Whether LLMs Know If They Know: Identifying Knowledge Boundaries via Debiased Historical In-Context Learning

Bo Lv[1,2,3], **Nayu Liu**[4*], **Yang Shen**[4], **Xin Liu**[1], **Ping Luo**[1,2,3], **Yue Yu**[1]
[1]Peng Cheng Laboratory, [2]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[3]University of Chinese Academy of Sciences
[4]Tianjin Laboratory Autonomous Intelligence Technology and Systems, School of Computer Science and Technology, Tiangong University
nayuliu@tiangong.edu.cn, lvbo19@mails.ucas.ac.cn

Abstract

In active retrieval (AR), large language models (LLMs) need first assess whether they possess knowledge to answer a given query, to decide whether to invoke a retrieval module. Existing methods primarily rely on training classification models or using the confidence of the model's answer to determine knowledge boundaries. However, training-based methods may have limited generalization, and our analysis reveals that LLMs struggle to reliably assess whether they possess the required information based on their answers, often biased by prior cognitive tendencies (e.g., tokens' semantic preferences). To address this, we propose **D**ebiased **H**istorical **I**n-**C**ontext **L**earning (DH-ICL) to identify knowledge boundaries in AR. DH-ICL aims to reframe this self-awareness metacognitive task as a structured pattern-learning problem by retrieving similar historical queries as high-confidence in-context examples to guide LLMs to identify knowledge boundaries. Furthermore, we introduce a historical bias calibration strategy that leverages deviations in the model's past response logits to mitigate cognitive biases in its current knowledge boundary assessment. Experiments on four QA benchmarks show that DH-ICL achieves performance comparable to full retrieval on LLaMA with only half the number of retrievals, without any additional training.

1 Introduction

In open-domain question answering (Zhang et al., 2023) and knowledge-intensive tasks (Yin et al., 2022; Zhao et al., 2024b), active retrieval (AR) (Asai et al., 2023) enhances the accuracy and efficiency of retrieval by having large language models (LLMs) (Yang et al., 2024; GLM et al., 2024) first assess whether they possess sufficient knowledge to answer a query before deciding whether to invoke an external retrieval module.

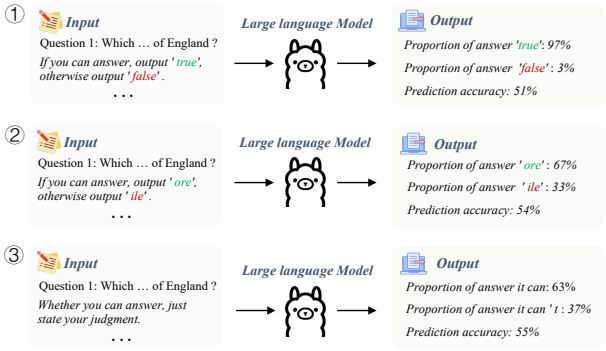

Figure 1: The diagram of LLM's self-knowledge assessment on KB-SAT, where known and unknown questions are evenly split. For different tokens, the model has obvious prior preferences. For case ③, we use ChatGPT to identify whether the model's output expresses "can" or "can't" based on its response.

Unlike traditional retrieval-augmented generation (RAG) methods (Gao et al., 2023; Lin et al., 2023; Yoran et al., 2024), which always call an external knowledge base, AR allows the LLM to autonomously decide whether to retrieve, avoiding unnecessary retrieval overhead while reducing the risk of hallucinations caused by overreliance on model knowledge.

A crucial challenge in AR is enabling the LLMs to accurately identify their knowledge boundaries. Current methods are primarily divided into two categories: (1) training independent classifiers (Cheng et al., 2024a; Asai et al., 2023; Wang et al., 2023; Liu et al., 2024), which use a small classification model to predict whether retrieval is needed based on the input query, and (2) according to the model response (Jiang et al., 2023; Wang et al., 2024; Yao et al., 2024), which directly assess whether the LLM has sufficient knowledge by evaluating the confidence of its generated answer.

Despite of the progress of these methods, two core issues remain. First, the generalization ability of training based methods is limited. Classifiers

*Corresponding author.

are typically trained on specific datasets or LLMs' hidden state features, leading to weak generalization across different domains or LLMs, which impacts retrieval decisions and deployment efficiency. Second, LLMs inherently struggle with this self-awareness metacognitive task. Our experimental analysis in Section 2 reveals that LLMs face difficulties in assessing their own knowledge boundaries based on answer confidence, and are heavily influenced by prior cognitive biases. For example, we found that label token semantics have a noticeable impact on model predictions, as shown in Figure 1. When using tokens such as "true/false" or "ore/ile" to indicate whether the model possesses knowledge, LLMs' judgment of its knowledge boundary changes considerably.

To address these issues, we propose **D**ebiased **H**istorical **I**n-**C**ontext **L**earning (DH-ICL) for self-knowledge boundary assessment in AR. Rather than requiring LLMs to directly assess their knowledge boundaries, a metacognitive task they inherently struggle with, DH-ICL transforms this into a structured pattern-learning task. Specifically, it retrieves historical queries that were previously answered as high-confidence in-context examples to prompt the LLM in identifying knowledge boundaries. This approach leverages the strengths of LLMs in pattern recognition rather than relying on their metacognitive abilities. Furthermore, to mitigate LLMs' prior cognitive biases in knowledge boundary prediction, we introduce a historical bias calibration strategy. This strategy corrects the model's bias in knowledge boundary assessment for the current query by collecting the logits biases induced by the prior knowledge encoded in LLMs' parameters when responding to historical queries.

We conducted tests using LLMs of sizes 7B and 13B on four different types of question-answering (QA) benchmarks. The experimental results demonstrate that our proposed DH-ICL effectively reduces the number of model retrievals while maintaining overall performance. Notably, on the TriviaQA dataset, the LLaMA-2-7B-Chat (Touvron et al., 2023) model achieved performance equivalent to full retrieval using only half the number of retrievals. Furthermore, DH-ICL achieves performance on par with training-based SOTA AR methods without requiring any additional training.

