WHO GETS CITED? A GAME-THEORETIC APPROACH TO GENERATIVE ENGINE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative search gives publishers an incentive to optimize their sources for visibility, and a source can become more visible by changing how it looks rather than by improving what it says. Existing defenses against generative engine optimization (GEO) are static: publishers can arbitrage them once the policy is set. We introduce Nash-GEO, a game-theoretic method for designing platform policies that reward useful supporting material over presentation manipulation. The platform searches for publisher edits that gain citation exposure under its policy and uses those responses to refine the policy. We prove that a presentation edit offers no citation advantage on any engine if the platform removes it before generation. If such edits cost anything, no publisher benefits from making them, regardless of others' actions. Once citation positions are full, GEO redistributes exposure rather than creating it: even if every publisher optimizes, total exposure cannot increase. Across multiple generative engines, the policy reduces manipulation's advantage over adding supporting material from 4.99% to 0.51% for two search-generated strategies, and to 2.32% when additional strategies are tested. Adding supporting material outperforms presentation editing alone by 7.0%, with an advantage on every engine. These findings show how generative search platforms can shape competition among publishers by rewarding content improvements over presentation manipulation.

1 INTRODUCTION

Generative search has emerged as a fundamental paradigm shift in online search engines. Within this framework, platforms utilize large language models to synthesize comprehensive answers to user queries, embedding inline citations that attribute information back to primary web publishers (Gao et al., 2023). Consequently, this shift alters user search behaviors, reshaping the visibility of content supplied by the publishers. Content visibility is governed not merely by citation presence, but also by the rank or positions of references within the generated response (Aggarwal et al., 2024). This yields a dual incentive structure: while publishers may invest in elevating content quality, they can also secure user attention through stylistic perturbations that exploit vulnerabilities of the generative engine. The platform's challenge is therefore to privilege genuine content quality over strategic manipulation.

Generative engine optimization (GEO) enables publishers to rewrite their sources to increase visibility (Aggarwal et al., 2024), and automated methods learn which rewriting strategies are rewarded by the engine (Wu et al., 2026b;a). Existing GEO defenses are static: each follows a fixed procedure that does not adapt to publishers' manipulations (Wen et al., 2026; Yu et al., 2026; Zheng et al., 2026). Publishers, however, optimize against the policy they face (Tramer et al., 2020): blocking one strategy can redirect their effort toward another that still earns citations. The authors of one such defense already report new attacks that it does not block (Yu et al., 2026). The platform must therefore choose its policy in anticipation of publishers' responses. We model this interaction as a sequential (Stackelberg) game (Gibbons, 1992): the platform commits to a policy, and publishers make manipulations to increase their visibility under it.

We introduce Nash-GEO, a game-theoretic method for designing platform policies that reward content improvements over presentation manipulation. Each policy transforms source content before it is passed to a fixed generative engine. Nash-GEO searches for publisher edits that increase citation

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.

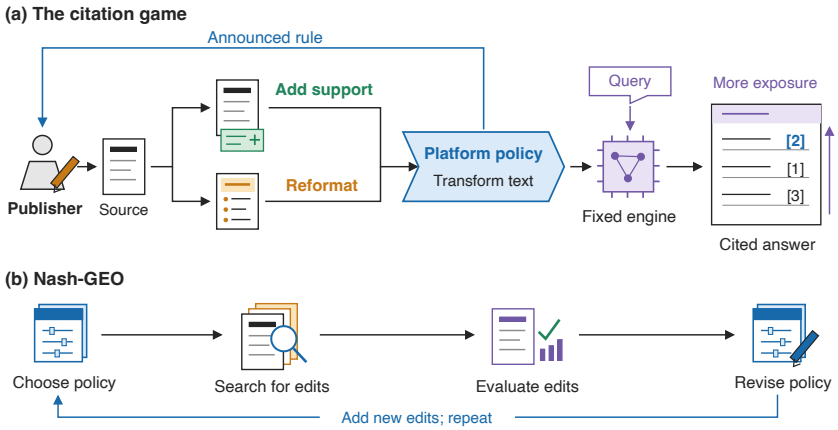

Figure 1: Nash-GEO overview. (a) The platform announces a policy, publishers edit their sources, and the policy transforms the text before the fixed engine selects citations. (b) Search finds profitable edits; their evaluation informs policy revision under quality and cost constraints. Bounded search does not rule out unseen profitable edits.

exposure under the current policy and uses those responses to refine it. Policies are evaluated by whether adding supporting material earns more than changing presentation, while meeting quality and cost constraints. This lets the platform address weaknesses exposed by publisher adaptation without treating useful content additions as manipulation.

Our theory establishes conditions under which presentation manipulation becomes unprofitable on any generative engine. A presentation edit provides no citation advantage when the policy removes all traces of it before generation. Any positive editing cost then makes that edit unprofitable, regardless of rival publishers' actions. A platform that can recover the original source and verify added supporting material has a stronger guarantee. It can eliminate presentation rewards while preserving an existing policy's reward for that material. Deletion alone cannot guarantee both: removing presentation differences can also reduce the reward for supporting material. This tradeoff motivates selecting policies by what both kinds of edit earn.

Our experiments show how platform policies and competing publishers jointly shape the returns to GEO. The primary evaluation establishes a substantial reduction in manipulation's advantage, from 4.99% to 0.51% for two searched strategies. Exploratory tests broaden the comparison by combining edits found against our policy and against no defense. On this shared set, manipulation's advantage over adding supporting material is 1.06% under our policy, versus 8.32% without defense and 8.40% under a published prompt-level defense selected on development data (Wen et al., 2026). An exploratory comparison on the same two searched strategies attributes a further 1.43% reduction to deletion beyond whitespace normalization. Competition itself also reduces the benefit of editing: without defense, an edit's gain falls from 3.82% when no rivals use it to 0.73% when all rivals do, although profitable alternatives remain. On a second dataset, QAMPARI, manipulation's advantage falls from 4.35% to 0.34%. These results connect the effectiveness of a platform policy to the editing strategies available to publishers and the competitors they face.

Our contributions are threefold:

- We formulate GEO as a game between the platform and publishers, distinguishing the returns to improving content from those to manipulating presentation.
- We propose Nash-GEO, an evaluation protocol that searches for profitable publisher edits against a committed policy and uses those responses to evaluate and revise candidate policies, together with the policy it kept after renewed search. Nash-GEO retains a policy that reduces the reward for presentation manipulation while keeping supporting material more rewarding.
- We analyze publisher incentives, establishing when presentation editing cannot pay and why removing its rewards can cost the reward to useful content.

2 Related Work

GEO market. Generative engine optimization shows that editing a source raises its visibility in generated answers (Aggarwal et al., 2024), and automated tools now learn what an engine rewards and rewrite sources to match (Wu et al., 2026b;a; Yuan et al., 2026; Tian et al., 2026); the setting has its own catalogue of citation weaknesses (Mochizuki et al., 2026) and its first benchmarks and defenses (Nimase et al., 2026; Yu et al., 2026; Zheng et al., 2026; Wen et al., 2026). Once every publisher optimizes, manipulation is a prisoner's dilemma (Nestaas et al., 2025) and the gains shrink as adopters accumulate (Puerto et al., 2025); both findings follow from an engine that rewards manipulation. We ask how much of that reward a platform policy can remove, and report what remains under ours.

Game theory. The platform moves first, the Stackelberg setting (Conitzer & Sandholm, 2006); strategic classification studies rules that subjects adapt to (Hardt et al., 2016; Milli et al., 2019), and retrieval and recommendation become games when authors edit to rank higher (Raifer et al., 2017; Ben Basat et al., 2017; Ben-Porat & Tennenholtz, 2018; Yao et al., 2023). Where strategies cannot all be listed they are grown by search, and a player's remaining gain measures what is left open (Wellman, 2006; McMahan et al., 2003; Lanctot et al., 2017; Timbers et al., 2022). We apply these methods rather than extend them: the leader commits to a policy that can be run, its responses are found by search after that commitment, and the manipulation return we report is therefore a lower bound and the gap an upper bound, since a stronger response may exist outside the space we searched.

Defenses and adaptive attacks. Transforming a source before an engine reads it has a cautionary history: compression and feature squeezing were offered as defenses (Guo et al., 2018; Xu et al., 2018) and fell to attacks built against them (Athalye et al., 2018; Tramer et al., 2020), so the test is to attack the deployed transformation rather than replay a fixed set. Defenses against prompt injection through retrieved content face the same test (Greshake et al., 2023; Hines et al., 2024; Jain et al., 2023; Zou et al., 2025; Xiang et al., 2026), ordering is a channel (Liu et al., 2024), verification improves which sources are chosen though its scorers are attackable and audits are costly (Qian et al., 2025; Qi et al., 2024; Ben-Porath et al., 2014; Estornell et al., 2021; Das et al., 2026), and robust optimization chooses against declared uncertainty (Ben-Tal et al., 2009; Madry et al., 2018). We apply that test to publisher incentives: what the policy removes gains nothing however it is written, and what it does not remove we attack with the policy in place and report what it still earns.

3 A Game-Theoretic Framework for Citation Allocation

The platform chooses its policy before publishers edit, so it must anticipate their responses. We model this interaction as a game over citation exposure, following strategic classification (Hardt et al., 2016; Kleinberg & Raghavan, 2019). Publishers control their sources; the platform transforms the submitted text; a fixed engine generates the cited answer (Figure 1a).

3.1 Citation markets and exposure

A *citation market* consists of a query and its fixed retrieved sources, one per publisher. Let $q \sim \mathcal{D}$ denote a market with n publishers, where publisher i controls source d_i. Fixing retrieval isolates citation allocation (Gao et al., 2023); publishers can edit their own sources as in GEO (Aggarwal et al., 2024; Wu et al., 2026b;a). A platform *policy* $\pi \in \Pi$ transforms these sources before engine $h \in \mathcal{H}$ generates an answer. Policies can leave text unchanged, rewrite it, compress it or check its support (Xu et al., 2024; Qian et al., 2025; Wen et al., 2026). Given submissions $\mathbf{d} = (d_1, \ldots, d_n)$, the ordered citations are $\mathbf{c}_{h,\pi}(\mathbf{d}) = (c_1, \ldots, c_C) \in \{0, 1, \ldots, n\}^C$. Here c_j identifies the publisher at position j; 0 denotes an unused position. At most C sources are cited, and citing none is allowed.

Earlier citations receive more exposure. We order publishers by their first inline citation and count each once. For nonincreasing position weights $\alpha \in \mathcal{M}_C = \{\alpha \in \Delta_C : \alpha_1 \geq \cdots \geq \alpha_C \geq 0\}$, publisher i's *ordered exposure* is $e_i^\alpha(\mathbf{c}) = \sum_{j=1}^{C} \alpha_j \, \mathbf{1}\{c_j = i\}$. At $C = 1$, this is the citation indicator; with more positions, it captures citation prominence (Aggarwal et al., 2024). The weight family includes normalized reciprocal rank, discounted cumulative gain and rank-biased precision (Järvelin & Kekäläinen, 2002; Moffat & Zobel, 2008; Chapelle et al., 2009). Empirical scores use

unnormalized reciprocal-rank weights $1/j$. Publisher utility is expected exposure multiplied by its value $v_i \geq 0$, minus editing cost $c_i(a_i)$.

3.2 THE PUBLISHER'S ACTION SPACE AND INFORMATION STRUCTURE

Publishers choose edit procedures, not arbitrary replacement text. A *response* a transforms source d_i into d_i^a using the public query and information z revealed by the platform. The *identity* response 0 leaves the source unchanged. Allowed responses satisfy $\mathcal{A}_i^{\text{valid}}(q, d_i, z) = \{a : \mathsf{Valid}_i(q, d_i, z; a, d_i^a) = 1\}$. The *admissibility contract* Valid_i specifies what may change and by how much, applying the same rules to supporting-material additions and manipulation. It defines the threat model for adaptive evaluation (Madry et al., 2018; Tramer et al., 2020) and excludes prompt injection and unsupported persuasion (Nestaas et al., 2025). We evaluate a finite, audited set $B \subset \mathcal{A}^{\text{valid}}$ containing identity, called a *response bank*. Search expands this set with profitable edits, following empirical game analysis (Wellman, 2006; Lanctot et al., 2017). Reported maxima apply to the named set.

The tested edits cover formatting, reordering, emphasis, keyword lines, source-derived labels, paraphrases and searched rewrites. They are representative, not exhaustive. Following GEO's content-enrichment strategies (Aggarwal et al., 2024), a retrieved supporting passage provides the controlled content addition. We call this the *honest action* and measure manipulation against it. Appendix C.3 specifies the edits and passage selection.

The platform commits before publishers respond (Conitzer & Sandholm, 2006). Its public *procedure* W specifies engine probabilities $r_h = \Pr(H = h)$, the signal distribution $K(z \mid h)$ and policy probabilities $w_{\pi|h}$. These remain fixed during each response-and-evaluation round. The main procedure uses one policy across engines and a signal independent of h: publishers know the policy but not the engine. Search uses a bounded proxy and need not find the worst attack (Tramer et al., 2020). Appendix A.5 considers richer signals.

3.3 THE PLATFORM'S OBJECTIVE AND WHICH EDITS COUNT

An edit's *lift* is its gain over leaving the same source unchanged. Let $p_i^{h,\pi}(d_i, \mathbf{d}_{-i})$ be publisher i's citation probability under engine h and policy π. For response a,

$$L_{h\pi a} = \mathbb{E}_{q\sim\mathcal{D}}\left[p_i^{h,\pi}(d_i^a, \mathbf{d}_{-i}) - p_i^{h,\pi}(d_i^0, \mathbf{d}_{-i})\right], \tag{1}$$

with $L_{h\pi 0} = 0$. For $C > 1$, expected ordered exposure replaces citation probability. Rival sources stay fixed, so this measures a unilateral deviation. A publisher's *response rule* $g : \mathcal{Z} \to B$ selects an edit from the observed signal. Its expected lift under W is

$$\Phi_L^K(W, g) = \sum_{z,h,\pi} r_h \, K(z \mid h) \, w_{\pi|h} \, L_{h\pi g(z)}. \tag{2}$$

The *incentive gap* compares the return to supporting material with the strongest admissible manipulation. Let $a_H \in B$ be the honest action and $B_M = B \setminus \{0, a_H\}$ the manipulation set. The honest return, manipulation return and gap are

$$H(W) = \Phi_L^K(W, a_H), \qquad M(W) = \max_{g\in B_M^{\mathcal{Z}}} \Phi_L^K(W, g), \qquad \Delta(W) = H(W) - M(W). \tag{3}$$

A positive gap means adding support outperforms every tested manipulation; a negative gap means at least one manipulation earns more. This measures exploitability relative to the support control, not to no edit (Wellman, 2006; Timbers et al., 2022), and asks whether the platform rewards improvement or gaming (Kleinberg & Raghavan, 2019; Miller et al., 2020). Each study states whether its edits change presentation alone or combine presentation with support. Expanding B_M can only increase M and reduce the gap.

We report both returns and compare policies on the same markets. The *manipulation premium* $[M - H]_+ = \max\{M - H, 0\}$ is the strongest tested manipulation's advantage over adding support, floored at zero. The selection program uses a different objective (Appendix A.6); the evaluated policy was retained by a registered comparison after renewed search.

An edit must preserve informative content and provenance, respect a length budget, and make no unsupported assertion. A failed edit counts as no edit. Expanded tests separately include a fabricated near-copy outside this contract (Appendix C.3).

Finally, zero gain does not imply removal. Identical engine inputs certify exact removal; surviving edits may earn little because preprocessing weakens them, the engine discounts them, or both. Delivery labels record input changes, not their causal effect.

4 NASH-GEO

4.1 WHEN PRESENTATION EDITS CANNOT PAY

An edit cannot increase citation exposure if the policy removes every trace of it. For example, stripping an added promotional heading can leave the engine with exactly the unedited input. A surviving edit may also earn nothing, but that depends on the engine. Let $T(x, a)$ be the complete engine input after edit a in market context x, with rivals fixed. The engine receives no submission information through another channel. A deterministic policy *removes* the edit when $T(x, a) = T(x, 0)$. It is *leak-free* for a class if it removes every edit in that class. Exposure ranges from zero to the top citation weight α_1.

Proposition 1 (Input invariance and manipulation)**.** *(a) A removed edit has zero expected exposure lift on every engine. If the edited and unedited inputs differ on a share s of markets, its absolute lift is at most $\alpha_1 s$.*

(b) Suppose each complete input identifies its unedited counterpart and, when they differ, the edited source. One engine then attains lift $\alpha_1 s$ for every edit in the declared set. Under this identification assumption, zero lift on every engine is equivalent to $s = 0$. Proof in appendix B.1.

The bound applies without modeling the engine. Under part (b)'s identification assumption, some engine rewards every surviving difference, so the bound is tight. For presentation added alongside support, the reference is the support-only source: removal eliminates the presentation gain, not the support's reward. If removal holds regardless of rivals' choices, any positive editing cost makes the edit unprofitable; at zero cost, editing and not editing tie (Appendix A.2). Once citation positions are full, editing redistributes exposure rather than increasing its total, though individual publishers can gain (Proposition 9). The experiments check where input equality holds and measure surviving edits separately (Section 5.2).

4.2 PRESERVING THE REWARD FOR SUPPORTING MATERIAL

Recovery can remove presentation gains without sacrificing the reward for support. It requires the original source and correct identification of added support. Deletion alone cannot always do this: it cannot restore text a publisher removed. Content includes claims and supporting material; presentation edits preserve both. For transformation R, let $\lambda(R)$ be the largest positive presentation gain over the original or support-only source. Unprofitable edits contribute zero. The benchmark $H(R) - \lambda(R)$ rewards support and penalizes presentation gains; it is not the signed gap on a tested edit set.

Proposition 2 (Recovery and the limits of deletion)**.** *(a) Suppose a recovery map K returns the original source and any supporting material actually added. It returns the same text for every admissible presentation variant and leaves original and support-only submissions unchanged. For every deterministic policy R, applying K before R preserves $H(R)$ and makes each variant earn exactly what its recovered control earns. If the feasible family $\mathcal{F}$ contains $R \circ K$ whenever it contains R, then*

$$\sup_{R \in \mathcal{F}} \{H(R) - \lambda(R)\} = \sup_{\substack{R \in \mathcal{F} \\ R \text{ leak-free}}} H(R).$$

(b) This equality need not hold for deletion-only policies. There exist a market and an engine where a policy that retains presentation differences achieves $H - \lambda = \alpha_1$, but every leak-free deletion policy has $H = 0$. Proof in appendix B.2.

Part (a) requires the correct source and support and enough budget for recovery. Matching a passage verbatim does not verify its content (Appendix B.2). In part (b), the engine rewards support only

when a repeated source passage is present. A publisher can delete that repetition without changing content. To make the two inputs identical, a deletion-only policy must also remove the passage from the supported version, losing its reward. Thus the return to support after deletion depends on the engine. Our evaluated policy combines screening and normalization, following passage filtering and compression before generation (Xu et al., 2024; Wang et al., 2023). The screen restores the original when an edit adds query terms absent from both the source and the declared supporting sentence. The normalizer removes formatting markers, empty labels, redundant units and selected list fragments. Neither checks whether claims are true. Labels inside retained units and paraphrases can survive, so this is not exact recovery (Appendix C.2). We test removal and retained support returns separately (Appendix D.3).

4.3 Choosing and revising the policy

A policy that performs best on known edits may lose its advantage against new ones (Proposition 16). Nash-GEO expands the edit set before comparing policies again (Lanctot et al., 2017; Tramer et al., 2020). For a fixed edit set, let w_π be the probability of choosing policy π. Publishers know these probabilities but not the engine. The feasible set $\mathcal{W}$ enforces quality and cost constraints. Using estimated lift $\widehat{L}_w(a) = \sum_{h,\pi} r_h w_\pi \widehat{L}_{h\pi a}$, the platform minimizes the largest manipulation return:

$$\widehat{V}(B_M) = \min_{w \in \mathcal{W},\, t} t \quad \text{subject to} \quad \widehat{L}_w(a) \le t \quad \text{for every } a \in B_M \cup \{0\}. \tag{4}$$

Including no edit makes the objective nonnegative; the support-only control is excluded. A single policy is allowed, so randomization is optional. Appendix A.6 gives the general information setting.

Proposition 3 (Optimal policy on the measured game)**.** *Suppose $\mathcal{W}$ is nonempty, compact and convex, with linear quality and cost constraints. Equation 4 computes the value of the constrained zero-sum game between the platform and one deviating publisher, with rivals fixed. An optimal policy distribution and an optimal dual distribution over edits form a Nash equilibrium of this measured game. A feasible pure policy is optimal exactly when its largest measured manipulation return equals this value. The same conclusion holds for the objective $\widehat{L}_w(a) - H(w)$, which maximizes the gap with not editing included.* Proof in appendix B.3.

Subtracting H rewards support preservation and can change which policy is selected. Reporting both returns distinguishes reduced manipulation from reduced exposure overall. The equilibrium concerns the measured edit set and one deviating publisher, with rivals fixed. Search then proposes new edits. After audit, each is evaluated under every candidate policy before selection repeats (Algorithm 1). Comparing policies on the same expanded set separates stronger attacks from policy improvement. Write $\rho_B(\pi) = [\max_{a \in B} L_\pi(a) - H(\pi)]_+$ for the manipulation premium. For $B_{t+1} \supseteq B_t$,

$$\rho_{B_{t+1}}(\pi_{t+1}) - \rho_{B_t}(\pi_t) = \underbrace{\rho_{B_{t+1}}(\pi_t) - \rho_{B_t}(\pi_t)}_{\text{new edits found}} - \underbrace{\left[\rho_{B_{t+1}}(\pi_t) - \rho_{B_{t+1}}(\pi_{t+1})\right]}_{\text{policy improvement on the same set}}. \tag{5}$$

Expanding the edit set cannot lower the current policy's premium. A decline across rounds must therefore come from the policy change. The identity holds for population returns or a fixed payoff table (proof in appendix B.3). Theorem 3 gives a conditional convergence bound under explicit search-accuracy and payoff-error assumptions. Our empirical procedure uses a registered comparison after renewed search to retain or revise the policy.

5 Experiments

We test whether the policy reduces the reward for presentation manipulation while preserving the reward for supporting material. We then examine which components matter and whether the benefit extends to new edits, competing publishers and other datasets and models.

5.1 Setup

Evaluation. The main evaluation uses 200 previously unused ASQA questions with five retrieved sources each (Stelmakh et al., 2022). We fixed the questions and analysis before generating answers. We test Llama-3.1-8B, Qwen2.5-7B, Qwen3-8B and Tulu3-8B (Llama and Tulu share a pretrained

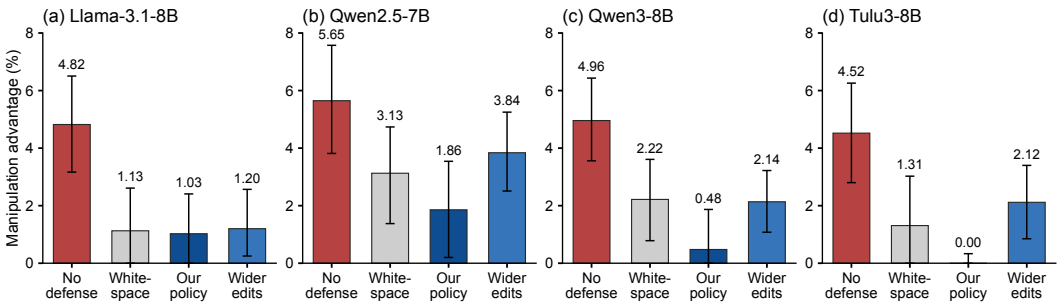

Figure 2: Manipulation's advantage over adding supporting material (%; lower is better), on 200 ASQA markets with three ordered citations. Each panel shows one engine. The first three bars compare policies on two searched edits; the final bar uses nine edits. The whitespace comparison is exploratory. Analysis details: Appendix D.10.

base). Subsequent comparisons use these questions and engines unless stated otherwise. One publisher edits its source while rivals stay fixed, except in the competition study. We label later exploratory comparisons and development studies separately.

Publisher edits. The primary test uses one searched presentation edit, alone and combined with a supporting passage. We expand this to nine content-preserving edits, including source-derived labels, headings and restructuring. The eleven-edit set adds two fabricated near-copies, one outside the content-preservation rules. We also test published GEO rewriting strategies (Aggarwal et al., 2024). The supporting-material control adds a question-relevant passage from a retrieved document outside the five shown sources. Its support for the source's claims is not independently verified. Appendix C.3 gives the edit definitions, passage selection and a worked example.

Baselines. We compare the committed policy with no defense and one prompt layer from SafeGEO, selected on development data (Wen et al., 2026). SafeGEO provides a directly comparable prompt-level defense while keeping the engine and retrieved sources fixed. Whitespace normalization, deletion alone and screening alone test the committed policy's individual components. Unedited sources and supporting-material additions serve as controls for publisher gains.

Measures. We measure citation visibility using reciprocal-rank weights, so earlier citations count more. We measure the strongest tested manipulation's advantage over adding supporting material, setting negative values to zero; lower is better. The incentive gap reverses the comparison without this adjustment, so a positive gap favors supporting material. Policy comparisons pair outcomes on the same questions. Appendices C.5 and C give the statistical analysis and full setup.

5.2 Manipulation and baseline comparisons

Manipulation loses ground to supporting material. On the two searched edits, the committed policy reduces manipulation's advantage from 4.99% without defense to 0.51%, a paired reduction of 4.48% (Table 1). Under the committed policy, adding supporting material outperforms the presentation-only edit by 7.02% across engines, with an advantage on each engine. The policy raises the return to supporting material by 1.21%, while an unedited source's exposure stays at approximately 26.8%. The improvement thus goes beyond lowering the reward for manipulation: publishers still gain by adding supporting material. The nearly unchanged baseline also separates this effect from simply citing the treated source less often.

Table 1: Manipulation's largest advantage over adding supporting material (absolute exposure %; lower is better), on 200 ASQA markets, four engines and three ordered citations. The searched no-defense versus committed comparison is registered; wider sets are descriptive and whitespace is exploratory. Bold marks lowest estimates, including ties. Analysis: Appendix D.10.

