Not All Citations Are Equal: Entropy-Guided Citation Selection for Noise-Resistant Medical LLM

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) provides external knowledge support for large language models (LLMs) in medical applications, but retrieved contexts often contain noisy or conflicting evidence that can degrade reasoning. We observe that when internal and external knowledge disagree, models systematically prefer external citations, inadvertently injecting retrieval noise. Our analyses further show that only a subset of retrieved citations consistently improves outcomes; these effective citations exhibit markedly lower token-level entropy, linking citation entropy to model accuracy. Building on these findings, we propose a complete pipeline consisting of a training-free multi-turn reasoning framework and a post-training methodology. The training-free framework elicits internal thought, external thought, and fusion thought, and applies conflict detection and explicit denoising for complex queries. For post-training, we distill structured supervised fine-tuning (SFT) data and apply GRPO with an entropy-based citation reward that encourages selective citation of beneficial external knowledge while penalizing noisy citations. Experiments across diverse benchmarks demonstrate consistent gains in noise-resistant medical reasoning, with larger improvements on harder cases.

1 Introduction

Large language models (LLMs) have demonstrated extensive potential in medical applications(Liu et al., 2024). However, their practical deployment in healthcare settings faces significant challenges: the inherent limitations of pre-training frequency, the stringent requirements for knowledge accuracy in medical tasks, and the temporal dynamics of medical knowledge—where emerging clinical guidelines or trial results may supersede established treatment protocols. These factors collectively render standalone LLMs inadequate for addressing the full spectrum of medical challenges.

Retrieval-Augmented Generation (RAG) is a common solution that supplies external knowledge to LLMs. However, in medical applications, retrieved contexts often contain substantial irrelevant or conflicting information(Cuconasu et al., 2024; Xu et al., 2025; Fang et al., 2024; Wu et al., 2024). We refer to this phenomenon as *retrieval noise*, and our preliminary experiments show that incorporating retrieved knowledge can degrade performance even when retrieval comes from high-quality search and query reformulation.

Prior work largely improves the problem from the retrieval side (e.g., better search engines, larger knowledge bases, or training for tool use)(Peng et al., 2025; Fang et al., 2024; Zheng et al., 2025). In contrast, we focus on the model's inference-time ability to discern task-relevant information. To isolate this capability from retrieval quality variations, we fix the retrieved contents for each query across all experiments, using results from general-purpose search engines. We argue that intervening at the reasoning level can enable *noise-resistant* inference: the model should decide when to rely on internal knowledge and when to adopt external evidence.

Through training-free experiments, we study how models use internal versus external knowledge. We find a critical tension: when internal and external knowledge conflict, models systematically prefer external knowledge, which can introduce retrieval noise and harm reasoning. Meanwhile, models already internalize substantial world knowledge through pre-training, and in our medical setting only a limited amount of critical or updated evidence is needed for correct resolution.

Motivated by these observations, we propose a training-free, test-time framework that encourages reflection and knowledge discrimination, and then internalize this capability through post-training.

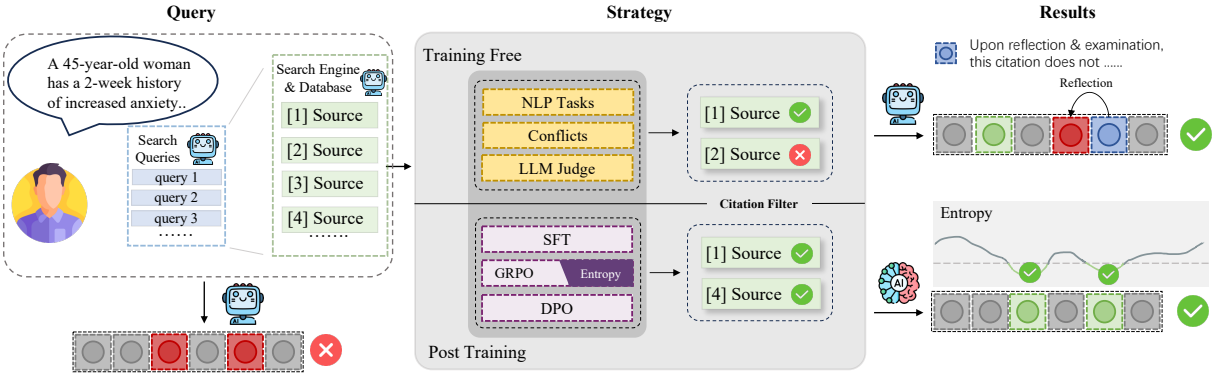

Figure 1: Overview of our proposed framework. Directly answering with retrieved search evidence can introduce retrieval noise and lead to errors. We first use a training-free framework to elicit citation selection behavior, then use citation entropy as a bridge for post-training to prefer appropriate citations and improve medical reasoning accuracy.

Concretely, we establish retrieval citation behavior via SFT cold-start, and use token-level entropy from citation actions to connect *entropy*, *key citations*, and *accuracy*. We further design an entropy-based reward that guides models to select appropriate citation content for improved medical reasoning. Figure 1 provides an overview of this pipeline.

The contributions of this paper are as follows:

- We empirically demonstrate that models produce erroneous outputs under external knowledge interference, revealing the fundamental tension between internal and external knowledge sources.
- We establish citation action entropy as a bridge, demonstrating that key correct citations contribute to improved problem-solving rates.
- We design an effective and stable entropy-based reward mechanism that enables models to develop noise-resistant reasoning capabilities in medical question answering.

