Can AI Assistants Know What They Don't Know?

Qinyuan Cheng [*1,2] **Tianxiang Sun** [*1,2] **Xiangyang Liu** [1,2] **Wenwei Zhang** [2] **Zhangyue Yin** [1] **Shimin Li** [1]
Linyang Li [1,2] **Zhengfu He** [1] **Kai Chen** [2] **Xipeng Qiu** [1]

Abstract

AI assistants powered by Large Language Models (LLMs) have demonstrated impressive performance in various tasks. However, LLMs still make factual errors in knowledge-intensive tasks such as open-domain question answering. These untruthful responses from AI assistants can pose significant risks in practical applications. Therefore, in this paper, we ask the question "**Can AI assistants know what they don't know and express this awareness through natural language?**" To investigate this, we construct a model-specific "I don't know" (Idk) dataset. This dataset includes Supervised Fine-tuning data and preference data, categorizing questions based on whether the assistant knows or does not know the answers. Then, we align the assistant with its corresponding Idk dataset using different alignment methods, including Supervised Fine-tuning and preference optimization. Experimental results show that, after alignment with the Idk dataset, the assistant is more capable of declining to answer questions outside its knowledge scope. The assistant aligned with the Idk dataset shows significantly higher truthfulness than the original assistant.

1. Introduction

Large language models (Brown et al., 2020; Chowdhery et al., 2023; Zeng et al., 2023; Touvron et al., 2023) possess extensive world knowledge and demonstrate capabilities in numerous natural language tasks, capabilities that smaller models lack (Wei et al., 2022b). Recently, many artificial intelligence chat assistants built on LLMs have emerged, capable of assisting users with a variety of tasks in daily life and providing satisfactory user experiences (Ouyang et al., 2022; OpenAI, 2022; Anthropic, 2023; Sun et al., 2023; Baichuan, 2023; Qwen-Team, 2023). However, despite their frequent interactions with users, these chat assistants are prone to generating hallucinations (Shuster et al., 2021; Zhang et al., 2023c; Cheng et al., 2023), such as responses with factual errors (Wang et al., 2023b) or mimicking human falsehoods found in their training corpus (Lin et al., 2022a), some of which are difficult for users to detect. These untruthful responses could potentially harm society and also diminish the credibility of AI assistants.

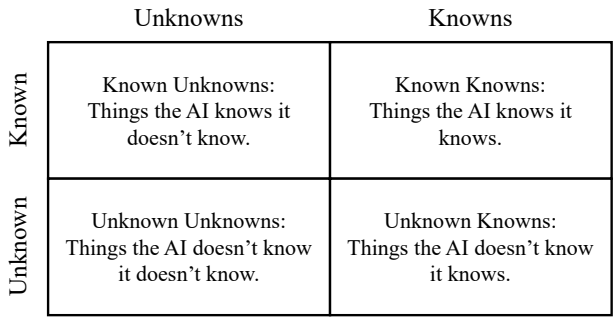

Figure 1. Knowledge quadrants of an AI assistant. "Unknowns" represents what the AI does not actually know. "Knowns" represents what the AI actually knows. "Known" represents what the AI believes it knows. "Unknown" represents what the AI believes it does not know.

An AI assistant aligned with human values must adhere to truthfulness (Evans et al., 2021), ensuring its information accurately reflects reality. When an assistant disseminates incorrect facts, it not only reveals a lack of knowledge but also an inability to acknowledge and communicate its limitations. A truthful AI, therefore, must recognize and convey its own knowledge boundaries clearly. It should provide precise information for what it knows and refrain from answering what it does not. This paper investigates the capability of AI assistants to discern their knowledge limits and articulate this uncertainty in natural language.

The AI assistant's understanding of its own knowledge

[*]Equal contribution [1]Department of Computer Science, Fudan University, Shanghai, China [2]Shanghai AI Laboratory, Shanghai, China. Correspondence to: Qinyuan Cheng <chengqy2019@foxmail.com>, Kai Chen <chenkai@pjlab.org.cn>, Xipeng Qiu <xpqiu@fudan.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

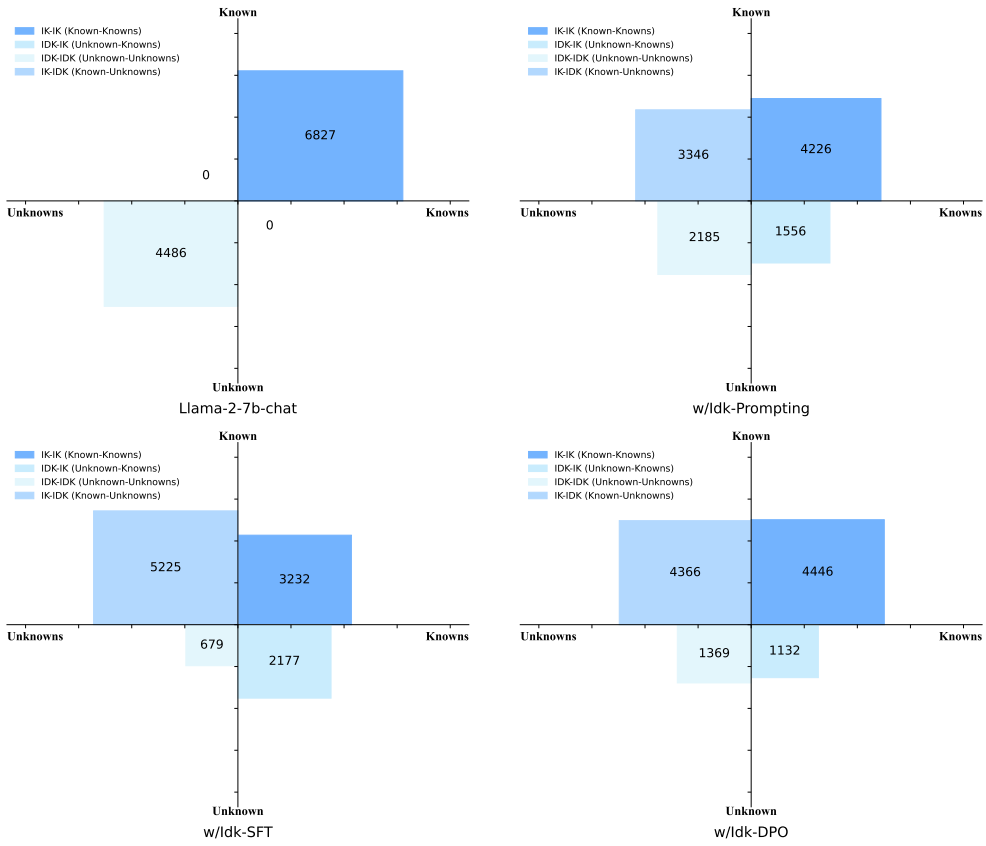

Figure 2. Knowledge quadrants of AI assistants on the Idk dataset (Ik threshold=1.0). **IK-IK** represents the AI answers the questions correctly. **IDK-IK** represents the AI knows the answer but refuses to respond to the question. **IDK-IDK** represents the AI answers the question incorrectly. **IK-IDK** represents the AI doesn't know the answer and refuses to respond to the question. **w/Idk-Prompting:** Using prompting can transform certain IDK-IDK questions to IK-IDK questions. **w/Idk-SFT:** Idk-SFT allows the model to refuse to answer more questions it does not know, but it also tends to make the model more convervative, leading to incorrect refusals to answer some questions that it actually knows. **w/Idk-DPO:** Using preference-aware optimization, like DPO, can alleviate the model's excessive conservatism and reduce the number of IDK-IK questions.

can be delineated through knowledge quadrants (Yin et al., 2023b). These quadrants categorize knowledge into four segments: Known Knowns, Known Unknowns, Unknown Knowns, and Unknown Unknowns, as depicted in Figure 1. Known Knowns are essential for an AI assistant's accuracy and reliability, with IK-IK (I know I know) symbolizing this category. The greater the extent of knowledge in the Known Knowns quadrant, the more helpful the AI assistant. Moreover, it is crucial for an AI assistant to recognize and communicate its limitations in knowledge, fitting into the Known Unknowns (IK-IDK: I know I don't know). Unknown Unknowns (IDK-IDK: I don't know I don't know) and Unknown Knowns (IDK-IK: I don't know I know) can lead to inaccuracies or helpless outputs. For AI assistants to be truthful, they must be programmed to distinguish between what they know and do not know, thereby transforming Unknown Knowns and Unknown Unknowns into Known Knowns and Known Unknowns, enhancing their truthfulness and utility.

