Knowledge Boundary Discovery for Large Language Models

Ziquan Wang[1,2], **Zhongqi Lu**[1,2*]

[1]College of Artificial Intelligence, China University of Petroleum-Beijing, China
[2]Hainan Institute of China University of Petroleum (Beijing), Sanya, Hainan, China
2021011537@student.cup.edu.cn, zhongqi@cup.edu.cn

Abstract

We propose **Knowledge Boundary Discovery (KBD)**, a reinforcement learning based framework to explore the knowledge boundaries of the Large Language Models (LLMs). We define the knowledge boundary by automatically generating two types of questions: (i) those the LLM can confidently answer (*within-knowledge boundary*) and (ii) those it cannot (*beyond-knowledge boundary*). Iteratively exploring and exploiting the LLM's responses to find its knowledge boundaries is challenging because of the hallucination phenomenon. To find the knowledge boundaries of an LLM, the agent interacts with the LLM under the modeling of exploring a partially observable environment. The agent generates a progressive question as the action, adopts an entropy reduction as the reward, receives the LLM's response as the observation and updates its belief states. We demonstrate that the KBD detects knowledge boundaries of LLMs by automatically finding a set of non-trivial answerable and unanswerable questions. We validate the KBD by comparing its generated knowledge boundaries with manually crafted LLM benchmark datasets. Experiments show that our KBD-generated question set is comparable to the human-generated datasets. Our approach paves a new way to evaluate LLMs.

1 Introduction

Large Language Models (LLMs) have made remarkable strides across diverse domains (Brown et al. 2020). Despite their impressive capabilities, a persistent challenge remains: LLMs often produce vague, inaccurate, or incorrect responses, commonly referred to as "hallucinations". Hallucinations arise primarily due to three reasons: (i) the LLM possesses relevant knowledge but generates unqualified responses, (ii) the LLM lacks the necessary knowledge but attempts to answer, and (iii) the LLM has no required knowledge. A key underlying issue is the LLM's inability to recognize its own *knowledge boundaries*, leading to overconfidence and unreliable answers beyond its expertise.

Prompt engineering is one major method to address hallucinations. While being effective when the LLM has sufficient knowledge, the prompt engineering approach struggles when the LLM does not have any related knowledge. In other words, prompt engineering can only handle the cases when the LLM possesses relevant knowledge but generates unqualified responses. The measure of LLMs' knowledge level is essential to the proper post processing. Therefore, various manually crafted benchmarks are published. However, even though many human efforts have been involved in generating these datasets, the static benchmarks fail to adapt dynamically to the LLM's evolving responses. To fully understand the knowledge capabilities of LLMs, it is necessary to engage in iterative interactions with the LLM, systematically probing and exploring its knowledge boundaries. This aligns naturally with reinforcement learning (RL) methods.

In this work, we define the knowledge boundary of LLMs by automatically generating two types of questions: those LLM can confidently and accurately answer (*within-knowledge boundary*) and those it cannot (*beyond-knowledge boundary*). This definition provides a more intuitive way to gauge the extent of an LLM's knowledge.

We propose **Knowledge Boundary Discovery (KBD)**, a novel reinforcement learning-based framework designed to detect LLMs' knowledge boundaries. Due to the *partially observable* nature of the environment, the agent must infer the hidden belief of the LLM's knowledge level based on the observation of its responses. Therefore, we adopt the Partially Observable Markov Decision Process (POMDP) modeling in our KBD framework. The KBD employs a reinforcement learning agent to iteratively generate progressive questions around the knowledge boundaries of the LLM, adopt an entropy reduction as the reward, receive the LLM's response as the observation and update the belief states.

To guide the search for the knowledge boundaries, we integrate *information gain* from information theory into the reward function. Information gain quantifies the reduction in uncertainty achieved through each interaction, aligning the rewards with the quality of the LLM's responses. Our intuition is that the information gain changes dramatically around the knowledge boundaries. This reward function allows the agent to identify areas where the LLM demonstrates stronger knowledge, as evidenced by decreasing the response entropy. Through iterative interactions, KBD dynamically detects the knowledge boundaries.

Moreover, our KBD-generated questions are *non-trivial*—they are not random or template-based, but are semantically meaningful and reside near the boundary where

*Corresponding author.

the LLM's knowledge becomes uncertain. This property is empirically supported by entropy-based clustering and embedding-space distance analysis, which shows that KBD-generated questions differ significantly from random ones while staying close to the knowledge boundary between known and unknown questions.

In summary, our contributions are as follows:

- We introduce a practical and intuitive definition of the knowledge boundary for LLMs, covering both confidently answerable and unanswerable questions, thereby offering a clear and operational criterion for evaluating LLM capabilities.
- We propose **Knowledge Boundary Discovery (KBD)**, a novel reinforcement learning framework that dynamically identifies knowledge boundaries via entropy-guided exploration. Experimental results demonstrate that KBD effectively discovers knowledge boundaries and generates question sets that are comparable in quality to human-curated benchmarks.