2 Motivation

In this section, we discuss several key findings that motivate our research. We first identify models' dependence on external knowledge when facing complex queries. We then demonstrate that external knowledge contributes both useful information and harmful noise. Finally, we discover an entropy-based method to precisely identify beneficial external knowledge.

2.1 Dependence on External Knowledge

When internal and external knowledge conflict, models tend to follow external knowledge(Huang et al., 2024). The extent to which models rely on contextual external knowledge remains an open question, as it determines the model's strategy for specific tasks—whether to maintain self-consistency or contextual coherence(Bi et al., 2024; Zhang et al., 2025b).

To investigate this issue, we designed preliminary experiments that decompose the knowledge components in responses to medical reasoning queries when both intrinsic and external information are available. More details about this experiment are shown in Section A. We observe that for easy queries—where internal and external knowledge do not conflict—the model predominantly relies on its internal knowledge; however, for complex queries where they conflict, the model shows a strong tendency to favor external knowledge.

This finding suggests that selecting different strategies based on query difficulty is essential, and underscores the importance of carefully discerning effective external knowledge in complex reasoning tasks.

2.2 Importance of Key External Knowledge

Among the large volume of external knowledge, beyond the portions that overlap with the model's existing knowledge or are incorrect or irrelevant, a certain proportion of key knowledge can provide decisive help for solving medical tasks(Jung et al., 2024; Xiao et al., 2025).

We designed a preliminary experiment to investigate the influence of introducing external knowl-

edge and differences among these external knowledge. More details and configurations about this experiment are recorded in Section B.

The task involves combining external knowledge in the input, which includes a series of labeled citations. We systematically controlled which citations were introduced and evaluated outcomes using the accuracy of final responses. For example, we defined a valid citation as one appearing most frequently in correct samples and tested the model with and without it. Other permutations of citation combinations yielded similar results: the highest accuracy was achieved when the input contained only those citations identified as valid and effective, surpassing even the accuracy when all retrieved citations were included. Conversely, removing these effective citations resulted in a marked decline in accuracy.

This confirms that not all external information benefits the model's reasoning process. Mandating search actions and citation of search results may lead to worse performance compared to reasoning without external knowledge. Also, these findings indicate that among all retrieved citations, only a subset positively influences model outcomes, while others may have detrimental effects. Enhancing the model's ability to selectively attend to beneficial citations while disregarding irrelevant or ineffective information is therefore critical for improving overall performance.

2.3 Entropy and Accuracy

In our subsequent training experiments, we discovered that token-level entropy serves as a bridge connecting key knowledge to accuracy. External test-time computation methods—such as computing similarity, uncertainty(Li et al., 2025b), or employing LLMs-as-a-judge(Zheng et al., 2023)—often incur excessive inference costs and latency.

Prior work has demonstrated that the internal states of LLMs can reflect whether the model has encountered relevant data during training or exhibits hallucinations(Azaria and Mitchell, 2023; Ji et al., 2024; Kossen et al., 2024). Accordingly, we hypothesize that the internal representations of LLMs may serve as reliable indicators for identifying effective external knowledge. Among these internal representations, the most accessible is the probability distribution (logits) of each token, from which entropy(Cheng et al., 2025) can be computed:

$$H(o) = -\sum_{v \in \mathcal{V}} p(v) \log p(v) \tag{1}$$

We employ entropy as a quantitative metric for each citation, using this value to assess their efficacy. In validation experiments, we aggregate the entropy values of all tokens belonging to the same citation across responses:

$$H_{\text{average}} = \frac{1}{|\text{citation}|} \sum_{o \in \text{citation}} H(o) \tag{2}$$

As shown in Table 1, citations that positively influence outcomes *core* consistently exhibit lower average token entropy than other retrieved citations. Combined with the evidence in Section 2.2, this result supports a substantive link between citation entropy and model accuracy: only a subset of citations effectively supports LLM reasoning, while the rest is irrelevant or even detrimental. Notably, effective citations are characterized by markedly lower entropy.

Model	**Citation Entropy**	
	Core	**Other**
Qwen3-30B-A3B	0.383	0.426
Llama3.3-70B	0.256	0.263
Qwen2.5-72B	0.244	0.255
Deepseek-v3	0.490	0.493

Table 1: Average token entropy H_{average} (Equation 2) aggregated within each citation, comparing effective citations (Core) against other retrieved citations across different base models.

This finding demonstrates that LLMs possess an internal capacity to distinguish among various sources of external knowledge. It also informs our subsequent experimental design aimed at enhancing the model's ability to discern and appropriately incorporate valuable external information.

3 Methods

We design a complete pipeline from training-free inference to SFT and then to RL, enabling the model to effectively leverage key external knowledge for complex reasoning.