Our approach aligns an AI assistant (like llama-2-7b-chat) with a model-specific **"I don't know" (Idk)** dataset, which catalogues the assistant's known and unknown questions. We construct the Idk dataset based on an existing knowledge-intensive open-domain question answering dataset, TriviaQA (Joshi et al., 2017). We assess if the assistant knows an answer by evaluating its average accuracy across several attempts at each question. Questions the assistant consistently answers incorrectly are identified as unknowns, and we annotate them with a template indicating lack of knowledge. Conversely, for questions answered correctly on multiple occasions, we use the assistant's responses as annotated answers. The accuracy threshold at which the assistant is deemed knowledgeable about a question is set as a hyperparameter, named the **Ik threshold**. The construction details of the Idk dataset are further elaborated in Section 3.1.

In order to teach AI assistants to know what they don't know, we conduct systematical experiments to exploit the most effective method, including prompting, supervised fine-tuning and preference-aware optimization [1]. For prompting, we instruct the assistant to refuse answering questions it does not know through a prompt. For supervised fine-tuning (SFT), we directly fine-tune the original assistant using our Idk datasets. For preference-aware optimization, we employ best-of-n sampling (BoN), proximal policy optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022), direct preference optimization (DPO) (Rafailov et al., 2023), and hindsight instruction relabeling (HIR) (Zhang et al., 2023b). Representative results are presented in Figure 2.

The original model (llama-2-7b-chat) can be considered as lacking the ability to recognize questions it does not know[2]. It often attempts to answer questions without the requisite knowledge, leading to numerous IDK-IDK instances and thus, diminishing its truthfulness. Instructing the model to refuse answering unknown questions through a prompt can be effective to some extent, but there are still numerous IDK-IK and IDK-IDK questions. After supervised fine-tuning using Idk dataset, the number of IDK-IK and IDK-IDK has significantly decreased, indicating that the model's ability to be aware of its own knowledge has been enhanced. Despite this improvement, there's an unintended side effect of the model declining to answer questions it actually knows, reducing the IK-IK responses. Compared to SFT model, preference-aware optimization (like DPO) can mitigate the phenomenon where the model incorrectly refuses to answer questions it knows. Besides, we conduct extensive ablation experiments to explore the effect of Ik threshold, data sources, model size and other settings.

Our findings can be summarized as follows:[3]

1. After aligning using Idk datasets, AI assistants are capable of largely knowing what they know and what they do not know and refusing their unknown questions. Llama-2-7b-chat can definitively determine whether it knows the answer to up to **78.96%** of the questions in the test set. And it exhibits good performance on out-of-distribution test sets.

2. Supervised fine-tuning tends to make the model overly cautious, leading to the erroneous rejection of known questions. Preference-aware optimization helps counteract this, increasing the proportion of accurately identified IK-IK and IK-IDK questions.

3. The Ik threshold used to define knowns and unknowns questions influences the behavior of the assistant. The more questions labeled as "I don't know," the more likely the assistant is to refuse to answer questions. In general, the higher the Ik threshold, the greater the total number of Ik-Ik and Ik-Idk questions, resulting in a more truthful assistant.

4. Larger model is more adept at distinguishing which questions it knows and which it doesn't know. The use of Idk-SFT on Llama-2-70b-chat, as compared to Llama-2-7b-chat, results in a 5.8% improvement in the total number of IK-IK and IK-IDK questions.

2. Related Work

Aligning LLMs with Human Values. To develop AI assistants utilizing large language models, it's essential to align these models with human values, ensuring they are helpful, truthful, and harmless (Askell et al., 2021; Bai et al., 2022; Ouyang et al., 2022). In this context, we highlight several prominent alignment methods pertinent to our study. The most common alignment method for pre-trained models is instruction tuning, also known as Supervised Fine-Tuning (SFT). Wei et al. (2022a); Sanh et al. (2022) fine-tune pre-trained models on a collection of NLP datasets combined with natural language instructions to enhance zero-shot performance on unseen tasks. Further, Chung et al. (2022); Longpre et al. (2023) expand task variety and model scale, fine-tuning on diversified data. Sun et al. (2023) employ Self-Instruct (Wang et al., 2023c) to generate SFT data encapsulating three key aspects: helpfulness, honesty, and harmlessness, for developing a conversational assistant. Beyond SFT, preference optimization emerges as a subsequent step. Bai et al. (2022); Ouyang et al. (2022) use Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020). They first train a reward model on the human preference data and then optimize the policy model using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with the trained reward model. Zhang et al. (2023b) propose a reward-free method named Hindsight Instruction Relabeling (HIR) to utilize preference data by converting feedback to instructions and training the model using supervised fine-tuning. Rafailov et al. (2023) propose Direct Preference Optimization (DPO), a method enabling direct fine-tuning of language models to align with human preferences, bypassing the need for reward modeling.

Discovering LLMs' Knowledge. Large language models encapsulate vast world knowledge during pre-training, sparking growing interest in exploring this knowledge. Ka-

[1] We use "preference-aware optimization" to refer to the method of using preference data for alignment, such as DPO, Reward Modeling, etc.

[2] We conducted a search for keywords such as "I don't know", "not sure", "Sorry" in the responses of Llama-2-7b-chat and found that only a very small number of responses contained these keywords.

[3] We release our code, data and models at https://github.com/OpenMOSS/Say-I-Dont-Know.

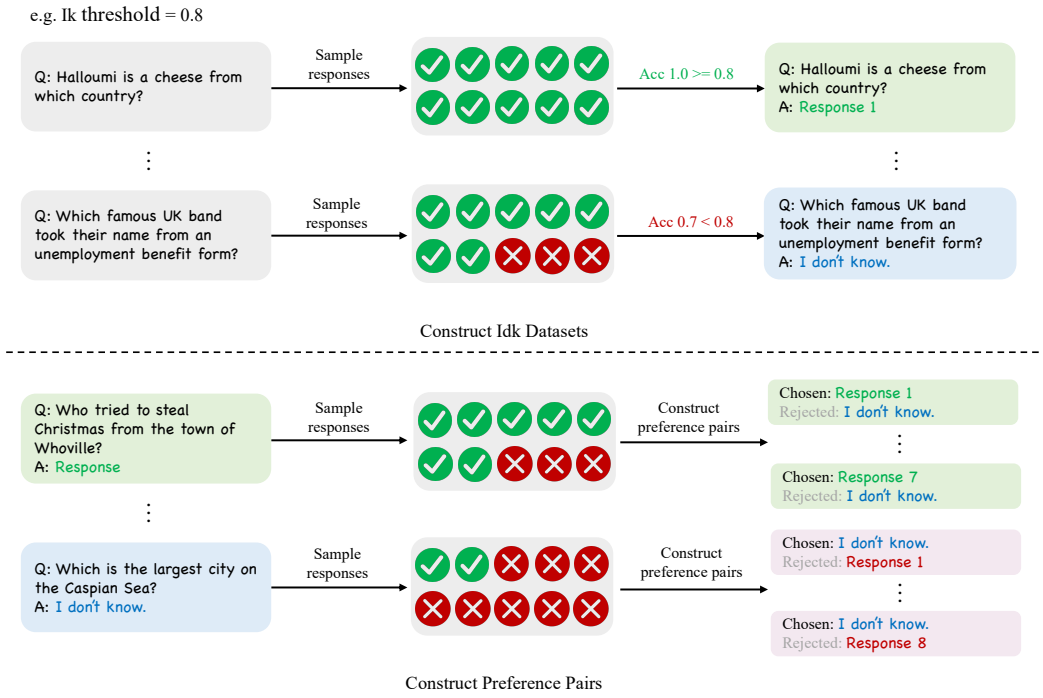

Figure 3. **Top**: Construction process of the Idk dataset. **Bottom**: Construction process of preference pairs. The green response indicates a correct answer, the red response indicates an incorrect answer, and "I don't know" represents the template for refusal to answer.

davath et al. (2022); Lin et al. (2022b) fine-tune language models using a classification head or verbalized confidence, yet these methods do not teach models to aware their knowledge boundary and refuse to answer the questions they don't know. Yin et al. (2023b), Amayuelas et al. (2023) and Liu et al. (2024) investigate whether large language models can identify unanswerable questions. The unanswerable question includes questions about the future we cannot know, questions about science , history or problems that we don't know the answer to, subjective questions, questions based on a hypothetical scenario, etc. These questions are unanswerable by all models, not unknown to a specific model. Burns et al. (2023) develop an unsupervised method to find latent knowledge inside the activations of a language model by answering yes-no questions given only unlabeled model activations. Ren et al. (2023) investigate whether LLMs can perceive their knowledge boundaries or not under retrieval-augmented setting and normal setting. Additionally, Zhang et al. (2023a) and Yang et al. (2023) try to teach language models to refuse unknown questions through supervised fine-tuning. In our work, we systematically explore whether it is possible to teach AI assistants to say "I don't know" to their unknown questions. We investigate the impact of prompting, supervised fine-tuning and preference-aware optimization. And we utilize the knowledge quadrant to track the changes in various types of knowledge within the model following the application of different alignment methods.

