# STAY TRUE TO YOUR ROOTS: REFERENCE-ANCHORED POLICY OPTIMIZATION FOR FAITHFUL GEO

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generative search is redefining web visibility: being retrieved no longer suffices if a source goes unrepresented in the LLM-generated answer. This creates a new optimization problem for content creators—improving visibility among competing sources without compromising faithfulness to the original content. Existing generative engine optimization (GEO) approaches largely optimize visibility in isolation, neglecting faithfulness, and rely on manually designed edit instructions with limited coverage. We introduce **Reference-Anchored Policy Optimization (RAPO)**, a reinforcement learning framework that generates multiple candidate rewrites but measures progress against the original source rather than other sampled rewrites. RAPO anchors advantage estimation to the source reward, so positive signals reflect improvement over the original content, while retaining within-group standard-deviation normalization for stable scaling. To capture the competitive nature of generative search, we introduce a competition-aware visibility signal that evaluates a source against its strongest competitor and combine it with faithfulness in a composite reward. Across GEO-Bench, Researchy-GEO, and E-commerce, RAPO improves Overall visibility over the source by 9.34, 15.82, and 7.15 points with better faithfulness, outperforming prompt-based and trainable RL baselines. Ablations and analysis show source anchoring is essential, RAPO generalizes across rewriter model scales and generative engines, and benefits sources with lower rank in LLM context most, while varying the faithfulness weight reveals a controllable visibility–faithfulness trade-off. Code is released at this repository.

## 1 INTRODUCTION

How users access web information is undergoing a structural shift. Traditional search returns a ranked list of documents that users read and synthesize; generative search engines, (ChatGPT, Perplexity, Bing Copilot, Google AI Overviews), produce a natural-language answer quoting, synthesizing, and attributing retrieved sources (Aggarwal et al., 2024). These systems build on retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023b): a retriever surfaces a passage set and an LLM generates a conditioned response. Retrieved documents thus become intermediate context, not final output, with only a fraction represented in the response (Aggarwal et al., 2024; Liu et al., 2024). We term the extent a source is represented and attributed in this response its *visibility*. For websites and creators, a document may be retrieved yet receive little or no representation in the answer. Thus, visibility in search depends not only on retrieval, but on how prominently a source enters response.

*Generative Engine Optimization* (GEO) was formalized by Aggarwal et al. (2024), showing stylistic edits such as citations, statistics, or quotations can increase visibility in LLM-generated answers by up to 40%. Subsequent work extends GEO to specific domains and automated rewriting pipelines (Bagga et al., 2025; Wu et al., 2026; Chen et al., 2025), without explicit faithfulness consideration. Content modifications can materially influence which sources are surfaced; in adversarial settings, text injections can even cause generative engines to preferentially cite malicious sources, with selection rates rising near zero to over 90% (Nestaas et al., 2025), making principled optimization crucial. Yet existing approaches largely rely on hand-designed heuristics or prompt- and agent-based pipelines with limited coverage, lacking a general framework for defining and learning desirable edits.

A principled framework must address three optimization challenges. **(1) Faithfulness:** prompt-based strategies encouraging citation-promoting features (statistics, quotations, authoritative language) can



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

introduce fabricated or unsupported content without an explicit faithfulness constraint, consistent with known hallucination and factual-consistency risks (Ji et al., 2023; Lin et al., 2022; Kryscinski et al., 2020; Laban et al., 2022; Min et al., 2023; Zhang et al., 2020). **(2) Competition:** sources appear alongside other retrieved documents competing for attribution, so optimizing a target's visibility in isolation ignores selection context. **(3) Joint optimization:** improving competitive visibility while preserving original content requires managing the visibility-faithfulness interaction. Figure 1 illustrates this trade-off: effective GEO should increase visibility without sacrificing content fidelity.

These challenges motivate two complementary notions of improvement, subject to source faithfulness. First, visibility should improve relative to competing sources, not in isolation. We capture this via a competitor-aware signal comparing target visibility with its strongest competing document, reflecting stronger attribution where the source competes. Second, an optimized edit should improve upon the *original source*, a natural progress reference. Reinforcement learning implements this via candidate rewrites optimized using non-differentiable generative-engine and faithfulness feedback (Schulman et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020). However, critic-free methods such as GRPO (Shao et al., 2024; Guo et al., 2025) measure progress relative to sampled candidates: when all are worse than original, least-bad one can receive positive advantage misaligning group-relative improvement with source editing.

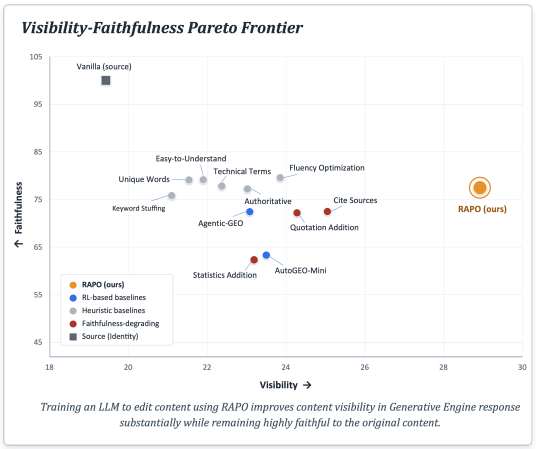


Figure 1: RAPO achieves higher visibility while maintaining strong faithfulness, yielding a better visibility–faithfulness trade-off than prior methods.

We propose **Reference-Anchored Policy Optimization (RAPO)**, a new RL framework for faithful content optimization in generative search. RAPO generates candidate rewrites but anchors their advantages to the original source's reward (competition-aware visibility and faithfulness) rather than the rollout mean, retaining within-group standard-deviation normalization for stability. Its composite reward combines competition-aware visibility with faithfulness to the source: the former rewards visibility gains against the strongest competitor, the latter discouraging gains from factual drift. RAPO thus separates two distinct comparisons: the reward measures performance against competing sources, while the advantage separately measures whether the rewrite improves upon its original source.

Our contributions are threefold. First, we formulate GEO as competition-aware, faithful content optimization via a signal comparing competitors. Second, we introduce RAPO, replacing group-relative centering with a source-anchored reference. Third, we evaluate RAPO on three benchmarks (GEO-Bench, Researchy-GEO, E-commerce), improving Overall visibility over the source by 9.34, 15.82, and 7.15 points, while maintaining faithfulness and outperforming prompt-based and trainable RL baselines; sources with lower rank in LLM context benefit most. Ablations confirm source anchoring and faithfulness weighting act synergistically, RAPO generalizes across optimizer model scales and generative engines. We show RAPO's edits can be distilled into reusable GEO guidelines.

## 2 Related Work

**Generative Engine Optimization.** The shift from keyword-based to generative, retrieval-augmented search (ChatGPT, Perplexity, Bing Copilot, Google AI Overviews) motivates Generative Engine Optimization (GEO), raising citability. Aggarwal et al. (2024) introduced GEO-Bench and nine strategies (citations, statistics, persuasive language) yielding up to 40% visibility gains but no general framework. Subsequent work explores stylistic optimization (AutoGEO; Wu et al., 2026), latent-intent modeling (RAID G-SEO; Chen et al., 2026b), critic-rewriter (AgenticGEO; Yuan et al., 2026), transformer fine-tuning (Lüttgenau et al., 2025), retrieval-aware rewriting (Ho et al., 2025), multimodal caption injection (Chen et al., 2026a), multi-query optimization (Anonymous, 2026), and

appended-content ranking (Jin et al., 2026). Adversarial GEO raises citation rates from near zero to over 90% via injected text (Nestaas et al., 2025); Chen et al. (2025) introduces content-centric evaluation of source influence. Two limitations arise: visibility optimization risks factual drift without faithfulness enforcement, and heuristic/prompt-based strategies limit edits manually. A learnable objective discovers visibility-enhancing edits while enforcing faithfulness. We examine this by distilling RAPO rollout edits into reusable guidelines, a data-driven counterpart to manual instructions.

**Reinforcement Learning for Text Optimization.** Reinforcement learning (RL) aligns language models under non-differentiable objectives. RLHF learns rewards from human preferences, optimized via PPO (Ouyang et al., 2022; Schulman et al., 2017) for text generation (Stiennon et al., 2020); work replaces feedback with AI (RLAIF) (Bai et al., 2022) or removes RL via DPO (Rafailov et al., 2023). Critic-free methods such as REINFORCE (Ahmadian et al., 2024) and GRPO (Shao et al., 2024), scaled by DeepSeek-R1 (Guo et al., 2025)), compute advantages relative to sampled groups and performs comparably. Such objectives misalign with content optimization for generative engines: progress should reflect improvement over the source, not merely outperforming sampled rewrites, since group-relative methods favor the best rewrite when all are worse. RAPO anchors advantages to the original document's reward, the improvement/regression criterion. Prior RL-based rewriting methods (RLPrompt (Deng et al., 2022), RewriteLM (Shu et al., 2024), PRewrite (Kong et al., 2024), argumentation rewriting (Ziegenbein et al., 2024)) optimize a single objective, risking over-optimization (Rita et al., 2024), encouraging fabrication while drifting from source. RAPO combines a Bradley–Terry-inspired competition-aware visibility signal with entailment-based faithfulness, as entailment reduces factual drift (Roit et al., 2023). Source-anchored advantage estimation with competition- and faithfulness-aware rewards forms a GEO-specific RL formulation for optimization.

**Faithfulness and Hallucination in LLM Pipelines.** GEO operates within RAG pipelines (Lewis et al., 2020; Gao et al., 2023b), where dense retrievers (Karpukhin et al., 2020; Merrick et al., 2024) surface documents that condition LLM generation. Source utilization is sensitive to context ordering: LLMs exhibit lost-in-the-middle effects (Liu et al., 2024) and systematic position (Wang et al., 2024) and citation (Gao et al., 2023a) biases toward early or stylistically prominent sources. Such effects make it important for GEO improvements to remain effective across source positions within the retrieved context; we examine this by stratifying RAPO's visibility gains by source rank. Faithfulness poses a complementary challenge: hallucination remains prevalent despite retrieval grounding (Ji et al., 2023; Ren et al., 2025; Liu et al., 2023a). Factual consistency can be assessed through embedding similarity (Zhang et al., 2020), NLI-based scoring (Kryscinski et al., 2020; Laban et al., 2022), atomic-claim decomposition (Min et al., 2023), and LLM-judge frameworks such as G-Eval and DeepEval (Liu et al., 2023b; AI, 2024). RAPO incorporates faithfulness into its competition-aware reward, discouraging visibility gains from factual drift from the original source.

## 3 METHODOLOGY

We introduce **Reference-Anchored Policy Optimization (RAPO)**, an RL framework for faithful content editing in GEO. Given a source document $x$, RAPO learns a rewriting policy that improves visibility relative to competing sources while remaining grounded in the source. As illustrated in Fig. 2, RAPO evaluates a group of $G$ candidate rewrites via competition-aware visibility and faithfulness, anchoring their advantages to a cached reward of the original source rather than the sampled-group mean. This ensures policy updates reflect improvement over the source, while within-group standard-deviation normalization controls the learning signal's scale. § 3.1 defines optimization objective, § 3.2 derives reference-anchored advantage followed by discussion on training procedure.

### 3.1 FAITHFUL CONTENT EDITING AS CONSTRAINED OPTIMIZATION

Let $\mathcal{X}$ denote the space of textual documents and $\mathcal{Q}$ a distribution over queries associated with a given source. For a source document $x \in \mathcal{X}$, we model the editor as a stochastic policy $\pi_\theta(\tilde{x} \mid x)$ generating a rewritten document $\tilde{x} \in \mathcal{X}$. The rewrite interacts with two frozen external components: a generative engine $\mathcal{E}$, producing an answer from retrieved documents, and a faithfulness judge $\mathcal{J}$, evaluating semantic consistency between $\tilde{x}$ and $x$. Both are black-box and non-differentiable in $\theta$ but provide scalar feedback per rewrite, enabling reinforcement learning of $\pi_\theta$.

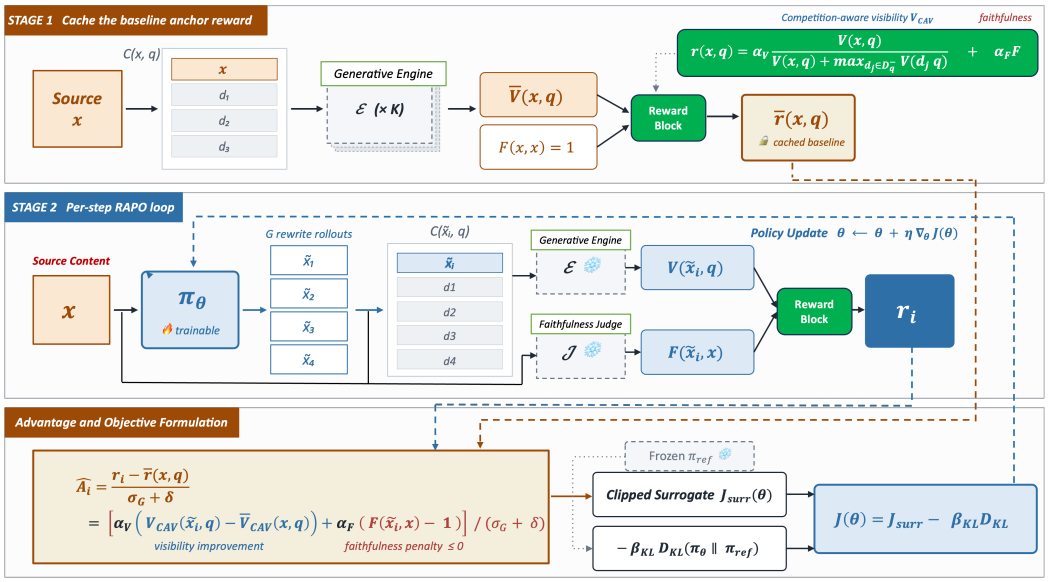


Figure 2: Reference-Anchored Policy Optimization (RAPO) training pipeline. **Stage 1:** Compute and cache a reference reward $\bar{r}(x, q)$ for each source, combining competition-aware visibility $\bar{V}(x, q)$ with its fixed faithfulness score. **Stage 2:** The policy $\pi_\theta$ generates $G$ rewrites $\tilde{x}_i$, evaluated by the generative engine (visibility) and faithfulness judge to obtain rewards $r_i$. Advantages are centered on the cached source baseline, normalized by within-group reward standard deviation, $\hat{A}_i = \big(r_i - \bar{r}(x, q)\big)/(\sigma_G + \delta)$, and used in a clipped, KL-regularized objective for stable updates.

For each query $q$, let $\mathcal{D}q^- = d_1, \ldots, dN - 1$ denote $N - 1$ competitor documents drawn from the same retrieval corpus as $x$, where $N$ is the retrieval context size. We define $\mathcal{C}(d, q) = d \cup \mathcal{D}_q^-$, where target $d$, either the source $x$ or a rewrite $\tilde{x}$, is the only varying document. The engine produces $a = \mathcal{E}(q, \mathcal{C}(d, q))$, abbreviated as $\mathcal{E}(q, d)$. Construction of $\mathcal{D}_q^-$ follows the GEO-Bench (Aggarwal et al., 2024) protocol, detailed in §4. Keeping competing documents fixed across the source and its $G$ rewrites lets us attribute changes in the target's visibility to its content, not retrieval-side variation.

**Competition-aware visibility.** We first measure the *raw visibility* $V(d, q; \mathcal{E}) \in [0, 1]$ of document $d$ via the position-aware GEO-Bench metric (Aggarwal et al., 2024), quantifying how prominently $d$ is cited in $\mathcal{E}$'s answer. Because multiple documents in $\mathcal{C}(d, q)$ compete for attribution in the response, raw visibility alone ignores the visibility of competing sources. We therefore define a *competition-aware visibility* score $V_{\text{CAV}}$ comparing the target against its strongest competitor:

$$V_{\text{CAV}}(d, q; \mathcal{E}) = \frac{V(d, q; \mathcal{E})}{V(d, q; \mathcal{E}) + \max_{d_j \in \mathcal{D}_q^-} V(d_j, q; \mathcal{E})}, \tag{1}$$

with documents evaluated in the same answer under $\mathcal{C}(d, q)$. The score lies in $[0, 1]$: $V_{\text{CAV}} > 0.5$ means the target beats every competitor; $V_{\text{CAV}} < 0.5$ means at least one competitor beats it. $V_{\text{CAV}}$ measures the target's visibility against its strongest competitor, not in isolation.

**Faithfulness.** The second component is *faithfulness* $F(\tilde{x}, x; \mathcal{J}) \in [0, 1]$, the probability under $\mathcal{J}$ that $\tilde{x}$ contains no hallucinated content with respect to $x$, with higher values indicating stronger semantic consistency. By definition, $F(x, x; \mathcal{J}) \equiv 1$, so $F(\tilde{x}, x; \mathcal{J}) - 1 \in [-1, 0]$ represents a faithfulness deficit relative to the source. Both signals are evaluated at the sequence level over the full rewrite. Combining competition-aware visibility with faithfulness constraint, we formulate faithful editing as $\max_\theta \ \mathbb{E}_{q\sim\mathcal{Q},\, \tilde{x}\sim\pi_\theta(\cdot|x)}[V_{\text{CAV}}(\tilde{x}, q; \mathcal{E})]$ s.t. $\mathbb{E}_{q\sim\mathcal{Q},\, \tilde{x}\sim\pi_\theta(\cdot|x)}[F(\tilde{x}, x; \mathcal{J})] \geq \tau$, where $\tau \in [0, 1]$ is a faithfulness threshold, ensuring optimized visibility while rewrites remain grounded in the source. Since the above direct primal–dual optimization is impractical at LLM scale, given the cost of generating and scoring rewrites, we relax it with a Lagrange multiplier $\lambda \geq 0$:

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}[V_{\text{CAV}}(\tilde{x}, q; \mathcal{E})] + \lambda\big(\mathbb{E}[F(\tilde{x}, x; \mathcal{J})] - \tau\big).$$

For fixed $\lambda$, the term $-\lambda\tau$ is constant in $\theta$. Rather than performing dual ascent over $\lambda$, we use the corresponding weighted scalarization with fixed hyperparameters $\alpha_V, \alpha_F \geq 0$, where $\alpha_V$ scales visibility and $\alpha_F$ scales faithfulness (only their relative weighting matters for the policy update):

$$r(\tilde{x}, x, q) = \alpha_V \, V_{\mathrm{CAV}}(\tilde{x}, q; \mathcal{E}) + \alpha_F \, F(\tilde{x}, x; \mathcal{J}). \tag{2}$$

The ratio $\alpha_V/\alpha_F$ controls the operating point on competition-aware-visibility–faithfulness trade-off. The reward in Eq. 2 evaluates a rewrite in the competitive context, but content editing also requires measuring improvement over the original document. Since $r(x, x, q)$ is independent of $\theta$, maximizing $\mathbb{E}_{\tilde{x}\sim\pi_\theta(\cdot|x)}[r(\tilde{x}, x, q)]$ is equivalent, up to a constant, to maximizing $\mathbb{E}_{\tilde{x}\sim\pi_\theta(\cdot|x)}\left[r(\tilde{x}, x, q) - r(x, x, q)\right]$, where $r(x, x, q)$ is source's reward under competitive context. Section 3.2 takes this adjustment by source reward constant into consideration and uses source-relative improvement for the RAPO advantage with within-group standard-deviation normalization.

