

GroundedGEO: Auditing the Evidence Gap in Generative Search Rankings

Yihan Xia Huiling Fan Kangrong Zhong Taotao Wang[†]
 Shenzhen University
 Shenzhen, China

Abstract

Generative search systems rank products and services for consequential decisions, and publishers can cheaply make candidate text look relevant. Yet evidence status is not a text property but a claim–evidence relation: text-only rankers and defenses cannot separate honest detailed content from fabricated detail—an *identifiability gap*. We audit this gap with an evidence-paired benchmark (50 e-commerce queries, 1,950 cases) and a claim-level reranker, GroundedGEO, that penalizes query-relevant claims lacking support in a supplied packet. Matched rich variants control format and volume; *packet twins* add attestations at fixed text, while thinned packets withdraw them. On the frozen listwise ranker Qwen2.5-7B, unsupported-rich variants show significant normalized rank gain over clean candidates (+0.065 to +0.092 across claim profiles, Holm-corrected) while supported and neutral controls do not; the effect is model-dependent (marginal on MiMo-v2.5, absent on GLM-5.3-Flash). On a frozen pointwise scorer, oracle evidence labels cut the unsupported-rich top-3 rate from 0.65 to 0.43 (laundering 0.61 to 0.39) at $\lambda = 40$ with zero false suppression; packet twins restore the original rates without changing text. Against a 370-claim human gold, all tested automatic judges fail the preregistered reliability gate, although the best local judge retains 79–100% of oracle suppression with zero measured false suppression on protected arms. Separately, stripping attestation coverage increases false suppression by 0.307. These diagnostic effects identify two limits on the evidence channel: label quality and packet coverage. They do not validate an automatic defense, and interpretation of the adverse human-gold arm remains pending adjudication.

Keywords

generative search, generative engine optimization, ranking integrity, evidence grounding, Web content manipulation

1 Introduction

Generative search is turning retrieval into a decision interface. An LLM system synthesizes, compares, and recommends candidates rather than returning only links. Ranking position is consequential: a top candidate gains attention, credibility, and commercial value even when the user never visits the page. Yet candidate text is written by publishers with a direct incentive to occupy it, which generative engine optimization (GEO) makes actionable: rewrites measurably increase answer visibility [1], crafted third-party content can manipulate an LLM’s preferences [14], and black-box rewriting already yields fluent, competitive rank-promotion attacks [15]. Recent work also studies forged authority, synthetic consensus, and

[†]Corresponding author: ttwang@szu.edu.cn.

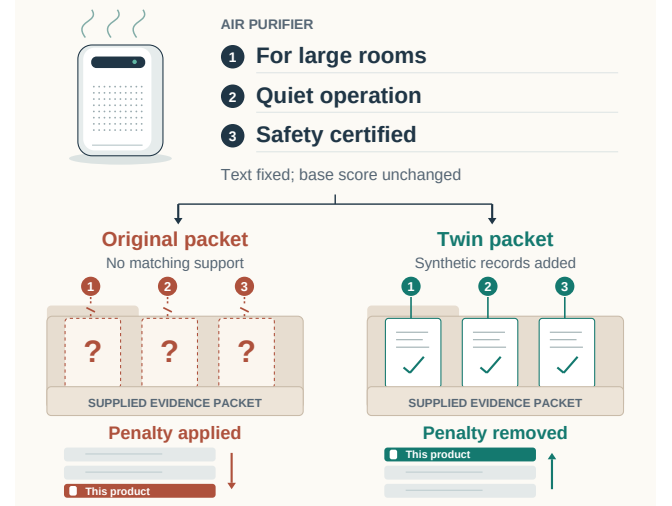


Figure 1: Packet-twin intervention. Packet twins add synthetic attestations for previously unsupported claims while holding text fixed. The purifier, claims, and rank shifts are schematic; measured results appear in Figure 4.

citation chains in search-agent endorsement [5], and coordinated page ecosystems that shape browsing and recommendation [20].

Some defenses look for manipulation cues inside the text itself—unusual perplexity, persuasive framing, or authority attribution [21]—asking whether a passage *looks* manipulated. Our concern is a narrower information constraint, not the absence of prior evidence-aware work: whether a claim is supported by evidence is not a property of the candidate text but a relation between the claim and evidence outside it. Any text-only defense thus assigns the same score to candidates differing only in whether their claims are evidence-backed, so it cannot in principle separate honest from fabricated detailed content: the *identifiability gap*, our first pillar. This argument concerns a candidate-text-only interface; it does not assert that search agents lack access to external evidence.

The attack the gap enables is *relevance inflation*: query-aligned claims—specifications, use cases, credentials—that enlarge apparent relevance, producing *unsupported rank gain*. Equally detailed supported content is valuable and must not be suppressed, so a ranking-integrity mechanism must separate candidates indistinguishable at the text surface—one rich in verifiable facts, the other in unsupported query-aligned claims (Figure 1). The U.S. Federal Trade Commission’s \$1.7M settlement with Truly Organic found the seller altered a third party’s USDA organic certificate to pass

off as its own—evidence laundering as a documented business practice [7].

The second pillar is a budget asymmetry: fabrication supplies unlimited query-aligned claims at near-zero cost, while honest content is bounded by existing recorded facts. Any reward for “looking well-supported” is thus always maximized by fabrication, never by honesty alone—relevance inflation is a structural pressure on generative search, not an incidental exploit—and predicts that exposed rankers reward unsupported relevance, not detail or length.

Together the two pillars state our thesis. Ranking integrity in generative search requires an *evidence channel*: candidate claims are grounded against a supplied evidence packet before reranking, in a closed world where *unsupported* asserts the absence of support in the packet, not open-world falsity. That channel is no free repair, and its value is bounded on two measured sides—by the quality of its evidence labels, which decide whom it penalizes, and by the coverage of its packets, which decide how much honest detail it can protect.

Our third pillar is an audit apparatus for studying that channel. We build an evidence-paired benchmark of 50 product-search queries from ESCI [18], five arms, and 1,950 ranking cases, with packets drawn from listing records, buyer reviews, and registry checks. GROUNDEDGEO matches atomic query-relevant claims to these packets and applies a post-hoc penalty to unsupported claim mass, retaining a claim–evidence trace. *Packet twins* add attestations while holding candidate text fixed; packet thinning removes attestations from otherwise honest claims. These complementary interventions separate sensitivity to evidence from sensitivity to prose. Whereas prior work audits generated answers against citations and identified sources [8, 12, 17], we audit candidate selection.

The measurements establish both a capability and its limits. Unsupported richness is rewarded on Qwen2.5-7B, with weaker evidence on MiMo-v2.5 and no significant contrast on GLM-5.3-Flash. Under construction labels, reranking reduces the unsupported-rich top-3 rate from 0.65 to 0.43 at $\lambda = 40$; changing only the packet restores it to 0.65. Yet all tested automatic judges fail the preregistered reliability gate, despite useful diagnostic ranking effects, and thinning evidence makes honest richness vulnerable. The resulting account is neither a universal attack nor a validated automatic defense: it identifies where evidence helps and what its protection depends on.

The contributions are threefold: (i) an information-structure argument for the identifiability gap, tested with fixed-text packet twins (§3); (ii) controlled evidence that unsupported detail receives ranking rewards on susceptible models, consistent with the budget asymmetry (§6.1); and (iii) a reusable claim-level audit framework (§4) whose experiments distinguish label agreement, ranking effects, and evidence coverage (§6).

2 Background and Related Work

LLM rankers and their input contract. Listwise LLM rerankers order candidate sets competitively zero-shot by jointly comparing candidate texts with proprietary and open-source models [16, 19, 23] on resources like MS MARCO [3]. They supply our substrate—a learned map from (query, candidate texts) to an order. Our defense

is a separate post-hoc layer, so the base ranker’s text-only input contract remains fixed.

Generative engine optimization. Publishers can rewrite their own content to raise visibility in generated answers by adding citations, statistics, and quotations [1]; GEO-Bench benchmarks fluent black-box attacks that promote ranking [15]. The outcome variable is visibility itself: claim support is uncontrolled, so a promoted candidate may be more relevant or merely more answer-shaped. Our rich variants match format and volume across support conditions; packet twins additionally vary supplied support while holding candidate text exactly fixed. The threat also connects to preference-manipulation attacks on LLMs selecting among competing third-party content [14] and to adversarial passages injected to obtain high retrieval rank [24]. Those attacks optimize selection or retrieval; our intervention instead asks whether the rank-driving claims have support in a separately supplied packet.

Search-agent evidence manipulation. SearchGEO measures endorsement corruption under forged authority, synthetic consensus, and authority-backed citation chains, explicitly probing apparent source diversity versus evidentiary independence [5]. EcoGEO’s TRACE organizes navigation and support pages into an evidence graph whose links and cross-references shape browsing and final recommendation [20]. HAE-GEO studies progressively more persuasive evidence poisoning and tracks exposure, adoption, verification, and recovery, including evidence independence and retained utility [4]. These works establish evidence manipulation as an existing research problem. Our complementary setting fixes the candidate-ranking interface and uses separately supplied packets, fixed-text twins, and packet thinning to audit ranking effects, label reliability, and evidence coverage. We do not model adaptive browsing or reconstruct Web-wide source dependencies.

