Efficient Prior-Guided Reasoning for Robust Retrieval-Augmented Generation under Conflicts

Xiaowei Yuan[1,2,3], **Ziyang Huang**[1,2], **Zhao Yang**[4], **Yequan Wang**[3], **Jun Zhao**[1,2], **Kang Liu**[1,2,*]

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

[3]Beijing Academy of Artificial Intelligence

[4]Meituan

Abstract

Retrieval-Augmented Generation (RAG) has become a standard paradigm for grounding Large Language Models (LLMs) with external knowledge. However, RAG performance often degrades substantially when faced with noisy, outdated, or conflicting retrieved information. In this work, we empirically demonstrate that Prior-Guided Reasoning—a strategy that explicitly elicits the model's parametric knowledge as prior information to guide reasoning on retrieved documents—effectively mitigates the impact of external conflicts. Building on this, we propose BrPr (Bernoulli-gated reinforcement learning for Prior-Guided reasoning), a framework that achieves robust performance across varying degrees of external inconsistency. Furthermore, by employing a Bernoulli-gated dropout mechanism during training, BrPr distills the prior-driven reasoning capability into the model parameters, enabling efficient latent reasoning without explicit prior generation. The experimental results demonstrate that BrPr consistently exhibits superior robustness to external conflicts and noise.

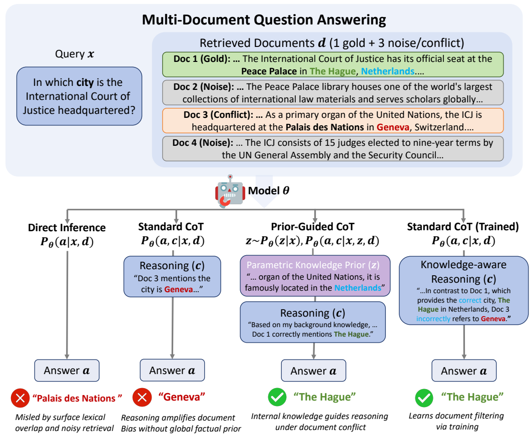

Figure 1: A comparison of RAG inference strategies under noisy and conflicting retrieval. Prior-Guided CoT leverages elicited parametric knowledge as background information, enabling prior-guided reasoning with correct answer generation where standard methods fail.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm for extending the knowledge boundaries of Large Language Models (LLMs) (Yang et al., 2025; Guo et al., 2025). By dynamically integrating external knowledge during inference, RAG empowers models to transcend the limitations of their static training data, producing outputs that are more factually precise and contextually grounded (Gao et al., 2024).

Despite the significant advances enabled by RAG, the dependency on external retrieval introduces a new set of vulnerabilities, particularly when the retrieved corpus contains conflicting or noisy information (Xu et al., 2024; Cuconasu et al., 2024). Previous studies (Shi et al., 2023; Mansurova et al., 2024; Amiraz et al., 2025) have empirically demonstrated that irrelevant or misleading passages significantly degrade performance in Question Answering (QA) tasks.

Existing research has explored many paradigms to mitigate knowledge conflicts, ranging from in-context prompting strategies (Zhou et al., 2023) that utilized manual heuristics to elicit model reasoning to decoding-based approaches (Yuan et al., 2024a; Shi et al., 2024) that leveraged contrastive decoding-like method to adjust the probabilities of output tokens in conflicting scenarios. Beyond these inference-time adjustments, other studies focused on enabling models to learn the discriminative features of gold and noisy documents through supervised fine-tuning or preference optimization (Fang et al., 2024; Zhang et al., 2025). Another line of methods (Wang et al., 2025; Hu et al., 2025b) introduced multi-agent frameworks that facilitate multi-round debates between LLM agents to resolve inter-document discrepancies.

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22149–22164
July 2-7, 2026

However, these approaches often fail to fully exploit internalized world knowledge (Yu et al., 2023; Zhang et al., 2023) for document conflict resolution. To bridge this gap, we propose **Prior-Guided Reasoning**, a strategy that explicitly elicits parametric knowledge as prior information to guide its reasoning on retrieved documents. As illustrated in Figure 1, under this paradigm, the model first elicits its internal knowledge (e.g., "*located in Netherlands*") to establish a semantic baseline. This baseline then functions as a critical evaluative anchor (e.g., "*Doc 1 correctly mentions The Hague in Netherlands*"), allowing the model to distinguish between misleading conflicts or irrelevant noise.

Through empirical (Sec. 2.1) and theoretical analysis (Sec. 2.2), we demonstrate the performance improvements yielded by naive prior-guided reasoning. However, we also identify a critical robustness trade-off: the model may ignore factual external documents in favor of its own parametric biases. This behavior becomes particularly detrimental when the model's internal knowledge is erroneous or outdated. Moreover, explicit externalization of parametric knowledge entails a high inference latency due to sequence expansion.

To this end, we propose **BrPr** (Bernoulli-gated reinforcement learning for Prior-Guided reasoning). We utilize Reinforcement Learning (RL) to optimize the utilization of the model's prior knowledge. Through trial-and-error exploration with explicit feedback, the model learns to better reconcile parametric priors with document conflicts. To improve inference efficiency, we introduce a Bernoulli-gated dropout mechanism that stochastically omits explicit prior generation during training. It compels the model to maintain performance in the absence of prior guidance. This objective effectively distills the prior-driven reasoning capability into the model parameters, fostering a robust and efficient latent reasoning process without explicit prior generation.

Experiments conducted on the synthetic conflict dataset RAG-Bench (Fang et al., 2024), which includes Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQuestions (WebQ) (Berant et al., 2013), demonstrate that BrPr substantially improves model robustness under varying levels of external inconsistency. Moreover, results on Wikipedia-based retrieval benchmarks indicate the strong generalization of BrPr across diverse retrieval scenarios.

The contributions of this paper are as follows:

- Through empirical and theoretical analysis on multi-document QA tasks, we demonstrate that prior-guided reasoning effectively mitigates the volatility of external knowledge by providing an evaluative anchor that grounds the model's reasoning in conflicting or noisy information.
- This paper proposes BrPr, a method that achieves robust and efficient performance across conflicts and noises. Experimental results demonstrate that BrPr provides superior robustness against external conflicts with low computational overhead.

2 Externalizing Internal Knowledge as prior information

Standard RAG models θ generate a response $y = (a, c)$—comprising an answer a and an optional Chain-of-Thought (CoT) c sequence (Wei et al., 2022)—as $P_\theta(y|x, d)$ given query x and documents d. To mitigate performance degradation from noisy or conflicting external data, we propose eliciting the model's internal knowledge $z \sim P_\theta(z|x)$ as prior information, thereby enhancing robustness against external knowledge volatility.

2.1 Analysis on Task Performance

To evaluate this hypothesis, we conduct a series of comparative analyses using the Qwen2.5-7B-Instruct (Yang et al., 2025) model on the conflict-centric RAG-Bench (Fang et al., 2024). We specifically examine the performance variations across following strategies. Detailed prompts and data description are provided in Appendix B.

(i) **Parametric Answer**: $P_\theta(a, c|x)$, generating the answer without documents.

(ii) **Direct Inference**: $P_\theta(a|x, d)$, generating the answer directly without reasoning.

(iii) **Standard CoT**: $P_\theta(a, c|x, d)$, incorporating a reasoning chain without explicit internal knowledge.

(iv) **Joint CoT:** $P_\theta(a, c, z|x, d)$, where background knowledge z is generated within a unified reasoning chain.

(v) **Prior-Guided CoT:** $P_\theta(a, c|x, z, d)$, where internal knowledge z is elicited in an initial stage $z \sim P_\theta(z|x)$ as prior information.

Table 1 illustrates a clear trade-off between strategy complexity and the volume of external distractors. As documents become noisier (from Doc 2 to Doc 10), externalizing internal knowledge as

Table 1: Performance comparison of different strategies on the RAG-Bench test set. For document count > 1, the other documents are sampled from noisy and conflicting ones corresponding to the same query.