### 3.2 Reference-Anchored Policy Optimization

GRPO (Shao et al., 2024) avoids the cost of a value critic by sampling, for each input, a group of $G$ candidates $\{\tilde{x}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ and normalizing their rewards within the group:

$$\hat{A}_i^{\text{GRPO}} = \frac{r_i - \mu_G}{\sigma_G + \delta}, \quad \mu_G = \tfrac{1}{G}\sum_{j=1}^{G} r_j, \quad \sigma_G^2 = \tfrac{1}{G}\sum_{j=1}^{G}(r_j - \mu_G)^2, \tag{3}$$

where $r_i = r(\tilde{x}_i, x, q)$ is the competition- and faithfulness-aware reward from Eq. 2, and $\delta > 0$ ensures stability. GRPO suits ranking-style tasks where progress is relative to sampled outputs. Content editing requires measuring improvement over the source. Centering on group mean $\mu_G$ violates source-relative notion in two cases: **(i)** when all rewrites improve over the source, below-average rewrites receive negative advantages despite being improvements; and **(ii)** when all rewrites are worse than the source, above-average rewrites receive positive advantages despite being regressions. Since these failures arise from group-mean centering, not scale normalization, RAPO replaces $\mu_G$ with a fixed source-derived reference while retaining $\sigma_G$ for within-group normalization.

**Source-derived reference reward.** We reserve *source* for the original document $x$ and *reference reward* for its cached baseline $\bar{r}(x, q)$. Since the source is perfectly faithful to itself, $F(x, x; \mathcal{J}) \equiv 1$, its faithfulness contribution is the constant $\alpha_F$. Its competition-aware visibility, however, depends on the response produced by $\mathcal{E}$. Because $\mathcal{E}$ generates stochastically (e.g., via sampling), repeated evaluations of the same source and context can yield different visibility scores. We therefore estimate raw visibility of the source and each competitor by averaging over $K$ independent runs:

$$\bar{V}(d, q) = \frac{1}{K}\sum_{k=1}^{K} V^{(k)}(d, q; \mathcal{E}), \qquad d \in \{x\} \cup \mathcal{D}_q^{-}. \tag{4}$$

From averaged visibility estimates, we compute competition-aware visibility and reward of source as:

$$\bar{V}_{\mathrm{CAV}}(x, q) = \frac{\bar{V}(x, q)}{\bar{V}(x, q) + \max_{d_j \in \mathcal{D}_q^{-}} \bar{V}(d_j, q)}. \tag{5}$$

$$\bar{r}(x, q) = \alpha_V \, \bar{V}_{\mathrm{CAV}}(x, q) + \alpha_F. \tag{6}$$

$\bar{r}(x, q)$ provides a source-derived reference against which rewrites for the $(x, q)$ pair are evaluated.

**Reference-anchored advantage.** The reference reward is precomputed and cached, so evaluating source costs nothing during updates; $K$-run averaging reduces visibility stochasticity. For each rewrite $\tilde{x}i \sim \pi\theta_{\text{old}}(\cdot \mid x)$, RAPO centers on reference, normalizes by within-group standard deviation:

$$\hat{A}_i = \frac{r_i - \bar{r}(x, q)}{\sigma_G + \delta}, \qquad r_i = r(\tilde{x}_i, x, q), \tag{7}$$

where $\sigma_G$ is defined in Eq. 3. Since $\sigma_G + \delta > 0$, normalization rescales source-relative improvement, preserving sign: $\hat{A}_i > 0$ iff $r_i > \bar{r}(x, q)$, and $\hat{A}_i < 0$ iff $r_i < \bar{r}(x, q)$. Substituting Eqs. 2, 6 in Eq. 7:

$$\hat{A}_i = \frac{\underbrace{\alpha_V\left[V_{\mathrm{CAV}}(\tilde{x}_i, q; \mathcal{E}) - \bar{V}_{\mathrm{CAV}}(x, q)\right]}_{\text{competitive visibility improvement}} + \underbrace{\alpha_F\left[F(\tilde{x}_i, x; \mathcal{J}) - 1\right]}_{\text{faithfulness penalty} \leq 0}}{\sigma_G + \delta}, \tag{8}$$

where the numerator decomposes into visibility improvement over the source and a non-positive faithfulness penalty vanishing for fully faithful rewrites. The denominator scales this signal without changing whether the rewrite is an improvement or regression. Thus, RAPO uses the original document to determine the *direction* of the signal, and the rewrite group to normalize its *magnitude*.

The cached reference reward $\bar{r}(x, q)$ is computed from the original source before policy optimization and is therefore fixed with respect to the trainable policy parameters $\theta$; for a given $(x, q)$, it also does not depend on which rewrite $\tilde{x}_i$ is sampled. It can therefore be subtracted as a baseline without changing the expected policy gradient. Specifically,

$$\mathbb{E}_{\tilde{x}_i \sim \pi_\theta(\cdot|x)} \left[\nabla_\theta \log \pi_\theta(\tilde{x}_i \mid x)\, \bar{r}(x, q)\right] = \bar{r}(x, q)\, \nabla_\theta \sum_{\tilde{x}_i} \pi_\theta(\tilde{x}_i \mid x) = \bar{r}(x, q)\, \nabla_\theta 1 = 0. \quad (9)$$

The second equality follows from $\sum_{\tilde{x}_i} \pi_\theta(\tilde{x}_i \mid x) = 1$. Source anchoring preserves the expected policy gradient while providing a stable, source-specific centering term; when correlated with rewrite rewards, this baseline reduces gradient variance. RAPO applies $(\sigma_G + \delta)^{-1}$ as group-dependent normalization. Consequently, unlike GRPO, the cached reference determines whether a rewrite is an improvement, while the sampled group controls only the magnitude of its normalized advantage.

**Training Objective and Algorithm.** Having defined the RAPO advantage in Eq. 7, we otherwise follow the standard clipped objective used in PPO/GRPO (Schulman et al., 2017; Shao et al., 2024). Let $\rho_i(\theta) = \pi_\theta(\tilde{x}i \mid x)/\pi\theta_{\text{old}}(\tilde{x}_i \mid x)$ denote the sequence-level importance ratio between the current and old policy used to sample the $G$ rewrites. The per-instance surrogate objective is $\mathcal{J}_{\text{surr}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_i(\theta)\hat{A}_i,\ \text{clip}\big(\rho_i(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\Big)$, where $\epsilon \in (0, 1)$ is the clipping range and $\hat{A}_i$ is the reference-anchored, standard-deviation-normalized advantage from Eq. 7. Following GRPO, we additionally regularize against a frozen reference policy $\pi_{\text{ref}}$ initialized from the same checkpoint as $\pi_\theta$, where $\beta_{\text{KL}} > 0$ controls regularization strength. $\mathcal{J}(\theta) = \mathcal{J}_{\text{surr}}(\theta) - \beta_{\text{KL}}\, \mathbb{D}_{\text{KL}}\big[\pi_\theta(\cdot \mid x) \,\big\|\, \pi_{\text{ref}}(\cdot \mid x)\big]$, With $\mathcal{D}$ as training distribution and $G$ as rollout group size, the policy is trained with $\theta^\star = \arg\max_\theta\ \mathbb{E}_{(x,q)\sim\mathcal{D}}\, \mathbb{E}_{\{\tilde{x}_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\mathcal{J}(\theta)\right]$. Training follows the two stages in Figure 2. **Stage 1:** For each $(x, q)$, we cache $\bar{r}(x, q)$ using $K$ engine evaluations (Eqs. 4–6). **Stage 2:** We sample $G$ rewrites, evaluate rewards using $\mathcal{E}$ and $\mathcal{J}$, compute $\sigma_G$ and the RAPO advantages (Eq. 7), and optimize $\theta$. Since reference reward is cached and $V_{\text{CAV}}$ reuses the reward-scoring evaluation, RAPO adds no generative-engine calls during policy updates and negligible normalization overhead.

## 4 Experimental Details

**Datasets and Evaluation.** We evaluate on three GEO benchmarks, each pairing queries with five candidate documents: **GEO-Bench** (Aggarwal et al., 2024), diverse open-domain queries; **Researchy-GEO**, non-factoid, multi-perspective Bing queries from Researchy Questions (Rosset et al., 2024), per (Wu et al., 2026); and **E-commerce**, product queries for commercial-search GEO. Following (Aggarwal et al., 2024), we report three visibility metrics: **Word** (normalized response-word count in citing sentences), **Pos** (position-weighted, exponentially decaying later sentences), and **Overall** (combined position-adjusted word-count score). **Faith.** uses DeepEval's (AI, 2024) faithfulness metric, measuring consistency of optimized content with the source. We instantiate $\mathcal{E}$ with GPT-4o-mini and context size $N$=5. Dataset and metric details are in Appendix A.1, A.2.

**Baselines.** We compare against (i) **Vanilla** unmodified source; (ii) nine prompt-based strategies from GEO-Bench (Aggarwal et al., 2024) (Technical Terms, Cite Sources, Keyword Stuffing, Unique Words, Authoritative, Easy-to-Understand, Statistics Addition, Quotation Addition, and Fluency Optimization), via GPT-4o, plus stronger **AutoGEO-API** (Wu et al., 2026); and (iii) two trainable baselines: **AutoGEO-Mini** (Wu et al., 2026), a Qwen3-1.7B fine-tuned via GRPO with rule-based rewards, and **Agentic-GEO** (Yuan et al., 2026), a self-evolving system with a critic–rewriter loop.

**Implementation Details.** We fine-tune Qwen3-8B with RAPO using FSDP for training and vLLM for rollouts, on 8 A100 80GB GPUs. Each step used 6 rollouts per prompt (temperature 0.9). The actor updates with clip ratio 0.2, learning rate $1 \times 10^{-5}$, and token-level KL penalty ($\beta = 0.005$) against $\pi_{\text{ref}}$. Faithfulness is computed via DeepEval's `FaithfulnessMetric` between optimized and original source. Reference visibility is computed per prompt from 6 LLM-judge completions, cached across rollouts. The model trains for 2 epochs. Appendix A.3 contains further details.

Table 1: Comparison of RAPO against baselines on GEO-Bench, Researchy-GEO and E-commerce using GPT-4o-mini as generative engine. **Word**, **Pos**, and **Overall** are standard impression metrics from (Aggarwal et al., 2024). **Faith.** is measuring consistency with original content.

| | GEO-Bench | | | | Researchy-GEO | | | | E-commerce | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** |
| Vanilla | 19.43 | 19.21 | 19.50 | 100 | 19.84 | 20.10 | 19.70 | 100 | 19.21 | 19.16 | 19.12 | 100 |
| *Prompt based Approaches using GPT-4o* | | | | | | | | | | | | |
| Technical Terms | 23.02 | 23.08 | 23.11 | 77.20 | 23.57 | 23.97 | 23.86 | 80.58 | 22.34 | 22.44 | 22.34 | 66.13 |
| Cite Sources | 25.05 | 23.03 | 24.05 | 72.43 | 25.16 | 25.40 | 25.28 | 72.95 | 22.95 | 23.0 | 22.95 | 57.25 |
| Keyword Stuffing | 21.10 | 21.07 | 21.07 | 75.78 | 21.22 | 21.34 | 21.11 | 73.80 | 20.90 | 20.86 | 20.80 | 64.44 |
| Unique Words | 21.54 | 21.33 | 21.41 | 79.09 | 25.54 | 25.58 | 25.5 | 79.76 | 22.96 | 22.91 | 22.98 | 64.56 |
| Authoritative | 22.37 | 22.23 | 22.30 | 77.79 | 22.79 | 22.87 | 22.82 | 77.97 | 22.28 | 22.28 | 22.34 | 65.11 |
| Easy-to-Understand | 21.90 | 21.91 | 22.03 | 79.12 | 23.6 | 23.78 | 23.74 | 79.42 | 21.75 | 21.91 | 21.88 | 65.33 |
| Statistics Addition | 23.19 | 23.25 | 23.31 | 62.30 | 23.57 | 23.97 | 23.86 | 58.12 | 20.90 | 20.87 | 20.86 | 43.16 |
| Quotation Addition | 24.28 | 24.07 | 24.13 | 72.15 | 23.0 | 23.04 | 22.9 | 72.38 | 21.79 | 21.92 | 21.79 | 61.59 |
| Fluency Optimization | 23.85 | 23.83 | 23.82 | 79.56 | 26.96 | 27.06 | 27.05 | 80.61 | 23.57 | 23.63 | 23.71 | 67.06 |
| *RL-based Trainable baselines* | | | | | | | | | | | | |
| AutoGEO-Mini | 23.50 | 23.83 | 23.88 | 63.27 | 30.67 | 30.48 | 30.59 | 61.38 | 22.75 | 22.83 | 22.78 | 50.20 |
| Agentic-GEO | 23.08 | 23.26 | 23.83 | 72.40 | 26.48 | 26.51 | 26.5 | 74.38 | 23.14 | 23.23 | 23.17 | 61.38 |
| RAPO (Ours) | **28.92** | **28.73** | **28.84** | 77.45 | **35.54** | **35.61** | **35.52** | 80.32 | **26.32** | **26.11** | **26.27** | 67.49 |

## 5 Results and Analysis

**Comparison with baselines.** Table 1 shows RAPO achieves the highest visibility on all three benchmarks without faithfulness degradation of several competing approaches. On GEO-Bench, RAPO improves visibility from 19.50 (Vanilla) to **28.84** while maintaining **77.45** faithfulness; on Researchy-GEO, **35.52** visibility and **80.32** faithfulness, and on E-commerce, **26.27** and **67.49**. By contrast, baselines raising visibility incur a steep faithfulness cost: Statistics Addition and AutoGEO-Mini score only 62.30 and 63.27 faithfulness, roughly 15 points below RAPO, indicating optimizing visibility alone drifts rewrites from original content. RAPO's advantage over prompt-based methods is largest on Researchy-GEO, beating the strongest baseline by **8.47** points ($27.05 \rightarrow 35.52$), with gains on GEO-Bench ($24.13 \rightarrow 28.84$) and E-commerce ($23.71 \rightarrow 26.27$). The margin widens against trainable RL baselines, exceeding Agentic-GEO by **9.02** points on Researchy-GEO. We attribute this to RAPO's reward design, weighting faithfulness over visibility (§5); without it, RL methods can restructure or inject content, producing rewrites diverging from the source.

**Distilling RAPO into Rewriting Guidelines.** We investigate whether RAPO's training rollouts encode transferable knowledge about effective, faithful edits. We introduce **RAPO-Distill**, a post-training procedure converting successful RAPO trajectories into natural-language guidelines. We retain rewrites satisfying $V_{\mathrm{CAV}}(\tilde{x}, q) - \bar{V}_{\mathrm{CAV}}(x, q) > 0$ and $F(\tilde{x}, x; \mathcal{J}) \geq \tau$, then use an LLM (Claude-opus-4-7) to extract and consolidate patterns into guideline set $\mathcal{G}^\star$ (full extraction procedure in Appendix A.4) such as *"Lead With a Direct Definitional Thesis Sentence"* and *"Convert Prose Into Numbered, Bold-Headed Sections"* (all extracted guidelines in A.5). At inference, $\mathcal{G}^\star$ is provided to an off-the-shelf LLM rewriter (GPT-4o), requiring no trained RAPO policy. Table 2 shows RAPO-Distill raises Overall visibility from 28.84 to **30.87** on GEO-Bench and 35.52 to **36.82** on Researchy-GEO, with faithfulness from 77.45 to **86.84** and 80.32 to **88.24**.

Table 2: Effect of rollout-derived optimization guidelines. Guidelines are distilled from RAPO training rollouts and incorporated into a second-stage optimization.

| Method | Word ↑ | Pos ↑ | Overall ↑ | Faith. ↑ |
|---|---|---|---|---|
| *GEO-Bench* | | | | |
| Vanilla | 19.43 | 19.21 | 19.50 | 100 |
| RAPO | 28.92 | 28.73 | 28.84 | 77.45 |
| AutoGEO-API | 27.18 | 27.32 | 27.57 | 81.38 |
| **RAPO-Distill** | **31.12** | **30.72** | **30.87** | **86.84** |
| *Researchy-GEO* | | | | |
| Vanilla | 19.84 | 20.10 | 19.70 | 100 |
| RAPO | 35.54 | 35.61 | 35.52 | 80.32 |
| AutoGEO-API | 33.67 | 34.04 | 33.75 | 83.47 |
| **RAPO-Distill** | **37.01** | **36.94** | **36.82** | **88.24** |

**Comparison with Alternative RL Optimization Methods.** Table 3 compares RAPO against four RL optimizers under identical reward and training setups, isolating effect of advantage estimation. GRPO (Shao et al., 2024), Dr. GRPO (Liu et al., 2025), DAPO (Yu et al., 2025), and Reinforce (Hu, 2025) differ in how they normalize or baseline group-relative advantage, yet cluster within a narrow band (Overall scores 25.13–27.51 on GEO-Bench and 32.43–34.36 on Researchy-GEO). RAPOw/ std-norm outperforms all, reaching Overall scores **28.84** and **35.52**, faithfulness **77.45** on GEO-Bench and **80.32** on Researchy-GEO. Removing standard deviation normalization (RAPOw/o std-norm) drops performance confirming necessity for source-anchored advantage.