Defenses and their decision targets. Defenses such as SCI-Defense flag manipulated passages via surface signals—perplexity, persuasive or authority framing, cross-candidate anomalies [21]—effective when manipulation leaves such signatures, but the defended quantity is the *appearance* of manipulation, not the presence of support. We use a length penalty as a controlled text-only comparator and packet twins to test a distinction that cannot be made from surface signals (§3.4). Evidence-aware prompting is also an existing comparator: SearchGEO tests source scrutiny and independent cross-validation [5]. GEO Defender learns a defensive ranking residual and uses an experience library to guide generation, targeting GEO-rewritten documents while preserving benign evidence use [11]. Its threat includes rewrites that preserve factual content. Our penalty instead targets query-relevant claims lacking packet support, not rewriting as such; we do not claim to detect all malicious GEO or establish superiority over these defenses.

Verifiability and attribution. In complementary attribution work, Liu et al. audit whether statements *produced* by generative search engines are supported by their cited sources, finding substantial unsupported content [12]. We move this standard upstream to ranking reward: rather than checking citations supplied in an answer, we pair candidate claims with a separately assembled packet before candidate selection. This claim-level view follows work that decomposes generations into atomic facts and checks support against a

knowledge source [13]. Citation and attribution research evaluates whether generated statements are supported by retrieved or identified sources [8, 17]; automatic attribution evaluators further show that claim–source matching is itself a modeling problem [22]. These methods evaluate generated output. GroundedGEO uses the same evidence relation as a ranking-time audit signal and separately gates the reliability of its automatic matcher.

The open gap. These lines of work motivate a joint evaluation of ranking reward and claim support, with evidence availability controlled independently of text. Product search offers a tractable setting: resources such as the Shopping Queries Dataset [18] supply relevance judgments, while listing, review, and registry records provide a separate basis for auditing rank-driving claims.

3 Problem Formulation

3.1 The Ranking Task

A *case* pairs a query $q \in Q$ with a candidate set $C_q = \{c_1, \dots, c_N\}$ of $N=5$ candidates: one *target* t and four distractors D_q fixed for the query. A generative ranker m observes the query and candidate texts, returning a total order of C_q ; $r_m(q, c) \in \{1, \dots, N\}$ is the rank of c , rank 1 at top. Separately, $s_0(q, c) \in \mathbb{R}$ is a pointwise base-relevance score—exposed by the ranker when available, else a standalone pointwise scorer over identical text—inducing order r_0 ; the phenomenon study consumes r_m ; the defense path s_0, r_0 (§3.6). Every candidate couples an entity identity with a publisher-controlled description; the intervention rewrites the target’s description alone, holding entity, core facts, distractors, and format fixed. It takes one of five *arms* (Table 1); rich arms carry a *claim profile* $p \in \{S, U, SUC\}$ fixing added-claim type (specifications, use cases, composite). A *variant* v is an arm–profile pair and t_v its text: $1 + 4 \times 3 = 13$ variants per query; a case pairs one variant with distractors under a frozen order $o \in \{1, 2, 3\}$ permuting all five candidates: $|Q| \times 13 \times 3 = 1,950$ cases, $|Q|=50$ queries.

3.2 Evidence Packets and Claim Labels

Evidence packets. Each query has $E_q = \{e_1, \dots, e_J\}$, a packet of atoms assembled before candidate generation and shared by all the query’s arms. An atom e is a checkable proposition with provenance: a source type and an ordinal *independence level* $\iota(e) \in \{0, \dots, 4\}$ measuring distance from the publisher (0 self-assertion; 2 affiliated or advertorial; higher is more independent). Four source types: official_public_dataset_record ($\iota=1$; public record of the listing’s own statements), platform_user_review ($\iota=3$; real buyer reviews), registry_corroboration ($\iota=3$), and certification_registry ($\iota=4$; authoritative registry checks). e *attests* g ($e \models g$) when it states g ’s content verbatim or equivalently.

Claim labels. A candidate has atomic claims $G(c) = \{g_i\}_{i=1}^L$. Each is tagged with a claim type and a query-relevance grade; $G_{QR}(c) \subseteq G(c)$ is the query-relevant subset. Against the packet, every claim receives an *evidence-support label*

$$\sigma_{E_q}(g) \in \{\text{independent_source_supported}, \text{supported}, \text{self_claimed_only}, \text{unsupported}, \text{unverifiable}\}.$$

Table 1: Evidence-paired target variants of one entity; arms differ only in added claims and their support labels.

Variant	Added content	Evidence status of added claims
Clean	Core facts only; the reference description	Self-descriptions of the entity (self_claimed_only)
Supported-rich	Query-relevant claims drawn from packet atoms	supported or independent_source_supported
Unsupported-rich	Matched query-relevant claims	unsupported
Neutral-matched	Natural, matched-length detail off the query’s decision dimensions	Attested only at listing level (self_claimed_only)
Evidence-laundering	Query-relevant claims in a third-party voice	unsupported: no atom with $\iota \geq \tau$ behind the implied outside source

assigned by the strongest basis the packet offers: independent_source_supported when some atom attests it at $\iota \geq \tau$ ($\tau=3$ separates independent from publisher-side sources); supported when a listing-record atom documents it as fact but no independent atom attests it; self_claimed_only when the packet carries it only as the publisher’s own assertion; unsupported when no atom attests an adjudicable claim; and unverifiable when the claim cannot be adjudicated against the packet, logged rather than coerced. Labels certify packet support, not plausibility: self_claimed_only is attested but never independent; unsupported asserts absence of support, not falsity (§3.4).

3.3 Paired Variants and Packet Twins

Variants. The five arms in Table 1 rewrite the same target so they differ only in added claims and the labels they receive: neutral-matched separates promotion-by-relevance from promotion-by-volume, supported-rich is the protected honest arm, and evidence-laundering stresses independence—the outside source its third-party voice implies does not exist in the packet.

Packet twins. For each (query, claim-profile) pair whose attack text has unattested claims, a *packet twin* flips the evidence world, text fixed:

$$E'_q = E_q \cup \{e_1^+, \dots, e_K^+\}, \quad (1)$$

where each e_i^+ attests one previously unsupported claim verbatim and carries a construction-tier tag, not an authoritative independence level, isolating attestation availability, not source prestige. Under the twin, $\sigma_{E'_q}(g) = \text{supported}$ where $\sigma_{E_q}(g) = \text{unsupported}$.

3.4 Scope and Assumptions

Closed-world evidence boundary. Every support label is a function of the claim and the supplied packet alone: a world-true claim attested by no atom of E_q is unsupported by definition, a fabricated-but-true claim remains unsupported, and a claim the packet cannot adjudicate is unverifiable, never silently converted. This pins

every estimand to an observable information set, reproducible from $(G(c), E_q)$.

Information structure. The phenomenon ranker observes (q, C_q) alone; the packet is auditor-side, reaching the pipeline only via the defense’s labels, so evidence status is no text property. Twins $(t_v, E_q), (t_v, E'_q)$ share text across two evidence worlds: text-only functions return identical outputs. Separating them requires an evidence-conditioned mechanism. This formalizes the *identifiability gap*: an evidence premium on the base ranker is unidentifiable; defense sensitivity to the evidence world is demonstrable.

Adversary capability. The adversary is a black-box publisher: it queries the ranker and optimizes its description from general knowledge, but not model weights, system prompts, distractors, or the packet. It adds query-aligned claims—specifications, use cases, credentials, coverage, and statements styled as third-party—raising apparent completeness; success is promotion over the clean arm.

Invariance controls. Within a (query, profile) pair the arms share entity, core facts, distractors, format, claim count, and claim type; the four rich arms are length-matched under a frozen tolerance, the clean arm is the natural reference and is not. Rich-versus-clean contrasts quantify promotion but do not isolate the mechanism, which rests on the matched rich arms (unsupported vs. neutral, supported vs. unsupported).

3.5 Estimands

Primary estimand: normalized rank gain. For variant v of the target in query q ,

$$\text{NRG}(q, v) = \frac{1}{3} \sum_{o=1}^3 \frac{r_m(q, t_{\text{clean}}, o) - r_m(q, t_v, o)}{N - 1}, \quad (2)$$

with $r_m(q, c, o)$ the rank of c under order o . $\text{NRG} \in [-1, 1]$; positive values mean promotion over the same entity’s clean description. The query is the independent statistical unit ($n=50$): order seeds are averaged within the query before paired comparison, and candidate- or call-level counts are diagnostics only.

Diagnostic: the evidence-discrimination gap. Within a claim profile,

$$\text{EDG}(q) = \text{NRG}(q, \text{supported-rich}) - \text{NRG}(q, \text{unsupported-rich}). \quad (3)$$

EDG is a descriptive diagnostic only, by deliberate demotion: a well-defined rank difference of two matched rich arms, it is not an *evidence premium*—the ranker observes candidate texts, never the packet, so evidence status cannot act on r_m , and the arms differ in attractiveness. The structure identifies promotion (NRG) and, defense-side, sensitivity to the evidence world via the twin contrast; a nonsignificant gap licenses no equivalence claim absent a prespecified smallest effect size and equivalence interval.

