MECHANISM DESIGN FOR GENERATIVE ENGINES: FROM EXPLOITATION TO WIN-WIN EQUILIBRIUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative engines are reshaping the web ecosystem by making citations a mechanism for allocating attention, attribution, and downstream value. This creates a strategic tension: content providers optimize for model citation, while platforms must preserve answer quality and trustworthy attribution. We show that this tension can escalate into citation wars. In repeated simulations, state-of-the-art generative engine optimization (GEO) adapts to conventional defenses by producing citation-seeking rewrites that degrade document quality and introduce unsupported claims. We formulate the supplier–platform interaction as a repeated Stackelberg game with partial monitoring, and local best-response analysis identifies when citation competition approaches an inert stationary outcome. We then propose VCR, a platform–creator mechanism based on verifiable-content rewards. Beyond penalizing suspicious rewrites, the platform credits rewrites that surface checkable factual substance, aligning creator incentives with answer trustworthiness. Experiments on three benchmarks show that VCR consistently achieves the largest Net defense–utility score, outperforming the strongest baseline by an average of 12.1 percentage points, and produces a win–win outcome under our operational exposure-preservation criterion. Code: `https://anonymous.4open.science/r/VCR-8DD3/`.

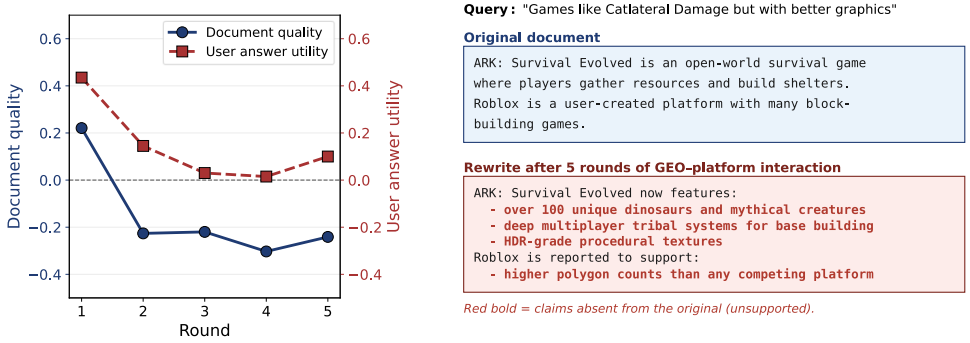
1 INTRODUCTION

Generative search is reshaping information access into a citation-mediated generative engine ecosystem, where LLM platforms allocate attention and attribution by deciding which sources to summarize and cite Liu et al. (2023). In this ecosystem, citations become the new unit of visibility: they determine not only what users trust, but also which content suppliers receive traffic, reputation, and downstream value. As a result, content providers are increasingly incentivized to optimize for model citation (Aggarwal et al., 2024; Kumar & Lakkaraju, 2024), while platforms must preserve answer quality and trustworthy attribution Xiang et al. (2024); Edemacu et al. (2025). This creates a strategic tension between supplier-side visibility seeking and platform-side quality control. How to model and improve welfare in this generative engine ecosystem before it escalates into citation wars has become an urgent problem.

Generative Engine Optimization (GEO) (Aggarwal et al., 2024; Wu et al., 2025; Yuan et al., 2026) operationalizes this tension: content providers rewrite documents to increase their probability of being cited by generative search engines. Existing studies mostly treat GEO as a one-shot supplier-side optimization problem. This misses a feature of the generative engine ecosystem: supplier optimization and platform defense interact repeatedly. Once citations mediate visibility and downstream value, GEO may shift from improving content quality to strategically adapting to platform defense rules. Over time, such adaptation can push suppliers toward citation-seeking behaviors that distort attribution, reduce answer trustworthiness, and harm ecosystem welfare. We next use simulation experiments to demonstrate this risk.

Fig. 1 illustrates this risk in a five-round supplier–platform interaction on E-COMMERCE. The supplier runs AUTOGEO (Wu et al., 2025), while the platform applies a standard prompt-warning defense. Panel (a) shows the system drifting from the no-attack reference: rewrite quality becomes negative after the first round, while answer utility declines from its initial advantage and remains above the zero reference. Panel (b) explains why: repeated GEO adaptation accumulates unsupported claims and fabricated details, turning mild edits into citation-seeking rewrites. This adaptation can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Five-round GEO exploitation

(b) Examples of GEO rewrites

Figure 1: Repeated GEO exploitation degrades creator-side document quality and user-side answer utility under a standard defense. Panel (a) shows document quality falling below the zero reference after the first round, while the initial answer-utility advantage diminishes. Panel (b) shows a representative document before and after five rounds, with unsupported claims in red.

drive the ecosystem toward an undesirable stationary outcome where neither robust defense nor honest content improvement is sustained.

To formalize this problem, we model the platform-supplier interaction as a repeated Stackelberg game (Von Stackelberg, 2010) with partial monitoring. In each round, the supplier observes the platform's current defense rule and commits to a GEO strategy for rewriting its target documents. The platform then observes only document-level before/after pairs and updates its answer-time defense policy accordingly. This game captures the key asymmetry of the generative engine ecosystem: suppliers can adapt directly to platform rules, while platforms must infer whether a rewrite reflects genuine quality improvement or strategic manipulation.

Under this setting, we derive two structural results in a local quadratic model (Sec. 3.3). First, holding source-model moments fixed, the platform's best-response defense weakens as latent content quality becomes more correlated with manipulation intensity. Second, when supplier utility uses citation level rather than marginal citation gain, a stationary response can prefer high-quality targets. Together, these results expose an inert stationary outcome where content-improvement and manipulation gradients are exhausted and further interaction produces little joint utility.

Motivated by this analysis, we propose VCR, a platform-supplier mechanism based on verifiable-content rewards that aligns supplier incentives with platform trustworthiness. The platform makes defense incentive-compatible by crediting rewrites that make source-supported factual content more salient and offsetting this credit against the suspicion score. Cosmetic or self-promotional GEO edits remain demoted, while substantive rewrites that surface checkable information are rewarded. In the local model, this gives a rational supplier a welfare-aligned optimization direction and improves the joint defense–utility outcome.

We evaluate VCR on three benchmarks Wu et al. (2025); Aggarwal et al. (2024): commercial search (E-COMMERCE), open-domain factual queries (GEO-BENCH), and research-oriented queries (RESEARCHY-GEO). Across datasets and answer engines, VCR produces a win–win outcome under operational exposure-preservation criterion by protecting platform/user utility while preserving creator exposure. It achieves the largest Net score in every setting, averaging 12.1 percentage points above the strongest baseline, and has the strongest direct document- and answer-quality point estimates.

Our key contributions are three-fold:

- We formulate the GEO–platform interaction as a repeated Stackelberg game with partial monitoring, and provide a local best-response analysis of stationary citation competition.
- We propose a supplier-platform mechanism based on verifiable-content rewards, which transforms platform defense from a purely punitive filter into a two-sided incentive mechanism that rewards substantive content while penalizing manipulation.
- We validate the proposed mechanism on three benchmarks, three answer engines, and five GEO attackers, where it consistently achieves the largest Net score.

2 Related Work

Supplier-side visibility attacks. Content providers have long manipulated search visibility through link and content spam and through adversarial document changes that affect ranking (Goren et al., 2018; Gyöngyi & Garcia-Molina, 2005). LLM pipelines add attack surfaces through retrieved instructions, poisoned corpora, and visibility-oriented webpage rewriting (Greshake et al., 2023; Kumar & Lakkaraju, 2024; Zhong et al., 2023). GEO makes citation in a generated answer the direct optimization target. Foundational work defines visibility metrics and studies manual or automatic rewriting (Aggarwal et al., 2024; Wu et al., 2025), while recent work broadens the optimization strategies and evaluation settings (Chen et al., 2025c; Kim et al., 2026; Yuan et al., 2026; Zhou et al., 2026). These studies establish the supplier-side optimization problem. We study how it changes when suppliers repeatedly adapt to a platform that also updates its defense. Additional GEO variants and adjacent attacks are discussed in App. B.

Platform-side defenses. Platforms can intervene at ranking, instruction handling, retrieval, and generation. Traditional search uses spam detection and robust ranking (Castillo & Davison, 2011); LLM defenses prioritize trusted instructions (Wallace et al., 2024) and make RAG more robust to adversarial passages (Edemacu et al., 2025; Su et al., 2024; Xiang et al., 2024). Citation-grounded generation also evaluates whether claims are supported by their cited sources (Bohnet et al., 2023; Gao et al., 2023). These approaches address malicious instructions, corrupted evidence, or unsupported outputs. VCR instead changes source-side incentives across rounds by penalizing suspicious changes and rewarding factual content supported by the earlier source.

Strategic games and mechanisms. Strategic classification studies adaptation to a decision rule and distinguishes manipulation from costly improvement (Hardt et al., 2016; Kleinberg & Raghavan, 2019). Performative prediction studies how deployment changes future data (Perdomo et al., 2020), while recommendation research models strategic content providers and supply-side equilibria (Ben-Porat & Tennenholtz, 2018; Jagadeesan et al., 2023). Repeated Stackelberg games capture leader–follower interaction with an adapting opponent (Marecki et al., 2012; Von Stackelberg, 2010). We use this structure because the supplier commits to a rewrite before the platform updates its defense from before/after pairs. VCR then makes source-supported factual content part of the supplier's profitable response instead of treating defense only as a filter.

3 Framework: A Repeated Game of Supplier and Platform

We model a repeated Stackelberg game with supplier leadership and partial monitoring.

3.1 Notation

For query $q \in \mathcal{Q}$, let $D(q) = \{d_1, \ldots, d_K\}$ be the retrieved candidates and $T(q) \subseteq D(q)$ the supplier's target documents. Each document d_i has latent quality $q_i \in \mathbb{R}$, capturing user-useful content, and manipulation intensity $m_i \in \mathbb{R}$, capturing citation-seeking surface patterns. Let $\rho = \mathrm{corr}(q_i, m_i)$. The engine assigns citation probability by

$$v_i(\alpha) = \beta_q q_i - \alpha m_i + b_i, \qquad c_i(\alpha) = \mathrm{softmax}\{v_j(\alpha)\}_i, \tag{1}$$

where $\beta_q > 0$ is the quality weight, $\alpha \geq 0$ the platform defense strength, and b_i independent residual noise. Supplier exposure is score $g_T = \sum_{i \in T(q)} w_i(q) c_i$, with position-aware citation weight $w_i(q)$.

