” defect [47]. The citation-selection/absorption framework likewise shows why counting displayed links and measuring source influence are different tasks [65]. For a production audit, automated entailment scores should be sampled against human judgments because both answer segmentation and model-based judging can introduce error.

3.7 Stage 5: presentation, interaction, and action

Even a faithful answer is not the end of the system. Citation placement, collapsing, label text, device viewport, follow-up prompts, and action buttons alter what a user can inspect. If a study’s outcome is a visit, sign-up, or purchase, the unit becomes a user–task interaction rather than a response. Attributing that outcome to a source edit requires a field design that addresses selection, seasonality, concurrent campaigns, and interference.

This final stage also changes the welfare question. A system can maximize source exposure while increasing user

Table 8: Examples of stage-specific observations and controls.

Stage	Direct or near-direct evidence	Useful control
Acquisition	server logs, fetch output, rendered DOM, canonical and status inspection	allow/disallow and static/rendered positive controls
Retrieval	open-system top- k , exposed search result, agent trace	fixed corpus with known relevant and irrelevant passages
Reranking	open-system scores or ordered candidate trace	order-swapped candidates and relevance perturbations
Context Generation	logged packed passages in a controlled system response, seed when available, model/version, prompt	equal-token and deduplicated context conditions repeated runs and counterfactual passage removal
Attribution	resolved citations and answer spans	supported, unsupported, and deliberately conflicting passages
Action	consented event logs and declared funnel	holdout or randomized exposure where feasible

effort or reducing answer quality. It can produce an accurate answer while withholding enough provenance that correction becomes difficult. Pipeline evaluation must therefore report both source-side and user-side outcomes.

3.8 Observability and diagnostic tests

A negative result is interpretable only when positive controls succeed. If a canary source is not fetched under the allow condition, “disallow had no effect” is not a valid conclusion; the experiment is inconclusive. If a retriever fails to recover a uniquely matching passage, downstream answer differences cannot isolate generation.

Reproduction Studio 3 · Build and audit a minimal end-to-end system

Input. A frozen corpus of 300–1,000 documents, 30 queries with relevance judgments, and a local or openly logged generator.

Task. Compare lexical, dense, and hybrid retrieval; apply one reranker; pack a fixed context budget; generate answers; and resolve citations. Preserve the candidate set, scores, packed passages, response, and attribution for every run. Add three positive controls and three failure injections.

Metrics. Recall@ k , nDCG@ k , evidence recall, answer correctness, citation entailment, attribution completeness, latency, and cost. Do not merge them into one score.

Deliverable. A trace bundle, stage-level error table, and a causal graph showing which comparisons identify which effects.

Stop rule. Stop when every downstream failure can be traced to the earliest observed failing stage or is explicitly labeled unresolved.

Chapter summary

1. The reference pipeline is a diagnostic abstraction, not a claim about a closed product’s internals.
2. Access, retrieval, reranking, context, generation, attribution, and action require different evidence and controls.
3. Document-level truth can be lost at passage and context level.
4. Citation selection, entailment, completeness, and absorption are separate evaluation problems.
5. End-to-end traces are necessary to identify where an intervention helps and where it fails.

4 Retrieval, Fusion, Reranking, and Context Allocation

Key idea

Retrieval quality is not a property of a page in isolation. It is a relation between a query representation, a corpus snapshot, a candidate-construction rule, a ranking objective, and a finite context budget. A source can be retrieved yet displaced, duplicated, truncated, or packed without the passage that supports the answer.

4.1 Questions this chapter answers

By the end of the chapter, a reader should be able to answer five questions without treating retrieval as an invisible prelude to generation:

1. Which corpus, representation, and retrieval unit define the population to which a top- k result refers?
2. How can lexical and dense lists be fused without pretending that their raw scores share a scale?
3. Which query, rank, duplicate, and provenance paths must survive fusion so that repeated copies are not mistaken for independent evidence?
4. What does a reranking improvement identify, and what remains unknown about answer use and attribution?
5. How should a finite context budget trade off relevance, material-claim coverage, redundancy, and source dependence?

4.2 Declare the retrieval unit and corpus state

Let $\mathcal{D}_t$ be a frozen corpus snapshot and let $\Pi(d) = \{p_{d1}, \dots, p_{dn_d}\}$ be the passages produced from document d . The retrieval unit may be a document, passage, image-text pair, collection page, or another representation. These units are not interchangeable: a document-level relevant judgment does not prove that the supporting passage is recoverable after rendering and segmentation.

For each run, the trace should record at least

$$\mathcal{T}_{\text{ret}} = \langle q, \hat{q}, \mathcal{D}_t, \Pi, f, k, \theta, \mathbf{r}, t \rangle, \quad (18)$$

where q is the user request, $\hat{q}$ the retrieval query, f the retriever, k the requested depth, θ its versioned configuration, and $\mathbf{r}$ the ordered results with scores. Without the corpus hash and segmentation version, two runs with the same model name need not be comparable.

Definition 4.1: Evidence recall

For a query with a preregistered set of material evidence units E_q , evidence recall at depth k is

$$\text{ER @}k(q) = \frac{|E_q \cap E(R_k(q))|}{|E_q|}, \quad (19)$$

where $E(R_k(q))$ is the set of evidence units recoverable from the top- k representations. This is stricter than document recall when one document contains several independent claims or qualifications.

Table 9: An observability contract for a retrieval study. “Observed” means recorded by the stated protocol, not universally exposed by a platform.

Class	Examples in a controlled system	Required treatment
Observable	query and rewrite; corpus and passage hashes; returned identifiers, ranks, scores, packed positions, and truncation spans	preserve raw values and timestamps before computing aggregate metrics
Latent	unlogged query expansion; complete index eligibility; hidden candidate pool; proprietary score; unobserved source contribution	name the estimand or proxy and do not fill missing stages from the visible answer
Assumed	relevance-label validity; segmentation adequacy; duplicate threshold; stability during a run; sufficiency of the chosen candidate depth	preregister, stress-test, and report sensitivity where the conclusion depends on it

4.3 Separate observables, latent variables, and assumptions

The boundary between observed and latent quantities is system-relative. In a controlled reconstruction, candidate scores and packed token spans can be logged directly; on a closed platform, the same quantities may be hidden. A retrieval report should therefore classify each field before interpreting it.

Three distinctions follow. A returned rank is an observation only for the logged retriever. Passage relevance is a judgment attached to a declared task, not an intrinsic property of the passage. Candidate absence may mean failure of acquisition, representation, matching, or depth; without stage logs it does not identify which mechanism failed.

4.4 Use executable baselines before learned fusion

A lexical baseline such as BM25 tests whether relevant wording survives the representation. A dense retriever tests semantic proximity under a locked encoder. Their raw scores usually have different scales, so adding them without normalization, calibration, or a learned fusion rule is not a neutral baseline. Rank-based fusion provides an inspectable alternative. For ranked lists $L_1, \dots, L_J$, reciprocal-rank fusion assigns

$$s_{\text{RRF}}(d) = \sum_{j=1}^J \frac{\mathbb{1}\{d \in L_j\}}{\kappa + \text{rank}_{L_j}(d)}. \quad (20)$$

The constant κ is part of the protocol, not a universal optimum. A learned fusion model requires a training split, loss, feature specification, and leakage audit.

Worked example 4.2: Why score addition can reverse a result

Suppose passage A receives lexical/dense scores (12, 0.62) and passage B receives (10, 0.81). Direct addition favors A because the lexical scale dominates. Min-max normalizing within this two-item pool makes the components (1, 0) and (0, 1), producing a tie. Neither outcome is intrinsically correct; the example shows that “hybrid score” is undefined until the scale and fusion rule are declared. Relevance judgments, not arithmetic convenience, select the rule.

Counterexample · topical relevance can conceal evidence loss

Consider two synthetic top-three lists evaluated with a binary “about the product” label. List A contains a product page, a distributor mirror, and a superseded launch manual. List B contains the current product page, an independent calibration passage, and the current safety notice. All six items can receive the same topical label, so both lists obtain Precision @3 = 1. Yet List A covers only the nominal-range evidence unit, giving ER @3 = 1/3, whereas List B covers the nominal range, validated range, and cold-weather condition, giving ER @3 = 3/3. The numbers are didactic. They demonstrate that a ranker can saturate a document-level metric while losing the qualifications needed for a decision.

4.5 Fuse queries without manufacturing consensus

Query planning can produce several lists for different subquestions. Candidate fusion must preserve which query recovered which passage. Otherwise, a passage returned repeatedly by near-duplicate queries may appear to have independent support. A useful candidate record is

$$c_i = \langle d_i, p_i, \mathcal{Q}_i, \mathbf{r}_i, h_i, o_i \rangle, \quad (21)$$

where $\mathcal{Q}_i$ is the set of recovering queries, $\mathbf{r}_i$ their ranks, h_i a content hash, and o_i a provenance origin. Exact duplicates, near duplicates, syndicated copies, and independently produced sources require different labels.

The current GEO literature includes content-centric and end-to-end testbeds that make source influence or retrieval stages observable [6, 23]. Those environments support claims about their specified corpora, systems, and metrics. They do not disclose the candidate fusion policy of every commercial answer surface.

Algorithm 2 Trace-preserving hybrid retrieval and evidence-aware packing

Inputs: query q ; corpus $\mathcal{D}_t$ and segmentation Π ; lexical and dense retrievers; depths k_j ; fusion constant κ
locked reranker g ; material-evidence labels E_q ; token budget B ; origin caps b_o

Outputs: ordered packed passages S and replayable trace $\mathcal{T}$

- 1: Verify and record hashes for $\mathcal{D}_t$, Π , model checkpoints, and configuration files
 - 2: Run each retriever on the recorded query representation and retain its complete ordered top- k_j list
 - 3: Form the candidate union C ; attach every recovering query, list, rank, raw score, passage hash, and origin to each candidate
 - 4: **for all** $p \in C$ **do**
 - 5: $s_{\text{RRF}}(p) \leftarrow \sum_j \mathcal{K}\{p \in L_j\} / (\kappa + \text{rank}_{L_j}(p))$
 - 6: **end for**
 - 7: Cluster exact, near-duplicate, and syndicated passages under a locked rule; preserve the cluster relation in $\mathcal{T}$
 - 8: Apply any representative-selection rule declared before evaluation; otherwise retain all cluster members
 - 9: Rerank the surviving candidates with g ; log input order, truncation, scores, ties, and output order
 - 10: Instantiate token costs ℓ_i , claim coverage e_{ic} , redundancy u_{ij} , and origin labels o_i
 - 11: Select S by solving the declared packing objective under B and b_o ; record rejected candidates and binding constraints
 - 12: Order S , serialize the exact token spans, and return $(S, \mathcal{T})$
-

Algorithm 2 is a protocol template, not a claim that a commercial system implements these steps. Its purpose is to make each transformation replayable in a controlled system and each unobserved transformation explicit elsewhere.

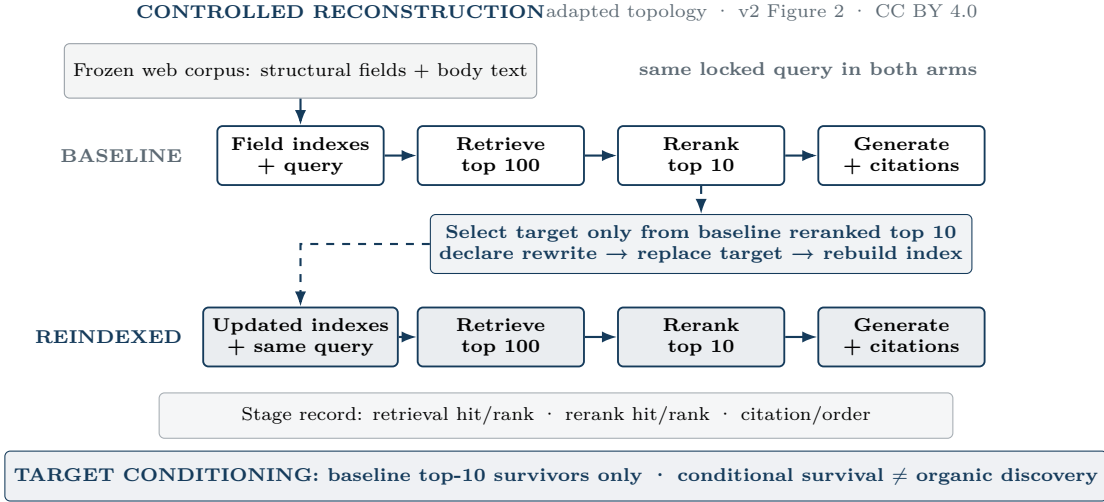


Figure 9: A controlled reconstruction adapted from Kim et al. [23, Fig. 2], under CC BY 4.0. The redraw compresses labels and removes vendor iconography while exposing the baseline top-ten target gate. It depicts conditional survival in the paper’s locked pipeline, not organic discovery or a disclosed production architecture.

4.6 Read an effect at the stage where it was measured

SAGEO Arena makes the survival funnel visible in a reconstructed system: 171,003 unique pages and 2,700 queries feed field-level BM25 and reciprocal-rank fusion, a top-100 candidate set, a Qwen3 reranker that retains ten items, and a GPT-5-mini generation stage [23]. Its stage results are not monotone. Relative to a baseline retrieval hit rate of 0.58, rerank hit rate of 1.00, and citation rate of 0.50, body-only rewrites yielded 0.53, 0.84, and 0.47; structural-only rewrites yielded 0.71, 0.83, and 0.52. The baseline rerank value of 1.00 follows from selecting targets that already survived the baseline top ten. These figures therefore describe conditional survival in the published reconstruction, not the fate of arbitrary web pages or a commercial ranking rule.

The contrast with fixed-candidate experiments is instructive. E-GEO holds a ten-product candidate list fixed and estimates how description rewrites change prompted LLM reranking; its source collection contains 13,747 query–list instances and its reported test set contains 2,000 [3]. That design cleanly studies candidate-conditional rank change, but cannot show that an edited item would be recovered from the underlying product corpus. Likewise, fixed conflicting-evidence studies show that relevance, composition, and order can alter generation after retrieval [48, 50]; they do not estimate retrieval effects.

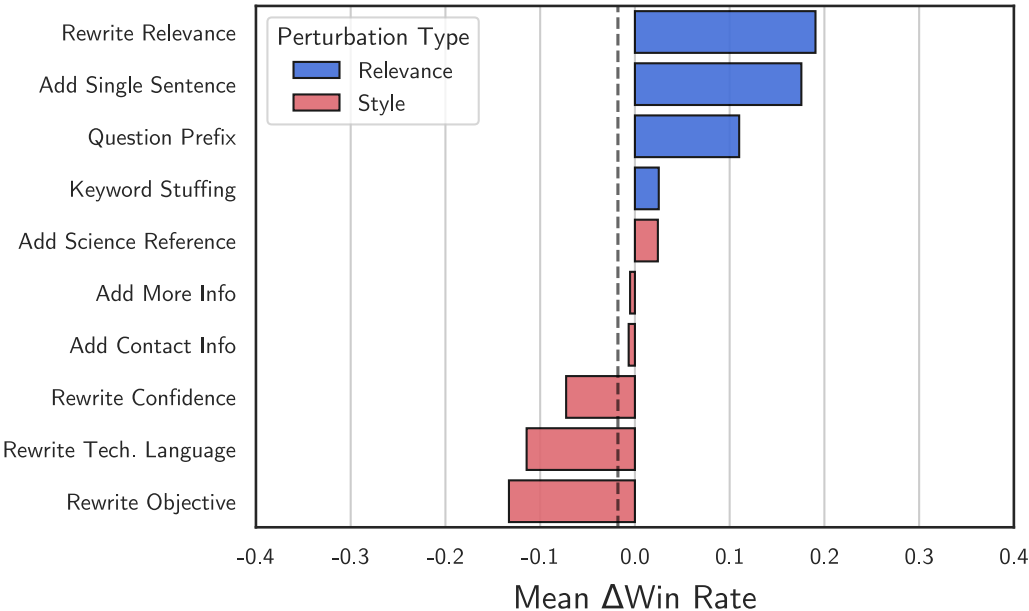


Figure 10: Mean win-rate change after editing one passage in a fixed conflicting-context experiment; the dashed line is a separate control, not the zero baseline. Reproduced from Wan et al. [50, Fig. 2], ACL 2024, under CC BY 4.0; vector crop and font-outline preflight only. The result is a model preference in supplied context, not retrieval, citation, truth, or human credibility.

Interpretation boundary · relevance sensitivity is not evidence quality

The experiment places two conflicting evidence passages in the model context, edits only the passage supporting “Yes,” and asks for a binary answer. Most counterfactual rewrites were generated with `claude-v1-instant`; the separate control is not independently proven neutral. Its win-rate response therefore identifies model sensitivity within that constructed comparison. A positive bar cannot be reported as a retrieval gain, a citation gain, a truth signal, or a human credibility judgment. The figure is useful precisely because it separates a model preference from the quality of the evidence being preferred. The source figure reports no uncertainty intervals or cell sample sizes, and filtered evaluation sizes differ by model.

Table 10: Three evidence environments and the estimands they can support.

Environment	Identifiable comparison	Claim ceiling
Reconstructed funnel	survival through retrieval, reranking, and generation in the locked pipeline	no universal platform mechanism or organic-discovery rate
Fixed candidate list	rank or citation choice conditional on supplied candidates	no corpus recall or acquisition claim
Fixed context	generation, order, conflict, and ablation under supplied evidence	no retrieval qualification

4.7 Rerank for the declared task

Table 11: A fully specified synthetic hybrid-retrieval trace. The RRF column is recomputable from the two ranks with $\kappa = 10$; v_i is a locked, task-specific reranker value used by the packing example.

ID	Controlled passage content	ℓ_i	Lexical rank	Dense rank	RRF	v_i
p_1	Product page: nominal range 0–50 mg/L	70	1	2	0.1742	0.80
p_2	Calibration report: validation limited to 5–35 mg/L at 10–25 °C	100	4	1	0.1623	1.00
p_3	Current safety notice: warmed enclosure required below 5 °C	60	3	3	0.1538	0.95
p_4	Superseded launch manual	80	2	5	0.1500	0.15
p_5	Syndicated near-duplicate of p_1	70	5	4	0.1381	—
p_6	Unrelated maintenance passage	50	6	6	0.1250	0.00

4.7 Rerank for the declared task

Reranking estimates an ordering conditional on a candidate pool. Pointwise, pairwise, and listwise objectives answer different questions; a cross-encoder score is not automatically calibrated across queries. At minimum, report the candidate depth, model and checkpoint, input truncation, ordering loss, tie policy, latency, and the difference between relevance labels and answer-level utility.

Let $g_{qi} \in \{0, 1, 2, 3\}$ be a graded relevance judgment for candidate i . The discounted cumulative gain at depth k is

$$\text{DCG}@k(q) = \sum_{i=1}^k \frac{2^{g_{qi}} - 1}{\log_2(i + 1)}, \quad (22)$$

and $\text{nDCG}@k$ divides by the ideal ordering. The metric rewards ranking quality under the chosen labels; it does not measure whether the final answer uses the passage faithfully.

4.8 Allocate context as a constrained evidence problem

Context allocation chooses passages, positions, and sometimes excerpts. For passage i , let v_i denote query relevance, e_{ic} coverage of material claim c , ℓ_i token cost, u_{ij} redundancy, and $z_i \in \{0, 1\}$ its selection indicator. One auditable formulation is

$$\max_{\mathbf{z}} \quad \sum_i z_i v_i + \alpha \sum_c \min\left(1, \sum_i z_i e_{ic}\right) - \beta \sum_{i < j} z_i z_j u_{ij}, \quad (23)$$

$$\text{subject to} \quad \sum_i z_i \ell_i \leq B, \quad \sum_{i: o_i = o} z_i \leq b_o. \quad (24)$$

The optional origin cap b_o limits domination by one dependency cluster. It is a design choice that can improve diversity while excluding the best source; its effect must be reported rather than presumed beneficial.

4.9 Worked running case: from two ranks to one context

Return to the synthetic Orchid Bay query introduced in [section 1.3](#). The requested answer must distinguish the nominal measurement range from the independently validated range and must retain the cold-weather qualification. Suppose a frozen six-passage corpus produces the two complete rankings in [table 11](#). Token costs are counts under one locked tokenizer. The values and passages in this example are didactic inputs, not observations about a real retriever, organization, or platform.

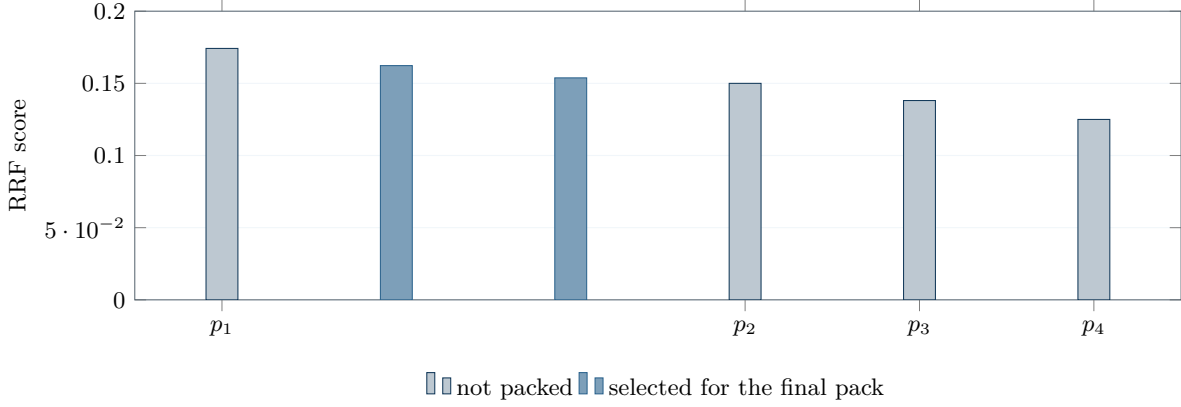


Figure 11: Rank-fusion scores and the later packing decision in the synthetic case. The highest RRF candidate is not selected under the declared budget and material-evidence objective. Every bar is recomputable from table 11; color represents the later decision, not an RRF threshold.

Table 12: Calculation for the non-dominated feasible two-passage packs in the synthetic example. Pairs containing zero-value p_6 are dominated. Utility is $\sum_i v_i + 0.8$ times the number of covered material claims; the budget is 180 tokens.

Pack	Tokens	$\sum_i v_i$	Material claims covered	Utility
$\{p_1, p_2\}$	170	1.80	c_N, c_V	3.40
$\{p_1, p_3\}$	130	1.75	c_N, c_T	3.35
$\{p_1, p_4\}$	150	0.95	c_N	1.75
$\{p_2, p_3\}$	160	1.95	c_V, c_T	3.55
$\{p_2, p_4\}$	180	1.15	c_V	1.95
$\{p_3, p_4\}$	140	1.10	c_T	1.90

For example,

$$s_{\text{RRF}}(p_1) = \frac{1}{10+1} + \frac{1}{10+2} = \frac{23}{132} \approx 0.1742,$$

$$s_{\text{RRF}}(p_2) = \frac{1}{10+4} + \frac{1}{10+1} = \frac{25}{154} \approx 0.1623.$$

Computing the remaining four scores yields the fused order $p_1, p_2, p_3, p_4, p_5, p_6$. This order is rank-based: it does not assert that p_1 supplies the most decision-relevant evidence. Under the locked duplicate rule, p_5 is recorded as a syndicated member of the p_1 cluster and removed before reranking; the relation is not erased from the trace. The task-specific reranker values then order the surviving relevant candidates as p_2, p_3, p_1, p_4, p_6 .

Now let the three material claims be nominal range c_N , independently validated range c_V , and cold-weather condition c_T . Set $e_{1N} = e_{2V} = e_{3T} = 1$, all other $e_{ic} = 0$, $\alpha = 0.8$, and $B = 180$. After duplicate removal, no retained pair is redundant, so the redundancy term is zero for this calculation. No origin cap binds. The non-dominated feasible multi-passage packs are enumerated in table 12.

Every three-passage set with positive marginal value exceeds the budget. The only feasible triple, $\{p_1, p_3, p_6\}$, has the same utility as $\{p_1, p_3\}$ because p_6 adds no claim coverage or relevance value; the best singleton has utility 1.80. Therefore the optimizer selects $S = \{p_2, p_3\}$, consumes 160 tokens, and leaves 20 tokens unused. The pack excludes the highest-RRF passage because the independent validation limit and cold-weather condition cover the two decision-critical qualifications. The trace must record that this conclusion depends on the registered claims,

Table 13: Stage-specific retrieval diagnostics and their evidentiary limits.

Observation	Supports	Does not by itself support
Top- k trace	recovery and ordering in the frozen system	source use in an answer
Reranker gain	ordering improvement under stated labels	user utility or attribution fidelity
Context log	admission, order, and truncation	causal influence without an ablation
Displayed citation	presentation of a resolved source object	retrieval rank, claim entailment, or absorption
Leave-one-out change	conditional influence in that context	unique authorship or global platform behavior

relevance values, and budget: changing any of them defines a different allocation problem. A generator may still fail to use either passage faithfully, so pack membership is not counted as answer contribution.

4.10 Estimate contribution without confusing it with ownership

A leave-one-source-out comparison measures how an answer changes when a source is removed from a controlled context. Shapley-style attribution averages marginal contribution over coalitions, but its interpretation depends on the utility function and on what counts as a player. Work on fair context attribution makes those assumptions explicit [35]; work on attention and exposure examines related mechanisms in bounded settings [2]. Neither an attention weight nor a marginal answer change proves intellectual ownership, legal authorship, or displayed citation.

Boundary and failure conditions

Open-system retrieval traces can identify failure in the reconstructed system. A commercial result page, citation panel, or network observation exposes only the corresponding surface and time. Do not infer a hidden candidate pool, score, reranker, or context order unless the platform exposes it or a controlled intervention identifies it.

4.11 Reproduction checkpoint

A screenshot of a result page is not a replayable retrieval record. Before a stage comparison is accepted, an independent reader should be able to recover the artifacts in [table 14](#) and obtain the same candidate identities, fusion scores, and packed token spans, allowing only documented nondeterminism.

Reproduction Studio 4 · Trace retrieval loss into context loss

Question. Which evidence disappears during representation, retrieval, fusion, reranking, and packing?

Input. A frozen, licensed corpus; 30 clustered queries; passage-level relevance and material-evidence labels; lexical and dense baselines; one locked reranker; and three context budgets.

Procedure. Preserve corpus and passage hashes. Compare single-query and fan-out retrieval, RRF and one learned fusion, reranking, deduplication, and evidence-aware packing. Inject one missing qualification, one syndicated duplicate cluster, and one long but highly relevant passage.

Output. Recall@ k , nDCG@ k , evidence recall, origin diversity, context coverage, latency, and a stage-transition table with uncertainty intervals matched to the sampling unit. Release the artifact bundle specified in [table 14](#).

Reference criterion. Every reported loss must resolve to a versioned document, passage, query list, candidate rank, and packing decision.

Independent check. A second implementation must reproduce document and passage identities, RRF scores,

Table 14: Minimum checkpoint for reproducing a retrieval-to-context result.

Artifact	Required contents	Pass condition
Corpus and representation	fetch timestamp, license boundary, document and passage hashes, renderer, segmenter, tokenizer	a clean rebuild produces the same identifiers or a documented migration map
Query and candidate trace	user query, every rewrite, retriever configuration, depth, raw ordered lists, scores, and recovering paths	recomputed RRF scores and tie decisions match the released trace
Judgment package	task definition, material evidence units, relevance rubric, assessor labels, disagreements, and adjudication	another assessor can apply the rubric without access to the reported outcome
Rerank and packing ledger	model checkpoint, truncation, candidate order, token costs, coverage and redundancy values, constraints, selected spans, and rejections	replay yields the same ordered context and identifies every binding constraint
Failure and sensitivity log	missing fetches, duplicate clusters, nondeterministic runs, alternative k , κ , budget, and label choices	conclusions are restated at the narrowest setting under which they remain supported

duplicate decisions, and final packed token spans before model-quality comparisons are interpreted.

Stop rule. Stop and mark the comparison unidentified if the corpus, segmentation, relevance labels, candidate path, or packed context cannot be reconstructed. Record a failed checkpoint as a result; do not silently replace the missing stage with a displayed-citation proxy.

Chapter summary

1. Retrieval is conditional on a representation, corpus, query, and versioned scoring rule; every report should separate observations, latent stages, and identifying assumptions.
2. Lexical and dense scores require an explicit fusion method; raw addition is not scale invariant, while rank fusion remains conditional on its lists, depths, tie rules, and κ .
3. Candidate fusion preserves query and provenance paths so duplication is not mistaken for independent support.
4. Reranking metrics evaluate the declared labels and candidate pool, not final answer use. High topical relevance can coexist with low material-evidence recall.
5. Context packing is a constrained evidence-allocation problem whose budget, claim weights, diversity, redundancy, and rejection decisions must remain visible. Highest fused rank need not imply admission.
6. Contribution measures are conditional system measurements, not proofs of ownership or platform-wide behavior.
7. A reproducible result resolves every aggregate loss to versioned documents, passages, candidate paths, judgments, and exact packed token spans.

5 Generation, Citation, Attribution, and Source Influence

Key idea

Context inclusion, answer use, displayed citation, and faithful attribution are four different events. Evaluation must segment answer claims, resolve citation objects, test support, and estimate source influence without treating a visible link as proof that the linked source caused or entails the text.

5.1 Questions this chapter answers

By the end of the chapter, a reader should be able to answer five questions:

1. Which context, source, claim, citation, and answer objects must be versioned before a support score is interpretable?
2. Which parts of generation and citation placement are observed, which are reconstructed, and which remain latent on a closed surface?
3. Why do “has a citation,” citation entailment, citation completeness, claim absorption, and counterfactual source influence answer different questions?
4. How can an automated support judge be calibrated without promoting its labels to ground truth outside the calibration strata?
5. Which ledger and context intervention would let a second analyst reproduce the conclusion without access to an imagined hidden prompt?

5.2 Model the effective context, not an imagined prompt

In a controlled system, the response can be represented as

$$y \sim p_{M,v}(y \mid q, h, C, \sigma, \eta), \quad (25)$$

where M, v identify the model and version, h is conversation state, C is the logged effective context, σ contains system and tool instructions, and η covers decoding state. For a closed surface, some terms are latent. The researcher records observed inputs and outputs and marks reconstructed components as hypotheses.

Parametric knowledge creates a bypass: an answer claim can appear even when no supporting source is present in C . Conversely, a source can influence phrasing or answer structure without receiving a displayed citation. This is why an end-to-end pipeline cannot be reduced to citation counting.

5.2.1 Declare observability before interpretation

The same field can be observed in a controlled reconstruction and latent on a commercial surface. The report therefore classifies fields relative to the study rather than attaching “observable” to a concept forever.

5.3 Segment answer claims before scoring citations

Let $A(y) = \{a_1, \dots, a_J\}$ be material answer claims. A claim is the smallest unit that can be assessed for support while retaining quantities, conditions, scope, time, and attribution. Splitting “Plan A supports 128 projects, except

Table 15: Minimum observability contract for a generation–attribution study.

Class	Examples	Required treatment
Directly observed	submitted query; displayed answer and anchors; resolved URLs; timestamps; logged context and decoding configuration in a controlled system	preserve raw objects before normalization and attach every derived label to their immutable identifiers
Reconstructed	canonical source identity; claim spans; passage attachment; likely context membership inferred from an authorized trace	state the resolver, annotator, error state, and alternative reconstruction
Latent	proprietary effective context; parametric contribution; hidden instruction hierarchy; causal source use; citation-placement policy	do not fill the field from a visible URL; estimate only under an explicit intervention or leave it unknown
Assumed	materiality weights; adequacy of claim segmentation; stable judge error; no context interference after removing a source	preregister, test on counterexamples, and show how the decision changes when the assumption fails

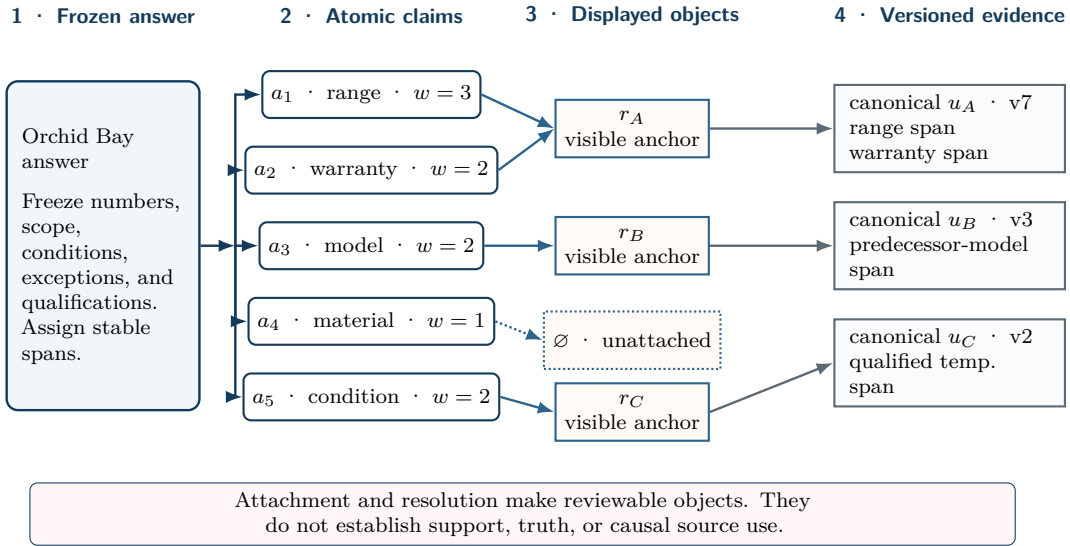


Figure 12: From a frozen answer to reviewable evidence objects. Segmentation preserves quantities, scope, conditions, and qualifications; citation attachment records visible or declared scope; resolution preserves redirects, canonical identity, source version, and the exact evidence span. None of these transformations alone establishes support, truth, or causal source use. The objects and labels are an original synthetic teaching construction.

under a legacy contract” into two unrelated fragments can falsely label the number supported while dropping the exception.

Let $R(y) = \{r_1, \dots, r_K\}$ be resolved citation objects. Resolution stores the displayed label, target URL, redirects, canonical URL, access time, quoted or linked passage when available, and provenance origin. The answer–citation graph has an edge $a_j \rightarrow r_k$ only when the interface attaches the citation to the claim or a documented resolver makes that relation.

Figure 12 keeps segmentation, visible attachment, URL resolution, and evidentiary adjudication as separate operations.

Definition 5.1: Citation entailment and completeness

For material claim a_j and resolved source passage r_k , let $S_{jk} \in \{0, 1\}$ denote human-adjudicated support under a published labeling guide. Then

$$\text{Entailment}(y) = \frac{\sum_{(j,k) \in E_y} w_j S_{jk}}{\sum_{(j,k) \in E_y} w_j}, \quad (26)$$

$$\text{Completeness}(y) = \frac{\sum_j w_j \mathbb{1}\{\exists k : (j, k) \in E_y, S_{jk} = 1\}}{\sum_j w_j}. \quad (27)$$

Entailment asks whether attached citations support their claims; completeness asks whether material claims have at least one supporting citation. Both depend on claim weights, segmentation, resolution, and adjudication policy.

Proposition 5.2: A response-level citation flag does not identify completeness

Let $H(y) = \mathbb{1}\{|R(y)| > 0\}$ indicate that response y displays at least one citation. For any $\epsilon > 0$, there exist two responses y_1, y_2 with $H(y_1) = H(y_2) = 1$ such that citation completeness is one for y_1 and below ϵ for y_2 .

Proof. Give y_1 one material claim of weight one and attach a supporting passage. Give y_2 that same supported claim plus n uncited material claims of weight one. Both responses display a citation, whereas their completeness values are 1 and $1/(n+1)$. Choosing $n > 1/\epsilon - 1$ proves the claim. $\square$

5.4 Keep selection, use, absorption, and attribution separate

For source s , define four binary or graded variables for a response:

Context selection $Z_{\text{ctx}}(s)$

a representation of s is in the logged effective context;

Source use $Z_{\text{use}}(s)$

a controlled ablation or another stated method attributes a material answer change to s ;

Claim absorption $Z_{\text{abs}}(s)$

one or more preregistered, source-specific verified claims appear faithfully in the answer;

Displayed attribution $Z_{\text{cite}}(s)$

the interface presents a citation resolved to s .

None logically entails all the others. A source can be selected but unused, absorbed but uncited, cited next to an unsupported claim, or cited because its URL represents a cluster of duplicated origins.

The distinction between citation selection and citation absorption is central to recent measurement work [65]. Citation-failure research also motivates stage-specific diagnosis and repair rather than a generic rewrite loop [47]. The exact taxonomies and results remain bounded to the locked paper versions and evaluated systems.

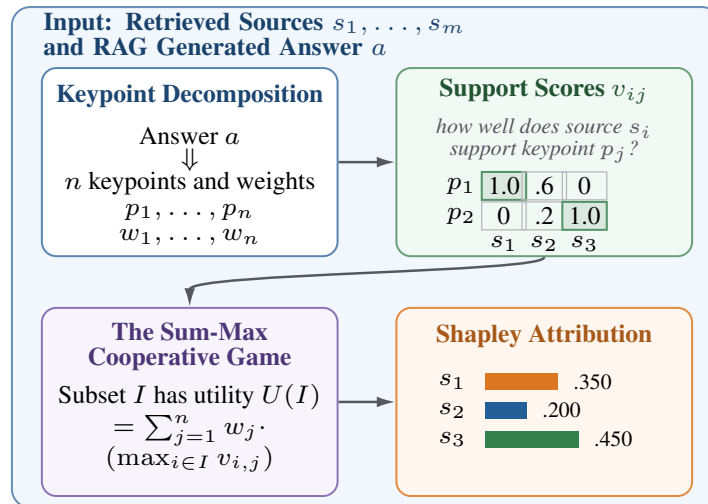
5.4.1 Three estimands that must not be merged

Recent studies make the separation concrete. A two-source experiment can estimate which supplied source receives the first displayed URL; a platform snapshot can describe citation breadth and a constructed textual-overlap proxy; a coalition method can allocate a declared answer-coverage utility. Calling all three *source influence* would erase their different interventions, denominators, and assumptions.

The two-source study reports that 3.4% of runs lacked a matching first URL and were excluded from its binary

Table 16: Citation choice, absorption proxies, and utility attribution answer different questions.

Evidence design	Directly measured object	Required boundary
Two-source citation choice [49]	first displayed URL in 252,000 trials across six models, with source order counterbalanced	exactly two supplied sources; no live retrieval; unstable large odds ratios are not portable effects
Platform snapshot [65]	citation breadth and an engineered score over fetchable cited pages for 602 prompts	descriptive, fetch-conditioned proxy; not hidden attention or causal dependence
Sum-max allocation [35]	exact Shapley credit for the declared best-source-per-answer-point utility	exact for that utility and player set; not authorship, training contribution, or recovered model use

**Figure 13:** A keypoint-support pipeline followed by Shapley allocation of a declared max-coverage utility. Reproduced from Patel et al. [35, Fig. 2], under CC BY 4.0; crop only. The allocation is exact for that utility and player set, not hidden attention, causal source use, authorship, training contribution, or universal fairness.

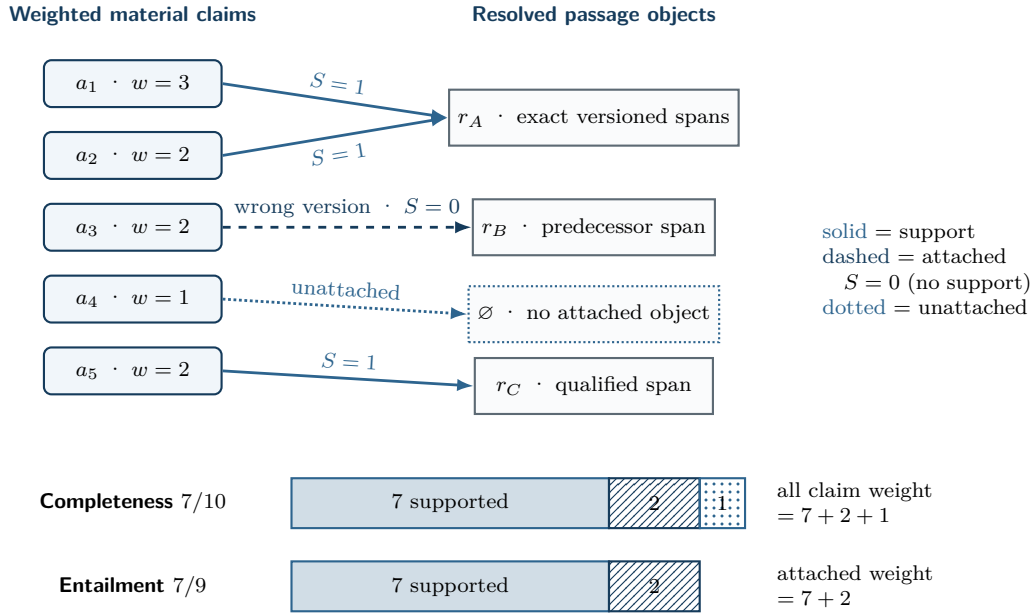
model; it also reports quasi-separation or singular fits for some factors. Directional evidence about order, topical match, specifications, and timestamp can be useful, while extremely large odds ratios should not be treated as precise, cross-system effect sizes. The platform snapshot likewise conditions its page-level proxy on successful fetches. Both exclusions belong in the estimand, not in a footnote.

Worked example 5.3: A visible citation with incomplete support

An answer says: “System K retains logs for 30 days and is certified under Standard X.” The attached page supports the retention period but says only that the organization uses a control framework related to Standard X. The citation object is real and one clause is supported; the certification claim is not. Response-level “has citation” therefore equals one while claim-weighted completeness is below one and the unsupported high-impact claim requires correction.

Table 17: Frozen claim–citation ledger for the synthetic Orchid Bay answer.

Claim	Weight	Attached object	Final support	Adjudicated reason
a_1	3	r_A	support	exact operating-range statement and version
a_2	2	r_A	support	stated warranty period with matching region
a_3	2	r_B	none	source discusses a predecessor, not this model
a_4	1	none	unresolved	material compatibility statement has no attachment
a_5	2	r_C	support	qualification and temperature condition are retained

**Figure 14:** Displayed attachment and evidentiary support in the synthetic Orchid Bay ledger. Solid edges are attached claim–passage pairs adjudicated as support; the dashed edge is attached but does not support the claim; the dotted claim has no attached object. The denominator ribbons show why completeness is 7/10 while entailment is 7/9. The graph is a teaching reconstruction, not a platform observation.

5.5 Worked trace: one answer, five claims, three citations

The following example is synthetic and didactic. It describes no production system. A frozen answer about the fictional Orchid Bay controller contains five material claims with preregistered integer weights summing to ten. Three resolved source passages are available. Two annotators independently segment and label the answer, then adjudicate the one disagreement shown below.

Four of five claims have a displayed attachment, so unweighted citation incidence is $4/5 = 0.80$. The attached-edge weight is $3 + 2 + 2 + 2 = 9$, of which $3 + 2 + 2 = 7$ is supported. Therefore

$$\widehat{\text{Entailment}} = \frac{7}{9} = 0.778, \quad \widehat{\text{Completeness}} = \frac{7}{10} = 0.700. \quad (28)$$

The missing weight is not silently coded as support or as an attached error: two units are attached to the wrong product version and one unit is uncited. If the report kept only “80% of claims are cited,” it would hide both failures.

