Thinking Forward and Backward: Multi-Objective Reinforcement Learning for Retrieval-Augmented Reasoning

Wenda Wei[1,2], Yu-An Liu[1,2], Ruqing Zhang[1,2*], Jiafeng Guo[1,2*], Lixin Su[3], Shuaiqiang Wang[3], Dawei Yin[3], Maarten de Rijke[4], Xueqi Cheng[1,2]

[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Baidu Inc., Beijing, China
[4]University of Amsterdam, Amsterdam, The Netherlands
{weiwenda25z, liuyuan21b, zhangruqing, guojiafeng, cxq}@ict.ac.cn, {sulixin, wangshuaiqiang, yindawei02}@baidu.com, m.derijke@uva.nl

Abstract

Retrieval-augmented generation (RAG) has proven to be effective in mitigating hallucinations in large language models, yet its effectiveness remains limited in complex, multi-step reasoning scenarios. Recent efforts have incorporated search-based interactions into RAG, enabling iterative reasoning with real-time retrieval. Most approaches rely on outcome-based supervision, offering no explicit guidance for intermediate steps. This often leads to reward hacking and degraded response quality. We propose Bi-RAR, a novel retrieval-augmented reasoning framework that evaluates each intermediate step jointly in both forward and backward directions. To assess the information completeness of each step, we introduce a bidirectional information distance grounded in Kolmogorov complexity, approximated via language model generation probabilities. This quantification measures both how far the current reasoning is from the answer and how well it addresses the question. To optimize reasoning under these bidirectional signals, we adopt a multi-objective reinforcement learning framework with a cascading reward structure that emphasizes early trajectory alignment. Empirical results on seven question answering benchmarks demonstrate that Bi-RAR surpasses previous methods and enables efficient interaction and reasoning with the search engine during training and inference.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al. 2020) has emerged as a prominent framework for mitigating hallucination in large language models (LLMs) (Achiam et al. 2023; Gemini Team et al. 2024; Zhao et al. 2023).

Integrating RAG with reasoning. While basic RAG methods are effective, they often struggle in real-world scenarios involving complex and heterogeneous data (Gao et al. 2023) that require multi-hop retrieval (Hendrycks et al. 2020). To address these limitations, recent research has increasingly focused on enhancing RAG with advanced reasoning capabilities. Specifically, LLMs can be prompted or trained to incorporate external tools, such as search engines, into a more dynamic and iterative reasoning process (Zhao et al. 2024).

A representative paradigm is Search-R1 (Jin et al. 2025), which has achieved strong performance across a range of question answering benchmarks. The key idea is to optimize LLM reasoning trajectories through multi-turn search interactions, using retrieved token masking to enable reinforcement learning (RL) training. The success of Search-R1 is largely attributed to its outcome-based reward function based on the correctness of the final answer. However, this form of supervision lacks explicit feedback for intermediate reasoning steps, making it difficult to control the reasoning process throughout. As a result, such optimizations may induce in-context reward hacking, where the model generates unnecessarily long or inefficient reasoning chains. These extended chains can accumulate hallucinations and ultimately compromise the final response. *Can we precisely supervise the information understanding at each reasoning step?*

Beyond unidirectional reasoning. Cognitive research has shown that humans reason not only in a forward deduction, from problem to solution, which reflects how the brain plans over unknown information, but also in a backward deduction, from solution to problem (Hawes, Vostroknutov, and Rustichini 2012). Bidirectional deductive reasoning enables the brain to evaluate the reliability of known information and to plan toward the unknown information, ensuring a reasoning process that bridges the gap between the question and the answer. A recent study has also demonstrated that LLMs can similarly benefit from integrating forward and backward reasoning in complex tasks (Chen et al. 2024). Inspired by these findings, *we explore optimizing each step through top-down planning over unknown information and bottom-up evaluation of known information.*

Our method: RAG with bidirectional reasoning. We propose a novel retrieval-augmented reasoning framework, Bi-RAR, which dynamically evaluates each reasoning step through both forward and backward guidance to determine whether it provides sufficient support for task-solving. To achieve this, we need to address two key challenges.

First, *how to quantify the information completeness of*

*Jiafeng Guo and Ruqing Zhang are the corresponding authors.

each step from both forward and backward perspectives? Kolmogorov complexity (Li and Vitanyi 1993), a foundational concept in information theory, defines the amount of information required to describe an object. Building on this, information distance (Bennett et al. 1998; Vitányi et al. 2009; Zhang et al. 2007) provides a universal, domain-agnostic metric for measuring the similarity between objects and has successfully been applied across a variety of domains (Li and Vitanyi 1993; Zhang et al. 2007; Li, Zhang, and Zhu 2008). In this work, we adopt a conditional normalized information distance under specified condition patterns. For forward completeness, we measure how far the current reasoning context is from the final answer; for backward completeness, we assess how well it addresses the input question. By approximating Kolmogorov complexity via language model generation probabilities, we estimate the information distance in both directions, thereby capturing the information completeness of each step.

