



RouteRAG: Efficient Retrieval-Augmented Generation from Text and Graph via Reinforcement Learning

Yucan Guo, Miao Su, Saiping Guan[†], Zihao Sun,
Xiaolong Jin[†], Jiafeng Guo, Xueqi Cheng

¹State Key Laboratory of AI Safety

²Institute of Computing Technology, Chinese Academy of Sciences

³University of Chinese Academy of Sciences

{guoyucan23z, sumiao22z, guansaiping, sunzihao18z, jinxiaolong, guojiafeng, cxq}@ict.ac.cn

Abstract

Retrieval-Augmented Generation (RAG) integrates non-parametric knowledge into Large Language Models (LLMs), typically from unstructured texts and structured graphs. While recent progress has advanced text-based RAG to multi-turn reasoning through Reinforcement Learning (RL), extending these advances to hybrid retrieval introduces additional challenges. Existing graph-based or hybrid systems typically depend on fixed or handcrafted retrieval pipelines, lacking the ability to integrate supplementary evidence as reasoning unfolds. Besides, while graph evidence provides relational structures crucial for multi-hop reasoning, it is substantially more expensive to retrieve. To address these limitations, we introduce RouteRAG, an RL-based framework that enables LLMs to perform multi-turn and adaptive graph-text hybrid RAG. RouteRAG jointly optimizes the entire generation process via RL, allowing the model to learn when to reason, what to retrieve from either texts or graphs, and when to produce final answers, all within a unified generation policy. To guide this learning process, we design a two-stage training framework that accounts for both task outcome and retrieval efficiency, enabling the model to exploit hybrid evidence while avoiding unnecessary retrieval overhead. Experimental results across five question answering benchmarks demonstrate that RouteRAG significantly outperforms existing RAG baselines, highlighting the benefits of end-to-end RL in supporting adaptive and efficient retrieval for complex reasoning.¹

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning, decision-making, and long-form generation (Zhao et al., 2023; Touvron et al., 2023; Team et al., 2024),

especially when further trained with Reinforcement Learning (RL) (OpenAI et al., 2023; DeepSeek-AI et al., 2025; Yang et al., 2025a). These abilities have enabled LLMs to follow complex instructions, emulate chain-of-thought reasoning, and solve complicated multi-hop questions (Zhou et al., 2023; Wei et al., 2022). However, the knowledge of LLMs remains static, bounded by the data available at pretraining time. As a result, LLMs often produce inaccurate or outdated outputs when faced with knowledge-intensive queries that require access to external or up-to-date information (Augenstein et al., 2024; Huang et al., 2025).

To overcome this limitation, Retrieval-Augmented Generation (RAG) has emerged as a core paradigm for enhancing LLMs with access to external knowledge sources (Lewis et al., 2020; Gao et al., 2023). By conducting retrieval at inference time, RAG decouples knowledge from model parameters, improving factual accuracy and generalization. Early RAG systems typically perform a single round of retrieval before generation (Guu et al., 2020; Wang et al., 2023). Recent work has shown the benefits of multi-turn retrieval, where the model interleaves retrieval and reasoning over multiple steps (Yao et al., 2023; Trivedi et al., 2023; Li et al., 2025). However, these prompt-based approaches often depend on large closed-source models with strong intrinsic reasoning and planning skills. Smaller open-source models struggle to determine when to retrieve, how to formulate retrieval queries, and how to analyze retrieved evidence. This gap has motivated a new line of research (Jin et al., 2025; Song et al., 2025) that employs RL to explicitly train models to make retrieval and reasoning decisions. By optimizing a learned policy over interleaved thinking and retrieval actions, these RL-based methods aim to equip models with adaptive, context-sensitive retrieval strategies that surpass static instructions.

In parallel, graph-based RAG systems (Edge

[†]Corresponding authors.

¹The code is publicly available at <https://github.com/YucanGuo/RouteRAG>.

et al., 2024; Gutiérrez et al., 2024, 2025) utilize structured knowledge graphs to integrate and reason over information scattered across multiple passages, thereby improving coverage of factual entities and relations. While graphs enable more accurate entity disambiguation and multi-hop path reasoning than text-only retrieval, retrieving and processing graph evidence is often more computationally expensive, especially in large-scale or dense graphs. Moreover, existing graph-based RAG systems typically operate in a one-shot retrieval setting, fetching graph evidence once before generation, and lack the ability to adaptively choose between graph and text retrieval based on the evolving information needs of the query. Consequently, the current architecture of graph-based RAG systems presents challenges in managing complex reasoning that necessitates multi-turn interactions, and can also lead to unnecessary retrieval overhead when the reasoning chain is too long.

We address these limitations with RouteRAG, an RL-based framework that enables LLMs to perform multi-turn and hybrid retrieval over both unstructured texts and structured knowledge graphs. Instead of passively executing preset instructions, RouteRAG actively orchestrates retrieval decisions, selecting when and where to access external knowledge. To overcome the challenges of managing complex reasoning and avoiding unnecessary retrieval overhead, RouteRAG learns to interleave reasoning, retrieval, and answer formulation through a unified generation policy, adapting its retrieval behavior to the evolving task context.

To enable RouteRAG to generate accurate answers while efficiently retrieving relevant knowledge, we adopt a two-stage Group Relative Policy Optimization (GRPO) (Shao et al., 2024) training framework. In the first stage, the model is rewarded solely for answer correctness, allowing it to acquire the core capability of generating accurate responses and establishing a solid starting point for further optimization. In the second stage, we introduce an additional efficiency reward that discourages unnecessary retrieval, guiding the model to strike a balance between accuracy and computational cost. With these designs, RouteRAG can achieve both high accuracy and retrieval efficiency in complex multi-hop reasoning tasks.

Our main contributions lie in three aspects:

- We propose RouteRAG, an RL-based framework for multi-turn and hybrid RAG. The

model learns a unified generation policy that interleaves reasoning, adaptive graph-text hybrid retrieval, and answer formulation through a two-stage training framework.

- We design a reward function that jointly optimizes answer accuracy and retrieval efficiency, encouraging the model to retrieve selectively and to reason effectively over retrieved evidence across multiple steps.
- Extensive experiments on five Question Answering (QA) benchmarks demonstrate that RouteRAG outperforms prior multi-turn and graph-based RAG systems significantly.

2 Related Work

2.1 RAG

RAG has become a key paradigm for enhancing LLMs with external knowledge, thus mitigating hallucination and improving factual grounding (Guu et al., 2020; Gao et al., 2023). Traditional RAG systems retrieve relevant text chunks from an external knowledge base according to the query, and then feed the query into the LLM together with those text chunks to generate a final answer (Lewis et al., 2020; Yu et al., 2022). Beyond such one-shot retrieve-then-generate pipelines, recent research has explored multi-turn retrieval to provide more fine-grained and incremental supplementation of external knowledge, interleaving reasoning with evidence acquisition. For instance, IRCot (Trivedi et al., 2023) shows that alternating chain-of-thought reasoning with retrieval improves the performance of LLM on knowledge-intensive multi-hop QA. Search-o1 (Li et al., 2025) further develops this line by introducing a reason-in-documents module to alleviate the issue of redundant information in retrieved documents.