3.1 Training-free Approach

The core of our training-free framework is a multi-turn reasoning paradigm consisting of three stages:

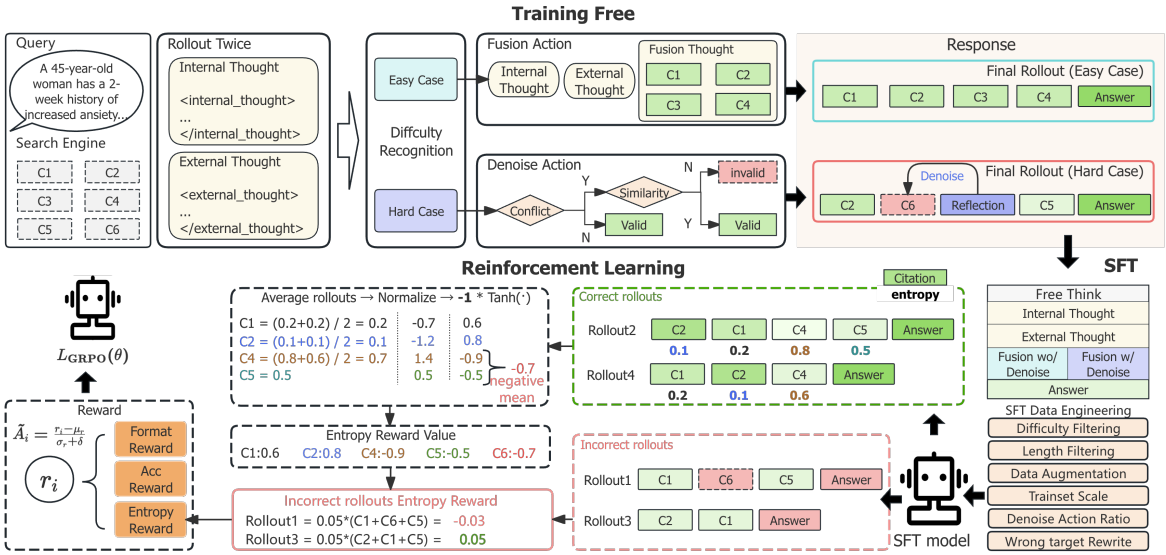

Figure 2: **Illustration of our method pipeline.** First we use an training-free method which can discern different citations during reasoning. Then we use the reasoning contents generated in training-free method to construct SFT data and get our cold start model. Finally we use our designed entropy-related GRPO to get the final model.

internal thought (prompted to use internal knowledge), external thought (exposed to citations for external reasoning), and fusion thought (integrating both sources).

In the first two stages, the model generates responses using internal knowledge and retrieved external knowledge, respectively.

In the third stage, the model assesses query difficulty and selects an appropriate strategy: for easy queries, it integrates results straightforwardly; for complex queries, it carefully discerns relevant external information.

When handling complex queries, we employ LLMs-as-a-judge to determine whether each citation conflicts with the model's internal knowledge. When a conflict occurs, we measure the macro-level similarity between the conflicting content and the internal knowledge topic. If the similarity is high, we assume the conflict lies in details, where external knowledge may provide updated or more precise information—in this case, we trust the citation. If the similarity is low, we consider the external knowledge to be highly noisy and execute a denoise action by adding a reflection sentence that instructs the model to distrust this citation during subsequent reasoning. Notably, we add a reflection sentence rather than directly removing the citation to prevent the model from reciting the excluded content in later generation.

3.2 Post-training

We aim to internalize this capability through post-training via a multi-stage methodology.

3.2.1 Supervised Fine-Tuning

We design SFT data based on the training-free framework. The prompt contains the medical query and retrieved external knowledge. The target is structured into *think*, *internal_thought*, *external_thought*, *fusion_thought*, and *answer*, demarcated by special tokens corresponding to the framework's steps.

We distill SFT data from GPT-4.1 using the training-free method with the following key improvements: 1) filter overly easy queries; 2) filter excessively long data; 3) language unification, open-ended question augmentation, and out-of-domain augmentation; 4) appropriate data scale for cold-start; 5) control the proportion of denoise actions; 6) rewrite with ground truth when incorrect.

3.2.2 GRPO with citation

To cultivate the model's capacity for recognizing effective external knowledge, we incorporate citation-related rewards into GRPO(Shao et al., 2024). As shown in Section 2.3, citations that positively influence model accuracy exhibit lower entropy. Leveraging this property, we design specialized rewards.

During rollout, we use correct responses to guide incorrect ones for effective citation identification.

For correct responses, we assign accuracy and format scores as rewards, compute citation entropy, and normalize via tanh:

$$H_{\text{std}}(c) = \frac{H_{\text{average}}(c) - \text{mean}(\mathbf{H}_{\text{average}})}{\text{std}(\mathbf{H}_{\text{average}})} \quad (3)$$

$$v(c) = -\tanh(H_{\text{std}}(c)) \quad (4)$$

For incorrect responses, the total reward is determined by summing the rewards of citations that also appear in correct samples:

$$r = \alpha \cdot \sum_{c \in \text{citations}} v(c) \quad (5)$$

α in Equation 5 means the weight of final entropy reward in training process. Citations that appear exclusively in incorrect responses are considered bad actions. Their rewards are set to the mean value of all negative-reward citations. This penalizes such actions while preventing excessive punishment that could induce reward hacking—where the model indiscriminately cites all available sources to avoid penalty.

4 Experiments

4.1 Training-free Experiments

We compare inference using only internal knowledge, only external knowledge, and our framework, and separately test the simple strategy for simple query in the 3rd stage to demonstrate their effectiveness on datasets of varying difficulty.

4.1.1 Models

We used GPT-4.1, one of the best-performing models on medical benchmarks, as a baseline to demonstrate that our method delivers stable improvements even on top-performing models.