Methods	Doc 1 (Gold)		Doc 2		Doc 5		Doc 10	
	EM	F1	EM	F1	EM	F1	EM	F1
NQ								
Parametric Answer	28.73	37.35						
Direct Inference	**60.46**	**69.81**	51.75	61.96	38.20	47.73	31.13	39.85
Standard CoT	59.46	69.11	52.35	62.14	42.14	52.79	36.73	47.79
Joint CoT	57.36	68.18	50.75	60.98	42.94	54.40	38.74	50.34
Prior-Guided CoT	52.95	62.64	**53.85**	**63.93**	**48.04**	**58.28**	**41.94**	**52.92**
TriviaQA								
Parametric Answer	59.86	64.89						
Direct Inference	79.18	84.68	74.17	79.56	64.56	70.86	53.54	62.96
Standard CoT	**79.68**	**85.02**	76.98	82.39	67.36	74.50	58.76	66.10
Joint CoT	79.34	84.87	76.18	81.49	70.07	76.49	61.36	69.14
Prior-Guided CoT	76.87	82.48	**77.17**	**82.35**	**75.18**	**81.12**	**69.37**	**75.70**
WebQ								
Parametric Answer	26.63	46.01						
Direct Inference	**40.44**	**58.39**	33.63	52.62	27.93	44.96	20.32	34.63
Standard CoT	38.53	55.24	33.43	51.83	28.63	47.30	23.92	42.75
Joint CoT	36.34	54.95	34.63	53.13	32.63	51.41	25.22	43.85
Prior-Guided CoT	37.24	54.81	**35.24**	**53.74**	**34.03**	**52.33**	**28.83**	**47.58**

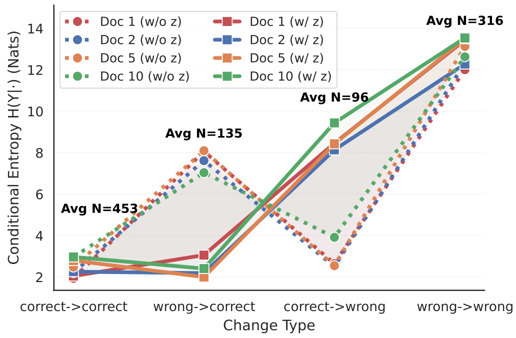

Figure 2: Impact of prior knowledge z on output uncertainty: entropy analysis across four data subsets following the transition from $p(y|x, d)$ to $p(y|x, z, d)$.

prior information significantly enhances robustness. *Prior-Guided CoT* effectively mitigates the volatility of external knowledge by providing a stabilizing anchor that grounds the model's reasoning in conflicting or noisy information. Conversely, in gold setting (Doc 1), simple strategies (*Direct* and *Standard CoT*) consistently achieve efficient performance while minimizing generation overhead (w/o z). This indicates that prior knowledge may induce a prejudicial reliance on internal biases, even when accurate external information is available. This issue becomes prominent when the model's internal knowledge is unreliable. For instance, *Prior-Guided CoT* significantly underperforms the *Direct* method demonstrated by the drop from 60.46% to 52.95% on the NQ benchmark.

2.2 Analysis on Information Gain

Furthermore, to quantify the contribution of z to the output y, we employ Monte Carlo estimation to approximate the conditional mutual information (CMI) (Cover and Thomas, 2006), defined as $I(Y; Z|X, D)$. CMI captures the reduction in output uncertainty when conditioning on internal knowledge z:

$$I(Y; Z|X, D) = H(Y|X, D) - H(Y|X, Z, D)$$

where $H(Y|X, D)$ is the conditional entropy of the output given only external contexts, and $H(Y|X, Z, D)$ denotes the entropy when both external contexts and model's internal knowledge are available.

CMI serves as an empirical measure of the marginal information gain contributed by the internal knowledge z. Within this framework, a positive information gain ($I > 0$) is manifested as a reduction in entropy ($\mathrm{H}_{w/\ z} < \mathrm{H}_{w/o\ z}$), indicating that the prior z effectively constrains the output space[1]. Detailed settings are provided in Appendix B.3.

As illustrated in Figure 2, the significant entropy drop observed in the "wrong → correct" category provides empirical evidence that z can serve as a high-quality prior, effectively reducing the predictive entropy and steering the model toward the correct response. Conversely, data where entropy increases suggest that the prior knowledge also introduces noise and the model lacks the capacity to reconcile the prior with the provided context.

To overcome this limitation, this paper introduces BrPr, which optimizes the model's ability to reconcile parametric priors with external information from multiple retrieved documents.

3 BrPr: Bernoulli-gated Reinforcement learning for Prior-Guided Reasoning

With the goal of achieving robust and efficient performance across conflicts and noises, we propose BrPr method, as illustrated in Figure 3.

First, within the prior-guided reasoning strategy (Sec. 3.1), we employ the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) within the framework of Reinforcement Learning from Verifiable Rewards (RLVR) to strengthen the model's ability to reconcile parametric priors with multiple retrieved documents (Sec. 3.2). This training objective explicitly optimizes the alignment between intermediate reasoning chains and verifiable outcomes. Through explorative trial-and-error

[1] To isolate the marginal information gain attributed to the internal knowledge z, we bypass the intermediate reasoning chain and constrain the response y to the target answer a (i.e., $y = a$).

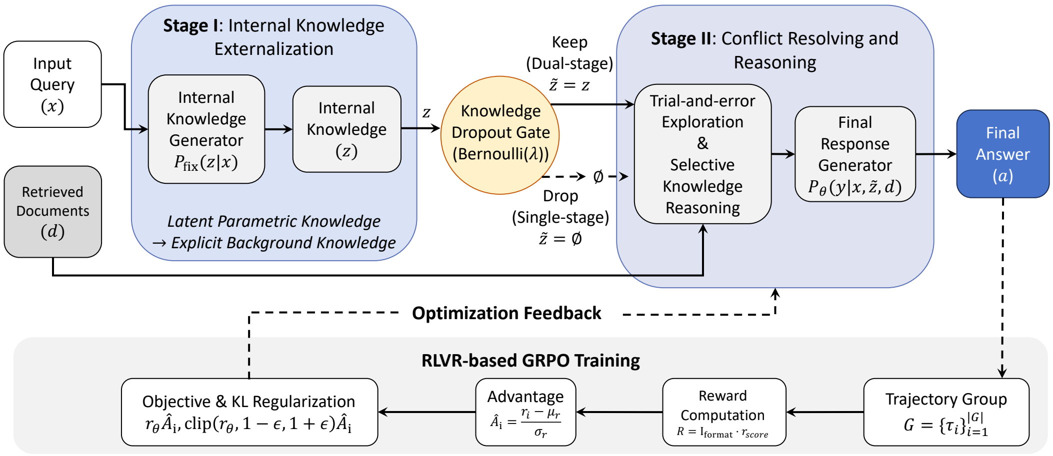

Figure 3: An overview of the BrPr framework. In the 1st stage, the model's prior background knowledge about the query is externalized as a latent variable z. In the 2nd stage, we employ the GRPO algorithm integrated with a Bernoulli-gated dropout mechanism on the prior variable z. This strategy encourages the model to distill the prior-driven reasoning capability into the model parameters.

guided by explicit feedback, the model learns to reconcile between its internal knowledge and external documents.

Second, to improve inference efficiency, we propose distilling the prior-guided reasoning capability into the model's latent reasoning process. Concretely, we introduce a Bernoulli-gated dropout mechanism on the prior variable z during training (Sec. 3.3). By stochastically omitting the explicit prior generation, we encourage the model to implicitly infer z and reason within its latent parameters, effectively consolidating the reasoning process without requiring additional generation steps.

3.1 Stage I: Externalizing Internal Knowledge

The model first externalizes its parametric background knowledge by generating a response $z \sim P_{\text{fix}}(z|x)$. This stage transforms latent parametric knowledge into an explicit textual representation. The prior-guided prompt template is described in Appendix B.2.

The parametric knowledge generator is kept **fixed** as the initialized model throughout the training phase to prioritize reasoning over memorization. It ensures that the model's performance stems from its ability to process external documents in the next stage.

3.2 Stage II: Reasoning under Conflicts

To augment the model's reasoning capabilities, we utilize GRPO for policy optimization within the knowledge-prior framework. Specifically, the model takes the prior knowledge z—obtained from Stage I—alongside the query x and retrieved documents d as inputs. The output trajectory τ consists of a structured reasoning trace and a final answer, following the structure in (Guo et al., 2025).