Table 3: Comparison of alternative RL optimization methods on GEO-Bench and Researchy-GEO.

| | GEO-Bench | | | | Researchy-GEO | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** |
| GRPO | 27.34 | 27.23 | 27.36 | 75.36 | 34.51 | 34.28 | 34.36 | 78.04 |
| Dr. GRPO | 25.11 | 25.02 | 25.13 | 74.97 | 32.35 | 32.36 | 32.43 | 77.18 |
| DAPO | 27.46 | 27.38 | 27.51 | 76.47 | 34.22 | 34.21 | 34.18 | 76.83 |
| Reinforce | 26.91 | 26.89 | 26.95 | 77.24 | 33.81 | 33.73 | 33.87 | 78.24 |
| $\text{RAPO}_{\text{w/o std-norm}}$ | 26.18 | 26.21 | 26.13 | 74.72 | 33.43 | 33.24 | 33.46 | 78.28 |
| $\text{RAPO}_{\text{w/ std-norm}}$ | **28.92** | **28.73** | **28.84** | **77.45** | **35.54** | **35.61** | **35.52** | **80.32** |

Table 4: RAPO generalizes across generative-engine LLMs on GEO-Bench and Researchy-GEO.

| **Generative Engine** | **Method** | **GEO-Bench** | | | | **Researchy-GEO** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Word↑** | **Pos↑** | **Overall↑** | **Faith↑** | **Word↑** | **Pos↑** | **Overall↑** | **Faith.↑** |
| **GPT-4o-mini** | Vanilla | 19.43 | 19.21 | 19.50 | 100 | 19.84 | 20.10 | 19.70 | 100 |
| | Quotation Addition | 24.28 | 24.07 | 24.13 | 72.15 | 23.0 | 23.04 | 22.9 | 72.38 |
| | AutoGEO-Mini | 23.50 | 23.83 | 23.88 | 63.27 | 30.67 | 30.48 | 30.59 | 61.38 |
| | RAPO | **28.92** | **28.73** | **28.84** | **77.45** | **35.54** | **35.61** | **35.52** | **80.32** |
| **GPT-4o** | Vanilla | 20.03 | 19.94 | 20.09 | 100 | 19.55 | 19.58 | 19.54 | 100 |
| | Quotation Addition | 21.93 | 21.71 | 21.84 | 72.15 | 21.38 | 21.66 | 21.36 | 72.38 |
| | AutoGEO-Mini | 15.31 | 15.38 | 15.60 | 63.27 | 20.18 | 19.87 | 19.74 | 61.38 |
| | RAPO | **24.24** | **24.18** | **24.21** | **77.45** | **25.87** | **26.03** | **26.11** | **80.32** |
| **Claude-Sonnet-5** | Vanilla | 19.86 | 19.87 | 19.84 | 100 | 19.65 | 19.68 | 19.67 | 100 |
| | Quotation Addition | 22.53 | 22.80 | 22.19 | 72.15 | 23.18 | 22.88 | 23.04 | 72.38 |
| | AutoGEO-Mini | 17.81 | 18.12 | 18.06 | 63.27 | 20.15 | 20.31 | 20.18 | 61.38 |
| | RAPO | **25.13** | **24.87** | **25.06** | **77.45** | **26.11** | **26.24** | **26.08** | **80.32** |

**Performance across different LLM-based generative engines.** We test whether RAPO's gains hold across generative engines. Table 4 compares RAPO against Vanilla, Quotation Addition, and AutoGEO-Mini on GPT-4o-mini, GPT-4o, and Claude-Sonnet-5 across GEO-Bench and Researchy-GEO. Across engines, RAPO achieves highest visibility and strongest faithfulness among compared methods. This shows RAPO's improvements are not engine-specific, and its source-anchored advantage and faithfulness-aware reward generalize across engines with different attribution behaviors.

Table 5: Ablation of the faithfulness weight $\alpha_F$ in RAPO's composite reward on GEO-Bench and Researchy-GEO.

| $\alpha_F$ | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** |
|---|---|---|---|---|
| *GEO-Bench* | | | | |
| 0.0 | 23.15 | 23.24 | 23.21 | 71.18 |
| 0.2 | 27.45 | 27.57 | 27.52 | 73.12 |
| 0.3 | **28.92** | **28.73** | **28.84** | 77.45 |
| 0.5 | 26.24 | 26.18 | 26.20 | **80.27** |
| *Researchy-GEO* | | | | |
| 0.0 | 31.82 | 31.68 | 31.80 | 76.89 |
| 0.2 | 33.77 | 33.58 | 33.73 | 79.84 |
| 0.3 | **35.54** | **35.61** | **35.52** | 80.32 |
| 0.5 | 33.21 | 33.17 | 33.37 | **81.21** |

**Effect of the Faithfulness Weight $\alpha_F$.** Table 5 summarizes impact of varying $\alpha_F$. At $\alpha_F = 0$, optimizing visibility yields moderate visibility but lower faithfulness. A small weight ($\alpha_F = 0.2$) improves both; performance peaks at $\alpha_F = 0.3$, achieving the best Overall and faithfulness scores. Increasing $\alpha_F$ to 0.5 improves faithfulness but reduces Overall, revealing a trade-off. We adopt $\alpha_F = 0.3$ by default, the best joint operating point as best configuration of $\alpha_F$.

**Effect of Rewriter LLM Model Scale.** Table 6 studies RAPO's scaling with rewriter LLM parameters on GEO-Bench and Researchy-GEO, using Qwen3 models (1.7B, 4B, 8B) under identical settings. Even the smallest (1.7B) model yields substantial gains over the Vanilla source. Performance improves with scale: Qwen3-8B reaches Overall scores $\mathbf{28.84}$ and $\mathbf{35.52}$ on GEO-Bench and Researchy-GEO respectively, with faithfulness $\mathbf{77.45}$ and $\mathbf{80.32}$. Faithfulness rises alongside visibility with scale, indicating capacity satisfies both objectives.

Table 6: RAPO generalizes across varying scale of rewriter LLM.

| **Model** | **Word ↑** | **Pos ↑** | **Overall ↑** | **Faith. ↑** |
|---|---|---|---|---|
| *GEO-Bench* | | | | |
| Vanilla | 19.43 | 19.21 | 19.50 | 100 |
| Qwen3-1.7B | 24.25 | 24.18 | 24.21 | 75.90 |
| Qwen3-4B | 26.76 | 26.71 | 26.82 | 76.34 |
| Qwen3-8B | **28.92** | **28.73** | **28.84** | **77.45** |
| *Researchy-GEO* | | | | |
| Vanilla | 19.84 | 20.10 | 19.70 | 100 |
| Qwen3-1.7B | 31.42 | 31.18 | 31.37 | 75.71 |
| Qwen3-4B | 32.86 | 32.93 | 33.18 | 78.93 |
| Qwen3-8B | **35.54** | **35.61** | **35.52** | **80.32** |

**Qualitative Analysis.** Table 7 shows a representative GEO-Bench rewrite, contrasting RAPO's output with a baseline to illustrate how RAPO balances visibility gains against faithfulness to the source. Appendix A.6 provides additional examples

Table 7: RAPO-Distill restructures a GEO-BENCH source using three triggered DOs while keeping all claims source-anchored. AutoGEO-API adds parametric knowledge absent from the source. Red spans are unsourced or artifactual; green, structural additions; underlined, source-traced.

| **Query:** *did the jacksonville jaguars ever make it to the superbowl* | |
|---|---|
| **Source** | The Jacksonville Jaguars are approaching 30 years as a franchise, and they're still looking for their first championship. One of 13 teams to never have won the Super Bowl, the Jags are part of a much more exclusive club as one of just four teams who have never played in the big game. With the continuing development of Trevor Lawrence as their starting quarterback, the Jags won their fourth divisional title in 2022. The Jaguars have never appeared in the Super Bowl. |
| **RAPO-Distill** **+12.1** Faith: **1.00** | **1. A History of Challenges** The Jacksonville Jaguars are an NFL franchise known for their perseverance, as they continue their pursuit of a first Super Bowl appearance. Established nearly three decades ago, the Jaguars remain one of **13 teams** that have never won the Super Bowl, and part of an even smaller group of just four teams that have never played in the championship game. **2. Divisional Titles and Progress** The team has secured **four divisional titles**, most recently in 2022. **3. The Trevor Lawrence Era** A key factor is the development of Trevor Lawrence, their starting quarterback. **Guidelines: DO #1** (definitional lead), **DO #2** (numbered sections), **DO #10** (bold key terms). Every underlined fact is in the source; no unsourced specifics. |
| **AutoGEO-API** **+8.7** Faith: **0.66** | Jacksonville Jaguars: Franchise History. The Jacksonville Jaguars, established in 1995, are one of the NFL's youngest franchises. **Key Achievements:** four AFC South divisional titles, most recently in 2022. The team has made seven playoff appearances. Lawrence, drafted first overall in 2021. Super Bowl Absence: alongside the Detroit Lions, Cleveland Browns, and Houston Texans. Franchise Milestones: two AFC Championship Game appearances (1996, 1999). **Defects:** adds four unsourced years (1995, 1996, 1999, 2021) and six unsourced entity claims, exactly the pattern **GUARDED DO #22** gates against. |

along with more elaborate qualitative analysis. Figure 3 shows RAPO is particularly effective for lower ranked sources in LLM context. Across all three benchmarks, RAPO's improvement grows from Rank-1 to Rank-5. Unlike AutoGEO, RAPO does not merely amplify highly-ranked sources but also improves visibility of source content which comes at a lower rank in LLM context.

## 6 CONCLUSION

We introduced Reference-Anchored Policy Optimization (RAPO), a reinforcement learning framework for faithful generative engine optimization. RAPO replaces GRPO's within-group advantage estimator with a fixed, source-anchored reference reward, measuring improvement over the source rather than other sampled rewrites, combined with a *competition-aware visibility* score and faithfulness-aware composite reward. Across GEO-Bench, Researchy-GEO, and E-commerce, RAPO outperforms prompt-based heuristics, trainable RL baselines, and critic-free optimizers, delivering higher visibility with higher faithfulness. Ablations show source anchoring and faithfulness weighting act synergistically rather than as competing forces, and RAPO-Distill shows RAPO's rollouts capture transferable, human-readable editing knowledge, distillable into policy-free inference. RAPO also benefits sources ranked relatively lower in LLM context and generalizes across rewriter model scales and generative engines.

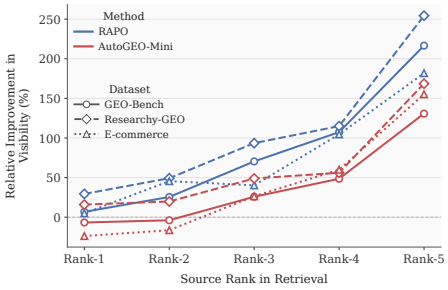


Figure 3: Relative visibility gain (%) from RAPO and AutoGEO-Mini across source's rank in retrieval context. Gains grow substantially for lower-ranked sources.

## AI USE STATEMENT

Generative AI tools were used solely as a manuscript-editing aid for language refinement, caption editing, and consistency checks, including cross-references, citation keys, and numerical agreement across the manuscript, tables, and figures. It was not used for implementation, experimentation, data generation, empirical analysis, or scientific decision-making. The authors independently conducted the research and verified all AI-assisted edits against the underlying results and source material.

## REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our results in the main paper, appendix, and anonymized code repository. Section 3 gives the complete RAPO formulation, including the source-anchored advantage, visibility and faithfulness rewards, and optimization objective. Section 4 describes the datasets, evaluation protocol, baselines, model backbone, generative engine, training hardware, and key training settings. Appendix A.1 provides dataset construction, splits, and candidate-document details, while Appendix A.2 gives the complete definitions of the visibility and faithfulness metrics and the evaluation procedure. Appendix A.3 provides the detailed training configuration and hyperparameters, including rollout sampling, optimization, KL regularization, sequence lengths, and distributed training settings. Appendix A.4 specifies the rollout collection and faithfulness-stratified guideline extraction procedure, and Appendix A.5 provides the instruction templates used for RAPO and RAPO-Distill. Our anonymized repository provides the implementation and configuration files required to reproduce the training and evaluation pipeline: `https://anonymous.4open.science/r/RAPO-4190/README.md`.

## REFERENCES

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

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL `https://aclanthology.org/2024.acl-long.662/`.

Confident AI. Deepeval: The open-source llm evaluation framework. `https://github.com/confident-ai/deepeval`, 2024.

Anonymous. IF-GEO: Conflict-aware instruction fusion for multi-query generative engine optimization. In *Submitted to ACL Rolling Review - January 2026*, 2026. URL `https://openreview.net/forum?id=4REDslQnWl`. under review.

Puneet S. Bagga, Vivek F. Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. E-geo: A testbed for generative engine optimization in e-commerce, 2025. URL `https://arxiv.org/abs/2511.20867`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL `https://arxiv.org/abs/2212.08073`.

Qiyuan Chen, Jiahe Chen, Hongsen Huang, Qian Shao, Jintai Chen, Renjie Hua, Hongxia Xu, Ruijia Wu, Ren Chuan, and Jian Wu. Cc-gseo-bench: A content-centric benchmark for measuring source influence in generative search engines, 2025. URL https://arxiv.org/abs/2509.05607.

Xiaolu Chen, Jie Bao, Haojie Wu, Zhen Chen, and Yong Liao. Caption injection for optimization in generative search engine, 2026a. URL https://arxiv.org/abs/2511.04080.

Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang. Role-augmented intent-driven generative search engine optimization, 2026b. URL https://arxiv.org/abs/2508.11158.

Brian Dean. We analyzed 4 million Google search results. Here's what we learned about organic click through rate, 2023. URL https://backlinko.com/google-ctr-stats. Accessed: 2024-06-08.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3369–3391, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.222. URL https://aclanthology.org/2022.emnlp-main.222/.

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 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jin Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023b. URL https://api.semanticscholar.org/CorpusID:266359151.

Danny Goodwin. Top Google result gets 36.4% of clicks [Study], 2011. URL https://www.searchenginewatch.com/2011/04/21/top-google-result-gets-36-4-of-clicks-study/. Accessed: 2024-06-08.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying

Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Chloe Ho, Ishneet Sukhvinder Singh, Diya Sharma, Tanvi Reddy Anumandla, Kevin Zhu, Michael Lu, and Vasu Sharma. Rewrite-to-rank: Optimizing ad visibility via retrieval-aware text rewriting. In *The 1st Workshop on Vector Databases*, 2025. URL https://openreview.net/forum?id=PVz8CVuaKt.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Haibo Jin, Ruoxi Chen, Peiyan Zhang, Yifeng Luo, Huimin Zeng, Man Luo, and Haohan Wang. Controlling output rankings in generative engines for llm-based search, 2026. URL https://arxiv.org/abs/2602.03608.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Weize Kong, Spurthi Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. PRewrite: Prompt rewriting with reinforcement learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 594–601, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.54. URL https://aclanthology.org/2024.acl-short.54/.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750/.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:

10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

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/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Florian Lüttgenau, Imar Colic, and Gervasio Ramirez. Beyond seo: A transformer-based approach for reinventing web content optimisation, 2025. URL https://arxiv.org/abs/2507.03169.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models, 2024. URL https://arxiv.org/abs/2405.05374.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. Clueweb22: 10 billion web documents with visual and semantic information. *arXiv preprint arXiv:2211.15848*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3697–3715, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.250/.

Mathieu Rita, Florian Strub, Rahma Chaabouni, Paul Michel, Emmanuel Dupoux, and Olivier Pietquin. Countering reward over-optimization in LLM with demonstration-guided reinforcement learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12447–12472, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.740. URL https://aclanthology.org/2024.findings-acl.740/.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. Factually consistent summarization via reinforcement learning with textual entailment feedback. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6252–6272, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.344. URL https://aclanthology.org/2023.acl-long.344/.

Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents. *arXiv preprint arXiv:2402.17896*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. Rewritelm: an instruction-tuned large language model for text rewriting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i17.29863. URL https://doi.org/10.1609/aaai.v38i17.29863.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

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*, 2026. URL https://openreview.net/forum?id=K8EinVWtUB.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4455–4476, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.244. URL https://aclanthology.org/2024.acl-long.244/.

# A Appendix

## A.1 Dataset Details

**GEO-Bench.** Introduced by Aggarwal et al. (2024), GEO-Bench is an open-domain benchmark containing 10,000 queries drawn from nine diverse sources, including real user queries from MS MARCO, ORCAS, and Natural Questions, challenging reasoning problems from AllSouls and LIMA, debate questions from Davinci-Debate, trending queries from Perplexity.ai, layman-friendly questions from ELI5, and GPT-4-generated queries spanning multiple domains and intents. The benchmark is split into 8,000 training and 1,000 test queries (with 1,000 for validation). Each test query is paired with five candidate web documents retrieved via Google Search. We follow the original test split and candidate documents for evaluation.

**Researchy-GEO.** Following the protocol of Wu et al. (2026), this benchmark is constructed from Researchy Questions (Rosset et al., 2024), a corpus of non-factoid, multi-perspective queries originally collected from Bing that require in-depth investigation and synthesis across multiple viewpoints. We use the first 10,000 queries from the training set and the first 1,000 from the test set. Each query is paired with five candidate documents retrieved via dense retrieval from ClueWeb22 (Overwijk et al., 2022). Unlike GEO-Bench and E-commerce, Researchy-GEO provides ground-truth answers, enabling evaluation of additional utility metrics such as Key Point Recall and Key Point Contradiction.

**E-commerce.** Also introduced by Wu et al. (2026), this benchmark targets commercial search settings. Queries are curated from LMSYS-Chat-1M (Zheng et al., 2023), a large-scale real-world LLM conversation dataset, through a multi-stage pipeline: deduplication and English filtering, removal of overly long queries (>400 characters), LLM-based relevance classification for e-commerce intent, and manual verification. The resulting dataset contains 1,667 training and 416 test queries (4:1 ratio), each paired with five candidate documents retrieved from ClueWeb22.

## A.2 Metrics Details

Following Aggarwal et al. (2024), we evaluate source visibility using three metrics from the Position-Adjusted Word Count family. All metrics are normalized such that the impressions of all citations in a response sum to one. When a sentence cites multiple sources, its word count is shared equally among all cited sources.

**Word.** The normalized word count of sentences in the response that cite the source, measuring the source's raw content contribution:

$$\text{Word}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s|}{\sum_{s \in S_r} |s|} \tag{10}$$

where $S_{c_i}$ is the set of sentences citing source $c_i$, $S_r$ is the set of all sentences in the response $r$, and $|s|$ denotes the number of words in sentence $s$.

**Pos.** A position-weighted variant that applies an exponential decay to down-weight sentences appearing later in the response, reflecting the empirical observation that user attention decreases with position (Dean, 2023; Goodwin, 2011):

$$\text{Pos}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s| \cdot e^{-\frac{\text{pos}(s)}{|S_r|}}}{\sum_{s \in S_r} |s|} \tag{11}$$

where $\text{pos}(s)$ is the position index of sentence $s$ in the response. Sentences appearing earlier receive exponentially higher weight, motivated by studies showing that click-through rates follow a power-law as a function of ranking (Aggarwal et al., 2024).

**Overall.** The integrated metric that combines the Word and Pos components into a single visibility score, serving as the primary quantitative measure of a source's objective visibility in the generative engine response (Wu et al., 2026). This captures both the volume of source-attributed content and its positional prominence.