2 Related Work

Knowledge Boundaries and Hallucinations in LLMs

Understanding the knowledge boundaries of Large Language Models (LLMs) is essential for recognizing their limitations and guiding their safe and effective deployment. These boundaries encompass prompt-agnostic, prompt-sensitive, and unanswerable knowledge types (Zhang, Xu, and Cai 2024; Yin et al. 2024; Li et al. 2024). Despite their fluency and impressive performance, LLMs frequently suffer from hallucinations, i.e., producing confident but factually incorrect responses, stemming from limitations in training data, model architecture, and inference mechanisms (Maynez et al. 2020; Ye et al. 2023; Zhang et al. 2023). Such issues are further exacerbated by outdated or noisy data used during training (Penedo et al. 2023; Reddy et al. 2024).

To characterize the knowledge limitations of LLMs, recent studies have investigated how LLMs handle known versus unknown information, including dynamically shifting knowledge states. These works show that LLMs often overestimate their abilities, failing to recognize when a question is unanswerable (Yin et al. 2023; Amayuelas et al. 2024; Liu et al. 2024). Additional research has analyzed the effect of retrieval augmentation, showing that the quality of retrieved information directly influences the accuracy and reliability of LLM outputs (Zhang et al. 2024b; Yin et al. 2023; Ren et al. 2024).

To detect and mitigate hallucinations, several approaches have been developed to detect when an LLM generates hallucinated content. Pacchiardi (2023) proposed a simple lie detector to detect inaccurate answers. Chen (2024a) introduced the EigenScore metric to assess the internal consistency of responses, while Zhang (2024a) developed SELF-EVAL, a mechanism that enables LLMs to self-verify their factual output using internal knowledge alone.

To explicitly model and teach knowledge boundaries, other works aim to improve how LLMs recognize and express the limits of their knowledge. Chen (2024b) proposed COKE, a method that leverages internal confidence signals to teach LLMs to better recognize their knowledge boundaries. Pezeshkpour (2023) employed information-theoretic measures like entropy and KL-divergence to more accurately model factual confidence, helping detect hallucinations and improve retrieval-augmented generation. Zheng (2024) introduced KGLens, a sampling-based framework that uses knowledge graphs and Thompson sampling to efficiently identify factual blind spots in LLMs.

Reinforcement Learning for Knowledge Exploration

Reinforcement learning (RL) has advanced knowledge exploration by integrating querying mechanisms, goal conditioning, and large language models (LLMs). Approaches like Asking for Knowledge (AFK) enable RL agents to efficiently query external knowledge sources, addressing sparse rewards and large action spaces (Liu et al. 2022). Goal-conditioned methods, such as ReenGAGE, use knowledge distillation to transfer information and prioritize goals in high-dimensional spaces (Levine and Feizi 2023). Model-free techniques, such as RandQL, employ randomized learning rates for efficient exploration without Bayesian complexity (Tiapkin et al. 2023). LLMs are increasingly integrated into RL workflows, as seen in TWOSOME and other frameworks that align language models with decision-making tasks through fine-tuning and structured prompting (Tan et al. 2024; Gholamian and Huh 2024). Extreme Q-Learning (X-QL) further enhances RL by applying Extreme Value Theory to model Q-values and mitigate function approximation errors (Garg et al. 2023). These innovations enable autonomous agents to query, reason, and solve complex, knowledge-intensive problems.

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances generative models by integrating external knowledge retrieval. Relevant to our work are contributions to data augmentation and selection strategies. Lewis (2021) proposed the RAG framework, using the Dense Passage Retriever (DPR) to dynamically retrieve Top-k documents for improved contextual coverage. Zhang (2024) introduced RetrievalQA, which adaptively identifies whether retrieval is necessary, reducing computational overhead. Yue (2024) designed a framework that retrieves evidence to synthesize contrastive arguments, supporting or refuting claims. Yu (2023) addressed few-shot learning with retrieval objectives that prioritize task-relevant examples. Chen (2023) proposed RETROPROMPT, which retrieves training examples from an internal knowledge store to reduce reliance on parametric memory. Su (2024) propose DRAGIN, a dynamic RAG framework that can dynamically determine when to trigger the retrieval.

In conclusion, our KBD method, combining reinforcement learning with entropy-based measures, demonstrates significant advantages in dynamic question generation and automated validation. It not only enhances the ability to explore knowledge boundaries, but also effectively reduces response uncertainty.

3 Methodology

Reinforcement Learning Framework for KBD

KBD formulates the discovery of LLM knowledge boundaries as a reinforcement learning (RL) task. In this setup, an agent interacts with the target LLM, treated as a black-box environment, by generating questions (actions) and receiving observations. Each response is evaluated through entropy-based reward signals, guiding the agent toward areas of high or low uncertainty in model confidence.

Since the internal state of an LLM is inaccessible, the agent maintains a belief state based on observed responses. This naturally aligns the problem with a Partially Observable Markov Decision Process (POMDP), enabling a structured exploration of the LLM's latent knowledge.