For each input x and d, GRPO generates a group G of trajectories $\{\tau_i\}_{i=1}^{|G|}$ using the current policy π_θ. It performs multiple rollouts per task and calculates the relative reward within the group as the advantage. GRPO optimizes the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{(x,d)\sim P, z\sim P_{ref}, \tilde{z}\sim \text{Bern}(\lambda; z), \{\tau_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \min \left(r_\theta \hat{A}_i, \text{clip}\left(r_\theta, 1-\epsilon, 1+\epsilon\right) \hat{A}_i \right) \right) \right],$$
$$r_\theta = \frac{\pi_\theta(y_{i,t}|x, \tilde{z}, d, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, \tilde{z}, d, y_{i,<t})} \quad (1)$$

where the prior $\tilde{z}$ is formulated through a Bernoulli-gated dropout mechanism, where $\tilde{z} = z$ with probability λ and $\tilde{z} = \emptyset$ with probability $1 - \lambda$. ϵ is the clipping ratio, and π_{ref} serves as the reference policy—acting as both a regularization constraint and the parametric knowledge generator.

The term $\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r}$ represents the standardized advantage of trajectory τ_i, where r_i is the reward assigned to the trajectory, and μ_r and σ_r are the mean and standard deviation of the rewards within group G, respectively. The total reward R is defined as:

$$R = \mathbb{I}_{\text{format}} \cdot r_{score}, r_{score} = \text{F1}_{score}(a_i, \hat{y}) \quad (2)$$

where $\mathbb{I}_{\text{format}}$ is an indicator function that equals 1 if the trajectory format is correct, and 0 otherwise. The F1_{score} measures the correctness of the model's prediction a_i compared to the ground-truth label $\hat{y}$.

3.3 Bernoulli-gated Knowledge Dropout

To bridge the gap between explicit prior-guided reasoning and implicit latent inference, we treat the prior knowledge z as a stochastic variable. Formally, the sampling process is governed by a Bernoulli distribution $b \sim \text{Bernoulli}(\lambda)$, such that the effective prior $\tilde{z}$ alternates between the explicit prior z and a null state $\emptyset$.

While Eq. 1 provides the operational GRPO objective based on advantage-weighted policy ratios, its optimization can be theoretically interpreted as maximizing a joint log-likelihood. Specifically, the behavior of the model under the Bernoulli gate aligns with the following objective $\mathcal{J}(\theta)$:

$$\begin{aligned}\mathcal{J}(\theta) &= P(\tilde{z} = z) \cdot \mathcal{J}_2(\theta) + P(\tilde{z} = \emptyset) \cdot \mathcal{J}_1(\theta) \\ &= \lambda \cdot \mathcal{J}_2(\theta) + (1 - \lambda) \cdot \mathcal{J}_1(\theta) \end{aligned} \tag{3}$$

where λ represents the probability of retaining z [2], and the objects can be reformulated as:

$$\mathcal{J}_2(\theta) = \log P_\theta(y|x, z, d), \mathcal{J}_1(\theta) = \log P_\theta(y|x, d) \tag{4}$$

Through theoretical derivations in Appendix C, this objective function can be approximated with the KL divergence (Kullback and Leibler, 1951). By taking the expectation of $\mathcal{J}(\theta)$ under the full-information distribution $P(y|x, z, d)$, we obtain:

$$\begin{aligned}\mathbb{E}_{y\sim P}[\mathcal{J}(\theta)] \approx\ & \mathbb{E}[\log P(y|x, z, d)] - \\ & (1 - \lambda) \cdot \text{KL}[P(y|x, z, d)||P(y|x, d)] \end{aligned} \tag{5}$$

By optimizing $\mathcal{L}(\theta)$, the model is encouraged to align the non-prior distribution with the prior-augmented distribution. This alignment facilitates the internalization of prior-guided reasoning within the model's parameters, allowing it to maintain high accuracy when the explicit prior z is omitted.

4 Experiments

4.1 Settings

Dataset Construction. To assess the robustness of LLMs against retrieval noise and conflict, we utilize the test set from RAG-Bench (Fang et al., 2024), which includes the NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013). For each query, this benchmark provides both conflicting documents and noisy documents that do not contain the correct answer. To analyze model performance under varying degrees of interference, we modulate the number of input documents $k \in \{1, 2, 5, 10\}$. In the $k = 1$ configuration, only the gold document is provided. For $k > 1$, additional documents are sampled from the respective noisy and conflicting document pools within RAG-Bench. Further details are provided in Appendix B.1.

To reflect distractor information in real-world retrieval, we employ the pre-processed English Wikipedia dump from December 2018, as released by Karpukhin et al. (2020), as our corpus. We utilize a supervised retriever, DPR (Karpukhin et al., 2020), followed by two distinct reranking methods—BGE-m3 (Chen et al., 2024) and UR³ (Yuan et al., 2024b)—to obtain the top k documents for each query, where $k \in \{5, 20, 50\}$.

We additionally employ the HotpotQA (Yang et al., 2018) and 2Wikimultihopqa (Ho et al., 2020) for a multi-hop reasoning evaluation in Appendix D.

Models. We conduct our experiments using the Qwen2.5-7B-Instruct (Yang et al., 2025) and Qwen3-8B (Team, 2025) models for both training and inference.

Baselines. We first evaluate several **training-free** method: (1) **Direct Inference**: The model generates answers directly based on the provided external knowledge (Lewis et al., 2020). (2) **Prompt-based Methods**: Based on CoT (Wei et al., 2022), we employ three strategies to evaluate model performance: Standard CoT, Joint CoT, and Prior-Guided CoT (see Sec. 2.1). (3) **Decoding-based Methods**: Self-Consistency (SC) (Wang et al., 2023) decoding enhances reasoning via diverse path sampling and majority voting[3]. COIECD (Yuan et al., 2024a) employs a contextual information-entropy constraint to mitigate contextual conflicts by adjusting token probabilities. (4) **Multi Agent Debate**: MADAM-RAG (Wang et al., 2025) uses a multi-agent framework to handle conflicts, misinformation, and noise in retrieved content[4].

Then our evaluation further extends to **training-based** methods: (5) **Fine-tuning**: RALM (Lin

[2] We assign a default value of 0.5 for λ, with a comparative study regarding the impact of varying this parameter detailed in Table 6.

[3] For all evaluations, we employ a sampling size of 10.

[4] For a fair comparison, we adapt the original 72B agent setup to the 7B/8B scale used in our other experiments.

Table 2: EM and F1 scores on the GAR-Bench test set. In the $k = 1$ configuration, only the gold document is provided; for $k > 1$, the input is augmented with documents sampled from the RAG-Bench noisy and conflicting pools. Best results for training-free and training-based models are indicated by underlining and **bolding**, respectively.