As an alternative route for deep knowledge integration, graph-based RAG methods incorporate structured knowledge graphs to aggregate evidence across passages and to make relational connections explicit (Peng et al., 2024; Zhao et al., 2023). By exposing entities and relations directly, these methods are particularly effective for multi-hop questions that require linking facts across disparate documents (Edge et al., 2024; Gutiérrez et al., 2024, 2025). However, graph retrieval is often more computationally expensive than text retrieval, and existing methods commonly perform one-shot retrieval. Although recent work such as HybGRAG (Lee

et al., 2025b) demonstrates that multi-turn hybrid text-graph retrieval is feasible through predefined multi-step procedures, these methods rely on fixed heuristics rather than a learnable policy.

2.2 RL for LLM Reasoning

RL has played a central role in improving the reasoning capabilities of LLMs. RL from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) has established a standard paradigm, where a reward model trained from human preferences directs the optimization of policies (Lambert et al., 2025), allowing models to adhere to instructions and reason with greater accuracy. Proximal Policy Optimization (PPO) (Schulman et al., 2017) remains the predominant algorithm for achieving these goals. More recently, GRPO (Shao et al., 2024) has been proposed as a more efficient variant, which leverages group-wise relative rewards to stabilize training and reduce variance. Building on these advances, researchers have begun to apply RL directly to the training of multi-turn RAG systems (Jin et al., 2025; Song et al., 2025). For instance, Search-R1 (Jin et al., 2025) trains LLMs with RL to decide when and what to search in the middle of reasoning, using only outcome rewards. While this reward design effectively improves correctness, it does not explicitly address retrieval cost or efficiency.

3 RouteRAG

In this section, we present RouteRAG, an RL-based framework for multi-turn hybrid RAG. We first describe the multi-turn workflow and the mechanism for hybrid knowledge access (Section 3.1). Subsequently, we introduce our two-stage RL framework (Section 3.2), encompassing the formulation of outcome and efficiency rewards, alongside the GRPO-based training algorithm.

3.1 Overall Framework

We begin by outlining the overall architecture of RouteRAG. This framework integrates LLMs with external retrievers in a multi-turn reasoning loop, where special tokens from the reasoning process can trigger retrieval actions from text and graph knowledge sources. We describe the multi-turn reasoning and retrieval workflow in Section 3.1.1 and the hybrid knowledge access mechanisms of the external retriever in Section 3.1.2.

Algorithm 1 RouteRAG Framework

Require: Input query q , policy model π_θ , retriever \mathcal{R} , maximum step budget B .
Ensure: Final response y .

- 1: Initialize response $y \leftarrow \emptyset$, step count $b \leftarrow 0$
- 2: **while** $b < B$ **do**
- 3: Initialize current rollout $y_b \leftarrow \emptyset$
- 4: **while** True **do**
- 5: Sample next token $y' \sim \pi_\theta(\cdot | q, y + y_b)$
- 6: $y_b \leftarrow y_b + y'$
- 7: **if** $y' \in \{\langle \text{/search} \rangle, \langle \text{/answer} \rangle, \langle \text{eos} \rangle\}$ **then**
- 8: **break**
- 9: $y \leftarrow y + y_b$ \triangleright Combine rollout with history
- 10: **if** $\langle \text{search} \rangle \dots \langle \text{/search} \rangle$ detected in y_b **then**
- 11: **if** $[\text{passage}]$ in y_b **then** $m \leftarrow$ Passage
- 12: **if** $[\text{graph}]$ in y_b **then** $m \leftarrow$ Graph
- 13: **if** $[\text{passage}]$ and $[\text{graph}]$ in y_b **then**
- 14: $m \leftarrow$ Hybrid
- 15: Extract query $q' \leftarrow$ ParseQuery(y_b)
- 16: $d \leftarrow \mathcal{R}(q', m)$ \triangleright Retrieve documents according to the retrieval mode m
- 17: $y \leftarrow y + \langle \text{information} \rangle d \langle \text{/information} \rangle$
- 18: **else if** $\langle \text{answer} \rangle \dots \langle \text{/answer} \rangle$ detected in y_b **then**
- 19: **return** final response y
- 20: $b \leftarrow b + 1$
- 21: **return** final response y

3.1.1 Multi-Turn Reasoning and Hybrid Retrieval Workflow

We formulate multi-turn retrieval-augmented generation as a sequential decision-making process, as shown in Algorithm 1. Given an input query q , the policy model π_θ interacts with external knowledge sources over a sequence of steps $b = \{1, \dots, B\}$, where B is the maximum step budget. At each step, the policy model conditions on the query and the current context to generate an action token. The action space includes continuing internal reasoning, triggering a retrieval operation ($\langle \text{search} \rangle \dots \langle \text{/search} \rangle$), or producing a final answer ($\langle \text{answer} \rangle \dots \langle \text{/answer} \rangle$). The retrieval operation further specifies a retrieval mode $m \in \{\text{Passage}, \text{Graph}, \text{Hybrid}\}$ by special tokens $y \in \{[\text{passage}], [\text{graph}]\}$ and a sub-query q' , which are used to obtain documents d from the retriever $\mathcal{R}(q', m)$. The retrieved information is then appended to the context and becomes available for subsequent reasoning.

This workflow enables the model to progressively refine its knowledge state by deciding what to retrieve and when to retrieve it, conditioned on the evolving reasoning trajectory. Moreover, the explicit action space over different retrieval modes allows the model to adaptively choose among passage, graph, and hybrid retrieval, depending on the requirements of the query.

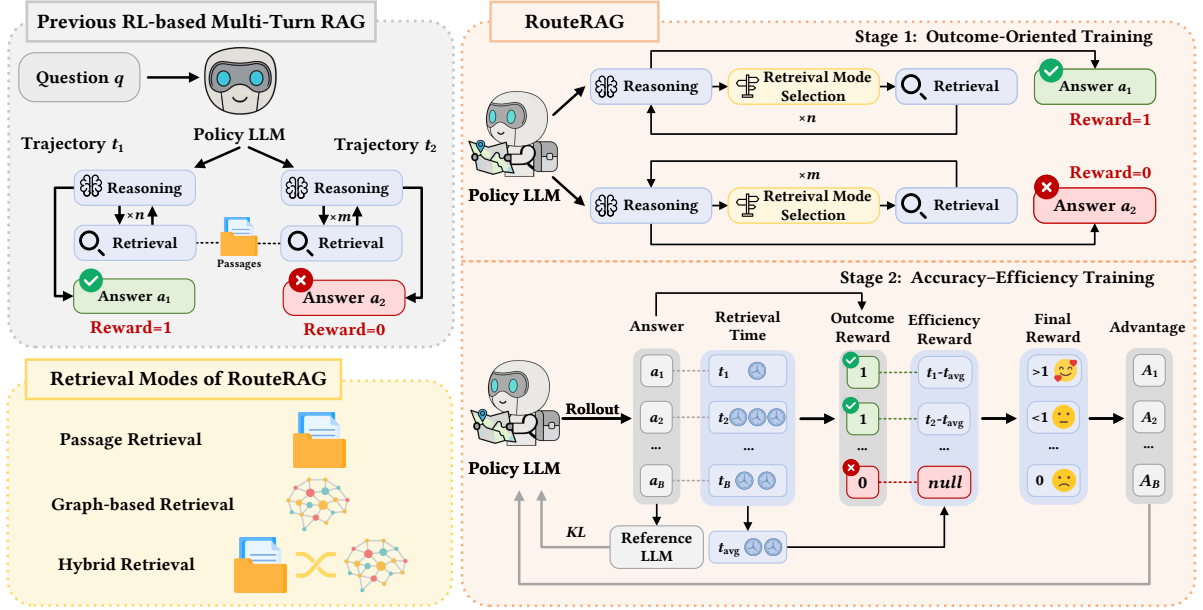


Figure 1: Previous RL-based multi-turn RAG vs. RouteRAG. Prior methods mainly focus on interleaving reasoning with passage retrieval and reward on answer correctness. RouteRAG extends retrieval to passage, graph, and hybrid modes, and is trained with a two-stage RL framework that optimizes both accuracy and efficiency.