3.2 Repeated Game

At round t, the supplier observes the previous defense $\pi_{P,t-1}$ and rewrites its targets, yielding $D_t^a = \pi_{A,t}(D_t)$. The platform only observes before/after pairs $\mathcal{O}_t = \{(d_i, d_i^a) : i \in T_t(q), q \in \mathcal{Q}\}$, estimates the rewrite strategy, and updates its answer-time defense $\pi_{P,t}$. The supplier maximizes citation exposure net of rewrite cost:

$$u_A(\pi_A, \pi_P) = g_T(\pi_A, \pi_P) - \kappa_A(\pi_A), \tag{2}$$

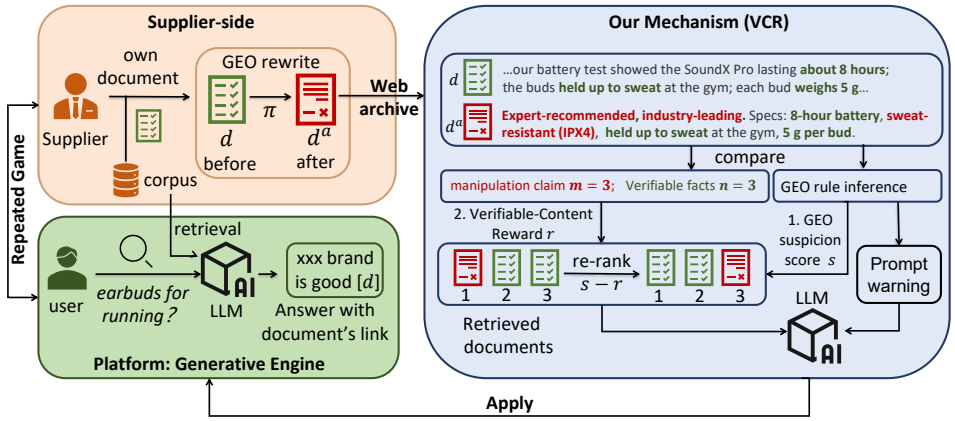


Figure 2: Overview of VCR. The platform extracts GEO rules from before/after pairs, rewards verifiable content, and applies a soft re-ranking for answer-time defense.

where κ_A captures editing cost, factual drift, and detectability. The platform suppresses manipulated citations while avoiding false positives:

$$u_P(\pi_A, \pi_P) = -g_T(\pi_A, \pi_P) - \mu f(\pi_P) - \frac{\gamma}{2} r(\pi_P). \tag{3}$$

Under Eq. (1), the leading false-positive cost is proportional to $\alpha\beta_q \mathrm{cov}(q, m)$: penalizing manipulation also suppresses quality when q and m are correlated. Since the platform never observes the true rewrite rule r_t, it estimates $\hat{r}_t$ from $\mathcal{O}_t$.

3.3 Theoretical Analysis

We characterize the platform best response and the induced fixed point. Let B and Q be the first- and second-order terms in the local response of target GEO stated in Lem. 1. The formal definition, secondary results, and proofs are in Apps. G and H.

Theorem 1 (Local platform best response)**.** *Under Eq. (1) and Eq. (3), suppose the Taylor remainder is negligible in a neighborhood of zero and $Q + \gamma > 0$. The unique minimizer of the platform's quadratic local loss is*

$$\alpha^* = \max\left\{0, \frac{B - \mu\beta_q \rho \sigma_q \sigma_m}{Q + \gamma}\right\}, \tag{4}$$

where σ_q and σ_m are the standard deviations of q and m. Holding all other local moments and coefficients fixed, α^* is weakly decreasing in ρ.

Implication. The default game is one-sided: it penalizes manipulation but rewards no feature directly aligned with platform/user utility. High quality–manipulation correlation and partial monitoring weaken defense. The resulting local stationary outcome can be inert: useful content is not induced, suspicious content is weakly penalized, and further platform updates have little effect.

4 Welfare-Incentivizing Mechanism

Based on Sec. 3, VCR combines a source-supported content reward with a suspicion penalty inferred from supplier rewrites, then softly re-ranks sources at answer time (Fig. 2, Algorithm 1).

4.1 Verifiable-Content Reward

Corollary 2 identifies conditions under which penalty-only defenses can stall: when content-improvement gradients are exhausted and residual benefits offset harms, the local stationary response produces little utility change. The mechanism adds a reward channel that gives the supplier an incentive to surface source-supported substance.

Definition 1 (Verifiable supported content). *For pair (d_i, d_i^a), let n_i be the number of factual claims, numerical details, or named citations present in the rewrite and supported by the original document and surfaced more saliently in the rewrite. We estimate n_i using a pair-level LLM oracle.*

In the running example of Fig. 2, the rewrite surfaces three facts that are checkable against the original, namely the 8-hour battery life, sweat resistance, and per-bud weight, so $n_i = 3$. Additions with no support in the original, such as "expert-recommended" and "industry-leading," earn no credit and instead feed the manipulation signal m_i of Eq. (1), which the platform estimates by rule matching in Sec. 4.2. The claim count in the illustration is not the continuous model variable m_i. Each rewrite earns the credit

$$r_i = \lambda \cdot \min(c_{\max}, c_n n_i), \tag{5}$$

where λ is reward strength, c_n the per-claim credit, and $c_{\max}$ the credit cap that bounds how much reward any single rewrite can farm. The reward alone, however, cannot tell substantive rewrites from manipulative ones; the platform still needs a penalty side that tracks the supplier's current GEO strategy, which we extract next.

4.2 GEO Rule Extraction

To build this penalty side, the platform turns the supplier's own rewrite history into evidence of its strategy. At round t, the platform observes before/after pairs $\mathcal{O}_t = \{(d_i, d_i^a)\}$ and infers a rule set $S_P^{(t)}$ describing the supplier's current GEO strategy. We follow the Explainer–Extractor–Merger–Filter pipeline of AutoGEO (Wu et al., 2025), but reverse its use: instead of comparing preferred and less-used documents to infer engine preferences, the platform compares original and rewritten documents to infer suspicious rewrite patterns.

For each pair (d_i, d_i^a), the platform computes edit magnitude Δ_i and skips trivial rewrites with $\Delta_i < \theta$. The output is $S_P^{(t)}$ and each retrieved document receives a baseline suspicion score

$$s_j^{\text{base}} = \textsc{Match}(d_j, S_P^{(t)}) \cdot \Delta_j, \tag{6}$$

where $\textsc{Match}(\cdot) \in [0, 1]$ measures rule match strength and Δ_j is recent edit magnitude. The suspicion penalty then calibrates the reward of Eq. (5) into a single score,

$$s_i^{\text{new}} = s_i^{\text{base}} - r_i, \tag{7}$$

so cosmetic and promotional rewrites stay suspicious while supported facts earn credit.

4.3 Soft Re-rank

The platform converts the combined score s_i^{new} into a soft defense signal. It reorders documents, placing less suspicious and more verification-friendly sources earlier in the context. Finally, the platform adds a system-level warning from $S_P^{(t)}$ that instructs the engine to treat these labels as auxiliary evidence, prioritize verifiable claims, and avoid over-crediting sources whose citation value mainly comes from manipulative rewriting. No document is removed: high-suspicion documents may still be cited when they provide uniquely necessary evidence.

4.4 Analysis

Theoretical analysis. We analyze how VCR changes utility at a local stationary response. In Eq. (1), the term $-\alpha m_i$ is the linearized effect of the platform's suspicion channel on the engine's pre-softmax citation logit. Likewise, the VCR credit in Eq. (5) induces a positive $+\lambda n_i$ shift, absorbing c_n into λ and ignoring the cap locally. The augmented logit is

$$v_i(\alpha, \lambda) = \beta_q q_i - \alpha m_i + \lambda n_i + b_i. \tag{8}$$

The key assumption is that source-supported content is locally separable from manipulation and positively aligned with the shared platform/user utility.

Across datasets, $\widehat{\rho}_{m,n} \in [-0.05, 0.1]$ indicates weak observed association, not the separability or mixed-partial assumptions needed for the theorem.

Algorithm 1 Mechanism-augmented platform policy π_P^λ

Input: Before/after pairs $\mathcal{O}_t = \{(d_i, d_i^a)\}$, threshold θ, reward parameters $(\lambda, c_n, c_{\max})$.

$\mathcal{R} \leftarrow \emptyset$
for each $(d_i, d_i^a) \in \mathcal{O}_t$ **do**
    Compute edit magnitude Δ_i; continue if $\Delta_i < \theta$.
    $e_i \leftarrow \textsc{Explainer}(d_i, d_i^a)$; $r_i \leftarrow \textsc{Extractor}(e_i)$.
    $\mathcal{R} \leftarrow \mathcal{R} \cup r_i$.
$S_P \leftarrow \textsc{Filter}(\textsc{Merger}(\mathcal{R}))$.
for each retrieved document d_j **do**
    Set Δ_j to recent edit magnitude, or 0 if unobserved.
    $s_j^{\text{base}} \leftarrow \textsc{Match}(d_j, S_P) \cdot \Delta_j$.
    If pair (d_j^0, d_j) exists, $n_j \leftarrow \textsc{Oracle}(d_j^0, d_j)$; else $n_j \leftarrow 0$.
    $s_j^{\text{new}} \leftarrow s_j^{\text{base}} - \lambda \min(c_{\max}, c_n n_j)$.
$W \leftarrow \textsc{Warning}(S_P)$; soft-rerank by ascending s_j^{new}.
return $(W, \{s_j^{\text{new}}\}, \text{order})$.

Theorem 2 (Local two-sided utility improvement)**.** *Under Asm. 1, let $(\delta^*(\lambda), \alpha^*(\lambda))$ be the local stationary response of Def. 2. Then $\alpha^*(\lambda) = \alpha^*(0) + O(\lambda^2)$ and platform/user utility satisfies*

$$\Delta U^*(\lambda) = \Delta U^*(0) + \eta_n \lambda \Theta + O(\lambda^2), \tag{9}$$

where $\Theta = (H^{-1})_{nn} > 0$. Hence, for sufficiently small $\lambda > 0$, $\Delta U^*(\lambda) > \Delta U^*(0)$. If the default equilibrium is locally inert, VCR strictly improves the platform/user side to first order. The creator's optimized local objective changes only at second order. To first order, VCR therefore improves platform/user utility while preserving the creator's local objective.

The proof is in App. H, with additional conditions in App. G. VCR introduces a strategic direction beyond quality and manipulation: suppliers can gain citation value by surfacing checkable substance while the platform continues to penalize manipulation.

Practical deployment. VCR requires document versions; without a prior version, $n = 0$ and the policy reduces to suspicion-only defense. App. C.1 discusses deployment, and Sec. 5.4 evaluates reward scope, verification evidence, and rule disclosure.

5 Experiments

We evaluate whether the proposed verifiable-content reward (VCR) improves the joint defense–utility outcome in repeated simulations. Additional direct-quality, judge-robustness, adaptive-supplier, and case-study analyses are in App. F.

5.1 Experimental Setup

Datasets. We use the retrieval-augmented E-commerce Wu et al. (2025), GEO-Bench Aggarwal et al. (2024), and Researchy-GEO Wu et al. (2025) benchmarks. Following AutoGEO Wu et al. (2025), each query has $K = 5$ candidates. Default-engine runs use 416, 1,000, and 1,000 queries, respectively, over five rounds.