Our contributions are summarized as follows:

(1) We provide a comprehensive analysis of LLMs' ability to self-assess knowledge boundaries and the impact of prior cognitive biases on this evaluation.

(2) We propose debiased historical in-context learning (DH-ICL) for self-knowledge assessment in AR, avoiding relying on LLMs' metacognitive abilities by reducing this task to a specific pattern-learning task.

(3) We introduce a historical bias calibration strategy, which collects logits deviations from historical QA in-context to correct LLMs' prior biases in current knowledge boundary recognition.

(4) Experiments show that, without requiring additional training, DH-ICL achieves performance comparable to training-based SOTA methods[1].

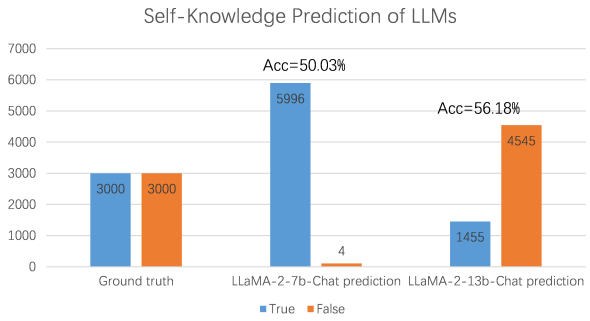

Figure 2: Self-knowledge boundary assessment of different LLMs on KB-SAT.

2 Analysis of Whether LLMs Know

2.1 Can LLMs Predict Their Own Knowledge Boundaries?

To examine whether LLMs can reliably assess their own knowledge boundaries, first, we follow prior works (Kadavath et al., 2022a) to directly prompt them to reflect on LLM confidence (i.e., logits) in answering a given question. We use the following prompt:

"You are a student being tested. For each question provided, first go through a thinking phase (no need to output specific content). Then, assess whether you can answer it correctly based on your knowledge. If you believe you can answer it correctly, output 'true'. If you feel your answer might be incorrect, output 'false'."

By comparing the logits of "true" and "false", we determine whether the model internally categorizes a question as known or unknown.

To simulate this scenario, we constructed a test set consisting of 6,000 samples based on the

[1]The codes are released at https://github.com/lvbotenbest/DH-ICL

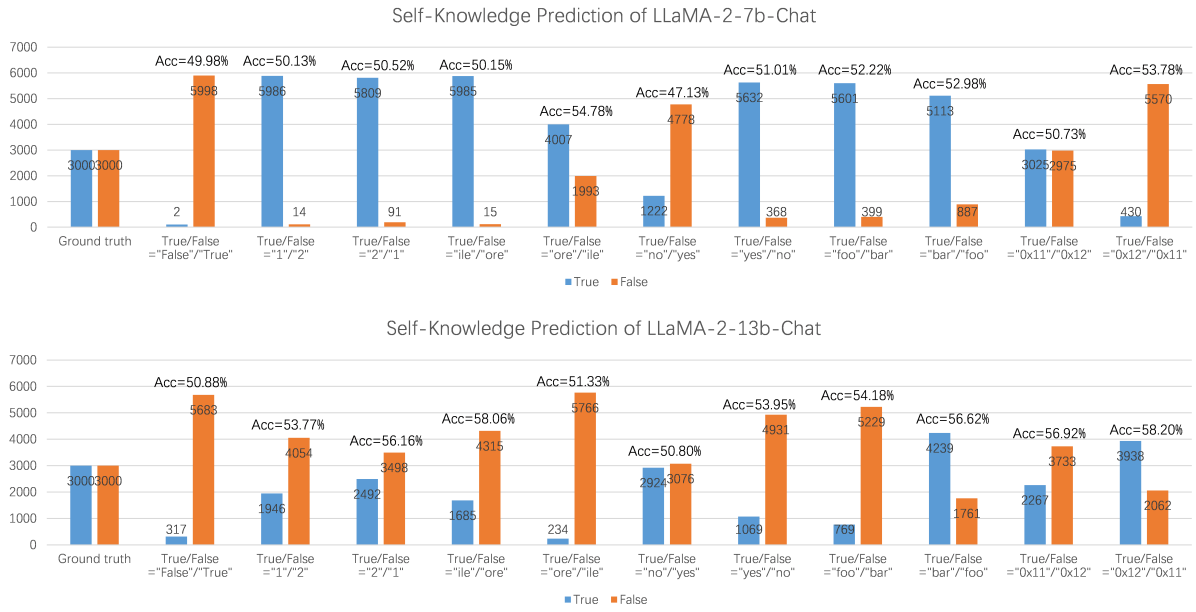

Figure 3: Impact of label token preferences on LLMs' self-knowledge boundary assessment.

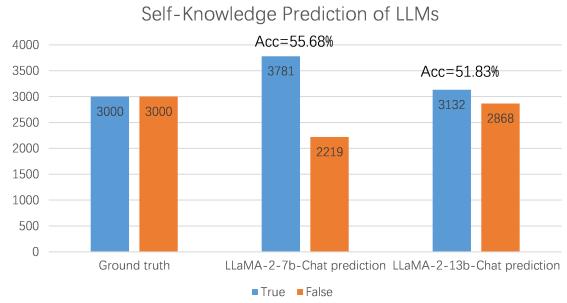

Figure 4: Self-knowledge boundary assessment of different LLMs, in which LLMs are not guided to reply to the preferred token, but are free to answer, and whether the LLM's reply is a positive or negative category is evaluated through GPT-4.

model's performance in answering TrivialQA without retrieval, with half of the model's responses being correct and the other half incorrect. This dataset is used to evaluate the model's self-assessment of its knowledge boundaries, which we refer to as the Knowledge Boundary Self-Assessment Test Set (KB-SAT).

