# KNOW YOUR TOKENS: STRUCTURE-AWARE GRADIENT ALLOCATION FOR AGENTIC SEARCH VIA REINFORCEMENT LEARNING

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

## ABSTRACT

Reinforcement Learning (RL) from verifiable rewards is a powerful paradigm for training Large Language Model (LLM)-based search agents. Applying RL to agentic search models faces a core technical bottleneck: (a) standard Group Relative Policy Optimization (GRPO) causes gradient instability under parameter-efficient training (LoRA), and (b) existing fixes that subsample tokens by position (Stochastic GRPO, S-GRPO) restore convergence but discard gradient signal from late-position decisions that are equally critical. We show that position is not a well suited proxy for importance in multi-step search trajectories where critical decisions can occur at any position. We propose SA-GRPO (Structure-Aware S-GRPO), a general methodology for training LLM-based search agents via RL that addresses both problems: each model-generated token receives an inclusion probability in the policy gradient loss based on its functional role in the trajectory; search decisions and queries always receive gradient; answer tokens at high probability; reasoning is subsampled; and redundant re-searches are rarely included. Role assignment requires only deterministic parsing of tool-use tags, adding zero compute overhead. We evaluate on the HotpotQA and MuSiQue multi-hop Question Answering (QA) benchmarks using Mistral-7B-Instruct and Qwen2.5-7B-Instruct with QLoRA (NF4 quantization with LoRA adapters), comparing the QLoRA baseline against four RL-trained arms: GRPO, S-GRPO, SA-GRPO, entropy-weighted baseline under identical training conditions. Using a normalized answer-quality score (0–1) on Mistral/QLoRA, all subsampled arms improve over the baseline (0.620) while full GRPO barely moves (0.642). On HotpotQA, S-GRPO leads (0.703) with SA-GRPO close behind (0.684, 50% fully correct vs 40% baseline), under the same expected token budget ($k = 100$). On genuine 2-search trajectories under Qwen, where role-based selection provides maximum benefit, SA-GRPO scores 0.820 versus S-GRPO's 0.777. On the harder compositional questions of MuSiQue, SA-GRPO leads in both regimes: under the resource-constrained Mistral setting it is the *only* arm to improve over the baseline (0.416 vs 0.349, McNemar $p = 0.001$ vs base), while GRPO falls below base and the entropy baseline collapses; under the stronger Qwen base all arms improve and SA-GRPO scores highest (0.590 vs 0.517) though within noise of GRPO (0.579). Beyond the method, we identify a key empirical finding: role-based selection requires multi-hop trajectory diversity to differentiate from position-based selection, a design guideline with implications for evaluation design in agentic RL.

## 1 INTRODUCTION

LLMs equipped with search tools can tackle multi-hop questions requiring iterative retrieval and reasoning. Training such agents via RL requires assigning credit across trajectories that interleave model-generated actions (queries, reasoning, answers) with environment observations (retrieved documents).

Standard GRPO Shao et al. (2024) includes every model-generated token in the policy gradient loss. Under full fine-tuning, where gradient magnitude is distributed across all model parameters, this is feasible. Under LoRA Hu et al. (2022) where only low-rank adapter matrices are trainable, the



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.

Table 1: S-GRPO vs SA-GRPO token selection comparison on a 3-search trajectory. S-GRPO assigns inclusion probability by *position* ($P = 1.0$ for pos $< \alpha$, $P = 0.3$ otherwise). SA-GRPO assigns by *functional role*, ensuring all search decisions receive gradient regardless of position.

| Token span | Role | Position | S-GRPO $P$ | SA-GRPO $P$ | Difference |
|---|---|---|---|---|---|
| <search> (1st) | search_decision | 0–5 | 1.0 | 1.0 | – |
| Query: "Scott Derrickson..." | query | 5–15 | 1.0 | 1.0 | – |
| *[retrieved documents]* | *(excluded)* | – | – | – | – |
| Post-retrieval reasoning | reasoning | 50–95 | **0.3** | **0.5** | +0.2 |
| <search> (2nd) | search_decision | 100–105 | **0.3** | **1.0** | **+0.7** |
| Query: "Ed Wood..." | query_2nd+ | 105–115 | **0.3** | **1.0** | **+0.7** |
| *[retrieved documents]* | *(excluded)* | – | – | – | – |
| Post-retrieval reasoning | reasoning | 150–180 | **0.3** | **0.5** | +0.2 |
| <answer>yes</answer> | answer | 180–190 | **0.3** | **0.8** | **+0.5** |
| **2nd search decision inclusion** | | | **30%** | **100%** | |
| **2nd query inclusion** | | | **30%** | **100%** | |

$P$ = probability of including token in the RL gradient loss. S-GRPO uses position cutoff $\alpha = 10$ (adapted from Lee & Tong's $\alpha = 50$; see Appendix A); SA-GRPO uses functional role (see Table 2).

gradient signal is concentrated into far fewer parameters, causing large magnitude oscillations that prevent convergence. Lee & Tong (2025) demonstrated this empirically on the SVAMP math benchmark (GRPO: 46.7% vs base: 45.0%). In our agentic search setting on HotpotQA under QLoRA, GRPO shows only marginal improvement over the baseline (0.642 vs 0.620), consistent with this instability, whereas position-based subsampling recovers a substantial gain (S-GRPO 0.703). On the harder compositional questions of MuSiQue, however, the position heuristic is no longer sufficient, motivating the role-based selection we introduce below.