3.1.2 Hybrid Knowledge Access

In RouteRAG, the retriever \mathcal{R} is responsible for providing external knowledge to support reasoning, with three different retrieval modes.

Passage Retrieval. The passage retriever is implemented with Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), which encodes both the sub-query and all passages in the corpus into a shared embedding space. Retrieval is performed by computing similarity scores between the query vector and passage vectors, and the top- k passages are selected as evidence.

Graph-based Retrieval. The graph retriever is implemented based on HippoRAG 2 (Gutiérrez et al., 2025), which first constructs a knowledge graph over passages. Given a sub-query, the retriever applies personalized PageRank over the graph to propagate relevance from query-linked nodes, thereby identifying passages that are related to the query through multi-hop connections.

Hybrid Retrieval. The hybrid retriever combines passage and graph retrieval using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Specifically, given two ranked lists, each document is assigned a fused score that decreases with its reciprocal rank in each list, which ensures that documents highly ranked by either retrieval mode are promoted in the

merged list. Formally, the fused score is defined as

$$\text{RRF}(d) = \sum_{m \in \{\text{Passage, Graph}\}} \frac{1}{k + \text{rank}_m(d)}, \quad (1)$$

where $\text{rank}_m(d)$ denotes the rank position of document d in retrieval mode m , and k is a smoothing hyperparameter. Documents are then re-ranked according to $\text{RRF}(d)$ to form the final hybrid list.

3.2 Two-Stage Reinforcement Learning

To optimize the unified generation policy, RouteRAG is trained with a two-stage RL framework based on GRPO. The motivation is to first ensure that the model acquires the basic ability to produce correct answers, and then to further refine its retrieval strategy to improve efficiency without sacrificing accuracy, as shown in Figure 1. In this section, we introduce the reward design that guides the learning objectives (Section 3.2.1) and the training algorithm that realizes the optimization procedure (Section 3.2.2).

3.2.1 Reward Design

RL optimization is fundamentally guided by the reward signal. To support the two-stage training, we devise different rewards for each stage, i.e., outcome-oriented reward and accuracy–efficiency reward.

Stage 1: Outcome-Oriented Reward. In the first stage, the reward is defined purely by the correct-

ness of the model output. Specifically, the reward is set to 1 if the generated answer y exactly matches the ground-truth label y^* , and 0 otherwise:

$$R_\phi(x, y) = \text{EM}(y, y^*). \quad (2)$$

Stage 2: Accuracy–Efficiency Reward. In the second stage, we extend the reward function to jointly optimize for correctness and retrieval efficiency. The reward is defined as

$$R_\phi(x, y) = \begin{cases} R_{\text{outcome}}, & R_{\text{outcome}} = 0 \\ R_{\text{outcome}} + R_{\text{efficiency}}, & R_{\text{outcome}} = 1 \end{cases}, \quad (3)$$

where $R_{\text{outcome}} \in \{0, 1\}$ denotes exact match accuracy. The efficiency reward $R_{\text{efficiency}}$ is computed from the total retrieval time across all reasoning steps, and only for those trajectories that correctly reach the answer. We apply a centered scaling by subtracting the average retrieval time t_{avg} , such that

$$R_{\text{efficiency}} = \frac{t_{\text{avg}} - t}{T}, \quad (4)$$

where t is the total retrieval time for the current trajectory, t_{avg} is the average retrieval time of the current batch, and T is a normalization constant ensuring the value of t and $t_{\text{avg}} \in [0, 0.5]$. This design provides positive reward for trajectories that achieve the correct answer at a pace exceeding the average, while imposing penalties on those that do not, thereby encouraging the model to retrieve more selectively without sacrificing answer quality.

3.2.2 Training Algorithm

We adopt GRPO (Shao et al., 2024; DeepSeek-AI et al., 2025) to train the unified generation policy π_θ over interleaved reasoning and retrieval actions. GRPO stabilizes learning by comparing trajectories within a group, thereby reducing variance in sparse-reward settings.

The policy model π_θ is optimized by maximizing the following objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim Q, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(Y|q)} \frac{1}{G} \sum_{i=1}^G \left[\min \left(r_i(\theta) A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i \right) - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta_{\text{old}}} \parallel \pi_\theta] \right], \quad (5)$$

where ϵ and β are hyperparameters, $\pi_{\theta_{\text{old}}}$ denotes the old policy, $r_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}$, A_i denotes the

group-relative advantage for the i -th trajectory, and the KL penalty $\mathbb{D}_{\text{KL}}[\pi_{\theta_{\text{old}}} \parallel \pi_\theta]$ regularizes the new policy against deviating excessively from the old policy. Further theoretical analysis on the effectiveness of the efficiency reward and GRPO is provided in Section A.

4 Experiments

4.1 Experimental Setting

Evaluation Datasets. Following (Gutiérrez et al., 2024) and Gutiérrez et al. (2025), we evaluate RouteRAG on five widely used benchmarks for simple and multi-hop QA, namely PopQA (Mallen et al., 2023), Natural Questions (NQ) (Kwiatkowski et al., 2019; Wang et al., 2024), HotpotQA (Yang et al., 2018), 2WikiMultihopQA (2Wiki) (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). PopQA is an open-domain QA dataset designed to evaluate factual recall over long-tail knowledge, and NQ contains naturally occurring queries paired with answers from Wikipedia. HotpotQA and 2Wiki focus on multi-hop reasoning across Wikipedia passages, while MuSiQue requires reasoning over compositional sub-questions.

Baselines. We compare RouteRAG against several types of representative approaches: (1) Vanilla RAG (Lewis et al., 2020), which performs single-shot dense passage retrieval and generation. (2) Multi-turn RAG methods, including Search-o1 (Li et al., 2025), Search-R1 (Jin et al., 2025), and R1-Searcher (Song et al., 2025), wherein the latter two methods utilize RL to enhance multi-turn passage RAG. (3) Graph-based RAG methods, including GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025), RAPTOR (Sarthi et al., 2024), HippoRAG (Gutiérrez et al., 2024), and HippoRAG 2 (Gutiérrez et al., 2025), which leverage structured knowledge graphs for retrieval.

Implementation Details. We conduct training using Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2025b) as the backbone models. The training data consists of 10k sampled queries from the HotpotQA training set (Yang et al., 2018), while the retrieval corpus is built from their associated documents. For retrieval, we adopt Contriever (Izacard et al., 2022) and NV-Embed-v2 (Lee et al., 2025a) as the dense retriever for 3B and 7B models, respectively. For baseline evaluations, text-based RAG systems are assessed under the same Qwen2.5 backbone, while graph-based RAG systems utilize the GPT-4o-mini backbone.




Method	Simple QA				Multi-hop QA						Average	
	PopQA		NQ		HotpotQA		2Wiki		MuSiQue		EM	F1
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1		
 <i>GPT-4o-mini</i>												
Direct Inference	16.1	22.7	35.2	52.7	28.6	41.0	30.2	36.3	11.2	22.0	24.3	34.9
<i>Graph-based RAG</i>												
GraphRAG	30.7	51.3	<u>38.0</u>	<u>55.5</u>	<u>51.4</u>	<u>67.6</u>	45.7	61.0	27.0	<u>42.0</u>	38.6	<u>55.5</u>
LightRAG	1.9	14.8	2.8	15.4	9.9	20.2	2.5	12.1	2.0	9.3	3.8	14.4
RAPTOR	<u>41.9</u>	55.1	37.8	54.5	50.6	64.7	39.7	48.4	<u>27.7</u>	39.2	39.5	52.4
HippoRAG	42.5	56.2	37.2	52.5	46.3	60.0	<u>59.4</u>	<u>67.3</u>	24.0	35.9	<u>41.9</u>	54.4
HippoRAG 2	41.7	<u>55.7</u>	43.4	60.0	56.3	71.1	60.5	69.7	35.0	49.3	47.4	61.2
 <i>Qwen2.5-3B</i>												
Vanilla RAG	30.3	41.6	18.1	31.8	29.5	41.8	19.7	27.4	10.3	17.5	21.6	32.0
<i>Graph-based RAG</i>												
HippoRAG 2	29.1	40.1	20.3	33.5	31.2	45.0	21.5	33.8	12.2	20.2	22.9	34.5
<i>Multi-turn RAG</i>												
Search-o1	17.1	23.8	19.9	29.1	18.7	26.3	16.9	20.9	3.9	10.5	15.3	22.1
Search-R1	<u>45.8</u>	<u>53.3</u>	46.2*	54.8*	<u>45.2*</u>	<u>56.9*</u>	<u>42.4</u>	<u>50.8</u>	<u>22.2</u>	<u>30.9</u>	<u>40.4</u>	<u>49.3</u>
RouteRAG-3B	49.4	56.8	<u>44.1</u>	<u>53.4</u>	53.2*	65.1*	57.5	64.1	30.7	39.3	47.0	55.7
 <i>Qwen2.5-7B</i>												
Vanilla RAG	26.3	37.8	11.0	30.6	29.1	47.5	17.2	33.3	11.4	23.0	19.0	34.4
<i>Graph-based RAG</i>												
HippoRAG 2	27.0	37.9	8.1	27.2	27.4	46.1	16.8	34.7	12.3	24.0	18.3	34.0
<i>Multi-turn RAG</i>												
Search-o1	4.7	7.2	18.1	27.5	13.5	19.1	6.4	7.9	2.9	7.7	9.1	13.9
R1-Searcher	28.4	41.0	41.6	52.2	46.6*	56.7*	41.7*	49.0*	29.3	37.6	37.5	47.3
Search-R1	51.3	57.1	56.8*	65.3*	<u>51.0*</u>	<u>62.0*</u>	<u>51.8</u>	<u>58.9</u>	<u>32.0</u>	<u>40.8</u>	<u>48.6</u>	<u>56.8</u>
RouteRAG-7B	<u>50.6</u>	<u>56.4</u>	<u>51.5</u>	<u>60.4</u>	60.8*	72.5*	57.1	64.6	39.6	49.3	51.9	60.6

Table 1: Main results on simple and multi-hop QA benchmarks. The best results within each backbone group are indicated in bold, while the underlined values represent the second-best results. * represents in-domain datasets.

In particular, HippoRAG 2 (Gutiérrez et al., 2025), the strongest graph-based baseline, is evaluated employing both Qwen2.5 and GPT-4o-mini backbones. We report Exact Match (EM) and F1 scores as evaluation metrics. Additional implementation details, including the training prompt template, hyperparameters, and training configuration, are provided in Section B.

4.2 Main Results

We conduct a comprehensive comparison of RouteRAG against all the baseline methods, as shown in Table 1. From the results, we make the following key observations:

(1) **RouteRAG substantially improves the performance of a small backbone, especially on multi-hop QA.** Graph-based methods such as Hip-

poRAG 2 perform well with the strong GPT-4o-mini backbone but drop sharply with the smaller Qwen2.5-3B and Qwen2.5-7B models, indicating that small LLMs struggle to handle complex reasoning chains. In contrast, RouteRAG achieves much better performance on small backbones by jointly learning reasoning, retrieval, and answer generation within a unified policy model.

(2) **RouteRAG approaches GPT-4o-mini-based graph-based RAG systems despite using a much smaller model.** Despite the large performance gap usually observed between GPT-4o-mini and Qwen2.5-3B/7B, RouteRAG narrows this gap substantially and even surpasses several graph-based systems built on GPT-4o-mini. This suggests that improving the policy can be as impactful as scaling up the backbone itself.

Method	PopQA		NQ		HotpotQA		2Wiki		MuSiQue		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
RouteRAG-3B	49.4	56.8	44.1	53.4	53.2	<u>65.1</u>	57.5	<u>64.1</u>	30.7	39.3	47.0	55.7
<i>w/o</i> Stage 2 training	47.0	54.6	39.1	49.4	46.4	58.6	48.5	55.6	24.3	33.0	41.1	50.2
<i>w/o</i> training	26.2	41.3	18.6	30.5	35.4	46.9	24.4	37.9	15.4	24.4	24.0	36.2
<i>w/ only</i> passage retrieval	48.7	55.2	43.9	<u>53.3</u>	53.1	64.5	53.0	58.9	28.0	36.2	45.3	53.6
<i>w/ only</i> graph-based retrieval	<u>49.3</u>	56.8	43.8	53.0	53.4	65.5	<u>57.4</u>	64.2	<u>29.7</u>	<u>38.3</u>	<u>46.7</u>	<u>55.6</u>
<i>w/ only</i> hybrid retrieval	48.7	<u>55.9</u>	43.1	52.5	<u>53.3</u>	<u>65.1</u>	53.7	59.8	28.8	37.0	45.5	54.0

Table 2: Ablation studies on RL training and hybrid retrieval. “*w/o* training” denotes the base model without any RL training, and “*w/o* stage 2 training” denotes training only with the first stage. For hybrid retrieval ablation, we compare RouteRAG with variants restricted to only passage retrieval, only graph retrieval, or only hybrid retrieval.

Method	PopQA		NQ		HotpotQA		2Wiki		MuSiQue		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
RouteRAG-3B	49.4	56.8	44.1	53.4	53.2	<u>65.1</u>	57.5	<u>64.1</u>	30.7	39.3	47.0	55.7
<i>w/o</i> efficiency reward	<u>41.5</u>	<u>54.1</u>	<u>41.1</u>	<u>51.7</u>	53.7	66.2	<u>56.9</u>	65.0	<u>30.2</u>	<u>38.9</u>	<u>44.7</u>	<u>55.2</u>
RouteRAG-7B	50.6	56.4	<u>51.5</u>	<u>60.4</u>	60.8	<u>72.5</u>	57.1	64.6	<u>39.6</u>	<u>49.3</u>	51.9	60.6
<i>w/o</i> efficiency reward	<u>46.0</u>	<u>54.1</u>	52.5	62.1	<u>59.9</u>	72.6	<u>49.5</u>	<u>56.8</u>	42.4	52.0	<u>50.1</u>	<u>59.5</u>

Table 3: Ablation study on efficiency reward. “*w/o* efficiency reward” denotes training only with EM-based outcome rewards.

(3) RouteRAG outperforms the strongest RL-trained multi-turn baseline with much smaller training cost. Search-R1, the prior strongest RL-based multi-turn system, is trained on 170k questions from NQ and HotpotQA. Despite being trained on a mere 10k HotpotQA instances, RouteRAG achieves the best average scores among all 3B and 7B methods, demonstrating that structured retrieval and retrieval mode selection can yield more effective and sample-efficient multi-turn RAG policies than scaling training data alone. While RouteRAG is slightly weaker on simple QA, this is expected because the training data is dominated by multi-hop questions. Nevertheless, its overall performance remains competitive given the smaller training cost.

4.3 Ablation Study

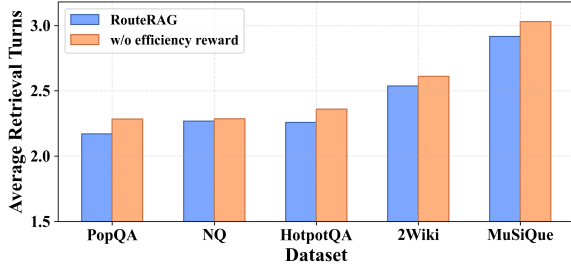
In this section, we conduct ablation experiments to validate the effectiveness of RL training, hybrid retrieval, and the efficiency reward.

RL Training. The upper part of Table 2 compares the full model against two ablated variants: (1) a model trained only with the first stage, and (2) the untrained backbone. The results show that RouteRAG already acquires a strong reasoning ability through the Stage 1 training, outperforming the

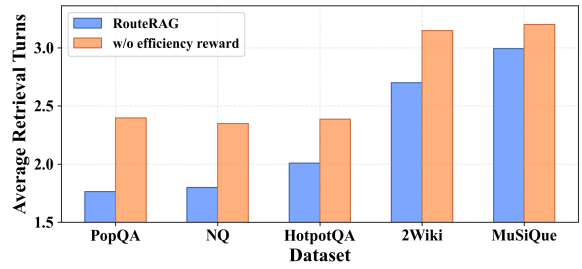
untrained backbone by a large margin. The Stage 2 training further improves the performance of RouteRAG, especially on multi-hop QA datasets.

Hybrid Retrieval. The lower part of Table 2 presents an ablation study comparing RouteRAG with variants restricted to a single retrieval mode. Passage retrieval performs well on simple QA, while graph retrieval is more effective for multi-hop reasoning, confirming their complementary strengths. The full RouteRAG achieves the highest average accuracy by dynamically selecting among the three modes, showing that adaptive retrieval yields stronger robustness and generalization than any fixed strategy.

Efficiency Reward. We evaluate whether RouteRAG learns to retrieve evidence more efficiently while keeping effectiveness with the efficiency reward by comparing it to a variant trained only with the outcome reward, using the same number of training steps. As shown in Table 3, RouteRAG maintains comparable or even higher accuracy than its outcome rewards-only counterpart. Figure 2 shows that both RouteRAG-3B and RouteRAG-7B consistently reduce the average retrieval turns across all datasets, with larger savings observed on the 7B model. Table 4 further assesses



(a) RouteRAG-3B.



(b) RouteRAG-7B.

Figure 2: Comparing the average retrieval turns of RouteRAG and its variant without efficiency reward.

Table 4: Average retrieval turns and accuracy of RouteRAG.

Method	Avg. Turns	Avg. F1
RouteRAG-3B	2.43	55.7
w/o efficiency reward	2.51 (+3.3%)	55.2 (-0.5)
RouteRAG-7B	2.25	60.6
w/o efficiency reward	2.70 (+20%)	59.5 (-1.1)

the trade-off between retrieval turns and task accuracy, highlighting that efficiency gains do not come at the cost of answer quality. This demonstrates that incorporating efficiency rewards encourages the policy to avoid unnecessary retrieval steps while still collecting sufficient evidence. A comprehensive analysis of this phenomenon is available in Section A.4.

4.4 Analysis on Reasoning Ability

To better understand how the multi-round reasoning ability of RouteRAG evolves during training, we analyze the average number of reasoning steps before and after training. As shown in Figure 3, both the 3B and 7B models exhibit increased reasoning depth after training. Compared with RouteRAG-3B, RouteRAG-7B shows a more pronounced improvement, with larger gains on the more complex MuSiQue and 2Wiki datasets, and smaller increases on PopQA and NQ. Figure 4 further expands the analysis by jointly examining performance, response token length, and reasoning turns across datasets. The results show that RouteRAG consistently achieves a more favorable balance among the three factors compared with Search-R1 and R1-Searcher, achieving high F1 with moderate token length. For a qualitative view of how the reasoning behavior changed in practice, case studies comparing model outputs before and after training are presented in Section D.2.

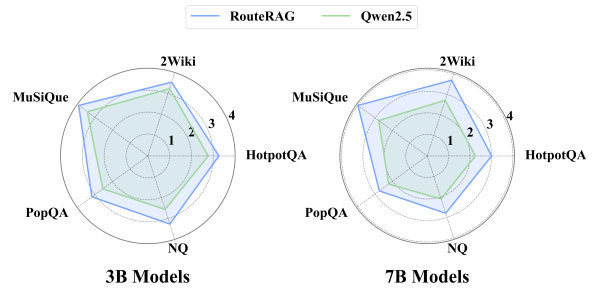


Figure 3: Comparison of average reasoning steps.

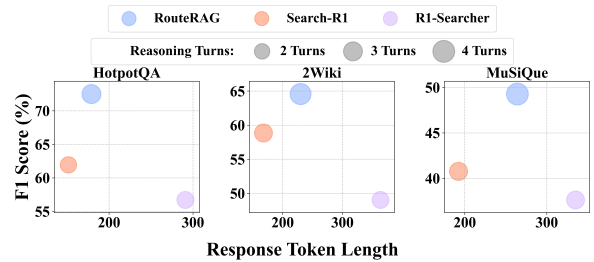


Figure 4: Comparison in terms of performance, response token length, and reasoning turns.

4.5 Analysis on Dense Retrievers

Table 5 shows the performance of RouteRAG and Search-R1 with two different dense retrievers, i.e., Contriever (Izacard et al., 2022) and NV-Embed-v2 (Lee et al., 2025a). Across both 3B and 7B model sizes, RouteRAG yields consistent gains on multi-hop QA benchmarks, regardless of which dense retriever is used. In contrast, Search-R1 is more sensitive to the quality of the dense retriever, benefiting substantially from the stronger NV-Embed-v2 retriever on simple QA tasks. This highlights that RouteRAG relies less heavily on the dense retriever, since its adaptive use of graph and hybrid retrieval provides accurate evidence, thereby stabilizing performance even when the dense retriever is relatively weak.



Method	Simple QA				Multi-hop QA						Average	
	PopQA		NQ		HotpotQA		2Wiki		MuSiQue		EM	F1
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1		
 <i>Qwen2.5-3B</i>												
<i>Contriever</i>												
Search-R1	45.8	53.3	46.2*	54.8*	45.2*	56.9*	42.4	50.8	22.2	30.9	40.4	49.3
RouteRAG	49.4	56.8	44.1	53.4	<u>53.2*</u>	<u>65.1*</u>	<u>57.5</u>	<u>64.1</u>	30.7	39.3	47.0	55.7
<i>NV-Embed-v2</i>												
Search-R1	53.5	60.1	54.8*	63.1*	51.0*	62.5*	45.7	54.5	<u>32.1</u>	<u>41.0</u>	<u>47.4</u>	<u>56.2</u>
RouteRAG	<u>53.0</u>	<u>59.6</u>	<u>47.0</u>	<u>57.3</u>	56.1*	68.2*	58.2	65.4	36.6	45.4	50.2	59.2
 <i>Qwen2.5-7B</i>												
<i>Contriever</i>												
Search-R1	47.1	52.7	49.5*	58.1*	45.2*	55.7*	45.9	52.7	24.4	31.8	42.4	50.2
RouteRAG	50.5	<u>56.7</u>	46.9	55.4	<u>53.8*</u>	<u>65.3*</u>	<u>55.3</u>	<u>62.7</u>	<u>34.3</u>	<u>43.7</u>	48.2	<u>56.8</u>
<i>NV-Embed-v2</i>												
Search-R1	51.3	57.1	56.8*	65.3*	51.0*	62.0*	51.8	58.9	32.0	40.8	<u>48.6</u>	<u>56.8</u>
RouteRAG	<u>50.6</u>	56.4	<u>51.5</u>	<u>60.4</u>	60.8*	72.5*	57.1	64.6	39.6	49.3	51.9	60.6

Table 5: Performance comparison among RouteRAG and Search-R1 with different dense retrievers. * represents in-domain datasets.

5 Conclusions

In this paper, we presented RouteRAG, an RL framework for efficient multi-turn hybrid RAG. Unlike prior multi-turn RAG systems that rely on static prompting or single-mode retrieval, our approach learns a unified policy that interleaves reasoning, retrieval mode selection, retrieval query generation, and answer generation. Our two-stage training framework further ensures that the model first acquires robust answer correctness and then improves retrieval efficiency without sacrificing accuracy. Experiments conducted on five knowledge-intensive QA benchmarks demonstrate that RouteRAG significantly outperforms existing graph-based and multi-turn RAG systems, highlighting that efficiency gains can be achieved without compromising answer quality.

Limitations

Despite the strong empirical results, our proposed RouteRAG has several limitations. First, due to computational constraints, we conduct RL training and evaluation only on 3B and 7B LLMs. Larger or more diverse model architectures may exhibit different behaviors under our training framework. Second, our experiments utilize HippoRAG 2 as the graph retriever. While we adopt it because it is

currently the strongest graph-based RAG system, this choice limits our evaluation of how RouteRAG interacts with alternative graph retrievers or graph construction pipelines.

Acknowledgments

This work is partially supported by the Strategic Priority Research Program of the CAS under Grant No. XDB0680102, the CAS Special Research Assistant Fund, the National Natural Science Foundation of China under Grants No. 62306299, 62441229, and 62406308, and the Innovation Funding of ICT, CAS under Grant No. E561010.

References

- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. [Factuality challenges in the era of large language models and opportunities for fact-checking](#). *Nature Machine Intelligence*, 6(8):852–863.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). *Advances in neural information processing systems*, 30.

- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. 2024. [From local to global: A graph rag approach to query-focused summarization](#). *arXiv preprint arXiv:2404.16130*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*, 2(1).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. [LightRAG: Simple and fast retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761, Suzhou, China. Association for Computational Linguistics.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: neurobiologically inspired long-term memory for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 59532–59569.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. [From RAG to memory: Non-parametric continual learning for large language models](#). In *Forty-second International Conference on Machine Learning*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. [Retrieval augmented language model pre-training](#). In *International conference on machine learning*, pages 3929–3938. PMLR.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*, 43(2):1–55.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *arXiv preprint arXiv:2503.09516*.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [RewardBench: Evaluating reward models for language modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. [NV-embed: Improved techniques for training LLMs as generalist embedding models](#). In *The Thirteenth International Conference on Learning Representations*.

- Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025b. Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 879–893.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-ol: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. [A survey of knowledge-enhanced text generation](#). *ACM Computing Surveys*, 54(11s):1–38.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*, 1(2).
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *arXiv preprint arXiv:2311.07911*.

A Analysis on Efficiency Reward

In this section, we provide a detailed theoretical analysis of why the efficiency reward designed in RouteRAG improves selective retrieval in GRPO-based training.

A.1 Batch-Level Efficiency Reward

Let τ_i denote the i -th trajectory in a group of G trajectories sampled from a batch of size B . The total reward for trajectory τ_i is

$$R_\phi(\tau_i) = \begin{cases} R_{\text{outcome}}(\tau_i), & R_{\text{outcome}}(\tau_i) = 0 \\ R_{\text{outcome}}(\tau_i) + R_{\text{efficiency}}(\tau_i), & R_{\text{outcome}}(\tau_i) = 1 \end{cases}, \quad (6)$$

where $R_{\text{outcome}}(\tau_i) \in \{0, 1\}$ indicates the correctness of the answer. The efficiency reward is centered on the batch-level average retrieval time rather than the group-level average:

$$R_{\text{efficiency}}(\tau_i) = \frac{t_{\text{avg}} - t_i}{T}, \quad t_{\text{avg}} = \frac{1}{B} \sum_{\tau \in \text{batch}} t_\tau, \quad (7)$$

where t_i is the total retrieval time of trajectory τ_i , and T is a normalization constant.

There are three main reasons for choosing batch-level efficiency reward together with GRPO-based training, instead of using group-level efficiency:

- **Variance reduction and stability.** Batch-level averaging reduces the impact of noisy fluctuations in retrieval time (e.g., hardware latency or network delays).
- **Mitigating anomalies across queries.** Although batch-level normalization may produce unusually high or low raw efficiency rewards for certain queries, GRPOs group-relative advantage compensates for this effect.
- **Encouraging selective retrieval.** Combining batch-level centering with GRPO advantage ensures that trajectories with unnecessary retrieval are penalized, while efficient yet accurate trajectories are favored.

A.2 Variance Reduction and Stability

Each GRPO group may contain only a few trajectories (e.g., $G = 5$). Raw retrieval times t_i can fluctuate due to hardware noise, network latency, or retriever stochasticity. If group-level averaging were used, such fluctuations could lead to unstable rewards. By computing t_{avg} across the entire batch, we obtain a more stable reference signal that smooths out these random variations.

This reduces the variance of the group-relative advantage, which in turn stabilizes policy gradient updates.

A.3 Mitigating Anomalies Across Queries

Batch-level normalization may produce unusually high or low raw efficiency rewards for certain queries. For example, a simple question requiring little retrieval could yield a disproportionately large positive $R_{\text{efficiency}}(\tau_i)$. This is indeed a potential drawback of using batch-level efficiency reward.

However, GRPO compensates for this issue through its group-relative formulation. Although $R_{\text{efficiency}}(\tau_i)$ is normalized at the batch level, the advantage A_i is computed relative to the group mean reward within each GRPO group, i.e.,

$$A_i = \frac{R_\phi(\tau_i) - \frac{1}{G} \sum_{j=1}^G R_\phi(\tau_j)}{\text{std}(\{R_\phi(\tau_j)\}_{j=1}^G)}. \quad (8)$$

Even if a particular query obtains an abnormally large batch-normalized efficiency reward, its influence on learning is moderated by this group-relative centering. As a result, the combination of batch-level normalization and group-level centering ensures that the learning signal remains consistent across diverse query types.

A.4 Encouraging Selective Retrieval

For trajectories that correctly answer the query ($R_{\text{outcome}} = 1$), the numerator of A_i can be decomposed into outcome and efficiency components, yielding

$$A_i = \frac{(1 - \overline{R_{\text{outcome}}}) + (R_{\text{efficiency}}(\tau_i) - \overline{R_{\text{efficiency}}})}{\text{std}(\{R_\phi(\tau_j)\}_{j=1}^G)}, \quad (9)$$

where $\overline{R_{\text{outcome}}}$ and $\overline{R_{\text{efficiency}}}$ are group means computed within the GRPO group, while $R_{\text{efficiency}}(\tau_i)$ itself is computed using the batch-level reference t_{avg} .