As shown in Figure 2, on KB-SAT, LLaMA-2-7B-Chat achieves an overall accuracy of 50%, yet exhibits a strong bias toward predicting "true" (over 95%) regardless of actual correctness. LLaMA-2-13B-Chat achieves a slightly higher accuracy of 56%, but it demonstrates an opposite bias, preferring to predict "false" (over 75%). These results indicate that different models exhibit distinct prior biases in self-knowledge assessment, and their predictions are unreliable.

2.2 Influence of LLMs' Prior Biases

To examine whether the observed biases stem from an intrinsic preference for specific tokens rather than genuine self-knowledge assessment, we conduct experiments using different tokens as labels: (a) Synonymous tokens (e.g., replacing "true/false" with "yes/no"); (b) Neutral numerical tokens (e.g., using "1/2" instead of "true/false"); (c) Reversed tokens (e.g., assigning "true" to denote "false"). The experimental results, shown in Figure 3, reveal that modifying the label tokens significantly impacts the model's prediction tendency. However, despite these label changes, the overall accuracy remains limited to 47%-58%.

2.3 Can LLMs Improve Self-Assessment Without Including Specific Tokens?

Given that explicit tokens introduce biases, we investigate whether LLMs can better assess their knowledge boundaries without being constrained by predefined tokens. To this end, we modify the self-assessment prompt,

"You are a student being tested. For each given question, assess based on your knowledge whether you can answer it correctly. Just state your judgment."

to prevent the model from explicitly generating "true" or "false" and instead allow it to freely ex-

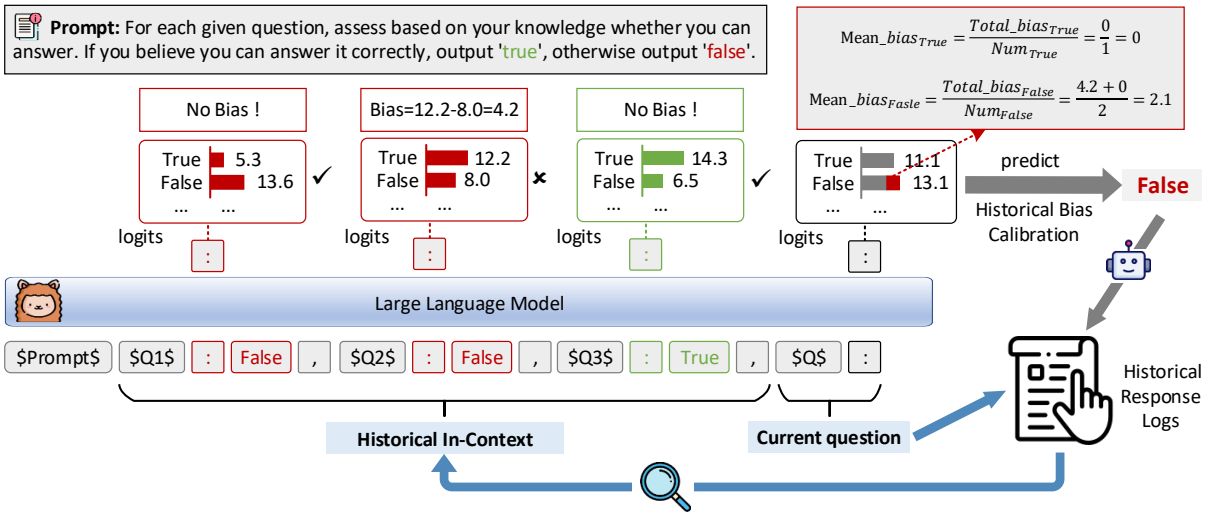

Figure 5: Overview of DH-ICL. Avoiding direct self-knowledge boundary assessment, DH-ICL reformulates this metacognitive task into a specific pattern-learning task. It acquires prior similar questions from historical QA logs as in-contexts, and corrects the model's bias in self-assessment for the current question by leveraging the biases collected in historical in-context instances.

press its confidence. We then use GPT-4 as an external evaluator to classify the responses into affirmative (true) or negative (false) categories.

As shown in Figure 4, even without introducing specific tokens, the overall prediction accuracy remains unsatisfactory (56%), reinforcing the notion that LLMs struggle with accurately assessing their own knowledge boundaries.

2.4 Observations and Insights

From the experimental analysis, we derive the following key insights: (a) LLMs' responses are heavily influenced by prior cognitive biases. For instance, LLaMA2-13B-Chat exhibits a strong preference for predicting true over false, whereas LLaMA2-7B-Chat favors the opposite. These biases likely stem from variations in the training data used during pretraining and fine-tuning. (b) LLMs inherently struggle with metacognitive reasoning (i.e., assessing whether they know something). Regardless of how we format the query that using "true/false", "yes/no", "0x11/0x12", etc., or even prohibiting explicit label generation, the proportion of true and false responses varies significantly, yet accuracy consistently hovers between 47%-57%.

These findings highlight a fundamental limitation: LLMs' self-knowledge assessment is unreliable and suffer from ingrained prior cognitive biases. This motivates the need for our proposed DH-ICL to reframe self-knowledge assessment as a structured pattern-learning task rather than direct self-reflection and mitigate these biases.

3 Debiased Historical In-Context Learning Method

3.1 Collecting Historical Response Logs

Let a widely used LLM, denoted as M, accumulate a history of responses to numerous queries over time. We simulate this scenario using publicly available QA datasets. For a given question Q_{new}, if M generates a response that matches the ground truth answer, this indicates that the question falls within the model's self-knowledge scope. In this case, we annotate it as a positive instance (e.g., labeled as "true"). Conversely, if M's response deviates from the ground truth, it signifies that the model cannot reliably answer the question within its self-knowledge scope, and we annotate it as a negative instance (e.g., labeled as "false"). Using this annotation scheme, we construct a historical response experience set for M, denoted as $H = \{Q_1 : R_1, Q_2 : R_2, ..., Q_n : R_n\}$, where Q_i represents a historical question, and R_i represents whether it can be answered (e.g. true or false).