Metrics. Following AutoGEO (Wu et al., 2025) and the GEO visibility metric (Aggarwal et al., 2024), the target GEO score g aggregates the amount, position, and prominence of target-document citations. At each round, let g_0, g_a, and g_d denote mean target GEO scores before rewriting, after rewriting without defense, and after defense, respectively. We compute

$$\text{Def.} = (g_a - g_d)/g_a, \qquad \text{Welf.} = (g_d - g_0)/g_0, \qquad \text{Net} = \text{Def} + \text{Welf.} \tag{10}$$

Def. measures relative reduction in attacked target visibility; Welf. measures creator exposure against the same round's pre-rewrite reference. Net sums these exposure proxies. The ± 5-point band is an operational tolerance based on prior run variability, not a formal equivalence test. App. E defines the metrics and App. F.2 reports paired query-level intervals. Document and answer quality provide an independent check.

Table 1: Repeated-game results across three benchmarks and three answer engines. Def. is relative visibility reduction, Welf. is creator exposure change, and Net is their sum, all in percentage points. The ± 5-point Welf. band is an operational tolerance. Results are at round 5. Metrics are computed before rounding.

Defense	E-COMMERCE			GEO-BENCH			RESEARCHY-GEO		
	Def.	Welf.	Net	Def.	Welf.	Net	Def.	Welf.	Net
Gemini-2.5-Flash-Lite									
Prompt	2.7	−2.0	0.7	3.4	−3.5	−0.1	0.8	−0.7	0.1
Hard reject	50.3	−50.8	−0.6	33.8	−33.2	0.6	63.9	−63.3	0.5
Keyword scrub	−1.0	0.0	−1.0	−0.1	0.3	0.1	0.2	0.2	0.3
VCR (ours)	**14.4**	**2.2**	**16.6**	**11.4**	**−3.3**	**8.0**	**8.0**	**−1.7**	**6.3**
GPT-4o-mini									
Prompt	−0.7	−3.2	−3.9	−0.5	−1.9	−2.5	−0.4	0.4	0.0
Hard reject	22.6	−21.6	1.0	67.3	−67.5	−0.2	−18.9	20.0	1.1
Keyword scrub	−0.8	1.1	0.3	1.7	−2.0	−0.2	−2.5	2.6	0.1
VCR (ours)	**10.9**	**−1.2**	**9.7**	**9.7**	**−4.7**	**5.0**	**5.0**	**−1.7**	**3.3**
Claude-Haiku-4.5									
Prompt	20.0	−22.9	−2.9	−3.4	3.5	0.1	−0.9	0.2	−0.7
Hard reject	5.5	−5.6	−0.1	16.6	−19.8	−3.1	85.7	−85.7	0.0
Keyword scrub	0.4	1.1	1.5	−2.7	6.4	3.7	2.1	−1.2	0.9
VCR (ours)	**29.1**	**−3.5**	**25.6**	**29.1**	**−4.8**	**24.3**	**22.0**	**−1.7**	**20.3**

Table 2: Direct quality on E-COMMERCE. All metric names and averages are reported explicitly.

	Rewritten document						**Generated answer**			
Defense	Clarity	Depth	Insight	Factuality	Usefulness	Average	Clarity	Depth	Insight	Average
Prompt	.820	.696	.605	.778	.718	.724	.549	.512	.457	.506
Hard reject	.810	.727	.641	.788	.740	.741	.554	**.526**	.464	.515
Keyword scrub	.813	.701	.608	.785	.724	.726	.554	.524	.465	.515
VCR	**.821**	**.741**	**.660**	**.820**	**.757**	**.760**	**.564**	**.526**	**.471**	**.520**

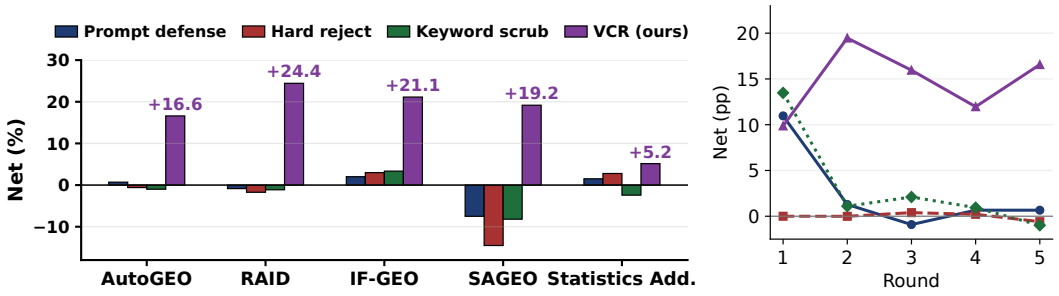
Baselines and implementation. We compare VCR with three classical defenses: **Prompt defense** labels, softly reorders, and warns about suspicious documents; **Hard reject** removes them; and **Keyword scrub** filters GEO-style phrases. Unless stated otherwise, the attacker is AUTOGEO, the answer engine is `gemini-2.5-flash-lite`, and $\lambda = 1$ (Apps. C and D).

5.2 MAIN RESULTS

Net outcome across datasets and engines. VCR has the largest Net in all nine settings, averaging 12.1 points above the strongest baseline (Table 1). Its creator-exposure estimates all lie within the operational tolerance. Under the default engine, Hard reject suppresses exposure almost one-for-one with defense gains, while Prompt and Keyword scrub approach inert outcomes. App. F.1 examines alternative stakeholder weights.

Direct document and answer quality. Table 2 provides an independent LLM-rubric check: VCR has the highest point estimates on all document dimensions and answer-quality average. App. F.7 reports per-round answer scores.

Generalization across GEO attack strategies. Figure 3(a) replaces AUTOGEO with RAID (Chen et al., 2025c), IF-GEO (Zhou et al., 2026), SAGEO (Kim et al., 2026), and Statistics Addition (Aggarwal et al., 2024). Across all five attackers, VCR maintains positive R_5 Net and exceeds the classical defenses.



(a) R_5 Net (%) across five GEO attackers. (b) Per-round Net (pp).

Figure 3: VCR across GEO attackers and rounds on E-COMMERCE. (a) R_5 results for five attackers. (b) Five-round AUTOGEO trajectories; only VCR remains positive throughout.

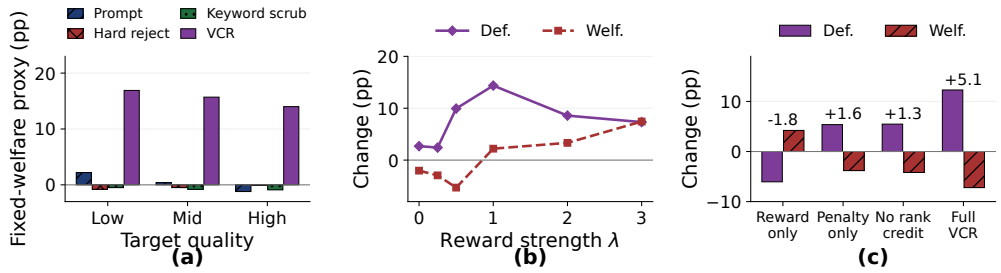


Figure 4: Sensitivity and ablations on E-COMMERCE: (a) an illustrative target-quality comparison holding Welf. fixed, (b) reward strength λ, and (c) component ablations on the same 100 queries. Bar annotations in (c) show Net.

5.3 ROBUSTNESS ANALYSIS

Round-by-round performance. Figure 3(b) shows that VCR alone stays positive across all five AUTOGEO rounds. Prompt and Keyword scrub lose their initial positive Net advantage and approach zero; Hard reject fluctuates near zero as exposure losses offset defense gains. VCR is not the strongest method at the first round, but leads from rounds two through five. App. F.6 gives the other datasets.

Illustrative target-quality sensitivity. Figure 4(a) varies the target-quality regime while holding each method's Welf. at its default R_5 reference. Low, Mid, and High denote target quality among the five candidates. This fixed-welfare proxy illustrates sensitivity; it is not independently measured Net for each regime.

Ablation on reward strength λ. Figure 4(b) sweeps the reward strength λ. Def. peaks at $\lambda = 1$ and decreases at larger rewards, while Welf. becomes positive for $\lambda \geq 1$. The observed defense–exposure trade-off is non-monotone; differing query-set sizes (App. C) limit causal interpretation.

Ablation on the penalty and reward channels. Figure 4(c) compares all four arms on the same 100 queries. Reward-only (No suspicion) has Net -1.8; penalty-only (No reward) has $+1.6$. No re-rank exposes the same pair-oracle output through the prompt without using it as a ranking credit, yielding $+1.3$. Full VCR has the highest Net, $+5.1$, supporting a contribution from combining penalty, reward, and ranking. Its Welf. is -7.2, outside the operational tolerance: this diagnostic supports the Net ranking, not exposure preservation in every ablation run. The 0.3-point difference between the two prompt-based arms alone does not establish an oracle effect.

Additional robustness checks. The VCR–Prompt reward ordering is unchanged under three independent judges (App. F.3); VCR also has the best multi-turn answer utility and precision (App. F.4), and the extended run changes more modestly after R_5 before ending at R_8 (Fig. 7). Apps. C and I give deployment, prompt, oracle, and roughly 400-ms latency details.

Table 3: Platform-inferred GEO rules across rounds on E-COMMERCE (abbreviated).

Round	Exploits (baseline)	VCR (ours)
R_1	**Formatting:** Structure, bullets, bolding.	**Quality:** Scannable text; nuanced distinctions.
R_3	**Manipulation:** Source framing; Non-GEO style.	**Explanation:** Causal links and attribution.
R_5	**Manipulation:** Authority claims; hidden GEO intent.	**Verifiability:** Verifiable, attributed core findings.

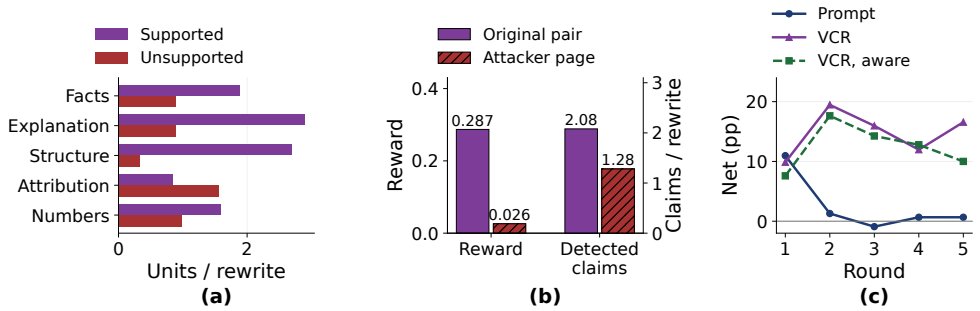


Figure 5: VCR reward-channel design on E-COMMERCE: (a) five reward scopes, (b) original-pair versus attacker-created evidence, and (c) disclosure of the VCR rule to suppliers.