From Equation (9) we see:

- If a trajectory answers correctly and its retrieval time is less than the group average, then $R_{\text{efficiency}}(\tau_i) - \overline{R_{\text{efficiency}}} > 0$, and consequently, the numerator increases, thereby rewarding the policy for selective retrieval.
- If a trajectory performs redundant retrieval, the efficiency term is negative and reduces A_i , discouraging unnecessary retrieval.

Thus, GRPO guides the policy towards trajectories that balance correctness with efficient retrieval.

B Implementation Details

B.1 Evaluation Datasets and Baselines

Evaluation Datasets. The statistics of evaluation datasets are shown in Table 6.

Table 6: Dataset statistics

Dataset	# of queries	# of passages
PopQA	1,000	8,678
NQ	1,000	9,633
HotpotQA	1,000	9,811
2Wiki	1,000	6,119
MuSiQue	1,000	11,656

Baselines.

- **Vanilla RAG** (Lewis et al., 2020) is a standard RAG framework that combines an LLM with a learned retriever to condition generation on retrieved documents.
- **Search-o1** (Li et al., 2025) is an agentic retrieval system that performs on-demand multi-step search with a “reason-in-documents” refinement module.
- **Search-R1** (Jin et al., 2025) is an RL-trained model that interleaves stepwise reasoning with autonomous search actions.
- **R1-Searcher** (Song et al., 2025) is an RL approach that teaches LLMs when and how to invoke external search for improved reasoning.
- **GraphRAG** (Edge et al., 2024) is a graph-based RAG method that aggregates evidence via local-to-global graph traversal.
- **LightRAG** (Guo et al., 2025) is a lightweight and efficient graph-based RAG method with simplified indexing and fast dual-level retrieval.
- **RAPTOR** (Sarathi et al., 2024) is a hierarchical, tree-structured RAG approach that uses recursive summarization to access multi-level abstractions.
- **HippoRAG** (Gutiérrez et al., 2024) is a memory-inspired RAG framework that builds a knowledge graph and retrieves via personalized PageRank.
- **HippoRAG 2** (Gutiérrez et al., 2025) is an enhanced graph-based RAG framework that extends HippoRAG with deeper passage integration.

B.2 Training Prompt Template

The training prompt template for the policy LLM is shown in Table 7.

B.3 Training Details

Hyperparameters. For GRPO training of RouteRAG, we set the policy LLM learning rate to 1×10^{-6} , a total batch size of 256, with a mini-batch size of 128 and a micro-batch size of 32. The KL divergence regularization coefficient β is set to 0.001, and the clip ratio ϵ is set to 0.2. The retrieval budget is fixed at $B = 4$, and the number of retrieved passages per call is $k = 3$. The maximum sequence length is set to 4,096 tokens, with a maximum response length of 500 tokens, a maximum start length of 2,048 tokens, and a maximum observation length of 500 tokens.

Open Information Extraction Models We employ Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Grattafiori et al., 2024) for Open Information Extraction (OpenIE) during graph construction. Following HippoRAG 2, the model is used to extract entities and relational triplets from corpus passages, which are then used to build the knowledge graph that supports graph and hybrid retrieval in both HippoRAG 2 and RouteRAG. For experiments based on 3B models, we use Llama-3.1-8B-Instruct as a more lightweight alternative to the larger OpenIE models (Llama-3.3-70B-Instruct and GPT-4o-mini) adopted in HippoRAG 2. For experiments involving 7B models, we adopt the same Llama-3.3-70B-Instruct model used in HippoRAG 2 to construct higher-quality graphs.

Training Configuration. Our training framework is adapted from the Search-R1 training framework (Jin et al., 2025), which builds upon the verl (Sheng et al., 2025). Training is conducted on a single node with $8 \times 80\text{GB}$ NVIDIA A100 GPUs. To improve memory efficiency, we enable gradient checkpointing and apply Fully Sharded Data Parallel (FSDP) with CPU offloading for parameters, gradients, and optimizer states. Rollouts are sampled with vLLM (Kwon et al., 2023) using a tensor parallel size of 1 and a GPU memory utilization ratio of 0.6. The rollout sampling temperature is set to 1.0.

Table 7: Training prompt template for multi-turn reasoning and retrieval.

Training Prompt Template for the Policy LLM

Answer the given question. You must conduct reasoning inside `<think>` and `</think>` first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine using the following strict format:

- You MUST first decide which retrieval mode to use (both modes will return relevant documents, but use different retrieval methods):
 - Use `[passage]` to find documents using semantic similarity-based dense retrieval
 - Use `[graph]` to find documents through graph-based retrieval, which performs retrieval on a structured knowledge graph constructed from documents using fact ranking and graph reasoning
 - You can combine them as `[graph][passage]` to get documents from both retrieval methods
- Then formulate your specific search query based on what information you need
- Finally, wrap everything in `<search>` and `</search>` tags

For example:

- Using dense retrieval: `<search> [passage] the capital of France </search>`
- Using graph-based retrieval: `<search> [graph] the capital of France </search>`
- Using both methods: `<search> [graph][passage] the capital of France </search>`

The search results (relevant documents) will be returned between `<information>` and `</information>` tags. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside `<answer>` and `</answer>`, without detailed illustrations. For example, `<answer> Paris </answer>`.
 Question: {question}

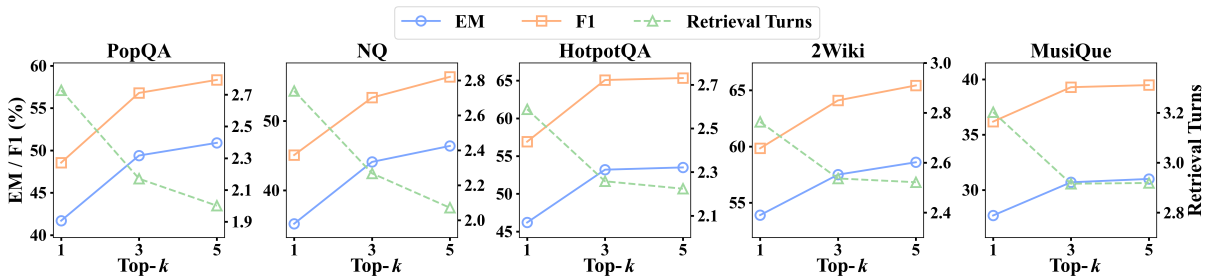


Figure 5: Performance of RouteRAG-3B with different numbers of retrieved documents.

Two-Stage Training. Stage 1 is trained for 20 steps (0.5 epoch) with EM-based rewards only, ensuring correctness. Stage 2 continues for an additional 20 steps (0.5 epoch) with the accuracy–efficiency reward introduced in Section 3.2. In both stages, we sample five responses per prompt during training to compute group-relative advantages. Checkpoints are saved every 10 steps, and the final checkpoint is used for evaluation.

C Findings on Learned Routing Behavior

We summarize key findings on the learned routing behavior to better understand how the model selects retrieval modes.

(1) Suppression of weak retrieval modes. Dur-

ing training, the policy naturally down-weights retrieval modes that provide limited marginal benefit. In our setup, standalone passage retrieval is often less informative than graph or hybrid exploration, and the learned policy concentrates decisions on graph-based and hybrid actions. This supports the motivation for learning routing decisions rather than relying on fixed pipelines.

