Perception of Knowledge Boundary for Large Language Models through Semi-open-ended Question Answering

Zhihua Wen[1]†, **Zhiliang Tian**[1]*, **Zexin Jian**[1], **Zhen Huang**[1],
Pei Ke[2], **Yifu Gao**[1], **Minlie Huang**[2], **Dongsheng Li**[1]*
[1] College of Computer, National University of Defense Technology, Hunan, China
[2] Tsinghua University, Beijing, China
{zhwen, tianzhiliang, jianzexin21, gaoyifu, huangzhen, dsli}@nudt.edu.cn
kepei1106@outlook.com, aihuang@tsinghua.edu.cn

Abstract

Large Language Models (LLMs) are widely used for knowledge-seeking purposes yet suffer from hallucinations. The knowledge boundary of an LLM limits its factual understanding, beyond which it may begin to hallucinate. Investigating the perception of LLMs' *knowledge boundary* is crucial for detecting hallucinations and LLMs' reliable generation. Current studies perceive LLMs' knowledge boundary on questions with concrete answers (close-ended questions) while paying limited attention to *semi-open-ended questions* that correspond to many potential answers. Some researchers achieve it by judging whether the question is answerable or not. However, this paradigm is not so suitable for semi-open-ended questions, which are usually "partially answerable questions" containing both answerable answers and ambiguous (unanswerable) answers. Ambiguous answers are essential for knowledge-seeking, but they may go beyond the knowledge boundary of LLMs. In this paper, we perceive the LLMs' knowledge boundary with semi-open-ended questions by discovering more ambiguous answers. First, we apply an LLM-based approach to construct semi-open-ended questions and obtain answers from a target LLM. Unfortunately, the output probabilities of mainstream black-box LLMs are inaccessible to sample for low-probability ambiguous answers. Therefore, we apply an open-sourced auxiliary model to explore ambiguous answers for the target LLM. We calculate the nearest semantic representation for existing answers to estimate their probabilities, with which we reduce the generation probability of high-probability existing answers to achieve a more effective generation. Finally, we compare the results from the RAG-based evaluation and LLM self-evaluation to categorize four types of ambiguous answers that are beyond the knowledge boundary of the target LLM. Following our method, we construct a dataset to perceive the knowledge boundary for GPT-4. We find that GPT-4 performs poorly on semi-open-ended questions and is often unaware of its knowledge boundary. Besides, our auxiliary model, LLaMA-2-13B, is effective in discovering many ambiguous answers, including correct answers neglected by GPT-4 and delusive wrong answers GPT-4 struggles to identify.

†Work done during internship at the CoAI Group.
*Corresponding Authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-2820

1 Introduction

Large language models (LLMs) have revolutionized our interactions with AI, enabling users to acquire knowledge by posing questions in natural language [12, 6]. However, LLMs are prone to hallucination and generate non-factual responses, hindering the development of trustworthy AI.

One main cause of LLM hallucination is its unfamiliarity with the long-tail knowledge that appears less frequently than common-sense knowledge in the training data. To alleviate this issue, many researchers collect more domain-specific training data [27] or incorporate external information [36, 29] via retrieval-augmented generation (RAG) during inference. Another line of work investigates the perception of `knowledge boundaries` for LLMs, which indicates the extent of knowledge that the LLM can grasp well, beyond which it may begin to hallucinate [18]. Studying the perception of knowledge boundaries for LLMs helps alleviate hallucinations in many ways. For example, 1) It helps detect the hallucinations of a target LLM and the extent and scope of its factual knowledge [16, 45]. 2) It helps align LLMs for more honest generation [43, 42].

Existing studies on the perception of knowledge boundaries are primarily in the form of Question-Answering (QA). Their methods mainly aim to judge whether a question is answerable or unanswerable and regard their border as the knowledge boundary. An answerable question refers to when the LLM is capable of generating a response matching the ground truth, and conversely, an unanswerable question means unable to answer correctly. These studies can be categorized into two groups. Prompt-based perception employs prompt engineering [45, 14] to assess whether the LLM can answer the question via LLM self-evaluation. They question whether the LLM knows the answer [32, 8, 46] or needs external knowledge to answer the question [36]. As LLMs tend to be overconfident [29, 32, 17, 49], more researchers explore representation-based perception. These studies optimize different representations for answers with different answerability [14, 10, 36, 49] or extract representations from a fixed encoder to train a classifier [23].

However, directly discriminating questions into answerable and unanswerable ones may not apply to some partially answerable questions. In many scenarios, the questions are relatively open-ended (i.e. having a list of correct answers) that may include (1) a subset of easy answerable answers, and (2) a subset of hard and unpopular answers, which may be unanswerable. These questions (referred to as "semi-open-ended questions") are particularly challenging and knowledge-extensive. Investigating the ambiguous answers to these semi-open-ended questions in various fields benefits knowledge-seeking. Ambiguous answers often go beyond the knowledge boundaries of LLMs and could lead to misinformation (see App. G). Therefore, we argue that investigating these questions with their ambiguous answers can augment the perception of the knowledge boundaries for LLMs.

In this paper, we propose to perceive the knowledge boundary for a target LLM with semi-open-ended questions by discovering pieces of unfamiliar knowledge where the LLM learns badly. Particularly, We first construct a dataset with semi-open-ended questions on the open domain and query the target LLM for their corresponding answers. We define the low-probability correct answers and delusive incorrect answers are the ambiguous answers corresponding to the LLM's unfamiliar knowledge.

A challenge is that obtaining LLMs' low-probability answers needs accessing LLMs' output probabilities (or violently sampling LLMs' outputs many times to approximate the probabilities), which is inaccessible (or expensive) for mainstream black-box LLMs, i.e. GPT-4. Therefore, we approximate the generation probabilities of the target LLM with an open-sourced auxiliary model. We use the Pseudo-inverse of model embedding to estimate the nearest semantic representation for the existing answers. Consequently, we obtain the probability distribution of existing answers and repetitively filter the existing answers (and their semantic-related counterparts) to obtain answers with low-probabilities. Finally, we recognize answers beyond the knowledge boundary of the target LLM by comparing its self-evaluation results against the ground truth answers obtained from RAG-based evaluation.

Empirically, we use our method to construct a dataset of approximately 1k samples and evaluate GPT-4's performance. We find that GPT-4 makes mistakes in 82.90% of questions and 40.15% of its ambiguous answers generated are unqualified. Besides, GPT-4 also makes inaccurate self-evaluation 28.77% of the time, indicating that these are beyond the knowledge boundary of GPT-4. Moreover,

For example, when asked to "Tell me about some exercise habits that are easy to overlook but are good for your health." there are many correct answers, yet the question remains constrained by the context of "exercise habits", "easy to overlook" and "good for your health".

https://doi.org/10.52202/079017-2820

we find nearly 50% of the candidate answers discovered by our auxiliary model, LLaMA-2-13B, are also beyond the knowledge boundary of GPT-4, including both factual answers that GPT-4 fails to produce and delusive wrong answers GPT-4 evaluates incorrectly.

Our contributions are as threefold: (1) We are the first to investigate the importance of semi-open-ended questions to the perception of knowledge boundaries for LLMs. (2) We propose an ambiguous answer discovery strategy that discovers many ambiguous answers with pieces of knowledge that are beyond the LLM's knowledge boundary. (3) Experimental results show the poor performance of an advanced LLM, GPT-4, on semi-open-ended questions and the effectiveness of our ambiguous answer discovery method in finding more pieces of knowledge which the LLMs are unfamiliar with.