	NQ								TriviaQA								WebQ							
	Doc 1		Doc 2		Doc 5		Doc 10		Doc 1		Doc 2		Doc 5		Doc 10		Doc 1		Doc 2		Doc 5		Doc 10	
Method	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Qwen2.5-7B-Instruct																								
w/o parameter update																								
Direct	60.46	69.81	51.75	61.96	38.20	47.73	31.13	39.85	79.18	84.68	74.17	79.56	64.56	70.86	53.54	62.96	40.44	58.39	33.63	52.62	27.93	44.96	20.32	34.63
COIECD	61.26	71.03	53.65	63.43	44.04	54.87	33.63	46.26	78.48	83.99	74.87	80.26	67.57	74.98	52.55	64.32	39.44	58.06	33.13	52.80	27.53	46.52	20.82	41.06
SC	61.96	71.23	55.06	61.71	46.85	56.97	39.94	50.02	81.78	87.05	78.38	83.56	70.87	77.39	62.86	69.98	38.83	57.36	34.83	53.20	31.13	49.74	24.82	42.27
MADAM-RAG	62.05	72.26	53.31	64.79	43.67	55.42	31.65	44.04	75.19	85.55	69.91	79.99	65.19	75.49	55.84	66.90	45.86	62.43	39.18	58.52	32.61	50.59	28.60	45.70
Standard CoT	59.46	69.11	52.35	62.14	42.14	52.79	36.73	47.79	79.68	85.02	76.98	82.39	67.36	74.50	58.76	66.10	38.53	55.24	33.43	51.83	28.63	47.30	23.92	42.75
Joint CoT	57.36	68.18	50.75	60.98	42.94	54.40	38.74	50.34	79.34	84.87	76.18	81.49	70.07	76.49	61.36	69.14	36.34	54.95	34.63	53.13	32.63	51.41	25.22	43.85
Prior-Guided CoT	52.95	62.64	53.85	63.93	48.04	58.28	41.94	52.92	76.87	82.48	77.17	82.35	75.18	81.12	69.37	75.70	37.24	54.81	35.24	53.74	34.03	52.33	28.83	47.58
w/ parameter update																								
RAAT	68.17	74.34	67.07	73.59	63.87	70.30	59.96	67.62	80.35	86.49	79.18	85.40	78.38	84.69	75.28	81.77	47.75	63.53	46.25	61.94	43.64	59.38	42.84	59.44
RALM	66.77	73.30	66.37	72.73	64.76	70.87	61.16	68.54	79.58	85.43	79.08	84.90	76.87	83.00	75.87	82.01	47.34	62.14	46.75	62.38	45.45	60.85	44.34	60.08
KnowPO	66.75	72.58	65.45	70.67	61.24	69.31	56.44	65.51	81.34	86.47	80.54	85.87	78.95	84.73	76.27	83.07	53.85	65.36	51.75	67.64	49.36	64.67	47.20	63.11
GRPO-RAG	69.67	75.31	69.07	75.61	65.37	73.26	62.96	70.67	83.88	88.96	83.34	89.05	82.88	87.72	80.88	86.03	62.66	70.57	62.03	71.42	62.46	72.46	61.96	71.98
BrPr (1-turn)	**69.92**	**75.46**	69.87	76.68	67.17	74.88	66.27	74.33	**85.08**	89.11	84.48	89.03	84.30	**89.23**	81.89	87.32	**64.66**	**73.86**	**63.56**	72.81	62.96	73.23	62.16	72.45
BrPr (2-turn)	69.41	75.25	**70.07**	**77.34**	**68.97**	**75.55**	**66.57**	**74.45**	84.71	**89.41**	**85.49**	**89.14**	**84.38**	88.96	**82.48**	**87.57**	63.95	71.22	63.36	**72.88**	**63.16**	**73.24**	**62.26**	**72.50**
Qwen3-8B																								
w/o parameter update																								
Direct	64.16	72.65	56.56	65.81	45.54	55.29	38.74	49.31	80.08	85.47	77.28	82.40	70.27	76.90	61.76	68.74	43.54	59.73	38.44	54.89	31.03	48.49	26.73	44.37
COIECD	63.96	72.68	56.66	65.43	46.05	55.67	37.34	48.09	81.28	86.13	77.98	82.90	70.77	77.63	61.96	68.71	43.24	59.66	38.84	54.72	30.33	47.54	26.23	43.83
SC	66.47	74.94	61.46	71.04	52.25	60.91	47.15	56.56	86.59	91.27	82.98	88.33	79.68	85.53	70.42	78.90	44.14	60.98	40.04	57.25	35.44	53.07	30.63	47.45
MADAM-RAG	66.36	74.06	57.22	68.22	45.74	60.43	40.28	49.27	84.41	90.06	80.83	86.96	78.60	86.04	69.13	78.42	49.57	65.26	43.36	58.60	33.12	54.87	31.04	46.81
Standard CoT	64.76	72.89	58.36	67.15	47.04	55.67	41.64	49.65	83.18	87.84	80.88	85.85	73.97	79.55	66.07	71.47	41.94	59.25	36.94	54.13	31.35	49.70	27.93	43.92
Joint CoT	62.06	71.30	57.26	66.72	48.35	57.59	42.64	51.59	84.48	89.16	80.18	85.54	76.27	81.20	68.97	74.77	40.24	58.39	37.84	55.73	32.63	50.77	29.03	45.13
Prior-Guided CoT	56.46	66.23	52.95	62.05	48.94	58.22	45.45	55.39	82.78	87.71	81.28	86.13	80.28	84.99	72.47	78.55	38.64	55.78	35.94	52.99	33.23	50.27	31.53	47.86
w/ parameter update																								
RAAT	67.87	74.50	66.97	72.92	64.86	71.85	58.26	65.66	80.68	86.35	80.18	85.71	79.88	85.69	76.48	83.12	49.75	65.07	48.95	64.20	45.75	62.13	44.64	60.94
RALM	69.17	**75.23**	68.77	74.12	66.17	72.30	63.67	69.59	80.18	85.87	80.78	86.04	79.97	85.13	77.34	84.50	51.05	66.28	50.45	65.17	49.55	64.04	48.95	64.06
KnowPO	67.34	74.02	66.42	72.39	65.48	72.11	62.90	68.73	81.44	86.38	81.03	85.75	80.33	85.27	78.94	84.86	60.53	69.77	59.83	68.05	55.45	65.21	53.91	63.38
GRPO-RAG	68.17	74.81	65.37	72.44	64.46	71.23	63.86	71.58	83.96	88.93	82.37	87.47	81.48	86.64	80.53	86.07	66.47	73.95	66.07	73.52	63.16	71.01	62.97	70.84
BrPr (1-turn)	**69.31**	75.19	68.57	74.62	**68.31**	**75.07**	66.06	73.72	**85.70**	**90.27**	**85.69**	**90.17**	82.38	87.98	81.68	86.71	**67.86**	**74.12**	66.47	74.10	64.86	73.84	64.06	**73.45**
BrPr (2-turn)	69.03	74.78	**69.40**	**75.21**	67.67	74.02	**67.28**	**74.92**	84.98	89.53	85.83	89.24	**84.57**	**88.02**	**82.75**	**87.12**	67.40	73.69	**67.57**	**75.45**	**65.37**	**73.54**	**64.36**	73.20

et al., 2024) involves fine-tuning on samples augmented with diverse retrieval noise. RAAT (Fang et al., 2024) utilizes adversarial training to distinguish between gold documents and noisy distractors. CARE-RAG (Chen et al., 2025) enhances model robustness through the conflict-driven summarization of all provided evidence[5]. (6) **Preference Optimization**: KnowPO (Zhang et al., 2025) achieves adaptive knowledge selection through a knowledge-aware preference optimization strategy. (7) **Reinforcement Learning**: GRPO-RAG employs rule-based RL to encourage the model to perform explicit reasoning prior to generating an answer. The training process is conducted using a Standard CoT strategy.

Notably, all trainable models are trained exclusively on the NQ training set from RAG-Bench to evaluate their generalization capabilities across different benchmarks. Our method can be implemented using two strategies: **BrPr (1-turn)**, which utilizes a Standard CoT prompt, and **BrPr (2-turn)**, which employs the Prior-Guided CoT strategy. The implementation details of BrPr are provided in the Appendix E.

Metrics. We use the Exact Match (EM) and F1 scores for evaluating the performance of LLMs.

[5]As the training data and model weights are not publicly available, we report the results directly from the original paper in Table 3 for comparison.

4.2 Overall Performance

As shown in the Table 2, both variants of BrPr (1-turn and 2-turn) consistently outperform all baselines across all tested models.

Importance of Prior Guidance. Among methods without parameter updates, Prior-Guided CoT exhibits superior scalability compared to heavy-sampling or multi-agent approaches like SC and MADAM-RAG. As the context expands from 1 to 10 documents, the performance of other methods often declines sharply due to the accumulation of retrieval noise and conflicting information. In contrast, Prior-Guided CoT consistently maintains higher scores in high-noise environments, demonstrating that explicit prior guidance is effective for stabilizing model reasoning.

Robustness to Noise and Conflicts. While baselines such as RAAT and KnowPO struggle to handle the increased complexity of multi-document settings, BrPr consistently yields the highest metrics regardless of the document count or interference level. For instance, on the NQ (Doc 10) benchmark, BrPr-7B (2-turn) achieves an EM of 66.57%, significantly outperforming other models like GRPO-RAG (62.96%). This sustained performance across conflicts and noises underscores the effectiveness of our method and strengthening the model's core reasoning capabilities.

Table 3: EM and F1 scores are reported on datasets retrieved from Wikipedia. All models are trained based on Qwen2.5-7B-Instruct. The best-performing results are highlighted in **bold**.