Second, *how to optimize step-wise reasoning using forward-backward information distances?* Given the effectiveness of RL (Kaelbling, Littman, and Moore 1996) in sequential decision-making, and the bidirectional signals introduced above, we propose to use multi-objective RL methods (Roijers et al. 2013; Li, Zhang, and Wang 2020) to explore the entire preference space. Concretely, we first design a cascading reward structure that prioritizes the early establishment of correct reasoning directions, based on the forward and backward information distances, respectively. These two reward signals serve as the primary supervision for guiding RL optimization. We then train specialized models with their respective rewards independently, using group relative policy optimization (GRPO) (Shao et al. 2024). During training, the model progressively learns to perform accurate and efficient multi-step reasoning, dynamically determining whether and how to invoke the search engine at each step, in order to optimize the forward or backward objective. Finally, we obtain a balanced solution through weight-space interpolation, which enables task-specific optimization by selecting appropriate interpolation settings.

Experiments conducted on seven widely-used question answering benchmarks demonstrate that Bi-RAR achieves strong overall performance, with particularly notable improvements of 18.2% (Qwen2.5-3B-Instruct) and 8.3% (Qwen2.5-3B-Base) over the strongest baseline Search-R1 (Jin et al. 2025), while using only one-fourth of Search-R1's training data. Further analyses show that Bi-RAR is more effective in both training and inference.

2 Preliminaries

In this section, we review Search-R1 (Jin et al. 2025), a representative method for enhancing retrieval-augmented generation with reasoning capabilities.

Search-R1. Recent advances, such as Search-R1, extend RAG to support multi-step reasoning with interleaved retrieval. In this paradigm, given a question Q, the LLM generates a reasoning trajectory $T = \{T_1, T_2, \ldots, T_n\}$. At each step i, the LLM (i) first generates a reasoning step T_i based on the current information; (ii) then, it issues a search query and retrieves relevant documents; and (iii) finally, it judges whether to move on to the next reasoning step T_{i+1} or generate the final answer A based on the current content. The process alternates between reasoning and search.

During training, RL is employed to encourage the LLM to interact effectively with the search engine. A reward function r_ϕ evaluates the correctness of the final answer extracted from the model's output. To ensure that the LLM generates valid and stable search engine calls, a structured prompting template is adopted to structure the model's output into three parts in an iterative fashion: reasoning process, search engine calling function, and the answer. Specifically, the RL policy π_θ is optimized by:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x;\mathcal{R})} \big(r_\phi - \beta D_{\mathrm{KL}}[\pi_\theta(y|x;\mathcal{R}) \| \pi_{\mathrm{ref}}(y|x;\mathcal{R})]\big), \tag{1}$$

where π_{ref} is a reference policy, D_{KL} is the KL-divergence measure, β controls the strength of the KL penalty, $\mathcal{R}$ is the search engine, x are input samples from dataset $\mathcal{D}$, and y are the generated outputs. And $\pi_\theta(\cdot|x;\mathcal{R})$ denotes the policy that generates text interleaved with the search engine.

Discussion. Search-R1 aims to teach LLMs when and how to interact with a search engine during reasoning. However, its outcome-based supervision focuses solely on the correctness of the final answer, which can easily lead to reward hacking by the model. This behavior is characterized by the LLM issuing a large number of loosely relevant queries in an attempt to improve the answer through excessive retrieval, rather than through deliberate and coherent reasoning. As shown in Section 5.4, this not only reduces efficiency due to unnecessarily lengthy reasoning trajectories, but also introduces redundant information that may accumulate across steps, increasing the risk of hallucinated content and ultimately derailing the reasoning process. In this paper, we explore how to provide fine-grained guidance at each intermediate step of the reasoning trajectory to support more efficient and accurate retrieval-augmented reasoning approach.

3 Method

3.1 Overview

In this section, we present Bi-RAR, a retrieval-augmented reasoning framework that uses bidirectional reasoning to optimize the intermediate steps in answering complex questions. As illustrated in Figure 1, our approach comprises two main components: (i) Bidirectional information quantification: at each reasoning step, we evaluate the information distance to both the final answer and the original question, assessing step-wise information completeness; and (ii) Multi-objective optimization: these distances serve as bidirectional rewards, we use a multi-objective strategy to balance answer-seeking and question-grounding, guiding the model toward well-structured reasoning.

3.2 Bidirectional information quantification

Motivation. Effective multi-step reasoning requires fine-grained supervision signals that can evaluate the quality of each intermediate step. The central challenge is to quantify the information completeness of each step, i.e., assess-

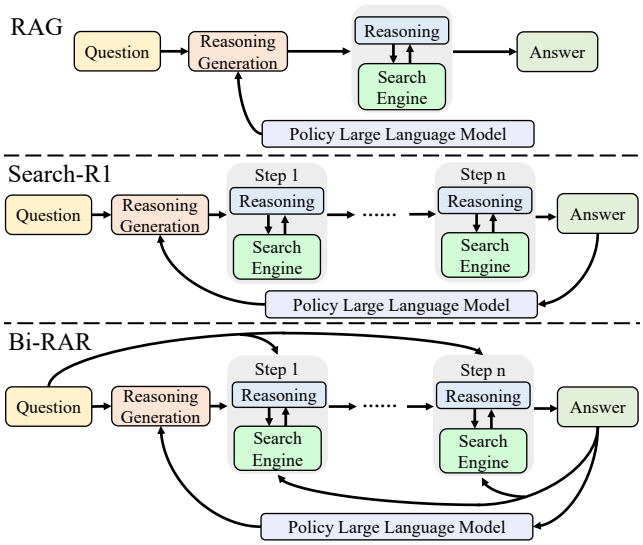


Figure 1: Framework of Bi-RAR compared with typical RAG (Lewis et al. 2020) and Search-R1 (Jin et al. 2025).

ing whether a step meaningfully advances problem-solving while remaining faithful to the original question.