2 Related Work

2.1 Perception of Knowledge Boundaries for LLMs

Existing studies on the perception of knowledge boundaries for LLMs can be categorized into prompt-based perception and pattern-based perception. Prompt-based perception perceives the knowledge boundary by querying the target LLM. Many researchers instruct the LLM before and after response generation, asking whether it can correctly answer the questions [32, 36, 8, 46] and if the generated answers are accurate [32, 43]. In addition, Yin et al. (2024) seek the optimal prompt for benchmarking LLM knowledge boundaries. Amayuelas et al. (2023) study the LLMs' ability to understand their knowledge and measure their uncertainty. Kadavath et al. (2022) also instruct LLMs to generate their confidence score for their responses. As studies find that LLMs tend to be overconfident [29, 32, 17, 49], many researchers explore representation-based perception. Researchers identify unknown questions [14] or evaluate correct and incorrect answers [10] by implicitly learning their different representations. Chen et al. (2023) train LLMs to identify incorrect answers via parameter-efficient tuning. Besides, Wang et al. (2023) extract representations of answerable and unanswerable questions to train a classifier to predict whether a question is answerable and assume questions with similar representations share the same answerability. Si et al. (2023) take token probability as the answer's confidence score during generation. Zhao et al. (2023) detect unanswerable questions by paraphrasing questions and checking the divergence of their answer distribution. The above studies primarily perceive knowledge boundaries for LLMs by distinguishing between answerable and unanswerable questions. This type of binary division does not apply to questions with both common easy answers and unpopular hard answers. Our study is the first to investigate the perception of knowledge boundaries on semi-open-ended questions.

2.2 Questions Answering for LLMs

Existing studies on Question Answering (QA) can be categorized into open-ended QA and close-ended QA based on the type of questions. Close-ended questions correspond to a limited number of correct answers, usually in the form of yes or no, true or false, or multiple-choice options, constraining the answers to a predetermined answer set [31, 15, 40, 33]. In addition, Researchers also study open-ended questions that allow the respondent to provide a more detailed and subjective response such as personal opinions and explanations [21, 40, 4, 3].

Researchers study the performance of LLMs on QA tasks mainly through various prompting strategies. Wei et al. (2023) explore "Chain-of-Thought" prompting (CoT), a simple and broadly applicable method for enhancing question answering ability of LLMs. Yao et al. (2023) and Besta et al. (2024) introduce similar frameworks for more complex QA tasks, namely, "Tree of Thought" and "Graph of Thoughts" prompting. As studies show that relying solely on an LLM's internal knowledge may lead to hallucinations [28], many researchers have also improved model performance in QA by incorporating external information (RAG systems [47] and knowledge graph [11, 1]). More recently, researchers have studied adaptive retrieval to avoid misinformation in the retrieved documents [41]. Ni et al. (2024) estimate the answerability of the given question and determines whether to retrieve [10, 29]. Xu et al. (2024) learn to identify the knowledge boundaries of LLMs and refuse to answer certain questions to avoid risks [14, 42].

Our code and data are available at https://github.com/araloak/LLM-knowledgeBoundary

https://doi.org/10.52202/079017-2820

3 Perception of Knowledge Boundary for LLM via Semi-open-ended QA

3.1 Overview

Our framework consists of three parts (see Fig. 1). We first exploit the instruction-following ability of a strong LLM to create a dataset consisting of semi-open-ended questions on various domains with LLM's answers. To discover more pieces of unfamiliar knowledge for the target LLM, we apply an open-sourced auxiliary model to incur more ambiguous answers by encouraging more distinctive generations. Finally, we evaluate whether the ambiguous answers to each question are beyond the knowledge boundary of the target LLM by comparing the self-evaluation results against RAG-based evaluation.

3.2 Semi-open-ended QA Dataset

3.2.1 Dataset Construction

To study the performance of semi-open-ended questions across various domains, we employ an LLM-based 2-step approach to obtain semi-open-ended questions and collect answers.

- **Domain selection.** We first prompt the LLM to generate a list of domains, encompassing world knowledge, which includes areas such as biology, geology, music, etc.
- **Question generation.** We prompt the LLM multiple times under each domain to generate a set of semi-open-ended questions Q. To ensure the quality of the generation quality, we provide human-written sample questions as demonstrations and specify the following requirements for the generation of candidate questions: 1. The question should correspond to multiple correct answers, making it challenging to answer. The question should also be relatively easy for non-expert users to understand. 2. The judgment of question answers should be based on objective standards in the real world instead of the subjective standards of the evaluator. 3. The truthfulness of an answer to a question should not change constantly over time. 4. The questions share the same template: *Tell me a list of .....* We use the same vanilla prompt to eliminate the influence of different question styles.
- **Answer collection.** For each question q in Q, we query the LLM I times and collect all responses $\mathsf{A} = \{a_0, a_1, ..., a_{I-1}\}$. In the i-th interaction, we inform the LLM of all previously generated responses $\mathsf{A}[:i]$ and obtain a_{i-1} by querying the LLM with question q', which repeats the same criteria specified in q and highlights the need for more answers.. Finally, we extract all answer entities in A and construct an answer list A.

3.2.2 Dataset Descriptions

We create a dataset to investigate the performance of mainstream LLM, GPT-4, on semi-open-ended questions. In dataset construction, we set I to 3 to exploit the knowledge of GPT-4 through multi-round conversations. Like humans, when faced with such questions (e.g., *What are the animals unique to Australia?*), LLMs tend first to give answers in which they hold high confidence (like the *red angaroo*). The latter answers are less certain and may have more mistakes (like *echidna*). We define the initial 75% of answer entities from GPT-4 generations as common-sense answers, while the remaining 25% as ambiguous answers. Our dataset comprises 953 questions covering 32 domains, with well-distributed data within each domain. On average, GPT-4 yields 52 answers for each question, including an average of 13 ambiguous answers. See the data samples in App. A. See the full prompts and demonstrations in App. C.

3.3 Ambiguous Knowledge Discovery

We apply an open-sourced auxiliary LLM to effectively discover more ambiguous answers that may be beyond the knowledge boundaries for black-box LLMs. Our intuition is that low-probability ambiguous answers reflect LLM's unfamiliarity with certain pieces of knowledge. However, it is challenging to collect low-probability ambiguous answers as 1) the generation probability of black-box LLMs (e.g. GPT-4) are inaccessible and their hyper-parameters (e.g. temperature) cannot target

For example, if q is: *Tell me a list of animals unique to Australia.*, then q' is *Tell me more animals unique to Australia.*

https://doi.org/10.52202/079017-2820

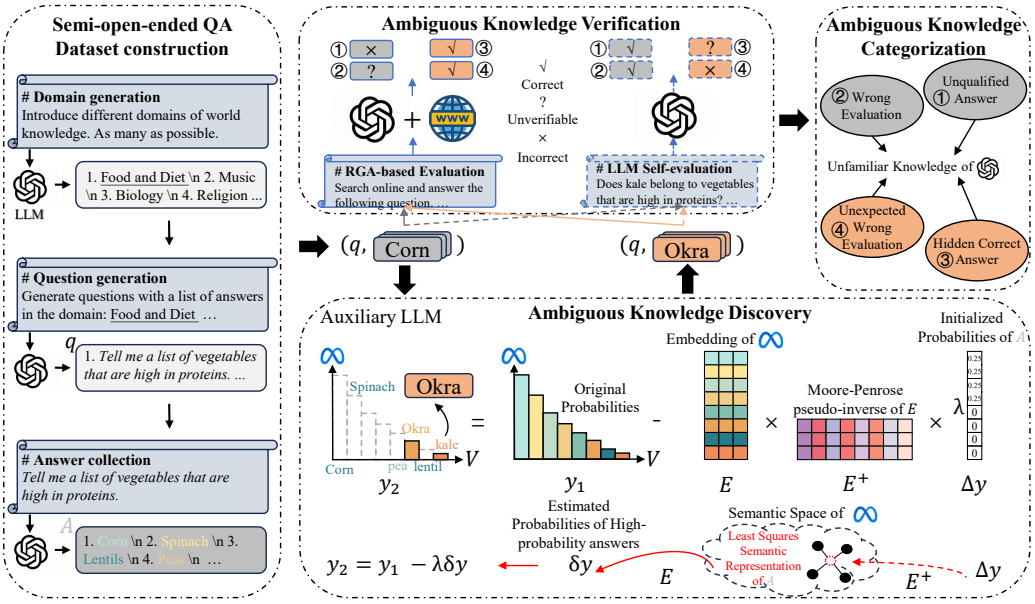

Figure 1: The overview of our framework. For the question q in the constructed dataset, The open-sourced auxiliary model prevent the high-probability answers based on existing answer entities A from the black-box LLM (i.e. GPT-4) and generate 4 categories of ambiguous knowledge that are unfamiliar knowledge for the target model.

ambiguous answer-related tokens (see Sec. 4.2). 2) violently prompting the LLM with question q many times to approximate the generation probability is inefficient as it prioritizes high-probability answers that may already be present in existing answers. Answers that are semantically similar to high-probability answers also tend to have a high probability during generation. Hence, we propose to prevent the generation of high-probability answers with their semantic-related counterparts on an open-sourced auxiliary LLM to incur more low-probability ambiguous answers for the perception of the knowledge boundary for the black-box LLM.