(2) Adaptation to task structure. The learned routing strategy varies with dataset characteristics. On multi-hop datasets, the policy performs more retrieval rounds and relies more heavily on hybrid exploration, whereas on simpler factoid datasets, it terminates earlier and uses fewer retrieval steps. This indicates that the model learns to adjust explo-

ration depth to query complexity.

(3) More refined strategies with larger backbones. We observe that smaller models primarily optimize the number of retrieval steps, while the larger model learns a more nuanced strategy that jointly optimizes both the number of steps and the choice of retrieval modes, preferring lower-cost modes for simpler queries.

D Additional Experiments

D.1 Analysis on Number of Retrieved Documents

In this experiment, we analyze the performance of RouteRAG-3B with varying numbers of retrieved documents. As shown in Figure 5, the increasing k generally improves both EM and F1 across all datasets, though the gains diminish beyond $k = 3$. Interestingly, retrieval turns consistently decrease as k grows, indicating that providing the model with richer evidence reduces the need for iterative retrieval. These results demonstrate that a moderate retrieval breadth yields a balance between accuracy and retrieval efficiency.

D.2 Case Study

To further illustrate how our training framework improves model behavior, we present several qualitative case studies comparing the outputs of the base model before training and our proposed RouteRAG after training, as shown in Tables 8–9. These cases reveal several representative issues of the base model and how our training addresses them.

(1) Over-reliance on internal knowledge without validation. The base model tended to rely on its parametric knowledge and give confident but unsupported answers, skipping the step of verifying correctness against external evidence (Case 1). Our model, by contrast, learns to cross-check candidate answers with retrieved documents before finalizing its response, which reduces hallucinations and increases factual accuracy.

(2) Failure to understand or decompose complex questions. The base model frequently struggled with multi-hop or compositional questions, failing to recognize their underlying structure or to plan effective retrieval steps (Case 2). Through training, our model acquires the ability to analyze the question, break it down into subproblems, and issue precise retrieval queries for each part, which leads to more accurate and complete answers.

Overall, after training, RouteRAG shows clear

improvements in its ability to understand complex questions, decompose them into manageable subproblems, and construct concise yet targeted retrieval queries that lead to well-grounded and correct answers.

Table 8: **Case study 1.** Before training, the model hallucinates incorrect internal knowledge, thus wrongly claiming that Johnny Pemberton played Bo Thompson in “That ’70s Show” and that it was created by Steven Molaro. After training, it avoids hallucinations, correctly decomposes the question into sub-steps, and retrieves the right evidence to identify the creator as Justin Spitzer.

Question: Who created the NBC sitcom that Johnny Pemberton appears in as the character Bo Thompson?

Ground Truth: Justin Spitzer

Before Training (Qwen2.5-3B-Instruct)

<think> To answer this question, I need to identify the NBC sitcom that Johnny Pemberton appeared in as the character Bo Thompson. Let’s start with the reasoning: 1. Johnny Pemberton is known for his role as Bo Thompson in the NBC sitcom “That ’70s Show.” 2. “That ’70s Show” aired from 1998 to 2006. 3. The show was created by Steven Molaro, who wrote the pilot and served as showrunner. Now, I will search for more detailed information about the creator of “That ’70s Show” to confirm the answer. </think>

<search> [passage] Steven Molaro That ’70s Show creator </search>

<information> Doc 1(Title: Bigfoot (TV series)) Bigfoot is an American animated series television pilot created by Seth Rogen, Matt McKenna, and Evan Goldberg... </information>

<think> Based on the search results, the information about Steven Molaro and “That ’70s Show” is not found in the documents. However, the documents do not contain any information about Johnny Pemberton or the show “That ’70s Show.” Since the documents do not provide the necessary information, I will use graph-based retrieval to find the creator of “That ’70s Show.” </think>

<search> [graph] That ’70s Show creator </search>

<information> Doc 1(Title: Parent show) Refers to a TV series that has spawned a spin-off... </information>

<think> The search results do not provide the creator of “That ’70s Show.” Given the lack of relevant information, I will conclude based on the initial reasoning: </think>

<answer> Steven Molaro </answer>

After Training (RouteRAG-3B)

<think> To answer this question, I need to first identify the NBC sitcom Johnny Pemberton appears in as the character Bo Thompson. Then, I’ll find out who created that show. Let’s start with the sitcom. </think>

<search> [graph] Johnny Pemberton Bo Thompson </search>

<information> Doc 1(Title: Johnny Pemberton) Johnny Pemberton (born in 1981) is an American actor and comedian from Rochester, Minnesota. He is best known for his role as the titular “Son of Zorn” in the short-lived Fox sitcom, and has also appeared as the recurring character Bo Thompson in the NBC sitcom “Superstore”... </information>

<think> From Doc 1, it’s clear that Johnny Pemberton appeared as Bo Thompson in “Superstore”. Now, I need to find out who created “Superstore”. </think>

<search> [graph] Superstore creator </search>

<information> Doc 1(Title: Superstore (TV series)) Superstore is an American single-camera sitcom television series that premiered on NBC on November 30, 2015. The series was created by Justin Spitzer... </information>

<think> From Doc 1, it’s evident that “Superstore” was created by Justin Spitzer. Therefore, the answer to the question is: </think>

<answer> Justin Spitzer </answer>

Table 9: **Case study 2.** Before training, the model failed to answer the question due to a lack of reasoning and question decomposition. After training, the model correctly decomposed the question into two steps, i.e., identifying the performer of “Hello Tomorrow” as Dave Koz, and then finding out who was featured on Smooth Jazz Stations along with Kenny G and Dave Koz.

Question: Along with Kenny G and the performer of Hello Tomorrow, what artist was featured on Smooth Jazz Stations?

Ground Truth: George Benson

Before Training (Qwen2.5-7B-Instruct)
<think> I need to find out which artist was featured on Smooth Jazz Stations along with Kenny G and the performer of Hello Tomorrow. I don't have this information directly, so I will need to search for it. </think>
<search> [graph][passage] artists featured on Smooth Jazz Stations with Kenny G and Hello Tomorrow performer </search>
<information> Doc 1(Title: Adult contemporary music) In its early years of existence, the smooth jazz format...artists like George Benson, Kenny G and Dave Koz...
Doc 2(Title: Hello Tomorrow (album)) Hello Tomorrow is the ninth studio album by saxophone player Dave Koz...
</information>

<answer> Dave Koz </answer>

After Training (RouteRAG-7B)
<think> To answer this question, I need to identify three artists: Kenny G, the performer of “Hello Tomorrow”, and the artist featured on Smooth Jazz Stations. I will start by identifying the performer of “Hello Tomorrow”. </think>
<search> [graph] performer of Hello Tomorrow </search>
<information> Doc 1(Title: Hello Tomorrow (album)) Hello Tomorrow is the ninth studio album by saxophone player Dave Koz...</information>

<think> From the information provided, Dave Koz performed “Hello Tomorrow”. Now I need to find out who was featured on Smooth Jazz Stations along with Kenny G and Dave Koz. </think>
<search> [graph][passage] artist featured on Smooth Jazz Stations along with Kenny G and Dave Koz </search>
<information> Doc 1(Title: Adult contemporary music) In its early years of existence, the smooth jazz format...artists like George Benson, Kenny G and Dave Koz...</information>

<think> From the information, it seems that the artist featured on Smooth Jazz Stations along with Kenny G and Dave Koz is George Benson. </think>
<answer> George Benson </answer>
