Know the Unknown: An Uncertainty-Sensitive Method for LLM Instruction Tuning

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate remarkable capabilities but face challenges from hallucinations, which typically arise from insufficient knowledge or context. While instructing LLMs to acknowledge knowledge limitations by responding with *"I don't know"* appears promising, we find that models consistently struggle with admitting knowledge gaps. This challenge may originate from current instruction datasets that emphasise answer generation over knowledge boundary awareness. To address this limitation, we introduce **U**ncertainty-and-**S**ensitivity-Aware **Tuning** (**US-Tuning**), a novel two-stage approach for contextual question answering (QA). The first stage enhances LLMs' ability to recognise their knowledge boundaries, while the second stage reinforces instruction adherence through carefully designed causal prompts. Our experimental results demonstrate that US-Tuning not only significantly reduces incorrect answers in contextual QA but also improves models' faithfulness to their parametric knowledge, mitigating hallucinations in general QA tasks. Our fine-tuned Llama2-7B model achieves up to a 34.7% improvement in handling out-of-knowledge questions and outperforms GPT-4 by 4.2% in overall performance.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks (Brown et al., 2020; Wei et al., 2022; Joshi et al., 2017). Despite their impressive performance, these models face significant challenges that limit their reliable deployment in real-world applications. One of the most critical challenges is hallucination, the tendency to generate factually incorrect or non-sensical content (Maynez et al., 2020). This phenomenon occurs when LLMs generate outputs that either contradict the input context or introduce factually unsupported claims (Ji et al., 2023; Ye et al., 2023). The root cause of this behaviour lies in the inherent limitations in how these models learn and store knowledge during training. Specifically, LLMs encode extensive knowledge from training corpora, this knowledge is inherently incomplete and outdated. When encountering queries that require information beyond their knowledge, these models often resort to generating plausible but factually incorrect responses (Huang et al., 2024a).

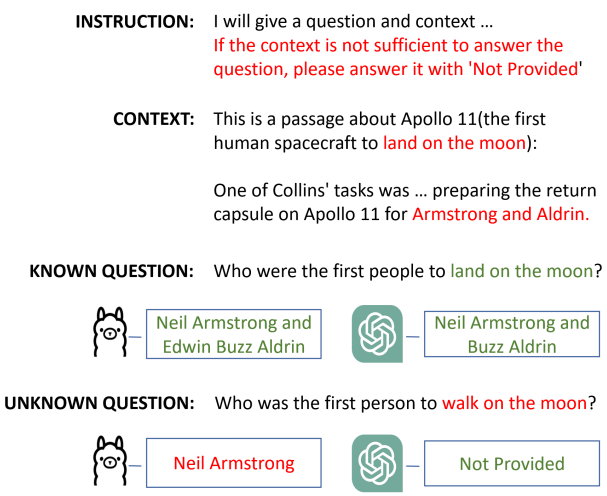

Figure 1: The intention of this paper is to address the inability of LLMs to recognise uncertain answers. We categorise questions into two types: **Known Questions**, which have specific answers, and **Unknown Questions**, which fall outside the provided context.

To solve this question, two approaches have emerged. The first involves further fine-tuning models with additional knowledge (Liu et al., 2023b; Gao et al., 2023; Liu et al., 2023a), while the second leverages retrieval-augmented generation techniques to incorporate external databases (Es et al., 2023). However, as demonstrated in Fig. 1, these approaches still struggle with unknown queries in real-world applications, often producing incorrect answers. Recent work suggests that LLMs should be capable of acknowledging their

knowledge limitations by explicitly stating "I don't know" when applicable (Cole et al., 2023; Yu et al., 2024a). However, there are two major challenges to this goal. First, current instruction datasets predominantly train LLMs to provide definitive answers, inadvertently discouraging models from recognising and expressing uncertainty—defined here as a model's awareness of knowledge beyond its training boundaries (Zhang et al., 2024). Second, models explicitly optimised for uncertainty recognition often exhibit degraded performance in zero-shot question answering (QA) (Kasai et al., 2023; Li et al., 2023a; Si et al., 2023). A fundamental barrier to addressing these challenges is the lack of high-quality datasets containing unknown questions for training and evaluation. Thus, in this work, we focus on constructing contextual QA training data, including a scenario where the provided context is intentionally insufficient. We prioritise this approach over regulating parametric knowledge due to its greater impact on reasoning processes (Huang et al., 2024b).

Our dataset development is motivated by research showing that subtle discrepancies between available knowledge and questions can trigger hallucinations (Shuster et al., 2021). Building on the ASQA data set (Stelmakh et al., 2022), we create a balanced collection of both in-context (known) and out-of-context (unknown) questions. For the latter, we deliberately introduce minor inconsistencies in the context, such as mismatched dates or objects, while maintaining overall contextual coherence. Unlike previous works (Li et al., 2022; Chen et al., 2023), these subtle discrepancies are particularly effective in exposing the tendency of LLMs to hallucinate, making our data set especially valuable for evaluating model performance.