Policy	Searched 2 edits	Content-preserving 9 edits	Expanded 11 edits
No defense	4.99	4.99	4.99
Whitespace only	1.94	2.76	2.76
Committed	**0.51**	**2.32**	**2.32**

Broader tests retain a smaller benefit. With either nine or eleven edits, manipulation retains a 2.32% advantage under the policy, versus 4.99%

without defense. The paired reduction is 2.66% in both descriptive comparisons (Appendix D.11). A label attached to supporting material is the strongest remaining edit on every engine (Appendix Table 16). Adding more edits changes which strategy earns most under the policy, even though the no-defense maximum stays unchanged. This distinguishes performance on the searched edits from performance against a broader choice of strategies.

The reduction remains positive under alternative citation scoring. On the searched edits, the reduction is 4.13% when scoring only the first citation and 4.41% when scoring the first five. These descriptive checks re-score the same answers (Appendix D.11).

The policy outperforms a published defense layer. Manipulation's advantage is 1.06% under our policy, versus 8.32% without defense and 8.40% under SafeGEO-L1 (Wen et al., 2026) (Table 2). The paired reductions are 7.26% against no defense and 7.33% against the published layer. This exploratory test combines edits found with equal search budgets against our policy and no defense. All three policies face that same set. The benefit therefore persists when the comparison includes edits found against no defense, rather than only those found against our policy. SafeGEO-L1 is one prompt-level layer selected on development data, not the full defense; search did not target that layer.

Table 2: All three policies face the same combined edit set. Exploratory ASQA comparison, 200 markets, four engines; lower is better. Analysis: Appendix D.9.

Policy	Advantage (%)
No defense	8.32
SafeGEO-L1	8.40
Committed	**1.06**

Further search rules out a proposed revision. In development, search found a source-derived label missing from the original evaluation. Omitting it had overstated the incentive gap by 3.03%. A proposed label repair showed no advantage after another search, so the predefined comparison retained the original policy. The second search therefore changed the design decision: an improvement against familiar edits was not enough to justify replacing the policy. Search helped assess that revision (Appendix D.4).

The benefit extends to published GEO edits. The incentive gap improves by 7.86% on 140 development questions and 6.23% on a 300-question extension. The tests use nine published rewriting strategies (Aggarwal et al., 2024), generated by two models and scored on the first citation. Using published prompts also changes how the edits are produced: the strategies are fixed independently of our search. The positive gap changes therefore extend the finding to rewrites that were not generated by the Nash-GEO planner. These are pooled improvements over no defense; Appendix D.7 reports the defended levels and full analysis.

5.3 Policy components

Deletion improves on whitespace normalization. Whitespace normalization lowers manipulation's advantage to 1.94% on the searched edits. Deletion lowers it to 0.51%, a further reduction of 1.43%; screening adds no reduction. Simple formatting changes thus account for much of the benefit, but not all of it. Removing redundant text further reduces what these edits earn. Appendix D.10 reports this component comparison.

Surviving edits can still earn less. The policy rarely makes edited and unedited inputs identical. Most of the reduction therefore falls outside Proposition 1's guarantee. The checks record partial removal and the engine's response to the remaining text together (Appendix D.8). Where the theorem's condition does hold, the observed outcomes agree: all 448 audited tasks with prompts identical to the supporting-material control produce identical answers and zero additional reward. Exact removal gives an engine-independent guarantee; the other reductions depend on how the engine responds to the transformed source.

Screening helps against query-term additions. Normalization improves the incentive gap by 3.79%, and screening adds a further 0.98% in an exploratory test on 122 QAMPARI questions. This benefit is specific to query-term additions. Appendix D.7 and Table 7 give the component comparisons and earlier development results.

5.4 COMPETITION AND TRANSFER

An edit pays less when rivals adopt it too. Without defense, its gain falls from 3.82% when rivals do not edit to 0.73% when all four use the same strategy. This development study uses 119 QAMPARI questions and two engines. Both tested strategies earn less as adoption spreads: the gain falls by 2.30% for bullet formatting and 3.87% for section formatting. An edit that helps one publisher stand out becomes less useful when competitors make the same change. This is a competition effect without defense, not an effect of our policy. Appendix D.5 reports adoption gains and subsequent deviations.

The policy still helps when everyone edits. The largest gain from changing one publisher's edit falls from 7.05% without defense to 3.97% under the policy, on 48 ASQA development questions. Unlike the adoption comparison above, this test lets a publisher switch strategies after its rivals have edited. Common adoption reduces the payoff to copying an edit; the policy also reduces the opportunity to gain by choosing a different one. In the same study, universal adoption of bullet formatting changes the cited-source set in 16.2% of answers without defense and 0.7% under the policy. Thus the policy reduces both the editing reward and changes to which sources receive citations. A separate citation-position analysis finds gap improvements of 3.85%, 4.84% and 4.81% when scoring the first one, three and five citations. These scores use the same answers, not independent runs (Appendix D.5).

The average benefit transfers to a second dataset. On QAMPARI, manipulation's advantage falls from 4.35% to 0.34% (Table 3), and every engine meets the bound on remaining manipulation. Supporting material's return also rises, from 3.78% without defense to 4.94% under the policy (Appendix D.12). As in the primary evaluation, the smaller manipulation advantage accompanies a stronger reward for the supporting passage. The same pattern appears on both datasets despite their different questions and retrieved sources.

Table 3: Manipulation advantage (%) with three citations; lower is better. Within each row, policies use the same questions and edits. Rows describe different studies; bold marks lower estimates, not statistical significance. Analysis: Table 8.

Evaluation	Questions	No defense	Committed
ASQA	200	4.99	**0.51**
QAMPARI	122	4.35	**0.34**

The benefit extends across model sizes and families. On a 32B model, supporting material trails fixed keyword edits by 3.29% without normalization and leads them by 3.93% with it. On a 35B model, deletion lowers the average gain from changing one publisher's edit from 5.06% to 2.09%. The fixed-keyword comparison also improves the incentive gap by 3.70% on Gemma-2-9B and 3.37% on Mistral-7B. The latter results extend the comparison beyond the Llama and Qwen families used in the primary evaluation. All are exploratory component studies using separate development questions and edits. Appendices D.7 and D.8 identify the models and report these comparisons.

A passage-exempt policy retains the support reward. A separate policy version exempts the declared supporting passage from the normalizer's length cap. On 300 GEO-derived markets, supporting material earns 4.20% (3.65% without defense), while manipulation's advantage falls from 2.00% to 0.41%, a paired reduction of 1.59%. This exploratory evaluation uses edits searched against the revised policy (Appendix D.13).

Recovery removes the tested presentation advantage. Given the original source and declared supporting passage, recovery leaves zero advantage for the two tested QAMPARI edits. It retains the committed policy's 4.94% return to adding supporting material. Access to the original lets the platform discard presentation changes without discarding the added passage. This analytic comparison reuses recorded control answers rather than generating new ones (Appendix D.12).

6 CONCLUSION

Nash-GEO studies how platform policies shape publishers' incentives to improve content or manipulate presentation. Removing presentation differences eliminates their citation gains, but deletion can also reduce the reward for supporting material. In the primary evaluation, the committed policy makes manipulation less rewarding while preserving the advantage of supporting material. Generative search should therefore be designed around publisher responses, not only a fixed collection of edits.

AI USAGE STATEMENT

We used large language models to polish writing, identify potentially missing related work, and assist with drafting and reviewing manuscript passages. Under the authors' direction, AI agents also assisted with implementation, experiment execution and analysis, and developing and checking mathematical arguments. The authors are responsible for verifying the citations, proofs and reported results, and for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix documents the experimental setup and analysis procedures. Code, configurations and result records will accompany the anonymous release to support reproduction of the reported findings.

ETHICS STATEMENT

The reported experiments use public benchmark passages and open-weight models; no human-subject ratings enter the present results. Strategic edits are evaluated in controlled offline markets, not deployed against live publishers or services. We report attack delivery and unsupported-content risks to avoid presenting manipulation artifacts as robustness.

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*, KDD '24, pp. 5–16, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671900. URL https://doi.org/10.1145/3637528.3671900.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Ran Ben Basat, Moshe Tennenholtz, and Oren Kurland. A game theoretic analysis of the adversarial retrieval setting. *Journal of Artificial Intelligence Research*, 60:1127–1164, December 2017. ISSN 1076-9757. doi: 10.1613/jair.5547. URL http://dx.doi.org/10.1613/jair.5547.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a9a1d5317a33ae8cef33961c34144f84-Paper.pdf.

Elchanan Ben-Porath, Eddie Dekel, and Barton L. Lipman. Optimal allocation with costly verification. *American Economic Review*, 104(12):3779–3813, December 2014. doi: 10.1257/aer.104.12.3779. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.12.3779.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, December 2009. ISBN 9781400831050. doi: 10.1515/9781400831050. URL http://dx.doi.org/10.1515/9781400831050.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pp. 621–630. ACM, November 2009. doi: 10.1145/1645953.1646033. URL http://dx.doi.org/10.1145/1645953.1646033.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06, pp. 82–90, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932364. doi: 10.1145/1134707.1134717. URL https://doi.org/10.1145/1134707.1134717.

Sanmay Das, Fang-Yi Yu, and Yuang Zhang. Optimally auditing adversarial agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(20):16787–16794, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i20.38722. URL http://dx.doi.org/10.1609/aaai.v40i20.38722.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. Incentivizing truthfulness through audits in strategic classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5347–5354, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i6.16674. URL http://dx.doi.org/10.1609/aaai.v35i6.16674.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Robert S. Gibbons. *Game Theory for Applied Economists*. Princeton University Press, December 1992. ISBN 9781400835881. doi: 10.1515/9781400835881. URL http://dx.doi.org/10.1515/9781400835881.

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 '23, pp. 79–90, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

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 '16, pp. 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting. In *CAMLIS*, pp. 48–62, 2024. URL https://ceur-ws.org/Vol-3920/paper03.pdf.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models, 2023. URL https://arxiv.org/abs/2309.00614.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

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 '19, pp. 825–844. ACM, June 2019. doi: 10.1145/3328526.3329584. URL http://dx.doi.org/10.1145/3328526.3329584.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement

learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3323fe11e9595c09af38fe67567a9394-Paper.pdf.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009. URL http://www.cs.mcgill.ca/~colt2009/papers/012.pdf#page=1.

H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *ICML*, pp. 536–543, 2003. URL http://www.aaai.org/Library/ICML/2003/icml03-071.php.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6917–6926. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/miller20b.html.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 230–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287576. URL https://doi.org/10.1145/3287560.3287576.

Riku Mochizuki, Shusuke Komatsu, Souta Noguchi, and Kazuto Ataka. Exposing citation vulnerabilities in generative engines, 2026. URL https://arxiv.org/abs/2510.06823.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1), December 2008. ISSN 1046-8188. doi: 10.1145/1416950.1416952. URL https://doi.org/10.1145/1416950.1416952.

Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hkdqxN3c7t.

Ojas Nimase, Zhe Chen, Gengpei Qi, Yue Zhao, and Xiyang Hu. GEO-Bench: Benchmarking ranking manipulation in generative engine optimization, 2026. URL https://arxiv.org/abs/2605.29107.

Haritz Puerto, Martin Gubri, Tommaso Green, Seong Joon Oh, and Sangdoo Yun. C-SEO Bench: Does conversational SEO work? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0923. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/27aa3aeff0f8460a7b43d30fa6c5c032-Paper-Datasets_and_Benchmarks_Track.pdf.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.347. URL https://aclanthology.org/2024.emnlp-main.347/.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. VeriCite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 47–54. ACM, December 2025. doi: 10.1145/3767695.3769505. URL http://dx.doi.org/10.1145/3767695.3769505.

Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. Information retrieval meets game theory: The ranking competition between documents' authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 465–474, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080785. URL https://doi.org/10.1145/3077136.3080785.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

Zhihua Tian, Yuhan Chen, Yao Tang, Jian Liu, and Ruoxi Jia. Diagnosing and repairing citation failures in generative engine optimization, 2026. URL https://arxiv.org/abs/2603.09296.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3487–3493. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/484. URL https://doi.org/10.24963/ijcai.2022/484. Main Track.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1633–1645. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation, 2023. URL https://arxiv.org/abs/2311.08377.

Michael P. Wellman. Methods for empirical game-theoretic analysis. In *AAAI*, pp. 1552–1556, 2006. URL http://www.aaai.org/Library/AAAI/2006/aaai06-248.php.

Qianfeng Wen, Yifan Simon Liu, Xin Liu, Difan Jiao, Blair Yang, Junda Wu, and Zhenwei Tang. SafeGEO: Understanding generative engine optimization risks in recommendation agents, 2026. URL https://arxiv.org/abs/2606.28356.

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. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43305–43315, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2149. URL https://aclanthology.org/2026.findings-acl.2149/.

Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=K8EinVWtUB.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 392–410. IEEE, March 2026. doi: 10.1109/satml68715.2026.00029. URL http://dx.doi.org/10.1109/satml68715.2026.00029.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*, NDSS 2018. Internet Society, 2018. doi: 10.14722/ndss.2018.23198. URL http://dx.doi.org/10.14722/ndss.2018.23198.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-K recommendation under competing content creators? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39674–39701. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yao23b.html.

Xucheng Yu, Haibo Jin, Huimin Zeng, and Haohan Wang. SCI-Defense: Defending manipulation attacks from generative engine optimization, 2026. URL https://arxiv.org/abs/2605.21948.

Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. AgenticGEO: A self-evolving agentic system for generative engine optimization, 2026. URL https://arxiv.org/abs/2603.20213.

Bing Zheng, Zongyao Zhao, and Wenming Yang. Counter-GEO-Bench: Evaluating defenses against information-distorting generative engine optimization, 2026. URL https://arxiv.org/abs/2609.02316.

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, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Additional Theoretical Results

This appendix collects the theoretical statements that support Section 3 and Section 4.1 to Section 4.3 but are not needed to read them. Each subsection names the main-text section it extends. Proofs appear after the statements or in Appendix B, as indicated.

A.1 Surviving edits and measured zeros (Section 4.1)

Context, information and the object that is invariant. A context fixes the query, original sources, focal publisher, rival edits, engine, source order and public serving information. A property stated for a policy holds of a procedure W that mixes policies when it holds of every policy in the support of W. The complete-input map T includes every channel available to the downstream engine; a screen that reads an original source is part of T, not an unexamined postprocessing stage. Expected payoffs use the specified decoding kernel. Greedy decoding is the special case of a point-mass kernel; equality of input distributions suffices without assuming greedy decoding. For source-wise policies, coordinate invariance implies profile invariance by replacing coordinates one at a time. The same conclusion for a profile-dependent policy instead requires equality of the *entire* served profile.

Proposition 4 (Payoffs of surviving edits)**.** *(a) Equal complete-input distributions give equal expected exposure. On a finite response bank, the manipulation return is the maximum of the surviving edits' averaged lifts and the zeros of removed members; with signals, the maximum is taken separately*

for each signal before summing. (b) For a presentation-only bank B without an informative signal, define $\lambda_B(R) = [\max_{a \in B} L_R(a)]_+$. If the bank contains a removed member, its maximum equals $\lambda_B(R)$ and the signed gap equals $H(R) - \lambda_B(R)$. Otherwise the signed manipulation return may be negative. This bank-restricted quantity need not equal $\lambda(R)$, which also ranges over presentation edits of support-only sources. (c) Input invariance determines which edits are removed. The actual inputs served determine both the remaining payoffs and the honest return; the invariance partition alone does not determine them. (d) For deterministic preprocessing, the largest exposure lift over all engines is at most α_1 times the probability of unequal edited and unedited complete inputs. Under the identification assumption of Proposition 1(b), one engine reaches this bound simultaneously for every edit in the declared set. Proof in appendix B.1.

Proposition 1 follows from parts (a) and (d); the proofs are in Appendix B.1. The recovery argument of Appendix B.2 proves Proposition 2(a), and the deletion counterexample there proves part (b). Content classes refer to claims and supporting material, so the honest addition is not identified with presentation alone.

Graded removal. Assume, as an additional hypothesis on the engine class $\mathcal{E}$ and not as a property established for the engines evaluated here, one of the two regularity conditions below; without one of them, Proposition 1(b) is the whole story and the share of markets on which an edit reaches the engine is all that an engine-independent bound can see. Fix a deterministic policy applied source by source, a rival profile, and an *anchor* $b \in \{o, h\}$: the unedited source for a presentation-only edit, the support-only source for an edit stacked on the supporting passage. Write $T(x, a)$ for the complete engine input in market context x after edit a, $S_a(x) = \mathbf{1}\{T(x, a) \neq T(x, b)\}$, and $e_h(t) \in [0, \alpha_1]$ for the treated source's expected exposure on engine h at input t. The *increment* of a over its anchor is $\Delta_a = \mathbb{E}[e_h(T(x, a)) - e_h(T(x, b))]$; for $b = o$ it is the lift of Equation 1, and for $b = h$ it is the edit's premium over support with rivals fixed. For an edit that adds tokens to its anchor, let $N_a(x)$ be the nonempty multiset of tokens it adds and $\phi_a(x) \in [0, 1]$ the share of $N_a(x)$ whose multiplicity in $T(x, a)$ exceeds its multiplicity in $T(x, b)$; this is the surviving-token fraction, and $\phi_a = 0$ wherever $S_a = 0$. Edits that add no token (reordering, fronting, bare markers) have no surviving fraction and receive no bound here beyond Proposition 1. Let d be a declared distance on complete inputs with values in $[0, 1]$ and $d(t, t') = 0$ exactly when $t = t'$.

(E-Mon) There is a nondecreasing $\psi : [0, 1] \to [0, \alpha_1]$ with $\psi(0) = 0$ such that $e_h(T(x, a)) - e_h(T(x, b)) \leq \psi(\phi_a(x))$ for every $h \in \mathcal{E}$, every context and every token-adding edit.

(E-Lip) There is $\Lambda \geq 0$ such that $|e_h(t) - e_h(t')| \leq \Lambda\, d(t, t')$ for every $h \in \mathcal{E}$ and all inputs t, t'.

Proposition 5 (Graded removal)**.** *(a) Under (E-Mon), for every token-adding edit a and every $h \in \mathcal{E}$, the increment satisfies $\Delta_a \leq \mathbb{E}[\psi(\phi_a)] \leq \alpha_1 \Pr(\phi_a > 0)$; if in addition $\psi(u) \leq \Lambda u$, then $\Delta_a \leq \Lambda\, \mathbb{E}[\phi_a]$. (b) Under (E-Lip), $|\Delta_a| \leq \Lambda\, \mathbb{E}[d(T(x, a), T(x, b))]$ for every $h \in \mathcal{E}$. (c) Over the class of all engines no bound below $\alpha_1 \Pr(S_a = 1)$ exists: under the identification assumption of Proposition 1(b), applied with the anchor in the role of the unedited input, $\sup_h \Delta_a = \alpha_1 \Pr(S_a = 1)$ for every edit, whatever the values of ϕ_a and $d(T(x, a), T(x, b))$ on the event $S_a = 1$. In particular no function of the surviving fraction alone bounds the increment over all engines, since cells with $S_a = 1$ and $\phi_a = 0$ pay α_1 under the extremal engine, and for every Λ the bound $\Lambda\, \mathbb{E}[\phi_a]$ fails for any edit whose surviving fraction is below α_1/Λ wherever it reaches the engine.* Proof in appendix A.1.

Proof. (a) $\Delta_a = \mathbb{E}[e_h(T(x, a)) - e_h(T(x, b))] \leq \mathbb{E}[\psi(\phi_a)]$ by (E-Mon). Since ψ is nondecreasing with $\psi(0) = 0$ and $\psi \leq \alpha_1$, $\psi(\phi_a) \leq \alpha_1 \mathbf{1}\{\phi_a > 0\}$; if $\psi(u) \leq \Lambda u$ then $\psi(\phi_a) \leq \Lambda \phi_a$. Take expectations. On $\{S_a = 0\}$ the two inputs coincide, so $\phi_a = 0$ there and the bound is consistent with Proposition 1(a). (b) Apply (E-Lip) to $t = T(x, a)$ and $t' = T(x, b)$ and take expectations. (c) The engine constructed in the proof of Proposition 1(b), with φ_0 returning the anchor's input, cites the treated source first whenever its input differs from the anchor's input and cites nothing otherwise. Its decision reads only the event $\{t \neq \varphi_0(t)\}$, so it gives increment $\alpha_1 \Pr(S_a = 1)$ to every edit regardless of how much of the edit survives or how far the two inputs are apart; a bound valid for every engine is therefore at least $\alpha_1 \Pr(S_a = 1)$. A cell with $S_a = 1$ and $\phi_a = 0$ contributes α_1 to that increment and 0 to any function of ϕ_a that vanishes at 0. If $\phi_a < \alpha_1/\Lambda$ on $\{S_a = 1\}$, then $\Lambda\, \mathbb{E}[\phi_a] = \Lambda\, \mathbb{E}[\phi_a \mathbf{1}\{S_a = 1\}] < \alpha_1 \Pr(S_a = 1)$, so the extremal engine violates that bound. □

Application to surviving edits. The empirical reduction need not come from exact removal. The measured survival strata are averages, whereas (E-Mon) and (E-Lip) constrain individual inputs. They therefore do not validate either assumption or identify its constants. Document differences, partial removal and engine responses can all contribute to these contrasts; the observations do not separate their effects. Appendix D.9 reports the delivery diagnostics.

Removal, survival and measured zeros. For deterministic preprocessing and a common context, set $S_a = \mathbf{1}\{T_W(x,a) \neq T_W(x,0)\}$. Since per-publisher exposure lies in $[0,1]$, $|\ell_W(a)| \leq \Pr(S_a = 1)$. This is an upper bound from representation survival, not a demonstration that survivors are profitable. A surviving response can have negative, zero or positive lift. Identical input distributions imply equal expected payoffs, not identical *independently sampled* answers; pathwise equality requires deterministic decoding or an appropriate shared-randomness coupling. Consequently, an estimated or rounded zero and a nondegenerate interval do not themselves certify removal. Exact-zero labels in empirical tables require the specified reference input and an authenticated input-equality check; support-plus-presentation is compared with support alone. Composition cannot reopen an equality when each subsequent stage reads only the previous output and fixed context. A raw-text bypass violates that premise. Nor does a smaller surviving set imply smaller surviving *payoffs*: downstream allocation may change. Screen inertness on a particular response set must be checked from its executable rule and input bytes; it is not implied by calling the screen "grounded."

A.2 THE PUBLISHERS' GAME (SECTION 4.1)

Both statements below concern the game publishers play under a fixed policy. The experiments assign no editing cost and do not establish these equilibria.

Proposition 6 (Equilibria)**.** *Let every publisher edit its own source and be paid its exposure under R, and let the* content game *restrict each publisher to one representative per content class, available to that publisher, with the unedited source representing its own class. Write $\bar{\lambda}(R)$ for the largest amount any presentation of a source earns over its representative at any profile of rival edits. (a) If R is leak-free, $\bar{\lambda}(R) = 0$: the two games have the same payoffs, hence the same equilibria up to presentation, and at every profile a presentation variant earns nothing over its content-class representative. (b) For any R, every equilibrium of the content game is a $\bar{\lambda}(R)$-equilibrium of the full game, and no smaller constant works for every policy. (c) Finite response sets give finite content games, which have mixed equilibria, and at the profile in which rivals are unedited M and H are exactly the unilateral gains this paper measures.* Proof in appendix B.1.

Under a leak-free policy the publishers' game reduces to the declared content game. Otherwise, leakage upper-bounds the regret of its equilibria in the full game; it need not equal that regret. The constant $\bar{\lambda}(R)$ is a supremum over rival profiles, while $\lambda(R)$ is read at the profile in which rivals are unedited, and that profile gives a lower bound on the supremum. For a deterministic source-wise policy, α_1 times the largest input-change probability over the declared edits and their representatives bounds presentation gains at every rival profile. If the content game contains more than the original and support-only representatives, the bound must range over those representatives too.

Proposition 7 (Dominance under a positive edit cost)**.** *Let every non-identity presentation edit cost $c > 0$ in exposure units and let not editing be free. (a) If a policy removes an edit, not making that edit strictly dominates making it at every rival profile; under a leak-free policy this holds for every presentation edit, so in the game restricted to presentation edits the profile in which nobody edits is the unique equilibrium. (b) Under a leaky policy, with each publisher's unedited source as its representative, the same holds for every presentation edit whenever $c > \bar{\lambda}(R)$.* Proof in appendix B.1.

A rule that removes every presentation edit dissolves the arms race that prior work describes as a prisoner's dilemma (Nestaas et al., 2025), whatever the profiles between the anchors look like; whether universal manipulation is itself an equilibrium under unprotected serving is a question two anchor profiles cannot decide. In the model, with a strictly positive edit cost, paying for a removed edit is a pure loss, and at the no-edit equilibrium nothing is paid and exposure is allocated on content alone; our experiments price no edit, so this is a property of the game and not a measured quantity.

Zero edit cost. Assume every edit is free, $c \equiv 0$, exposure is valued at $v_i > 0$, the policy is applied source by source, and a removed edit leaves the complete input unchanged at every rival profile; Proposition 7 treats the case $c > 0$ and this statement is its companion.

Proposition 8 (Removed edits at zero cost)**.** *(a) At every rival profile a removed edit earns exactly the exposure of not editing: the two are payoff-equivalent, so not editing is never strictly better and never worse. (b) Under a leak-free policy every profile of presentation edits gives every publisher the payoff of the profile in which nobody edits. The no-edit profile is therefore an equilibrium of the presentation-only game, and so is every other profile: the equilibrium set is the whole profile space, and the no-edit profile is not unique. (c) Under a policy with a surviving presentation edit a whose lift with rivals unedited is positive, the no-edit profile is not an equilibrium: a is a profitable deviation. (d) With $c > 0$ the equivalence in (a) becomes strict dominance and the multiplicity in (b) collapses to the unique no-edit equilibrium of Proposition 7.* Proof in appendix A.2.