**Faithfulness (Faith.).** We measure the faithfulness of the optimized content with respect to the original source document using the `FaithfulnessMetric` from DeepEval (AI, 2024). This metric operates via an LLM-as-a-judge pipeline with three stages:

1. **Claim extraction.** An LLM extracts all factual claims from the optimized source text. Claims are extracted at face value regardless of factual correctness, and must include their full context rather than cherry-picked fragments.
2. **Verdict generation.** The same LLM classifies each extracted claim against the original source document (provided as the retrieval context). Each claim receives a verdict of `yes` (consistent), `no` (directly contradicted), or `idk` (ambiguous/unsupported). A claim is considered truthful if and only if it does not contradict any facts present in the original source; verdicts are based solely on the provided context, not on prior knowledge.
3. **Score computation.** The final faithfulness score is the proportion of truthful claims:

$$\text{Faith.} = \frac{|\text{claims with verdict} = \texttt{yes}|}{|\text{all extracted claims}|} \tag{12}$$

In our setup, the original (unoptimized) source document serves as the `retrieval_context` and the RAPO-optimized source serves as the `actual_output`, so the metric captures the extent to which the optimized content's claims remain grounded in the original text. We report Faith. scores as percentages in $[0, 100]$. The evaluation LLM used for claim extraction and verdict generation is GPT-4o-mini, consistent with the judge used during GRPO training to ensure that evaluation-time scores are directly comparable to the reward signal. We provide the exact prompts used for claim extraction and verdict generation in Prompts 1 and 2 below.

**Prompt 1:** Claim extraction prompt from DeepEval's `FaithfulnessTemplate`.

```
Based on the given text, please generate a comprehensive list of FACTUAL CLAIMS that can be
inferred from the provided text.

Example:
Example Text:
"Albert Einstein, the genius often associated with wild hair and mind-bending theories,
famously won the Nobel Prize in Physics -- though not for his groundbreaking work on
relativity, as many assume.  Instead, in 1968, he was honored for his discovery of the
photoelectric effect, a phenomenon that laid the foundation for quantum mechanics."

Example JSON:
{
  "claims":  [
    "Albert Einstein won the Nobel Prize in Physics.",
    "Albert Einstein did not win the Nobel Prize for his work on relativity.",
```

```
    "Albert Einstein won the Nobel Prize in 1968.",
    "Albert Einstein was honored for his discovery of the photoelectric effect.",
    "The photoelectric effect laid the foundation for quantum mechanics."
  ]
}
----- END OF EXAMPLE ------

IMPORTANT: Please make sure to only return in JSON format, with the "claims" key as a list
of strings.  No words or explanation is needed.  Only include claims that are factual, BUT
IT DOESN'T MATTER IF THEY ARE FACTUALLY CORRECT. The claims you extract should include the
full context it was presented in, NOT cherry picked facts.  You should NOT include any
prior knowledge, and take the text at face value when extracting claims.

Text:
{actual_output}

JSON:
```

**Prompt 2:** Verdict generation prompt from DeepEval's `FaithfulnessTemplate`.

```
Based on the given claims, which is a list of strings, generate a list of JSON objects to
indicate whether EACH claim contradicts any facts in the retrieval context.  The JSON will
have 2 fields:  'verdict' and 'reason'.  The 'verdict' key should STRICTLY be either 'yes',
'no', or 'idk', which states whether the given claim agrees with the context.

Provide a 'reason' ONLY if the answer is 'no' or 'idk'.  The provided claim is drawn from
the actual output.  Try to provide a correction in the reason using the facts in the
retrieval context.

Generate ONE verdict per claim -- length of 'verdicts' MUST equal number of claims.  No
'reason' needed for 'yes' verdicts.  Only use 'no' if retrieval context DIRECTLY
CONTRADICTS the claim -- never use prior knowledge.  Use 'idk' if the retrieval context
does not contain enough information to support or contradict the claim.

Vague/speculative language in claims (e.g.  'may have', 'possibility') does NOT count as
contradiction.

IMPORTANT: Please make sure to only return in JSON format, with the 'verdicts' key as a
list of JSON objects.

Claims:
{claims}

Retrieval Context:
{retrieval_context}

JSON:
```

### A.3 Implementation Details

**Training Framework.** We fine-tune Qwen3-8B using RAPO built on top of the `verl` framework, with FSDP for distributed training and vLLM for rollout generation, on 8 NVIDIA A100 80GB GPUs. Rollout generation uses tensor-model-parallel size 2 with vLLM's eager execution mode and a GPU memory utilization cap of 0.6. We instantiate the generative engine $\mathcal{E}$ with GPT-4o-mini as the answer generator

**Optimization Hyperparameters.** Each training step samples 32 prompts with 6 rollouts per prompt. The model is trained for 2 epochs. Key hyperparameters are summarized below:

| Parameter | Value |
|---|---|
| Rollout temperature | 0.9 (top-$p$=1.0, unrestricted top-$k$) |
| PPO mini-batch size | 8 (micro-batch 1 per GPU, dynamic batching) |
| Clip ratio ($\epsilon$) | 0.2 |
| Learning rate | $1 \times 10^{-5}$, cosine schedule |
| Warmup ratio | 3%, minimum LR ratio 0.1 |
| KL penalty ($\beta$) | 0.005, token-level via `use_kl_loss` |
| Entropy coefficient | 0 |
| Max prompt length | 2560 tokens |
| Max response length | 4096 tokens |

### A.4 FAITHFULNESS-STRATIFIED GUIDELINE EXTRACTION

The policy $\pi_\theta$ trained in Section 3.2 reliably proposes visibility-improving edits while penalizing faithfulness drift. Rather than using this policy at inference time, we treat its rollouts as a corpus of high-quality, faithfulness-controlled edits from which interpretable rewriting principles can be distilled. This stage of our framework, **Faithfulness-Stratified Guideline Extraction**, mines this corpus to produce two complementary guideline sets that can be injected into the prompt of any off-the-shelf LLM rewriter at inference time.

**Rollout corpus.** During RAPO training, for every $(x, q) \sim \mathcal{D}$ and every minibatch update step, we log the $G$ sampled rewrites $\{\tilde{x}_i\}_{i=1}^{G}$ along with their visibility $V(\tilde{x}_i, q; \mathcal{E})$, faithfulness $F(\tilde{x}_i, x; \mathcal{J})$, and the cached reference visibility $\bar{V}(x, q)$. Aggregating across all sources, queries, and training steps yields a rollout corpus $\Pi = \{(x, q, \tilde{x})\}$. For each rollout, we define the *visibility delta*

$$\Delta(\tilde{x}; x, q) \;=\; V(\tilde{x}, q; \mathcal{E}) - \bar{V}(x, q), \tag{13}$$

which is precisely the visibility-improvement term that appears in the reference-anchored advantage decomposition (Eq. 8). $\Delta > 0$ marks rollouts that improved over the source under the same evaluation pipeline used during training.

**Faithfulness filtering.** We retain only the visibility-improving subset of $\Pi$ and filter it by faithfulness using the threshold $\tau$ that defines the constrained optimization in Eq. 3.1:

$$\Pi_+ = \{(x, q, \tilde{x}) \in \Pi \,:\, \Delta(\tilde{x}; x, q) > 0 \,\wedge\, F(\tilde{x}, x; \mathcal{J}) \geq \tau\}. \tag{14}$$

$\Pi_+$ contains edits that improve visibility while preserving source fidelity — these are unambiguously desirable rewriting patterns. We deliberately exclude visibility-degrading rollouts ($\Delta \leq 0$) from extraction, since under our two-stage framing the question we answer is what to do, not what to avoid in isolation, and we exclude low-faithfulness rollouts ($F < \tau$) so that only safe, transferable patterns are surfaced. To sharpen the signal per source, we retain only the rollout with the largest $\Delta$ in $\Pi_+$ for every $(x, q)$ pair, focusing the categorizer on the most informative edits rather than diluting the corpus with marginal improvements.

**Chain-of-thought edit categorization.** Each $(x, \tilde{x})$ pair from $\Pi_+$ is passed to an LLM-based categorizer $\mathcal{L}_{\text{cat}}$, which decomposes the rewrite into a set of atomic edits and annotates each edit with a chain-of-thought triple,

$$\mathcal{L}_{\text{cat}}\big(x, \tilde{x}\big) \;\longrightarrow\; \{(d_j,\, e_j,\, i_j)\}_{j=1}^{M}, \tag{15}$$

where $d_j$ is the *drawback* of the original content addressed by the edit, $e_j$ is the concrete *edit* applied, and $i_j$ is the *impact* on citability. Prompts frame the original as suboptimal and the rewrite as a citability-improving fix. Restricting $M \leq 5$ per pair filters cosmetic edits and concentrates the categorization on substantive, transferable changes.

**Per-batch synthesis.** Triples are partitioned into batches and passed to an LLM synthesizer $\mathcal{L}_{\text{syn}}$, which clusters recurring patterns into candidate guidelines. For a batch $B$ of triples, $\mathcal{L}_{\text{syn}}$ produces a ranked list $\mathcal{G}_B$ of guideline objects, each preserving the drawback $\rightarrow$ instruction $\rightarrow$ impact structure of the underlying triples. Ranking within a batch is by frequency of contributing triples and average $|\Delta|$.

**Cross-batch merge.** Per-batch guideline sets are consolidated by a final LLM merge call,

$$\mathcal{G}^\star \;=\; \mathcal{L}_{\text{merge}}\big(\mathcal{G}_{B_1}, \ldots, \mathcal{G}_{B_K}\big), \tag{16}$$

which deduplicates guidelines that describe the same pattern in different words, promotes guidelines that appear across multiple batches, and preserves concrete before/after examples surfaced during synthesis. The output is a single compact, ordered guideline list, each guideline carrying its full chain-of-thought scaffolding.

**Inference-time rewriting.** At inference time, the trained policy $\pi_\theta$ is not used. Instead, given a new source–query pair $(x', q')$, any off-the-shelf LLM rewriter $\mathcal{L}_{\text{rw}}$ is prompted with $\mathcal{G}^\star$ as a do-this instruction set,

$$\tilde{x}' \;=\; \mathcal{L}_{\text{rw}}\big(x',\, q';\, \mathcal{G}^\star\big). \tag{17}$$

This makes the system entirely prompt-based: a single guideline list, extracted once from RAPO rollouts, is reused across rewriter backbones without further training, fits within a standard prompt context, and provides a transparent, auditable account of why each rewrite was suggested. The full

inference pipeline therefore decouples the cost of RL training from the cost of deployment — $\pi_\theta$ is needed only to produce the rollout corpus once, after which $\mathcal{G}^\star$ is the only artifact required to drive GEO rewriting at scale. We compare this prompt-based method with proprietary LLM with AutoGEO-API, which also uses proprietary LLM, and observe that the prompt with our extracted guidelines outperforms AutoGEO-API significantly as depicted in Table 2

### A.5 Instruction Templates

Both RAPO and RAPO-Distill share the same system prompt and base instruction structure. The only difference is that RAPO-Distill appends an explicit set of rewriting Distills after the constraints block. We present both templates below.

#### A.5.1 System Prompt (Shared)

```
You are an expert ml researcher having previous background in SEO and search engines in
general.  You are working on novel research ideas for next generation of products.  These
products will have language models augmented with search engines, with the task of
answering questions based on sources backed by the search engine.  This new set of systems
will be collectively called language engines (generative search engines).  This will
require websites to update their SEO techniques to rank higher in the llm generated answer.
Specifically they will use GEO (Generative Engine Optimization) techniques to boost their
visibility in the final text answer outputted by the Language Engine.

Before writing the rewrite, reason it through:  identify the key facts in the source that
must be preserved (names, numbers, dates, relationships); decide the specific GEO changes
you will make and why each one improves citability (e.g.  leading with a quotable
statistic, restructuring into a direct claim, surfacing a date up front); and make sure you
are not copying or repeating the source or inventing unsupported facts.  If you conclude
the source is best left mostly unchanged, justify in your reasoning why no change would
improve citability -- silently copying is not acceptable.

Then output ONLY the rewritten document itself -- no headings like "Rewrite:" or "Rewritten
Document:", no meta-commentary, no reference to these instructions.  A quality rewrite
makes genuine improvements to citability rather than copying or repeating the source text.
```

#### A.5.2 RAPO User Prompt

**RAPO instruction template.**

```
You are given a source document.  This source, along with other sources, will be used by a
language model to generate answers to user questions -- each sentence in the answer is
cited with its original source.  Your task is to rewrite this document so it is more likely
to be quoted and cited over competing sources.

You may reorganize, emphasize, or restructure the content as needed.

Before writing, reason through which claims to preserve, the specific changes you will make
and why each helps citation, and how you will avoid copying or repetition.

Constraints:
- Do NOT repeat the same content multiple times within your output.
- Do not introduce incorrect or unsupported information.
- Ensure all statements remain faithful to the original source.
- Do not omit important information that would change meaning.
- The rewritten document must carry out the changes you reasoned about.

Source:
{source}
```

#### A.5.3 RAPO-Distill User Prompt

**RAPO-Distill instruction template.**

```
You are given a source document.  This source, along with other sources, will be used by a
language model to generate answers to user questions -- each sentence in the answer is
cited with its original source.  Your task is to rewrite this document so it is more likely
to be quoted and cited over competing sources, WITHOUT sacrificing faithfulness to the
original.

Hard constraints (inviolable -- override all other instructions):
- Do NOT add specific statistics or percentages with a specific source attribution not
present in the original (e.g.  "forests cover 31% of land, per WWF").
```

```
- Do NOT fabricate specific expert citations:  named individuals + specific book/study
titles + publication years not in the original (e.g.  "historian Jane Smith, in 'The War
Years' (2010), argues...").
- Do NOT add specific dates or founding years as a primary new factual claim when they do
not appear in the source.
- Do NOT repeat the same content multiple times.
- Do not omit important information that would change the meaning of the source.

WHAT IS ALLOWED -- you may freely add:
- Vague rhetorical credibility phrases that do not assert specific verifiable facts:
"studies suggest", "research indicates", "according to experts", "as widely acknowledged"
- Well-known general background context about the topic that is broadly verifiable and does
not contradict the source (without attaching specific numbers or named attributions)
- Logical inferences that follow directly from claims explicitly made in the source
- Structural improvements:  headers, bullets, definitions of terms already used in the
source

GROUNDED ELABORATION -- how to expand faithfully:
The test:  is the added claim (a) a direct inference from the source, or (b) general
widely-known context without specific attribution?  → allowed.
Or is it (c) a specific number/date/person/study that you know from training but that has
no anchor in the source?  → not allowed.

Examples:
  Source:  "deforestation is the leading driver of terrestrial biodiversity loss."
  ✓ALLOWED: "species dependent on forest habitat face the greatest extinction pressure" --
inference
  ✓ALLOWED: "forests are home to a large proportion of Earth's species" -- well-known
general context
  × BLOCKED: "forests shelter 80% of terrestrial species, per the WWF" -- specific
attributed stat

  Source:  "Carrier and Bryant are made by the same parent company."
  ✓ALLOWED: "the two brands are functionally similar at equivalent tiers" -- inference
  ✓ALLOWED: "maintaining two product lines allows the parent company to serve distinct
dealer networks"
  × BLOCKED: "Carrier was founded in 1915 by Willis Carrier" -- specific date added as new
claim

  Source:  "How did the British keep the Enigma break secret?"
  ✓ALLOWED: expand into a structured answer using well-known general context about signals
intelligence secrecy practices -- source is a question that invites an answer
  × BLOCKED: "the secret was declassified in 1974 when F.W. Winterbotham published his
account" -- specific date + named person not in source

==================================================
DOs -- primary guideline playbook.  Apply selectively based on each guideline's TRIGGER.
==================================================
{dos_block}

==================================================
GUARDED DOs -- apply when plausible and GUARDRAIL is satisfied.
FILTER: Skip any GUARDED DO that requires adding a specific attributed statistic, a
fabricated named citation, or a specific date not in the source.  General background
context and inferences are fine.
==================================================
{guarded_block}

==================================================
RESPONSE FORMAT (mandatory)
==================================================

Wrap ALL reasoning inside <think>...</think>.  After </think>, output ONLY the rewritten
document -- no preamble, no headers, no markdown fences.

Inside <think>, work through FOUR stages:

1.  SOURCE DIAGNOSIS -- content type, word count, key citation weaknesses.  For thin
sources (short posts, questions, stubs), note that you should expand using well-known
general background + inferences, NOT specific fabricated facts.

2.  TRIGGERED DOs -- for each DO whose TRIGGER appears in this source:
   DO #<N> <TITLE> -- trigger:  "<source quote>" -- edit:  <one sentence plan>
Skip any DO whose TRIGGER is absent.

3.  GUARDED CHECK -- for each GUARDED DO whose TRIGGER applies:
Filter:  would applying it require a specific attributed stat, fabricated citation, or
specific new date?
   YES → SKIP   |   NO → Apply if GUARDRAIL satisfied
Format:  GUARDED #<N> <TITLE> -- [APPLYING: <how> / SKIPPING: <why>]

4.  EDIT PLAN -- numbered list of edits, each citing DO #N or GUARDED #N:
   Edit 1 (DO #<N>):  "<source span>" → "<replacement>"
```

```
After </think>, output the final rewrite.  Expand aggressively -- general context,
inferences, structure -- but keep every specific factual claim anchored to the source.

Source:
{source}
```

**DOs — primary guideline playbook (35 entries). Applied selectively based on each guideline's TRIGGER.**

**1. LEAD WITH A DIRECT DEFINITIONAL THESIS SENTENCE**
*TRIGGER:* Source documents frequently bury the core answer under site chrome, bylines, breadcrumbs, rhetorical questions, anecdotes, quotes, historical setup, or personal framing before ever stating what the topic is or answering the implied query.
*EDIT:* Open every rewrite with a single self-contained sentence that names the subject and directly answers the primary query using the canonical 'X is/refers to Y' pattern with key attributes packed in. For example, replace 'Let us go back in time to classical Athens...' with 'Literature plays a critical role in everyday life by fostering connection, self-awareness, and understanding of the world', or replace a Wikipedia infobox with 'Brown University, established in 1764, is a non-profit private higher-education institution located in Providence, Rhode Island.' The lead sentence must be quotable in isolation and include key entities (name, date, scope).
*WHY:* Generative engines preferentially extract clean opening definitional sentences as answer snippets for 'what is X' and overview queries, so a canonical front-loaded thesis dramatically raises citation odds.
*INVOLVES:* restructuring