S-GRPO Lee & Tong (2025) restores convergence by subsampling *which model-generated tokens are included in the RL loss computation*: it always includes the first $\alpha$ tokens in the loss (where $\alpha = 50$ is a fixed position cutoff), then for each remaining token, includes it with probability $P = 0.3$ (a fixed Bernoulli rate), up to a maximum budget of $k = 100$ tokens per rollout. Tokens not selected are excluded from the policy gradient loss; they receive zero gradient. This position-based heuristic was designed for mathematical reasoning, where earlier tokens establish the proof structure and are disproportionately important (Lee & Tong, 2025, §3.2).

In agentic search, this assumption breaks: critical decisions can occur at any token position in the model's output. A second-hop query at token position 180, reformulating the search after the first retrieval failed, is no less important than a first-hop query at position 8, yet S-GRPO gives it only 30% inclusion probability while always including the first 50 tokens regardless of their content.

The fundamental issue is that *position is a poor proxy for importance in multi-step search*. A search agent's trajectory has known functional structure: the model must (1) decide to invoke a search tool, (2) formulate a query, (3) reason over retrieved results, (4) optionally reformulate and search again, and (5) synthesize a final answer. The agent is prompted to use structured tool-use tags: <search>...</search> to invoke retrieval and <answer>...</answer> to emit a final response. These tags are model-generated as part of the trajectory, providing a natural parse structure for role assignment. Each role has a different functional relationship to the outcome, independently of where it appears in the token sequence (see Table 1).

We propose **SA-GRPO** (Structure-Aware Stochastic GRPO): replace position-based selection with role-based selection using the trajectory's tool-use structure. Tokens not selected are excluded from the policy gradient loss (receive zero gradient). The objective function is identical to S-GRPO (Lee & Tong, 2025, Eq. 5); the budget constraint $k = 100$ is the same; the only change is the selection rule. Role assignment is deterministic; a simple parser over <search> and <answer> tags, requiring zero additional model calls or compute. This makes SA-GRPO a zero-cost upgrade to any existing S-GRPO training pipeline.

We evaluate on HotpotQA Yang et al. (2018) and MuSiQue Trivedi et al. (2022). Multi-hop QA mirrors enterprise knowledge search, where answering a question requires synthesizing information across multiple documents, directly relevant to customer use cases such as enterprise Q&A,