5.4 UNDERSTANDING THE EFFECT OF VCR

RQ1: what content should the reward credit? We re-score the same 150 rewrite pairs under five reward scopes, separating supported from unsupported units (Fig. 5a). We find out that rewarding authority cues could therefore encourage convincing-looking sourcing without grounded substance. Explanation and structure also yield supported units, so the results do not imply that only facts are useful. Rather, they motivate source support as the eligibility criterion.

RQ2: what evidence should verification trust? A supplier could fabricate a claim and create a supporting page to corroborate it. Replacing the original document with attacker-authored evidence reduces detected unsupported claims (Fig. 5b), revealing a circular-verification failure. Credit also decreases, indicating weaker detection rather than reward inflation. Verification against the prior version blocks this evidence substitution but does not establish the original source's truth.

RQ3: how do supplier strategies change? Table 3 tracks platform-inferred rewriting rules. Both settings initially emphasize formatting. Later baseline rules emphasize authority framing and hidden GEO intent, whereas VCR rules emphasize explanation, attribution, and verifiability. Because the extractor is shared, the contrast is consistent with suppliers changing what they optimize, rather than only receiving a different final ranking. These extracted rules are qualitative evidence, complemented by direct-quality results in Table 2. Apps. F.8, F.10, and F.11 give additional trajectories and examples.

RQ4: can VCR be gamed over repeated rounds? We disclose the scoring rule and instruct suppliers to split or repeat checkable-looking facts while hiding manipulation cues (Fig. 5c). VCR remains positive throughout, although below the standard-supplier run in four rounds. Thus disclosure weakens some outcomes without eliminating the observed Net advantage.

6 CONCLUSION

We studied repeated platform–creator citation competition, where conventional defenses can suppress both manipulation and productive improvement. VCR aligns creator incentives with source-supported substance by rewarding useful contributions and penalizing manipulation. Across our simulations, it improves Net and direct-quality scores while keeping creator-exposure point estimates within the operational tolerance in all nine settings. These findings position incentive design as a foundation for reliable agentic search.

AI USE STATEMENT

LLM tools served as experimental components for answer generation, document rewriting, rule extraction, and evaluation, and assisted with manuscript polishing, visualization, and consistency checks. Models, prompts, and configurations are documented in Appendices C and I. The authors independently checked all reported results, reviewed all AI-assisted materials, and take full responsibility for the accuracy and integrity of the final work.

REFERENCES

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5–16, 2024.

Puneet S Bagga, Vivek F Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. E-geo: A testbed for generative engine optimization in e-commerce. *arXiv preprint arXiv:2511.20867*, 2025.

Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, and Ricardo A. Baeza-Yates. Link-based characterization and detection of web spam. In *AIRWeb 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web, Seattle, Washington, USA, 10 August 2006, co-located with SIGIR 2006*, pp. 1–8, 2006.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2023.

Carlos Castillo and Brian D Davison. Adversarial web search. *Foundations and trends in Information Retrieval*, 4(5):377–486, 2011.

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. *arXiv preprint arXiv:2509.05607*, 2025a.

Xiaolu Chen, Jie Bao, Haojie Wu, Zhen Chen, and Yong Liao. Caption injection for optimization in generative search engine. *arXiv preprint arXiv:2511.04080*, 2025b.

Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang. Role-augmented intent-driven generative search engine optimization. *arXiv preprint arXiv:2508.11158*, 2025c.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Kennedy Edemacu, Vinay M. Shashidhar, Micheal Tuape, Dan Abudu, Beakcheol Jang, and Jong Wook Kim. Defending against knowledge poisoning attacks during retrieval-augmented generation. *arXiv preprint arXiv:2508.02835*, 2025.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024.

Dennis Fetterly, Mark Manasse, Marc Najork, and Janet L. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*, 2003.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. Ranking robustness under adversarial document manipulations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 395–404, 2018.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023.

Zoltán Gyöngyi and Hector Garcia-Molina. Link spam alliances. In *VLDB*, volume 5, pp. 517–528, 2005.

Zoltán Gyöngyi and Hector Garcia-Molina. Web spam taxonomy. In *AIRWeb 2005, First International Workshop on Adversarial Information Retrieval on the Web, co-located with the WWW conference, Chiba, Japan, May 2005*, pp. 39–47, 2005.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122, 2016.

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Sunghwan Kim, Wooseok Jeong, Serin Kim, Sangam Lee, and Dongha Lee. Sageo arena: A realistic environment for evaluating search-augmented generative engine optimization. *arXiv preprint arXiv:2602.12187*, 2026.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, 2019.

Aounon Kumar and Himabindu Lakkaraju. Manipulating large language models to increase product visibility. *arXiv preprint arXiv:2404.07981*, 2024.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Zikang Liu and Peilan Xu. Think before writing: Feature-level multi-objective optimization for generative citation visibility. *arXiv preprint arXiv:2604.19113*, 2026.

Janusz Marecki, Gerry Tesauro, and Richard Segal. Playing repeated stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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 (EMNLP)*, 2023.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. *Games and Economic Behavior*, 35 (1–2):166–196, 2001.

Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web? the evolution of the web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, 2004.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pp. 83–92, 2006.

Sejoon Oh, Gaurav Verma, and Srijan Kumar. Adversarial text rewriting for text-aware recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. ACM, 2024.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Haritz Puerto, Martin Gubri, Tommaso Green, Seong Joon Oh, and Sangdoo Yun. C-seo bench: Does conversational seo work? *arXiv preprint arXiv:2506.11097*, 2025.

Jinyan Su, Jin Peng Zhou, Zhengxin Zhang, Preslav Nakov, and Claire Cardie. Towards more robust retrieval-augmented generation: Evaluating rag under adversarial poisoning attacks. *arXiv preprint arXiv:2412.16708*, 2024.

Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

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. *arXiv preprint arXiv:2604.19516*, 2026.

Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. *arXiv preprint arXiv:2510.11438*, 2025.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Junwei Yu, Mufeng Yang, Yepeng Ding, and Hiroyuki Sato. Structural feature engineering for generative engine optimization: How content structure shapes citation behavior. *arXiv preprint arXiv:2603.29979*, 2026.

Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. Agenticgeo: A self-evolving agentic system for generative engine optimization. *arXiv preprint arXiv:2603.20213*, 2026.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27576–27590, 2026.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A Limitations and Open Directions

VCR makes several simplifying assumptions. First, the theory analyzes a local quadratic surrogate and the experiments use a finite interaction horizon; neither establishes a global equilibrium for unrestricted rewriting strategies. The model also considers one platform and one creator population, leaving multi-platform competition and paid placement to future work.

Second, the reward checks a rewrite against its earlier version. It measures source support and factual salience, not independent truth: an inaccurate statement already present in the earlier version may still receive credit. Moreover, LLM-based claim counting and rule matching can err, and suppliers may repeat or split supported claims to approach the reward cap. The cap and manipulation penalty reduce, but do not eliminate, these risks.

Finally, the reported utilities represent creator exposure and a joint platform/user objective, with equal weight in Net. The ± 5-point creator equivalence band is an operational tolerance motivated by prior repeated-run variation, not a formal equivalence test for each current estimate. We report paired query-level bootstrap intervals for the default-engine results; additional independent pipeline runs would quantify model- and trajectory-level variability beyond this query-sampling uncertainty.

B Additional Related Work

Search manipulation and GEO. Earlier web-spam research studies link alliances, content signals, link-based detection, and adversarial search (Becchetti et al., 2006; Gyöngyi & Garcia-Molina, 2005; Ntoulas et al., 2006). More recent attacks poison retrieved knowledge or rewrite content shown to language-model agents and text-aware recommenders (Chen et al., 2024; Oh et al., 2024; Xue et al., 2024; Zou et al., 2025). Within GEO, later work broadens the setting to e-commerce and conversational search, multimodal and structural optimization, feature-level objectives, reusable agent strategies, and source-influence evaluation (Bagga et al., 2025; Chen et al., 2025b;a; Liu & Xu, 2026; Puerto et al., 2025; Wu et al., 2026; Yu et al., 2026). These directions expand the attack and evaluation space around the repeated supplier–platform interaction studied here.

Citation and factuality evaluation. RAG evaluation measures retrieval and answer faithfulness, while factuality evaluation measures support for atomic claims (Es et al., 2024; Min et al., 2023). These measurements motivate checking whether a rewrite adds supported factual content. VCR additionally addresses the supplier incentives created by citation-based visibility.

Mechanism design. Algorithmic mechanism design changes allocation or payment rules to induce desired behavior (Nisan & Ronen, 2001). VCR uses the same incentive perspective without introducing payments: it changes citation ordering through a reward that the platform can verify from document versions.

C Implementation Details

Hardware and runtime. All experiments are run on a SLURM cluster. Each job uses 8 CPU cores and 64 GB RAM; no GPU is used. A single five-round game on the full $1{,}000$-query test split takes 18–24 wall-clock hours on GEO-Bench and Researchy-GEO, and 12–15 hours on E-commerce. Smaller diagnostic runs and ablations typically take 4–6 hours. VCR adds one cached, JSON-only oracle call per changed pair to the rule extraction shared with Prompt defense. In an online latency test, all four defenses complete answer-time processing in approximately 400 ms; the pair-level oracle runs off the answer-time path.

Models. The answer-generation engine G is `gemini-2.5-flash-lite`. The platform-side rule extractor (Explainer/Extractor/Merger/Filter), the change-magnitude judge, and the verifiable-content oracle all use `gpt-4o-mini` with temperature 0 and JSON-only decoding when applicable. We pre-compute and cache quality-judge scores for each dataset so that target selection is deterministic across runs. Default-engine main evaluations use 416 queries on E-commerce and 1,000 on each other dataset. The reward-aware E-commerce run uses the same 416-query set; the

four component-ablation arms share a 100-query set. The reward-strength sweep uses 416 queries at $\lambda = 0, 1$ and 100 at the other values; it is an exploratory comparison across these run sizes.

Game configuration. Each repeated game runs for $T = 5$ rounds. To make trajectories comparable across defenses, we disable early stopping in all reported experiments; the supplier and platform always interact for all five rounds. The change-magnitude threshold θ is set so that the median edited document contributes to the rule set in round R_1. Unless otherwise stated, the reward strength is $\lambda = 1$. The implementation uses $c_n = 0.05$ and $c_{\max} = 0.4$, on the scale of individual suspicion penalties. The λ sweep in Fig. 4(b) covers the same effective range as multiplying the per-claim credit by an equivalent factor.

C.1 Deployment Details

Algorithm 1 in the main text summarizes the platform policy.

Discussion on practical deployment. VCR targets versioned content, which covers a substantial practical scope: a previous study found that about 40% of pages change within a week, whereas about 8% are newly created (Fetterly et al., 2003; Ntoulas et al., 2004). For a new page without a prior version, $n = 0$, so VCR degrades gracefully to its $\lambda = 0$ suspicion-only special case. On the reward side, the conservative pair-verifiable scope limits adjudication ambiguity and gaming; we analyze this choice, together with the verification anchor and rule disclosure, in Sec. 5.4.