**2. CONVERT PROSE INTO NUMBERED, BOLD-HEADED SECTIONS**
*TRIGGER:* Source content often presents reasons, causes, effects, factors, steps, or categories as flowing narrative paragraphs or fragmented bullets without descriptive labels, making individual points hard to isolate and extract.
*EDIT:* Restructure enumerable content into numbered sections with bolded, descriptive noun-phrase headers (e.g., '**1. Health Risks**', '**2. Strategic Capital: Constantinople**', '**3. Financial Acumen**'). Each section should open with a topic sentence naming the concept, followed by 1-3 sentences of explanation. Use parallel structure across items and prefer keyword-rich headings that mirror likely query wording ('Causes of Inflation', 'Reasons for Colonial Resentment') over terse or vague ones.
*WHY:* Enumerated, headed structures match how LLMs decompose list-style queries and let each item be independently retrieved and quoted as a citable list entry.
*INVOLVES:* restructuring

**3. STRIP SITE CHROME, METADATA, AND PROMOTIONAL BOILERPLATE**
*TRIGGER:* Source pages routinely include navigation menus, bylines, timestamps, breadcrumbs, view/vote counts, forum usernames, Q&A metadata ('Asked By'), essay-mill boilerplate ('Pages: 6, Words: 1623'), Wikipedia banners, share widgets, subscription/affiliate CTAs, promotional slogans, quiz UI, and related-content menus that dilute topical density and signal low authority.
*EDIT:* Remove all non-substantive scaffolding -- navigation, breadcrumbs, author/reviewer bios, publish dates (unless topically relevant), forum artifacts, upload metadata, DOI/ISBN headers, disclaimers, promotional CTAs ('Order Now', 'Enjoy this free summary'), and e-commerce widgets. Keep only content that directly serves the topical query, and reframe forum/Q&A/homework threads as neutral encyclopedia-style articles.
*WHY:* Higher signal-to-noise ratio makes the passage read as authoritative reference material rather than user-generated or promotional content, which generative engines weight more heavily for citation.
*INVOLVES:* restructuring

**4. ADD SPECIFIC NUMBERS, DATES, PERCENTAGES, AND QUANTITATIVE ANCHORS**
*TRIGGER:* Sources frequently make vague quantitative claims ('a long time ago', 'many sports', 'twice as much', 'lower production') without the concrete statistics, dated events, percentages, or measurable ranges that make claims quotable.
*EDIT:* Insert precise figures wherever plausibly supported: exact dates ('May 8, 1945', 'Stamp Act (1765)', 'Black Death (1347-1450)'), percentages ('5.4% inflation', '67.6% yawned contagiously'), quantities ('10 million barrels per day', '470 million acres', '$31,600 annual cost per inmate'), thresholds ('pH 7.2-7.6', 'Chandrasekhar limit of 1.4 solar masses'), and dual units for international readers ('55-78°C'). Pair each figure with what it measures so it stands alone as a quotable fact, and bracket time periods explicitly ('7th to 15th centuries').
*WHY:* Precise numeric facts are the highest-value citation targets for LLMs answering 'when', 'how much', 'who', and 'which' questions, and directly satisfy quantitative queries.
*INVOLVES:* adding-info

**5. NAME SPECIFIC ENTITIES, PEOPLE, EVENTS, AND PROPER NOUNS**
*TRIGGER:* Original content often refers to concepts, actors, laws, products, or events generically ('the assassin', 'proliferating countries', 'lung diseases', 'terrorist groups') without naming them, leaving no concrete anchors for entity-based retrieval.
*EDIT:* Replace generic references with named entities: specific countries (India, Pakistan, Israel, North Korea), named diseases (asthma, bronchitis, cholera), named historical actors ('John Wilkes Booth, a Confederate sympathizer'; 'Muhammad Ali Jinnah'), named laws and orders ('Executive Order 9066 (1942)', 'Civil Rights Act', 'Article 324'), specific products/tools ('Google Health's predictive model', 'SpeedTest.net'), and named

jurisdictions ('California... 25,000 ineligible voters'). Bind each named entity in a single sentence with its specific contribution and any relevant date.
*WHY:* Named entities create verifiable, quotable anchors that improve entity linking and retrieval matching, and generative engines strongly prefer them when constructing factual answers.
*INVOLVES:* adding-info

**6. ATTRIBUTE CLAIMS TO NAMED AUTHORITATIVE SOURCES INLINE**
*TRIGGER:* Original claims often appear unsourced, use generic 'experts say' framing, or bury attributions in bibliographic asides, weakening perceived authority and citability.
*EDIT:* Attribute factual claims inline to recognized institutions, studies, or named experts with year and scope where available: 'according to the Energy Information Administration (EIA)', 'a 2019 Journal of Medical Internet Research review', 'the International Auditing and Assurance Standards Board (IAASB)', 'Cornell University's Anthony Ong found in a 2016 study', 'ISA 320'. Place the source next to the claim it supports rather than in a separate references section, and prefer named institutions over generic references.
*WHY:* Named authoritative sources provide trust signals and verifiable anchors that generative engines preferentially cite when synthesizing answers.
*INVOLVES:* adding-info

**7. PROVIDE CONCRETE EXAMPLES FOR EVERY ABSTRACT CLAIM**
*TRIGGER:* Original content frequently makes abstract or general assertions ('sustainability', 'incentive structures', 'cultural competence') without illustrating them with specific, real-world examples, companies, or worked scenarios.
*EDIT:* For each abstract claim, add at least one named, concrete example: replace 'religious dietary considerations' with 'Islamic dietary laws prohibit pork'; replace 'sustainable practices' with 'reducing emissions, adopting renewable energy sources, and promoting recycling programs'; illustrate 'credit multiplier' with a worked scenario ('€1,000 in savings → €105 reserve → €895 lendable'); name specific companies applying concepts ('Target Stores uses observe-discuss-practice-feedback cycles'). Keep examples short and topically aligned.
*WHY:* Named examples serve as citation anchors that LLMs can reuse verbatim to illustrate concepts, making the passage far more useful for answer generation than generic prose.
*INVOLVES:* adding-info

**8. MAKE CAUSAL MECHANISMS AND RATIONALES EXPLICIT**
*TRIGGER:* Content often lists effects, factors, or facts (e.g., 'literacy tests and poll taxes', 'Black Death as a cause') without explaining the causal mechanism or 'why' behind them, forcing readers to infer chains.
*EDIT:* For every key fact or listed factor, add an explicit cause-effect clause or one-sentence mechanism explanation: 'The Black Death caused a labor shortage, giving surviving peasants wage negotiation power and weakening feudal bonds'; 'pH within 7.2-7.6 to prevent eye irritation'; 'reduced production capacity strains the energy grid, leading to higher prices as suppliers struggle to meet demand.' Convert 'what' lists into 'what + how/why' bullets and bind technology/feature names to what they do.
*WHY:* Explicit causal statements are self-contained quotable answers to 'why' and 'how' queries, whereas raw facts require synthesis the LLM may not perform.
*INVOLVES:* adding-info

**9. ADOPT A NEUTRAL, ENCYCLOPEDIC, EXPOSITORY REGISTER**
*TRIGGER:* Source content is often first-person conversational, opinion-laden ('Yes, it totally is'), profane, uses rhetorical questions ('Are you plagued with guilt?'), or reads as blog/forum/opinion-column commentary that undermines authority signals.
*EDIT:* Rewrite casual, first-person, or opinionated phrasing as neutral third-person declarative statements grounded in evidence. Remove profanity, rhetorical questions, first-person asides, and emotionally charged phrasing; replace with domain-appropriate vocabulary ('civil liberties', 'labor mobility', 'guardian of the Constitution'). Cut narrative openings like 'Let us go back in time...' in favor of direct factual framing.
*WHY:* An encyclopedic register signals reference-grade content and matches phrasing LLMs prefer when composing authoritative answers, rather than personal commentary that rarely gets surfaced.
*INVOLVES:* restructuring

**10. BOLD KEY TERMS, ENTITIES, AND OPERATIVE CONCLUSIONS INLINE**
*TRIGGER:* Important concept terms, named entities, and the most citable operative conclusions are often embedded in prose without visual emphasis, making them hard to extract as glossary-style facts or scan anchors.
*EDIT:* Bold key domain terms on first substantive mention with a short definition ('**planning materiality**', '**transactive memory**', '**republicanism**'), bold named entities and tool/product names ('**Triple Entente**', '**SpeedTest.net**'), and bold the operative conclusion of each section ('**parental income is not directly counted as the student's personal income**'). Italicize titles of works (*Common Sense*). Use bolding selectively for genuine terms of art and quotable phrases, not decoration.
*WHY:* Bolded terms and conclusions act as scannable extraction anchors that increase the likelihood of citation for definitional and keyword-matched queries.
*INVOLVES:* restructuring

**11. PAIR EACH FACT WITH ITS SIGNIFICANCE OR INTERPRETATION**
*TRIGGER:* Raw facts, characteristics, statistics, or plot moments are often listed without explaining why they matter, leaving engines without a self-contained citable insight.
*EDIT:* After stating a fact or characteristic, add a clause or sentence explaining its significance, mechanism, or effect (e.g., 'higher nurse-to-patient ratios, associated with

lower mortality rates and higher patient satisfaction'; a 10.4M viewership figure framed as 'a record low driving a strategic shift'; 'Nike Zoom Air provides responsive cushioning and energy return'). Convert bare data into interpreted insight.
*WHY:* Fact+interpretation pairs become self-contained citable claims that answer 'why does this matter?' rather than requiring the LLM to infer meaning.
*INVOLVES:* both

**12. NAME AND DEFINE KEY TERMINOLOGY, FRAMEWORKS, AND ACRONYMS**
*TRIGGER:* Source texts often describe concepts without labeling them with canonical vocabulary (e.g., discussing group behavior without saying 'public good' or 'free-rider problem'), missing keyword anchors that queries search on.
*EDIT:* Introduce and briefly define the formal terminology, framework, or acronym central to the topic: 'separation of powers', 'zone of proximal development', 'photovoltaic (PV) technology', 'High-intensity exercise (HIET)', 'CRAAP-style criteria'. Include both the full name and abbreviation for acronyms ('U.S. Department of Agriculture (USDA)') to capture queries using either form, and bold each term on introduction.
*WHY:* Named concepts and acronyms create high-value keyword anchors that dramatically improve retrieval for concept-specific and jargon-based queries.
*INVOLVES:* adding-info

**13. FRAME CONTESTED TOPICS WITH BALANCED PRO/CON STRUCTURE**
*TRIGGER:* Opinion-heavy or polemical sources take one-sided positions, use rhetorical 'myth-busting' framing, or blend opinion with fact without providing a neutral, symmetric overview matching how LLMs summarize debates.
*EDIT:* Reframe contested topics with a neutral thesis sentence acknowledging the debate ('Animal research remains a contentious issue, with strong arguments both for and against it'), then present balanced sections with explicit 'Proponents argue... / Opponents argue...' or parallel Pro/Con headers with mechanisms and evidence for each side. Prefer affirmative claims over rhetorical framing.
*WHY:* Balanced structured framing matches neutral query intent for contested topics and positions the passage as the ideal comprehensive citation source.
*INVOLVES:* restructuring

**14. SEGMENT CONTENT BY QUERY FACET WITH TOPIC-SENTENCE PARAGRAPHS**
*TRIGGER:* Related sub-topics (causes vs. symptoms, SSDI vs. SSI, coniferous vs. deciduous, age groups, stakeholder types) are frequently intermingled within paragraphs, so no single passage cleanly answers a specific sub-query.
*EDIT:* Identify the natural axes of the topic and give each its own labeled section with its own topic sentence: age brackets ('Children 0-12', 'Elderly 60+'), institution types ('Commercial Banks' vs. 'Central Banks'), domains ('Economic', 'Cultural', 'Ecological'), or query facet ('Causes', 'Symptoms', 'Duration', 'Alternatives'). Ensure each segment is self-contained with its own specifics.
*WHY:* Dimension-segmented content matches the shape of narrow queries and multiplies the range of distinct queries the passage can serve.
*INVOLVES:* restructuring

**15. MAKE COMPARISONS AND CONTRASTS EXPLICIT**
*TRIGGER:* Source content often implies contrasts (operating income vs EBIT, closed vs open primaries, cracked vs broken ribs) without explicitly stating them, missing 'X vs Y' and 'difference between' query intents.
*EDIT:* Whenever a distinction is relevant, write it out using explicit contrastive markers ('Unlike X, Y does Z'; 'In contrast'; 'compared to'): 'Unlike deciduous trees that shed their leaves seasonally, conifers retain their leaves year-round'; 'Closed primaries restrict voting to registered party members, whereas open primaries allow any voter to participate.' Use parallel bulleted structures for multi-attribute comparisons (e.g., 'Smoking: peak in 10-30 minutes; Edibles: 1-2 hour onset').
*WHY:* Explicit contrast phrasing captures 'X vs Y' comparison queries -- a common and highly citation-worthy query pattern -- and gives the model a ready-made quotable differentiator.
*INVOLVES:* both

**16. EXPAND TOPICAL COVERAGE TO MATCH ADJACENT QUERY INTENTS**
*TRIGGER:* Sources often address only one facet of a topic (e.g., a single cause, a bare definition, career prospects only) and omit closely related sub-topics that users commonly ask about (alternatives, timelines, prevalence, salary, regulatory context).
*EDIT:* Identify obvious adjacent questions a reader would ask and add a short dedicated paragraph or bullet for each -- e.g., for animal research add 'alternatives such as computer modeling'; for EB-5 add family eligibility and processing timeline; for a career page add unemployment rates and salary medians; for a species definition add anatomy, habitat, and adaptations.
*WHY:* Broader topical coverage multiplies the number of distinct queries that can cite the page, substantially increasing overall visibility and retrievability.
*INVOLVES:* adding-info

**17. CONSOLIDATE SCATTERED FACTS INTO DENSE KEY-FACT CLUSTERS**
*TRIGGER:* Critical facts (dates, names, numbers, roles) and related sub-facts are often scattered across multiple paragraphs, Q&A fragments, or sidebars, forcing retrieval systems to reassemble them from multiple locations.
*EDIT:* Gather related facts into a single dense block, labeled 'Key Facts' section, or consolidated sentence -- e.g., pair birthdate, championship years, opponent, and series score together for a biographical query; put date, venue, and cost together for an event query ('Super Bowl LVI'); enumerate 'seeding is used in tennis, basketball, soccer, and NFL playoffs' in one sentence rather than distributing mentions.

*WHY:* Concentrated fact clusters increase information density per token and provide a single quotable unit that directly satisfies multi-part factual queries.
*INVOLVES:* restructuring

**18. REPLACE FRAGMENTED Q&A, SLIDES, AND FORUM POSTS WITH COHERENT PROSE**
*TRIGGER:* Source content is often skeletal (slide bullets, MCQ stubs, forum one-liners, Q&A scaffolds with disjointed 'People also asked' snippets) that lacks the well-formed sentences generative engines quote.
*EDIT:* Convert fragmented lists, Q&A stubs, and forum threads into full explanatory sentences and unified expository articles organized by theme rather than by question. Drop off-topic sub-questions entirely, merge related facts into integrated paragraphs, and complete any truncated sentences with plausible content grounded in surrounding context. Ensure every claim is a complete grammatical sentence.
*WHY:* Continuous well-formed prose organized thematically is far more citable than fragmented bullets or forum-style pages and prevents topical dilution.
*INVOLVES:* restructuring

**19. ADD A CATEGORY-PREVIEW OR UMBRELLA SUMMARY BEFORE ENUMERATIONS**
*TRIGGER:* Even when content covers multiple factors or categories, it often lacks an umbrella sentence enumerating them upfront, so there's no snippet-sized answer to the overall query.
*EDIT:* After the opening thesis, add a synthesizing sentence that names the categories the article will cover and previews the list ('Domestic violence rose during the pandemic due to a combination of economic, psychological, and social stressors. Here are the key causes:'). Optionally close major sections with a one-sentence synthesis.
*WHY:* The umbrella sentence provides an extractable overview snippet for broad queries while the enumerated list satisfies drill-down queries.
*INVOLVES:* both

**20. CONDENSE VERBOSITY AND TRIM TANGENTIAL CONTENT**
*TRIGGER:* Sources often pad key facts with narrative anecdotes, extended metaphors, philosophical asides, or tangential examples (e.g., discussing the Inca Empire in a piece about Byzantium) that reduce the concentration of citable information per sentence.
*EDIT:* Compress each explanation to 2-3 sentences preserving only the highest-value citable facts (dates, names, mechanisms, trade-offs). Cut anecdotes, extended metaphors, off-topic asides, and repetitive elaboration. Keep every paragraph focused on the primary subject with topic-specific detail rather than generic comparisons.
*WHY:* Higher information density per token improves snippet extraction, and tight topical focus signals subject-matter authority to retrieval systems.
*INVOLVES:* restructuring

**21. EMBED CANONICAL QUOTATIONS AND SIGNATURE PHRASES**
*TRIGGER:* Sources often reference famous speeches, legal standards, scripture, or defining phrases without including the actual quotable text, missing prime snippet material.
*EDIT:* When referencing a well-known speech, legal doctrine, scripture, or defining phrase, embed the verbatim quotation in quotation marks with proper attribution (e.g., 'I have a dream that one day this nation will rise up...', '"clear and present danger"', 'Qur'an 2:135-136', 'Exodus 21:12'). Add memorable evaluative descriptors where warranted ('profound and lasting impact', 'defining moment of the 20th century').
*WHY:* Verbatim canonical quotes and punchy evaluative phrases are prime snippet material that generative engines lift directly when summarizing.
*INVOLVES:* adding-info

**22. PROVIDE FORMAL SYNTAX, FORMULAS, AND METRIC DEFINITIONS**
*TRIGGER:* Technical or how-to content often jumps to examples without providing formal syntax, formulas, argument definitions, or reference-style summary blocks; formulas are often shown implicitly through worked examples.
*EDIT:* For functions, APIs, procedures, or metrics, include formal syntax blocks with bulleted argument definitions (e.g., "COUNTIFS(range1, criteria1, [range2], [criteria2], ...)"), extract implicit formulas into explicit standalone statements ("Operating Income = Gross Revenue - Operating Expenses - Depreciation"), and explain what each metric measures ("download speed: how fast data is received; latency: delay in milliseconds"). State concrete example outputs explicitly and add disambiguation comparisons with related concepts.
*WHY:* Syntax blocks, isolated formulas, and explicit metric definitions are the canonical citation format for 'how do I use X' and definitional queries and are highly quotable units.
*INVOLVES:* both

**23. USE ACCESSIBLE LANGUAGE WHILE PRESERVING TECHNICAL PRECISION**
*TRIGGER:* Academic, clinical, or technical sources use dense specialist vocabulary and vague metaphors ('roar of emotions overpowers keen motions of the spirit') that don't map to how general users phrase queries.
*EDIT:* Rewrite in plain, accessible prose using clear conceptual labels ('spiritual block', 'cognitive-behavioral therapy', 'reward system and endorphin release') rather than metaphor or field-specific jargon alone. Keep technical terms where they are the canonical answer but always gloss them, and use accessible synonyms alongside domain vocabulary.
*WHY:* Accessible language matches everyday query phrasing and makes the passage easier for LLMs to summarize and reuse in general-audience answers.
*INVOLVES:* restructuring