4.1.2 Benchmarks

We utilized numerous common medical benchmarks, including MedQA(Jin et al., 2021), MedMCQA(Pal et al., 2022), MedBullets(Chen et al., 2025a), MedXpertQA(Zuo et al., 2025), CARE-QA(Arias-Duart et al., 2025), GPQA(Rein et al., 2024), SuperGPQA(Du et al., 2025), MLEC-QA(Li et al., 2021), MMLU(Hendrycks et al., 2020), CMMLU(Li et al., 2024), and MMLU-Pro(Wang et al., 2024). These benchmarks consist entirely of multiple-choice questions, including both Chinese and English datasets, covering medical and general domains. For all experiments, retrieved citations are obtained from general web search and kept fixed for each query to ensure fair comparison across methods.

4.1.3 Results

Our method improves on almost all datasets compared with baseline. For simple benchmarks like MLEC-QA, the simple fusion strategy achieves better results; for more difficult benchmarks, the denoising strategy yields even better results, demonstrating the necessity of difficulty-based strategy selection.

4.2 Post-training Experiments

4.2.1 Datasets

We used MedQA-train, MedQA-mainland-train, MedMCQA-train, CM-EXAM, MMLU-Pro (excluding medical questions), and MLEC-QA. We selected datasets based on difficulty for SFT data, and used multi-LLM voting to select challenging subsets for RL training.

4.2.2 Models

We trained Qwen3-30B-A3B(Yang et al., 2025) using our dataset and method.

4.2.3 Results

Our models achieve better performance, especially on complex reasoning benchmarks (MedXpertQA, SuperGPQA). Both SFT and GRPO with entropy reduce the negative impact of external citations and outperform standard GRPO.

4.2.4 Ablation Studies on GRPO

Regarding the reward design for GRPO experiments, we designed a series of ablation experiments to verify the effectiveness and superiority of our reward.

Citation reward from where to where. To verify whether citations for entropy calculation should be selected from all samples or only correct samples, we conducted an ablation experiment. The training curves in Figure 3 show that selecting citations from all samples results in less stable training and worse validation performance.

Query whose rollouts are all right or all wrong. To validate our reward design for samples that are all incorrect, we compared not applying citation-related rewards to all-incorrect samples versus calculating rewards from all samples. The experiment results in Figure 3 show that the latter method significantly lags behind the first.

Benchmark	Internal	External	Fusion wo/ Denoise	Fusion w/ Denoise	Q Len	Opt Len	Max Opt
Complex Benchmarks							
Care-QA	92.11	90.74	93.50	**93.74**	379	59	4
GPQA	63.10	51.82	**63.92**	63.67	696	57	4
MedBullets-Op4	86.46	84.23	**86.82**	86.57	1021	25	4
MedQA	94.01	91.22	94.21	**94.61**	869	30	4
MedXpertQA-Text	**40.01**	35.60	38.77	39.79	1060	38	10
MMLU-Pro-Health	81.02	79.07	**82.05**	81.16	587	31	10
Super-GPQA	58.45	48.51	58.22	**60.05**	547	30	10
ALL (Complex)	68.14	64.19	67.91	**68.78**	673	38	10
Simple Benchmarks							
CM-Exam	75.00	82.23	**83.18**	82.95	81	9	5
CMMLU-Anatomy	86.75	**89.27**	**89.27**	**89.27**	45	7	4
CMMLU-Clinical-Knowledge	86.08	**90.08**	89.02	89.22	133	9	4
CMMLU-College-Medical-Statistics	82.53	83.41	**85.59**	**85.59**	99	14	4
CMMLU-College-Medicine	92.41	92.24	**93.40**	93.23	75	10	4
CMMLU-Professional-Medicine	85.97	87.64	87.53	**87.65**	62	9	4
CMMLU-Traditional-Chinese-Medicine	77.72	88.86	**89.85**	89.60	53	6	4
MedMCQA	81.13	79.69	**82.22**	82.09	184	24	4
MLEC-Clinic	87.67	84.45	**89.53**	89.07	104	10	5
MLEC-CWM	80.55	83.75	**84.21**	83.98	80	10	5
MLEC-Public-Health	84.03	84.03	**84.49**	84.26	92	11	5
MLEC-Stomatology	78.89	84.45	**85.15**	**85.15**	94	11	5
ALL (Simple)	84.48	85.68	**87.00**	86.92	117	15	5
All (Simple+Complex)	75.15	73.44	75.94	**76.14**	552	30	10

Table 2: Accuracy of our training-free framework on medical benchmarks. **Internal** and **External** answer with intrinsic knowledge only or with retrieved evidence, respectively; **Fusion** is the third-stage rollout integrating both. **w/ Denoise** adds an extra denoising step, triggered by the model's own query-level difficulty judgment at inference time. For reporting, we separately group benchmarks into *Simple* vs. *Complex* using a heuristic based on question length (Q Len), option length (Opt Len), and the maximum number of options (Max Opt). Best results are in **bold**; second best are underlined.

Model	All-Case	MedXpertQA	CM-Exam	GPQA	Super-GPQA
Qwen3-30B-A3B	57.43	26.57	84.80	**62.63**	55.71
Qwen3-30B-A3B With Citation	57.05	26.73	**87.40**	59.34	54.71
Qwen3-30B-A3B-SFT	57.10	27.06	85.00	**62.63**	53.71
Qwen3-30B-A3B-SFT-GRPO	58.66	33.59	84.80	58.97	57.28
Qwen3-30B-A3B-SFT-ours	**59.85**	**34.04**	**87.40**	58.97	**59.00**

Table 3: Accuracy of post-training methods on medical benchmarks. All-Case is the average over the four benchmarks. Best in **bold**, second best underlined.