To enhance LLMs' capability to know the unknown and reject uncertain answers, we introduce a novel training framework termed Uncertainty-and-Sensitivity-Aware Tuning (US-Tuning). This approach contains a two-stage training process designed to balance the trade-off between uncertainty recognition and zero-shot instruction adherence. By doing so, it enhances the ability to identify and acknowledge uncertainty while preserving its original QA performance. In the first stage, we focus on awareness of uncertainty, guiding LLMs to effectively identify questions outside the knowledge boundaries. The second stageemphasisess the sensitivity of the instruction, teaching the model to reject answering unknown questions and restoring the compromised QA performance through additional fine-tuning.

Our approach addresses several fundamental challenges in developing uncertainty-aware language models for question-answering tasks. The primary challenge lies in the delicate balance between admitting the knowledge boundary and general QA performance—models that are overly sensitive to uncertainty often experience significant degradation in their ability to answer standard questions. Additionally, when fine-tuning uncertainty-aware models on conventional QA datasets, which contain questions with supporting evidence, models frequently lose their ability to effectively recognise and reject unknown queries. We attribute this degradation to the model's weak sensitivity to uncertain instructions and address it through carefully designed causal instructions in our approach.

Experimental results demonstrate that US-Tuning significantly improves the performance of prevalent LLMs in acknowledging the unknown. Notably, it achieves a 34.7% improvement in addressing unknown questions and surpasses GPT-4 (OpenAI, 2023) with an overall performance increase of up to 4.2%. Furthermore, it not only reduces the frequency of incorrect answers in contextual QA but also encourages LLMs to remain faithful to their parametric knowledge, thereby mitigating hallucinations across various benchmark assessments. Our key contributions are as follows:

- We construct a novel dataset and benchmark for uncertainty recognition, enabling the evaluation of the models' awareness of knowledge gaps.
- We investigate why LLMs tuned to prioritise uncertainty fail to adhere to essential instructions, attributing this behaviour to their weak sensitivity to uncertain prompts.
- We propose a novel two-stage fine-tuning paradigm for instructing the model to remain faithful to the context and reject unknown questions while exploring the relationship between faithfulness and hallucinations.

2 Related Work

In this section, we analyse the former works about hallucinations and instruction datasets for training.

2.1 Uncertainty in Hallucinations

Although the large language models (LLMs) have demonstrated strong performance in downstream tasks by generalising and leveraging encoded

knowledge within the parameters (Liu and Demberg, 2023; Zhang et al., 2023), the uncertainty of such knowledge can also mislead models to generate untrustworthy outputs (Yu et al., 2023; Ye et al., 2023; Manakul et al., 2023). Generally, the uncertainty comes from training data and overestimation (Zhang et al., 2024). Research shows that models tend to mimic the output in the training set (Kang and Hashimoto, 2020), leading to hallucinations that generate reasonable answers for insufficient question-context pairs. Furthermore, models could be overconfident in their capacities and fail to identify unknown questions (Yin et al., 2023; Ren et al., 2023; Kadavath et al., 2022).

There are studies focusing on uncertainty measurement to mitigate hallucinations. Lu et al. (2023) conclude that a correlation exists between the uncertainty and the accuracy. CAD (Shi et al., 2023) proposes a contrastive method for measuring the uncertainty of generated knowledge, restricting models to be context-awarded by amplifying output probabilities when the context is provided. SelfCheckGPT (Manakul et al., 2023) utilises sampling to identify and exclude uncertain information.

2.2 Faithfulness to the External Knowledge

Hallucination is defined as generations that are nonsensical or unfaithful to the provided source content (Ji et al., 2023; Filippova, 2020), encompassing both context and paremetic knowledge. While most prior research has concentrated on the model's faithfulness to parametric knowledge, the aspect of contextual faithfulness as a specific and significant form of hallucination has received comparatively less attention. This gap is underscored by findings indicating that the incorporation of up-to-date and relevant knowledge within prompts can effectively mitigate fact-conflicting hallucinations (Zhou et al., 2023; Liu et al., 2022). However, these studies (Vu et al., 2023; Lewis et al., 2020) operate under the assumption that the given context is always sufficient for generating accurate answers. To address this limitation, various approaches utilise LLMs for post-generation detection (Shen et al., 2023) or editing (Chen et al., 2023) to ensure the faithfulness and consistency of the generated responses with the provided contexts. Self-RAG (Asai et al., 2023) leverages LLMs to screen the provided context, avoiding the disruptions of irrelevant information. However, models struggle to accurately determine whether the provided knowledge is sufficient for answering, especially when the domains of query and context exhibit similarities. Furthermore, some research suggests that reliance on 'unknown' external knowledge can significantly impair performance, potentially exacerbating hallucinations (Lee et al., 2024). Thus, there is a pressing need for an LLM capable of knowing the 'unknown'.