This ledger measures displayed attachment and support, not source influence. A paired leave-one-source-out

construction later in the chapter tests a different estimand while exposing the collateral changes that can invalidate it.

5.6 Calibrate automated judges

Model judges can assist with claim segmentation, entailment, and contradiction, but their labels are measurements with error. A calibration set should be sampled across claim type, source length, citation position, answer model, language, and expected difficulty. Report the confusion matrix, class prevalence, inter-rater agreement, adjudication rule, and sensitivity of final estimates to plausible judge error.

For binary support labels, let sensitivity be Se , specificity Sp , and observed positive rate $\tilde{p}$. Under stable error rates and $\text{Se} + \text{Sp} > 1$, a corrected prevalence is

$$\hat{p} = \frac{\tilde{p} + \text{Sp} - 1}{\text{Se} + \text{Sp} - 1}. \quad (29)$$

This correction is not valid when error rates shift across models, languages, or claim strata. The uncorrected labels, calibration sample, and stratum-level diagnostics remain part of the release artifact.

5.7 Test influence with counterfactual contexts

In a controlled environment, compare the full context C with a context C^{-s} that removes source s while holding order, token budget, and other passages as stable as possible. For a declared answer utility U , define

$$\Delta_s(C) = \mathbb{E}[U(Y \mid C) - U(Y \mid C^{-s})]. \quad (30)$$

The contrast measures contribution under that constructed context. It can be confounded by replacement passages, position shifts, decoding variance, and semantic redundancy. Coalition methods can distribute contribution across sources, but they inherit the chosen utility and player definition.

For a reproducible synthetic fixture, define $U_i = 1$ only when all three preregistered claims distinctive to r_A appear faithfully in paired run i . The full and ablated contexts use the same query, instruction hash, model configuration, surviving passage bytes, and paired seed. They also record every token-count, absolute-position, replacement-policy, and truncation difference; these collateral changes are not assumed away.

Table 18: Ten paired outcomes in the synthetic Orchid Bay source-ablation fixture. A one means that all three declared r_A -distinctive claims appear faithfully; the values are invented for arithmetic and protocol inspection.

Paired seed	01	02	03	04	05	06	07	08	09	10	Mean
Full C	1	1	1	1	1	1	1	1	0	0	.80
Without r_A , C^{-A}	0	0	0	1	0	1	0	1	0	0	.30
Paired loss	1	1	1	0	1	0	1	0	0	0	.50

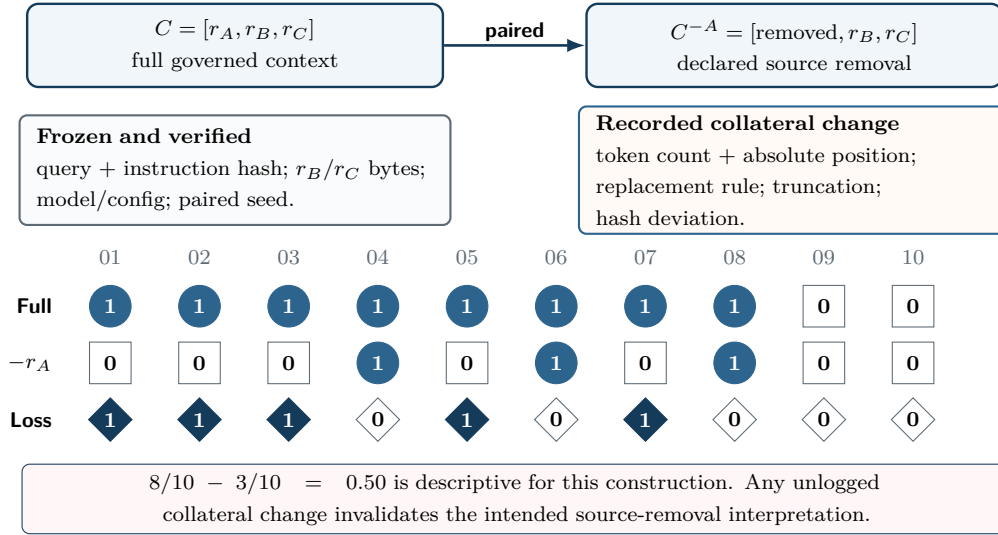


Figure 15: A paired leave-one-source-out construction for the synthetic Orchid Bay context. Query, configuration, surviving passages, and seed are frozen; every collateral token, position, replacement, or truncation change is recorded rather than assumed away. Five of ten paired runs lose the declared source-distinctive utility, giving a descriptive contrast of 0.50 for this constructed context only.

Research on citation vulnerabilities and caption-based interventions shows why adversarial and multimodal conditions belong in this evaluation [7, 30]. Such studies provide bounded attack evidence, not a license to assume that every cited or captioned source is vulnerable in the same way.

Boundary and failure conditions

A citation metric is not an authority score. A support judgment is not a truth judgment when the source itself is wrong. A source-influence estimate is not proof of authorship. A closed-surface citation observation identifies the surface, account state, location, and date—not the platform’s universal policy.

Reproduction Studio 5 · Audit citation and absorption separately

Question. When do displayed citations agree with supported and absorbed claims?

Input. A frozen context set with 20 verified material claims, 30 questions, at least two source-dependency clusters, and a locked generator.

Procedure. Generate repeated answers; segment claims; resolve URLs; double-label support; identify source-specific claims; run leave-one-source-out and leave-one-claim-out comparisons; and inject one wrong-source citation and one omitted qualification.

Output. Entailment, completeness, absorption, displayed-citation rate, judge calibration, counterfactual influence, uncertainty, and a failure matrix that never merges the measures into one score.

Reference criterion. A second reviewer can reproduce every claim span, source passage, resolution path, label, and context contrast.

Stop rule. Stop automated scoring if the calibration set shows a predeclared severe-error rate or if citation resolution loses the source version needed for adjudication.

Algorithm 3 Trace-preserving claim, citation, and influence audit

Inputs: raw answer y ; versioned answer state e ; displayed citation objects R ; source snapshots S ; annotation guide G ; calibration set K

Outputs: claim ledger, support metrics, judge diagnostics, and optional bounded influence contrasts

- 1: hash and freeze y, e, R, S, G, K before normalization
- 2: resolve redirects and canonical source identities; retain every resolution failure as a state
- 3: have two annotators independently segment material claims $A(y)$
- 4: preserve both segmentations, reconcile boundaries, and assign stable claim identifiers
- 5: attach citation objects only through visible scope or a declared resolver
- 6: label each claim–passage edge as support, qualify, contest, none, or unresolved
- 7: adjudicate disagreements without overwriting the pre-adjudication labels
- 8: compute entailment and completeness with declared weights and denominator flows
- 9: evaluate any automated judge on K by stratum; retain its confusion matrix and severe-error rate
- 10: **if** the preregistered severe-error or resolution-failure threshold is exceeded **then**
- 11: stop automated scoring and route affected strata to human review
- 12: **end if**
- 13: **for** each preregistered source s eligible for a controlled influence test **do**
- 14: construct C^{-s} and record all replacement, order, position, and budget changes
- 15: run paired seeds under C and C^{-s} ; retain missing and refused runs
- 16: estimate the declared utility contrast and cluster uncertainty
- 17: **end for**
- 18: export immutable raw objects, derived ledger, decisions, code/config hashes, and claim boundary

Table 19: Minimum claim–citation audit record.

Field group	Required record	Failure made visible
Answer state	raw answer, model/surface, version, locale, time, repeat	silent product or decoding drift
Claim unit	exact span, normalized proposition, materiality, weight	denominator and segmentation changes
Citation object	displayed anchor, resolved/canonical URL, passage, origin	redirects, wrong-source attachment, laundering
Support label	entail/qualify/contest/none, rationale, annotators	partial or contradictory support
Influence test	context condition, ablation, outcome, uncertainty	citation mistaken for causal use
Disposition	accept, qualify, correct, remove, escalate	unactioned audit findings

5.7.1 Reproduction checkpoint

An independent analyst receives the frozen answers and context variants, raw anchors, redirect log, source snapshots, annotation guide, both independent label files, adjudication record, materiality weights, calibration sample, generation configuration, paired seeds, and a manifest of SHA-256 hashes. The checkpoint passes only if that analyst can reconstruct every row of Table 17, reproduce Equation (28) and figure 14, and explain why the ablation contrast does not identify global platform behavior. Matching the final two decimals without reproducing claim boundaries and resolution states is not a pass.

Chapter summary

1. The effective context and system state define the generation condition.
2. Answer claims and citation objects must be segmented and resolved before support can be measured.
3. Selection, use, absorption, and displayed attribution are distinct.
4. Entailment and completeness have different denominators and failure modes.
5. Automated judges require calibration and sensitivity analysis.
6. Counterfactual influence is conditional on the constructed context and does not prove truth, authorship, or global platform behavior.
7. A release preserves raw objects, independent labels, denominator flows, calibration error, context changes, and failed resolutions; a scalar score is not a substitute for that trace.

6 Metrics, Labels, Judges, and Benchmark Design

Key idea

A GEO dashboard is not a measurement protocol. A protocol declares the target population, query distribution, system state, repetition scheme, unit of analysis, estimand, missing-data policy, uncertainty, and decision rule before responses are collected.

6.1 Questions this chapter answers

By the end of the chapter, a reader should be able to answer five questions without treating a dashboard value as self-explanatory:

1. What population, system state, outcome unit, and contrast does a GEO estimand describe?
2. How should query intents, paraphrases, stochastic repetitions, systems, and time windows be sampled and recorded as evaluation cells?
3. Why are repeated answers within one query intent not additional independent draws from the population of user needs?
4. How should unresolved labels, judge disagreement, collection failures, and other missing observations enter the analysis?
5. Which uncertainty interval, counterfactual, guardrail, and stopping rule make an intervention decision reproducible rather than retrospective?

6.2 Begin with the estimand

Suppose a permitted intervention δ changes source s into s^δ . For outcome Y under system m and query q , the ideal query-level treatment effect is

$$\tau(q, m) = \mathbb{E}[Y(s^\delta, q, m) - Y(s, q, m)]. \quad (31)$$

The target average is taken over an explicitly weighted query distribution Q and system distribution M :

$$\tau_{Q,M} = \sum_{q \in Q} \sum_{m \in M} w_q w_m \tau(q, m). \quad (32)$$

If the baseline and intervention cannot be shown concurrently, model drift and index updates can be confounded with δ . If only queries favorable to the brand are sampled, the estimand is visibility on that favorable panel, not on the user's demand space.

Definition 6.1: Evaluation cell

An evaluation cell is a versioned tuple

$$e = \langle q, m, v, d, \lambda, g, a, r, s \rangle, \quad (33)$$

where q is the query, m the system, v the product or model version when observable, d date and time, λ locale/language, g geography or endpoint, a account state, r repetition index, and s interface surface. The response and all derived labels are attached to this cell.

Every reported rate also needs a denominator flow. At minimum, retain planned cells, returned responses, citation-bearing responses, resolved citation objects, successfully fetched pages, segmented material claims, and adjudicated claim–source pairs. In one 602-prompt observational corpus, 23,745 citation feature records yielded 18,151 fetch-success pages, a 76.44% conditional subset [65]. A page-level score on those records estimates a fetch-conditioned object; silently treating it as the full citation population changes the estimand.

6.3 Build the query distribution before collecting answers

A query universe should represent tasks, not a bag of keywords. Useful strata include factual lookup, category exploration, comparison, troubleshooting, recommendation, procurement, reputation, and high-risk decisions. Within each stratum, queries should vary in specificity, entity naming, constraints, and language. The final sample records both design weights and observed usage weights if the latter are available.

Let stratum h contain N_h eligible queries and a sample of size n_h . A stratified estimate of mean outcome is

$$\hat{\mu} = \sum_{h=1}^H W_h \hat{\mu}_h, \quad W_h = \frac{N_h}{\sum_j N_j}. \quad (34)$$

Equal numbers per stratum are useful for diagnostics but do not imply equal population prevalence. Reports should show both unweighted diagnostic results and the declared weighted estimate.

6.3.1 Leakage and query tuning

Prompts used to develop an intervention cannot also be the only evaluation prompts. Separate development, validation, and held-out panels by underlying intent, not merely by surface paraphrase. If several paraphrases share the same fact need, they form a cluster for splitting and uncertainty estimation.

6.4 Metrics with declared units

For entity b , source s , response y_{qmr} , and R repetitions, define simple incidence measures:

$$\widehat{\text{MentionRate}}(b) = \frac{1}{|Q||M|R} \sum_{q,m,r} \mathbb{1}\{b \text{ is resolved in } y_{qmr}\}, \quad (35)$$

$$\widehat{\text{CitationRate}}(s) = \frac{1}{|Q||M|R} \sum_{q,m,r} \mathbb{1}\{s \in \mathcal{A}_{qmr}\}. \quad (36)$$

These estimators require an alias-resolution policy, a citation URL normalization policy, and a missing-response policy.

Citation frequency is not absorption. Let C_s be a set of source-distinctive, verified claims and let $z(c, y)$ indicate faithful presence of claim c in answer y . A transparent absorption estimator is

$$\widehat{\text{Absorb}}(s, y) = \frac{\sum_{c \in C_s} \omega_c z(c, y)}{\sum_{c \in C_s} \omega_c}, \quad (37)$$

with predeclared claim weights ω_c . If z is assigned by a model judge, a labeled human sample estimates judge error and disagreement.

Table 20: A minimum multi-objective GEO scorecard.

Dimension	Unit	Example metric	Essential companion
Exposure	entity–response	mention incidence, span position	sentiment and correctness
Attribution	source–claim–response	citation incidence, entailment, completeness	resolved URL and source independence
Absorption	claim–response	verified claim coverage	contradiction and scope preservation
Retrieval	query–passage	Recall@ k , nDCG@ k	downstream answer utility
Stability	query–system–time	variance, transition rate, survival	sampling frequency and missingness
Business	user–task	qualified action rate	experimental or quasi-experimental identification
Risk	claim, source, or response	unsupported claim rate, manipulation flags	severity and review outcome

6.5 Generated answers are repeated measures

One query issued once is a screenshot, not an estimate of a stochastic response distribution. Repeated runs are nested within query, system, and time. A useful hierarchical description is

$$Y_{qmt} = \mu + \alpha_q + \beta_m + \gamma_t + (\alpha\beta)_{qm} + (\beta\gamma)_{mt} + \varepsilon_{qmt}. \quad (38)$$

The purpose is not always to fit this exact model; it is to recognize that variation across prompts, systems, days, and stochastic repeats has different meanings. Treating every response as an independent row produces overconfident intervals.

Work focused on longitudinal GEO measurement argues that visibility should be described as a distribution and monitored over time rather than measured once [40]. A practical report therefore includes:

- the number of unique query intents and paraphrases;
- repetitions per cell and the interval between them;
- missing, refused, search-disabled, and error responses;
- within-query and between-query variability;
- sensitivity to URL, entity, and claim-resolution rules;
- a versioned raw-response archive where terms permit it.

The empirical lesson is to estimate a precision–cost curve, not to copy a repeat count. In the cited panel, same-window source-set Jaccard means ranged from 0.321 to 0.434, while adjacent-day means ranged from 0.336 to 0.423. Its brand-detection standard error fell from 0.370 with one run to 0.081 with seven and 0.062 with eight under the paper’s finite-panel calculation. These values are properties of that German-language, four-vertical, four-engine design; they are not a universal rule that seven or eight calls suffice.

6.5.1 Failure case: more calls do not create more query intents

Consider a synthetic study that sends one query 100 times to each arm. The entity appears in 50 baseline responses and 70 intervention responses. If all 200 responses are treated as independent population draws, the estimated difference is 0.20 and the usual two-sample Bernoulli standard error is

$$\sqrt{\frac{0.5(1-0.5)}{100} + \frac{0.7(1-0.7)}{100}} = 0.0678, \quad (39)$$

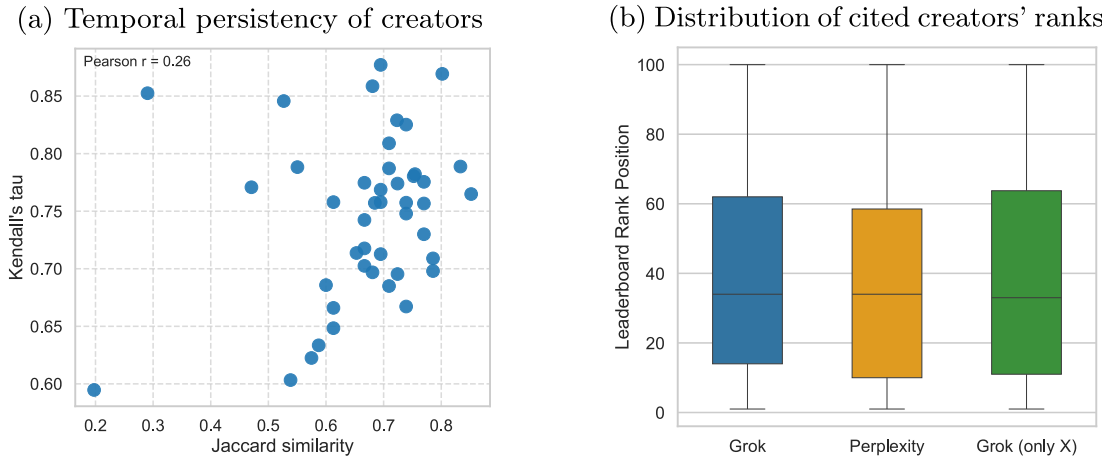


Figure 16: Two different exposure summaries: persistence of creator communities over time and the distribution of cited-creator ranks. Neither panel collapses the repeated observation into a single timeless visibility score.

Source note. Rectangular crop of Alipour et al. [2, Fig. 1], physical PDF page 5, CC BY 4.0; two Type 3 font programs were converted to vector outlines, with marks, values, colors, and geometry unchanged. The study reports a 20-day creator panel and engine-specific ranks; the plots do not identify a causal ranking mechanism or a population-wide exposure law.

giving a nominal 95% Wald interval of approximately [0.067, 0.333]. The calculation describes repeated generation for that one query under strong independence and stationarity assumptions. It does *not* estimate between-intent variation: for inference to a population of query intents the study has one cluster, so a cluster-based standard error is unavailable. More calls can reduce Monte Carlo error conditional on that query; they cannot repair the absence of independently sampled user needs.

6.6 Worked example: repeated measures with clustered uncertainty

This section uses a fully synthetic teaching dataset. It does not report the behavior of a production platform. In a frozen reconstructed pipeline, the Orchid Bay team compares a baseline source ($a = 0$) with a permitted, evidence-preserving revision ($a = 1$). The held-out panel contains eight procurement query-intent clusters. Each cluster has five paraphrases and two predeclared generation seeds, for ten planned responses per arm and 160 planned responses in total. Source version is paired within paraphrase-seed blocks; all other controlled components are held fixed.

Only the fields declared as observable may enter the label and estimator. An instrumented retrieval trace can be analyzed as a separate outcome, but an unlogged retrieval or generation state is latent and must not be back-filled from the presence or absence of the entity in the final response.

Let $Y_{har} \in \{0, 1\}$ be the resolved-mention label for returned repetition r in intent cluster h and arm a . If n_{ha} responses return and $x_{ha} = \sum_r Y_{har}$ contain the entity, define

$$\hat{p}_{ha} = \frac{x_{ha}}{n_{ha}}, \quad \hat{d}_h = \hat{p}_{h1} - \hat{p}_{h0}, \quad \hat{\tau}_{\text{intent}} = \frac{1}{H} \sum_{h=1}^H \hat{d}_h. \quad (40)$$

The response is the labeling unit, but the intent is the unit whose variation supports the declared generalization. Equal cluster weights prevent an intent with more valid repetitions from silently receiving more population weight.

Table 21: Measurement contract for the synthetic clustered example. The contract separates the response-level label from the intent-level unit of inference.

Protocol field	Frozen declaration
Target population	held-out procurement query intents represented by the eight-cluster teaching panel; clusters receive equal target weight
System and contrast	one frozen reconstructed pipeline; revised source minus baseline source within matched paraphrase–seed blocks
Observable record	intent ID, paraphrase ID, seed, arm, collection status, response, resolved-entity label, label provenance, and adjudication state
Outcome and unit	binary resolved mention for an entity–response; reported effect in percentage points
Inferential unit	query-intent cluster; paraphrases and repeated generations remain nested within their intent
Primary estimand	equal-intent average difference in mention probability among valid returned responses in this panel
Missingness	collection failures remain missing, never “no mention”; arm-specific counts and binary-outcome extreme bounds accompany the primary estimate

Table 22: Synthetic response counts for eight intent clusters. Each arm planned ten responses. “Missing B/I” gives baseline/intervention collection failures; the final column is computed only from valid responses.

Intent	Baseline x_{h0}/n_{h0}	Revision x_{h1}/n_{h1}	Missing B/I	$\hat{d}_h$
I1	3/10	5/10	0/0	0.200
I2	4/10	5/9	0/1	0.156
I3	2/9	3/10	1/0	0.078
I4	5/10	8/10	0/0	0.300
I5	4/10	5/10	0/0	0.100
I6	6/10	6/9	0/1	0.067
I7	5/9	4/10	1/0	−0.156
I8	3/10	6/10	0/0	0.300
Returned total	32/78	42/78	2/2	—

Using the unrounded ratios, the eight cluster differences sum to 1.0444, so

$$\hat{\tau}_{\text{intent}} = \frac{1.0444}{8} = 0.1306, \quad (41)$$

or a 13.1 percentage-point difference on this teaching panel. The sample standard deviation of the eight cluster differences is

$$s_d = \sqrt{\frac{1}{H-1} \sum_{h=1}^H (\hat{d}_h - \hat{\tau}_{\text{intent}})^2} = 0.1476, \quad \widehat{\text{SE}}_{\text{cluster}} = \frac{s_d}{\sqrt{8}} = 0.0522. \quad (42)$$

For a transparent small-sample illustration, treating the eight intents as independent draws gives

$$0.1306 \pm t_{0.975,7}(0.0522) = 0.1306 \pm 2.365(0.0522) = [0.007, 0.254]. \quad (43)$$

The interval is wide despite 156 returned answers because the relevant sample size for query-population uncertainty is eight. If the eight intents were a fixed census rather than a sample from a target population, this t interval would not have a superpopulation interpretation; the panel distribution would instead be reported descriptively or analyzed under a justified assignment design.

The four failed collections are a separate source of uncertainty. If each missing revision response is assigned outcome zero and each missing baseline response outcome one, restoring denominator ten within every arm gives

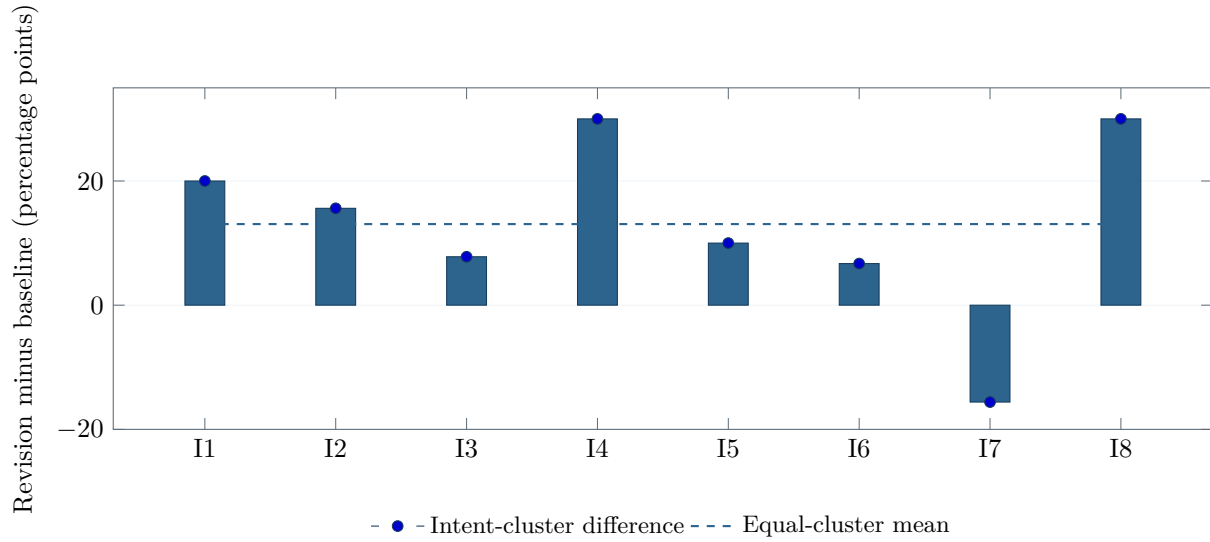


Figure 17: Intent-level effects recomputed from [table 22](#). One negative cluster is not erased by the positive mean. The dataset is synthetic and the dashed line is descriptive of this panel, not an estimate for a named platform.

an equal-cluster effect of 0.100. Reversing those assignments gives 0.150. Thus the binary-outcome extreme range is $[0.100, 0.150]$. This range is not a confidence interval; it asks how far the point estimate can move under extreme missing labels. A credible report shows both sampling uncertainty and missingness sensitivity rather than hiding one inside the other.

For a larger evaluation with enough independently sampled clusters, the paired cluster bootstrap in [algorithm 4](#) preserves the sampling unit. It must be adapted when the query design uses strata, unequal inclusion probabilities, or a fixed finite population.

Algorithm 4 Paired cluster bootstrap for a repeated-response GEO metric

Inputs: Frozen response ledger; intent-cluster IDs $h = 1, \dots, H$; paired arms $a \in \{0, 1\}$; outcome, validity, weight, and missingness rules; bootstrap replicates B ; random seed.

Outputs: Point estimate; cluster-respecting interval; heterogeneity, missingness, and deviation records.

- 1: Validate unique evaluation-cell IDs and reconcile planned, returned, valid, labeled, and adjudicated denominators by arm and cluster.
- 2: Apply the frozen entity resolver and adjudication rule; retain unresolved labels and collection failures as statuses rather than zeros.
- 3: **for** $h = 1$ to H **do**
- 4: Compute arm rates $\hat{p}_{h0}$ and $\hat{p}_{h1}$ from valid responses and the paired contrast $\hat{d}_h$.
- 5: Record repetitions, missingness, and within-cluster dispersion; flag an arm with no valid response under the predeclared policy.
- 6: **end for**
- 7: Compute $\hat{\tau} = \sum_h w_h \hat{d}_h / \sum_h w_h$ and retain the full distribution of $\hat{d}_h$.
- 8: **for** $b = 1$ to B **do**
- 9: Sample H cluster IDs with replacement under a rule matched to the cluster design (uniformly for an equal-probability cluster sample); each draw carries both arms and all nested responses.
- 10: Recompute the normalized weighted mean $\hat{\tau}^{*(b)}$; a cluster drawn twice contributes twice, but its rows are never split across arms.
- 11: **end for**
- 12: Form the predeclared bootstrap interval from $\{\hat{\tau}^{*(b)}\}_{b=1}^B$ and verify numerical stability across additional seeded replicates.
- 13: Recompute the estimate under missing-outcome bounds, alternate resolver rules, and any declared exclusions; do not select a specification by its sign.
- 14: Export the estimate, interval, cluster plot, denominator flow, code and environment versions, seed, and protocol deviations.

6.7 Judge calibration is a subgroup problem

Aggregate judge accuracy can conceal severe subgroup failure. In a controlled GEO-detection benchmark, a word-level TF-IDF detector reported overall F1 of 0.880 but worst-group accuracy of 0.375; an intervention-aware ModernBERT configuration reported F1 of 0.944 and worst-group accuracy of 0.883 [9]. Those numbers are benchmark-specific and do not certify a deployed detector. They demonstrate why a calibration report should include per-class confusion matrices, worst-group performance, false-positive gaps, abstention, and an independently adjudicated sample rather than one aggregate score.

6.8 Three experimental environments**6.8.1 Fixed-context experiments**

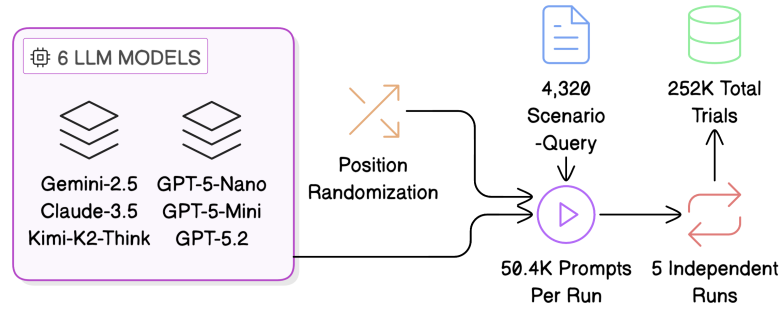
The researcher supplies the candidate documents or passages directly. This environment offers strong control over generation and attribution, making it appropriate for passage ablations, order effects, and content-use tests. It cannot show that a modified page would be retrieved from a large corpus.

6.8.2 Reconstructed end-to-end systems

The researcher controls a corpus, retriever, reranker, context builder, and generator. This supports stage-level traces and repeated counterfactuals. It is well suited to comparing algorithms and debugging mechanisms. Its

Table 23: What the three environments identify most credibly.

Environment	Strongest use	Main limit
Fixed context	causal source-use and generation comparisons	no retrieval qualification
Reconstructed pipeline	component traces and end-to-end benchmark effects	uncertain transfer to proprietary systems
Live platform	time-specific behavior of an actual surface	hidden confounding, drift, cost, and limited reproducibility

**Figure 18:** Nested execution scale in a competitive first-citation experiment: six models, scenario–query cells, randomized source position, repeated runs, and 252,000 trials. The diagram makes clear which repetition does and does not create a new scenario.

Source note. Rectangular crop of Vishwakarma et al. [49, Fig. 2], physical PDF page 3, CC BY 4.0. Each trial receives exactly two supplied source texts and models only the first citation; this is not open-web retrieval or a production-engine traffic experiment.

external validity depends on the resemblance between the reconstruction and the target ecosystem. Benchmarks that explicitly include retrieval and reranking address some limitations of fixed-candidate designs [6, 23].

6.8.3 Live-platform observations and interventions

The researcher queries a production surface. Ecological validity is higher, but internal states, drift, personalization, and concurrent updates are harder to control. A live before/after comparison should use concurrent or interleaved controls where possible. It should never violate platform terms, access controls, privacy requirements, or disclosure obligations.

6.9 Counterfactuals, ablations, and interference

A leave-one-source-out test compares $y(C)$ with $y(C \setminus \{s\})$; a leave-one-claim-out test removes a particular evidence unit. These counterfactuals can reveal source influence in controlled contexts, but natural language outputs make distance functions consequential. Reports should include claim-level differences, not only an embedding similarity score.

Interference is central to competitive visibility. Editing source s_i can change whether source s_j is selected, cited, or trusted. The stable unit treatment value assumption is often implausible. Competitive experiments should report total context composition and displacement, not only the target’s gain. Recent work on competitive GEO and repeated platform–supplier adaptation makes this ecosystem effect a first-class research problem [49, 59].

6.10 Decision rules and honest stopping

An experiment should declare three thresholds before data collection:

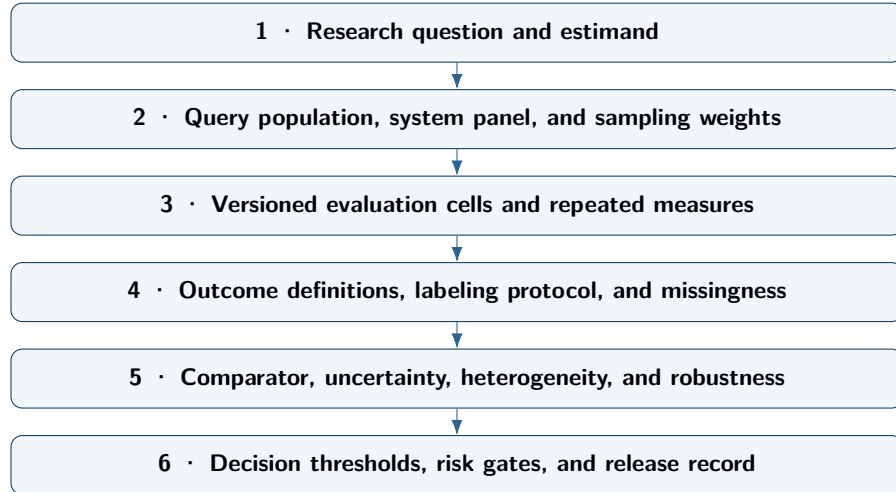


Figure 19: The six-layer measurement protocol. A dashboard begins only after the first five layers are specified.

1. a **minimum worthwhile effect** on the primary outcome;
2. a **non-inferiority margin** for fidelity and user utility;
3. a **maximum acceptable risk** for unsupported or manipulative content.

A statistically detectable mention increase is not a win if attribution accuracy falls beyond its margin. A noisy null is not proof of no effect. An inconclusive study should trigger a decision about more data, a better control, or stopping the intervention—not an invented narrative.

Reproduction Studio 6 · Design a measurement protocol

Input. One source, one permitted content intervention, a frozen query universe, and either a controlled pipeline or a live surface whose terms permit the study.

Protocol build. Before collection, write the target population, sampling frame, query-intent clusters, assignment unit, observation unit, analysis unit, estimand and units, repetition schedule, system/time cells, minimum worthwhile effect, guardrails, and missingness policy. Freeze positive and negative controls and a held-out intent split.

Measurement build. Create one row for every planned evaluation cell. Preserve collection status separately from the outcome. Double-code a predeclared stratified label sample, adjudicate disagreements without erasing them, and report judge performance by material subgroup.

Analysis build. Reconcile the denominator flow; compute the declared cluster estimate, interval, heterogeneity display, and missing-outcome sensitivity; then run the alternate resolver and exclusion analyses named in the protocol. Do not add analyses because their result is favorable.

Deliverable. A time-stamped protocol, source/query/system manifests, raw and adjudicated cell ledgers, executable analysis, generated tables and figures, deviation log, and a one-page decision memo containing uncertainty, guardrails, and failure cases.

Stop rule. Stop or redesign when controls fail, the system state cannot be versioned, or the intervention violates an evidence or risk gate.

6.11 Reproduction checkpoint

The studio is reproducible only if a second analyst can begin with the released ledger rather than a screenshot and recover the reported result. Before the chapter’s measurement protocol is considered complete, verify all of the following:

1. source versions, query manifests, cluster assignments, system cards, protocol, analysis environment, and permitted raw responses have stable IDs and cryptographic hashes;
2. planned, returned, valid, labeled, resolved, and adjudicated counts reconcile exactly by arm, cluster, system, and time block;
3. a clean run recomputes every estimate, interval, table, and plot from the released ledger, with no manually transcribed chart values;
4. the seeded resampling result is stable to more replicates, and an independently written calculation reproduces the point estimate and the direction of each cluster contrast;
5. unresolved labels, outages, refusals, exclusions, and protocol deviations remain visible, while missingness bounds and alternate resolver rules reproduce the reported sensitivity range; and
6. the decision memo applies the frozen worthwhile-effect, fidelity, utility, and risk thresholds even when doing so yields an inconclusive or negative decision.

Passing these checks establishes computational traceability for the declared study. It does not establish transfer to a different platform, query population, locale, model state, or time window.

6.11.1 A clean-run canary for the checkpoint

The abstract checkpoint above is exercised here against the bundled L06 repeated-measurement fixture. The build first invokes the Lab analyzer in a clean temporary directory and verifies its declared outputs. A separate standard-library program then reads the frozen event ledger, forms paired surface contrasts within each query, resamples whole query clusters, and emits the plot tables and a hash-bearing run manifest. The run uses 10,000 bootstrap replicates with seed 20,260,825; both surfaces and all repetitions nested under a sampled query travel together.

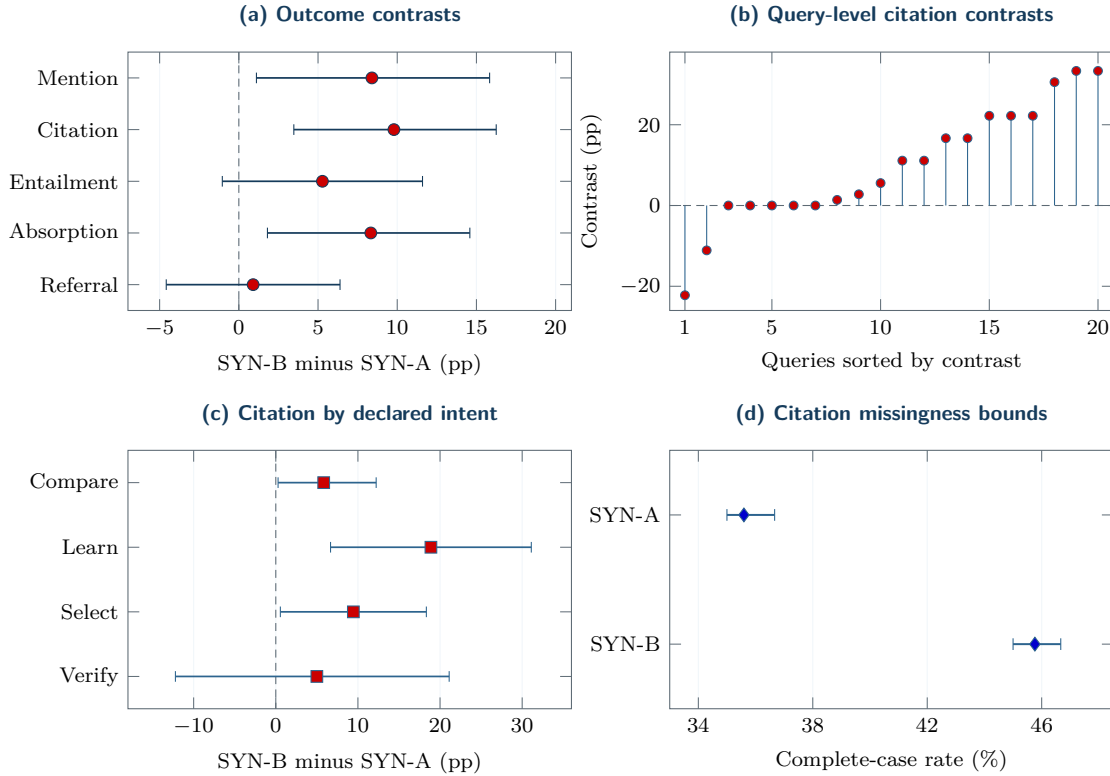


Figure 20: Deterministic local reproduction of a frozen synthetic repeated-measurement panel. The ledger contains 360 designed events in 20 query clusters, of which 354 are complete and 6 are noncomplete. Panels (a) and (c) show equal-query mean contrasts with percentile intervals from 10,000 seeded cluster-bootstrap replicates; they are descriptive of the authored query panel, not population confidence intervals. Panel (b) retains all query-level citation contrasts, including negative and zero values. Panel (d) shows the complete-case citation rate and the logical binary-outcome range obtained by assigning every noncomplete outcome to zero or one; these bounds are not sampling intervals.

The synthetic panel's citation contrast is 9.79 percentage points; its seeded query-cluster interval is 3.47 to 16.25 points. The aggregate retains two negative and five exactly zero query contrasts. The Verify interval crosses zero, as do the overall Entailment and Referral intervals. From the first to last time block, citation rates decrease by 2.87 points on SYN-A and 2.48 on SYN-B. These null and negative diagnostics are required, not post hoc defects to remove.

The clean run establishes only computational traceability. The data are a deterministic synthetic teaching fixture, SYN-A and SYN-B are fictional labels, and the 20 authored queries are not a probability sample of demand. The plate therefore estimates no named product, live engine, causal intervention, business outcome, or population prevalence. Its reproducibility identity is the frozen panel SHA-256 `c013f9ae...5d7`; the complete hash, environment, script identities, seed, output hashes, and claim ceiling are machine-checked in the figure manifest and run manifest.

Chapter summary

1. Define the estimand, target population, contrast, outcome unit, and inferential unit before collecting responses.
2. Mention, citation, absorption, retrieval, stability, business, and risk metrics have different units and must not be collapsed blindly.
3. Repeated responses are nested observations. More generations can estimate within-query stochasticity but do not increase the number of independently sampled query intents.
4. Missing responses are statuses, not negative outcomes. Sampling intervals and missing-outcome sensitivity answer different questions and should be reported together.
5. Judge quality is a subgroup property as well as an aggregate property; unresolved cases and adjudication must remain auditable.
6. Fixed-context, reconstructed, and live studies identify different effects and offer different forms of control and external validity.
7. A win requires a worthwhile primary effect and acceptable fidelity, utility, and risk outcomes under a frozen decision rule.

7 Experiments, Causal Inference, and Temporal Uncertainty

Key idea

An observed visibility change is causal only relative to a declared treatment, assignment mechanism, comparison condition, estimand, and set of assumptions. Generated-answer systems add repeated stochastic outputs, platform drift, interference among sources, and missing cells; a before/after screenshot cannot separate these processes.

7.1 Questions this chapter answers

- Q1.** What exactly is assigned, observed, and estimated when a source is revised but the generative system is only partially observable?
- Q2.** Which comparison design can separate an intervention from query mix, system drift, index propagation, and ordinary stochastic output variation?
- Q3.** How should repeated runs, query-intent clusters, time blocks, and missing cells enter the estimator and its uncertainty?
- Q4.** Which diagnostics can challenge—but never prove—parallel trends, no carryover, stable measurement, and absence of interference?
- Q5.** When must an apparently positive result be reported as descriptive, partially identified, or inconclusive rather than causal?

7.2 Define treatment versions and potential outcomes

Let $D_i \in \{0, 1\}$ assign evaluation unit i to baseline source state $\mathcal{W}^0$ or permitted intervention state $\mathcal{W}^1$. For a versioned system state S_t , define potential outcome

$$Y_i(d, S_t), \quad d \in \{0, 1\}. \quad (44)$$

The observed outcome is $Y_i = Y_i(D_i, S_t)$. A finite-sample average treatment effect under fixed S_t is

$$\tau(S_t) = \frac{1}{N} \sum_{i=1}^N [Y_i(1, S_t) - Y_i(0, S_t)]. \quad (45)$$

The system state is explicit because an index update, model revision, user interface change, or source-corpus change can alter both potential outcomes. If treatment and control are observed in different periods, additional assumptions are required.

The treatment includes the complete content version, not a label such as “added statistics.” Record the changed claims, markup, linked evidence, rendered output, content hash, publication time, propagation lag, and rollback state. Compound interventions identify their bundle unless a factorial or ablation design separates components.

7.3 Write the identification contract before collecting outcomes

Table 24: A minimum identification contract. A diagnostic can expose a failure; passing it does not establish the assumption.