Additionally, certain knowledge evolves over time. For example, the question "Who is the current President of the United States?" changes over time, whereas the model's knowledge remains fixed at the point of its last training data update. Therefore, our historical response experience set includes time-sensitive questions, specifically selecting instances where the model provided incorrect answers. The prompt used is as follows:

You are a student being tested. For each given

question, assess based on your knowledge whether you can answer it correctly. If you believe you can answer it correctly, output 'true'. If you are unsure whether you can answer it correctly, output 'false'. Additionally, if the question is asking about a recent event, for example, if words like recently, latest, or currently appear, also output 'false'.

The advantage of our method lies in the continuous updating of historical experience logs as the user interacts with the LLM, which leads to increasingly accurate judgments.

3.2 Historical In-Context Learning

As shown in Figure 5, we utilize historical response logs H to guide in-context learning, enabling the LLM to assess whether a given question falls within its knowledge boundary. This approach leverages the model's pattern-learning capabilities, transforming a metacognitive task into a specific pattern task. Specifically, given a current question Q_{new}, retrieve historical logs from H and select the top k most similar previously answered questions to construct the in-context examples. By incorporating historical questions with high similarity to Q_{new}, we provide relevant contextual cues to the model. These retrieved queries may contain similar entities or exhibit comparable question structures, thereby enhancing the model's ability to assess whether it can correctly answer. A detailed example is provided in Appendix C.

Notably, pairing historical in-context learning with historical bias calibration in Section 3.3 is particularly effective, as models may exhibit biases in their self-assessments for similar queries. If a model has errors in evaluating historical queries, it is more likely to exhibit similar biases when judging Q_{new}.

3.3 Historical Bias Calibration

In the historical in-context learning approach described in Section 3.2, despite providing QA exemplars, the model's prior cognitive bias may still misalign with the intended responses. For example, as illustrated in Figure 5, we provide an in-context example where question Q_2 is labeled as "false", indicating that the model lacks the necessary knowledge to answer it. However, at the prediction position (i.e. the token before "true/false"), the logits value for "true" exceeds that for "false", implying that the LLM mistakenly believes it can answer the question.

To address this issue, we introduce a historical bias calibration strategy, which corrects the LLM's current prediction bias by leveraging its historical in-context prediction logits deviations. Specifically, for the i-th QA pair in the in-context examples, we compute the bias correction values as follows:

$$\delta_i^{\text{true}} = \begin{cases} z_i^{\text{false}} - z_i^{\text{true}}, & \text{if True and } (z_i^{\text{true}} - z_i^{\text{false}}) < 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\delta_i^{\text{false}} = \begin{cases} z_i^{\text{true}} - z_i^{\text{false}}, & \text{if False and } (z_i^{\text{true}} - z_i^{\text{false}}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where z_i^{true} and z_i^{false} denote the logits produced by the LLM for tokens corresponding to "true" and "false" in the i-th example, respectively. Similarly, δ_i^{true} and δ_i^{false} represent their logits' biases.

Then the average prediction bias $\bar{\delta}^{\text{true}}, \bar{\delta}^{\text{false}}$ across all k historical in-context examples are computed as follows:

$$\bar{\delta}^{\text{true}} = \frac{\sum_i \delta_i^{\text{true}}}{N^{\text{true}}} \tag{3}$$

$$\bar{\delta}^{\text{false}} = \frac{\sum_i \delta_i^{\text{false}}}{N^{\text{false}}} \tag{4}$$

where N^{true} and N^{false} denote the number of ground truth "true" and "false" cases in the in-context examples, respectively.

Finally, for the current question Q_{new}, the logits of tokens $z_{\text{new}}^{\text{true}}, z_{\text{new}}^{\text{false}}$ for true and false in LLM could be calibrated by adjusting them based on the historical mean bias:

$$z_{\text{new}}^{\text{true}} = z_{\text{new}}^{\text{true}} - \bar{\delta}^{\text{true}} \tag{5}$$

$$z_{\text{new}}^{\text{false}} = z_{\text{new}}^{\text{false}} - \bar{\delta}^{\text{false}} \tag{6}$$

By collecting similar biases in historical in-context learning, LLM's prior cognitive preference for the current query could be mitigated, thereby improving its ability to accurately predict its own knowledge boundaries.

4 Experiments

4.1 Datasets

We select four different types of QA datasets to evaluate the effectiveness of our proposed DH-ICL in real-world downstream tasks and its adaptability to various active retrieval scenarios: TriviaQA (Joshi et al., 2017), WebQuestions (WQ) (Berant et al., 2013), TAQA (Zhao et al., 2024a), and FreshQA (Vu et al., 2023). Notably, TAQA and FreshQA contain time-sensitive questions.

To simulate historical response logs, we construct a query set from a subset of the training sets

of TriviaQA and TAQA, consisting of 50K samples from TriviaQA and the entire 10K samples from TAQA. The tested LLMs are then prompted to answer these queries, generating ten responses per question. If all responses are correct, the question is labeled as true; otherwise, it is labeled as false.

4.2 Metrics

Following prior work (Cheng et al., 2024a; Asai et al., 2023; Schick et al., 2023), we evaluate accuracy by matching whether the golden answer is in the generated text on TriviaQA and WQ. For TAQA and FreshQA, since gold answers are too long for direct lexical matching, we use ChatGPT for correctness evaluation, where the evaluation template are adopted from prior work and provided in Appendix A. In addition, we report the rate of samples where retrieval was triggered.

4.3 Baselines

We compare our proposed method against several recent baselines: **(a) Training-based methods: Self-RAG** (Asai et al., 2023), which trains LLMs to generate special tokens during inference to reflect on whether retrieval is necessary for knowledge-intensive tasks; **SKR** (Wang et al., 2023), which trains a BERT (Devlin, 2018) classifier to identify knowledge boundaries based on the collected self-knowledge data (known and unknown) and triggers external retrieval for unknown questions; **UAR** (Cheng et al., 2024a), which employs four orthogonal classifiers to make comprehensive retrieval decisions. **(b) Non-training-based methods: FLARE** (Jiang et al., 2023), which determines the need for external retrieval based on the model's confidence in the next sentence generation. In addition, we also include **Never-Ret** (never using retrieving) and **Always-Ret** (always using retrieving) for LLMs as baselines.