To tackle this, we draw inspiration from Kolmogorov complexity theory (Li and Vitanyi 1993), which offers a domain-independent, information-theoretic foundation for assessing semantic relevance based on minimal description length. We propose a mechanism to provide efficient feedback during LLM training by quantifying the information distance between each reasoning step and both the final answer and the original question.

Information distance based on Kolmogorov complexity. Kolmogorov complexity (Li and Vitanyi 1993) measures the amount of information contained in an individual object. Given a string a, its Kolmogorov complexity $K(a)$ is defined as the length of the shortest binary program that outputs a under a fixed universal computational model. The conditional complexity $K(a|c)$ refers to the shortest program that generates a given some auxiliary input string c, capturing the information in a that is not already present in c. More generally, $K(a|b,c)$ quantifies the information required to produce a when both strings b and c are known.

Here, we adopt the normalized information distance (NID) (Zhang et al. 2007), which uses Kolmogorov complexity to define a universal, context-aware similarity metric between two pieces of content. Formally, given two strings a and b with background context string c, the conditional normalized information distance between a and b is:

$$d(a,b|c) = \frac{\min\{K(a|b,c), K(b|a,c)\}}{\min\{K(a|c), K(b|c)\}}. \tag{2}$$

Since Kolmogorov complexity is uncomputable (Li and Vitanyi 1993), we approximate the it in Eq. (2), using the generation probabilities of a language model:

$$\begin{aligned} K(u|v) &\approx -\log_2 P_{\text{LM}}(u|v), \\ K(u|v,w) &\approx -\log_2 P_{\text{LM}}(u|v,w), \end{aligned} \tag{3}$$

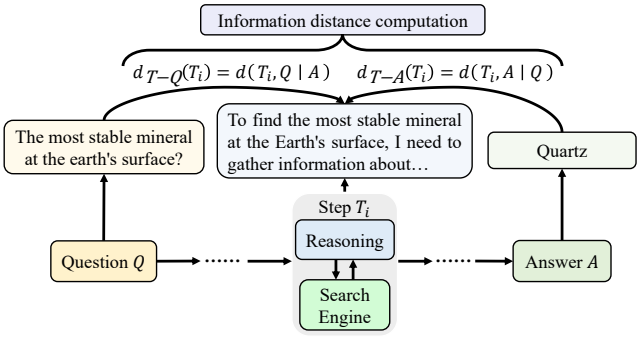


Figure 2: Sample of bidirectional distances computation.

where $P_{\text{LM}}(u|v)$ and $P_{\text{LM}}(u|v,w)$ denote the likelihood of generating u given the context v or the joint context (v,w), as computed by a language model. This approximation is grounded in Shannon's information theory (Shannon 1948), aligning with the concept of entropy. In our implementation, we use Qwen2.5-3B (Yang et al. 2024b) as the underlying language model. By employing the same language model as the generative model, this approximation can reasonably estimate the Kolmogorov complexity, which is the shortest program length required to generate the object given the contextual information.

Bidirectional distances. Building on the normalized information distance defined in Eq. (2), with conditional Kolmogorov complexity approximated by Eq. (3), we propose two complementary metrics to quantify the bidirectional informativeness of each reasoning step, which is generated by the LLM based on the question, previous reasoning steps, and retrieved documents. As shown in Figure 2, for each step T_i, we compute:

$$\begin{cases} d_{\text{T-A}}(T_i) = d(T_i, A \mid Q), & \text{step-to-answer distance} \\ d_{\text{T-Q}}(T_i) = d(T_i, Q \mid A), & \text{step-to-question distance.} \end{cases} \tag{4}$$

For each reasoning step T_i, these two distances reflect different aspects of information completeness:

1. *Step-to-answer distance* $d_{\text{T-A}}(T_i)$ quantifies how much the current step T_i contributes toward the final answer A, indicating its solution progress; and
2. *Step-to-question distance* $d_{\text{T-Q}}(T_i)$ assesses how well T_i remains grounded in the original question Q, ensuring contextual relevance and fidelity to the task.

This bidirectional formulation enables a comprehensive evaluation of each reasoning step, allowing the model to dynamically balance between deep exploration and consistent alignment with the question.

3.3 Multi-objective optimization with RL

Motivation. In this work, the LLM performs multi-step reasoning guided by bidirectional distances at each step. This frames the task as a multi-objective, multi-step sequential decision problem. To optimize this, we adopt RL to train the entire inference sequence, with three main components: (i) Designing bidirectional rewards derived from the information distances to supervise training; (ii) Independently

training models with the search engine, each optimized for a single reward, to mitigate conflicts between forward and backward objectives; and (iii) Combining the two models via weighted interpolation to obtain a balanced solution that guides the model to generate reasoning steps both relevant to the question and progressively closer to the correct answer.