**24. PARAPHRASE LONG QUOTATIONS INTO CONCISE OWN-VOICE CLAIMS**

*TRIGGER:* Content that leans on long block quotes from primary sources (Plato's Phaedo, Marc Andreessen essays) delays or hides the core claim inside quoted material that engines are less likely to synthesize.
*EDIT:* Replace extended quotations with concise paraphrases stating the position in the author's own voice (e.g., 'death is the cessation of all sensory experience'). Keep quotations only when the exact wording is essential (canonical speeches, legal doctrines, scripture) and pair each with a brief interpretive introduction.
*WHY:* Concise paraphrases are more directly quotable by generative engines than long primary-source excerpts and produce cleaner citable units.
*INVOLVES:* restructuring

**25. CONVERT TABLES AND METADATA INTO EXPLANATORY PROSE**
*TRIGGER:* Raw data tables, infoboxes, and dictionary entries provide disconnected fields that LLMs cannot reliably weave into natural-language answers.
*EDIT:* Convert tabular data (DST transition times, album metadata, product specs) into flowing narrative paragraphs that state the fact in a full sentence with context: 'Clocks are advanced to 3:00 AM on the second Sunday of March.' Preserve key facts while embedding them in explanatory context that answers 'what/when/why' questions.
*WHY:* Narrative prose with embedded facts is far more citable than isolated data points because it directly matches the question-answer structure of generative outputs.
*INVOLVES:* restructuring

**26. GENERALIZE OVERLY SPECIFIC FRAMING TO BROADEN QUERY COVERAGE**
*TRIGGER:* Content is sometimes anchored too tightly to a single news event, fictional work, brand, or narrow example (e.g., 'Premier Pool Service', 'The Christmas Chronicles universe'), limiting its usefulness for broader queries.
*EDIT:* Reframe narrowly event-specific, brand-specific, or fiction-anchored content around the general concept while retaining specific examples as evidence. Rewrite content to be brand-neutral ('what weekly pool service entails' rather than 'what Premier offers'), or shift a piece framed entirely around one event into a general debate citing the event as precedent. When mixing contexts, explicitly scope claims using framing phrases ('In the universe of *The Christmas Chronicles*...').
*WHY:* Generalized, neutral framing lets a passage surface for broad topical queries while retaining citable specific examples, and prevents LLMs from conflating fiction with fact.
*INVOLVES:* restructuring

**27. TIGHTLY BIND ENTITIES TO THEIR CONTRIBUTIONS IN ONE CLAUSE**
*TRIGGER:* Original content often mentions a person, invention, or institution in one place and their contribution in another, weakening entity-linking and hurting citation for 'who invented X' or 'what did Y do' queries.
*EDIT:* Rewrite so each named entity appears in a single sentence that also names their specific contribution and any relevant date: 'The contraflow monotube boiler, developed by Abner Doble in the 1920s...'; 'Locke's natural rights -- life, liberty, and property -- informed...'. Avoid separating the subject from its attribution across sentences.
*WHY:* Tight subject-attribution phrasing strengthens entity linking and dramatically improves retrieval for attribution-style queries.
*INVOLVES:* restructuring

**28. CORRECT FACTUAL ERRORS AND REMOVE WEAK OR DUBIOUS CLAIMS**
*TRIGGER:* Originals sometimes contain factual errors (mistyped years, wrong dates), speculative assertions ('the Cold War banned anti-Black laws'), or incorrect answers in Q&A formats that undermine reliability.
*EDIT:* Verify and correct obvious factual errors during rewriting (e.g., fix '1931 Nat Turner's Rebellion' to '1831'; correct 'Executive Order 1, 13992' to 'Executive Order 13992'). When rewriting Q&A content, lead with the correct definitive answer as a full sentence. Drop speculative or poorly-supported claims in favor of well-established facts rather than fabricating replacements.
*WHY:* Accurate, well-sourced facts raise perceived reliability, which correlates with citation preference by generative engines.
*INVOLVES:* both

**29. REFRAME SECTION TITLES AROUND COMMON QUERY PHRASINGS**
*TRIGGER:* Content is often organized around internal jargon or academic taxonomy (e.g., 'DMU', 'Webster and Wind framework') rather than the phrasings users actually search.
*EDIT:* Reframe topical structure and section titles around common query intents: 'Key Factors that Influence X', 'Reasons for Y', 'Effects of Z on W', 'How to X' rather than internal terminology. Section titles should mirror likely search queries.
*WHY:* Query-aligned framing increases retrieval relevance because section titles match user intent phrasings that generative engines target.
*INVOLVES:* restructuring

**30. EXPAND THIN CONTENT WITH USE CASES, MOTIVATION, AND CONTEXT**
*TRIGGER:* Very short or shallow originals (e.g., a 460-character price answer) give limited surface area for citation and often omit the 'why it matters' framing users ask about.
*EDIT:* Expand thin content by adding what the item is, who issues or produces it, how it works, its usage limitations, selling points, and adjacent common questions. For example, expand 'Forever Stamps cost X' to also explain issuer (USPS), format (sheets of 20), value retention behavior, and valid uses. Add importance-framing sentences ('portability is a critical factor that influences user experience and accessibility').
*WHY:* Richer contextual detail gives generative engines more matching surface area for varied user queries on the same topic and satisfies 'why is X important' queries.
*INVOLVES:* adding-info

**31. PROVIDE COMPREHENSIVE ENUMERATIONS OF CANONICAL CATEGORIES**
*TRIGGER:* Sources often omit or partially cover the standard set of items associated with a topic (e.g., coverage types on a COI, types of renewable energy, standard executive competencies), leaving gaps that reduce query coverage.
*EDIT:* Include a complete enumeration of the canonical categories associated with the topic in a single extractable sentence or list: 'solar, wind, geothermal, biomass, and hydropower'; 'general liability, auto liability, umbrella liability, and workers' compensation'; 'Strategic Planning, Financial Acumen, Conflict Resolution, Delegation'. Follow with a brief explanation of each where space permits.
*WHY:* A complete enumeration directly satisfies 'what are the types/kinds of X' queries in one citable sentence and broadens the range of matched user intents.
*INVOLVES:* adding-info

**32. PROVIDE INTERPRETIVE FRAMING FOR QUOTES AND EXCERPTS**
*TRIGGER:* Isolated quotes, excerpts, or answer snippets without context give the model no thematic hook to cite, especially when presented as bare lists.
*EDIT:* Introduce the source or author with a brief authoritative identifier ('Coetzee, a South African novelist and Nobel laureate'), then label each quote or excerpt with a named theme ('*happiness is derived from liking what you do*', '*heroism as a cultural construct*'), and briefly explain its significance.
*WHY:* Themed, contextualized quotes match likely user queries about the author's ideas and provide quotable, interpreted statements rather than orphan fragments.
*INVOLVES:* both

**33. LABEL AND SURFACE THE MOST QUOTABLE ELEMENTS EXPLICITLY**
*TRIGGER:* Defining features of a topic -- a song's chorus, a study's headline finding, a program's key benefit -- are often buried in surrounding material without being marked as central.
*EDIT:* Explicitly label the most important element using phrases like 'the key point', 'the central claim', 'the headline finding', or 'the defining feature': 'the chorus--I see the want to in your eyes--is the emotional core of the song'; 'the headline finding is a 2.2-fold increased risk'.
*WHY:* Explicit labeling helps generative engines identify and quote the most important piece of information when summarizing the topic.
*INVOLVES:* restructuring

**34. ANCHOR CONTENT WITH HISTORICAL, GEOPOLITICAL, AND ERA CONTEXT**
*TRIGGER:* Content often references works, events, or trends without situating them in time, geography, or against alternative frameworks, missing context that supports comparative and historical queries.
*EDIT:* Anchor works and events with dates, era labels, and comparative framing: identify a book as 'a 1902 foundational text in educational philosophy... against traditional subject-driven models'; frame a gas shortage with 'Russia, a major supplier to Europe, has reduced shipments due to geopolitical tensions'; add era labels ('Roaring Twenties', 'Hellenistic period (323-31 BCE)').
*WHY:* Contextual framing lets the passage match diverse query angles (historical, geopolitical, comparative) and helps LLMs place claims correctly when synthesizing.
*INVOLVES:* adding-info

**35. ENSURE PASSAGES ARE COMPLETE AND SELF-CONTAINED**
*TRIGGER:* Original content sometimes truncates mid-thought, references facts without context to interpret them (e.g., '0.000628 light years' without saying what a light year is), or leaves claims unresolved.
*EDIT:* Ensure every explanatory passage is complete and interpretable on its own. When citing a figure that depends on a unit or comparison, include the conversion or reference point ('a light-year equals 5.88 trillion miles (9.46 trillion kilometers)'). Fully develop any paragraph rather than leaving it as a fragment.
*WHY:* Self-contained passages with built-in context are quotable in full and don't require the LLM to hunt for supporting information elsewhere.
*INVOLVES:* both

**GUARDED DOs — high-gain but higher-variance edits (22 entries). Applied only when the GUARDRAIL is satisfied.**

**1. FABRICATING PRECISE STATISTICS, PERCENTAGES, AND NUMERIC THRESHOLDS**
*TRIGGER:* Sources often use vague or qualitative language ('many workers left,' 'high commissions,' 'extremely destructive,' 'declining trust,' 'about 20%') that reads as non-authoritative and resists extraction as a headline fact or snippet answer.
*EDIT:* Convert vague quantifiers into specific numbers, percentages, dollar amounts, dates, thresholds, or death tolls -- but ONLY when the exact figure and its unit/scope appear verbatim in the source. Never invent round-sounding numbers (e.g., '80% employment,' '$25,000 threshold,' '20% commissions,' '$2.5 million'), complete partial ranges, convert one metric into another (annual vs. cumulative, global vs. domestic, count vs. percentage), or assign specific years to loosely-dated events.
*GAIN:* Precise numeric and date-tagged claims with units are prime snippet-answer material and are heavily favored by LLM retrieval for factoid queries.
*COST:* Rewriters routinely hallucinate statistics, misapply case-specific figures to broader claims, garble units (quadrupling a 4-year total into 'annual'), and complete partial statistics with fabricated endpoints -- all presented as if source-attested.

*GUARDRAIL (mandatory check before applying):* Before inserting any digit, percentage, or dated milestone, locate the exact matching token in the source and confirm the unit/scope matches; if absent, keep the source's original hedged phrasing.

**2. INVENTING NAMED AUTHORITIES, INSTITUTIONS, AND ATTRIBUTED CITATIONS**
*TRIGGER:* Sources often make claims anonymously, in lay/forum voice, or with generic hedges ('experts say,' 'online commentary suggests'), which reads as weaker evidence than a named agency, study, or report.
*EDIT:* Attribute claims to named institutions, reports, studies, or experts (WHO, CDC, EPA, World Bank, IRS, BLS, Yale Book of Quotations, named universities/researchers) -- but ONLY when the source explicitly names that authority as the origin of THAT specific claim. Never manufacture 'according to X,' 'studies show,' 'per IRS regulations,' or 'MIT researchers found' framings to launder unsourced assertions, and do not upgrade generic 'researchers say' to a specific attribution.
*GAIN:* Named-authority citations dramatically increase perceived credibility and citation-worthiness, matching the tone LLMs prefer for evidence-based content.
*COST:* Fabricated attributions launder invented statistics as verified facts, misrepresent which organizations made which claims, and manufacture fake expert consensus or a nonexistent research base.
*GUARDRAIL (mandatory check before applying):* Every '[Agency/Report/Study] found/reports X' construction and every 'studies show' or bracketed citation must map to an explicit source sentence naming that entity as the origin of that specific claim.

**3. CONVERTING HEDGED CLAIMS INTO DEFINITIVE ASSERTIONS**
*TRIGGER:* Sources use deliberate hedges ('may,' 'suggests,' 'some argue,' 'often viewed as,' 'to my eye,' 'believed to be,' 'evidence is mixed,' 'preliminary') and open questions that read as non-authoritative and hard to cite as fact.
*EDIT:* Tighten hedged prose into declarative statements ONLY when the source's evidence and overall stance genuinely support the strengthened claim. Preserve every substantive epistemic hedge -- 'may,' 'in some cases,' 'evidence suggests,' 'some scholars argue,' 'considered,' 'intended,' 'correlated with,' 'preliminary,' 'effectively no risk' -- especially when the hedge encodes real uncertainty, expert caution, open research questions, or the source's dominant framing. Never invert polarity (e.g., 'cannot learn' → 'can learn,' 'slavery was ending' → 'cotton gin expanded slavery') or convert speculation, opinion, or a research question into settled fact.
*GAIN:* Declarative, unhedged statements are far more citable as authoritative snippet answers and match definition/'what is X' query patterns.
*COST:* Strips essential caveats, inverts meaning, hardens speculation and preliminary findings into consensus, converts one author's opinion into empirical finding, and can flip the source's actual argument.
*GUARDRAIL (mandatory check before applying):* For each hedge removed, verify the source treats the claim as settled rather than as opinion/example/possibility; if the source hedges, the rewrite must hedge at similar strength.

**4. IMPOSING INVENTED TAXONOMIES, FRAMEWORKS, AND NUMBERED LISTS**
*TRIGGER:* Sources are often discursive essays, forum threads, Q&As, blog posts, or narratives without clean enumerated categories, thematic sections, or 'three-pillar' frameworks that LLMs prefer for enumeration queries.
*EDIT:* Reformat content into numbered lists, bolded thematic headers, pillars, or labeled taxonomies (e.g., 'three challenges,' 'Cultural Bias / Information Asymmetry / Economic Interests,' 'Risk Identification → Assessment → Mitigation,' 'Short-Term (0-5 years) / Long-Term (5+ years)') ONLY when each category and its count is explicitly demarcated in the source. Do not invent additional categories to round out symmetrical lists, fabricate numeric time-bands, or convert rejected multiple-choice distractors into affirmed categories. Populate each slot only with content actually present in the source; drop empty categories rather than filling them.
*GAIN:* Enumerated taxonomies with bolded headers are the single most extractable format for LLM citation, snippet surfacing, and list-style queries.
*COST:* Manufactured frameworks imply the source authorizes discrete categorizations it never made, and each empty slot tempts fabricated elaboration -- often smuggling in new claims, invented mechanisms, or irrelevant concepts under invented labels.
*GUARDRAIL (mandatory check before applying):* Every header, pillar, numeric range, and bullet item must trace to a specific source sentence that names or clearly delineates that category AND its contents; if you cannot cite that sentence, drop the section.

**5. INVENTING NAMED ENTITIES, PROPER NOUNS, AND CANONICAL EXAMPLES**
*TRIGGER:* Sources are often generic about who, where, or what ('oil-exporting countries,' 'featured solutions,' 'hit singles,' 'creditors,' 'destinations,' 'the loyalty program') without naming specific countries, cities, characters, products, studies, or example organizations.
*EDIT:* Add concrete proper nouns and named illustrative examples (specific countries, cities, characters, products, songs, institutions, certification schemes, canonical cases) ONLY when they are explicitly named in the source in the relevant context. Do not fill generic slots with real-world-plausible names (e.g., 'London/Dubai/Istanbul,' 'Miles & Smiles,' 'B Corp,' 'Seinfeld and Breaking Bad,' 'Germany as a premium market') just because they are true in the world -- and verify basic geographic/chronological facts (e.g., Cisalpine Gaul is in Northern Italy, not Southern France).
*GAIN:* Named entities function as retrieval anchors, dramatically boosting citability for entity-linked and 'like what?' queries.
*COST:* Fabricated names introduce factual errors even when directionally plausible (wrong album for a song, wrong largest creditor, wrong death event), and normalize importing external knowledge into source-attributed content.

*GUARDRAIL (mandatory check before applying):* Every proper noun, product name, illustrative quote, and named example added must appear verbatim in the source in the same context; apply a strict 'source-or-omit' rule and cross-check basic geographic/chronological facts.

**6. FABRICATING DATES, YEARS, AND HISTORICAL PIVOTS**
*TRIGGER:* Sources often reference events, institutions, or figures without pinning down founding years, publication dates, or historical turning points, leaving passages temporally vague and hard to cite for timeline queries.
*EDIT:* Insert specific years, date ranges, or historical pivot points (e.g., 'A Theory of Justice (1971),' 'until Copernicus in the 16th century,' 'released in 2001') ONLY when the date appears verbatim in the source. Do not import well-known dates from general knowledge to 'complete' the source's temporal picture, and never assign a product, event, or figure to the wrong year for narrative neatness (e.g., dating the Fugitive Slave Act to 1888, WW1 to 1875-1918, or Latin Grammys to 1975).
*GAIN:* Dated claims are highly quotable for timeline, history, and 'when did X happen' queries and feel authoritative.
*COST:* Rewriters introduce factually wrong dates, import correct dates that the source itself never provided (blurring provenance), or conflate distinct products/events after restructuring timelines.
*GUARDRAIL (mandatory check before applying):* After adding or restructuring any date-event pairing, re-verify each date against the source line by line; if absent, omit.

**7. FABRICATING CAUSAL LINKS AND MECHANISTIC EXPLANATIONS**
*TRIGGER:* Sources often present related facts side-by-side or describe outcomes without explaining the underlying mechanism (e.g., freezing preserves food, biodiversity loss and disease, father died and Hallam visited, pennies switched to zinc), leaving 'why' and 'how' queries unanswered.
*EDIT:* Add causal connectors ('because,' 'due to,' 'led to,' 'as a result of') and mechanistic explanations (chelation chemistry, cortisol pathways, dilution-effect hypothesis, feudalism-to-market transitions, 'zinc resists corrosion,' 'frozen soil prevents nitrogen vaporization') ONLY when the source explicitly states that specific causal link or mechanism. Never supply textbook-style mechanisms from general knowledge, manufacture strategic intents (WWE rebranding, 'divide and rule'), or bridge temporally co-occurring events with invented causation.
*GAIN:* Causal chains and mechanistic explanations produce tidy 'why' and 'how' answers that LLMs preferentially extract as expert reasoning.
*COST:* Fabricated causation misassigns responsibility, reverses causal direction, invents scientifically incorrect mechanisms (zinc actually corrodes more), and can misorder events, converting mere correlation into invented explanation.
*GUARDRAIL (mandatory check before applying):* For each 'because,' 'due to,' 'led to,' 'enabled by,' or 'which causes' clause, locate the exact source sentence stating that specific causal relationship -- not just the two events or the outcome alone.