2.3 Instruction Dataset for Training

Aligning LLMs necessitates substantial training data, prompting a trend toward synthesising instruction data to enhance performance. Self-Instruct (Wang et al., 2023) proposes generating diverse instructions using ChatGPT. To improve the query's complexity in different dimensions, WizardLM (Xu et al., 2023) uses five prompts, including depth search and with search. Conversely, AttrPrompt (Yu et al., 2024b) generates various instructions from a feature perspective without relying on class-conditional prompts. Most existing methods concentrate on improving answer quality by exploring a variety of questions with definitive answers, rather than addressing where answers are uncertain. Recent research (Zhang et al., 2024; Cole et al., 2023) has led LLMs to reject unknown questions. R-Tuning (Zhang et al., 2024), for example, trains models to recognise their knowledge limits and to respond with "I don't know". However, identifying the boundaries of parametric knowledge remains challenging due to factors such as latent space compression and hallucination. Therefore, in this study, we build a dataset based on contextual question answering and propose a two-step training method that enables models to reject unknown questions while preserving performance in other tasks.

3 Uncertainty-and-Sensitivity-Aware Tuning

Our research centres on the open-book contextual question-answering (QA), which aims to generate an answer a based on three inputs: i_t, q, and c. Here, i_t denotes the task instructions, q represents the question, and c refers to the provided context. The generation process G can be formulated as:

$$a = G(i_t, q, c)$$

To induce the model to analyse uncertainty, we will implement two explicit constraints. First, we instruct the model not to utilise knowledge beyond the context by stating in i_{task}: *"Your answer must not use any additional knowledge that is not mentioned in the given contexts"*. Second, we require

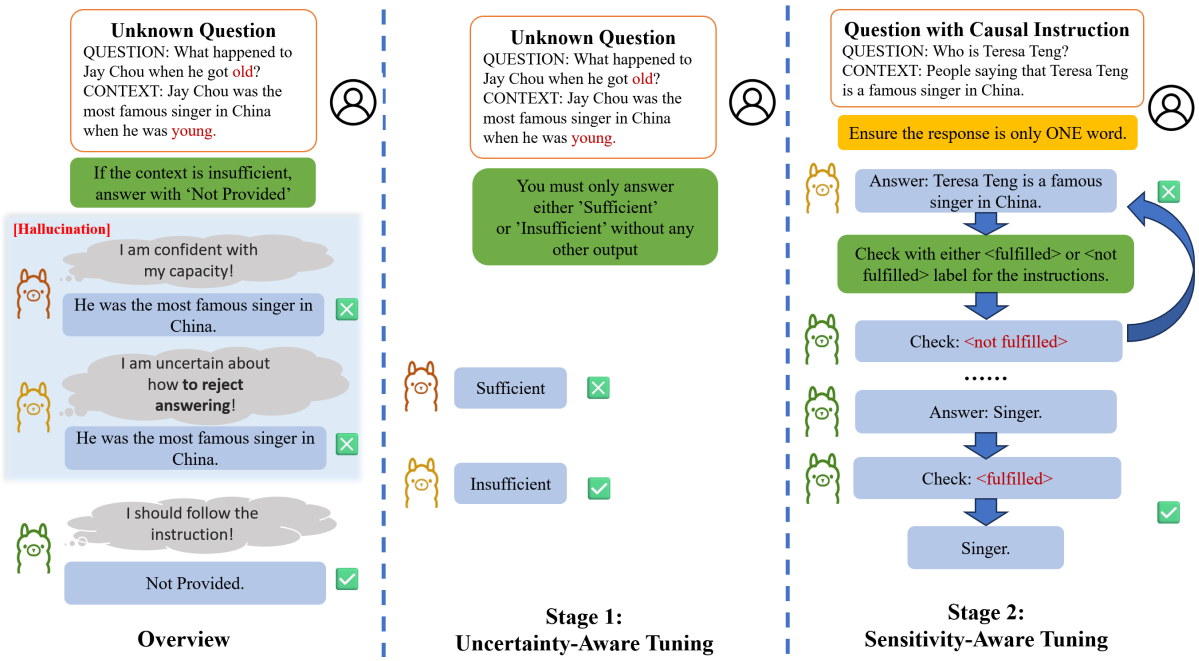

Figure 2: Illustration of our US-Tuning. The green dialog boxes represent task-oriented instructions, while the yellow box indicates additional causal instructions influencing the output. **Overview:** The models include the vanilla model, the Uncertainty-Aware Tuned (UT) model, and the Sensitivity-Aware Tuned (ST) model. We highlight that hallucinations stem from weak cognition of uncertainty and ignorance of instructions. **UT (Stage 1):** teaching the model to know the unknown. **ST (Stage 2):** instructing the model to effectively follow provided instructions.

the model to reject uncertain answers with the directive: *"If the context is not sufficient to answer the question, please answer it with 'Not Provided'"*. This process relies on the model G to evaluate whether the context c is adequate to answer the question q. Based on this assessment, G either generate an appropriate response (a) or acknowledge the insufficiency of c.