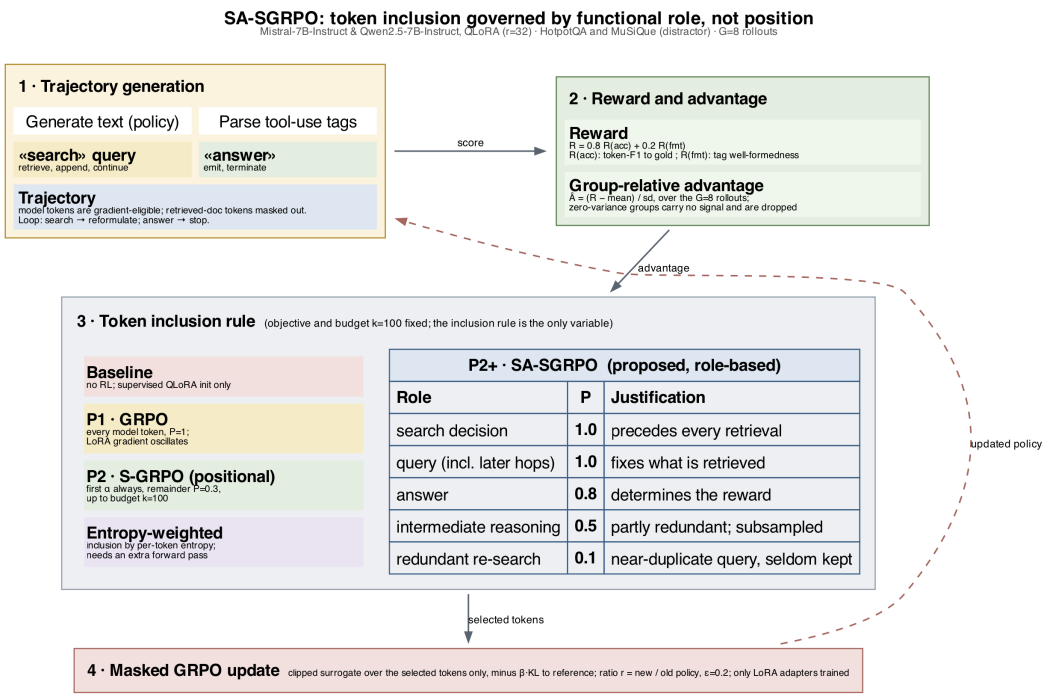



Figure 1: Proposed framework

AI-assisted product search, and support ticket resolution. MuSiQue additionally provides harder compositional multi-hop questions that require deeper multi-step search, the regime where token-role selection matters most.

**Contributions:**

1. **SA-GRPO**: role-based stochastic token selection for agentic search RL, extending S-GRPO with identical budget, objective, and zero implementation overhead.
2. **Regime characterization**: we identify and quantify the conditions under which role-based selection differentiates from position-based selection, showing that trajectory action **diversity** (multi-hop behavior) is a necessary condition.
3. **Controlled four-arm comparison** (GRPO, S-GRPO, SA-GRPO, and an entropy-weighted ablation) on HotpotQA and MuSiQue across two model families (Mistral-7B-Instruct-v0.3 and Qwen2.5-7B-Instruct) under identical QLoRA configuration (NF4, LoRA rank 64), training data (1,350 stratified questions, 300 held-out eval), and hyperparameters ($G$=8 rollouts, $k$=100 token budget for the subsampled arms); the single independent variable is the token selection rule within the GRPO loss.
4. **Resource-constrained RL**: SA-GRPO enables effective policy gradient training under QLoRA on a single GPU (ml.g5.2xlarge, 1×A10G 24 GB), demonstrating that role-based token selection works in the parameter-efficient regime.

## 2 Background

### 2.1 GRPO

For question $x$ with $G$ rollouts $\{y_1, \ldots, y_G\}$, GRPO (Shao et al., 2024, Eq. 2–3) computes group-relative advantages:

$$\hat{A}_i = \frac{R_i - \mu}{\sigma}, \quad \mu = \frac{1}{G}\sum_{i=1}^{G} R_i, \quad \sigma = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(R_i - \mu)^2} \tag{1}$$

where $\hat{A}_i$ is the normalized advantage of rollout $i$, $R_i$ is the scalar reward for rollout $i$ (see Eq. 5), $\mu$ is the mean reward across all $G$ rollouts for this question, $\sigma$ is the standard deviation, and $G$ is the number of rollouts per question ($G = 8$ in our experiments). Questions where all $G$ rollouts share identical reward ($\sigma = 0$) are excluded. The policy gradient objective is:

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|T_i|}\sum_{t\in T_i}\left[\min\Big(r_t\hat{A}_i,\ \text{clip}(r_t, 1\pm\varepsilon)\hat{A}_i\Big)\right] - \beta\, D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) \tag{2}$$

where $\theta$ denotes the trainable parameters (LoRA adapter weights), $r_t = \pi_\theta(y_t)/\pi_{\text{old}}(y_t)$ is the importance sampling ratio at token $t$, $T_i$ is the set of model-generated tokens in rollout $i$ (retrieved document tokens are excluded, following KARL (Chang et al., 2026, §4.2)), $\text{clip}(r_t, 1\pm\varepsilon)$ clamps $r_t$ to $[1-\varepsilon, 1+\varepsilon]$ ($\varepsilon = 0.2$), $\beta = 0.01$ is the KL penalty coefficient, and $D_{\text{KL}}$ measures divergence from the frozen reference model $\pi_{\text{ref}}$. Implementation details for the KL estimator are in Appendix A.

### 2.2 S-GRPO Token Selection

S-GRPO (Lee & Tong, 2025, Eq. 6) replaces the full token set $T_i$ with a stochastic subset $\tilde{T}_i \subseteq T_i$:

$$t \in \tilde{T}_i \iff \begin{cases} t < \alpha & \text{(always included)} \\ t \geq \alpha \text{ and } |\tilde{T}_i| < k & \text{with probability } P \end{cases} \tag{3}$$

where $\alpha = 10$ is the early-token cutoff (adapted from Lee & Tong's $\alpha = 50$; see Appendix A), $k = 100$ is the maximum token budget, and $P = 0.3$ is the Bernoulli sampling probability.

## 3 Method: SA-GRPO

We observe that in agentic search trajectories, position is an unreliable proxy for token importance. A query formulation token at position 150 (second search) most directly influences the success of the second retrieval step; a filler token at position 10 does not. Yet S-GRPO always includes the latter and gives the former only 30% probability.