POMDP Formulation

We formalize the task of exploring LLM knowledge boundaries as a Partially Observable Markov Decision Process (POMDP), represented by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O})$:

- $\mathcal{S}$: The set of states representing the knowledge level of the target LLM. Note that the true state $s \in \mathcal{S}$ is hidden and cannot be directly observed (e.g., whether the question within-knowledge boundary or beyond-knowledge boundary). Instead, we use the belief state b as an estimate of s for the computation. b is a probability distribution over $\mathcal{S}$, representing the probability distribution of the true states for each historical question, which is iteratively updated based on actions and observations.
- $\mathcal{A}$: The set of possible actions(generated questions).
- $\mathcal{T}(b'|b, a)$: The state transition function.
- $\mathcal{R}(b, a)$: The reward function(entropy-based).
- Ω: The set of all possible observations corresponding to the responses provided by the target LLM.
- $\mathcal{O}(o|b, a)$: The observation function.

This framework captures two essential aspects of interacting with the target LLMs:

- **Partial Observability**: The agent cannot access the internal computations or knowledge representation of the target LLM and must infer its state based on observed responses.
- **Sequential Decision-Making**: Each action (question) influences future observations and rewards, necessitating a strategy to balance exploration (investigating new areas) and exploitation (focusing on areas where the target LLM performs well).

Belief State Encoder To model the belief state b_t, we define an encoder function $f(\cdot)$ that computes a probability distribution over the hidden true state s (i.e., whether each question belongs to the within-knowledge boundary or the beyond-knowledge boundary) based on the entropy of the responses. For each interaction pair (a_i, o_i), E_i denote the response entropy of o_i. The probability that this interaction corresponds to a within-knowledge boundary is defined as:

$$P(s_i = \texttt{within} \mid E_i) = \sigma(\beta(E_{\text{th}} - E_i)), \tag{1}$$

where:

- $\sigma(\cdot)$ is the sigmoid function.
- E_{th} is the entropy threshold that separates within and beyond boundaries (e.g., 40 and 170 in our experiments).
- β is a scaling hyperparameter that controls the sharpness of the boundary.

The belief state at time t, denoted as b_t, is then a collection of these probabilities for each interaction up to time t:

$$\begin{aligned} b_t = \{&P(s_1 = \texttt{within} \mid E_1), \ldots, \\ &P(s_{t-1} = \texttt{within} \mid E_{t-1})\}. \end{aligned} \tag{2}$$

This probabilistic representation allows the agent to track and update its belief about the LLM's knowledge boundary based on the entropy trends in the dialogue history.

Action Space The action space $\mathcal{A}$ consists of all possible questions the agent can pose to the target LLM. At each time step t, the agent maintains a belief state b_t, and uses a learned policy $\pi(a_t \mid b_t)$ to select an action $a_t \in \mathcal{A}$. This policy ensures that the generated questions are contextually relevant and strategically guide the exploration of the target LLM's knowledge boundaries, by aiming to refine the agent's belief about the LLM's internal knowledge state.

Observation An observation o_t is the response provided by the target LLM when the agent's question a_t is posed:

$$o_t = \text{LLM}_{\text{target}}(a_t|b_t). \tag{3}$$

Observations include the textual content of the target LLM's response and its log-probability (logprob) in the vocabulary. This logprob is crucial for calculating the response entropy, which serves as a measure of the uncertainty of the target LLM.

Relationship between Belief State, Action, and Observation The belief state b_t summarizes the agent's understanding of the target LLM's knowledge level up to time t. After the agent takes action a_t and observes the response o_t, the belief state transitions to b_{t+1}, incorporating the new information:

$$b_{t+1} \leftarrow b_t \cup \{(a_t, o_t)\}. \tag{4}$$

This iterative structure ensures that the agent's decisions are informed by the cumulative interaction history, enabling adaptive refinement of its questioning strategy.

Entropy-Based Reward To encourage discovery of boundaries, we reward changes in entropy:

$$r_t = \left|E_{\text{prev}} - E_{\text{current}}\right|, \tag{5}$$

- E_{current} is the entropy of the target LLM's response at time t.
- E_{prev} is the entropy of the target LLM's previous response at time $t-1$.

The entropy E uses the logprob over the target LLM's vocabulary V:

$$E = -\sum_{w \in V} P(w) \log P(w), \tag{6}$$

where:

$P(w)$ is the probability assigned to word w. By incentivizing entropy reduction, the agent is motivated to pose questions that clarify the target LLM's responses and reduce uncertainty. Lower entropy indicates high LLM confidence (*within*), while high entropy signals uncertainty (*beyond*).

Q-Learning with Belief States

In a partially observable setting, the true state s of the environment is hidden and cannot be directly observed. Instead, the agent maintains a *belief state* b, which is an estimated probability distribution over all possible states $s \in \mathcal{S}$. The belief state serves as an approximation of the true state, summarizing the agent's knowledge of the environment based on its actions and observations. We define the Q-function as:

$$Q^*(b,a) \;=\; E\big[R(b,a) + \gamma \max_{a'} Q^*(b',a') \mid b\big], \qquad (7)$$

where $\gamma \in [0,1)$ is the discount factor, and $E[\cdot \mid b]$ takes the expectation with respect to the belief $b \sim$ the hidden state s. At each time step t:

1. The agent has a current belief state b_t.
2. Choose an action a_t using ϵ-greedy policy over $Q(b_t, a)$.
3. It executes a_t and receives observation o_t (the LLM's response).
4. Calculate the immediate reward r_t (e.g., the change in response entropy).
5. Update its belief state to b_{t+1}.
6. Update the Q-value $Q_t(b_t, a_t)$:

$$\begin{aligned} Q_{t+1}(b_t,a_t) \;\leftarrow\; Q_t(b_t,a_t) \;+\; \alpha\Big[r_t + \\ \gamma \max_{a'} Q_t(b_{t+1},a') \;-\; Q_t(b_t,a_t)\Big], \end{aligned} \qquad (8)$$

where α is the learning rate.