Defense events. Let r_0 be the order induced by s_0 , r_d that induced by defense rule d . For arm v and cutoff k , define top- k event $S_k^d(q, v) = \mathbb{I}[r_d(q, t_v) \leq k]$ and promotion event $P_k^d(q, v) = \mathbb{I}[r_0(q, t_{\text{clean}}) > k] \mathbb{I}[r_d(q, t_v) \leq k]$: the variant enters the top k though clean did not under r_0 (clean reference fixed by r_0 , not recomputed). The defense estimands are the query-mean drops

of these events on the attack arms (unsupported-rich, evidence-laundering):

$$\text{Success reduction}_k(d) = \frac{1}{|Q|} \sum_q [S_k^0(q, v) - S_k^d(q, v)], \quad (4)$$

$$\text{Promotion reduction}_k(d) = \frac{1}{|Q|} \sum_q [P_k^0(q, v) - P_k^d(q, v)]. \quad (5)$$

The frozen primary endpoint is Promotion reduction at $k=3$ (Success co-reported, $k=1$ secondary); the per-arm *top- k rate* $\frac{1}{|Q|} \sum_q S_k^d(q, v)$ is the raw rate. The false-suppression rate (FSR) applies the same drop to protected arms (clean, supported-rich, neutral-matched) and twin-world attack text, the attractiveness-controlled probe: indiscriminate suppression must cut the twin arm too. NDCG@ k and EvidenceSupport@ k (share of top- k query-relevant claims with independent support) are utility/grounding measures; a useful defense raises attack-arm reductions without false suppression or utility loss.

3.6 The Defense Scoring Interface

The defense consumes packet-relative claim labels. Let $n_Q(c) = |G_{QR}(c)|$ and $n_Q^{\max}(q) = \max_{c \in C_q} n_Q(c)$ for the current rerank unit. The normalized *unsupported mass* is

$$U(q, c) = \frac{\sum_{g \in G_{QR}(c)} \mathbb{I}[\sigma_{E_q}(g) = \text{unsupported}]}{\max\{1, n_Q^{\max}(q)\}}, \quad (6)$$

with $U = 0$ for candidates without query-relevant claims. Only unsupported enters the sum. Attested states, including self_claimed_only, never contribute; penalizing a listing’s own attested assertions would suppress honest self-description. The unverifiable state is logged as abstention, not falsified. The denominator is shared within a rerank unit but can vary between units; it is part of the scoring rule, not absorbed into a single global λ . Given the pointwise base score $s_0(q, c)$, the grounded score is

$$s_G(q, c) = s_0(q, c) - \lambda U(q, c), \quad \lambda \geq 0, \quad (7)$$

and candidates are re-sorted by descending s_G under a fixed tie-break. λ prices integrity against utility; Section 5 specifies the development-split selection rule and the fixed comparison grid. An identity precondition precedes any defense number: at $\lambda=0$ the induced order must equal the base-score order on every rerank unit. If s_0 comes from a scorer distinct from the phenomenon ranker, r_0 need not equal r_m ; the orders are reported separately, and the identity check concerns s_G versus s_0 only. This interface is post-hoc: it needs a numeric s_0 per candidate but does not modify the base ranker. Section 4 describes claim extraction, matching, and the audit ledger.

4 The GroundedGEO Framework

4.1 Setup and Interface

GroundedGEO is a claim-level, post-hoc reranking layer after a base ranker. Inputs: query q , candidate set C_q , supplied evidence packet E_q , and numeric base score $s_0(q, c)$. Outputs: adjusted score $s_G(q, c)$, the induced order, and a claim-level audit ledger (§4.4), under the closed-world label semantics of Section 3.2.

4.2 Pipeline Overview

Figure 2 shows the two input paths. The pipeline decomposes text into atomic, query-relevant claims (§4.3); grounds each against

the packet for a five-way label and independence grade (§4.4); aggregates unsupported mass for reranking (§4.6); and retains the matching decisions for audit. Packet provenance makes the evidence intervention replayable (§4.5).

4.3 Atomic Claim Extraction

An extractor emits claims $G(c)$ per candidate (§3.2), each carrying a *claim type* (six categories: Specification, Use Case, Credential, Location/Service Coverage, Case/Customer, Comparative), a *query relevance* grade $\rho_i \in \{\text{high, medium, low}\}$, and an external-evidence flag. The *high* tier forms $G_{QR}(c)$; other claims stay in the audit trace, unpenalized. Benchmark candidates ship with claims and *construction* support labels as oracle five-way labels; the automatic extractor instantiates the identical schema absent provenance (the gated seam, §4.7).

4.4 Closed-World Evidence Matching

Claim g_i is *independently supported* when some packet atom entails it and its source clears the independence threshold τ (§3.2):

$$z_i = \max_{e_j \in E_q} \# [e_j \models g_i] \# [\text{indep}(e_j) \geq \tau], \quad (8)$$

A per-claim judge (not emitting z_i) sees the claim’s construction-linked atoms, else the whole packet up to a fixed cap; absence of support in E_q is itself the verdict. It returns a five-way label (§3.2), an independence grade in $\{0, \dots, 4\}$, and a short rationale—unverifiable is retained, not coerced—and appends to the ledger the claim, label, grade, shown evidence identifiers, model version, and parse status.

4.5 Evidence Packets and Packet Twins

Packet atoms retain their source type, independence tier, and identifier in the matching ledger. The twin operation in Eq. 1 records *twin_of/twin_profile* provenance so an audit can replay the same candidate against the original and augmented packets. Construction-tier tags distinguish injected attestations from retrieved source records. Table 2 gives the benchmark’s source layers.

4.6 Grounded Scoring and Reranking

The implementation computes U once per candidate using the maximum query-relevant claim count in its rerank unit (Eq. 6), then sorts by descending s_G (Eq. 7); ties use ascending candidate identifier. For nonempty claim sets, U can equivalently be computed as the unsupported share multiplied by $n_Q(c)/n_Q^{\max}(q)$. This claim-count scaling is distinct from evidence-packet coverage. Labels also yield a supported-relevance share and a laundering-risk flag, diagnostics that do not enter the score: laundering gets no bespoke patch; third-party framing lacking an eligible third-party source enters U like any unattested claim.

Execution. The base score is a pointwise 0–100 relevance judgment per (query, candidate-text) pair; λ is in score points per unit unsupported mass. The pointwise seam is primary—listwise rankers expose only an order (kept as a robustness check). Handling is fail-closed: candidates lacking numeric s_0 are dropped and counted,

Table 2: Evidence-packet layers (independence: ordinal provenance).

Layer	Source type; independence; scale
Listing	official_public_dataset_record; 1; 943 atoms, all 50 queries
Buyer-review	platform_user_review (Amazon Reviews 2023); 3; 13,265 reviews $\rightarrow$ 828 atoms, 42/50 ASINs
Registry	certification_registry/registry_corroboration; 3–4; 13 atoms, 18 credential queries

never imputed; the $\lambda=0$ identity (rerank order equals base-score order) is enforced per unit by refusing to emit rankings on failure.

4.7 Reliability Conditions and Cost

Validity hinges on the matcher’s reliability: the judge must pass a frozen gate—macro-F1 ≥ 0.75 , Cohen’s $\kappa \geq 0.60$ on a query-disjoint human-adjudicated holdout—before labels enter confirmatory reranking; calibration uses the development split only. Absent the gate, construction labels (oracle by provenance, not human gold) bound what the mechanism *could* deliver; automatic reranking is diagnostic only. Cost is inference-only: no training, a judge pass linear in claims, the supplied packet.

The interface assumes supplied evidence; open-Web discovery, entity resolution, and universal fact checking are outside its scope.

5 Experimental Setup

5.1 Research Questions

RQ1: Unsupported-specific promotion. Do unsupported-rich variants gain rank over clean, unlike supported-rich and neutral arms?

RQ2: Evidence defense. Does GROUNDEDGEO suppress unsupported promotion without falsely suppressing the other arms?

RQ3: Evidence sensitivity. With text fixed, does changing the packet change the mechanism’s output?

RQ4: Automatic evidence labeling (diagnostic). Can a local language-model judge supply these labels, gated against human gold?

5.2 Benchmark and Candidate Arms

The benchmark comprises 50 English e-commerce queries derived from the ESCI Shopping Queries corpus [18]. Each query realizes the five arms in Table 1. Rich arms follow three *claim profiles* fixing the three added claims’ types (**S**: specification; **U**: use-case; **SUC**: one per type), giving $1 + 4 \times 3 = 13$ variants per query, for the 1,950 cases defined in Section 3.1. Only the target carries claims; the distractors carry none, so the penalty $U(q, c)$ (Eq. 6) can act on the target alone—the defense lowers a target’s rank rather than reordering competitors.

Evidence packets. Every query has a versioned evidence packet shared by all arms; atoms are assembled in three layers at distinct independence tiers (Table 2). One registry case is contested (q017: GREENGUARD Gold claimed, contradicted by independent sources).