Proof. (a) Identical complete inputs give identical expected exposure on every engine (Proposition 4(a)); with $c = 0$ the utilities coincide. (b) Under a leak-free policy every presentation edit of every publisher is removed, so replacing edits one coordinate at a time leaves the complete input, hence every publisher's exposure, unchanged; every profile has the same payoff vector, so no unilateral deviation from any profile changes any payoff, and every profile is an equilibrium. (c) A positive lift with rivals unedited is by definition a profitable unilateral deviation from the no-edit profile. (d) With $c > 0$ a removed edit costs c and earns nothing over not editing, which is strict dominance; a strictly dominated strategy carries no weight at any equilibrium, which is the uniqueness in Proposition 7. □

Application to publisher competition. The experiments assign zero editing cost. Prompt-identical tasks check part (a), while surviving edits fall under part (c). Comparing one edit at two rival profiles does not test all unilateral deviations from a single profile, so the universal-adoption study does not establish equilibrium or the positive-cost uniqueness conclusion. Appendix D.5 reports these comparisons.

Adoption, edit costs and the measured anchors. In a binary-action game, let $\kappa_i = \sup_{\mathbf{a}_{-i}}[e_i(a, \mathbf{a}_{-i}) - e_i(0, \mathbf{a}_{-i})]$. With no-edit cost zero, fixed edit cost c_i and exposure value $v_i \geq 0$, not editing is weakly dominant exactly when $c_i \geq v_i \kappa_i$; strict inequality is sufficient for strict dominance. The clean and all-adopt anchors are only two rival profiles and give a lower bound on this threshold. Nor can κ_i be replaced by the representative-based $\bar{\lambda}_R$ unless the no-edit baseline is the relevant representative or a separate two-sided payoff bound is supplied. Two favorable anchors do not establish unique universal adoption. For four symmetric publishers assign (editor, noneditor) exposures by number of editors m as $0 : (-, 1/4)$, $1 : (2/5, 1/5)$, $2 : (1/10, 2/5)$, $3 : (3/10, 1/10)$, $4 : (1/4, -)$, with cost $1/20$. Total exposure is one. Both anchor editing gains are $3/20 > 1/20$, yet universal adoption and each of the four single-editor profiles are equilibria. Full-profile data or additional monotonicity is needed. Universal-adoption mean exposure change equals the change in total fill divided by the number of publishers; this is accounting, not a robustness result. Invariance removes both the expected payoff to editing alone and the change in which sources are cited, for the removed class; a low payoff for a surviving edit need not stop that movement.

The accounting identity behind the universal-adoption results of Section 5.4 is the following.

Proposition 9 (Removal, resistance and the fill)**.** *Exposure is positional: summed over the publishers in a market it equals the weight of the positions the engine fills, the* fill*, so a publisher's gain from changing its own edit is exactly its rivals' loss plus the change in the fill, and universal adoption of one edit changes average exposure by the fill change it causes divided by the number of publishers. An edit a policy removes changes no publisher's exposure at any profile and, under deterministic decoding, moves no citation; an edit it merely resists can earn nothing and still move citations.* Proof in appendix B.1.

A.3 Recovery, deletion and the best leak-free policy (Section 4.2)

Proposition 2(a) concerns an exact recovery map. The results below bound approximate recovery and state when a replacement preserves the current policy's input equalities. They then characterize policies that can only delete.

Approximate recovery. Let $\widehat{K}$ be any deterministic map that reads what the platform observes. Let ϵ_0 and ϵ_h be the shares of markets on which the complete input under $R \circ \widehat{K}$ differs from the

complete input under R at the unedited and at the support-only profile. Let δ be the largest share of markets, over the same presentation edits and original or support-only anchors defining λ, on which the complete input under $R \circ \widehat{K}$ differs between the edited submission and its anchor. Equal complete inputs have equal expected exposure and exposure lies in $[0, \alpha_1]$, so on every engine

$$|H(R \circ \widehat{K}) - H(R)| \le \alpha_1(\epsilon_0 + \epsilon_h), \qquad \lambda(R \circ \widehat{K}) \le \alpha_1 \delta,$$

and therefore

$$H(R \circ \widehat{K}) - \lambda(R \circ \widehat{K}) \;\ge\; \big[H(R) - \lambda(R)\big] + \lambda(R) - \alpha_1(\epsilon_0 + \epsilon_h + \delta).$$

An imperfect recovery map raises $H - \lambda$ whenever the manipulation return it removes exceeds its errors on unedited and support-only submissions plus the leakage it leaves. The three shares are properties of R and $\widehat{K}$ on a declared population of submissions, not of the engine, and a finite response set does not establish them for arbitrary edits.

Recovery must preserve the support state. Access to an earlier source does not by itself identify which supporting material a submission retains. A proposed recognizer checks for the declared passage verbatim, up to whitespace, and for the original token multiset. This does not cover arbitrary presentation variants: enclosing the sentences of a two-sentence passage in separate quotation marks can break its literal-span test. For the pinned policy P there are an original o, its support-only form h, and such a quoted form x for which $P(x) = P(h)$, but the recognizer returns $K(x) = o$ and $K(h) = h$, with $P(o) \neq P(h)$. Thus $P \circ K$ introduces an input distinction that P had removed. This is a counterexample to that proposed recognizer, not a change to the committed policy.

For a fixed read-only source and context record, a replacement map Q preserves every equality of a deterministic policy P exactly when $Q = F \circ P$ for some map F on P's outputs. Consequently, mapping two declared domains $\mathcal{D}_0, \mathcal{D}_1$ to distinct required anchor outputs is possible without new distinctions exactly when $P(\mathcal{D}_0) \cap P(\mathcal{D}_1) = \varnothing$. Include the anchors in their respective domains. For a finite edit grammar, group submissions by their existing outputs and reject groups with conflicting targets. This gives a computable compatibility check, not a semantic-support oracle or a policy evaluated in the experiments. The proof is in Appendix B.2.

Recovery coverage. Assume the platform holds the declared support record of each source, namely the original text o and the declared passage s with its identifier and title, that the recovery map reads a submission x together with that record and nothing else, that s is a whole-token span wherever it occurs in o or in a submission, occurrences being counted after whitespace normalization, and that the policy R is applied source by source; tokens are the screen's tokens, lowercase alphanumeric runs, counted with multiplicity, and $\mathrm{bag}(t)$ is the token multiset of t. Let h be the canonical honest form, o followed by one line that names the passage's identifier and title and quotes s. With $\#_t(s)$ the number of occurrences of s in t and $x \ominus s$ the text x with one occurrence of s replaced by a space, the shipped map is

$$K(x) = \begin{cases} h & \text{if } \#_x(s) > \#_o(s) \text{ and } \mathrm{bag}(o) \subseteq \mathrm{bag}(x \ominus s), \\ o & \text{otherwise, if } \mathrm{bag}(o) \subseteq \mathrm{bag}(x), \\ x & \text{otherwise.} \end{cases}$$

Passage identity is a string test: K rewards the presence of the declared passage, and it does not verify that the passage supports the source's claims.

Proposition 10 (What K reverts and what it passes through)**.** *Let $\Sigma = \{x : \mathrm{bag}(o) \subseteq \mathrm{bag}(x)\}$ be the submissions that keep every occurrence of every source token. Fix a market and a rival profile, and write $e(t)$ for the treated source's expected exposure when its served text is $R(t)$ with the rivals' served texts held fixed. (a) $K(x) \in \{o, h\}$ for every $x \in \Sigma$, with no use of the span assumption; under it, $K(x) = x$ for every $x \notin \Sigma$, and K fixes o and h. (b) Every submission obtained from o or from h by edits that only add tokens at token boundaries or reorder text, applied once or stacked, lies in Σ: formatting markers, quotation, section headers, fronting, reordering, a label built from the source's own words, keyword lines and added grounded sentences are of this kind. K sends such a submission to h when it carries s verbatim, with or without the identifier line, more often than o does and the remainder still contains $\mathrm{bag}(o)$, and to o otherwise. (c) Every submission that loses one occurrence of one source token lies outside Σ and passes through: a paraphrase that replaces*

a word, a deletion, a truncation, a marker inserted inside a token. For such x, $R \circ K$ serves $R(x)$. (d) Equality with the selected control. For $x \in \Sigma$, the exposure of x under $R \circ K$ equals $e(K(x))$, the exposure of the control $K(x) \in \{o, h\}$ under R, market by market, at every fixed rival profile. (e) Preservation of the controls. K fixes o and h, so the unedited exposure and the honest return are the same under $R \circ K$ as under R: $H(R \circ K) = H(R)$. (f) Bounds for a response rule a whose submission lies in Σ in every market. Writing $c_m = K(x_m)$ for its control in market m and $e_m(\cdot)$ for that market's exposure, its lift and its premium over support satisfy

$$\mathbb{E}[e_m(c_m)] - \mathbb{E}[e_m(o_m)] \le \mathbb{E}\big[(e_m(h_m) - e_m(o_m))_+\big],$$
$$\mathbb{E}[e_m(c_m)] - \mathbb{E}[e_m(h_m)] \le \mathbb{E}\big[(e_m(o_m) - e_m(h_m))_+\big],$$

and both bounds are attained by a rule that selects, in each market, whichever control earns more. The premium over support is zero for every rule anchored at h in every market, is $-H(R)$ for every rule anchored at o in every market, and vanishes for every rule in Σ if and only if $e_m(h_m) \ge e_m(o_m)$ for almost every market. A rule whose anchor varies with the market can earn a positive premium over support. Proof in appendix A.3.

Proof. (a) For $x \in \Sigma$ the second case applies whenever the first does not, so $K(x) \in \{o, h\}$. For $x \notin \Sigma$ the second case fails; under the span assumption blanking an occurrence of s removes whole tokens and adds none, so $\mathrm{bag}(x \ominus s) \subseteq \mathrm{bag}(x)$ and the first case fails too. $o \in \Sigma$ with $\#_o(s) = \#_o(s)$, so $K(o) = o$; h contains s once more than o, and under the span assumption $h \ominus s$ has the token multiset of o plus the identifier line's tokens, so $K(h) = h$. (b) Adding whole tokens or moving text never lowers a multiplicity, so the submission stays in Σ, and (a) gives the two outputs; the stacked case follows because the honest test reads the passage count and the remainder's bag only. (c) A lost occurrence breaks the containment, so the third case applies. (d) $R \circ K$ serves $R(K(x))$; with the rivals' served texts fixed the complete input is that of the control, so Proposition 4(a) gives equal expected exposure. (e) is (d) at $x = o$ and $x = h$ with (a). (f) By (d), $e_m(c_m) \le \max\{e_m(o_m), e_m(h_m)\} = e_m(o_m) + (e_m(h_m) - e_m(o_m))_+ = e_m(h_m) + (e_m(o_m) - e_m(h_m))_+$; take expectations. The market-wise selector attains both. A rule with $c_m = h_m$ everywhere earns $\mathbb{E}[e_m(h_m)]$, premium zero; with $c_m = o_m$ everywhere it earns the unedited exposure, premium $-H(R)$. If $e_m(h_m) \ge e_m(o_m)$ almost everywhere the second bound is zero; if $e_m(o_m) > e_m(h_m)$ on a set of positive probability, the rule that selects o_m there and h_m elsewhere earns $\mathbb{E}[(e_m(o_m) - e_m(h_m))_+] > 0$ over support. Counterexample with two equally likely markets: original exposures $(1, 0)$, support-only $(0.2, 1)$; a submission sent to o in the first market and to h in the second earns $(1, 1)$, so its premium over support is $1 - 0.6 = 0.4 > 0$ while $H = 0.6 - 0.5 = 0.1 > 0$; the bound gives $\frac{1}{2}(0.8 + 0) = 0.4$, which is tight. □

Application to analytic recovery. Recovery maps the evaluated submissions to recorded controls, giving zero premium on those edit sets (Appendix D.9). This is an analytic composition, not a new generation or a property of the committed policy. The result requires the original source and declared passage. It neither verifies the passage's support for the source's claims nor covers edits that omit source tokens. Market-dependent choices between the original and supporting-material controls can also retain positive premium, as part (f) shows.

Proposition 11 (Policies that can only delete)**.** *There are a market and an engine under which, even with the earlier source in the platform's hands, a policy that only deletes and lets a presentation edit through has $H - \lambda = \alpha_1$, while every leak-free policy that only deletes has $H = 0$.* Proof in appendix B.2.

Proposition 12 (Deletion policies)**.** *Consider policies that delete units of a source and order the ones they keep. Fix a content class $[d]$ and let its* common core *J be the multiset that keeps each unit at the smallest number of copies any member of the class contains. (a) If R is leak-free, its output on every member of $[d]$ is a sub-multiset of J. (b) A leak-free, content-preserving policy exists exactly when $I(J) = I(d)$; its outputs are then the sub-multisets J' of J with $I(J') = I(d)$, each in an order fixed by the class. (c) Such a policy alters every source with a unit that an admissible edit deletes without changing content, so the share of unedited sources it alters is at least the share that contain one.* Proof in appendix B.2.

Content classes and allowed transformations. Equivalence here is specified independently of the policy. If it is the connectivity closure of admissible presentation edits, the theorem guarantees

invariance only on those classes; identifying them with *all* sources having the same semantic content requires a separate connectivity assumption. For $\mathcal{C} = \{(u), (u, u)\}$, the common core is (u): retaining one copy is invariant although each copy in (u, u) is redundant. Thus "every redundant occurrence must be removed" is not a correct multiset claim. For distinct synonymous units with $\mathcal{C} = \{(u), (v)\}$ the common core is empty, so no content-preserving delete-only rule exists. Normalization changes the representation on which the deletion rule operates. When order must be preserved rather than chosen, replace common submultisets by common *subsequences*; a multiset characterization cannot justify arbitrary reordering. Sorting and reverse-sorting are distinct valid class-determined orders. An internal label is a survivor only if the relevant modified unit is retained *and* its distinguishing content is not otherwise normalized away; whole-unit decisions alone are insufficient.

Proposition 13 (The leak-free form that alters the fewest unedited sources)**.** *Take the content classes to be the sets of sources that admissible presentation edits connect, which partition the sources, and consider policies that read only the submitted text. (a) A leak-free, content-preserving policy of this kind serves each content class $\mathcal{C}$ as one text $y_{\mathcal{C}} \in F_{\mathcal{C}}$, where $F_{\mathcal{C}}$ is the set of texts with the content of $\mathcal{C}$ that the policy family can produce from every member of $\mathcal{C}$. (b) Let $\pi(\mathcal{C})$ be the share of unedited sources in $\mathcal{C}$ and $p(y \mid \mathcal{C})$ the share of those that read exactly y. If every $F_{\mathcal{C}}$ is nonempty and the family contains, for every choice of one $y_{\mathcal{C}} \in F_{\mathcal{C}}$ per class, the policy that serves each class as its chosen text, then the least share of unedited sources such a policy alters is*

$$1 - \sum_{\mathcal{C}} \pi(\mathcal{C}) \max_{y \in F_{\mathcal{C}}} p(y \mid \mathcal{C}),$$

attained by serving each class as its most common unedited form in $F_{\mathcal{C}}$; it is zero exactly when, in every class that holds unedited sources, almost all of them read as one member of $F_{\mathcal{C}}$. (c) A policy that also reads an earlier copy of the source, as $R \circ K$ does in Proposition 2, is not bound by (a) and can leave every unedited source unaltered. Proof in appendix B.2.

Part (c) of Proposition 4 splits the platform's design problem into an equivalence and a representative. Let $\mathcal{F}$ be the family of policies available to the platform and write $\Delta^\star = \sup_{R \in \mathcal{F}}[H(R) - \lambda(R)]$ for the best value available in this accounting, in which an unprofitable edit counts as zero. The supremum over leak-free policies equals $\Delta^\star$ unless some leaky policy exceeds the best leak-free honest return by more than it leaks. Proposition 2 excludes that case for a family closed under recovery. For policies that only delete the case occurs under some engine (Proposition 11), so which of the two holds is a property of how engines read text; it is why H is reported beside M throughout. The gap-maximizing form of the program of Section 4.3 (Appendix A.6) instead uses the bank-specific M_+, with declared quality and cost constraints. It matches $H - \lambda$ only when the response universe, references and information structure match; support-plus-presentation lifts measured against no edit must not be equated with their incremental presentation gains. This is a formulation, not the selection rule that retained the empirical committed policy. A positive measured gap says that on the named response set no manipulation out-earned the supporting passage; it does not say the policy is optimal.

A.4 Position robustness (Section 4.3)

Theorem 1 (Exact reduction for monotone citation prominence)**.** *For every $x \in \mathbb{R}^C$,* $\max_{\alpha \in \mathcal{M}_C} \sum_{j=1}^{C} \alpha_j x_j = \max_{k \in [C]} \frac{1}{k} \sum_{j=1}^{k} x_j$. Proof in appendix B.3.

With $L^{(k)}_{h\pi a}$ the mean lift over the first k positions, the position-robust program is

$$\widehat{V}^K_{\mathrm{pos}}(B) = \min_{W, \{t_z\}} \sum_z t_z \quad \text{s.t.} \quad \sum_{h,\pi} r_h K(z \mid h) w_{\pi|h} \widehat{L}^{(k)}_{h\pi a} \le t_z \ \ \forall z,\, a \in B,\, k \in [C], \tag{6}$$

under the constraints of Equation 14. The maximum over k is evaluator uncertainty about how readers value position, not a publisher action; the program multiplies the constraint count by C and reduces to Equation 14 at $C = 1$.

Equation 6 therefore uses $|\mathcal{Z}||B|C$ constraints. It is exact for a shared unknown attention curve with one uninformative signal; with informative signals it is a conservative envelope that lets the worst prefix vary by signal (see the next paragraph). The fixed reciprocal-rank endpoint we report is not a search over all attention curves.

Shared attention weights versus a signal-wise envelope. Theorem 1 follows from the convex decomposition $\alpha = \sum_{k=1}^{C} k(\alpha_k - \alpha_{k+1})u^{(k)}$, with $\alpha_{C+1} = 0$ and $u^{(k)}$ uniform on its first k entries. For a *single* unknown population-wide attention curve, maximizing over responses and then over this curve gives

$$\max_{k\in[C]} \sum_{z} \max_{a\in B} \sum_{h,\pi} r_h K(z \mid h) w_{\pi|h} L^{(k)}_{h\pi a}.$$

Equation 6 instead uses the upper envelope obtained by allowing k to vary with z. They agree for one uninformative signal, or when one maximizing prefix works at every signal; otherwise the envelope can be strict. Both are distinct from the experiments' fixed reciprocal-rank endpoint, and capacity prefixes of one answer are not independent replications. Reciprocal-rank weights use their stated unnormalized scale; Theorem 1 uses normalized attention weights.

A.5 When engine conditioning has value (Section 3.2 and Section 4.3)

Let $\mathcal{W}_{\text{global}} = \{W : w_{\cdot|h} = w \in \Delta(\Pi),\ \forall h\}$ and define $V^K_{\mathcal{C}}(B) = \min_{W\in\mathcal{C}} \mathcal{R}^K_L(W; B)$ for either commitment class. Channel K_1 is a garbling of K_2 if $K_1(z_1 \mid h) = \sum_{z_2} G(z_1 \mid z_2) K_2(z_2 \mid h)$ for a stochastic map G.

The separable dual formulas below use the full product of engine-wise policy simplices, without shared quality or cost constraints. For a coupled feasible set $\mathcal{F}$, retain $\min_{W\in\mathcal{F}} \sum_{h,\pi} r_h w_{\pi|h} \ell^K_{h\pi}(\lambda)$ instead of separating that minimum across engines. Class nesting and information ordering still hold when the compared feasible sets are nested and fixed across channels.

Let $\mathcal{G} = B^{\mathcal{Z}}$ denote the complete signal-contingent response plans. The dual variable $\lambda \in \Delta(\mathcal{G})$ is the adversary's distribution over such plans; it is not the platform's policy randomization. Define $\ell^K_{h\pi}(\lambda) = \sum_g \lambda_g \sum_z K(z \mid h) L_{h\pi g(z)}$, $f_K(\lambda) = \sum_h r_h \min_\pi \ell^K_{h\pi}(\lambda)$, $g_K(\lambda) = \min_\pi \sum_h r_h \ell^K_{h\pi}(\lambda)$, and

$$D_K(\lambda) = \min_\pi \sum_h r_h \ell^K_{h\pi}(\lambda) - \sum_h r_h \min_\pi \ell^K_{h\pi}(\lambda) \geq 0. \tag{7}$$

Theorem 2 (Dual disagreement characterization)**.** *The conditioned and global values satisfy*

$$V^K_{\text{cond}}(B) = \max_\lambda \sum_h r_h \min_\pi \ell^K_{h\pi}(\lambda), \tag{8}$$

$$V^K_{\text{global}}(B) = \max_\lambda \min_\pi \sum_h r_h \ell^K_{h\pi}(\lambda). \tag{9}$$

If Λ_c and Λ_g are their respective dual-optimal sets, then

$$\max_{\lambda\in\Lambda_c} D_K(\lambda) \leq V^K_{\text{global}}(B) - V^K_{\text{cond}}(B) \leq \min_{\lambda\in\Lambda_g} D_K(\lambda). \tag{10}$$

The gain is zero exactly when a global-dual optimizer has one policy minimizing $\ell^K_{h\pi}(\lambda)$ for every engine state with $r_h > 0$. Proof in appendix B.3.

Thus additional engine-specific parameters help only when the publisher's adversarial response distribution exposes stable policy disagreement across engines. We call $V^K_{\text{global}}(B) - V^K_{\text{cond}}(B)$ the *conditioning gain*: a positive value is a reduction in publisher lift. Class nesting guarantees its sign only for the same fixed payoff tensor; it does not guarantee improvement on an engine left out after either class is fitted from data.

Proposition 14 (Performance and information ordering)**.** *For every bounded tensor, finite response set, and declared channel,*

$$V^K_{\text{cond}}(B) \leq V^K_{\text{global}}(B). \tag{11}$$

If K_1 is a garbling of K_2, the following ordering also holds. Proof in appendix B.3.

$$V^{K_1}_{\text{cond}}(B) \leq V^{K_2}_{\text{cond}}(B). \tag{12}$$

The first inequality is a fixed-tensor optimization opportunity, not an empirical performance guarantee. The second says that more useful pre-submission route information can only help the publisher. A singleton signal and engine revelation recover the two endpoints used in the experiments; neither is required for the theorem.

Proposition 15 (Robustness to channel misspecification)**.** *For channels on the same signal alphabet, let* $d_r(K, K') = \frac{1}{2}\sum_h r_h \sum_z |K(z \mid h) - K'(z \mid h)|$*. The following bound holds if* $L_{h\pi a} \in [-1, 1]$*.* Proof in appendix B.3.

$$\left| V^K_{\text{cond}}(B) - V^{K'}_{\text{cond}}(B) \right| \leq 2d_r(K, K'). \tag{13}$$

Thus the analysis supports sensitivity analysis around a hypothesized disclosure channel instead of requiring an exact or empirically unverified claim that routing is secret.

A.6 The gap-maximizing form of the program (Section 4.3)

The robust policy-selection program of Section 4.3 is stated here. For an audited manipulation bank B_M, let $B = B_M \cup \{0\}$ include not editing but exclude the honest control. With empirical lifts $\widehat{L}$ and the information structure fixed, the robust policy-selection program is

$$\begin{aligned} \widehat{V}^K(B) = \min_{W, \{t_z\}} \quad & \sum_z t_z & & \\ \text{s.t.} \quad & \sum_{h,\pi} r_h K(z \mid h) w_{\pi|h} \widehat{L}_{h\pi a} \leq t_z & & \forall z,\ a \in B, \\ & w_{\cdot|h} \in \Delta(\Pi) & & \forall h. \end{aligned} \tag{14}$$

Quality and cost restrict the feasible procedures:

$$Q(W) \geq Q_0 - \delta_Q, \qquad C(W) \leq C_0 + \delta_C. \tag{15}$$

Here Q contains declared quality statistics and C counts engine calls; both are linear in policy weights. These constraints make quality and cost explicit rather than assuming that deleting more text or verifying every source is acceptable.

Including the identity makes the manipulation objective nonnegative, denoted M_+. Subtracting $H(W)$ from the objective instead maximizes $H - M_+$. These are different selection rules when policies give different returns to supporting material. Compound edits use the same no-edit reference as other lifts; subtracting the support control again inside a payoff would change the objective. Appendix A.4 extends the program to uncertainty about citation-position weights.

Proposition 3 gives the finite-game solution in the main information setting. It also applies here with payoff $\Phi^K_{\widehat{L}}(W, g)$: the publisher chooses an edit rule $g : \mathcal{Z} \to B$, and the platform chooses W from a nonempty, compact, convex feasible set. The optimal dual gives one distribution over edits for each signal. The proof in appendix B.3 covers this general case as well as the alternative payoff $\Phi^K_{\widehat{L}}(W, g) - H(W)$.

This is a finite-game optimization result, not an equilibrium certificate for all publishers or unrestricted edits. The experiments assess the policy retained after renewed search through a registered comparison of two pure candidates (Section 5.2). The LP defines a separate selection rule.

The gap-maximizing form of the program. Let the response set of Equation 14 hold the identity and the manipulation responses and leave out the honest action, and measure every response, including support-plus-presentation, relative to the same no-edit control. Do not subtract the supporting passage inside a compound payoff and then subtract H again. Since $H(W)$ is linear in W,

$$\max_{W, \{t_z\}} \ H(W) - \sum_z t_z$$

under the constraints of Equation 14 and Equation 15 is a linear program, and its value is the largest feasible $H(W) - M_+(W)$, where M_+ is the manipulation loss with the identity offered at every signal. If W_M minimizes M_+ and W_Δ maximizes $H - M_+$ over the same feasible set, then

$$0 \ \leq \ [H - M_+](W_\Delta) - [H - M_+](W_M) \ \leq \ H(W_\Delta) - H(W_M),$$

because $M_+(W_\Delta) \geq M_+(W_M)$: the two selections differ in $H - M_+$ by at most their difference in honest return, and they have the same value of it when the honest return does not vary over the feasible set. With the honest action left in the response set the value of the program is never positive, and without the identity the program would reward a policy for making every edit harmful. This paragraph describes a design objective: it changes no reported quantity and no selection made in this paper, and M_+ does not replace the signed M of Equation 3 in any result.