4.4 Implementation Details

Backbones. We use LLaMA2-7B-Chat[2] and LLaMA2-13B-Chat[3], following the same setup as UAR (Cheng et al., 2024a).

Historical In-Context Acquisition. We encode each question in historical logs using the bge-large-en-v1.5 (Xiao et al., 2023) model and store their embeddings using NumPy[4]. During inference, historical in-context examples are retrieved by encoding the new query with bge-large-en-v1.5 and performing top-k nearest neighbor search via FAISS (Johnson et al., 2019). Through testing on the TriviaQA development set with k=10, 20, 30, 40, 50, we found that retrieving 20 historical examples yielded the best overall performance, while maintaining a relatively low retrieval rate. Therefore, we set k=20. The average retrieval time per query is 0.0083s. Please refer to Appendix B for RAG and inference details.

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[3]https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
[4]https://github.com/numpy/numpy

4.5 Overall Performance

The overall performance comparison of different methods is shown in Table 1. In non-training-based methods, DH-ICL substantially outperforms FLARE by an average of 3.24 points on the 7B model and 6.23 points on the 13B model. Moreover, in training-based methods, DH-ICL achieves performance comparable to the SOTA method UAR and clearly outperforms Self-RAG and SKR across all benchmarks. DH-ICL reduces retrieval overhead while outperforming Always-Ret in terms of average accuracy, particularly on the TriviaQA dataset, where it achieves comparable performance with only half the retrieval rate. Table 2 lists the inference time comparison between DH-ICL and UAR. It can be seen that both our proposed method and UAR, for both the 7B and 13B models, require less than 0.1s to decide whether retrieval is needed.

4.6 Performance of Self-Knowledge Boundary Assessment

This section evaluates the accuracy of knowledge boundary identification before AR on KB-SAT. Specifically, we incorporate DH-ICL into the model-confidence-based self-assessment described in Section 2 and compare it with UAR. As shown in the experimental results in Table 3, DH-ICL significantly improves the accuracy of direct self-assessment and effectively mitigates the prior tendency of models to bias self-knowledge assessment toward either overly positive or negative categories. Moreover, DH-ICL achieves performance close to that of UAR. More importantly, different from UAR, which requires training a classifier on LLM-generated logits, the proposed method operates without additional training, offering better generalization across different domains and models.

4.7 Ablation Analysis

The following ablation studies were conducted to analyze the contribution of each component in DH-

Type	Method	TriviaQA	WQ	TAQA	FreshQA	AVG
		Llama-2-7B Models				
-	Never-Ret	62.15 (0%)	59.79 (0%)	16.43 (0%)	35.64 (0%)	43.50
	Always-Ret	68.73 (100%)	53.99 (100%)	34.49 (100%)	65.35 (100%)	55.64
		Active Retrieval				
Training	Self-RAG (Asai et al., 2023)	61.68 (53.5%)	43.01 (61.9%)	11.09 (42.1%)	44.88 (51.2%)	40.17
	SKR (Wang et al., 2023)	65.39 (48.9%)	58.96 (26.8%)	30.63 (79.9%)	48.84 (39.3%)	50.96
	UAR (Cheng et al., 2024a)	**69.02** (50.1%)	**60.53** (25.0%)	**34.46** (99.7%)	**59.74** (78.5%)	**55.94**
Non-Training	FLARE (Jiang et al., 2023)	65.98 (58.8%)	55.46 (67.9%)	28.08 (63.5%)	57.76 (57.4%)	51.82
	DH-ICL (Ours)	**68.52** (49.8%)	**59.74** (29.1%)	**32.89** (91.6%)	**59.10** (77.2%)	**55.06**
		Llama-2-13B Models				
-	Never-Ret	63.18 (0%)	57.63 (0%)	11.14 (0%)	34.98 (0%)	41.73
	Always-Ret	71.02 (100%)	54.08 (100%)	34.20 (100%)	62.05 (100%)	55.34
Training	Self-RAG (Asai et al., 2023)	62.53 (30.0%)	42.37 (51.9%)	15.42 (37.0%)	39.60 (39.3%)	39.98
	SKR (Wang et al., 2023)	67.21 (49.2%)	56.20 (31.5%)	31.66 (89.2%)	50.17 (45.9%)	51.31
	UAR (Cheng et al., 2024a)	**71.71** (48.5%)	**59.20** (31.2%)	**34.14** (99.6%)	**55.45** (73.3%)	**55.13**
Non-Training	FLARE (Jiang et al., 2023)	68.00 (54.9%)	53.64 (69.6%)	25.40 (60.9%)	50.17 (55.8%)	49.30
	DH-ICL (Ours)	**70.42** (46.8%)	**58.71** (35.2%)	**33.60** (97.5%)	**58.40** (83.8%)	**55.28**

Table 1: Comparisons of downstream tasks performance. Never-Ret means that retrieval augmentation is never used during generation, while Always-Ret means that retrieval augmentation is used in every generation.

Method	Time per Sample (s)
UAR (Llama-2-7b)	0.0256
UAR (Llama-2-13b)	0.0481
DH-ICL (Llama-2-7b)	0.0440
DH-ICL (Llama-2-13b)	0.0827

Table 2: Time cost to determine if retrieval is required by the model.

Method	Acc	Pred True Num	Pred False Num
Ground Truth	-	3000	3000
UAR	70.18	3287	2713
Label (true/false)	50.03	5996	4
+DH-ICL	66.87	3522	2478
Label (1/2)	50.13	5986	14
+DH-ICL	67.77	2352	3648
Label (bar/foo)	52.98	5113	887
+DH-ICL	68.02	2417	3583

Table 3: Evaluate the accuracy of self-knowledge assessment on LLaMA-2-7B-Chat. Pred True/False Num represents the number of instances where the model predicts whether it knows the answer.