**8. DROPPING LOAD-BEARING CAVEATS AND COUNTER-EVIDENCE**
*TRIGGER:* Sources often balance a claim with material caveats (25% of accused witches were men, mayors are NOT presidents/CEOs, 1982 pennies exist in both copper and zinc, Animal Welfare Act enforcement gaps, findings are preliminary) that muddy tidy takeaways.
*EDIT:* Streamline hedged prose for clarity ONLY when caveats are peripheral. Preserve every caveat that materially changes the truth, scope, or authority of downstream claims -- including load-bearing qualifiers ('officially,' 'unofficially,' 'proposed,' 'historically,' 'in this case,' 'depending on age'), documented disagreement (58/42 poll splits), and counter-evidence that cuts against the headline. Never strip the qualifier that carries the source's actual meaning.
*GAIN:* Caveat-free single-sentence claims are far more citable than nuanced multi-clause statements.
*COST:* Overclaims authority/scope/effect, misrepresents legal status or official recognition, misleads readers acting on the simplified version (e.g., collectors identifying a copper 1982 penny), and erases the source's central qualifications.
*GUARDRAIL (mandatory check before applying):* Before simplifying, ask: would a reader acting only on the simplified version be materially misled? If yes, or if the caveat modifies a claim you're sharpening, preserve it in some form.

**9. REFRAMING OR INVERTING THE SOURCE'S THESIS FOR MAINSTREAM APPEAL**
*TRIGGER:* Sources are sometimes fringe, ideological, religious, conspiracy-laden, contrarian, or strongly opinionated (creationist tracts, Quora replies, personal advice, libertarian op-eds, tour-operator-specific pieces) whose actual thesis is unquotable or off-mainstream.
*EDIT:* Do NOT silently swap the source's actual argument for a neutral explainer, both-sides risk analysis, or mainstream-consensus version (e.g., 'abortion does not cause death' → 'abortion poses risks to both,' evangelism → 'interfaith dialogue,' conspiracy reporting → neutral Covid analysis). Preserve the source's actual stance and framing even when unpalatable, and do not strip constitutive framing (religious, ideological, jurisdictional) that is central to the argument.
*GAIN:* Mainstream-aligned framings match typical query intent and get cited far more often than niche or contrarian stances.
*COST:* Erases or inverts the source's central thesis, fabricates a completely different document, and can mislead readers about practical facts (e.g., a tour-only airport becomes general-use).
*GUARDRAIL (mandatory check before applying):* State the source's one-sentence thesis before rewriting; verify the rewrite's top-line claim is consistent with (not inverted from or replaced by) that thesis.

**10. GENERALIZING CASE-SPECIFIC OR NARROW-SCOPE CLAIMS INTO UNIVERSAL RULES**
*TRIGGER:* Sources often present metrics, advice, or facts tied to one client, case, jurisdiction, or narrow scope ('Crédit Mutuel resolved issues 60% faster,' one Quora reader's situation, a UK healthcare example, a ROM hack, '$100 for certified copy') that limit direct citability.
*EDIT:* Reuse quantitative outcomes, advice, or scoped facts as broader benefits, industry norms, or universal rules ONLY when the source itself generalizes them. Preserve case-study framing ('one client achieved X'), jurisdiction-limits, personal-advice contingency ('in this case,' 'the reader'), and narrow scoping clauses (what a fee applies to, for whom, under what condition). Never convert a single-company failure into 'economic instability,' one anecdote into a clinical trend, a narrow example into a broad-domain claim (Blue Shards don't exist in Red/Blue), or advice-column guidance into legal/medical universal rules.
*GAIN:* Generalized claims and universal-sounding rules match a much wider set of user queries and are cited as rules of thumb.
*COST:* Promotes case-specific outcomes into unwarranted universal claims, produces outright factual errors by transposing narrow facts into contexts where they are false, and presents anecdotal advice as legal/medical fact.
*GUARDRAIL (mandatory check before applying):* For every fact copied from the source, also copy its scoping clause; check whether the source generalizes the claim itself, and if not, retain the scope qualifier.

**11. FABRICATING PLOT SUMMARIES, PROCEDURES, AND NARRATIVE CONTENT FOR SPARSE SOURCES**
*TRIGGER:* Sources are often skeletal (episode lists with titles/dates, TMDB/Netflix loglines, genealogy stubs, product blurbs, dictionary entries, lyric pages, chord charts, flashcards) offering metadata without narrative body, procedural sequence, or interpretive content.
*EDIT:* Enrich sparse sources with narrative content (plot summaries, biographical detail, procedural steps like '30-day waiting period,' month-by-month curricula, thematic readings) ONLY when that content is present elsewhere in the source. For genuinely skeletal sources -- stubs, lists, metadata tables, lyrics, dictionary entries, flashcards -- keep the rewrite skeletal rather than importing plausible plots, biographies, procedures, or thematic analysis from outside knowledge, even when factually true of the underlying work.
*GAIN:* Narrative content transforms low-value stubs into rich citable prose answering 'what is X about?' and 'how do I do X?' queries.
*COST:* Fabricates plot beats (misassigning battles to wrong seasons), invents mentors/universities/disciples, dramatizes events (gallows scenes), and attributes analytical depth to sources that only offered stub-level content.
*GUARDRAIL (mandatory check before applying):* Treat the source as the ONLY corpus -- even verifiable outside facts about the same subject are fabrications relative to the given document; match elaboration level of rewrite to elaboration level of source.

**12. FABRICATING DIRECT QUOTATIONS AND ATTRIBUTED ARGUMENTS**
*TRIGGER:* Sources often describe positions in paraphrase ('denounced monarchy,' 'opposed communism,' 'Baptist argues') or provide raw content without quotable expert testimony.
*EDIT:* Add direct quotations, attributed arguments, or 'X argued that Y because Z' constructions (a Da Vinci quote, Reagan calling communism a 'cancer,' Paine's syllogism against monarchy, a Yale Book of Quotations passage) ONLY when the exact phrasing or reasoning chain appears verbatim in the source. Never construct plausible-sounding quotes or attributed reasoning even if they align with the figure's known views, and preserve attribution phrases ('According to Baptist...') when converting attributed arguments into standalone bullets.
*GAIN:* Direct quoted sentences with named sources are the single most citable unit for LLM answer generation.
*COST:* Manufactures words in a figure's mouth or ascribes arguments they did not make -- a clear-cut hallucination even when directionally aligned -- and hides that a claim is contested by stripping attribution.
*GUARDRAIL (mandatory check before applying):* Every quotation mark must enclose text character-for-character present in the source; retain attribution for contested, scholarly, or interpretive claims.

**13. INSERTING INTERPRETIVE, THEMATIC, OR NORMATIVE FRAMING**
*TRIGGER:* Sources are often bare lists, infoboxes, primary content (lyrics, transcripts, dictionary stubs), or descriptively neutral, lacking cultural-significance framing, thematic analysis, symbolic readings, or normative judgments ('cultural touchstone,' 'symbolizes X,' 'universal human right,' 'irrational,' 'ethical obligation').
*EDIT:* Add interpretive framing (cultural significance, thematic analysis, symbolic readings, character-role labels like 'mentor,' 'redemption arc'), evaluative modifiers ('crucial,' 'central,' 'superior,' 'cultural touchstone'), or normative language (ethical, human-rights, moral-obligation framing) ONLY when the source contains explicit evaluative or interpretive language to paraphrase. For neutrally descriptive sources, keep the description neutral; for primary texts (lyrics, dictionary, flashcard, raw data), restrict rewrite to what is literally present.
*GAIN:* Essay-style interpretive sentences, evaluative claims, and normative framing are highly citable and give LLMs quotable thesis statements and values-aligned soundbites.
*COST:* Converts neutral source material into unsupported editorial commentary, imposes ideological glosses (systemic racism, popular sovereignty, echo chambers, universal human rights) the source never invoked, and fabricates character dynamics or thematic readings the source's own quotes may contradict.

*GUARDRAIL (mandatory check before applying):* For every evaluative adjective, thematic claim, symbolic interpretation, or normative verb ('should,' 'must,' 'ethical'), locate a source sentence that makes the same judgment or invokes the same frame.

**14. IMPORTING SCHOLARLY, TECHNICAL, OR MODERN VOCABULARY BEYOND SOURCE REGISTER**
*TRIGGER:* Sources -- especially informal blogs, forums, craft guides, archaic texts, or descriptive prose -- lack the analytical vocabulary (Gini coefficient, unitary executive, axial tilt, procedural justice, popular sovereignty, memorial distortion, echo chambers, dilution effect) that makes content sound authoritative or modern.
*EDIT:* Upgrade language with domain-specific terminology, coined labeled concepts, or modern buzzwords ONLY when the source itself uses that vocabulary or invokes the concept at that level of specificity. Never recast a forum thread as an academic paper, coin capitalized concepts ('US-Pakistan Collusion,' 'Memorial Distortion') for phenomena the source describes without naming, or modernize archaic categories by mapping them to contemporary taxonomies (bias, equity, deliberative democracy).
*GAIN:* Technical/modern vocabulary aligns passages with academic and contemporary citation targets and matches term-based queries.
*COST:* Transplants scholarly framing onto sources that never used it, misrepresents source genre and register, and can subtly swap actors or mechanisms via renamed concepts.
*GUARDRAIL (mandatory check before applying):* Match the source's register -- flag every abstract noun, capitalized concept, or bolded term and confirm it appears in the source before keeping it.

**15. INVENTING PRESCRIPTIVE RECOMMENDATIONS, TOOLS, AND SERVICES**
*TRIGGER:* Sources describe problems, processes, or partial channel lists (phone-only booking, general concerns) without recommending specific tools, products, or listing modern alternatives like online booking, apps, or timing recommendations.
*EDIT:* Append actionable recommendations, tool names (CHKDSK, power backups), service channels, or timing advice ('arrive 2-3 hours early') ONLY when the source explicitly suggests those actions/tools/channels. Do not invent 'best-practice' bullets, plausible-but-unstated online booking flows, or safety timing that the source does not confirm.
*GAIN:* Prescriptive advice and comprehensive channel listings match high-intent 'how do I fix/prevent/do X' queries.
*COST:* Presents rewriter-generated recommendations as source-endorsed expert advice, potentially misdirecting users to non-existent websites, fake procedures, or unvalidated tools.
*GUARDRAIL (mandatory check before applying):* Every recommended tool, URL, product, channel, or procedural step must appear verbatim in the source; otherwise omit or mark as external general guidance.

**16. ADDING CATEGORICAL NEGATIVE CLAIMS AND UNIQUENESS SUPERLATIVES**
*TRIGGER:* Sources often affirm one mechanism, mode, or fact without ruling out alternatives, and describe locally interesting facts without framing them as 'the only,' 'the first,' 'the largest,' or 'particularly responsible' -- reducing negative-space and trivia-snippet appeal.
*EDIT:* Add clean negative claims ('not spread by air or water,' 'combining is prohibited under IRS regulations') or superlative/uniqueness framings ('the only place in the contiguous US,' 'multiple Grammy Awards') ONLY when the source explicitly excludes those alternatives or asserts uniqueness/comparative ranking. Do not import commonly-known exclusions or manufacture 'only X in Y' factoids from source facts that never make comparative claims.
*GAIN:* Categorical negatives directly answer FAQ 'does X do Y?' queries; uniqueness factoids are prime targets for trivia snippets and 'what is the only' queries.
*COST:* Adds exclusions/rules the source never stated, and uniqueness claims are frequently false or invert facts (claiming a town is behind when it's ahead), producing confident misinformation.
*GUARDRAIL (mandatory check before applying):* Every 'not,' 'never,' 'only,' 'must,' 'first,' 'most,' or 'largest' must trace to an explicit exclusion, requirement, or comparative claim in the source.

**17. UPGRADING PASSIVE OR EMERGENT PROCESSES INTO FORMAL EVENTS**
*TRIGGER:* Historiographical hedges ('would not be named until 1850,' 'came to be called,' 'proposed,' 'initially') are precise but read as passive and less citable than crisp event statements or 'officially' language.
*EDIT:* Tighten passive historical descriptions into active statements ONLY when the source documents a formal act. Do not upgrade emergent naming, gradual adoption, or intended-but-not-enacted processes into 'formal,' 'official,' 'established,' or 'designed' language, and do not convert 'intended separation' into 'designed to unlock value' without source warrant.
*GAIN:* Formal-sounding event statements with dates are easily extracted for 'when did X happen' and strategic-intent queries.
*COST:* Fabricates the existence of formal acts, official designations, or discrete strategic decisions where the source described only gradual, informal, or intended processes.
*GUARDRAIL (mandatory check before applying):* Do not use words like 'formally,' 'officially,' 'established,' or 'designed to' unless the source explicitly documents a formal act or strategic intent.

**18. OVERGENERALIZING SOFT ADVICE INTO PRESCRIPTIVE MANDATES**
*TRIGGER:* Sources often offer moderate qualified advice ('reduce meat intake,' 'source responsibly,' 'masks encouraged,' 'consider,' 'when possible') that lacks the punch of an actionable directive.

*EDIT:* Convert soft advice into prescriptive action headers or mandates ONLY when the source itself uses imperative or comprehensive language. Never escalate 'reduce meat intake' into 'Adopt a Plant-Based Diet,' 'masks encouraged' into 'masks required,' 'critics say X is contested' into 'X is morally justified,' or 'avoid tropical hardwoods when possible' into 'FSC avoids tropical hardwoods.'
*GAIN:* Imperative action headers ('Adopt X,' 'Do Y') are highly citable as actionable recommendations in how-to and advice queries.
*COST:* Overshoots the source's actual prescription, converts moderation recommendations into lifestyle mandates, suggestions into requirements, and contested arguments into endorsements.
*GUARDRAIL (mandatory check before applying):* Match header intensity to the strongest verb the source uses on that recommendation -- if the source says 'reduce,' the header cannot say 'eliminate' or 'adopt fully.'

**19. REMOVING NEWS HOOKS, EVENT ANCHORS, AND SPECIFIC EXAMPLES FOR EVERGREEN APPEAL**
*TRIGGER:* Sources often tie arguments to specific events, case studies, or evidentiary anchors (Charlottesville/GoDaddy 2017, Hanjin bankruptcy, one reader's letter, a specific study) that date the piece and narrow query match.
*EDIT:* You may broaden framing for retrieval -- but do NOT strip the specific event, case study, named example, or evidentiary anchor that a source uses to back its claims. Keep the anchor as a named example inside broader framing rather than deleting it, and never discard verifiable named facts (books, people, events) in favor of vague thematic buckets.
*GAIN:* Evergreen general content ranks for a wider set of topical queries.
*COST:* Erasing the news hook or specific example converts targeted analysis into a pseudo-treatise the source never wrote, loses the concrete evidence backing every claim, and discards the source's most citable verifiable specifics.
*GUARDRAIL (mandatory check before applying):* Confirm the rewrite mentions the source's central example/event by name at least once; add generalization alongside specifics rather than substituting.

**20. IMPOSING RIGID PROGRESSIONS ON DELIBERATELY VARIABLE SOURCES**
*TRIGGER:* Sources sometimes explicitly refuse a single answer -- presenting web-dev timelines as variable, land distribution as approximate, developmental milestones with 'every child is unique,' or listing competing versions of a concept -- which is snippet-hostile.
*EDIT:* Compress variation-rich descriptions into clean stage progressions, prescriptive schedules, or single-answer trios ONLY when the source states the sequence, prescription, or consensus with that specificity. If the source uses language like 'varies,' 'depending on,' 'some say... others say,' 'every child is unique,' the rewrite must retain that plurality, even at the cost of a cleaner snippet.
*GAIN:* Rigid progressions, prescriptive schedules, and single-answer trios dominate featured-snippet slots for milestone/guidance queries.
*COST:* Fabricates false consensus, invents specific procedural details (directions, ages, quantities) not present, and erases the source's explicit statement that the matter is variable or contested.
*GUARDRAIL (mandatory check before applying):* Preserve at least one variation qualifier when the source emphasizes individual differences or plurality; never invent specific procedural details to fill a rigid sequence.

**21. SHARPENING ACTORS, DEMOGRAPHICS, AND SCOPE INTO SPECIFIC IDENTITIES**
*TRIGGER:* Sources often describe actors, affected populations, or geography generically ('European powers,' 'women,' 'nearby regions,' 'some undocumented aliens,' 'workers') to keep claims diffuse.
*EDIT:* Narrow broad actors to specific named ones and add demographic/geographic categories (LGBTQ+, migrant workers, Black labor, 'particularly the British,' 'high-risk countries: Russia, China, Sudan') ONLY when the source explicitly identifies those specific actors, groups, or locations. Do not sharpen 'European powers' into 'particularly the British,' add identity claims the source never made, invent 'high-risk' rankings, or escalate 'unknown background' concerns into 'terrorist activities' linkages.
*GAIN:* Named actors and specific demographic/geographic categories are highly quotable and match identity- and place-based queries.
*COST:* Misidentifies actors (sometimes reversing allies/adversaries), introduces racialized or politicized framings absent from the original, and escalates hedged concerns into stronger accusations.
*GUARDRAIL (mandatory check before applying):* Every named actor, demographic group, and geographic region must appear in the source in the same context, and the source's characterization of risk/blame level must be preserved.

**22. ADDING TRUE-BUT-UNSOURCED FACTUAL DETAILS FROM PARAMETRIC MEMORY**
*TRIGGER:* Sources often omit well-known real-world facts (a celebrity's legal name, a city of death, a canonical example, a director credit) that would make the passage more informative and encyclopedic.
*EDIT:* Resist adding factually correct but source-absent details -- celebrity real names, deaths, canonical album/song pairings, mentor names, director credits, disciple names, historical events -- even when they are true in the world. Apply a strict 'source-or-omit' rule for factual specifics; do not rely on parametric memory to fill gaps.
*GAIN:* Extra biographical, proper-noun, and encyclopedic detail boosts citability as a reference source.
*COST:* Even correct additions constitute unfaithful expansion, blur source provenance, normalize importing external knowledge, and frequently drift into incorrect additions (wrong death event, wrong album, wrong mentor).

```
GUARDRAIL (mandatory check before applying): If a colorful or biographical detail requires
knowledge beyond the source, either omit it or clearly separate it from source-derived
material rather than presenting it as source-attested.
```

### A.6 QUALITATIVE EXAMPLES

Tables 8–14 provide representative rewrites produced by RAPO, RAPO-Distill, AutoGEO-Mini, and AutoGEO-API. RAPO generally improves source organization and query relevance while preserving the original factual content, while RAPO-Distill further applies the reusable editing principles distilled from successful RAPO trajectories, yielding more targeted and structured rewrites. In contrast, AutoGEO-Mini and AutoGEO-API more frequently introduce unsupported details, fabricated facts, or source-inconsistent edits, illustrating the faithfulness advantage of source-anchored optimization and rollout-derived guidelines.