Entropy of citation, whole contents or only label. To verify the scope of citation entropy calculationwhether it includes the entire content corresponding to the citation or only the initial labelwe designed an ablation experiment. The results show that using only labels to calculate entropy leads to more unstable rewards because the entropy of a single label often has very high variance. This ultimately results in poor training stability shown in Figure 3, causing model performance to decline at very early steps.

Reward only or both reward and punishment. To verify whether to reward only low-entropy citations or to reward low-entropy and penalize high-entropy citations separately, we designed an ablation experiment. The results in Figure 3 show that the latter strategy leads to more stable training and better model performance, encouraging the model to favor better citations within a stable total number of citations. The model output reveals that considering only rewards or penalties often leads to severe reward hacking, causing the model's citation count to skew towards two extremes. At a deeper level, examining entropy and citation entropy shows that the former strategy only encourages lower absolute entropy values, rather than considering the relative levels of entropy between different citations, resulting in severe entropy collapse in the model's citation content. Figure 4 further visualizes this effect via the evolution of response entropy over training steps.

Mapping method from citation entropy to reward. We conducted ablation experiment using our method, along with other methods that used

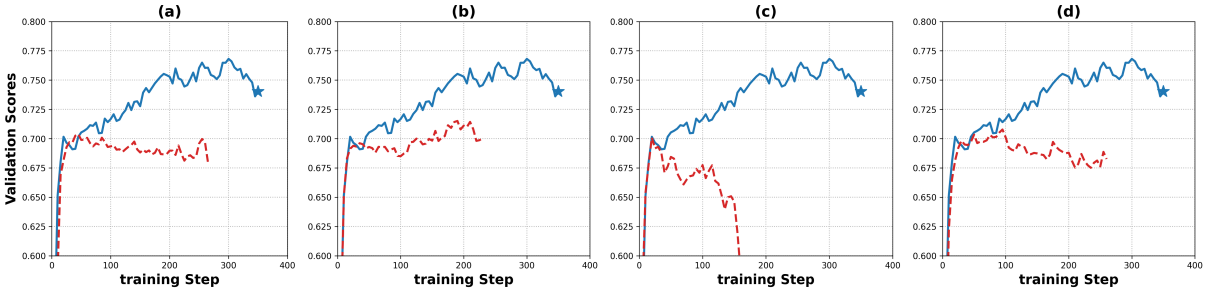

Figure 3: Ablation studies on GRPO reward design. Blue curves (starred endpoints) denote our method; red dashed curves denote ablation baselines. (a) Citation source: selecting citations from correct samples only (ours) vs. from all samples. (b) All-correct/all-wrong queries: excluding citation reward for all-incorrect rollouts (ours) vs. applying it uniformly. (c) Entropy scope: computing entropy over entire citation content (ours) vs. label tokens only. (d) Reward strategy: both rewarding low-entropy and penalizing high-entropy citations (ours) vs. reward only. Our design consistently leads to more stable training and better final performance.

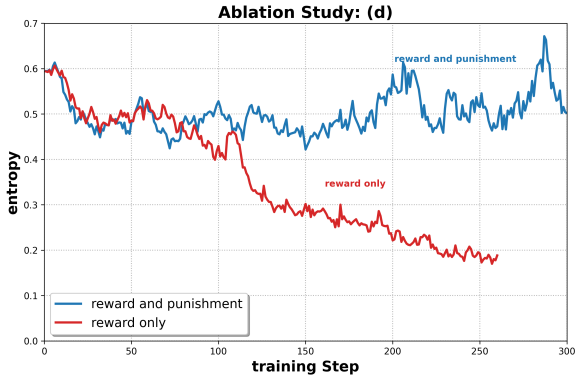

Figure 4: Evolution of overall response entropy over training steps under two reward shaping strategies. The blue curve (reward-and-punishment) maintains relatively high entropy throughout training, preserving exploration. In contrast, the red curve (reward-only) rapidly collapses, exhibiting entropy collapse due to one-sided reward.

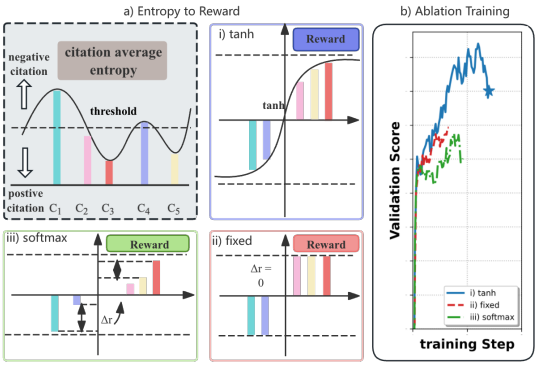

Figure 5: (a) Comparison of three entropy-to-reward mapping functions: tanh preserves absolute magnitude while mitigating extreme values; rank-based uses fixed thresholds; softmax normalizes across citations. (b) Training curves show tanh (blue, starred) achieves best performance over softmax (green) and rank-based (red).

the difference between the number, or the softmax value of the corresponding entropy, of low-entropy citations and high-entropy citations as the reward whose low-entropy and high-entropy citations are obtained by sorting by entropy value and using a fixed 30% threshold. The experiments demonstrate that our method leads to better model performance. Intuitively, using the tanh-mapped value as the value for different citations preserves the natural relative relationship between them, while using a fixed ratio to define different types of citations disrupts this relative relationship. Furthermore, using tanh-mapped method makes values smoother at the boundaries, unlike softmax which can produce very extreme values. Figure 5 compares three mapping functions from citation entropy to reward and shows that tanh better preserves absolute magnitude information.