Assumption	Operational meaning	Predeclared challenge or diagnostic
Consistency and versions	assignment maps to one recorded source state, not an unnamed family of edits	content hashes, rendered diffs, uptake checks, and version-specific exclusions
Exchangeability	assignment or the comparison design breaks the association between treatment and untreated potential outcomes	randomization audit, baseline balance, allocation concealment, or explicit adjustment set
Stable measurement	collection, parsing, and judging mean the same thing across arms and periods	frozen prompts and parsers, blinded re-annotation, state fingerprints, and instrument-change epochs
No anticipation or carryover	a cell is not affected before activation or after the declared washout	propagation-lag study, lead effects, washout sensitivity, and treatment-history fields
Comparison trend validity	absent treatment, treated and control aggregates would have evolved comparably	multiple pre-period trajectories, placebo dates, event-study contrasts, and bounded trend-deviation analysis
Limited interference	an assigned source change does not alter another unit’s outcome except through declared exposure mappings	candidate-set snapshots, spillover outcomes, exposure mapping, and displacement analysis
Positivity and attrition control	every target stratum can receive the needed condition and observation is not selected by post-treatment events	support tables, planned-cell flow, arm-specific missingness, and worst-case bounds

7.3 Write the identification contract before collecting outcomes

A causal analysis is reproducible only if its observable record is separated from the latent quantities supplied by assumptions. For cluster c , time block b , and repeat r , a minimum observed cell can be written

$$O_{cbr} = (D_{cb}, X_c, t_b, H_b, U_{cb}, R_{cbr}, Y_{cbr}, G_{cbr}, J_{cbr}), \quad (46)$$

where D is assigned treatment, X is the frozen cluster descriptor, H is a system-state fingerprint, U is measured treatment uptake, R indicates whether the planned outcome is observed, Y is the primary outcome, G is a vector of guardrails, and J identifies the judge or annotation state. The counterfactual outcome $Y_{cbr}(d, S_t)$, undocumented retrieval state, latent query demand, unmeasured competitor changes, and the judge’s error-free label are not observed merely because a row exists in the archive.

These assumptions define the identification ceiling. A state fingerprint may include the observable product or model label, interface mode, locale, search setting, collection code version, and timestamp. It is not evidence that all hidden system components were stable. When a hard state break occurs, the analysis should split epochs or stop; silently averaging across it changes the target estimand.

7.4 Choose an environment that matches the claim

A fixed-context study has strong control over available evidence and weak qualification for web acquisition. A reconstructed pipeline can expose every stage but remains a model of a target environment. A live observation improves ecological relevance while hiding internal treatment uptake. Reports should state which stage the treatment can reach and which conclusions remain outside the design.

Three current evidence designs illustrate the resulting claim ceilings. The 252,000-trial citation-choice study counterbalances two supplied sources and estimates the first displayed URL, but never invokes web retrieval [49]. SAGEO Arena reruns a 171,003-page reconstructed retrieval–reranking–generation funnel for 2,700 queries, but

Table 25: Design families for different causal questions.

Design	Most credible target	Central failure condition
Fixed-context randomization	effect of source expression on use/generation	no end-to-end retrieval claim
Frozen reconstructed pipeline	component and end-to-end effect in an open system	uncertain transfer to a closed surface
Concurrent page or source experiment	time-specific live-surface effect	contamination, propagation, policy, or assignment failure
Switchback or interleaving	reversible short-window comparison	carryover and nonstationary demand
Difference-in-differences	comparative change with untreated units	violated parallel trends or spillovers
Interrupted time series	level/slope change after a dated event	concurrent shocks and insufficient pre-period

selects targets from the baseline top ten and reports no run-level confidence intervals [23]. A live four-engine panel repeats 32 German prompts over 45–46 days, but cannot isolate model decoding from routing, retrieval, cache, and product drift [40]. None is a weaker version of the others; each identifies a different object.

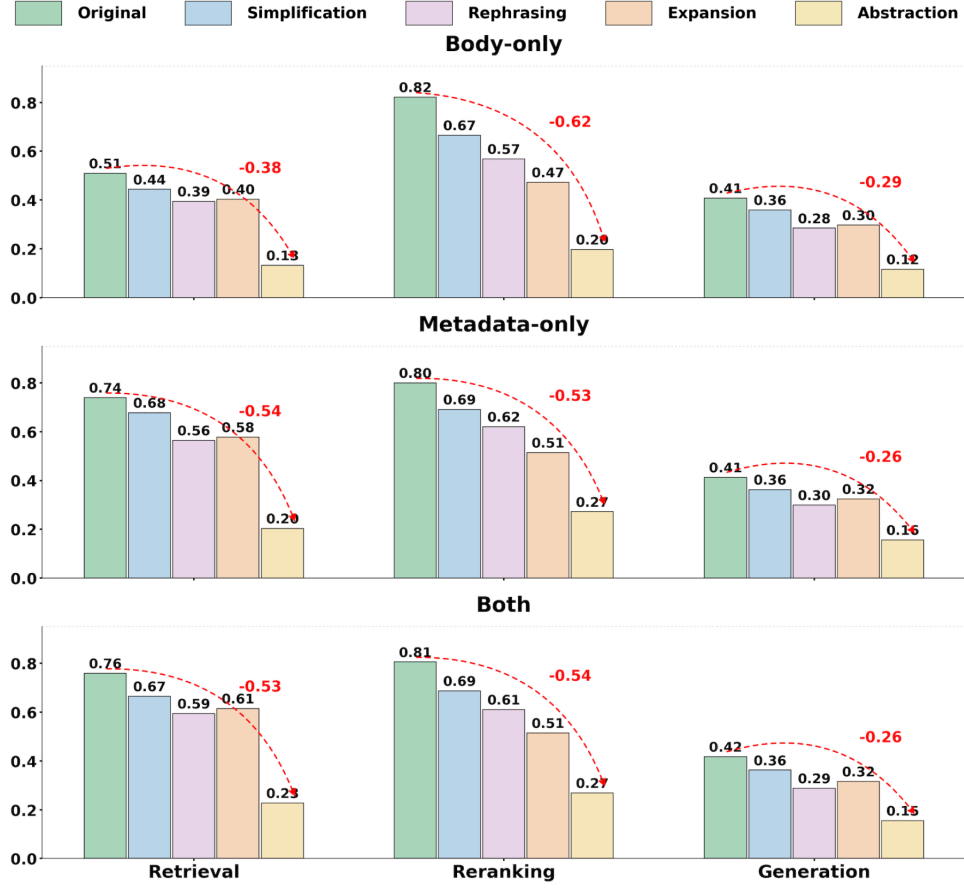


Figure 21: Query variation can move retrieval hit, reranking hit, and generation citation rate differently. Reproduced from Kim et al. [23, Fig. 7], under CC BY 4.0; crop only. The result is bounded to the paper’s reconstructed corpus, target gate, strategies, and perturbations, not a universal platform or causal effect.

7.5 Randomize and analyze at the correct unit

Queries that express the same intent are not independent experimental units. Repeated runs under one query are measurements within a cluster. If assignment occurs by page, domain, query intent, or time block, uncertainty should respect that assignment unit.

For cluster c with n_c observations, let $\bar{Y}_c(d)$ be its mean outcome under treatment d . A transparent cluster-level estimator is

$$\hat{\tau}_{cl} = \frac{1}{C_1} \sum_{c:D_c=1} \bar{Y}_c - \frac{1}{C_0} \sum_{c:D_c=0} \bar{Y}_c. \quad (47)$$

Randomization inference or a cluster bootstrap can provide intervals aligned with the assignment. A mixed-effects model can improve precision and describe heterogeneity, but model assumptions do not repair a flawed assignment.

Table 26: Synthetic two-period intervention audit. Each cell contains 120 planned observations nested in query-intent clusters. The outcome is a faithful, source-attributed answer under a frozen coding rule. These counts are a teaching construction, not measurements from a production platform.

Group	Source state	Pre-period successes	Post-period successes
Concurrent control	unchanged, version checked	30/120 = 0.250	36/120 = 0.300
Intervention group	baseline, then governed revision	24/120 = 0.200	45/120 = 0.375

Worked example 7.1: Five repeats are not five independent queries

Suppose 30 query intents are each run five times on three systems. The archive contains 450 responses, but the treatment was assigned at the 30-intent level. Treating 450 rows as independent can make intervals far too narrow. The primary uncertainty resamples or models the 30 clusters; the five runs estimate within-intent stochastic variation, and the three systems define either strata or a target distribution with declared weights.

7.5.1 Allocate precision across three variance components

Query or document sampling, repeated pipeline runs, and system drift are different uncertainty sources. More same-query calls estimate run variance; they do not create new query clusters or observe a later product state. For approximately equal cluster sizes $\bar{m}$, the familiar design-effect diagnostic

$$\text{DE} = 1 + (\bar{m} - 1)\rho \quad (48)$$

shows why adding repeats has diminishing independent-information value when within-cluster correlation ρ is positive. Power planning should use a pilot to estimate the query-cluster variance, run variance, base rate, and drift window, then simulate the actual assignment and missingness policy. Neither this approximation nor a mixed-effects model repairs too few independently assigned clusters.

7.6 Separate time drift from treatment uptake

Let G_i indicate a treated group and T_t the post-intervention period. A two-period difference-in-differences contrast is

$$\hat{\tau}_{\text{DiD}} = (\bar{Y}_{G=1, T=1} - \bar{Y}_{G=1, T=0}) - (\bar{Y}_{G=0, T=1} - \bar{Y}_{G=0, T=0}). \quad (49)$$

It identifies a causal contrast only under assumptions such as parallel trends in the absence of treatment, no anticipation, stable measurement, and appropriate handling of interference. Plot pre-period trajectories and placebo dates; do not use the formula as a decorative adjustment.

Worked example 7.2: From a 17.5-point change to a 12.5-point controlled contrast

The intervention group's raw before/after change is $0.375 - 0.200 = 0.175$, or 17.5 percentage points. Over the same windows, the unchanged control rises by $0.300 - 0.250 = 0.050$. Under the stated difference-in-differences assumptions,

$$\hat{\tau}_{\text{DiD}} = (0.375 - 0.200) - (0.300 - 0.250) \quad (50)$$

$$= 0.125. \quad (51)$$

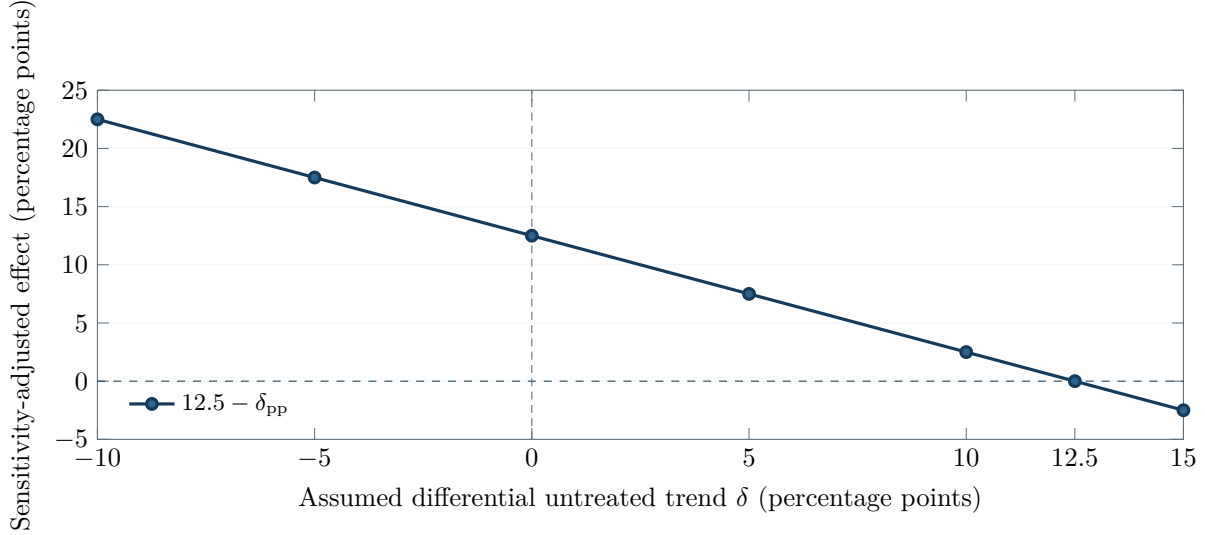


Figure 22: A recomputable trend-deviation analysis for table 26. The plotted values follow $100\hat{\tau}(\delta) = 12.5 - 100\delta$; they are not an empirical confidence interval. The sign reversal at 12.5 points makes the identifying assumption visible instead of hiding it in a footnote.

The implied untreated post-period counterfactual for the intervention group is $0.200 + 0.050 = 0.250$, so the same contrast is $0.375 - 0.250 = 0.125$. Thus 5.0 of the 17.5 observed percentage-point rise is assigned to the concurrent background trend and 12.5 points to the intervention—but only under the identification contract. The cell counts permit the point estimate; uncertainty must still be computed at the query-intent assignment level rather than from 480 independent Bernoulli rows.

The parallel-trends condition is counterfactual, not a testable equality. Let δ be the untreated trend in the intervention group minus the untreated trend represented by the control. The sensitivity-adjusted effect for the worked example is

$$\hat{\tau}(\delta) = 0.125 - \delta. \quad (52)$$

For example, an untreated differential trend of $+0.05$ reduces the contrast to 0.075 ; a differential trend of $+0.125$ explains the entire observed contrast. Negative values imply that the control trend was more favorable than the intervention group's untreated trend.

Proposition 7.3: A two-period difference-in-differences contrast is not point identified under unrestricted differential drift

Let κ denote the observed difference-in-differences contrast, τ the causal effect, and δ the difference between the intervention group's untreated trend and the control trend. Then $\kappa = \tau + \delta$. Observed group means alone do not identify τ when δ is unrestricted. If the design justifies $|\delta| \leq b$, then τ is partially identified in $[\kappa - b, \kappa + b]$.

Proof. Write the observed intervention-group change as its untreated change plus τ . Subtracting the control change gives $\kappa = \tau + \delta$. For every candidate effect τ^* , setting $\delta^* = \kappa - \tau^*$ reproduces the same observed contrast, so the group means cannot distinguish the pairs. Under $|\delta| \leq b$, $-b \leq \kappa - \tau \leq b$, which rearranges to $\kappa - b \leq \tau \leq \kappa + b$. $\square$

Multiple pre-periods can reveal grossly incompatible trajectories and support a bounded b , but they cannot

observe the post-period untreated trajectory of the treated units. Report the full sensitivity interval, its scientific basis, and whether the decision changes before presenting the parallel-trends point estimate.

Longitudinal GEO measurement work emphasizes repeated observation and temporal variation [40]. That evidence motivates a time-aware protocol, not a universal sampling interval. Sampling frequency should reflect expected product changes, index latency, outcome volatility, cost, and the minimum decision horizon.

7.6.1 Use switchbacks only when treatment is reversible on the study clock

A switchback assigns conditions to time blocks rather than permanently to units. It is useful for a reversible fixed-context treatment or controlled pipeline configuration when concurrent randomization is impractical. It is a poor default for a public-page revision if indexing, caching, or learned adaptation can persist after the nominal condition changes. Randomize the order within matched block pairs, balance day-of-week and hour strata, declare a washout, and retain treatment history for every cell. For pair p , let $\bar{Y}_{p1}$ and $\bar{Y}_{p0}$ be the cluster-weighted outcomes observed under the two conditions. A transparent paired estimator is

$$\hat{\tau}_{\text{SB}} = \frac{\sum_{p=1}^P w_p (\bar{Y}_{p1} - \bar{Y}_{p0})}{\sum_{p=1}^P w_p}, \quad (53)$$

with randomization inference based on the permitted within-pair assignment flips. Exclude neither transition blocks nor adverse states after looking at their outcomes. Apply the preregistered washout or classify the comparison as carryover-contaminated.

7.7 Define missingness as an outcome or a censoring process

Refusal, timeout, search-disabled response, no citation panel, parser failure, and policy block are different states. Let $R_i = 1$ when the planned outcome is observed. A complete-case estimand conditions on $R_i = 1$ and can be biased when missingness depends on treatment, query, or system. The protocol specifies whether each state is:

- a valid outcome, such as a refusal rate;
- a recoverable operational error with a bounded retry rule;
- censored with an explicit assumption and sensitivity analysis; or
- a protocol failure that invalidates the comparison.

Never retry until the preferred result appears. Preserve the first attempt, retry reason, count, delay, and all terminal states.

7.8 Account for judge error and multiple outcomes

If treatment effects are estimated from automated labels, judge calibration is part of the measurement model. Blind a human calibration sample to treatment and expected direction; stratify it by system, language, claim type, and model confidence. Recompute the treatment decision under plausible error matrices.

GEO interventions usually report mention, citation, absorption, fidelity, prominence, user utility, and risk. Declare one primary outcome or a formal decision rule, designate guardrails, and control multiplicity for confirmatory claims. A post hoc “win” on any one of many outcomes is exploratory evidence.

7.9 Treat competition and adaptation as interference

Changing source i can alter retrieval or context capacity available to source j . The potential outcome may therefore be

$$Y_i(D_i, \mathbf{D}_{-i}, S_t), \tag{54}$$

not merely $Y_i(D_i)$. Competitive GEO studies and repeated strategy-learning systems make this dependence central [49, 56]. Experiments should record the complete candidate and context composition and report displacement, not only the target source’s gain.

7.10 Run a drift-aware analysis without conditioning on success

Treatment uptake is often imperfect or delayed. Because uptake can be affected by assignment and system state, dropping cells with $U_{cb} = 0$ or replacing assignment with observed uptake generally changes the randomized estimand. The primary analysis should therefore preserve the intention-to-treat contrast; an uptake-adjusted or per-protocol estimate is secondary and requires its own identification assumptions. Likewise, a state break discovered after outcomes arrive is not permission to select the more favorable epoch.

Algorithm 5 Drift-aware intervention analysis

Inputs: Preregistered clusters $\mathcal{C}$, blocks $\mathcal{B}$, assignment D , source hashes, primary outcome, guardrails, state-break rule, missingness policy, and analysis code.

Outputs: Versioned cell ledger, assignment-respecting effect estimates, diagnostics, sensitivity envelope, deviations, and decision.

- 1: Freeze cluster weights, estimand, multiplicity rule, retry budget, washout, exclusions, and stop thresholds.
- 2: Materialize every planned (c, b, r) cell before collecting outcomes.
- 3: **for all** $b \in \mathcal{B}$ in chronological order **do**
- 4: Record the pre-collection observable state fingerprint H_b^{pre} .
- 5: **for all** planned cells (c, b, r) **do**
- 6: Collect the first attempt; preserve timestamp, raw response, attributions, trace, and terminal status.
- 7: Verify assigned condition, delivered source hash, and uptake U_{cb} ; retain assignment even when uptake fails.
- 8: Classify refusal, timeout, policy state, absent panel, and parse failure under the frozen missingness rule.
- 9: Apply the blinded outcome and guardrail rules; append judge state J_{cbr} and candidate/context composition.
- 10: **end for**
- 11: Record H_b^{post} and compare both fingerprints with the current epoch.
- 12: **if** either fingerprint crosses the preregistered hard-break rule **then**
- 13: Close the epoch; execute the declared stop-or-split action.
- 14: **end if**
- 15: **end for**
- 16: Lock exclusions and reconcile planned, collected, observed, and analyzed cell counts by arm, block, cluster, and state epoch.
- 17: Audit assignment, baseline support, treatment versions, uptake, attrition, and contamination before estimating effects.
- 18: Estimate the primary intention-to-treat contrast within each comparable epoch at the assignment unit.
- 19: Compute randomization or cluster-respecting intervals; aggregate epochs only under the preregistered target weights.
- 20: Run lead and placebo-date tests, negative controls, carryover checks, and candidate-displacement analysis.
- 21: Re-estimate across declared bounds for trend drift, missing outcomes, judge error, uptake, and spillovers.
- 22: Apply multiplicity and guardrail rules; label the result RETAIN, REVERT, or INCONCLUSIVE.
- 23: Release manifests, cell ledger, code, diagnostics, sensitivity results, and all deviations with the versioned claim ceiling.

Boundary and failure conditions

Statistical significance cannot rescue an unidentified comparison. A live before/after increase may reflect treatment, platform drift, changing query demand, index propagation, competitor edits, retry rules, or judge drift. State the causal assumptions and show the diagnostics that make each one plausible; otherwise report a versioned association.

7.11 Reproduction checkpoint: re-create every denominator

The checkpoint is passed only when an analyst who did not build the study can recover the planned population, treatment versions, analytic population, and decision from frozen artifacts. Reproducing a final coefficient from a cleaned table is insufficient: the exercise must expose planned cells that became refusals, timeouts, parser failures, carryover exclusions, or unobserved uptake. The default course path uses a reconstructed or synthetic environment. A live surface is optional and does not expand the permitted claim beyond its observed time, interface, query distribution, and versioned conditions.

Reproduction Studio 7 · Pre-register and analyze a drift-aware intervention

Question. Does a bounded evidence-preserving change improve a primary visibility outcome without reducing fidelity or increasing risk?

Inputs. Thirty query-intent clusters; baseline, intervention, and concurrent-control source manifests; four time blocks with five planned repeats per cluster-condition; a fixed annotation guide; treatment and state hashes; and a seeded synthetic drift/missingness injector.

Pre-registration. Freeze the cell IDs, assignment seed, cluster weights, primary intention-to-treat estimand, minimum detectable effect, missing-state and retry policies, one fidelity and one risk guardrail, multiplicity rule, drift threshold, washout, stopping rule, and decision table. Do not reveal the injected drift or missingness parameters to the analyst.

Execution. Preserve every first attempt. Produce a CONSORT-style flow from planned to analyzed cells; verify delivered hashes and uptake; analyze at the cluster level; add a hierarchical model only as a declared sensitivity analysis. Plot pre-period trajectories and effect intervals. Run a placebo date, a negative-control outcome, a carryover check, an interference/displacement audit, and sensitivity grids for differential trends, missing outcomes, and judge error.

Required outputs. Preregistration; version and environment manifests; raw cell ledger; immutable exclusion/deviation log; analysis code and seed; planned/collected/observed/analyzed flow; effect and interval plots; missingness and uptake tables; blinded judge-calibration record; diagnostic results; and a RETAIN, REVERT, or INCONCLUSIVE decision.

Reproduction tests. A second analyst must (i) reconstruct every denominator from cell IDs, (ii) recover 0.175 as the descriptive intervention change and 0.125 as the controlled contrast in [table 26](#), (iii) recover 0.075 at $\delta = 0.05$ and the zero crossing at $\delta = 0.125$, (iv) show that intervals use query-intent clusters rather than rows, and (v) trigger the declared response to one injected state break or guardrail breach.

Stop rule. Stop or redesign if treatment uptake cannot be verified, the control is contaminated, the cell ledger cannot reconcile, severe guardrail harm crosses threshold, or the system state leaves the declared comparability window. A failed diagnostic yields an inconclusive or narrower claim, not an unregistered replacement analysis.

Chapter summary

1. A causal result requires versioned treatments, an assignment mechanism, potential outcomes, a target system state and population, an estimand, and a declared identification contract.
2. Observed cells contain assignments, timestamps, state fingerprints, uptake, outcomes, guardrails, missing states, and judge states; counterfactual outcomes and hidden system mechanisms remain latent.
3. The experimental environment limits the stage and population identified. Fixed-context, reconstructed-pipeline, live concurrent, switchback, and longitudinal designs answer different questions.
4. Repetitions are nested measurements. Randomization, estimation, resampling, and degrees of freedom must respect query, source, or time-block assignment rather than treating response rows as independent units.
5. In the worked audit, the raw 17.5-point rise becomes a 12.5-point difference-in-differences contrast after subtracting the concurrent 5-point trend. That causal interpretation still depends on parallel trends and the other stated assumptions.
6. With unrestricted differential drift, the treatment effect is not point identified. A scientifically justified drift bound yields a sensitivity interval; pre-trend diagnostics alone do not prove the counterfactual trend.
7. Switchbacks require reversible treatments, randomized block order, and defensible washout. Missing cells, retries, judge error, multiplicity, uptake, carryover, and interference are parts of the design, not cleanup details.
8. A drift-aware release includes planned-cell reconciliation, state epochs, intention-to-treat estimates, assignment-respecting intervals, negative and placebo controls, guardrails, sensitivity envelopes, and all deviations. When identification fails, report a versioned association or an inconclusive experiment rather than a causal narrative.

8 Content and Evidence Engineering

Key idea

Responsible content optimization makes verified information easier to retrieve, interpret, attribute, and update. It does not manufacture authority. The unit of work is a governed claim and its evidence path, not a keyword inserted into fluent prose.

8.1 From an asset library to a fact system

Organizations often begin with slide decks, campaign copy, product pages, press coverage, and analyst reports. These assets were produced for different dates, audiences, and approval regimes. Publishing them more widely can amplify conflicts: two prices, an expired certification, a capability that applies only to one region, or a customer result stripped of its experimental context.

The Claim–Evidence–Source graph from [chapter 2](#) becomes operational through a fact ledger. Each row contains:

- canonical entity identifier and human-readable name;
- minimal predicate and value;
- time interval, locale, plan, audience, and other conditions;
- evidence spans, source identities, and independence status;
- owner, reviewer, approval state, and next review date;
- known contradictions, superseded values, and correction history;
- permitted public formulations and prohibited overstatements.

The ledger is not a public knowledge graph by default. It is a governance layer from which pages, documentation, feeds, structured data, media responses, and evaluation fixtures can be generated consistently.

Observation 8.1: Completeness is typed, not scalar

A source can be complete at document level and incomplete at answer-adjacent passage level. It can contain every product attribute but omit the relation between a metric and its test condition. It can be current but lack independent support. A single “content completeness” percentage hides these failure modes.

8.2 A page is an evidence interface

A well-engineered page aligns human comprehension, machine acquisition, passage retrieval, and governance. [Table 27](#) describes a reusable evidence page rather than a guaranteed ranking template.

The design principle is **co-located qualification**: a number and its unit, population, date, and exception should survive passage segmentation together. This is good scientific writing and good retrieval hygiene. It is not a request to repeat the same phrase mechanically.

8.3 Structural signals without structural superstition

Headings, lists, tables, captions, definitions, and question–answer blocks can make information easier to navigate and segment. Recent papers study content structure, cooperative optimization, and source influence under different

Table 27: Evidence-page components and their diagnostic purpose.

Component	Human and governance purpose	Retrieval and evaluation purpose
Identity header	names the entity, owner, version, and last substantive review	resolves aliases and separates current from superseded pages
Answer-first summary	states the scoped proposition and its main condition	creates a coherent passage without hiding exceptions elsewhere
Evidence table	pairs values with units, methods, dates, and sources	supports claim-level extraction and comparison
Method and limitations	describes how results were produced and where they fail	prevents unsupported transfer and supplies negative evidence
Definitions	disambiguates technical and commercial terms	stabilizes query and entity matching
Change log	makes corrections and version transitions visible	supports freshness and temporal audits
Source list	resolves primary and independent evidence	enables attribution and circular-citation checks
Next action	offers documentation, data, contact, or correction routes	distinguishes informational visibility from a declared business outcome

GEO setups [6, 57, 61]. Their existence does not justify a universal checklist in which every page must use the same format.

Structure has at least four potential pathways:

1. **acquisition:** critical text appears in parseable HTML and has stable identifiers;
2. **retrieval:** headings and local context improve passage meaning;
3. **generation:** claims, evidence, and conditions can be combined with less ambiguity;
4. **attribution:** source identity and evidence boundaries remain resolvable.

An evaluation should isolate the candidate pathway. If a Q&A rewrite improves human scanning but does not improve absorption in a particular platform study, the correct result is pathway-specific, not “FAQs do not work.” The selection–absorption study in the current library reports precisely such a negative result for Q&A formatting under its protocol [65].

8.3.1 Replace tactic lists with stage-conditioned evidence

The present corpus contains direct counterexamples to a universal writing recipe. Table 28 retains the intervention point, negative result, and claim ceiling alongside each positive result.

One structural paper in the supplied corpus reports a 17.3% citation-rate increase, but does not adequately reconcile its query/case counts, statistical unit, platform versions, repetitions, evaluation scale, or reproduction artifact [61]. These Notes quarantine that number: it may motivate a preregistered structural ablation, but it is not an admitted design rule or an empirical headline.

8.4 Intervention families and guardrails

The early GEO study evaluated stylistic and evidence-related modifications, including citations, quotations, and statistics. It reported heterogeneous effects across domains and methods [1]. Subsequent work expands the design space toward content-centric benchmarks, intent-aware rewriting, structural features, multi-objective methods, agentic systems, and trajectory-aware evidence ecosystems [6, 8, 26, 49, 60, 61, 63]. For practice, these methods are grouped by what is permitted to change.

The factor-by-model matrix in figure 23 is a useful counterexample to a universal tactic list. Formatting

Table 28: Three content-intervention studies with non-interchangeable estimands.

Study and controlled object	Reported contrast retained in these Notes	Negative result and admissible ceiling
E-GEO; one rewritten product inside a cached ten-item slate [3]	Across five rerankers, simple-prompt row means vary by rewriter: Claude +0.57, GPT-5 +0.56, and GPT-4o-mini −0.36 in paired rank change.	Eleven of fifteen hand-written prompts underperform the GPT-4.1 simple-prompt mean; storytelling is −4.36. The treatment starts after retrieval and says nothing about candidate inclusion, live shopping, clicks, or sales.
SAGEO Arena; reconstructed retrieval–reranking–citation pipeline [23]	With 171,003 documents and 2,700 queries, body-only scores are .53/.84/.47, structure-only .71/.83/.52, and StageAware .75/.80/.58 across the three reported stages.	The gains do not peak at the same stage; a reported quality score falls from 88.9 to 87.3. This supports stage-aware evaluation in the benchmark, not a production guarantee for structured pages.
FeatGEO; fixed five-source context plus one advertiser page [26]	Across GPT-4o-mini, Gemini-2.5-Flash, and Qwen-plus, visibility changes from 13.34/8.89/5.20 without optimization to 18.31/15.35/10.17 for the reported multi-objective method.	Most single heuristics are below the no-optimization condition, and feature ablations change by engine and domain. Automated visibility and quality judges do not establish human quality, factual fidelity, or live-platform effects.

contrasts do not show stable significance across the six evaluated configurations, while very large point odds ratios elsewhere sometimes coincide with degenerate or singular fits. The scientifically admissible lesson is heterogeneity plus a need for diagnostics—not a portable ranking of “best GEO factors.”

Interpretation boundary · a fixed citation choice is not end-to-end visibility

The experiment does not call a live search engine. It models only the first URL that matches one of two injected sources, and drops roughly 3.4% of runs with no URL or an unmatched first URL. Model snapshots, API dates, temperatures, seeds, cell-level valid sample sizes, and uncertainty intervals are not locked. The paper reports no multiplicity correction for the 108 factor–model tests. The matrix therefore cannot establish crawl, indexing, retrieval, later-citation use, factual quality, user utility, or commercial outcomes. Values above 10,000 indicate near-separation in this finite design; they are not precise transferable effect sizes.

Expression-preserving edits

Reorganize or clarify already supported claims without changing their truth conditions.

Evidence-surfacing edits

Add traceable methods, data, qualifications, citations, or comparisons that were previously absent or inaccessible.

Information-expanding edits

Conduct new analysis, testing, or research to support a new claim. These require a workflow for producing and reviewing evidence, not only an editor.

Ecosystem edits

Correct databases, documentation, third-party profiles, or independent sources through their legitimate processes.

Prohibited edits

Fabricate statistics, imitate independent authority, hide sponsorship, inject instructions, create circular citation networks, or degrade competitors.

Factor (A vs B)	Diverse Providers			GPT Models		
	Gemini 2.5 Flash	Claude 3.5 Son-net	Kimi K2 Thinking	GPT 5 Nano	GPT 5 Mini	GPT 5.2
On-Topic vs Off-Topic	>10k	>10k	>10k	221 [†]	>10k	>10k
Query Terms vs Missing	9.41	17.0 [†]	16.4	5.99	14.4	40.0
Price vs No Price	>10k	>10k	36.1	7.82	6.26	30.4
Specs vs No Specs	8.63	>10k	238	11.5	15.4	243
With vs No Comparisons	7.45*	4.55*	4.72	2.14	1.78*	1.61
Confident vs Hedged	599	754	10.6	2.67*	5.44	4.75
Evidence vs No Evidence	8.00*	>10k	46.3	2.73	5.91	2.09
Consistent vs Contradictory	2.81	2.81	2.72	2.19	4.09	1.74
Neutral vs Promotional	1.45*	362*	2.02	1.31	1.81	2.08*
Structured vs Dense	1.68	1.03 [†]	0.79	0.90	0.78	1.25
Organized vs Scattered	2.21	3.87	2.49	1.19	1.13	1.57
Strong vs Weak Value Prop	5.22	5.66	7.24	2.79	2.68	1.53*
Deep vs Shallow Coverage	1,480 [†]	>10k	132	5.33	19.5	3.98
Strong vs Weak Social Proof	6.61*	>10k	25.4	3.13	4.78	2.14
Position 1 vs 2	>10k*	>10k	>10k	2,002	1,795	>10k
Recent vs Old Timestamp	>10k	>10k	68.7	14.4	1,494	>10k
No vs Old Timestamp	2.32	2.28	1.33*	1.31	1.55	1.48*
Recent vs No Timestamp	1.99	13.0	3.67	1.15	3.53	1.99

Bold = statistically significant ($p < 0.05$). OR > 1 favors variant A.

Convergence warnings: *Degenerate Hessian, [†]Singular fit.

Effect Size Categories: Very strong (OR > 100), Strong (OR 10–100), Moderate (OR 3–10), Weak (OR 1.5–3), Negligible (OR < 1.5)

Figure 23: Factor-specific first-citation odds ratios in a fixed two-source experiment. Reproduced from Vishwakarma et al. [49, Table 2], SIGIR '26, under CC BY 4.0. The crop omits the source heading; values, emphasis, and convergence warnings are unchanged. Each cell is a point odds ratio after exactly two documents have already been supplied.

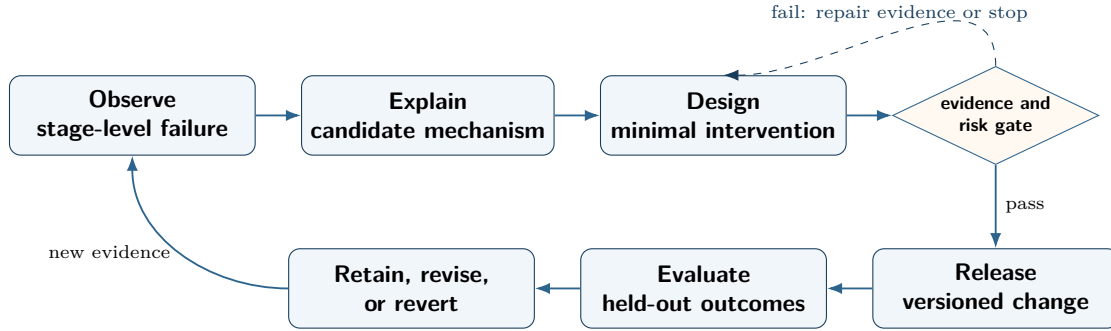


Figure 24: The evidence-first intervention lifecycle. Every change is versioned, evaluated on held-out outcomes, and reversible.

8.4.1 A constrained multi-objective formulation

Let x be the current source and x' a proposed revision. Define exposure $V(x')$, evidentiary coverage $E(x')$, factual fidelity $F(x', x)$, readability $H(x')$, and manipulation risk $R(x')$. A responsible optimization is

$$\max_{x' \in \mathcal{X}_{\text{permitted}}} \lambda_V V(x') + \lambda_E E(x') + \lambda_H H(x') \quad (55)$$

$$\text{subject to } F(x', x) \geq \tau_F, \quad (56)$$

$$R(x') \leq \tau_R, \quad (57)$$

$$\text{Approve}(x') = 1. \quad (58)$$

The approval constraint represents organizational accountability. An agent may propose a revision, score retrieval, and identify missing evidence; it does not authorize unsupported public claims.

Boundary and failure conditions

Optimization against a small prompt panel can overfit phrasing, entities, or judge preferences. Hold out query intents, audit claim drift, compare stage-level metrics, and test whether improvements persist after the source is placed back into a competitive corpus. A revision that wins only when its own page is preselected has not demonstrated end-to-end visibility.

8.5 The intervention lifecycle

1. **Observe.** Collect stage-specific evidence; do not begin with a preferred tactic.
2. **Explain.** State at least two plausible mechanisms and what would distinguish them.
3. **Design.** Choose the smallest permitted revision that tests the explanation.
4. **Gate.** Validate every changed claim, disclosure, license, accessibility requirement, and high-risk review.
5. **Release.** Record content hash, date, owner, target pages, and a rollback version.
6. **Evaluate.** Use a held-out query panel, guardrail metrics, and repeated observations.
7. **Decide.** Retain, revise, or revert according to the predeclared rule. Archive null and harmful outcomes.

Work on diagnosing and repairing citation failures is especially relevant to this lifecycle because it makes

diagnosis precede iterative revision [47]. Agentic optimization research can automate parts of the search over revisions, but its proposals still require the same evidence, holdout, and release gates.

8.6 External authority and circular evidence

First-party sources are often the correct authority for current specifications, official policy, and version history. Independent sources are often stronger for comparisons, reputation, and market claims. A robust evidence ecosystem uses both without pretending that duplicated first-party language has become independent confirmation.

Construct a directed provenance graph whose nodes are sources and whose edges represent material reliance. A cycle in which three articles all trace back to the same unverified press release has one origin, not three independent pieces of evidence. The public page should expose the primary origin, and the internal ledger should record the dependency.

Reproduction Studio 8 · Run a controlled evidence intervention

Input. A versioned evidence page, ten approved claims, a fixed corpus, and a held-out query set.

Task. Diagnose one retrieval or absorption failure. Produce two expression-preserving revisions and one evidence-surfacing revision. Verify that no claim scope changes. Compare the versions in fixed-context and end-to-end settings, including retrieval, citation, absorption, fidelity, readability, and risk outcomes.

Deliverable. CES diff, rendered-page diff, experiment trace, uncertainty table, failure analysis, and retain/revert decision.

Stop rule. Stop immediately if a version introduces an unsupported claim, hides a material condition, or cannot be rolled back reproducibly.

Chapter summary

1. A governed fact ledger precedes content optimization.
2. Evidence pages co-locate claims with units, scope, methods, limitations, and sources so they survive passage segmentation.
3. Structural elements are mechanisms to test, not a universal checklist.
4. Content optimization is multi-objective and constrained by fidelity, evidence, risk, and approval.
5. Every intervention is minimal, versioned, held-out, and reversible.

9 Optimization Algorithms and Agentic GEO

Key idea

Optimization is a constrained search over permitted source states, not an unbounded request to maximize mentions. The objective must retain factual equivalence, provenance, user utility, disclosure, and risk guardrails while remaining robust to unseen queries, competing sources, stochastic outputs, and platform drift.

9.1 Questions this chapter answers

By the end of the chapter, a reader should be able to answer five questions without confusing optimization with unrestricted persuasion:

- Q1.** Which source transformations are permitted, and which claims, evidence paths, approvals, and rollback records must accompany each one?
- Q2.** Which quantities are directly observable in a controlled optimizer, which platform mechanisms remain latent, and which conclusions depend on declared sampling or measurement assumptions?
- Q3.** How should hard fidelity, risk, accessibility, licensing, and cost constraints be separated from compensable performance objectives?
- Q4.** How can a finite candidate trace be validated, evaluated, reduced to a Pareto set, and selected without recycling the held-out test panel?
- Q5.** Which stopping and reproduction conditions justify *no change*, REVERT, or INCONCLUSIVE instead of a favorable release?

9.2 Define a permitted change space

Let x denote the released source state and $\mathcal{X}_{\text{perm}}(x)$ the set of candidate revisions allowed by claim, style, accessibility, licensing, security, and organizational policy. A proposed revision x' is admissible only if every changed proposition has an approved evidence path and all material qualifications remain locally recoverable.

Represent a revision as a typed operation rather than an opaque text diff:

$$\delta = \langle \text{target}, \text{operation}, \text{claim IDs}, \text{evidence IDs}, \text{scope}, \text{owner}, \text{rollback} \rangle. \quad (59)$$

Expression-preserving edits, evidence-surfacing edits, evidence-expanding edits, and ecosystem changes have different approval requirements. Adding a new claim is not merely a stronger rewrite.

9.3 Separate observables, latent variables, and assumptions

Observability is environment-specific. A reconstructed pipeline can expose a candidate's exact diff and every evaluation response; a closed answer surface may expose only a rendered result. The latter does not license reconstruction of hidden retrieval, ranking, or confidence states from the final answer.

The distinction determines the claim ceiling. A candidate's validation score is observable in the frozen testbed; its performance on an unseen platform version is not. A passing claim-drift classifier is also an instrument output, not proof of semantic equivalence. Human adjudication and independently replayable evidence paths remain part of the gate.

Table 29: Observability contract for a governed optimization run. “Observed” means captured by the stated protocol, not universally visible on a commercial platform.

Class	Examples in a controlled run	Required treatment
Observable	baseline and candidate hashes; typed diffs; claim and evidence IDs; gate outcomes; frozen query-cell IDs; raw responses; scores; timestamps; measured token, reviewer, and compute cost	preserve the row-level trace and the code, rubric, versions, seeds, and rejection reason that produced it
Latent	a closed platform’s eligible corpus, hidden candidates, ranking state, counterfactual response, future drift, and user utility outside the sampled task	name the missing quantity; report only a bounded proxy or redesign the environment rather than imputing a mechanism from the visible answer
Assumed	claim-equivalence rubric validity; query-panel relevance to the target population; judge calibration; within-epoch stability; no test-panel feedback; adequacy of the candidate and evaluation budgets	preregister the assumption, challenge it with sensitivity or independent review, and narrow the claim when the challenge fails

9.4 Use a vector objective and explicit constraints

For a query distribution Q , system distribution M , and revision x' , let

$$\mathbf{J}(x') = [J_{\text{ret}}, J_{\text{cite}}, J_{\text{abs}}, J_{\text{fid}}, J_{\text{user}}, -J_{\text{risk}}, -J_{\text{cost}}]. \quad (60)$$

A scalar reward $\lambda^\top \mathbf{J}$ is legitimate only after weights, scales, and non-compensable constraints have been declared. Severe unsupported claim risk should not be offset by a small visibility gain.

One constrained problem is

$$\max_{x' \in \mathcal{X}_{\text{perm}}(x)} \mathbb{E}_{q \sim Q, m \sim M} [J_{\text{primary}}(x'; q, m)] \quad (61)$$

$$\text{subject to } \text{ClaimDrift}(x', x) = 0, \quad (62)$$

$$\Delta J_{\text{fid}} \geq -\epsilon_F, \quad J_{\text{risk}} \leq \tau_R, \quad (63)$$

$$\text{EditCost}(x', x) \leq B, \quad \text{Approve}(x') = 1. \quad (64)$$

The thresholds are governance decisions. Report the Pareto frontier and guardrail failures rather than only the selected point.

Counterexample · a weighted score can prefer a forbidden candidate

Consider two didactic candidates evaluated on the same frozen panel. Candidate z_A has primary gain 0.07 and risk 0.04, while z_B has gain 0.18 and risk 0.24. The declared hard risk limit is $\tau_R = 0.10$, so z_B is inadmissible. Nevertheless, the apparently risk-adjusted score $s(z) = J_{\text{primary}}(z) - 0.25J_{\text{risk}}(z)$ gives

$$s(z_A) = 0.07 - 0.25(0.04) = 0.060, \quad s(z_B) = 0.18 - 0.25(0.24) = 0.120.$$

Maximizing the weighted score selects the risk-violating candidate. These numbers are synthetic and demonstrate a logical failure of compensation, not an effect on any platform. Apply non-compensable gates before scoring the survivors.