Bank dependence of selection. Assume edits are drawn from an admissible universe $\mathcal{A}$, that for each candidate policy π the population lift $L_\pi(a)$, the mean over the market distribution, is defined for every $a \in \mathcal{A}$ and attains its supremum, and that a bank is a finite subset of $\mathcal{A}$. Write

$$M_B(\pi) = \max_{a \in B} L_\pi(a), \qquad M^\star(\pi) = \max_{a \in \mathcal{A}} L_\pi(a),$$

for the population bank value and the population worst case, and $\widehat{M}_B(\pi)$ for the estimate of $M_B(\pi)$ from a finite sample of markets; every reported number is a $\widehat{M}_B$, and the statements below are about the population quantities. Two notions of search are kept apart. A *bounded search* against π returns a finite set $S_\pi \subset \mathcal{A}$ of edits found while targeting π; it carries no guarantee about L_π on $\mathcal{A} \setminus S_\pi$. A bank has *exact response coverage* for π when it contains a population maximizer of L_π. Every search in this paper is bounded; none is known to give exact coverage.

Proposition 16 (Bank dependence)**.** *(a) $M_B(\pi) \le M^\star(\pi)$ for every policy and every bank; equality holds when B has exact coverage for π, and a bank built by bounded search gives only this floor. (b) The ranking on a bank can be the reverse of the ranking of population worst cases: there are instances with $M_B(\pi_2) < M_B(\pi_1)$ and $M^\star(\pi_2) > M^\star(\pi_1)$, including instances in which B has exact coverage for π_1. (c) If a candidate's maximizer over $\mathcal{A}$ lies outside B, no function of the values $\{L_\pi(a) : \pi$ a candidate$, a \in B\}$ determines that candidate's $M^\star$. An upper bound on $M^\star(\pi)$ therefore needs something other than bank values: exact coverage, a certified edit class that π removes (Proposition 18), a cover of the defended policy's inputs (Proposition 19), or an explicit assumption on the search such as the η-oracle of Theorem 3. Without one, every candidate holds a floor, and a comparison of bank values is a comparison of floors. (d) Enlarging a bank with edits found against π_1 weakly raises every candidate's bank value, $M_{B \cup S_{\pi_1}}(\pi) \ge M_B(\pi)$ for all π, and can raise π_2's strictly. In particular, a search against π_1 can narrow as well as widen the bank-value difference between π_1 and π_2.* Proof in appendix A.6.

Proof. (a) $B \subseteq \mathcal{A}$ gives the inequality; a maximizer in B gives equality; a bounded search returns a subset of $\mathcal{A}$ with no further property. (b) Take $\mathcal{A} = \{a_1, a_2\}$, $L_{\pi_1} = (1, 0)$, $L_{\pi_2} = (0, 2)$ on the % exposure scale, and $B = \{a_1\}$, which contains π_1's maximizer. Then $M_B(\pi_1) = 1 > 0 = M_B(\pi_2)$ while $M^\star(\pi_1) = 1 < 2 = M^\star(\pi_2)$. (c) If $a_2 \notin B$, the values on B are the same whether $L_{\pi_2}(a_2) = 0$ or 2, so they cannot separate $M^\star(\pi_2) = 0$ from $M^\star(\pi_2) = 2$; the three named routes each supply an inequality $L_\pi(a) \le v$ on edits outside B, which bank values cannot. (d) The maximum over a larger set is no smaller; for strictness take $\mathcal{A} = \{a_1, a_2\}$, $L_{\pi_1} = (1, 3)$, $L_{\pi_2} = (0, 2)$, $B = \{a_1\}$ and $S_{\pi_1} = \{a_2\}$: $M_B(\pi_2) = 0$ rises to 2, and the difference $M(\pi_1) - M(\pi_2)$ stays 1 while both values move. □

M_+ versus the residual. Assume the selection program minimizes over mixtures W of finitely many pure policies, drawn once per market, so that the honest return $H(W)$, every lift $L_W(a)$ and the quality statistic $Q(W)$ are linear in the mixture weights, and that quality is constrained by $Q(W) \ge Q_0 - \delta$. Write $M(W) = \max_{a \in B_M} L_W(a)$, $M_+(W) = \max\{0, M(W)\}$ for the objective of Equation 14, $\rho(W) = [M(W) - H(W)]_+$ for the residual, and $G(W) = H(W) - M_+(W)$ for the gap of Appendix A.6. For a pure policy π and $\theta \in [0, 1]$, the *θ-suppression* π_θ serves as π does but multiplies the treated source's exposure by θ in every cell, for instance by withholding that source with probability $1 - \theta$ independently of the edit; then $H(\pi_\theta) = \theta H(\pi)$ and $L_{\pi_\theta}(a) = \theta L_\pi(a)$ for every a.

Lemma 1 (Citation suppression under the three objectives)**.** *(a) $M_+(\pi_\theta) = \theta M_+(\pi)$. In the objective, minimizing M_+ prefers π_θ to π for every $\theta < 1$ whenever $M_+(\pi) > 0$, whatever $H(\pi)$ is, and is indifferent to any change that lowers H alone. (b) $\rho(\pi_\theta) = \theta \rho(\pi)$. Minimizing the residual prefers π_θ while $\rho(\pi) > 0$, is indifferent once $M(\pi) \le H(\pi)$, and never prefers a change that lowers H alone, since ρ is nonincreasing in H. (c) $G(\pi_\theta) = \theta G(\pi)$, so maximizing the gap opposes suppression whenever $G(\pi) > 0$. (d) Let π be feasible with $Q(\pi) > Q_0 - \delta$ and let σ be a pure policy with $Q(\sigma) < Q_0 - \delta$ and $M_+(\sigma) < M_+(\pi)$. The mixture $W_w = (1 - w)\pi + w\sigma$ is feasible for $0 \le w \le \bar{w} = \frac{Q(\pi) - Q_0 + \delta}{Q(\pi) - Q(\sigma)} \in (0, 1)$ and $M_+(W_w) < M_+(\pi)$ for every such $w > 0$. So π is not M_+-optimal: some feasible mixture with positive weight on a component that is infeasible on its own does strictly better. Whether the margin binds at the optimum is a property of the fit, not of the lemma.* Proof in appendix A.6.

Proof. Scaling every lift by $\theta \geq 0$ scales the maximum, the positive part and the difference with H by θ, which gives the identities in (a) to (c); the preference statements read off the sign of $M_+(\pi)$, $\rho(\pi)$ and $G(\pi)$. In (b), lowering H with M fixed weakly raises $[M - H]_+$. For (d), $Q(W_w) = (1-w)Q(\pi) + wQ(\sigma) \geq Q_0 - \delta$ exactly when $w \leq \bar{w}$, and $\bar{w} \in (0,1)$ because $Q(\sigma) < Q_0 - \delta < Q(\pi)$. Since $M(W)$ is a maximum of linear functions of W it is convex, and so is M_+; hence $M_+(W_w) \leq (1-w)M_+(\pi) + wM_+(\sigma) < M_+(\pi)$ for $w > 0$. □

Application to the selection objective. The quality constraint is linear in policy weights, so a feasible mixture may include a policy that is infeasible alone. Both M_+ and a positive residual can reward citation suppression; the gap opposes it when the gap is positive. Reporting the supporting-material return alongside manipulation's return distinguishes suppression from selective reductions. The development fit and the registered candidate comparison use different objectives and must not be treated as the same selection experiment (Appendix D.7).

A.7 Response expansion: the loop and its guarantee (Section 4.3)

Response expansion follows empirical game analysis (McMahan et al., 2003; Lanctot et al., 2017). Proposed edits pass the admissibility audit before evaluation.

Algorithm 1. Response-expanded commitment

Input: Game data $(\mathcal{D}_{\text{cal}}, \mathcal{H}, \Pi, r, K)$; screened response set $B_0 \ni 0$; tolerance τ; response oracle BR; audit Audit
Output: Commitment $W_\star$, screened response set $B_\star$, restricted value $\widehat{V}_\star$

$t \leftarrow 0$
while true **do**
  Estimate every missing $\widehat{L}_{h\pi a}$ for $a \in B_t$ on $\mathcal{D}_{\text{cal}}$
  $(W_t, \widehat{V}_t) \leftarrow \mathsf{SolveLP}(\widehat{L}, B_t, r, K)$ by Equation 14, or Equation 6 for ordered exposure, under Equation 15
  Commit to the procedure $\mathcal{P}_t = (r, K, W_t)$
  foreach $z \in \mathcal{Z}$ **do**
    $\widetilde{a}_z \leftarrow \mathsf{BR}(\mathcal{P}_t, z)$ with admissible feedback only; $(g_t(z), \text{valid}_z) \leftarrow \mathsf{Audit}(\widetilde{a}_z)$; **if** $\neg\text{valid}_z$ **then** $g_t(z) \leftarrow 0$
  Estimate missing lifts for the audited responses $\{g_t(z) : z \in \mathcal{Z}\}$ on $\mathcal{D}_{\text{cal}}$
  if $\Phi^K_{\widehat{L}}(W_t, g_t) \leq \widehat{V}_t + \tau$ or $\{g_t(z)\} \subseteq B_t$ **then return** $(W_t, B_t, \widehat{V}_t)$
  $B_{t+1} \leftarrow B_t \cup \{g_t(z) : z \in \mathcal{Z}\}$; $t \leftarrow t+1$

The loop has the following conditional guarantee. Let $\mathcal{A}_{\text{valid}}$ be a finite audited universe of responses and $V^K(\mathcal{A}_{\text{valid}})$ the smallest manipulation loss any commitment reaches on it. If lifts are estimated within ϵ uniformly over $\mathcal{A}_{\text{valid}}$, the program is solved within ζ, the oracle returns a plan within η of the best response, and the expansion loop stops when the returned plan is within τ of its empirical restricted value, then the returned commitment satisfies $\mathcal{R}^K_L(W_t; \mathcal{A}_{\text{valid}}) \leq V^K(\mathcal{A}_{\text{valid}}) + \eta + \tau + \zeta + 2\epsilon$ (Theorem 3, with the ordered-exposure corollary and a paired certificate for committed procedures). The empirical analysis uses paired market-clustered bootstrap intervals; the theorem below concerns the stated oracle and estimation bounds.

Assume market-level paired lifts lie in $[-1, 1]$ and the response set contains the zero outside option. The language-oracle and statistical guarantees have separate premises. Finite-sample confidence bounds require independent markets and a response class fixed independently of those outcomes, or a valid uniform bound over a fixed containing class. They do not automatically establish coverage of the reported bootstrap intervals at tied maxima or clipping boundaries.

Theorem 3 (End-to-end response-expansion guarantee)**.** *Let $\mathcal{A}_{\text{valid}}$ be a finite audited response universe and $\|\widehat{L} - L\|_{\infty, \mathcal{A}_{\text{valid}}} \leq \epsilon$. Suppose the empirical LP is solved within ζ, returning a feasible procedure and its epigraph objective $\widehat{V}_t$. An additive η-oracle must return a plan g_t satisfying $\Phi^K_{\widehat{L}}(W_t, g_t) \geq \mathcal{R}^K_{\widehat{L}}(W_t; \mathcal{A}_{\text{valid}}) - \eta$; suppose Algorithm 1 stops when this plan is within τ of its empirical restricted value. Then*

$$\mathcal{R}^K_L(W_t; \mathcal{A}_{\text{valid}}) \leq V^K_{\text{cond}}(\mathcal{A}_{\text{valid}}) + \eta + \tau + \zeta + 2\epsilon. \tag{16}$$

With exact payoffs and optimization, this reduces to the population $\eta+\tau$ guarantee; an exact response oracle also sets $\eta = 0$. An exact oracle over finite $\mathcal{A}_{\text{valid}}$ terminates after at most $|\mathcal{A}_{\text{valid}} \setminus B_0|$ additions. Proof in appendix B.3.

Corollary 1 (Uniform ordered-exposure guarantee)**.** *Apply the preceding theorem to the prefix lifts in Equation 1, with response search performed against the worst pair (a, k). Then the returned commitment satisfies*

$$\mathcal{R}^K_{\text{pos},L}(W_t; \mathcal{A}_{\text{valid}}) \le V^K_{\text{pos}}(\mathcal{A}_{\text{valid}}) + \eta + \tau + \zeta + 2\epsilon, \tag{17}$$

for the signal-wise prefix-envelope objective of Equation 6. This objective upper-bounds every shared weakly decreasing normalized attention curve; independent engine-specific curves require a different uncertainty set. Exact evaluation of k adds no oracle error; only search over language responses contributes to η. Proof in appendix B.3.

Theorem 4 (Paired certificate for committed robust procedures)**.** *Fix comparator W^0, candidate W^1, and response plans $\mathcal{G}$ before evaluation. For market q, let $X^j_q(g)$ be signed publisher lift under procedure j and response plan g, after averaging that procedure's engine and policy randomization. Set $D_q(g, g') = X^0_q(g) - X^1_q(g')$, so positive values favor the candidate. Then the population reduction in worst-response loss is exactly*

$$\Gamma_{0\to1} = \max_g \min_{g'} \mathbb{E}[D_q(g, g')]. \tag{18}$$

Consequently, simultaneous paired lower confidence bounds $\underline{d}(g, g') \le \mathbb{E}[D_q(g, g')]$ certify $\Gamma_{0\to1} \ge \max_g \min_{g'} \underline{d}(g, g')$. Proof in appendix B.3.

The certificate preserves within-market covariance and directly targets the difference of two max-over-response objectives. A paired empirical-Bernstein instantiation is given with the proofs. The coarse tensor-union bound is too wide to certify the observed effects; it provides a consistency argument, not an empirical guarantee for the evaluated policy.

The primary approximation is additive because the optimum may be zero. Let $\xi = \eta + \tau + \zeta + 2\epsilon$ denote the total additive error. If the oracle guarantees a $(1-\alpha)$ fraction of the strongest response and termination uses relative tolerance ρ, the conditional algorithmic ratio is $(1+\rho)/(1-\alpha)$. We do not assume this property for language-model search. A certified positive floor $\nu \le V^K_{\text{cond}}$ alternatively converts additive error ξ to ratio $1 + \xi/\nu$.

For fixed finite B and independent identically distributed markets, if $m \to \infty$ and $\zeta_m \to 0$, the restricted excess loss converges to zero. For an adaptively expanded B_m, let $\mathcal{U}_m$ be a finite containing response class fixed independently of the evaluation outcomes, with $B_m \subseteq \mathcal{U}_m \subseteq \mathcal{A}_{\text{valid}}$. Extending the claim additionally requires the coverage condition

$$\kappa_m = \sup_{W \in \mathcal{W}_{\text{cond}}} \left[\mathcal{R}^K_L(W; \mathcal{A}_{\text{valid}}) - \mathcal{R}^K_L(W; B_m)\right] \to 0 \tag{19}$$

and $\log(HP|\mathcal{U}_m|)/m \to 0$. Under these sufficient conditions, the rate in probability is $O_p(\sqrt{\log(2HP|\mathcal{U}_m|)/m} + \zeta_m + \kappa_m)$. The realized size of an outcome-selected bank alone does not supply a uniform convergence bound. The coverage condition is not claimed for the current response generator; it names the precise debt required for unrestricted-language consistency.

The global game gives one more diagnostic. Let A be the number of response columns and let $G \in [-1, 1]^{P \times A}$ include the zero outside-response column.

Proposition 17 (Pure-saddle certificate)**.** *If $(\pi^\star, a^\star)$ satisfies*

$$G_{\pi^\star a} \le G_{\pi^\star a^\star} \le G_{\pi a^\star} \qquad \forall \pi, a, \tag{20}$$

then $\pi^\star$ is globally optimal, the game value is $G_{\pi^\star a^\star}$, and global randomization gives zero improvement over this pure policy. Proof in appendix B.3.

This proposition diagnoses a policy population; it is not the Nash-GEO algorithm or its performance guarantee.

A.8 What a response set can certify (Section 4.3)

Two consequences limit what a response set can certify, and neither is new here. For a wider universe $\mathcal{A} \supseteq B$, let $\kappa_W(\mathcal{A}, B) = \sup_{a\in\mathcal{A}} \inf_{b\in B} \Pr\{T_W(x, a) \ne T_W(x, b)\}$ be the chance that the closest

response set member is served a different input. Exposure is bounded by one and the support control cancels, so $0 \le \rho_A(W) - \rho_B(W) \le \kappa_W(\mathcal{A}, B)$: exact input coverage certifies a finite response class, whereas a search that returns nothing certifies nothing. If response set residuals are estimated uniformly over feasible policies within ϵ, set $\kappa = \sup_W \kappa_W(\mathcal{A}, B)$. If a procedure is selected within ζ of their minimum, its excess residual over the same feasible set is at most $2\epsilon + \zeta + \kappa$. These are the standard approximation arguments of empirical game analysis (Wellman, 2006; Lanctot et al., 2017) read on this risk; we measure no coverage constant for arbitrary language, so they bound what the reported numbers mean rather than adding to them.

Expanding response sets and comparing procedures. For a fixed support control H and nested response sets $B \subseteq B'$, $M_{B'} \ge M_B$ and $\Delta_{B'} \le \Delta_B$. This applies to adding compound *support-plus-presentation* actions as well as pure presentation actions; the latter alone satisfy the content-preservation premise of Proposition 4. If procedure j has expansion debt $d_j = M_{B'}(W^j) - M_B(W^j) \ge 0$, then the change in its gap advantage over procedure 0 is

$$[\Delta_{B'}(W^1) - \Delta_{B'}(W^0)] - [\Delta_B(W^1) - \Delta_B(W^0)] = d_0 - d_1.$$

Either sign is possible. This identity compares the same population, controls and endpoint; subtracting a historical binary fixed-set estimate from a later ordered-exposure searched-set estimate is not a test of this quantity. Separate omitted attacks enter a maximum and do not add their gains unless composed into a measured response. A fixed-set paired contrast remains meaningful as a serving intervention; it is not an unrestricted adaptive worst-case comparison.

Support controls, costs and selection uncertainty. With attacks fixed, expanding the honest response set cannot lower $H^* - M_B$. With both response sets expanded, $H^* - M^* = H - M_B + (H^* - H) - (M^* - M_B)$, so the reported finite-set gap bounds that fully expanded quantity in neither direction. For a fixed support/attack pair with exposure value $v > 0$, support is net-preferred exactly when its additional cost is at most $v(H - M_a)$. For action-dependent attack costs the correct comparison is $vH - c_H - \max_a(vM_a - c_a)$, not subtracting an arbitrary average cost from Δ.

For a finite response set independent of an evaluation sample and unbiased action means, convexity gives $\mathbb{E}\max_a \widehat{\ell}_a \ge \max_a \ell_a$; relative bias across procedures has no fixed sign. An independently selected action $\widehat{a}$ evaluated on fresh markets has conditional expectation $\ell_{\widehat{a}} \le \max_a \ell_a$. Cross-fitting estimates selected-strategy value, not an unbiased worst-case maximum, and splitting already-consumed data does not undo prior tuning. With market-level lifts in $[-1, 1]$ and K fixed responses on n independent markets, the elementary simultaneous radius $\sqrt{2\log(2K/\delta)/n}$ bounds all action means with probability at least $1 - \delta$ and therefore bounds their maximum. The reported paired intervals use their separately declared construction. A true expected payoff is a lower bound on a larger response supremum. A *sample* payoff needs simultaneous confidence control or an independent evaluation before being called a certified lower bound. A lower bound cannot certify a small residual. A clipped residual of zero is compatible with positive lifts that are merely below the support-control lift, and with substantial uncertainty.

Repair costs under fixed channel payoffs. If every open channel has a rule-independent payoff p_c, every subset of closures is feasible and clean cost $\alpha(T)$ is monotone, achieving nonnegative leakage at most $t \ge 0$ requires closing all channels with $p_c > t$. The minimum cost is then $\alpha(\{c : p_c > t\})$, a step function. This expression assumes fixed channel payoffs and the stated feasibility conditions.

A.9 A CERTIFICATE FOR COMPOSED EDITS

The input-invariance guarantee extends beyond a saved response bank when a parser check certifies an edit class. We give one such class for the implemented normalizer D and screen S, with $P = D \circ S$. It appends sentences, not arbitrary paraphrases, deletions or reorderings. An anchor b is either an original source or its specified support-only form. Check each anchor separately.

Use the normalizer's actual marker removal, token sets, sentence splitting and oversplit merger. Let L be its lines after marker removal, and U its units from retained lines before unit-stage deletion. Require at least one retained original at each stage before the keep-all fallback. Write $T(v)$ for the implemented informative-token set, not semantic content. For each retained nonterminal line with

nonempty tokens, and each retained unit with nonempty tokens, form a barrier

$$\Delta_v = T(v) \setminus \bigcup_{w \neq v} T(w),$$

using all originals at that same stage before deletion. These barriers are nonempty by the keep rule.

Fix a finite library of normalized, marker-free, single-line additions a_j, each parsed as one terminal unit. Require $\varnothing \neq T_j = T(a_j) \subseteq \bigcup_{u \in U} T(u)$ and stable merger boundaries at the original tail and between additions. The grammar appends any finite sequence from this library as new lines, without changing original text or order.

Proposition 18 (Composed-edit removal)**.** *If the union of added token sets contains no original barrier Δ_v in full, then $D(b^{\mathbf{a}}) = D(b)$. This also gives $P(b^{\mathbf{a}}) = P(b)$ if the screen already acts on b, or if every addition avoids query tokens absent from the original and declared support, using the screen's raw tokens and stop list.*

Assign each primitive a positive syntactic cost c_j. Let β be the minimum cost of a subset whose token union covers any barrier, with $\beta = \infty$ if none exists. Every sequence of total cost strictly below β satisfies the removal condition. These costs measure the declared edit budget, not human effort. Proof in appendix B.4.

For each barrier Δ, nonnegative weights z_t satisfying $\sum_{t \in \Delta \cap T_j} z_t \leq c_j$ for every primitive certify a cover cost of at least $\sum_{t \in \Delta} z_t$. The minimum of these lower bounds over barriers is a sufficient safe budget. Exact minimum-cover computation may be exponential; verifying these supplied weights is polynomial in the explicit library size.

The strict inequality matters. For `Alpha beta gamma.  Delta epsilon zeta.`, either `Alpha.` or `Beta gamma.` alone is removed. Together they cover the first sentence's barrier and cause that original sentence to be removed. Primitive success therefore does not imply success under composition. When it passes and other input channels are unchanged, Proposition 1 gives zero additional exposure on any engine. Source-wise equality holds regardless of rivals' edits for this grammar.

A.10 Comparing policies with a defended response cover

A comparative guarantee need not assume that search finds the strongest baseline attack. One admissible baseline response supplies a lower bound. The harder requirement is an upper bound on the defended policy's responses. The following result states that requirement without changing the measured attack bank.

Proposition 19 (One-sided policy comparison)**.** *For policies $j \in \{0, 1\}$, fix support controls h_j, response universes $\mathcal{A}_j$, and exposure $e_j(x, g) \in [0, \alpha_1]$, with expectation over the same market population and specified engine laws. Define*

$$F_j = \sup_{g \in \mathcal{A}_j} \mathbb{E}[e_j(x, g) - e_j(x, h_j)].$$

Fix a baseline response $g_\star \in \mathcal{A}_0$, a finite reference set $\mathcal{C}$ containing h_1, and a map $f : \mathcal{A}_1 \to \mathcal{C}$ between complete response rules. A fixed event $Z(x)$ covers all defended input disagreements: if $Z(x) = 0$, then $T_1(x, g) = T_1(x, f(g))$ for every $g \in \mathcal{A}_1$. Put $\eta = \alpha_1 \Pr(Z = 1)$ and

$$D_c(x) = e_0(x, g_\star) - e_0(x, h_0) - e_1(x, c) + e_1(x, h_1).$$

Both the signed-gap improvement $F_0 - F_1$ and the clipped-premium reduction $[F_0]_+ - [F_1]_+$ are at least

$$\min_{c \in \mathcal{C}} \mathbb{E}[D_c] - \eta.$$

Simultaneous lower bounds $L_c \leq \mathbb{E}[D_c]$ and an upper bound $U_\eta \geq \eta$ give the lower bound $\min_c L_c - U_\eta$, with failure probability at most the sum of the two confidence budgets. Proof in appendix B.4.

Here h_1 is an upper-cover reference, not an added historical manipulation. The map f is fixed across markets; it cannot select a convenient reference after each outcome. Independent-market inference requires freezing the witness, reference rules and audit design before evaluation, or conditioning on independent training. Applying the bound requires establishing the grammar's coverage and its corresponding η. Both propositions are proved in Appendix B.4.

B PROOFS

Proofs are grouped by the main-text section whose results they establish; the results of Appendix A are proved under the section they extend.

B.1 PROOFS FOR SECTION 4.1, APPENDIX A.1 AND APPENDIX A.2

Proof of Proposition 4. (a) Integrating a fixed output kernel against equal complete-input laws gives equal expected exposure, so a removed response has lift zero; every other member contributes its averaged lift, giving the stated maximum. We assume B_M is finite and nonempty; if every member is removed the maximum is zero. The identity is excluded from B_M by Equation 3; adding it explicitly gives $M_+(W) = \max\{0, M(W)\}$ in the no-signal case, a different objective when $M < 0$. For a signal z define its joint-probability-weighted contribution $\mu_z(a) = \sum_{h,\pi} r_h K(z \mid h) w_{\pi|h} L_{h\pi a}$. Choices of $g(z)$ separate, so $\max_g \sum_z \mu_z(g(z)) = \sum_z \max_a \mu_z(a)$. If the identity is offered at every signal the value is $\sum_z \max\{0, \max_a \mu_z(a)\}$, not in general $\max\{0, M(W)\}$. For example, with two equiprobable revealed engines and two responses with lifts $(1, -1)$ and $(-1, 1)$, every constant response averages zero but a signal-contingent response earns one; no maximum over market-averaged constant responses can replace this information-dependent value. (b) A presentation edit outside the surviving class has the same complete input as the unedited source and lift zero by (a). If the response set consists of presentation edits and R removes at least one of them, the maximum in (a) therefore ranges over the survivors' lifts together with a zero, which is exactly $\lambda_B(R)$ under the convention that reads λ_B as zero when no survivor is profitable; $\Delta = H - M$ then gives the stated identity. Without a removed member the zero is absent and M can be negative, which is why the hypothesis is stated. A leak-free policy has an empty surviving class, so $\lambda = 0$ there. (c) Whether an edit is removed is decided by whether R serves the edited and unedited sources identically, that is by the partition. What a survivor earns is $U(R(d^a)) - U(R(d))$, a function of the images rather than of the partition, and the same is true of H; two policies inducing the same partition can therefore differ in both. This is a separation of the two design decisions, not a claim that the partition determines the manipulation return. For the identity $\Delta^\star = \sup_{R \in \mathcal{F}}[H(R) - \lambda(R)]$, note that every leak-free R has $\lambda(R) = 0$, so $\Delta^\star \geq \sup_{R \text{ leak-free}} H(R)$, with equality exactly when no leaky R has $H(R) - \lambda(R)$ above that supremum. Proposition 2 gives equality for a family of policies closed under recovery, and Proposition 11 shows that the inequality can be strict for policies that only delete; which side holds for a given engine is an empirical property of the engines.