Retriever	Generator	NQ						TriviaQA						WebQ					
		Top 5		Top 20		Top 50		Top 5		Top 20		Top 50		Top 5		Top 20		Top 50	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
(Oracle) DPR w/o reranker	CARE-RAG*	62.20	62.20	-	-	-	-	75.40	75.70	-	-	-	-	-	-	-	-	-	-
	RAAT	57.05	64.72	49.58	57.68	43.41	51.30	75.11	81.04	70.32	76.46	62.92	68.93	42.17	58.13	37.20	52.90	34.71	49.68
	RALM	56.41	64.52	49.42	57.46	45.16	53.26	75.52	81.25	70.61	76.27	66.40	72.32	43.07	57.70	38.53	53.22	38.46	52.78
	KnowPO	57.34	65.97	50.68	58.37	45.60	53.85	76.38	82.09	72.56	79.48	66.79	73.03	50.83	62.94	44.52	58.77	42.59	54.31
	GRPO-RAG	58.79	66.59	52.91	60.97	47.34	55.89	79.52	84.49	77.02	82.27	72.83	76.88	54.66	66.30	50.59	61.40	47.03	59.08
	BrPr (1-turn)	59.91	67.41	**53.24**	61.35	48.81	57.55	78.94	84.25	76.04	81.39	72.74	77.23	54.81	**66.65**	**50.99**	62.41	48.28	59.74
	BrPr (2-turn)	**61.52**	**68.10**	53.20	**61.43**	**49.40**	**58.15**	**81.10**	**86.11**	**78.12**	**83.05**	**74.55**	**79.75**	**55.87**	66.53	50.40	**62.43**	**49.91**	**60.33**
(Normal) DPR w/ BGE reranker	RAAT	42.50	50.69	41.37	49.62	37.76	46.03	61.97	69.54	60.53	67.65	53.60	60.83	28.94	42.92	29.48	43.50	27.31	41.41
	RALM	41.72	50.58	40.13	48.78	39.61	48.43	62.30	69.47	61.13	67.91	57.67	64.70	29.77	43.52	30.12	43.64	30.66	44.22
	KnowPO	41.34	50.74	40.94	49.03	40.57	49.29	63.45	70.74	61.58	68.10	58.97	65.46	32.44	45.81	31.82	44.39	30.53	44.24
	GRPO-RAG	43.00	52.11	43.43	52.45	41.89	51.39	65.53	72.32	64.93	71.55	62.63	69.31	37.29	49.92	38.98	49.76	38.98	49.47
	BrPr (1-turn)	43.57	52.09	42.47	51.48	41.66	51.04	66.13	73.85	**65.47**	**72.02**	63.03	70.00	37.84	49.76	**39.03**	**50.61**	**39.71**	49.03
	BrPr (2-turn)	**45.29**	**54.38**	**43.77**	**52.62**	**42.44**	**52.15**	**66.97**	**74.02**	65.33	71.82	**63.73**	**70.44**	**37.99**	**50.01**	38.70	49.93	39.07	**50.57**
(Normal) DPR w/ UR³ reranker	RAAT	41.72	50.31	42.23	50.07	36.93	44.62	62.83	70.18	60.93	67.94	54.23	61.26	29.13	43.44	29.43	43.68	27.76	41.32
	RALM	40.33	49.43	40.67	49.27	38.72	47.33	62.97	69.88	61.53	68.40	58.19	65.20	29.72	43.23	30.27	44.20	30.61	44.25
	KnowPO	42.52	51.07	41.92	50.86	40.37	49.27	64.35	71.47	62.82	69.46	60.82	66.43	34.57	45.69	34.96	45.81	32.51	46.03
	GRPO-RAG	43.13	51.88	44.27	53.08	41.87	50.72	66.03	72.82	65.20	71.86	63.77	70.16	38.48	49.93	38.81	50.53	37.92	48.91
	BrPr (1-turn)	43.50	52.12	44.59	**53.32**	**42.73**	**51.78**	66.70	73.28	66.07	73.00	63.54	70.48	38.29	50.21	38.63	50.46	**39.07**	**50.38**
	BrPr (2-turn)	**44.20**	**53.34**	**44.62**	53.27	42.03	51.26	**67.41**	**74.21**	**66.71**	**73.41**	**64.31**	**70.62**	**38.53**	**50.55**	**39.12**	**50.81**	38.61	49.59

* As the training data and model weights are not publicly available, we report the results directly from the original publication.

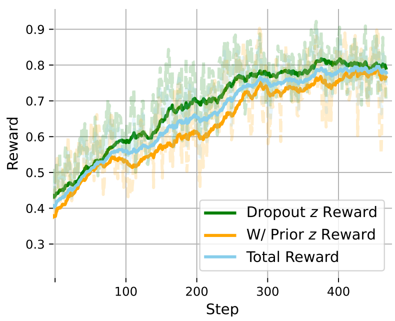

Figure 4: Training rewards under explicit or absent prior within bernoulli-gated mechanism.

Table 4: Performance comparison under no-document and all-incorrect-document settings on BGE-reranked NQ benchmark.

Methods	Top 5		Top 10		Top 20		Top 50	
	EM	F1	EM	F1	EM	F1	EM	F1
Qwen2.5-7B-Instruct								
Parametric Answer (w/o doc)	21.07	30.35						
W/o gold document	12.40	21.77	9.66	19.68	9.82	19.29	9.95	19.74
GRPO-RAG-7B								
Parametric Answer (w/o doc)	25.23	33.28						
W/o gold document	16.17	25.96	16.96	26.50	17.13	26.66	15.00	25.28
BrPr-7B								
Parametric Answer (w/o doc)	25.70	34.05						
W/o gold document (1-turn)	17.30	26.87	17.63	27.69	18.29	28.03	15.52	25.48
W/o gold document (2-turn)	20.67	30.32	20.95	30.49	21.45	31.02	19.24	28.91

The Effects of KL Divergence. As shown in Table 2, the 1-turn model yields performance competitive with the explicit 2-turn model, thereby demonstrating the effectiveness of our training objective in implicitly regularizing the distribution via KL divergence. Specifically, under the gold setting, the 2-turn model exhibits inferior performance compared to the one-turn variant, mirroring the trend observed in training-free methods. This result indicates that the divergence constraint merely reduces the distributional discrepancy between the two models, rather than achieving full alignment. As shown in Figure 4, by stochastically omitting the prior knowledge z during training, the model learns to implicitly infer factual priors within its internal representations, resulting in steady performance gains for the 1-turn variant.

4.3 Performance on Retrieved Documents

We evaluate the models under two distinct configurations: **Normal** and **Oracle**. In the **Normal** setting, the model retrieves top-k documents that do not necessarily include the gold document. Conversely, in the **Oracle** setting, the gold document is guaranteed to be included within the retrieved set.

Generalization Across Diverse Retrieval. The experimental results in Table 3 across different retrieval configurations demonstrate the robust generalization of the BrPr framework. Whether utilizing the *Oracle* or *Normal* retrieval settings, BrPr consistently achieves the highest performance across all benchmarks. As the retrieval depth increases from Top 5 to Top 50 documents, BrPr (2-turn) maintains a substantial performance margin over strong baselines like GRPO-RAG and KnowPO.

Robustness to Real-World Retrieval Noise. The results of *Normal* setting reveals that BrPr is particularly effective at mitigating the performance degradation typically associated with lower-quality retrieval. While all models experience a decline in performance when moving from *Oracle* to *Normal* retrieval settings, both BrPr variants exhibit significantly higher resilience compared to other methods. This indicates that our method's reasoning process is robust enough to handle the increased noise and conflicting information in large-scale, real-world retrieval scenarios.

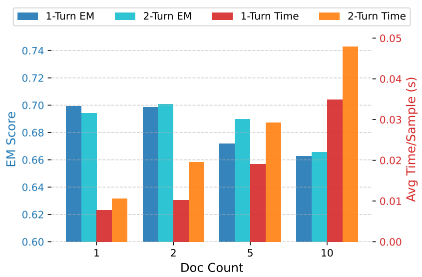

Figure 5: Comparison of BrPr-7B on the NQ test set (RAG-Bench). Accuracy is measured by EM score, while efficiency is evaluated based on inference latency.

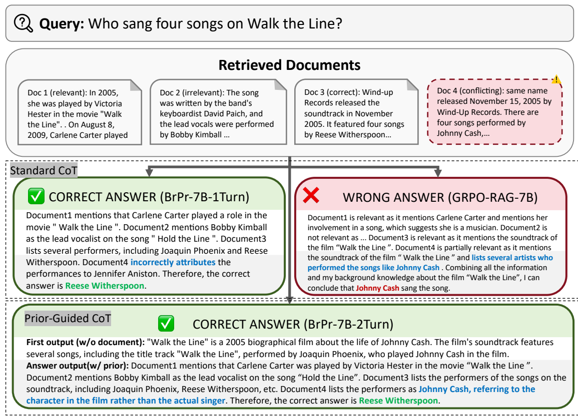

Figure 6: Comparative visualization of reasoning paths. Specifically, in the Standard CoT setting, GRPO-RAG is misled by conflicting external evidence, whereas BrPr leverages implicit internal knowledge to identify the correct answer.