After hard constraints have been applied, positive weights can be used to choose a supported point on the

feasible Pareto frontier. They do not make the constraints unnecessary.

Proposition 9.1: A positive weighted-sum maximizer over a feasible set is Pareto efficient

Let $\mathcal{F}$ be a finite nonempty feasible candidate set and let $\mathbf{u}(z) \in \mathbb{R}^d$ orient every retained objective so that larger is better. If $w_j > 0$ for every j and $z^* \in \arg \max_{z \in \mathcal{F}} \mathbf{w}^\top \mathbf{u}(z)$, then no $z \in \mathcal{F}$ weakly improves every component of $\mathbf{u}(z^*)$ and strictly improves at least one component.

Proof. Suppose such a candidate z existed. Componentwise dominance gives $u_j(z) - u_j(z^*) \geq 0$ for every j , with a strict inequality for at least one component. Since every w_j is strictly positive, $\mathbf{w}^\top [\mathbf{u}(z) - \mathbf{u}(z^*)] > 0$. This contradicts the maximality of z^* over $\mathcal{F}$. $\square$

The converse need not hold: a non-convex frontier can contain efficient points that no fixed weighted sum selects. More importantly, the proposition begins *after* feasibility has been established; it supplies no empirical claim about the quality of a gate, metric, or candidate generator.

9.5 Separate development queries from the target population

Let Q_{dev} , Q_{val} , and Q_{test} be split by underlying intent clusters. Surface paraphrases of one fact need remain in the same split. A revision optimized on Q_{dev} is selected using Q_{val} and reported once on held-out Q_{test} . If test results cause another revision, that panel has become development data.

For per-query gain $g_q(x')$, define held-out regret relative to the best per-query candidate in the evaluated set:

$$\text{Regret}_Q(x') = \frac{1}{|Q|} \sum_{q \in Q} \left[\max_{z \in \mathcal{Z}} g_q(z) - g_q(x') \right]. \quad (65)$$

Mean gain alone can hide a revision that substantially harms a minority query class. Report lower-tail performance, subgroup results, and risk separately.

9.6 Resolve multi-query conflict before editing

Queries can request concise definitions, detailed comparisons, recent specifications, or skeptical counterevidence. Their preferred revisions may conflict under a fixed content budget. Multi-query GEO work explicitly studies conflict-aware instruction fusion [67]; output-control and feature-level approaches explore related optimization spaces [21, 26]. Each method remains bounded by its benchmark, models, objective, and permitted manipulation assumptions.

Worked example 9.2: A mean gain that fails a material query class

Revision A raises a declared score by 0.10 on eight descriptive queries and lowers it by 0.20 on two safety queries. The unweighted mean gain is positive (0.04), but the revision fails if safety queries carry a non-inferiority guardrail. The decision is determined by the predeclared constraint, not by retrospectively changing query weights.

9.7 Build an auditable proposal–evaluation loop

An agentic optimizer can propose edits, retrieve evidence, run controlled evaluations, and update a strategy library. It cannot approve unsupported claims or silently publish. The two-part procedure in algorithms 6 and 7 separates development search from frozen validation, one-shot testing, and the release decision.

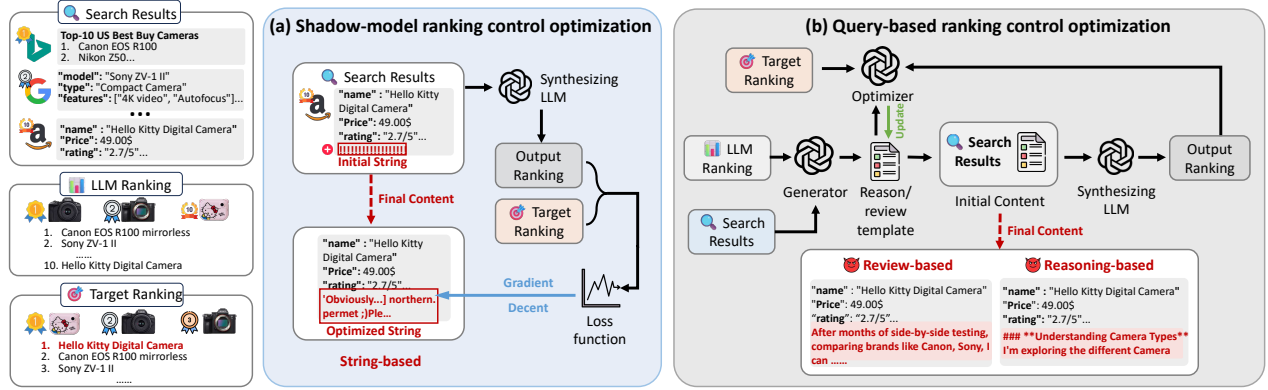


Figure 25: Two output-ranking control formulations: a differentiable shadow-model route and an iterative query-based generator–optimizer route. Their observable feedback and access assumptions are materially different.

Source note. Rectangular crop of Jin et al. [21, Fig. 2], physical PDF page 4, CC BY 4.0. The diagram concerns the paper’s product-ranking testbed; it does not authorize deceptive rewriting or establish transfer to a closed search platform.

Algorithm 6 Constrained optimization, Part I: development search and hard gates

Inputs: baseline x ; permitted typed operations Ω ; approved claim and evidence registry E ; intent-clustered development panel Q_{dev}
 locked system and judge configurations; hard gates $\mathcal{G}$; evaluation budget B ; minimum search improvement Δ ; patience p

Outputs: complete development trace $\mathcal{T}$ and frozen eligible set A

```

1: Freeze and hash  $x, E, \Omega, Q_{\text{dev}}$ , rubrics, models, judges, seeds, budget, gate thresholds, and rollback target
2:  $A \leftarrow \{x\}$ ;  $n \leftarrow 0$ ;  $s \leftarrow 0$ ;  $v_{\text{best}} \leftarrow 0$ ; halt  $\leftarrow$  false
3: while  $n < B \wedge s < p \wedge \neg \text{halt}$  do
4:   Diagnose the earliest observed failure on  $Q_{\text{dev}}$  and record at least one competing explanation
5:   Propose a finite batch using only  $Q_{\text{dev}}, E$ , and typed operations in  $\Omega$ ; assign immutable candidate IDs and hashes
6:   for all new candidates  $z$  in deterministic ID order do
7:     Materialize the typed diff, affected claim IDs, evidence spans, owner, estimated cost, and rollback record in  $\mathcal{T}$ 
8:     Validate schema, claim equivalence, evidence scope, disclosure, accessibility, licensing, security, and policy gates
  in  $\mathcal{G}$ 
9:     if a non-compensable gate fails then
10:       Log the gate, threshold, observed value, and rejection; do not score or repair  $z$  with validation/test feedback
11:       if the failure is a declared severe-risk event then
12:         halt  $\leftarrow$  true
13:       end if
14:     else
15:       Evaluate every planned  $Q_{\text{dev}}$  cell under the locked configurations; preserve raw outputs, missing states, guardrails, and cost
16:       Revalidate output-dependent gates; append  $z$  to  $A$  only if every gate passes; increment  $n$  by the charged evaluations
17:     end if
18:   end for
19:   Compute the best eligible development value using the frozen rule
20:   if its improvement over  $v_{\text{best}}$  is less than  $\Delta$  then
21:      $s \leftarrow s + 1$ 
22:   else
23:     update  $v_{\text{best}}$  and set  $s \leftarrow 0$ 
24:   end if
25:   Set halt  $\leftarrow$  true if the trace becomes unreconstructable or the next complete batch would exceed  $B$ 
26: end while
27: Freeze the development-Pareto shortlist from  $A$  and archive every rejected, null, harmful, unchanged, and unselected candidate in  $\mathcal{T}$ 

```

Algorithm 7 Constrained optimization, Part II: frozen selection and release

Inputs: baseline x ; frozen shortlist A and trace $\mathcal{T}$ from [algorithm 6](#); locked Q_{val} and Q_{test}
 hard gates $\mathcal{G}$; minimum worthwhile gain η ; objective orientation, dominance, tie, approval, and decision rules

Outputs: selected state $z^* \in \mathcal{X}_{\text{perm}}(x)$ or x ; RETAIN, REVERT, or INCONCLUSIVE decision

- 1: Prohibit further proposal or strategy updates from Q_{val} or Q_{test}
- 2: Evaluate A on Q_{val} ; reapply $\mathcal{G}$ and compute the validation Pareto set with the registered tie rule
- 3: Select z^* , taking $z^* = x$ if no survivor reaches η or accountable human review withholds approval
- 4: Evaluate z^* on Q_{test} exactly once; never generate a replacement candidate from this result
- 5: **if** a test guardrail, severe-risk, reconstruction, or approval condition fails **then**
- 6: Return the frozen rollback state and label REVERT or INCONCLUSIVE under the registered decision table
- 7: **else**
- 8: Release the approved hash with monitoring and rollback; label RETAIN
- 9: **end if**
- 10: Archive validation and test rows, approval, selected/reverted hash, monitoring rule, and exact stopping trigger in $\mathcal{T}$

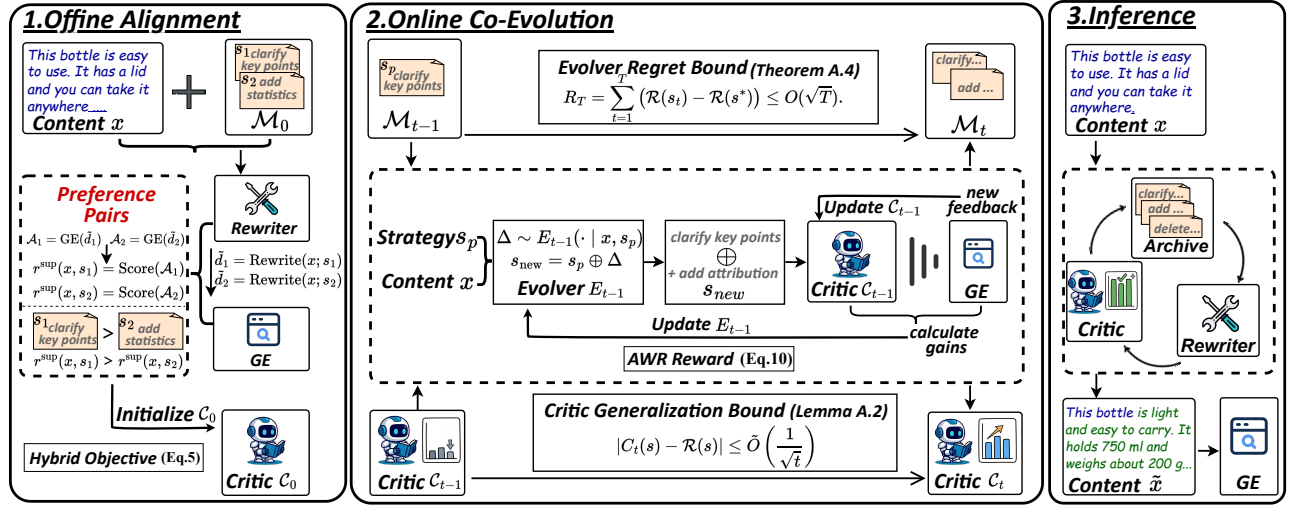


Figure 26: AgenticGEO separates offline critic alignment, online archive–critic co-evolution, and inference-time planning. The visual exposes where feedback, memory, and a surrogate score enter the loop.

Source note. Rectangular crop of Yuan et al. [63, Fig. 3], physical PDF page 4, CC BY 4.0. Each benchmark query uses a fixed five-document candidate set; the critic is a learned proxy, not evidence that a production engine self-evolves or validates factuality.

Budget exhaustion, p stale rounds, a severe-risk event, an unreconstructable trace, or insufficient validation gain are legitimate stopping outcomes. A stopped run is not automatically a failed run: when no admissible candidate beats the registered threshold, the scientifically supported action is no change. Conversely, a favorable test result does not authorize another search round; doing so converts the test panel into development data.

Agentic GEO and reusable strategy-learning studies motivate adaptive proposal and memory components [56, 63]. A deterministic or confidence-aware platform model can support controlled analysis [66]; it does not establish that a named closed platform uses the same confidence state or decision rule.

9.7.1 Worked synthetic trace: gates before the Pareto frontier

Consider a didactic documentation task for the synthetic Orchid Bay instrument. The baseline contains only a nominal range. Six typed revisions expose a definition, evidence comparison, qualification, unsupported superlative, dense evidence grid, or scope-first FAQ. The primary quantity in [table 30](#) is an absolute change in a declared,

Table 30: Complete synthetic candidate ledger. All values are invented for a recomputable teaching example. Lower L_F , risk, and cost are preferred.

ID	Typed change	Dev. gain	Val. gain	L_F	Risk	Cost	Gate/result
z_0	unchanged baseline	0.000	0.000	0.000	0.020	0	pass; no change
z_1	definition card with approved claim	0.060	0.045	0.010	0.060	8	pass; below η
z_2	two-column comparison with evidence anchors	0.085	0.070	0.010	0.050	10	pass; selected
z_3	concise qualification card	0.075	0.064	0.008	0.040	7	pass
z_4	unsupported “best” superlative	0.160	0.125	0.080	0.350	5	fail: drift, fidelity, risk
z_5	three-panel evidence grid	0.095	0.073	0.012	0.060	14	fail: cost
z_6	scope-first FAQ with qualification	0.068	0.062	0.004	0.025	9	pass

unitless evaluation proportion; for example, 0.070 means seven percentage points in this frozen synthetic panel. It is not a visibility estimate for a commercial system.

Before seeing outcomes, the controller registers

$$\text{ClaimDrift} = 0, \quad L_F \leq 0.020, \quad J_{\text{risk}} \leq 0.100, \quad C \leq 12, \quad \eta = 0.050, \quad (66)$$

where L_F is fidelity loss, C is a synthetic review-cost unit, and η is the minimum worthwhile validation gain. Accessibility and evidence-path checks are binary hard gates. The registered selection rule is: filter hard failures; remove dominated candidates; require gain at least η ; then take the largest validation gain. Values within 0.005 are tied and resolve first to lower risk and then to lower cost.

The hard gates first remove z_4 and z_5 , regardless of their larger validation gains. Thus

$$\mathcal{F} = \{z_0, z_1, z_2, z_3, z_6\}.$$

On the oriented vector $\mathbf{u} = (\text{validation gain}, -L_F, -\text{risk}, -\text{cost})$, z_3 dominates z_1 : $0.064 > 0.045$, $0.008 < 0.010$, $0.040 < 0.060$, and $7 < 8$. No remaining candidate improves every component of another, so the full four-objective Pareto set is

$$\mathcal{P} = \{z_0, z_2, z_3, z_6\}.$$

The minimum-gain rule excludes z_0 . Candidate z_2 has the largest surviving validation gain. Its margin over z_3 is $0.070 - 0.064 = 0.006$, which exceeds the registered 0.005 tie band, so z_2 is selected without invoking the risk or cost tiebreaker.

Only after this decision is frozen is z_2 run on eight held-out synthetic intent clusters. The gains are $(0.05, 0.04, 0.07, 0.06, 0.08, 0.03, 0.06, 0.05)$, whose arithmetic mean is $0.44/8 = 0.055$ and minimum is 0.03. The held-out trace records $L_F = 0.009$, risk 0.050, cost 10, zero claim drift, and passing binary gates. Under the didactic decision table this yields RETAIN. These values show how to replay a decision; they provide no confidence interval, causal effect, or claim about an external platform.

9.7.2 Read adaptive-system results as diagnostic ceilings

Table 31 separates evidence for adaptation from claims of autonomy, truthfulness, or production stability. The comparison is not a leaderboard because the studies use different data, candidate sets, models, objectives, and reporting completeness.

The operational conclusion is deliberately weaker than the system names: adaptive proposal, diagnosis, and memory can be useful components in a frozen testbed. They do not remove the need for a no-change action, an

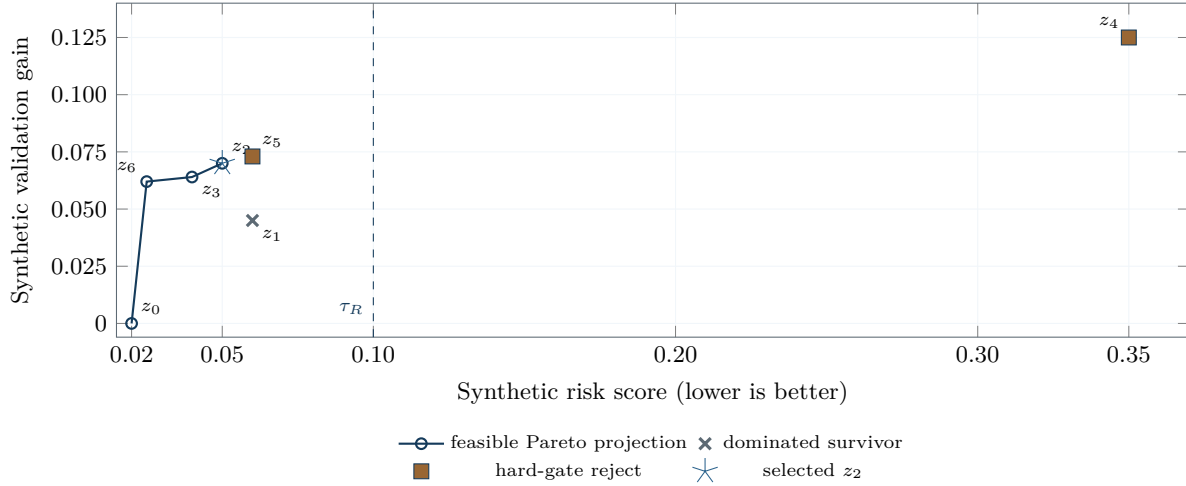


Figure 27: Risk–gain projection of the synthetic candidate trace. The purple polyline is the non-dominated envelope after all four objectives are considered; the projection alone is insufficient to recover the cost rejection of z_5 . The dashed line is the predeclared risk limit, not an estimated platform boundary. Every point is reproduced from table 30.

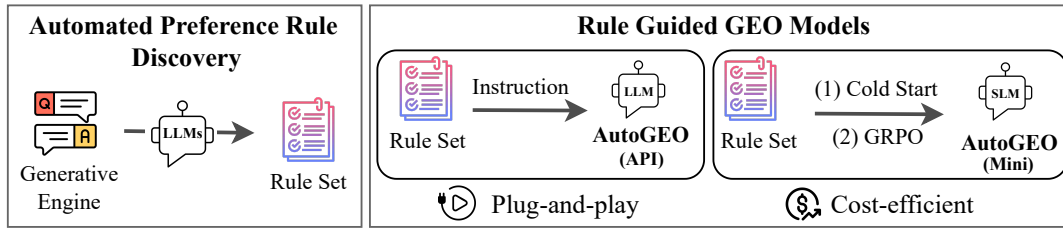


Figure 28: AutoGEO separates automatic preference-rule discovery from two rule-guided optimization routes, one prompt-based and one trained. The split is an implementation topology, not evidence that its learned rules are universal.

Source note. Rectangular crop of Wu et al. [57, Fig. 1], arXiv:2510.11438v1 physical PDF page 3, CC BY 4.0. Results remain bounded to the paper’s datasets, models, visibility objective, and quality checks.

unreparable failure class, held-out selection, factual review, cost accounting, or an accountable release decision.

9.8 Audit the optimizer, not only its best candidate

The optimization trace records every candidate, prompt or policy, evidence set, rejection reason, evaluation cells, scores, costs, and selected release. This prevents survivorship bias in which only the best revision remains visible. Report:

- number of proposals, evaluations, retries, and total compute/API cost;
- false acceptance and false rejection of claim-drift controls;
- development–validation–test gaps and subgroup failures;
- sensitivity to judge, generator, retriever, seed, and context budget;
- unchanged, null, harmful, and rolled-back candidates;
- human overrides with role, rationale, and timestamp.

9.9 Reproduction checkpoint: replay the losing candidates

Table 31: What three adaptive GEO systems establish—and do not establish.

System	Bounded result	Failure evidence and claim ceiling
AgentGEO [47]	In its reconstructed MIMIQ setting, in-context citation rate changes from 56.58 (vanilla) and 68.80 (AutoGEO) to 79.52; OOD and HTML cells are 70.62 vs. 64.18 and 67.19 vs. 45.96.	After training, 163 queries remain uncited; in one GPT setting recall is 67.10 vs. 68.99 and clarity 62.88 vs. 66.08 for AutoGEO. Some authority failures cannot be repaired by rewriting, and no live platform or human quality study is provided.
AgenticGEO [63]	On GEO-Bench, reported visibility is 25.48 vs. 23.71 for AutoGEO with Qwen2.5-32B and 24.52 vs. 22.78 with Llama-3.3-70B; removing the archive reduces three reported cells to 20.40/21.65/20.08.	The larger quoted percentage is relative to no optimization, not the strongest baseline. BERTScore is not factual verification, and generated reference-like text exposes a claim-integrity failure mode.
MAGEO [56]	The paper reports word-level visibility 4.52 for its full GPT configuration versus 1.33 for the strongest displayed heuristic, and 5.30 in the Gemini cell; a 100-example human comparison reports rank correlations.	Core benchmark size/splits and run uncertainty are missing, the composite weight is selected on the test set, and the skill bank later declines. The result is retained as an exploratory system report, not a scalable, trustworthy, or state-of-the-art claim.

Table 32: Optimization objects that must remain separately versioned.

Object	Required identity	Leakage or governance risk
Claim set	claim IDs, scope, evidence, freshness, approvals	model invents or broadens a proposition
Query panels	intent clusters, split, weights, dates	test-panel reuse and favorable-query selection
Candidate edit	typed diff, author/agent, prompt/policy, hash	only winning candidates retained
Evaluation stack	corpus, models, versions, judges, seeds	score drift mistaken for content improvement
Strategy memory	source runs, validity window, exclusions	benchmark-specific tactic treated as doctrine
Release	approvals, target, time, monitoring, rollback	experimental candidate published without accountability

Boundary and failure conditions

An optimizer can discover a reliable intervention in a frozen benchmark and still fail on a live surface. It can also optimize a proxy while reducing truthfulness, source diversity, accessibility, or user welfare. Never describe append-only steering, instruction injection, synthetic consensus, or concealed affiliation as ordinary evidence engineering merely because a benchmark score improves.

9.9 Reproduction checkpoint: replay the losing candidates

A released page and final score cannot show whether the optimizer exceeded its permitted space, leaked test feedback, or discarded unfavorable trials. A checkpoint passes only when an independent reader can rebuild the candidates, rerun every gate and metric, and recover the frozen no-change or release decision.

For the worked example, the minimal arithmetic check recovers $\mathcal{F} = \{z_0, z_1, z_2, z_3, z_6\}$, removes z_1 by dominance, selects z_2 because the 0.006 margin exceeds the 0.005 tie band, and obtains the held-out mean 0.055. Failure to reproduce any of these transitions is a failed checkpoint, not permission to report only the favorable endpoint.

Reproduction Studio 9 · Compare constrained optimization policies

Question. Which policy improves a held-out outcome under strict claim, fidelity, risk, and cost gates?

Input. Ten approved claims, three equivalent source states, 60 intent-clustered queries split 30/15/15, a frozen reconstructed pipeline, and a budget of 40 candidate evaluations.

Procedure. Compare a human-authored baseline, random typed edits, one black-box search policy, and one

Table 33: Minimum checkpoint for reproducing a constrained optimization run.

Artifact	Required contents	Pass condition
Baseline and permission manifest	rendered source, source hash, typed operation schema, claim/evidence registry, licenses, owners, approvals, and rollback hash	every candidate diff resolves to permitted fields and approved evidence
Panel and environment manifest	intent-cluster IDs and split map, query weights, corpus/system/judge versions, prompts, parsers, seeds, timestamps, and missing states	a clean run materializes the same planned cells without moving a paraphrase or failed row across splits
Complete candidate ledger	proposal policy and state, immutable IDs, hashes, typed diffs, raw outputs, retries, costs, gate values, and rejection reasons	proposal counts reconcile with evaluated, rejected, eligible, and unselected counts; no candidate is represented only by an aggregate
Decision implementation	hard-gate code and tests, objective orientation, thresholds, dominance and tie rules, minimum gain, budget, patience, and exact stop trigger	an independent implementation recovers the same feasible set, Pareto set, selected state, and stopping reason
Held-out and release package	one-shot test trace, human approval or override, decision label, released hash, monitoring window, incident threshold, and rollback event	the released or reverted state matches the decision table and no post-test candidate appears in the run

proposal-critic agent. Preserve all rejected candidates. Apply [algorithms 6 and 7](#); keep validation and test panels inaccessible to proposal policies; inject one unsupported superlative, one accessibility failure, one over-budget candidate, and one apparent mean gain that harms a protected query class. Test one judge swap and one corpus shift.

Output. Permission manifest, candidate funnel, complete ledger, development and validation Pareto plots, held-out effect with a cluster-respecting interval, claim-drift confusion matrix, accessibility and licensing gate report, cost ledger, null/harm table, exact stop event, and a signed RETAIN, REVERT, or INCONCLUSIVE decision.

Reference criterion. The selected release can be reconstructed from approved claims and the complete trace without access to hidden conversational state.

Independent check. A second implementation must reconstruct every candidate denominator, recover all hard-gate rejections, reproduce the Pareto set and tie rule, verify that no test query affected a proposal, and obtain the same decision. It must also reproduce the arithmetic in [table 30](#) and [figure 27](#).

Stop rule. Stop at the first registered trigger: budget exhaustion, p consecutive development rounds below Δ , severe risk, trace failure, no validation candidate above η , withheld approval, or held-out guardrail failure. Archive the no-change or rollback outcome; never lower a threshold or reuse the test panel merely to obtain a winner.

Chapter summary

1. Optimization searches a versioned, governed change space. A typed diff binds each changed proposition to evidence, scope, ownership, approval, and rollback; adding a claim is not an expression-preserving rewrite.
2. Candidate hashes, diffs, evaluation rows, and gate outputs may be observed in a controlled reconstruction. Hidden platform candidates, counterfactual responses, future drift, and population utility remain latent unless a design identifies them.
3. Fidelity, claim drift, disclosure, accessibility, licensing, security, risk, and budget can be non-compensable constraints. A weighted score can prefer an inadmissible candidate and therefore cannot replace hard gates.
4. Positive weighted sums select Pareto-efficient points only within the already feasible set and under their stated objectives. They neither validate a metric nor establish empirical superiority.
5. Development, validation, and test panels are split by intent cluster. Proposal policies see development data; selection uses a frozen validation rule; the chosen state receives one held-out test with no feedback search.
6. Mean gain can hide material subgroup harm. Lower-tail results, non-inferiority guardrails, missing states, uncertainty at the assignment unit, and cost remain visible alongside the primary objective.
7. In the fully synthetic trace, hard gates remove the high-scoring z_4 and over-budget z_5 ; dominance removes z_1 ; the registered rule selects z_2 . The arithmetic is pedagogical, not a platform-effect claim.
8. Agentic components may diagnose, propose, retrieve, and evaluate in a frozen testbed. Accountable humans approve publication, and unrepairable or no-change states are first-class outcomes.
9. The scientific object is the complete trace—including rejected, null, harmful, unchanged, and rolled-back candidates—plus the exact stop event. A result is reproducible only when an independent implementation recovers the feasible set, Pareto set, selected hash, and decision.

10 Multimodal, Platform, Brand, and Evidence Ecosystems

Key idea

Images, captions, pages, entities, platforms, languages, and source networks change the accessible evidence and the unit of intervention. Results transfer only when the representation, retrieval surface, population, geography, date, and dependency structure are sufficiently comparable.

10.1 Define the representation path for every modality

For source asset a , let

$$\Pi(a) = \{\pi_{\text{text}}(a), \pi_{\text{image}}(a), \pi_{\text{caption}}(a), \pi_{\text{metadata}}(a), \pi_{\text{page}}(a)\} \quad (67)$$

denote the representations available to a controlled system. A production surface may expose only a subset, transform them asynchronously, or create learned embeddings that cannot be inspected. Experiments specify which representation is indexed, queried, reranked, packed, and shown to the model.

Image pixels, alt text, a nearby caption, structured metadata, and the surrounding collection page are distinct treatments. Changing several at once identifies a bundle. Caption-based optimization research and production-scale visual-platform work motivate studying these paths [7, 64]; their reported mechanisms and business outcomes remain specific to the audited versions, deployments, and methods.

Figure 30 makes the governed object path explicit before any cross-modal score is computed.

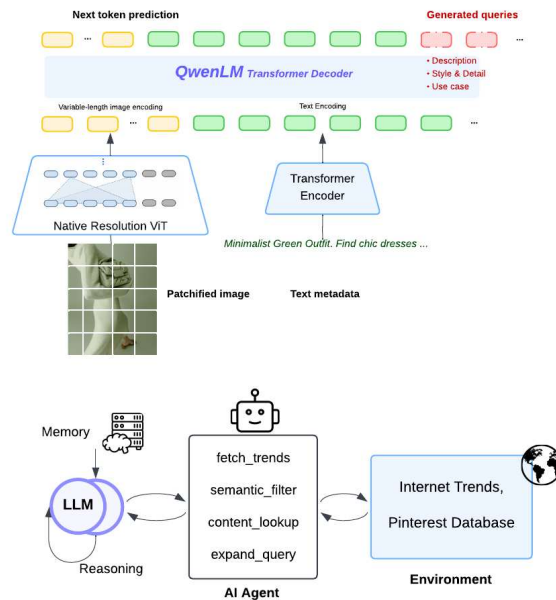


Figure 29: A visual-content case study joins a VLM query generator with an agent that expands candidate intents from an external trend environment. It makes the image, text-metadata, agent, memory, and environment paths separately visible.

Source note. Rectangular crop of Zhang et al. [64, Fig. 2], physical PDF page 3, CC BY 4.0. The diagram documents one industrial content-acquisition architecture; it does not prove causal business growth, citation gain, or transfer beyond the disclosed deployment.

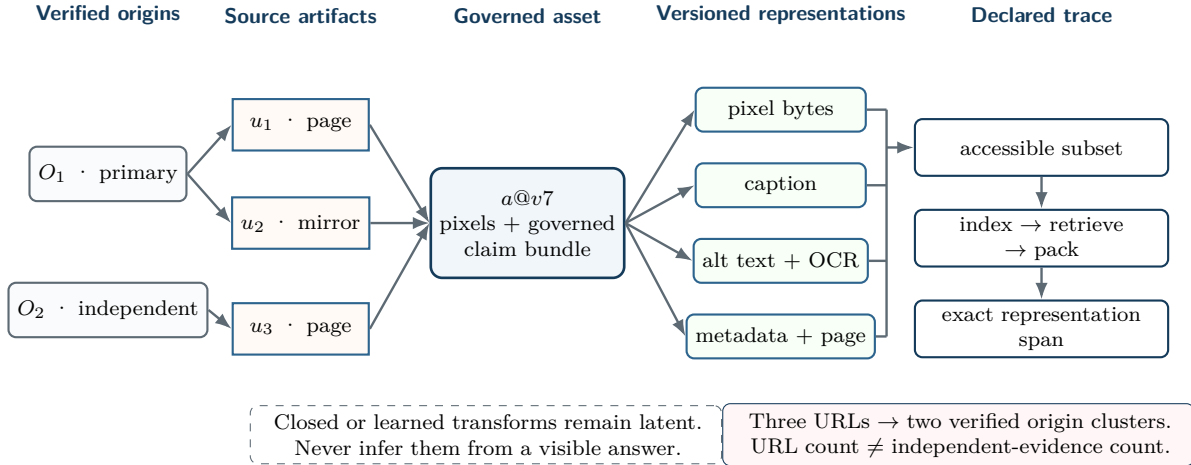


Figure 30: Asset, representation, and provenance dependencies in a controlled multimodal trace. Three source artifacts resolve to two origin clusters and one governed asset version. Only a declared accessible subset of the versioned representations proceeds into an inspectable index–retrieve–pack trace. The diagram is an original research abstraction, not a disclosure of a closed platform’s internal architecture.

10.2 Measure cross-modal agreement and contribution

Let c be an approved claim about an asset. Define support labels $S_m(c, a)$ for modality or representation m . A cross-modal conflict occurs when one accessible representation supports c while another materially contests or narrows it. A system may retrieve the image using a caption but generate from OCR text, or rank a product by a joint embedding while citing the collection page.

A multimodal audit therefore includes:

- representation availability and hashes for each modality;
- OCR, captioner, embedding, and parser versions;
- image–text relevance and claim–support labels;
- modality ablations and equal-budget controls;
- attribution from answer claims to exact representation spans;
- accessibility, provenance, disclosure, and manipulation tests.

Adversarial multimodal GEO research shows that coordinated image and text perturbations can alter ranking in evaluated VLM-ranker settings [13]. That evidence establishes a bounded threat class, not a universal success rate, invisibility guarantee, or permission to reproduce attack artifacts against a live service.

Table 34: Multimodal and production evidence separated by representation and outcome.

Evidence object	Reported observation	Required boundary
Joint text-image ranking attack [13]	On one Qwen2.5-VL-7B static-list setting, mean rank change is -0.73 for text, -1.30 for image, and -2.25 for the joint method; an OpenEvents transfer cell reports -2.71 vs. -1.84 for text.	Weaker regularization produces more visible distortion. One open VLM, small static lists, and offline data do not establish stealth or current commercial platform vulnerability.
Caption proxy [7]	In the reported multimodal proxy, traditional, fluency, quotation, statistics, and caption-injection changes are -0.66 , $+0.71$, -0.08 , -2.12 , $+1.12$; multi-shot injection is approximately $+0.01$.	The target VLM does not directly consume the image; the experiment inserts a caption into context. All single-modal means are negative, so neither a more-captions rule nor a direct-image-attack conclusion follows.
Pinterest content pipeline [64]	Among 140 human-reviewed items, relevance is 4.47 vs. 3.28 while diversity is 3.43 vs. 3.54 ; reported login rate is $.361$ vs. $.365$, alongside positive use-case, intent, signup, and CTR cells.	Online randomization unit, sample size, intervals, and several denominators are not reported; some labels and evaluations use the same model family. This is an industrial architecture case, not a causal business-growth result or a generative-citation study.

10.3 Treat platforms as versioned surfaces

A platform comparison cell is

$$e = \langle p, s, v, q, \ell, g, a, t, r, \omega \rangle, \quad (68)$$

where p is provider or system, s the surface, v the observable version, q the query, ℓ language/locale, g geography, a account state, t time, r repeat, and ω search/tool state. Missing version information is recorded as missing, not guessed from a model family name.

Different surfaces can use different accessible worlds, query plans, retrieval partners, citation resolvers, safety policies, and interface affordances. “Platform effect” is therefore a descriptive contrast until sampling, measurement equivalence, and confounding are addressed.

An API alias is not necessarily a version. Perplexity’s Agent API documentation, for example, states that a named preset is unversioned and may resolve to updated models, tools, prompts, and parameters; copying the documented explicit fields creates a frozen caller-visible configuration instead [36]. This supports a narrow experimental control, not a claim that the provider’s index, serving stack, model implementation behind an identifier, or other unexposed state has been frozen. A defensible comparison therefore records the full request and response metadata, pairs dynamic and explicit cells at the same time, repeats both, and still labels backend state as unobserved.

Table 35: Minimum disclosure for platform or brand visibility comparisons.

Dimension	Required disclosure	Common invalid shortcut
Demand	query-frame, sampling frame, strata, weights, paraphrase clusters	convenient favorable prompt list
System	provider, surface, date, observable version, search state	one model label for several products
User state	language, locale, geography, account/history controls	assuming identical personalization
Entity	alias resolution, category, maturity, region, source footprint	raw string mention as brand knowledge
Outcome	unit, denominator, missing policy, repeated uncertainty	single screenshot or vendor composite score
Transfer	comparable mechanisms and representations	extrapolating one platform/date to the market

10.3.1 Worked paired representation audit

The disclosure contract becomes operational only when it produces an inspectable comparison. Consider a declared synthetic audit of two open retrieval surfaces, two locales, and two query intents. For each planned pair j , arm $z = 0$ withholds a governed caption from the indexed representation and arm $z = 1$ exposes that same caption. Pixels, alt text, metadata, rendered page, query, locale, account state, search configuration, time block, and all non-treatment hashes remain fixed. The visible page does not change.

Let $Y_j^{(z)} = 1$ when the frozen top- k trace recovers all material claim qualifications required by the registered evidence-recall rubric, and let

$$d_j = Y_j^{(1)} - Y_j^{(0)}, \quad \bar{d}_{s\ell} = \frac{1}{|\mathcal{P}_{s\ell}|} \sum_{j \in \mathcal{P}_{s\ell}} d_j, \quad (69)$$

where $\mathcal{P}_{s\ell}$ contains only complete, admissible pairs for surface s and locale ℓ . A missing or unauditable arm remains missing; it is not recoded as a failure or a zero difference. The endpoint is evidence recall, not answer contribution, displayed citation, recommendation, conversion, or a business outcome.

Table 36: Eight planned pairs in the synthetic representation audit. The last pair preserves a parser failure as missing rather than converting it to zero.

Pair	Surface	Locale	Intent	Caption off	Caption on	d_j	Status
OB-01	Open-R	en-HK	specification	0	1	1	complete
OB-02	Open-R	en-HK	support	1	1	0	complete
OB-03	Open-R	zh-HK	specification	0	1	1	complete
OB-04	Open-R	zh-HK	support	1	1	0	complete
OB-05	Open-S	en-HK	specification	1	1	0	complete
OB-06	Open-S	en-HK	support	0	1	1	complete
OB-07	Open-S	zh-HK	specification	0	0	0	complete
OB-08	Open-S	zh-HK	support	1	NA	NA	missing parser

The seven complete pairs have $\sum_j d_j = 3$. The four descriptive stratum means are

$$\bar{d}_{R,en} = .500, \quad \bar{d}_{R,zh} = .500, \quad \bar{d}_{S,en} = .500, \quad \bar{d}_{S,zh} = .000.$$

The final mean uses one complete pair, not two. Pooling the complete pairs gives $3/7 \approx .429$, but neither this number nor the stratum means estimate a live platform effect. The small table is an arithmetic fixture designed to expose pairing, denominator, and missingness errors.

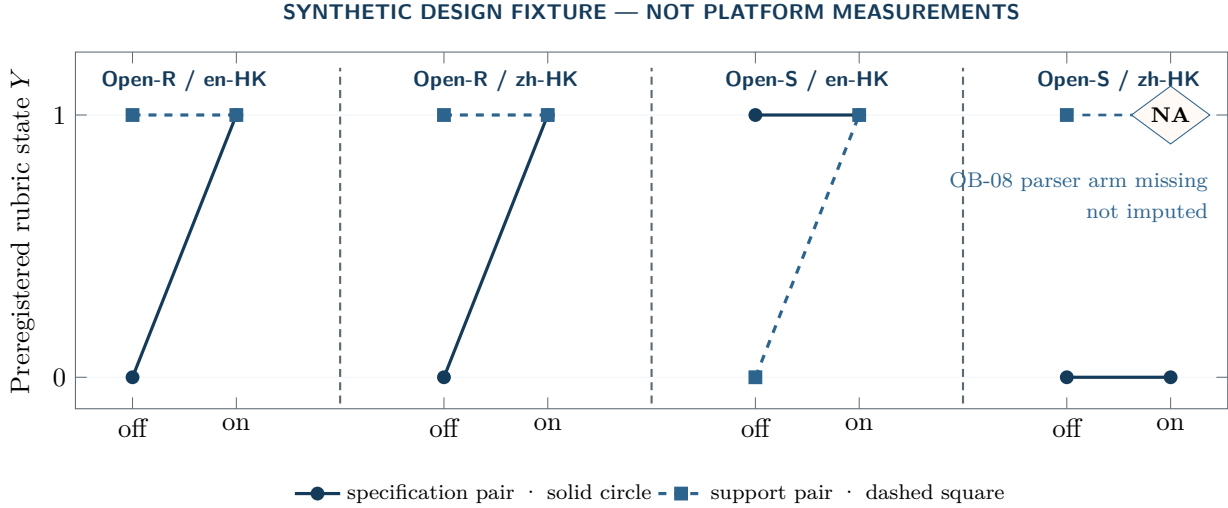


Figure 31: Paired caption-off/on outcomes reconstructed without aggregation from Table 36. Each panel retains its specification and support pair; solid circles and dashed squares remain distinguishable without color. The Open-S/zh-HK support pair ends in an explicit missing-parser state, not a zero. All surface names, observations, and outcomes are synthetic; the figure is not an empirical cross-platform comparison.

Algorithm 8 Trace-preserving paired representation audit

Inputs: Governed asset bundle and caption; frozen surfaces $\mathcal{S}$, locales $\mathcal{L}$, intents $\mathcal{I}$, repeats, outcome rubric, pairing rule, missingness rule, and counterbalancing seed

Outputs: Immutable pair ledger, stratum contrasts, missingness report, and bounded interpretation

- 1: Hash the asset, caption, non-caption representations, queries, renderer, index configuration, annotation guide, and execution environment
 - 2: Construct $z = 0$ by withholding only the caption representation and $z = 1$ by exposing that same caption
 - 3: **for all** $s \in \mathcal{S}$, $\ell \in \mathcal{L}$, and registered intent-repeat pairs **do**
 - 4: Verify that every non-treatment field and hash agrees across arms
 - 5: Execute the arms in seeded counterbalanced order and preserve request, candidate, context, response, error, and version traces
 - 6: **if** one arm is missing or unauditable, or any non-treatment field differs **then**
 - 7: retain the planned pair with a reason code; do not impute Y or d
 - 8: **else**
 - 9: apply the frozen rubric blind to z and append $Y^{(0)}$, $Y^{(1)}$, and $d = Y^{(1)} - Y^{(0)}$
 - 10: **end if**
 - 11: **end for**
 - 12: Compute each $\bar{d}_{s\ell}$ on complete admissible pairs and report planned, complete, missing, and deviating pair counts separately
 - 13: Export raw traces, hashes, deviations, arithmetic reconstruction, figure source, and the evidence boundary
-

Evidence boundary · paired arithmetic is not a platform effect

This module contains no live-platform observation. Open-R and Open-S are synthetic protocol fixtures, and every row in Table 36 is invented. In a future execution, a within-surface contrast supports a caption-representation interpretation only if caption exposure is the sole changed channel, pairing remains intact, and time, order, measurement, and missingness follow the registered protocol. A cross-surface difference can still mix indexes, retrievers, candidate sets, parsers, ranking policies, and other mechanisms. The endpoint cannot establish generated-answer use, citation, brand value, training-data presence, or hidden closed-system behavior.

10.4 Separate brand resolution from brand value

A brand mention requires entity resolution, including aliases, products, subsidiaries, transliterations, and homonyms. Let $B(y)$ be resolved entity mentions and let $C_b(y)$ be claims attached to brand b . Exposure metrics do not establish correctness, sentiment, prominence, recommendation, referral, or commercial value.

At minimum, report the vector

$$\mathbf{V}_b = [p_{\text{mention}}, p_{\text{correct}}, p_{\text{cite}}, p_{\text{abs}}, p_{\text{prominent}}, p_{\text{action}}, R_b], \quad (70)$$

where R_b is a separately defined risk profile. Do not average the components unless weights and scales serve a declared decision. A high mention rate with incorrect attributes can be harmful.

Figure 32 inserts a fail-closed entity gate before any brand-level outcome is computed.