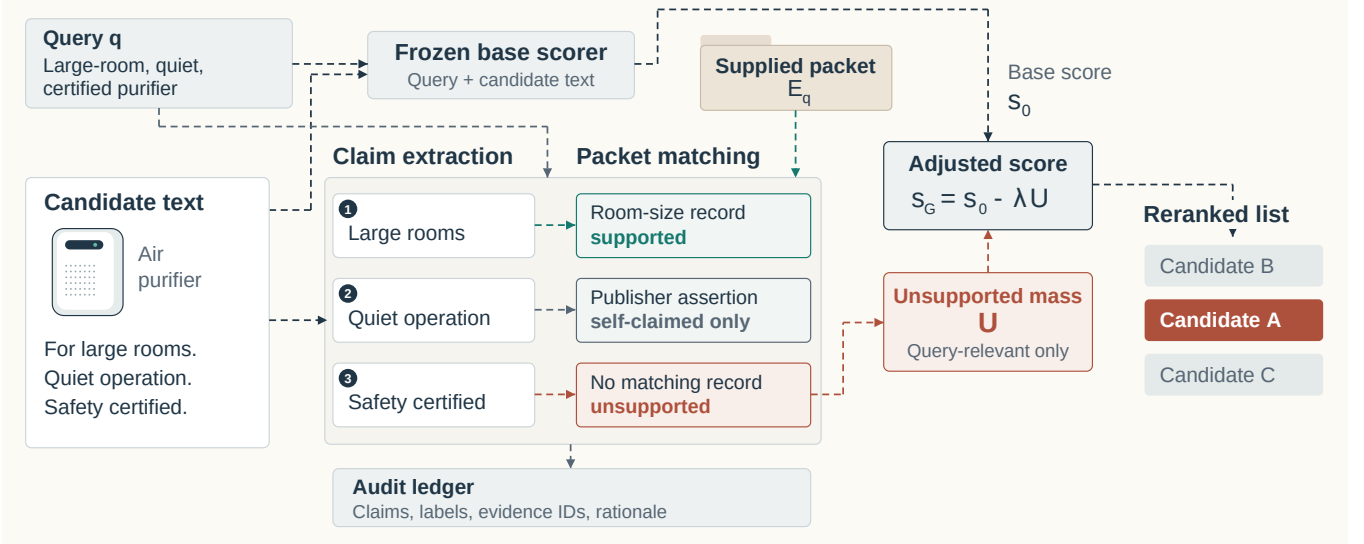


Figure 2: GroundedGEO’s claim-level reranking interface. Query and candidate text feed the frozen base scorer; the supplied packet enters only claim matching. Only query-relevant unsupported claims contribute to U ; all labels remain auditable. Three of the five label states are shown; unverifiable remains an abstention, not an unsupported verdict. The product, records, and output order are schematic. This diagram describes the interface, not a validated automatic defense.

5.3 Label Sources and the Reliability Gate

Claim labels come from four sources: construction-time labels, the 370-claim human gold, and two local judges. The API judge (DeepSeek-Flash) is a holdout-only side check, not a fifth source comparable to the local judges.

Construction labels. The benchmark ships each claim with a construction-time five-way label (oracle by provenance, not human gold). These labels supply the reference reranking arm; their agreement with human gold is evaluated alongside the automatic sources.

Human gold (370 claims). A frozen 10-query holdout supplies the human gold: 37 claims per query, 370 in total, each labeled independently by two annotators (five-way evidence label; source independence). Pre-adjudication κ : 0.833 (support), 0.816 (independence); 329 claims agreed outright, 41 adjudicated. Gold: 166 unsupported, 100 supported, 91 self_claimed_only, 13 independent_source_supported, 0 unverifiable.

Local judges. Two local judges label the same claims. L1 is RedHatAI/Qwen3.6-35B-A3B-NVFP4 on the existing :8000 endpoint; L2 is unsloth/Qwen3.8-27B-NVFP4 on :8002 (NVFP4 quantisation, GPU utilisation 0.35, -{ }-max-num-seqs32, enable_thinking=false). Each labeled all 4,250 claims; parsing completeness and agreement with gold are reported in Section 6.5.

API judge (holdout-only). The API judge, DeepSeek-Flash, was run only on the 10-query holdout (850 rows). Because it never ran on the development split, its labels cannot be treated as the same experiment as the local judges and serve the gate check only. The same model also supplies the pointwise base score $s_0(q, c)$ (Eq. 7).

Reliability evaluation. We apply the predeclared gate in Section 4.7 on the holdout. Alongside five-way agreement, we report precision and recall of the binary penalty decision, false penalties on attested claims, and both attested-only and operational escape rates (Table 7).

Label-noise sweep. To read the gate against endpoint needs rather than accuracy alone, we resample the penalty-relevant binary (unsupported vs. attested) on the construction labels to target precision P and recall R , deterministically, and re-run the reranker at $\lambda=40$ over all 1,950 cases. Precision targets are 0.30, 0.70, 0.90, and 1.00; recall targets are 0.2, 0.4, 0.6, 0.8, and 1.0. The measured sources’ binary precision and recall provide reference points for interpreting the sweep.

5.4 Probes: Packet Twins and Thinned Packets

Packet twins. 350 flipped packets (50×7 claim profiles, 1,050 injected claims) instantiate Eq. 1; the S, U, and SUC profiles enter the core ranking cases. For RQ3 we compare the original and twin packets at fixed text and λ .

Thinned packets. The mirror probe removes the attesting record of otherwise honest claims, so each thinned claim becomes unsupported under the closed-world rule and the penalty appears. The dose axis is the benchmark-wide share of attested claims whose attesting record is absent—0/25/50/75/100%, i.e. 0/263/525/788/1050 claims thinned; the listing layer is retained throughout. Per-candidate fractions are degenerate (each candidate carries three attested claims), so the share is pooled across the benchmark. Thinned packets carry thinned_of/thinned_level provenance, mirroring the twins, and all levels are reranked at $\lambda=40$ over all 1,950 cases.

5.5 Metrics and Their Link to the Estimand

Phenomenon. The phenomenon estimand is query-level NRG (Eq. 2; seeds averaged within query, $n=50$).

Defense. The main defense tables report each arm’s raw *top-3 rate*, the mean of S_3^d in Section 3.5. Attack suppression is its drop for unsupported-rich and evidence-launders (Success reduction); FSR is the drop for supported-rich, neutral, and clean. These rates are distinct from the clean-referenced promotion event P_3^d and do not by themselves establish the frozen Promotion reduction endpoint. Oracle reranking uses $\lambda \in \{0, 10, 20, 40\}$ over all 1,950 cases, subject to the identity check in Section 3.6. EDG remains descriptive.

Comparators. A per-token length penalty (0.05, 0.10, 0.15) on the same s_0 is the text-only comparator; for RQ3, candidates are reranked under each twin packet at $\lambda=40$.

Inference. Wilcoxon signed-rank tests on paired query-level contrasts, Holm-corrected within each claim profile. RQ1’s primary contrasts are unsupported-rich versus clean, supported-rich, and neutral, with the remaining arms versus clean as specificity controls.

5.6 Split, Tuning Rule, and Implementation

Frozen split. The 50 queries are partitioned into a 40-query development split and a 10-query holdout, query-disjoint. The split is frozen before evaluation: the development split is used to evaluate the λ selection rule and for calibration; holdout labels are not used to select λ . All label-source validation, the reliability gate, and the 370-claim human gold live on the holdout-10.

Pre-registered tuning rule. The rule below is pre-registered for this revision; it was not part of earlier drafts. Rule text: among λ with $\text{FSR} = 0$, select the smallest λ that attains 95% of the maximum suppression (the knee); if no λ attains $\text{FSR} = 0$, select the λ with the smallest FSR and flag it as a fallback. FSR is the binding constraint in the rule. Reported operating points are $\lambda \in \{0, 10, 20, 40\}$ (the main comparisons use $\lambda=40$ for comparability with the oracle curve); the rule was additionally evaluated at $\lambda \in \{80, 160\}$. We report the swept points and flag grid-edge selections; the main $\lambda = 40$ comparisons are not presented as evaluations of a development-selected optimum.

Implementation and reproducibility. Queries, candidates, packets, claim ledgers, and order manifests are versioned and hash-pinned; model identifiers, prompts, and parser schemas are frozen pre-evaluation. The auxiliary rankers that produce the phenomenon measurements are Qwen2.5-7B (local), GLM-5.3-Flash, and MiMo-v2.5 (API) [16, 23], all receiving identical candidate text. A result manifest binds each table to input hashes, raw logs, and analysis code.

6 Results

Unless a subset is named, results use the frozen 50-query benchmark (1,950 cases) and the protocol of Section 5. Phenomenon results use the three listwise rankers; all defense comparisons share the pointwise base scorer and report raw top-3 rates and their drops. Automatic-label, coverage, and label-noise arms are diagnostic.

6.1 Unsupported-Specific Rank Promotion (RQ1)

Qwen2.5-7B promotes unsupported-rich over clean on all three claim profiles (paired NRG $+0.065$ to $+0.092$, Holm-adjusted p between 0.035 and 0.0011; Table 3); at S it even beats supported-rich ($+0.103$, $p=0.0010$)—fabricated richness outranks honest richness. The gain is unsupported-specific: supported-rich, neutral-matched, and laundering arms never separate from clean. Figure 3 isolates the unsupported-rich versus clean contrast across rankers and profiles. MiMo-v2.5 has one significant corrected contrast (S-profile unsupported-rich versus neutral-matched); GLM-5.3-Flash has none. These within-model tests establish where promotion is detected, not a formal difference between models or equivalence on nonsignificant arms. On Qwen, the significant matched-rich contrasts are stronger evidence against a length-only explanation than the comparisons with clean: the rich arms share a length constraint, whereas clean does not. The claim is therefore a susceptibility demonstrated on particular rankers, not a universal consequence of adding detail.