In this Q-learning framework, the agent not only benefits from the immediate reward r_t, but also from the potential future rewards, captured by the term $\gamma \max_{a'} Q_t(b_{t+1}, a')$.

ϵ-Greedy Action Selection To balance exploration and exploitation, the agent adopts an ϵ-greedy action selection strategy:

$$a_t \;=\; \begin{cases} \text{random action from } \mathcal{A}, & \text{with probability } \epsilon, \\ \arg\max_{a \in \mathcal{A}} Q(b_t,a), & \text{with probability } 1-\epsilon, \end{cases} \qquad (9)$$

where $\mathcal{A}$ denotes the action space (i.e., candidate questions), and $Q(b_t, a)$ is the estimated value of taking action a under belief state b_t.

Exploration (ϵ branch) injects stochastic questions that can push the dialogue into regions the agent has not yet sampled, which is crucial to uncovering *beyond-knowledge boundary* gaps. Exploitation (greedy branch) focuses on high-value questions expected to maximize entropy reduction, refining the *within-knowledge boundary* area. Each episode begins with a randomly initialized question, which may fall either within or beyond the LLM's knowledge boundary. Even if a single episode fails to reach the beyond-knowledge region, the aggregated effect across many episodes enables a comprehensive boundary exploration.

Algorithm 1: Updating Strategy for KBD

Require: Learning rate α, exploration rate ϵ, discount factor γ, number of episodes N, maximum steps per episode M, entropy threshold $E_{\text{threshold}}$ and topic T

Initialize $Q(b,a) \leftarrow 0$ for all belief states b and actions a
for episode $= 1$ to N **do**
  Initialize prior belief b_1 (e.g., uniform or from domain knowledge) with topic T
  Initialize previous entropy $E_{\text{prev}} \leftarrow \infty$
  for $t = 1$ to M **do**
    With probability ϵ, select a random action $a_t \in \mathcal{A}$
    Otherwise, select $a_t \leftarrow \arg\max_a Q(b_t, a)$
    Execute a_t, **receive observation** o_t from the LLM
    Compute current entropy E_{current} from o_t
    Calculate reward: $r_t = |E_{\text{prev}} - E_{\text{current}}|$
    Update Q-value by Eq. 8
    Update belief state $b_{t+1} \leftarrow b_t \cup \{(b_t, o_t)\}$
    Update previous entropy: $E_{\text{prev}} \leftarrow E_{\text{current}}$
    if $E_{\text{current}} < E_{\text{threshold}}$ **then**
      break
    end if
  end for
end for

Moreover, our method supports flexible reward shaping and policy guidance. By modifying the reward function (e.g., rewarding high-entropy responses) and adjusting the exploration bias, the agent can be directed to actively search for questions that fall into the *beyond-knowledge boundary*. This design ensures that the agent systematically probes both sides of the knowledge boundary, reinforcing the dual discovery capability of our proposed algorithm.

4 Experiments

In this section, we first validate the effectiveness of using entropy to define and identify the knowledge boundaries of LLMs, and demonstrate that the discovered boundaries are meaningful and non-trivial. We then show that the question set generated by our KBD method achieves results comparable to human-generated datasets used in previous work.

Experimental Setup

To evaluate the effectiveness of our proposed KBD framework, we adopt a dual-model architecture comprising: (1) a powerful supervised model responsible for question generation and (2) smaller target LLMs whose knowledge boundaries are to be explored. Importantly, the supervised model only generates questions and does not participate in answering them, ensuring that the identified knowledge boundaries reflect the limitations of the target models alone.

- **Target LLMs:** We examine the knowledge boundaries of three target models: ChatGLM3-6B, ChatGLM2-6B

Method	KUQ			Sware			Infeasible Benchmark		
	$\mathbf{K_{aware}}$	$\mathbf{U_{aware}}$	$\mathbf{S_{aware}}$	$\mathbf{K_{aware}}$	$\mathbf{U_{aware}}$	$\mathbf{S_{aware}}$	$\mathbf{K_{aware}}$	$\mathbf{U_{aware}}$	$\mathbf{S_{aware}}$
Min-Prob	78.8	24.0	51.4	82.5	17.3	49.9	76.3	25.6	51.0
Fst-Prob	51.0	36.9	43.9	94.4	5.2	49.8	65.5	27.1	41.3
Prod-Prob	48.6	61.0	54.8	95.4	30.6	63.0	86.4	18.7	52.5
Prior	83.0	40.1	61.5	80.8	37.7	59.2	74.8	42.9	58.9
Posterior	76.0	17.3	46.6	83.0	16.1	49.5	85.2	17.4	51.3
IC-IDK	88.0	13.5	50.7	80.5	19.2	49.8	90.9	18.5	54.7
Verb	95.5	15.9	55.7	99.0	20.8	59.9	96.1	30.2	63.1
Entropy	61.4	78.2	**69.8**	76.7	56.5	**66.6**	80.1	60.3	**70.2**