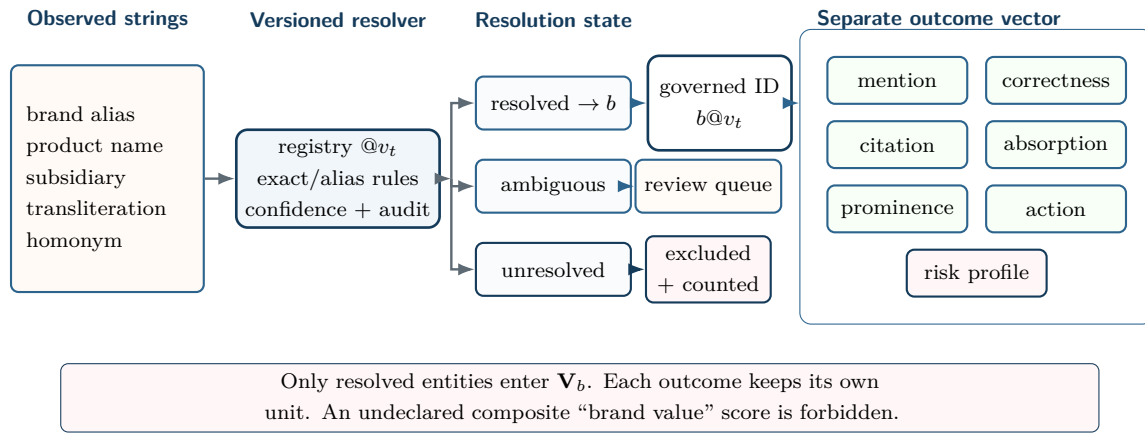


Figure 32: Entity resolution is a measurement gate, not a cosmetic cleanup. Aliases, products, subsidiaries, transliterations, and homonyms pass through a versioned resolver; ambiguous and unresolved strings remain explicit error states. Only a resolved governed entity enters seven separately reported outcomes. The diagram is an original protocol design and contains no observed brand or platform data.

Cross-cultural brand research and multi-platform visibility studies are routed to this chapter [16, 25]. Their sample construction, entity sets, languages, system dates, and causal language require full-text evidence cards before exact findings enter stable doctrine.

Worked example 10.1: Translation is not a controlled locale experiment

An English procurement prompt is translated into Chinese and issued to a different product surface under a different account and geography. A brand appears only in the second answer. The contrast mixes language, locale, surface, account, time, and potentially retrieval availability. It is a useful observation and hypothesis generator, not an identified language effect.

10.5 Model evidence as a dependency ecosystem

Let $G = (V, E)$ be a provenance graph whose nodes are claims, source artifacts, publishers, and versions. Directed edge $u \rightarrow v$ means v materially derives from u . Several URLs that reproduce one press release form a dependency cluster; their agreement does not provide independent confirmation.

For source set S , let $o(s)$ map each source to its earliest verified origin cluster. An effective-origin count is

$$N_{\text{origin}}(S) = |\{o(s) : s \in S\}|. \quad (71)$$

This count is descriptive and depends on the completeness of provenance tracing. It should accompany, not replace, source-quality and claim-support assessment.

Trajectory-aware ecosystem research treats discovery and evidence acquisition as a sequence rather than a flat list [60]. That framing is useful for experiments that log search actions and provenance; it must not be presented as the internal trajectory of a closed system without a trace.

10.6 Evaluate production outcomes with causal restraint

Traffic, active users, qualified actions, and revenue are downstream outcomes. A deployment report can document scale and operational feasibility, but causal attribution requires a credible counterfactual, stable instrumentation, and a record of concurrent product and market changes. For any production number, retain the population, denominator, window, baseline, treatment uptake, uncertainty, source identity, and whether the result was independently audited.

Vendor reports and practitioner studies can suggest hypotheses, denominators, or monitoring designs. They cannot become population-level cross-platform laws when their sample frame, parser, missingness, or causal design is proprietary or incomplete.

Boundary and failure conditions

Do not reuse a paper figure merely because its PDF is available. Record the exact version, page, figure number, data provenance, and license. Prefer an original mechanism diagram or a replot from licensed data and code. Never infer training-corpus presence, platform architecture, or cultural causality solely from a brand's observed mention rate.

Reproduction Studio 10 · Audit a multimodal, multilingual evidence ecosystem

Question. Which representation and provenance path explains a bounded visibility difference across two controlled surfaces?

Input. A licensed set of 100 image–page pairs, verified captions and claims, two languages, two open retrieval configurations, origin clusters, and one synthetic brand/entity alias set.

Procedure. Ablate pixels, captions, alt text, metadata, and collection context; hold query intents and account state fixed; test cross-modal conflict; trace duplicate origins; and repeat all cells. Include one adversarial caption only in the isolated safety sandbox.

Output. Retrieval and evidence-recall matrix, modality-ablation plot, entity-resolution errors, provenance graph, cluster-aware intervals, accessibility audit, and bounded transfer statement.

Reference criterion. Every result resolves to exact asset and representation hashes, parser/model versions, query cell, provenance origin, and license record.

Stop rule. Stop external comparison if measurement equivalence fails, provenance or license is unresolved, or a treatment cannot be isolated from language, locale, platform, or account state.

Chapter summary

1. Pixels, captions, metadata, and pages are distinct representations and treatments.
2. Multimodal studies require support checks, ablations, provenance, accessibility, and attack resistance.
3. Platform comparisons are versioned observations, not automatic mechanism comparisons.
4. Mention, correctness, citation, absorption, prominence, action, and risk are separate outcomes.
5. Source counts must cluster duplicated and provenance-dependent origins.
6. Production and cross-cultural findings retain causal and transfer limits.

11 Manipulation, Detection, Defense, and Mechanism Design

Key idea

The same openness that lets a source correct missing evidence can let an attacker steer retrieval, ranking, generation, or attribution. GEO is therefore both an optimization problem and a security-and-governance problem. A mature program specifies who may change what, how a system distinguishes evidence improvement from manipulation, and how harms are detected and reversed.

11.1 Threat modeling the source-to-answer path

A threat model names assets, actors, capabilities, trust boundaries, and harms. For generative search, protected assets include answer fidelity, source integrity, fair exposure, attribution, user autonomy, privacy, and the ability to correct errors. Relevant actors include honest creators, aggressive optimizers, competitors, affiliates, content farms, compromised publishers, platforms, and users with conflicting goals.

Let the attacker choose a source-side perturbation δ under budget B to maximize a target outcome while avoiding detector D :

$$\max_{\delta} \mathbb{E}_{q \sim Q} [Y_t(\mathcal{G}(q; \mathcal{W} \oplus \delta))] \quad (72)$$

$$\text{subject to } c(\delta) \leq B, \quad D(\delta) < \tau_D. \quad (73)$$

The target Y_t may be increased citation of the attacker’s source, decreased visibility of a competitor, inclusion of a false claim, or a downstream recommendation. A defense that only recognizes one known string can be adapted around; a defense that suppresses all optimized content can punish legitimate clarification.

11.2 Attack surfaces

Adversarial SEO research demonstrates that LLM-mediated search and ranking can be manipulated under experimental conditions and studies both promotion and competitor-oriented threats [34]. Other work examines the dynamics of such attacks [15], multimodal rank manipulation [13], and citation vulnerabilities [30]. These studies establish a research agenda and specific attack evidence; they do not imply that every production system shares the same vulnerability or success rate.

Table 37: Representative attack surfaces and required defenses.

Surface	Example capability or failure	Defense direction
Acquisition	cloaked content, compromised page, hidden text, rapid domain churn	fetch/render comparison, provenance, reputation and ownership change monitoring
Retrieval	keyword or embedding manipulation, corpus poisoning, synthetic consensus	robust retrieval, source diversity, duplicate and dependency analysis
Reranking	preference manipulation and competitor displacement	adversarial evaluation, counterfactual rank tests, utility guardrails
Context	instruction-bearing passages, context flooding, strategic redundancy	data/instruction separation, passage isolation, budget and diversity controls
Generation	persuasive unsupported claims, recommendation steering	evidence-constrained generation, claim verification, uncertainty and refusal policies
Attribution	citation laundering, wrong-source attachment, source substitution	span-level entailment, URL resolution, provenance and completeness checks
Multimodal	captions or images crafted to influence vision-language rankers	modality-specific parsing, cross-modal consistency, image provenance
Measurement	cherry-picked prompts, repeated-until-success screenshots, judge gaming	preregistration, held-out panels, raw-cell audit, human validation

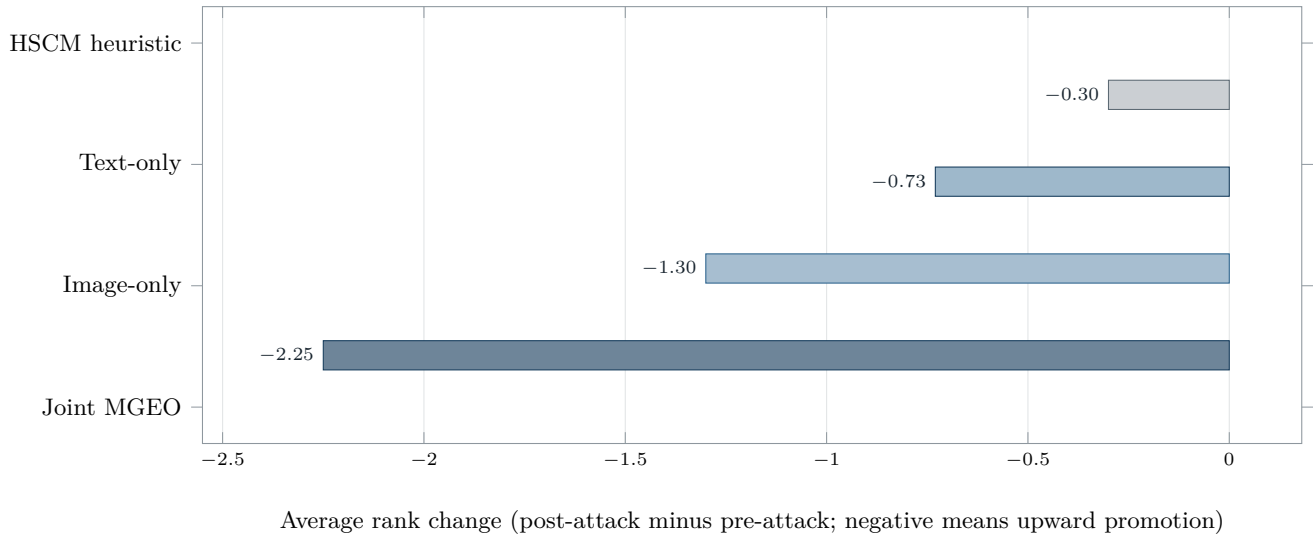


Figure 33: Original redraw from Du et al. [13, Table 1], *Multimodal Generative Engine Optimization*, KnowFM 2026, physical PDF page 6 / proceedings page 120, under CC BY 4.0. Values are the reported average target-rank change for Qwen2.5-VL-7B: HSCM -0.30 , text-only -0.73 , image-only -1.30 , and joint MGEO -2.25 . The redraw changes presentation, not values. More-negative values mean greater upward promotion by an attack, not improved information quality.

Evidence boundary · a larger attack point estimate is not a production effect

The source study uses Amazon-derived items in ten categories, ranks ten fixed candidates at a time, holds competitors fixed, and attacks one target with white-box gradients in a static offline Qwen2.5-VL-7B setting. Table 1 reports no interval, significance test, or repeated-seed uncertainty. The joint point estimate is more negative than the two unimodal point estimates and their arithmetic sum, but that pattern does not establish statistical synergy, transfer to a closed reranker, attack prevalence, persistence, user harm, or a defense effect. The paper’s product and generated images are not reproduced.

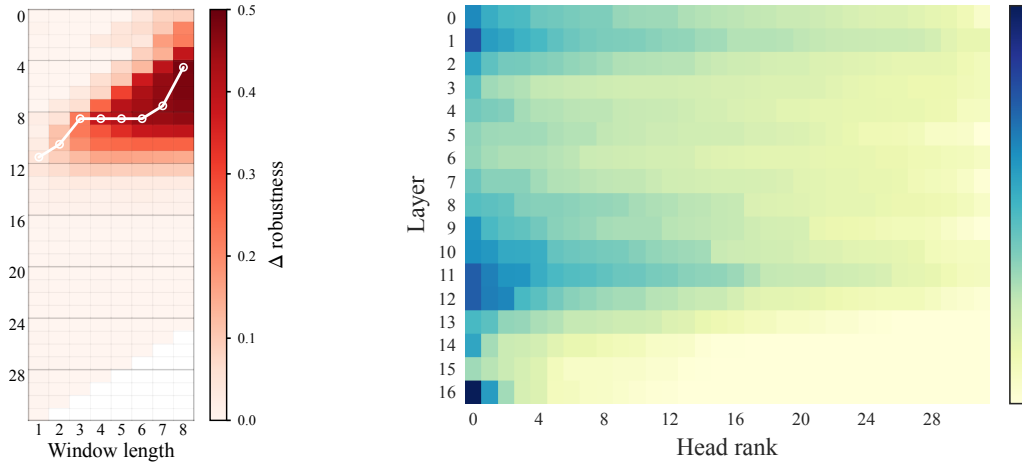


Figure 34: Mechanistic localization of an option-routing feature in a constrained persuasion task: layer-window patching and head-circuit composition both concentrate in a middle-layer band.

Source note. Rectangular crop of Sun et al. [46, Fig. 6], physical PDF page 8, CC BY 4.0. The result uses four-option, single-token decisions and selected persuasion-flip cases; it must not be generalized to free-form generation, real search, or human persuasion.

11.3 A boundary between optimization and manipulation

Intent alone cannot draw the boundary. A creator may intend to clarify a fact yet publish a misleading comparison; an attacker may publish only technically true sentences arranged to create a false implication. The revision itself is audited along five dimensions:

1. **truth:** are explicit and implied claims supported within scope?
2. **provenance:** can evidence origins and dependencies be traced?
3. **disclosure:** are sponsorship, affiliation, and material interests visible where a user would reasonably need them?
4. **system integrity:** does the content contain instructions or structures intended to override a consuming system rather than inform a user?
5. **competitive fairness:** does it fabricate consensus, impersonate independence, or degrade another party through unsupported content?

Definition 11.1: Evidence-preserving intervention

A source revision is evidence-preserving only if every changed proposition is either logically equivalent to an approved claim or is supported by a newly approved evidence path, all material qualifications remain locally available, and provenance and disclosure are not weakened. Improved visibility is not part of this definition.

11.4 Defense must protect utility, not only reject edits

Purely punitive filters create two problems. First, attackers can adapt to known signals. Second, legitimate sources may be penalized for making evidence more explicit. Recent mechanism-design work models repeated supplier–platform interaction and evaluates a reward for verifiable content alongside manipulation penalties [59]. Its particular game and benchmark results require independent scrutiny, but the design principle is valuable: a

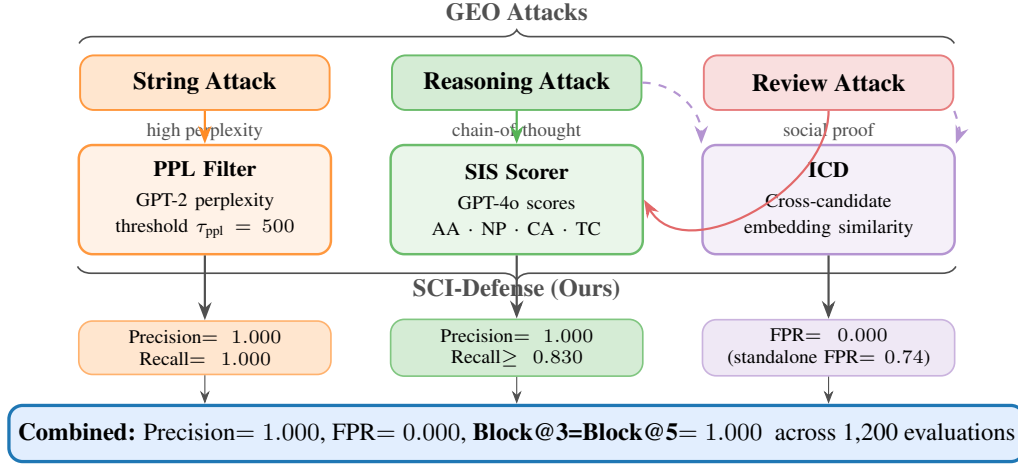


Figure 35: SCI-Defense routes three known attack families to distinct detectors before combining their decisions. The architecture itself reveals its template and threshold dependencies.

Source note. Rectangular crop of Yu et al. [62, Fig. 1], physical PDF page 4, CC BY 4.0. The displayed perfect aggregate precision/FPR cells are benchmark-specific; cross-domain, novel-attack, and small-sample results prevent a claim of universal defense.

defense should optimize joint utility rather than celebrate a high block rate.

For a defense D , define a vector

$$\mathbf{U}(D) = [U_{\text{answer}}, U_{\text{attribution}}, U_{\text{honest source}}, -R_{\text{attack}}, -R_{\text{false positive}}]. \quad (74)$$

Evaluation reports every component and the distribution across source classes. A weighted net score may support a declared decision, but it must not replace the components or conceal a severe harm.

Defense-oriented papers in the current library include SCI-Defense, SafeGEO, evidence aggregation under controlled GEO-style poisoning, and detection of optimized web content [9, 51, 52, 62]. They span different tasks—recommendation agents, fact verification, manipulation defense, and detection—so they belong in a defense matrix rather than one leaderboard.

11.4.1 Base-rate-aware detector calibration and defense routing

Benchmark F1, sensitivity, or false-positive rate describes discrimination under a sampled class balance. It does not answer the operational question: among sources flagged in the intended population, how many satisfy the frozen target label? Let $M = 1$ denote that independently adjudicated condition, $Z = 1$ a detector flag, $\pi = \Pr(M = 1)$ target-population prevalence, $s = \Pr(Z = 1 \mid M = 1)$ sensitivity, and $f = \Pr(Z = 1 \mid M = 0)$ the false-positive rate. Bayes’ rule gives

$$\text{PPV}(\pi; s, f) = \Pr(M = 1 \mid Z = 1) = \frac{s\pi}{s\pi + f(1 - \pi)}. \quad (75)$$

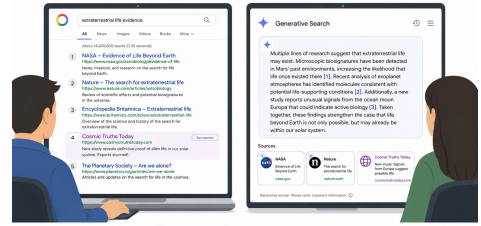
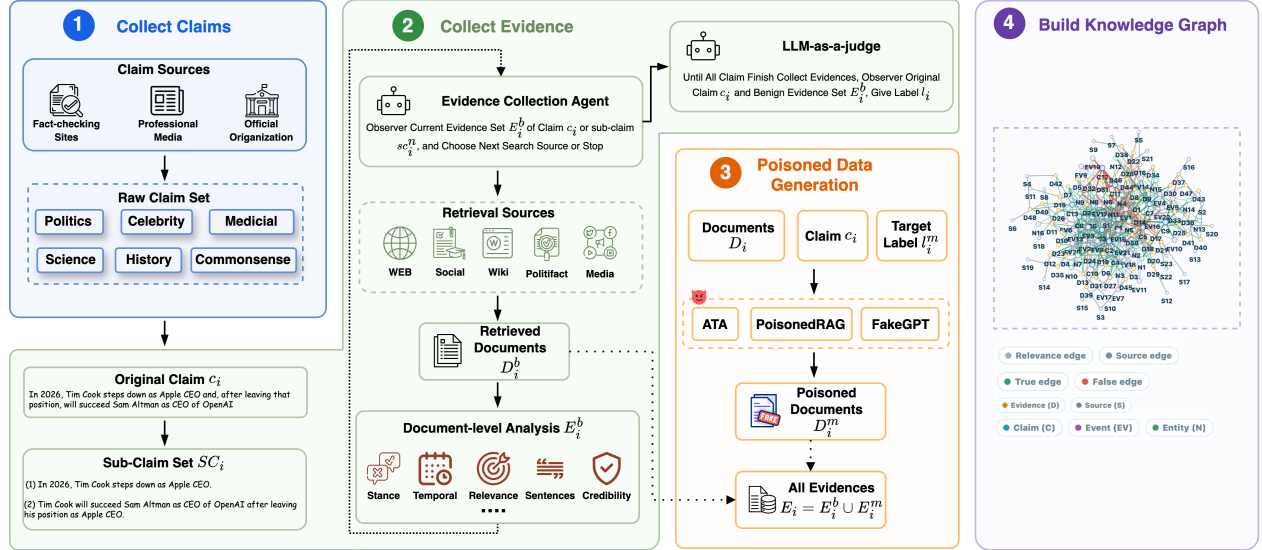


Figure 37: Why synthesized answers can make source provenance less immediately visible than a ranked result page.

Source note. Crop of Chu et al. [9, Fig. 1], physical PDF page 1, CC BY 4.0. It is an illustration, not empirical evidence that users never inspect sources or that detected GEO content is false.

Table 38: Defense evidence must retain its threat family, outcome, and failure mode.

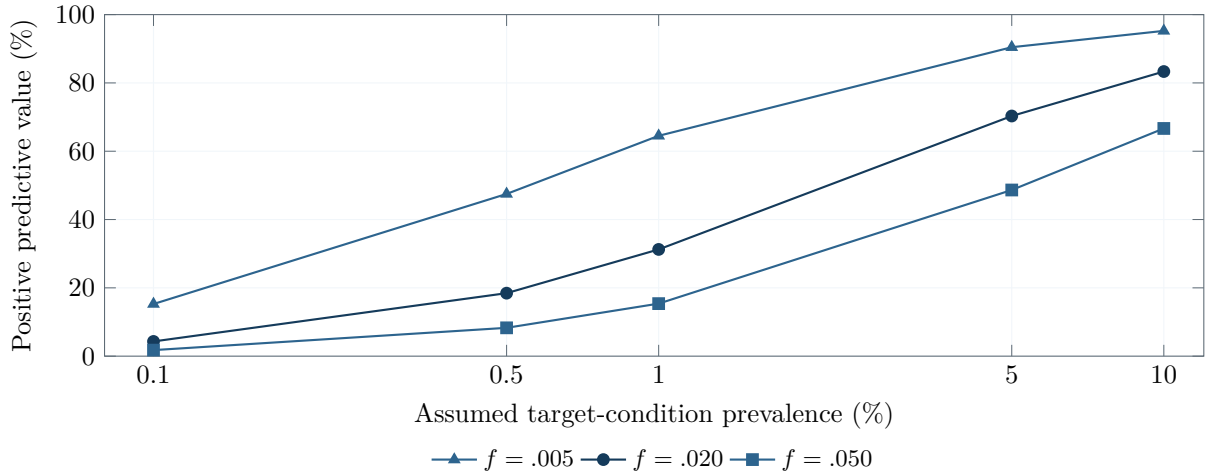
Evidence object	What the reported experiment supports	Negative result or claim ceiling
SCI-Defense [62]	For known ProductBench templates, reported recall is 1.000/.952/.830 for string/reasoning/review attacks, with an aggregate zero FPR in the main table.	An independent $n = 50$ table reports FPR .02. Five of six new black-box attack families have detection 0-.067 and Block@3 0; known-template performance is not strategy-external defense.
SafeGEO [52]	Across 600 synthetic recommendation cases and 40,800 instances, one Gemma cell moves target recommendation 3.4 $\rightarrow$ 79.6, hard-constraint violation 16.9 $\rightarrow$ 75.6, and utility NDCG 74.4 $\rightarrow$ 68.6.	The strongest reported evidence decomposition still leaves target rates 49.9/39.1/73.2 across three models, far above truthful baselines; rationale elicitation worsens some cells. This is single-turn, text, open-model evidence, not a production prevalence estimate.
GPE evidence poisoning [51]	With 638 constructed claims and three evidence items per claim, a Direct cell under ATA poisoning falls 52.7 $\rightarrow$ 39.2 $\rightarrow$ 30.4 $\rightarrow$ 8.8 as contamination rises.	No aggregator wins across every attack. Configurations appear to run once, and data dates, annotation agreement, implementation, and uncertainty are insufficient; exact robustness rankings remain quarantined.
GEO-Flag [9]	On its benchmark, ModernBERT with in-domain pretraining reports F1 .944, worst-group accuracy .883, and FPR gap .108, improving substantially over word TF-IDF's .880/.375/.558.	Live citations have no GEO ground truth, the gate recall is 88.97%, and metadata are incomplete. A detected style is not proof of falsehood, malicious intent, policy violation, or a valid basis for automatic public accusation.

**Figure 36:** GPE benchmark construction separates claim collection, iterative evidence collection, controlled poisoning, and knowledge-graph assembly. This topology is useful because the poisoned and original evidence remain explicit.

Source note. Rectangular crop of Wang et al. [51, Fig. 1], physical PDF page 3, CC BY 4.0. The benchmark contains constructed claims and controlled evidence environments; it does not estimate real-world poisoning prevalence or prove one aggregation method robust across attacks.

Table 39: Synthetic PPV sensitivity under $s = .900$. Each entry is a direct evaluation of Equation (75), not an empirical estimate.

Prevalence π	PPV at $f = .005$	PPV at $f = .020$	PPV at $f = .050$
.001	.152672	.043103	.017699
.005	.474934	.184426	.082949
.010	.645161	.312500	.153846
.050	.904523	.703125	.486486
.100	.952381	.833333	.666667

**Figure 38:** Deterministic PPV sensitivity for one synthetic sensitivity and three false-positive rates. The logarithmic horizontal axis emphasizes the low-base-rate region. Joined points are arithmetic traces, not fitted curves; no line estimates a named detector, corpus, provider, or deployment population.

The expected alert fraction is $a(\pi; s, f) = s\pi + f(1 - \pi)$. Both quantities become deployment claims only when the label, population, threshold, abstention rule, error rates, and time window transport to the intended use. Benchmark class balance and the detector-positive fraction are not substitutes for π .

Worked example 11.2: A 10,000-source base-rate calculation

Consider a declared synthetic population of 10,000 sources with $\pi = .01$, $s = .90$, and $f = .02$. It contains 100 target-positive and 9,900 target-negative sources, producing

$$\begin{aligned} \text{TP} &= .90(100) = 90, & \text{FN} &= 100 - 90 = 10, \\ \text{FP} &= .02(9,900) = 198, & \text{TN} &= 9,900 - 198 = 9,702. \end{aligned}$$

There are $90 + 198 = 288$ alerts and $\text{PPV} = 90/288 = .3125$. Thus 68.75% of alerts are false positives even though sensitivity is 90% and specificity is 98%. These are arithmetic fixtures, not estimates for GEO-Flag, a named platform, or live-web manipulation.

A release gate also retains answer utility, burden on honest sources, and the worst declared group rather than optimizing detector performance alone. For the following synthetic policy, require $s \geq .850$, worst-group $f \leq .030$, answer-utility loss at most .050, and honest-source burden at most .040. The thresholds are teaching inputs, not

Table 40: Three synthetic defense-gate fixtures under the declared teaching policy.

ID	Sensitivity	Worst FPR	Utility loss	Honest burden	Decision
D-A	.900	.020	.020	.030	PASS_INTERNAL_GATE
D-B	.920	.060	.010	.020	DENY_FPR
D-C	.910	.020	.080	.020	DENY_UTILITY

Algorithm 9 SELECTDEFENSEGATE: utility-preserving detector routing

Inputs: Frozen label and target population; held-out sensitivity s ; group FPRs f_g ; answer-utility loss L_U ; honest-source burden L_H ; registered limits $s_{\min}, f_{\max}, L_{U,\max}, L_{H,\max}$

Outputs: Deterministic internal decision code and reconstructable audit record

```

1: Verify label, sampling frame, time window, threshold, abstention rule, subgroup map, adjudication process, and their
   hashes
2: if any required object is absent, selected after outcomes, or unmatched to the target population then
3:   return DENY_CONTRACT
4: end if
5: if  $s < s_{\min}$  then
6:   return DENY_TPR
7: end if
8: if  $\max_g f_g > f_{\max}$  then
9:   return DENY_FPR
10: end if
11: if  $L_U > L_{U,\max}$  then
12:   return DENY_UTILITY
13: end if
14: if  $L_H > L_{H,\max}$  then
15:   return DENY_HONEST_SOURCE_BURDEN
16: end if
17: Archive all component values, intervals, denominators, gates, reviewer, appeal, monitoring, and rollback owners
18: Route flags to accountable review with abstention; do not automate a public accusation, removal, or penalty
19: return PASS_INTERNAL_GATE

```

universal safety limits.

Evidence boundary · calibration is not accusation

Every prevalence, error rate, count, PPV, threshold, and gate decision in this subsection is a declared synthetic teaching value. A benchmark, enriched audit queue, or detector-positive sample does not estimate prevalence in the intended population. A style or anomaly flag is not evidence of falsehood, malicious intent, authorship, policy violation, causal influence, or user harm. PPV alone does not establish utility: recall, abstention, coverage, subgroup error, alert load, reviewer error, adaptive shift, and every component of $\mathbf{U}(D)$ remain separate release objects. Detector output alone must never authorize a public accusation or another consequential action.

The longer-run incentive question is different again. A repeated-game model shows that reducing a high attack-success parameter need not monotonically increase cooperation under its local assumptions [15]. A separate finite-round Stackelberg study reports positive values for its equally weighted Defense–Welfare Net under a combined mechanism, but rankings depend on that proxy, its weights, and a local experimental population [59]. These results justify parameter sensitivity and mechanism experiments; neither is a policy proof or a global win–win guarantee.

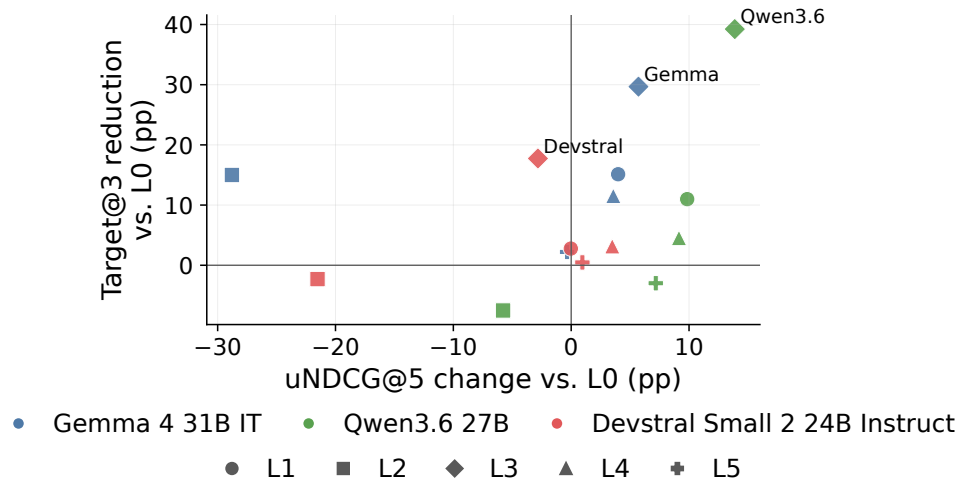


Figure 39: Mitigation trade-off between target-promotion reduction and utility change across model/layer packages. Several points improve one axis while worsening the other, so a defense cannot be chosen from block rate alone.

Source note. Vector crop of Wen et al. [52, Fig. 20], physical PDF page 38, CC BY 4.0. The result uses a fixed, single-turn, post-retrieval text environment with 22 documents; it does not establish a truthful baseline recovery or production transfer.

Table 41: A minimal governance pattern for GEO changes.

Role	Accountable for	Cannot unilaterally approve
Claim owner	factual scope, freshness, correction	experimental method or legal sufficiency
Evidence reviewer	source identity, support, conflict, independence	public release outside assigned domain
Experiment owner	protocol, traces, analysis, uncertainty	unsupported claim changes
Security reviewer	attack surface, abuse cases, red-team evidence	business acceptance of residual risk
Legal/compliance reviewer	disclosure, privacy, sector and advertising obligations	scientific effect claims without method review
Release owner	version, approvals, deployment, rollback, incident trigger	bypass of mandatory evidence or safety gate

11.5 Governance as a release system

Responsible practice requires named decision rights. Table 41 shows a minimal role pattern; actual legal and regulatory responsibilities vary by jurisdiction and sector.

Each release record includes the source hash before and after revision, changed claims, supporting evidence, reviewers, experiment identifier, jurisdictions, deployment time, monitoring window, and rollback action. High-risk domains add professional review appropriate to the claim. A language model can assist with extraction and consistency checking, but it is not the accountable approver.

11.5.1 Monitoring and incident response

Monitoring watches both desired and harmful outcomes:

- unsupported-claim and wrong-attribution rates;
- unusual citation concentration or source-network dependence;

- drift across systems, locales, and time;
- spikes in optimized or instruction-like content;
- complaints, correction requests, and business-impact anomalies;
- false-positive burden on legitimate and less-established sources.

An incident workflow is *detect* → *preserve* → *assess* → *contain* → *correct* → *notify* → *learn*. Preservation includes the exact response, URLs, timestamps, account and surface state, source versions, and permissible logs. Correcting only the public page is insufficient if stale copies, syndicated sources, or internal ledgers retain the error.

11.6 Standards and regulation without status inflation

Standards can contribute terminology, lifecycle controls, risk management, measurement procedures, privacy controls, and citation practice. They enter the manuscript only after the identity record in [definition 2.1](#) verifies issuer, number, edition, status, effective date, and relevant clauses. A draft, project announcement, association white paper, national standard, and mandatory law are different institutional objects.

These Notes therefore use a **status-first standard audit**: the researcher first reproduces the official identity page, then maps applicable clauses to a specific system or research artifact. A generic principle is not proof that an intervention is compliant. [Chapter C](#) provides the initial versioned register and records scope without reproducing paywalled clauses.

Boundary and failure conditions

These Notes provide a research and engineering framework, not legal advice. A deployment involving personal data, consumer recommendations, health, finance, employment, education, or other high-impact decisions requires qualified review for the relevant jurisdiction, sector, and current system design.

Reproduction Studio 11 · Red-team a GEO intervention

Input. The three revisions from Reproduction Studio 8, an attacker capability budget, a controlled retrieval/-generation pipeline, and the release record.

Task. Implement benign evidence clarification, unsupported persuasion, instruction injection, synthetic consensus, and competitor displacement test conditions. Measure target exposure, answer utility, attribution, attack success, and false positives. Run the governance gate and one simulated incident.

Deliverable. Threat model, adversarial test corpus, defense matrix, residual-risk decision, release/deny record, and incident postmortem.

Stop rule. Deny release when claim provenance is broken, a material interest is undisclosed, severe harm exceeds threshold, or the rollback and evidence-preservation procedure fails.

Chapter summary

1. GEO creates attack surfaces from acquisition through measurement.
2. Truth, provenance, disclosure, system integrity, and competitive fairness distinguish evidence improvement from manipulation.
3. Defenses should protect joint utility and measure false positives, not only block known attacks.
4. Governance is a versioned release system with named accountable roles, monitoring, rollback, and incident response.
5. Standards and regulation must be identified by official status and applicable scope before entering a compliance claim.

12 Governance, Standards, Regulation, and Incident Response

Key idea

Governance is a versioned decision system: it maps claims, evidence, risks, controls, accountable roles, approvals, monitoring, and corrective actions to a specific release. Standards and regulations enter only through verified identity, status, scope, jurisdiction, date, and applicability—not through a generic “compliant” label.

12.1 Govern releases, not aspirations

A governance record concerns a concrete object r : a source revision, measurement protocol, model-assisted workflow, dataset, public result, or course artifact. Define

$$\Gamma(r) = \langle I_r, C_r, E_r, R_r, K_r, A_r, M_r, B_r \rangle, \quad (76)$$

where I_r is identity and version, C_r changed claims, E_r evidence, R_r risks, K_r controls, A_r accountable approvals, M_r monitoring, and B_r rollback or correction. A policy statement without a mapping to these objects cannot demonstrate that the release was governed.

Definition 12.1: Release evidence package

A release evidence package is an immutable or tamper-evident bundle containing the before/after source hashes, claim and rendered diffs, supporting evidence versions, experimental protocol and results, approvals, applicable-control crosswalk, deployment time, monitoring window, incident triggers, and tested rollback. It records what was reviewed; it does not by itself prove legal compliance or control effectiveness.

12.1.1 Fail closed on release state

Package completeness does not authorize a state transition. Let the canonical payload root be computed over a sorted manifest of non-approval artifacts. Approval records bind that root, the release identity, declared scope, decision time, and validity interval; the complete archive may have a separate outer digest. This avoids a circular construction in which an approval would need to sign itself.

For release r at evaluation time t , retain eight non-compensable gates: identity g_I , changed claims g_C , evidence g_E , risk g_R , controls g_K , approvals g_A , monitoring g_M , and rollback g_B . Each gate is one only when its status is explicitly *pass*; missing, stale, failed, ambiguous, unauthenticated, or unresolved evidence cannot inherit a previous pass. Define

$$G_{\text{release}}(r, t) = g_I g_C g_E g_R g_K g_A g_M g_B. \quad (77)$$

Thus $G_{\text{release}} = 1$ only when all eight gates pass for the same release identity and evaluation time. No weighted score may compensate for a non-pass gate, and a machine validator cannot create an accountable approval.

Table 42: Eight non-compensable release gates. Passing is a bounded internal control state, not a finding of compliance, safety, scientific validity, accessibility, publication acceptance, or production readiness.

Gate	Evidence object	Minimum pass condition
g_I	identity and version	normalized paths and IDs are unique; declared bytes and hashes reproduce the exact payload root
g_C	changed claims	every material change has an ID, source and rendered diff, evidence route, limitation, owner, and disposition
g_E	evidence	required objects resolve to declared versions and locators, stay in scope, pass integrity checks, and remain fresh
g_R	risks	no unresolved severe trigger or threshold breach remains; residual-risk decisions are current
g_K	controls	applicable controls name owners, artifacts, tests, thresholds, results, and unexpired exceptions
g_A	approvals	every required accountable role has a valid decision bound to the payload root, identity, scope, and interval
g_M	monitoring	outcomes, denominators, thresholds, owners, preservation, review windows, and escalation triggers are executable
g_B	rollback and correction	a version-specific target exists and its dependencies and preservation steps have been exercised

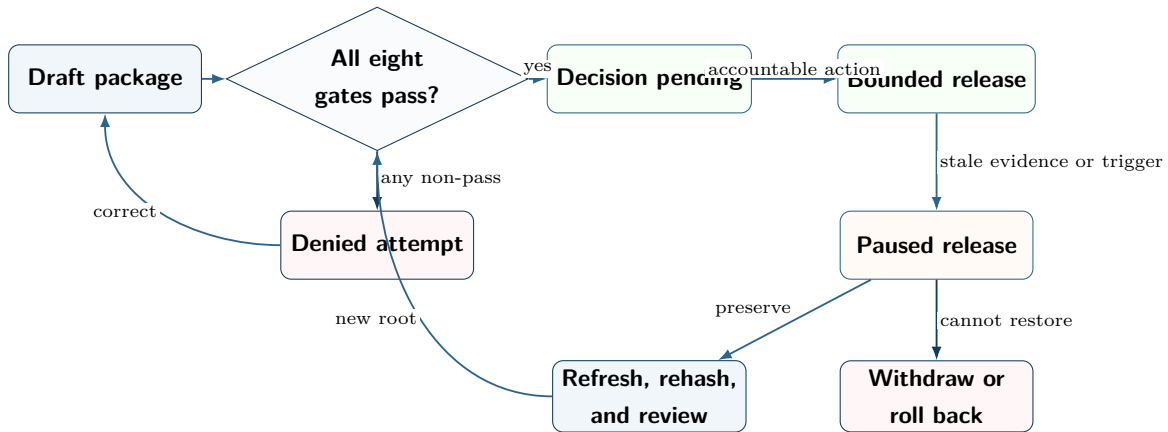
**Figure 40:** Original fail-closed release state machine. A pre-release non-pass denies the attempted transition; a non-pass after bounded release pauses continued use. Repair changes the payload root and requires a new review. A machine result never supplies the accountable release decision.

Figure 40 distinguishes denial of a proposed transition from suspension of an already active release. Neither state deletes adverse evidence, deviations, or prior approvals. A corrected or refreshed payload receives a new root and returns to review; an approval bound to the old root cannot be copied forward.

Evidence boundary · an internal gate is not external approval

$G_{\text{release}} = 1$ is necessary only under this project's declared gate policy. It is not sufficient evidence of truth, causal effect, scientific acceptance, lawful applicability, conformity, control effectiveness, security, accessibility, publication acceptance, deployment authorization, or production readiness. A digest establishes byte identity, not truth or lawful use. An approval record establishes a bounded recorded decision, not reviewer competence or correctness. A freshness limit is a project rule, not a universal legal interval. DENIED and PAUSED are routing states, not findings of wrongdoing.

Algorithm 10 VALIDATERELEASEEVIDENCEPACKAGE: fail-closed validation

Inputs: Release package $\Gamma(r)$; canonical manifest and package bytes; current state; evaluation time t ; frozen gate policy

Outputs: Append-only audit, next state, and deterministic decision code; the procedure never creates or infers an approval

- 1: Parse normalized paths and IDs; reject duplicates, traversal, unsupported hashes, and byte-count or digest mismatch
- 2: **if** payload identity or a declared digest fails **then**
- 3: **return** DENY_HASH_MISMATCH before release, otherwise pause
- 4: **end if**
- 5: Derive gate states in frozen order I, C, E, R, K, A, M, B from evidence; never trust an unverified manifest assertion of pass
- 6: **if** $g_C = 0$ **then**
- 7: **return** DENY_UNRESOLVED_CLAIM before release, otherwise pause
- 8: **end if**
- 9: **if** $g_E = 0$ **then**
- 10: **return** DENY_EVIDENCE before release; otherwise preserve the package and pause, using PAUSE_STALE_EVIDENCE when evidence is stale
- 11: **end if**
- 12: **if** $g_R = 0$ **then**
- 13: **return** DENY_RISK before release, otherwise pause
- 14: **end if**
- 15: **if** $g_K = 0$ **then**
- 16: **return** DENY_CONTROL_TEST before release, otherwise pause
- 17: **end if**
- 18: **if** $g_A = 0$ **then**
- 19: **return** DENY_MISSING_APPROVAL before release, otherwise pause
- 20: **end if**
- 21: **if** $g_M = 0$ **then**
- 22: **return** DENY_MONITORING before release, otherwise pause
- 23: **end if**
- 24: **if** $g_B = 0$ **then**
- 25: **return** DENY_ROLLBACK before release, otherwise pause
- 26: **end if**
- 27: Record all tested predicates, locators, values, thresholds, times, approvals, and the ordered non-pass set
- 28: **return** PASS_INTERNAL_RELEASE_GATE; route the exact root to the accountable release owner for a separate decision

Table 43: Two synthetic fail-closed fixtures. One denies a draft with missing approval; the other pauses an active release whose evidence has become stale.

ID	State	<i>I</i>	<i>C</i>	<i>E</i>	<i>R</i>	<i>K</i>	<i>A</i>	<i>M</i>	<i>B</i>	Decision	Next
R-A	draft	1	1	1	1	1	0	1	1	DENY_MISSING_APPROVAL	draft
R-B	released	1	1	0	1	1	1	1	1	PAUSE_STALE_EVIDENCE	paused

Fixture R-A retains a missing accountable approval as $g_A = 0$; the validator does not invent a signer or copy assent from another payload. Fixture R-B has seven current pass records but stale evidence, so the already active release is paused and preserved. A refreshed evidence object changes the root and returns the package to review. Both rows are invented and contain no real person, organization, product, jurisdiction, or production system.