SA-GRPO replaces position with *functional role* as the selection criterion. The objective and budget are unchanged; only the selection rule changes.

### 3.1 Role Assignment

Each model-generated token is assigned a functional role by parsing the trajectory's tool-use tags.

Table 2: SA-GRPO role assignment and inclusion probabilities.

| Role | Definition | $P$ | Rationale |
|---|---|---|---|
| `search_decision` | <search> tag tokens | 1.0 | Upstream of all retrieval |
| `query` | Inside <search>...</search> | 1.0 | Determines retrieval result |
| `query_2nd+` | 2nd+ query, ≤50% overlap | 1.0 | Novel multi-hop query |
| `answer` | Inside <answer>...</answer> | 0.8 | Directly determines reward |
| `reasoning` | Other model-generated tokens | 0.5 | Partially redundant |
| `verification` | 2nd+ query, >50% overlap | 0.1 | Redundant (Chang et al., 2026, §8.2.2) |

Role assignment is deterministic and zero-cost: the parser reads character-level tag spans, maps them to token positions, and computes Jaccard token overlap to distinguish novel multi-hop queries from redundant re-searches. The inclusion probabilities are domain-informed heuristics, not optimized hyperparameters.

### 3.2 Token Selection Rule (Novel)

SA-GRPO replaces S-GRPO's position-based rule (Eq. 3) with:

$$P(t \in \tilde{T}_i) = P_{\text{role}}(t) \tag{4}$$

where $P_{\text{role}}(t)$ is the inclusion probability from Table 2. Each token is independently included via a Bernoulli draw, matching S-GRPO's sampling mechanism to ensure the single independent variable is the selection criterion.

### 3.3 BUDGET CONTROL

To ensure a fair comparison, we enforce the same *expected* token budget $k = 100$. If $\mathbb{E}[|\tilde{T}_i|] > k$, non-mandatory probabilities are scaled by $\gamma = (k - n_{\text{always}})/\sum_{t:P<1} P_{\text{role}}(t)$.

## 4 EXPERIMENTAL SETUP

### 4.1 DATASET

We use the **HotpotQA** Yang et al. (2018) and **MuSiQue** Trivedi et al. (2022) distractor settings, each providing multiple context paragraphs per question (gold paragraphs mixed with distractors). We use dense retrieval via sentence-transformer embeddings (cosine similarity, top-3 results), run on CPU to preserve GPU memory for online RL training on a single A10G. The embedding model encodes each document as a title-plus-text concatenation, and queries are ranked by cosine similarity against the document embeddings.

**Training data.** We generate $G = 8$ rollouts per question using online rollout generation with the current policy. Training uses 1,350 stratified questions with 1 pass (450 optimizer steps); 3 questions per mini-batch with no gradient accumulation (24 rollouts per optimizer step). An early-stopping criterion monitors the moving-average F1 and halts training if accuracy decays for 4 consecutive checkpoint windows (25 steps each).

**Evaluation.** 300 held-out evaluation questions, distinct from training set.

### 4.2 REWARD FUNCTION

$$R(x, y) = \max(0,\ 0.8 \cdot R_{\text{accuracy}}(y, \text{gold}) + 0.2 \cdot R_{\text{format}}(y) - 0.2 \cdot \mathbb{1}[\text{no search}]) \tag{5}$$

$R_{\text{accuracy}}$ is the token-level F1 score between predicted and gold answer. $R_{\text{format}} \in \{0.0, 0.5, 1.0\}$: 1.0 if the trajectory contains both `<search>` and `<answer>` tags. A no-search penalty of 0.2 is subtracted when the trajectory issues no search, discouraging the model from answering without retrieval, and the total is clamped to be non-negative.

### 4.3 MODEL AND TRAINING

**Base models.** Mistral-7B-Instruct-v0.3 Jiang et al. (2023) and Qwen2.5-7B-Instruct Team (2025), each loaded in NF4 quantization (QLoRA) Dettmers et al. (2023). Only LoRA adapter weights are updated during RL training.

**QLoRA configuration.** Rank $r = 64$, $\alpha = 32$, dropout$= 0.05$, targeting {q_proj, k_proj, v_proj, o_proj}. Gradient checkpointing enabled.

**Optimizer.** AdamW, $\text{lr} = 1 \times 10^{-5}$, gradient norm clipping 1.0, 1 epoch per batch. Batch: 3 questions $\times$ $G = 8$ rollouts $= 24$ rollouts per mini-batch, with no gradient accumulation (24 rollouts per optimizer step). $\beta = 0.01$ KL penalty.