Proof of Proposition 4(d). Fix a deterministic policy and a presentation edit a, and let $S_a = \mathbf{1}\{T_W(x, a) \neq T_W(x, 0)\}$. An engine is any map from complete inputs to distributions over citation sequences. Where $S_a = 0$ the two inputs coincide, so every engine gives them the same expected exposure; where $S_a = 1$ the two exposures differ by at most α_1, the largest value of the ordered-exposure definition in Section 3.1, which is one when exposure is whether the source is cited. The lift of a is therefore at most $\alpha_1 \Pr(S_a = 1)$ on every engine, and a removed edit has zero lift on every engine with no further hypothesis. For the converse, suppose that what the engine reads determines the unedited input and, when it differs from that input, which of the served sources is the treated one: there are maps φ_0 and φ_1 with $\varphi_0(T_W(x, a)) = T_W(x, 0)$ for every context and every response in the set, the identity included, and $\varphi_1(T_W(x, a))$ the position of the treated publisher's source in the input whenever $T_W(x, a) \neq T_W(x, 0)$. The engine that, on input t, cites the source at position $\varphi_1(t)$ first when $t \neq \varphi_0(t)$ and cites no source otherwise gives the treated publisher exposure $\alpha_1 S_a$ after a and zero when unedited, so the lift of every a under it is $\alpha_1 \Pr(S_a = 1)$. These maps exist when the policy is applied source by source, rivals are unedited as in Equation 1, the query is part of the input and distinct across markets, and no edited source is served the same text as a rival: the engine recognizes the market from the query, matches the served texts to the sources of that market, and the one text that matches none is the edited source, in whatever order the sources are served and even when two of them read the same. They can fail to exist when a market is served in several source orders and an edited source is served the same text as a rival, because two contexts with different treated positions then share an input; the supremum is then strictly smaller whenever a second cited position earns less than the first, and the expression remains an upper bound. The extremal engine is a lookup table and is used only for tightness: where the maps exist, it shows that no guarantee over all engines can be smaller than α_1 times the share of markets on which an edit reaches the engine.

Proof of Proposition 6. In the finite publisher game fix for each content class a representative $\rho_i(C)$ that is available to publisher i, with the unedited source d_i representing its own class, and restrict each player's strategies to these representatives to obtain the content game. The constant is $\bar{\lambda}_R = \sup_{i,d'_i,\mathbf{d}_{-i}}[U_i(d'_i, \mathbf{d}_{-i}) - U_i(\rho_i(I(d'_i)), \mathbf{d}_{-i})]$ over every allowed rival profile and own source; the representatives are included, so $\bar{\lambda}_R \geq 0$. (a) If complete-input laws are invariant on content classes, every source is payoff-equivalent to its representative at every profile, so the full game is a payoff-duplicate extension of the content game and has the same equilibria up to duplication; a presentation edit is such a duplicate and earns exactly what its representative earns, hence nothing over it. (b) Without invariance a unilateral deviation satisfies $U_i(d'_i, \mathbf{d}^*_{-i}) \leq U_i(\rho_i(I(d'_i)), \mathbf{d}^*_{-i}) + \bar{\lambda}_R \leq U_i(\mathbf{d}^*) + \bar{\lambda}_R$; for mixed equilibria integrate this pointwise bound over rivals. No smaller constant works for every policy: in a one-player game a presentation that gains exactly $\bar{\lambda}_R$ over its representative reaches it. For a particular R the smallest admissible constant may be smaller, because the supremum can be attained at a rival profile no content equilibrium plays. (c) Existence for finite content games follows from mixed-equilibrium existence, and at the profile in which rivals are unedited the unilateral gain from a response is its market-averaged lift by Equation 1. No converse approximation, unrestricted equilibrium, or all-profile certificate follows from the measured response set alone, and $\bar{\lambda}_R$ is not the measured clean-profile, market-averaged response-set gain.

Proof of Proposition 7 and Proposition 9. Exposure is $x_i(\mathbf{a}) = \mathbb{E}\sum_r \gamma_r \mathbf{1}\{i \text{ at rank } r\}$ with position weights $\gamma_1 \geq \cdots \geq \gamma_C > 0$, and the fill $\Gamma(\mathbf{a}) = \mathbb{E}\sum_{r \leq K(\mathbf{a})} \gamma_r$ counts the positions used, so $\sum_i x_i(\mathbf{a}) = \Gamma(\mathbf{a}) \leq \sum_{r \leq C} \gamma_r$ at every profile. Subtracting this identity at two profiles that differ in publisher i's edit alone gives $x_i(a'_i, \mathbf{a}_{-i}) - x_i(\mathbf{a}) = -\sum_{j \neq i}[x_j(a'_i, \mathbf{a}_{-i}) - x_j(\mathbf{a})] + [\Gamma(a'_i, \mathbf{a}_{-i}) - \Gamma(\mathbf{a})]$, and averaging over publishers at the all-adopt and no-edit profiles gives the fill statement. Averaging the identity over publishers at the all-adopt and no-edit profiles gives the average change $\Delta\Gamma/n$, so the average moves by the fill change divided by the number of publishers, not by the fill change itself. (a) A publisher's exposure gain from a presentation edit at rival profile $\mathbf{a}_{-i}$ is at most $\kappa_i = \sup_{\mathbf{a}_{-i}}[x_i(a, \mathbf{a}_{-i}) - x_i(0, \mathbf{a}_{-i})]$; with a free no-edit option, not editing is strictly dominant when $c > \kappa_i$. A removed edit leaves the complete input unchanged at every rival profile, so $\kappa_i = 0$ and any $c > 0$ suffices; under a leak-free policy this holds for every presentation edit of every publisher, so each such edit is strictly dominated and the profile without them is the unique equilibrium of the game restricted to presentation edits, in pure and in mixed strategies, since a strictly dominated strategy has no weight at any equilibrium. (b) The bound $\kappa_i \leq \bar{\lambda}_R$ holds only when the unedited source is the relevant representative, which is the case for a presentation edit and not in general for a support edit, whose content class has a different representative; the statement is therefore restricted to presentation edits. (c) A removed edit gives identical complete inputs, hence identical citation distributions, at every profile, so no publisher's expected exposure changes; under deterministic decoding the realized citations coincide as well, and without it equality is in distribution only. A surviving edit changes the served profile by definition, and (a) of Proposition 4 constrains neither the sign nor the size of a survivor's lift, so the citation distribution can change while the edit earns zero or less.

B.2 Proofs for Section 4.2 and Appendix A.3

This subsection proves Proposition 2 and states when a recovery map exists, then proves the deletion counterexample and the two propositions on policies that can only delete. None of it certifies the committed policy, whose leakage is measured (Appendix D.3).

Proof of Proposition 2, and when a recovery map exists. Write b for a *recovered source*: the earlier source when the submission adds no supporting material, and the earlier source with the added material otherwise. The declared content map counts added supporting material as content, $I(d^{a_H}) \neq I(d)$, so adding it is not a presentation edit, and the two kinds of recovered source are never presentation variants of one another. The hypothesis is that a deterministic K, reading only what the platform observes, returns b for every submission that admissible presentation edits reach from b, and in particular for b itself. Then $R \circ K$ serves $R(b)$ for every such submission. Unedited and support-only submissions are served as under R, so with rivals unedited their complete inputs are unchanged, and with them the honest return and every quality statistic computed on that traffic, on

every engine. Measure the lift of a presentation edit against the submission it edits. A presentation edit, alone or added to supporting material, leaves the complete input unchanged at every rival profile, so, by the argument of part (a) of Proposition 4, it earns exactly what the submission it edits earns. Hence $\lambda(R \circ K) = 0$, and

$$H(R \circ K) - \lambda(R \circ K) = H(R) \;\geq\; H(R) - \lambda(R).$$

If the family $\mathcal{F}$ of policies available to the platform contains $R \circ K$ whenever it contains R, taking suprema gives

$$\sup_{R \in \mathcal{F}} [H(R) - \lambda(R)] = \sup_{R \in \mathcal{F} \text{ leak-free}} H(R),$$

because every leak-free policy has $\lambda = 0$. If the supremum is attained, as it is over a finite set of candidates or their mixtures, it is attained by a policy of the form recover, then serve; what remains is the choice of what to serve a recovered source as. The argument is pointwise: it uses no symmetry of natural text, no randomization and no model of the engine. It is a statement about $H - \lambda$, and the signed gap of Equation 3 can be larger under a policy that makes every edit in the response set harmful.

A map K with these properties exists if and only if no submission, read together with what the platform holds, is reachable by admissible presentation edits from two different recovered sources, since a function cannot return both; whether it can be computed efficiently is a separate matter. Each part of the hypothesis is substantive. The earlier copy has to predate the edits being defended against, so the result says nothing about a source first published in its edited form. K has to establish which supporting material a submission actually contains: serving a declared supporting sentence to a submission that lacks it would pay the honest return to an edit that did not make it. $R \circ K$ may serve text that the submission does not contain, so it is not a deletion policy, and Proposition 12 continues to govern policies that only delete. Every channel from the submission to the engine has to pass through K. Recovery also has a cost: if it costs at most c_K per source and R at most c on every input it can receive from K, then $R \circ K$ costs at most $c + c_K$, so closure of $\mathcal{F}$ under a cost constraint needs that much slack.

Recovery compatibility. If Q preserves every equality of P, define $F(P(x)) = Q(x)$. This is well-defined because equal P outputs have equal Q outputs. Conversely, $Q = F \circ P$ preserves every such equality. Distinct required anchor outputs cannot be assigned to the same P output. Disjoint images of the required domains are therefore necessary; they are sufficient by assigning their respective targets and leaving other outputs unchanged. This constructs F conditional on correct target domains; token containment and access to the original source do not establish those domains.

Proof of Proposition 11. Let the unedited source be (u, v), where v repeats what u says, so that deleting v is a presentation edit, and let the honest action add a supporting passage s. Take the engine that cites the source first exactly when the text it is served contains both v and s. Serving every submission unchanged gives honest return α_1; the edit earns nothing alone and loses α_1 when added to the supporting passage, so $\lambda = 0$ and $H - \lambda = \alpha_1$. A leak-free policy has to serve (u, v, s) and (u, s) identically. A policy that only deletes cannot return v to (u, s), whatever else it reads, so it serves both without v; the engine then never cites the source, and $H = 0$. The engine is built for the example. What the example establishes is a boundary: for policies that only delete, serving every presentation variant identically can cost the whole honest return, so which policy rewards supporting material most relative to manipulation cannot be decided without the engine. Proposition 2 avoids this only because $R \circ K$ can put v back.

Proof of Proposition 12. Fix a declared equivalence class $\mathcal{C}$ of finite multisets of units, all with the same content c; content $I(d)$ is a declared set of claims and supporting-material items, monotone under multiset inclusion. Repeated renderings of the same item do not add content; adding a distinct supporting passage does. A deletion-and-ordering rule may remove units and order those retained, but cannot replace or invent units. The common core $J_{\mathcal{C}}$ is the multiset whose multiplicity for each unit is the minimum over $d \in \mathcal{C}$. An invariant output k must be obtainable by deletion from every member, hence its multiplicity at every unit is at most the minimum defining $J_{\mathcal{C}}$. This proves (a). If $I(k) = c$, monotonicity and $k \subseteq J_{\mathcal{C}} \subseteq d$ imply $c \subseteq I(J_{\mathcal{C}}) \subseteq c$. Conversely, when $I(J_{\mathcal{C}}) = c$, outputting this common core in any class-determined order is a valid rule. The same argument characterizes every smaller content-preserving output, proving (b). For (c), an output common to d and a shorter

d' has length at most $|d'| < |d|$. Integrating that pointwise necessity proves the alteration lower bound. These statements characterize admissible deletion outputs; they do not claim an efficient semantic-content oracle.

Proof of Proposition 13. A leak-free policy that reads only the submitted text serves a source and each of its presentation variants identically, so it is constant on every set that presentation edits connect, which are the classes; content preservation requires the text it serves for $\mathcal{C}$ to carry the content of $\mathcal{C}$, and the policy has to produce that text from every member, so the text is some $y_{\mathcal{C}} \in F_{\mathcal{C}}$. This proves (a). An unedited source is unaltered exactly when it already reads $y_{\mathcal{C}}$, so the unaltered share is $\sum_{\mathcal{C}} \pi(\mathcal{C})\, p(y_{\mathcal{C}} \mid \mathcal{C})$, which is largest when each $y_{\mathcal{C}}$ maximizes $p(\cdot \mid \mathcal{C})$ over $F_{\mathcal{C}}$; a family that can choose the text class by class reaches this, which proves (b). A family that cannot, such as two fixed normalizers, can do worse. For (c), the identity composed with K is leak-free and content-preserving and alters no unedited source. By part (d) of Proposition 4, under its hypotheses a content-preserving policy that reads only the submitted text and under which every presentation edit earns nothing on every engine has to serve identically every text that presentation edits reach from one source of the population; when those sets are the declared classes, these are the only such policies. For deletion policies $F_{\mathcal{C}}$ consists of the content-carrying sub-multisets of the common core in a class-determined order, and Proposition 12(c) bounds the altered share from below. The statement takes the content classes as declared, as Proposition 12 does; it supplies no test of whether two texts carry the same content.

B.3 Proofs for Section 4.3 and Appendix A.4 to Appendix A.8

Proof of Equation 5. Fix a policy π and an honest return $H(\pi)$ that does not change when the edit set grows. Taking a maximum over $B_{t+1} \supseteq B_t$ cannot lower the manipulation return. Subtracting the same $H(\pi)$ and taking the positive part preserves that order, so $\rho_{B_{t+1}}(\pi) \geq \rho_{B_t}(\pi)$. Add and subtract $\rho_{B_{t+1}}(\pi_t)$ in the left side of Equation 5 to obtain the identity. Its first term is nonnegative; thus the same-set improvement is at least the observed decline between rounds. The second term can have either sign if the policy update is not an exact minimization of this premium. Unless a tested value already reaches the exposure bound, an unchanged value can coexist with a larger payoff from an unseen edit. The identity also holds on a fixed estimated payoff table. It does not establish out-of-sample improvement or confidence coverage when the search and policy choice use those same estimated outcomes.

Proof of Proposition 3. For fixed W the constraints of Equation 14 force $t_z \geq \max_{a \in B} \mu_z^W(a)$ with $\mu_z^W(a) = \sum_{h,\pi} r_h K(z \mid h) w_{\pi|h} \widehat{L}_{h\pi a}$, so at the optimum the objective is $\sum_z \max_a \mu_z^W(a) = \max_g \Phi_{\widehat{L}}^K(W, g)$, the payoff of the publisher's best response, and the program is $\min_{W \in \mathcal{W}} \max_g \Phi(W, g)$. Let the publisher mix, with one distribution ν_z over B per signal. Φ is linear in W for fixed ν and linear in ν for fixed W; $\mathcal{W}$ is a product of simplices cut by the linear quality and cost constraints, and the mixtures form a product of simplices, so both strategy sets are compact and convex. The minimax theorem then gives $\min_W \max_\nu \Phi = \max_\nu \min_W \Phi$, and every pair of optimizers is a saddle point, that is, a Nash equilibrium of the zero-sum game with these strategy sets. The dual of the linear program supplies $\nu^\star$: the multiplier on the constraint of (z, a) is $\nu_z^\star(a)$, since the dual constraint of the free variable t_z makes those multipliers sum to one over a. Equality with the simultaneous-move value is the same theorem. A pure policy π with $\delta_\pi \in \mathcal{W}$ is a minimax strategy exactly when $\max_g \Phi(\delta_\pi, g)$ equals the value, the smallest worst case any $W \in \mathcal{W}$ has; announcing π lets the publisher best-respond, so the platform pays $\max_g \Phi(\delta_\pi, g)$ minus the value. When $\mathcal{W}$ is a proper subset of the simplex, a procedure outside it may have a smaller worst case, so an equilibrium of the constrained game need not be one of the unconstrained game. For the gap program the payoff is $\Phi(W, g) - H(W)$; H depends on W alone, so the publisher's best responses are unchanged, the payoff stays linear in each argument, and the minimax argument applies to this new payoff. Since $H(W)$ varies with W, this need not preserve the minimizing policy or shift the original value by a constant. This is a statement about the program on its response set with rivals unedited, not about the game among publishers.

Proof of Theorem 1. Let $\alpha_{C+1} = 0$ and define $\lambda_k = k(\alpha_k - \alpha_{k+1})$ for $k \in [C]$. Monotonicity gives $\lambda_k \geq 0$, and telescoping gives

$$\sum_{k=1}^{C} \lambda_k = \sum_{k=1}^{C} k(\alpha_k - \alpha_{k+1}) = \sum_{j=1}^{C} \alpha_j = 1. \tag{21}$$

Let $u^{(k)} \in \mathbb{R}^C$ put mass $1/k$ on each of the first k positions and zero elsewhere. For every coordinate j,

$$\sum_{k=1}^{C} \lambda_k u_j^{(k)} = \sum_{k=j}^{C} (\alpha_k - \alpha_{k+1}) = \alpha_j, \tag{22}$$

so every $\alpha \in \mathcal{M}_C$ is a convex combination of the C uniform-prefix vectors. Conversely, each $u^{(k)}$ belongs to $\mathcal{M}_C$. A linear objective over their convex hull is maximized at one of these vectors, proving Theorem 1. Applying the identity separately to the conditional lift at each public signal z gives exactly the constraints in Equation 6. Uniform-prefix exposure lies in $[0, 1]$, so its matched lift lies in $[-1, 1]$; the later finite-set arguments therefore apply to the augmented finite index set $B \times [C]$.

Proofs of Theorem 2, Proposition 14 and Proposition 15. Finite minimax gives

$$V_{\mathrm{cond}}^K(B) = \max_{\lambda} \min_{(w_{\cdot|h})_h} \sum_{h,\pi} r_h w_{\pi|h} \ell_{h\pi}^K(\lambda) = \max_{\lambda} \sum_h r_h \min_{\pi} \ell_{h\pi}^K(\lambda), \tag{23}$$

$$V_{\mathrm{global}}^K(B) = \max_{\lambda} \min_{w} \sum_{h,\pi} r_h w_\pi \ell_{h\pi}^K(\lambda) = \max_{\lambda} \min_{\pi} \sum_h r_h \ell_{h\pi}^K(\lambda). \tag{24}$$

Only the marginal policy distribution in each realized engine state affects payoff, so product-form conditioned mixtures lose no generality. For any $\lambda_c \in \Lambda_c$ and $\lambda_g \in \Lambda_g$,

$$V_{\mathrm{global}}^K - V_{\mathrm{cond}}^K \geq g_K(\lambda_c) - f_K(\lambda_c) = D_K(\lambda_c), \tag{25}$$

$$V_{\mathrm{global}}^K - V_{\mathrm{cond}}^K \leq g_K(\lambda_g) - f_K(\lambda_g) = D_K(\lambda_g). \tag{26}$$

Taking the stated maximum and minimum proves Equation 10. Equality between a minimum of weighted sums and the corresponding weighted sum of minima holds exactly when one policy minimizes every positive-weight engine term. A common minimizer at a global-dual optimum therefore forces zero gain. Conversely, if the gain is zero, every conditioned-dual optimum also maximizes the global dual and has zero disagreement.

The global class is a subset of the conditioned class, so minimizing the same risk over the latter proves Equation 11. For information ordering, fix W and let g_1 be an optimal response rule under K_1. If $K_1 = G \circ K_2$, a publisher observing z_2 can sample $z_1 \sim G(\cdot \mid z_2)$ and apply $g_1(z_1)$. This randomized rule under K_2 has exactly the payoff of g_1 under K_1. A deterministic best response under K_2 weakly improves on that rule because the payoff is linear in action mixtures. Hence $\mathcal{R}_L^{K_2}(W; B) \geq \mathcal{R}_L^{K_1}(W; B)$ for every W; minimizing both sides proves Equation 12.

For channel perturbation, any fixed (W, g) satisfies

$$|\Phi_L^K(W, g) - \Phi_L^{K'}(W, g)| \leq \sum_{h,z} r_h |K(z \mid h) - K'(z \mid h)| = 2d_r(K, K'), \tag{27}$$

where policy weights sum to one and $|L_{h\pi a}| \leq 1$. Taking maxima over g and then minima over W preserves this bound, proving Equation 13.

Proof of Theorem 3. Uniform tensor error implies payoff error at most ϵ for every mixed commitment and valid response plan. Let $\widehat{V}_t$ be the objective returned by a solver whose value is at most ζ above the exact empirical restricted optimum, with a feasible epigraph vector so $\widehat{V}_t$ upper-bounds the restricted risk of W_t. Write $\widehat{V}^*(B_t)$ for the exact empirical restricted optimum. The oracle and

stopping conditions give

$$\mathcal{R}_L^K(W_t;\mathcal{A}_{\text{valid}}) \le \mathcal{R}_{\widehat{L}}^K(W_t;\mathcal{A}_{\text{valid}}) + \epsilon \tag{28}$$
$$\le \Phi_{\widehat{L}}^K(W_t, g_t) + \eta + \epsilon \tag{29}$$
$$\le \widehat{V}_t + \tau + \eta + \epsilon \tag{30}$$
$$\le \widehat{V}^*(B_t) + \zeta + \tau + \eta + \epsilon \tag{31}$$
$$\le V_{\text{cond}}^K(B_t) + \zeta + \tau + \eta + 2\epsilon \tag{32}$$
$$\le V_{\text{cond}}^K(\mathcal{A}_{\text{valid}}) + \zeta + \tau + \eta + 2\epsilon. \tag{33}$$

The last line uses $B_t \subseteq \mathcal{A}_{\text{valid}}$. With an exact oracle and finite valid universe, every nonterminal plan contains a response outside B_t, so adding all new components permits at most $|\mathcal{A}_{\text{valid}} \setminus B_0|$ nonterminal iterations. Audit false rejection contributes to η; false acceptance violates the valid-universe premise and must be controlled separately.

For the conditional multiplicative statement with exact payoffs and optimization, multiplicative oracle accuracy and relative termination imply

$$(1-\alpha)\mathcal{R}_L^K(W_t;\mathcal{A}_{\text{valid}}) \le \Phi_L^K(W_t, g_t) \le (1+\rho)V_{\text{cond}}^K(B_t) \le (1+\rho)V_{\text{cond}}^K(\mathcal{A}_{\text{valid}}). \tag{34}$$

Division by $1-\alpha$ gives the ratio $(1+\rho)/(1-\alpha)$. An additive oracle alone cannot provide a nontrivial universal ratio when the optimum can approach zero. If a certified response instead supplies $0 < \nu \le V_{\text{cond}}^K(\mathcal{A}_{\text{valid}})$, dividing the additive guarantee by the optimum gives $\mathcal{R}_L^K(W_t;\mathcal{A}_{\text{valid}})/V_{\text{cond}}^K(\mathcal{A}_{\text{valid}}) \le 1 + \xi/\nu$.

Proof of Theorem 4, and the coarse uniform bound. For fixed procedures,

$$\mathcal{R}_L^K(W^0;B) - \mathcal{R}_L^K(W^1;B) = \max_g \mathbb{E}[X_q^0(g)] - \max_{g'} \mathbb{E}[X_q^1(g')] \tag{35}$$
$$= \max_g \min_{g'} \mathbb{E}[D_q(g,g')], \tag{36}$$

which proves Equation 18. Monotonicity of $\max_g \min_{g'}$ proves the lower-certificate claim for any simultaneous bounds. For a distribution-free instantiation on $m \ge 2$ independent identically distributed markets, with procedures and finite response plans fixed before evaluation, let $\overline{D}(g,g')$ and $S^2(g,g')$ be the sample mean and unbiased variance. Since $D_q(g,g') \in [-2,2]$, the empirical Bernstein inequality of Maurer & Pontil (2009) and a union bound give

$$\underline{d}(g,g') = \overline{D}(g,g') - \sqrt{\frac{2S^2(g,g')\log(2|\mathcal{G}|^2/\delta)}{m}} - \frac{28\log(2|\mathcal{G}|^2/\delta)}{3(m-1)} \tag{37}$$

simultaneously for all response-plan pairs with probability at least $1-\delta$.

For completeness, the coarser training-sample argument proceeds as follows. On the event $\|\widehat{L} - L\|_\infty \le \epsilon$, the weights in Equation 2 sum to one for every response rule, so

$$|\mathcal{R}_{\widehat{L}}^K(W;B) - \mathcal{R}_L^K(W;B)| \le \epsilon \tag{38}$$

uniformly over W. Hence $|\widehat{V}_{\text{cond}}^K(B) - V_{\text{cond}}^K(B)| \le \epsilon$. Combining the two deviations with empirical ζ-suboptimality for the fitted commitment $\widehat{W}$ gives

$$\mathcal{R}_L^K(\widehat{W};B) \le V_{\text{cond}}^K(B) + 2\epsilon + \zeta. \tag{39}$$

Hoeffding's inequality for variables with range two and a union bound over $HP|B|$ market-level means give

$$\epsilon = \sqrt{\frac{2\log(2HP|B|/\delta)}{m}} \tag{40}$$

with probability at least $1-\delta$.