Mitigating LLMs' Factual Errors. There are some studies focus on eliminating factual errors in AI assistants. Asai et al. (2023) propose a framework named SELF-RAG to enhance an LM's factuality by retrieval augmentation and self-reflection. Li et al. (2023) first find truthful directions through probing and then do inference-time intervention in these truthful directions. Zou et al. (2023) use representation engineering to enhance factuality of the model's output. Chuang et al. (2023) propose a simple decoding strategy for reducing hallucinations by contrasting the differences in tokens' logits obtained from different layers. Tian et al. (2023) directly fine-tune language models to learn factuality from preference dataset using direct preference optimization. However, there is currently no method that can guarantee the complete elimination of factual errors. In practical applications, it is a necessary feature for AI assistants to refuse to answer questions they do not know.

3. Methodology

3.1. Construction of the Idk Dataset

Assessing an AI model's knowledge on a question-answering dataset poses challenges, particularly in determining the questions for which the model truly knows the answers. This complexity arises from the model's varying knowledge mastery levels across different knowledge do-

mains. Therefore, following the approach of previous work (Kadavath et al., 2022; Lin et al., 2022b), we evaluate the model's knowledge by sampling multiple responses to each question and calculating the accuracy rate across these responses. This accuracy can be used to measure the model's mastery of a certain knowledge. We then define a particular accuracy as the Ik threshold, which helps ascertain whether the model knows or does not know the answer to a question. To construct the QA pairs in the Idk dataset, for questions that the model does not know, we use a template for refusal to reply as the answer. For questions that the model knows, we select a correct response generated by the model itself as the answer. The procedure is demonstrated in Figure 3 (top). Our refusal to answer template is:

```
This question is beyond the scope of my
knowledge, and I am not sure what the
answer is.
```

We use both "I don't know" and "Idk template" to refer to this template in our subsequent discussion.

Determine whether the output of a model is correct To develop the Idk dataset, an automatic method is required to assess the accuracy of the model's responses. Experimental evidence from Wang et al. (2023a) indicates that lexical matching, which involves verifying the presence of the golden answers within the model-generated responses, achieves approximately a 90% consistency rate with human evaluations when applied to a portion of the TriviaQA validation set (Joshi et al., 2017). This suggests that lexical matching serves as a sufficiently reliable method for automatic evaluation on the TriviaQA dataset. Given TriviaQA's status as a prominent knowledge-intensive open-domain question answering dataset, it forms the foundation for our Idk dataset construction.

Meaning of different Ik thresholds The model's different response strategies are determined based on the relationship between its level of knowledge mastery and the Ik threshold. Notably, varying the Ik threshold alters the composition of the Idk dataset: a higher threshold necessitates greater mastery for the model to respond, embodying a conservative strategy, whereas a lower threshold permits responses at reduced mastery levels, reflecting an aggressive approach. In this work, we sample ten responses for each question and derive ten discrete Ik thresholds based on different accuracy rates. For simplicity, and to ensure the model's responses are highly reliable, we set the Ik threshold at 1.0. This means the model is deemed knowledgeable on a question only if it correctly answers all ten times. Unless specifically stated otherwise, the Idk dataset mentioned hereafter is constructed based on an Ik threshold of 1.0. We discuss the impact of different Ik thresholds in Section 4.4.

In the following sections, we introduce our methods to teach AI assistants to say "I don't know" when encounter unknown questions. Since the AI assistant we discuss is based on large language models, we will interchangeably use the terms "model" and "assistant" in the following sections.

3.2. Idk Prompting

For models capable of following human instructions, such as Llama-2-7b-chat, We can directly instruct an assistant to say "I don't know" to unknown questions by adding a prompt in front of the input question. We call this method Idk-Prompting. This requires the model to have a high capability for following instructions, but the advantage is that it eliminates the need for additional training. We call such a prompt an Idk prompt. Our Idk prompt is as follows:

```
Answer the following question, and if you
don't know the answer, only reply with "I
don't know": <Question>
```

As for pre-trained models lacking the ability to follow instructions, Idk-Prompting may not yield satisfactory results.

3.3. Idk Supevised Fine-tuning

Supervised Fine-tuing is a simple yet effective alignment method. We directly use the Idk dataset for Supervised Fine-tuning of the model. Since the Idk dataset contains both questions and responses, this constitutes a conditional generation task. We input the questions into the model and require the model to predict the responses. We perform the standard sequence-to-sequence loss to train our model. SFT details are demonstrated in Appendix B.1.

3.4. Preference-aware Optimization

In this section, we introduce how we conduct preference-aware optimization to help the model perceive its internal knowledge better.

Direct Preference Optimization (DPO) To implement DPO, we first train a SFT model on one half of the Idk dataset as a warm up. Subsequently, we gather responses from this model by randomly sampling multiple answers from the remaining half of the Idk dataset. This process is aimed at compiling preference data from the generated responses, as illustrated in Figure 3 (bottom)[4]. Each preference data entry comprises a question, a chosen response, and a rejected response. The questions in the Idk dataset can be categorized into two types: those the model knows and those it does not know. For questions the model knows, we use the correct response generated by it as the chosen response and "I don't know" as the rejected response. For questions the model does not know, we use "I don't know" as the

[4]We omit the cases where the model responds with 'I don't know' in the figure.

chosen response and its incorrectly generated response as the rejected response. Additionally, we found that only using the DPO loss Rafailov et al. (2023) can occasionally result in the model's inability to accurately generate the Idk template. To counteract this, in addition to the original DPO loss, we also incorporate SFT loss for the chosen responses and multiply it by a coefficient α. The details of the DPO are demonstrated in Appendix B.2.

Best-of-n Sampling (BoN) We also try to determine if the model knows the answer to a certain question by training a reward model to score the candidate responses. We first train a SFT model using a half of the Idk data and then use the SFT model to initialize the reward model. After collecting responses on the other half of the Idk dataset and constructing preference data using the same procedure as 3, we train the reward model using a pairwise loss. During inference, we employ the Best-of-10 strategy. First, we sample ten responses using the SFT model, then we score these candidate responses with the reward model. The response with the highest reward score is selected as the final response. The details of reward modeling are demonstrated in Appendix B.3.

Proximal Policy Optimization (PPO) Based on our reward model, we can use proximal policy optimization to optimize the model. We use the same inputs for PPO training as we do for reward modeling, but sample responses in an online manner. The details of the PPO are demonstrated in Appendix B.4.

Hindsight Instruction Relabeling (HIR) So far, our Idk dataset is constructed based on a fixed Ik threshold. In order to utilize all Idk datasets constructed with different Ik thresholds, inspired by Hindsight Instruction Relabeling (Zhang et al., 2023b), we design an instruction format to relabel all Idk datasets. Specifically, we prepend the following instruction to each question in the Idk datasets:

```
Your current knowledge expression
confidence level is <X>, please answer the
user's question: <Question>
```

where $< Question >$ is a question from an Idk dataset and $< X >$ is the value of model's knowledge expression confidence level ranging from 0 to 1.0, derived from the Ik threshold corresponding to the Idk dataset. The lower the knowledge expression confidence level, the more inclined the model is to refuse answering questions. Then we use the combined Idk dataset to perform supervised fine-tuning. The advantage of using instruction relabeling is that we can control the model to adopt either a conservative or aggressive response strategy through the instruction, without the need to retrain the model. The details of HIR are demonstrated in Appendix B.5.

4. Experiments

4.1. Dataset

TriviaQA (Joshi et al., 2017), originally a reading comprehension dataset, is utilized here for open-domain question answering tasks, forming the basis of our Idk dataset with 87,622 training samples and an 11,313 sample test set derived from TriviaQA's development set, due to the absence of ground truth in its test set. Further details on the Idk dataset are provided in Appendix A.

For out-of-distribution (OOD) evaluation, we incorporate the Natural Questions (NQ) (Kwiatkowski et al., 2019) and ALCUNA (Yin et al., 2023a) datasets. NQ, featuring real queries from the Google search engine, contributes 3,610 development set samples to our OOD test set. Lexical matching, demonstrating over 80% consistency with human assessments according to Wang et al. (2023a), is employed for automatic evaluation of model responses against the NQ dataset.

ALCUNA is a benchmark to assess LLMs' abilities in new knowledge understanding. It creates new artificial entities by altering existing entity attributes and generates questions about these artificial entities. Since these entities are artificially created, the model cannot possibly possess this knowledge. Therefore, we use a portion of the questions from ALCUNA to test whether the model can refuse to answer, totaling 8,857 samples.