ICL: a. Removing the historical bias calibration; b. Excluding time-sensitive data from the historical logs; c. Removing the entire historical in-context learning and directly performing self-assessment. The ablation results are shown in Table 4. As seen from the results, all components contribute positively to improving active retrieval accuracy. Notably, when historical context retrieval is entirely removed and the model relies solely on its self-awareness to decide whether to trigger external retrieval, there is a clear drop in performance. These

Method	TriviaQA	WQ	TAQA	AVG
DH-ICL	**68.52**	**59.74**	**32.89**	**53.72**
w/o Bias Calibration	67.96	59.06	32.32	53.11
w/o Time sensitive	68.50	59.35	31.65	53.17
w/o Historical in-context	62.17	59.79	16.43	46.13

Table 4: Ablation results on LLaMA-2-7B-Chat.

Method	TriviaQA	WQ	TAQA	AVG
Never-Ret	62.15 (0%)	59.79 (0%)	16.43 (0%)	46.12
Always-Ret	68.73 (100%)	53.99 (100%)	34.49 (100%)	52.40
DH-ICL	68.52 (49.8%)	59.74 (29.1%)	32.89 (91.6%)	53.71
DH-ICL*	66.83 (25.1%)	59.06 (37.7%)	31.50 (84.0%)	52.46

Table 5: Cold-start robustness analysis on LLaMA-2-7B-Chat. DH-ICL* uses 20 fixed randomly selected logs from TriviaQA training set as in-context examples without any similarity-based historical case retrieval.

results demonstrate the effectiveness of DH-ICL in helping the LLM recognize its knowledge boundaries.

4.8 Cold-Start Robustness Analysis

Can the proposed DH-ICL method still be effective when similar historical cases are unavailable, especially in cold-start scenarios? To investigate this, we conducted a controlled experiment: 20 historical logs were randomly selected from the TriviaQA training set and fixed as the in-context examples for all benchmarks across the TriviaQA, WQ, and TAQA test sets, without performing any similarity-based historical case retrieval. As shown in Table 5, even with only 20 fixed random logs as context, DH-ICL achieves performance close to that of

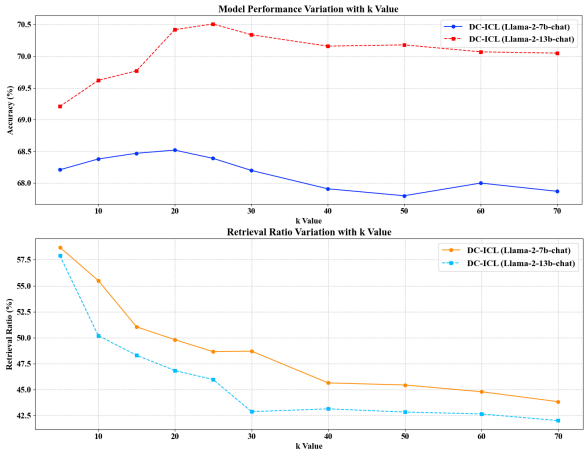

Figure 6: Variation in accuracy and retrieval ratio of DH-ICL on TriviaQA with changing in-context count.

full similarity-based retrieval and the Always-Ret baseline, and clearly outperforms the Never-Ret baseline. This suggests that the core advantage of DH-ICL lies in reformulating the metacognitive task into an exemplar-driven pattern learning problem, rather than depending on high-quality, domain-specific historical data. Consequently, DH-ICL is naturally well-suited for cold-start settings, with generalization and deployment potential.

4.9 Effect of Historical In-Context Count

How does the length of historical in-context affect the model's self-knowledge boundary assessment? To answer this question, we adjusted the number of in-context examples to conduct experiments, ranging from k=5 to 70. As can be observed from the experimental results in Figure 6 that, LLaMA-7B and 13B models achieve the best performance when using 20 and 25 in-context examples, respectively. Additionally, as the value of k increases, the retrieval rate decreases. After k=40, both accuracy and retrieval rate tend to stabilize, indicating that an excessive amount of context does not help improve the accuracy of the model's predictions.

4.10 Effect of Using Various True/False Tokens

As analyzed in Section 2, considering the model's inherent prior biases, how does the use of different tokens as labels impact the performance of DH-ICL? To answer this, we conducted experiments on DH-ICL using different tokens to represent whether the model knows the answer. As can be observed from experimental results in Table 6, DH-ICL achieves commendable performance

True/False	7b Models		13b Models	
	Acc	Ret. (%)	Acc	Ret. (%)
true/false	68.52	49.80	70.42	46.80
false/true	66.78	46.95	70.62	80.63
yes/no	67.86	43.16	69.80	39.49
bar/foo	68.38	57.69	69.84	39.70
foo/bar	68.77	62.95	70.10	47.09
1/2	**68.81**	57.71	**70.89**	59.65
2/1	68.41	54.92	70.37	46.91
ile/ore	68.59	62.47	69.59	42.43
ore/ile	68.77	68.38	70.22	46.96

Table 6: Accuracy and retrieval rate of DH-ICL using different tokens as labels on TriviaQA.

across various tokens, suggesting that the proposed method helps resist model bias to some extent. However, it is inevitable that the LLM is still influenced by its prior semantic knowledge, as using a "false" token to represent "true" semantics leads to a performance decrease. Notably, using tokens "1/2" tends to yield better performance, which we attribute to these tokens avoiding the introduction of prior preferences in the model's responses.