At the framework level, the NIST Generative AI Profile foregrounds governance, content provenance, pre-deployment testing, and incident disclosure as four cross-cutting considerations [32]. In this book they become separate evidence-producing activities. A provenance record cannot substitute for a red-team result; a pre-deployment test cannot substitute for incident readiness; and none of the four is a legal safe harbor or certification.

12.2 Verify the institutional object before mapping it

For every law, regulation, standard, framework, technical specification, platform policy, or project announcement, retain an identity record:

$$s = \langle \text{issuer, identifier, title, edition, status, publication, effective date, scope, official URL} \rangle. \quad (78)$$

Status distinctions are substantive. A voluntary framework, published international standard, recommended national standard, mandatory national standard, draft, project, law, product policy, and community proposal do not create the same obligations or evidentiary authority.

The distinctions are concrete at this research cutoff. GB 45438–2025 is recorded by the national standards platform as a current mandatory Chinese national standard, while the 2025 labeling measures are a separate joint official measure with their own actor and service scope [12, 42]. ISO/IEC 42005:2025 is a published International Standard on AI-system impact assessment [19]; NIST AI RMF 1.0 is a voluntary government framework [31]. Their official identities support these status and scope statements. They do not, without applicability analysis and evidence of implementation, establish a duty for a particular release or prove that a control is effective.

Appendix C is a dated identity and project-control register. It deliberately avoids reproducing licensed clauses and does not certify conformity. Before public release, an owner rechecks every official status page and qualified reviewers assess applicability for the actual organization, jurisdiction, sector, data, and system role.

12.3 Map requirements to testable controls

A control crosswalk is not a list of citations. Each row records:

- source identity and exact provision or officially stated function;
- applicability rationale and unresolved interpretation;
- risk and control objective;
- accountable owner and implementation artifact;
- test method, frequency, threshold, and evidence location;

Table 44: Status-first admission test for governance sources.

Question	Required evidence	Invalid inference blocked
Identity	official issuer page, identifier, edition, stable URL	title match treated as verification
Status	published/draft/project/withdrawn and effective date	proposal treated as current requirement
Scope	subject, actors, exclusions, jurisdiction, sector	general principle treated as universal duty
Text access	licensed clause or authoritative public text	metadata page treated as full requirements
Applicability	system/data/role facts and qualified analysis	crosswalk treated as legal conclusion
Evidence of control	owner, implementation, test, exception, artifact	policy statement treated as effectiveness

- exception, compensating control, expiry, and approval;
- last review and next required refresh.

Technical protocols have similarly bounded roles. Robots rules can express crawler access preferences; sitemaps can advertise URLs; structured data can encode supported page information; provenance vocabularies can record derivation; accessibility standards can define testable presentation criteria. None guarantees crawling, indexing, retrieval, citation, truth, or business outcome.

12.4 Assign decision rights and preserve independence

Claim owners, evidence reviewers, experiment owners, security reviewers, legal/privacy/compliance reviewers, and release owners provide different forms of assurance. One person may hold several roles in a small team, but the record still identifies which judgment was made and where independence is absent.

High-impact claims receive domain-qualified review. A language model can extract candidate provisions, compare versions, detect missing fields, or draft a crosswalk. It is not the accountable approver and cannot determine legal applicability from generic context.

Worked example 12.2: Why a standards citation is not a release decision

A team links to an AI risk-management framework and marks a content agent “compliant.” The link may verify that the framework exists. It does not show which functions apply, who owns the risks, whether controls were implemented, how they were tested, what exceptions remain, or whether the word “compliant” is meaningful for that voluntary framework. A release package replaces the label with a dated, bounded crosswalk and evidence of the team’s own controls.

12.5 Monitor both desired effects and harms

Monitoring preserves the same units and denominators used in evaluation. A release dashboard should not collapse the following into one health score:

- claim correctness, unsupported-claim severity, and correction latency;
- citation entailment, completeness, wrong-source attachment, and source concentration;
- visibility, fidelity, user utility, and qualified downstream action;
- attack indicators, false positives, and disparate burden on legitimate source classes;
- platform, model, parser, corpus, query-demand, and judge drift;

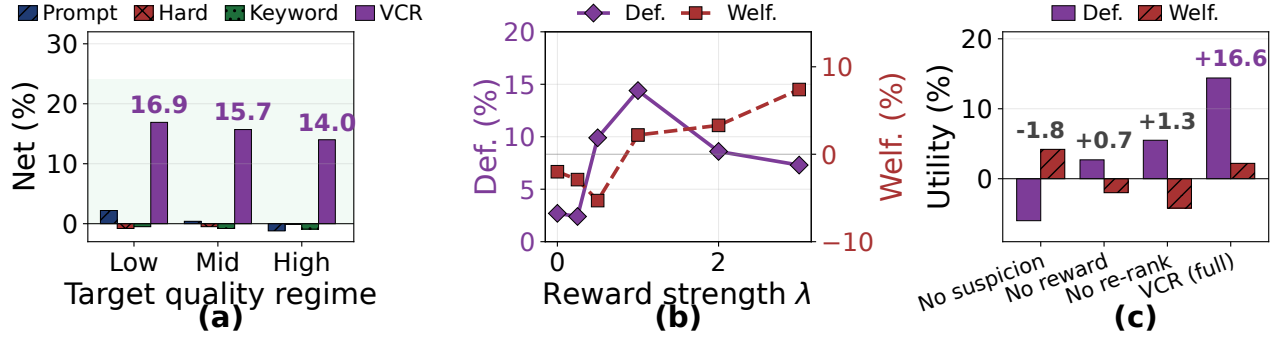


Figure 41: Robustness and ablation panels for a verifiable-content reward: quality regimes, reward strength, and component removal all change the authors’ declared Defense–Welfare objectives.

Source note. Vector crop of Xu et al. [59, Fig. 4], physical PDF page 12, CC BY 4.0. “Net” is the paper’s declared equal-weight proxy in a finite, fixed-top-5 benchmark; it is not a global equilibrium proof, legal decision rule, or universal win–win guarantee.

- complaints, opt-outs, privacy events, accessibility failures, and evidence or license expiry.

Each alert has a threshold, severity, owner, response time, preservation rule, and escalation path. Absence of an alert is not evidence of safety when coverage or detection sensitivity is unknown.

12.6 Run incident response as an evidence-preserving protocol

The response sequence is

$$\text{detect} \rightarrow \text{preserve} \rightarrow \text{triage} \rightarrow \text{contain} \rightarrow \text{correct} \rightarrow \text{notify} \rightarrow \text{learn}. \quad (79)$$

Preservation includes the exact answer, interface and citation state, URLs, timestamps, locale, account and search state, source versions, model and parser versions when available, experiment trace, and permissible logs. Containment may pause a release, revert a source, disable an agent action, quarantine an evidence path, or add human review.

Correction propagates across the fact ledger, public page, syndicated copies, structured representations, course materials, benchmark labels, and published errata. The incident closes only after corrective action is verified and the monitoring gap is addressed.

12.7 Use mechanism design to align incentives

Pure suppression can punish legitimate evidence improvements and invite adaptation. Governance should reward verifiable, useful, independently traceable evidence while increasing the cost of hidden instruction, fabricated consensus, unsupported persuasion, and competitor harm. Mechanism-design work in GEO studies such incentives under formal and experimental assumptions [59]; risk-focused scholarship defines broader open questions [53]. These sources inform design and research priorities, not an off-the-shelf guarantee for a live platform.

Boundary and failure conditions

These Notes provide a research, engineering, and documentation framework, not legal advice, certification, or a conformity assessment. Current obligations depend on jurisdiction, sector, system role, data, affected people, and the latest official text. High-impact, privacy, consumer, advertising, employment, education, health, finance, and security deployments require appropriately qualified review.

Reproduction Studio 12 · Build and exercise a release evidence package

Question. Can a bounded GEO intervention pass a status-first governance gate and survive a simulated incident without losing evidence?

Input. One approved intervention, its experiment bundle, three official source identities from Appendix C, a synthetic organization and jurisdiction profile, and two injected incidents.

Procedure. Verify status and scope; build a clause/function-to-control crosswalk without reproducing restricted text; assign roles; test controls; approve or deny release; then simulate wrong-source attribution and stale high-impact evidence. Preserve, contain, correct, notify, and run a postmortem.

Output. Source-identity records, applicability questions, control matrix, signed decision, immutable manifest, monitoring plan, incident timeline, correction propagation map, residual-risk register, and lessons learned.

Reference criterion. An independent reviewer can distinguish official source status, project control, test evidence, interpretation, exception, and unresolved legal question in every row.

Stop rule. Deny or pause release when identity/status cannot be verified, required review is absent, evidence preservation fails, severe harm exceeds threshold, or rollback is not executable.

Chapter summary

1. Governance attaches evidence, controls, decisions, monitoring, and rollback to a concrete versioned release.
2. Standards, regulations, policies, and proposals enter through verified identity, status, scope, date, and applicability.
3. A crosswalk is not compliance; each control needs an owner, artifact, test, exception path, and refresh date.
4. Monitoring retains separate visibility, fidelity, utility, risk, fairness, and operational outcomes.
5. Incident response preserves the evidence chain and propagates corrections across every dependent artifact.
6. Qualified accountable reviewers—not a model or a generic checklist—make legal, safety, and release decisions.

References

- [1] Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization, 2024. URL <https://arxiv.org/abs/2311.09735>.
- [2] Shayan Alipour, Mehdi Kargar, and Morteza Zihayat. When attention becomes exposure in generative search, 2026. URL <https://arxiv.org/abs/2601.01750>.
- [3] Puneet S. Bagga, Vivek F. Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. E-GEO: A testbed for generative engine optimization in E-Commerce, 2025. URL <https://arxiv.org/abs/2511.20867>.
- [4] Norman E. Breslow and David G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993. doi: 10.1080/01621459.1993.10594284.
- [5] Mahe Chen, Xiaoxuan Wang, Kaiwen Chen, and Nick Koudas. Generative engine optimization: How to dominate AI search, 2025. URL <https://arxiv.org/abs/2509.08919>.
- [6] Qiyuan Chen, Jiahe Chen, Hongsen Huang, Qian Shao, Jintai Chen, Renjie Hua, Hongxia Xu, Ruijia Wu, Ren Chuan, and Jian Wu. CC-GSEO-Bench: A Content-Centric benchmark for measuring source influence in generative search engines, 2025. URL <https://arxiv.org/abs/2509.05607>.
- [7] Xiaolu Chen, Jie Bao, Haojie Wu, Zhen Chen, and Yong Liao. Caption injection for optimization in generative search engine, 2026. URL <https://arxiv.org/abs/2511.04080>.
- [8] Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang. Role-Augmented Intent-Driven generative search engine optimization, 2026. URL <https://arxiv.org/abs/2508.11158>.
- [9] Junjie Chu, Ye Leng, Mingjie Li, Yun Shen, Xinyue Shen, and Yang Zhang. GEO-Flag: Detecting and measuring GEO-Optimized web content, 2026. URL <https://arxiv.org/abs/2608.16824>.
- [10] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. Association for Computing Machinery, 2009. doi: 10.1145/1571941.1572114.
- [11] Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration. Interim measures for the management of generative artificial intelligence services, 2023. URL https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Order No. 15; issued 10 July 2023; effective 15 August 2023; Chinese official text.
- [12] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration. Measures for labeling artificial-intelligence-generated and synthetic content, 2025. URL https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm. Joint official measure; issued 7 March 2025; effective 1 September 2025; Chinese official text.
- [13] Yixuan Du, Chenxiao Yu, Haoyan Xu, Ziyi Wang, Yue Zhao, and Xiyang Hu. Multimodal generative engine optimization: Rank manipulation for Vision-Language model rankers, 2026. URL <https://arxiv.org/abs/2601.12263>.
- [14] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993. ISBN 978-0-412-04231-7. doi: 10.1007/978-1-4899-4541-9.
- [15] Xiyang Hu. Dynamics of adversarial attacks on large language Model-Based search engines, 2025. URL <https://arxiv.org/abs/2501.00745>.
- [16] Junyao Huang, Ruimin Situ, and Renqin Ye. Cultural encoding in large language models: The existence gap in AI-Mediated brand discovery, 2025. URL <https://arxiv.org/abs/2601.00869>.
- [17] International Organization for Standardization. ISO/IEC 23894:2023: Information technology—artificial intelligence—guidance on risk management, 2023. URL <https://www.iso.org/standard/77304.html>. Edition 1; International Standard.

-
- [18] International Organization for Standardization. ISO/IEC 42001:2023: Information technology—artificial intelligence—management system, 2023. URL <https://www.iso.org/standard/42001>. Edition 1; International Standard.
- [19] International Organization for Standardization. ISO/IEC 42005:2025: Information technology—artificial intelligence (AI)—AI system impact assessment, 2025. URL <https://www.iso.org/standard/42005>. Edition 1; International Standard; published May 2025.
- [20] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.
- [21] Haibo Jin, Ruoxi Chen, Peiyan Zhang, Yifeng Luo, Huimin Zeng, Man Luo, and Haohan Wang. Controlling output rankings in generative engines for LLM-based search, 2026. URL <https://arxiv.org/abs/2602.03608>.
- [22] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentaoh Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL <https://aclanthology.org/2020.emnlp-main.550/>.
- [23] Sunghwan Kim, Wooseok Jeong, Serin Kim, Sangam Lee, and Dongha Lee. SAGEO arena: A realistic environment for evaluating Search-Augmented generative engine optimization, 2026. URL <https://arxiv.org/abs/2602.12187>.
- [24] Arlen Kumar and Leanid Palkhouski. AI answer engine citation behavior an empirical analysis of the GEO16 framework, 2025. URL <https://arxiv.org/abs/2509.10762>.
- [25] Pratyush Kumar. Generative engine optimization at scale: Measuring brand visibility across AI search engines, 2026. URL <https://arxiv.org/abs/2606.20065>.
- [26] Zikang Liu and Peilan Xu. Think before writing: Feature-Level Multi-Objective optimization for generative citation visibility, 2026. URL <https://arxiv.org/abs/2604.19113>.
- [27] Florian Lüttgenau, Imar Colic, and Gervasio Ramirez. Beyond SEO: A Transformer-Based approach for reinventing web content optimisation, 2025. URL <https://arxiv.org/abs/2507.03169>.
- [28] Ministry of Industry and Information Technology of the People’s Republic of China. SJ/T 12116–2026: Artificial intelligence—key technology—general technical requirements for retrieval-augmented generation, 2026. URL <https://std.samr.gov.cn/hb/search/stdHBDetailed?id=594CFDC622F68BADE06397BE0A0A2CD3>. Recommended electronic-industry standard; published and effective 28 April 2026; public metadata only.
- [29] Ministry of Industry and Information Technology of the People’s Republic of China. SJ/T 12117–2026: Artificial intelligence—key technology—evaluation specification for retrieval-augmented generation systems, 2026. URL <https://std.samr.gov.cn/hb/search/stdHBDetailed?id=594CFDC623018BADE06397BE0A0A2CD3>. Recommended electronic-industry standard; published and effective 28 April 2026; public metadata only.
- [30] Riku Mochizuki, Shusuke Komatsu, Souta Noguchi, and Kazuto Ataka. Exposing citation vulnerabilities in generative engines, 2025. URL <https://arxiv.org/abs/2510.06823>.
- [31] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023. URL <https://doi.org/10.6028/NIST.AI.100-1>.
- [32] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024. URL <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>.
- [33] National Institute of Standards and Technology. Guidance and templates for public-facing ai documentation: Ai standards “zero draft” initial public draft, 2026. URL <https://www.nist.gov/publications/guidance-and-templates-public-facing-ai-documentation-ai-standards-zero-draft-initial>. Initial public draft, 30 July 2026; preliminary proposal rather than a consensus standard.
- [34] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models, 2024. URL <https://arxiv.org/abs/2406.18382>.

- [35] Sara Patel, Mingxun Zhou, and Giulia Fanti. MaxShapley: Towards incentive-compatible generative search with fair context attribution, 2025. URL <https://arxiv.org/abs/2512.05958>.
- [36] Perplexity AI. Agent api presets, 2026. URL <https://docs.perplexity.ai/docs/agent-api/presets>. First-party platform documentation; accessed 24 August 2026.
- [37] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/15000000019.
- [38] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.
- [39] Uri Samet. From noise to narrative: Building reputation in AI dominated search and discovery environments. *Journal of Emerging Technologies and Innovative Research*, 12(8):e35–e49, August 2025. doi: 10.56975/jetir.v12i8.568287. URL <https://www.jetir.org/papers/JETIR2508406.pdf>.
- [40] Julius Schulte, Malte Bleeker, and Philipp Kaufmann. Don’t measure once: Measuring visibility in AI search (GEO), 2026. URL <https://arxiv.org/abs/2604.07585>.
- [41] Standardization Administration of China. GB/T 43782–2024: Artificial intelligence—technical requirements for machine learning systems, 2024. URL <https://std.samr.gov.cn/gb/search/gbDetailed?id=14156507D1D40337E06397BE0A0AE656>. Chinese recommended national standard; current status verified on the National Public Service Platform for Standards.
- [42] Standardization Administration of China. GB 45438–2025: Cybersecurity technology—labeling method for content generated by artificial intelligence, 2025. URL <https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcnno=F32EA2A561F1886CD8D606513512D547>. Mandatory Chinese national standard; published 28 February 2025; effective 1 September 2025.
- [43] Standardization Administration of China. GB/T 45288.1–2025: Artificial intelligence—large-scale model—part 1: General requirements, 2025. URL <https://std.samr.gov.cn/gb/search/gbDetailed?id=2FF37940EB12D753E06397BE0A0A413F>. Chinese recommended national standard; current.
- [44] Standardization Administration of China. Project 20252041-z-469: Artificial intelligence—evaluation indicators and methods for retrieval-augmented generation systems, 2025. URL <https://std.samr.gov.cn/gb/search/gbDetailed?id=37FC03D2E1446322E06397BE0A0AA17F>. Registered national guiding-technical-document project; status under approval at the 24 August 2026 research cutoff; not a published standard.
- [45] Standardization Administration of China. GB/T 47507–2026: Artificial intelligence—trustworthiness—general principles, 2026. URL <https://std.samr.gov.cn/gb/search/gbDetailed?id=511EBC5967DA9318E06397BE0A0AFBD5>. Chinese recommended national standard; published 30 April 2026; effective date 1 August 2026.
- [46] Xiangkun Sun, Ling kai Kong, Aoqi Zhang, Liang Zeng, and Tonghan Wang. How LLMs are persuaded: A few attention heads, rerouted, 2026. URL <https://arxiv.org/abs/2605.09314>.
- [47] Zhihua Tian, Yuhua Chen, Yao Tang, Jian Liu, and Ruoxi Jia. Diagnosing and repairing citation failures in generative engine optimization, 2026. URL <https://arxiv.org/abs/2603.09296>.
- [48] Harold Triedman and Vitaly Shmatikov. MillStone: How Open-Minded are LLMs?, 2025. URL <https://arxiv.org/abs/2509.11967>.
- [49] Rahul Vishwakarma, Shushant Kumar, and Ratnesh Jamidar. What gets cited: Competitive GEO in AI answer engines. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4950–4954, 2026. doi: 10.1145/3805712.3808445.
- [50] Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing?, 2024. URL <https://arxiv.org/abs/2402.11782>.
- [51] Zhaoqi Wang, Zijian Zhang, Xiaomei Yuan, Pengtao Kou, Jiamou Liu, Zhen Li, and Liehuang Zhu. GPE: Evaluating robust evidence aggregation for fact verification under controllable GEO-Style poisoning, 2026. URL <https://arxiv.org/abs/2607.20730>.

-
- [52] Qianfeng Wen, Yifan Simon Liu, Xin Liu, Difan Jiao, Blair Yang, Junda Wu, and Zhenwei Tang. SafeGEO: Understanding generative engine optimization risks in recommendation agents, 2026. URL <https://arxiv.org/abs/2606.28356>.
- [53] Yizhu Wen, Nan Zhang, Haohan Yuan, Xun Chen, Haopeng Zhang, and Hanqing Guo. Position: On the risks of generative engine optimization in the era of LLMs, 2025.
- [54] World Wide Web Consortium. PROV-O: The PROV ontology, 2013. URL <https://www.w3.org/TR/prov-o/>. W3C Recommendation, 30 April 2013.
- [55] World Wide Web Consortium. Web content accessibility guidelines (WCAG) 2.2, 2024. URL <https://www.w3.org/TR/WCAG22/>. W3C Recommendation, 12 December 2024.
- [56] Beining Wu, Fuyou Mao, Jiong Lin, Cheng Yang, Jiaxuan Lu, Yifu Guo, Siyu Zhang, Yifan Wu, Ying Huang, and Fu Li. From experience to skill: Multi-Agent generative engine optimization via reusable strategy learning, 2026. URL <https://arxiv.org/abs/2604.19516>.
- [57] Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=K8EinVWtUB>.
- [58] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: Visions and challenges, 2024. URL <https://arxiv.org/abs/2407.00128>.
- [59] Chen Xu, Zitian Guo, and Chenyan Xiong. Mechanism design for generative engines: From exploitation toward Win-Win outcomes, 2026. URL <https://arxiv.org/abs/2608.11390>.
- [60] Hengwei Ye, Jiasheng Mao, Zhenhan Guan, and Zheng Tian. EcoGEO: Trajectory-Aware evidence ecosystems for Web-Enabled LLM search agents, 2026. URL <https://arxiv.org/abs/2605.12887>.
- [61] Junwei Yu, Mufeng Yang, Yepeng Ding, and Hiroyuki Sato. Structural feature engineering for generative engine optimization: How content structure shapes citation behavior, 2026. URL <https://arxiv.org/abs/2603.29979>.
- [62] Xucheng Yu, Haibo Jin, Huimin Zeng, and Haohan Wang. SCI-Defense: Defending manipulation attacks from generative engine optimization, 2026. URL <https://arxiv.org/abs/2605.21948>.
- [63] Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. AgenticGEO: A Self-Evolving agentic system for generative engine optimization, 2026. URL <https://arxiv.org/abs/2603.20213>.
- [64] Faye Zhang, Qianyu Cheng, Jasmine Wan, Vishwakarma Singh, Jinfeng Rao, and Kofi Boakye. Generative engine optimization: A VLM and agent framework for pinterest acquisition growth, 2026. URL <https://arxiv.org/abs/2602.02961>.
- [65] Kai Zhang, Xinyue He, and Jingang Yao. From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms, 2026. URL <https://arxiv.org/abs/2604.25707>.
- [66] XinYu Zhao, ChengYou Li, XiangBao Meng, Kai Zhang, and XiaoDong Liu. Beyond retrieval: Modeling confidence decay and deterministic agentic platforms in generative engine optimization, 2026. URL <https://arxiv.org/abs/2604.03656>.
- [67] Heyang Zhou, JiaJia Chen, Xiaolu Chen, Jie Bao, Zhen Chen, and Yong Liao. IF-GEO: Conflict-Aware instruction fusion for Multi-Query generative engine optimization, 2026. URL <https://arxiv.org/abs/2601.13938>.

A Versioned Research Map

Key idea

This appendix is a routing layer between a versioned paper library, the Notes, and the course. Inclusion means that a record has entered the review queue. It does not mean that every claim in the paper has been accepted, replicated, or absorbed into the main text.

A.1 Snapshot and provenance

The GEO-42 library is derived from the user-provided Zotero CSV with SHA-256

```
c149a69140b0dcedc1855f8cbb02fc5cb1e1346470306c76cd68fbafc2e50468.
```

The source CSV is preserved read-only. A normalization script produces a JSON catalog, this table, and the BibTeX file. The snapshot contains 42 unique normalized titles, 42 DOI fields, 41 URL and abstract fields, and 42 local PDF attachments.

A.2 Version and status controls

Before a figure, table, or exact numerical result is admitted to the manuscript, the local PDF version is compared with the current official paper record. The first audit found eight local PDFs behind the current arXiv version: records 2, 11, 12, 13, 17, 23, 25, and 35. Record 23 is especially important: the local SAGEO Arena v1 still contains template placeholders, whereas the current arXiv record is v2 and reports an updated acceptance status. The table flags all eight records; they may inform taxonomy, but no old-version figure or result enters the Notes.

Publication status is a separate field. “Has a DOI,” “is on arXiv,” “was accepted at a workshop,” and “appeared in a main conference or journal” are not collapsed. Status is verified against an official venue or current paper record before it appears in public-facing course metadata.

A.3 Initial editorial route

Table 45: The GEO-42 research library and its initial editorial route. Routes are curatorial decisions, not paper conclusions.

#	Paper	Route and editorial state
1	Cultural Encoding in Large Language Models: The Existence Gap in AI-Mediated Brand Discovery (2025) 16	Platform empirics Ch. 1 source year 2025 differs from arXiv-ID year 2026
2	Multimodal Generative Engine Optimization: Rank Manipulation for Vision-Language Model Rankers (2026) 13	Security Ch. 10 current arXiv v2 locked 2026-08-24

Continued on the next page

Table 45 continued from the previous page

#	Paper	Route and editorial state
3	IF-GEO: Conflict-Aware Instruction Fusion for Multi-Query Generative Engine Optimization (2026) 67	Intervention Ch. 9 metadata ingested
4	Controlling Output Rankings in Generative Engines for LLM-based Search (2026) 21	Intervention Ch. 9 metadata ingested
5	Generative Engine Optimization: A VLM and Agent Framework for Pinterest Acquisition Growth (2026) 64	Multimodal Ch. 10 metadata ingested
6	When Search Engine Services meet Large Language Models: Visions and Challenges (2024) 58	Systems foundation Ch. 1 metadata ingested
7	Adversarial Search Engine Optimization for Large Language Models (2024) 34	Security Ch. 11 metadata ingested
8	AI Answer Engine Citation Behavior An Empirical Analysis of the GEO16 Framework (2025) 24	Intervention Ch. 6 metadata ingested
9	Beyond SEO: A Transformer-Based Approach for Reinventing Web Content Optimisation (2025) 27	Intervention Ch. 9 metadata ingested
10	CC-GSEO-Bench: A Content-Centric Benchmark for Measuring Source Influence in Generative Search Engines (2025) 6	Benchmark Ch. 4 metadata ingested
11	Dynamics of Adversarial Attacks on Large Language Model-Based Search Engines (2025) 15	Security Ch. 11 current arXiv v3 locked 2026-08-24
12	E-GEO: A Testbed for Generative Engine Optimization in E-Commerce (2025) 3	Intervention Ch. 6 current arXiv v2 locked 2026-08-24
13	Exposing Citation Vulnerabilities in Generative Engines (2025) 30	Security Ch. 5 current arXiv v2 locked 2026-08-24

Continued on the next page

Table 45 continued from the previous page

#	Paper	Route and editorial state
14	From Noise to Narrative: Building Reputation in AI Dominated Search and Discovery Environments (2025) 39	Context Ch. 2 missing URL; missing abstract
15	Generative Engine Optimization: How to Dominate AI Search (2025) 5	Systems Ch. 2 metadata ingested
16	GEO: Generative Engine Optimization (2024) 1	Foundation Ch. 1 source year 2024 differs from arXiv-ID year 2023
17	MaxShapley: Towards Incentive-compatible Generative Search with Fair Context Attribution (2025) 35	Systems Ch. 4 current arXiv v2 locked 2026-08-24
18	MillStone: How Open-Minded Are LLMs? (2025) 48	Systems Ch. 4 metadata ingested
19	Position: On the Risks of Generative Engine Optimization in the Era of LLMs (2025) 53	Risk and agenda Ch. 11 metadata ingested
20	What Evidence Do Language Models Find Convincing? (2024) 50	Evidence behavior Ch. 2 metadata ingested
21	What Generative Search Engines Like and How to Optimize Web Content Cooperatively (2025) 57	Intervention Ch. 8 metadata ingested
22	When Attention Becomes Exposure in Generative Search (2026) 2	Systems Ch. 4 metadata ingested
23	SAGEO Arena: A Realistic Environment for Evaluating Search-Augmented Generative Engine Optimization (2026) 23	End-to-end benchmark Ch. 4 current arXiv v2 locked 2026-08-24
24	Diagnosing and Repairing Citation Failures in Generative Engine Optimization (2026) 47	Diagnosis and repair Ch. 8 metadata ingested

Continued on the next page

Table 45 continued from the previous page

#	Paper	Route and editorial state
25	Caption Injection for Optimization in Generative Search Engine (2026) 7	Systems Ch. 5 current arXiv v4 locked 2026-08-24; source year 2026 differs from arXiv-ID year 2025
26	Role-Augmented Intent-Driven Generative Search Engine Optimization (2026) 8	Systems Ch. 3 source year 2026 differs from arXiv-ID year 2025
27	Beyond Retrieval: Modeling Confidence Decay and Deterministic Agentic Platforms in Generative Engine Optimization (2026) 66	Intervention Ch. 3 metadata ingested
28	Structural Feature Engineering for Generative Engine Optimization: How Content Structure Shapes Citation Behavior (2026) 61	Intervention Ch. 8 metadata ingested
29	Don't Measure Once: Measuring Visibility in AI Search (GEO) (2026) 40	Measurement Ch. 6 metadata ingested
30	AgenticGEO: A Self-Evolving Agentic System for Generative Engine Optimization (2026) 63	Systems Ch. 3 metadata ingested
31	Think Before Writing: Feature-Level Multi-Objective Optimization for Generative Citation Visibility (2026) 26	Intervention Ch. 9 metadata ingested
32	From Experience to Skill: Multi-Agent Generative Engine Optimization via Reusable Strategy Learning (2026) 56	Measurement Ch. 7 metadata ingested
33	From Citation Selection to Citation Absorption: A Measurement Framework for Generative Engine Optimization Across AI Search Platforms (2026) 65	Measurement Ch. 5 metadata ingested
34	How LLMs Are Persuaded: A Few Attention Heads, Rerouted (2026) 46	Security Ch. 11 current arXiv v1 locked 2026-08-26

Continued on the next page

Table 45 continued from the previous page

#	Paper	Route and editorial state
35	EcoGEO: Trajectory-Aware Evidence Ecosystems for Web-Enabled LLM Search Agents (2026) 60	Intervention Ch. 8 current arXiv v2 locked 2026-08-24
36	What Gets Cited: Competitive GEO in AI Answer Engines (2026) 49	Intervention Ch. 7 metadata ingested
37	SCI-Defense: Defending Manipulation Attacks from Generative Engine Optimization (2026) 62	Security Ch. 11 metadata ingested
38	Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines (2026) 25	Systems Ch. 1 metadata ingested
39	SafeGEO: Understanding Generative Engine Optimization Risks in Recommendation Agents (2026) 52	Intervention Ch. 11 metadata ingested
40	GPE: Evaluating Robust Evidence Aggregation for Fact Verification under Controllable GEO-Style Poisoning (2026) 51	Security Ch. 11 metadata ingested
41	GEO-Flag: Detecting and Measuring GEO-Optimized Web Content (2026) 9	Intervention Ch. 11 metadata ingested
42	Mechanism Design for Generative Engines: From Exploitation toward Win-Win Outcomes (2026) 59	Mechanism design Ch. 11 metadata ingested

A.4 Absorption states

Each record eventually receives one of four states:

Core

Its verified contribution is necessary to the main argument and is discussed in the Notes and course.

Supporting

It supplies a method, comparison, counterexample, or extension used in one chapter or lab.

Frontier

It is recent, promising, or unresolved and appears in a versioned research-frontier box rather than as stable doctrine.

Catalog only

It remains discoverable with an explicit reason for not being substantively absorbed, such as scope mismatch, missing method detail, stale artifact, or insufficient evidence.

The public Reading Atlas will expose both route and absorption state. This makes it possible to distinguish “we have the paper” from “this idea has survived editorial review and entered the curriculum.”

B Reproducibility and Release Protocols

B.1 Minimum experiment manifest

Every empirical figure or table in the Notes has a manifest containing:

1. research question, estimand, and preregistered decision rule;
2. repository commit, environment lock, random seeds, and execution date;
3. dataset identity, license, source hash, exclusions, and splits;
4. query panel, intent clusters, weights, and development/holdout boundary;
5. system, model, endpoint, interface, account, locale, and version state;
6. intervention and comparator hashes;
7. raw evaluation cells, missingness, labeling rubric, and adjudication;
8. analysis script, uncertainty method, robustness checks, and failures;
9. figure-generation script and exact output hash;
10. evidence, safety, privacy, and release approvals.

B.2 Figure checklist

-
- ☐ The caption states the unit, population, system, date, treatment, comparator, metric, and uncertainty.
 - ☐ Every visual encoding is explained without relying only on color.
 - ☐ Axis limits, missing values, and exclusions are visible.
 - ☐ A script or source file regenerates the figure from frozen inputs.
 - ☐ External content has a source, version, figure/page, license, crop, modification status, and alt text in the source ledger.
 - ☐ The figure remains legible in grayscale and at final print width.
 - ☐ The PDF build has no clipped labels, broken glyphs, or unresolved references.
-

The audit in [figure 20](#) is a computational replay example. Regenerating frozen arithmetic and graphics establishes internal traceability only; it must not be described as an independent empirical replication unless a separately governed study independently recollects or reproduces the relevant observations.

B.3 Closed-platform observation card

For a closed product, every observation reports at least:

Field	Required value
Surface	product and specific mode, page, API, or feature
Time	timestamp and timezone
Context	locale, language, geography/endpoint, account state, device
Input	exact prompt, conversation history, uploaded content, and settings
Repetition	run count, spacing, and retry/error policy
Output	response, citations, UI state, and permissible trace
Interpretation	observation, proxy, inference, and alternative explanations
Retention	storage, privacy, platform-term, and redaction decision

B.4 Manuscript release gate

An arXiv candidate is released only when:

- every citation key and cross-reference resolves;
- bibliography identities and high-impact paper versions are verified;
- all figures pass provenance, license, alt-text, and print checks;
- every quantitative claim is traceable to an exact table, figure, page, dataset, or regenerated analysis;
- claims about closed systems include a versioned observation card;
- no TODOs, template placeholders, hidden author notes, or stale outputs remain in the source package;
- the PDF passes full-page rendering, font, metadata, link, overflow, and accessibility-oriented review;
- the source package builds in a clean TeX Live environment without proprietary fonts or undeclared local files.

These checks are necessary but not sufficient for scientific validity. A final release also requires domain review of the argument, methods, and risk claims.

C Standards and Official-Guidance Register

C.1 Purpose and admission rule

This register identifies standards and official guidance that can constrain a GEO research or deployment workflow. It is not a compliance checklist. An entry is admitted only when the issuer, identifier, edition or publication date, status, and official identity page have been verified. The register records a document's scope; it does not reproduce paywalled clauses or infer that a particular organization, system, or intervention conforms.

The register deliberately separates published international and national standards, industry standards, public technical recommendations, government guidance, legal measures, registered projects, and consultation drafts. These classes have different normative force. A W3C Recommendation is not an ISO management-system requirement; a voluntary NIST framework is not a statute; a recommended Chinese standard is not automatically a mandatory rule; and a registered project or zero draft is not a published consensus standard.

Table 46: Status-first register at the research cutoff. *Use* names a course or manuscript control, not a claim of conformity.

Identity	Verified status	Relevant scope	Bounded use in this project
NIST AI 100-1, AI RMF 1.0	Final publication, 2023; voluntary framework	AI risk governance, mapping, measurement, and management across the lifecycle	Risk register, role assignment, measurement plan, and release review; not a certificate or legal safe harbor [31]
NIST AI 600-1	Final profile, 2024; companion to AI RMF 1.0	Risks that are novel to or exacerbated by generative AI and suggested actions	Threat-model prompts and control coverage for Labs 7–8; not evidence that a specific product has implemented the actions [32]
ISO/IEC 42001:2023	Published International Standard, edition 1	Requirements for an AI management system and continual improvement	Governance-process crosswalk and evidence-of-control inventory; no clause-level conformity claim without licensed text and qualified audit [18]
ISO/IEC 23894:2023	Published International Standard, edition 1	Guidance for integrating AI-specific risk management into organizational activities	Risk taxonomy and treatment-record crosswalk; not a substitute for sector or jurisdictional analysis [17]
W3C PROV-O	W3C Recommendation, 30 April 2013	Interoperable representation and exchange of provenance information	Vocabulary reference for Claim–Evidence–Source lineage; not a complete evidence-quality or truth model [54]
WCAG 2.2	W3C Recommendation; current Recommendation dated 12 December 2024	Testable web-content accessibility success criteria	Course-site and public-artifact acceptance criteria; conformance still requires automated and human evaluation [55]
GB/T 43782–2024	Chinese recommended national standard; current	Technical requirements for machine-learning systems	System-card prompts for controlled pipelines; not a direct standard for GEO content or commercial answer-engine ranking [41]
GB/T 45288.1–2025	Chinese recommended national standard; current	General requirements for large-scale AI models	Model and system evaluation context for agentic and generative surfaces; not a publisher optimization rule [43]
GB/T 47507–2026	Published Chinese recommended national standard; official effective date 1 August 2026	General principles for trustworthy AI	Trustworthiness vocabulary and governance crosswalk. Reverify the live status label before release because cached official search text may lag the effective date [45]

Table 47: RAG-specific official metadata and status watch. No clause-level requirement is inferred from a title or public identity page.

Identity	Verified status at cutoff	Bounded use	Prohibited inference
SJ/T 12116–2026	Current Chinese recommended electronic-industry standard; MIIT; published and effective 28 April 2026; filing 107753–2026	Metadata-only RAG technical-requirements status card [28]	No clause, metric, threshold, conformity, certification, or product-scope claim before lawful full-text review
SJ/T 12117–2026	Current Chinese recommended electronic-industry standard; MIIT; published and effective 28 April 2026; filing 107764–2026	Metadata-only RAG evaluation-specification status card [29]	The title does not reveal a test corpus, metric, threshold, applicable product, or proof that GEO works
Project 20252041-Z-469	Registered Chinese national guiding-technical-document project; registered 20 June 2025; official status “under approval” at the cutoff	Standards-lifecycle counterexample [44]	Not a published standard; no final identifier, clause, effective date, or conformity claim
NIST public-facing AI documentation “Zero Draft”	Initial public draft, 30 July 2026; preliminary NIST proposal open to input at the cutoff	Candidate documentation-template and draft-to-standardization lifecycle example [33]	Not consensus, law, certification, or evidence that public documentation makes a system trustworthy or compliant

Table 48: Additional status-verified governance and disclosure objects. Legal applicability remains a separate review.

Identity	Verified status	Relevant scope	Bounded use in this project
ISO/IEC 42005:2025	Published International Standard, edition 1; May 2025	Guidance for organizational AI-system impact assessment across foreseeable effects on individuals, groups, and society	Impact-assessment dossier and stakeholder review prompts; public metadata does not supply clause-level conformity evidence [19]
GB 45438–2025	Current mandatory Chinese national standard; published 28 February 2025; effective 1 September 2025	Technical methods for explicit and implicit labeling of AI-generated and synthetic content	Disclosure and metadata test cases within its applicable scope; no ranking, citation, truth, or authorship claim [42]
AI-generated-content labeling measures	Joint official measure; issued 7 March 2025; effective 1 September 2025	Duties for specified generation and distribution service providers, including explicit and implicit labels	Role-and-artifact applicability card; not a universal duty for every author, research prototype, locale, or publishing action [12]
Interim Measures for Generative AI Services	Official Order No. 15; effective 15 August 2023	Specified services that provide generated text, images, audio, or video to the public in mainland China, with stated exclusions	Jurisdiction and provider-role gate; legal review is required before operational advice [11]

C.2 Applicability card

Before a standard enters a claim, lab rubric, or release decision, the analyst completes the following card:

Field	Required record
Identity	issuer, identifier, full title, edition, language, and official URL
Status	draft, recommendation, published standard, withdrawn, superseded, or law
Temporal scope	publication, effective, review, withdrawal, and access dates
Institutional force	voluntary guidance, contractual requirement, certification basis, recommended standard, mandatory rule, or other declared class
Applicability	organization, system, role, lifecycle stage, geography, sector, and excluded cases
Exact control	licensed clause or public section used, with a paraphrase and artifact-level acceptance test
Evidence	document, log, decision, or measurement that demonstrates the control
Reviewer	domain owner and, where needed, qualified legal, security, privacy, accessibility, or conformity reviewer
Residual boundary	what the mapped control does not establish

C.3 Course crosswalk

Artifact	Standards route	Acceptance question
Claim–Evidence–Source graph	W3C PROV-O plus the project evidence rubric	Can every derivation, revision, and responsible agent be reconstructed without treating provenance as proof of truth?
Closed-system observation card	NIST AI RMF and NIST AI 600-1	Are context, affected parties, failure modes, measurements, and treatment decisions explicit?
Intervention release record	ISO/IEC 42001 and ISO/IEC 23894	Are ownership, risk treatment, monitoring, corrective action, and retained evidence integrated into a repeatable process?
Model/system card	GB/T 43782–2024 and GB/T 45288.1–2025	Does the artifact name the technical object, boundary, evaluation context, and limitations instead of converting general requirements into ranking claims?
Public course artifact	WCAG 2.2	Can a keyboard and screen-reader user obtain the same information, status, and controls, with content still understandable without color alone?
Trustworthiness review	GB/T 47507–2026, NIST AI RMF, and ISO/IEC 23894	Are trustworthiness terms mapped to testable project controls and residual risks rather than used as promotional adjectives?
AI-system impact dossier	ISO/IEC 42005:2025 plus the project stakeholder map	Are foreseeable individual, group, and societal effects documented throughout the lifecycle without claiming conformity from public metadata alone?
Generated-content disclosure	GB 45438–2025 and the 2025 labeling measures, with C2PA considered separately where applicable	Are role, jurisdiction, content type, explicit label, embedded metadata, distribution state, test evidence, and exceptions recorded without treating provenance as truth?
RAG standard-status card	SJ/T 12116–2026, SJ/T 12117–2026, and project 20252041-Z-469	Are a current industry standard, an evaluation specification, and an under-approval national project kept in separate status classes, with clause-level use blocked until lawful full-text review?

Boundary and failure conditions

The register is versioned evidence, not legal advice. Standards may be revised, superseded, adopted into contracts, or incorporated into law after the research cutoff. Public release requires rechecking every official identity page and the rules applicable to the actual jurisdiction, sector, organization, and system.

D Exercises and Reproduction Problems

Key idea

These problems turn the Notes into an inspectable research apprenticeship. Each chapter contributes one conceptual or derivational problem and one empirical, experimental, or audit problem. A successful submission states the estimand, preserves denominators and versions, reports nulls and failures, and stops at the claim ceiling of the design. Completing a problem is not evidence that a commercial platform implements the mechanism used in the exercise.

D.1 Problem contract and submission conventions

The 24 problems are original to these Notes. They can be solved independently of a particular vendor, paid endpoint, or unpublished dataset. Problems marked “C” emphasize concepts or derivations; problems marked “R” require a small reproduction, experiment design, or audit. Every response must include a one-sentence result, an uncertainty or ambiguity statement, and the strongest claim that the work *does not* establish.