4.2. Metrics and Evaluation

We evaluate the model using the following metrics:

- IK-IK Rate: I know what I know (Ik-Ik) rate represents the proportion of questions answered correctly by the model out of all questions, indicating its knowledge accuracy.
- IK-IDK Rate: I know what I don't know (Ik-Idk) rate represents the proportion of questions that the model correctly refuses to answer out of all questions, showcasing its ability to recognize limitations.
- TRUTHFUL Rate: Truthful rate is the sum of Ik-Ik rate and Ik-Idk rate. It represents the proportion of questions for which the model provides truthful responses. The higher the value of TRUTHFUL rate, the clearer the model's perception of what it knows and does not know, which also indicates a higher level of truthfulness. An ideal model achieves a TRUTHFUL rate of 100%, denoting complete accuracy and self-awareness.

The higher these three metrics are, the better. The primary metric, the TRUTHFUL rate, is crucial as it signifies the likelihood of the model providing a truthful response. Detailed

Table 1. Overall results on the test set of the Idk dataset constructed based on TriviaQA and out-of-distribution test sets.

	TriviaQA			Natural Questions			ALCUNA
	IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL	IK-IDK
Idk-Dataset$_{test}$	45.05	54.95	100.00	24.65	75.35	100.00	100.00
Idk-Prompting	37.36	29.58	66.93	19.75	41.72	61.47	91.67
Idk-SFT	28.57	46.19	74.75$_{\uparrow 7.82}$	15.93	53.99	69.92$_{\uparrow 8.45}$	98.01
Idk-DPO	**39.30**	38.59	77.89$_{\uparrow 10.96}$	20.91	45.60	66.51$_{\uparrow 5.04}$	98.08
Idk-BoN$_{N=10}$	38.37	40.59	**78.96**$_{\uparrow 12.03}$	20.55	47.40	67.95$_{\uparrow 6.48}$	98.32
Idk-PPO	35.90	40.57	76.47$_{\uparrow 9.54}$	**23.13**	42.08	65.21$_{\uparrow 3.47}$	92.66
Idk-HIR	27.36	**48.55**	75.91$_{\uparrow 8.98}$	15.40	**56.90**	**72.30**$_{\uparrow 10.83}$	**98.96**

metric calculations are available in Appendix B.6.

We use Llama-2-7b-chat as our initial model for further training, with specific training details introduced in Appendix B. We test the trained model on the test set of the Idk dataset to evaluate whether the model can distinguish between questions it knows and does not know. Except for Idk-BoN, we use greedy decoding in all tests. For Idk-BoN, we set the temperature coefficient to 1.0 and top_p to 0.9, sample ten responses, and then score them using the reward model. The response with the highest reward score is selected as the final model response.

4.3. Main Results

The overall results are in Table 1. The Idk-Dataset used for evaluation contains 45.05% IK-IK questions and 54.95% IK-IDK questions, which can be seen as two upper bounds of IK-IK and IK-IDK rate. Simply using an Idk prompt to let the model refuse to answer questions it doesn't know can have a certain effect, but the model's TRUTHFUL rate is still only 66.93%. The Idk-SFT can increase the TRUTHFUL rate to 74.75%, but this will result in a decrease in the IK-IK rate, which can be considered a form of "alignment tax". We find that preference optimization can encourage the model to answer questions, thereby mitigating the alignment tax. DPO, PPO, and BoN can all reduce the loss of IK-IK while maintaining a relatively high IK-IDK rate. Idk-BoN achieves the highest TRUTHFUL rate. Idk-HIR combines all Idk datasets, which can improve IK-IDK rate but help less for IK-IK rate. However, Idk-HIR provides an switching method for Ik-threshold that does not need to retrain the model. Overall, by aligning with the Idk dataset, we can transform IDK-IK and IDK-IDK questions into IK-IK and IK-IDK questions. The model can have a clear perception of whether it knows the answers to most questions in the test set, significantly increasing truthfulness compared to before the alignment. The overall experimental results of all knowledge quadrants are represented in Appendix C.2. We also include comparisons with some logits-based methods in Appendix C.6.

Evaluation on out-of-distribution data We also test whether the aligned model is capable of refusing to answer questions it does not know when encountering out-of-distribution (OOD) data. We first construct the Idk dataset for testing based on Natural Questions, setting the Ik threshold to 1.0. As shown in Tabel 1, the Idk dataset contains 24.65% IK-IK questions and 75.35% IK-IDK questions, which means Natural Questions is more challenging than TriviaQA. The results on Natural Questions are similar to those on TriviaQA. The algined models show improvements in all metrics compared to using prompts. In contrast to the results on TriviaQA, Idk-HIR achieves the highest TRUTHFUL rate, rather than Idk-BoN. This is because the proportion of IK-IDK questions in Natural Questions test set is higher than TrivialQA. We demonstrate the results of a sampled test set which has the same proportion as TriviaQA in Appendix C.3 and Idk-BoN gets the highest TRUTHFUL rate. Furthermore, the models aligned using preference optimization methods exhibit a reduction in the TRUTHFUL rate compared to the Idk-SFT. We believe this is due to the fact that preference optimization encourages the model to answer more questions. We can observe that, compared to the Idk-SFT model, preference-optimized models have more IK-IK questions but less IK-IDK questions. In addition to this, we utilize ALCUNA to construct the Idk dataset, which only contains ID-IDK questions. The results from Table 1 indicate that the prompting method can already enable the model to refuse answering most unanswerable questions. After alignment, the model achieves an even higher IK-IDK rate[5]. The model aligned with TriviaQA demonstrates a high TRUTHFUL rate on Natural Questions and a high IK-IDK rate on ALCUNA, suggesting that the model's behavior of refusing to answer unknown questions can be generalized to OOD data.

4.4. Ablation Study

Effect of model size The efficacy of LLMs often correlates with their parameter count, with larger models typically exhibiting enhanced capabilities. We apply Idk-SFT

[5]We find that the DPO model, when refusing to answer questions within ALCUNA, occasionally rephrases our Idk template. Consequently, we utilize a substring of the original Idk template: "I am not sure what the answer is" to detect whether the model refuse to answer the question.

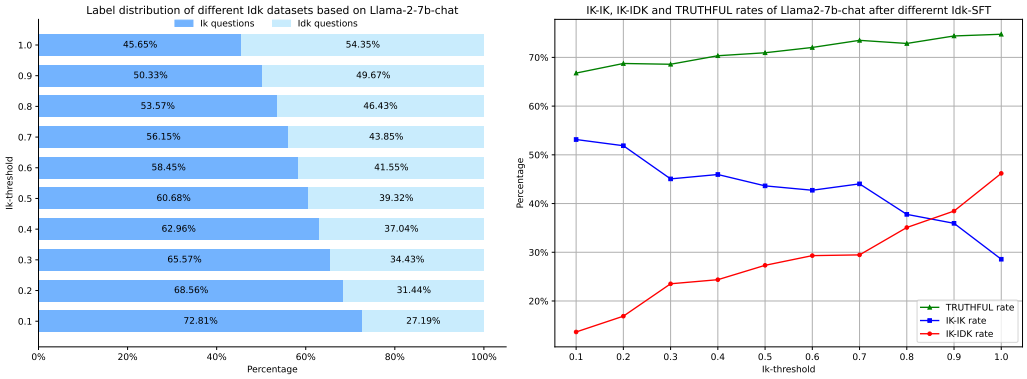

Figure 4. **Left:** Variation in the proportions of Ik and Idk questions within the Idk datasets constructed based on different Ik thresholds. **Right:** The changes in IK-IK rate, IK-IDK rate, and TRUTHFUL rate after conducting Idk-SFT with different Idk datasets.

Table 2. Results of ablation experiments.

	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-SFT$_{7b}$	28.57	46.19	19.24	6.00	74.75
w/Llama-2-13b-chat	33.92	41.43	17.45	7.20	$75.35_{\uparrow 0.60}$
w/Llama-2-70b-chat	57.78	22.68	10.78	8.66	$80.55_{\uparrow 5.8}$
w/Idk-Mistral	18.35	50.65	27.68	3.31	$69.00_{\downarrow 5.75}$
w/Idk-Baichuan	8.85	53.07	36.37	1.71	$61.92_{\downarrow 12.83}$

to Llama-2-7b-chat, Llama-2-13b-chat, and Llama-2-70b-chat, examining the influence of model size on Idk-SFT's effectiveness. Table 2 details the knowledge quadrant proportions for each model. Notably, the label distribution of the Idk dataset corresponding to different initial models is inconsistent (the larger the model, the more IK-IK questions), as shown in Appendix A.3. This results in the IK-IK rate and IK-IDK rate being incomparable. Therefore, we mainly focus on the TRUTHFUL rate of different models. The TRUTHFUL rate of the 13B model is slightly higher than that of the 7B model. The TRUTHFUL rate of the 70B model is significantly higher than that of the 13B and 7B models. This indicates that larger models are more adpet at distinguishing between questions they know and do not know.