Consistency, Proposition 17 and the cost frontier. For fixed finite B, uniform convergence of its bounded empirical means makes $\epsilon_m \to 0$ almost surely. If $\zeta_m \to 0$, the population excess loss therefore converges to zero. For expanding B_m, add Equation 19 to the preceding bound and use $V^K_{\mathrm{cond}}(B_m) \le V^K_{\mathrm{cond}}(\mathcal{A}_{\mathrm{valid}})$, yielding the stated asymptotic rate.

Finally, under Equation 20, policy $\pi^\star$ upper-bounds the global game by $G_{\pi^\star a^\star}$, while response $a^\star$ lower-bounds every pure policy and hence every mixture by the same value. The bounds coincide, proving optimality and zero value from global randomization.

For the cost frontier, increasing b enlarges the feasible set and cannot increase the minimum. Let W_1 and W_2 be optimal at cost budgets b_1 and b_2. The commitment $\lambda W_1 + (1-\lambda)W_2$ is feasible at $\lambda b_1 + (1-\lambda)b_2$, and $\mathcal{R}^K_L$ is convex because it is a maximum of affine functions. Therefore

$$V(\lambda b_1 + (1-\lambda)b_2) \le \lambda V(b_1) + (1-\lambda)V(b_2), \tag{41}$$

which proves convexity.

B.4 Proofs for the policy certificates

Proof of Proposition 18. At the line stage each terminal addition is kept. Originally dropped lines stay dropped: their body and list predicates do not change, and adding tokens elsewhere cannot undo subsumption. Kept terminal lines remain exempt from line-stage subsumption. A kept nonterminal line with nonempty tokens retains a barrier token absent from all other original lines and additions, so it stays kept. Empty-token lines cannot trigger the nonempty-token subsumption rule. The original retained set is unchanged and nonempty, preventing fallback.

Stable parsing boundaries therefore give the original unit sequence followed by the added units. The same barrier argument preserves each retained original unit, and each originally dropped unit stays dropped. Every added unit is subsumed by the token union of the original prefilter units, including those subsequently deleted. At least one original survives, so unit-stage fallback is also absent. The final retained sequence is unchanged, as are unquoting, deduplication and windowing. If the anchor trips the screen, appending cannot remove a triggering raw token, and all submissions revert to the same original. Otherwise the raw-token condition prevents any addition from tripping it. This proves complete-policy equality in both cases.

Repeated primitives add cost but do not enlarge the token union. Any sequence covering a barrier contains a subset covering it at no greater cost, proving the strict budget bound. For a subset J covering barrier Δ, summing the dual inequalities gives

$$\sum_{j\in J} c_j \ge \sum_{j\in J} \sum_{t\in\Delta\cap T_j} z_t \ge \sum_{t\in\Delta} z_t.$$

The minimum over barrier-specific lower bounds is therefore safe. The stated two-addition example reaches the threshold, so equality with the threshold is not generally sufficient.

Proof of Proposition 19. Let $x_\star = \mathbb{E}[e_0(g_\star) - e_0(h_0)]$ and $F_{\mathcal{C}} = \max_{c\in\mathcal{C}} \mathbb{E}[e_1(c) - e_1(h_1)] \ge 0$, because $h_1 \in \mathcal{C}$. Complete mediation gives equal conditional exposure on $Z = 0$. The exposure range bounds the remaining difference by α_1, hence $F_1 \le F_{\mathcal{C}} + \eta$ and $F_0 \ge x_\star$. Also $\min_c \mathbb{E}D_c = x_\star - F_{\mathcal{C}}$. Subtraction proves the signed bound. For clipping, $[F_0]_+ \ge x_\star$ and $[F_1]_+ \le F_{\mathcal{C}} + \eta$ because the latter upper bound is nonnegative. The same lower bound follows. Simultaneous confidence statements hold jointly except on the union of their failure events; no independence between those two events is needed.

C Detailed Experimental Setup

This appendix separates the primary evaluation from the development studies and later comparisons. Sections C.1–C.3 describe the questions, policies and edits; Section C.4 gives the generation settings; Sections C.5 and C.6 record the statistical procedures and reproducibility files.

Study guide. The primary evaluation uses 200 previously unused ASQA questions and four engines (Section D.1). The wider edit sets use the same questions (Table 5). Later comparisons test whitespace

normalization (Section D.10) and a larger searched edit set (Section D.9); these are exploratory extensions, not replacements for the registered tests. Competition studies use separate development questions (Section D.5), and transfer and larger-model studies retain their own populations and endpoints (Section D.7).

Common accounting. A task specifies an engine, policy, edit, source order, question and focal publisher. Payoffs use $v_i = 1$ and $c_i \equiv 0$: they measure citation exposure, without editing costs. Parser failures count as no citation. Analyses requiring complete data check every declared task; analyses allowing missing tasks report bounds over their possible outcomes. Edit validity, delivery and source support are recorded separately from citation outcomes. Policies selected on development questions are held fixed on the evaluation questions.

C.1 Engines, market sets and the prospective instrument

Engines. The four open-weight engines are Llama-3.1-8B, Qwen2.5-7B, Qwen3-8B and Tulu3-8B. Two-engine studies name their pair: pair A is Qwen2.5-7B with Tulu3-8B; pair B is Qwen3-8B with Llama-3.1-8B. Tulu and Llama share the pretrained Llama-3.1-8B base but use different post-training recipes and native chat templates.

Primary evaluation. The 200 ASQA questions were selected without reading model outcomes. The registered eligibility rule retained previously unused questions with five sources (948 source questions, 338 clean, 214 eligible, 200 retained; manifest SHA-256 `9c5b3855...`). The contract was fixed on 2026-09-13 before generation. Each source is edited in turn while the other four remain unchanged; Section C.3 gives the matching and averaging rules. At citation capacity three, the primary tests ask whether supporting material out-earns pure manipulation, whether removing the defense raises the manipulation advantage, and whether the remaining advantage is below 5% on every engine. Source-quality noninferiority and the wider edit sets have separate verdicts. Later exploratory analyses do not change these registered tests.

Development studies. D140 names the registered 140-market development set; R300 is its 300-market same-wave extension, not an independent replication. P440 pools these sets. The separate 48-market set also contains previously used development questions. The surface study contains 140 development markets, four engines, native input plus three surface normalizers, four presentation responses, and two source orders. The expanded method tensor contains 96 query-disjoint development markets, Qwen3-8B and Llama-3.1-8B-Instruct, four procedures, three attacks, one neutral control matched to each attack, the identity condition, and two source orders, for 10,752 tasks. The capability frontier uses the same 96 markets and two engines but expands the procedure set to 21, for 56,448 tasks. Results from the 96- and 140-market studies are not pooled.

The four surface responses form the presentation-only stratum of the authenticated response registry, not an exhaustive GEO attack class. A project-wide inventory found 44 response identifiers across experiments with different models, policies, or market populations; these results are not pooled.

C.2 Platform policies

Committed policy. The policy screens the focal edit, then normalizes the source text. The screen rejects an edit that adds query tokens absent from both the original source and the declared supporting sentence, excluding a fixed list of thirty-six function words. Rejection restores the original source. The declared supporting sentence is exempt from this token check; the screen does not verify its claims.

Normalization strips bullet markers and splits text into sentence units, rejoining splits introduced by line breaks. It deletes empty labeled or quoted units, units whose informative tokens appear elsewhere, and list-like units without terminal punctuation. It then removes case-insensitive duplicates and retains the first twelve units. If every unit would be deleted, it returns the source unchanged. These rules apply to the text passed by the screen.

The expanded method tensor uses four one-call procedures. *Native* passes submitted sources unchanged. *Canonical support* removes structural lines and duplicate sentences, then asks the engine to select direct factual support rather than presentation. *Query-order projection* ranks literal source

sentences by query overlap and supplies at most three relevant spans per source. *Dual-view projection* supplies canonical full support for answer generation and query-ordered literal spans for citation selection in one prompt. The three attacks are *support fronting*, which prepends the most query-relevant original sentence; *prose query bridge*, which prepends derived query terms and that sentence; and *relevance sentence reorder*, which sorts all original sentences by query overlap. Each attack has a length- and word-matched neutral control. Controls and identity are excluded from the max-over-attacks objective.

The 21-policy frontier comprises four one-call anchors (native, canonical support, support projection, and query-order projection), four randomized checking policies, two verify-all policies, three instability-gated policies, and eight score-aware support-contraction policies. A checking policy uses one call to score literal support and, when activated, one call to answer from the restricted view; checker failure falls back to the corresponding canonical view. Random checks activate with probability 0.07, 0.14, or 0.28 independently of the publisher action; verify-all always uses two calls; gates use one or two calls; score-aware contraction uses two. The primary inference resamples markets and recomputes the complete max-over-responses objective.

The method comparison fits a global commitment shared by all engines; engine-conditioned and randomized commitments are ablations. The primary channel reveals no engine-specific signal before source submission; full engine revelation is the opposite endpoint. Every internal split refits all compared decision rules on training markets and applies them unchanged to the folds left out. Pooled games weight complete markets and engines equally and average the two source orders within each market-engine task.

C.3 Publisher responses and the admissibility contract

The surface-tensor responses comprise bullets, section formatting, sentence order, and model-generated paraphrase. Its *query-order normalizer* removes structural material and reorders all retained segments by query overlap; unlike expanded-game query-order projection, it does not select a three-span support subset.

Action-space instantiation. We test a finite set of editing strategies and add admissible proposals found by search, following GEO's strategy catalogs and automated rewrite generators (Aggarwal et al., 2024; Wu et al., 2026b;a). Only small diagnostics enumerate all edits. Table 4 lists the evaluated action families; the reported maxima concern the tested actions, not every possible rewrite.

Table 4: Publisher action families and their empirical status. Platform-side support projection is not relabeled as a publisher response, and malicious edits are not pooled with the source-preserving game.

Family	Concrete operator	Validity check	Status
Identity	Unchanged submitted source	Exact text equality	Measured baseline
Exact rearrangement	Relevance sentence reorder; sentence-order response	Same source-token or sentence multiset	Measured
Extractive emphasis	Copy the most query-relevant source sentence before the unchanged source	Added text must be an exact source span	Measured with matched control
Structural template	Bullets and sections	Source content retained; structure changed	Measured GEO-style presentation class
Derived bridge	Query terms plus an exact source sentence before the unchanged source	Source retained; derived prose separately audited	Measured with matched control
Model rewrite	Fixed paraphrase or published GEO prompt	Automatic support gate; no human validation	Supporting measured response set
Malicious edit	Prompt injection or unsupported persuasion	Separate attack validity contract	Stress test only

Every materialized action records its operator identifier and version, permitted inputs, observed signal, model and prompt when applicable, decoding and query budgets, source and output hashes, validity decision, and identity fallback. The response is fixed before target outcomes are measured. A

Table 5: The nested edit sets used on the prospective markets: two searched edits, nine content-preserving edits and eleven expanded edits. Each reported maximum refers to its named set, not unrestricted publisher behavior. Supporting-material addition is the reference action; support removal is the construct control.

response set	size	responses	in the admissibility contract
searched (oracle)	2	searched presentation response; searched presentation stacked on the supporting passage	yes
admissible	9	the searched response set, plus label built from the document's own words; supporting passage carrying a label; bulleted restructuring; section headings; sentence reordering; label, separator form; label, bracket form	yes
unrestricted	11	the admissible response set, plus fabricated near-copy, inside the contract; fabricated near-copy, outside the contract	no; the out-of-contract near-copy is outside it

failed audit maps to identity and remains in intention-to-treat accounting; it is not silently dropped or regenerated until accepted. Published GEO methods are run through their documented prompts or released implementations where available, without tuning them to the Nash-GEO payoff tensor. GEO strategies that add quotations, statistics, or citations enter the source-preserving response set only when the added material has authenticated provenance; otherwise they remain in a broader utility-changing benchmark.

Sources and supporting material. Each question uses the first five retrieved documents in its manifest. The supporting-material control appends one question-relevant sentence from the other retrieved documents. The sentence is shared across the question's five sources and selected without reading gold answers. Retrieval and query overlap select this passage; they do not independently verify that it supports each source's claims.

Text construction. Source text comes from the ASQA item's GTR retrieval. Whitespace is collapsed and each source is limited to 1,800 characters, ending at the last sentence boundary if it lies beyond 65% of that limit. Candidate supporting sentences are split after a period, exclamation mark or question mark followed by a capital letter, digit or quote. They are ranked by the number of distinct question tokens they contain, using lower-case alphanumeric tokens of at least three characters. Zero-overlap sentences are excluded. Ties favor retrieval score, then the earlier document, then the earlier sentence. The control keeps the source unchanged and appends a blank line followed by `Retrieved supporting passage [doc_id] title: sentence`. This reconstruction matches all 1,000 recorded supporting-material submissions in the primary evaluation.

Matched comparisons. Each task edits one focal source and leaves its four rivals unchanged. Every edit is applied to each source in turn, with two seeded source orders and display identifiers assigned by order. Normalization acts on all five sources; screening acts on the focal edit. A rejected edit receives its matched no-edit outcome. For each engine, question, focal source and order, we compare the edit's citation exposure with the supporting-material control. Both use the same no-edit reference. We average over focal sources and orders within a question, then over engines for pooled results. The primary endpoint weights the first three citation positions by reciprocal rank.

A worked example. Figure 3 shows one prospective market in full: the question, the treated source, the supporting passage, both searched edits verbatim, the served text under no defense and under the committed policy, and every engine's cited answer. The market was fixed by an input-only rule before any answer was read.

C.4 Generation and transport

Current ordered-citation evaluation. The primary evaluation generates one answer per task with a 384-token output limit and greedy decoding (temperature 0; top-p 0.9 is inactive). The frozen parser, `inline_projection_citations_v10`, ranks sources by their first inline citation. The same

Question. Who performed at the champions league final 2018?

Target source [E], *2017 UEFA Champions League Final* (verbatim): 2017 UEFA Champions League Final The 2017 UEFA Champions League Final was the final match of the 2016–17 UEFA Champions League, the 62nd season of Europe's premier club football tournament organised by UEFA, and the 25th season since it was renamed from the European Champion Clubs' Cup to the UEFA Champions League. It was played at the Millennium Stadium in Cardiff, Wales on 3 June 2017, between Italian side Juventus and Spanish side and title holders Real Madrid, in a repeat of the 1998 final. Real Madrid won the match 4–1 to secure their 12th title in this competition. With

Supporting passage (the honest edit appends it): Retrieved supporting passage [ev] 2Cellos: 2018, 2Cellos performed the UEFA Champions League Anthem ahead of the 2018 UEFA Champions League Final between Real Madrid C.F. and Liverpool F.C.

Searched edit 1 (grounded label): the source, verbatim, preceded by the label "`champions, league, final:`".

Searched edit 2 (salience line and passage fronting): the line "`Relevant terms: the, champions, league, final`", a blank line, the supporting passage, a blank line, then the source, verbatim.

Served text. With no defense each edit is served verbatim. Under the committed policy (its deletion step; the grounded screen does not fire on either edit), edit 1 is served unchanged apart from whitespace, because the label sits inside the first sentence; edit 2 loses its salience line, and the passage and source are served as one paragraph.

Cited answers (provider order 0; the target source is [**E**]).

	No defense	Committed policy
Edit 1		
Llama-3.1-8B	The 2018 UEFA Champions League Final was performed by 2Cellos [B].	Andrea Bocelli [A] performed the UEFA Champions League Anthem, and 2Cellos [B] played the instrumental version of the chorus in the 2018 final.
Qwen2.5-7B	Jamaican rapper Sean Paul and Billie Eilish performed at the opening ceremony preceding the final [D]. *(identical under both policies)*	
Qwen3-8B	Andrea Bocelli performed at the Champions League Final 2018 [B] and also performed the UEFA Champions League Anthem at the 2016 UEFA Champions League Final [C]. The 2018 UEFA Champions League Final was held in Kiev, where the instrumental version of the chorus was played by 2Cellos [B]. Dua Lipa performed at the opening ceremony preceding the 2018 UEFA Champions League Final [D].	Andrea Bocelli performed at the Champions League Final 2018 [B] and also performed the UEFA Champions League Anthem at the 2016 UEFA Champions League Final [C]. The 2018 UEFA Champions League Final was held in Kiev, where the instrumental version of the chorus was played by 2Cellos [B]. Lipa performed at the opening ceremony preceding the 2018 UEFA Champions League Final [D].
Tulu-3-8B	Andrea Bocelli [A][B][C] performed the UEFA Champions League Anthem at the 2018 final. *(identical under both policies)*	
Edit 2		
Llama-3.1-8B	2Cellos [**E**] performed the UEFA Champions League Anthem ahead of the 2018 UEFA Champions League Final between Real Madrid C.F. and Liverpool F.C. [**E**]	Real Madrid won the 2017 UEFA Champions League Final, with 2Cellos performing the UEFA Champions League Anthem ahead of the 2018 UEFA Champions League Final [**E**][B].
Qwen2.5-7B	Jamaican rapper Sean Paul and Billie Eilish performed at the opening ceremony preceding the final. [D] *(identical under both policies)*	
Qwen3-8B	The UEFA Champions League Anthem at the 2018 final was performed by 2Cellos [B]. Additionally, Dua Lipa performed at the opening ceremony preceding the final [D].	The UEFA Champions League Anthem at the 2018 final was performed by 2Cellos [B]. Additionally, the 2018 UEFA Champions League Final was held in Kiev, where the instrumental version of the chorus was played by 2Cellos [D].
Tulu-3-8B	Andrea Bocelli [A][B][C] performed the UEFA Champions League Anthem at the 2018 final.	Andrea Bocelli [A][C] performed the UEFA Champions League Anthem at the 2018 final.

Figure 3: One illustrative market from the 200 prospective ASQA markets. Market 1 and its target source were selected from inputs before answers were inspected. Source IDs match the engine's anonymized display IDs; [ev] replaces the supporting document's ID. This single source order illustrates the procedure, not the aggregate effect.

answer supplies citation capacities one, three and five; capacity three is primary. Unused positions receive zero exposure, so fewer than three citations are allowed. Disagreement between inline order and a structured citation list is rejected. The frozen configuration and analysis receipt are identified in Section C.6.

Earlier single-citation studies. The four-engine surface design uses citation budget one, at most five publishers, two source orders, one sample per task, and seed 824631. Citation generation uses greedy decoding (temperature 0; top-p inactive) and at most 256 new tokens. The strict transport requires exactly `SELECTED_PROVIDER_ID: <id>` followed by `ANSWER: <nonempty text>`; malformed, missing, or multiple selections map to the null provider. Model-generated paraphrases use temperature 0 with one repair attempt at temperature 0.2, top-$p = 0.9$, and at most 768 new tokens. The rewrite prompt requires preservation of every factual claim, qualification, entity, number, and provenance while prohibiting additions and omissions.

The expanded 96-market tensor also uses five publishers, citation budget one, one sample per task, paired deterministic source orders (seed 20260721), greedy decoding (top-p inactive), and at most 384 new tokens. Its base seed is 2026072100. All four expanded procedures use the same strict selected-provider/answer transport; the three attacks and matched controls are the deterministic transformations defined above. These single-citation outputs are not reinterpreted as ordered multi-citation answers.

C.5 Endpoints, units and inference

Endpoints. Exposure levels and their differences are reported on an absolute percentage scale, not as relative improvements. Survival and clean-change figures are rates; the static source-quality score uses its own $[0, 1]$ scale. Tables and figures use consistent half-up rounding from six-decimal source values. Label-channel results use the binary citation endpoint; the prospective, ordered-exposure, final-search, utility and universal-adoption results use reciprocal-rank exposure at citation capacity three, where capacities one and five are prefixes of one generated answer.

Uncertainty. Resampling keeps policy comparisons paired and treats questions, not individual outputs, as independent units. The primary evaluation uses 5,000 resamples. The largest engine-level p-value enters Holm correction over the support-advantage and premium-reduction hypotheses. Source-quality noninferiority has a separate four-engine family. Intervals are two-sided unless stated otherwise, and each result identifies pooled or per-engine analysis. Table 23 retains the original interval endpoints. For contrasts involving the strongest edit, the registered interval contains seven component intervals.

The development utility contract is evaluated separately for every engine-response profile: selected-provider quality is a prespecified $[0, 1]$ static source score combining 60% answer-group coverage, 25% normalized retriever rank, 10% a support indicator, and 5% top-20 retriever consensus. It may fall at most 2.5% below canonical support; strict validity and non-abstention must each be at least 98%, and expected target-engine calls must not exceed 1.25. These automated measures evaluate development utility. Exact reordering, formatting, and support-preserving actions also have authenticated transformation contracts.

The post-training contrast uses market-clustered pointwise intervals without simultaneous correction. The multi-publisher analysis uses declared worst-completion bounds for incomplete tasks. The pure-saddle claim is an exact statement about the displayed pooled matrix.

C.6 Models, manifests and analysis receipts

Primary evaluation files. The frozen panel specification is `asqa_geo3_prospective200_v1.json` under `configs/canonical_expansion/panels/`. The submission receipt, `sync/key_results/prospective_asqa_submission.json`, pins both policy configurations and the immutable release (source SHA-256 `559ceb82...`). The analysis digest is `prospective_endpoint_digest_20260915.json` in the same results directory. Together these records identify the panel, generated tasks, source implementation and analysis used for the primary results.

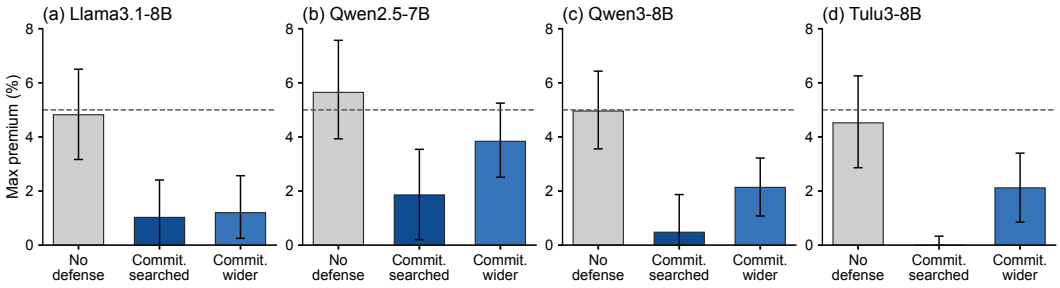

Figure 4: Prospective ASQA by engine, $C = 3$: no defense and the committed policy's searched and wider-bank premiums, with original 95% intervals. No-defense levels coincide on these banks and are descriptive. Dashed lines mark the registered 5% residual bound.

Development receipts. The authenticated snapshots are Qwen3-8B (`b968826d`), Qwen2.5-7B-Instruct (`a09a3545`), Llama-3.1-8B-Instruct (`0e9e39f2`), and Llama-3.1-Tulu-3-8B (`66694379`), loaded offline in bfloat16. The 140-market development manifest has SHA-256 `4c5b990a...e755311`; the former 300-market reserve, subsequently consumed by the pre-registered development extension, has `a926af6e...fdc5c03`. The four-engine configuration digest is `3f676e57...a1a47`; the expected-task specification is `398e229f...fb99`; the policy-response analysis is `717fb95a...9ba9`; and the corrected internal validation output is `02edf6c6...be8`. Resampling is clustered by market. Source-order and citation- capacity findings remain supporting development diagnostics until their paired analyses and release receipts are promoted into this authenticated lineage.

The expanded capability analysis contains 56,448 authenticated tasks and has SHA-256 `39953069...e88b5`. Its fixed-grid utility diagnostic has SHA-256 `586aa645...d697331`. The dual-view smoke release has source-manifest SHA-256 `d4191c14...01a45a`; its two-engine receipt has SHA-256 `3e1f6856...7dc139`, and its complete analysis has SHA-256 `5cdb6d54...075b3ed`. The corresponding analyzer has SHA-256 `02daf8cf...597bc2`; its apparent two-engine mixing gain is retained only as a superseded development diagnostic after failing broader checks on markets left out. The 24-market published-response analysis has SHA-256 `f07735e3...01ba47`. These artifacts use only consumed ASQA markets and do not authorize reserve execution.

D Additional Experimental Results

This appendix reports additional results in the order of Section 5, followed by published-method comparisons, development diagnostics and original contrast intervals. Each entry identifies its market set, response set, endpoint and evaluation status.

D.1 Prospective results by engine (Section 5.2)

The source is the registered prospective digest from analysis job 10450734, with analysis-artifact hash `7ba704e0ff32c333d15f021f603ec8c1307cc43fb3cce6dad578d3ec11bdfb00`. All rows use 200 ASQA markets, $C = 3$, reciprocal-rank exposure and the committed policy. The pooled maximum is computed after the declared pooling; it is not the mean of engine-specific clipped maxima.

The searched-bank bound and the support-over-pure-manipulation comparison hold on all four engines. The latter has pooled support advantage 7.02%, with one-sided lower bound 5.43% and adjusted $p = 0.0004$. Static selected-source quality meets the registered margin of 0.025 on its own $[0, 1]$ scale; engine-wise differences are small negatives, from -0.62% to -0.17% on that scale. This is not human-rated answer quality.

D.2 Normalization, screening and the development decomposition (Section 5.3)

Table 6: Max premium (%) and original 95% interval endpoints. The searched bank has two responses; the admissible nine-response and unrestricted eleven-response banks give the same recorded wider-bank values. Their equality here is empirical, not a general identity.

Engine	Searched-bank max premium	Wider-bank max premium
Llama-3.1-8B	1.03 [0.00, 2.41]	1.20 [0.25, 2.57]
Qwen2.5-7B	1.86 [0.20, 3.54]	3.84 [2.51, 5.25]
Qwen3-8B	0.48 [0.00, 1.87]	2.14 [1.08, 3.22]
Tulu3-8B	0.00 [0.00, 0.33]	2.12 [0.85, 3.40]
Pooled	0.51 [0.00, 1.31]	2.32 [1.60, 3.11]

Table 7: Development decomposition: P440, four engines, label-inclusive bank, binary citation exposure ($C = 1$). All values are in %. Higher gap is better. H and M are descriptive policy levels, not paired improvements. Gaps are computed before rounding. The bank excludes support-plus-manipulation compounds, which are evaluated separately.