3.1 Motivation

As demonstrated in Table 1, our benchmark indicates that vanilla large language models (LLMs) exhibit limited efficacy in rejecting questions beyond their knowledge boundaries. Through systematic experimentation, we identify two core challenges underlying this limitation. First, models frequently generate speculative answers to satisfy perceived user expectations, attributable to standard QA training paradigms that prioritise definitive responses over uncertainty acknowledgement. Second, models fine-tuned for uncertainty recognition demonstrate weakened adherence to the zero-shot instructions, creating a trade-off between rejecting unknown questions and generalisable instruction-following capabilities. This trade-off arises from the scarcity of highly confusing unknown question-context pairs. To preserve the integrity of these rare but critical samples, we avoid direct fine-tuning on unknown questions. Instead, our proposed two-stage training framework addresses these challenges synergistically. The first stage emphasises training the model to identify and reject uncertain questions, thereby preventing inaccurate responses. The second stage involves a systematic instruction review process with answer refinement, contrasting conventional QA tuning by emphasising instruction adherence in response generation.

3.2 Stage 1: Uncertainty-Aware Tuning (UT)

The first stage fine-tunes the model to accurately recognise its knowledge boundaries and identify the known questions. To safeguard the ground truth in the benchmark, we formalise this task as a binary classification problem, as shown in Figure 2. Questions are categorised into two groups: known questions and unknown questions. Known questions are defined as queries with sufficient contextual support to yield accurate answers. Conversely, unknown questions are characterised by lacking adequate contextual information, often exhibiting subtle differences from the query. The model learns to evaluate contextual adequacy and classify its confidence as either "Sufficient" or "Insufficient" for response generation.

Formally, given a contextual QA dataset $D = \{(q_i, c_i), (q_i, c'_i)\}_{i=1}^{n}$ comprising n known question-context pairs and n unknown pairs, we fine-tune the LLM to perform binary classification, where responses are restricted to two categories: "Sufficient" and "Insufficient." The instruction for tuning is recorded in Appx. B.2.

3.3 Stage 2: Sensitivity-Aware Tuning (ST)

Although UT enables models to delineate knowledge boundaries and reject unanswerable queries, Table 1 reveals two critical challenges. First, UT-trained models exhibit heightened uncertainty sensitivity, which affects their ability to answer known questions with confidence. Second, conventional QA tuning exacerbates the model's inability to reject unknown questions, as UT reduces sensitivity to uncertain instructions. We hypothesise that this stems from a conflict in objective alignment: instructions for rejecting unknowns (applicable only to out-of-distribution queries) are not effective on the training data. Consequently, enforcing these instructions during evaluation introduces a misalignment between uncertainty recognition and instruction adherence, degrading overall performance.

To address this, our proposed ST is motivated by explicitly distinguishing the instructions into causal and non-causal ones.

- **Causal** instructions directly affect the response content, whereas non-causal instructions provide auxiliary guidance without affecting answer semantics. For example, instructions that constrain the format or tense of responses serve as typical causal ones. Conversely, extra instructions, such as *"answering with 'Not Provided' if the context is insufficient"*, function as non-causal instructions when fine-tuning known questions, as they do not contribute directly to the answer.

- **Non-causal** instructions risk being disregarded, despite their critical importance to the overall task.

Our ST is designed to enhance the model's sensitivity and adherence to all instructions by ensuring that even non-causal instructions are prioritised. As shown in Fig. 2, it comprises two synergistic components: additional causal instructions and instruction review synthesis.

Causal Instruction Synthesis: By instructing GPT-4 to produce controlling conditions that directly influence response properties, such as tense, length, or output format, we obtain additional causal instructions. These causal instructions are then randomly integrated into the original QA prompts, ensuring the model learns to prioritise and comply with diverse task requirements. The prompt for generation is presented in Appx. B.3.

Review Instruction Synthesis: The instruction review module employs the model itself to verify the fulfilment of all instructions. The model will recursively regenerate until it gets a perfect answer by utilising the prompts in Appx. B.4. The process of the instruction review is illustrated in Algo. 1.

As shown in Fig. 2, given a question-answering dataset $\{(q_1, c_1), ...(q_n, c_n)\}$ and additional causal instructions, the entire process is formulated as $a = R(G(i_t + i_c, q, c))$, where i_c is a randomly selected casual instruction and i_t is the original task description. R is the loop function for instruction review. We employ GPT-4 and record the conversation from the loop to fine-tune the smaller model.

4 Experiments

In this section, we describe the data construction and the associated experiments. Table 1 shows that the suboptimal performance of LLMs in rejecting unknown questions can be attributed to two primary factors: weak uncertainty-recognition capacity and the instruction-sensitivity reduction. We assess the effectiveness of US-Tuning using prevalent LLMs on our proposed benchmark, as well as on traditional QA hallucination benchmarks.

4.1 Data Construction

We create a benchmark that balances known and unknown questions for evaluation, along with two specific datasets designed for US-Tuning.