We denote that the open-sourced auxiliary model's final layer is $E \in \mathbb{R}^{|V| \times d}$, where d and V is the dimension size and V is the output vocabulary. In each generation step, LLM encodes the semantic representation of the input context as $x \in \mathbb{R}^{d \times 1}$ and uses it to calculate the next token generation probability as $y_1 = Ex$. Specifically, we take the following 3 steps to estimate and decrease the probability of high-probability answers.

1. Probability initialization for existing answers. We choose the first token in an answer entity $a \in A$ as "anchor token", which is suppose to represent the primary information about a since the first token is indispensable to its generation[9]. For a given question q, we define a vector $\Delta y \in \mathbb{R}^{V \times 1}$, indicating the existence of unique anchor tokens in all answer entities to the question. In Δy, we assign the value $\frac{1}{n}$ to the anchor tokens' position and assign 0 to other positions, where n is the number of all anchor tokens. As $\sum_i \Delta y_i = 1$, we deem Δy as the initialized probability distribution of all existing answers for the given question.
2. Semantic estimation of high-probability answers. We estimate the semantic representation δx for high-probability answer entities from initialized existing answer probability Δy. We calculate $\delta x = E^+ \Delta y$, where $E^+ E = I$. Here, E^+ is the left Moore-Penrose pseudo-inverse, and I is the identity matrix. Because the pseudo-inverse of a non-square matrix is often used to find the least squares solution of equations, In this way, we obtain δx as the least squares semantic representation that is the nearest to approximate the real semantic representation (which is unknown) of anchor tokens.
3. Probability Reduction. We calculate the probability of high-probability answers as $\delta y = E\delta x$ and ultimately obtain an adjusted generation probability y_2:

$$y_2 = y_1 - \lambda \delta y = y_1 - \lambda E \delta x = y_1 - \lambda E E_l^+ \Delta y,$$

https://doi.org/10.52202/079017-2820

where λ is a scaler that controls the extension of the reduction. Since anchor tokens and their semantic tokens almost share the same semantics, interpreting the estimated semantic representation of anchor tokens propagates the semantic information to their semantic-related ones. Thereby, we obtain an estimated probability distribution of high-probability tokens δy and use it to prevent the generation of high-probability answers.

During generation, the auxiliary LLM samples words on the adjusted probability distribution y_2 instead of the original distribution y_1. In this way, we only reduce the probability of answers that are semantically related to existing high-probability answers. while preserving the remaining probability distribution almost intact. The black-box model has already exhausted almost all high-probability common answers in A, forcing the auxiliary model to generate new ambiguous answers with low probabilities. These ambiguous answers are either low-probability correct answers neglected by the black-box model or delusive incorrect answers.

3.4 Ambiguous Knowledge Verification

We compare the results of LLM self-evaluation against the ground truth from RAG-based evaluation to verify the truthfulness of ambiguous answers, thereby identifying answers beyond the knowledge boundary for the target LLM. We conduct self-evaluation on the target LLM with well-designed instructions. We craft a prompt template including 1) an incentive statement that encourages better performance: *I'll pay you $100 for a factually correct answer* [7]; 2) an instruction *think step by step* which prompts the LLM to analyze before reaching to a conclusion [37, 24]. 3) multiple human-crafted examples as in-context learning demonstrations.

We believe an answer a in each test case $(q, \mathcal{A})$ to be factually correct to q if verified on public, trustworthy sources. As questions in our dataset correspond to numerous low-probability hard answers, it is cost-prohibitive to annotate the truthfulness of each answer with expert knowledge. Inspired by Web-GPT [28], we adopt a cost-efficient approach to mimic human behavior when faced with unfamiliar knowledge. We instruct an Internet-connected LLM with the same prompt for self-evaluation and require it to search online for related information before making judgments.

For both evaluations, we evaluate each answer a to its corresponding question as 1) *incorrect*, which contradicts reliable sources; 2) *correct*, which is supported by reliable sources, and 3) *unverifiable*, which is for cases that cannot be verified based on available information. Finally, for each answer a, we compare the differences between LLM self-evaluation and RAG-based evaluation to categorize different types of ambiguous knowledge.

3.5 Ambiguous Knowledge Categorization

We classify the ambiguous answers into four categories based on the above two evaluation results and posit that they are beyond the knowledge boundary of the target LLM. For a question, we categorize the following types of answers to be the LLM's ambiguous answers:

- Unqualified answers: answers from the target LLM's response that are identified as incorrect or unverifiable according to the ground truth.
- Inaccurate evaluations: answers from the target LLM's response whose self-evaluation results contradict their ground truth.
- Hidden correct answers: answers that are neglected by the target LLM, yet supplemented by the auxiliary model, which are correct according to ground truth.
- Unexpected wrong evaluations: answers that are neglected by the target LLM but generated by the auxiliary model whose self-evaluation results misalign with their ground truth.

The above categorization helps us understand different types of misunderstanding of the target LLM regarding specific pieces of unfamiliar knowledge.

Near-duplicate tokens roughly share the same semantic meaning but are different under tokenization due to typos, capitalization, or whitespace marking. For example, *Pea* and *peas* are two near-duplicate answers.

4 Experiments

4.1 Settings

In our experiment, we investigate the knowledge boundary of GPT-4 on our constructed dataset. We use two LLaMA-2-13b [35] models, as our auxiliary models in Sec. 3.3. Our method sets λ in Sec. 3.3 to 80. See more implementation details in App. B.

We compare our method with several baselines. Following prompt-based approaches [36],*Prompt* instructs the auxiliary model to generate more answers via prompt-engineering. Inspired by Zhang et al. (2023), *MASK* belongs to the representation-based perception that uses an average initialization for the probability of tokens from existing answers to represent the likelihood of high-probability answers and reduce their generation probabilities in the auxiliary model.

For the evaluation metrics, we use widely adopted Exact Match [32] (EM) and F1 scores [32, 36] to measure the performance of in discovering ambiguous answers. Different from previous research, our semi-open-ended questions correspond to a large number of correct answers, making it hard to build a comprehensive answer set for evaluation. Instead, we select ambiguous answers identified by the RAG-evaluation and confirmed by our human annotators as their ground truth and compare them with the full response to calculate the **EM** and **F1**. In this way, EM is an entity-level metric that measures the percentage of ambiguous answers within the response. F1 is a word-level metric that quantifies the word overlap between the ambiguous answers and the ground truth. We adopt Bleu [30] to measure word-level overlap between responses from the GPT-4 response and the auxiliary model. We also use answer overlap rate (AOR) to evaluate the efficiency of generating distinctive ambiguous answers. AOR is an entity-level metric that calculates the proportion of words in a list of answer entities that duplicate the reference response. See more evaluation details in App. B.

4.2 Overall Performance

Table 1: The performance of different auxiliary models with various strategies in discovering more ambiguous answers.