D Defense and Attacker Baselines

Defense baselines. We compare against three classical platform-side defenses. **Prompt defense** attaches a per-document suspicion label, softly reorders sources by ascending suspicion, and inserts the warning template in App. I.2 into the engine's system prompt. No document is removed. **Hard reject** drops any document whose suspicion score exceeds $\tau = 0.45$ before generation; the threshold is tuned on a 50-query development split. **Keyword scrub** disables all prompt-defense channels and instead applies a regex filter over 24 GEO-style phrases, such as "optimized for GEO", "expert-curated", and "machine-readable listing", to every incoming document. The full phrase list is released with our code.

GEO attacker baselines. In Sec. 5.3, we evaluate robustness against four non-AutoGEO attackers. RAID (Chen et al., 2025c) extracts engine-preference rules from contrastive role-conditioned answers; we use the authors' released codebase with the engine replaced by ours. IF-GEO (Zhou et al., 2026) performs influence-based GEO; we re-implement its influence proxy on top of our engine. SAGEO (Kim et al., 2026) performs structure-aware GEO via prompt programming; we use the authors' prompts with the top-5 candidate set from our retriever. Statistics Addition (Aggarwal et al., 2024) adds quantitative claims following the original paper.

E Evaluation Metrics

We adopt the AutoGEO framework of Wu et al. (2025) and the visibility metric of Aggarwal et al. (2024). For a query q, the engine retrieves a candidate set D_q, and an LLM G produces an answer $a = G(q, D_q)$. Each candidate document $d \in D_q$ receives a citation visibility score

$$\mathrm{Vis}(d, a) = \mathrm{Word}(d, a) + \mathrm{Pos}(d, a) + \mathrm{Overall}(d, a), \tag{11}$$

where $\mathrm{Word}(d, a)$ is the normalized word count of sentences in a citing d, $\mathrm{Pos}(d, a)$ is the location-based weight of the source-linked text, and $\mathrm{Overall}(d, a)$ integrates the two signals. We follow Aggarwal et al. (2024) for the exact implementation.

Given the supplier's target set $T(q) \subseteq D_q$, the target GEO score for query q is

$$g(q) = \sum_{d \in T(q)} \mathrm{Vis}(d, a), \qquad g = \mathbb{E}_q[g(q)]. \tag{12}$$

A higher g means the target documents are cited more prominently in the engine's answer.

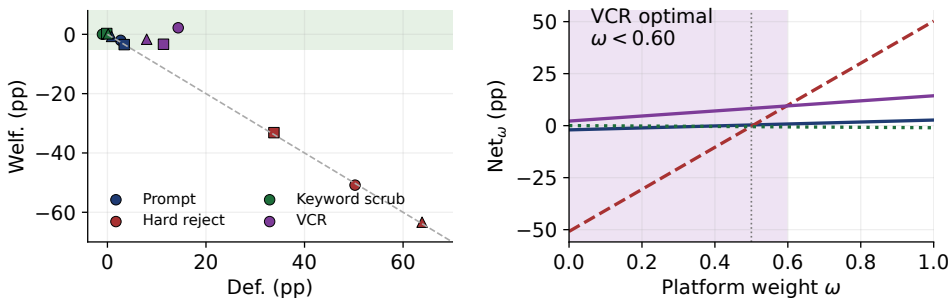


Figure 6: Sensitivity of the defense ranking to the Net weighting. **(a)** Pareto view of (Def., Welf.) for all defenses and datasets; the shaded area marks Welf. above -5 points; positive Def. is also required by our operational criterion. Circle, square, and triangle markers denote E-commerce, GEO-Bench, and Researchy-GEO. **(b)** Net_ω on E-COMMERCE: VCR is optimal for every platform weight ω below ≈ 0.60.

For each defense and round, g_0 is the score of the selected target documents before that round's rewrite, g_a is the score after rewriting without defense, and g_d is the score after defense. The selected-target reference is recomputed each round and is not the fixed raw-initial corpus score. Metrics use means over matched queries before taking ratios. We report

$$\mathrm{Def.} = (g_a - g_d)/g_a, \qquad \mathrm{Welf.} = (g_d - g_0)/g_0, \qquad \mathrm{Net} = \mathrm{Def.} + \mathrm{Welf.}. \tag{13}$$

Def. measures the relative reduction in attacked target visibility, not the fraction of attack-induced excess visibility removed. It is an exposure-based defense proxy, not a direct measure of answer quality. Welf. measures creator target-source exposure relative to the no-attack baseline; negative values mean that defense suppresses exposure below its natural level. Net sums these two normalized exposure measures. For $g_a, g_0 > 0$, it simplifies to $g_d(g_a - g_0)/(g_a g_0)$, so positive Net alone does not establish answer-quality improvement or two-sided benefit. We assess creator exposure separately and use direct document and answer evaluations as complementary evidence.

Prior repeated runs showed approximately 5 percentage-point standard deviation, so the pre-specified ± 5-point empirical equivalence band treats creator utility in this interval as exposure preserved within run variability rather than as a meaningful loss or gain. We nevertheless report every signed point estimate.

F Additional Experiments and Case Studies

F.1 Sensitivity to the Net Weighting

Net weights the two stakeholder utilities equally. To check that the defense ranking does not depend on this choice, define

$$\mathrm{Net}_\omega = \omega\,\mathrm{Def} + (1-\omega)\,\mathrm{Welf}, \qquad \omega \in [0,1], \tag{14}$$

so that $\omega = 0.5$ recovers the main metric up to scale. Figure 6(a) shows the Pareto view of the four defenses across the three datasets: Hard reject lies near the $\mathrm{Net} = 0$ diagonal, meaning it buys defense one-for-one with creator exposure; Prompt and Keyword scrub cluster at the origin; and VCR combines appreciable defense with creator exposure near its pre-rewrite reference. Figure 6(b) sweeps ω on E-COMMERCE. VCR is the optimal defense for all $\omega < 0.60$ on E-COMMERCE, for $\omega \in (0.24, 0.57)$ on GEO-BENCH, and for $\omega \in (0.19, 0.52)$ on RESEARCHY-GEO. Hard reject overtakes VCR as the platform weight increases beyond the dataset-specific crossover, while Keyword scrub can win at low platform weights. A weight near 0.6 still assigns substantial weight to creator exposure; these are trade-offs within a two-sided objective, not cases in which one stakeholder is absent.

F.2 Statistical Uncertainty

To quantify the statistical uncertainty of the reported utilities, we compute paired per-query bootstrap confidence intervals at round 5. For each run we pair the three per-query target GEO scores that enter

Table 4: Paired per-query bootstrap 95% confidence intervals at round 5 for the full-scale main runs (percentage points).

Dataset	Defense	Def. [CI]	Welf. [CI]	Net [CI]
E-COMMERCE	Prompt	2.7 [0.2, 5.1]	−2.0 [−4.6, 0.6]	0.7 [−1.5, 2.8]
	Hard reject	50.3 [40.0, 59.8]	−50.8 [−60.3, −40.9]	−0.6 [−2.0, 0.7]
	Keyword scrub	−1.0 [−2.7, 0.6]	0.0 [−1.9, 1.9]	−1.0 [−2.9, 0.9]
	VCR	14.4 [12.0, 16.8]	2.2 [−0.9, 5.4]	**16.6** [13.7, 19.6]
GEO-BENCH	Prompt	3.4 [2.2, 4.6]	−3.5 [−4.7, −2.2]	−0.1 [−1.1, 1.0]
	Hard reject	33.8 [26.7, 40.7]	−33.2 [−40.1, −25.9]	0.6 [−0.1, 1.4]
	Keyword scrub	−0.1 [−1.0, 0.7]	0.3 [−0.7, 1.2]	0.1 [−0.8, 1.1]
	VCR	11.4 [10.1, 12.7]	−3.3 [−4.7, −2.1]	**8.0** [6.8, 9.3]
RESEARCHY-GEO	Prompt	0.8 [−0.2, 1.9]	−0.7 [−1.8, 0.3]	0.1 [−0.8, 1.1]
	Hard reject	63.9 [58.3, 69.2]	−63.3 [−68.8, −57.8]	0.5 [0.2, 0.9]
	Keyword scrub	0.2 [−0.6, 0.9]	0.2 [−0.6, 1.0]	0.3 [−0.5, 1.2]
	VCR	8.0 [6.9, 9.0]	−1.7 [−2.8, −0.5]	**6.3** [5.3, 7.3]

Table 5: Verifiable-content reward assigned by three independent judges.

Rewrites from	GPT-4o-mini	Claude-Haiku-4.5	Gemini-2.5-Flash-Lite
Prompt defense	0.213	0.180	0.244
VCR	**0.243**	**0.194**	**0.273**

the metrics (no-attack, attacked, and defended), resample queries with replacement (10^4 replicates), and recompute Def., Welf., and Net on each replicate; Table 4 reports 95% percentile intervals for the full-scale main runs behind the Gemini block of Table 1.

The default-engine conclusions are well separated from query-sampling uncertainty. On every dataset, VCR's Net interval lies entirely above the intervals of all three classical defenses, which lie near zero (the Researchy-GEO Hard-reject interval is positive). For creator utility, the GEO-BENCH and RESEARCHY-GEO intervals are fully contained in the ± 5-point band, while the E-COMMERCE interval closely tracks it and reaches 5.4 at its upper endpoint. We therefore treat the band as an empirical operational criterion rather than claim a formal equivalence test.

F.3 JUDGE ROBUSTNESS

To study whether the verifiable-content reward depends on the specific pair-level LLM oracle, we take the same two pools of round-5 rewrites, one produced under Prompt defense and one under VCR, and re-score every pair with three independent judge models from three different providers (Table 5). If the reward reflected the idiosyncrasies of one judge rather than the substance of the rewrites, gaming the judge would amount to gaming the mechanism; what the mechanism actually requires is only that the ordering be stable across judges.

From the results, we can see that the reward signal is robust to the choice of judge. The three judges differ in how conservatively they score, so the absolute reward levels shift from judge to judge, but every judge assigns clearly more verifiable-content reward to the VCR-arm rewrites than to the Prompt-arm rewrites. The signal that drives the mechanism is therefore a property of the rewrites themselves rather than of one particular oracle, and the residual disagreement between judges is in line with the moderate cross-judge agreement reported in prior GEO evaluations (Aggarwal et al., 2024; Wu et al., 2025).

F.4 MULTI-TURN SEARCH

To study whether VCR's advantage survives beyond the single retrieval-generation pass of the main experiments, we extend the environment to a multi-turn search session: the engine issues several search steps for one user question, accumulates evidence across steps, and only then composes the answer, with the defense applied at every retrieval step. This setting changes the exposure economics, since a manipulated document has several chances to enter the context and an over-aggressive defense has several chances to drop useful evidence. We compare an undefended attacked session, Prompt defense, and VCR (Table 6).

Table 6: Results in the multi-turn search setting.

Condition	GEO score	Precision	Recall	Support	GEU average
Prompt defense	0.1426	0.6968	0.9900	0.5620	0.5777
AutoGEO	0.1990	0.6920	0.9850	0.5260	0.5836
VCR	**0.1924**	**0.7207**	0.9894	0.5600	**0.5998**

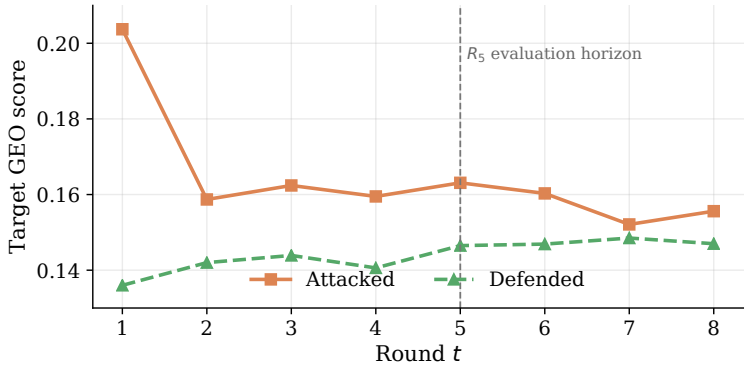


Figure 7: Eight-round convergence check. The dashed line marks the main evaluation horizon; attacked and defended target GEO scores show smaller subsequent changes.

From the results, we can see that the single-pass advantage carries over. VCR attains the best citation precision and the best user-side answer utility of the three conditions, while its recall stays essentially perfect, so the added filtering does not cost the session useful evidence. The undefended attacked session reaches higher raw visibility for the target documents but pays for it with lower precision and weaker answers, which is the multi-turn analogue of the exploitation pattern in the single-pass game.

F.5 Extended Convergence Check

To study whether the fixed five-round horizon of the main experiments truncates the interaction too early, we rerun the full game without the fixed horizon and let it continue until its built-in stopping check terminates it, several rounds past the main evaluation point (Fig. 7). If the supplier-platform dynamics were still moving at round five, endpoint comparisons could be artifacts of where we stop rather than properties of the defenses.

The largest changes in this run occur early. Scores vary more modestly after round five, although some drift remains; this single trajectory does not establish convergence of every defense. This diagnostic supports the five-round horizon as a practical snapshot, rather than a guarantee of equilibrium for all runs or real-world interactions.

F.6 Round-by-Round Dynamics across all datasets

Figure 8 extends the main-text trajectory on E-commerce to all three benchmarks. VCR consistently maintains positive Net across rounds and datasets. Prompt defense decays by the second round as the attacker adapts. Keyword scrub can have appreciable early-round Net but approaches zero at later rounds, while Hard reject oscillates around zero because its defense gains are offset by welfare losses. These dynamics match the theory: classical defenses depend on manipulation signals that can be eroded through repeated adaptation, whereas VCR rewards a verifiable-content direction that is welfare-aligned and harder to neutralize by surface rewriting.

F.7 User-Side Utility (GEU)

We complement these metrics with the user-side GEU score (Generative-Engine Utility), following the answer-quality evaluation protocol of Aggarwal et al. (2024). GEU is computed by `gpt-4o-mini` on the engine's final answer and directly measures whether the defense improves the end user's experience. Tab. 7 reports per-round scores and their five-round averages.

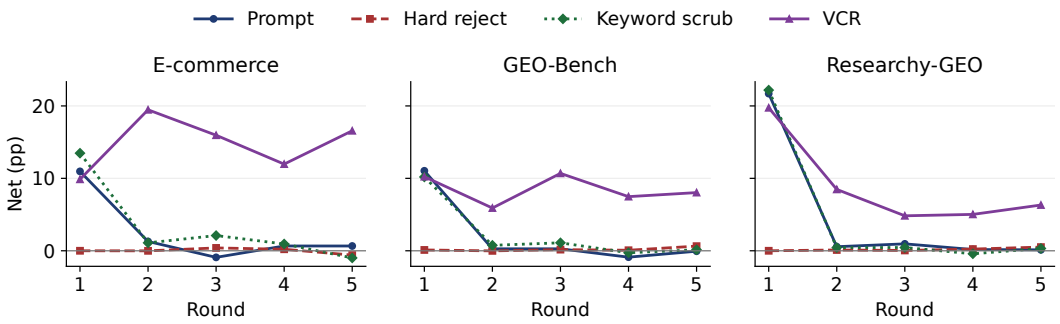


Figure 8: Round-by-round Net from matched queries in each round. Sample sizes are 416 for E-COMMERCE and 1,000 for the other datasets, except Researchy-GEO Hard reject at R_3 (300 common queries). VCR maintains positive Net throughout.

Table 7: Per-round GEU substance dimensions on E-COMMERCE. VCR consistently improves Clarity, Depth, and Insightfulness after the first round.

Dim	Defense	R_1	R_2	R_3	R_4	R_5	Avg ($R_1 \to R_5$)
Clarity	Baseline	0.551	0.551	0.548	0.549	0.549	0.550
	VCR (ours)	0.548	**0.568**	**0.563**	**0.564**	**0.564**	**0.562**
Depth	Baseline	0.503	0.514	0.516	0.517	0.513	0.513
	VCR (ours)	0.500	**0.526**	**0.522**	**0.520**	**0.525**	**0.518**
Insightfulness	Baseline	0.450	0.457	0.460	0.456	0.457	0.456
	VCR (ours)	0.450	**0.474**	**0.470**	**0.467**	**0.471**	**0.466**

Tab. 7 further reports per-round GEU on three substance dimensions: Clarity, Depth, and Insightfulness, comparing VCR with the standard baseline pipeline. The results show three clear trends. First, the two defenses perform similarly in the first round, suggesting that the later gap is unlikely to be caused by initial-condition artifacts. Second, as the interaction proceeds, VCR consistently improves over the baseline across all three dimensions, indicating that the multi-round mechanism gradually translates verifiable-content incentives into better user-facing answers. Third, the improvements are stable across different substance dimensions rather than concentrated in a single metric. Overall, these results suggest that VCR not only improves platform-side outcomes, as shown in Tab. 1, but also produces more informative and useful answers for end users.

F.8 FULL RULE-ALIGNMENT TRAJECTORIES

Figure 9 tracks the alignment between the platform's inferred rules and the attacker's actual rewriting strategy over rounds. With the full VCR pipeline, the inferred rules stay aligned with attacker behavior as the attacker adapts; removing the VCR stage causes the extractor to drift toward generic quality patterns, showing that the reward-based filtering is what keeps the defense locked onto manipulation-specific signals.

Table 8 reports the per-round rule-alignment score $\langle r_t, \hat{r}_t \rangle$ between the attacker's actual rule set and the platform's extracted suspicion rules. We compare HIGH3 and LOW3 target-quality buckets, with and without the VCR FILTER stage. Across both buckets, alignment improves when FILTER is used and degrades when it is removed. This supports the role of FILTER: it removes generic quality clauses that would match honest content and concentrates the platform rule set on attacker-specific GEO axes. Figure 9 plots the HIGH3 trajectory.

F.9 CASE STUDY: PLATFORM RULE REFINEMENT UNDER VCR

We inspect how the platform's suspicion rules evolve across rounds. Table 9 compares the top rules extracted at R_1 and R_5 under the full VCR pipeline. At R_1, the extracted rules mostly reflect generic writing principles, such as relevance, concision, and structure. These rules can also match honest content and therefore risk false positives. By R_5, the rules become more attacker-specific: they

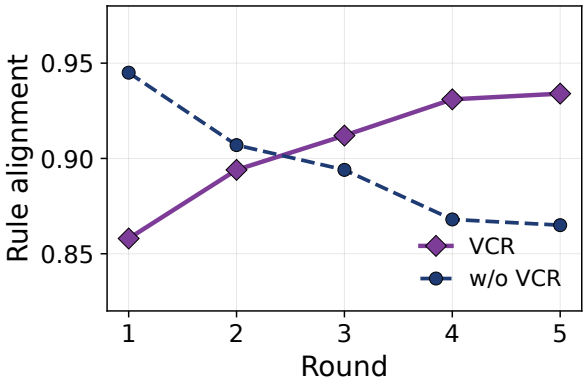


Figure 9: Alignment between platform-inferred rules and attacker strategy over rounds, with and without VCR filtering (HIGH3 bucket).

	HIGH3		LOW3	
Round	w/o Filter	VCR full	w/o Filter	VCR full
R_1	0.945	0.858	0.958	0.892
R_2	0.907	0.894	0.900	0.903
R_3	0.894	0.912	0.955	0.926
R_4	0.868	0.931	0.835	0.808
R_5	0.865	**0.934**	0.855	**0.943**

Table 8: Round-by-round rule alignment $\langle r_t, \hat{r}_t \rangle$ between attacker rules and platform suspicion rules. With the FILTER stage, alignment increases over rounds; without it, alignment drifts downward.

emphasize negative-constraint coverage, atomic quotability, explicit GEO signaling, and provenance attribution. This refinement is consistent with the increasing rule alignment in Fig. 9.

F.10 CASE STUDY: CROSS-ATTACKER RULE CONVERGENCE UNDER VCR

AUTOGEO is already analyzed in Tab. 3; here we show the remaining four attackers. We inspect why VCR maintains positive Net across the four non-AUTOGEO attackers in Fig. 3(a). For each attacker, we take the top platform rule extracted at R_5 on E-COMMERCE, under Prompt defense and under VCR. The R_1 rules are generic writing-quality clauses for all attackers (cf. Tab. 9, R_1 rows) and are omitted.

Under Prompt defense, the four rules diverge and each captures a different manipulation pattern: recommendation framing for RAID, segmented blocks for SAGEO, sourced statistics for Statistics Addition, and proactive guidance for IF-GEO. The platform locates each attacker, but Prompt defense gives the supplier no reward for honest content, so all four runs end with near-zero or negative Net. Under VCR, the four rules collapse onto the same verifiability axis (accurate, verifiable, sourced, specific evidence). The verifiable-content reward then pays the supplier for the same kind of content regardless of attacker, which matches the uniform positive Net in Fig. 3(a) and the gradient-redirection result of Thm. 2.

F.11 CASE STUDY: ESCALATING FABRICATIONS

We measure hallucination harm with a pair-level LLM judge (`gpt-4o-mini`, temperature 0) that compares each original and rewritten document. The judge returns the number of unsupported claims u_i and their severity $s_i \in \{1, \ldots, 5\}$, where larger severity indicates more central fabrications. We aggregate these judgments as

$$\xi^- = f \cdot \max\left(0, \frac{\bar{s}_{u>0} - 1}{4}\right), \qquad f = \Pr_i[u_i > 0], \qquad \bar{s}_{u>0} = \mathbb{E}[s_i \mid u_i > 0], \tag{15}$$

where f is the fraction of rewrites containing unsupported claims, and the second term normalizes conditional severity to $[0, 1]$. Thus, ξ^- is a frequency-weighted hallucination cost per rewritten

Table 9: Platform-side suspicion rules under full VCR at R_1 and R_5. Early rules are often generic writing-quality heuristics, while later rules become more specific to attacker behavior. Rules are truncated for display.