Uncertainty-Recognition Benchmark To comprehensively evaluate the model's cognitive ability to identify knowledge gaps, we construct a test dataset using the ASQA (Stelmakh et al., 2022) dataset, which consists of ambiguous questions. Each question is divided into multiple sub-questions with their corresponding contexts. For example, as recorded in Appx. A.10, one pair may discuss the discovery of the photoelectric effect in 1887, while another may cover the theoretical development in 1905. To generate the unknown questions, we shuffle these pairs, reassigning the questions to different but related contexts. As a result, there are two significant advancements in our benchmark. First, the context is closely rele-

vant to the query, featuring partial mismatches in dates or objects, thereby challenging the model's ability to handle uncertainty. Second, the context is definitely insufficient for the query. Such samples are rare and valuable, as ASQA is the only dataset we have found that could yield sufficient samples that satisfy the requirement. We generate 3,320 known questions and 3,320 unknown questions to construct our benchmark.

In the evaluation, we design the QA template for uncertainty recognition by instructing the model to reject unknown questions, as presented:

- QA Uncertainty-Recognition: *If the context is not sufficient to answer the question, please answer it with 'Not Provided'.*

US-Tuning Datasets Two distinct instruction datasets are used for separate stages. For the UT, we construct a binary dataset comprising 646 samples from the ASQA (Stelmakh et al., 2022) with the ground truth concealed to prevent overlap with the evaluation data. Here is a demonstration of the prompt we used for tuning on this dataset:

- Uncertainty-Aware Tuning: *You must only answer either 'Sufficient' or 'Insufficient' without any other output*

To protect our valuable benchmark, the dataset for ST is derived from HotpotQA (Yang et al., 2018), a dataset designed for multi-hop QA. We generate causal instructions using GPT-4 (OpenAI, 2023) and manually select the 28 most robust instructions, as listed in Appx. C. These instructions were then integrated into 300 randomly selected samples from HotpotQA. Subsequently, we utilised GPT-4, following the methodology outlined in Section 3.3, to synthesise the final ST dataset.

4.2 Experiment Setting

Training Details. We evaluate our US-Tuning on prevalent open-sourced LLMs, including Llama2-7B-Chat (Touvron et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Gemma-2-9B-Instruct (Team et al., 2024). We also test GPT-4-1106-preview (OpenAI, 2023), GPT-3.5 Turbo (OpenAI, 2023), Vicuna-7B v1.5 (Zheng et al., 2024a) and Self-RAG-7B (Asai et al., 2023) on our benchmark. Our fine-tuning bases on an RTX3090 GPU in conjunction with LLaMA-Factory (Zheng et al., 2024b), with Lora (Hu et al., 2021) in a rank of 8, a batch size of 4, and a learning rate of 5e-5. We configured the epochs to 1 and 5 for the two stages, respectively. This research integrates the instruction-based and attributed prompts, which demonstrate to effectively mitigate hallucinations (Zhou et al., 2023), as provided in Appx. B.

Evaluation Metric. We use Acc_{known} for representing the accuracy of questions with specific answers, and $Acc_{unknown}$ for unknown questions.

Benchmark Result. As summarised in Table 1, our analysis (Appx. A.3) reveals that prevalent LLMs struggle to reliably identify unknown questions, achieving modest accuracy rates of 60%.

4.3 Analysis

4.3.1 Weak Uncertainty-Recognition Capacity

Tables 1 and 7 reveal a persistent performance gap of up to 21.0% between known and unknown questions for Llama2, indicating the challenge associated with models' capacity to recognise uncertainty. By leveraging uncertainty-aware tuning (UT), as evidenced in Table 1, there is a notable improvement of up to 26.1% in the accuracy of responses to unknown questions ($Acc_{unknown}$), surpassing baseline performances and being comparable to GPT-4. However, this increased awareness of uncertainty leads to a decrease in the QA capability. Specifically, models demonstrate an excessive sensitivity to the varied phrasing of similar questions.

4.3.2 Instruction-Sensitivity Reduction Problem

According to Table 1, further fine-tuning on HotpotQA results in a degradation in the model's ability to reject unknown questions, primarily due to a decline in its adherence to instructions. This is evidenced by a low $Acc_{unknown}$ of 20.9%, despite the uncertainty recognition capacity being maintained at 66.7% (Table 7). We term this phenomenon the "instruction-sensitivity reduction problem."