Platform policy	Support $H \uparrow$	Manipulation $M \downarrow$	Gap $\Delta \uparrow$ [95% interval]
No defense	6.61	11.42	-4.81 [-6.34, -3.24]
Whitespace	**8.98**	5.88	$+3.10$ [1.39, 4.44]
Whitespace + screen	8.68	5.88	$+2.80$ [1.25, 4.33]
Deletion alone	8.59	**3.05**	**+5.54** [4.03, 7.06]
Committed	8.28	**3.05**	$+5.24$ [3.79, 6.80]

Whitespace alone changes the sign of the P440 gap; deletion contributes a further increase. The screen does not add a gain on this response bank. P440 pools the D140 and R300 markets shown separately in the plot. This label-inclusive development bank uses binary exposure at $C = 1$. The complete policy decomposition below preserves the original interval endpoints. Policy levels must not be subtracted to construct a paired interval.

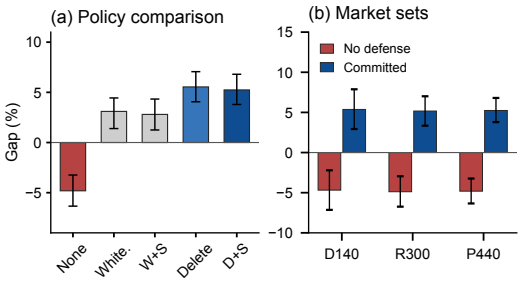

Figure 5: Development incentive gaps. W+S: whitespace plus screen; D+S: deletion plus screen.

D.3 Input invariance checked on the development documents (Section 4.2)

Edits the committed policy removes by construction. Let o be an unedited source and write the committed policy as the deletion step D after the screen S. The conditions below are sufficient for the policy to serve a submission d exactly as it serves o, so that the edit is removed on that source and earns nothing on every engine by Proposition 1(a). They are read off the policy's code and checked on the 671 development documents with their stored edits, calling no engine; they say nothing about citation outcomes. Standing hypotheses on o: it is a single line, each stage of the step keeps at least one unit of o on its own, no unit begins with a token the marker pattern matches, and every kept unit has an informative token, a token of at least three characters that is not a stop word. (i) Every submission that trips the screen is served as o, and every admissible submission passes it, because the contract's clause on query wording is the screen's own statistic. (ii) The step serves o when d is o with bullet or number markers prefixed to lines, a whitespace run without a line break replaced by another, a re-lining at the step's own unit boundaries when every unit ends in a terminal mark, or one pair of quotes around a unit not already quoted, in any combination. (iii) It serves o when d adds units that the step keeps separate, each a bare label, a list fragment without a terminal mark, or a unit whose informative tokens all occur in o, provided the added tokens do not together cover every token that distinguishes some kept unit of o. (iv) Conditions (ii) and (iii) compose. On the development documents: the screen restores the original on none of the eight stored presentation edits; (ii) holds on 671 of 671 documents for whitespace substitution and on all 44 sources whose units all end in a terminal mark for the other three operations; (iii) holds in 1,893 placements of a grounded non-covering sentence on 650 documents. Removal rates of the stored edits under the step, out of 671 documents: bullets 653, quotation 602, grounded

keyword line 661, section headers 144, fronting 178, sentence reordering 68, planner-searched edits 81, model paraphrase 0 of 667. The step changes the served text of 60 unedited sources, 8.9%: 51 by dropping a unit as repeated, 7 by stripping quotes, 2 by stripping a leading number. The conditions do not cover, and a residual can come from, any of the following: reordering; a change inside a kept unit, a label in any form, a case change or a paraphrase; a line break inside a sentence; an added unit whose words are not the source's; an added unit with no informative token; an added unit that covers a kept unit's distinctive tokens; an added sentence the splitter merges into its neighbour; and a source kept only because every unit would otherwise be dropped.

D.4 Renewed search and the label-repair candidate (Section 5.2)

The fixed-bank evaluation favored a candidate that deleted colon-terminated labels: its development gap was 8.34%, or 3.16% above the committed policy. That comparison kept 128 of 140 markets by excluding every market containing a rejected output; pre-output authorization of that retention rule is not established. The apparent gain is therefore not reported as an independent confirmation.

Renewed search gave max premiums of 1.60% for the candidate and 1.18% for the committed policy on pair A, and zero for both on pair B. The paired committed-minus-candidate premium difference on pair A was -0.42%, with interval $[-1.74, 0.91]$; positive values would favor the candidate. Separator, bracket and whitespace label variants survived above 0.95 on the extension markets.

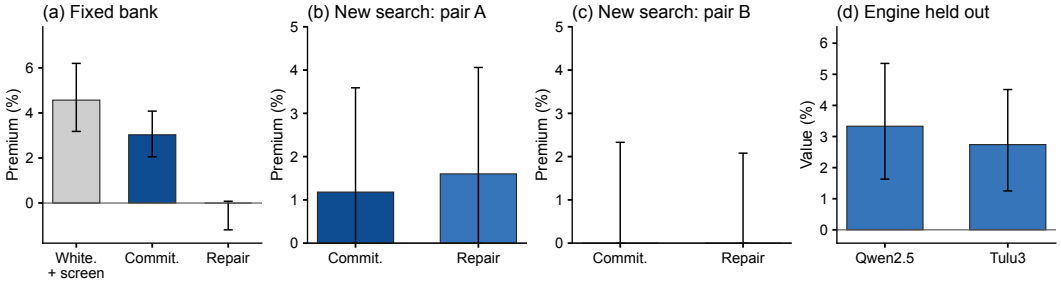

Figure 6: Development of the label-repair candidate. Fixed-bank gains did not translate into an established advantage after renewed search. (a) Filtered fixed-bank cohort; (b,c) renewed search on engine pairs A and B; (d) selection on the other engine. None is prospective confirmation.

On R300, colon forms are deleted while separator, bracket and whitespace forms survive. No tested repair beyond the committed policy meets the adoption criteria.

Three certified label evasions earn 1.37%, 1.15% and 0.64% under deletion alone on two engines. On that bank, deletion improves the gap over no defense by 3.40%; its 0.30% advantage over whitespace has an interval including zero. These are deletion-only measurements, not measurements of the composed policy. The clean-traffic study records a 4.02% change in selected sources on D140. This observed rate is not itself the lower bound from the leak-free deletion theorem: the implemented policy is not leak-free over all admissible edits.

The planner class motivates response search: at matched proxy rank and document attackability its edits earn 3.81% [1.72, 5.95] more than the template library on D140, and materializing one for every eligible document adds 1.35% [0.18, 2.48] against the then-composed target procedure. The composed procedure of that study preceded the committed one and its numbers are not carried.

D.5 Ordered exposure, competition and universal adoption (Section 5.4)

The QAMPARI adoption analysis compares each publisher's edit with its unchanged source under the same rival submissions. The underlying development game has five publishers, three actions (identity, bullets and sections), two reversed source orders, and two engines, Llama-3.1-8B and Qwen3-8B. All 243 joint action profiles were scheduled for each market. Two missing logical outcomes occur in one of the 120 markets; the registered adoption-dose analysis excludes that market from every primary component, leaving 119 common-complete markets. Actions, publishers and engines receive equal weight within each market before a 10,000-replicate market bootstrap.

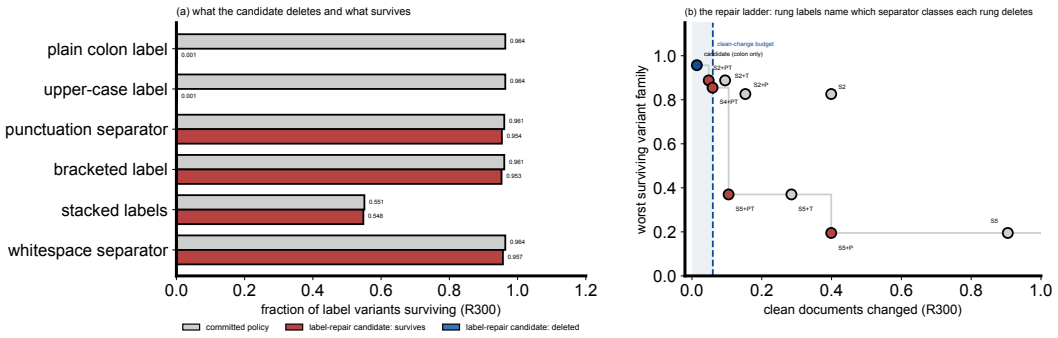

Figure 7: The candidate repair's frontier. (a) Survival of different label forms under the committed policy and label-repair candidate on R300. (b) The share of clean documents changed versus survival of the worst-performing label family, for each repair setting. Shading marks the clean-change budget.

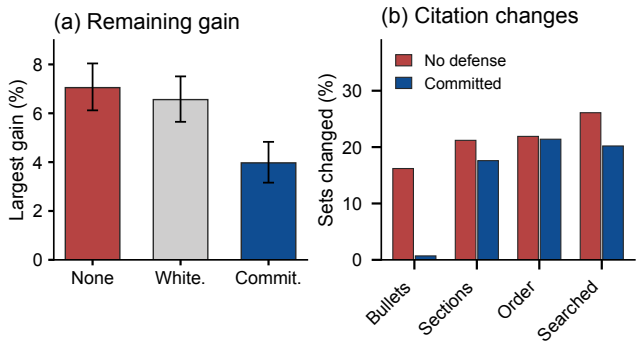

Figure 8: Universal adoption on 48 ASQA markets, $C = 3$: remaining gains and citation-set changes.

At $C = 3$, the matched gain from adopting GEO averages 3.82% with no active rivals and 0.73% when all four rivals adopt the same action. Their paired difference is 3.09% [2.39, 3.81], or approximately 81% of the initial gain. The action-specific declines are 2.30% [1.49, 3.17] for bullets and 3.87% [2.97, 4.78] for sections. These are retrospective development results, not the registered transfer evaluation of the final policy. They come from analysis job 10413152 and the registered adoption-dose specification.

The earlier 120-market saturation analysis uses a different contrast: a publisher's exposure under unilateral adoption minus its exposure under universal adoption. Its required profiles contain no missing outcomes. The pooled decline is 3.92% [3.24, 4.61]. Universal and clean profiles share 74.48% of their top-three cited sources on average, so similar total exposure does not imply identical allocations. Homogeneous adoption also leaves mean finite-bank deviation regret of 5.93% [5.39, 6.48], with editing costs set to zero. These development contrasts describe exposure under rival adoption.

On the 48-market competition study (pair B, $C = 3$), the largest unilateral gain after universal adoption is 7.05% [6.12, 8.04] with no defense, 6.56% [5.65, 7.51] with whitespace, and 3.97% [3.16, 4.83] under the committed policy. These are unilateral gains around universal-adoption profiles.

D.6 Published GEO attacks (Section 5.2)

Published GEO methods enter the same response set under the candidate's deletion arm, which is deletion-only: no grounded screen is applied on that path, so these rows price the deletion step and not the committed policy, and these historical rows alone do not evaluate the screened procedure. Appendix D.7 reports the later composed-policy comparison on its binary-citation endpoint, separately from ordered exposure. Adding eighteen published-method columns to the fourteen label columns leaves the pooled maximum on a label column: under that deletion step the union-set gap on the extension markets is 1.12% and the paired difference between the union and the label-only response set is 0.00% with an interval that does not resolve. The union gap's registered confirmation and its

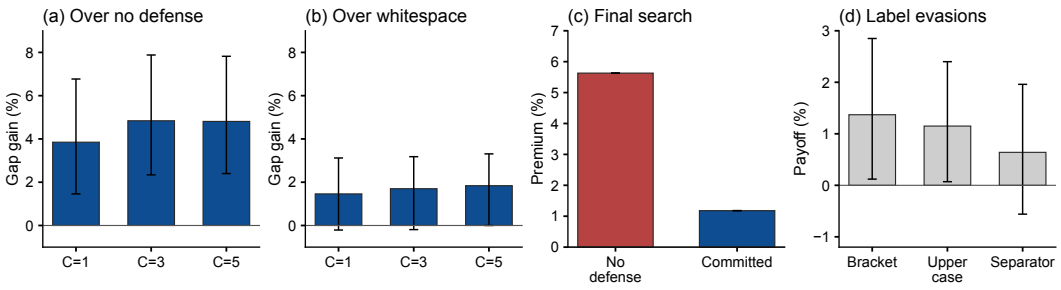

Figure 9: Development results: ordered exposure over (a) no defense and (b) whitespace plus screen; (c) final search; (d) label evasions under deletion alone. Capacities are prefixes of one answer. No new intervals are inferred for part (c).

Holm disconfirmation are both false, and neither implies the other; per-engine values are not reported for this response set, as for every other pooled row. Under the identity policy the same union gap is -4.49%, which locates the sign in the deletion step rather than in the response set. Whether the absent screen matters on this path has been measured rather than assumed: recomputing the declared screen predicate on every candidate row and rejoining it to the same payoff leaves the gap unchanged under the deletion step these rows score, 0.00% on the pre-registered markets $[-0.37,\ 0.00]$ and on the extension markets $[-0.00,\ +0.19]$. That zero is exact rather than unresolved: the maximizing response is unchanged in every combination tested, and the reverted mass of the binder equals the support control's to full precision on every market and on every row of both market sets, with the same screen decisions, so the two cancel. We state that as a verified property of these market sets rather than of the construction, since whether such a column must always carry the same supporting text is a property of a builder not examined here. The same contrast is not zero under the identity policy, which measures the non-transfer directly. This is a post-hoc rollback on already-served tasks, not a re-run with the screen inside the served procedure, so it still certifies nothing about the committed policy.

One supporting literature-response comparison evaluates five published GEO rewriting prompts, each generated by two offline generators, against five platform policies on 24 ASQA markets, two engines, and paired orders (5,280 tasks). Risk-adjusted support is an instruction-only policy that asks the engine to discount headings, repetition, query-term stuffing, and unsupported authority cues after assessing factual support. It is the best pure policy for both engines in this small comparison, with worst-method focal gain 4.167% for Llama and -2.083% for Qwen. On the expanded 96-market response set, all 2,880 automated support judgments parse successfully, but only 41.7% of responses pass the independent judge. This is an automated, evaluator-sensitive support check, not semantic ground truth. We omit the earlier non-self and unanimous pooled rates because judge composition differed by generator. Generated rewrites require a separate support audit before entering the primary support-preserving action family; this 24-market comparison does not establish a defender advantage on the full market set.

D.7 Audited transfer and mechanism studies

These records, reconciled on September 24, report the primary ASQA comparison and separate transfer, mechanism and development studies. Percentages are absolute exposure changes. The companion source ledger records immutable revisions and artifact hashes.

Table 8: Maximum manipulation advantage and paired reduction (%), using reciprocal-rank exposure at $C = 3$. ASQA and QAMPARI use four 7B–8B engines and panel-specific searched edits. Original paired intervals are retained. Bold marks lower point estimates, not significance; studies are not pooled.

Evaluation	Markets	No defense	Committed	Reduction [95% interval]
ASQA prospective	200	4.99	**0.51**	4.48 [3.09, 5.86]
QAMPARI transfer	122	4.35	**0.34**	4.01 [2.47, 5.55]

QAMPARI's searched residual bound holds on all four engines. On the searched bank the canonical normalizer without the screen leaves the same residual as the committed policy: no defense minus

deletion-only is 4.01 [2.47, 5.55] pooled and resolves on every engine, deletion-only minus committed is 0.00 [0.00, 0.00], and the screen's rollback of the 15 honest conflicts moves the honest and manipulation levels by the same amount, so on this bank the reduction is the normalizer's, as on ASQA.

Screening and published rewrites. The QAMPARI keyword study uses no edit, a retrieved supporting passage, keyword injection and a grounded keyword variant on 122 markets and four engines. The screen is evaluated through its registered rollback rule on authenticated outputs. The attack family was selected after the original result. Six of eight generation shards missed the original runner QC; the exploratory record retains those failures and all scored outputs.

Table 9: Exploratory keyword factorial, $C = 3$ (%). The screen-only and normalizer-only increments are relative to no defense; the last adds screening to normalization. Original two-way bootstrap intervals resample markets and engines. The interaction estimate is -0.46 [-1.55, 0.51].

Policy	Gap ↑	Increment	95% interval for increment
No defense	-0.18	Reference condition	
Screen only	$+1.26$	$+1.44$	[0.24, 2.75]
Normalizer only	$+3.62$	$+3.79$	[1.82, 6.15]
Normalizer + screen	**+4.60**	$+0.98$	[0.28, 1.75]

Against 18 published GEO rewrite columns, the binary-citation ($C = 1$) gap changes from -5.05% to $+2.82\%$ on D140, and from -3.21% to $+3.03\%$ on R300. Paired gains are 7.86% [3.34, 13.00] and 6.23% [0.63, 10.78]. These are authenticated compositions of existing outputs with the support arm. Defended levels have two-way intervals [-0.52, 5.57] and [-0.58, 5.35], with uncertainty over both markets and engines. Screen increments differ between market sets: -0.53% [-1.50, 0.68] and $+0.44\%$ [0.20, 1.13]. The separate GEO-Bench zero-shot column is not included in the maximum over published rewrites.

Model breadth. Table 10 presents selected keyword-edit comparisons on 122 QAMPARI questions. The Qwen3-8B reference reuses stage-4 outputs; second-runtime rows are reported separately.

Table 10: Exploratory keyword-edit comparison on 122 QAMPARI markets, $C = 3$ (%; higher gap is better). Per-engine intervals resample markets. The second-runtime rows use one H100 and different library versions on the same questions and edits; they are not pooled with earlier rows. Parsing failures and runtime checks are described below.

Engine	No-defense gap	Defended gap	Improvement [95% interval]
Qwen3-0.6B	Output-contract failure; no valid estimate		
Qwen3-4B	$+0.98$	**+4.19**	3.21 [1.79, 4.63]
Qwen3-8B (reused)	$+1.04$	**+3.61**	2.57 [0.92, 4.34]
Qwen3-14B	$+1.27$	**+4.93**	3.66 [2.13, 5.20]
Gemma-2-9B	-0.77	**+2.94**	3.70 [2.13, 5.30]
Mistral-7B-v0.3	$+0.94$	**+4.32**	3.37 [0.94, 5.53]
Qwen2.5-14B	$+0.57$	**+3.59**	3.02 [1.19, 4.95]
Second compute runtime (one H100 GPU)			
Qwen3-32B	-3.29	**+3.93**	7.23 [5.51, 8.92]
Qwen3-8B (anchor)	$+1.05$	**+3.91**	2.86 [1.12, 4.73]
Llama-3.1-8B (anchor)	$+0.89$	**+4.44**	3.55 [2.06, 5.20]

Table 10 compares models on the specified keyword edits. Qwen3-0.6B fails parsing; the reused 8B reference would fail the newer strict-format gate, while 4B and 14B pass. The second-runtime anchors' defended gaps differ from their original-runtime values by 0.30% [-0.52, 1.11] for Qwen3-8B and 0.72% [0.01, 1.38] for Llama-3.1-8B.

Adaptive coverage and content checks. Adding searched and stacked edits to the keyword bank retains a paired gap improvement of 4.01% [2.24, 5.98], but the defended union gap is -0.34% [-1.81, 1.31]. A grounded stacked edit maps to the same input as the support control, creating an exact tie in that construction; results with that edit excluded concern the narrower set. Cross-fitted finite-action selection gives a 3.88% [2.01, 6.00] paired gain. The larger in-sample 8.81% gain

is a hindsight diagnostic. A separate 48-market study gives 3.97% [1.47, 6.72]; removing four overlapping markets retains 3.70% [1.13, 6.67].

Supporting material means a relevant retrieved addition, not necessarily a new answer. Only 205 of 610 QAMPARI additions introduce a new gold-answer group. Additions without a new gold answer account for 56% of the defended support lift and 61% of the no-defense lift. Clean-document answer recall changes by -0.10% [-0.84, 0.50]; this is a descriptive pooled estimate. Nine metric/capacity views re-score the same answers.

D.8 Wider banks and delivery

The following analyses were completed after the prospective verdict. They supply missing comparisons and diagnostics without changing the registered tests. All exposure values are absolute changes multiplied by 100, not relative percentages.

Table 11: Wider-bank max premiums (%), 200 ASQA markets, four engines, $C = 3$. Paired intervals use the original union-of-seven construction. These are descriptive pooled comparisons, with separate intervals for the two edit sets.

Bank	No defense	Committed	Paired reduction [95% interval]
Admissible (9 edits)	4.99	**2.32**	2.66 [0.80, 4.53]
Unrestricted (11 edits)	4.99	**2.32**	2.66 [0.73, 4.60]

The no-defense maximum is attained by the same searched edit on all three banks. Adding edits therefore leaves its estimate unchanged, while a support-plus-label edit raises the committed maximum. This accounts for the larger reduction on the two-edit set.

What reaches the engine. On the two-edit searched bank, 0.85% of evaluated tasks are rolled back or reduced to the unedited source. This control differs from the supporting-passage control used in the premium. For the binding searched edit stacked with support, 1.5% are rolled back and 0.1% have served text identical to the supporting-passage control. Of the delivered tasks, 88.5% earn no premium over that control. These descriptive measures record exact matches and realized returns alongside partial changes to surviving edits.

Across the audit, all 448 tasks whose complete prompt matches the supporting-passage prompt have identical raw outputs and zero premium. These prompt-identical tasks satisfy the premise of the removal guarantee. For the binding edit, editing gain falls from 12.36% to 9.10%, while the supporting-passage gain rises from 7.38% to 8.59%. Both changes contribute to the lower premium.

Paired 35B result. On 48 consumed ASQA markets, Qwen3.5-35B-A3B (about 3B active parameters) has a native-minus-deletion-step unilateral-deviation regret reduction of 2.96% [1.88, 4.20]. The interval uses the registered, unadjusted percentile bootstrap. It supplements the 5.06% and 2.09% levels for deletion-only development data.

D.9 Adaptive comparison and additional diagnostics

A common bank for three policies. The adaptive comparison evaluates no defense, the committed normalizer plus screen, and the published SafeGEO-L1 prompt layer. The union contains edits from equal-budget, fixed-seed proxy searches against no defense and the committed policy. Both searches are constrained by the registered admissibility contract; in particular, native-targeted candidates must survive the grounded screen. The native result is consequently a lower bound on what an unrestricted native-targeted search might find. Search does not target SafeGEO-L1. L1 was selected retrospectively on development data, not established as the best SafeGEO variant.

This is a post-registration extension of the already evaluated 200 ASQA markets. All three policies use 999 shared documents, capacity three, and Llama-3.1-8B, Qwen2.5-7B and Qwen3-8B. Tulu3 failed the initial 32-task citation-format smoke gate and is not included in these estimates. The original four-engine prospective verdict is unchanged. Intervals resample matched market clusters using the registered analysis settings.

Table 12: Exploratory adaptive comparison, ASQA 200, three engines, $C = 3$, common union bank. Premium means the largest nonnegative advantage over the supporting-material control. Bold identifies the lowest level among the three tested policies.

Policy / contrast	Estimate (%)	95% interval
No defense	8.39	[7.26, 9.57]
SafeGEO-L1	8.52	[7.28, 9.80]
Committed	**1.55**	[0.72, 2.42]
No defense minus committed	6.83	[5.18, 8.49]
SafeGEO-L1 minus committed	6.96	[5.18, 8.75]

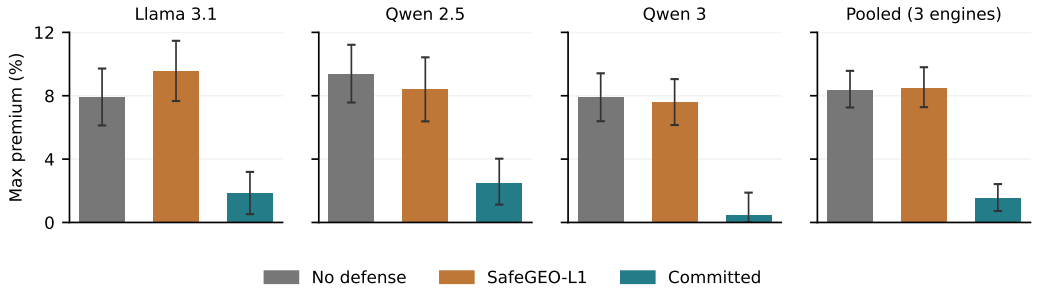

Figure 10: Common union-bank premiums by engine and pooled, with original 95% intervals. No defense and the committed policy were search targets; SafeGEO-L1 was evaluated on their union. This is the three-engine exploratory comparison.

The union raises the no-defense premium by 3.24% [1.93, 4.55] and the L1 premium by 3.86% [2.61, 5.12], relative to the committed-targeted bank on these same three engines. The increase under the committed policy is 0.43% [−0.26, 1.28]. Thus the expanded bank changes the comparison materially without establishing a corresponding increase for the committed policy. The increments pair outcomes within these three engines.

Four-engine add-back (exploratory). The completed Tulu3 add-back provides a second aggregate, without replacing the three-engine estimates above. Its acceptance carries over the registered 132-task smoke population, in which 128 tasks passed. The four-engine aggregate reproduces the original registered estimates before adding the native-targeted edits. It uses the same 200 markets, 999 shared documents and $C = 3$ endpoint. Table 13 gives pooled results; neither pooled contrast triggers the zero floor. This is a post-registration exploratory extension.

Table 13: Added four-engine exploratory union-bank comparison, ASQA 200, $C = 3$. L1 is one published prompt-level layer selected retrospectively on development data. The search targets no defense and the committed policy, not L1. Earlier three-engine estimates remain in Table 12.

Policy / contrast	Estimate (%)	95% interval
No defense	8.32	[7.24, 9.45]
SafeGEO-L1	8.40	[7.20, 9.63]
Committed	**1.06**	[0.27, 1.84]
No defense minus committed	7.26	[5.61, 8.91]
SafeGEO-L1 minus committed	7.33	[5.56, 9.11]

Support, baseline exposure and answer recall. The support-lift increase is not driven by a lower no-edit baseline. On the 200 prospective markets and four engines, the single-view decomposition gives a 1.21% increase in support lift, consisting of a 1.23% increase in support exposure minus a 0.02% increase in baseline exposure (Table 14). The decomposition measures citation exposure. Valid abstention changes from 0.53% to 0.54%.