For an empirical problem, submit a compact research bundle with `manifest.yaml`, immutable input hashes, a tidy cell-level result file, an analysis script or calculation sheet, one decision memo, and a `limitations.md`. The full-compute path may use an open retriever or generator. The low-compute path is always valid and is designed for a laptop or pencil-and-paper audit. Synthetic examples must be labeled synthetic; missing cells remain in the denominator ledger rather than disappearing during cleanup.

Table 49: Coverage map for the 24-problem sequence. “Sketch” identifies the problem with a solution outline later in this appendix.

Chapter	Concept or derivation	Reproduction, experiment, or audit	Sketch
1	Conditional visibility chain	Visibility-profile denominator audit	1C
2	Evidence adequacy and claim ceilings	Claim–Evidence–Source audit	2C
3	Query fan-out stopping rule	Earliest-failure trace localization	3C
4	Rank fusion and context allocation	Retrieval-stage loss audit	4C
5	Citation and absorption estimands	Counterfactual source-influence audit	5C
6	Repeated-measure uncertainty	Benchmark protocol audit	6C
7	Cluster effects and temporal change	Drift-aware preregistration audit	7C
8	Constrained content decisions	Claim-preserving revision audit	8C
9	Multi-query constrained optimization	Optimizer trace audit	9C
10	Provenance dependence across modalities	Paired platform-surface design	10C
11	Defense utility and false positives	Controlled red-team audit	11C
12	Control evidence versus conformity	Incident evidence-preservation drill	12C

D.2 Chapter 1: objects and conditional visibility

Problem D.1: 1C – Diagnose a conditional visibility chain

For one query–source stratum, suppose the estimated conditional stage probabilities in equation (2) are

$$(p_D, p_R, p_C, p_G, p_T, p_A) = (0.90, 0.60, 0.75, 0.80, 0.50, 0.20).$$

Compute the joint probability of the all-stage outcome. Then derive the elasticity of the joint probability with respect to each positive stage probability. Compare two proposed interventions: intervention a changes p_R from 0.60 to 0.69; intervention b changes p_T from 0.50 to 0.60. Hold the other conditionals fixed only for this calculation. State why the comparison is a sensitivity exercise rather than an identified forecast. Finally, construct a two-stratum example in which the intervention with the larger pooled change harms the lower-baseline stratum.

Required inputs. The six conditional probabilities, the two hypothetical changes, and a declared weighting rule for the two-stratum counterexample.

Required outputs. The baseline joint probability, both counterfactual arithmetic changes, elasticities, the subgroup table, and a four-sentence interpretation.

Forbidden extrapolation. Do not infer independence, a platform funnel, a causal intervention effect, or a downstream business effect from the multiplication alone.

Low-compute path. Use a calculator or show exact products and choose any two synthetic strata whose weighted averages demonstrate the reversal.

Problem D.2: 1R – Build a denominator-safe visibility profile

Create or use a synthetic cell table with 12 queries, two repeats, and the following columns: `query_id`, `intent`, `eligible`, `retrieved`, `cited`, `absorbed`, `attribution_correct`, `claim_faithful`, `timestamp`, and `missing_reason`. Define the denominator for each component of $\mathbf{V}$ before inspecting the outcomes. Estimate

each observable component by intent and overall, attach an interval appropriate to the small sample, and report temporal disagreement between the two repeats. Show how a complete-case filter changes at least one estimate. Explain why a single arithmetic mean of the profile is not a neutral summary.

Required inputs. A 24-row synthetic or permissioned table, a data dictionary, a preregistered missing-state policy, and query weights.

Required outputs. A denominator ledger, stratified profile, interval table, complete-case sensitivity analysis, and one bounded result paragraph.

Forbidden extrapolation. Do not call citation retrieval, call mention fidelity, or present the panel as the visibility of a brand, language, geography, or platform population.

Low-compute path. Use six queries and two repeats; compute Wilson intervals with a spreadsheet or report raw fractions plus their exact denominators.

D.3 Chapter 2: claims, evidence, and source identity

Problem D.3: 2C – Reason with a bottleneck evidence gate

Let the five components of $A(c)$ in [equation \(9\)](#) be scored on $[0,1]$. Claim c_1 has component vector $(0.9, 0.8, 0.4, 0.9, 0.7)$ and claim c_2 has $(0.65, 0.65, 0.65, 0.65, 0.65)$. Rank the claims under the minimum gate, the arithmetic mean, and a release rule requiring every component to be at least 0.60. Prove that the minimum gate is monotone in every component but cannot reward an improvement to a non-minimum component until the bottleneck changes. Give one example in which adding a new source increases coverage while reducing consistency. Explain which part is mathematics and which part is an editorial rubric.

Required inputs. The two score vectors, definitions of the three decision rules, and one explicit rule for updating consistency when evidence conflicts.

Required outputs. All rankings, the short monotonicity proof, a conflicting-evidence example, and a release recommendation with a named unresolved field.

Forbidden extrapolation. Do not interpret the scores as calibrated probabilities of truth, source prestige, or scientific validity.

Low-compute path. Complete the arithmetic by hand; for the conflict example use one claim and two evidence edges.

Problem D.4: 2R – Audit a Claim–Evidence–Source packet

Choose one non-sensitive, time-bounded factual claim from an official public page and one related statement from a research, practitioner, or vendor item. Freeze both source objects. Decompose the target into a scoped claim tuple, record each source identity, extract only the evidence span needed for the claim, and label every edge as supporting, qualifying, contradicting, or non-informative. Assign an evidence-layer ceiling and list every unknown field. Repeat the audit after changing one scope element, such as date, locale, product version, or comparator.

Required inputs. Two lawfully accessible source snapshots, hashes or archival identities, access dates, a target claim, and a declared evidence-edge rubric.

Required outputs. Two identity records, two scoped claim tuples, a typed CES graph, evidence ceilings, an unknown-field register, and a before/after scope diff.

Forbidden extrapolation. Do not treat an official product description as effectiveness evidence, a paper abstract as full-method verification, or provenance as proof of truth.

Low-compute path. Use one claim, two sources, and a hand-drawn three-node graph; omit automation and preserve unresolved fields explicitly.

D.4 Chapter 3: acquisition, planning, and pipeline traces

Problem D.5: 3C – Apply a constrained fan-out stopping rule

A query planner proposes five subqueries in a fixed order. Their marginal unique-evidence gains are $(8, 5, 2, 1, 3)$ units and marginal costs are $(1, 1, 1, 1, 3)$ units. Set $\epsilon_E = 2$ and $\tau = 1.5$. First apply the rule in Chapter 3 literally. Then impose two coverage constraints: at least one counterevidence item must be recovered, first appearing at subquery 4, and at least two provenance origins must be represented, first achieved at subquery 3. State precisely whether a threshold is tested before or after executing a proposed subquery. Show why the fifth proposal creates a non-monotone marginal-gain sequence and why a greedy stopping rule can miss it.

Required inputs. The ordered gains and costs, both thresholds, the two coverage constraints, and an explicit pre-test or post-test convention.

Required outputs. The stopping index under each convention, cumulative coverage and cost, a constraint-feasibility check, and one alternative look-ahead rule.

Forbidden extrapolation. Do not claim that the subqueries were issued by a closed product or that more fan-out necessarily improves answers.

Low-compute path. Solve the five-row table by hand; no retrieval model or API is required.

Problem D.6: 3R – Localize the earliest observable failure

Construct five synthetic end-to-end traces containing source eligibility, query fan-out, retrieval ranks, reranker ranks, packed passages, answer claims, displayed citations, and action state. Inject one acquisition failure, one segmentation failure, one packing displacement, one generation omission, and one wrong-source attribution. For each trace, identify the earliest *observed* failing stage and distinguish it from latent or reconstructed explanations. Add a positive control whose evidence survives every observable stage.

Required inputs. A frozen toy corpus, five query–source traces plus one positive control, passage hashes, and stage-specific pass/fail definitions.

Required outputs. A trace bundle, stage-transition matrix, earliest-failure labels, unresolved alternatives, and one diagnostic decision tree.

Forbidden extrapolation. Do not infer hidden candidate scores, context order, causal mechanisms, or organic discovery from a displayed answer or citation alone.

Low-compute path. Use six index cards or a spreadsheet with one row per trace and manually specified ranks; no generator is needed.

D.5 Chapter 4: retrieval, fusion, and context allocation

Problem D.7: 4C – Compare rank fusion with evidence-aware packing

Two retrievers return $L_1 = (A, B, C, D)$ and $L_2 = (C, A, E, B)$. Compute RRF scores and the fused ordering for $\kappa = 60$, using document ID to break exact ties. Prove that RRF is unchanged by any strictly increasing transformation of the underlying raw scores that preserves each list order, whereas raw-score addition need not be. Now consider passages A, C, E with $(v_i, \ell_i) = (7, 5), (6, 4), (4, 2)$, evidence sets $\{x\}, \{y\}, \{x, y\}$, and budget $B = 6$. With coverage reward $\alpha = 3$ and no redundancy penalty, enumerate feasible packs and select the optimum under equation (24). Then state how an origin cap could reverse the decision.

Required inputs. Both ranked lists, κ , the passage utilities, lengths, evidence sets, budget, tie rule, and objective coefficients.

Required outputs. RRF table, invariance argument, complete feasible-pack table, optimum, and one origin-cap counterexample.

Forbidden extrapolation. Do not equate fused rank with answer contribution, evidence coverage with truth, or the toy optimum with a production context policy.

Low-compute path. Enumerate at most eight subsets by hand and retain fractions to six decimal places for tie checking.

Problem D.8: 4R – Audit stage-conditioned retrieval loss

Using an open or synthetic corpus, define 12 queries, material evidence qrels, and an immutable segmentation. Compare lexical retrieval, dense retrieval, and RRF at one fixed depth. Apply one transparent reranking rule and two context budgets. Track evidence recall, nDCG, origin diversity, and material-evidence coverage at every transition. Inject a syndicated duplicate cluster and one long passage containing a necessary qualification. Attribute each loss to the first observed transition where it occurs.

Required inputs. Corpus and passage hashes, 12 queries, graded and material-evidence labels, three retrieval traces, one reranking rule, two budgets, and origin IDs.

Required outputs. Metric table with denominators, transition Sankey or tidy loss table, duplicate audit, qualification survival record, intervals across queries, and failures.

Forbidden extrapolation. Do not convert a reconstructed retrieval gain into a claim about commercial acquisition, final answer quality, or user behavior.

Low-compute path. Use six queries and ten passages; replace the dense model with a supplied ranked list and enumerate context selection in a spreadsheet.

D.6 Chapter 5: generation, citation, and influence

Problem D.9: 5C – Separate citation entailment, completeness, and absorption

An answer contains four material claims. Claims 1–3 are supported by source S ; claim 4 is supported only by source T . Displayed citation S follows claims 1 and 4, citation T follows claim 3, and claim 2 has no citation. A source-specific phrase from S appears faithfully in claims 1 and 2. Under an explicit span-assignment rule, compute citation entailment, citation completeness, displayed-citation rate, and S -specific absorption. Report every denominator. A binary entailment judge has sensitivity 0.90 and specificity 0.80 on a representative calibration

set. Derive the standard matrix correction for an observed positive rate $\tilde{p}$ and evaluate it at $\tilde{p} = 0.65$. State the feasibility condition for the corrected estimate.

Required inputs. The four claim–source relations, displayed citation placements, an assignment rule, the absorption definition, and judge sensitivity and specificity.

Required outputs. Four separate metric estimates with denominators, the correction derivation and value, a feasibility check, and a failure interpretation.

Forbidden extrapolation. Do not infer retrieval, factual truth, unique authorship, or platform-wide citation quality from the constructed answer.

Low-compute path. Use the four claims as the full table and perform the correction with a basic calculator.

Problem D.10: 5R – Audit counterfactual source influence

Create a fixed context with three sources, two provenance origins, and six verified material claims. Generate or manually construct a baseline answer and leave-one-source-out answers under a locked rubric. Segment answer claims, resolve citations, and score support, completeness, and source-specific absorption independently. Include one wrong-source citation and one omitted qualification. Compare leave-one-source-out changes with leave-one-origin-out changes.

Required inputs. A frozen context, source and origin IDs, six claim labels, answer construction rule or locked generator, removal conditions, and a double-label rubric.

Required outputs. Claim-level comparison table, displayed-citation and absorption matrices, source- and origin-level deltas, disagreements, uncertainty, and failure cases.

Forbidden extrapolation. Do not interpret a changed answer as proof of truth, ownership, necessary causation outside the fixed context, or the behavior of an unobserved retriever.

Low-compute path. Write three short answers manually from a six-claim fact sheet and have a second reader adjudicate only the two injected failures.

D.7 Chapter 6: measurement and benchmark design

Problem D.11: 6C – Quantify repeated-measure pseudo-replication

Suppose 30 query clusters are each evaluated with four repeated generations in each of two conditions. Within a condition, approximate the intraclass correlation among repeats from the same query as $\rho = 0.35$. Compute the nominal response count, the design effect $1 + (r - 1)\rho$, and the corresponding effective sample size when only repeat-level dependence is considered. Explain why this correction still does not account for clustering across related queries, system drift, or paired treatment assignment. Design a minimal hierarchical model that separates query, time, and residual variation and name the unit for the primary contrast.

Required inputs. Cluster count, repeats, conditions, ρ , treatment pairing information, and definitions of the response and primary contrast.

Required outputs. Design effect, effective sample size, naïve-versus-clustered warning, a model formula with indexed random components, and declared analysis unit.

Forbidden extrapolation. Do not claim that effective sample size is literal information content or that the stated ρ transports to another query panel or system.

Low-compute path. Perform the design-effect calculation by hand and write the model symbolically; no mixed-model software is required.

Problem D.12: 6R – Repair an under-specified benchmark protocol

Audit this intentionally defective protocol: “Use 50 popular prompts, tune the rewrite until average GEO score improves, retry errors, remove refusals, and let one LLM judge quality.” Identify at least 12 distinct threats spanning target population, estimand, leakage, treatment version, missingness, repetition, system state, judge validity, multiplicity, uncertainty, guardrails, and stop rules. Rewrite the protocol as an executable manifest and construct a tiny planned-versus-observed cell ledger containing at least one error, refusal, and late response.

Required inputs. The defective protocol, a benchmark-manifest schema, a synthetic query frame, and a declared resource and ethics boundary.

Required outputs. Threat register, corrected manifest, cell ledger, missing-state policy, judge calibration plan, primary and guardrail metrics, and decision rule.

Forbidden extrapolation. Do not claim that protocol repair validates the metric, judge, sample, or treatment effect; it only makes the study auditable.

Low-compute path. Use eight synthetic prompts, two conditions, and two repeats; calibrate the judge plan rather than calling an external model.

D.8 Chapter 7: experiments and temporal uncertainty

Problem D.13: 7C – Distinguish cluster effects from temporal change

Four equal-weight query clusters are randomized in pairs: clusters A and B receive treatment, while C and D remain controls. Their post-period cluster means are $(0.70, 0.50, 0.45, 0.35)$ and their pre-period means are $(0.40, 0.45, 0.40, 0.30)$. Compute the post-only difference in means and the difference-in-differences estimate. Show the contribution of each cluster and state the potential-outcome estimand corresponding to the randomization. Then give one treatment-uptake pattern and one interference pattern under which the two arithmetic estimates cease to have the intended interpretation.

Required inputs. Assignment, cluster weights, pre/post outcomes, treatment-version definition, and assumptions for the post-only and temporal contrasts.

Required outputs. Both estimates, cluster contribution table, estimand statement, uptake and interference counterexamples, and an identification verdict.

Forbidden extrapolation. Do not call the difference-in-differences estimate causal without parallel trends and stable treatment/control definitions, or generalize beyond the four synthetic clusters.

Low-compute path. Complete the eight-cell calculation by hand and draw the two counterexamples as arrows between clusters.

Problem D.14: 7R – Preregister a drift-aware intervention

Design a synthetic study with 12 intent clusters, baseline and intervention pages, a concurrent control, three time windows, and three planned repeats. Preregister treatment hashes, assignment, primary estimand, minimum worthwhile effect, missing-state handling, system-state comparability window, guardrail thresholds, and stop rule.

Inject one platform-state break, one contaminated control cluster, one missing-not-at-random refusal pattern, and one placebo date. Produce the analysis and decision that the protocol requires even when the result is inconclusive.

Required inputs. A synthetic cluster frame, treatment versions, assignment, planned cells, state-break rules, outcomes, guardrails, and four declared injections.

Required outputs. Preregistration, planned/observed flow, cluster estimates and intervals, missingness table, placebo and interference checks, exclusions with reasons, and retain/revert/inconclusive decision.

Forbidden extrapolation. Do not repair comparability after seeing results, recycle contaminated clusters, or describe an unidentified before/after difference as a treatment effect.

Low-compute path. Use four clusters and two time windows; calculate cluster means and mark the state-break cell as non-comparable rather than fitting a time-series model.

D.9 Chapter 8: content and evidence engineering

Problem D.15: 8C – Select under fidelity and risk constraints

Four revisions have metric vectors (V, F, U, R, C) for visibility, fidelity, user utility, risk, and normalized cost:

revision	V	F	U	R	C
r_0	0.50	0.98	0.70	0.04	0.10
r_1	0.62	0.96	0.72	0.05	0.18
r_2	0.68	0.88	0.76	0.04	0.15
r_3	0.61	0.97	0.69	0.09	0.12

Admit only candidates with $F \geq 0.95$, $U \geq 0.70$, $R \leq 0.06$, and $C \leq 0.20$. Identify the feasible set, Pareto relations among feasible candidates, and the release decision if the minimum worthwhile visibility gain over r_0 is 0.08. Then show why a weighted sum can select an inadmissible revision unless constraints are enforced before optimization.

Required inputs. The metric table, threshold directions, baseline, minimum worthwhile effect, and an explicit Pareto dominance convention.

Required outputs. Feasibility table, Pareto set, release decision, and one weighted-sum counterexample with declared weights.

Forbidden extrapolation. Do not treat synthetic scores as calibrated user benefit, proof of evidence quality, or a universal trade-off rate among the five dimensions.

Low-compute path. Inspect the four rows by hand and choose any nonnegative weights that make the counterexample transparent.

Problem D.16: 8R – Produce a claim-preserving revision diff

Start from a one-page synthetic evidence sheet containing six approved claims, two qualifications, one table, and three sources. Diagnose one passage-level retrieval or absorption failure. Produce two expression-preserving revisions and one evidence-surfacing revision. For every version, align claim spans to the approved claim ledger, compare rendered text, and test whether units, dates, population, comparator, uncertainty, and exceptions remain co-located. Evaluate all versions under a fixed manually supplied retrieval and answer trace.

Required inputs. A synthetic source page, approved claim ledger, source and passage IDs, fixed query set, supplied stage traces, and a claim-equivalence rubric.

Required outputs. CES diff, rendered diff, claim confusion table, stage-specific metrics, failure analysis, version hashes, and retain/revert decision.

Forbidden extrapolation. Do not add a stronger claim to improve retrieval, infer organic discovery, or report the manual trace as an end-to-end platform experiment.

Low-compute path. Use three claims, one qualification, four queries, and manually assign retrieval/absorption outcomes to each version.

D.10 Chapter 9: optimization and agentic systems

Problem D.17: 9C – Prevent average gain from hiding query harm

Three query clusters have weights $(0.5, 0.3, 0.2)$. Relative to baseline, candidate a changes visibility by $(0.10, 0.08, -0.12)$, candidate b by $(0.06, 0.05, 0.03)$, and candidate c by $(0.15, -0.04, 0.01)$. Compute each weighted mean. The release gate requires mean gain at least 0.04, no cluster loss below -0.02 , fidelity at least 0.95, and risk at most 0.05; the respective (F, R) pairs are $a : (0.97, 0.03)$, $b : (0.96, 0.04)$, and $c : (0.94, 0.02)$. Determine eligibility and explain why the best average is not necessarily the release. Propose a minimax or regret objective that exposes subgroup conflict.

Required inputs. Cluster weights and changes, fidelity and risk values, all release thresholds, and a tie rule.

Required outputs. Weighted means, cluster-harm checks, full gate table, release decision, and a formal alternative objective.

Forbidden extrapolation. Do not interpret a weighted query panel as the user population, or synthetic visibility gain as causal, durable, or commercially valuable.

Low-compute path. Calculate the three dot products by hand; express the alternative objective without running an optimizer.

Problem D.18: 9R – Audit an optimizer’s complete candidate trace

Construct a 20-candidate trace for a bounded edit space with proposal parent, edit type, development score, validation score, holdout status, fidelity, risk, cost, judge version, and rejection reason. Inject repeated holdout access, a silent judge change, two null candidates, one high-development/low-validation candidate, one guardrail violation, and one duplicate proposal. Reconstruct the candidate funnel and determine whether any selected release remains valid after quarantining contaminated evaluations.

Required inputs. A typed change space, 20-row synthetic trace, fixed splits, budgets, gates, judge identities, and the six injected audit events.

Required outputs. Lineage graph, funnel, contamination register, Pareto plot or table, cost ledger including rejects, null/harm table, and signed release/revert decision.

Forbidden extrapolation. Do not present the surviving candidate as globally optimal, autonomous proof, or evidence that an agentic loop should hold release authority.

Low-compute path. Use ten candidates and a spreadsheet sort/filter; draw the lineage as a simple parent list rather than running an agent.

D.11 Chapter 10: multimodal and platform ecosystems**Problem D.19: 10C – Correct a provenance-dependent evidence count**

Six assets appear to support a claim: three pages syndicate one original report, two images derive from one shared dataset, and one independent regulatory record supplies a qualification. First compute the naive source count and the number of provenance origins. Define a binary asset–origin incidence matrix and show why treating six assets as independent overstates corroboration. Then construct a cross-modal support matrix for text, pixels, caption, alt text, and metadata with one deliberate conflict. Propose an ablation sequence that can localize representation contribution without erasing the conflict.

Required inputs. Six asset identities, three origin identities, dependency edges, five representation channels, one scoped claim, and one qualification.

Required outputs. Naive and origin-aware counts, incidence and support matrices, dependency graph, ablation order, and an unresolved-conflict statement.

Forbidden extrapolation. Do not call origin count an effective statistical sample size, infer truth from independence, or infer a platform’s multimodal weights from the ablation plan.

Low-compute path. Draw the two matrices by hand with binary entries and compare six assets with three origins; no vision model is required.

Problem D.20: 10R – Design a paired platform-surface observation

Design a non-destructive paired observation across two versioned surfaces or two explicit open-system configurations. Use eight intent-matched queries, two languages, two repeats, and both a dynamic preset alias and an explicitly frozen caller-visible configuration where the interface permits it. Predefine entity aliases, answer and citation outcomes, missing states, account and locale controls, and measurement-equivalence tests. Estimate within-query contrasts and surface-by-language interaction descriptively.

Required inputs. Observation cards, exact prompts, query-pair map, language review, surface and configuration identities, timestamps, outputs, and a missing/retry policy.

Required outputs. Cell archive, disclosure table, paired contrasts with intervals or raw pairs, equivalence audit, configuration-drift log, and bounded transfer statement.

Forbidden extrapolation. Do not infer hidden retrieval or ranking mechanisms, merge dynamic and frozen configurations, or attribute surface differences to brand value.

Low-compute path. Use four bilingual query pairs and manually supplied synthetic outputs; report pairwise signs and magnitudes without fitting a model.

D.12 Chapter 11: manipulation, defense, and mechanism design**Problem D.21: 11C – Evaluate defense utility with false positives**

Two defenses are evaluated on 100 benign and 40 attacked cases. Defense *A* blocks 32 attacks and 18 benign cases; defense *B* blocks 26 attacks and 4 benign cases. Compute attack recall, benign false-positive rate, precision of a block decision, and residual attack rate for each defense. Define $U = 3TP - 2FP - 5FN$ and select under that declared utility. Then perform a sensitivity analysis in which the false-negative cost ranges from 1 to 8 and identify any preference reversal. Explain why the utility function is a policy choice, not an empirical law.

Required inputs. Both confusion matrices, class counts, metric definitions, utility weights, and the false-negative sensitivity range.

Required outputs. All metrics with denominators, utility values, reversal threshold or interval, and a decision that reports benign utility loss separately.

Forbidden extrapolation. Do not infer robustness to unknown attacks, transport class prevalence, or claim that blocking accuracy establishes end-user safety.

Low-compute path. Use a four-cell confusion table and evaluate integer false-negative costs from 1 through 8 with a calculator.

Problem D.22: 11R – Run a bounded red-team and defense audit

In an isolated synthetic pipeline, compare a benign clarification, an unsupported persuasion condition, a prompt-like instruction embedded in content, a duplicated-origin consensus condition, and a competitor-displacement condition. Use only fictional entities and inert strings. Test two transparent defenses and preserve benign utility, evidence recall, attribution, attack success, false positives, and abstentions. Include one attack family withheld from defense tuning and one simulated rollback failure.

Required inputs. Synthetic corpus and entities, threat model, attacker capability budget, five high-level conditions, two defenses, outcomes, safety boundary, and stop rules.

Required outputs. Adversarial test manifest, per-family results, benign-utility table, known-versus-withheld comparison, residual-risk decision, and rollback postmortem.

Forbidden extrapolation. Do not publish operational payloads, test real targets without authorization, generalize from known attacks, or treat rejection rate as joint utility.

Low-compute path. Use paper cards representing 20 cases, one rule-based defense, and one manual review defense; simulate model outcomes without executing adversarial content.

D.13 Chapter 12: governance, standards, and incidents

Problem D.23: 12C – Distinguish a control map from conformity evidence

Consider four rows in a governance crosswalk: (1) official identity and status verified, but no project control; (2) a control exists, but has no owner or test; (3) owner, artifact, and passing test exist, but applicability is unresolved; (4) applicability, control, owner, artifact, test, exception path, and refresh date are all recorded. Define a completeness predicate over these fields and classify each row. Show formally that provenance completeness is necessary for reconstructing the decision but not sufficient for truth, legal applicability, or conformity. Add one counterexample in which every field is populated yet the test measures the wrong object.

Required inputs. The four rows, required field set, a status-first source identity, and an explicit completeness predicate.

Required outputs. Row classifications, logical necessity/sufficiency argument, wrong-object counterexample, unresolved-review flags, and bounded governance conclusion.

Forbidden extrapolation. Do not label a crosswalk, checklist, public metadata record, or internal test as certification, compliance, legal advice, or a safe harbor.

Low-compute path. Represent each row as an eight-bit vector and evaluate the predicate by hand.

Problem D.24: 12R – Preserve evidence through an incident drill

Run a tabletop exercise for a released synthetic evidence page whose displayed answer suddenly attributes a stale high-impact claim to the wrong source. At time zero, the team has an immutable release manifest but incomplete downstream cache visibility. Assign accountable, responsible, consulted, and informed roles. Preserve the observation, assess scope, contain the affected release, correct the claim and dependent artifacts, notify according to a fictional policy, verify recovery, and conduct a blameless postmortem. Inject one failed rollback and one source-status change during the exercise.

Required inputs. Synthetic release bundle, CES graph, role map, monitoring alert, observation card, rollback procedure, notification rule, and two injected events.

Required outputs. Timestamped incident log, evidence holds, scope graph, decision records, correction-propagation matrix, rollback evidence, residual-risk register, refresh tasks, and postmortem.

Forbidden extrapolation. Do not infer real legal duties, conceal uncertainty, overwrite the original evidence, or claim recovery before all declared acceptance tests pass.

Low-compute path. Use one claim, two dependent artifacts, four roles, and a paper timeline; mark legal and external-notification questions for qualified review.

D.14 Selected solution sketches

The sketches below cover the conceptual problem in each chapter. They identify the central reasoning step and a reference result, not the only acceptable argument. Empirical problems deliberately have no single numerical answer: their reference output is the contract, traceability, and claim boundary stated in the problem.

Sketch for Problem D.1 (1C)

The baseline joint probability is $0.90(0.60)(0.75)(0.80)(0.50)(0.20) = 0.0324$. Because $P = \prod_j p_j$, $\partial \log P / \partial \log p_j = 1$ for every positive conditional; equal *proportional* changes have equal local proportional effects. Intervention a gives 0.03726, an absolute gain of 0.00486; intervention b gives 0.03888, a gain of 0.00648. The calculation holds other conditionals fixed and therefore does not account for treatment-induced dependence or selection. For the subgroup reversal, any weighted construction in which a large gain in a high-weight, high-baseline stratum masks a negative change in the other stratum is sufficient; the harm must remain visible rather than being canceled in the pooled result.

Sketch for Problem D.3 (2C)

The minimum scores are 0.40 and 0.65, while the means are 0.74 and 0.65. Thus the mean ranks c_1 first, but the bottleneck gate and the 0.60 componentwise release rule prefer only c_2 . If $m(\mathbf{x}) = \min_j x_j$ and $x'_j \geq x_j$ for every j , then $m(\mathbf{x}') \geq m(\mathbf{x})$, proving monotonicity. Increasing a component strictly above the current minimum leaves the set's minimum unchanged. A new source can cover a missing subclaim while contradicting the existing value or time scope, so coverage rises and consistency falls. The function and scores are an editorial release rubric; they are not a truth-probability model.

Sketch for Problem D.5 (3C)

The gain/cost ratios are 8, 5, 2, 1, and 1. Under a post-test convention and the literal “gain below 2 or ratio below 1.5” rule, subquery 4 triggers stopping because its gain is 1, below $\epsilon_E = 2$, and its gain/cost ratio is 1, below $\tau = 1.5$. The coverage constraints make execution through subquery 4 necessary: origin diversity is first met at 3 and counterevidence at 4. A pre-test convention needs a predicted gain and must say whether the proposal that fails the threshold is executed. Since gain returns to 3 at query 5, the marginal sequence is not monotone. A bounded two-step look-ahead or a reserved coverage query can avoid treating the first low marginal gain as proof that all later proposals are unproductive.

Sketch for Problem D.7 (4C)

The RRF scores are $A = 1/61 + 1/62$, $C = 1/63 + 1/61$, $B = 1/62 + 1/64$, $E = 1/63$, and $D = 1/64$, giving $A > C > B > E > D$. RRF uses only rank positions, so any strictly increasing score transform that preserves each ordering leaves every term unchanged. Raw addition changes under shifts or rescaling of one retriever. Under budget 6, feasible nonempty packs are A , C , E , and $C + E$. Their objective values are respectively 10, 9, 10, and 16 when a covered evidence type earns 3 once; $C + E$ is optimal. If C and E share an origin and its cap is one, that pack becomes infeasible, illustrating that a governance constraint can alter the arithmetic optimum.

Sketch for Problem D.9 (5C)

Under the stated local placement rule, only displayed S after claim 1 is entailed; S after claim 4 and T after claim 3 are not. Entailment is $1/3$ of displayed claim–citation links, completeness is $1/4$ of material claims, the displayed-citation rate is $3/4$ of claims, and S -specific absorption is $2/3$ of claims supported by S . With sensitivity Se and specificity Sp , $\tilde{p} = Se p + (1 - Sp)(1 - p)$, so $p = (\tilde{p} + Sp - 1)/(Se + Sp - 1) = 0.45/0.70 \approx 0.643$. Feasibility requires $Se + Sp > 1$, a corrected value in $[0, 1]$, and representative calibration.

Sketch for Problem D.11 (6C)

There are $30 \times 4 \times 2 = 240$ response cells. For four repeats, $DE = 1 + 3(0.35) = 2.05$, giving the rough repeat-adjusted effective size $240/2.05 \approx 117.1$. This is only a warning against treating 240 correlated responses as independent. A minimal model is $Y_{iqtr} = \mu + \tau T_i + a_q + b_t + (aT)_{qi} + \varepsilon_{iqtr}$, with query and time components chosen to match the design; the primary contrast should be formed at the randomized or paired query-cluster unit. Related-query clusters, drift, and assignment dependence require design-specific treatment beyond the single intraclass-correlation approximation.

Sketch for Problem D.13 (7C)

The treated post mean is $(0.70 + 0.50)/2 = 0.60$ and the control post mean is $(0.45 + 0.35)/2 = 0.40$, so the post-only difference is 0.20. Treated change is $[(0.70 - 0.40) + (0.50 - 0.45)]/2 = 0.175$; control change is $[(0.45 - 0.40) + (0.35 - 0.30)]/2 = 0.05$, so difference-in-differences is 0.125. The randomization estimand is the finite-sample average effect of assignment to the declared treatment version across the four clusters. If a treated cluster never receives the revision, assignment and uptake effects diverge. If treated A changes evidence retrieved for control C , stable-unit assumptions fail. The temporal estimate additionally needs a credible parallel-trends argument.

Sketch for Problem D.15 (8C)

Rows r_0 and r_1 pass every threshold. Row r_2 fails fidelity; r_3 fails utility and risk. Relative to r_0 , r_1 gains 0.12 visibility, above the 0.08 minimum, while preserving the declared gates, so it is the release candidate subject to uncertainty and protocol checks. Within the feasible set, neither row dominates the other because r_1 improves visibility and utility while r_0 has higher fidelity, lower risk, and lower cost. A weighted sum placing sufficient weight on visibility and utility can rank r_2 first; this does not repair its failed fidelity constraint. Feasibility precedes preference aggregation.

Sketch for Problem D.17 (9C)

The weighted gains are 0.050 for a , 0.051 for b , and 0.065 for c . Candidate a fails the subgroup-loss gate because $-0.12 < -0.02$; candidate c fails both the second-cluster loss and fidelity gates. Candidate b clears the mean, cluster, fidelity, and risk thresholds and is therefore the only eligible release. One transparent alternative is to maximize $\min_k \Delta V_k$ subject to the other gates, or minimize maximum weighted regret relative to the best feasible candidate for each cluster. The example shows why the largest weighted mean is not equivalent to admissibility or stakeholder acceptability.

Sketch for Problem D.19 (10C)

The naive count is six assets, but the dependency graph contains three origins: one report, one dataset, and one independent regulatory record. The incidence matrix has one row per asset and one column per origin; the three syndicated pages occupy the same report column and the two images the same dataset column. It preserves access paths without pretending they are independent corroborations. The support matrix should distinguish support, silence, and conflict across representation channels. Ablate one representation at a time, then coherent channel groups, while retaining a conflict-only condition. A change under ablation is conditional contribution in this construction, not a truth test or a disclosed platform weight.

Sketch for Problem D.21 (11C)

For A , $(TP, FP, FN, TN) = (32, 18, 8, 82)$; recall is $32/40 = 0.80$, benign FPR is $18/100 = 0.18$, precision is $32/50 = 0.64$, and residual attack rate is $8/40$. For B , the tuple is $(26, 4, 14, 96)$, yielding 0.65, 0.04, $26/30 \approx 0.867$, and $14/40$. At false-negative cost 5, $U_A = 20$ and $U_B = 0$, so A wins under the declared policy. With false-negative cost c , the difference $U_A - U_B = 3(6) - 2(14) + c(6) = 6c - 10$, so A is preferred for $c > 5/3$ and B for $c < 5/3$. The reversal is a consequence of stakeholder loss weights, not a property that the confusion matrices determine alone.

Sketch for Problem D.23 (12C)

Let $K(r)$ be true only when identity/status, applicability, control, owner, artifact, test, exception path, and refresh date are populated with the correct object types. Rows 1–3 are incomplete for the missing fields named in the prompt; row 4 is structurally complete. Completeness permits a reviewer to reconstruct who mapped what, to which version, with what evidence. It does not entail that the source is true, applicable law has been interpreted correctly, or a conformity body would accept the control. A fully populated row whose test checks a staging page while the release concerns a production API satisfies a syntactic predicate but measures the wrong object.

Semantic review and qualified applicability judgment therefore remain separate gates.

Boundary and failure conditions

The empirical problems specify auditable work; they do not claim that the work has already been run, independently reproduced, or approved for a real target. Before public release, each reference output needs a versioned artifact bundle, license review, safety review, and an independent check against the stated problem contract.

E Notation, Symbols, and Terminology

Key idea

Notation in these Notes is a type system, not decoration. A symbol is usable only after its domain, unit, epistemic status, conditioning set, and chapter scope are known. In particular, an observed interface object is not silently substituted for a hidden platform state, an estimator is not its estimand, and a displayed citation is not evidence of source use or factual support.

E.1 How to read this appendix

This appendix is the normative notation and terminology contract for the English manuscript. It consolidates symbols that first appear throughout the twelve chapters and records a small amount of compatibility debt retained in the current working edition. It does not turn a conceptual variable into an observable one, validate a proxy, or add evidence to an empirical claim.

Every mathematical object is described along four independent axes:

1. **domain:** the set of values the object may take;
2. **unit:** the entity, claim, source, passage, response, user task, time interval, or other unit on which it is defined;
3. **epistemic status:** observed, constructed, estimated, latent, or counterfactual;
4. **scope:** the query, system, corpus, surface, locale, time, and protocol under which the object has meaning.

An object can be observed and still require a constructed label. For example, the answer text is observed, whereas whether one of its claims is faithfully supported is a reviewed measurement. Conversely, a quantity can be precisely defined while remaining latent on a closed surface. Precision of definition must not be mistaken for availability of data.

E.1.1 Epistemic-status legend

Code	Status	Meaning in these Notes	Minimum reporting requirement
O	Observed	Captured directly at the declared interface or in a controlled trace	identity, timestamp, protocol cell, raw value, and missing-state rule
C	Constructed	Produced by a documented transformation or human labeling rule	input version, transformation or guide, reviewer, and reproducibility
E	Estimated	Computed from sampled or assigned observations	estimand, estimator, sampling or assignment unit, uncertainty, and missingness
L	Latent / proxy-sensitive	Not directly exposed in the study setting	the observable proxy, assumptions, alternatives, and an explicit non-observation statement
CF	Counterfactual	Defined under a treatment or state not jointly observed for the same unit	treatment versions, assignment or identification argument, and assumptions
N	Normative / governed	Set by a policy, protocol, approval, or decision rule rather than learned from response data	owner, version, applicability, threshold, exception, and review date

E.2 House conventions**E.2.1 Random objects, realizations, estimands, and estimates**

Uppercase Roman letters normally denote random objects and lowercase letters their realized values. Thus Y is a response-valued random variable and y is one realized response; $\mathcal{A}$ is a random set of displayed citation objects and $\mathcal{A}$ is the realized set. This convention is applied when the distinction affects an argument. A fixed source identifier s or query object q remains lowercase because it is an input to the relevant protocol.

Greek letters without hats may denote parameters or estimands. A hat denotes an estimator or estimate: τ is a treatment estimand and $\hat{\tau}$ is an estimate produced by a declared analysis. A tilde is reserved for a noisy, proxy, or pre-correction measurement when that distinction is explicit, as in $\tilde{p}$ for an observed judge-positive rate. A subscript names the population, stage, stratum, or protocol; it must not be used merely to decorate an otherwise undefined score.

Boundary and failure conditions

The manuscript never uses a hat to imply truth or causal identification. An estimate can be exactly computed from the available cells and still estimate the wrong population, condition on treatment-dependent missingness, or rely on an invalid identification assumption.

E.2.2 Sets, sequences, vectors, and indicators

Calligraphic capitals denote sets, spaces, collections, or stochastic kernels, such as $\mathcal{D}_t$, $\mathcal{W}_t$, and $\mathcal{G}_\theta$. Parentheses preserve order or conditioning when order is material; braces denote an unordered set unless the text states otherwise.

Bold symbols such as $\mathbf{J}$ and $\mathbf{V}$ are vectors. Their components retain separate units and may be aggregated only after scaling, weighting, and non-compensable constraints have been declared.

The indicator $\mathbb{I}\{A\}$ equals one when event A holds and zero otherwise. A graded label is not converted to an indicator without a recorded threshold. Cardinality is $|S|$ for a finite set S ; token, financial, latency, or capability budgets are named explicitly rather than inferred from a bare number.

E.2.3 Conditioning, time, and versioning

Every probability is conditional on the variables shown in its expression or on a versioned evaluation cell referenced in the surrounding text. Omitting a conditioning variable for readability does not license averaging over an unknown platform, locale, time, or account distribution. The timestamp t is one observation time; T is a declared window. A provider label, dynamic API alias, or copied prompt is not by itself a version lock.

E.3 Master symbol register

The locator column names the first substantive use. Later uses inherit the same meaning unless an entry is explicitly marked *local compatibility use*. Units are semantic units, not merely numeric data types.

Table 50: System, source, and evidence objects.

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
q	text or structured request; query-task	O	User request supplied to the surface. It is not necessarily the string sent to retrieval.	chapter 1
u	versioned user-state record	O / L	Account, locale, conversation, and personalization state; only captured fields are observed.	chapter 1
t, T	timestamp; analysis window	O / N	One observation time and a declared target or monitoring window.	chapter 1
$\mathcal{W}_t$	set of accessible source representations	L in closed systems; O in a frozen stack	Information environment accessible to the system at t ; not identical to the public Web.	chapter 1
θ	configuration / version record	O / L	The identifiable portion of system state. Unexposed backend state remains latent.	chapter 1
$\mathcal{G}_\theta$	conditional stochastic kernel	C	Research model of a generative-search system; not a reverse-engineered commercial architecture.	chapter 1
Y, y	response-valued random object; realized response	O at the surface	Y is the conditional output object; y is one captured realization.	chapter 1
$\mathcal{A}, \vdash$	set of citation / attribution objects	O after resolution	Random and realized displayed attributions. Neither proves source use or support.	chapter 1
L, ℓ	interaction or action trace	O subject to consent and retention	Random and realized observable trace; it excludes imagined internal calls.	chapter 1
s	governed source ID; source artifact	O / C	Version-resolved source, not merely a URL string or publisher name.	chapter 1

Continued on next page

Table 50 continued

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
c	claim ID; scoped proposition	C	Minimal reviewed proposition with scope, time, locale, and status.	chapter 2
e^{ev}	evidence ID; evidence unit	C	Relevant span, table, record, dataset, or observation linked to a claim. The superscript is the preferred disambiguation of the legacy bare e .	chapter 2
$I(s)$	source-identity record	C / O	Issuer, type, title, version, status, method, artifacts, license, and URL; identity does not establish claim support.	chapter 2
$A(c)$	adequacy gate output	C / N	Project editorial bottleneck over declared evidence dimensions; not a universal scientific quality score.	chapter 2

Table 51: Acquisition, retrieval, generation, and attribution objects.