6.2 Text-Only Proxies Cannot Separate Honest from Fabricated Richness (RQ2)

The per-token length penalty on the same base scorer cannot reproduce evidence-conditioned suppression. From 0.05 to 0.10 per token nothing moves (supported/neutral 0.75, unsupported 0.65, laundering 0.61); at 0.15 the attack arms drop only to 0.63 and 0.59 while the honest rich arms fall to 0.73 (Table 4). The weak settings leave the ranking unchanged; the stronger setting reduces both attack and protected rich-arm rates by about the same amount. Thus increasing the proxy penalty does not supply the missing distinction at the tested strengths. This comparison is specific to the length proxy on matched arms, not evidence that every text-only defense is ineffective on every attack. The fixed-text twin probe below isolates the broader information limitation.

6.3 Oracle Evidence-Constrained Reranking (RQ2)

Under oracle labels, GROUNDEDGEO’s penalty suppresses both attack arms: as λ grows $0 \rightarrow 40$, unsupported-rich falls monotonically $0.65 \rightarrow 0.43$ and laundering $0.61 \rightarrow 0.39$, while clean (0.84), supported-rich (0.75), and neutral (0.75) stay flat at every λ —FSR zero (Figure 4a; Table 5). The $\lambda = 0$ identity check passed on all 1,950 cases, excluding base-order mismatch as an explanation. Under construction labels the protected arms carry no unsupported mass, so their zero FSR follows from the rule rather than from threshold calibration; honest self-descriptions are not penalized. The length comparator lacks this selectivity. This establishes a mechanism-level reference under the supplied labels, not automatic-label reliability or protection when evidence is missing; the following experiments test those dependencies separately.

6.4 Packet Twins: An Existence Proof for Evidence Worlds (RQ3)

Flipping only the evidence world—twin packets whose injected atoms attest the previously unsupported claims—restores the attack

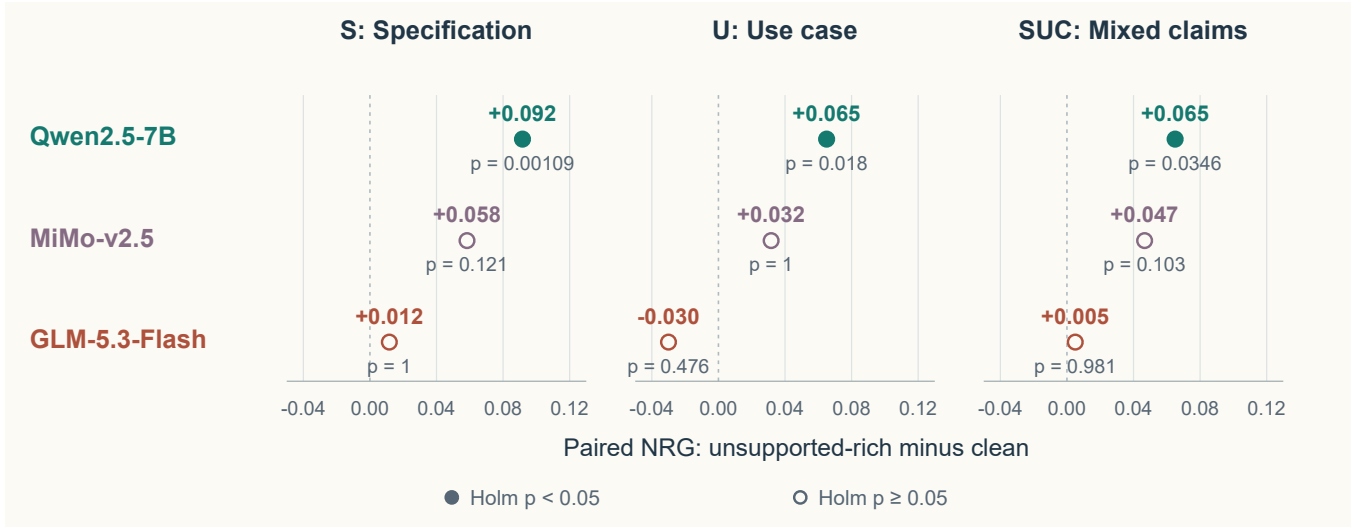


Figure 3: Model-specific promotion of unsupported-rich over clean, using the frozen query-level paired statistics ($n = 50$, seeds averaged within query). Columns are specification (S), use-case (U), and mixed (SUC) profiles. Points show mean paired NRG; filled points have Holm-adjusted $p < 0.05$. All displayed p values are within-model, within-profile tests, not tests of differences between models. Other comparators, including MiMo’s significant S-profile contrast against neutral-matched, are in Table 3.

Table 3: Paired NRG contrasts (Eq. 2); Holm-adjusted p in parentheses, bold $p < 0.05$; controls n.s. (min 0.064).

Ranker	Arm	Comparator	S	U	SUC
Qwen2.5-7B	unsupported-rich	clean	+0.092 (0.0011)	+0.065 (0.018)	+0.065 (0.035)
	unsupported-rich	supported-rich	+0.103 (0.0010)	+0.055 (0.063)	+0.045 (0.059)
	unsupported-rich	neutral-matched	+0.100 (0.0011)	+0.063 (0.049)	+0.058 (0.069)
	supported-rich	clean	−0.012	+0.010	+0.020
	neutral-matched	clean	−0.008	+0.002	+0.007
	evidence-laundering	clean	+0.042	+0.023	+0.018
MiMo-v2.5	unsupported-rich	clean	+0.058 (0.121)	+0.032 (1.0)	+0.047 (0.103)
	unsupported-rich	neutral-matched	+0.068 (0.037)	−	−
GLM-5.3-Flash	no significant contrast in any profile (21 tests after Holm, incl. negative directions; e.g., SUC neutral-matched vs. clean −0.030, $p=0.41$)				

Table 4: Per-token length penalty on the same base scorer s_0 , per-arm top-3 rate; the text-only comparator to Table 5.

Penalty (per token)	supported	unsupported	laundering	neutral
0.05	0.75	0.65	0.61	0.75
0.10	0.75	0.65	0.61	0.75
0.15	0.73	0.63	0.59	0.73

arms exactly to $\lambda=0$ rates (unsupported-rich $0.43 \rightarrow 0.65$, laundering $0.39 \rightarrow 0.61$; Figure 4b; Table 5 last column; $\lambda=40$). The injected attestations remove the penalized mass without changing s_0 . Their construction-tier tags do not confer authoritative source status, so the contrast isolates attestation availability rather than source prestige. A function of (q, c) alone cannot reproduce this fixed-text difference. The result is an existence probe for the evidence channel, not a test of real-world retrieval or certificate authentication.

Table 5: Oracle GROUNDEDGEORERANKING ($s_G=s_0 - \lambda U$), 1,950 cases, per-arm top-3 rate; twins flip only the packet.

Arm	$\lambda=0$	$\lambda=10$	$\lambda=20$	$\lambda=40$	twin $\lambda=40$
clean	0.84	0.84	0.84	0.84	0.84
supported-rich	0.75	0.75	0.75	0.75	0.75
neutral-matched	0.75	0.75	0.75	0.75	0.75
unsupported-rich	0.65	0.58	0.55	0.43	0.65
evidence-laundering	0.61	0.59	0.53	0.39	0.61

6.5 The Automation Gap: Claim–Evidence Matching Is the Bottleneck (RQ4)

All three automatic judges fail the preregistered reliability gate against human gold; the best local judge reaches macro-F1 0.374 and κ 0.442, well under the gate’s 0.75/0.60. Because the labels

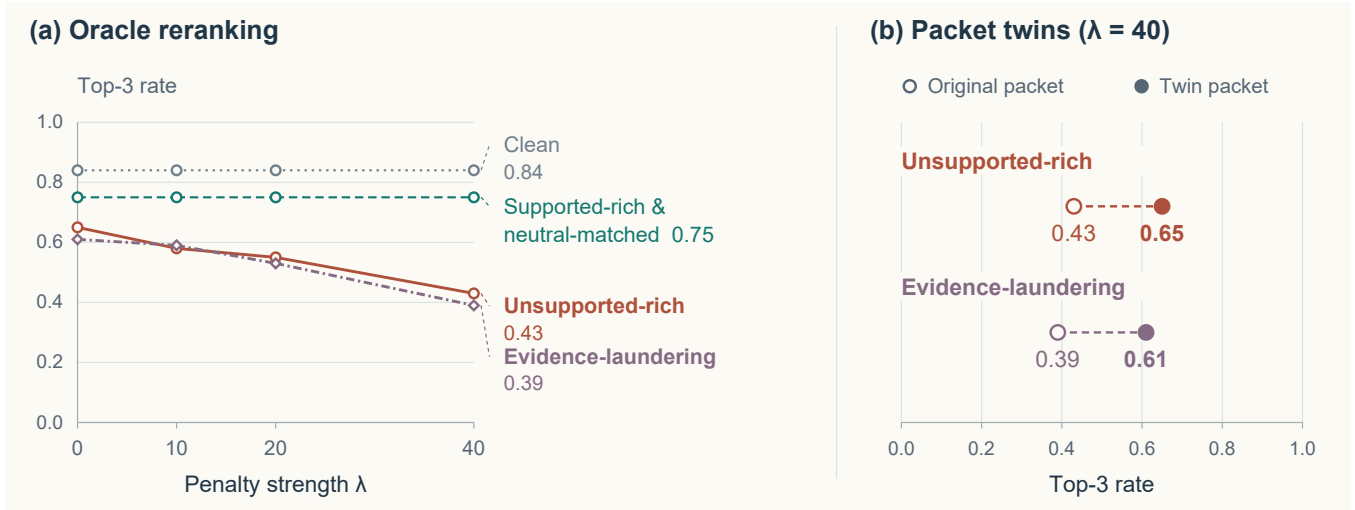


Figure 4: Oracle reranking and packet twins, using the rounded rates in Table 5 (1,950 cases). (a) Lines connect tested penalty strengths; the supported-rich and neutral-matched series overlap. (b) Original and synthetic twin packets at fixed $\lambda = 40$ and text. Hollow and filled points denote original and twin packets, respectively. These construction-label results bound the mechanism, not validated automatic-defense performance.