Bidirectional rewards design. Based on the computed bidirectional distances, we define corresponding bidirectional reward functions. To account for the varying importance of reasoning steps, we introduce a cascading reward structure that prioritizes early establishment of correct reasoning directions. Specifically, the forward reward R_{forward} and backward reward R_{backward} are defined as:

$$R_{\text{forward}} = \mathbb{1}[\text{correct}] \cdot \sum_{i=1}^{n} \left[\prod_{j=1}^{i-1} (1 - r_j^{\text{T-A}}) \right] r_i^{\text{T-A}}, \quad (5)$$

$$R_{\text{backward}} = \mathbb{1}[\text{correct}] \cdot \sum_{i=1}^{n} \left[\prod_{j=1}^{i-1} (1 - r_j^{\text{T-Q}}) \right] r_i^{\text{T-Q}}, \quad (6)$$

where $\mathbb{1}[\text{correct}]$ equals 1 if the final answer is correct, and 0 otherwise; and

$$r_i^{\text{T-A}} = e^{-d_{\text{T-A}}(T_i)}, \quad r_i^{\text{T-Q}} = e^{-d_{\text{T-Q}}(T_i)}, \quad (7)$$

represent the rewards derived from the step-to-answer and step-to-question distances at step i.

The exponential mapping ensures rewards increase as distances decrease, normalizing values between 0 and 1. The cascading factor $\prod_{j=1}^{i-1}(1 - r_j)$ diminishes the contribution of later steps if earlier steps already show strong alignment, thereby encouraging efficient reasoning paths that establish correct directions as early as possible.

Independent training with the search engine. To mitigate convergence issues from conflicting optimization objectives between the two rewards in early training, we initialize and train two models independently from the same pretrained checkpoint. Specifically, (i) θ_{forward} is only optimized for the forward reward R_{forward}; (ii) θ_{backward} is only optimized for the backward reward R_{backward}. During training, the model can autonomously interact with the retriever at each reasoning step based on its current needs. RL guides the model to perform accurate multi-step reasoning and streamlined retrieval to optimize the forward or backward objective.

Each model is trained using group relative policy optimization (GRPO) (Shao et al. 2024), which enhances training stability by employing group-wise baselines instead of value networks. For each input question x, we sample G candidate responses $\{y_i\}_{i=1}^{G}$ from the current policy π_{old} with the search engine $\mathcal{R}$, and optimize the objective:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|x;\mathcal{R})} \Big[& \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \\ & \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) \\ & - \beta D_{\text{KL}}[\pi_\theta||\pi_{\text{ref}}]\Big], \end{aligned} \quad (8)$$

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t}; \mathcal{R})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t}; \mathcal{R})}, \quad (9)$$

where $\hat{A}_{i,t}$ is the standardized advantage computed from group-relative rewards, ϵ controls the trust region size, and β weights the KL penalty term.

Multi-objective optimization. After training the two models θ_{forward} and θ_{backward}, we seek a balanced solution that integrates the strengths of both reward directions. Inspired by linear mode connectivity (Neyshabur, Sedghi, and Zhang 2020; Frankle et al. 2020), we apply linear weight interpolation to combine the parameters of the two models, enabling the resulting model to simultaneously incorporate forward and backward reasoning capabilities. The final interpolated model $\theta_{\text{Bi-RAR}}$ is defined as:

$$\theta_{\text{Bi-RAR}} = (1-\lambda)\cdot\theta_{\text{forward}} + \lambda\cdot\theta_{\text{backward}}, \quad \lambda \in [0,1], \quad (10)$$

where λ controls the interpolation ratio. By varying λ, we can explore a continuum of models that trade off between answer accuracy and question relevance, allowing flexible adaptation to different task requirements without the need for additional retraining.

4 Experimental Settings

Datasets. We evaluate Bi-RAR on seven question answering benchmarks split into two groups: (i) **General QA** datasets focus on factual questions that require accurate retrieval and understanding of real-world knowledge, generally involve single-hop reasoning: NQ (Kwiatkowski et al. 2019), TriviaQA (Joshi et al. 2017), and PopQA (Mallen et al. 2022). (ii) **Multi-hop QA** datasets are specifically designed to evaluate a model's ability to integrate multiple pieces of evidence across documents to answer a question, making them ideal for testing complex reasoning: HotpotQA (Yang et al. 2018), 2WikiMultiHopQA (Ho et al. 2020), Musique (Trivedi et al. 2022), and Bamboogle (Press et al. 2022).

Baselines. The baselines are grouped by how they incorporate retrieval into the reasoning process: (i) **Reasoning without retrieval**: These methods rely solely on the model's parametric knowledge to perform reasoning without retrieval, including Direct inference, Chain-of-Thought (CoT) reasoning (Wei et al. 2022), Supervised fine-tuning (SFT) (Chung et al. 2024) and RL-based fine-tuning without retrieval (R1) (Guo et al. 2025). (ii) **One-step retrieval and reasoning**: These approaches retrieve external evidence once before generating the answer, including Retrieval-Augmented Generation (RAG) (Lewis et al. 2020). (iii) **Multi-step retrieval and reasoning**: These methods perform iterative retrieval interleaved with reasoning, enabling the model to gather new information at each step, including IR-CoT (Trivedi et al. 2023), Search-o1 (Li et al. 2025), and Search-R1 (Jin et al. 2025) trained with GRPO. All baseline results are taken from Search-R1. To ensure a fair comparison, all methods use the same retriever, retrieval setting, knowledge corpus, training dataset, and pre-trained LLMs.