5 Related Work

Different from passive retrieval, which applies retrieval for every query in RAG methods (Lewis et al., 2020; Meng et al., 2022; Sachan et al., 2022; Pradeep et al., 2023), active retrieval (AR) (Gosselin and Cord, 2008; Su et al., 2024) aims to selectively retrieve external knowledge only when necessary. AR reduces unnecessary retrieval overhead and prevents potential interference from irrelevant or low-quality retrieved information. Some recent studies train lightweight classifiers to predict whether retrieval is needed based on input queries (Cheng et al., 2024a; Asai et al., 2023; Wang et al., 2023; Liu et al., 2024). For instance, Asai et al. (2023) uses ChatGPT to generate training examples where retrieval should be avoided for non-knowledge-intensive queries. Cheng et al. (2024a) introduces four orthogonal classification criteria, including intent, time, knowledge, and self-aware, to evaluate whether an LLM requires RAG. Another line of research explores model-internal confidence estimation, where LLMs determine their knowledge sufficiency based on response confidence scores (Jiang et al., 2023; Wang et al., 2024; Yao et al., 2024). For instance, Jiang et al. (2023) triggers retrieval only when the model exhibits high

uncertainty in its predictions. (Yao et al., 2024) evaluates self-aware uncertainty by computing the determinant of the regularized Gram matrix of hidden representations from multiple sampled generations. (Wang et al., 2024) assesses whether the model already knows the answer by explicitly prompting the LLM to output confidence scores.

A key challenge in active retrieval is enabling LLMs to recognize their knowledge boundaries (Chen et al., 2024; Kadavath et al., 2022b; Zhang et al., 2024; Cheng et al., 2024b), ensuring efficient retrieval by minimizing redundancy while enhancing knowledge augmentation when necessary. Despite progress made by approaches based on training classification models and leveraging model confidence scores, these methods often suffer from limited generalization ability and deployment efficiency across different datasets and LLMs. Our experiments further reveal that LLMs generally struggle with self-assessing their knowledge boundaries and exhibit prior cognitive biases. To this end, we propose DH-ICL to improve knowledge boundary recognition.

6 Conclusion

In this work, we introduced Debiased Historical In-Context Learning (DH-ICL) to enhance the reliability of knowledge boundary identification in active retrieval. Rather than relying on direct self-assessment, which LLMs inherently struggle with, DH-ICL reframes the problem as a structured pattern-learning task, leveraging historical high-confidence examples to guide retrieval decisions. Additionally, we proposed a historical bias calibration strategy, which adjusts logits deviations from historical responses to mitigate biases in current knowledge boundary identification. Extensive experiments across multiple QA benchmarks demonstrated that DH-ICL effectively reduces unnecessary retrievals while maintaining strong performance. Notably, our method is both efficient and generalizable, requiring no additional training. As users continue to interact with the LLM system, the historical experience log is continuously updated, leading to increasingly accurate knowledge boundary judgments over time. Furthermore, cold-start analysis shows that DH-ICL retains its effectiveness even in the absence of similar historical logs.

7 Limitations

Our approach benefits from continuously updating historical logs generated through sustained LLM usage, leading to increasingly precise knowledge boundary judgments. However, determining whether a historical query truly falls within the LLM's knowledge boundary still requires human verification. Automating this process through feedback-driven mechanisms remains an challenge. Future work could explore methods for automatically constructing and validating historical ICL data, reducing reliance on human intervention while maintaining accuracy in knowledge boundary recognition.

Acknowledgements

We thank the support of the National Natural Science Foundation of China (Grant No. 62406223), the Natural Science Foundation of Tianjin (Grant No. 24JCZDJC00130), and the research funding from Cangzhou Institute of Tiangong University (Grant No. TGCYY-Z-0303).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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*, pages 1533–1544.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024. Teaching large language models to express knowledge boundary from their own signals. *arXiv preprint arXiv:2406.10881*.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024a. Unified active retrieval for retrieval augmented generation. *arXiv preprint arXiv:2406.12534*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024b. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, and Jiayi Gui et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Philippe Henri Gosselin and Matthieu Cord. 2008. Active learning methods for interactive image retrieval. *IEEE Transactions on Image Processing*, 17(7):1200–1211.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel S 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)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022a. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022b. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke S. Zettlemoyer, and Scott Yih. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *ArXiv*, abs/2310.01352.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback. *arXiv preprint arXiv:2403.06840*.

Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Moorthy Bhat, and Yingbo Zhou. 2022. Augtriever: Unsupervised dense retrieval and domain adaptation by scalable data augmentation.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *ArXiv*, abs/2309.15088.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joëlle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Hongjin Su, Shuyang Jiang, Yuhang Lai, Haoyuan Wu, Boao Shi, Che Liu, Qian Liu, and Tao Yu. 2024. Evor: Evolving retrieval for code generation. In *Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross

Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam-fai Wong. 2024. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *arXiv preprint arXiv:2402.13514*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, and Haoran Wei et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215*.

Da Yin, Li Dong, Hao Cheng, Xiaodong Liu, Kai-Wei Chang, Furu Wei, and Jianfeng Gao. 2022. A survey of knowledge-intensive nlp with pre-trained language models. *ArXiv*, abs/2202.08772.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A Smith. 2024a. Set the clock: Temporal alignment of pretrained language models. *arXiv preprint arXiv:2402.16797*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024b. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858.

A ChatGPT Evaluator

As shown in Figure 7, We follow (Cheng et al., 2024a) to use the following prompt to evaluate whether the model-generated answers are correct on the TAQA (Zhao et al., 2024a) and FreshQA (Vu et al., 2023) datasets.

Prompt for ChatGPT Evaluation

In the following task, you are given a Question, a model Prediction for the Question, and a Ground-truth Answer to the Question. You should decide whether the model Prediction implies the Ground-truth Answer.

Question:
{question}

Prediction:
{predicted answer}
Ground-truth Answer:
{ground-truth answer}

Does the Prediction imply the Ground-truth Answer? Output Yes or No:

Figure 7: Prompt to evaluate accuracy on the TAQA and FreshQA datasets.

B Implementation details

RAG Settings: When retrieval is triggered, we follow prior works (Cheng et al., 2024a; Vu et al., 2023; Asai et al., 2023) to perform RAG. For TAQA and FreshQA, we use Google Search to retrieve the top-5 relevant documents. For TriviaQA and WQ, the Contriever-MS MARCO (Izacard et al., 2021) model are employed to retrieve the top-10 Wikipedia passages.

Inference Settings: Inference hyperparameters are listed in Table 7. All the experimental results in

this paper are obtained by testing three times and averaging the values.