Table 1: Performance comparison of entropy confidence estimation with several baselines on three human-generated datasets: KUQ, Sware and Infeasible Benchmark. The metrics include: K_{aware} (accuracy on answerable questions), U_{aware} (rate of correctly indicating "unknown" for unanswerable questions), and $S_{\text{aware}} = \frac{1}{2}(K_{\text{aware}} + U_{\text{aware}})$, which reflects overall self-awareness. Entropy consistently achieves the highest S_{aware} on all datasets, and shows notably strong performance in U_{aware}, demonstrating its effectiveness at identifying unanswerable questions and delineating knowledge boundaries.

(both from Zhipu AI (GLM et al. 2024)), and LLaMA 7B-Chat. These models serve as the environments that interact with the RL agent to reveal their knowledge boundaries.

- **Supervised Model:** We use the more powerful GLM-4-9b model (GLM et al. 2024) as the supervised model to generate questions. Due to its larger parameter scale and broader coverage, we assume that GLM-4-9B has a wider knowledge boundary than the target models.

In each episode, the supervised model proposes a question to the target LLM, then the target LLM answers. And based on the entropy of the response, the agent updates its strategy. Through repeated interaction, the agent progressively identifies regions where the target model answers with high confidence (within-knowledge boundary) versus high uncertainty (beyond-knowledge boundary).

Effectiveness of Entropy as a Confidence Estimation Metric

Confidence estimation serves as a critical tool for evaluating whether an LLM's response falls within or beyond its knowledge boundaries. In our framework, entropy plays a central role as a unified metric that captures both the confidence of the model and the depth of its knowledge.

We first evaluate entropy as a confidence estimation metric by comparing it with several existing baselines using three human-generated datasets: KUQ (Amayuelas et al. 2024), Sware (Yin et al. 2023) and Infeasible Benchmark (Zhang, Xu, and Cai 2024). Following the evaluation setup of Chen (2024b), we include the following baseline methods: Min-Prob, Prod-Prob, Fst-Prob, Prior Prompt (Ren et al. 2024), Posterior Prompt (Kadavath et al. 2022), In-Context IDK (IC-IDK) (Cohen et al. 2023), and Verbalized Uncertainty (Verb) (Tian et al. 2024).

As shown in Table 1, entropy achieves competitive or superior performance on all metrics. In particular, it excels in identifying unanswerable questions (U_{aware}) and achieves the highest overall self-awareness score (S_{aware}) on both datasets. These results demonstrate the effectiveness of entropy in identifying unanswerable questions and delineating knowledge boundaries, outperforming probability-based confidence scores in capturing model uncertainty.

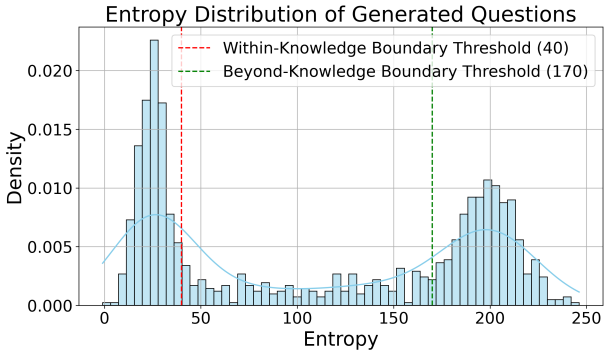

Figure 1: Distribution of entropy values for over 2,000 responses generated by our KBD algorithm across various topics and parameter configurations. The histogram and KDE curve reveal two prominent peaks: one in the low-entropy region ($\leq$ 40), and another in the high-entropy region ($\geq$ 170). These correspond to within-knowledge and beyond-knowledge boundaries, respectively. Only about 20% of the questions fall into the mid-entropy range (40 $<$ entropy $<$ 170), suggesting that the transition zone (i.e. where the model's knowledge is ambiguous) is narrow.

Furthermore, to illustrate how entropy delineates the knowledge boundary, we analyze the entropy distribution over a large set of KBD-generated questions. Figure 1 overlays a histogram with a kernel density estimation (KDE) curve (i.e., a smooth nonparametric density estimate obtained by summing Gaussian kernels over all samples). The result is a clean bimodal distribution: one mode cen-

Within-Knowledge Boundary	Beyond-Knowledge Boundary
Portico of the Aetolians and Delphi refer to what culture? Entropy: 26.30	What challenges might future telemedicine face? Entropy: 205.71
Which Enlightenment thinker was against the separation of powers? Entropy: 27.84	Are there limits to human creativity? Entropy: 211.29
What treaty ended the Russo-Persian War? Entropy: 26.63	What sort of life exists in the center of the cosmos? Entropy: 197.64

Table 2: Random sample from the KBD-generated dataset. Questions with entropy values $\leq$ 40 are classified as within-knowledge boundary, while those with entropy values $\geq$ 170 are categorized as beyond-knowledge boundary.

tered around entropy $\leq$ 40 (answerable) and the other beyond $\geq$ 170 (unanswerable), with a narrow transition region between. This empirical evidence reinforces our threshold selection and validates entropy's effectiveness in marking clear knowledge boundaries. The detailed procedure for selecting the boundary thresholds based on entropy is provided in Appendix A.