Model variants. Bi-RAR includes two variants: (i) **Forward-RAR**, trained only with the forward reward R_{forward}, as in θ_{forward}; and (ii) **Backward-RAR**, trained only with the backward reward R_{backward}, as in θ_{backward}. For these two variants, we only train a single model using the correspond-

Methods	General QA			Multi-Hop QA				
	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
Reasoning without retrieval								
Direct Inference	0.106	0.288	0.108	0.149	0.244	0.020	0.024	0.134
CoT	0.023	0.032	0.005	0.021	0.021	0.002	0.000	0.015
SFT	0.249	0.292	0.104	0.186	0.248	0.044	0.112	0.176
R1-base	0.226	0.455	0.173	0.201	0.268	0.055	0.224	0.229
R1-instruct	0.210	0.449	0.171	0.208	0.275	0.060	0.192	0.224
One-step reasoning with retrieval								
RAG	0.348	0.544	0.387	0.255	0.226	0.047	0.080	0.270
Multi-step reasoning with retrieval								
IRCoT	0.111	0.312	0.200	0.164	0.171	0.067	0.240	0.181
Search-o1	0.238	0.472	0.262	0.221	0.218	0.054	0.320	0.255
Search-R1-base	0.421	0.583	0.413	0.297	0.274	0.066	0.128	0.312
Search-R1-instruct	0.397	0.565	0.391	0.331	0.310	0.124	0.232	0.336
Bi-RAR-base	**0.442**	**0.614**	**0.432**	0.317	0.297	0.073	0.188	0.338
Bi-RAR-instruct	0.438	0.608	0.421	**0.391**	**0.402**	**0.153**	**0.363**	**0.397**

Table 1: Main results of Bi-RAR and baselines on QA benchmarks. The best performance is highlighted in bold.

ing rewards, without performing interpolation.

Implementation details. We use both the Qwen2.5-3B-Base and Qwen2.5-3B-Instruct model (Yang et al. 2024b). Following Search-R1 (Jin et al. 2025), we train our model on a combined training set of NQ and HotpotQA, adopt the same training and evaluation prompt template as Search-R1, and use **Exact Match (EM)** as the evaluation metric. For retrieval, we adopt the 2018 Wikipedia dump (Karpukhin et al. 2020) as the knowledge source and use E5 (Wang et al. 2022) to simulate a search engine.

The training batch size is set to 128, and the validation batch size is set to 256, using only one-fourth of the training data compared to Search-R1. To manage memory usage efficiently, we use gradient checkpointing and fully sharded data parallel (FSDP) with CPU offloading. For efficient response generation, we use vLLM with a tensor parallel size of 1 and a GPU memory utilization ratio of 0.6. Sampling is performed with a temperature of 1.0 and top-p of 1.0. We set the KL divergence regularization coefficient to $\beta = 0.001$, and the clipping ratio to $\epsilon = 0.2$. In GRPO training, we follow the implementation from Verl (Sheng et al. 2025). Training runs for 200 steps. We set the policy model's learning rate to 1e-6 and sample 5 responses per prompt. In multi-objective optimization, we tested λ values of 0.25, 0.5, and 0.75, which emphasize different objectives. We select $\lambda = 0.25$ for both the Qwen2.5-3B-Base and Qwen2.5-3B-Instruct models as it achieved the best performance.

5 Experimental Results

In this section, we report the experimental results to demonstrate the effectiveness of Bi-RAR.

5.1 Main results

Table 1 presents the overall performance of Bi-RAR compared to baseline methods across seven question answering benchmarks. Observations on the baselines are: (i) Overall, models equipped with retrievers achieve better performance than those without, indicating that access to external knowledge sources can effectively complement the model's internal knowledge. (ii) Models perform better on general QA datasets than multi-hop QA datasets. This discrepancy indicates that multi-hop reasoning and evidence aggregation remain challenging for the models. (iii) Among the baselines, Search-R1 performs best, benefiting from iterative retrieval and outcome-based supervision that improve accuracy.