5 Related Work

5.1 RAG in Various Scenarios

In the RAG scene, some works use metrics such as perplexity or entropy to consider changes before and after incorporating relevant context, detecting the effectiveness of relevant context to compress external information, and then using this to modify the context or generate data for SFT training(Jia et al., 2025; Jiang et al., 2024; Lin et al., 2025). In contrast, other works employ complex reasoning frameworks or train with factual data to keep faithful to simple factual queries with fixed external knowledge, mitigating problems related to knowledge conflicts(Huang et al., 2025b; Zhang

et al., 2025b; Bi et al., 2025; Chen et al., 2025b; Zhang et al., 2025d). However, these efforts either require complex external systems and significant inference costs, or they can only improve the model's performance on simple factual datasets and cannot effectively improve the model's ability in complex medical inference scenarios.

5.2 Medical Reasoning

Evidence-based medicine requires medical reasoning models to ensure the accuracy of reasoning and knowledge citation at every step. Some works use PRM to score intermediate reasoning steps, using RL to teach the model reasonable reasoning processes(Yun et al., 2025; Jiang et al., 2025a; Yu et al., 2025b; Zhang et al., 2025e). Other works focus on providing reliable external knowledge for reasoning through customized search engines or knowledge graphs(Yu et al., 2025a; Wang et al., 2025b; Wu et al., 2025; Liu et al., 2025). Others design multi-agent reasoning frameworks based on different domain tasks(Zhao et al., 2025). However, these works aim to consider the rationality of the reasoning steps and how to introduce high-quality external knowledge, lacking consideration to how to identify and optimize these external knowledge.

5.3 Improvements in RL Training

Balancing a model's exploration and exploitation capabilities has always been a critical issue in RL(Zhang et al., 2025a; Cui et al., 2025). Some works encourage low entropy by making entropy the optimization objective or incorporating it into rewards and advantages, thereby achieving more confident responses(Agarwal et al., 2025; Prabhudesai et al., 2025). Other works argue that a high-entropy distribution implies more exploration action space and better performance. Therefore, they choose to add rewards related to entropy or semantic entropy at certain tokens or whole response to mitigate the entropy collapse problem(Wang et al., 2025a; Chen et al., 2025c; Zhang et al., 2025c; Jiang et al., 2025b). Furthermore, some works have proposed using different training strategies based on the difficulty of the problem, such as dynamic learning rates, advantages or reward strategies(Huang et al., 2025a; Zhang et al., 2025f; Li et al., 2025a). However, most of these works rarely focus on a specific part of the outcomes. At the same time, the debates about whether to reward high or low entropy during training lack rigorous analysis and validation in specific domains. Moreover, none of these training methods have been tested in RAG scenarios.

Limitations

Our work has several limitations. First, although we mix general web search and medical databases for generality, our study does not aim to compare retrieval sources; we instead fix retrieved contents for each query throughout the experimental cycle to avoid confounding factors.

Second, while token-level entropy provides a useful signal for citation quality, it is not a perfect indicator. Encouraging low-entropy citations can lead to safer use of external knowledge, but high entropy may also correspond to beneficial exploration; thus, entropy alone cannot fully capture the optimal citation behavior. Moreover, we find that extracting entropy at finer granularity can be unstable, and our effective strategy aggregates entropy over the full citation sentence (including the citation marker and cited content), which limits how precisely the signal can localize useful evidence.

Finally, our current training design applies rewards at the citation level and optimizes at the rollout level via advantage, rather than assigning token-level rewards or selecting fine-grained evidence within each citation (e.g., key sentences or spans). More stable designs that analyze prompt and evidence tokens may enable finer-grained supervision, but are beyond the scope of this work.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare llms beyond question-answering. *arXiv preprint arXiv:2502.06666*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2025. Context-dpo: Aligning language models for

context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *arXiv preprint arXiv:2404.00216*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

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

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025c. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

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.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

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. *arXiv preprint arXiv:2405.20978*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Chen Huang, Wei Lu, and Wenxuan Zhang. 2025a. Pear: Phase entropy aware reward for efficient reasoning. *arXiv preprint arXiv:2510.08026*.

Yin Huang, Yifan Ethan Xu, Kai Sun, Vera Yan, Alicia Sun, Haidar Khan, Jimmy Nguyen, Mohammad Kachuee, Zhaojiang Lin, Yue Liu, and 1 others. 2025b. Confqa: Answer only if you are confident. *arXiv preprint arXiv:2506.07309*.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2024. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. *arXiv preprint arXiv:2410.14675*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*.

Junlong Jia, Ziyang Chen, Xing Wu, Chaochen Gao, Zijia Lin, Debing Zhang, Songlin Hu, and Binghui Guo. 2025. Entropylong: Effective long-context training via predictive uncertainty. *arXiv preprint arXiv:2510.02330*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025a. Meds^3: Towards medical slow thinking with self-evolved soft dual-sided process supervision. *arXiv preprint arXiv:2501.12051*.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025b. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*.