Hyperparameters	Value
LLM float type	bf16
Top-k	50
Top-p	0.9
Temperature	0.6
Max input length	1,024
Max new tokens	512
Retrieval Question Max length	256
Retrieval Max Passage Length	256
GPU	2xA6000

Table 7: Inference hyperparameters.

C Case Study

Figure 10 presents a case from DH-ICL on LLaMA-2-7B-Chat. The number of retrieved similar in-context examples from historical response experience logs is 10. After concatenating the prompt with the retrieved historical instances, they are input into the model, which outputs its judgment with the historical bias calibration mechanism. We can see that, with the inspiration from the historical experience data, the model correctly acknowledges that it cannot answer the question, overcoming the issue of prediction inaccuracies caused by model bias.

D Analysis of Whether LLMs Know

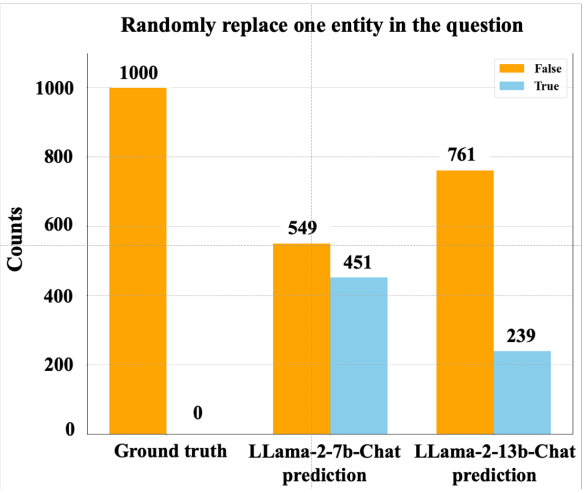

Figure 8: Randomly select an entity from the question and replace it with an entity extracted from other questions.

Although we have analyzed the impact of prior biases on model predictions in the main text, we conducted some additional interesting tests, which

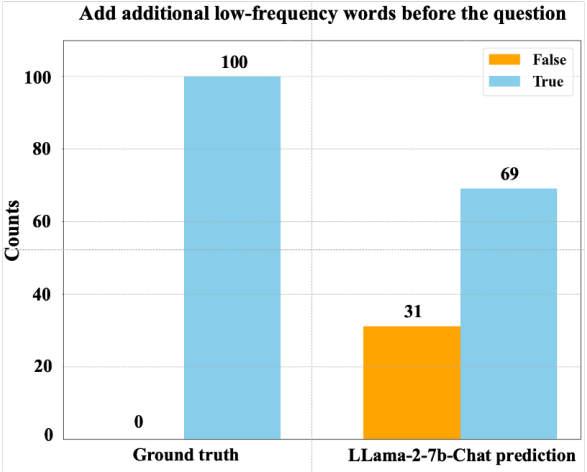

Figure 9: Add an irrelevant low-frequency word before each question, such as 'Atramentous' or 'Brobdingnagian'; these words do not alter the meaning of the question. The selected questions are those that the model can answer correctly without the low-frequency word.

are presented in this section. These include: (1) randomly replacing known entities in each known question to turn it into a fictitious unknown question, as shown in Figure 8; and (2) adding a low-frequency word before each known question as an interference factor, as shown in Figure 9.

Figure 8. We selected 1000 questions that the model can answer from KB-SAT, and used ChatGPT to extract the entities from these known questions to create an entity set X. We then randomly replaced an entity in each question with an entity from the set X. Using the set of fictitious questions generated by entity replacement, we evaluated the model's self-knowledge boundary as described in Section 2. Intuitively, the fictitious questions, created by random entity replacement, were almost impossible for the model to answer. However, 45% of these fictitious questions were still considered answerable by LLaMA-2-7B-Chat, and 24% were considered answerable by LLaMA-2-13B-Chat. We attribute this to the model's tendency to give positive responses to familiar entities or question structures, even when the questions are fabricated.

Figure 9. We selected 100 questions from KB-SAT that the model can answer, and used ChatGPT to randomly generate a low-frequency word, which was added before each question and separated by a separator. We then evaluated the model's self-knowledge boundary using these modified questions. In this experiment, 69% of the questions that were originally known to the model became

unanswerable after adding a low-frequency word. We hypothesize that the addition of low-confidence patterns caused the model to lean towards negative responses, even when the questions were known.

Case Study of the DH-ICL

You are a student being tested. For each given question, assess based on your knowledge whether you can answer it correctly. If you believe you can answer it correctly, output 'true'. If you are unsure whether you can answer it correctly, output 'false'. Additionally, if the question is asking about a recent event, for example, if words like recently, latest, or currently appear, also output 'false'.

Question: Who was President of the Screen Actor's Guild in the 40s and 50s?
Answer:false

Question: Which New York business woman manufactured 'Aramis' for men (1943)?
Answer:false

Question: Name the 1942 movie and character/actor that featured the quote "Of all the gin joints, in all the towns, in all the world, she walks into mine."
Answer:false

Question: The film which won the Oscar for Best Picture in 1942 also featured an Academy Award winning performance from Greer Garson. What was its title?
Answer:true

Question: Who were the two principal actors starring in the 1942 film 'Holiday Inn'?
Answer:false

Question: In which New York borough is Greenwich Village?
Answer:true

Question: Who made the longest Oscar acceptance speech, at five minutes and 30 seconds, when collecting her Oscar for Best Actress in 1942 for her role as Mrs. Miniver?
Answer:false

Question: Who plays Mrs Kay Miniver in the 1942 film 'Mrs Miniver'?
Answer:true

Question: Who starred opposite Walter Pidgeon in the 1942 film 'Mrs. Miniver'?
Answer:true

Question: Who was the actress wife of Clark Gable, killed in an air crash in 1942?
Answer:false

Question: Which actress was voted Miss Greenwich Village in 1942?
Answer:false ⇒ ***new query***

Figure 10: A case study of our DH-ICL method.