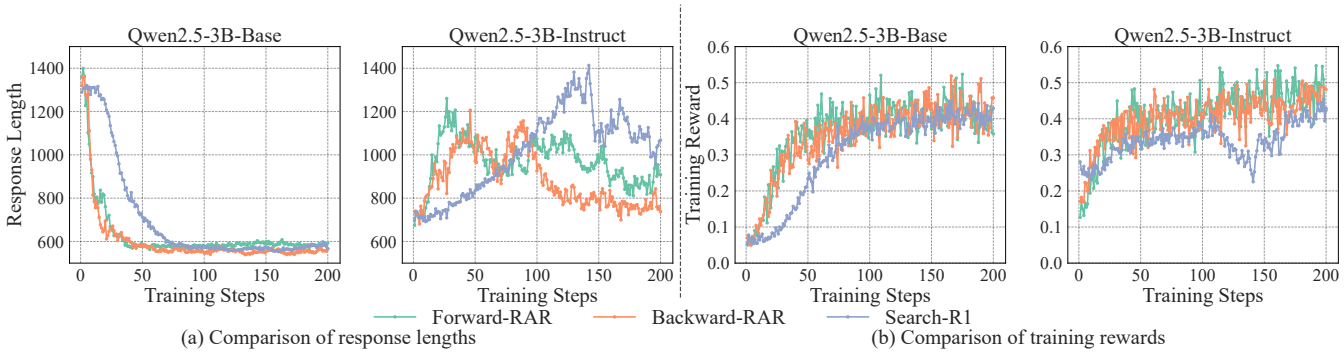
When we look at Bi-RAR, we find that: (i) Bi-RAR achieves the best overall performance among all evaluated models, with an average relative improvement of 18.2% and 8.3% over the strongest baseline when using Qwen2.5-3B Instruct and Base, respectively. This demonstrates that our multi-objective optimization approach based on bidirectional information quantification supervision effectively constrains the reasoning trajectory, guiding the model to generate more accurate and compact answers. (ii) Compared to the strongest baseline Search-R1, Bi-RAR delivers consistent gains across diverse datasets, despite being trained on only one-fourth of Search-R1's training data. For example, Bi-RAR improves the performance on HotpotQA by 0.06 and 2Wiki by 0.092, corresponding to 18.1% and 29.7% relative increase, under the instruct-tuned model. This indicates that bidirectional distances offer precise step-level optimization signals, leading to more efficient training and better inference quality. (iii) Bi-RAR demonstrates effectiveness on both base and instruction-tuned models, suggesting strong generalization across model types.

5.2 Ablation study

We conduct ablation experiments comparing the variants of Bi-RAR. The results shown in Table 2 demonstrate that: (i) Forward-RAR performs better than Backward-RAR, with relative improvements of 3.7% and 1.9% on the base and instruct variants. This result indicates that reward signals propagated from the answer side contribute more directly to final answer correctness. This aligns with our intuition, as anchoring generation on the expected answer better constrains the

Methods	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
Qwen2.5-3B-Base								
Forward-RAR-base	0.440	0.613	**0.435**	0.313	0.293	0.066	0.169	0.333
Backward-RAR-base	0.435	0.599	0.423	0.313	0.282	0.069	0.125	0.321
Bi-RAR-base	**0.442**	**0.614**	0.432	**0.317**	**0.297**	**0.073**	**0.188**	**0.338**
Qwen2.5-3B-Instruct								
Forward-RAR-instruct	0.432	0.598	0.418	0.376	0.375	0.144	0.347	0.384
Backward-RAR-instruct	0.436	0.602	0.391	0.380	0.339	0.145	0.347	0.377
Bi-RAR-instruct	**0.438**	**0.608**	**0.421**	**0.391**	**0.402**	**0.153**	**0.363**	**0.397**

Table 2: Ablation results of forward, backward, and bidirectional reasoning in Bi-RAR with different backbone LLMs.



(a) Comparison of response lengths (b) Comparison of training rewards

Figure 3: Trends in response lengths and rewards change during RL training for Forward/Backward-RAR and Search-R1.

reasoning path. (ii) Bi-RAR achieves the best performance across most datasets under both base and instruct backbone models. This demonstrates the effectiveness of our multi--objective optimization framework in integrating forward and backward objectives, allowing the model to incorporate complementary reasoning signals and achieve stronger overall accuracy.

5.3 Training analysis

We compare the training dynamics of Forward-RAR, Backward-RAR, and Search-R1 on both the Qwen2.5-3b-Base and Qwen2.5-3b-Instruct models, focusing on response length and train reward trends.

Response length. As shown in Figure 3(a), on models initialized from the base model, the response lengths of Forward-RAR and Backward-RAR decrease faster than Search-R1. This demonstrates that the cascading reward structure, which emphasizes early trajectory alignment, leads to more efficient reasoning by guiding the model to eliminate unnecessary steps early in training. On the instruction-tuned models, all methods show an initial increase followed by a decrease in response length. This is because the instruct model has a stronger instruction following ability, initially attempts to find correct answers via longer reasoning chains. As training progresses, the model learns that shorter responses can omit redundant steps while improving answer accuracy, leading to a response length reduction. By the end of training, both Forward-RAR and Backward-RAR produce shorter responses than Search-R1 on the instruct model, indicating that our forward and backward information distance supervision effectively guides the model to generate accurate and concise reasoning steps.

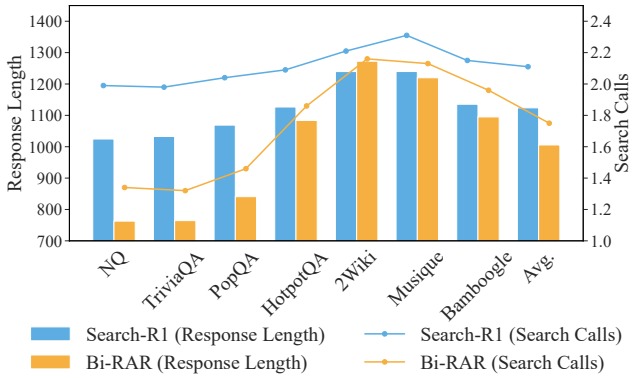


Figure 4: Response lengths and search calls in inference.