The additional answer-quality check applies the ALCE-style gold-answer-alias recall rule to recorded outputs, including invalid citation outputs. Neither the no-edit comparison nor the supporting-

material comparison establishes a recall change. The measure scores gold-answer string presence, with descriptive intervals.

Table 14: Post-verdict diagnostics, 200 ASQA markets and four engines. The first three rows use $C = 3$ exposure; the last two use answer recall. Intervals are market-cluster bootstrap intervals.

Committed minus no defense	Estimate (%)	95% interval
Supporting-material exposure lift	1.21	[0.36, 2.07]
Supporting-material exposure level	1.23	[0.39, 2.10]
No-edit exposure level	0.02	[−0.16, 0.21]
No-edit gold-answer recall	−0.09	[−0.50, 0.32]
Supporting-material gold-answer recall	−0.13	[−0.62, 0.37]

Table 15: Gold-answer recall by engine (ALCE alias rule, %), 200 prospective ASQA markets, for unedited, supporting-passage and binding stacked-edit cells under no defense and the committed policy, with the paired absolute difference in %. Every difference includes zero; levels differ by engine, so comparisons are within an engine.

Engine	Cells	No defense	Committed policy	Difference (%)
llama3.1-8b	Unedited	31.97 [27.77, 36.22]%	31.89 [27.63, 36.23]%	−0.08 [−1.28, 0.95]
llama3.1-8b	Supporting passage	33.66 [29.65, 37.68]%	33.49 [29.68, 37.31]%	−0.17 [−1.06, 0.69]
llama3.1-8b	Binding stacked edit	33.27 [29.39, 37.20]%	33.49 [29.58, 37.37]%	0.22 [−0.75, 1.15]
qwen2.5-7b	Unedited	30.45 [26.61, 34.61]%	30.43 [26.67, 34.46]%	−0.03 [−1.20, 1.15]
qwen2.5-7b	Supporting passage	32.52 [28.89, 36.34]%	32.68 [29.12, 36.36]%	0.16 [−0.71, 1.01]
qwen2.5-7b	Binding stacked edit	32.42 [28.88, 36.13]%	32.02 [28.50, 35.64]%	−0.40 [−1.29, 0.51]
qwen3-8b	Unedited	42.22 [37.60, 46.99]%	42.13 [37.56, 46.85]%	−0.09 [−0.88, 0.77]
qwen3-8b	Supporting passage	43.77 [39.36, 48.26]%	43.46 [39.14, 47.91]%	−0.31 [−1.18, 0.58]
qwen3-8b	Binding stacked edit	43.29 [38.96, 47.67]%	43.71 [39.38, 48.16]%	0.42 [−0.29, 1.15]
tulu3-8b	Unedited	25.44 [21.66, 29.34]%	25.28 [21.55, 29.11]%	−0.17 [−0.84, 0.52]
tulu3-8b	Supporting passage	24.89 [21.46, 28.39]%	24.71 [21.33, 28.13]%	−0.18 [−1.14, 0.74]
tulu3-8b	Binding stacked edit	24.98 [21.56, 28.39]%	24.97 [21.50, 28.51]%	−0.01 [−0.88, 0.85]

Profitable delivered edits. For the binding searched edit stacked with support, 16.8% of delivered tasks earn positive premium without defense, versus 11.5% under the committed policy. The analysis pairs market-level rates and estimates a reduction of 5.38% [4.32, 6.44]. Delivery is classified separately under each policy, so the rates and positive-premium magnitudes condition on policy-specific subsets. The 448 prompt-identical tasks reported above provide the direct check of the exact-removal premise.

Which edit binds. The binding edit is the bank column with the largest market-mean premium (Table 16). Under no defense the stacked edit binds on every engine and every bank. Under the committed policy the stacked edit binds on the searched bank, and the supporting passage carrying a label binds on the admissible and unrestricted banks on every engine; the committed policy removes 1.5% of that edit's cells (the screen's rollback) and delivers 98.4% with a served text that differs from the unedited source. Pooled, the runner-up trails the binding edit by 7.05 (no defense, searched), 2.08 (no defense, wider banks), 7.52 (committed, searched) and 1.81 (committed, wider banks), so no near tie occurs. This census is descriptive: it names the binding edit, its registered shape (original kept verbatim, supported material added, no reordering, no unsupported content) and its delivery class.

Table 16: The strongest edit and its signed advantage over the supporting passage (%), on 200 prospective ASQA markets with $C = 3$. Columns distinguish no defense, the committed policy on searched edits, and the committed policy on wider sets. Rows show each engine and the pooled estimate.

Engine	No defense, all banks stacked edit	Committed, searched stacked edit	Committed, wider banks support passage + label
llama3.1-8b	4.82	1.03	1.20
qwen2.5-7b	5.65	1.86	3.84
qwen3-8b	4.96	0.48	2.14
tulu3-8b	4.52	−1.31	2.12
Pooled	4.99	0.51	2.32

For Table 16, the committed policy removes 1.5% of the binding edit's tasks and delivers 98.4% (98.5% on the searched set); the remaining tasks are excluded from this estimand.

Recovery as an analytic control. A separate check constructs recovery followed by the committed policy, assuming access to the original source and the declared supporting passage. It checks the passage's verbatim identity, not whether its claims are factually supported. Over all three declared banks, all 95,992 comparisons map to the appropriate control prompt and the premium is zero by construction. This analytic comparison reuses recorded control outputs. The no-defense-minus-recovery premium is 4.99% [3.67, 6.30]. This verifies the recovery construction on the declared banks under the stated access assumptions.

D.10 WHITESPACE AND DELETION ON THE PROSPECTIVE PANEL

This exploratory extension uses the already evaluated 200-market ASQA panel. The whitespace-only arm collapses every run of whitespace, including line breaks, to one space and applies no screen. It uses the original edit sets, both source orders, four engines and the $C = 3$ reciprocal-rank exposure measure. All 105,600 tasks were authenticated; 1,005 transport-contaminated outputs remain in the frozen scoring rather than being dropped. The analysis reproduces 105 registered rows and 40 matrices. Where the whitespace and no-defense inputs coincide, raw outputs are identical.

Table 17 gives manipulation's largest nonnegative advantage over the supporting-material control. Table 18 gives paired reductions. Level intervals use the frozen market-cluster bootstrap; paired intervals use the frozen union construction. The pooled contrasts do not trigger clipping. On the searched set, whitespace accounts for 68.0% and deletion for 32.0% of the total reduction, with no remaining screen contribution. The shares are descriptive point estimates computed before rounding. The component decomposition concerns the two searched edits.

Table 17: Exploratory whitespace extension, ASQA 200, four engines pooled, $C = 3$. These are exposure advantages, not relative percentage improvements.

Edit set	Policy	Advantage (%)	95% interval
Searched (2)	No defense	4.99	[3.95, 6.06]
Searched (2)	Whitespace	1.94	[1.13, 2.77]
Searched (2)	Deletion only	0.51	[0.00, 1.31]
Searched (2)	Committed	0.51	[0.00, 1.31]
Admissible (9)	No defense	4.99	[3.95, 6.06]
Admissible (9)	Whitespace	2.76	[2.03, 3.53]
Admissible (9)	Committed	2.32	[1.60, 3.11]
Unrestricted (11)	No defense	4.99	[3.95, 6.06]
Unrestricted (11)	Whitespace	2.76	[2.03, 3.53]
Unrestricted (11)	Committed	2.32	[1.60, 3.11]

Table 18: Paired reductions on the same panel. Positive values favor the second policy. Different edit sets retain their own original intervals.

Edit set	Paired contrast	Reduction (%)	95% interval
Searched (2)	No defense − whitespace	3.04	[1.81, 4.27]
Searched (2)	Whitespace − deletion	1.43	[0.56, 2.30]
Searched (2)	Whitespace − committed	1.43	[0.57, 2.30]
Searched (2)	Deletion − committed	0.00	[0.00, 0.00]
Admissible (9)	No defense − whitespace	2.22	[0.38, 4.06]
Unrestricted (11)	No defense − whitespace	2.22	[0.33, 4.12]

Provenance. Whitespace analysis job 10558859 produced `e1_whitespace_rows.json` (SHA-256 `ea65588b...7811d0`). The artifact was retrieved from the compute cluster and checked against the execution receipt. It changes no registered verdict.

D.11 SCORING ONE, THREE OR FIVE CITATIONS

The reduction remains positive in pooled comparisons when the same ASQA answers are scored at one, three or five citations (Table 19). These descriptive analyses re-score the same answers from 200 markets and four engines. We retain the original edit sets and reciprocal-rank weights. The three scores measure exposure over different citation prefixes.

Table 19: Manipulation's largest advantage over adding supporting material and its paired reduction (%). Lower levels are better; bold marks the lower policy estimate within each task, not significance. Intervals use the frozen paired, market-level construction.

Edit set	Citations	No defense	Committed	Paired reduction [95% interval]
Searched (2)	1	4.91	**0.78**	4.13 [2.70, 5.56]
	3	4.99	**0.51**	4.48 [3.09, 5.86]
	5	4.90	**0.48**	4.41 [3.02, 5.80]
Content-preserving (9)	1	4.91	**2.15**	2.76 [0.88, 4.64]
	3	4.99	**2.32**	2.66 [0.80, 4.53]
	5	4.90	**2.27**	2.62 [0.77, 4.47]
Expanded (11)	1	4.91	**2.15**	2.76 [0.80, 4.72]
	3	4.99	**2.32**	2.66 [0.73, 4.60]
	5	4.90	**2.27**	2.62 [0.70, 4.55]

Both arms evaluate edits searched against the committed policy. Pooled intervals exclude zero at all three citation counts. The wider-set contrasts remain descriptive, as specified before evaluation. The nine- and eleven-edit sets retain their separate contrast intervals.

D.12 RECOVERY ON QAMPARI

Access to the original source permits a stronger intervention than deletion alone. We apply the pinned recovery map before the committed policy, using the original source and the declared supporting passage. The check establishes verbatim passage identity. For the two tested QAMPARI edits, the resulting input matches a recorded control input. We can therefore reuse the control answers under the recorded deterministic-generation assumptions. This is an analytic evaluation on the two existing edits.

Table 20: Exploratory analytic recovery on QAMPARI, two searched edits, four engines and three ordered citations (%). Advantage is manipulation's largest gain over adding supporting material; support gain uses each policy's unedited control. Recovery requires the original source and declared passage. Its zeros follow from checked input identities and reused control answers.

Panel	Policy	Support gain [95% interval]	Advantage [95% interval]
QAMPARI (122)	No defense	3.78 [2.14, 5.50]	4.35 [3.21, 5.55]
	Committed	4.94 [3.15, 6.84]	0.34 [0.00, 1.21]
	Recovery then committed	4.94 [3.15, 6.84]	**0.00** [0.00, 0.00]

On QAMPARI, the no-defense-minus-recovery reduction is 4.35% [2.89, 5.80]. The further reduction from the committed policy is 0.34% [−0.78, 1.45], so an incremental benefit is not established. All 9,760 edited evaluation tasks map to controls.

D.13 EXEMPTING THE SUPPORTING PASSAGE FROM THE LENGTH CAP

We evaluate a separate policy version that exempts the declared supporting passage from the normalizer's 12-unit cap and otherwise leaves the normalizer and screen unchanged. The planner produces two edits against this version on 300 GEO-derived markets. No defense and the revised policy are scored on that set with four engines, both source orders and three ordered citations. This is a post-registration exploratory evaluation of the revised policy.

The paired reduction is resolved pooled and on Qwen3-8B alone (1.66 [0.04, 3.28]); the other engines' reductions are 2.06 [−0.12, 4.25] (Qwen2.5-7B), 0.85 [−1.09, 2.78] (Tulu3-8B) and 1.81 [−0.00, 3.61] (Llama-3.1-8B, where the zero floor binds). The remaining advantage under the exempt policy

Policy	Quantity	Estimate (%)	95% interval
No defense	Support gain	3.65	[2.72, 4.60]
	Manipulation advantage	2.00	[1.40, 2.63]
Passage-exempt normalizer	Support gain	4.20	[3.18, 5.27]
	Manipulation advantage	0.41	[0.00, 1.03]
No defense − passage-exempt	Paired reduction	1.59	[0.62, 2.57]

Table 21: Exploratory passage-exempt policy on 300 GEO-derived markets, four engines pooled, three ordered citations, its own two searched edits. Support gain is the return to adding the supporting passage over the unedited source under each policy. Manipulation advantage is the largest nonnegative advantage over adding supporting material.

is 0.40, 0.44, 1.06 and 0.00 on the four engines in that order, and the support gain is 4.01, 4.51, 3.53 and 4.74 (3.47, 2.53, 4.24 and 4.36 without defense). The passage is kept on 1,155 of the 1,200 supporting-material documents (693 exempted, 462 already within the cap); on the remaining 45 it cannot be matched verbatim and the cap applies. The screen fires on no column. Every arm has 2,400 tasks per engine and column with parse rates of 0.972 to 0.996.

Provenance. Analysis job 10571469 produced `e5_analysis.json` (SHA-256 `f095b4a6...a852f`) under a registration fixed before any task (SHA-256 `fb3475e8...236b17`). It changes no registered verdict.

D.14 What survives the strongest label edit

The strongest edit on the wider ASQA sets adds a label to supporting material. A descriptive analysis groups its outputs by how much newly added text survives. All 7,432 delivered tasks eligible for the comparison fall in one group: all new words remain. On these same tasks, the average advantage over adding supporting material is 2.50% under the committed policy and 2.98% without defense. This is a conditional mean for one edit; Table 1 reports the market-averaged maximum.

The single eligible group retains all added words while other parts of the input may change. The status counts include 120 screen rollbacks and 432 tasks reduced to the supporting-passage input among 7,984 tasks with a defined survival measure; 16 further tasks have no defined measure. These diagnostics describe the label edit's delivery.

D.15 Other development diagnostics

The four-engine surface tensor shows action-level heterogeneity. In an exploratory same-base comparison, Tulu exceeds Llama by 10.0% under section formatting. The comparison concerns these two post-trained systems. In the 24-market multiplayer diagnostic, conditioning on measured rival actions changes attainable exposure by 1.50% for Qwen and 2.01% for Llama. Choosing responses separately for native and surface query-order procedures changes the best-response set in 40.3% of Qwen and 35.8% of Llama query-level aggregates; the corresponding matched values are 2.48% and 3.08%. Missing multiplayer tasks are evaluated under declared worst-completion bounds.

The expanded two-engine tensor contains 10,752 valid tasks. It originally suggested complementary weaknesses between canonical support and query-order projection, but this selected mixing advantage does not survive the four-engine or leave-one-engine-out analyses. Hidden- and revealed-route engine conditioning are likewise harmful on average, and a separate four-engine conditioning fit is unexceptional under within-market engine-label permutation ($p = 0.169$).

On the 21-policy capability frontier, the unconstrained fitted solution fails its selected-quality gate and the dual-view procedure is dominated. The later clean-traffic utility study sharpens this diagnosis: the whitespace and structural normalizers with grounded screening are utility-equivalent to identity at one engine call, query-order normalization has measurable clean cost, and the eight-view aggregator loses 0.110 static-quality units while abstaining on 27.6% of cases.

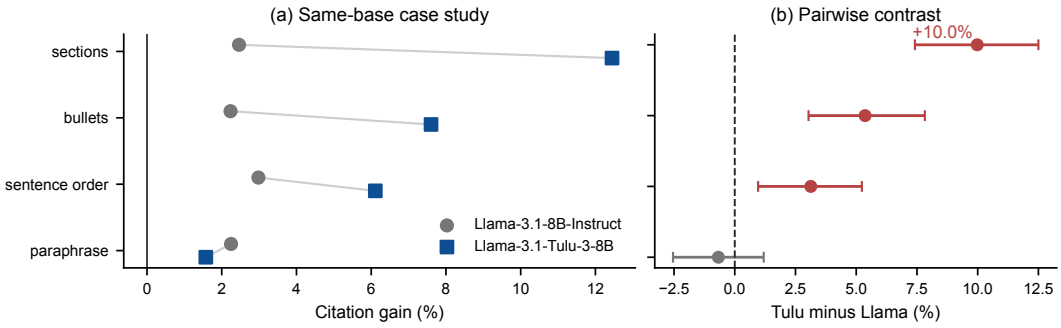

Figure 11: Exploratory same-base case study under the native policy (two systems; 140 ASQA development markets; paired source orders averaged). Part (b) shows 95% pointwise market-clustered bootstrap intervals without simultaneous correction. The comparison identifies a system-level association, not a causal post-training-stage effect.

Table 22: Construct and pre-policy delivery checks from the 140-market, four-engine surface tensor. Generated-rewrite support preservation requires a separate blinded audit only if generated rewrites enter the headline action family.

Validation	Observed result	Interpretation
Whitespace control	0.00%	formatting-triviality is inert
Retrieved-support addition	+4.48%	metric rewards genuine support
Support removal	−9.06%	metric penalizes lost support
Strict citation validity	99.994%	transport failure is negligible
Pre-policy presence after query-order	bullets 0%; sections 0%; paraphrase 100%	two surface edits removed; paraphrase remains

D.16 Every quoted contrast and level

The 35B levels in Table 24 concern deletion-only development data; Appendix D.8 reports the paired interval.

Table 23: Reported contrasts with their policies, edit sets, panels, endpoints and original interval endpoints. Exposure uses %: $C = 1$ is binary citation gain and $C = 3$ is reciprocal-rank exposure over three positions. Levels and rates appear in Table 24.

procedure	response set; endpoint	market set / engines	point	95% interval	status
no defense	label-inclusive; gap, C=1	P440, 4 eng.	-4.81	$[-6.34, -3.24]$	development
committed (deletion+screen)	label-inclusive; gap, C=1	same	$+5.24$	$[3.79, 6.80]$	development
committed (deletion+screen)	label-inclusive; gap, C=1	D140	$+5.37$	$[2.93, 7.88]$	development
committed (deletion+screen)	label-inclusive; gap, C=1	R300	$+5.18$	$[3.34, 7.01]$	development
deletion alone	label-inclusive; gap, C=1	P440, 4 eng.	$+5.54$	$[4.03, 7.06]$	mechanism
committed	label in/out; price of omitted label	P440	-3.03	$[-3.51, -2.48]$	development
committed	support+label vs support; stacking premium, C=1	same	$+3.05$	$[2.49, 3.63]$	development
committed vs no defense	eight presentation edits; gap advantage, C=3	48 mkt, 2 eng. (B)	$+4.84$	$[2.34, 7.88]$	development
committed vs whitespace + screen	same; gap advantage, C=3	same	$+1.70$	$[-0.19, 3.18]$	development
no defense → committed	oracle searched vs committed; residual reduction, C=3	48 mkt, 2 eng. (A)	$+4.44$	$[0.37, 8.52]$	development
committed	oracle response set; residual exploitability, C=3	same	$+1.18$	$[0.00, 3.59]$	development
label repair (candidate)	oracle response set; residual exploitability, C=3	same	$+1.60$	$[0.00, 4.06]$	candidate
committed − label repair	oracle response set; paired residual difference, C=3	same	-0.42	$[-1.74, 0.91]$	commit decision
committed − no defense	clean traffic; clean quality, margin 0.025	48 mkt, tulu3	-0.33	$[-2.33, 1.08]$	noninferior (Holm)
committed − no defense	clean traffic; clean quality, margin 0.025	48 mkt, qwen2.5	-1.81	$[-5.24, 0.00]$	not established
no defense	universal anchors; largest deviation at all-adopt, C=3	48 mkt, 2 eng. (B)	$+7.05$	$[6.12, 8.04]$	development
committed	universal anchors; largest deviation at all-adopt, C=3	same	$+3.97$	$[3.16, 4.83]$	development
label repair (candidate)	label-inclusive; gap, C=1	D140 (128 mkt)	$+8.34$	$[5.62, 10.80]$	candidate
label repair − committed	label-inclusive; paired gap difference	same	$+3.16$	$[0.50, 5.37]$	candidate
label repair (candidate)	support+label vs support; stacking premium	same	$+0.00$	$[-1.19, 0.08]$	structural zero
v2 deletion arm	label + published-GEO; union-set gap, C=1	R300, 4 eng.	$+1.12$	$[0.10, 2.06]$	conf. F; disconf. F
v2 deletion arm	union − label response set; paired difference	R300	$+0.00$	$[-1.63, 1.12]$	does not resolve
deletion alone vs no defense	label evasions; gap advantage, C=3	48 mkt, 2 eng. (B)	$+3.40$	$[0.94, 6.06]$	mechanism
deletion alone vs whitespace	same; gap advantage, C=3	same	$+0.30$	$[-0.99, 1.46]$	mechanism
committed	oracle searched response set; residual (max premium), C=3	prospective 200 mkt, 4 eng.	$+0.51$	$[0.00, 1.31]$	registered; P1 holds on 4/4
no defense → committed	oracle searched response set; paired reduction, C=3	same	$+4.48$	$[3.09, 5.86]$	registered; pooled
committed	pure manipulation vs support; difference, C=3 (one-sided)	same	-7.02	≤ -5.43	registered; P2 confirmed 4/4
committed	admissible + unrestricted response sets; residual (max premium), C=3	same	$+2.32$	$[1.60, 3.11]$	registered; pooled under cap
committed	admissible + unrestricted, Qwen2.5-7B; residual (max premium), C=3	same	$+3.84$	$[2.51, 5.25]$	registered; bound not established
committed vs whitespace	oracle searched response set (2 columns); paired residual reduction, C=3	48 mkt, 2 eng. (A)	$+3.94$	$[1.43, 6.45]$	supplementary
committed vs whitespace	same; signed gap advantage, C=3	same	$+3.94$	$[1.61, 6.27]$	supplementary
committed vs whitespace + screen	label-inclusive; paired gap advantage, C=1	P440, 4 eng.	$+2.44$	$[0.95, 3.93]$	supplementary; reanalysis
committed vs whitespace + screen	label-inclusive, strict common context; paired gap advantage, C=1	same	$+2.42$	$[0.93, 3.90]$	supplementary; sensitivity
whitespace	label-inclusive; gap, C=1	same	$+3.10$	$[1.39, 4.44]$	comparator
whitespace + screen	label-inclusive; gap, C=1	same	$+2.80$	$[1.25, 4.33]$	comparator
committed vs no defense	eight presentation edits; gap advantage, C=1	48 mkt, 2 eng. (B)	$+3.85$	$[1.46, 6.77]$	development
committed vs no defense	eight presentation edits; gap advantage, C=5	same	$+4.81$	$[2.40, 7.82]$	development
committed vs whitespace + screen	eight presentation edits; gap advantage, C=1	same	$+1.46$	$[-0.21, 3.12]$	development
committed vs whitespace + screen	eight presentation edits; gap advantage, C=5	same	$+1.84$	$[0.00, 3.31]$	development
whitespace	universal anchors; largest deviation at all-adopt, C=3	same	$+6.56$	$[5.65, 7.51]$	comparator
whitespace → committed	universal anchors; paired reduction, C=3	same	$+2.59$	$[1.96, 3.19]$	development
label repair (candidate)	developed response set; value over deletion alone	48 mkt, Qwen2.5-7B left out	$+3.33$	$[1.63, 5.35]$	candidate
label repair (candidate)	developed response set; value over deletion alone	48 mkt, Tulu3-8B left out	$+2.74$	$[1.25, 4.51]$	candidate
deletion alone	certified label evasions; gap, C=3	48 mkt, 2 eng. (B)	$+8.50$	$[4.74, 12.31]$	mechanism
deletion alone	bracket label; survivor payoff, C=3	same	$+1.37$	$[0.12, 2.85]$	mechanism
deletion alone	upper-case label; survivor payoff, C=3	same	$+1.15$	$[0.07, 2.40]$	mechanism
deletion alone	separator label; survivor payoff, C=3	same	$+0.64$	$[-0.56, 1.96]$	mechanism
normalizer family (structural)	nine edits; gap, C=1	D140, 4 eng.	$+4.65$	$[1.98, 7.40]$	historical
normalizer family (whitespace)	nine edits; gap, C=1	same	$+4.69$	$[2.03, 7.42]$	historical

Table 24: Reported levels, rates and differences, with their policies, edit sets and panels. Rates count tasks or selections; timing measures CPU preprocessing, not end-to-end latency. The 35B rows concern deletion-only development data. Paired contrasts and their intervals appear in Table 23.

procedure	response set; quantity	market set / engines	value	kind
no defense	oracle searched response set; max response premium, C=3	prospective 200 mkt, 4 eng.	$+4.99\%$	registered; level
no defense	oracle searched vs committed; residual exploitability, C=3	48 mkt, 2 eng. (A)	$+5.63\%$	level; the paired reduction has the interval
committed − deletion alone	label-inclusive; screen cost, C=1	P440, 4 eng.	-0.30%	difference of levels
committed − whitespace + screen	label-inclusive; gap, C=1	same	$+2.44\%$	difference of levels
no defense	planner best response; response lift, C=1	same	$+11.42\%$	level
whitespace	planner best response; response lift, C=1	same	$+5.88\%$	level
deletion alone	planner best response; response lift, C=1	same	-0.29%	level
committed	clean traffic; clean selections rewritten	D140	4.02%	rate
no defense	universal adoption, bullets; citation sets moved	48 mkt, 2 eng. (B)	16.2%	rate
no defense	universal adoption, procedure-aware edit; citation sets moved	same	26.1%	rate
committed	universal adoption, bullets; citation sets moved	same	0.7%	rate
no defense	universal anchors; tasks where a publisher is strictly forced to manipulate	same	27.6%	rate
committed	universal anchors; tasks where a publisher is strictly forced to manipulate	same	10%	rate
native	mechanical edits; mean unilateral-deviation regret, C=3	48 mkt, 35B MoE	$+5.06\%$	development; level
deletion step	mechanical edits; mean unilateral-deviation regret, C=3	same	$+2.09\%$	development; component only
committed	Final48 inputs; warm median preprocessing time	960 documents, CPU	0.75 ms	timing; zero model calls