**Infrastructure.** SageMaker Studio, ml.g5.2xlarge (1×A10G, 24 GB GPU).

### 4.4 EVALUATION METRICS

We use **LLM-judged answer quality** as the primary evaluation metric. A three-model judge council (Haiku4.5, Qwen3-32B, and Gemma-3-12B) independently scores each predicted answer on a 1–5 scale (1=Wrong, 2=Related but wrong, 3=Partial, 4=Mostly correct, 5=Correct) given the question and gold answer. The council score is the mean of valid judge scores, normalized to 0–1 ($\text{llm_norm} = \overline{\text{score}}/5$). This captures semantic correctness that token-level F1 misses; for example, "American" and "He is an American citizen" both correctly answer a nationality question, but F1 penalizes the verbose form.

**Secondary metrics.** We also report token-level F1 (HotpotQA standard, word-overlap after normalization), exact match (EM), whole-word containment (*contains*), and mean searches per question.

From the judge scores we derive %Correct ($\mathrm{round}(\overline{\mathrm{score}}) = 5$), %Good ($\mathrm{round}(\overline{\mathrm{score}}) \geq 4$), and %Poor ($\mathrm{round}(\overline{\mathrm{score}}) \leq 2$).

**Stratified analysis.** We report results both overall and stratified by the number of search actions in the trajectory (0, 1, 2, 3+). The 2-search stratum is of particular interest, as it represents the canonical multi-hop retrieval pattern where role-based token selection provides maximum differentiation from position-based methods.

**Statistical tests.** Arm-vs-baseline comparisons use McNemar's test on paired binary correctness (%Good) with Holm correction for multiple comparisons. Effect sizes are reported as paired bootstrap 95% confidence intervals on $\mathrm{llm_norm}$ differences (10,000 resamples).

# 5 RESULTS

## 5.1 MAIN RESULTS

Table 3: Overall results ($N = 300$ eval). LLM Score = normalized 3-judge council mean (0–1). Bold = best per model–dataset block.

| | | HotpotQA | | | | | MuSiQue | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Arm** | **LLM** | **%Corr** | **%Good** | **%Poor** | **Srch** | **LLM** | **%Corr** | **%Good** | **%Poor** | **Srch** |
| Mistral | Base | 0.620 | 40% | 44% | 50% | 0.81 | 0.349 | 13% | 15% | 82% | 1.10 |
| | GRPO | 0.642 | 43% | 47% | 46% | 1.70 | 0.331 | 12% | 14% | 84% | 1.05 |
| | S-GRPO | **0.703** | **51%** | **56%** | **37%** | 1.65 | 0.358 | 15% | 18% | 81% | 1.64 |
| | SA-GRPO | 0.684 | 50% | 53% | 42% | 1.89 | **0.416** | **20%** | **24%** | **73%** | 1.99 |
| | Entropy | 0.674 | 48% | 53% | 41% | 1.81 | 0.240 | 4% | 4% | 95% | 0.57 |
| Qwen | Base | 0.750 | 59% | 64% | 32% | 1.17 | 0.517 | 34% | 36% | 60% | 1.89 |
| | GRPO | 0.763 | 61% | 66% | 29% | 1.17 | 0.579 | 41% | 44% | 52% | 2.04 |
| | S-GRPO | 0.762 | 59% | 65% | 29% | 1.15 | 0.562 | 40% | 42% | 54% | 1.99 |
| | SA-GRPO | 0.759 | 61% | 65% | 31% | 1.18 | **0.590** | **43%** | **46%** | **51%** | 1.84 |
| | Entropy | **0.772** | **62%** | **66%** | **28%** | 1.26 | 0.558 | 40% | 42% | 55% | 1.93 |

Under Mistral (resource-constrained), all RL arms improve over the base (0.620). S-GRPO leads on HotpotQA (0.703, 51% correct vs 40% base). GRPO shows minimal improvement (0.642), consistent with gradient instability under LoRA. Under Qwen (stronger base at 0.750), all arms cluster between 0.759–0.772 with marginal differentiation.

On MuSiQue, SA-GRPO leads in both regimes but the degree of separation depends on base strength. Under Mistral (evaluated at step-200, where training was stopped by design short of a full pass), SA-GRPO is the *only* arm to improve over base (0.416 vs 0.349, McNemar $p = 0.001$): uniform GRPO falls below base (0.331, $p = 0.79$), S-GRPO ties it (0.358, $p = 0.32$), and the entropy baseline collapses into degenerate output (0.240). Under Qwen (full training) all four arms beat base (0.517) and SA-GRPO scores highest (0.590, $p = 0.001$) but is statistically tied with GRPO (0.579). The harder compositional multi-hop questions in MuSiQue are the setting where role-based selection separates from position-based selection, most decisively in the resource-constrained regime, whereas on the shallower HotpotQA questions S-GRPO is sufficient.

RL training under Mistral reduces zero-search behavior from 50% to <5% across all arms. Under Qwen, the base already searches effectively (10% zero-search), leaving less room for behavioral change.