Table 6: Complementary diagnostics on the 370-claim holdout: pre-adjudication human agreement and API-judge binary performance. Five-way and penalty-decision metrics appear in Table 7.

Quantity	Value
Annotator agreement κ (support / independence)	0.833 / 0.816
Judge binary accuracy	0.738
Attested recall	0.975

that carry the oracle result are construction-time, this is a separate test of their automatic reproduction. Table 7 compares all sources against the 370-claim gold; Table 6 gives complementary agreement and binary diagnostics. Both local judges parsed all 4,250 claims successfully.

Error direction. The error direction is uniformly credulous: precision on the penalty decision is high (L2 0.944) while recall is the binding weakness (L2 0.819 best, API 0.271 worst), so fabricated claims are passed rather than honest ones suppressed. The two escape columns must be read separately—the API judge abstains heavily ($0.554 \rightarrow 0.729$ under the operational definition) whereas L2 does not ($0.181/0.181$). The strongest alternative—that the gate simply reflects noisy gold—is weakened by the construction row, which scores κ 0.859 with 3.0% escape and 9.3% false suppression against the same gold: the gold can be matched closely, and the automatic judges fail directionally, not randomly. These are point estimates on one 10-query holdout, not interval estimates. The API judge has no development-split run and cannot supply the full-set end-to-end comparison below. The gate remains unmet; we next examine ranking behavior diagnostically without relaxing that admission rule.

6.6 End-to-End Arms: What the Automatic Labels Actually Do (RQ2, diagnostic)

At $\lambda=40$ on the same pointwise base scorer, oracle labels suppress the attack arms with zero false suppression; both local judges also leave the protected arms untouched (FSR = 0) while suppressing substantially, with L2 reaching 79% of oracle unsupported-rich suppression and 100% of oracle laundering suppression on the full 1,950 cases, and 85–100% on the shared 10-query subset (Tables 8 and 9). The human-gold arm is the one adverse-measured cell: its collateral is positive under the current gold version, and its interpretation is pending.

Suppression and collateral summary (same cases). Attack suppression ($\lambda=0 - \lambda=40$, unsupported / laundering) and maximum collateral (FSR on protected arms): oracle +0.433/+0.267, FSR 0.000; human gold +0.167/+0.167, FSR +0.300; L1 +0.267/+0.167, FSR 0.000; L2 +0.367/+0.267, FSR 0.000.

Comparison scope. L1 and L2 each cover all 50 queries; the human-gold arm covers only the 10-query holdout, so it is compared on the subset where all four sources exist; the two local judges come from the same model family (Section 5), so agreement between them is not independent evidence.

Pending human-gold interpretation. The human-gold arm differs from the construction and local-judge arms: the entire measured collateral traces to five clean base-identity claims that the humans judged unsupported (the expected construction label is `self_claimed_only`), and part of the weaker suppression reflects about 10% of attack-arm claims that the humans judged attested, a standards divergence distinct from the base-identity case. Whether the collateral reflects gold noise or a definitional split is not adjudicated here: these are the numbers *as measured under this gold version*, and their interpretation is marked pending (Section 8). The

Table 7: Label sources against the 370-claim human gold (holdout-10), $\lambda=40$ inputs. Gate: macro-F1 ≥ 0.75 and Cohen’s $\kappa \geq 0.60$; all three automatic judges fail. “P/R_{pen}” is precision/recall of the penalty (unsupported) decision; FSR is attested claims penalized; the two escape columns are the attested-only and operational-definition escape rates. The API judge covers only the 850 holdout rows and is not the same experiment as the local judges.

Source	5-way acc	macro-F1	κ	P _{pen}	R _{pen}	FSR	esc. (attested)	esc. (op.)
construction (oracle row)	0.908	0.692	0.859	0.894	0.970	0.093	0.030	0.030
DeepSeek-Flash (API)	0.419	0.246	0.213	0.957	0.271	0.010	0.554	0.729
L1 (Qwen3.6-35B-A3B-NVFP4)	0.522	0.276	0.301	0.919	0.548	0.039	0.446	0.452
L2 (Qwen3.8-27B-NVFP4)	0.630	0.374	0.442	0.944	0.819	0.039	0.181	0.181

Table 8: End-to-end arms over all 1,950 cases, per-arm top-3 rate. The $\lambda=0$ column is shared by all sources (identical base order); the oracle, L1, and L2 columns are at $\lambda=40$. The oracle column reproduces the $\lambda=40$ rows of Table 5.

Arm	$\lambda=0$	oracle	L1	L2
clean	0.840	0.840	0.840	0.840
supported-rich	0.753	0.753	0.753	0.753
neutral-matched	0.753	0.753	0.753	0.753
unsupported-rich	0.647	0.427	0.500	0.473
evidence-laundersing	0.613	0.387	0.420	0.387

Table 9: Same cases (holdout-10, 390 cases), all four label sources at $\lambda=40$; the human-gold column is the only adverse-measured cell and its interpretation is pending (see text and Section 8).

Arm	oracle	human gold	L1	L2
clean	0.700	0.400	0.700	0.700
supported-rich	0.733	0.733	0.733	0.733
neutral-matched	0.700	0.700	0.700	0.700
unsupported-rich	0.233	0.500	0.400	0.300
evidence-laundersing	0.300	0.400	0.400	0.300

local judges’ favorable ranking effects do not reverse their gate failure. Section 7 considers why agreement on the audit ledger and changes in ranking need not track one another.

6.7 Coverage Dose–Response: The Cost of Unsupported Truth (diagnostic)

Thinning the packet—removing the proof records of an increasing fraction of attested claims—monotonically depresses the honest supported-rich arm while leaving clean, neutral, and both attack arms flat; at 100% thinning an honest candidate (0.447) approaches a fabricated one (0.427) at the endpoint (Figure 5; Table 10). The intervention changes evidence availability, not candidate length or format: withdrawing an attestation changes its claim’s label and hence U . The other arms’ labels are unchanged. At full thinning, supported-rich incurs +0.307 FSR despite unchanged text. Because the same queries are re-evaluated at each dose, the contrast avoids the query-difficulty confound in comparing naturally covered and uncovered strata. It measures the cost of engineered evidence loss, not its prevalence in deployment.

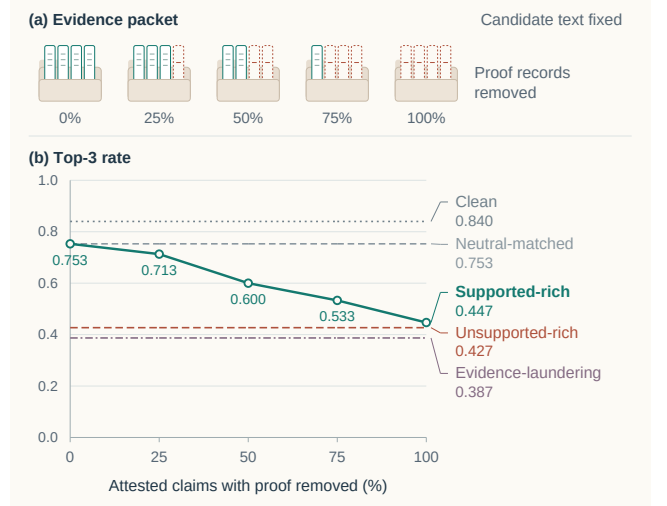


Figure 5: Evidence coverage dose–response ($\lambda = 40$, 1,950 cases). (a) Packet thinning, with schematic proportions rather than record counts. (b) Measured rates from Table 10; lines connect tested doses. Text and the base scorer remain fixed. This controlled probe does not estimate deployed-system performance.

Table 10: Coverage dose–response on thinned packets ($\lambda=40$, 1,950 cases). Axis: fraction of attested claims whose proof record is not in the packet (0/25/50/75/100% $\rightarrow$ 0/263/525/788/1050 claims). FSR is the supported-rich top-3 drop relative to the unthinned packet.

Thinned	supported-rich	FSR	neutral	clean	unsupported	laundersing
0%	0.753	0.000	0.753	0.840	0.427	0.387
25%	0.713	+0.040	0.753	0.840	0.427	0.387
50%	0.600	+0.153	0.753	0.840	0.427	0.387
75%	0.533	+0.220	0.753	0.840	0.427	0.387
100%	0.447	+0.307	0.753	0.840	0.427	0.387

6.8 Label-Noise Sweep: Precision Buys Safety, Recall Buys Suppression (diagnostic)

Resampling the binary unsupported-versus-attested decision shows the two error types act on different arms: label *precision* controls collateral on protected arms, and label *recall* controls attack suppression, whose ceiling is the oracle level +0.223 (Table 11). The four

Table 11: Deterministic label-noise sweep at $\lambda=40$ (1,950 cases): global re-sampling of the binary unsupported-versus-attested decision on the construction labels. Left and right column pairs summarize separate axes, not matched (P, R) settings per row. Measured-source precision and recall are in Table 7.