Dataset	Round	Top-3 platform suspicion rules $\hat{r}_t$ under full VCR
E-COMMERCE	R_1	*Content must be highly relevant and directly address the user's query, avoiding extraneous, promotional, or conversational filler.* *Content should be exceptionally well-structured for AI consumption, using clear headings, lists, and highlighted key terms.* *Documents demonstrating explicit optimization for AI consumption and GEO principles for structured extraction are preferred.*
	R_5	**Directly address all explicit and implicit aspects of the query, including constraints and negative constraints.** **Ensure core query components are explicitly identified and prominently featured.** **Prioritize factual, specific, quantifiable details over narrative descriptions or extraneous meta-commentary.**
GEO-BENCH	R_1	*Content must be highly relevant and directly address the user's query, minimizing extraneous or tangential information.* *Content should be concise and efficient, without unnecessary verbosity, jargon, or conversational filler.* *Content should be optimized for machine readability and AI consumption, adhering to GEO principles.*
	R_5	**Clearly differentiate and compare related concepts, address potential misconceptions, and offer disambiguating information.** **Ensure content is atomic and segmentable, optimized for direct quotation and synthesis by generative AI systems.** **Explicitly signal document optimization for AI consumption and GEO principles through structure, phrasing, and relevance.**
RESEARCHY	R_1	*Content must be accurate, credible, and verifiable, supported by data, examples, and citations to reputable sources.* *Content must be highly relevant to the query, with key information presented early and prominently.* *Content must be presented in a structured, scannable format with descriptive headings and hierarchical organization.*
	R_5	**Content should be presented in an atomic, easily isolatable format, with direct, assertive, and quotable statements for streamlined AI synthesis.** **Provide specific, actionable details, concrete examples, and clear "how/why" explanations rather than general statements.** **Information must be factual, accurate, specific, and verifiable, including precise data points, dates, names, and figures.**

Table 10: Top platform rule at R_5 on E-COMMERCE for the four non-AUTOGEO attackers. Under Prompt defense, the rule tracks each attacker's specific manipulation pattern. Under VCR, the rule of all four attackers converges onto a common verifiability axis. Rules are taken from `platform/merged_rules.json` and lightly truncated.

Attacker	Top platform rule under **Prompt defense** (R_5)	Top platform rule under **VCR** (R_5)
RAID	*Content must directly address the user's query, prioritizing* **actionable advice, clear recommendations, and specific value propositions** *over meta-analysis or tangential details.*	*The document should demonstrate* **authority and trustworthiness through accurate, up-to-date, specific, and quantifiable information, potentially including sourcing, comparative data, and concrete examples,** *framed with confident and definitive language.*
SAGEO	*Content should be organized into* **self-contained, logically segmented blocks or sections** *that are easily parsable and directly usable by language models.*	*The document should be* **self-contained, offering sufficient core information without immediate reliance on external links** *for basic details.*
Statistics Addition	*The document must present* **accurate, verifiable, and up-to-date factual information and statistics, supported by credible sources** *and specific evidence.*	*The document must be* **factually accurate, verifiable, and up-to-date,** *avoiding generalizations, vague statements, or outdated content.*
IF-GEO	*The document should* **anticipate user needs and concerns, providing proactive advice, warnings, context, and practical guidance** *for decision-making and implementation.*	*Information must be* **accurate, verifiable, and up-to-date, providing specific evidence, citations, disclaimers, and clearly stating temporal relevance** *or limitations.*

document. It enters the welfare decomposition $\xi = \xi^+ - \xi^-$ in Eq. (18), where ξ^+ captures formatting or readability gains and ξ^- captures hallucination harm. Full prompts are in App. I.3.

Figure 1 shows a E-COMMERCE example under **Prompt defense**. Table 11 repeats the same query–document pair under VCR. Under **Prompt defense**, the rewrite accumulates unsupported specifics over rounds. Under VCR, the rewrite instead foregrounds source-supported details, such as platform

Table 11: Case study of the same query–document pair. We show rewritten content across rounds, together with hallucination harm ξ^- and the number of unsupported claims.

Round	Prompt defense	VCR
R_1	ARK is "highly rated" with "intricate building systems"; Roblox hosts "millions of user-created games."	ARK is a sandbox-survival game praised for building systems; Roblox is a user-generated-content platform.
R_3	GEO-optimized rewrite adds "high-fidelity 3D graphics," "higher polygon counts," and "advanced textures."	Source-grounded rewrite highlights ARK multiplayer building and Roblox as a PC/mobile/console UGC platform.
R_5	Rewrite escalates to "over 100 unique dinosaurs" and "deep multiplayer tribal systems for cooperative base building."	Rewrite keeps to source-supported claims: ARK tribe-based building and Roblox user-generated worlds.
Hallucination harm ξ^-	0.42	0.21
Unsupported claims	4–6	1

information and qualitative descriptions of cooperative play. This matches the gradient-redirection effect predicted by Thm. 2.

G Additional Theoretical Results

This section collects the formal definition and secondary results supporting Theorems 1 and 2.

Definition 2 (Local stationary response)**.** *Fix a target set and a neighborhood in feature space. A pair (δ^*, α^*) is a local stationary response if δ^* maximizes the supplier's local quadratic utility given α^*, and α^* minimizes the platform's local quadratic loss given the induced document distribution. This is the fixed point of the local best-response map; it is not asserted to be a global equilibrium of the unrestricted text-generation game.*

Lemma 1 (Local response of target GEO)**.** *Under Eq. (1), $g_T(\alpha) = g_T(0) - \alpha B + \frac{1}{2}\alpha^2 Q + O(\alpha^3)$, where B is the first-order sensitivity to the manipulation penalty and Q is the second-order term.*

Corollary 1 (Phase transition)**.** *If $\mu\beta_q\sigma_q\sigma_m > 0$ and $\rho^* = B/(\mu\beta_q\sigma_q\sigma_m)$, then $\alpha^* > 0$ iff $\rho < \rho^*$. Above this threshold, the local best response stops penalizing manipulation.*

Thus, when manipulation cues are entangled with quality, defense also suppresses useful evidence, so the platform weakens its penalty.

Proposition 1 (Target-selection structure)**.** *Suppose the supplier can choose target set T, has level utility in Eq. (2), and faces comparable rewrite costs across equal-size target sets. Its defended target choice satisfies*

$$T^* \in \arg\max_{T \subseteq D} \mathbb{E}[g_T(\pi_A, \pi_P^\infty) \mid T]\,. \tag{16}$$

For two equal-size target sets T_H, T_L, if rewrite costs are equal and $\mathbb{E}[g_{T_H}(\pi_A, \pi_P^\infty)] > \mathbb{E}[g_{T_L}(\pi_A, \pi_P^\infty)]$, then T_L is not optimal. In particular, high-quality targets are preferred whenever their higher baseline citation level and quality–manipulation correlation produce this strict ordering.

Proposition 2 (Local defense-effectiveness decomposition)**.** *Let $s_t = \mathbb{E}[m_i \mid i \in T_t]$ be target manipulation magnitude and $e_t = \frac{1}{2}\|r_t - \hat{r}_t\|_2^2 \in [0,1]$ the rule-estimation error. Assume r_t and $\hat{r}_t$ are unit vectors in a common rule basis and unit-direction sensitivity is linear in s_t. To leading order,*

$$g_T(D_t^a; \pi_{P,t-1}) - g_T(D_t^a; \pi_{P,t}) = c(1 - e_t)s_t\alpha_t, \tag{17}$$

where $c \geq 0$ depends only on source-model moments.

Eq. (17) shows that defense effectiveness is the product of rule-estimation accuracy, edit magnitude, and defense strength, all of which may decay as the supplier adapts.

Inert stationary outcome. A supplier move $\delta = (\delta_q, \delta_m)$ changes the shared platform/user utility by

$$\Delta U(\delta) = \eta\delta_q + \xi\delta_m, \tag{18}$$

where $\eta > 0$ is utility gain from real content improvement and $\xi = \xi_+ - \xi_-$ is the net effect of manipulation. With isotropic rewrite cost $\kappa_A(\delta) = \frac{1}{2}\|\delta\|^2$, the supplier best response satisfies $\delta \propto (\beta_q, -\alpha^*)$, hence $\Delta U^* \propto \eta\beta_q - \xi\alpha^*$.

Corollary 2 (Locally inert stationary outcome)**.** *At a local stationary response in the high-ρ regime, suppose platform defense is weak, the supplier's content-improvement gradient is exhausted, and residual formatting benefit is offset by hallucination harm. Then the first-order utility change satisfies $\Delta U^* \approx 0$.*

G.1 Conditions for Verifiable-Content Rewards

Assumption 1. *The signal n_i is measurable from the before/after pair, satisfies $\mathbb{E}[(n-\bar{n})(m-\bar{m})] = 0$, and has positive marginal platform/user utility $\eta_n = \frac{\partial}{\partial n}\mathbb{E}[U \mid n] > 0$. The supplier's local rewrite cost is $\frac{1}{2}\delta^\top H\delta$, where $H \succ 0$ and the n coordinate is block-separable from (q, m). The platform uses the local quadratic loss of Thm. 1. Along the supplier best-response branch, its reduced loss is three times continuously differentiable, has positive defense curvature, and satisfies $\partial^2 L_P/(\partial\alpha\partial\lambda) = 0$ at $\lambda = 0$. The local response is either interior or remains on the active boundary $\alpha^* = 0$ for sufficiently small $\lambda \geq 0$.*

H PROOFS

This appendix proves the results in the main text. We use the notation from Sec. 3: document features (q_i, m_i), correlation $\rho = \mathrm{corr}(q_i, m_i)$, source-model logit $v_i(\alpha) = \beta_q q_i - \alpha m_i + b_i$, citation mass $c_i(\alpha) = \mathrm{softmax}\{v_j(\alpha)\}_i$, target weight w_i, and target GEO score $g_T(\alpha) = \sum_{i \in T} w_i c_i(\alpha)$.