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

Dongwon Jung, Qin Liu, Tenghao Huang, Ben Zhou, and Muhao Chen. 2024. Familiarity-aware evidence compression for retrieval-augmented generation. *arXiv preprint arXiv:2409.12468*.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Ang Li, Zhihang Yuan, Yang Zhang, Shouda Liu, and Yisen Wang. 2025a. Know when to explore: Difficulty-aware certainty as a guide for llm reinforcement learning. *arXiv preprint arXiv:2509.00125*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. Mlec-qa: A chinese multi-choice biomedical question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025b. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157.

Yuhao Lin, Wenya Guo, Ying Zhang, Chengyi Yang, and Zengxiang Li. 2025. Prompt compression based on key-information density. *Expert Systems with Applications*, page 127738.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. 2025. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *arXiv preprint arXiv:2509.15279*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Zhiyuan Peng, Jinming Nian, Alexandre Evfimievski, and Yi Fang. 2025. Eloq: Resources for enhancing llm detection of out-of-scope questions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3509–3519.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Ziyu Wang, Elahe Khatibi, and Amir M Rahmani. 2025b. Medcot-rag: Causal chain-of-thought rag for medical question answering. *arXiv preprint arXiv:2508.15849*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Liang Xiao, Wen Dai, Shuai Chen, Bin Qin, Chongyang Shi, Haopeng Jing, and Tianyu Guo. 2025. Retrieval-augmented generation by evidence retroactivity in llms. *arXiv preprint arXiv:2501.05475*.

Yongxin Xu, Ruizhe Zhang, Xinke Jiang, Yujie Feng, Yuzhen Xiao, Xinyu Ma, Runchuan Zhu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. Parenting: Optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11643–11662.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun

Yue, and 1 others. 2025a. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*.

Heng Yu, Junfeng Kang, Rui Li, Qi Liu, Liyang He, Zhenya Huang, Shuanghong Shen, and Junyu Lu. 2025b. Ca-gar: Context-aware alignment of llm generation for document retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5836–5849.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, and 1 others. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025a. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025c. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025d. 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.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025e. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, Kejiang Chen, and Xing Hu. 2025f. Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning. *arXiv preprint arXiv:2510.10959*.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Xin Sun, Ya Zhang, and 1 others. 2025. An agentic system for rare disease diagnosis with traceable reasoning. *arXiv preprint arXiv:2506.20430*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Qiaoyu Zheng, Yuze Sun, Chaoyi Wu, Weike Zhao, Pengcheng Qiu, Yongguo Yu, Kun Sun, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2025. End-to-end agentic rag system training for traceable diagnostic reasoning. *arXiv preprint arXiv:2508.15746*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Faithfulness Experiments

To explore the conflict between internal and external knowledge during complex reasoning and to verify the model's faithfulness and preference for knowledge from different sources, we designed an experiment. During the reasoning process, we allowed the model to autonomously decide whether to use a search engine to retrieve external knowledge to assist in the reasoning, until the model autonomously decided that all the current reasoning content and external information were sufficient to reach the final conclusion, and then reasoned out the final answer.

In this process, we break down both the external knowledge and the internal knowledge involved to calculate the model's faithfulness and preference. First, for external knowledge, we use LLM to break it down into knowledge forms similar to triples as the external knowledge set. Next, we extract the knowledge used in the reasoning process and break it down into triplets as overall knowledge set. Then, we remove one element from these triples and let LLM continue to complete these triples, thus extracting the model's internal knowledge set.

By using these knowledge triples, we can determine, through LLM, whether there is a conflict between internal and external knowledge regarding the same fact. Furthermore, we will calculate the proportion of external (or internal) knowledge followed by LLM in different situations relative to the total number of knowledge points involved in the entire reasoning process, to compare the degree to which LLM adheres to both.

As can be seen from Table 4: 1) When internal and external knowledge conflict, the model's

Knowledge Source	no conflict	conflict
internal	0.89	0.16
external	0.73	0.56

Table 4: Proportion of different knowledge that LLM adhere to during reasoning in different situations

ability to follow the knowledge decreases, verifying that inconsistency between internal and external knowledge will damage the model's ability; 2) When internal and external knowledge conflict, the model's performance on internal knowledge decreases more significantly, indicating that the model tends to rely more on external knowledge in this situation.

This reveals the importance of external knowledge when models engage in complex reasoning and when internal and external knowledge conflict. Incorrect external knowledge can be detrimental, while good external knowledge can improve outdated or erroneous knowledge within the model, enhancing the accuracy of reasoning. This also demonstrates the necessity of our work in improving the model's ability to discern external knowledge.

B Citation Absence Experiments

To verify the role of key citations in reasoning, as presented in Section 2.2, a simple experiment was designed. For a number of queries containing a series of retrieved citations, we first sampled them directly using LLM, repeating this process 20 times. Then, we counted the frequency of citations in the correct samples, simply using citation frequency as a value indicator, selecting the top 50% of citations as high-frequency citations. Next, we repeat the above sampling process using these three strategies: removing these citations from the input prompt, retaining only these high-frequency citations in the prompt, and removing all citations. We then statistically calculate the accuracy of the sampling results under each strategy.