Performance under Incorrect Contexts. Table 4 reports the performance of three models when no external documents are provided and when all retrieved documents are incorrect. RL–based training methods improve the model's ability to understand and utilize parametric knowledge, leading to a modest increase in the accuracy of parametric answers. BrPr demonstrates a stronger capability to select correct parametric knowledge in the presence of incorrect documents, particularly under the 2-turn strategy. These findings suggest that our approach effectively reconciles parametric priors with information from external documents.

4.4 Accuracy and Efficiency Comparison

In this section, we evaluate the impact of different BrPr variants on both QA performance and computational efficiency. As illustrated in Figure 5, both variants exhibit a marginal degradation in performance as the number of input documents increases. Conversely, while inference latency per query scales linearly with the document count for both methods, the 1-turn variant maintains highly competitive reasoning stability while simultaneously mitigating inference latency, achieving a temporal reduction on the order of 10^{-2} seconds.

4.5 Case study

We present a qualitative comparison between BrPr and a RL-trained model (GRPO-RAG) in Figure 6. It demonstrates that BrPr effectively resolves document conflicts and noise by utilizing parametric knowledge as a semantic anchor, whereas GRPO-RAG remains susceptible to external inconsistencies. A detailed analysis with complete input–output examples is provided in Appendix G.

5 Related Work

While prior work has made strides in addressing parametric and single-document conflicts (Zhou et al., 2023; Yuan et al., 2024a), recent studies highlight the difficulty of resolving disagreements across multiple documents (Su et al., 2024). Most approaches optimized the contextual knowledge prioritization in LLMs via supervised fine-tuning strategy (Yoran et al., 2024; Fang et al., 2024). And KnowPO (Zhang et al., 2025) aimed to achieve adaptive knowledge selection based on contextual relevance by preference optimization strategy. More recently, MADAM-RAG (Wang et al., 2025) and DRAG (Hu et al., 2025b) introduced a multi-agent framework to mitigate inter-document conflicts and misinformation. Furthermore, CARE-RAG (Chen et al., 2025) derived parameter-aware evidence by comparing internal records, but its efficacy remains heavily contingent upon a pre-trained summary generator, which may limit its flexibility and overall performance.

DeepRAG (Guan et al., 2025) and R1-Searcher++ (Song et al., 2025) are motivated by similar objectives; however, R1-Searcher++ emphasizes retrieval efficiency, whereas DeepRAG focuses on structured, adaptive retrieval. In contrast, BrPr addresses a distinct failure mode in RAG, namely performance instability under conflicting documents. Consequently, the retrieval action in our framework is fixed—meaning the external contexts remain unchanged during experiments—to specifically improve the model's discriminative capability and accuracy under conflicts.

6 Conclusion

We empirically demonstrate that prior-guided reasoning substantially mitigates the impact of external distractors. Building on this, this paper proposes BrPr method, which achieves robust and efficient performance across conflicts and noises. Experimental results demonstrate that BrPr significantly improves performance and stability across varying degrees of external inconsistency.

Limitations

While the Bernoulli-gated mechanism and the implicit KL divergence constraint effectively bridge the performance gap between the 1-turn and 2-turn strategies, the 1-turn model's performance is inherently upper-bounded by its explicit counterpart. Our empirical results suggest that the KL divergence serves as a distributional alignment tool that facilitates approximation rather than transcendence. Consequently, while the 1-turn model achieves superior efficiency, it fails to surpass the 2-turn model in reasoning depth, indicating that implicit latent optimization may not fully capture the entire expressive power provided by explicit, step-by-step reasoning sequences.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (L243006), the National Natural Science Foundation of China (No. U24A20335), the independent research project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Juan Chen, Baolong Bi, Wei Zhang, Jingyan Sui, Xiaofei Zhu, Yuanzhuo Wang, Lingrui Mei, and Shenghua Liu. 2025. Rethinking all evidence: Enhancing trustworthy retrieval-augmented generation via conflict-driven summarization. *CoRR*, abs/2507.01281.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory (2. ed.)*. Wiley.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieve step by step for large language models. *arXiv preprint arXiv:2502.01142*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. 2025a. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *Preprint*, arXiv:2405.11143.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. 2025b. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15839–15853. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. RA-DIT: retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Aigerim Mansurova, Aiganym Mansurova, and Aliya Nugumanova. 2024. Qa-rag: Exploring llm reliance on external knowledge. *Big Data and Cognitive Computing*, 8(9):115.

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. *Preprint*, arXiv:2402.03300.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence. *CoRR*, abs/2504.13079.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024.

Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

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. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024a. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3903–3922. Association for Computational Linguistics.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Jun Zhao, and Kang Liu. 2024b. Improving zero-shot LLM reranker with risk minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17967–17983.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25895–25903.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain qa. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14544–14556. Association for Computational Linguistics.

A LLM Usage Disclosure

We use LLM for paper writing to check grammar and boost the clarity. We do not use LLM for research or generate experiment code.

B Experimental Settings

B.1 Detailed Description of RAG-Bench

The test set of RAG-Bench (Fang et al., 2024) incorporates 1,000 samples from each of three subsets: NQ, TriviaQA, and WebQ. For each query, the benchmark provides multiple noisy and conflicting retrieval contexts. Specifically, **noisy contexts** consist of documents that are topically relevant to the query but do not contain the correct answer. **Conflicting contexts** are constructed by selecting one of the two gold passages and substituting the correct answer entity with an incorrect one.

For Section 4, the RAG-Bench training set comprises 4.5k samples (1.5k per subset), where each instance includes a gold retrieval context augmented with three distractors: two noisy documents and one conflicting document. In our experiments, we utilize **only the NQ training subset**. To analyze model performance across varying context lengths, we sample 1.5k instances for each document count $k \in \{1, 2, 3, 4\}$, resulting in a total of 6k training samples. While the $k = 1$ configuration provides only the gold document, settings where $k > 1$ incorporate a mixture of the gold document and noisy distractors. This allows for a robust evaluation of the model's performance across varying levels of contextual interference. In this setup, each unique query is replicated four times, each time paired with a different number of documents. For a fair comparison, all baseline methods follow this identical training configuration.

B.2 Instruction Template

We adopt the following instruction template for each prompting strategy:

Parametric Answer: $P_\theta(a, c|x)$, generating the answer without documents.

Parametric Answer

System
You are a helpful assistant that answers factual questions.
User
You are given a query. Please use your background knowledge to answer the query.
Format your response strictly as follows:
`<think> [step-by-step reasoning] </think>`
`<answer> [final answer in one or two words] </answer>`
Input
[query]: $\{query\}$
Assistant

Direct Inference: $P_\theta(a|x, d)$, generating the answer directly without reasoning.

Direct Inference

System
You are a helpful assistant that answers factual questions.
User
You are given a query and several documents.
Use the provided documents combined with your background knowledge to determine the correct factual answer.
Format your response strictly as follows:
`<answer> [final answer in one or two words] </answer>`
Input
[query]: $\{query\}$
[documents]: $\{docs\}$
Assistant

Standard CoT: $P_\theta(a, c|x, d)$, incorporating a reasoning chain without explicit internal knowledge.

Standard CoT

System
You are a helpful assistant that answers factual questions.
User
You are given a query and several documents.
Use the provided documents combined with your background knowledge to determine the correct factual answer.
Format your response strictly as follows:
`<think> [step-by-step reasoning] </think>`
`<answer> [final answer in one or two words] </answer>`
Input
[query]: $\{query\}$
[documents]: $\{docs\}$
Assistant

Joint CoT: $P_\theta(a, c, z|x, d)$, where background knowledge z is generated within a unified reasoning chain.

Joint CoT

System
You are a helpful assistant that answers factual questions.
User
You are given a query and several documents.
First, please provide relevant background knowledge based on query using your own parametric knowledge.
Then use the provided documents combined with your background knowledge to determine the correct factual answer.
Format your response strictly as follows:
`<background> [background knowledge] </background>`
`<think> [step-by-step reasoning] </think>`
`<answer> [final answer in one or two words] </answer>`
Input
[query]: $\{query\}$
[documents]: $\{docs\}$
Assistant

Prior-Guided CoT: $P_\theta(a, c|x, z, d)$, where internal knowledge z is elicited in an initial stage $z \sim P_\theta(z|x)$ as prior information.