## 5.2 2-SEARCH TRAJECTORIES

On 2-search trajectories under Qwen, SA-GRPO scores 0.820, a 5.5% improvement over S-GRPO (0.777). The 2nd search decision at a late token position receives full gradient under SA-GRPO but only stochastic inclusion under S-GRPO. Under Mistral, the 2-search improvement is more diffuse across arms, with entropy (0.701) and SA-GRPO (0.689) both exceeding S-GRPO (0.681). On MuSiQue, SA-GRPO leads the 2-search stratum in *both* regimes (Mistral 0.412, Qwen 0.682), consistent with role-based selection helping most where a second search decision must receive gradient.

Table 4: LLM Score on 2-search trajectories (the multi-hop retrieval pattern). $n$ = trajectory count.

| | HotpotQA | | MuSiQue | |
|---|---|---|---|---|
| **Arm** | **Mistral** | **Qwen** | **Mistral** | **Qwen** |
| Base | 0.662 (79) | 0.770 (76) | 0.396 (109) | 0.581 (136) |
| GRPO | 0.671 (137) | 0.770 (61) | 0.356 (73) | 0.647 (172) |
| S-GRPO | 0.681 (159) | 0.777 (62) | 0.405 (142) | 0.620 (169) |
| SA-GRPO | 0.689 (153) | **0.820** (58) | **0.412** (200) | **0.682** (163) |
| Entropy | 0.701 (150) | 0.793 (79) | 0.376 (25) | 0.624 (172) |

### 5.3 Head-to-Head vs Base

Table 5: Per-question wins/losses vs base ($N = 300$). Win = arm scored strictly higher.

| | HotpotQA (Mistral) | | | HotpotQA (Qwen) | | | MuSiQue (Mistral) | | | MuSiQue (Qwen) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **W** | **L** | **T** | **W** | **L** | **T** | **W** | **L** | **T** | **W** | **L** | **T** |
| GRPO | 86 | 73 | 141 | 31 | 26 | 243 | 45 | 69 | 186 | 69 | 39 | 192 |
| S-GRPO | **107** | **49** | 144 | 31 | 30 | 239 | 59 | 69 | 172 | 65 | 47 | 188 |
| SA-GRPO | 102 | 77 | 121 | 31 | 29 | 240 | **75** | **47** | 178 | **75** | **46** | 179 |
| Entropy | 95 | 67 | 138 | 35 | 29 | 236 | 12 | 88 | 200 | 65 | 47 | 188 |

Under Mistral on HotpotQA, S-GRPO has the strongest win rate (107W/49L). Under Qwen, all arms are closely matched against the base (net $+1$ to $+6$), consistent with the tight overall scores. On MuSiQue the pattern is regime-dependent: under Mistral, SA-GRPO is the only arm with a positive margin (75W/47L, $+0.067$, $p = 0.001$), while GRPO falls below base (45W/69L, $p = 0.79$) and entropy collapses (12W/88L, $p < 0.001$ worse); under Qwen all arms beat base and SA-GRPO's lead (75W/46L, $+0.074$, $p = 0.001$) is matched by GRPO (69W/39L, $p = 0.003$). After Holm correction for the four per-condition comparisons, SA-GRPO and GRPO remain significant under Qwen whereas S-GRPO ($p = 0.031$) and entropy ($p = 0.054$) do not.

### 5.4 Regime Boundary

These results identify a regime boundary for token selection:

1. **Resource-constrained** (Mistral/QLoRA): GRPO barely improves on HotpotQA (0.642 vs 0.620 base). All subsampled arms improve substantially, with S-GRPO leading (0.703). Token subsampling is necessary because full GRPO fails.
2. **Harder multi-hop** (Mistral/MuSiQue): SA-GRPO is the *only* arm to improve over base (0.416 vs 0.349, $p = 0.001$); GRPO falls below base and the entropy baseline collapses. When questions require genuine compositional retrieval in the resource-constrained regime, role-based selection is the only rule that keeps RL improving.
3. **Strong base** (Qwen): Arms cluster at 0.759–0.772 on HotpotQA and 0.558–0.590 on MuSiQue. SA-GRPO scores highest on MuSiQue (0.590) but is statistically tied with GRPO (0.579). The selection rule provides diminishing returns when the base is already competent.
4. **Multi-hop subset** (2-search): SA-GRPO leads under Qwen (0.820 vs 0.777 S-GRPO). Role-based selection, which guarantees gradient to both search decisions regardless of position, provides maximum benefit on genuine multi-hop trajectories.

Structure-aware token selection is most valuable when training is difficult (weak base, parameter-efficient) or when evaluation targets multi-hop reasoning, and less critical when the base model is already competent.