Effect of data sources Different pre-trained models harbor unique knowledge bases due to their distinct pre-training processes. During training, we construct model-specific Idk dataset for different pre-trained models. This is because we want the model to determine whether it knows the answer to a question based on its internal knowledge, rather than learning to recognize some specific patterns of questions. The model-specific Idk dataset can connect the model's internal knowledge with the labels of the Idk dataset. To explore the impact of using a non-model-specific Idk dataset on training, we construct two Idk training sets using Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) and Baichuan2-7B-chat (Baichuan, 2023) respectively, named "Idk-Mistral" and "Idk-Baichuan". We present label distributions of these Idk datasets in Appendix A.3 As shown in Tabel 2, using non-model-specific Idk datasets like "Idk-Mistral" or "Idk-Baichuan" does result in a TRUTHFUL rate loss. Due to the numerous Idk questions in the Idk-Mistral and Idk-Baichuan datasets, the trained model tends to be more inclined towards refusing to answer questions, which has resulted in a significant reduction in Ik-Ik related queries, far below their proportion in the dataset. This indicates that constructing a model-specific Idk dataset is necessary for enabling the model to learn to refuse to answer questions it does not know.

Effect of Ik threshold Here, we discuss the impact of different Ik thresholds on model behaviors. We mainly focus on the impact of Ik threshold on Idk-SFT. The Ik threshold primarily affects the distribution of labels in the Idk dataset, with a higher Ik threshold indicating that more questions will be labeled as "I don't know". As demonstrated in Figure 4 (left), the higher the value of the Ik threshold, the greater the proportion of Idk questions. This is because when the Ik threshold is high, only questions with a high knowledge mastery will be annotated as questions known to the model. As shown in Figure 4 (right), increasing the Ik threshold results in a decrease in the IK-IK rate and an increase in the IK-IDK rate. As the Ik threshold is raised, the model's TRUTHFUL rate will continue to improve. In other words, setting a high Ik threshold aids the model in better distinguish between knowledge it knows and does not know, making the model more truthful. In contrast, setting a low Ik threshold can make the model more helpful, since the number of IK-IK questions will increase. Besides, we find that as the proportion of Idk questions in the dataset increases, the model tends to refuse to answer questions more frequently. We report the F1 scores of Idk and Ik questions in different Idk datasets in Appendix C.4 and the knowledge quadrants under different Ik thresholds in C.1.

5. Conclusion

In this study, we address the question, "Can AI assistants know what they don't know?" Our findings indicate that

by aligning an AI assistant, such as Llama-2-7b-chat, with a tailored "I don't know" (Idk) dataset, which catalogues both its known and unknown questions, the AI assistant can, to a significant extent, identify what it does not know. In open-domain question-answering tests, Llama-2-7b-chat is able to accurately determine its knowledge boundaries for 78.96% of the questions, opting to refrain from answering those it could not confidently address. To accomplish this, we employ a variety of alignment strategies with the Idk dataset, including supervised fine-tuning and preference-aware optimization. Our analysis reveal that the Ik threshold which determines knowns and unknowns influences the model's tendency to decline responses. Moreover, using Idk datasets derived from different models tends to diminish performance, while larger models, like Llama-2-70b-chat, attain a superior TRUTHFUL rate. This capability of an AI assistant to decline answering questions beyond its knowledge effectively mitigates hallucinations, a trait we deem vital for maintaining the truthfulness of AI assistants.

Impact Statement

This research contributes to the field of LLM-based AI assistants by enhancing the ability of AI assistants to acknowledge their knowledge limits, offering a pathway towards more trustworthy and transparent machine learning applications. Ethically, it emphasizes the importance of developing AI systems that prioritize accuracy and truthful, particularly in sectors where reliable information is crucial. The societal implications include reducing misinformation and fostering a cautious approach to AI reliance, ensuring that AI aids rather than misleads human decision-making.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No.2022ZD0160102). The computations in this research were performed using the CFFF platform of Fudan University.

References

Amayuelas, A., Pan, L., Chen, W., and Wang, W. Y. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *CoRR*, abs/2305.13712, 2023. doi: 10.48550/ARXIV.2305.13712. URL https://doi.org/10.48550/arXiv.2305.13712.

Anthropic. Introducing claude, 2023. URL https://www.anthropic.com/index/introducing-claude.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511, 2023. doi: 10.48550/ARXIV.2310.11511. URL https://doi.org/10.48550/arXiv.2310.11511.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 2021. URL https://arxiv.org/abs/2112.00861.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., Showk, S. E., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. URL https://arxiv.org/abs/2309.10305.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=ETKGuby0hcs.

Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., He, J., Huang, M., Yin, Z., Chen, K., and Qiu, X. Evaluating hallucinations in chinese large language models. *CoRR*, abs/2310.03368, 2023. doi: 10.48550/ARXIV.2310.03368. URL https://doi.org/10.48550/arXiv.2310.03368.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017.

Chuang, Y., Xie, Y., Luo, H., Kim, Y., Glass, J. R., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883, 2023. doi: 10.48550/ARXIV.2309.03883. URL https://doi.org/10.48550/arXiv.2309.03883.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., and Saunders, W. Truthful AI: developing and governing AI that does not lie. *CoRR*, abs/2110.06674, 2021. URL https://arxiv.org/abs/2110.06674.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., Showk, S. E., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221. URL https://doi.org/10.48550/arXiv.2207.05221.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A. P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, K., Patel, O., Viégas, F. B., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *CoRR*, abs/2306.03341, 2023. doi: 10.48550/ARXIV.2306.03341. URL https://doi.org/10.48550/arXiv.2306.03341.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022a. doi: 10.18653/v1/2022.acl-long.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022b. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Liu, G., Wang, X., Yuan, L., Chen, Y., and Peng, H. Examining llms' uncertainty expression towards questions outside parametric knowledge, 2024.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The flan collection: Designing data and methods for effective instruction tuning, 2023.

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Qwen-Team. Qwen technical report. 2023. URL https://qianwen-res.oss-cn-beijing.aliyuncs.com/QWEN_TECHNICAL_REPORT.pdf.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290, 2023. doi: 10.48550/ARXIV.2305.18290. URL https://doi.org/10.48550/arXiv.2305.18290.

Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., Wu, H., Wen, J., and Wang, H. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *CoRR*, abs/2307.11019, 2023. doi: 10.48550/ARXIV.2307.11019. URL https://doi.org/10.48550/arXiv.2307.11019.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., Datta, D., Chang, J., Jiang, M. T., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Févry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In Moens, M., Huang, X., Specia, L., and Yih, S. W. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 3784–3803. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://doi.org/10.18653/v1/2021.findings-emnlp.320.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

Sun, T., Zhang, X., He, Z., Li, P., Cheng, Q., Yan, H., Liu, X., Shao, Y., Tang, Q., Zhao, X., Chen, K., Zheng, Y., Zhou, Z., Li, R., Zhan, J., Zhou, Y., Li, L., Yang, X., Wu, L., Yin, Z., Huang, X., and Qiu, X. Moss: Training conversational language models from synthetic data. 2023.

Tian, K., Mitchell, E., Yao, H., Manning, C. D., and Finn, C. Fine-tuning language models for factuality. *CoRR*, abs/2311.08401, 2023. doi: 10.48550/ARXIV.2311.08401. URL https://doi.org/10.48550/arXiv.2311.08401.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Wang, C., Cheng, S., Guo, Q., Xu, Z., Ding, B., Wang, Y., Hu, X., Zhang, Z., and Zhang, Y. Evaluating open-qa evaluation, 2023a. URL https://arxiv.org/abs/2305.12421.

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Cheng, J., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., and Zhang, Y. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521, 2023b. doi: 10.48550/ARXIV.2310.07521. URL https://doi.org/10.48550/arXiv.2310.07521.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13484–13508. Association for Computational Linguistics, 2023c. doi: 10.18653/V1/2023.ACL-LONG.754. URL https://doi.org/10.18653/v1/2023.acl-long.754.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022b. URL https://openreview.net/forum?id=yzkSU5zdwD.

Yang, Y., Chern, E., Qiu, X., Neubig, G., and Liu, P. Alignment for honesty. *CoRR*, abs/2312.07000, 2023. doi: 10.48550/ARXIV.2312.07000. URL https://doi.org/10.48550/arXiv.2312.07000.