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
$\hat{q}$	emitted retrieval-query object	O in a logged stack; L on a closed surface	One query produced by a query plan. The hat marks a derived query here, not a statistical estimate.	chapter 3
$\mathcal{Q}(q, u)$	ordered or dependency-aware query collection	O / L	Fan-out queries conditioned on request and state. It must preserve query order or dependencies when they affect evidence acquisition.	chapter 3
$\Pi(d)$	ordered or set-valued representation transform	O in a controlled stack	Versioned passages or representations derived from document d .	chapter 4
$\mathcal{D}_t$	frozen corpus snapshot	O / C	Candidate corpus with time, hash, license, and segmentation version.	chapter 4
$r_{\text{ret}}(d, q)$	real-valued retrieval score	O in a controlled stack	Score under a named retriever. It is not a probability and need not be comparable across queries or retrievers.	chapter 3
$R_k(q)$	ordered top- k list	O in a controlled stack	Retrieved representations for q at declared depth k .	chapter 4
g_{qi}	usually $\{0, 1, 2, 3\}$; query-candidate	C	Preregistered graded relevance judgment for candidate i . The scale and guide are part of the measurement.	chapter 4
C	ordered effective context	O in a logged stack; L on a closed surface	Passages or representations actually available to generation, including order and truncation.	chapter 3
$A(y)$	set of atomic material claims	C	Smallest support-assessable claims in answer y , retaining quantities, scope, conditions, time, and attribution.	chapter 5

Continued on next page

Table 51 continued

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
$R(y)$	set of resolved citation objects	O / C	Displayed anchors resolved with canonical identity, version, access time, and exact passage when available. Resolution does not imply support.	chapter 5
E_y	set of claim–citation pairs	C	Visible or rule-declared attachment edges for answer y . An edge records attachment, not evidentiary support.	chapter 5
w_j	nonnegative claim-weight units	N / C	Preregistered materiality weight for atomic claim a_j ; the scale and sensitivity analysis are disclosed.	chapter 5
C^{-s}	ordered counterfactual context	C	Context formed by removing source s while recording every order, budget, position, replacement, and truncation change.	chapter 5
U, U_i	declared answer-utility units	C / E	Preregistered utility for a context contrast. The Orchid Bay binary fixture is local and cannot be exported as a universal utility.	chapter 5
B_{ctx}	tokens or capability units	N / O	Declared context or tool budget. Use a typed subscript; do not reuse it for an edit or attack budget.	chapter 3
$Z_{\text{ctx}}(s)$	binary or graded source–response event	O / L	A representation of source s enters the effective context.	chapter 5
$Z_{\text{use}}(s)$	binary or graded source–response event	E	A named ablation or attribution method assigns material answer influence to s .	chapter 5
$Z_{\text{abs}}(s)$	binary or graded claim–source–response event	C / E	Preregistered source-distinctive claims appear faithfully.	chapter 5
$Z_{\text{cite}}(s)$	binary or graded source–response event	O / C	A displayed citation resolves to source s .	chapter 5
S_{jk}	support label; claim–passage pair	C	Human or calibrated-judge assessment that passage k supports answer claim j under a stated guide.	chapter 5
$\Delta_s(C)$	answer-utility units	E	Conditional leave-one-source-out contrast in constructed context C ; not authorship or global platform use.	chapter 5

Table 52: Measurement, causal, intervention, and governance objects.

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
Q, M	finite weighted panels or target distributions	N / E	Query and system populations. Capital M denotes a distribution here; use M_{gen} for a generator model.	chapter 6

Continued on next page

Table 52 continued

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
w_q, w_m	nonnegative normalized weights	N / E	Design or target weights for queries and systems; their provenance must be stated.	chapter 6
e^{cell}	versioned evaluation cell	C / O	Tuple of query, system, version, time, locale, geography, account, repeat, and surface.	chapter 6
$\mu, \hat{\mu}$	outcome units	estimand / E	Population mean and its estimate. A weighted target mean differs from an equal-stratum diagnostic mean.	chapter 6
D_i	binary or multivalued assignment	O	Treatment version assigned to experimental unit i ; not a detector.	chapter 7
S_t	versioned system state	O / L	State under which potential outcomes are defined; the unobserved portion remains explicit.	chapter 7
$Y_i(d, S_t)$	outcome units	CF	Potential outcome for unit i under treatment d and system state S_t ; only the assigned state is observed.	chapter 7
$\tau, \hat{\tau}$	declared outcome units	estimand / E	Treatment effect target and its estimate. Population, comparison, and time state are part of the subscript or surrounding protocol.	chapter 7
x, x'	versioned source states	O / C	Released baseline and proposed or candidate source state.	chapter 8
δ	typed change record	O / C	Permitted intervention between source states, including claim IDs, evidence IDs, owner, scope, and rollback.	chapter 9
$\mathcal{X}_{\text{perm}}(x)$	set of source states	N / C	Candidate revisions allowed by evidence, policy, safety, accessibility, and release gates.	chapter 9
$\mathbf{J}(x')$	vector; component-specific units	E / N	Retrieval, citation, absorption, fidelity, utility, risk, and cost objectives. Components are not averaged without declared scales and weights.	chapter 9
$B_{\text{edit}}, B_{\text{atk}}$	edits, tokens, cost, or capability units	N	Optimization and attacker budgets. The unit must accompany every value.	chapter 9, chapter 11
a^{asset}	asset ID	O / C	Multimodal asset used inside $\Pi(a^{\text{asset}})$; distinct from account state.	chapter 10
$\Pi(a^{\text{asset}})$	set of versioned representations	O in a controlled stack; L otherwise	Pixel, caption, alt/OCR, metadata, and page representations derived from the governed asset; only the declared accessible subset is scored.	chapter 10
$B(y), C_b(y)$	entity IDs; scoped claim set	C	Resolved brand entities in answer y and claims attached to governed entity b . Ambiguous and unresolved strings remain explicit error states.	chapter 10

Continued on next page

Table 52 continued

Symbol	Domain / unit	Status	Meaning and restrictions	First locator
$\mathbf{V}_b$	seven-component outcome vector	E / C	Mention, correctness, citation, absorption, prominence, action, and separately defined risk. No composite exists without declared scaling and decision weights.	chapter 10
$o(s)$, $N_{\text{origin}}(S)$	origin-cluster mapping and count	C	Earliest verified provenance origin for source s and the number of distinct mapped origins in S ; both depend on trace completeness.	chapter 10
$G = (V, E)$	directed provenance graph	O / C	Claims, sources, publishers, versions, and material derivation edges. Coverage depends on the tracing protocol.	chapter 10
$D_{\text{det}}(\delta)$	detector-score units	E	Declared detector output for intervention δ ; threshold and calibration are part of its meaning.	chapter 11
$\Gamma(r_{\text{rel}})$	structured release record	O / N	Identity, changed claims, evidence, risks, controls, approvals, monitoring, and rollback for a concrete release.	chapter 12

E.4 Compatibility map for overloaded symbols

The current edition retains several locally overloaded symbols so that the chapter sources remain traceable to earlier drafts. New derivations and future revisions use the typed forms below. A local chapter definition remains valid, but it does not propagate beyond its stated scope.

Legacy form	Existing local meanings	Preferred typed forms	Rule
e	evidence item; evaluation cell	$e^{\text{ev}}; e^{\text{cell}}$	Never join the two tables on bare e .
B	context, edit, tool, or attacker budget	$B_{\text{ctx}}, B_{\text{edit}}, B_{\text{tool}},$ B_{atk}	Every value states its unit; budgets of different kinds are not added.
a	multimodal asset; account state; local answer-claim index	$a^{\text{asset}}, a^{\text{acct}}, a_j^{\text{claim}}$	Local indices do not become globally identified assets or accounts.
M	generator model; target system distribution	$M_{\text{gen}}; M$ for the distribution	A model object is not a random draw from an unspecified system population.
D	treatment assignment; detector	$D_i; D_{\text{det}}(\delta)$	Detector output is not treatment status.
s	source artifact; locally, surface	$s; s_{\text{surf}}$	A source ID and a product surface must not share a database field.
r	repeat index; retrieval score; release object	$r_{\text{rep}}, r_{\text{ret}}, r_{\text{rel}}$	Infer meaning only from an explicit local definition, never from letter shape.
C	effective context; claim set or cluster count in local derivations	$C; \mathcal{C}_s; n_{\text{cl}}$	Context order is preserved; a set of claims and a count of clusters use different types.

The derived query $\hat{q}$ is the one intentional exception to the usual “hat means estimator” convention. It is retained because the chapters use it to distinguish the user’s request from an emitted retrieval query. Its domain and trace status must always be stated.

E.5 Forbidden substitutions

The following pairs are related but not interchangeable. The right-hand column gives the minimum repair when a draft sentence or analysis crosses the boundary.

Do not substitute	For	Required repair
crawler request or successful fetch	indexed, retrieved, cited, mentioned, or acted upon	Name the observed acquisition event and mark downstream stages unobserved.
document contains a claim	relevant evidence	Name the rendering, segmentation, representation, query, and retrieval depth.
retrieval score	passage is recoverable	Report the retriever and score scale; calibrate only against declared judgments.
top- k inclusion	calibrated relevance probability	Report the retriever and score scale; calibrate only against declared judgments.
displayed citation	effective-context inclusion	Preserve reranking, packing, order, token budget, and truncation evidence.
mention rate	source use, claim support, truth, authorship, or authority	Resolve the citation, segment the claim, label support, and use a separate influence design when needed.
model-judge label	correctness, recommendation, sentiment, prominence, or business value	Report each outcome and denominator separately.
number of response rows	human ground truth	Calibrate by stratum, preserve disagreement, and propagate plausible error.
before–after change	number of independent units	Identify the sampling and assignment cluster; treat repeats as nested measurements.
statistical detectability	treatment effect	Define treatment versions, comparator, assignment or identification assumptions, uptake, and concurrent changes.
several URLs	practical or ethical success	Predeclare a minimum worthwhile effect and fidelity, utility, and risk gates.
standard or policy identity	several independent origins	Trace derivation and report origin clusters.
platform or model name	applicability, conformity, compliance, or control effectiveness	Verify status and scope, obtain qualified applicability review, and attach implementation and test evidence.
	frozen system state	Record surface, version evidence, request, locale, account, time, search state, and all unknowns.

E.6 Cross-chapter locator

Ch.	Primary objects	Main distinction introduced	Later dependency
1	$q, u, t, \mathcal{W}_t, \theta, Y, \mathcal{A}, L$	system object versus realized interface observation; distinct visibility events	conditions every measurement and intervention
2	$s, c, e^{\text{ev}}, I(s), A(c)$	source identity versus claim support; provenance versus truth	evidence gates in Chapters 5, 8, 11, and 12
3	$\hat{q}, \mathcal{Q}(q, u), r_{\text{ret}}, C, B_{\text{ctx}}$	observed trace versus reconstructed query or hidden context	retrieval and generation protocols
4	$\mathcal{D}_t, \Pi(d), R_k(q), g_{qi}$	document versus passage evidence; rank score versus relevance judgment	benchmark and context-allocation metrics
5	$S_{jk}, Z_{\text{ctx}}, Z_{\text{use}}, Z_{\text{abs}}, Z_{\text{cite}}$	selection, use, absorption, support, and display	citation audit and source-influence designs
6	$Q, M, w_q, w_m, e^{\text{cell}}, \hat{\mu}$	target estimand versus sample estimate; planned versus observed denominator	experiment design and dashboards
7	$D_i, Y_i(d, S_t), \tau, \hat{\tau}$	assignment, potential outcome, estimand, and estimator	causal and temporal claims
8	x, x', δ	claim-preserving expression change versus a new claim	release validation and optimization
9	$\mathcal{X}_{\text{perm}}, \mathbf{J}, B_{\text{edit}}$	vector objective versus non-compensable constraints	agentic proposal and selection
10	$a^{\text{asset}}, \Pi(a), G = (V, E)$	modality representation versus asset; URL count versus origin dependence	multimodal and ecosystem audits
11	$B_{\text{atk}}, D_{\text{det}}, \mathbf{U}$	attack outcome versus detector measurement and honest-source utility	risk gates and monitoring
12	$\Gamma(r_{\text{rel}})$	institutional source status versus applicable, tested project control	release, incident, correction, and errata

E.7 Terminology glossary

The glossary defines how a term is used in this manuscript. It is not a claim that every platform, paper, or professional community uses the term identically.

E.7.1 Generative search and visibility

Generative search surface

The specific product mode, page, API, or interface at which a generated response is observed. A provider may expose several non-equivalent surfaces.

Generative search system

The conditional stochastic object in [chapter 1](#); it may search, retrieve from a controlled collection, use tools, rely on parametric knowledge, or combine these paths.

Generative engine optimization (GEO)

The study and governed intervention of source visibility, use, attribution, and fidelity in generative-search settings. In these Notes, GEO is neither a guarantee of exposure nor a synonym for content persuasion.

Visibility profile

A vector of separately estimated outcomes such as retrieval, citation, absorption, attribution, fidelity, and temporal variation. It is not a universal composite score.

Mention

A resolved entity occurs in a response under a declared alias policy. Mention does not establish correctness, source use, or sentiment.

Citation

A displayed attribution object resolves to a source. Citation is an interface event, not a support or causal-use judgment.

Attribution

A relation that assigns an answer claim or contribution to a displayed or inferred source under a stated resolver or method.

Absorption

Faithful presence of preregistered, source-distinctive evidence in an answer. Its unit is claim–source–response, not URL count.

Fidelity

Preservation of the scoped proposition, including material units, conditions, exceptions, and time bounds.

Prominence

A declared positional or share property of an entity or span in a response. It is not equivalent to user attention.

Action

A predeclared user–task event such as a visit or qualified request, measured under consent and a suitable identification design.

E.7.2 Information retrieval and RAG**Acquisition**

The processes by which a representation becomes discoverable, fetchable, renderable, eligible, or otherwise available. These are distinct states.

Representation

A versioned object derived from a source, such as rendered text, a passage, image, caption, metadata record, or embedding.

Passage

A retrievable segment produced by a declared segmentation procedure. It may omit conditions present elsewhere in its source document.

Retrieval

Candidate recovery and ordering for a query over a frozen or versioned corpus. Retrieval is measured before answer generation.

Lexical retrieval

Retrieval based on discrete term evidence, commonly an inverted index and a term-frequency / document-frequency score.

Dense retrieval

Retrieval based on learned vector representations and a declared similarity function under a locked encoder.

Hybrid retrieval

A protocol that combines lexical and dense evidence through normalization, calibration, rank fusion, or a learned rule. Raw score addition is not the default meaning.

Query fan-out

A dependency-aware sequence or set of retrieval queries derived from one user request, with cost and stopping rules.

Rank fusion

Combination of ordered candidate lists. Fusion retains recovering-query and provenance paths so repetition is not treated as independent confirmation.

Reranking

Conditional reordering of a candidate pool for a declared relevance or utility task. It does not expand the underlying candidate set.

Context allocation

Selection, ordering, and possible truncation of representations under a finite budget and declared coverage, redundancy, origin, and conflict objectives.

Effective context

The ordered representations actually made available to generation. On a closed surface it is usually latent.

Retrieval-augmented generation (RAG)

Generation conditioned on externally retrieved representations. A controlled RAG stack does not imply that a commercial surface follows the same pipeline.

Citation resolver

The process that maps a displayed anchor or label to a canonical, versioned source object and relevant passage when available.

Evidence recall

The fraction of preregistered material evidence units recoverable from the top- k representations. It is stricter than document recall when one document contains several independent claims.

E.7.3 Measurement, statistics, and causal inference**Target population**

The sources, queries, systems, users, locales, times, or tasks to which the estimand refers. A convenient test panel is not automatically the target population.

Unit of analysis

The object represented by one analytic outcome. It may differ from the sampling, labeling, and assignment units.

Unit of assignment

The independently randomized or otherwise assigned cluster. Uncertainty follows this unit when treatment is assigned by query, page, domain, or time block.

Estimand

The precisely defined population quantity or treatment contrast the study seeks to learn.

Estimator

A rule mapping observed data to an estimate of an estimand. Good computation does not repair an ill-defined estimand.

Evaluation cell

A versioned combination of query, system, surface, time, locale, geography, account, and repeat to which raw outputs and labels attach.

Denominator ledger

The flow of planned cells through response, resolution, fetching, segmentation, labeling, exclusion, and analysis. It makes conditional populations and missingness visible.

Repeated measure

Another observation of the same underlying unit or cluster. Repeats estimate within-unit variation; they do not create new independently sampled queries.

Cluster

A set of observations sharing an intent, source, assignment, system state, or other dependence structure relevant to uncertainty.

Bootstrap

A resampling procedure whose sampling unit and target must match the data-generating or assignment structure. Row bootstrap and cluster bootstrap answer different questions.

Judge

A human or model-based measurement procedure that assigns a label. A judge has error, abstention, and possible subgroup shift.

Calibration

Assessment and, where justified, adjustment of a measurement or probability against a declared reference set. Aggregate calibration does not establish subgroup calibration.

Potential outcome

The outcome a unit would have under one specified treatment version and system state. Only one mutually exclusive state is observed for a unit at a time.

Treatment version

The complete, hashed source or system state assigned in an experiment, including all bundled changes and propagation conditions.

Interference

Dependence of one unit's outcome on treatments assigned to other units, for example through competition for retrieval or context.

Drift

Change in the response-generating, measurement, corpus, query, or product process over time. Drift is not interchangeable with within-cell stochastic variation.

Non-inferiority margin

A predeclared largest tolerable degradation on a guardrail outcome. It is a decision parameter, not a property

discovered after observing the result.

E.7.4 Evidence, risk, and governance

Claim–Evidence–Source graph (CES)

A typed, versioned graph linking a scoped proposition to relevant evidence units and the sources that contain or govern them. Support, qualification, contest, and supersession are distinct edges.

Source identity

The issuer, object type, title, version, date, status, method, artifacts, license, and stable locator of a source. Identity is not proof of truth or methodological adequacy.

Evidence ceiling

The strongest class of claim a source can normally support within its disclosed method and scope. It is an editorial boundary, not a prestige ranking.

Provenance

A record of origins, agents, entities, activities, derivations, and versions. Provenance can expose dependence; it does not prove that a claim is true.

Origin cluster

A group of source manifestations that materially derive from the same earliest verified origin. Several URLs in one cluster are not independent confirmations.

Evidence-preserving intervention

A revision whose changed propositions remain equivalent to approved claims or obtain newly approved evidence, while qualifications, provenance, and disclosure are preserved.

Threat model

A versioned account of protected assets, actors, capabilities, access, budget, adaptation, trust boundaries, and harms.

False positive

A benign or honest source, claim, or intervention that a detector or policy incorrectly flags under the declared reference label.

Release evidence package

A tamper-evident bundle of identities, before/after hashes, claim and rendered diffs, evidence, protocol, results, approvals, controls, monitoring, and tested rollback.

Institutional status

The verified class and lifecycle state of a law, regulation, mandatory or recommended standard, voluntary framework, technical specification, platform policy, or project. These classes are not synonyms.

Applicability

A qualified determination that a particular institutional object or control relates to the actual organization, role, system, data, jurisdiction, sector, and lifecycle stage.

Control effectiveness

Evidence that an implemented control meets its declared objective under a specified test. A policy statement or crosswalk is not effectiveness evidence.

Rollback

A tested, authorized action that restores or contains a released state while preserving evidence required for investigation.

Incident

A detected event that triggers preservation, triage, containment, correction, notification, and learning under a declared response protocol.

E.8 Acronym and metric index

Short form	Expansion	Use in these Notes
GEO	generative engine optimization	governed study and intervention of visibility, use, attribution, fidelity, and associated risks
IR	information retrieval	representation, candidate recovery, ranking, and evaluation before generation
RAG	retrieval-augmented generation	generation conditioned on retrieved external representations
BM25	Best Matching 25 family	lexical scoring family; implementation and parameters are part of the protocol
RRF	reciprocal-rank fusion	inspectable rank-based fusion of candidate lists
DCG / nDCG	discounted cumulative gain / normalized DCG	graded ranking utility under declared relevance labels and cutoff
MRR	mean reciprocal rank	mean inverse rank of the first relevant item
AP	average precision	mean precision at ranks containing relevant items, under a binary relevance convention
GLMM	generalized linear mixed model	link-scale regression with fixed and random effects for non-Gaussian repeated or clustered outcomes
ATE	average treatment effect	mean contrast of two specified potential outcomes over a declared target population
DiD	difference-in-differences	temporal comparison requiring, among other conditions, a defensible untreated trend
ICC	intraclass correlation coefficient	within-cluster dependence measure under a stated variance model
SUTVA	stable unit treatment value assumption	treatment-version and no-interference conditions; often questionable in competitive GEO
CES	Claim–Evidence–Source	project graph for scoped claims, evidence units, source identity, conflict, and provenance
RACI	responsible, accountable, consulted, informed	role-assignment aid; not evidence that the assigned work occurred
WCAG	Web Content Accessibility Guidelines	accessibility source whose applicability and conformance require separate evaluation
PROV-O	PROV Ontology	provenance vocabulary; not a truth or evidence-quality model

E.9 Notation release checklist

Before a chapter, appendix, figure, or empirical artifact is released, its editor checks that:

1. every new symbol has one definition, domain, unit, epistemic status, and first-use locator;
2. every estimate points to an estimand, estimator, independent unit, uncertainty method, and missing-state rule;
3. every latent platform state is marked latent and any proxy is named;
4. every vector preserves component units and non-compensable constraints;
5. every use of e , B , a , M , D , s , r , or C is locally disambiguated or migrated to its typed form;
6. a source ID, claim ID, evidence ID, citation object, and publisher are never joined by string coincidence alone;
7. terminology respects the forbidden substitutions in [section E.5](#); and
8. any symbol or terminology change records affected claims, equations, figures, data schemas, and cross-references in the version history.

Chapter summary

The governing rule is simple: if two quantities have different domains, denominators, epistemic status, or decision roles, they receive different names. If a platform state cannot be observed, the notation must preserve that fact. If a metric does not answer the research question, typographic precision cannot make it do so.

F Prerequisite Bridges

Information Retrieval, Statistics, and Causal Inference

Key idea

This appendix supplies the minimum technical bridge needed to read the main chapters without treating an external course as a hidden prerequisite. Each bridge explains why the idea is needed, defines the object, works one small synthetic example, identifies a common error, and ends with a checkpoint. The examples teach arithmetic and interpretation; they are not completed empirical studies or evidence about a production platform.

F.1 Reader contract and route through the appendix

The bridges are deliberately compact. They do not replace a full course in information retrieval, probability, statistics, experimental design, or causal inference. They provide enough structure to identify what an equation measures, what data it requires, and which conclusion it cannot support. The notation follows [chapter E](#).

Readers interested primarily in system mechanics should study BM25, dense and hybrid retrieval, and ranking metrics before [chapter 4](#). Readers preparing an evaluation should add sampling, bootstrap, and repeated measures before [chapter 6](#). Readers interpreting intervention claims should complete the potential-outcomes and generalized-linear-mixed-model bridges before [chapter 7](#). No section reports an external effect size; every number below is constructed for instruction.

Bridge	Main question	Primary manuscript use
BM25	How can exact term evidence produce an inspectable lexical rank?	Acquisition and retrieval baselines in Chapters 3–4
Dense and hybrid retrieval	How are semantic vectors and heterogeneous ranks combined without hiding scale?	Candidate construction and fusion in Chapter 4
Ranking metrics	What does a top- k ordering recover under declared labels?	Retrieval and benchmark evaluation in Chapters 4 and 6
Sampling and bootstrap	Which population is averaged, and what is resampled?	Query panels, denominators, and uncertainty in Chapter 6
Repeated measures	Which rows share a query, system, or time state?	Longitudinal measurement in Chapters 6–7
Potential outcomes and estimands	What causal contrast is the intervention supposed to identify?	Experiments and temporal designs in Chapter 7
GLMM intuition	How can non-Gaussian clustered outcomes be modeled without pretending rows are independent?	Sensitivity and heterogeneity analyses in Chapters 6–7

F.2 BM25: an inspectable lexical baseline

Why this is needed

Lexical retrieval asks whether query terms, variants, and local qualifications survive rendering, segmentation, and indexing. A strong semantic model can hide a broken representation; an inspectable lexical baseline makes exact term evidence and document-length effects visible. In chapter 4, BM25 is therefore a diagnostic baseline, not a claim that every generative surface uses BM25 [37].

Definitions

Let $\mathcal{D}$ be a corpus of N documents. For term t and document d , let $f(t, d)$ be term frequency, $|d|$ document length, $\overline{|d|}$ mean document length, and n_t the number of documents containing t . One common BM25 convention is

$$\text{IDF}(t) = \log \left(1 + \frac{N - n_t + 0.5}{n_t + 0.5} \right), \quad (80)$$

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \frac{f(t, d)(k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b|d|/\overline{|d|} \right)}. \quad (81)$$

Here $k_1 > 0$ controls term-frequency saturation and $b \in [0, 1]$ controls length normalization. Tokenization, field weighting, repeated query terms, stop-word handling, and the IDF convention vary across implementations. They belong in the experiment manifest.

The score is an ordering device for one query and one locked index. It is not a probability that a document is relevant, true, cited, or used by a generator.

Worked micro-example

Consider a synthetic corpus with $N = 3$, average length $\overline{|d|} = 100$, and one query term that occurs in two documents. Under the convention above,

$$\text{IDF}(t) = \log(1 + 1.5/2.5) \approx 0.470.$$

Take $k_1 = 1.2$ and $b = 0.75$. Document A has length 100 and contains the term three times; document B has length 50 and contains it once. Their single-term components are

$$\begin{aligned} \text{BM25}(t, A) &\approx 0.470 \times \frac{3(2.2)}{3 + 1.2(0.25 + 0.75)} \approx 0.738, \\ \text{BM25}(t, B) &\approx 0.470 \times \frac{1(2.2)}{1 + 1.2(0.25 + 0.75 \times 0.5)} \approx 0.591. \end{aligned}$$

The calculation explains why A ranks above B under this exact convention and parameter choice. It does not establish that A is better evidence, that the parameters are optimal, or that the ordering transfers to another query or corpus.

Common error

A BM25 score is sometimes read as an absolute measure and compared across queries or indexes. Corpus frequency, tokenization, length distribution, fields, and implementation conventions all change the scale. Another error is to repeat a term to increase $f(t, d)$ and call the resulting score change an evidence improvement. The equation contains no truth, provenance, readability, or downstream-use term.

Checkpoint

With N , document lengths, and all other parameters fixed, what happens to the IDF component in [equation \(81\)](#) as n_t increases? Why is that a corpus statement rather than a claim about the term's real-world importance?

Answer cue. The stated IDF decreases as the term appears in more corpus documents. It measures rarity under the locked index convention, not social, scientific, or commercial importance.

F.3 Dense and hybrid retrieval

Why this is needed

Exact wording is often insufficient for paraphrases, multilingual expressions, or conceptually related passages. Dense retrieval supplies a learned semantic representation, while hybrid retrieval combines complementary lexical and dense evidence. The combination must remain auditable: two score columns on different scales do not become commensurate merely because they are added.

Definitions

A dual encoder maps a query and candidate representation into vectors [22],

$$\mathbf{h}_q = f_q(q), \quad \mathbf{h}_d = f_d(d).$$

A common score is cosine similarity,

$$s_{\text{dense}}(q, d) = \frac{\mathbf{h}_q^\top \mathbf{h}_d}{\|\mathbf{h}_q\|_2 \|\mathbf{h}_d\|_2}. \quad (82)$$

The encoder, checkpoint, pooling rule, text truncation, and vector normalization are part of the retriever’s identity. Similarity is not entailment and does not establish source authority.

Hybrid retrieval may use a calibrated or normalized weighted score, a learned fusion function, or rank fusion. Reciprocal-rank fusion (RRF) combines ranked lists without pretending their raw scores share a scale:

$$s_{\text{RRF}}(d) = \sum_{j=1}^J \frac{\mathbb{1}\{d \in L_j\}}{\kappa + \text{rank}_{L_j}(d)}. \quad (83)$$

The constant κ and the input-list depths are protocol choices [10]. RRF rewards repeated high-rank recovery; it does not create independent evidentiary support from repeated or near-duplicate queries.

Worked micro-example

For a two-dimensional synthetic example, let the normalized query vector be $\mathbf{h}_q = (1, 0)$. Candidate vectors $\mathbf{h}_A = (0.8, 0.6)$ and $\mathbf{h}_B = (0.6, 0.8)$ are also normalized. Equation (82) gives similarities 0.8 and 0.6, so the dense component ranks A above B. The numbers describe geometry under this invented encoder output; they say nothing about factual support.

Now suppose a lexical list is (A, B) and a dense list is (C, A, B) . With $\kappa = 10$, RRF gives

$$s_{\text{RRF}}(A) = 1/11 + 1/12 \approx 0.174,$$

$$s_{\text{RRF}}(B) = 1/12 + 1/13 \approx 0.160,$$

$$s_{\text{RRF}}(C) = 1/11 \approx 0.091.$$

The fused order is A, B, C . A remained first because it was recovered near the top in both lists. That repetition is retrieval agreement, not evidence-source independence.

Common error

The most common hybrid error is $s_{\text{lex}} + s_{\text{dense}}$ with no normalization, calibration, or learned rule. Whichever component happens to have the larger numeric scale can dominate. A second error is to treat embedding proximity as a claim-support label; two passages can discuss the same topic while contradicting each other.

Checkpoint

Two rankers produce $L_1 = (A, B, C)$ and $L_2 = (C, A, B)$. Under equation (83) with $\kappa = 10$, which candidate ranks first? What extra record is required before the two appearances of A can be described as independent support?

Answer cue. A ranks first because $1/11 + 1/12$ exceeds the corresponding sums for B and C. Independence requires source-origin and derivation tracing; rank-list agreement alone is insufficient.

F.4 Ranking metrics: recovery is not answer quality

Why this is needed

Retrieval and reranking require stage-appropriate metrics. A top- k metric can show whether labeled evidence was recovered or ordered well, but it cannot show that a generator used the evidence faithfully. The cutoff, judgment unit, relevance scale, and denominator define the metric.

Definitions

Let $R_k(q)$ be the ordered top- k candidates for query q , and let $\text{Rel}(q)$ be the preregistered relevant set. Binary recall is

$$\text{Recall @}k(q) = \frac{|R_k(q) \cap \text{Rel}(q)|}{|\text{Rel}(q)|}. \quad (84)$$

Precision at k divides the same numerator by k . The reciprocal rank is the inverse rank of the first relevant candidate, or zero under a declared no-relevant-result convention. Mean reciprocal rank (MRR) averages that value over the target query panel.

For graded relevance $g_i \geq 0$ at rank i [20],

$$\text{DCG @}k = \sum_{i=1}^k \frac{2^{g_i} - 1}{\log_2(i + 1)}, \quad (85)$$

$$\text{nDCG @}k = \frac{\text{DCG @}k}{\text{IDCG @}k}, \quad (86)$$

where IDCG is DCG for the ideal ordering under the same labels. The zero-IDCG case needs an explicit policy. For binary labels, average precision (AP) averages precision at the ranks where relevant items occur, using the declared relevant-item denominator.

Document relevance, passage relevance, material-evidence recovery, source diversity, citation support, and answer correctness are different label spaces.

Worked micro-example

Suppose the top three synthetic candidates have graded labels (3,0,2) and the two nonzero candidates are the complete relevant set. Then

$$\text{Recall @}3 = 1, \quad \text{MRR} = 1,$$

and

$$\begin{aligned} \text{DCG @}3 &= 7 + 0 + 3/\log_2(4) = 8.5, \\ \text{IDCG @}3 &= 7 + 3/\log_2(3) \approx 8.893, \\ \text{nDCG @}3 &\approx 0.956. \end{aligned}$$

Under binary relevance, relevant items occur at ranks 1 and 3, so

$$\text{AP @}3 = \frac{1 + (2/3)}{2} \approx 0.833.$$

The metrics disagree because they reward different aspects of the ordering. None evaluates whether a later answer cites or preserves the two items.

Common error

Reporting “nDCG improved” without the relevance guide, cutoff, query panel, ideal-order convention, and uncertainty makes the number uninterpretable. A second error is to compare document-level nDCG with claim-level evidence recall as though they shared a denominator. A relevant document can still lose the material passage during segmentation or packing.

Checkpoint

For a three-item ranking with grades $(0, 2, 1)$, what is the reciprocal rank of the first relevant item? Would that value change if the second item supported the wrong claim in the final answer?

Answer cue. The reciprocal rank is $1/2$. It does not change because the ranking label and the answer-support label belong to different stages; the latter requires a separate citation or claim-support audit.

F.5 Sampling, weighting, and bootstrap**Why this is needed**

A prompt list defines a population only when the study says how that list was constructed and how it represents the target demand space. Equal numbers of factual and comparative queries can be useful diagnostically while badly misrepresenting usage prevalence. Bootstrap uncertainty is meaningful only when the resampled unit respects sampling and assignment.

Definitions

Let a target query population be partitioned into strata $h = 1, \dots, H$ with target weights W_h , where $W_h \geq 0$ and $\sum_h W_h = 1$. If $\hat{\mu}_h$ is the stratum estimate, a stratified target estimate is

$$\hat{\mu} = \sum_{h=1}^H W_h \hat{\mu}_h. \quad (87)$$

The weights may come from the sampling design, a declared decision population, or estimated usage; those sources define different estimands.

A nonparametric bootstrap repeatedly resamples units with replacement and recomputes the estimator. If query intent is the sampling unit and each intent has repeated responses, a cluster bootstrap resamples query-intent IDs and retains their associated rows. A row bootstrap instead treats response rows as exchangeable. These procedures generally represent different uncertainty models. Number of replicates, interval construction, stratification, and random seed belong in the manifest [14].

Worked micro-example

Suppose a synthetic target population is 80% factual tasks and 20% comparative tasks. A balanced diagnostic sample yields stratum means $\hat{\mu}_1 = 0.25$ and $\hat{\mu}_2 = 0.60$. The target-weighted estimate is

$$0.8(0.25) + 0.2(0.60) = 0.32,$$

whereas an equal-stratum diagnostic mean is

$$0.5(0.25) + 0.5(0.60) = 0.425.$$

Neither arithmetic result is intrinsically preferable. They answer different questions and must be labeled accordingly.

For a cluster bootstrap, imagine four query intents with all their repeated rows. One bootstrap draw might select intent IDs (2, 2, 4, 1). The estimator is recomputed from every row attached to those selected clusters, including the duplicate copy of cluster 2. This is a procedural illustration, not a claim about interval coverage for a particular dataset.

Common error

Analysts often resample hundreds of response rows even though treatment or sampling occurred at a few dozen query intents. The resulting calculation can represent within-intent variation while omitting uncertainty from which intents entered the study. Another error is to retrofit target weights after seeing which stratum favors the intervention.

Checkpoint

Thirty query intents are sampled, each is run five times, and the target is the mean over query intents. Which unit should the primary nonparametric bootstrap resample? What variation do the five repeats contribute?

Answer cue. Resample the thirty intent clusters and retain each selected intent's repeated rows. The repeats describe within-intent stochastic variation; they do not create 150 independently sampled intents.

F.6 Repeated measures and dependence

Why this is needed

Generated answers vary across stochastic runs, queries, systems, and time. Repeating the same prompt estimates one component of that variation but does not create a new demand sample or freeze the platform. A useful analysis separates within-cell variation from between-query, between-system, and temporal change.

Definitions

Let Y_{qmr} denote the outcome for query intent q , system m , time block t , and repeat r . A descriptive decomposition is

$$Y_{qmr} = \mu + \alpha_q + \beta_m + \gamma_t + \xi_{qm} + \zeta_{mt} + \varepsilon_{qmr}. \quad (88)$$

The terms ξ_{qm} and ζ_{mt} denote query–system and system–time interactions. Together, the terms distinguish query, system, time, interaction, and run-level variation; this equation need not be the final fitted model.

The intraclass correlation coefficient (ICC) describes similarity within a cluster under a stated variance model. For approximately equal cluster size $\bar{m}$, the design-effect diagnostic

$$1 + (\bar{m} - 1)\rho$$

shows how positive within-cluster correlation reduces the independence gained from additional rows. It is an approximation tied to its assumptions, not a universal correction factor.

Worked micro-example

Two synthetic query intents are each run three times. Their binary outcomes are

$$A : (1, 1, 0), \quad B : (0, 0, 0).$$

The row mean is $2/6 = 1/3$, and the equal-query mean is also $[(2/3) + 0]/2 = 1/3$. The point estimates coincide, but the information structure does not: there are two sampled query intents and three measurements within each. If treatment were assigned by query, there would be two assignment units, not six. With only two clusters, a stable treatment-effect analysis is not supported; more repetitions cannot repair the lack of independently assigned queries.

Common error

Treating every response as independent produces a sample-size claim that ignores shared query, system, or time state. The opposite error is to average all repeats immediately and discard run-level variability and missing states. The cell archive should preserve both levels; the estimator chooses the level appropriate to the target.

Checkpoint

Forty query intents are assigned to two source versions, and each intent is run five times. How many independent assignment units are available if assignment occurs by intent? Can a sixth repeat compensate for losing ten query intents?

Answer cue. There are forty assignment units. Another repeat can refine within-intent variation but does not replace independently assigned intents or recover the same target-population coverage.

F.7 Potential outcomes and estimands

Why this is needed

An intervention claim is incomplete until it identifies the treatment versions, units, comparison, target population, system state, and assumptions. Potential outcomes separate the causal object from the observed data and make the missing counterfactual explicit. They do not make an observational comparison causal by notation alone.

Definitions

For unit i , treatment assignment $D_i \in \{0, 1\}$, and versioned system state S_t , let $Y_i(1, S_t)$ and $Y_i(0, S_t)$ be the two potential outcomes. The observed outcome is [38]

$$Y_i = D_i Y_i(1, S_t) + (1 - D_i) Y_i(0, S_t). \quad (89)$$

A finite-sample average treatment effect is

$$\tau(S_t) = \frac{1}{N} \sum_{i=1}^N [Y_i(1, S_t) - Y_i(0, S_t)]. \quad (90)$$

The **estimand** is the target in equation (90) or another declared contrast. The **estimator** is the rule applied to observed outcomes, such as a randomized difference in means. Identification requires an assignment or design argument, treatment uptake, and assumptions about interference, missingness, and system-state comparability.

Treatment is a complete versioned source state, not a label such as “added a table.” If several elements change together, the estimand concerns the bundle unless the design separates them.

Worked micro-example

For instruction only, suppose the complete potential-outcome table for four synthetic units were visible:

i	$Y_i(0)$	$Y_i(1)$	Individual contrast
1	0	1	1
2	0	0	0
3	1	1	0
4	0	1	1

The finite-sample ATE is $(1 + 0 + 0 + 1)/4 = 0.5$. Now assign units 1 and 2 to treatment and units 3 and 4 to control. The observed outcomes are $(1, 0, 1, 0)$, so the observed treated and control means are both 0.5 and their difference is zero. This one assignment does not reveal every unit’s missing potential outcome; its difference happens not to equal the finite-sample ATE. Randomization-based reasoning concerns the estimator’s behavior over the assignment mechanism, not recovery of individual counterfactuals.

Common error

A before–after difference is often labeled $\hat{\tau}$ without defining what would have happened after the intervention time under the baseline source state. Platform updates, index propagation, query-demand change, competitor actions, and judge drift are alternative causes. A causal symbol does not remove them. Another error is to report an average effect while concealing material negative effects for a protected or high-risk query stratum.

Checkpoint

Why can the individual effect $Y_i(1, S_t) - Y_i(0, S_t)$ not normally be computed from one experimental run? What extra condition is needed before an effect estimated at state S_t is transported to state $S_{t'}$?

Answer cue. Mutually exclusive treatment states reveal at most one potential outcome for the same unit. Transport to $S_{t'}$ needs a defensible stability or transport argument and evidence that relevant system, population, and measurement changes do not alter the target contrast.

F.8 Generalized linear mixed models: disciplined intuition

Why this is needed

Many GEO outcomes are binary, counts, or ordered labels, and observations are nested within query, system, source, or time. A generalized linear mixed model (GLMM) can represent a non-Gaussian outcome while allowing cluster-specific variation. It is a sensitivity or modeling tool, not a substitute for adequate sampling, randomized assignment, or transparent cluster-level summaries.

Definitions

A generalized linear model relates an outcome mean to a linear predictor through a link function. A mixed model adds random effects for clusters. For a binary outcome, one illustrative random-intercept specification is [4]

$$\text{logit Pr}(Y_{qmr} = 1 \mid b_q, v_m) = \beta_0 + \beta_1 D_q + \beta_2 P_t + b_q + v_m, \quad (91)$$

with, for example, $b_q \sim \mathcal{N}(0, \sigma_q^2)$ and $v_m \sim \mathcal{N}(0, \sigma_m^2)$ as modeling assumptions, and where P_t is a declared post-period indicator. Fixed effects such as β_1 describe systematic link-scale contrasts under the model; random effects describe modeled heterogeneity across sampled clusters.

For the logit link, $\exp(\beta_1)$ is a conditional odds ratio, not a probability difference. A subject- or cluster-

conditional contrast need not equal the population-marginal contrast. The analysis must report which one is the estimand and how predictions are averaged over random effects and target weights.

Other GLMM families and links may be appropriate for counts or ordered outcomes, but the outcome support, dispersion, zero inflation, and link are modeling choices requiring diagnostics. A complicated model does not create information that the design did not collect.

Worked micro-example

Set all random effects and other covariates to zero for a synthetic teaching cell. Suppose baseline probability is 0.20, so baseline odds are $0.20/0.80 = 0.25$. Let $\beta_1 = \log 2$, a conditional odds ratio of two. The treatment odds are 0.5, which corresponds to probability

$$\frac{0.5}{1 + 0.5} = \frac{1}{3} \approx 0.333.$$

Thus an odds ratio of two does not mean a twenty-percentage-point increase and does not double probability from 0.20 to 0.40. The calculation is purely a link-function translation; it is not an estimated intervention effect.

Common error

Three errors recur. First, a log-odds coefficient is described as a probability change. Second, a random-intercept model is assumed to repair confounding, contaminated assignment, or too few independent clusters. Third, convergence, separation, variance estimates near a boundary, and alternative correlation structures are omitted. Model output is evidence only after design, fit, and sensitivity are reported.

Checkpoint

Under a logit model, a baseline probability of 0.20 and a conditional odds ratio of two imply what conditional probability when other terms are zero? Why can that number differ from a population-average treatment probability?

Answer cue. The conditional probability is $1/3$. Population averaging integrates or averages over the distribution of query, system, and other random effects; the nonlinear inverse-logit means that evaluating at zero random effects is not generally the same operation.

F.9 Putting the bridges together

A complete study moves through the objects in a fixed order:

1. freeze the source representations, corpus, query population, and system state;
2. build inspectable lexical and dense candidates, then declare any fusion and reranking rule;
3. evaluate recovery under stage-appropriate labels and cutoffs;
4. define the target query and system distribution before weighting;
5. preserve repeated responses and resample or model at the independent sampling or assignment unit;
6. define the intervention estimand, treatment versions, and assumptions; and
7. use a GLMM only when its link, dependence structure, diagnostics, and target interpretation add information beyond transparent summaries.

Question	Minimum method object	Claim that remains outside the method
Was material evidence recoverable?	frozen corpus, passage labels, Recall@ k or evidence recall	source use or answer fidelity
Was the candidate order useful?	graded labels, nDCG / AP / MRR, cutoff, query panel	context admission or user utility
Does a rate describe target demand?	sampling frame, strata, target weights, missing policy	a population not covered by the frame
How uncertain is the estimate?	independent unit, resampling or model, assumptions, sensitivity	validity of the estimand or causal identification
Did the intervention cause change?	treatment versions, assignment or design, potential-outcome estimand, uptake, interference and drift checks	transfer to an unstudied system state or population
Does a GLMM coefficient describe the target?	outcome family, link, fixed and random effects, diagnostics, marginalization rule	repair of a weak or confounded design

Boundary and failure conditions

These bridges license no universal numerical conclusion. A worked value is valid only for its displayed synthetic inputs. An empirical manuscript claim still requires a versioned dataset or source, a declared protocol, uncertainty, evidence and license records, and the chapter-specific release gates.

Chapter summary

BM25 and dense retrieval describe candidate evidence under different representations; hybrid methods require an explicit fusion rule. Ranking metrics evaluate declared labels at a stage, not final answers. Sampling and bootstrap define which population and variation an estimate represents. Repeated measures preserve dependence. Potential outcomes define the missing causal contrast, and GLMMs provide one model for clustered non-Gaussian outcomes without replacing design or evidence.