## 6 CONCLUSIONS

We present **SA-GRPO**, a role-based token selection strategy for policy gradient training of agentic search models. Its advantage is regime-dependent. On the shallower HotpotQA questions, S-GRPO's position-based selection is sufficient and leads (0.703), with SA-GRPO close behind (0.684). On the harder compositional questions of MuSiQue, SA-GRPO leads in both regimes: under the resource-constrained Mistral setting it is the *only* arm to improve over base (0.416 versus 0.349, McNemar $p = 0.001$, while GRPO falls below base and the entropy baseline collapses), and under Qwen it scores highest (0.590) though within noise of GRPO (0.579); on genuine 2-search trajectories under Qwen it reaches 0.820 against S-GRPO's 0.777. Role-based selection separates from position-based selection most clearly where multi-hop structure is present and training is hard.

The implementation overhead is zero: role assignment requires only a deterministic parser over tool-use tags. SA-GRPO is a drop-in replacement for S-GRPO's position-based selection in any existing training pipeline.

Our empirical results reveal that single-search trajectory dominance limits the differentiation between role-based and position-based selection. **Type-aware selection requires action diversity to differentiate from position-based selection.** This provides a design guideline: role-based token selection is most valuable in settings with functionally diverse trajectories. More broadly, role-based token selection exploits the agent's structured action space, a pattern that generalizes to any tool-using LLM agent (code execution, API calls, database queries) whose action structure provides a natural token role taxonomy.

## 7 LIMITATIONS AND ANALYSIS

**Multi-hop trajectory scarcity.** On HotpotQA under Qwen, only 20–28% of evaluation trajectories contain 2+ searches; under Mistral, RL training increases this to 50–62%, but single-search trajectories still dominate Qwen's distribution. Role-based selection provides limited differentiation when most trajectories follow a single search-then-answer pattern. Evaluation on settings that naturally produce deeper multi-hop behavior (e.g., full-wiki retrieval with dense passage retrieval Karpukhin et al. (2020)) would better isolate the SA-GRPO advantage.

**Retrieval ceiling.** Embedding-based retrieval over the distractor pool does not always surface the gold paragraphs, placing an upper bound on answer quality that is independent of the token selection strategy. This is visible at the tail of the search-count distribution: a fraction of questions exhaust the five-search budget, and score near zero, due to repeating queries on innacurate passages. **Additional searches cannot compensate when the retriever never surfaces the evidence**. Because this ceiling sits upstream of the policy, it caps all arms equally and does not change their relative ranking, but it bounds absolute accuracy. Stronger retrieval, such as dense passage retrieval Karpukhin et al. (2020), cross-encoder reranking, or explicit query decomposition, is a promising direction for lifting this ceiling.

**Hand-set role probabilities.** SA-GRPO's inclusion probabilities (1.0, 0.8, 0.5, 0.1) are heuristics without sensitivity analysis. The entropy arm provides a data-driven alternative but does not consistently outperform SA-GRPO on multi-hop trajectories, suggesting the hand-set values are reasonable but not necessarily optimal.

**LoRA rank.** All experiments use rank $r = 64$; higher ranks may reduce the gradient instability that makes token subsampling necessary, potentially narrowing the gap between GRPO and subsampled methods.

### AI USE STATEMENT

The methodology of this work incorporates a large language model as a core experimental component: a three-model judge council (Haiku 4.5, Qwen3-32B, and Gemma-3-12B, accessed via Amazon Bedrock) serves as the automated judge that scores answer quality on a 1–5 scale (Section 4). This use is a measured variable in our evaluation pipeline, not an aid to the research process, and is described in full in the main text.

We additionally used generative AI tools to assist with manuscript preparation, specifically for copy-editing and improving the clarity of prose. All AI-assisted text was reviewed and edited by the authors, and we take full responsibility for the final content. We did not use generative AI to generate research ideas, design experiments, produce or analyze results, or write code for the core method; the research contributions, experimental design, results, and their interpretation are the authors' own.

## REFERENCES

Jonathan D. Chang, Andrew Drozdov, Shubham Toshniwal, et al. KARL: Knowledge agents via reinforcement learning. *arXiv preprint arXiv:2603.05218*, 2026.

Zhe Chen, Jian Yang, Tong Xiao, et al. Reinforcement learning for tool-integrated interleaved thinking towards cross-domain generalization. *arXiv preprint arXiv:2510.11184*, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

Rui Ding, Yang Lv, Xiang Meng, et al. PRPO: Aligning process reward with outcome reward in policy optimization. *arXiv preprint arXiv:2601.07182*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, et al. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. arXiv:2004.04906.

Alex Lee and Hanghang Tong. Token-efficient RL for LLM reasoning. *arXiv preprint arXiv:2504.20834*, 2025.

Jie Lu, Jian Wu, Jing Li, et al. Bridging perception and reasoning: Token reweighting for RLVR in multimodal LLMs. *arXiv preprint arXiv:2603.25077*, 2026.

Prasanna Parthasarathi, Mathieu Reymond, Boyuan Chen, Yuchen Cui, and Sarath Chandar. GRPO-$\lambda$: Credit assignment improves LLM reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Hao Sang, Yang Xu, Zhi Zhou, Ran He, and Zhi Wang. Not all tokens are needed (NAT): Token efficient reinforcement learning. *arXiv preprint arXiv:2603.06619*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jian Wu, Yannis George, Jin Ye, Yi Wu, Daniel F. Schmidt, and Jian Cai. SPINE: Token-selective test-time reinforcement learning with entropy-band regularization. *arXiv preprint arXiv:2511.17938*, 2025.

Guang Xie, Yu Shi, Hao Tian, Ting Yao, and Xin Zhang. CAPO: Towards enhancing LLM reasoning through generative credit assignment. *arXiv preprint arXiv:2508.02298*, 2025.

Tao Xu, Yang Chen, and Ming Li. CLEANER: Self-purified trajectories boost agentic reinforcement learning. *arXiv preprint arXiv:2601.15141*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

# APPENDICES

## A IMPLEMENTATION DETAILS

**LoRA vs QLoRA terminology.** We use QLoRA Dettmers et al. (2023) to refer to LoRA adapters applied to an NF4-quantized base model. The gradient instability arises from LoRA's low-rank parameterization, independent of quantization.

**KL divergence estimator.** The KL term in Eq. 2 uses the Schulman approximation: $\hat{D}_{\mathrm{KL}} = \frac{1}{n}\sum_t[(r_t - 1) - \log r_t]$ where $r_t = \pi_\theta(y_t)/\pi_{\mathrm{ref}}(y_t)$, guaranteed non-negative by $x - 1 \geq \log x$.

**S-GRPO $\alpha$ adaptation.** Lee & Tong used $\alpha = 50$ for SVAMP (100–500 token outputs). For HotpotQA's shorter outputs (25–300 tokens), $\alpha = 50$ makes S-GRPO equivalent to GRPO for outputs under 50 tokens. We reduce $\alpha$ to 10.

## B WHY WE COMPARE TO GRPO AND S-GRPO, NOT KARL

KARL Chang et al. (2026) trains full-precision models via OAPL-the gradient instability motivating SA-GRPO does not arise. KARL asks "model vs environment tokens?" (binary); SA-GRPO asks "what role does this model token play?" (graded). The two are composable.

## C DISCUSSION

### C.1 FORMAT REWARD PROVIDES NO GRADIENT SIGNAL

With >98% format compliance, the 0.2 format weight contributes a constant that cancels in advantage normalization. The RL signal is purely accuracy-driven.

### C.2 REGIME CHARACTERIZATION

SA-GRPO's advantage scales with the multi-hop fraction $f_{\mathrm{multi}}$. In our settings $f_{\mathrm{multi}}$ is roughly one third, limiting but not eliminating differentiation. The improvement on 2-search trajectories, up to about 10% on Qwen/MuSiQue and 5.5% on Qwen/HotpotQA, confirms the mechanism, while the dominance of single-search trajectories limits the aggregate effect.

### C.3 PATH TO MULTI-HOP RL

A multi-hop reward bonus and larger retrieval corpus would increase trajectory diversity, amplifying SA-GRPO's advantage.

## D RELATED WORK

**S-GRPO/T-SPMO** Lee & Tong (2025): SA-GRPO extends S-GRPO, changing only selection criterion. **NAT** Sang et al. (2026): Horvitz–Thompson correction compatible as future extension. **SPINE** Wu et al. (2025): entropy-based (forward pass needed) vs our tag-based (string parsing).

**KARL** Chang et al. (2026): binary mask; SA-GRPO adds graded selection within model tokens. **RITE** Chen et al. (2025), **CLEANER** Xu et al. (2026): trajectory-level, complementary. **GRPO-$\lambda$** Parthasarathi et al. (2025), **CAPO** Xie et al. (2025), **PRPO** Ding et al. (2026): improve per-token advantages; complementary to SA-GRPO's selection. **ToR** Lu et al. (2026): concurrent type-aware differentiation for multimodal; validates the principle.

## E SA-GRPO Role Assignment Example

**Multi-hop trajectory:**

```
<search>Scott Derrickson nationality</search>[results]
<search>Ed Wood nationality</search>[results]
<answer>yes</answer>
```

"Ed Wood nationality" has Jaccard $\approx 0.33 \leq 0.5$ with the prior query $\rightarrow$ classified as `query_2nd+` ($P = 1.0$), not verification. S-GRPO gives this 2nd search $\sim$30% inclusion; SA-GRPO gives it 100%.