Yin, X., Huang, B., and Wan, X. ALCUNA: large language models meet new knowledge. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 1397–1414. Association for Computational Linguistics, 2023a. URL https://aclanthology.org/2023.emnlp-main.87.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. Do large language models know what they don't know? In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 8653–8665. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.findings-acl.551. URL https://doi.org/10.18653/v1/2023.findings-acl.551.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., and Tang, J. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

Zhang, H., Diao, S., Lin, Y., Fung, Y. R., Lian, Q., Wang, X., Chen, Y., Ji, H., and Zhang, T. R-tuning: Teaching large language models to refuse unknown questions. *CoRR*, abs/2311.09677, 2023a. doi: 10.48550/ARXIV.2311.09677. URL https://doi.org/10.48550/arXiv.2311.09677.

Zhang, T., Liu, F., Wong, J., Abbeel, P., and Gonzalez, J. E. The wisdom of hindsight makes language models better instruction followers. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41414–41428. PMLR, 2023b. URL https://proceedings.mlr.press/v202/zhang23ab.html.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023c. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

A. Idk Dataset Construction Details

A.1. Data Statistics

We use the training set of TriviaQA, consisting of 87,622 samples, to construct the training and development sets of the Idk dataset. We partition 10% of the training set of TriviaQA to serve as the validation set of the Idk dataset, with the other 90% as the training set. Therefore, the validation set contains 8,763 samples and the training set contains 78,899 samples. We use the development set of TriviaQA to construct the test set for the Idk dataset, which comprises a total of 11,313 samples. The number of samples in each part of the Idk dataset for different models is the same, it is only the distribution of the labels that varies.

A.2. Sampling Parameters

When constructing the Idk dataset through sampling model responses, our sampling parameters are set as follows: top_p=0.9, temperature=1.0, max_new_tokens=512, repetition_penalty=1.0 (no penalty). We use this set of parameters for all random sampling in this work.

A.3. Label Distribution of Idk Datasets

In Figure 5 and Figure 6, we present the label distribution in the Idk datasets constructed using different Ik thresholds across various models. It is evident that different models possess varying knowledge reserves, as indicated by the distinct differences in the label distribution of their Idk datasets. As shown in Figure 6, the larger the size of the model, the more extensive its knowledge, resulting in fewer questions being labeled as "I don't know".

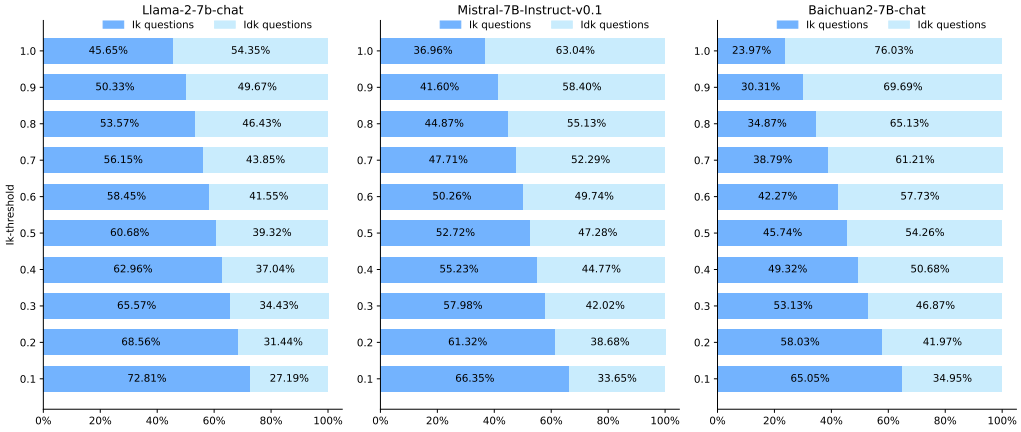

Figure 5. Label distribution in the Idk dataset across different models.

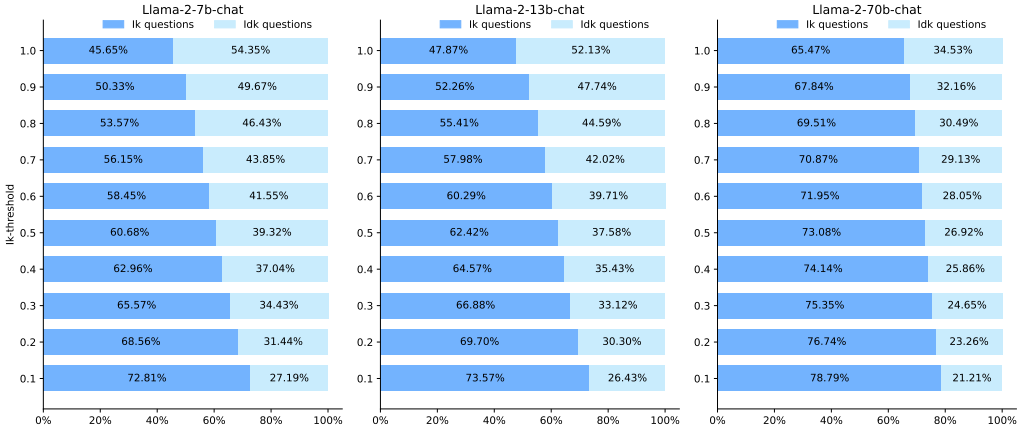

Figure 6. Label distribution in the Idk dataset across different sizes.

B. Training and Evaluation Details

B.1. Supervised Fine-tuning

We organize our Idk dataset into single-turn dialogues following the conversation format of Llama-2-7b-chat and then use the standard SFT loss to train the model:

$$\mathcal{L}_{SFT} = -E_{(x,y)\sim D}[\frac{1}{N}\sum_{t}^{N}\log p(y_t|x, y_{<t}; \theta)] \tag{1}$$

(x, y) is a question-answering pair in the Idk dataset, where x represents the question, and y represents the answer. N represents the length of the answer y, and θ represents the model parameters. During training, we employ a packing strategy to combine multiple samples into a single sequence with a maximum length of 4096. Following the settings of llama-recipes, our batch size is set to 32, with a learning rate of 1e-4 and train 10 epochs. During training, we save a checkpoint at the end of each epoch, and select the checkpoint that performs the best on the validation set as the final model. We employed Fully Sharded Data Parallelism (FSDP) to conduct SFT training on eight A100 80G GPUs. For Llama-2-70b-chat, we train 10 epochs using 32 A100 80G GPUs and select the checkpoint of the last epoch as the final model. The decision to forego the use of a validation set for model selection was based on our observation that the model exhibiting the lowest loss on the validation set tended to erroneously reject numerous Ik questions. We speculate that this may be attributed to the inherent alignment training of the Llama-2-70b-chat itself.

B.2. Direct Preference Optimization

The original DPO loss proposed by Rafailov et al. (2023) is:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]. \tag{2}$$

where π_{ref} is the SFT model trained with half of the Idk data, π_θ is the policy model, y_w is the chosen response and y_l if the rejected response. To alleviate the problem of the DPO model sometimes failing to fully generate the Idk template, we additionally incorporate the SFT loss. Our final loss function of direct preference optimization is:

$$\mathcal{L}_{DPO-SFT} = \mathcal{L}_{DPO} + \alpha * \mathcal{L}_{SFT} \tag{3}$$

In the experiment, we set the coefficient α of the SFT loss to 0.01. The hyperparameters of our SFT model training are the same as Appendix B.1. During DPO training, following DPO's official implementation, we set our batch size to 64, the learning rate to 5e-7, β to 0.1 and train for one epoch. We partition 10% of the preference data to construct a validation set to select the best checkpoint. We use 8 A100 80G GPUs for DPO training. We present the impact of different α values on the model's TRUTHFUL rate in Table 3.

Table 3. The impact of the coefficient α of the SFT loss on the model's TRUTHFUL rate.

	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$
Ik-threshold=0.5	74.28	72.39	72.06	72.31	72.08
Ik-threshold=1.0	66.14	77.89	76.68	75.55	75.72

As shown in Table 4, when using the Idk dataset constructed with Ik-threshold=0.5 for DPO training, the model is capable of accurately generating the Idk template. In this scenario, incorporating SFT loss reduces the model's TRUTHFUL rate. However, when using the Idk dataset constructed with Ik-threshold=1.0 for DPO training, the model occasionally fails to accurately generate the Idk template. In such cases, employing a coefficient of 0.01 yields the most effective mitigation.

B.3. Best-of-n Sampling

We train the reward model using a pairwise loss:

$$\mathcal{L}_{RM} = -E_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(r(x_i, y_w) - r(x_i, y_l)\right)\right] \tag{4}$$

where (x, y_w, y_l) is a question-chosen-rejected triplet from the preference dataset. During training of the reward model, we set batch size to 128, learning rate to 9e-6, and train for one epoch. We partition 10% of the preference data to construct a validation set to select the best checkpoint. We use 4 A100 80G GPUs for reward model training.