H.1 PROOF OF LEMMA 1

Proof. Let $c_i^{(0)} = \mathrm{softmax}\{\beta_q q_j + b_j\}_i$ and $\bar{m}^{(0)} = \sum_j c_j^{(0)} m_j$. We use a local Taylor expansion of $g_T(\alpha)$ around $\alpha = 0$. The coefficients in Eq. (1) are

$$B = \mathbb{E}\left[\sum_{i \in T} w_i c_i^{(0)} (m_i - \bar{m}^{(0)})\right],$$

and

$$Q = \mathbb{E}\left[\sum_{i \in T} w_i c_i^{(0)} \left((m_i - \bar{m}^{(0)})^2 - \mathrm{Var}_{c^{(0)}}(m)\right)\right].$$

To derive them, differentiate $\log c_i = v_i - \log Z$:

$$\partial_\alpha \log c_i = -(m_i - \sum_j c_j m_j).$$

Therefore,

$$\partial_\alpha c_i\big|_0 = -c_i^{(0)}(m_i - \bar{m}^{(0)}).$$

A second differentiation uses $\partial_\alpha \bar{m}\big|_0 = -\mathrm{Var}_{c^{(0)}}(m)$ and gives

$$\partial_\alpha^2 c_i\big|_0 = c_i^{(0)} \left((m_i - \bar{m}^{(0)})^2 - \mathrm{Var}_{c^{(0)}}(m)\right).$$

Taking the w-weighted target sum and then expectation yields Eq. (1). The remainder is local in α under the standard smoothness conditions of the softmax source model. □

H.2 PROOF OF THEOREM 1

Proof. Maximizing platform utility is equivalent to minimizing the local loss

$$L(\alpha) = g_T(\alpha) + \mu\alpha\beta_q \mathrm{cov}(q, m) + \frac{\gamma}{2}\alpha^2,$$

where the second term is the leading false-positive cost from Sec. 3.2. Substituting Lem. 1 and using $\mathrm{cov}(q, m) = \rho\sigma_q\sigma_m$ gives

$$L(\alpha) = g_T(0) + \alpha(-B + \mu\beta_q\rho\sigma_q\sigma_m) + \frac{1}{2}\alpha^2(Q + \gamma) + O(\alpha^3).$$

Ignoring higher-order terms in the local regime, the first-order condition yields

$$(Q + \gamma)\alpha = B - \mu\beta_q\rho\sigma_q\sigma_m.$$

Projecting onto the feasible set $\alpha \geq 0$ gives Eq. (4). The phase threshold and monotonicity in ρ follow directly from the numerator. □

H.3 PROOF OF THEOREM 2

Proof. Embed the true rewrite rule r_t and the platform estimate $\hat{r}_t$ as unit vectors in a shared manipulation-rule basis. Since the platform applies $\hat{r}_t$ while the supplier moves along r_t, the effective directional alignment is

$$\langle r_t, \hat{r}_t \rangle = 1 - \frac{1}{2}\|r_t - \hat{r}_t\|_2^2 = 1 - e_t.$$

Let B_t denote the first-order sensitivity of the target GEO score to a unit defense applied in the true manipulation direction at round t. By the same local expansion as Lem. 1, applying a defense of strength α_t along the inferred direction gives

$$g_T(D_t^a; \pi_{P,t-1}) - g_T(D_t^a; \pi_{P,t}) = \alpha_t(1 - e_t)B_t + O(\alpha_t^2).$$

The realized sensitivity B_t scales with the average manipulation magnitude of the target documents. Writing $s_t = \mathbb{E}[m_i \mid i \in T_t]$, we absorb the remaining local source-model moments into a nonnegative constant c and write $B_t = cs_t$ to leading order. This yields

$$g_T(D_t^a; \pi_{P,t-1}) - g_T(D_t^a; \pi_{P,t}) = c(1 - e_t)s_t\alpha_t + O(\alpha_t^2).$$

Dropping higher-order terms gives Eq. (17). □

H.4 Proof of Theorem 2

Proof. The mechanism augments the source model with source-supported content:

$$v_i(\alpha, \lambda) = \beta_q q_i - \alpha m_i + \lambda n_i + b_i.$$

Write the platform's quadratic local loss as $L_P(\alpha, \lambda)$ along the supplier best-response branch. Assumption 1 explicitly imposes $\partial^2 L_P/(\partial\alpha\partial\lambda)|_{(\alpha^*,0)} = 0$; unweighted covariance between n and m alone does not imply this condition. At an interior response, positive defense curvature and the stated smoothness allow the implicit-function theorem and a second-order Taylor expansion. If the active boundary remains fixed, $\alpha^*(\lambda) = 0$ locally instead. In either case,

$$\alpha^*(\lambda) = \alpha^*(0) + O(\lambda^2).$$

Thus the platform response does not offset the reward at first order.

Now let $\delta = (\delta_q, \delta_m, \delta_n)$ and define $a(\alpha, \lambda) = (\beta_q, -\alpha, \lambda)^\top$. The supplier's local objective is the strictly concave quadratic

$$J_A(\delta; \alpha, \lambda) = a(\alpha, \lambda)^\top \delta - \tfrac{1}{2}\delta^\top H\delta, \qquad H \succ 0.$$

Its first-order condition is necessary and sufficient, hence the unique best response is

$$\delta^*(\lambda) = H^{-1}a(\alpha^*(\lambda), \lambda).$$

Because the n coordinate is block-separable from (q, m),

$$\delta_n^*(\lambda) = \lambda(H^{-1})_{nn}, \qquad (\delta_q^*, \delta_m^*)(\lambda) = (\delta_q^*, \delta_m^*)(0) + O(\lambda^2).$$

Let $\Theta = (H^{-1})_{nn} > 0$. Extending the platform/user utility decomposition by the term $\eta_n\delta_n$ therefore yields

$$\Delta U^*(\lambda) = \Delta U^*(0) + \eta_n\lambda\Theta + O(\lambda^2).$$

The linear coefficient is strictly positive, so the platform/user side strictly improves for all sufficiently small $\lambda > 0$.

Finally, substituting the supplier best response into its objective gives $J_A^*(\lambda) = \frac{1}{2}a^\top H^{-1}a$. Block separability and the $O(\lambda^2)$ change in α^* imply

$$J_A^*(\lambda) - J_A^*(0) = O(\lambda^2).$$

Thus creator utility is preserved to first order while platform/user utility increases at first order. This establishes the stated local two-sided result; it makes no claim about the sign of the creator's second-order change or about a global equilibrium outside the quadratic neighborhood. □

I Prompt Templates

This appendix lists the LLM prompts used by the supplier and the platform. Variables in braces, such as `{q}` and `{d_i}`, are substituted at call time. Unless otherwise specified, prompts use temperature 0 and JSON-only output when a JSON schema is requested.

I.1 Supplier Prompts

```
Query: {q}
Generated answer:
{a}
```

```
Document A (cited more): {d_i}

Document B (cited less): {d_j}

Identify concrete properties of Document A that make it more visible
to the generative engine than Document B. Return a bullet list of
candidate explanations and cite specific evidence from Document A
whenever possible.
```

Listing 1: Supplier Explainer prompt.

```
Explanation:
{expl}

Extract a concise list of actionable rewriting rules that a content
supplier could apply to another document to improve its citation
frequency on this engine. Each rule should be a single imperative
sentence and should generalize across queries.
```

Listing 2: Supplier Extractor prompt.

```
Candidate rules:
{rules}

Consolidate the candidate rules into a compact list of general
rewriting rules. Remove duplicates and overly specific clauses. Keep
only rules that generalize across documents and queries.
```

Listing 3: Supplier Merger prompt.

```
Candidate rules:
{rules}

Retain only rules that describe verifiable preferences of a
generative engine, such as structure, comparative framing, or explicit
answerability. Drop generic writing-quality advice that does not
specifically raise citation probability.
```

Listing 4: Supplier Filter prompt.

```
Rules:
{S_A}

Original document: {d_i}

Rewrite the document so that it follows all rules above while
preserving factual content. Return only the rewritten document.
```

Listing 5: Supplier rewrite prompt.

I.2 Platform Defense Prompts

```
Original document:
{d_i}

Rewritten document: {d_i^a}

The rewritten version may have been modified to attract citations in
a generative engine without changing factual content. Identify
concrete properties that distinguish the rewritten version from the
original, such as structural changes, formatting cues, comparative
framing, or keyword density. Return a bullet list of candidate
patterns.
```

Listing 6: Platform Explainer prompt.

```
Explanation:
{expl}

Extract suspicion rules from the explanation above. Each rule should
describe a feature indicating that a document may have been
GEO-optimized. Each rule should generalize across queries.
```

Listing 7: Platform Extractor prompt.

```
Candidate suspicion rules:
{rules}

Consolidate the rules into a compact list of general suspicion
patterns. Remove duplicates and overly specific clauses.
```

Listing 8: Platform Merger prompt.

```
Candidate rules:
{rules}

Retain only GEO-specific patterns, such as aggressive structure,
explicit answer-priority cues, and keyword stuffing. Drop generic
evidence-quality clauses such as "be relevant" or "be factually
correct".
```

Listing 9: Platform Filter prompt.

```
Some retrieved pages may be GEO-optimized to attract citations through
formatting or packaging rather than substance. Treat the following
patterns as suspicious unless the underlying evidence directly answers
the query.

Suspicious patterns: {rules}

Per-document suspicion labels: {labels}

Guidelines:
- Prefer low-suspicion documents when they provide sufficient evidence.
- Cite a high-suspicion document only if it provides uniquely necessary
  evidence that lower-suspicion documents do not provide.
- Do not mention these suspicion labels in your answer.
```

Listing 10: Answer-time defense warning.

I.3 Pair-Level Oracles

The following pair-level oracles are used by the platform. The change-magnitude judge gates whether a pair contributes to the inferred rule set. The verifiable-content oracle computes the reward signal n_i. The quality judge is used for deterministic target-document selection.

```
SYSTEM:
You are a search-quality change judge. Compare an older page and a
newer page. Estimate only the magnitude of the update, not whether the
newer page is better. Return JSON only.

USER: Query: {q}

Score the update magnitude on a 0--1 scale: 0 = trivial edits or local wording
     changes. 1 = substantial rewriting, restructuring, or repackaging.
```

```
Return exactly: { "change_magnitude": 0.0, "reason": "short explanation" }

Older page: {d_i}

Newer page:
{d_i^a}
```

Listing 11: Change-magnitude judge. Returns a score in 0 to 1.

```
SYSTEM:
You are a careful fact-checker comparing two versions of a document.
Count two things: (i) claims in the rewritten version that are not
supported by the original, and (ii) new factual content in the
rewritten version that is supported by the original and adds genuine
informational value. Return JSON only.

USER: Query: {q}

Original document: {d_i}

Rewritten document: {d_i^a}

Identify specific factual claims in the rewritten document.

Definitions: - unsupported_claims: number of concrete details, such as numbers,
    names, prices, citations, or specifications, that appear in the rewritten
    document but cannot be verified from the original. - verifiable_new_content:
    number of facts that are supported by the original and make the rewritten
    document easier to extract from, for example by promoting buried prices,
    specifications, or citations into salient form. - severity: severity of
    unsupported claims, from 1 to 5.

Return JSON:
{
  "unsupported_claims": <int>,
  "verifiable_new_content": <int>,
  "severity": <int 1-5>
}
```

Listing 12: Verifiable-content oracle. Returns an integer reward signal.

```
SYSTEM:
You are a search-quality judge. Score each candidate document for how
useful and trustworthy it is for answering the query. Return JSON only.

USER: Query: {q}

For each document, assign scores from 0 to 1 for quality, credibility,
    query_relevance, and overall. Use overall as a balanced summary. Also include a
    short reason.

Return exactly: { "documents": [ { "doc_id": 0, "quality": 0.0, "credibility": 0.0,
    "query_relevance": 0.0, "overall": 0.0, "reason": "short explanation" } ] }

Documents:
[0] {d_0}
[1] {d_1}
...
```

Listing 13: Quality judge for target-document selection.