Method	Auxiliary Model	EM ↑	F1 ↑	AOR ↓	Bleu1↓	Bleu2↓	Bleu3↓	Bleu4↓
Prompt	LLaMA-2-13B	0.300	0.461	0.490	0.252	0.118	0.052	0.023
MASK	LLaMA-2-7B	0.458	0.570	0.342	0.185	0.075	0.021	0.004
	LLaMA-2-13B	0.470	0.587	0.344	0.189	0.075	0.021	0.004
Ours	LLaMA-2-13B	**0.481**	**0.587**	**0.326**	**0.181**	**0.071**	**0.018**	**0.004**

We analyze the effectiveness of our auxiliary model in exploring the knowledge boundary of GPT-4 by comparing it with multiple baselines. We randomly sample 200 questions from our dataset and discover ambiguous answers with an auxiliary model. Then, we verify the truthfulness of these answers using RAG-based evaluation with human annotation following Sec 3.4 and measure their performance with our evaluation metrics. Tab 1 shows the results of ambiguous answers on different evaluation metrics when using different auxiliary models and strategies. *Prompt* directly prompts the auxiliary model to generate answers, achieving the worst performance on all metrics. This suggests that directly prompting the auxiliary model may generate many repetitive answers (results in high AOR and Bleu scores) and, therefore inefficient in discovering new ambiguous answers (results in low EM and F1). *MASK* reduces the generation probabilities of anchor tokens during generation. When employing *MASK* on the same auxiliary model, its EM and F1 increase to 0.47 and 0.587 respectively while achieving a lower AOR (0.344). It indicates that reducing the probability of the generation of anchor words effectively achieves a more diverse generation. Replacing the auxiliary model with LLaMA-7B results in a slightly lower EM of 0.458. This marginal decrease implies that while a larger model can offer a broader knowledge base, reducing anchor word probabilities is more influential in generating distinctive answers. Our strategy estimates and reduces the generation probability of near-duplicate answers. This approach achieves the best performance on all metrics. It

There are many low-frequency answers for these questions. For example, there are maybe hundreds of accurate answers to the question:*What animals are native to Australia?*

https://doi.org/10.52202/079017-2820

underscores the effectiveness of our strategy in generating a diverse set of ambiguous answers that are less likely to duplicate existing ones, thus exploring the knowledge boundary for the target LLM. See our case study in App.G.

4.3 Ablation Study

Table 2: Ablation study on the key components in our method. We use the same metrics as in Sec. 4.2, apart from those that require manual annotation. AOR and Bleu measure the entity- and word-level overlap respectively between answers from GPT-4 and different model variants. − Auxiliary Model prompt GPT-4 with existing answers as examples for more ambiguous answers. − Inverse Matrix keeps the near-duplicate tokens during generation. Besides, we adjust the probability influence scaler (λ) to verify its impact on the generation results.

Variants	AOR↓	Bleu1↓	Bleu2↓	Bleu3↓	Bleu4↓
− *Auxiliary Model*	0.535	0.267	0.106	0.037	0.010
− *Inverse Matrix*	0.344	0.189	0.075	0.021	0.004
Ours(λ=60)	0.419	0.224	0.098	0.038	0.013
Ours(λ=70)	0.352	0.192	0.076	0.021	0.005
Ours	**0.326**	**0.181**	**0.071**	**0.018**	**0.004**

We conduct an ablation study on our proposed method to verify the importance of each component in eliciting more distinctive ambiguous answers (as shown in Tab 2). − *Auxiliary Model* abandons the auxiliary model and use existing answers as in-context learning examples to prompt GPT-4 for more answers. It achieves the highest on all metrics, indicating that prompting the black-box model violently for more answers is inefficient as it results in many repetitive answers. We also try to encourage a more diverse generation by increasing the generation temperature of GPT-4. However, we find that GPT-4 starts generating scrambled texts after just a few words and the perplexity of these texts exceeds 1000, while normally GPT-4 generates a low perplexity of around 10. This indicates that increasing the sampling temperature results in the generation of scrambled texts. Although adjusting the generation temperature can change the generation probabilities, it does not alter the original probability relationships, nor can it specifically target tokens related to ambiguous answers. − *Inverse Matrix* only reduces the probability of existing answers without considering their near-duplicate answers. It performs better than − *Auxiliary Model* while underperforms Ours on all metrics. It shows that estimating and reducing the probability of near-duplicate tokens augment the auxiliary model for higher generation diversity. We increase the intervention on the generation probability by lowering λ, the probability influence scaler. From row 3 to row 5 in Tab. 2, λ increases from 60 to 80, resulting in a decrease on all metrics. This suggests that the extent to which we intervene in the generation probability is positively correlated with the diversity of ambiguous answers produced by the auxiliary model.

4.4 Results of Perceiving the Knowledge Boundary for GPT-4

Table 3: Percentages of different categories of answers comparing the LLM self-evaluation and ground truth labels. Following the categorization in Sec. 3.5, we calculate that the percentage of unqualified answers is 40.15% by adding up the underlined results that are incorrect or unverifiable according to the ground truth. We also obtain the percentage of inaccurate evaluations as 28.47%, by adding up the results highlighted in red where self-evaluation is inconsistent with the ground truth.

Ground Truth\ Self-evaluation	Incorrect	Correct	Unverifiable
Incorrect	20.98	8.37	2.25
Correct	9.30	47.77	2.78
Unverifiable	3.30	3.73	1.52

By analyzing the ambiguous answers in our dataset, we arrive at the following findings: (1) **GPT-4 performs poorly on the semi-open-ended questions and generates many unqualified answers.**

We calculate the percentage of questions where at least one of the GPT-4's answers is unverifiable or incorrect according to the ground truth. We find that GPT-4 generates incorrect or unverifiable answers in 82.90% of questions. By adding up the proportion of incorrect answers (row 1 in Tab.3) and unverifiable answers (row 3 in Tab.3), we identify that 40.15% of ambiguous answers belong to unqualified answers. (2) **GPT-4 makes many inaccurate evaluations regarding the truthfulness of ambiguous answers, indicating that the LLM lacks understanding of the relevant knowledge.** We identify answers whose ground truth misaligns with self-evaluations in Tab.3 and find that 28.47% (by adding up the results in red color from Tab.3) of ambiguous answers belong to inaccurate evaluations for GPT-4. It indicates that GPT-4 is unfamiliar with these pieces of knowledge, and retrieval is helpful for LLM to draw the correct conclusion. (3) **GPT-4 has limited ability to recognize its knowledge boundary, while in most cases it continues to produce unqualified answers.** We search for keywords in all responses that reflect GPT-4's admission of its knowledge boundary (e.g. *I apologize* and *I'm afraid*) and calculate the proportion of corresponding questions in the dataset. We find that in about 7% of questions, GPT-4 admits that it has listed all the answers and refuses to provide more answers (it generates a response like *I apologize for any confusion, but to the best of my knowledge, the list I provided includes all the correct answers.*). However, it fails to recognize its knowledge boundary in the rest questions and continues to generate unqualified answers. Our findings indicate that advanced LLM (i.e. GPT-4) is easy to hallucinate on semi-open-ended questions, indicating the importance of detecting the LLM knowledge boundary via these questions.

4.5 Results of the Auxiliary Model on Perceiving the Knowledge Boundary for GPT4

Table 4: Percentages of different categories of ambiguous answers comparing the GPT-4 self-evaluation results and their ground truth. Following the categorization in Sec.3.5, we calculate that the percentage of hidden correct answers is 75.12% by adding up the starred (*) results that are correct according to the ground truth. We also obtain the percentage of unexpected wrong evaluations as 62.43%, by adding up the results highlighted in orange where GPT-4-evaluation is inconsistent with the ground truth.