As shown in Tables 1 and 7, UT equips the model with the ability to recognise and reject uncertain questions. However, the absence of unknown questions in HotpotQA means that the instruction to reject uncertain answers is never effectively implemented during training. This creates a conflict that adherencing to zero-shot instructions can inadvertently increase uncertainty, counteracting the objectives of UT and diminishing performance. Consequently, the model often disregards instruction constraints, generating hallucinated answers for unknown questions. Our proposed ST (US-Tuning in Table 1) addresses this issue by ensuring adherence

Category	Model	QA Uncertainty-Recognition Acc_{known}	$Acc_{unknown}$	$F1$
Benchmark	**GPT-4**	79.6	**83.6**	**81.6**
	GPT-3.5	**82.1**	51.8	63.5
	Vicuna-7B v1.5	74.6	43.8	55.2
	Self-RAG-7B	67.9	48.1	56.3
Llama2	Vanilla	79.3	58.3	67.2
	UT (Stage 1)	52.4	84.4	64.6
	UT+HotpotQA	77.0	20.9	32.8
	US-Tuning	**79.7**	**93.0**	**85.8**
Mistral	Vanilla	85.1	63.0	72.4
	UT (Stage 1)	77.5	**75.8**	76.6
	UT+HotpotQA	87.1	52.4	65.5
	US-Tuning	**87.3**	75.3	**80.9**
Gemma	Vanilla	86.1	74.1	73.5
	UT (Stage 1)	76.1	**86.2**	**80.8**
	UT+HotpotQA	**91.3**	20.8	33.9
	US-Tuning	87.6	81.2	**84.3**

Table 1: Results (in %) for prevalent LLMs on QA uncertainty-recognition benchmark. The overall best results for each category are highlighted in **bold**. Results that are more than 5% higher or lower than the baseline are highlighted in green and orange, respectively.

to all instructions, bridging the gap between uncertainty recognition and instruction compliance.

4.4 Effectiveness on Contextual QA

Among the models tested on our benchmark, the US-Tuned Llama2 ranks the highest, achieving an F1 score of 85.8%, which surpasses GPT-4 by 4.2% and exceeds the baseline by 18.6% (as shown in Table 1). This impressive performance can be attributed to the model's optimal balance between uncertainty recognition and adherence to zero-shot instructions. Notably, it achieves a remarkable 93.0% accuracy on unknown questions, the highest among prevalent LLMs, while maintaining a 79.7% accuracy on known questions. Additionally, Gemma-2 and Mistral exhibit improvements of 13.1% and 5.6%, respectively, highlighting the robustness and effectiveness of our US-Tuning approach in enhancing performance among prevalent LLMs. This tuning method effectively mitigates the risk of generating incorrect answers without compromising the original question-answering capabilities.

Model	**US-Tuning (ours)** Cor.	Wro.	Unk.	Vanilla Cor.	Wro.	Unk.
GPT-4	-	-	-	79.6	4.4	16.0
Llama2	79.7	**1.4**	18.9	79.2	8.5	12.2
Mistral	87.3	**2.5**	10.2	85.1	5.1	9.8
Gemma	87.6	**1.6**	10.8	86.1	3.9	10.0

Table 2: The portions of correct, wrong, and unknown responses among the responses for known questions.

Our model effectively supports high-stakes decision-making. For unknown questions, in addition to the significantly increased $Acc_{unknown}$, the case study in Appx. A.9 demonstrates that our model prioritises uncertainty analysis, acknowledging limitations rather than hallucinating responses. For known questions, Table 2 presents a detailed distribution of responses. The data indicate that US-Tuning substantially reduces the occurrence of wrong answers by up to 7.1%, albeit with a modest increase in the proportion of unknown responses.

4.5 Comparison with SOTA Approaches

We evaluate our method against SOTA approaches within our uncertainty-recognition benchmark, as detailed in Appx. A.4. Table 3 shows that Honesty (Yang et al., 2023) and Calibration (Kapoor et al., 2024), which target noncontextual QA tasks, face significant instruction-sensitivity reduction, evidenced by the low $Acc_{unknown}$. Despite being fine-tuned with unknown questions, these methods prioritise uncertainty but struggle with uncertain zero-shot instructions related to contextual uncertainty identification. As a result, they exhibit limited robustness in contextual QA. However, when integrated with our proposed ST, as experimented in Appx. A.5, Honesty exhibits significantly improved compliance with instructions and outperforms the baseline. This highlights the effectiveness of our ST in generalising uncertainty recognition capacity across diverse tasks. The results of C-DPO (Bi et al., 2024) indicate that Direct Preference Optimisation (Rafailov et al., 2024) effectively enhances the overall capabilities of the model in both QA and instruction adherence, but a gap persists compared to our tailored method. Additionally, post-generation methods face challenges in recognising unknown questions due to their limited capacity for uncertainty detection.

Category	Method	$Acc_{kno.}$	$Acc_{unk.}$	$F1$
Vanilla	Llama2	79.3	58.3	67.2
Post-Gen.	Validation	82.5	53.8	65.1
	Sampling	79.7	66.5	72.5
Prompt	CFP	**87.4**	47.6	61.6
Tuning	Calibration	67.1	63.2	65.1
	Honesty	74.7	61.1	67.2
	C-DPO	77.7	69.6	73.4
	US-Tuning	79.7	**93.0**	**85.8**

Table 3: Comparison results with SOTA methods on QA uncertainty-recognition benchmark.