B.4. Proximal Policy Optimization

We employ the SFT model and reward model obtained from B.3 fro PPO training. We use DeepSpeed-Chat for PPO training. The SFT model and reward model used in PPO training are obtained from the BoN's supervised fine-tuning and reward modeling. For PPO (Schulman et al., 2017), the loss function of the actor model is:

$$\mathcal{L}_{PPO-Actor} = -\hat{E}_t[max(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t], r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{5}$$

And the loss function of the critic model is:

$$\mathcal{L}_{PPO-Critic} = 0.5 * \hat{E}_t[max((V_\phi(s_t) - \hat{R}_t)^2, clip(V_\phi(s_t), V_{old}(s_t) + \epsilon, V_{old}(s_t) - \epsilon))] \tag{6}$$

We set the learning rate for both the actor model and the critic model to 1e-6. The generation batch size is 64 and the training batch size is 32. Each training step, we train a single inner epoch. We utilize DeepSpeed ZeRO-3 to train one epoch on 32 A100 80G GPUs.

B.5. Hindsight Instruction Relabeling

We combine 10 Idk datasets using the HIR method, constructed from 10 distinct Ik thresholds ranging from 0.1 to 1.0. These Ik thresholds correspond to knowledge expression confidence level from 1.0 to 0.1, respectively. The lower the knowledge expression confidence level, the less confident the model is in its own knowledge, resulting in a more conservative response strategy. Besides, we also add a dataset consisting entirely of refusals to respond, corresponding to situations where the knowledge expression confidence level is 0 and its Ik threshold can be seen as 1.1. We utilize the following formula to convert from the Ik threshold to the knowledge expression confidence level:

$$Knowledge\_expression\_confidence\_level = 1.1 - Ik\_threshold \tag{7}$$

For example, we prepend the following instruction to questions in the Idk dataset corresponding to an Ik threshold of 1.0:

```
Your current knowledge expression confidence level is 0.1, please answer the user's
question: <Question>
```

We set the batch size to 256, the learning rate to 2e-5 and we train for 3 epochs using 8 A100 80G GPUs. The advantage of this method is that it allows users to control the model's response strategy without the need to retrain the model. For instance, in scenarios where there is a low tolerance for factual errors, we can set the knowledge expression confidence level to 0.1. This setting prompts the model to answer only those questions it is particularly certain about, thereby ensuring truthfulness. Conversely, in situations where there is a higher tolerance for factual errors, we can adjust the knowledge expression confidence level to 1.0. This adjustment encourages the model to respond to a wider range of questions, enhancing its helpfulness. We show the comparison between Idk-HIR and Idk-SFT in Appendix C.5.

B.6. Calculation of Metrics

To calculate these metrics, we categorize the inference results into four knowledge quadrants using the following method.

- IK-IK: If a question model does not refuse to answer and the answer is correct, then the question belongs to the Ik-Ik category. We determine whether the model's answer is correct by checking if the ground truth appears in the model's response.
- IK-IDK: If a question model refuses to answer, and the question is marked as one that the model does not know, then this question belongs to Ik-Idk category. We determine whether the model refuses to answer a question by checking whether the refusal template appears in the model's response.
- IDK-IK: If a question model refuses to answer, but the question is not marked as one the model does not know, then this question falls into the Idk-Ik category.
- IDK-IDK: If a question model does not refuse to answer but provides an incorrect response, then the question belongs to the Idk-Idk category.

C. Additional Experimental Results

C.1. Knowledge Quadrants Under Different Ik Thresholds

In Figure 7, we present the distribution of the model's knowledge quadrants after Idk-SFT when the Ik threshold ranges from 0.1 to 0.9.

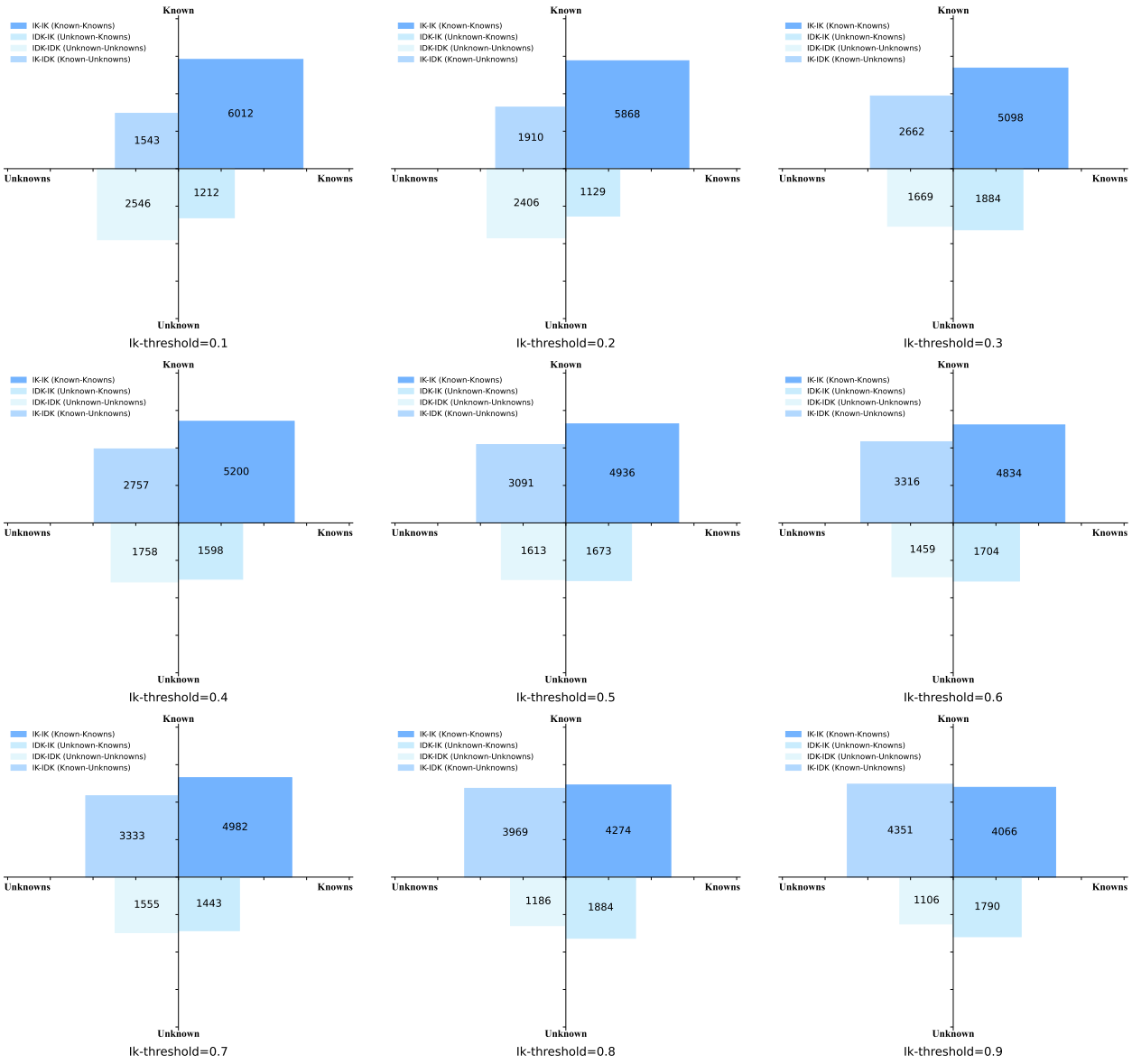

Figure 7. Knowledge quadrants under different Ik thresholds.

C.2. Overall Results of All Knowledge Quadrants

We present the overall results of all knowledge quadrants here.

Table 4. Overall results of all knowledge quadrants on TriviaQA.

	TriviaQA				
	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-Dataset$_{test}$	45.05	54.95	0.00	0.00	100.00
Idk-Prompting	37.36	29.58	13.75	19.31	66.93
Idk-SFT	28.57	46.19	19.24	6.00	74.75$_{\uparrow 7.82}$
Idk-DPO	**39.30**	38.59	**10.01**	12.10	77.89$_{\uparrow 10.96}$
Idk-BoN$_{N=10}$	38.37	40.59	11.53	9.51	**78.96**$_{\uparrow 12.03}$
Idk-PPO	35.90	40.57	13.85	9.68	76.47$_{\uparrow 9.54}$
Idk-HIR	27.36	**48.55**	20.35	**5.66**	75.91$_{\uparrow 8.98}$

Table 5. Overall results of all knowledge quadrants on Natural Questions.