Then to verify that the questions generated by our KBD algorithm are non-trivial, that is, they are not random or easily separable but instead lie meaningfully near the model's decision boundary, we analyze their distribution in the semantic embedding space. Specifically, we embed 1,000 questions from each of three categories: *answerable* (within-knowledge boundary), *unanswerable* (beyond-knowledge boundary), and *randomly generated*. We use a sentence-level embedding model to obtain high-dimensional representations, which we then cluster before applying t-SNE for 2D visualization. As shown in Figure 2, the structure supports the validity of our identified knowledge boundary: The questions form distinct clusters that meaningfully distributed near the knowledge boundary, demonstrating that KBD-generated samples are informative and non-trivial.

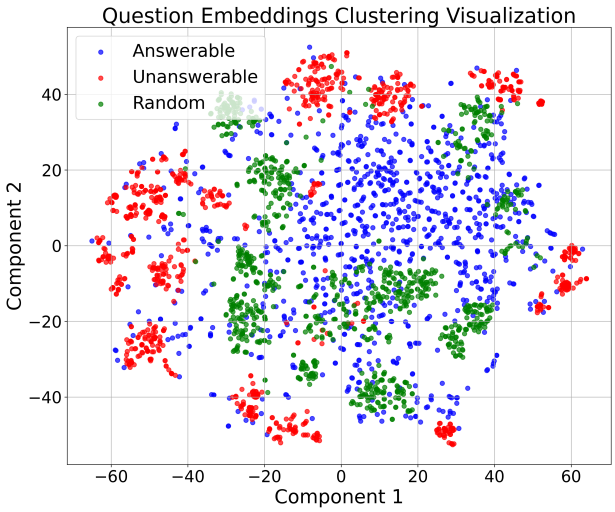

Figure 2: t-SNE visualization of question embeddings. Blue: answerable questions (within-knowledge boundary) form a central cluster. Red: unanswerable questions (beyond-knowledge boundary) form a surrounding band. Green: random questions are diffusely scattered. This structure supports the non-triviality of KBD-generated samples.

KBD-Generated Examples and Boundary Analysis

Using KBD, we automatically generate questions that the target LLM can or cannot answer with high confidence. Table 2 shows typical *within-knowledge* and *beyond-knowledge* items, demonstrating that our algorithm accurately identifies the knowledge frontier.

We applied the method to both *specialized technical domains*—such as clinical medicine, biotechnology (Singhal et al. 2022), and fundamental science concepts (Wang et al. 2024)—and four broad academic areas: social sciences, natural sciences, applied sciences and humanities. In every setting, KBD located coherent boundaries; further examples are provided in Appendix B.

The questions whose entropies lie in the narrow transition band ($40 < E < 170$) reveal the upper and lower bounds of the model's knowledge, where the answers become ambiguous. Some examples are provided in Appendix C.

Comparison of KBD-generated Dataset with Other Human-generated Dataset

Our KBD-generated dataset achieves performance comparable to that of other human-generated datasets. To validate this, we compared our question set with the KUQ dataset (Amayuelas et al. 2024). Following KUQ, they define a similarity function, f_{sim}, as a binary metric between the generated text (t_i) and some reference text (ref_i) to be 1 if they express the same content or 0 if they do not (the reference texts are a predefined set of phrases that encompass general uncertainty, and the full list can be found in Appendix D). Specifically, if the reference text is contained in the generated text or the similarity measured with SimCSE (2022) exceeds a threshold τ, the function returns 1:

$$\text{Sim}_i = f_{\text{sim}}(\text{t}_i, \text{ref}_i) \tag{10}$$

In our evaluation, two metrics are adopted: the F1 score and the Equal Error Rate (EER). The F1 score, derived from the similarity metric, evaluates the positive class (e.g., unknown questions or the chosen category). The EER measures the balance between false acceptance and false rejection rates, providing a holistic view of the LLM's performance on the dataset.

As shown in Table 3, models exhibit similar performance

Model	KBD(Ours)		KUQ	
	EER	F1	EER	F1
LlaMA 7B-Chat	0.239	0.725	0.301	0.732
ChatGLM3-6B	0.267	0.507	0.325	0.484
ChatGLM2-6B	0.508	0.431	0.514	0.449

Table 3: Performance comparison of different LLMs on KBD-generated dataset and KUQ dataset. Metrics include Equal Error Rate (EER) and F1 score (lower EER, better performance, higher F1 scores, better performance). The table demonstrates that the EER and F1 scores of our KBD-generated dataset show similar performance to the KUQ dataset across different models, validating that our KBD-generated dataset comparable to human-generated datasets.

in the KBD-generated and KUQ datasets. This validates the effectiveness of our automated question generation framework. Additionally, consistent with previous findings, models with larger parameter counts(e.g., LlaMA 7B-Chat) tend to perform better in boundary recognition tasks.

Comparison of KBD Algorithm with Human Expert Questioning and Random Questioning

We compare our KBD algorithm with two baselines: expert questioning and random questioning, evaluated using the entropy metric. Expert questioning represents the upper bound, while random questioning serves as the lower bound.

For expert questioning, domain experts iteratively asked questions to explore the LLM's knowledge boundary. For random questioning, questions were randomly selected from the science book *Hundred Thousand Whys* and posed without considering context or relevance.

As shown in Figure 3, both our proposed KBD algorithm and the expert questioning baseline exhibit a decreasing trend in entropy over successive rounds of interaction. They successfully identify the within-knowledge boundary of the target LLM (entropy ≤ 40) and maintain convergence in subsequent responses. In contrast, random questioning fails to reach the boundary within 50 rounds and exhibits significant entropy fluctuations.

This comparison demonstrates the effectiveness of the KBD in identifying knowledge boundaries. Although it is slower than expert questioning, it significantly outperforms random questioning, providing a scalable and automated solution for exploring LLM knowledge boundaries.

Convergence of KBD Algorithm

This section highlights the effectiveness of our algorithm in learning an optimal strategy during training. As shown in Figure 4, the cumulative reward stabilizes between 110 and 120 after approximately 50 episodes, indicating convergence. This shows that the algorithm progressively optimizes its strategy to achieve higher rewards and effectively explore the knowledge boundaries of the LLM.

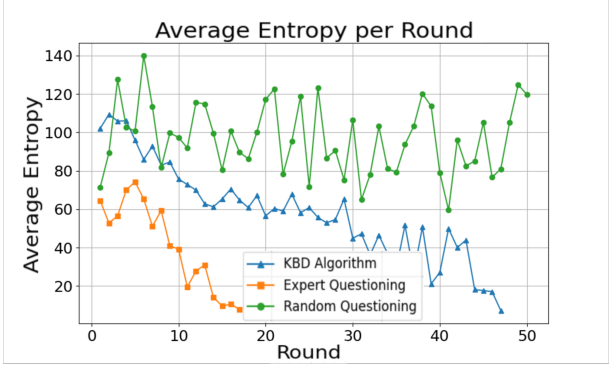

Figure 3: The figure illustrates the average entropy over 50 rounds across 1000 episodes for our KBD algorithm, the expert questioning baseline, and the random questioning baseline. It shows that both KBD algorithm and expert questioning effectively explore the LLM's knowledge boundary. However, the random questioning baseline fails to converge or reach the knowledge boundary.

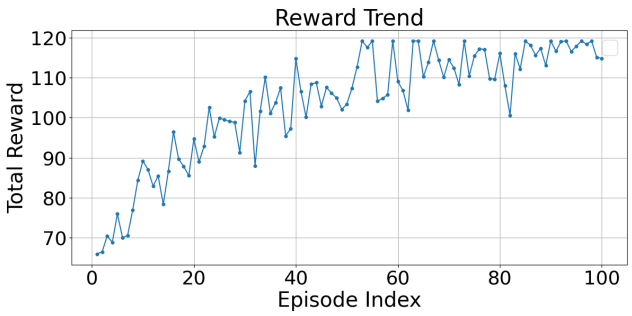

Figure 4: The cumulative reward per episode increases with training episodes and eventually converges. This indicates that KBD is consistently learning to optimize strategies.

5 Conclusion

In this work, we introduce a practical and operational definition of the knowledge boundary for LLMs, distinguishing between the questions that the model can confidently answer (*within-knowledge boundary*) and those it cannot (*beyond-knowledge boundary*). Building upon this foundation, we propose **Knowledge Boundary Discovery (KBD)**, the first reinforcement learning framework that dynamically explores and uncovers these boundaries through entropy-guided interaction. Through extensive experiments, we demonstrate the effectiveness and superiority of entropy as a confidence estimation metric, outperforming traditional probability-based baselines in identifying unanswerable questions and capturing model uncertainty. Furthermore, the KBD-generated dataset that we construct is both well structured and meaningful, achieving performance comparable to human-generated datasets. These findings validate the reliability of our proposed framework and the soundness of our knowledge boundary definition for LLMs.

Acknowledgments

This work is supported by the Science Challenge Project (Grant No. TZ2025008), the Science Foundation of China University of Petroleum, Beijing (Grant No. 2462023YJRC024) and the Frontier Interdisciplinary Exploration Research Program of China University of Petroleum, Beijing (Grant No. 2462024XKQY003). Zhongqi Lu is the corresponding author.

References

Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; and Wang, W. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. arXiv:2305.13712.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, C.; Liu, K.; Chen, Z.; Gu, Y.; Wu, Y.; Tao, M.; Fu, Z.; and Ye, J. 2024a. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. arXiv:2402.03744.

Chen, L.; Liang, Z.; Wang, X.; Liang, J.; Xiao, Y.; Wei, F.; Chen, J.; Hao, Z.; Han, B.; and Wang, W. 2024b. Teaching Large Language Models to Express Knowledge Boundary from Their Own Signals. arXiv:2406.10881.