All Case	All Citation	Only Top Citation	No Top Citation	No All Citation
GPT-4.1	0.477	**0.486**	0.455	0.4855
GPT-4o-0806	0.385	**0.411**	0.380	0.4025
o4-mini	0.390	**0.430**	0.400	0.420
Qwen2.5-72B-Instruct	0.363	**0.389**	0.340	0.386
Qwen3-30B-A3B-sft-1112b	0.522	**0.536**	0.444	0.458
Llama3.3-70B-Instruct	0.407	**0.447**	0.382	0.409
Deepseek-v3	0.419	**0.4485**	0.3875	0.420

Table 5: Accuracy of different models under various citation absence strategies on the test dataset. The test dataset contains 100 randomly selected samples from the benchmarks listed in Table 2.

As shown in Table 5, the final results showed a consistent trend across all models: sampling results with only high-frequency citations had the highest accuracy, exceeding those with all citations, while sampling results with only high-frequency citations removed had the lowest accuracy. This result confirms our previous conclusion that only a subset of key citations have a positive impact on the results, while other citations act as interference or even have a negative effect.

It's also worth mentioning that our key citation selection strategy here simply uses the frequency of occurrence in the correct samples, rather than the lower entropy mentioned earlier in Section 2.3. This is because this part of the experiment was our initial exploratory attempt. After validating the effectiveness of key citations using high-frequency, we subsequently discovered the generally lower entropy characteristic of these citations, revealing the connection between lower entropy and higher accuracy of citations, which guided us in designing the final training strategy.

C Training Hyperparameters

Here, we show our detailed hyperparameter configurations employed in our training method in Table 6.

rollout.n	mbs	lr	entropy reward weight
16	32	5e-6	0.05
16	32	3e-6	0.05
16	32	6e-6	0.05
16	32	5e-6	0.1

Table 6: Hyperparameters used for GRPO training. mbs means mini batch size. lr means learning rate

We experimented with these different hyperparameter configurations and randomly selected 100 queries from MedXpertQA as the test dataset. The results shown in Figure 6 indicate the best hyperparameters of our method. What's more, it shows that LLMs are very sensitive to the size of the entropy reward weight, as excessive entropy-related rewards may affect model training by causing its entropy to drop rapidly.

D Prompts and Data Examples

Here are the prompt examples.

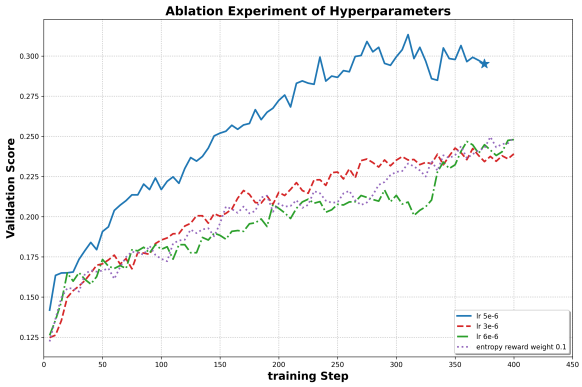

Figure 6: Evolution of validation score during training using different hyperparameter configurations through our methods. These curves correspond to different hyperparameters configurations in Table 6

input template

You are an Evidence-Based Medical AI Assistant.
Given a question and a series of texts related to the question from a search engine, please carefully read the question and the related texts, provide a detailed thought process following the given requirements and the present your final answer.
As the questions may be too complex or need some related documents, your responses must be grounded in high-quality given evidence and follow a systematic reasoning approach.
Your thought process must include the analysis of your own knowledge, the analysis of given external text and the combination of these two thought process.

CITATION FORMAT REQUIREMENTS:
You MUST use the following special citation format to mark ALL content that comes from or relates to the search results:
1. **Single Citation Format**:
- Use: [$Citation_3_STarT$]content related to citation 3[$Citation_3_FInisH$]
- Note the capitalization pattern:
- Start marker: STarT (first two letters S,T and last letter T are capitalized)
- End marker: FInisH (first two letters F,I and last letter H are capitalized)
2. **Multiple Citations Format**:
- When content relates to multiple citations simultaneously, use:
- [$Citation_3_7_STarT$]content related to both citations 3 and 7[$Citation_3_7_FInisH$]
- List all citation numbers separated by underscores
3. **Citation Rules**:
- EVERY piece of information from search results MUST be wrapped in citation markers
- Split citations when content can be separated, even within the same sentence
- Only merge citations when they share EXACTLY the same content
- Citation numbers correspond to the [number] in the search results

Here is the question: query

Here are the texts: texts

The formatted output should be as follows:
<think>
[Think deeply and freely explore all plausible hypotheses step by stepexamining evidence, alternative explanations, and counterarguments.]
[Analysis based on Citations in this part is NO need to follow CITATION FORMAT.]
</think>
<internal_thought>
[Your reasoning process using internal knowledge]
</internal_thought>
<external_thought>
[Your reasoning process using the given text]
[If any citation is helpful or critical, follow the CITATION FORMAT to cite here.]
</external_thought>
<fusion_thought>
[Your analysis combining both internal and external think, plan the optimal fusion strategy]
[If any citation is helpful or critical, follow the CITATION FORMAT to cite here.]
</fusion_thought>
</think>

<answer>
[Detailed analysis of the query based on all the think process]
[Final answer for the given question]
[If this question is a multiple-choice question, please give the final option with format as: boxedinsert your option label]
</answer>

Let's think step by step and then give your response: