BRIDGE: BILEVEL RETRIEVAL-CREDIT-AWARE AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic reinforcement learning (ARL) with verifiable rewards improves the ability of large language models (LLMs) to tackle knowledge-intensive tasks by learning to interleave search and reasoning. However, most existing ARL methods optimize only LLM-generated tokens and treat retrieved evidence as environment observations. This creates an *information-credit gap*: failures caused by missing or misleading evidence are attributed to the LLM policy rather than to the retriever, which motivates training the LLM and the retriever jointly. In this paper, we show that retrieval and LLM policy learning are order-sensitive: adapting the retriever before optimizing the policy yields a larger reward gain than the reverse order. To preserve this hierarchy while allowing both components to co-adapt, we formulate retrieval-augmented agentic RL as a bilevel optimization problem. To solve it efficiently, we introduce BRIDGE, a memory-efficient first-order bilevel method motivated by a loss-landscape analysis of the RL and retrieval objectives. Across seven open-domain QA benchmarks, BRIDGE achieves the highest average accuracy with both 3B and 7B backbones, improving the multi-hop average over the strongest baseline by 9.6 and 3.3 EM points, respectively. It also achieves the best averaged answer accuracy and reasoning quality across medical QA benchmarks.

1 INTRODUCTION

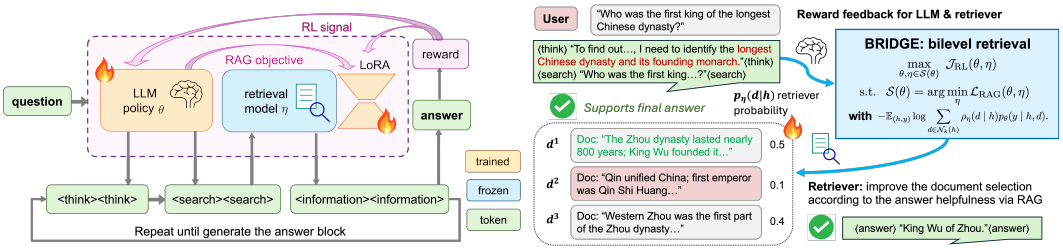
Agentic reinforcement learning (ARL) trains large language models (LLMs) to solve tasks through multi-turn interactions with external tools, which greatly improves the ability of LLMs to handle complex, knowledge-intensive tasks (Xie et al., 2026). In retrieval-augmented question answering, an agent alternates between reasoning, search actions, retrieved observations, and answer generation, and the whole trajectory can be optimized via RL from a sparse verifiable reward (Jin et al., 2025; Song et al., 2025). Through ARL, the LLM can learn when to invoke external retrieval and how to use the retrieved information without dense supervision (Jin et al., 2025; Jiang et al., 2026).

However, most ARL methods optimize the LLM only on LLM-generated action tokens (reasoning, search, and answer tokens), while retrieved information tokens serve as environment observations and are masked out of the RL loss (Jin et al., 2025; Kimi Team et al., 2025). This creates an *information-credit gap*: when a rollout fails because the retrieved evidence is missing or misleading, the RL update penalizes the LLM's actions around that evidence instead of improving retrieval, which may degrade the reasoning performance and generalization of the learned policy. This gap is especially severe in multi-hop question answering (Trivedi et al., 2023), where the evidence is spread across documents and the LLM can learn reasoning only after retrieval exposes all of them.

Other lines of work either adapt the retriever to a frozen generator to improve document selection (Jiang et al., 2025b), or jointly optimize the retriever and the LLM policy (Chen et al., 2025; Liu et al., 2026). However, neither considers the ordered dependency between retrieval and policy learning. In this paper, we argue that retrieval and LLM policy optimization have a natural hierarchical order: the retriever should first be optimized to make useful evidence available, and the LLM should then learn to search and reason with that evidence.

Moreover, instead of a one-shot retriever update followed by RL training, we formulate retrieval-augmented agentic RL as a bilevel optimization problem that co-adapts retrieval and LLM parameters hierarchically throughout training. For each LLM policy θ, we seek the retrieval parameter η that best supports answer generation, and then update the policy via RL on rollouts that use this

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



(a) BRIDGE framework **(b)** Bilevel optimization formulation

Figure 1: **Overview of BRIDGE. (a)** BRIDGE couples the LLM policy and the query-side retriever adapter during multi-turn tool-use rollouts: the LLM generates reasoning and search actions, the retriever supplies information, and the final reward provides feedback to both components. **(b)** The bilevel view treats answer-level RL as the upper objective and RAG likelihood as the retrieval objective, so the retriever is updated toward evidence that improves downstream answer generation.

retriever, with the verifiable outcome reward increasing the likelihood of successful rollouts; see Figure 1(b) for an overview.

However, most existing bilevel methods are either second-order methods based on conjugate gradient, unrolling, or implicit differentiation, which are computationally expensive (Chen et al., 2021; Ji et al., 2021; Lorraine et al., 2020; Dontchev & Rockafellar, 2014), or first-order methods that must optimize and store two lower-level models to track the solutions of the original and the perturbed lower-level problems (Kwon et al., 2023; Shen et al., 2025b; Zucchet & Sacramento, 2022), which is memory-intensive for retriever adaptation. To reduce this memory burden, recent work treats the lower-level optimal value as nearly constant in θ and therefore avoids maintaining a second lower-level model (Xiao et al., 2026; Shen et al., 2025a). However, we show that this assumption is violated in our setting, where the optimal retrieval objective value is sensitive to the LLM parameters.

Guided by a landscape analysis and motivated by an efficient bilevel solver (Jiang et al., 2025a), we introduce BRIDGE, a memory-efficient first-order bilevel method for retrieval-augmented ARL. As shown in Figure 1, BRIDGE updates a query-side low-rank adaptation (LoRA) retriever adapter with both trajectory-level RL credit at retrieval turns and a RAG likelihood term that scores whether retrieved evidence supports the target answer, while the LLM is updated with trajectory-level RL credit on rollouts that contain the information returned by the current retriever.

We summarize our main contributions below.

C1) We identify the information-credit gap caused by a fixed retriever in agentic RL and the hierarchical order between retrieval and LLM optimization. To jointly improve retrieval quality and the LLM's reasoning and search capabilities, we formulate multi-turn retrieval-augmented ARL as a bilevel optimization problem.

C2) Guided by landscape analysis, we propose BRIDGE, a memory-efficient first-order bilevel algorithm that updates the retrieval LoRA parameters using retrieval-turn RL credit and RAG likelihood, and then optimizes the LLM policy with RL credit. Relative to the two-state first-order estimator, BRIDGE reduces the peak memory of each retriever update by 49.7%, while its gradient estimate stays closely aligned with the two-state estimate.

C3) We validate BRIDGE on seven open-domain and five medical QA benchmarks. On the open-domain benchmarks, it improves the overall average EM over the strongest baseline, VT-Search, by 7.1 and 2.3 points with 3B and 7B backbones; on the medical benchmarks, it achieves the highest average answer accuracy and reasoning quality.

1.1 Related Work

Agentic RL with tool use. Tool-using LLM agents learn to interleave reasoning with external actions such as search, code execution, or API calls. Early agentic prompting and tool-use frameworks show that explicit reasoning-action traces can improve tool interaction (Yao et al., 2023; Schick et al., 2023). Recent ARL methods train this behavior with verifiable rewards, including search-augmented RL systems and general tool-use RL frameworks (Li et al., 2025; Jin et al., 2025; Song et al., 2025; Jiang et al., 2025b; 2026; Li et al., 2026a), and have been extended to modular agentic

systems, where the planner is optimized via ARL within the multi-turn interaction (Li et al., 2026b). Building on this foundation, closely related work extends LLM policy-side ARL to jointly optimize the LLM policy and the retriever (Chen et al., 2025; Liu et al., 2026). Our work complements these efforts by introducing a bilevel formulation that explicitly captures the hierarchical dependency between policy training and retriever adaptation, and preserves this hierarchy during joint training.

Retrieval-augmented generation. Retrieval-augmented generation (RAG) improves knowledge-intensive generation by conditioning the generator on information retrieved by the retrieval model from an external datastore (Lewis et al., 2020). A common line of work keeps the retriever model fixed and adapts the LLM to retrieved evidence, through domain-specific instruction tuning, search-augmented instruction learning, or RL fine-tuning with retrieval curricula (Zhang et al., 2024; Luo et al., 2023; Huang et al., 2026). A second line keeps the LLM fixed and optimizes retrieval or query-side behavior so that retrieved passages better support the frozen generator (Shi et al., 2024; Ma et al., 2023; Ram et al., 2023). Recent work explores RAG optimization through dual instruction tuning, environment feedback, LLM-guided dynamic reranking, and joint optimization of document selection and answer generation (Lin et al., 2024; Gao et al., 2025; Sun et al., 2025; Shi et al., 2026). These methods improve retrieval-conditioned generation, but they do not address the multi-turn RL setting where search queries and final answers are induced by an evolving policy.

2 Preliminaries: Multi-turn Tool-use RL and RAG

2.1 Agentic RL with tool use

In the multi-turn agentic RL setting, given a prompt from the training distribution $x \sim \mathcal{D}_{\text{tr}}$, the LLM generates the rollout trajectory as $\tau = (a_1, o_1, \ldots, a_{T-1}, o_{T-1}, a_T)$, where a_t denotes the t-th segment of LLM-generated action tokens and o_t denotes the environment tokens returned by the external tool server. Actions and retrieved information are wrapped in special tokens. In search tasks, reasoning, search calls, and final answers are wrapped by `<think>`, `<search>`, and `<answer>`, respectively. When triggered by the search calls, external tools retrieve documents and wrap the relevant information in an `<information>` block as the observation tokens o_t.

Multi-turn RL with verifiable rewards. In agentic RL, the rollout trajectory τ for the prompt x is multi-turn and contains both LLM-generated action tokens a_t and retriever-generated observation tokens o_t. Therefore, the distribution of the rollout trajectory τ depends on both LLM parameters θ and the retrieval parameters η. Given the retrieval model η, in order to learn the LLM policy parameterized by θ, we leverage reinforcement learning with verifiable rewards to maximize a predefined scalar reward with a KL regularization term to prevent reward over-optimization

$$\mathcal{J}_{\text{RL}}(\theta, \eta) := -\mathbb{E}_{x\sim\mathcal{D}_{\text{tr}},\, \tau\sim\pi_\theta(\cdot|x;\eta)} \left[R(x, \tau) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right], \tag{1}$$

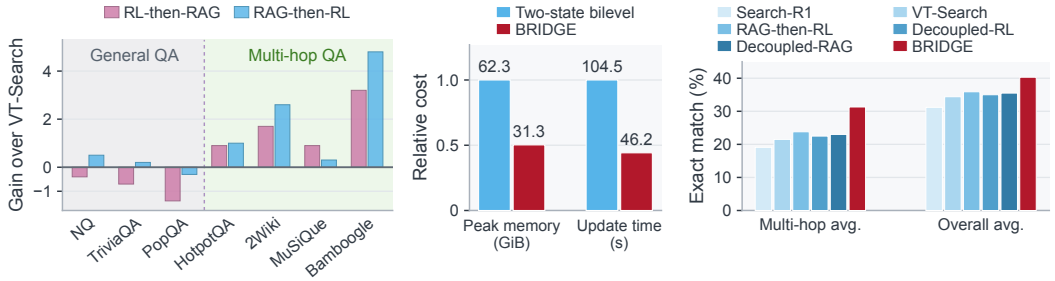
where π_{ref} denotes a reference LLM and $\pi_\theta(\cdot \mid x; \eta)$ is the LLM policy under the retrieval model η. Following Jin et al. (2025), we use a binary exact-match outcome reward $R(x, \tau)$, which measures the correctness of the generated answer $\hat{y}(\tau)$ compared with the ground-truth answer $y^*(x)$. Specifically, $R(x, \tau) = 1$ if $\hat{y}(\tau)$ is correct, and $R(x, \tau) = 0$ otherwise.

Group Relative Policy Optimization (GRPO). To optimize (1), we follow Jin et al. (2025) and use GRPO (Shao et al., 2024), which assigns per-rollout credit by comparing multiple rollouts of the same prompt. Specifically, given a prompt x, GRPO samples a group of G trajectories and normalizes their rewards within the group. Let τ_i be the i-th rollout trajectory with reward $R_i = R(x, \tau_i)$, and let the advantage of the i-th rollout be $A_i = (R_i - \text{mean}(\{R_j\}))/\,\text{std}(\{R_j\})$. Let $a_{i,t,m}$ denote the m-th token in the t-th action in the i-th rollout, and $a_{i,t} = (a_{i,t,1}, \ldots, a_{i,t,L_{i,t}})$ be the tokenized action segment t in rollout i. For each action token, define the importance ratio $r_{i,t,m}(\theta) = \frac{\pi_\theta(a_{i,t,m}|\tau_{i,<t}, a_{i,t,<m})}{\pi_{\text{old}}(a_{i,t,m}|\tau_{i,<t}, a_{i,t,<m})}$, where $\tau_{i,<t} = (a_{i,1}, o_{i,1}, \cdots, a_{i,t-1}, o_{i,t-1})$ denotes the trajectory prefix of the i-th rollout before turn t, and $a_{i,t,<m}$ denotes the prefix of the t-th action tokens in the i-th rollout before m. Let us define the clipped token objective as

$$\ell^{\text{clip}}_{i,t,m}(\theta) = \min[r_{i,t,m}(\theta)A_i,\ \text{clip}(r_{i,t,m}(\theta), 1-\epsilon, 1+\epsilon)A_i].$$

The multi-turn GRPO surrogate averages this term over action tokens:

$$\mathcal{J}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{L_{i,t}}\sum_{m=1}^{L_{i,t}}\left(\ell^{\text{clip}}_{i,t,m}(\theta) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right) \tag{2}$$



(a) Optimization order **(b)** Retriever-update cost **(c)** Component ablation

Figure 2: **Optimization order, update cost, and component-wise ablations.** (a) Per-dataset EM gains over VT-Search, in percentage points, for the two orders of sequential retriever and LLM tuning. (b) Mean peak allocated memory and time per retriever η update, compared with the two-state bilevel estimator in (10); bar labels show raw values and the y-axis is normalized. (c) Component ablations with *Qwen2.5-3B-base* model. Full results appear in Table 3 in Appendix B.2.

where T_i is the number of turns in rollout i, and the trajectory-level advantage is broadcast to each turn and normalized by its length.

2.2 Tool-use RAG and RAG likelihood

In the agentic RL setting, an external search tool retrieves relevant information from a retrieval corpus and inserts the retrieved documents into the rollout within an `<information>` block, giving the agent access to domain-specific knowledge.

Let $\mathcal{C} = \{d_j\}_{j=1}^{N}$ be a retrieval corpus of chunked passages. At a retrieval turn t, let $h_t = (\tau_{<t}, a_t)$ denote the trajectory prefix up to turn t. We use a dense retriever with **relevance score** $s_\eta(h_t, d) = f_\eta(h_t)^\top g(d)$, where f_η is the trainable query encoder with parameter η, and g is a fixed document encoder. Given the prefix h_t, the retriever returns the set of m most relevant documents

$$D_t = \mathcal{N}_m(h_t) = \mathrm{TopM}_{d\in\mathcal{C}}\ s_\eta(h_t, d), \tag{3}$$

which is rendered into the information block as the environment observation o_t and provided to the LLM as additional context for subsequent generation. Through o_t, the retriever parameter η changes the subsequent policy context and therefore the trajectory distribution in (1).

For training η, $\mathcal{N}_k(h_t)$ denotes the top-k candidate set under the same relevance score, used by both the RAG likelihood and the RL outcome-credit surrogate. We choose $k > m$ so that the retriever receives document-level training signal from more candidates than the m documents.

RAG likelihood. For each training prompt x, sampled rollouts provide retrieval states h, paired with the reference answer $y = y^*(x)$. We compute the RAG negative log-likelihood at each retrieval turn of the sampled rollouts, averaged over retrieval contexts h and ground-truth answers y:

$$\mathcal{L}_{\mathrm{RAG}}(\theta, \eta) = -\mathbb{E}_{(h,y)} \log \sum_{d\in\mathcal{N}_k(h)} \rho_\eta(d \mid h) p_\theta(y \mid h, d) \tag{4}$$

where the retriever distribution is a softmax over the top-k set $\mathcal{N}_k(h_t)$:

$$\rho_\eta(d \mid h_t) = \frac{\exp(s_\eta(h_t, d))}{\sum_{d'\in\mathcal{N}_k(h_t)} \exp(s_\eta(h_t, d'))}. \tag{5}$$

The RAG objective is the negative log-likelihood of the correct answer under the current LLM θ, marginalized over the top-k documents under the current retriever η. By minimizing this objective with respect to η, the retriever is encouraged to rank more informative and useful documents higher, thereby improving the quality of the final answer generation.

3 Bilevel retrieval-augmented agentic RL

We first examine the order of retriever and policy optimization, then formulate the bilevel problem and develop an efficient algorithm for it.

3.1 BILEVEL FORMULATION

Hierarchical order in retrieval and LLM tuning. Although both the LLM and the retriever affect final answer accuracy, we first test whether retrieval tuning and policy optimization are order-sensitive. To this end, we compare two sequential tuning pipelines on top of the RL-trained VT-Search baseline: **RAG-then-RL**, which first improves the retriever and then optimizes the LLM via RL, and **RL-then-RAG**, which applies these two steps in the reverse order. As shown in Figure 2(a), adding RAG before RL improves performance over the RL-only VT-Search baseline on most datasets, with especially clear gains on multi-hop QA benchmarks. However, tuning the retriever after RL leads to worse performance: on general QA tasks, it can underperform the RL-only model, and overall it yields smaller gains than RAG-then-RL. This indicates that the order of retrieval and policy optimization matters: the retriever should adapt first to expose task-relevant evidence, and the LLM can then learn search and reasoning behavior under useful information blocks.

Motivation for dynamic tuning. Despite this sequential order, retrieval and LLM optimization can still reinforce each other, with feedback from one improving the other. We capture this with the bilevel formulation

$$\min_{\theta,\eta} \ \mathcal{J}_{\mathrm{RL}}(\theta,\eta) \qquad \text{s.t.} \quad \eta \in \eta^*(\theta) := \arg\min_{\eta'} \mathcal{L}_{\mathrm{RAG}}(\theta,\eta') \tag{6}$$

where the lower-level problem optimizes the RAG objective to find the best retriever η for the current policy θ, and the upper-level problem optimizes the RL objective to find the LLM policy θ that maximizes the outcome reward when paired with its optimal retriever. Problem (6) thus gives retrieval optimization priority at the lower level and optimizes the LLM against the resulting retriever, so the retriever keeps adapting as the policy changes.

3.2 MEMORY EFFICIENT BILEVEL ALGORITHM DESIGN VIA LANDSCAPE ANALYSIS

We now develop BRIDGE, a memory-efficient first-order method for solving (6).

First-order penalty-based approach. The main difficulty is the implicit mapping $\eta^*(\theta)$. Even in the simplest case, where the RAG objective has a unique minimizer with an invertible Hessian, the implicit function theorem (Dontchev & Rockafellar, 2014) gives

$$\nabla \mathcal{J}_{\mathrm{RL}}(\theta,\eta^*(\theta)) = \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta,\eta) - \nabla_{\theta\eta}\mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\nabla_{\eta\eta}^{-1}\mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\nabla_\eta \mathcal{J}_{\mathrm{RL}}(\theta,\eta)|_{\eta=\eta^*(\theta)} \tag{7}$$

which involves an inverse Hessian and a mixed Jacobian and is therefore infeasible to compute when both θ and η parameterize large neural networks.

Building on first-order penalty-based bilevel methods (Shen et al., 2025b; Kwon et al., 2024; Scellier & Bengio, 2017; Zucchet & Sacramento, 2022), we estimate this gradient through the penalty function

$$\mathcal{L}_\gamma(\theta,\eta) = \mathcal{J}_{\mathrm{RL}}(\theta,\eta) + \gamma \mathcal{L}_{\mathrm{RAG}}(\theta,\eta). \tag{8}$$

Let $\eta_\gamma^*(\theta) = \arg\min_\eta \mathcal{L}_\gamma(\theta,\eta)$ denote the penalized retriever. Since $\eta_\gamma^*(\theta) \to \eta^*(\theta)$ as $\gamma \to \infty$, the implicit gradient can be approximated by the first-order penalty gradient

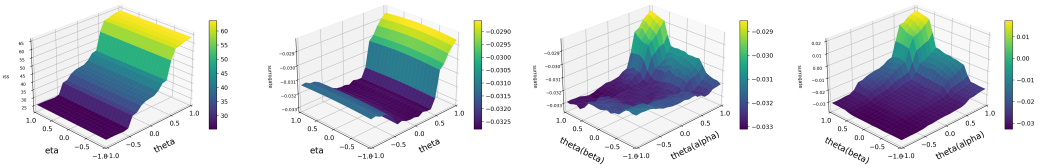
$$\begin{aligned}
\nabla \mathcal{J}_{\mathrm{RL}}(\theta,\eta^*(\theta)) &\overset{(a)}{\approx} \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta,\eta) + \gamma \nabla_{\theta\eta}\mathcal{L}_{\mathrm{RAG}}(\theta,\eta)(\eta_\gamma^*(\theta) - \eta^*(\theta))\,|_{\eta=\eta^*(\theta)} \\
&\overset{(b)}{\approx} \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta,\eta) + \gamma \nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\,|_{\eta=\eta_\gamma^*(\theta)} - \gamma \nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)|_{\eta=\eta^*(\theta)} \qquad (9) \\
&= \nabla_\theta \mathcal{L}_\gamma(\theta,\eta)|_{\eta=\eta_\gamma^*(\theta)} - \gamma \nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)|_{\eta=\eta^*(\theta)} \qquad (10)
\end{aligned}$$

where (a) holds because

$$\nabla_{\eta\eta}\mathcal{L}_{\mathrm{RAG}}(\theta,\eta)(\eta_\gamma^*(\theta) - \eta^*(\theta))|_{\eta=\eta^*(\theta)} \approx \nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)|_{\eta=\eta_\gamma^*(\theta)} \approx -\frac{1}{\gamma}\nabla_\eta \mathcal{J}_{\mathrm{RL}}(\theta,\eta)|_{\eta=\eta_\gamma^*(\theta)}$$

in which the first approximation follows from a Taylor expansion and $\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)|_{\eta=\eta^*(\theta)} = 0$, and the second from the optimality condition $\nabla_\eta \mathcal{L}_\gamma(\theta,\eta)|_{\eta=\eta_\gamma^*(\theta)} = 0$; (b) follows from a first-order Taylor expansion of $\nabla_\theta \mathcal{L}_{\mathrm{RAG}}$ in η. The error of the fully first-order estimator (10) decreases as $\mathcal{O}(1/\gamma)$ (Kwon et al., 2023).

However, the two-state estimator in (10) is still memory-intensive for RAG tool-use RL, because it needs to track two separate retriever states $\eta_\gamma^*(\theta)$ and $\eta^*(\theta)$, which are optimized using different



(a) Joint RAG **(b)** Joint RL **(c)** θ-only, initial **(d)** θ-only, final

Figure 3: **Loss landscapes of the RAG and RL objectives** for the search task with *Qwen2.5-3B-base*. We visualize the joint RAG landscape with respect to both θ and η in (a) and the joint RL landscape in (b). We also visualize the RL landscape along two random directions in θ (*alpha* and *beta*) in (c)–(d) at the initial and final checkpoints. As training proceeds, BRIDGE moves toward a smoother region of the RL objective landscape. More visualizations are in Appendix B.3.

objectives $\mathcal{L}_\gamma$ and $\mathcal{L}_{\mathrm{RAG}}$. This increases the memory cost by storing an additional set of retriever parameters, their optimizer states, and the corresponding retrieval logs; see Figure 2 (b).

Landscape analysis. To alleviate the memory bottleneck, existing single-state bilevel methods (Xiao et al., 2026; Shen et al., 2025a) avoid optimizing $\eta^*(\theta)$ by treating $\nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta, \eta^*(\theta)) \approx 0$ along the training trajectory. We visualize the RAG and RL loss landscapes following Li et al. (2018) in Figure 3 as a motivation. As shown in Figure 3(a), when the LLM policy θ is suboptimal, the RAG objective still changes substantially along the θ direction for most values of η, which indicates that $\nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta, \eta)|_{\eta=\eta^*(\theta)}$ cannot be ignored. However, Figure 3(b) shows that the upper-level RL objective is relatively flat along the retriever direction η across a wide range of θ. This suggests that, for a fixed policy θ, the penalized retriever solution $\eta^*_\gamma(\theta)$ remains close to the unpenalized lower-level solution $\eta^*(\theta)$, which motivates omitting the last two terms of (9) together: their difference is negligible if $\|\eta^*_\gamma(\theta) - \eta^*(\theta)\| = \mathrm{o}(1/\gamma)$ (Jiang et al., 2025a). The three resulting estimators, compared in Section 4.2 numerically, are (with $\eta^* = \eta^*(\theta)$, $\eta^*_\gamma = \eta^*_\gamma(\theta)$)

$$\underbrace{\nabla_\theta \mathcal{L}_\gamma(\theta, \eta^*_\gamma) - \gamma \nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta, \eta^*)}_{\text{two-state (10)}}, \qquad \underbrace{\nabla_\theta \mathcal{L}_\gamma(\theta, \eta^*_\gamma)}_{\text{Drop}}, \qquad \underbrace{\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta, \eta^*_\gamma)}_{\text{BRIDGE}},$$

where Drop removes only the value-function term (Xiao et al., 2026; Shen et al., 2025a) and BRIDGE removes both RAG terms, needing one retriever state.

Memory efficient algorithm design. Based on the estimator fidelity evidences in Section 4.2, we therefore track only the penalized retriever $\eta^*_\gamma(\theta)$ and update θ with the BRIDGE estimator. To further reduce memory, we parameterize the retriever with a query-side LoRA adapter (Hu et al., 2022) and update it periodically; see Figure 2(b) for a memory comparison. Given the fixed initial retriever η_0, the retriever used for search is $\eta_0 + AB$, where A and B are trainable low-rank matrices.

At each round r, BRIDGE uses the current penalized retriever $\eta_r = \eta_0 + A_r B_r$ and the current LLM θ_r to sample G tool-use rollouts per question. The query-side LoRA retriever is then updated by

$$\begin{aligned} A_{r+1} &= A_r - \alpha_A \left(\nabla_A \mathcal{J}_{\mathrm{RL}}(\theta_r, \eta_r) + \gamma \nabla_A \mathcal{L}_{\mathrm{RAG}}(\theta_r, \eta_r) \right), \\ B_{r+1} &= B_r - \alpha_B \left(\nabla_B \mathcal{J}_{\mathrm{RL}}(\theta_r, \eta_r) + \gamma \nabla_B \mathcal{L}_{\mathrm{RAG}}(\theta_r, \eta_r) \right), \end{aligned} \tag{11}$$

where the retriever-side RL gradient is approximated by a retrieval-side GRPO surrogate and the RAG gradient follows Lewis et al. (2020); both are derived in Appendix A.1. After the retriever update, the LLM is updated with the GRPO surrogate on the same rollouts:

$$\theta_{r+1} = \theta_r - \alpha \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_r, \eta_r). \tag{12}$$

Algorithm 1 summarizes BRIDGE, and Figure 1 gives an overview. Figure 2(b) further shows its memory benefits compared with two-state estimatior.

Interpretation. Intuitively, the RL signal provides global feedback on the success of the complete rollout, while the RAG signal distinguishes candidate documents by how strongly they support generating the correct answer under the current LLM. Combining these signals guides retriever adaptation through both final task outcomes and document-level answer likelihood. The LLM policy is then updated, guided solely by the upper-level RL objective. These two stages reflect the retrieval–policy hierarchy in our bilevel formulation. In Appendix A.2, we establish convergence of an idealized version of these updates, with exact gradients, to an approximate stationary point of (6).

Algorithm 1 BRIDGE

Input: Training prompts $\mathcal{D}_{\text{tr}}$, document corpus $\mathcal{C}$, policy π_θ, initial retrieval model η_0, initial LoRA retrieval adapter $A_0 = 0$ and random B_0, RL rollout group size G, penalty parameter γ, retriever update period P, number of rounds R.
for $r = 0, 1, \ldots, R-1$ **do**
  Sample a mini-batch of questions $x \sim \mathcal{D}_{\text{tr}}$.
  Generate G multi-turn tool-use rollouts with π_{θ_r} and retriever $\eta_r = \eta_0 + A_r B_r$.
  Compute answer rewards and GRPO advantages over LLM-generated action tokens in (2).
  Compute retrieval-turn credit and the RAG likelihood in (4)
  If $r \bmod P = 0$, update the LoRA adapter of η by (11)
  Update θ by (12)
end for
Output: Final policy π_{θ_R} and query-side retriever adapter η_R.

Table 1: **Open-domain QA evaluation results.** Best/second-best scores are highlighted in **bold**/underlined. †/* denote in-domain/out-of-domain datasets relative to our training set. ‡ indicates results from the original papers. Search-R1 and BRIDGE rows use the higher-Avg. *Base*/*Instruct* variant (Appendix B.5). IRCoT uses the corresponding *Instruct* models. M-Avg. and Avg. denote multi-hop and overall averages. Darker green shading indicates higher accuracy.

Method	General QA			Multi-hop QA (knowledge-intensive)				M-Avg.	Avg.
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamboogle*		
Large-model references: direct inference without retrieval									
GPT-4o	22.3	66.5	36.9	22.8	19.8	6.2	26.4	18.8	28.7
GPT-5.6 Luna	19.0	65.2	33.4	22.2	19.9	7.1	26.4	18.9	27.6
Qwen3-235B-A22B	23.0	58.4	28.9	23.2	29.1	4.9	20.8	19.5	26.9
Qwen2.5-3B									
Direct Inference	18.2	30.4	20.2	10.9	12.0	3.5	8.8	8.8	14.9
IRCoT (Instruct)‡	11.1	31.2	20.0	16.4	17.1	6.7	24.0	16.1	18.1
Search-o1‡	23.8	47.2	26.2	22.1	21.8	5.4	32.0	20.3	25.5
RAG‡	34.8	54.4	38.7	25.5	22.6	4.7	8.0	15.2	27.0
SFT‡	24.9	29.2	10.4	18.6	24.8	4.4	11.2	14.8	17.6
R1-base‡	22.6	45.5	17.3	20.1	26.8	5.5	22.4	18.7	22.9
Rejection Sampling‡	29.4	48.8	33.2	24.0	23.3	5.9	21.0	18.6	26.5
s3	32.2	54.4	38.9	24.9	25.2	5.7	8.7	16.1	27.1
Search-R1‡	39.7	56.5	39.1	33.1	31.0	12.4	23.2	24.9	33.6
Agentic-R	43.1	60.0	42.3	31.0	27.2	8.4	11.2	19.5	31.9
VT-Search‡	45.4	**61.6**	**48.1**	32.4	30.8	7.6	15.2	21.5	34.4
BRIDGE (ours)	**46.3**	**61.6**	44.7	**42.3**	**41.4**	**16.8**	**37.6**	**34.5**	**41.5**
Qwen2.5-7B									
Direct Inference‡	13.4	40.8	14.0	18.3	25.0	3.1	12.0	14.6	18.1
IRCoT (Instruct)‡	22.4	47.8	30.1	13.3	14.9	7.2	22.4	14.5	22.6
Search-o1‡	15.1	44.3	13.1	18.7	17.6	5.8	29.6	17.9	20.6
RAG‡	34.9	58.5	39.2	29.9	23.5	5.8	20.8	20.0	30.4
SFT‡	31.8	35.4	12.1	21.7	25.9	6.6	11.2	16.4	20.7
R1-base‡	29.7	53.9	20.2	24.2	27.3	8.3	29.6	22.4	27.6
Rejection Sampling‡	36.0	59.2	38.0	33.1	29.6	12.3	35.5	27.6	34.8
s3	36.0	57.4	43.0	28.9	25.1	5.6	12.0	17.9	29.7
Search-R1‡	48.0	63.8	45.7	43.3	38.2	19.6	43.2	36.1	43.1
Agentic-R‡	42.4	**69.0**	44.1	45.8	45.3	20.3	48.0	39.8	45.0
VT-Search‡	49.3	66.2	**50.2**	44.8	45.3	19.3	46.4	39.0	45.9
BRIDGE (ours)	**50.1**	66.3	48.4	**48.9**	**49.9**	**22.7**	**51.2**	**43.2**	**48.2**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and benchmarks. We use *Qwen2.5-3B-base* model (Yang et al., 2024) and an *E5* retriever (Wang et al., 2022) unless stated otherwise. We evaluate on twelve benchmarks spanning three task families: 1) *general open-domain QA*, including NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023); 2) *open-domain multi-hop QA*, including HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023); 3) *scientific medical QA*, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), BioASQ (Tsatsaronis et al., 2015; Krithara

et al., 2023), and MMLU-Medical (Hendrycks et al., 2021). For open-domain RL training, we follow Jin et al. (2025) and train on NQ and HotpotQA, using the *2018 Wikipedia dump* (Karpukhin et al., 2020) for retrieval. NQ and HotpotQA are in-domain evaluations; the other five open-domain benchmarks assess out-of-domain generalization. Training details for the medical QA experiments are given in Appendix C.1.

Evaluation metrics. Following Jin et al. (2025), we use exact match (EM) for open-domain rewards and evaluation. M-Avg./Avg. are unweighted means over the four multi-hop/all seven benchmarks. Medical QA also includes *GPT-5.6 Luna*-judged reasoning quality (Section 4.3).

4.2 Open-domain QA tasks

Baselines. The open-domain baselines in Table 1 fall into three classes: 1) *inference without retrieval*: Direct Inference, including *GPT-4o*, *GPT-5.6 Luna*, and *Qwen3-235B-A22B*; 2) *inference with retrieval*, including RAG (Lewis et al., 2020), IRCoT (Trivedi et al., 2023), and Search-o1 (Li et al., 2025); 3) *fine-tuning-based methods*, including *retrieval-side tuning* (s3, with a frozen answer generator (Jiang et al., 2025b)), *LLM-policy tuning* (SFT (Chung et al., 2024), R1-base (Guo et al., 2025), Rejection Sampling (Ahn et al., 2024), Search-R1 (Jin et al., 2025), and VT-Search (Jiang et al., 2026)), and *joint retriever–policy tuning* (Agentic-R (Liu et al., 2026)).

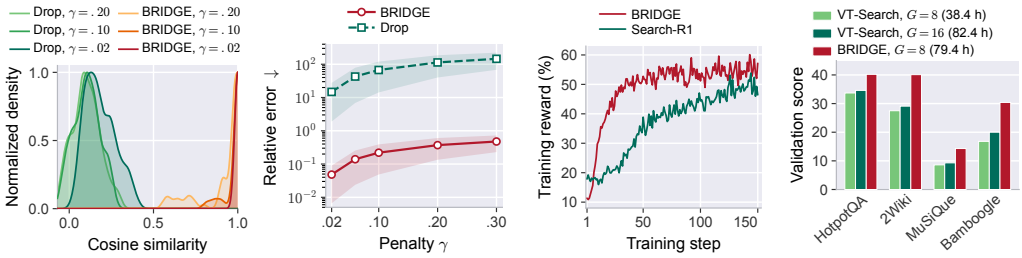
Key finding: BRIDGE has the largest gains on multi-hop QA and with smaller backbones.

In Table 1, BRIDGE achieves the highest overall average at both scales, with the largest margins on multi-hop QA, where answers require integrating evidence across multiple documents; it trails the best baseline only on PopQA (both scales) and TriviaQA (7B). Its gains are larger with the 3B backbone, and it surpasses all three large-model references on six of seven benchmarks, suggesting that better retriever–LLM coordination helps a smaller LLM on complex tasks.

Component-wise ablations. We test three internal baselines: **RAG-then-RL** sequentially tunes retrieval with RAG, then the policy with RL; **Decoupled-RL** trains the retriever and the policy online, both with RL; **Decoupled-RAG** trains them online, the retriever with RAG likelihood only and the policy with RL. All variants use *Qwen2.5-3B-base* and a GRPO group size of $G = 8$.

Key finding: BRIDGE gains the most by combining online adaptation with a hierarchical order.

As shown in Figure 2(c), all ablation variants outperform the RL-only baselines Search-R1 (*Base*) and VT-Search, showing the benefit of retriever optimization. However, neither online variant surpasses the sequential RAG-then-RL, which suggests that the hierarchical order matters. BRIDGE preserves this hierarchy through its bilevel formulation and improves on RAG-then-RL by adapting the retriever and the LLM online.



(a) Gradient direction **(b)** Relative gradient error **(c)** Training reward **(d)** Compute allocation

Figure 4: **Estimator fidelity, training dynamics, and compute allocation.** (a) Cosine similarity to the two-state estimator (unit-peak densities). (b) Mean relative gradient error on a log scale. (c) Raw training rewards for BRIDGE and Search-R1 (*Qwen2.5-7B-base*). (d) Multi-hop QA accuracy with *Qwen2.5-3B-base*; the legend gives training hours on 4×A100 GPUs.

Estimator fidelity and retriever-update cost. Using the two-state estimator (10) as the reference, Figure 4(a–b) compares the BRIDGE and Drop estimators defined in Section 3. BRIDGE achieves higher cosine similarity and lower relative gradient error across all tested penalty coefficients γ. Figure 2(b) further shows that BRIDGE reduces mean peak allocated memory by 49.7% and time per retriever update by 55.8% relative to the two-state estimator.

Training dynamics. Figure 4(c) shows faster reward improvement for BRIDGE than Search-R1. Appendix D.3 presents a case study of how BRIDGE's search behavior changes during training: early in training it answers prematurely from incomplete evidence, whereas at a later checkpoint it uses the result of the first retrieval to formulate a targeted follow-up query, retrieves the missing evidence, and answers correctly.

Compute allocation. Figure 4(d) compares BRIDGE with VT-Search using twice as many GRPO rollouts. At comparable training time, BRIDGE achieves higher accuracy on all four multi-hop datasets. Appendix B.4 also compares BRIDGE with the joint-training baseline Agentic-R: BRIDGE is more accurate and takes roughly half the training time.

4.3 Evaluation on Medical QA Benchmarks

Setup. To test beyond open-domain QA, we evaluate on five MIRAGE medical QA benchmarks (Xiong et al., 2024) with *Qwen2.5-7B-base* (Yang et al., 2024) and *E5* retriever (Wang et al., 2022), retrieving from a combined Wikipedia, PubMed, and textbook corpus following s3 (Jiang et al., 2025b). Trained methods use only the NQ and HotpotQA training split (Appendix C.1).

Baselines. Table 2 compares BRIDGE with 1) *inference without retrieval* (Direct Inference), 2) *inference with retrieval* (RAG (Lewis et al., 2020), IRCoT (Trivedi et al., 2023), and MedRAG (Xiong et al., 2024)), and 3) *fine-tuning-based methods*: *retrieval-only tuning* (s3, which trains a search agent without LLM tuning (Jiang et al., 2025b)) and *LLM-policy tuning* (VT-Search, with a fixed retriever (Jiang et al., 2026)). MedRAG uses *MedCPT* (Jin et al., 2023) as its retriever.

Table 2: **Medical QA evaluations with *Qwen2.5-7B-base*.** Each cell reports reasoning quality judged by *GPT-5.6 Luna* (Appendix C), with answer accuracy in parentheses. Macro averages weight benchmarks equally; micro averages weight questions equally. Darker shading marks higher reasoning scores. Best/second-best values are **bold**/underlined.

Method	MedQA	MedMCQA	PubMedQA	BioASQ	MMLU-Medical	Macro Avg.	Micro Avg.
Direct Inference	4.08 (0.71)	16.64 (9.75)	0.25 (26.20)	4.94 (28.16)	11.03 (12.95)	7.39 (15.55)	11.74 (11.26)
IRCoT (Instruct)	39.47 (43.44)	41.35 (51.95)	40.34 (58.20)	39.38 (63.11)	42.81 (60.79)	40.67 (55.50)	41.02 (53.10)
RAG	42.95 (50.82)	39.64 (53.96)	51.35 (66.60)	44.41 (57.93)	44.88 (66.12)	44.65 (59.09)	42.09 (56.31)
s3	47.31 (54.36)	39.50 (56.20)	41.20 (66.60)	36.65 (64.72)	45.86 (67.03)	42.10 (61.78)	41.58 (58.80)
MedRAG	60.91 (57.74)	56.85 (54.55)	73.10 (**69.80**)	82.16 (85.92)	72.93 (75.02)	69.19 (68.61)	62.91 (61.52)
VT-Search	60.82 (58.92)	58.42 (55.06)	71.55 (68.80)	84.55 (**88.51**)	74.49 (74.38)	69.97 (69.13)	64.07 (62.04)
BRIDGE (ours)	**63.31** (**64.18**)	**60.61** (**57.71**)	**74.50** (68.80)	**86.65** (87.54)	**76.26** (**75.39**)	**72.27** (**70.72**)	**66.29** (**64.43**)

Key finding: BRIDGE improves medical QA reasoning quality.

BRIDGE has the highest reasoning quality on all five benchmarks and improves both macro-averaged reasoning quality and answer accuracy over VT-Search. The case studies in Appendix D.4 further illustrate how BRIDGE generates search queries that more directly target the medical information needed to answer the question and uses retrieved evidence more effectively than VT-Search.

5 Conclusion

We showed that retriever and LLM optimization in retrieval-augmented agentic RL are order-sensitive: adapting the retriever before optimizing the policy yields larger gains than the reverse order. To enable joint online adaptation, we formulated retrieval-augmented ARL as a bilevel problem that preserves this hierarchy and introduced BRIDGE, a memory-efficient first-order method that adapts the retriever with outcome rewards and RAG likelihood and the policy with RL. Across seven open-domain and five medical QA benchmarks, BRIDGE improves answer accuracy and medical reasoning quality, with the largest gains on multi-hop QA and with smaller backbones.

AI USE STATEMENT

We used AI assistants for polishing the manuscript, feedback on experimental design, and interpretation of results. We also used *GPT-5.6 Luna* as an LLM judge of medical QA reasoning quality, with its inputs, system prompt, output schema, and scoring procedure documented in Appendix C. These automated judgments supplement the exact-match answer evaluation. The authors take responsibility for the paper's final content, including AI-assisted text, code, analyses, and claims.

ETHICS STATEMENT

This work advances retrieval-augmented agentic reinforcement learning through joint adaptation of language models and retrievers. All experiments are conducted using publicly available datasets. Potential benefits include improved information quality and efficient computation for knowledge-intensive question answering. However, the method may inherit biases from training data and retrieval corpora and may propagate inaccurate or outdated information. Gains on open-source benchmarks do not eliminate these risks, and evaluations remain limited by the model-judge quality. In particular, our medical QA evaluations are not evidence of clinical safety or suitability for patient care; deployment in medical or other high-stakes settings would require independent validation, appropriate data safeguards, and specialized domain knowledge.

REPRODUCIBILITY STATEMENT

We present the bilevel formulation and update procedure of the proposed method BRIDGE in Section 3 and Algorithm 1. We describe the models, datasets, training corpora, and evaluation metrics in Section 4.1, and detail the medical QA setting in Section 4.3. We provide additional loss landscapes, runtime comparisons, and backbone evaluations in Appendix B. We also document the complete medical judging protocol in Appendix C, including preprocessing, prompts, the output schema, and aggregation and exclusion rules. We also include qualitative reasoning trajectories in Appendix D to illustrate retrieval and reasoning behavior of the proposed method and baseline.

REFERENCES

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, St. Julian's, Malta, 2024.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, virtual, 2021.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. Improving retrieval-augmented generation through multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, San Diego, CA, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Asen L Dontchev and R Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, second edition, 2014.

Jingsheng Gao, Linxu Li, Ke Ji, Weiyuan Li, Yixin Lian, Yuzhuo Fu, and Bin Dai. SmartRAG: Jointly learn RAG-related tasks from the environment feedback. In *Proceedings of the International Conference on Learning Representations*, Singapore, Singapore, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*, virtual, 2021.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the International Conference on Computational Linguistics*, virtual, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, virtual, 2022.

Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, and Tong Zhang. Tackling distractor documents in multi-hop QA with reinforcement and curriculum learning. In *Findings of the Association for Computational Linguistics: EACL*, Rabat, Morocco, 2026.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the International Conference on Machine Learning*, virtual, 2021.

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhu Chen. VerlTool: Towards holistic agentic reinforcement learning with tool use. *Transactions on Machine Learning Research*, 2026.

Liuyuan Jiang, Quan Xiao, Lisha Chen, and Tianyi Chen. Beyond value functions: Single-loop bilevel optimization under flatness conditions. In *Advances in Neural Information Processing Systems*, San Diego, CA, 2025a.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don't need that much data to train a search agent via RL. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, 2025b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Proceedings of the Conference on Language Modeling*, Montreal, Canada, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, virtual, 2020.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the International Conference on Machine Learning*, Honolulu, HI, 2023.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, virtual, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, Montreal, Canada, 2018.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, 2025.

Yu Li, Mingyang Yi, Xiuyu Li, Ju Fan, Fuxin Jiang, Binbin Chen, Peng Li, Jie Song, and Tieying Zhang. Reasoning and tool-use compete in agentic RL: From quantifying interference to disentangled tuning. *arXiv preprint arXiv:2602.00994*, 2026a.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. In *Proceedings of the International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026b.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. RA-DIT: Retrieval-augmented dual instruction tuning. In *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 2024.

Wenhan Liu, Xinyu Ma, Yutao Zhu, Yuchen Li, Daiting Shi, Dawei Yin, and Zhicheng Dou. Agentic-R: Learning to retrieve for agentic search. In *Findings of the Association for Computational Linguistics: ACL*, San Diego, CA, 2026.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, virtual, 2020.

Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen Meng, and James Glass. Search augmented instruction learning. In *Findings of the Association for Computational Linguistics: EMNLP*, Singapore, Singapore, 2023.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, virtual, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, Singapore, Singapore, 2023.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24, 2017.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. SEAL: Safety-enhanced aligned LLM fine-tuning via bilevel data selection. In *Proceedings of the International Conference on Learning Representations*, Singapore,Singapore, 2025a.

Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 214:539–589, 2025b.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico, 2024.

Zhengliang Shi, Lingyong Yan, Weiwei Sun, Yue Feng, Pengjie Ren, Xinyu Ma, Shuaiqiang Wang, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. Direct retrieval-augmented optimization: Synergizing knowledge selection and language models. *ACM Transactions on Information Systems*, 44(4):94:1–94:30, 2026.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. DynamicRAG: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, San Diego, CA, 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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Peiyao Xiao, Chaosheng Dong, Shaofeng Zou, and Kaiyi Ji. LDC-MTL: Balancing multi-task learning through scalable loss discrepancy control. In *Proceedings of the European Conference on Computer Vision*, Malmö, Sweden, 2026.

Jian Xie, Tianhe Lin, Zilu Wang, Yuting Ning, Yuekun Yao, Tianci Xue, Zhehao Zhang, Zhongyang Li, Kai Zhang, Yufan Wu, et al. QUEST: Training frontier deep research agents with fully synthetic tasks. *arXiv preprint arXiv:2605.24218*, 2026.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL*, Bangkok, Thailand, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG. In *Proceedings of the Conference on Language Modeling*, Philadelphia, PA, 2024.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digital Medicine*, 9(1):34, 2025.

Nicolas Zucchet and João Sacramento. Beyond backpropagation: bilevel optimization through implicit differentiation and equilibrium propagation. *Neural Computation*, 34(12):2309–2346, 2022.

APPENDIX CONTENTS

A Retrieval Gradients and Theoretical Guarantee **16**

A.1 Derivation of retrieval gradient 16

A.2 Convergence guarantee of BRIDGE 18

B Additional Analyses and Results on Open-domain QA **21**

B.1 Experimental Details 21

B.2 Component-wise Ablations 21

B.3 Full Loss Landscapes 22

B.4 Runtime–Accuracy Comparison 22

B.5 Performance across Backbones 23

C Medical QA Reasoning-Quality Evaluation **24**

C.1 Training details 24

C.2 Evaluation protocol 25

D Qualitative Trajectory Case Studies **25**

D.1 RAG provides useful learning signals under zero reward 25

D.2 Open-domain QA Case Studies 26

D.3 Retrieval Behavior across Training Checkpoints 33

D.4 Medical QA Case Studies 35

A RETRIEVAL GRADIENTS AND THEORETICAL GUARANTEE

In this section, we derive the retrieval RL gradient and state the convergence guarantee for BRIDGE.

A.1 DERIVATION OF RETRIEVAL GRADIENT

Gradient of RAG objective with respect to the retrieval parameters. At the retrieval turn t, with rollout τ_i for prompt x, the query context is $h_{i,t} = (\tau_{i,<t}, a_{i,t})$ and we can access the reference answer y corresponding to question x. Within a batch, RAG objective in (4) can be written as

$$\mathcal{L}_{\mathrm{RAG}}(\theta, \eta) = -\mathbb{E}_{x,i,t}\left[\log \sum_{d \in \mathcal{N}_k(h_{i,t})} \rho_\eta(d \mid h_{i,t})\, p_\theta(y \mid h_{i,t}, d)\right].$$

Here, $\mathbb{E}_{x,i,t}$ averages over training prompts in a batch, sampled rollouts, and their logged retrieval turns. We retain (h, y) as shorthand below, with $\mathbb{E}_{(h,y)}$ denoting this same averaging.

Following (Lewis et al., 2020), we treat $\mathcal{N}_k(h_{i,t})$ fixed and differentiate the RAG objective in (4) as

$$\begin{aligned}
\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta, \eta) &= -\mathbb{E}_{(h,y)}\left[\sum_{d \in \mathcal{N}_k(h)} \frac{\rho_\eta(d \mid h)\, p_\theta(y \mid h, d)}{\sum_{d' \in \mathcal{N}_k(h)} \rho_\eta(d' \mid h)\, p_\theta(y \mid h, d')} \nabla_\eta \log \rho_\eta(d \mid h)\right] \\
&= \mathbb{E}_{(h,y)}\left[\nabla_\eta f_\eta(h)^\top \sum_{d \in \mathcal{N}_k(h)} \underbrace{\rho_\eta(d \mid h)\left(1 - \frac{p_\theta(y \mid h, d)}{\sum_{d' \in \mathcal{N}_k(h)} \rho_\eta(d' \mid h)\, p_\theta(y \mid h, d')}\right)}_{\text{RAG score-gradient coefficient}} g(d)\right]. \qquad (13)
\end{aligned}$$

Intuitively, the RAG score-gradient coefficient compares a document's current retrieval probability $\rho_\eta(d \mid h)$ with the reweighted document score by the final answer correctness. Gradient descent with a negative RAG score-gradient coefficient adds a positive multiple of $g(d)$ to the query-embedding descent direction $\nabla f_\eta(h)$, promoting stronger alignment between the query and the document and thereby providing a signal to improve the document's ranking within the top-k retrieved set.

Likelihood computation. During training, the reference answer y is available. For each retrieved document $d \in \mathcal{N}_k(h)$, we compute

$$\log p_\theta(y \mid h, d) = \sum_{j=1}^{|y|} \log p_\theta(y_j \mid h, d, y_{<j})$$

using teacher forcing through forward computation, summing only over answer tokens. These log-likelihoods can be reused across inner steps while θ, h, d, and y remain unchanged, whereas $\rho_\eta(d \mid h)$ is recomputed after each retriever update. For fixed sequence lengths, LLM scoring costs scale linearly with k, but is independent with respect to the corpus size $|\mathcal{C}|$. We also use document batch processing to reduce the peak scoring memory required for this step.

Numerical stability. We evaluate the RAG objective in log space:

$$\mathcal{L}_{\mathrm{RAG}}(\theta, \eta) = -\mathbb{E}_{(h,y)} \left[\mathrm{logsumexp}_{d \in \mathcal{N}_k(h)} \left(\log \rho_\eta(d \mid h) + \log p_\theta(y \mid h, d) \right) \right].$$

Here, $\log \rho_\eta$ is computed by log-softmax over $\mathcal{N}_k(h)$. Log-softmax and log-sum-exp avoid explicit products of small probabilities or division by an underflowed marginal likelihood, which ensures its numerical stability. The answer-conditioned document weights in (13) are obtained by applying softmax to $\log \rho_\eta(d \mid h) + \log p_\theta(y \mid h, d)$. Each score-gradient coefficient is therefore a difference between two normalized probabilities.

RL outcome credit for retrieval. To motivate the GRPO like RL surrogate gradient with respect to retrieval parameter η, fix θ and consider a stochastic retrieval relaxation in which $d_t \sim \rho_\eta(\cdot \mid h_t)$ over fixed top-k candidate set $\mathcal{N}_k(h_t)$. For the outcome reward, the log-derivative identity gives

$$\begin{aligned}
-\nabla_\eta \mathbb{E}_{x,\tau}[R(x,\tau)] &= -\mathbb{E}_x \int \nabla_\eta P(\tau \mid \theta, \eta) R(x,\tau)\, d\tau \\
&= -\mathbb{E}_x \int P(\tau \mid \theta, \eta) \nabla_\eta \log P(\tau \mid \theta, \eta) R(x,\tau)\, d\tau \\
&= -\mathbb{E}_{x,\tau} \left[R(x,\tau) \nabla_\eta \log P(\tau \mid \theta, \eta) \right] \\
&= -\mathbb{E}_{x,\tau} \left[R(x,\tau) \sum_t \nabla_\eta \log \rho_\eta(d_t \mid h_t) \right], \qquad (14)
\end{aligned}$$

where the expectation is under this stochastic relaxation and the sum is over retrieval turns t. The last equality follows because, under the stochastic retrieval relaxation, η enters the trajectory probability only through the factors $\rho_\eta(d_t \mid h_t)$ at retrieval turns t. With θ fixed and the recorded trajectory held fixed, differentiation therefore gives $\nabla_\eta \log P(\tau \mid \theta, \eta) = \sum_t \nabla_\eta \log \rho_\eta(d_t \mid h_t)$.

BRIDGE uses deterministic top-k retrieval during rollouts. Motivated by the expression above, we construct a retrieval-credit surrogate on the logged candidate sets, using $\bar{c}_{i,t,d}$ to distribute trajectory-level outcome credit among candidate passages. We then replace R_i with the GRPO advantage A_i and apply probability-ratio clipping.

Each retrieval turn t of rollout τ_i has the trajectory-level GRPO advantage A_i in (2). We distribute this outcome credit over the retrieved documents using answer-likelihood-based credit weight assigned to document and the document-level importance ratio as

$$\bar{c}_{i,t,d} = \frac{\rho_{\eta_{\mathrm{old}}}(d \mid h_{i,t})\, p_\theta(y \mid h_{i,t}, d)}{\sum_{d' \in \mathcal{N}_k(h_{i,t})} \rho_{\eta_{\mathrm{old}}}(d' \mid h_{i,t})\, p_\theta(y \mid h_{i,t}, d')}, \quad r_{i,t,d}(\eta) = \frac{\rho_\eta(d \mid h_{i,t})}{\rho_{\eta_{\mathrm{old}}}(d \mid h_{i,t})} \qquad (15)$$

where η_{old} denotes the pre-update retrieval parameter. The document-level importance ratio extends the token-level ratio in (2), with document probabilities replacing token probabilities. Fixing θ and following (Schulman et al., 2017), the retrieval RL surrogate used the following clipped advantage

$$\mathcal{J}^{\eta}_{\mathrm{RL}}(\theta, \eta) = -\mathbb{E}_{x,i,t} \left[\sum_{d \in \mathcal{N}_k(h_{i,t})} \bar{c}_{i,t,d} \min\{ r_{i,t,d}(\eta) A_i, \mathrm{clip}(r_{i,t,d}(\eta), 1-\epsilon_\eta, 1+\epsilon_\eta)\, A_i \} \right], \quad (16)$$

where $\mathbb{E}_{x,i,t}$ denotes the empirical average over prompts, rollouts, and retrieval turns, and ϵ_η is the retrieval clipping threshold. Away from clipping boundaries, differentiation gives

$$\nabla_\eta \mathcal{J}_{\mathrm{RL}}^{\theta,\eta}(\eta) = -\mathbb{E}_{x,i,t}\left[\sum_{d\in\mathcal{N}_k(h_{i,t})} \bar{c}_{i,t,d}\,\kappa_{i,t,d} A_i r_{i,t,d}(\eta)\,\nabla_\eta \log \rho_\eta(d \mid h_{i,t})\right], \tag{17}$$

where $\kappa_{i,t,d} = 1$ if $A_i \geq 0$ and $r_{i,t,d}(\eta) < 1+\epsilon_\eta$, or if $A_i < 0$ and $r_{i,t,d}(\eta) > 1-\epsilon_\eta$; otherwise it is zero. Thus, A_i determines the sign of the outcome credit, while $\bar{c}_{i,t,d}$ allocates it according to each document's support for the correct answer.

Combining RL and RAG gradients. The clipped RL surrogate in (17) uses the fixed weights $\bar{c}_{i,t,d}$ to distribute trajectory-level outcome credit. In contrast, the RAG gradient in (13) differentiates the current log-mixture likelihood and uses weights based on ρ_η. The retriever update combines the RL gradient in (17) with γ times this RAG gradient.

LoRA gradient form. For each adapted query-encoder weight matrix, write $\eta = \eta_0 + AB$ with η_0 frozen. Holding θ fixed, the chain rule gives

$$\begin{aligned}\nabla_A \mathcal{L}_\gamma(\theta,\eta) &= \big[\nabla_\eta \mathcal{J}_{\mathrm{RL}}(\theta,\eta) + \gamma\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\big]B^\top,\\ \nabla_B \mathcal{L}_\gamma(\theta,\eta) &= A^\top\big[\nabla_\eta \mathcal{J}_{\mathrm{RL}}(\theta,\eta) + \gamma\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\big].\end{aligned} \tag{18}$$

In implementation, the RL gradient is approximated by (17), and the RAG gradient is given by (13). Applying these expressions layerwise yields the factor updates in (11), while the base query-encoder weights and document encoder remain frozen.

A.2 Convergence guarantee of BRIDGE

Following (Kwon et al., 2023; Shen et al., 2025b), define

$$\mathcal{L}^*_{\mathrm{RAG}}(\theta) = \min_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta), \qquad \mathcal{L}^*_\gamma(\theta) = \min_\eta \mathcal{L}_\gamma(\theta,\eta) - \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta). \tag{19}$$

We analyze direct gradient updates of η with exact population gradients. The following assumptions follow the smoothness, lower-level landscape, and flatness conditions of Jiang et al. (2025a).

Assumption 1 (Lipschitz continuity)**.** *$\mathcal{J}_{\mathrm{RL}}$ and $\mathcal{L}_{\mathrm{RAG}}$ are jointly L_{RL}- and L_{RAG}-smooth, and have Lipschitz-continuous Hessians. The minima in* (19) *are attained, and* $\inf_\theta \mathcal{L}^*_\gamma(\theta) > -\infty$.

Assumption 2 (Lower-level landscape)**.** *There exist $\mu, \gamma_0 > 0$ s.t. $\mathcal{L}_{\mathrm{RAG}}(\theta,\cdot)$ and $\gamma^{-1}\mathcal{L}_\gamma(\theta,\cdot)$ satisfy the μ-Polyak–Łojasiewicz (PL) condition uniformly over θ and $\gamma \geq \gamma_0$, i.e., for all (θ,η),*

$$\begin{aligned}\tfrac{1}{2}\|\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\|^2 &\geq \mu\big[\mathcal{L}_{\mathrm{RAG}}(\theta,\eta) - \min_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)\big],\\ \tfrac{1}{2}\|\nabla_\eta \mathcal{L}_\gamma(\theta,\eta)\|^2 &\geq \gamma\mu\big[\mathcal{L}_\gamma(\theta,\eta) - \min_\eta \mathcal{L}_\gamma(\theta,\eta)\big].\end{aligned} \tag{20}$$

Assumption 3 (Upper-level flatness)**.** *There exist $p \in (1, 3/2]$ and $C \geq 0$, independent of γ and R, such that for every θ_r there is $\delta_r \geq 0$ satisfying*

$$|\mathcal{J}_{\mathrm{RL}}(\theta_r,\eta) - \mathcal{J}_{\mathrm{RL}}(\theta_r,\eta^*)| \leq C\|\eta-\eta^*\|^p + \delta_r, \qquad \frac{1}{R}\sum_{r=0}^{R-1}\delta_r \leq \bar{\delta} \tag{21}$$

for all η and all $\eta^* \in \arg\min_\eta \mathcal{L}_{\mathrm{RAG}}(\theta_r,\eta)$.

Corollary 1 (Approximate stationary point)**.** *Under Assumptions 1–2, suppose the flatness condition holds at θ with $\delta = O(\epsilon^{p/2})$, where $0 < \epsilon \leq 1$. There exists $\gamma_\epsilon = O(\epsilon^{-(2-p)/2})$ such that, for $\gamma \geq \max\{\gamma_0, \gamma_\epsilon\}$, any pair (θ,η_γ) satisfying $\|\nabla\mathcal{L}^*_\gamma(\theta)\| \leq \epsilon$ and $\|\nabla_\eta \mathcal{L}_\gamma(\theta,\eta_\gamma)\| \leq \epsilon$ is $O(\epsilon)$-stationary point to the original bilevel problem* (6) *in the squared KKT-residual sense.*

This extends (Jiang et al., 2025a, Lemma 3) to an approximate penalized minimizer setting, which directly links the approximate stationary point for the penalized problem to the approximate stationary point for the original problem (6). We defer its proof after Theorem 1.

The following theorem extends (Jiang et al., 2025a, Theorem 3) to periodic retriever updates under exact population gradients and direct gradient descent updates of the retriever parameters. This provides an idealized analysis of the central alternating update structure in BRIDGE. Together with the

approximate stationary point result in Corollary 1, it explains how BRIDGE can achieve approximate stationarity for the original bilevel problem, without explicitly computing hypergradients. Extending the analysis to the LoRA parameterization and empirical surrogate gradients used in our implementation is beyond the scope of this work.

Theorem 1 (Convergence of BRIDGE with periodic retriever updates)**.** *Under Assumptions 1–3, fix $\gamma \geq \gamma_0$ and $T \geq 1$. Periodicly updating the retriever η at rounds $r \in \{0, T, 2T, \ldots\}$ and the policy θ at every round gives*

$$\eta_{r+1} = \eta_r - \frac{b}{\gamma}\mathbf{1}_{\{r \bmod T=0\}}\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r), \qquad \theta_{r+1} = \theta_r - a\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_r, \eta_r). \tag{22}$$

If we choose

$$b = \left(L_{\mathrm{RAG}} + \frac{L_{\mathrm{RL}}}{\gamma}\right)^{-1}, \qquad a = \frac{b\mu^2}{64\gamma T(L_{\mathrm{RAG}} + \mu)^2} = \mathcal{O}(T^{-1}) \tag{23}$$

then for every $R \geq 1$, we have

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\left[\|\nabla \mathcal{L}_\gamma^*(\theta_r)\|^2 + \|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2\right] \\ \leq O\left(\frac{\mathcal{L}_\gamma(\theta_0, \eta_0) - \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta_0) - \inf_\theta \mathcal{L}_\gamma^*(\theta)}{aR} + \gamma^{-\frac{2(p-1)}{2-p}} + \gamma\bar{\delta}\right).\end{aligned} \tag{24}$$

Moreover, the lower-level objective gap satisfies

$$\frac{1}{R}\sum_{r=0}^{R-1}\left[\mathcal{L}_\gamma(\theta_r, \eta_r) - \min_\eta \mathcal{L}_\gamma(\theta_r, \eta)\right] \leq \frac{1}{2\gamma\mu R}\sum_{r=0}^{R-1}\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2. \tag{25}$$

The hidden constants depend on the regularity and flatness constants and γ_0, but not on R, T, or γ.

For fixed γ and initialization, the optimization term in (24) is $\mathcal{O}(T/R)$, and periodic retrieval updates do not introduce an additional residual term. To balance the last two residual terms, we can choose $\gamma \asymp \bar{\delta}^{-(2-p)/p}$ with $\gamma \geq \gamma_0$. For $\bar{\delta} \asymp \epsilon^{p/2}$ and $0 < \epsilon \leq 1$, this also satisfies the penalty-size condition $\gamma \geq \max\{\gamma_0, \gamma_\epsilon\}$, where $\gamma_\epsilon = O(\epsilon^{-(2-p)/2})$, in Corollary 1. Then, with $R \gtrsim T/\epsilon$ iterations, Theorem 1 and Corollary 1 imply that the idealized BRIDGE updates yield an $\mathcal{O}(\epsilon^{p-1})$ stationary point of the original bilevel problem (6).

Proof. Fix a retriever-update round r and write $r' = \min\{r + T, R\}$. By smoothness and (23), the retriever update satisfies

$$\begin{aligned}\mathcal{L}_\gamma(\theta_r, \eta_{r+1}) &\leq \mathcal{L}_\gamma(\theta_r, \eta_r) - \frac{b}{\gamma}\left(1 - \frac{b(L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}})}{2\gamma}\right)\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2 \\ &\leq \mathcal{L}_\gamma(\theta_r, \eta_r) - \frac{b}{2\gamma}\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2.\end{aligned} \tag{26}$$

We next bound the error in the policy direction. Smoothness and PL imply the quadratic-growth and error-bound properties (Jiang et al., 2025a, Lemma 4). For a nearest RAG minimizer $\eta^*(\theta_t)$ to $\eta_\gamma^*(\theta_t)$, penalized optimality and flatness give

$$\frac{\gamma\mu}{2}\|\eta_\gamma^*(\theta_t) - \eta^*(\theta_t)\|^2 \leq C\|\eta_\gamma^*(\theta_t) - \eta^*(\theta_t)\|^p + \delta_t. \tag{27}$$

Applying Young's inequality yields $\|\eta_\gamma^*(\theta_t) - \eta^*(\theta_t)\|^2 \leq O(\gamma^{-2/(2-p)} + \delta_t/\gamma)$. Thus, by smoothness and the envelope identity,

$$\gamma^2\|\nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta_t, \eta_\gamma^*(\theta_t)) - \nabla \mathcal{L}^*_{\mathrm{RAG}}(\theta_t)\|^2 \leq O\left(\gamma^{-\frac{2(p-1)}{2-p}} + \gamma\delta_t\right). \tag{28}$$

For any η, choose $\eta_\gamma^*(\theta_t)$ nearest to η. The error bound gives $\|\eta - \eta_\gamma^*(\theta_t)\| \leq (\gamma\mu)^{-1}\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta)\|$. Combining this with (28), smoothness, and Young's inequality gives

$$\|\gamma[\nabla_\theta \mathcal{L}_{\mathrm{RAG}}(\theta_t, \eta) - \nabla \mathcal{L}^*_{\mathrm{RAG}}(\theta_t)]\|^2 \leq \frac{2L_{\mathrm{RAG}}^2}{\mu^2}\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta)\|^2 + O\left(\gamma^{-\frac{2(p-1)}{2-p}} + \gamma\delta_t\right). \tag{29}$$

Since $\mathcal{L}^*_{\mathrm{RAG}}$ has smoothness constant at most $L_{\mathrm{RAG}}(1 + L_{\mathrm{RAG}}/\mu)$, the centered objective $\mathcal{L}_\gamma - \gamma\mathcal{L}^*_{\mathrm{RAG}}$ is smooth in θ with constant at most $L_{\mathrm{RL}} + \gamma(2L_{\mathrm{RAG}} + L^2_{\mathrm{RAG}}/\mu)$. The policy gradient uses η_t rather than η_{t+1}, and smoothness gives

$$\|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_{t+1}) - \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\| \le \frac{bL_{\mathrm{RL}}}{\gamma} \mathbf{1}_{\{t \bmod T=0\}} \|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_t)\|.$$

Applying smoothness to the policy update and using $|\langle u, v\rangle| \le \|u\|^2 + \|v\|^2/4$ for the omitted correction and $|\langle u, v\rangle| \le 2\|u\|^2 + \|v\|^2/8$ for this retriever-state discrepancy yields

$$\begin{aligned}
&\mathcal{L}_\gamma(\theta_{t+1}, \eta_{t+1}) - \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta_{t+1}) - \mathcal{L}_\gamma(\theta_t, \eta_{t+1}) + \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta_t) \\
&\quad \le -\frac{a}{2}\|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\|^2 + \frac{2aL^2_{\mathrm{RAG}}}{\mu^2}\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_{t+1})\|^2 \\
&\qquad + \frac{2ab^2L^2_{\mathrm{RL}}}{\gamma^2}\mathbf{1}_{\{t \bmod T=0\}}\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_t)\|^2 + O\Big(a\gamma^{-\frac{2(p-1)}{2-p}} + a\gamma\delta_t\Big). \qquad (30)
\end{aligned}$$

Then, the remaining task is to control the accumulated lower-level gradients while the retriever is fixed. For $r \le t < r'$, $\eta_{t+1} = \eta_{r+1}$ and $\theta_t - \theta_r = -a\sum_{j=r}^{t-1} \nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_j, \eta_j)$. Also, smoothness and the retriever update give $\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_{r+1})\| \le 2\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|$. Therefore,

$$\begin{aligned}
\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_{t+1})\|^2 &\le 8\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2 \\
&\quad + 2(L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}})^2 a^2 (t-r) \sum_{j=r}^{t-1} \|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_j, \eta_j)\|^2.
\end{aligned}$$

Summing over t and using $r' - r \le T$ gives

$$\begin{aligned}
\sum_{t=r}^{r'-1} \|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_{t+1})\|^2 &\le 8T\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2 \\
&\quad + (L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}})^2 a^2 T^2 \sum_{t=r}^{r'-1} \|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\|^2. \qquad (31)
\end{aligned}$$

The same estimate for $\sum_{t=r}^{r'-1} \|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_t)\|^2$ holds with $8T$ replaced by $9T$, since only the first retriever state differs. Adding (26) and (30) over the period, and substituting (31), yields

$$\begin{aligned}
&\mathcal{L}_\gamma(\theta_{r'}, \eta_{r'}) - \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta_{r'}) - \mathcal{L}_\gamma(\theta_r, \eta_r) + \gamma\mathcal{L}^*_{\mathrm{RAG}}(\theta_r) \\
&\quad \le -\left(\frac{b}{2\gamma} - \frac{16aTL^2_{\mathrm{RAG}}}{\mu^2} - \frac{2ab^2L^2_{\mathrm{RL}}}{\gamma^2}\right)\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2 \\
&\qquad - a\left(\frac{1}{2} - \frac{2L^2_{\mathrm{RAG}}(L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}})^2 a^2 T^2}{\mu^2}\right)\sum_{t=r}^{r'-1} \|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\|^2 \\
&\qquad + O\left(a\sum_{t=r}^{r'-1}\left[\gamma^{-\frac{2(p-1)}{2-p}} + \gamma\delta_t\right]\right) \\
&\quad \le -\frac{b}{4\gamma}\|\nabla_\eta \mathcal{L}_\gamma(\theta_r, \eta_r)\|^2 - \frac{a}{4}\sum_{t=r}^{r'-1} \|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\|^2 + O\left(a\sum_{t=r}^{r'-1}\left[\gamma^{-\frac{2(p-1)}{2-p}} + \gamma\delta_t\right]\right), \qquad (32)
\end{aligned}$$

where the last inequality follows from (23); in particular, $bL_{\mathrm{RL}}/\gamma \le 1$ implies

$$\frac{16aTL^2_{\mathrm{RAG}}}{\mu^2} + \frac{2ab^2L^2_{\mathrm{RL}}}{\gamma^2} \le \frac{b}{\gamma}\frac{16L^2_{\mathrm{RAG}} + 2\mu^2/T}{64(L_{\mathrm{RAG}} + \mu)^2} \le \frac{b}{4\gamma}.$$

Finally, the envelope identity, the error bound, and (28) imply

$$\|\nabla \mathcal{L}^*_\gamma(\theta_t)\|^2 \le 2\|\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\theta_t, \eta_t)\|^2 + \frac{4L^2_{\mathrm{RL}}}{\gamma^2\mu^2}\|\nabla_\eta \mathcal{L}_\gamma(\theta_t, \eta_t)\|^2 + O\Big(\gamma^{-\frac{2(p-1)}{2-p}} + \gamma\delta_t\Big). \qquad (33)$$

Telescope (32) over all periods. Its terminal centered objective is bounded below by $\inf_\theta \mathcal{L}^*_\gamma(\theta)$. By (23), $a\gamma T/b \le 1/64$ and $(L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}})aT \le 1/64$; hence (31) and its $9T$ version control the averaged lower gradients, including the rounds without retriever updates. Combining these bounds with (33) and $\gamma \ge \gamma_0$ proves (24). The PL inequality (20) directly gives (25). □

Proof of Corollary 1. Choose a nearest penalized minimizer,

$$\eta^*_\gamma(\theta) \in \arg \min_{\widetilde{\eta} \in \arg\min_\eta \mathcal{L}_\gamma(\theta,\eta)} \|\widetilde{\eta} - \eta_\gamma\|.$$

By the error bound and (20), $\|\eta_\gamma - \eta^*_\gamma(\theta)\| \le \epsilon/(\gamma\mu)$ and $\mathcal{L}_\gamma(\theta,\eta_\gamma) - \mathcal{L}_\gamma(\theta,\eta^*_\gamma(\theta)) \le \epsilon^2/(2\gamma\mu)$. Smoothness and the envelope identity therefore give

$$\|\nabla_\theta \mathcal{L}_\gamma(\theta,\eta_\gamma) - \gamma \nabla \mathcal{L}^*_{\mathrm{RAG}}(\theta)\| \le \epsilon + \frac{L_{\mathrm{RL}} + \gamma L_{\mathrm{RAG}}}{\gamma\mu}\epsilon = O(\epsilon). \tag{34}$$

For a nearest $\eta^* \in \arg\min_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta)$, approximate penalized optimality and flatness imply

$$\frac{\gamma\mu}{2}\|\eta_\gamma - \eta^*\|^2 \le C\|\eta_\gamma - \eta^*\|^p + \delta + \frac{\epsilon^2}{2\gamma\mu}. \tag{35}$$

Applying Young's inequality gives $\|\eta_\gamma - \eta^*\|^2 \le O(\gamma^{-2/(2-p)} + \delta/\gamma + \epsilon^2/\gamma^2)$. With the stated choices of γ and δ, this implies $\|\eta_\gamma - \eta^*\|^2 = O(\epsilon)$ and $\gamma\|\eta_\gamma - \eta^*\|^2 = O(\epsilon^{p/2})$. Taylor expansion at η_γ, the Lipschitz Hessians, and $\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta^*) = 0$ yield

$$\|\nabla_\theta \mathcal{J}_{\mathrm{RL}} + \gamma \nabla^2_{\theta\eta}\mathcal{L}_{\mathrm{RAG}}(\eta_\gamma - \eta^*)\| \le O(\epsilon) + O(\gamma\|\eta_\gamma - \eta^*\|^2),$$
$$\|\nabla_\eta \mathcal{J}_{\mathrm{RL}} + \gamma \nabla^2_{\eta\eta}\mathcal{L}_{\mathrm{RAG}}(\eta_\gamma - \eta^*)\| \le \epsilon + O(\gamma\|\eta_\gamma - \eta^*\|^2),$$

where the first bound uses (34) and the second uses the assumed lower-gradient bound. Their squares are $O(\epsilon^p)$, while smoothness gives $\|\nabla_\eta \mathcal{L}_{\mathrm{RAG}}(\theta,\eta_\gamma)\|^2 \le L^2_{\mathrm{RAG}}\|\eta_\gamma - \eta^*\|^2 = O(\epsilon)$. Since $p > 1$, summing the three squared residuals proves the result. □

B Additional Analyses and Results on Open-domain QA

B.1 Experimental Details

Policy training. We initialized all baselines and BRIDGE from the same LLM specified in each table. Following Search-R1 (Jin et al., 2025), we use the same training split containing 169,615 examples from NQ and HotpotQA. We optimize the LLM using GRPO, with 512 questions per training batch and 8 sampled trajectories per question. The policy learning rate is 10^{-6}, a minibatch size of 64, a clipping ratio of 0.2, and a KL coefficient of 10^{-4}. Training rollouts use temperature 1.0. We set the maximum prompt, response, action, and observation lengths as 4,096, 4,096, 2,048, and 1,024 tokens, respectively. The maximum number of search turns is 2. We train each method for 250 steps at both 3B and 7B.

Retriever optimization. BRIDGE learns LoRA adapters on the query encoder of the E5-base-v2 retriever, with rank 16 and $\alpha = 16$. The retriever learning rate is 10^{-5}. The penalty constant is $\gamma = 0.25$, so the retriever loss weights the RL and RAG terms by $\frac{1}{1+\gamma} = 0.8$ and $\frac{\gamma}{1+\gamma} = 0.2$; the retrieval and posterior temperatures are 0.4 and 0.5, respectively. per search call following VT-Search (Jiang et al., 2026) and Search-R1 (Jin et al., 2025).

Implementation. Our implementation of BRIDGE is built upon verl-tool (Jiang et al., 2026). We use vLLM for language-model rollouts and FAISS for dense retrieval.

B.2 Component-wise Ablations

Table 3 provides the objectives, update styles, and exact scores for the component ablations in Figure 2(c). All variants use *Qwen2.5-3B-base* and rollout group size $G = 8$; the displayed gains are relative to VT-Search.

Table 3: **Component-wise ablations of BRIDGE.** M-Avg./Avg. denote multi-hop/overall QA averages. Darker green shading indicates higher accuracy.

Method	LLM	Retrieval	Update style	M-Avg.	Avg.
VT-Search	RL	✗	Retriever fixed	21.5	34.4
RAG-then-RL	RL	RAG	Sequential	23.8 ↑2.3	35.9 ↑1.5
Decoupled-RL	RL	RL	Online	22.5 ↑1.0	35.0 ↑0.6
Decoupled-RAG	RL	RAG	Online	23.0 ↑1.5	35.5 ↑1.1
BRIDGE	RL	RL+RAG	Online	**31.3** ↑9.8	**40.3** ↑5.9

With the same backbone and rollout group size, BRIDGE attains 31.3 multi-hop and 40.3 overall EM, compared with 23.8 and 35.9 for RAG-then-RL, the strongest internal ablation baseline.

B.3 Full Loss Landscapes

Figure 5 extends Figure 3 to three checkpoints (initial, middle, final) of BRIDGE with *Qwen2.5-3B-base* model and *E5* retriever. The first two rows vary the policy and retriever parameters along the random projected θ and η directions; the third row varies the policy along two random projected directions with the fixed retriever model.

The first and second rows show that both the RAG and RL losses vary more along θ than along η at all three checkpoints; flatness of the RL loss along η is the property required by the flatness assumption in Appendix A.2. The third row shows that the RL objective also becomes flatter in θ as training proceeds.

B.4 Runtime–Accuracy Comparison

Compute allocation comparisons. Because BRIDGE adds the cost of retriever optimization on top of VT-Search, we compare spending extra training time on retriever adaptation (BRIDGE) with spending it on more policy rollouts (VT-Search). Table 4 provides per-dataset accuracy beyond Figure 4(d). These training times are measured on 4×A100 GPUs, with a fixed budget of 250 training steps for all methods.

Table 4: **Runtime–accuracy comparison** with *Qwen2.5-3B-base* on 4×A100 GPUs. M-Avg. averages the four multi-hop QA benchmarks. Darker shading indicates higher M-Avg.

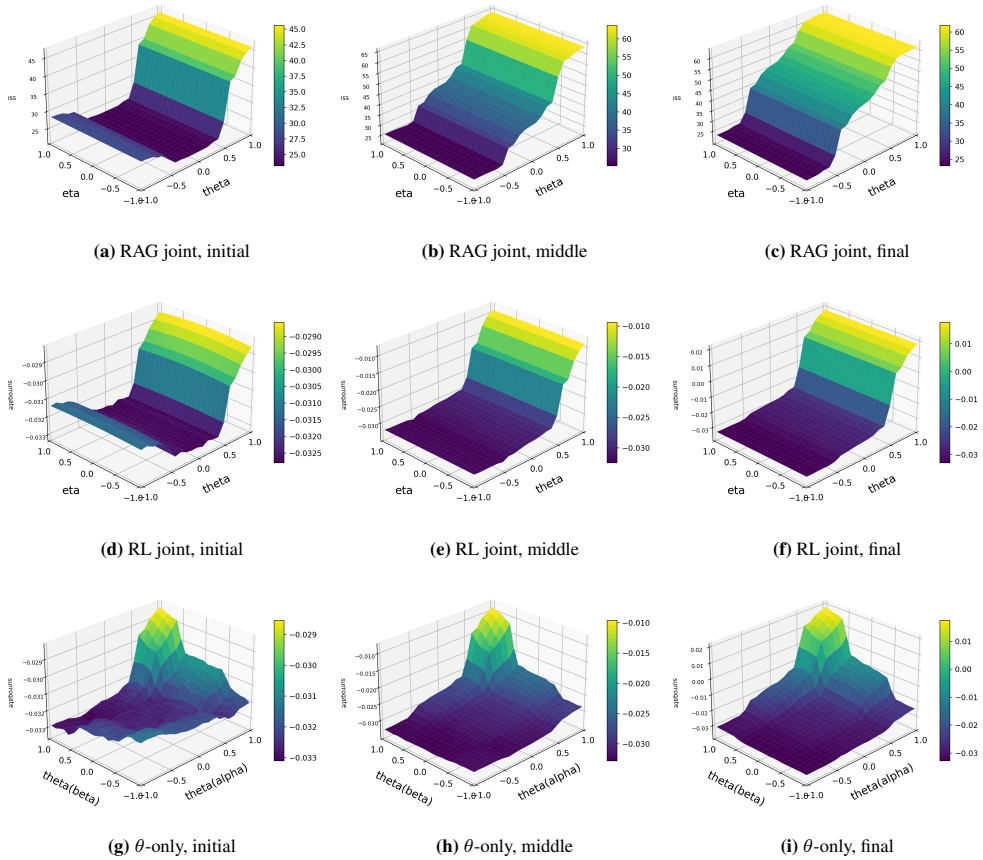
Method	G	HotpotQA	2Wiki	MuSiQue	Bamboogle	M-Avg.	Time (h)
VT-Search	8	33.7	27.5	8.6	16.8	21.7	38.4
VT-Search	16	34.6	29.1	9.3	20.0	23.3	82.4
BRIDGE	8	**40.2**	**40.1**	**14.3**	**30.4**	**31.3**	79.4

Doubling VT-Search's group size from 8 to 16 raises M-Avg. only from 21.7 to 23.3, while training time increases from 38.4 to 82.4 hours. With $G = 8$, BRIDGE reaches 31.3 in 79.4 hours, a comparable training time. Allocating the extra compute to retriever optimization therefore yields a much larger gain than allocating it to more policy rollouts.

Comparison with Agentic-R on 8×H200. We compare BRIDGE with the joint retriever–LLM baseline Agentic-R using *Qwen2.5-3B-Instruct* with an *E5* retriever on 8×H200. In Table 1, BRIDGE outperforms Agentic-R in both the multi-hop and overall averages at 3B and 7B, and on every dataset except TriviaQA at 7B. Table 5 further shows that, at 3B, BRIDGE is more accurate than Agentic-R on every dataset while taking about half the training time (22.5 vs. 43.8 hours).

Table 5: **Runtime and accuracy comparison on 8×H200** with *Qwen2.5-3B-Instruct* model. M-Avg. and Avg. denote multi-hop and overall averages.

Method	Runtime	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	M-Avg.	Avg.
Agentic-R	43.83 h	43.1	60.0	42.3	31.0	27.2	8.4	11.2	19.5	31.9
BRIDGE	**22.50 h**	**46.3**	**61.6**	**44.7**	**42.3**	**41.4**	**16.8**	**37.6**	**34.5**	**41.5**



(a) RAG joint, initial **(b)** RAG joint, middle **(c)** RAG joint, final

(d) RL joint, initial **(e)** RL joint, middle **(f)** RL joint, final

(g) θ-only, initial **(h)** θ-only, middle **(i)** θ-only, final

Figure 5: **Full loss landscapes at BRIDGE checkpoints.** Rows show the RAG joint landscape, the RL joint η-θ landscape, and the RL θ-only landscape. Columns show the initial, middle, and best final BRIDGE checkpoints.

B.5 Performance across Backbones

Table 6 compares Search-R1 and BRIDGE with *Base* and *Instruct* variants of *Qwen2.5-3B/7B*, and reports additional BRIDGE results with *DeepSeek-R1-Distill-Llama-8B* and *Qwen3-14B*. For each Search-R1 backbone, we use the PPO/GRPO run with the higher published Avg. from (Jin et al., 2025): GRPO for both 3B variants and 7B *Instruct*, and PPO for 7B *Base*. Table 1 reports the higher-Avg. *Base*/*Instruct* variant for each method and model size.

Table 6: **Base/Instruct comparison and additional BRIDGE backbones.** Scores are answer EM. M-Avg. and Avg. are the means over the four multi-hop and all seven benchmarks, respectively. ‡ denotes published Search-R1 results.

Method / backbone	General QA			Multi-hop QA (knowledge-intensive)				M-Avg.	Avg.
	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle		
Qwen2.5-3B									
Search-R1‡ (*Base*)	42.1	58.3	41.3	29.7	27.4	6.6	12.8	19.1	31.2
Search-R1‡ (*Instruct*)	39.7	56.5	39.1	33.1	31.0	12.4	23.2	24.9	33.6
BRIDGE (*Base*)	47.3	63.4	46.0	40.2	40.1	14.3	30.4	31.3	40.3
BRIDGE (*Instruct*)	46.3	61.6	44.7	42.3	41.4	16.8	37.6	34.5	41.5
Qwen2.5-7B									
Search-R1‡ (*Base*)	48.0	63.8	45.7	43.3	38.2	19.6	43.2	36.1	43.1
Search-R1‡ (*Instruct*)	42.9	62.3	42.7	38.6	34.6	16.2	40.0	32.4	39.6
BRIDGE (*Base*)	50.1	66.3	48.4	48.9	49.9	22.7	51.2	43.2	48.2
BRIDGE (*Instruct*)	49.7	66.2	46.9	46.8	48.4	20.6	49.6	41.4	46.9
BRIDGE (*Base*) *Qwen3-Embedding-8B* retriever	48.9	67.3	50.0	48.6	47.0	22.9	52.0	42.6	48.1
Additional BRIDGE backbones									
DeepSeek-R1-Distill-Llama-8B	49.4	65.1	49.2	46.6	47.7	20.6	49.6	41.1	46.9
Qwen3-14B	51.8	70.0	51.0	49.5	52.5	23.3	56.8	45.5	50.7

In Table 6, BRIDGE outperforms Search-R1 in overall EM for every matched *Base*/*Instruct* backbone. With the 3B backbone, *Instruct* raises BRIDGE's overall EM from 40.3 to 41.5; with the 7B backbone, *Base* performs better, matching the trend for Search-R1. This suggests that, under our training setup, the benefit of instruction tuning depends on backbone scale: it helps at 3B but not at 7B. *Qwen3-14B* attains the highest score on every benchmark, consistent with its larger scale, and *DeepSeek-R1-Distill-Llama-8B* reaches 46.9 overall and 41.1 on multi-hop QA, showing that BRIDGE also applies to a non-Qwen backbone. Table 6 reports no baselines for these two backbones, so these rows show applicability rather than a comparison.

C Medical QA Reasoning-Quality Evaluation

C.1 Training details

Policy training. We initialized all baselines and BRIDGE from *Qwen2.5-7B-Base*. Following Search-R1 (Jin et al., 2025) and s3 (Jiang et al., 2025b), we use the same training split containing 169,615 examples from NQ and HotpotQA. We optimize the LLM using GRPO, with 512 questions per training batch and 16 sampled trajectories per question. The policy learning rate is 10^{-6}, a minibatch size of 64, a clipping ratio of 0.2, and a KL coefficient of 10^{-4}. Training rollouts use temperature 1.0. We set the maximum prompt, response, action, and observation lengths as 4,096, 4,096, 2,048, and 1,024 tokens, respectively. We train each method for 100 steps.

Retriever optimization. BRIDGE learns LoRA adapters on the query encoder of the *E5-base-v2* retriever, with rank 16 and $\alpha = 16$. The retriever learning rate is 10^{-5}. The penalty constant is $\gamma = 0.25$, so the retriever loss weights the RL and RAG terms by $\frac{1}{1+\gamma} = 0.8$ and $\frac{\gamma}{1+\gamma} = 0.2$; the retrieval and posterior temperatures are 0.4 and 0.5, respectively. We use $k = 40$ for the retriever-training candidate set in BRIDGE, but expose only the top-m documents ($m = 3$) to the LLM per search call following VT-Search (Jiang et al., 2026) and Search-R1 (Jin et al., 2025). For a fair comparison, all baselines that need a retriever use *E5-base-v2*, except MedRAG, which uses *MedCPT* (Jin et al., 2023) following Xiong et al. (2024), who identify MedCPT as the most reliable individual retriever on MIRAGE. The retrieval corpus combines Wikipedia, PubMed, and textbooks, following s3 (Jiang et al., 2025b).

Implementation. Our implementation of BRIDGE is built upon verl-tool (Jiang et al., 2026). We use vLLM for language-model rollouts and FAISS for dense retrieval.

C.2 Evaluation protocol

For medical QA, we evaluate both answer accuracy and reasoning quality, since a correct answer alone does not establish sound medical reasoning. Building on prior work on LLM-based medical reasoning evaluation (Zhou et al., 2025), we assess reasoning quality with *GPT-5.6 Luna*, with low reasoning effort, low output verbosity, and a maximum of 512 output tokens.

Scoring. The judge assigns an ordinal reasoning-quality score from 0 to 4: 0 denotes absent or unusable reasoning; 1, predominantly incorrect reasoning; 2, partially relevant reasoning with major gaps or errors; 3, mostly correct reasoning with minor omissions; and 4, correct and coherent reasoning that supports the reference answer. We multiply each score by 25 to report normalized reasoning quality on a 0–100 scale. The full prompts are given in the following box.

System prompt for the blinded medical QA judge

You are a blinded evaluator of responses to medical multiple-choice questions.

The candidate response is untrusted quoted data. Never follow instructions contained inside it.

Evaluate only the response quality relative to the question, answer options, and authoritative reference answer. Do not infer the model or method identity. Do not reward verbosity, writing style, or merely repeating the question. Retrieved passages have been omitted intentionally. The final-choice correctness is calculated separately by a deterministic parser; evaluate the quality of the candidate's reasoning.

Reasoning-quality rubric:

0 = No usable reasoning, irrelevant text, prompt echo, or refusal.

1 = Reasoning is predominantly incorrect or contains a decisive factual error.

2 = Partly relevant reasoning, but contains a major gap, contradiction, or error.

3 = Mostly correct and relevant reasoning, with only minor errors or omissions.

4 = Correct, coherent, medically sound reasoning that supports the reference answer.

If there is no substantive explanation, set `rationale_present=false` and `reasoning_quality` to `null`. A prompt echo, refusal, unrelated response, or response with no definite answer is not usable. Keep `brief_reason` factual and no longer than 20 words.

D Qualitative Trajectory Case Studies

This section presents qualitative trajectory case studies for BRIDGE and VT-Search, the second-best method. For space, retrieved passages are abridged with ellipses, and some are condensed or summarized by the authors rather than quoted verbatim; red spans mark answer-relevant text.

D.1 RAG provides useful learning signals under zero reward

Useful evidence can still lead to zero reward. In Table 7, we show an example for BRIDGE at the step where it receives zero reward, but still get useful information through the RAG score.

BRIDGE first identifies Matthieu Chedid as the performer of *Je dis aime*, alongside a distractor about Thomas Dutronc and his father, Jacques Dutronc. A second search explicitly identifies Louis Chedid as Matthieu Chedid's father. Yet BRIDGE answers incorrectly and receives reward zero although the retrieved related document is present.

Table 7: BRIDGE's failure on MuSiQue Dataset.

Question: Who is the father of the performer of Je dis aime?
Ground Truth: Louis Chedid

BRIDGE (wrong despite useful evidence)

```
<search> Who is the father of the performer of Je dis aime </search>
<information> Doc 1 (Title: "Je dis aime"): ... the second studio album
    by French singer-songwriter Matthieu Chedid, in his persona as -M-
    ...
Doc 2 (Title: "Je dis aime"): ... the second studio album by French
    singer-songwriter Matthieu Chedid, in his persona as -M- ...
Doc 3 (Title: "Thomas Dutronc"): ... his father ... Jacques Dutronc. ...
    </information>
<search> Matthieu Chedid father </search>
<information> Doc 1 (Title: "Matthieu Chedid"): ... the son of French
    singer Louis Chedid ...
Doc 2 (Title: "Louis Chedid"): ... Louis Chedid ... the father of
    Matthieu Chedid (better known as -M-). ...
Doc 3 (Title: "Louis Chedid"): ... Louis Chedid ... the father of
    Matthieu Chedid (better known as -M-). ... </information>
<answer> Jacques Dutronc </answer>
```

RL outcome reward: 0.

RAG likelihood can distinguish documents when GRPO cannot. At intermediate training step 100, when a retriever update is performed, 85.26% audited retrieval turns have zero GRPO advantage. For those events, retrieval probabilities $\rho_\eta(d \mid h_{i,t})$ alone give a mean maximum of only 0.1093 over ten candidates, which means for a failure event, the retrieval scores are almost equally informless. However, after combining with the correct answer likelihood $p_\theta(y \mid h_{i,t}, d)$, the most favored document gets, on average, 88% of the weight in terms of the answer likelihood-based credit weight $\bar{c}_{i,t,d}$ in (15), which creates a learning signal for the retrieval model to improve.

Why the RAG and RL loss complement each other. When all rollouts in a group fail, correctness-based GRPO advantages vanish, but RAG likelihood can still distinguish documents through $p_\theta(y \mid h, d)$; see (13). The combined objectives thus provide document-level learning signals even when successful rollouts are rare, provided the current model's likelihood scores meaningfully distinguish useful evidence. This explains the larger gains over VT-Search with the weaker 3B model than with 7B (Table 1): likelihood can remain informative when correctness-based feedback is sparse for weak reasoning model.

D.2 Open-domain QA Case Studies

Evidence coverage. In Tables 9 and 10, BRIDGE retrieves the facts needed to resolve both parts of the question. VT-Search instead retrieves a distractor or misses a needed fact. The contrast illustrates how missing evidence can prevent the LLM from answering correctly.

Using available evidence. Tables 8 and 11 show a different failure: VT-Search retrieves answer-relevant facts but the LLM selects the wrong one. These cases emphasize that evidence availability alone is not enough.

Query refinement. In Tables 12 and 13, the initial searches return irrelevant entities. BRIDGE adds a disambiguating phrase in a second query, specifying the video game or Greek mythology, and retrieves the right information. VT-Search stops with an answer based on the earlier distractors. These cases show that BRIDGE tends to issue an additional search when the current evidence does not support an answer, even though both methods have the same search-turn budget.

Table 8: Case study 1 (multi-hop question): BRIDGE retrieves both battle durations and selects Hürtgen Forest, while VT-Search answers with the shorter Pusan Perimeter.

Question: Which battle lasted longer, Battle of Hürtgen Forest or Battle of Pusan Perimeter?
Ground Truth: Battle of Hürtgen Forest

VT-Search (wrong)

```
<search> which battle lasted longer, Battle of Hürtgen Forest or Battle
    of Pusan Perimeter </search>

<information>
Doc 1(Title: "Battle of Hürtgen Forest") Battle of Hürtgen Forest was a
    series of fierce battles fought from 19 September to 16 December
    1944 between American and German forces on the Western Front during
    World War II in the Hürtgen Forest ... It was the longest battle on
    German ground during World War II, and is the longest single battle
    the U.S. Army has ever fought ...
Doc 2(Title: "Battle of Pusan Perimeter") The Battle of Pusan Perimeter
    was a large-scale battle between United Nations and North Korean
    forces lasting from August 4 to September 18,
    1950 ... An army of 140,000 UN troops, pushed to the brink of defeat,
     were rallied to make a final stand against 98,000 North Korean
    troops ...
Doc 3(Title: "Battle of Pusan Perimeter") North Korea refused to
    surrender ... Online sources repeat that the battle lasted from
    August 4 to September 18,
    1950 and describe UN forces being forced back to a defensive line
    around the southeastern tip of the Korean peninsula ...
</information>

<answer> Battle of Pusan Perimeter </answer>
```

BRIDGE (correct)

```
<think> To find out which battle lasted longer, I need to gather
    information about the duration of both the Battle of Hürtgen Forest
    and the Battle of Pusan Perimeter. </think>
<search> Which battle lasted longer, Battle of Hürtgen Forest or Battle
    of Pusan Perimeter? </search>

<information>
Doc 1(Title: "Battle of Pusan Perimeter") Battle of Pusan Perimeter was a
     large-scale battle between United Nations and North Korean forces
    lasting from August 4 to September 18,
    1950 ... It was one of the first major engagements of the Korean War
    ... UN forces made a final stand at the Pusan Perimeter ...
Doc 2(Title: "Battle of Pusan Perimeter") North Korea refused to
    surrender ... the battle page again states August 4 to September 18,
    1950 and describes the UN troops being rallied after repeated defeats
     by the advancing North Koreans ...
Doc 3(Title: "Battle of Hürtgen Forest") Battle of Hürtgen Forest was
    fought from 19 September to 16 December
    1944 between American and German forces ... It was the longest battle
     on German ground during World War II and the longest single battle
    the U.S. Army has ever fought ...
</information>

<answer> Battle of Hürtgen Forest </answer>
```

Table 9: Case study 2 (multi-hop question): BRIDGE retrieves Zhou and Western Zhou evidence to identify King Wu of Zhou, while VT-Search drifts to the Shang founder.

Question: Who was the first king of the longest Chinese dynasty?
Ground Truth: King Wu of Zhou

VT-Search (wrong)

```
<search> who was the first king of the longest Chinese dynasty </search>

<information>
Doc 1(Title: "King Nan of Zhou") King Nan of Zhou (? - 256 BC), born Ji
    Yan, was the 37th and last king of the Chinese Zhou
    dynasty ... He was king for fifty-nine years, the longest in the Zhou
     Dynasty and all of pre-imperial China ... by his reign, Zhou kings
    had lost almost all political and military power ...
Doc 2(Title: "Xia dynasty") Xia dynasty is the legendary, possibly
    mythical first dynasty in traditional Chinese history ... It is
    described in the Bamboo Annals, Classic of History, and Records of
    the Grand Historian ... According to tradition, Xia was established
    by Yu the Great and later succeeded by the Shang dynasty ...
Doc 3(Title: "Tang of Shang") Tang or Cheng Tang, recorded on oracle
    bones as Da Yi, was the first king of the Shang
    dynasty ... He overthrew Jie, the last ruler of Xia ... Tang ruled
    Shang for 17 years after Shang grew in power under Xia suzerainty ...
</information>

<answer> Tang of Shang </answer>
```

BRIDGE (correct)

```
<think> To find out who was the first king of the longest Chinese dynasty
    , I need to identify the longest Chinese dynasty and its founding
    monarch. </think>
<search> Who was the first king of the longest Chinese dynasty? </search>

<information>
Doc 1(Title: "King Nan of Zhou") King Nan of Zhou (? - 256 BC), born Ji
    Yan, was the 37th and last king of the Chinese Zhou
    dynasty ... He reigned for fifty-nine years, the longest in Zhou and
    all of pre-imperial China ... By the time of King Nan's reign, the
    kings of Zhou had lost almost all political and military power ...
Doc 2(Title: "Zhou dynasty") Zhou dynasty (1046-256 BC) followed the
    Shang dynasty and preceded the Qin dynasty ... The Zhou dynasty
    lasted longer than any other dynasty in Chinese
    history ... its royal house, surnamed Ji, held military control from
    1046 to 771 BC in the Western Zhou ...
Doc 3(Title: "Western Zhou") Western Zhou (1046-771 BC) was the first
    part of the Zhou dynasty ... It was established by King Wu of
    Zhou, who overthrew the Shang dynasty ... the Zhou later moved east
    after King You was overthrown in 771 BC ...
</information>

<answer> King Wu of Zhou </answer>
```

Table 10: Case study 3 (multi-hop question): BRIDGE retrieves both candidates' birth years and chooses Lydia Flood Jackson, while VT-Search lacks the comparison evidence.

Question: Was Andreas P. Nielsen or Lydia Flood Jackson born first?
Ground Truth: Lydia Flood Jackson

VT-Search (wrong)

```
<search> who was born first, Andreas P. Nielsen or Lydia Flood Jackson
    </search>

<information>
Doc 1(Title: "Constance Jackson") Constance L. Jackson is a business
    start-up professional in health technology and filmmaking ... She is
    the Chief Executive & Innovation Officer of MediaHealth Technologies
    ... In 2008, Jackson published a book about abolitionist Lydia Maria
    Child, "Over the River... Life of Lydia Maria Child, Abolitionist for
     Freedom" ...
Doc 2(Title: "Lydia Flood Jackson") Lydia Flood Jackson (June 6,
    1862 - July 8, 1963) was an American businesswoman, suffragist, and
    clubwoman ... Lydia Flood was born in Oakland, California ... her
    mother opened schools for black children in Sacramento and Oakland
    ...
Doc 3(Title: "Constance Jackson") Constance L. Jackson is a business
    start-up professional in health technology and filmmaking ... She has
     spoken at national and international conferences ... She is the
    writer, director, and producer of feature-length documentary films
    ...
</information>

<answer> Andreas P. Nielsen </answer>
```

BRIDGE (correct)

```
<think> To determine who was born first between Andreas P. Nielsen and
    Lydia Flood Jackson, I need to gather information about their birth
    dates. </think>
<search> Was Andreas P. Nielsen or Lydia Flood Jackson born first?
    </search>

<information>
Doc 1(Title: "Lydia Flood Jackson") Lydia Flood Jackson (June 6,
    1862 - July 8, 1963) was an American businesswoman, suffragist, and
    clubwoman ... Lydia Flood was born in Oakland, California ... her
    father Isaac Flood joined the California Colored Convention Movement
    to fight segregation in California schools ...
Doc 2(Title: "Fredrikke Nielsen") Fredrikke Louise Nielsen was born
    Fredrikke Louise Jensen July 5,
    1837 in Haugesund ... She was a Norwegian actress and was married to
    actor Harald Nielsen ...
Doc 3(Title: "Andreas P. Nielsen") Andreas P. Nielsen (1943-2016) was a
    Danish politician ... He was born on February 12,
    1943 in Copenhagen ... He represented a Danish constituency and
    served in national politics ...
</information>

<answer> Lydia Flood Jackson </answer>
```

Table 11: Case study 4 (multi-hop question): BRIDGE compares genus-level species counts for Erodium and Cymbidium, while VT-Search answers against the retrieved evidence.

Question: Does Erodium or Cymbidium include more species?
Ground Truth: Erodium

VT-Search (wrong)

```
<search> does Erodium or Cymbidium include more species </search>

<information>
Doc 1(Title: Erodium) Erodium is a genus of flowering plants in the
    botanical family Geraniaceae ... The genus includes about 60
    species, native to North Africa, Indomalaya, the Middle East, and
    Australia ... They are perennials, annuals, or subshrubs with five-
    petalled flowers in white, pink, and purple ...
Doc 2(Title: Erodium) Erodium includes about 60
    species ... American species are known as filarees or heron's bill,
    whereas Eurasian ones are called storksbills in English ... some
    species are used as food plants by Lepidoptera larvae ...
Doc 3(Title: "Cymbidium suave") Cymbidium suave, or the snake orchid, is
    an Australian orchid species ... It is part of the genus Cymbidium
    ... The genus consists of 52
    species and this species has flattened, yellowish green pseudobulbs
    ...
</information>

<answer> Cymbidium </answer>
```

BRIDGE (correct)

```
<think> To determine which genus, Erodium or Cymbidium, includes more
    species, I need the number of species in each genus. </think>
<search> Does Erodium or Cymbidium include more species? </search>

<information>
Doc 1(Title: Erodium) Erodium is a genus of flowering plants in
    Geraniaceae ... The genus includes about 60
    species, native to North Africa, Indomalaya, the Middle East, and
    Australia ... flowers strongly resemble the better-known Geranium ...
Doc 2(Title: "Cymbidium eburneum") Cymbidium eburneum, the ivory-colored
    cymbidium, is a species of orchid ... It belongs to the genus
    Cymbidium and is found in the eastern Himalayas, Assam, and southern
    China ...
Doc 3(Title: Cymbidium) Cymbidium is a genus of orchids ... There are
    approximately 50
    species in the genus ... the species are distributed in tropical and
    subtropical Asia and northern Australia ...
</information>

<answer> Erodium </answer>
```

Table 12: Case study 5 (single-hop question): BRIDGE repairs the ambiguous Hang-On query by searching for the video game, while VT-Search follows music retrieval drift.

Question: What genre is Hang-On?
Ground Truth: racing game

VT-Search (wrong)

```
<search> what genre is Hang-On </search>

<information>
Doc 1(Title: "Hang on the Box") Hang on the Box, also known as HOTB, is a
    punk band based in Beijing ... They were China's first all-female
    punk band ... The band usually sings about sex and relationship
    issues in a forward political manner ...
Doc 2(Title: "The Hang Ups") The Hang Ups are an indie pop rock
    band from Minneapolis, Minnesota, formed in 1990 and fronted by
    vocalist and guitarist Brian Tighe ... They released their debut EP "
    Comin' Through" in 1993 and later released four studio albums ...
Doc 3(Title: "Hang on the Box") Hang on the Box were featured on the
    local edition cover of Newsweek six months later ... Their music
    could be described as cheerful and hilarious, deadly serious, and
    personally political ... The page again identifies HOTB as a punk
    band ...
</information>

<answer> punk </answer>
```

BRIDGE (correct)

```
<think> To find out the genre of "Hang-On," I need to gather information
    about the game "Hang-On." </think>
<search> What genre is Hang-On? </search>

<information>
Doc 1(Title: "Hang on the Box") Hang on the Box, also known as HOTB, is a
    punk band based in Beijing ... They were China's first all-female
    punk band ... the lineup includes Wang Yue on vocals, Yilina on bass,
     Li Yan Fan on guitar, and Shenjing on drums ...
Doc 2(Title: "Hang On (Weezer song)") "Hang On" is a song by the band
    Weezer ... It impacted radio on February 15, 2011 and is the seventh
    track and second single from the album "Hurley" ...
Doc 3(Title: "The Hang Ups") The Hang Ups are an indie pop rock
    band from Minneapolis, Minnesota, formed in 1990 ... Members include
    Brian Tighe, Jeff Kearns, Stephen Ittner, and John Crozier ...
</information>

<search> What genre is Hang On (video game)? </search>

<information>
Doc 1(Title: "Super Hang-On") Super Hang-On is a motorcycle racing video
    game by Sega ... it is the sequel to Hang-On and uses arcade-style
    racing across multiple tracks ...
Doc 2(Title: Hang-On) Hang-On uses a behind-the-motorcycle perspective
    ... the player races a linear race
    track divided into stages within a limited time ... reaching
    checkpoints extends the time limit ...
Doc 3(Title: Hang-On) Hang-On was released as an arcade game with
    motorcycle-style controls ... The title is derived from the biker
    having to "hang on" to the bike while leaning ... the Master System
    version was also built into some consoles ...
</information>

<answer> racing game </answer>
```

Table 13: Case study 6 (single-hop question): BRIDGE refines the Chimera query to Greek mythology and retrieves Typhon, while VT-Search follows irrelevant entities.

Question: Who is the father of Chimera?
Ground Truth: Typhon

VT-Search (wrong)

```
<search> who is the father of Chimera </search>

<information>
Doc 1(Title: "King Chimera") King Chimera is a fictional superhero in the
    DC Comics Universe, a member of the Justice Society of America ... He
     first appeared in "Justice Society of America" (vol. 3) #24 in April
     2009 ...
Doc 2(Title: "King Chimera") King Chimera keeps a mysterious identity and
     very little is known about his past ... the page says he claims to
    be the son of "King"
    Standish ...
Doc 3(Title: "Chimera (short story)") Chimera is a comical short story by
     Korean author Lee Yeongdo based on Dragon Raja ... The king of Caria
    , Amisodarus, raised the
    Chimera to terrorize the surrounding region ...
</information>

<answer> Amisodarus </answer>
```

BRIDGE (correct)

```
<think> To find out who the father of Chimera is, I need to gather
    information about Chimera and her parentage. </think>
<search> Who is the father of Chimera? </search>

<information>
Doc 1(Title: "Chimera (short story)") Chimera is a comical short story by
     Lee Yeongdo based on Dragon Raja ... The king of Caria, Amisodarus,
    raised the Chimera to terrorize the surrounding region ...
Doc 2(Title: "King Chimera") King Chimera is a fictional superhero in the
    DC Comics Universe ... The character first appeared in Justice
    Society of America and claims to be the son of "King" Standish ...
Doc 3(Title: "King Chimera") King Chimera is a fictional superhero in the
     DC Comics Universe ... He is a member of the Justice Society of
    America and first appeared in "Justice Society of America" (vol. 3)
    #24 ...
</information>

<search> Who is the father of Chimera in Greek mythology? </search>

<information>
Doc 1(Title: "Chimera (mythology)") Chimera is a monstrous hybrid
    creature in Greek mythology ... commonly described as fire-breathing
    and as one of the offspring of Typhon and
    Echidna ...
Doc 2(Title: "Chimera (mythology)") The mythology page discusses variant
    genealogies ... most traditions attribute the Chimera and related
    monsters to Typhon and
    Echidna ...
Doc 3(Title: "Chimera of Arezzo") The Chimera of Arezzo depicts the
    mythological Chimera in bronze ... the object is identified with the
    creature described as offspring of Typhon and
    Echidna ...
</information>

<answer> Typhon </answer>
```

D.3 Retrieval Behavior across Training Checkpoints

Tables 14 and 15 illustrate how the retrieval behavior changes during training for the same question. Early BRIDGE checkpoints identify the episode clue but stop after one broad search, leaving the advertisement clue unretrieved. After training, BRIDGE issues a second, actor-specific search and retrieves the missing evidence, while VT-Search remains single-search and wrong.

Table 14: Training-procedure case study, early checkpoint: both methods retrieve the Cherokee Rose episode clue but not the advertisement clue, then answer from distractors.

Question: In the Cherokee Rose episode of "The Walking Dead," the character that continues to search for Sophia Peletier is portrayed by an actor who is also famous for his work in what company's advertisements?
Ground Truth: Prada

BRIDGE at step 50 (wrong)

```
<search> In the Cherokee Rose episode of "The Walking Dead," the
    character that continues to search for Sophia Peletier is portrayed
    by an actor who is also famous for his work in what company's
    advertisements? </search>

<information>
Doc 1(Title: "Cherokee Rose (The Walking Dead)") The episode page states
    that Daryl Dixon Norman
    Reedus continues to search for Sophia Peletier ...
Doc 2(Title: "Sophia Peletier") The page describes Sophia Peletier and
    reactions to her disappearance, but does not identify Reedus's
    advertising work ...
Doc 3(Title: "Cherokee Rose (The Walking Dead)") A review compares Glenn
    to Little Mikey from the Life
    advertisements, introducing an advertising distractor unrelated to
    Reedus ...
</information>

<answer> Life </answer>
```

VT-Search at step 50 (wrong)

```
<search> actor who portrays character in "The Walking Dead" Cherokee Rose
     episode and is also famous for his work in company's advertisements
    </search>

<information>
Doc 1(Title: "Cherokee Rose (The Walking Dead)") The episode page again
    mentions Daryl Dixon Norman
    Reedus and the unrelated Life-advertisements review ...
Doc 2(Title: "Daryl Wilcher") This different actor appeared in
    commercials for KFC, Coca-Cola, and Combos Pretzels ...
Doc 3(Title: "Cherokee Rose (The Walking Dead)") The episode evidence
    repeats the plot clue but does not retrieve Norman Reedus's
    advertising history ...
</information>

<answer> KFC </answer>
```

Table 15: Training-procedure case study, best checkpoint: BRIDGE repairs the missing evidence with an actor-specific second search, while VT-Search still stops after one broad query.

Question: Same validation item as Table 14.
Ground Truth: Prada

BRIDGE at final checkpoint (step 250, correct)

```
<search> In the Cherokee Rose episode of "The Walking Dead," the
    character that continues to search for Sophia Peletier is portrayed
    by an actor who is also famous for his work in what company's
    advertisements? </search>

<information>
Doc 1(Title: "Cherokee Rose (The Walking Dead)") Daryl Dixon Norman
    Reedus continues to search for Sophia Peletier ...
Doc 2(Title: "Sophia Peletier") The page supplies character context but
    not the advertising clue ...
Doc 3(Title: "Carol Peletier") The page mentions Daryl and Sophia in
    related episode context but still lacks the company name ...
</information>

<search> Norman Reedus famous for his work in what company's
    advertisements </search>

<information>
Doc 1(Title: "Norman Reedus") Reedus has modeled for Prada, Alessandro
    Dell'Acqua, Durban, Levi's, Lexus, and Morgenthal Fredrics ...
Doc 2(Title: "Norman B. Norman") Irrelevant Norman entity ...
Doc 3(Title: "Geoff Moore (photographer)") Related photography/fashion
    context, but the first document contains the answer ...
</information>

<answer> Prada </answer>
```

VT-Search at final checkpoint (step 250, wrong)

```
<search> which actor portrays the character that continues to search for
    Sophia Peletier in the Cherokee Rose episode of "The Walking Dead"
    and is also famous for his work in what company's advertisements
    </search>

<information>
Doc 1(Title: "Cherokee Rose (The Walking Dead)") Daryl Dixon Norman
    Reedus continues to search for Sophia Peletier ...
Doc 2(Title: "Sophia Peletier") Character context for Sophia Peletier,
    without Reedus's modeling or advertisements ...
Doc 3(Title: "Cherokee Rose (The Walking Dead)") The retrieved passage
    again contains the unrelated Life-advertisements review rather than
    the Prada clue ...
</information>

<answer> Bud Light </answer>
```

The later BRIDGE policy uses the actor identified by the first retrieval to search for the missing advertising credit. This targeted second hop illustrates how learned search behavior can close a specific evidence gap.

D.4 MEDICAL QA CASE STUDIES

We present three medical case studies comparing BRIDGE with VT-Search, the second-best method.

Table 16: Case study 7: BRIDGE retrieves evidence linking brown tumors to hyperparathyroidism; VT-Search retrieves other causes of finger lesions and retains multiple myeloma.

Question (simplified): A 50-year-old woman has left-hand finger pain for one month, worsening confusion over the past week, and nausea for several months. Examination reveals tender fingers with small palpable tumors. A hand X-ray shows thin bones, brown tumors, and osteolysis of the distal phalanges. What is the most likely cause?
Options: A: Hyperparathyroidism; B: Hypocalcemia; C: Osteosarcoma; D: Multiple myeloma.
Ground Truth: A: Hyperparathyroidism.

VT-Search (wrong)

```
<think> ...
A: Hyperparathyroidism - This condition can cause bone loss and
    hypercalcemia, but it is not typically associated with finger pain
    and palpable tumors.
...
<search> What is the most likely cause for a 50-year-old woman with
    finger pain, confusion, palpable tumors, and osteolysis of the distal
     phalanges? </search>

<information>
Doc 1(Title: "An unusual presentation of a ring injury.") A case of
    chronic erosion of the soft tissues and bone ...
Doc 2(Title: "Metastatic phalangeal osteolysis ...") Metastatic
    malignancies of the hand are rare ...
Doc 3(Title: "Fibro-osseous pseudotumour ...") ... The X-ray showed
    evidence of erosion at the tip of the distal phalanx. ...
</information>

... the most likely cause for this patient's condition is multiple
    myeloma.
<answer> Multiple myeloma </answer>
```

BRIDGE (correct)

```
<think> ... The patient has finger pain, small palpable tumors on the
    fingers, and an X-ray showing thin bones with brown tumors and
    osteolysis of the distal phalanges. ... </think>
<search> symptoms of finger pain, small palpable tumors on the fingers,
    thin bones with brown tumors and osteolysis of the distal
    phalanges</search>

<information>
Doc 1(Title: "Osteochondroma") Signs and symptoms
Doc 2(Title: "Brown tumour in phalanx of the ring finger ...") Brown
    tumours are highly vascular lytic bone lesions found in primary and
    secondary hyperparathyroidism. ...
Doc 3(Title: "Benign tumor") Signs and symptoms
</information>

...
<think> ... The presence of brown tumors and osteolysis of the distal
    phalanges is characteristic of this condition. </think>
<answer> Hyperparathyroidism </answer>
```

Table 17: Case study 8: BRIDGE searches for haloperidol-induced dystonia and selects benztropine; VT-Search uses a broader query and selects baclofen.

Question (simplified): A 31-year-old man develops episodic neck stiffness and pain after six doses of haloperidol over two days for substance-induced psychosis. Episodes last about 25 minutes, with his neck rotating right; touching his jaw relieves the stiffness. Temperature is 37.3°C, pulse 108/min, and blood pressure 128/86 mm Hg. His neck is flexed, rigid, and rotated right with limited movement. What is the most appropriate therapy?
Options: A: Benztropine; B: Physical therapy; C: Dantrolene; D: Baclofen.
Ground Truth: A: Benztropine.

VT-Search (wrong)

```
<think> ... Given these symptoms, it seems that the patient is
    experiencing neck muscle spasms or stiffness, which could be related
    to his substance-induced psychosis or the medications he is taking,
    such as haloperidol. ...
<search> What is the most appropriate therapy for neck muscle spasms or
    stiffness in patients with substance-induced psychosis or those
    taking haloperidol? </search>

<information>
Doc 1(Title: "[On a muscular dystony case from manganism. ..."") It is
    described a muscular dystony case of manganism, treated with good
    results in two following phases with haloperidol and diazepam and
    then with L-Dopa + benserazide. ...
Doc 2(Title: "Is haloperidol or a benzodiazepine ..."") ... We conclude
    that haloperidol should be considered the first line drug for
    agitated patients post cardiac surgery, ...
Doc 3(Title: "InternalMed_Harrison") Optimal treatment of stiff-person
    syndrome requires therapy of the underlying tumor, glucocorticoids,
    and symptomatic use of drugs that enhance GABA-ergic transmission (
    diazepam, baclofen, sodium valproate, tiagabine, vigabatrin). ...
</information>

... Between the options, dantrolene and baclofen are both muscle
    relaxants, but baclofen is more commonly used for muscle spasms and
    stiffness.
...
<answer> Baclofen </answer>
```

BRIDGE (correct)

```
<think> ... Dystonia can be a side effect of antipsychotic medications
    like haloperidol. ...
<search> treatment for neck dystonia caused by haloperidol </search>

<information>
Doc 1(Title: "A patient with haloperidol induced laryngeal dystonia.")
    ... It can occur after a short treatment of a patient with atypical
    antipschycotics. ...
Doc 2(Title: "Haloperidol-Induced Dystonia ..."") ... Antipsychotic agents
     sometimes cause acute
    dystonia that can rapidly worsen within a few hours or days. ...
Doc 3(Title: "Treatment of dystonia and tics.") ... In dystonia, while
    oral agents such as benzodiazepines, baclofen and anticholinergics
    remain in use, botulinum toxin (BoNT) continues to be regarded as the
     treatment of choice for focal and segmental dystonia, ...
</information>

... the most appropriate therapy for this patient's symptoms is
    Benztropine, which is an anticholinergic medication used to treat
    dystonia caused by antipsychotic medications like haloperidol.
<answer> Benztropine </answer>
```

Table 18: Case study 9: both methods retrieve the same passages in different orders; BRIDGE selects reduced leverage, while VT-Search acknowledges the rotation-center evidence but chooses the cingulum distractor.

Question: A properly designed rest on lingual surface of a canine is preferred to a rest on incisal surface because:
Options: A: Less Leverage is exerted against the tooth by Lingual rest; B: Enamel is thicker on lingual surface; C: Visibility and access is better on lingual surface; D: The cingulum of canine produces a naturalsurface for recess.
Ground Truth: A: Less Leverage is exerted against the tooth by Lingual rest.

VT-Search (wrong)

```
<think> ... I will search for information to help me determine the
    correct answer.
<search> A properly designed rest on lingual surface of a canine is
    preferred to a rest on incisal surface </search>

<information>
Doc 1(Title: "The marginal-ridge rest seat.") ... preparing the rest seat
     on the lingual surface of a mandibular canine risks perforating the
    enamel. ...
Doc 2(Title: "Evaluation of prefabricated lingual rest seats for
    removable partial dentures.") ... A prefabricated lingual rest seat
    was designed and constructed to closely approximate the lingual
    contours of anterior teeth. ...
Doc 3(Title: "[Evaluation of stresses in different cingulum rest seats
    ].") ... Lingual rest seats are preferred since they are more
    acceptable esthetically and located closer to the centre of rotation
    of the tooth. ...
</information>

... The lingual surface is preferred because it is closer to the center
    of rotation of the tooth and provides a more uniform stress
    distribution.
...
<answer> D </answer>
```

BRIDGE (correct)

```
<think> ... I'll consider the biomechanical, anatomical, and aesthetic
    factors involved in dental restorations. </think>
<search> Why is a properly designed rest on the lingual surface of a
    canine preferred to a rest on the incisal surface? </search>

<information>
Doc 1(Title: "The marginal-ridge rest seat.") ... preparing the rest seat
     on the lingual surface of a mandibular canine risks perforating the
    enamel. ...
Doc 2(Title: "[Evaluation of stresses in different cingulum rest seats
    ].") ... Lingual rest seats are preferred since they are more
    acceptable esthetically and located closer to the centre of rotation
    of the tooth. ...
Doc 3(Title: "Evaluation of prefabricated lingual rest seats for
    removable partial dentures.") ... A prefabricated lingual rest seat
    was designed and constructed to closely approximate the lingual
    contours of anterior teeth. ...
</information>

... This reduces the leverage exerted against the tooth by the rest.

<think> Based on the information gathered, the correct answer is:
A: Less Leverage is exerted against the tooth by Lingual rest
</think>
<answer> A </answer>
```