Chen, X.; Li, L.; Zhang, N.; Liang, X.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2023. Decoupling Knowledge from Memorization: Retrieval-augmented Prompt Learning. arXiv:2205.14704.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12621–12640. Singapore: Association for Computational Linguistics.

Gao, T.; Yao, X.; and Chen, D. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821.

Garg, D.; Hejna, J.; Geist, M.; and Ermon, S. 2023. Extreme Q-Learning: MaxEnt RL without Entropy. arXiv:2301.02328.

Gholamian, S.; and Huh, D. 2024. Reinforcement Learning Problem Solving with Large Language Models. arXiv:2404.18638.

GLM, T.; :; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; and Lai, H. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Levine, A.; and Feizi, S. 2023. Goal-Conditioned Q-Learning as Knowledge Distillation. arXiv:2208.13298.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, M.; Zhao, Y.; Deng, Y.; Zhang, W.; Li, S.; Xie, W.; Ng, S.-K.; and Chua, T.-S. 2024. Knowledge Boundary of Large Language Models: A Survey. arXiv:2412.12472.

Liu, G.; Wang, X.; Yuan, L.; Chen, Y.; and Peng, H. 2024. Examining LLMs' Uncertainty Expression Towards Questions Outside Parametric Knowledge. arXiv:2311.09731.

Liu, I.-J.; Yuan, X.; Côté, M.-A.; Oudeyer, P.-Y.; and Schwing, A. G. 2022. Asking for Knowledge: Training RL Agents to Query External Knowledge Using Language. arXiv:2205.06111.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. arXiv:2005.00661.

Pacchiardi, L.; Chan, A. J.; Mindermann, S.; Moscovitz, I.; Pan, A. Y.; Gal, Y.; Evans, O.; and Brauner, J. 2023. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. arXiv:2309.15840.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116.

Pezeshkpour, P. 2023. Measuring and Modifying Factual Knowledge in Large Language Models. arXiv:2306.06264.

Reddy, R. G.; Lee, D.; Fung, Y. R.; Nguyen, K. D.; Zeng, Q.; Li, M.; Wang, Z.; Voss, C.; and Ji, H. 2024. SmartBook: AI-Assisted Situation Report Generation for Intelligence Analysts. arXiv:2303.14337.

Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; Wen, J.-R.; and Wang, H. 2024. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. arXiv:2307.11019.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Scharli, N.; Chowdhery, A.; Mansfield, P.; y Arcas, B. A.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2022. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138.

Su, W.; Tang, Y.; Ai, Q.; Wu, Z.; and Liu, Y. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Information Needs of Large Language Models. arXiv:2403.10081.

Tan, W.; Zhang, W.; Liu, S.; Zheng, L.; Wang, X.; and An, B. 2024. True Knowledge Comes from Practice: Aligning LLMs with Embodied Environments via Reinforcement Learning. arXiv:2401.14151.

Tian, K.; Mitchell, E.; Yao, H.; Manning, C. D.; and Finn, C. 2024. Fine-Tuning Language Models for Factuality. In *The Twelfth International Conference on Learning Representations*.

Tiapkin, D.; Belomestny, D.; Calandriello, D.; Moulines, E.; Munos, R.; Naumov, A.; Perrault, P.; Valko, M.; and Menard, P. 2023. Model-free Posterior Sampling via Learning Rate Randomization. arXiv:2310.18186.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. arXiv:2307.10635.

Ye, H.; Liu, T.; Zhang, A.; Hua, W.; and Jia, W. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv:2309.06794.

Yin, X.; Zhang, X.; Ruan, J.; and Wan, X. 2024. Benchmarking Knowledge Boundary for Large Language Models: A Different Perspective on Model Evaluation. arXiv:2402.11493.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? arXiv:2305.18153.

Yu, G.; Liu, L.; Jiang, H.; Shi, S.; and Ao, X. 2023. Retrieval-Augmented Few-shot Text Classification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6721–6735. Singapore: Association for Computational Linguistics.

Yue, Z.; Zeng, H.; Shang, L.; Liu, Y.; Zhang, Y.; and Wang, D. 2024. Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments. arXiv:2406.09815.

Zhang, W.; Xu, Z.; and Cai, H. 2024. Defining Boundaries: A Spectrum of Task Feasibility for Large Language Models. arXiv:2408.05873.

Zhang, X.; Peng, B.; Tian, Y.; Zhou, J.; Jin, L.; Song, L.; Mi, H.; and Meng, H. 2024a. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation. arXiv:2402.09267.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219.

Zhang, Z.; Fang, M.; and Chen, L. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 6963–6975. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, Z.; Wang, X.; Jiang, Y.; Chen, Z.; Mu, F.; Hu, M.; Xie, P.; and Huang, F. 2024b. Exploring Knowledge Boundaries in Large Language Models for Retrieval Judgment. arXiv:2411.06207.

Zheng, S.; Bai, H.; Zhang, Y.; Su, Y.; Niu, X.; and Jaitly, N. 2024. KGLens: Towards Efficient and Effective Knowledge Probing of Large Language Models with Knowledge Graphs. arXiv:2312.11539.