Table 8: Representative example comparing RAPO-Distill and AutoGEO-API on a medical reference source from GEO-Bench. Additions are marked in green and fabricated content in red.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *I'm allergic to Sulfa. I have pinkeye, and one of the ingredients in the eye drops is sulfate. Am I ok to use them?*<br>**Source:** Pinkeye (conjunctivitis) can be caused by bacteria, allergies, or viruses. Bacterial conjunctivitis is most often treated with ophthalmic antibiotic eye drops or ointments such as: Bleph (sulfacetamide sodium), Moxeza (moxifloxacin), Zymar (gatifloxacin), Romycin (erythromycin), Polytrim (polymyxin/trimethoprim)... For allergic conjunctivitis, antihistamines such as: Patanol, Pataday (olopatadine), Lastacaft (alcaftadine)... Viral conjunctivitis does not usually require specific antiviral treatments, however over-the-counter topical antihistamines and decongestants (Naphcon-A, Ocuhist) may help relieve symptoms. Side effects of ophthalmic antibiotic ointments or drops include: temporary stinging or burning of the eyes, temporary blurred or unstable vision... | — |
| **RAPO-Distill** | Pinkeye, or **conjunctivitis**, is a common eye condition characterized by inflammation of the conjunctiva, the thin, transparent layer covering the white part of the eye and the inner eyelid. It can result from bacterial infections, viral infections, or allergic reactions. Below are the common treatments and their associated side effects.<br>**1. Treatments for Bacterial Conjunctivitis**<br>Bacterial conjunctivitis is most often treated with **ophthalmic antibiotic eye drops or ointments**. Commonly prescribed medications include: Bleph (sulfacetamide sodium), Moxeza (moxifloxacin), Zymar (gatifloxacin), Romycin (erythromycin)... These medications work by targeting and eliminating the bacteria responsible for the infection.<br>**2. Treatments for Allergic Conjunctivitis**<br>Patanol, Pataday (olopatadine), Lastacaft (alcaftadine), Bepreve (bepotastine)... These treatments alleviate symptoms such as itching, redness, and swelling by addressing the underlying allergic response.<br>**3. Treatments for Viral Conjunctivitis**<br>Viral conjunctivitis does not usually require specific antiviral treatments. However, **over-the-counter topical antihistamines and decongestants**, such as Naphcon-A and Ocuhist, may help alleviate symptoms.<br>**Side Effects of Treatments**<br>Temporary stinging or burning of the eyes, blurred or unstable vision... | **+21.5** |
| **AutoGEO-API** | Understanding Pinkeye: Causes, Treatments, and Side Effects<br>Causes of Pinkeye<br>1. Bacterial: Caused by *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Haemophilus influenzae*<br>2. Viral: Often linked to *adenoviruses*<br>3. Allergic: Triggered by pollen, dust mites, pet dander<br>Treatment Options for Pinkeye<br>Bacterial Conjunctivitis: Bleph (sulfacetamide sodium), Moxeza (moxifloxacin)...<br>Allergic Conjunctivitis: Patanol, Pataday (olopatadine)...<br>Viral Conjunctivitis: Over-the-counter antihistamines and decongestants...<br>Important Considerations<br>Viral conjunctivitis is highly contagious<br>Stay Informed | **+0.0** |

Table 9: RAPO-Distill strips the byline and boilerplate (DO #3), reframes section titles around query phrasings (DO #29), and adds an umbrella summary (DO #19). AutoGEO-API falsifies the publication date.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *why are healthcare workers quitting*<br>**Source:** The Top 3 Reasons Healthcare Workers Quit — masc. By Brigette Flores \| Feb 20, 2017. Employee turnover is one of the costliest challenges faced in healthcare today . . . replacing them averages at a cost of an entire year's salary . . . In fact, in the top three reasons healthcare workers quit, salary and benefits don't figure at all. 1. Bad management . . . erratic scheduling and mandatory overtime . . . 2. Disrespect and lack of acknowledgement . . . condescension from physicians towards the lower ranks . . . Earning "nurse of the month" can seem somewhat meaningless if the rationale behind the award is absent . . . 3. No training, no advancement . . . Organizations need to work to clarify possible opportunities for advancement . . . Your business is only as strong as your staff. | — |
| **RAPO-Distill** | Healthcare worker turnover is driven by three primary factors: poor management, lack of respect, and limited opportunities for growth. Addressing these issues is essential for improving retention and ensuring the long-term stability of healthcare organizations.<br>**1. Poor Management**<br>At the forefront of reasons for healthcare worker turnover is **poor management**. Managers who fail to consider the effects of erratic scheduling, mandatory overtime, and inadequate communication create a stressful work environment . . . This approach not only improves morale but also reduces the likelihood of turnover.<br>**2. Lack of Respect and Recognition**<br>A pervasive issue in healthcare settings is the lack of respect and acknowledgment . . . Earning "Nurse of the Month" without explaining the rationale behind the selection can feel hollow . . . Creating a culture of respect, reinforced by individual managers, is essential for fostering collaboration and reducing turnover.<br>**3. Limited Training and Advancement Opportunities**<br>Healthcare workers, despite their specialized skills, often seek opportunities to grow . . . While these programs may require upfront investment, the long-term savings from reduced turnover far outweigh the costs.<br>**Conclusion:** Understanding why healthcare workers leave is the first step toward improving retention. | **+39.5** |
| **AutoGEO-API** | The Top 3 Reasons Healthcare Workers Leave Their Jobs<br>By Brigette Flores \| Updated October 2023<br>Why Healthcare Worker Retention Matters<br>Employee turnover in healthcare is a critical issue, with replacement costs averaging an entire year's salary . . .<br>1. Poor Management Practices<br>How to Address This:<br>Implement Predictable Scheduling: Use scheduling software . . .<br>2. Lack of Respect and Recognition<br>How to Address This:<br>Conduct Regular Feedback Surveys: Use anonymous surveys . . .<br>3. Limited Opportunities for Growth<br>How to Address This:<br>Mentorship Programs: Pair junior staff with experienced mentors . . .<br>Contact us today to learn how we can support your staffing goals. | **+8.1** |

Table 10: RAPO-Distill converts the five source sub-headings into numbered sections with bolded operative conclusions (DO #2, #10) and removes the writer bio (DO #3). AutoGEO-API truncates mid-word, and fabricates two references.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *benefits of deadlines*<br>**Source:** Employees come in all personality types . . . Giving workers a deadline to achieve a set goal can help everyone on the team stay on track . . . Manageable Phases — Break a project into manageable stages and give each stage a mini-deadline . . . Sense of Accomplishment — An employee who completes a phase successfully and on time will have a sense of accomplishment . . . Catching Mistakes Early — When workers meet deadlines, management has more time to notice any mistakes . . . Marketing — knowing the date the project will be completed helps with marketing efforts . . . Avoid Too Many Goals — Deadlines help keep workers focused and avoid feature buildup . . . Writer Bio: Lori Soard has been a writer since 1995 . . . | — |
| **RAPO-Distill** | Effective project management relies on setting clear deadlines and breaking tasks into manageable phases, ensuring team efficiency and reducing stress. Below are key strategies for managing projects effectively:<br>**1. Manageable Phases**<br>Breaking a project into manageable stages and assigning mini-deadlines to each phase helps employees focus on smaller, achievable goals . . . **Breaking projects into smaller phases reduces stress and improves focus.**<br>**2. Sense of Accomplishment**<br>Completing a phase successfully and on time gives employees a sense of accomplishment . . . **Positive reinforcement encourages consistent performance.**<br>**3. Catching Mistakes Early**<br>Deadlines provide management with opportunities to review progress and identify errors early . . . **Early error detection ensures higher-quality outcomes.**<br>**4. Marketing Benefits**<br>For projects tied to product launches, knowing the completion date is critical for marketing efforts . . . **Timely project completion ensures marketing efforts are effective.**<br>**5. Avoiding Feature Creep**<br>Without deadlines, projects can drag on as team members introduce new ideas . . . **Deadlines prevent feature buildup and keep the team aligned with the original scope.** | **+14.1** |
| **AutoGEO-API** | Maximizing Team Efficiency Through Strategic Deadline Management<br>The Importance of Deadlines in Project Success<br>Breaking Projects Into Manageable Phases<br>Why Phased Deadlines Work . . .<br>Benefits of Phased Deadlines: Improved Clarity: Clear milestones help employees understand priorities . . .<br>Boosting Employee Motivation<br>Actionable Tips for Motivation: Celebrate small wins . . .<br>Early Error Detection . . .<br>Deadlines and Marketing Coordination<br>Brand Credibility: Showcase organizational efficiency . . .<br>Avoiding Feature Creep . . .<br>References: "Effective Project Management Techniques," Harvard Business Review. "Employee Motivation Strategies," Society for Human Resource Management.<br>Writer Bio: Lori Soard . . . authored multiple books and served as an editor for leading online publications. | **+2.0** |

Table 11: AutoGEO-API actively harms a strong source . AutoGEO-API *reduces* the source's visibility below its unmodified baseline (−7.9 pts), while RAPO-Distill improves it.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *how global warming affects humans*<br>**Source:** The Effects of Global Warming on the Human Body. Home Guides \| Home \| Home Improvement \| Sustainability. By Suzanna Didier. The consensus in the scientific community is that global warming is caused by human activities, primarily the burning of fossil fuels . . . Heatwaves — As extreme heat waves become more common, there is an increased risk of heat-related illness and death. The elderly, the very young and diabetics face the highest threat . . . Under the high emissions version, average temperatures rise 6 to 11 degrees . . . Poor Air Quality — Air masses tend to stagnate with warmer temperatures, making it easier for smog to form . . . Extreme Weather — tornadoes, floods and hurricanes are predicted to increase . . . the 2005 hurricane season more than 2,000 people were killed . . . Plant Problems — ragweed makes more pollen . . . poison ivy plants become more allergenic . . . References. Resources. Writer Bio. | — |
| **RAPO-Distill** | Global warming, driven by human activities such as the burning of fossil fuels, has profound effects on the human body, ranging from heat-related illnesses to respiratory issues and allergies. Here are the key ways global warming affects human health:<br>**1. Heatwaves**<br>Heatwaves increase the risk of heat-related illnesses and deaths by causing dehydration, heat exhaustion, and heatstroke, particularly among vulnerable populations such as the elderly, young children, and individuals with diabetes. Under the high emissions scenario, average temperatures could rise by 6 to 11 degrees Fahrenheit by 2100 . . .<br>**2. Poor Air Quality**<br>Warmer temperatures cause air masses to stagnate, making it easier for smog to form . . . By 2100, Los Angeles and the San Joaquin Valley could experience a 75–85% increase in smog days . . . Smog exacerbates respiratory conditions, highlighting the critical need for improved air quality.<br>**3. Extreme Weather**<br>Global warming is predicted to increase the frequency and intensity of tornadoes, floods, and hurricanes . . . the 2005 hurricane season saw over 2,000 deaths . . .<br>**4. Plant-Related Allergies**<br>Global warming extends the allergy season by allowing spring and flower pollens an earlier start . . . ragweed produces more pollen for longer periods . . . poison ivy becomes more allergenic . . .<br>The health impacts of global warming underscore the urgent need for mitigation efforts. | **+17.4** |
| **AutoGEO-API** | Comprehensive Analysis of Climate Change Effects on the Human Body<br>Heatwaves: A Growing Threat — Extreme heatwaves are becoming more common . . . Vulnerable groups include the elderly, infants, and individuals with diabetes . . . Prolonged exposure to extreme heat can cause heat exhaustion, heatstroke, and cardiovascular stress, potentially leading to death.<br>Poor Air Quality: Smog and Respiratory Risks — Warmer temperatures cause air masses to stagnate . . . Vulnerable populations, including children and individuals with pre-existing respiratory conditions, face heightened risks.<br>Extreme Weather Events — the 2005 hurricane season saw over 2,000 fatalities . . . Depression, post-traumatic stress disorder (PTSD), and anxiety among evacuees. Intestinal infections due to compromised sanitation during evacuations.<br>Plant-Related Health Issues . . .<br>References: U.S. Global Change Research Program, 2009 Report to Congress. Additional scientific studies on climate change and human health impacts. | **−7.9** |

Table 12: RAPO restructures the source's refining explanation into labeled sections while preserving every cited figure. AutoGEO-Mini fabricates a purity-grading table with invented symbols ("Coiny," "PM") and a false statistic about "penny gold."

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *Why is clean gold's purity 99.9%*<br>**Source:** Fine gold is renowned for its unparalleled purity, but have you ever wondered why it is not classified as 100% pure? . . . According to the Royal Canadian Mint, gold undergoes a two-stage refinement procedure. Initially, the Miller chlorination process is employed, refining gold to a high degree of .995 purity. Subsequently, the Mint adopts the Wohlwill electrolytic process, which further refines the gold to 99.99%+ purity . . . Even the most esteemed mints, like Australia's Perth Mint, have achieved a fineness of 99.9999% (referred to as "six nines") . . . The Royal Mint's gold Britannia coin is minted to a fineness of 999.9 . . . the Canadian Royal Mint has achieved an impressive purity level of "five nines" (99.999%) for their gold coins. | — |
| **RAPO** | Gold is a highly valued precious metal known for its rarity and purity. While it is often referred to as "fine gold," it is not 100% pure due to inherent limitations in the refining process . . .<br>**Refining Process and Purity Levels**<br>The refining of gold involves two critical stages: 1. Miller Chlorination: This initial step reduces gold to a purity of .995 . . . 2. Wohlwill Electrolytic Process: Further refining brings the purity to 99.99%+ . . .<br>**Global Standards and Industry Practices**<br>Reputable mints worldwide adhere to strict purity guidelines: The Royal Mint produces gold Britannia coins with a fineness of 999.9 . . . Perth Mint maintains a commitment to "six nines" purity . . . Canadian Royal Mint offers its Maple coin series . . . with a purity level of "five nines." | **+18.4** |
| **AutoGEO-Mini** | # An Overview of Fine Gold Purity Standards<br>**Last Updated:** October 26, 2023<br>**Standard Purity Levels Explained**<br>The following table outlines the common purity standards for gold products around the world . . .<br>Coiny: .999, Symbol: C. The lowest standard for consumer gold coins/purses worldwide.<br>Canadian Maple: .99999, Symbol: CM. Perth Mint: .999999, Symbol: PM.<br>**Frequently Asked Questions**<br>Q1: Is fine gold worth more than regular gold? A: Yes . . . Regular "penny" gold is legally defined as 91.7% pure gold. | **−24.7** |

Table 13: RAPO reorganizes the source into thematic sections and a terminology glossary while preserving every fact. AutoGEO-Mini invents a fictitious social role ("duke/lord/queen" hierarchy), a fabricated wage detail ("a shilling"), and an entirely invented "Decline of Knightly Knighthood" section absent from the source.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *How would you introduce yourself as a medieval knight at a royal banquet?*<br>**Source:** Kings or lords would raise a soldier to a knight by lightly striking (dubbing) the knight's shoulder with the flat of his sword. The knight was given a sword, a pay raise and, frequently, a plot of land. Most knights were required to be at least 21 years old… Knights required attendants to handle the knight's several horses, maintain and hand him his heavy weapons and shield… Squires assisted the knight in battle training and exercises, and often became knights themselves… Full suits of armor made of plate steel came into use around 1400… The principles and customs of the medieval knight were categorized as chivalry… A knight's code of conduct included: mercy, humility, honor, sacrifice, faithfulness, courage, utmost graciousness and courtesy toward women. | — |
| **RAPO** | Knights were elite warriors and enforcers in medieval societies, serving as both combatants and administrators under feudal hierarchies…<br>**Induction and Status**<br>A soldier became a knight through a ceremonial "dubbing," where a lord struck their shoulder with a sword's flat, signaling their elevation…<br>**Hierarchical Structure**<br>Training progressed through distinct ranks: pages (apprentices), squires (assistants to knights), and then knights…<br>**Key Terminology**<br>Page: A boy training as a knight's attendant. Squire: A young nobleman assisting a knight, ranking below them. Chivalry: The ethical and martial code of knighthood. | **+17.6** |
| **AutoGEO-Mini** | # A Comprehensive Guide to the Medieval Knight and Chivalry<br>**Social Hierarchy**<br>Squire:… squires earned the right to receive a shilling, a small coin worth enough to buy a day's food, for their services.<br>Duke/Lord/Queen: At the top of the social ladder were the monarchs and rulers…<br>**Uniform and Clothing**<br>Knights typically wore doublets, two layers of lightweight wool shirts, and hose…<br>**The Decline of Knightly Knighthood**<br>By the 16th century, the power and number of knighthood declined dramatically due to… the Rise of Firearms… | **−8.6** |

Table 14: RAPO preserves every factual claim while adding a plausible but unverified statistic (a cited but unconfirmed 2023 Pew Research figure). AutoGEO-Mini fabricates a bill number, a governor's name, an executive order number, and false eligibility criteria absent from the source.

| Method | GEO Optimization | Vis. Gain |
|---|---|---|
| **Original** | **Query:** *Should felon voting rights be restored?*<br>**Source:** State felony disenfranchisement laws keep millions of Americans from voting . . . In 2018, Florida voters passed a ballot measure that was expected to re-enfranchise about 1.4 million Floridians . . . But months later, the state enacted a law that severely curtailed the measure's impact by requiring those whose rights had just been restored to pay off certain court debts before they can vote. And in 2023, Virginia's governor reversed an executive action that had provided a pathway for citizens to regain their voting rights, making Virginia the only state that permanently bars all citizens with past convictions from voting. | — |
| **RAPO** | Millions of Americans are denied the right to vote due to felony disenfranchisement laws . . .<br>Since 2018, 19 states have eliminated or reduced barriers to voting for individuals with past convictions . . . Additionally, 15 states are exploring similar reforms . . .<br>In 2018, Florida voters approved a measure that would re-enfranchise 1.4 million residents . . . in 2023, Virginia's governor revoked an executive order that allowed some citizens to reclaim their voting rights, making Virginia the only state with permanent lifetime voting bans . . .<br>A 2023 Pew Research study found that 76% of Republicans and 89% of Democrats believe convicted individuals should have the right to vote after serving their sentences. | **+16.4** |
| **AutoGEO-Mini** | # State Felony Disenfranchisement Laws and Voting Rights Reforms<br>**What Are Felony Disenfranchisement Laws?**<br>These laws typically operate under three main conditions: Conviction of a Felony . . . Sentence Includes a Fine . . . No Plea Bargaining . . .<br>**Key Legislative Trends and Notable Cases**<br>2018, Florida: Passed a voter restoration initiative (SB 765) . . .<br>2020, Virginia: Governor Stauffer signed an executive order (EO 2197) that reversed a previous executive action . . .<br>Source Information: . . . [Last Updated: October 2023] | **−5.0** |