4.6 Ablation Study

To further investigate the impact of US-Tuning, we decompose it into three distinct components.

- **UT**: 646 samples for uncertainty-aware tuning.
- **HP**: 300 samples from HotpotQA with QA prompts provided in Appx. B.1.
- **CI**: HP with causal instructions, termed ST.

As illustrated in Table 4 and Fig. 3, our findings indicate that models without UT exhibit a weak capacity for uncertainty recognition, presented by low $Acc_{unknown}$. Furthermore, QA fine-tuning that does not incorporate causal instructions contradicts the objectives of UT, resulting in a decline in $Acc_{unknown}$. In contrast, our ST approach not only enhances performance on known answers, achieving the highest Acc_{known} reported in the table. But also, when effectively integrated with UT, our method attains optimal performance across both known and unknown questions.

Component			QA Uncertainty-Recognition		
UT	HP	CI	Acc_{known}	$Acc_{unknown}$	$F1$
			79.3	58.3	67.2
✓			52.4	84.4	64.6
	✓		77.5	58.3	66.5
	✓	✓	**84.8**	59.0	69.6
✓	✓		77.0	20.9	32.8
✓	✓	✓	79.7	**93.0**	**85.8**

Table 4: Results of ablation on our QA benchmark with significant values highlighted.

4.7 Relationship between Faithfulness and Hallucination

We also conduct the experiment of our approach within a traditional QA setting. To our knowledge, it is the first work to elucidate the relationship between the faithfulness to context and to parametric knowledge (hallucination). R-Tuning (Zhang et al., 2024) preconstructs the tuning datasets to explicitly convey uncertainty for unknown questions, while we directly tune our pre-trained model on raw samples, as detailed in Appx. A.7. According to Table 5, while our US-Tuning shows lower effectiveness compared to the SOTA approaches specifically designed for noncontextual QA tasks, it represents a significant improvement over the vanilla model, with increases of 11.30%, 10.38%, and 6.26% in accuracy, respectively. Our findings indicate that our model can leverage uncertainty recognition as a metacapacity, effectively applying it in both contextual and noncontextual QA scenarios.

Furthermore, CoCoNot (Brahman et al., 2024) provides 302 unknown noncontextual QA pairs and suggests employing GPT-3.5 (OpenAI, 2023) to assess compliance. We test our pre-trained models on a subset of CoCoNot, and our results indicate that US-Tuning can also slightly improve the performance in rejecting noncontextual questions.

Tuning	Model	ParaRel	MMLU	HaluEval
Vanilla	Llama2	43.38	38.56	76.22
NC	Honesty	-	49.28*	**88.11***
	Calibration	-	53.00	87.78
	R-Tuning	**69.54**	**55.56**	77.17
C	**US-Tuning**	54.68	48.94	82.48

* Based on Llama2-13B-Chat (Touvron et al., 2023)

Table 5: Accuracies (%) of SOTA methods separately designed for noncontextual (NC) and contextual (C) QA tasks on QA hallucination detection benchmarks.

Model	Vanilla	Model	Vanilla	**US-Tuning**
GPT-4	92.05	Llama2	94.04	**94.37**
GPT-3.5	77.81	Mistrial	96.36	**96.70**
Vicuna	82.62	Gemma	**95.28**	95.04

Table 6: Compliance rate (%) of prevalent LLMs on the CoCoNot noncontextual unknown QA benchmark.

5 Conclusion

This paper investigates a prevalent issue in large language models (LLMs), where insufficient contextual information results in plausible yet incorrect responses. Our research reveals that LLMs often struggle with unknown questions, primarily due to their limited uncertainty recognition capacity and weak robustness to zero-shot instructions. Notably, tuning the models to focus on uncertainty will adversely weaken adherence to zero-shot instructions. To address these issues, we propose a novel two-stage training framework, termed "uncertainty-and-sensitive-aware tuning." The first stage guides the LLM to identify unknown questions, while the second stage aims to recover diminished question-answering performance through carefully designed causal instructions. This approach enhances the model's reliability and reduces hallucinations. Our methodology distinguishes itself by fine-tuning the uncertainty recognition as a metacapacity, rather than direct training on unknown question samples, thereby enabling effective adaptation across various tasks. By open-sourcing this work, we aim to advance the development of automatic instruction synthesis datasets, emphasising data diversity and the critical reduction of hallucinations.

Limitations

In this study, we identify two key areas for future refinement. First, the LLM encounters a long-tail problem when tuned with datasets that contain a limited number of unknown questions, necessitating further adaptation of our US-Tuning. Second, we have not analysed the parametric knowledge acquired by Llama2 during its pre-training phase, and our fine-tuning dataset may overlap with this pre-training data, potentially affecting performance. To address these challenges, future research will investigate methods for measuring model uncertainty through internal parameter monitoring, as proposed by Lu et al. (2023). By quantifying uncertainty across various inputs, we aim to identify knowledge gaps and long-tail weaknesses, informing targeted fine-tuning strategies to enhance the LLM's performance across diverse queries.