Rewards. As shown in Figure 3(b), Forward-RAR and Backward-RAR converge faster than Search-R1 on both Base and Instruct models. This suggests that forward and backward reward signals provide more precise guidance for optimizing each step, enabling more effective training supervision. On instruction-tuned models, the rewards of Forward/Backward-RAR exhibit no severe fluctuations as observed in Search-R1 during training, which reflects the robustness and consistency of our bidirectional reward design.

5.4 Inference analysis

To analyze the response efficiency during the inference phase, we compare Bi-RAR and Search-R1 in terms of response length and number of search calls, both of which di-

rectly affect inference efficiency. We used the Qwen2.5-3B-Instruct model for comparison, with similar observations on other backbones. The results across seven datasets and their average are shown in Figure 4.

We observe that: (i) For response length, Bi-RAR generates shorter responses than Search-R1 on most datasets, with the reduction notable on the general QA datasets. This is attributed to the cascading reward structure that emphasizes early trajectory alignment, enabling Bi-RAR to generate more concise and less redundant responses. (ii) For search calls, Bi-RAR reduces the number of retrievals compared to Search-R1 across all datasets. This is because the bidirectional distances supervision mitigates invalid reasoning paths and corresponding redundant searches, leading to more efficient inference. (iii) On 2WikiMultiHopQA, Bi-RAR produces longer responses than Search-R1 while using fewer search calls. This is due to the complex multi-hop reasoning required by the dataset, where our bidirectional supervision better guides the model to maintain coherent long-range inference with fewer but more targeted retrievals. As a result, Bi-RAR achieves a substantial relative performance gain of 29.7%.

6 Related Work

Retrieval-augmented generation. Retrieval-augmented generation (Lewis et al. 2020) is a widely adopted framework that enhances large language models (LLMs) (Achiam et al. 2023; Gemini Team et al. 2024; Zhao et al. 2023) by incorporating external knowledge sources. This technique effectively reduces hallucination (Zhang et al. 2023) and improves task performance (Gao et al. 2023; Shuster et al. 2021; Jiang et al. 2023). Building on this foundation, many studies have explored improving the performance of RAG systems by optimising prompts or training objectives, such as Self-RAG, REPLUG, and RA-DIT (Asai et al. 2023; Shi et al. 2024; Lin et al. 2023). However, this single-round framework of retrieval-then-answering makes LLMs difficult to capture sufficient information and perform complete reasoning, leading to poor performance in handling complex multi-hop reasoning tasks. To address this, recent approaches incorporate multi-step reasoning and retrieval, retrieval-augmented reasoning, to further enhance the model's capability in complex scenarios: (i) IRCoT (Trivedi et al. 2023) interleaves retrieval within the chain-of-thought reasoning process; (ii) Search-o1 (Li et al. 2025) enhances LLMs by integrating agentic search capabilities that dynamically retrieve and incorporate external knowledge during the reasoning process; (iii) Search-R1 (Jin et al. 2025) uses reinforcement learning to train LLMs to autonomously generate search queries and use real-time retrieval during step-by-step reasoning.

These methods are primarily guided by the final answer, encouraging LLMs to interact more with search engines. However, such a "distant" supervision signal cannot provide precise guidance for each interaction, leading to over or distorted reasoning directions by LLMs. An effective reasoning trajectory should continuously progress toward the solution while remaining grounded in the original problem context.

Therefore, we propose a bidirectional information quantification to define the optimization objective for each reasoning step, enabling LLMs to determine the reasoning direction based on the current information sufficiency.

LLMs and reinforcement learning. Reinforcement learning (Kaelbling, Littman, and Moore 1996) has fundamentally reshaped how we align LLMs with human preferences, evolving from computationally intensive strategies to more elegant and efficient solutions. Early implementations such as PPO required both a reward and a critic model (Ouyang et al. 2022; Schulman et al. 2017). DPO simplified this process by removing the reward model and directly optimizing on preference data (Rafailov et al. 2023), while GRPO further simplifies the pipeline by dropping the critic model and using sampled responses to estimate advantages (Shao et al. 2024). These advances have significantly enhanced LLM reasoning capabilities, as evidenced by models such as OpenAI's o1, DeepSeek-R1 and Qwen2.5 (Jaech et al. 2024; Guo et al. 2025; Yang et al. 2024b).

In practice, LLM training often involves multiple optimization objectives that need to be effectively balanced during reinforcement learning. Multi-objective reinforcement learning (MORL) (Barrett and Narayanan 2008; Roijers et al. 2013; Li, Zhang, and Wang 2020) extends standard RL by replacing the single scalar reward signal with multiple feedback signals, each corresponding to a different objective. Recent work (Rame et al. 2023) has begun exploring multi-objective optimization for LLMs.

In this paper, we adopt a multi-objective reinforcement learning approach to simultaneously optimize the forward and backward objectives in our framework to achieve an effective balanced solution.

7 Conclusion

We have proposed Bi-RAR, a novel retrieval-augmented reasoning framework designed to enhance multi-step reasoning ability of LLMs. We introduce bidirectional information quantification grounded in Kolmogorov complexity theory, which jointly measures how far each reasoning step is from the final answer and how well it addresses the original question. To effectively use these signals, we adopt a multi-objective reinforcement learning framework, enabling a smooth trade-off between the two objectives. Experiments demonstrate that bidirectional reasoning guidance can significantly improve the accuracy of LLMs in solving complex problems, while achieving efficient interaction and reasoning with the search engine during training and inference.