Prior-Guided CoT (1st Stage)

System
You are a helpful assistant that answers factual questions.
User
Please provide relevant background knowledge based on query using your own parametric knowledge.
Format your response strictly as follows:
`<think> [background knowledge] </think>`
Input
[query]: $\{query\}$
Assistant

Prior-Guided CoT (2nd Stage)

User
You are given a query and several documents.
Use the provided documents combined with your background knowledge to determine the correct factual answer.
Format your response strictly as follows:
`<think> [step-by-step reasoning, combining parametric knowledge and the provided documents] </think>`
`<answer> [final answer in one or two words] </answer>`
Input
[query]: $\{query\}$
[documents]: $\{docs\}$
Assistant

B.3 Analysis on CMI

Experiments are conducted on the Natural Questions subset of the RAG-Bench test suite, utilizing a document retrieval depth of $k \in \{1, 2, 5, 10\}$. As illustrated by the x-axis in Figure 2, we stratify the data into four distinct partitions based on the F1 score transitions between the baseline distribution, $P(a|x, d)$, and the prior-guided distribution, $P(a|x, z, d)$. For this binary classification of correctness, a prediction is deemed correct if its F1 score relative to the ground truth exceeds a threshold of 0.5; otherwise, it is classified as incorrect.

C Theoretical Derivations

C.1 The Gradient of the GRPO Objective

The GRPO objective in Eq. 1 utilizes a clipped surrogate loss. In the neighborhood of the current policy $\pi_{\theta_{old}}$, and assuming the clipping parameter ϵ is not triggered (i.e., for small updates), the gradient of the objective with respect to θ can be approximated by the standard policy gradient:

$$\nabla_\theta \mathcal{L}(\theta) \approx \mathbb{E}_{(x,d),z,\tilde{z}\sim\text{Bern}(\lambda),\{y_i\}_{i=1}^G\sim\pi_{\theta_{old}}} \left[\sum_{i=1}^{G} \hat{A}_i \nabla_\theta \log \pi_\theta(y_i|x, \tilde{z}, d)\right] \quad (6)$$

where $\hat{A}_i$ is the advantage of the i-th trajectory within the group G.

C.2 Marginalization over the Bernoulli Gate

The variable $\tilde{z}$ is sampled from a Bernoulli distribution such that $\tilde{z} = z$ with probability λ and $\tilde{z} = \emptyset$ with probability $1 - \lambda$. Using the *Law of Total Expectation*, we can decompose the objective:

$$\mathcal{J}(\theta) = P(\tilde{z} = z) \cdot \mathbb{E}_{\pi_\theta}[\hat{A} \mid \tilde{z} = z]+ P(\tilde{z} = \emptyset) \cdot \mathbb{E}_{\pi_\theta}[\hat{A} \mid \tilde{z} = \emptyset] \quad (7)$$

Then,

$$\mathcal{J}(\theta) = \lambda \cdot \mathbb{E}_{y\sim\pi_\theta(\cdot|x,z,d)}[\hat{A}] + (1 - \lambda) \cdot \mathbb{E}_{y\sim\pi_\theta(\cdot|x,\emptyset,d)}[\hat{A}] \quad (8)$$

C.3 Mapping Advantages to Log-Likelihoods

According to Eq. 2, the reward R (and the advantage $\hat{A}$) is a function of the correctness (F1 score) of the generated output y relative to the ground-truth $\hat{y}$. In the optimal limit of RL, maximizing the expected advantage $\mathbb{E}_{y\sim\pi_\theta}[\hat{A}]$ where the reward is sparse and peaked at the ground-truth y^* is mathematically equivalent to maximizing the log-likelihood of the ground-truth under the policy:

$$\arg\max_\theta \mathbb{E}_{y\sim\pi_\theta}[\hat{A}] \implies \arg\max_\theta \log P_\theta(y^*|\text{context})$$

Thus, we define the corresponding terms as negative log-likelihoods:

(i) For the case $\tilde{z} = z$: $\mathcal{J}_2 = \log P_\theta(y|x, z, d)$

(ii) For the case $\tilde{z} = \emptyset$: $\mathcal{J}_1 = \log P_\theta(y|x, d)$

C.4 Theoretical Objective

Combining the results from C.2 and C.3, the expected behavior of the model optimization under the Bernoulli gate aligns with the weighted sum of these likelihoods:

$$\mathcal{J}(\theta) = P(\tilde{z} = z) \cdot \mathcal{J}_2(\theta) + P(\tilde{z} = \emptyset) \cdot \mathcal{J}_1(\theta) \quad (9)$$

Decompose the log-likelihood term $\mathcal{J}_1$ as follows:

$$\log P_\theta(y|x, d) = \log P(y|x, z, d) - \log \frac{P(y|x, z, d)}{P(y|x, d)}$$

Then, we obtain:

$$\begin{aligned}
&\mathcal{J}(\theta) \\
&= \lambda\mathcal{J}_2 + (1 - \lambda)\left(\mathcal{J}_2 - \log \frac{P(y|x, z, d)}{P(y|x, d)}\right) \\
&= [\lambda + (1 - \lambda)] \log P(y|x, z, d) \\
&\quad - (1 - \lambda) \log \frac{P(y|x, z, d)}{P(y|x, d)} \\
&= \log P(y|x, z, d) - (1 - \lambda) \log \frac{P(y|x, z, d)}{P(y|x, d)}
\end{aligned} \quad (10)$$

The KL divergence is defined as:

$$\text{KL}(P||Q) = \mathbb{E}_{y\sim P}\left[\log \frac{P(y)}{Q(y)}\right] \quad (11)$$

By taking the expectation of $\mathcal{J}(\theta)$ with respect to the full-information distribution $P(y|x, z, d)$, we obtain:

$$\mathbb{E}_{y\sim P}[\mathcal{J}(\theta)] \approx \mathbb{E}[\log P(y|x, z, d)] - (1 - \lambda) \cdot \text{KL}[P(y|x, z, d)||P(y|x, d)] \quad (12)$$

The first term ensures the model maintains high predictive fidelity under prior guidance, while the second term—the KL divergence—serves as a distributional constraint. This formulation encourages the model to internalize the explicit reasoning process into its latent parameters, effectively bridging the performance gap between the prior-guided (2-turn) and standard (1-turn) configurations.

D Performance on Multi-hop QA

To evaluate the generalization of our method on multi-hop reasoning tasks, we conduct experiments on the development sets of HotpotQA (Yang et al., 2018) (7k samples) and 2WikiMultihopQA (Ho et al., 2020) (12k samples). For each query, we

Table 5: Performance on Multi-hop QA Datasets. All models are trained on Qwen2.5-7B-Instruct.

Method	**HotpotQA (Doc 10)**		**2Wiki (Doc 10)**	
	EM	F1	EM	F1
Para. Answer	16.09	21.29	18.03	23.01
RAAT	28.51	38.44	26.94	35.29
RALM	28.82	38.55	30.12	38.69
KnowPO	29.10	39.72	29.46	38.03
GRPO-RAG	32.86	42.94	**31.74**	**41.05**
BrPr (1-turn)	30.94	41.40	29.02	39.49
BrPr (2-turn)	**33.01**	**43.88**	31.52	40.77

provide the model with all available documents in the original datasets.

Table 5 shows that BrPr (1-turn) underperforms relative to the 2-turn version and GRPO-RAG in multi-hop tasks. This disparity primarily stems from the inherent complexity of multi-hop tasks, which necessitate multiple intermediate reasoning steps; compressing such intricate logic into latent parameters within a 1-turn framework inevitably incurs information loss. Additionally, the low accuracy of the model's parametric answer acts as a bottleneck for BrPr, preventing it from attaining superior results over the GRPO-RAG baseline.

E Implementation Details

We employ Qwen2.5-7B-Instruct and Qwen3-8B as the initial models. We utilize the OpenRLHF (Hu et al., 2025a) framework for training. GRPO (Shao et al., 2024) is used as the reinforcement learning algorithm. We use the train set of NQ from RAG-Bench (Fang et al., 2024) to construct training datasets. We set the number of rollouts as 16 for one task. We set the learning rate as 5e-7, batch size as 32, training steps as 480. We set λ as 0.5 in Bernoulli-gated Knowledge Dropout. We conduct the ablation study regarding the impact of varying this parameter detailed in Table 6. We use 8 A100 GPUs for all the experiments.