	Natural Questions				
	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-Dataset$_{test}$	24.65	75.35	0.0	0.0	100.00
Idk-Prompting	19.75	41.72	9.75	28.78	61.47
Idk-SFT	15.93	53.99	12.38	17.70	69.92$_{\uparrow 8.45}$
Idk-DPO	20.91	45.60	8.48	25.01	66.51$_{\uparrow 5.04}$
Idk-BoN$_{N=10}$	20.55	47.40	8.81	23.24	67.95$_{\uparrow 6.48}$
Idk-PPO	**23.13**	42.08	**7.34**	27.45	65.21$_{\uparrow 3.47}$
Idk-HIR	15.40	**56.90**	13.38	**14.32**	**72.30**$_{\uparrow 10.83}$

C.3. Results on Resampled Natural Questions

We downsample the Idk questions in the NQ dataset by randomly discarding some Idk questions, making the ratio of Ik questions to Idk questions in the dataset consistent with that in TriviaQA (45.05 : 54.95). Then we recalculate the IK-IK rate, IK-IDK rate, and Truthful rate. As shown in 6, after adjusting the question ratio in NQ, the performance of each method on both datasets became more consistent, with Idk-BoN achieving the highest Truthful rate.

Table 6. Overall results of all knowledge quadrants on Resampled Natural Questions.

	Natural Questions				
	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-Dataset$_{test}$	45.05	54.95	0.0	0.0	100.00
Idk-Prompting	30.41	29.81	17.81	21.96	60.22
Idk-SFT	24.85	38.06	22.62	14.47	62.90$_{\uparrow 2.68}$
Idk-DPO	31.48	32.19	15.49	20.85	63.66$_{\uparrow 3.44}$
Idk-BoN$_{N=10}$	31.58	33.76	16.09	18.57	**65.33**$_{\uparrow 5.11}$
Idk-PPO	**34.87**	29.55	**13.41**	22.17	64.42$_{\uparrow 4.2}$
Idk-HIR	24.19	**40.44**	24.44	**10.93**	64.63$_{\uparrow 4.41}$

C.4. Effect of Ik Threshold

Answer F1 and Refusal F1. We report Answer F1 score and Refusal F1 score of different Idk-SFT models to reflect changes in the model's behavior influenced by the Ik threshold. Regarding Answer F1, we only consider whether the model answer the question, without taking into account the accuracy of the answer.

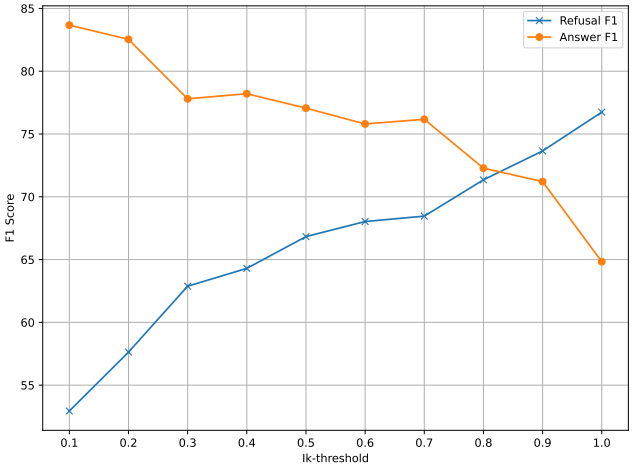

Figure 8. Refusal F1 and Answer F1 scores at different Ik thresholds after Idk-SFT.

As shown in Figure 8, when the Ik threshold raises, the model tends to refuse to answer questions, resulting in an increase in Refusal F1. Conversely, when the Ik threshold is low, the model in more inclined to answer questions, leading to an increase in Answer F1.

C.5. Idk-HIR vs Idk-SFT

In this section, we compare the effects of Idk-HIR and Idk-SFT.

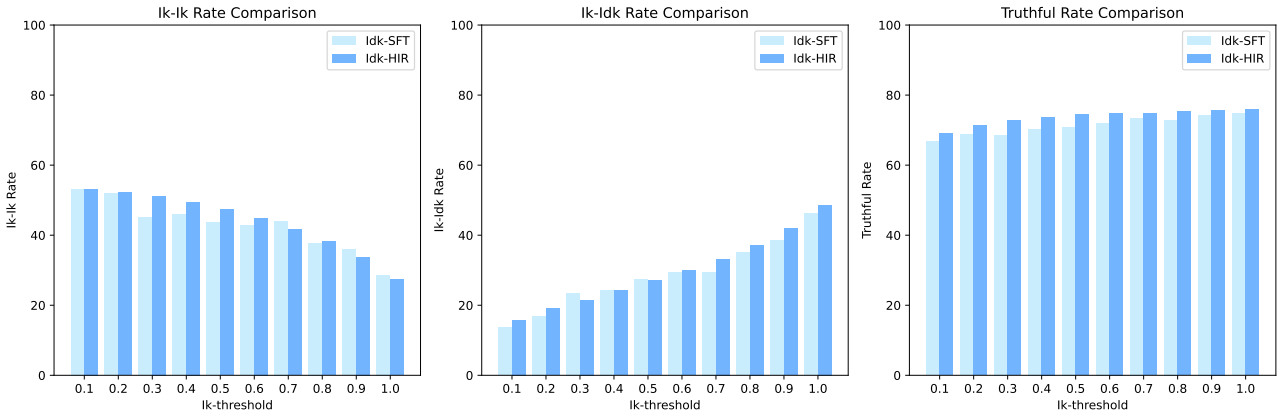

Figure 9. Comparison between Idk-SFT and Idk-HIR.

As shown in Figure 9, the IK-IK rate and IK-IDK rate of the Idk-HIR model are comparable to those of the Idk-SFT model across various Ik thresholds, and the TRUTHFUL rate is consistently higher than that of the Idk-SFT. Therefore, in certain scenarios, the flexible and controllable Idk-HIR model serves as an excellent alternative to the Idk-SFT model.

C.6. Comparisons with Logits-based Baselines

We follow the settings in Kadavath et al. (2022), using our Idk dataset (Ik-threshold=1), and conduct three experiments on Llama2-7b-chat to investigate the effect of the external classifier based on logits.

Using logits from the multiple-choice task. According to the conclusions in Kadavath et al. (2022), using a multiple-choice format can lead to better calibration. In Kadavath et al. (2022), to evaluate the model's calibration, the authors change the question-answering pairs into True or False questions, and then let the model judge whether the given answer is correct. In our experiments, we need the model to directly judge whether it knows the answer to a question, so we modify their prompt, turning the question into a yes or no question about whether it knows the answer. Our modified prompt is as follows:

```
[INST] Questions: Who was the first president of the United States?
Do you know the answer of this question:
A Yes
B No
[/INST] My choice is
```

[INST] and [/INST] are the special tokens of Llama2-7b-chat. We will choose the model's final answer based on the probabilities of A and B in prediction of the next token. We select the option with the highest probability of becoming the next token from A and B as the final answer. The model choosing B represents a refusal to answer. We use "mc_logits" to refer to this method.

Training logits of an additional value head. Following (Kadavath et al., 2022), we add an additional value head (nn.Linear(hidden_size, 2)) to llama2-7b-chat to classify whether the model knows the given question. We use our Idk dataset (Ik-threshold=1.0) to train this binary classification task. The model needs to determine whether a question is an Ik question or an Idk question. During training, we freeze other parameters except for the new value head. We use "value_head" to refer to this method.

Training the whole LLM for classification. In addition to training only the newly added value head, we also attempt to train the entire Llama2-7b-chat for this classification task (we don't freeze any parameters). We use the same training hyperparameters as 2 for a fair comparison. We use "whole_llm" to refer to this method.

Table 7. Comparisons with logits-based baselines.

	TriviaQA				
	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-Prompting	37.36	29.58	13.75	19.31	66.93
Idk-SFT	28.57	46.19	19.24	6.00	$74.75_{\uparrow 7.82}$
Idk-DPO	39.30	38.59	10.01	12.10	$77.89_{\uparrow 10.96}$
mc_logits	40.01	22.65	14.30	23.04	$62.65_{\downarrow 4.28}$
value_head	35.44	39.83	15.09	9.64	$75.27_{\uparrow 8.34}$
whole_llm	38.18	42.65	12.06	7.12	$80.83_{\uparrow 13.90}$

Experimental results. The three methods here essentially belong to independent classifiers. During evaluation, we use the above three methods to classify questions. If the classification result is that the model does not know the answer to the question, then we consider the model refuse to answer the question. Otherwise, we allow Llama2-7b-chat to generate responses in a normal manner using greedy decoding. We compare these three logits-based methods with Idk-Prompting, Idk-SFT, and Idk-DPO in Table 7. The experimental results indicate that compared to methods based on fine-tuning, methods relying on calibration (mc_logits) present more IDK-IDK questions, and the overall Truthful rate is lower than Idk-Prompting. However, training an additional value head can achieve a Truthful rate comparable to that of IDK-SFT but lower than preference optimization method like Idk-DPO. Furthermore, training a complete LLM for classification can achieve good classification performance, but the downside is that it introduces an amount of extra parameters equivalent to the original AI assistant (Llama2-7b-chat).