Ground Truth\ GPT-4-evaluation	Incorrect	Correct	Unverifiable
Incorrect	0.04	9.53	11.31
Correct	23.97*	37.54*	13.61*
Unverifiable	3.18	0.83	0.00

Tab.4 shows the fine-grained results of GPT-4 self-evaluation and the ground truth of ambiguous answers discovered by our auxiliary model. By analyzing the results, we conclude some interesting findings: (1) **LLaMA-2-13B effectively supplements GPT-4 by identifying hidden correct answers.** We add up the starred results in Tab.4 and obtain the proportion of LLM neglected hidden correct answers (75.12%). Notably, 23.97% and 13.61% of correct ambiguous answers are both neglected by GPT-4 and deemed to be incorrect or unverifiable under GPT-4 self-evaluation, showcasing the GPT-4's unfamiliarity with the corresponding knowledge. (2) **LLaMA-2-13B is easy to incur unexpected wrong evaluations during GPT-4 self-evaluation.** We add up the results highlighted in orange from Tab.4 and find that 62.43% of the GPT-4 self-evaluations are inconsistent with the ground truth. It implies that GPT-4's self-evaluation mechanism may not be fully aligned with the actual correctness of the ambiguous answers, especially when they are supplemented by the auxiliary model. (3) **LLaMA-2-13B is also able to discover situations where GPT-4 admits for its knowledge boundary.** We add up the results in column 3 where GPT-4 admits that it cannot make a judgment (unanswerable) during self-evaluation and obtain 24.92% of aligned ambiguous answers. It means that GPT-4 aligns well with these ambiguous answers because it neglected these answers during generation and deems them as unknown knowledge during self-evaluation.

4.6 Practical Implications

Perceiving LLMs' knowledge boundaries is important to understand and alleviate hallucination [18, 49]. Ambiguous answers for semi-open-ended questions are highly likely beyond the knowledge boundaries of LLMs (see Sec 4.4). Discovering ambiguous answers benefits many applications,

https://doi.org/10.52202/079017-2820

including: 1) It helps detect the knowledge scope of LLMs more faithfully. While many close-ended hallucination evaluation benchmarks face the danger of data contamination [25, 13], semi-open-ended questions are easy to design and correspond to a large number of undocumented answers; 2) Flagging ambiguous answers with higher uncertainty enhances the LLM outputs [22, 19]; 3) Identifying ambiguous answers helps achieve selective retrieval that augments LLM with indispensable external information while reducing the distraction of irrelevant data [36, 26, 39]; 4) It helps align LLMs for a more honest generation by teaching the LLM to admit its knowledge limit on the knowledge it is unfamiliar with (ambiguous answers) [43, 8, 42].

5 Conclusion

We investigate the perception of knowledge boundary for LLMs with semi-open-ended questions, an important yet underexplored type of question corresponding to a large number of accurate answers. We introduce an LLM-based approach to construct semi-open-ended questions and collect LLM answers from the target LLM. Then, we discover more pieces of unfamiliar knowledge for the target LLM by eliciting ambiguous answers from an auxiliary model that the LLM neglects. To achieve a more effective generation, we estimate and reduce the generation probability of existing answers with their near-duplicate counterparts. With our methods, we construct a dataset to evaluate the performance of GPT-4 and discover many ambiguous answers with our auxiliary model, LLaMA-2-13B. Our findings reveal that GPT-4 produces many unqualified answers and suffers from inaccurate evaluations. Besides, we verify that LLaMA-2-13B is effective in discovering more unpopular correct answers and delusive wrong answers neglected by GPT-4. Our findings underscore the importance of semi-ended questions and the effectiveness of our method in assisting in perceiving knowledge boundaries for LLMs.

6 Acknowledgement

This work is supported by the following findings: Young Elite Scientist Sponsorship Program by CAST (2023QNRC001) under Grant No. YESS20230367, the National Natural Science Foundation of China under Grant No. 62306330, No. 62106275, No. 62025208, No. 62421002, and the Grant of No. WDZC20235250103.

References

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Can knowledge graphs reduce hallucinations in llms? : A survey, 2024.

[2] Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*, 2023.

[3] Yejin Bang, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Samuel Cahyawijaya, Dan Su, Bryan Wilie, Romain Barraud, Elham J. Barezi, Andrea Madotto, Hayden Kee, and Pascale Fung. Towards answering open-ended ethical quandary questions, 2023.

[4] Man Luo;Shailaja Keyur Sampat;Riley Tallman;Yankai Zeng;Manuha Vancha;Akarshan Sajja;Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in the development and evaluation of open-ended visual question answering (vqa) tasks. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[5] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar 2024. doi: 10.1609/aaai.v38i16.29720. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29720`.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[7] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4, 2024.

[8] Lang Cao. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. *arXiv preprint arXiv:2311.01041*, 2023.

[9] Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci, and Andrew McCallum. Multi-CLS BERT: An efficient alternative to traditional ensembling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–854, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.48. URL https://aclanthology.org/2023.acl-long.48.

[10] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.345. URL https://aclanthology.org/2023.findings-emnlp.345.

[11] Zhuo Chen;Jiaoyan Chen;Yuxia Geng;Jeff Z. Pan;Zonggang Yuan Huajun Chen. Zero-shot visual question answering using knowledge graph. In *The Semantic Web – ISWC 2021*, 2021.

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[13] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *ACL*, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482.

[14] Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. Gotcha! don't trick me with unanswerable questions! self-aligning large language models for responding to unknown questions. *arXiv preprint arXiv:2402.15062*, 2024.

[15] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107.

[16] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9OevMUdods.

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232, 2023. URL https://api.semanticscholar.org/CorpusID:265067168.

[18] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[19] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *TACL*, 9, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

[20] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-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, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

[21] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

[22] Kimiya Keyvan and Jimmy Xiangji Huang. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Comput. Surv.*, 55(6), December 2022. ISSN 0360-0300. doi: 10.1145/3534965. URL https://doi.org/10.1145/3534965.

[23] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

[24] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[25] Yucheng Li, Frank Guerin, and Chenghua Lin. An open source data contamination report for large language models, 2024. URL https://arxiv.org/abs/2310.17589.

[26] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

[27] Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. Injecting new knowledge into large language models via supervised fine-tuning, 2024.

[28] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022.

[29] Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. *arXiv preprint arXiv:2402.11457*, 2024.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[31] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487.

[32] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*, 2023.

[33] Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering, 2023.

[34] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=98p5x51L5af.

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[36] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.691. URL https://aclanthology.org/2023.findings-emnlp.691.

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[39] Di Wu, Wasi Uddin Ahmad, Dejiao Zhang, Murali Krishna Ramanathan, and Xiaofei Ma. Repoformer: Selective retrieval for repository-level code completion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pages 53270–53290. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24a.html.

[40] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021.

[41] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, 2024.

[42] Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback, 2024.

[43] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

[44] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models, 2023.

[45] Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. Benchmarking knowledge boundary for large language model: A different perspective on model evaluation. *arXiv preprint arXiv:2402.11493*, 2024.

[46] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551.

[47] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL https://aclanthology.org/2023.findings-emnlp.307.

[48] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*, 2023.

[49] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what llms do not know: A simple yet effective self-detection method. *arXiv preprint arXiv:2310.17918*, 2023.

A Dataset Information

Table 5: Example of data samples in our dataset. Each question corresponds to many ambiguous answers. Different colors of the answers represent different ground truth truthfulness labels: yellowgreen represents that the answer is verified as being factually correct, red is verified as being incorrect, and salmon is for unverifiable answers.