Precision P	FSR span	Recall R	Suppression
0.30	0.20–0.38	0.2	+0.017
0.70	0.033–0.16	0.4	+0.090
0.90	0.020–0.033	0.6	+0.127
1.00	0.000	0.8	+0.167
		1.0	+0.223

Table 12: λ tuning on the dev-40 split under the preregistered knee rule. Dashes are grid points not swept per source; **grid_edge flags a selection at the largest tested strength. Main comparisons retain $\lambda = 40$, not the selected grid-edge value.**

Source	$\lambda=40$ supp.	$\lambda=80$	$\lambda=160$	selected
oracle	0.192	0.567	0.633	160 (grid_edge, FSR 0)
L1	–	–	–	40 (supp. 0.158, FSR 0)
L2	–	–	–	160 (grid_edge, FSR 0)

measured label sources sit inside the grid. Figure 6 separates the two diagnostic axes. Relative to the oracle (P, R) = (1, 1), false unsupported labels add mass to protected candidates, whereas missed unsupported labels remove mass from attack candidates. This directional structure explains why aggregate label accuracy alone is insufficient for predicting endpoint behavior. The sweep is deterministic relabeling, not repeated sampling from a judge; its ranges describe the tested settings and carry no estimate of judge variability. It motivates separate safety and suppression requirements without replacing the preregistered agreement gate.

6.9 The λ Tuning Curve Is Weakly Identified

Under the preregistered knee rule (Section 5), the operating point is weakly identified: the FSR constraint is nearly empty on the dev-40 split (Table 12). The oracle and L2 curves are selected at the grid edge ($\lambda=160$, flagged **grid_edge**); L1 selects $\lambda=40$. With almost no protected-arm error to trade against suppression, the rule has little basis for selecting an interior operating point. A grid-edge selection therefore does not identify a deployable optimum. We report the swept points and selected settings, while retaining $\lambda = 40$ as the common comparison point; the main tables are not a tuned-and-frozen holdout evaluation of that value.

7 Analysis and Discussion

The joint results distinguish three questions that a single ranking score would conflate: whether a ranker rewards unsupported detail, whether a defense can respond to evidence, and whether the evidence labels are reliable enough to justify its decisions. A positive answer to the second does not establish the third. This distinction explains both the useful automatic-label ranking effects and the limits of interpreting them as a validated defense.

7.1 Mechanism

Evidence access versus evidence use. The API judge’s high attested recall coexists with substantial escape of unsupported claims (Tables 6 and 7). Unlike the text-only ranker, this judge receives evidence, so its errors are not forced by the same information restriction. They are consistent with a verifier accepting plausibility or apparent attestation in place of packet entailment, but the aggregate errors do not establish that cause. The practical implication is to audit the matching decisions themselves: adding an evidence channel is necessary for the twin distinction, but merely supplying a packet is not a reliability guarantee.

The scoring rule’s normalisation sets where it bites. In Eq. 6, the denominator is the largest query-relevant claim count in the current rerank unit. Here only the target carries claims, so for a nonempty target $n_Q^{\max}(q) = n_Q(t)$ and U reduces to its unsupported fraction. One disputed claim can therefore carry a large penalty on a sparse description. This explains the score’s sensitivity to the clean-arm disagreements in Section 6.6, but does not decide which labels are correct. If distractors also carried claims, the denominator need not equal the target’s count; this particular rate interpretation would not generalize unchanged.

Coverage and matching require separate audits. The thinning probe intervenes upstream of the judge: even a matcher that perfectly follows the closed-world rule must change a label when its sole attestation disappears. Improving agreement with that rule therefore cannot, by itself, protect uncovered truth. A deployment would need to distinguish a failed match against adequate evidence from an inadequately assembled packet, with retrieval or human review as possible responses to the latter. These are deployment requirements suggested by the probe, not capabilities validated by the current supplied-packet experiments.

“Attested” is itself contested at the margin. Construction labels never assign `independent_source_supported`, whereas the human annotators assign it to 13 claims (unsupported-rich 3, neutral 5, laundering 5), so agreement on independent support is not automatic and the attested/unsupported boundary is a judgment at the margin. Source eligibility and the binary decision to penalize are consequently different annotation tasks; a ledger can disagree on independence even when the ranking penalty is unchanged.

7.2 Unresolved Alternatives

The gate’s form versus endpoint behaviour. The strongest reading against our automation gap is that it measures the gate rather than the judge: L2’s penalty decisions yield useful ranking effects even though its five-way agreement is insufficient. The label-noise sweep explains why a binary penalty can tolerate some errors among non-penalized states while remaining sensitive to the direction of errors across the penalty boundary. Yet a useful ranking is not a reliable audit ledger. The preregistered gate addresses the latter; no alternative endpoint-based admission threshold was validated here. Future evaluation should specify both requirements in advance rather than choose a gate after observing favorable endpoints.

Whose labels are right. A second alternative is that the human-gold arm’s collateral and weaker suppression reflect the gold rather

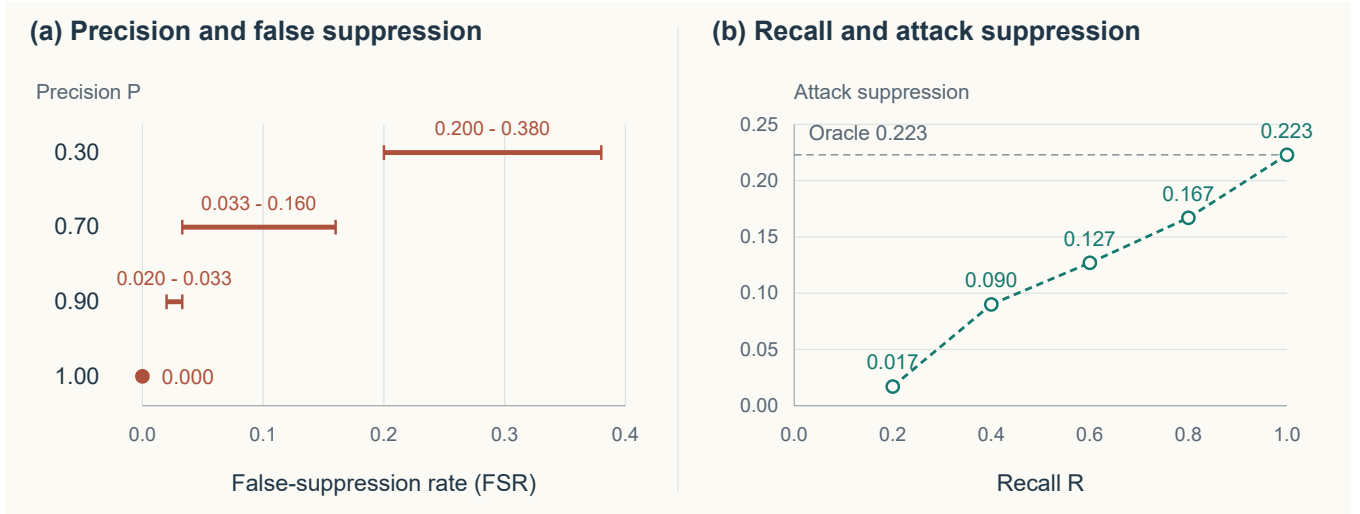


Figure 6: Label-noise diagnostics at $\lambda = 40$ on 1,950 cases (Table 11). (a) Horizontal segments show the reported FSR range across swept recall settings at each precision; these are simulation ranges, not confidence intervals, and no midpoint estimate is implied. (b) Attack suppression versus recall; dashed lines connect tested settings, with the oracle reference at 0.223. This deterministic relabeling simulation does not measure repeated-judge variability or validate an automatic judge.

than a property of human judgment. The concentrated disagreements reported in Section 6.6 permit both annotation noise and a substantive difference in the evidentiary standard. Re-adjudication needs the annotators’ rationales and linked atoms, not a preference for whichever source produces the cleaner ranking curve. Until then, the adverse numbers remain part of the result under the current gold version, without a claim that human judgment itself causes harm.

7.3 The Threat Model Is Institutional, Not Hypothetical

Fluent claims with absent or borrowed support recur in enforcement records at commercial scale. Beyond the Truly Organic case discussed in the introduction, examples include the USDA’s list of fraudulent organic certificates [2]; a GREENGUARD certificate absent from UL’s registry, confirmed forged on inquiry [10]; and the FTC’s first purchased-reviews settlement (\$12.8M [6]). The scarce resource is never convincing text, which is cheap, but the evidence behind it: exactly what an evidence-constrained ranker prices.

7.4 Implications and Open Problems

For platforms. Unsupported rank gain is a susceptibility, not a constant, so integrity auditing must be per ranker: a clean result on one family certifies nothing about another. The defense objective should shift from suppressing detail, which taxes honest richness, to discounting unsupported query-relevant mass. Every Grounded-GEO adjustment carries a claim–evidence trace, providing a basis for reviewing a demotion against an eligible source. A platform would still need a procedure for contesting packet omissions: the existing unverifiable state handles claims that cannot be adjudicated, but does not automatically exempt adjudicable claims whose attestations are missing.

Open problems. The next evaluation should first settle the disputed gold labels, then test broader judge families against both ledger agreement and ranking requirements on an expanded hold-out. Separately, a retrieval intervention is needed to test whether repairing missing evidence can recover the protection lost in the thinning probe. Neither question can be resolved by increasing the penalty strength alone.