Broader impact and limitations. We aim to make an initial exploration into multi-step retrieval-augmented reasoning, and to inspire the community to further advance this line of research. As to the limitations of our work, we approximate Kolmogorov complexity using generation probabilities from an LLM when computing bidirectional information quantification, which could be time-consuming. In future work, we plan to explore more efficient approximation methods and extend our framework to larger models. Investigating efficient reasoning through low-resource model training in complex real-world search scenarios represents another promising direction.

Acknowledgments

This work was funded by the Strategic Priority Research Program of the CAS under Grants No. XDB0680102, the National Natural Science Foundation of China (NSFC) under Grants No. 62472408 and 62441229, the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union under grant agreements No. 101070212 (FINDHR) and No. 101201510 (UNITE). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*.

Barrett, L.; and Narayanan, S. 2008. Learning All Optimal Policies with Multiple Criteria. In *Proceedings of the 25th international conference on Machine learning*, 41–47.

Bennett, C. H.; Gács, P.; Li, M.; Vitányi, P. M.; and Zurek, W. H. 1998. Information Distance. *IEEE Transactions on information theory*, 44(4): 1407–1423.

Chen, J. C.-Y.; Wang, Z.; Palangi, H.; Han, R.; Ebrahimi, S.; Le, L.; Perot, V.; Mishra, S.; Bansal, M.; Lee, C.-Y.; et al. 2024. Reverse Thinking Makes LLMs Stronger Reasoners. *arXiv preprint arXiv:2411.19865*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling Instruction-finetuned Language Models. *Journal of Machine Learning Research*, 25(70): 1–53.

Frankle, J.; Dziugaite, G. K.; Roy, D.; and Carbin, M. 2020. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In *International Conference on Machine Learning*, 3259–3269. PMLR.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Gemini Team; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Hawes, D. R.; Vostroknutov, A.; and Rustichini, A. 2012. Experience and Abstract Reasoning in Learning Backward Induction. *Frontiers in neuroscience*, 6: 23.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing a Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. *arXiv preprint arXiv:2011.01060*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*.

Jiang, Z.; Xu, F. F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; and Neubig, G. 2023. Active Retrieval Augmented Generation. In *EMNLP*, 7969–7992.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv preprint arXiv:1705.03551*.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4: 237–285.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*, 6769–6781.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented Generation for Knowledge-intensive NLP Tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, F.; Zhang, X.; and Zhu, X. 2008. Answer Validation by Information Distance Calculation. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, 42–49.

Li, K.; Zhang, T.; and Wang, R. 2020. Deep Reinforcement Learning for Multiobjective Optimization. *IEEE Transactions on Cybernetics*, 51(6): 3103–3114.

Li, M.; and Vitanyi, P. 1993. *An Introduction to Kolmogorov Complexity and its Applications*. New York: Springer-Verlag.

Li, X.; Dong, G.; Jin, J.; Zhang, Y.; Zhou, Y.; Zhu, Y.; Zhang, P.; and Dou, Z. 2025. Search-o1: Agentic Search-enhanced Large Reasoning Models. *arXiv preprint arXiv:2501.05366*.

Lin, X. V.; Chen, X.; Chen, M.; Shi, W.; Lomeli, M.; James, R.; Rodriguez, P.; Kahn, J.; Szilvasy, G.; Lewis, M.; et al. 2023. Ra-dit: Retrieval-augmented Dual Instruction Tuning.

In *The Twelfth International Conference on Learning Representations*.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2022. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-parametric Memories. *arXiv preprint arXiv:2212.10511*.

Neyshabur, B.; Sedghi, H.; and Zhang, C. 2020. What is Being Transferred in Transfer Learning? *Advances in Neural Information Processing Systems*, 33: 512–523.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and Narrowing the Compositionality Gap in Language Models. *arXiv preprint arXiv:2210.03350*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2023. Rewarded Soups: Towards Pareto-optimal Alignment by Interpolating Weights Fine-tuned on Diverse Rewards. *Advances in Neural Information Processing Systems*, 36: 71095–71134.

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A Survey of Multi-objective Sequential Decision-making. *Journal of Artificial Intelligence Research*, 48: 67–113.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3): 379–423.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 1279–1297.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2024. Replug: Retrieval-augmented Black-box Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8371–8384.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-thought Reasoning for Knowledge-intensive Multi-step Questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 10014–10037.

Vitányi, P. M.; Balbach, F. J.; Cilibrasi, R. L.; and Li, M. 2009. Normalized Information Distance. In *Information Theory and Statistical Learning*, 45–82. Springer.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Text Embeddings by Weakly-supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024b. Qwen2.5 Ttechnical Rreport. *arXiv preprint arXiv:2412.15115*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *arXiv preprint arXiv:1809.09600*.

Zhang, X.; Hao, Y.; Zhu, X.; Li, M.; and Cheriton, D. R. 2007. Information Distance from a Question to an Answer. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 874–883.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-augmented Generation for AI-generated Content: A Survey. *arXiv preprint arXiv:2402.19473*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.