Domain	Percentage	Question	Ambiguous Answer
Biology	4.93%	Tell me a list of trees that produce fruit with hard shells and are native to tropical regions.	Malabar Chestnut, Miracle Fruit Tree, Guapinol Tree, Peanut Tree, Sapodilla Tree, Desert Date, Castor Oil Plant, Chiclé Tree, Oriental Persimmon, Dika Tree, Langsat Tree, Pequi Tree, Karite Tree, Imbu Tree, Caraipa Tree
Music	4.20%	Tell me a list of films whose plot is centered on a historically significant event, but the narrative is from the perspective of an imagined protagonist.	The Hurt Locker, Zulu, Letters from Iwo Jima, Doctor Zhivago, Black Hawk Down, Les Miserables, Lawrence of Arabia, The Wind that Shakes the Barley, The Longest Day, Anthony Adverse, The Guns of Navarone, Lion of the Desert
Geology	4.09%	Tell me a list of current deserts across the world that were previously covered by an ancient sea.	Qaidam Basin, Dasht-e Kavir, Dasht-e Lut, The Registan Desert, The Franklin Basin, The Makran Desert, The Sonoran Desert, The Jornada Del Muerto, Great Karoo, Tanami Desert
Food and Diet	5.67%	Tell me a list of fruits that are sources of healthy fats, not including avocados and coconuts.	Sacha Inchi Nuts, Hemp Seeds, Chokeberries, Elderberries, Pine Nuts, Passion Fruit Seeds, Tibetan Goji Berries
Literature	1.68%	Tell me a list of female writers from the Victorian era whose work focuses on social reform.	Helen Taylor, Catherine Helen Spence, Octavia Hill, Rhoda Broughton, Elizabeth Missing Sewell, Emmeline Pankhurst, Alice Meynell, Elizabeth Robins, Mary Augusta Ward, Constance Garnett

Our dataset covers the following topics: Environment and Climate, Technology and Industry, Political Science, History and Archaeology, Sociology, Economy and Finance, Philosophy, Languages, Art, Architecture, Music, Physics, Astronomy, Chemistry, Biology, Geology, Computer Science, Anthropology and Cultures, Education, Psychology and Mental Health, Fitness and Physical Health, Literature, Religion, Law and Criminology, Military and War Agriculture, Tourism, Film and Television, Sports and Athletics, Food and Diet, Energy and renewable resources, Mathematics and Statistics, Medicine and Health, Games, Clothing and Fashion.

B Implimentation Details

In our experiments, we query GPT-4-Turbo through API calls. For our RAG-based evaluation in Sec. 3.4, we first employ Microsoft Copilot to search the Internet to find evidence and draw a conclusion. For our evaluation to calculate EM and F1, we hire 11 human annotators with Master's degrees to manually review the responses of Copilot, ensure the credibility of the references, and determine the truthfulness of each answer.

During the generation of our auxiliary model, We apply nucleus sampling (with p=0.9) during generation, setting the generation temperature to 0.7, and the repetition penalty to 1.15.

For our metrics, EM and F1 are common in QA tasks and knowledge boundary detection tasks. The Bleu metric is often used to measure the n-gram overlap between the model response and the ground truth. A lower Bleu score means a more dissimilar response generated by the auxiliary model. Before calculating the above metrics, we normalize answer entities with the NLTK library by lowercasing the answers and turning them into the singular form.

C Important Instructions

Here we provide important instructions in building our dataset, guiding LLM for self-evaluation, and show guidelines for human annotators in the supplementary materials.

- Domain Generation. *I hope to test my students' knowledge in different domains. Which domains can I use to create questions?*

https://doi.org/10.52202/079017-2820

- Question Generation. *I am a professor of [CATEGORY] and need to test students' understanding of [CATEGORY] by asking a series of challenging questions. These questions require respondents to list entities that they know meet a series of certain conditions. You need to create more different and diverse challenging questions according to the requirements. Read the following requirements carefully. I'm going to tip $100 for a perfect list of questions! \n The questions should meet the following criteria: \n 1. Each question should start with "Tell me a list of"; \n 2. To make the question challenging enough, each question should contain multiple limiting conditions. \n 3. The requirement of the question should not involve specific numbers (which makes the question too hard to answer) or vague descriptions (which makes it hard to evaluate the truthfulness of the answer), like "long lifespan", "quick speed", "popular", and "important"; \n 4. The boundaries of the question should be very clear, making it easy to evaluate its truthfulness; \n 5. The answers to the questions should be consistent through a relatively long time and not change frequently, for example, yearly. \n Refer to the style in the following two examples from an exemplary subject, biology. \n Question 1: Tell me a list of land animals unique to Australia. \n Question 2: Tell me a list of fruits that grow on trees in tropical regions.*
- Self-evaluation. *Does [AMBIGUOUS ANSWER] belong to [QUESTION REQUIREMTNTS]? I'll tip $100 for the factually correct answer. Think step by step and then give your answer.*
- RAG-based evaluation. *Search online for highly credible information related to the following question, and answer the question based on the search results. \n Does [AMBIGUOUS ANSWER] belong to [QUESTION REQUIREMTNTS]? I'll tip $100 for the factually correct answer. Think step by step and then give your answer.*

D Future Work

Our approach of modifying the LLM representations to guide answer generation may provide insight for different kinds of normal QA tasks: It may help alleviate the hallucinations in knowledge-extensive QA tasks via representation engineering. Editing LLM representations considering existing answers can reduce the probability of semantically related words, helping to generate more diverse answers for open-ended QA tasks. In our future work, we plan to investigate representation engineering for more diverse and honest responses for different QA tasks. Additionally, we plan to evaluate the performance of more cutting-edge LLMs on semi-open-ended questions and integrate user feedback mechanisms to guide LLMs in recognizing their limitations.

E Limitations

We only investigate the performance of one black-box LLM, GPT-4 on our dataset due to expensive human annotation. As a result, we have not engaged in a comparative analysis with other black-box models, such as Falcon-180B, which could have provided a broader perspective on the performance metrics. Our dataset also does not contain all unpopular answers to semi-open-ended questions because there may be hundreds and thousands of potential answers that require expert knowledge. Given the complexities and the extensive scope involved, it was deemed unfeasible to incorporate this aspect within the purview of our current study.

F Social Impact

We discuss the social impact of our research on the knowledge boundaries of LLMs as follows. On one hand, our work could enhance the reliability of AI systems by identifying their knowledge limits, thereby improving user trust and the accuracy of machine-generated information. It also contributes to reliable AI development by emphasizing the importance of factual information in AI outputs, which can be particularly beneficial in educational and professional settings. On the other hand, there is a risk that the ambiguous answers generated by LLMs could lead to misinformation if not properly resolved. Additionally, an overreliance on LLM for knowledge-seeking could undermine human critical thinking.

G Case Study

Table 6: Examples of two ambiguous answers with their related questions. The Raspberry is an unqualified answer generated by GPT-4, which yields a different answer in another question. Plantain is an answer that is neglected by GPT-4, yet supplemented by the auxiliary model, which is correct according to ground truth. However, GPT-4 believes it to be wrong and generates wrong information for another question. Texts in yellowgreen are truthful information, while texts in red are non-factual.

Semi-open-ended Question	Tell me a list of foods that are rich in Vitamin A but low in fat.	
GPT-4 Response for Semi-open-ended Question	1. Carrots \n 2. Spinach ... 46. Raspberries \n 47. Red Leaf Lettuce ...	
Auxiliary Model Response	1. Bell peppers \n 2. Liver ... 14. Meat such as beef liver \n 15. Plantains ...	
Ambiguous Answer	Raspberries	Plantains
Answer Type	Unqualified Answer	Hidden Correct Answer
Related Question	What vitamins are rich in raspberries?	Is it a good choice to eat Plantains for many Vitamin A?
GPT-4 Response	Raspberries are rich in vitamins C, K, E, and B-complex. They also contain small amounts of vitamin A.	Plantains do contain vitamin A, but not in very high amounts. ...

We showcase the importance of ambiguous answers in perceiving the knowledge boundary for LLMs. Tab. 10 shows an example with two ambiguous answers for the same question. First, we sample a semi-open-ended question with its GPT-4 responses and ambiguous answers augmented by the auxiliary model (row 2 and 3 in table Tab.10). Then, we manually construct two related questions, each involving different types of ambiguous answers, and request new responses from GPT-4. For the case displayed on the left side of Tab.10, GPT-4 falsely deems the raspberry as food that is rich in vitamin A but low in fat, yet it answers correctly in another related question. It indicates that GPT-4 is inconsistent in answering different questions involving the same ambiguous answer. For the case on the right side of Tab.10, given the same semi-open-ended question, the auxiliary model discovers another ambiguous answer, Plantain, which is correct to the question. Interestingly, GPT-4 generates misinformation regarding this answer entity. It shows that GPT-4 falsely believes plantain is incorrect. It also indicates that the auxiliary model helps discover ambiguous answers that elicit misinformation in the target LLM. It strengthens the importance of perceiving knowledge boundaries for LLMs by discovering ambiguous answers to semi-open-ended questions. See App. H for more examples.