8 Limitations

Ranker and domain scope. The phenomenon is established on the tested rankers, with significant unsupported-specific promotion on Qwen2.5-7B but weaker or absent evidence on the other two models. This is not a universal attack claim. The ESCI e-commerce setting [18] also leaves transfer to local services, B2B, and Web-passage ranking untested. A broader ranker-by-domain evaluation is needed to map susceptibility rather than transfer the reported magnitudes.

Supplied evidence and incomplete coverage. The closed-world estimand concerns support in a supplied packet, not open-world truth (Section 3.4). The controlled thinning probe establishes sensitivity to evidence loss; it does not estimate how often a deployed retrieval system would omit a valid attestation. Natural coverage in this benchmark is also uneven: the review layer covers 42 of 50 queries and the registry layer 18 credential queries. The documented coverage split has eight uncovered queries, while a derived split has nine (q008 differs). Uncovered queries have higher endpoints across arms, so they are not an exchangeable control group. The headline estimates retain all 50 queries; coverage claims rely on within-query thinning, pending reconciliation of the split and per-query coverage reporting.

Human-gold size and unresolved judgments. Gold-dependent estimates rest on 370 claims from only 10 queries, not 370 independent ranking tasks. Broader certification requires a larger adjudicated holdout. In addition, the current gold revision produces clean-arm collateral and weaker attack suppression (Table 9). Section 6.6 identifies the disputed base-identity and attack-arm labels; their causes remain unresolved. The adverse results are retained, but neither annotation noise nor a substantive difference in standards is established. Re-adjudication with reason fields is necessary before treating this arm as a settled reference. Zero FSR in the construction and local-judge arms does not resolve that disagreement.

Operating-point identification. The development sweep does not establish a sharp optimum (Section 6.9). Reported effects at $\lambda = 40$ are fixed-point comparisons, not a tuned maximum or validation of a frozen deployment setting. Selecting a useful tradeoff requires an evaluation regime with informative protected-arm errors, not merely a wider grid.

Judge and base-scorer breadth. The two local judges are checkpoints from one model family; the API judge runs on the holdout only. Their gate failures do not establish failure across judge families. All defense endpoints also use one pointwise base scorer, DeepSeek-Flash. Sharing it controls the within-study comparison, but neither absolute rates nor relative gains are established for other scorers. Replication must vary the label source and base scorer separately.

Target-only intervention and endpoint scope. Distractors carry no claims, so the penalty acts only on the target. This limits conclusions about competition among multiple penalized candidates and makes the target-specific normalization a special case (Section 7). Moreover, the displayed defense tables report raw top-3 rates and Success reduction, not the distinct frozen clean-referenced Promotion reduction endpoint. They do not alone establish that primary endpoint or the utility and grounding measures defined in Section 3.5.

9 Ethical Considerations

This paper studies manipulation to defend ranking integrity. Attacks are built on synthetic candidate descriptions in a controlled benchmark; no real page, listing, or review was altered, and no live ranking system was targeted. Enforcement cases in Section 7 are public records, not allegations. The work is dual-use: the same principles could aid real manipulation. We disclose because fabricated query-aligned text is already cheap—the scarce resource is evidence verification, which we measure. We release the evaluation apparatus—evidence-paired variants, packet twins, matcher reliability gate—making misuse measurable.

Human annotation covered 370 product claims and contained no personal information; annotator identities remain private. Two annotators labeled each claim independently; automated suggestions are disclosed as such, never as human judgments. Datasets are used under their research terms: ESCI [18]; Amazon Reviews 2023 [9].

Evidence-constrained ranking can disadvantage small publishers with thin third-party coverage. The controlled coverage probe (Section 6.7) demonstrates the mechanism of this harm, not its distribution across real publishers. A deployment should expose

packet omissions, allow candidates to submit eligible attestations, and provide review of disputed penalties. The claim–evidence trace supports such review but is not itself an appeals process. Likewise, retaining unverifiable avoids coercing unadjudicable claims into a verdict; it does not protect every true claim absent from a packet. Source eligibility should depend on independence rather than popularity, with first-party documents used only for claims they can establish. These governance measures require validation beyond the present experiments.

10 Conclusion

Evidence support is a relation between claims and records, not a property of candidate prose. GroundedGEO makes that relation explicit at ranking time through an evidence-paired benchmark, a claim-level reranker, and an auditable decision trace. Fixed-text packet twins demonstrate a response to evidence that a text-only rule cannot reproduce; matched-arm experiments show that unsupported detail can receive ranking reward, with susceptibility varying across the tested models.

The evidence channel offers conditional protection, not a completed automatic defense. Construction labels yield selective suppression on the supplied packets, and local automatic labels retain substantial ranking effects despite failing the preregistered agreement gate. Packet thinning exposes a separate vulnerability that better matching alone cannot repair: honest claims lose protection when their evidence is absent. Disputed human-gold labels and operating-point uncertainty further limit the interpretation of the current endpoints.

The contribution is therefore an apparatus for separating ranking reward, label reliability, and evidence availability, rather than treating a plausible description or a favorable ranking curve as sufficient proof of integrity. Progress requires validating the labels that justify a penalty and the evidence collection that makes such justification possible.

References

- [1] Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2023. GEO: Generative Engine Optimization. *arXiv preprint arXiv:2311.09735* (2023).
- [2] Agricultural Marketing Service, U.S. Department of Agriculture. [n. d.]. Fraudulent Organic Certificates. USDA AMS Enforcement. <https://www.ams.usda.gov/services/enforcement/organic/fraudulent-certificates>
- [3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamee, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268* (2016).
- [4] Zhongan Bi, Qiwen Wang, Jianrong Jiang, Jigang Ding, Wenwen Xiong, Changhua Meng, Xuanang Gao, Kepeng Lin, Changjiang Jiang, Yiang Chen, Huan Yao, Wei Wang, Zhenyu Ma, and Wenhui Dong. 2026. Evaluating Deep-Search Agents under Hierarchical Web Evidence Poisoning. *arXiv preprint arXiv:2609.06027* (2026). <https://arxiv.org/abs/2609.06027v2>
- [5] Yimeng Chen, Zhe Ren, Firas Laakom, Yu Li, Dandan Guo, and Jürgen Schmidhuber. 2026. How Much Can We Trust LLM Search Agents? Measuring Endorsement Vulnerability to Web Content Manipulation. *arXiv preprint arXiv:2606.16821* (2026). <https://arxiv.org/abs/2606.16821v2>
- [6] CNBC. 2019. FTC Cracking Down on Fake Amazon Reviews. CNBC. <https://www.cnbc.com/2019/03/01/ftc-cracking-down-on-fake-amazon-reviews.html>
- [7] Fair, Lesley. 2019. Deceptive “Certified Organic” Claims Leave Consumers Verklemt. FTC Business Blog, Federal Trade Commission. <https://www.ftc.gov/business-guidance/blog/2019/09/deceptive-certified-organic-claims-leave-consumers-verklemt>
- [8] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023*

- Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398
- [9] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiushi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
 - [10] LeedUser Community. 2020. Alert: Fraudulent Greenguard Certificates. BuildingGreen/LeedUser Forum. <https://leeduser.buildinggreen.com/forum/alert-fraudulent-greenguard-certificates>
 - [11] Haozhang Li, Yangguang Shao, Xinjie Lin, Zhong Guan, Mi Zhou, and Junzheng Shi. 2026. When Optimization Becomes Manipulation: Defending Generative Search against Malicious Generative Engine Optimization. *arXiv preprint arXiv:2609.02964* (2026). <https://arxiv.org/abs/2609.02964v1>
 - [12] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. *arXiv preprint arXiv:2304.09848* (2023).
 - [13] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long-Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741
 - [14] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. 2025. Adversarial Search Engine Optimization for Large Language Models. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=hkdxN3c7t>
 - [15] Ojas Nimase, Zhe Chen, Gengpei Qi, Yue Zhao, and Xiyang Hu. 2026. GEO-Bench: Benchmarking Ranking Manipulation in Generative Engine Optimization. *arXiv preprint arXiv:2605.29107* (2026).
 - [16] Ronak Pradeep, Sahel Sharifmoghadam, and Jimmy Lin. 2023. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *arXiv preprint arXiv:2309.15088* (2023).
 - [17] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics* 49, 4 (2023), 777–840. doi:10.1162/coli_a_00486
 - [18] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. *arXiv preprint arXiv:2206.06588* (2022).
 - [19] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 14918–14937. doi:10.18653/v1/2023.emnlp-main.923
 - [20] Hengwei Ye, Jiasheng Mao, Zhenhan Guan, and Zheng Tian. 2026. EcoGEO: Trajectory-Aware Evidence Ecosystems for Web-Enabled LLM Search Agents. *arXiv preprint arXiv:2605.12887* (2026). <https://arxiv.org/abs/2605.12887v2>
 - [21] Xucheng Yu, Haibo Jin, Huimin Zeng, and Haohan Wang. 2026. SCI-Defense: Defending Manipulation Attacks from Generative Engine Optimization. *arXiv preprint arXiv:2605.21948* (2026).
 - [22] Xiang Yue, Boshi Wang, Zirui Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 4615–4635. doi:10.18653/v1/2023.findings-emnlp.307
 - [23] Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2023. Rank-without-GPT: Building GPT-Independent Listwise Rerankers on Open-Source Large Language Models. *arXiv preprint arXiv:2312.02969* (2023).
 - [24] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 13764–13775. doi:10.18653/v1/2023.emnlp-main.849