All models employ identical sampling settings during inference, with a temperature of 0.6 and a top-p value of 0.95.

E.1 Ablation Study

As demonstrated in Table 6, optimal performance is achieved when λ is set to 0.5. Notably, we observe that while the removal of the knowledge gate leads to a decrease in 1-turn reasoning performance, its efficacy still closely rivals that of the GRPO-RAG model. A comprehensive analysis of this phenomenon is provided in Appendix F, where we demonstrate that the optimization of prior-guided strategy can implicitly bolster the performance of direct, 1-turn inference.

Table 6: Ablation study on the BrPr method. Evaluation is conducted on retrieved NQ dataset in Oracle setting. All method are compared in standard CoT strategy.

Method	**Top 5**		**Top 20**		**Top 50**	
	EM	F1	EM	F1	EM	F1
BrPr-7B ($\lambda = 0.5$)	**59.91**	**67.41**	**53.24**	**61.35**	**48.81**	**57.55**
GRPO-RAG	58.79	66.59	52.91	60.97	47.34	55.89
Dropout Variants						
$\lambda = 0.8$	58.84	66.93	52.79	60.94	49.23	58.20
$\lambda = 0.2$	59.75	67.37	53.04	61.22	48.55	57.03
Training Variants						
w/o knowledge gate	59.28	67.05	52.37	60.84	47.89	56.33

F Theoretical Analysis for Optimization of Prior-Guided Reasoning

The prediction task in standard RAG paradigm can be viewed as a marginalization over the intermediate latent variable z:

$$P_\theta(y|x,d) = \int P_\theta(y|x,z,d)P_\theta(z|x,d)dz \quad (13)$$

where $P_\theta(z|x,D)$ represents the prior distribution of the model's internal knowledge. To optimize the parameters θ, we consider the gradient of the log-marginal likelihood. Applying the chain rule and the log-derivative trick ($\nabla_\theta P = P\nabla_\theta \log P$), the gradient can be expanded as follows:

$$\begin{aligned} \nabla_\theta \int P_\theta(y|x,z,d)P_\theta(z|x,d)dz = \\ \int \big[\nabla_\theta P_\theta(y|x,z,d) \cdot P_\theta(z|x,d) + \\ P_\theta(y|x,z,d) \cdot \nabla_\theta P_\theta(z|x,d)\big]dz \end{aligned} \quad (14)$$

The first term optimizes the generative parameters given the prior, while the second term refines the prior parameters themselves. Assuming the internal knowledge prior $P_\theta(z|x,D)$ remains relatively stable (w/o training), we focus on the first term:

$$\nabla_\theta \log P_\theta(y|x,d) = \int \frac{P_\theta(y|x,z,d)P_\theta(z|x,d)}{P_\theta(y|x,d)} \nabla_\theta \log P_\theta(y|x,z,d)dz \quad (15)$$

By applying Bayes' rule, the ratio in the integrand simplifies to the posterior distribution $P_\theta(z|y,x,D)$. Consequently, the gradient of the marginal likelihood is equivalent to the expected gradient of the conditional likelihood under the posterior:

$$\nabla_\theta \log P_\theta(y|x,d) = \mathbb{E}_{z\sim P_\theta(z|y,x,d)} [\nabla_\theta \log P_\theta(y|x,z,d)] \quad (16)$$

This derivation proves that optimizing the model to generate y given z effectively enhances the direct 1-turn performance $P_\theta(y|x,d)$ by utilizing the posterior distribution as a weighting mechanism for training.

G Case Study

Figure 6 and Table 7 provide a qualitative comparison between BrPr and GRPO-RAG. In the Standard CoT strategy, the GRPO-RAG model erroneously relies on conflicting information within external documents, resulting in an incorrect prediction. Conversely, BrPr demonstrates the capacity to implicitly leverage parametric knowledge acquired during training to navigate these conflicts—such as correctly identifying misattributions (e.g., "*incorrectly attributes the performances to Jennifer Aniston*")—to obtain the correct answer. Similarly, in the Prior-Guided CoT strategy, the provision of explicit background knowledge enables the model to effectively differentiate between conflicting and noisy documents (e.g., recognizing that "*Johnny Cash*" denotes the cinematic character rather than the actual singer), thereby facilitating more robust reasoning.

Query: Who sang the songs on walk the line?

Document 1: tape collection in Hendersonville and uncovered in 2003. It was then remastered by her brother John Carter Cash. In the remastered version John added his and his wife Laura's (her sister-in-law) backing vocals and a guest appearance from Carlene herself—more than 25 years after she wrote and recorded the song. On June 26, 2001, a New Mexico police officer pulled over Carter and Epstein. A search of the vehicle revealed that it was stolen and drugs were found. In 2005, she was played by Victoria Hester in the movie ""Walk the Line"". On August 8, 2009, Carlene Carter played a

Document 2: "Hold the Line" is a song by the American rock band Toto. The song was written by the band's keyboardist David Paich, and the lead vocals were performed by Bobby Kimball. The song was released as the band's debut single, and was featured on their debut 1978 eponymous album. The song was a huge success in the U.S.; it reached number 5 on the ""Billboard"" Hot 100 chart during the winter of 1978–79, and number 14 on the official UK chart. Jeff Porcaro, the band's drummer, gave a definition for the song: ""'Hold the Line' was a

Document 3: The feature on disc one is 17 minutes longer than the theatrical release, and disc two features eight extended musical sequences with introductions and documentaries about the making of the film. The film has been released on Blu-ray Disc in France, Sweden and the UK in the form of its extended cut. The American Blu-ray features the shorter theatrical cut. Wind-up Records released the soundtrack in November 2005. It featured nine songs performed by Joaquin Phoenix, four songs by Reese Witherspoon, two songs by Tyler Hilton, and one song each by Waylon Payne, Johnathan Rice, and Shooter Jennings. The album

Document 4: Walk the Line (soundtrack) Walk the Line: Original Motion Picture Soundtrack is the soundtrack album to the 2005 biographical drama film of the same name released November 15, 2005 by Wind-Up Records. There are four songs by June Cash, one song by Jerry Lee Lewis, one song by Roy Orbison, two songs by Elvis Presley, and one song by Waylon Jennings. At the Golden Globe Awards, Jennifer Aniston was awarded the Best Actor - Musical

GRPO-RAG Output:

<think> Document1 is relevant as it mentions Carlene Carter and mentions her involvement in a song, which suggests she is a musician. Document2 is not relevant as it mentions the song without the singer. Document3 is relevant as it mentions the soundtrack of the film "Walk the Line". Document4 is partially relevant as it mentions the soundtrack of the film "Walk the Line" and lists several artists who performed the songs like Johnny Cash. Combining all the information and my background knowledge about the film "Walk the Line", I can conclude that Johnny Cash sang the song. </think>

<answer> Johnny Cash </answer>

BrPr 1-Turn Output:

<think> Document1 mentions that Carlene Carter played a role in the movie "Walk the Line". Document2 mentions Bobby Kimball as the lead vocalist on the song "Hold the Line". Document3 lists several performers, including Joaquin Phoenix and Reese Witherspoon. Document4 incorrectly attributes the performances to Jennifer Aniston. Therefore, the correct answer is Reese Witherspoon. </think>

<answer> Reese Witherspoon </answer>

BrPr 2-Turn Output:

First Output: <background> "Walk the Line" is a 2005 biographical film about the life of Johnny Cash. The film's soundtrack features several songs, including the title track "Walk the Line", performed by Joaquin Phoenix, who played Johnny Cash in the film. </background>

Second Output: <think> Document1 mentions that Carlene Carter was played by Victoria Hester in the movie "Walk the Line ". Document2 mentions Bobby Kimball as the lead vocalist on the song "Hold the Line". Document3 lists the performers of the songs on the soundtrack, including Joaquin Phoenix, Reese Witherspoon, etc. Document4 lists the performers as Johnny Cash, referring to the character in the film rather than the actual singer. Therefore, the correct answer is Reese Witherspoon. </think>

<answer> Reese Witherspoon </answer>

Table 7: The Case comparison using the GRPO-RAG and BrPr models.