H Cases of Different Types of Ambiguous Answers with Misinformation on the Related Questions

We show different types of ambiguous answers and GPT-4's performance on their related questions from Tab.7 to Tab.8. See full cases in Tab.9, Tab.10 and Tab.11.

Table 7: Examples of two ambiguous answers with their related questions. The Malabo, Equatorial Guinea is an answer that is neglected by GPT-4, yet supplemented by the auxiliary model, which is correct according to ground truth. However, GPT-4 believes it to be wrong and generates wrong information for another question. Rome, Italy is an answer that is neglected by GPT-4 but generated by the auxiliary model, whose self-evaluation results misalign with the ground truth. However, GPT-4 believes it to be wrong and generates wrong information for another question. Texts in yellowgreen are truthful information, while texts in red are non-factual.

Semi-open-ended Question	Tell me a list of world capitals where more than two languages are considered official.	
GPT-4 Response for Semi-open-ended Question	1. London, United Kingdom \n 2. Montreal, Canada ... 52. Zagreb, Croatia \n 53. Prague, Czech Republic ...	
Auxiliary Model Response	1. Brussels, Belgium \n 2. Ottawa, Canada ... 9. Rome, Italy \n 10. Malabo, Equatorial Guinea ...	
Ambiguous Answer	Malabo, Equatorial Guinea	Rome, Italy
Answer Type	Hidden Correct Answer	Unexpected Wrong Evaluations
Related Question	How many official languages are there in Malabo?	Is Latin the official language of Rome?
GPT-4 Response	There are two official languages in Malabo: Spanish and Pidgin English.	No, the official language of Rome is Italian. ...

https://doi.org/10.52202/079017-2820

Table 8: Examples of two ambiguous answers with their related questions. The Sand Island Light is an unqualified answer generated by GPT-4, which yields a different answer to another question. The Portsmouth Harbor Light is an answer from GPT-4, whose self-evaluation results contradict the ground truth. However, GPT-4 believes it to be true and generates wrong information for another question. Texts in yellowgreen are truthful information, while texts in red are non-factual.

Semi-open-ended Question	Tell me a list of famous lighthouses located on islands in the Atlantic Ocean.	
GPT-4 Response for Semi-open-ended Question	1. Pemaquid Point Light \n 2. Portland Head Light ... 46. Sand Island Light \n 47. Portsmouth Harbor Light \n ...	
Auxiliary Model Response	1. Maine Coast Guard Station and Pier Light \n 2. Minot's Leading Light ... 10. Woodenboat School Light \n 11. Chappaquiddick Light ...	
Ambiguous Answer	Portsmouth Harbor Light	Sand Island Light
Answer Type	Inaccurate Evaluations	Unqualified Answer
Related Question	Is Portsmouth Harbor Light located on which island in the Atlantic Ocean?	Where is Sand Island Light located?
GPT-4 Response	Portsmouth Harbor Light is located on New Castle Island in the Atlantic Ocean, in the state of New Hampshire, USA.	Sand Island Light is located in Alabama, USA, specifically on Sand Island off the coast of Dauphin Island. ...

Table 9: The full version of the cases in Tab.6

Semi-open-ended Question	Tell me a list of foods that are rich in Vitamin A but low in fat.
GPT-4 Response for Semi-open-ended Question	1. Carrots \n 2. Spinach \n 3. Sweet Potato \n 4. Kale \n 5. Butternut Squash \n 6. Red Bell Peppers \n 7. Apricots \n 8. Broccoli \n 9. Romaine Lettuce \n 10. Cantaloupe \n 11. Yellow Corn \n 12. Tangerines \n 13. Tomatoes \n 14. Papaya \n 15. Peaches \n 16. Mangoes \n 17. Watermelon \n 18. Green Peas \n 19. Kiwi \n 20. Honeydew Melon \n 21. Asparagus \n 22. Black-Eyed Peas \n 23. Plums \n 24. Pineapple \n 25. Blueberries \n 26. Grapefruit \n 27. Oranges \n 28. Green beans \n 29. Brussel sprouts. \n 30. Guava \n 31. Dandelion Greens \n 32. Swiss Chard \n 33. Mustard Greens \n 34. Arctic Char \n 35. Prunes \n 36. Nectarines \n 37. Rose Hips \n 38. Parsley \n 39. Watercress \n 40. Radish \n 41. Red Cabbage \n 42. Beet Greens \n 43. Endives \n 44. Arugula \n 45. Pomegranates \n 46. Raspberries \n 47. Red Leaf Lettuce \n 48. Pumpkin \n 49. Bok Choy \n 50. Turnip Greens \n 51. Red Grapes \n 52. Passionfruit \n 53. Collard Greens \n 54. Zucchini \n 55. Strawberries \n 56. Clementines \n 57. Lima Beans \n 58. Cucumbers \n 59. Yellow Bel Peppers \n 60. Pawpaw \n 61. Winter squash \n 62. Cantaloupes \n 63. Tuna \n 64. Persimmon \n 65. Pink Grapefruit \n 66. Honeydew melon \n 67. Guavas \n 68. Papayas \n 69. Passion fruit \n 70. Kiwi fruit \n 71. Blackberries \n 72. Cherries \n 73. Goji berries \n 74. Raspberry \n 75. Fenugreek leaves \n 76. Okra \n 77. Chayote \n 78. Eggplant
Auxiliary Model Response	1. Bell peppers \n 2. Liver \n 3. Collards \n 4. Carrots \n 5. Spinach \n 6. Yogurt \n 7. Salmon \n 8. Whole grains \n 9. Fortified cereals \n 10. Clams \n 11. Milk \n 12. Shrimp \n 13. Whole eggs \n 14. Meat such as beef liver \n 15. Plantains \n 16. Fresh herbs like basil, parsley and dill \n 17. Cheese (low-fat varieties) \n 18. Vegetable oils such as soybean oil or corn oil

Table 10: The full version of the cases in Tab.7

Semi-open-ended Question	Tell me a list of world capitals where more than two languages are considered official.
GPT-4 Response for Semi-open-ended Question	1. London, United Kingdom \n 2. Montreal, Canada \n 3. Brussels, Belgium \n 4. Ottawa, Canada \n 5. Bern, Switzerland \n 6. Johannesburg, South Africa \n 7. Suva, Fiji \n 8. Biel/Bienne, Switzerland \n 9. Singapore, Singapore \n 10. Luxembourg City, Luxembourg \n 11. New Delhi, India \n 12. Addis Ababa, Ethiopia \n 13. Nairobi, Kenya \n 14. Helsinki, Finland \n 15. Vienna, Austria \n 16. Maputo, Mozambique \n 17. Oslo, Norway \n 18. Reykjavik, Iceland \n 19. Valletta, Malta \n 20. Nicosia, Cyprus \n 21. Dublin, Ireland \n 22. Vaduz, Liechtenstein \n 23. Tirana, Albania \n 24. Andorra la Vella, Andorra \n 25. Baku, Azerbaijan \n 26. Manama, Bahrain \n 27. Dhaka, Bangladesh \n 28. Saraevo, Bosnia and Herzegovina \n 29. Sofia, Bulgaria \n 30. Copenhagen, Denmark \n 31. Tallinn, Estonia \n 32. Tbilisi, Georgia \n 33. Baghdad, Iraq \n 34. Jerusalem, Israel \n 35. Astana, Kazakhstan \n 36. Beirut, Lebanon \n 37. Tripoli, Libya \n 38. Kuala Lumpur, Malaysia \n 39. Abuja, Nigeria \n 40. Islamabad, Pakistan \n 41. Doha, Qatar \n 42. Bucharest, Romania \n 43. Riyadh, Saudi Arabia \n 44. Madrid, Spain \n 45. Lome, Togo \n 46. Ankara, Turkey \n 47. Kiev, Ukraine \n 48. Dubai, United Arab Emirates \n 49. Canberra, Australia \n 50. Praia, Cape Verde \n 51. N'Djamena, Chad \n 52. Zagreb, Croatia \n 53. Prague, Czech Republic \n 54. Beijing, China \n 55. San Jose, Costa Rica \n 56. Santo Domingo, Dominican Republic \n 57. Quito, Ecuador \n 58. Cairo, Egypt \n 59. Port Vila, Vanuatu \n 60. Windhoek, Namibia \n 61. Kathmandu, Nepal \n 62. Wellington, New Zealand \n 63. Asuncion, Paraguay \n 64. Lisbon, Portugal \n 65. Moscow, Russia \n 66. Dakar, Senegal \n 67. Ljubljana, Slovenia
Auxiliary Model Response	1. Brussels, Belgium \n 2. Ottawa, Canada \n 3. Panama City, Panama \n 4. Lima, Peru \n 5. Toronto, Canada \n 6. Paris, France \n 7. Hong Kong, China \n 8. Moskva, Russia \n 9. Rome, Italy \n 10. Malabo, Equatorial Guinea \n 11. Stockholm, Sweden \n 12. Athens, Greece \n 13. Tokyo, Japan \n 14. Jakarta, Indonesia \n 15. Mexico City, Mexico \n 16. Bangkok, Thailand \n 17. São Paulo, Brazil \n 18. Buenos Aires, Argentina \n 19. Caracas, Venezuela \n 20. Bogota, Colombia

Table 11: The full version of the cases in Tab.8

Semi-open-ended Question	Tell me a list of famous lighthouses located on islands in the Atlantic Ocean.
GPT-4 Response for Semi-open-ended Question	1. Pemaquid Point Light \n 2. Portland Head Light \n 3. Cape Hatteras Lighthouse \n 4. Tybee Island Light \n 5. Boston Light \n 6. St. George Reef Light \n 7. Lindau Lighthouse \n 8. Ouessant Lighthouse \n 9. St. Mary's Lighthouse \n 10. Belle Tout Lighthouse \n 11. Lighthouse of Genoa \n 12. Tower of Hercules \n 13. Lighthouse of Alexandria \n 14. Capelinhos Lighthouse \n 15. Start Point Lighthouse \n 16. Cape Spear's Lighthouse, Newfoundland \n 17. Sambro Island Lighthouse \n 18. Cape Borda Lighthouse \n 19. Amédée Lighthouse \n 20. Roches-Douvres Lighthouse \n 21. Bishop Rock Lighthouse \n 22. Skerryvore Lighthouse \n 23. Fastnet Rock Lighthouse \n 24. Thacher Island Twin Lights \n 25. Cape Lookout Lighthouse \n 26. Point Judith Lighthouse \n 27. Bass Harbor Head Lighthouse \n 28. Sandy Hook Lighthouse \n 29. Montauk Point Lighthouse \n 30. Cape May Lighthouse \n 31. Lundy Island Lighthouse \n 32. Lighthouse at Ile Vierge \n 33. Maughold Head Lighthouse \n 34. Ballycotton Island Lighthouse \n 35. Tory Island Lighthouse \n 36. Cape Wrath Lighthouse \n 37. Eilean Glas Lighthouse \n 38. Ardnamurchan Lighthouse \n 39. Longstone Lighthouse \n 40. Happisburgh Lighthouse \n 41. Kinnaird Head Lighthouse \n 42. The Needles Lighthouse \n 43. Corbiere Lighthouse \n 44. Casquets Lighthouse \n 45. The Round Island Lighthouse \n 46. Sand Island Light \n 47. Portsmouth Harbor Light \n 48. Neist Point Lighthouse \n 49. Flannan Isles Lighthouse \n 50. Cromarty Lighthouse \n 51. Fair Isle South Lighthouse \n 52. Copinsay Lighthouse \n 53. North Ronaldsay Lighthouse \n 54. Noup Head Lighthouse \n 55. Sule Skerry Lighthouse \n 56. Skokholm Lighthouse \n 57. Muckle Flugga Lighthouse \n 58. Brier Island Lighthouse \n 59. Articled Rock Lighthouse \n 60. Nubble Lighthouse \n 61. Cape Recife Lighthouse \n 62. Terrible Tilly \n 63. Butt of Lewis Lighthouse \n 64. Seaham Lighthouse \n 65. Girdle Ness Lighthouse \n 66. Walney Lighthouse \n 67. Europa Point Lighthouse \n 68. Lighthouse of the Whales \n 69. Hirtshals Lighthouse \n 70. Cordouan Lighthouse \n 71. Sankaty Head Light \n 72. Gay Head Light \n 73. Cape Romain Lighthouses \n 74. Ram Island Ledge Light \n 75. Boca Grande Lighthouse \n 76. Petit Manan Light \n 77. Cape Elizabeth Light \n 78. Cape Ann Light Station \n 79. Rose Island Lighthouse \n 80. Saint Augustine Light \n 81. Egmont Key Lighthouse \n 82. Dry Tortugas Light \n 83. Nobska Point Light \n 84. Sullivan's Island Lighthouse \n 85. Hatteras Island Lighthouse \n 86. Hunting Island Lighthouse \n 87. Långe Erik \n 88. Långe Jan \n 89. Ushant Island Lighthouse \n 90. Blackhead Lighthouse \n 91. Hook Lighthouse \n 92. Cape Clear Island Lighthouse \n 93. Rathlin West Lighthouse \n 94. Copeland Lighthouse \n 95. Fisgard Lighthouse \n 96. West Quoddy Head Light \n 97. Cape Jourimain Lighthouse
Auxiliary Model Response	1. Maine Coast Guard Station and Pier Light \n 2. Minot's Leading Light \n 3. Monhegan Island Light \n 4. Eastern Breakwater Light \n 5. Little River Light \n 6. Baldhead Island Light \n 7. Oak Bluffs Harbor Light \n 8. Brenton Reef Light\n 9. Block Island Southwest Light \n 10. Woodenboat School Light \n 11. Chappaquiddick Light \n 12. Race Rock Light \n 13. Plum Gut Light \n 14. Great Round Shoal Light \n 15. New London Island Lighthouse \n 16. Smith's Island Light \n 17. Highfield Hall Beach Light \n 18. Scituate Light House

https://doi.org/10.52202/079017-2820

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The last paragraph of our Introduction accurately reflects the paper's contributions and scope. See Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of our work in the appendix. See App. E

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete (and correct) proof in Perception of Knowledge Boundary for LLM. See Sec. 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose the implementation details, Dataset Information, Important Instructions and the results of experiment in Experiments and appendix. See Sec. 4 and App. B, A, H, C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-2820

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our data and code with instructions for other researchers to reproduce our results. We share these resources in https://github.com/araloak/LLM-knowledgeBoundary.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We fully disclose the implementation details, Important Instructions in Experiments and appendix. See Sec. 4.1 and App. B, A, C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: I would like to clarify that the validation of our experimental results currently requires manual assistance, and the cost of repetition is relatively high. Therefore, we have not conducted significance analysis at this stage. We plan to address this in future work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information on the computer resources (type of compute workers, memory, time of execution) in the appendix. See App. B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: When we generated the dataset, we added a number of requirements to make the dataset comply with the NeurIPS Code of Ethics. See App C.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the social impact of our work from both positive and negative perspectives in App F.

    Guidelines:

https://doi.org/10.52202/079017-2820

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Yes, we describe guidelines for the safe use of our data in the readme file in our GitHub repository.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we have cited the creator and releaser of our used LLMs in the reference.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we provide a detailed description of our dataset in App.A. Besides, we also introduce our dataset in the repository where we release the data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We introduce the settings for our human evaluation in the paper and include the full instructions in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes, the paper thoroughly addresses potential risks to study participants, ensuring that all ethical considerations are met. We confirms that all necessary Institutional Review Board (IRB) approvals, or equivalent reviews as required by the country or institution, were obtained prior to conducting the study. This demonstrates a strong commitment to ethical research practices and the protection of participants' rights and well-being.

https://doi.org/10.52202/079017-2820

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.