Ethics Statement

The benchmark and datasets utilised in this study are derived from public datasets. Additionally, the US-Tuning dataset incorporates refinements using GPT-4, which may introduce inherent biases. However, the methodologies in this research are designed to avoid introducing any additional biases beyond those already inherent in the datasets.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. 2024. Context-dpo: Aligning language models for context-faithfulness. *arXiv preprint arXiv:2412.15280*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, et al. 2024. The art of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2024b. Enhancing large language models' situated faithfulness to external contexts. *arXiv preprint arXiv:2410.14675*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. 2024. Crafting in-context examples according to lms' parametric knowledge. *Preprint*, arXiv:2311.09579.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Dongqi Liu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. *Preprint*, arXiv:2104.08704.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023b. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4549–4560, New York, NY, USA. Association for Computing Machinery.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2023. How are prompts different in terms of sensitivity? *arXiv preprint arXiv:2311.07230*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

OpenAI. 2023. GPT4 (Nov 7 version). https://chat.openai.com/chat. gpt-4-1106-preview.

OpenAI. 2023. Language models are few-shot learners. gpt-3.5-turbo-1106.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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

Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. 2023. "why is this misleading?": Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1662–1672, New York, NY, USA. Association for Computing Machinery.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *Preprint*, arXiv:2305.18153.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2024a. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1333–1351.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024b. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Supplementary Material

A.1 Algorithm for Instruction Review Module

Here we provide the algorithm chart for the Review Instruction Synthesis in Section 3.3.

Algorithm 1: Instruction Review Module

Data: context c, query q, task instruction i_t, causal instructions i_c

while *not fulfilled* **do**
  answer = generate(c, q, i_t, i_c);
  check = review(answer, i_t, i_c);
  if *"<not fulfilled>" not in check* **then**
    fulfilled = True;
  end
end

A.2 Postfix Uncertainty-Recognition

In addition to the question-answering (QA) Uncertainty-Recognition Benchmark mentioned in Section 4.1, we further develop a postfix template specifically for uncertainty recognition. Different from the QA one, the postfix template emphasises the assessment of uncertainty by evaluating the sufficiency of the responses and generating a tag after the corresponding answer. The prompt template is recorded as follow:

- Postfix Uncertainty-Recognition: *You must append either '<Sufficient>' or '<Insufficient>' after your answer.*

Category	Model	Postfix Uncertainty-Recognition Acc_{known}	$Acc_{unknown}$	$F1$
Benchmark	**GPT-4**	88.9	**78.3**	**83.3**
	GPT-3.5 Turbo	**97.0**	33.4	49.7
	Vicuna-7B v1.5	93.5	14.3	24.8
	Self-RAG-7B	46.0	74.9	57.0
Llama2	Vanilla	85.2	29.5	43.9
	UT (Stage 1)	81.3	**84.0**	**82.6**
	UT+HotpotQA	87.1	66.7	75.5
	US-Tuning	**88.0**	66.0	75.4
Mistral	Vanilla	82.8	43.1	56.7
	UT (Stage 1)	**86.1**	81.9	**84.0**
	UT+HotpotQA	80.7	75.1	77.8
	US-Tuning	82.5	**82.2**	82.4
Gemma	Vanilla	86.3	57.6	69.1
	UT (Stage 1)	93.4	**76.2**	**83.9**
	UT+HotpotQA	**99.4**	58.7	73.8
	US-Tuning	96.1	55.1	70.1

Table 7: Results (in %) for prevalent LLMs on postfix uncertainty-recognition benchmark. The overall best results are highlighted in **bold**. Results that are more than 5% higher or lower than the baseline are highlighted in green and orange, respectively.

Figure 7 presents the evaluation results from our benchmark using the postfix template, focusing solely on the accuracy of the sufficiency tags rather than the correctness of answers. The findings indicate that most prevalent large language models (LLMs) struggle to effectively identify uncertainty. Furthermore, our proposed uncertainty-aware tuning (UT) shows potential to mitigate this challenge.

A.3 Illustration to the Benchmark Results

Table 1 presents the QA performance on our benchmark. Coupled with the uncertainty recognition performance detailed in Table 7, our findings indicate that prevalent LLMs face challenges in accurately identifying unknown questions, achieving only approximately 60% accuracy. Notably, GPT-4 and Gemma-2 achieve higher accuracies of 83.6% and 74.1%, respectively. Mistral and Llama-2 rank highest among the remaining models, surpassing GPT-3.5 despite its larger parameter size. Nevertheless, a significant performance gap persists between GPT-4 and other models. Ongoing experiments aim to explore the underlying factors contributing to this disparity. The analysis further reveals that different models respond differently to insufficient queries. Models fine-tuned on dialogue tasks tend to overly rely on and trust the given information. Self-RAG, which is fine-tuned for QA tasks involving unknown questions, demonstrates a strong ability to identify uncertainty, as indicated in Table 7, but still struggles to acknowledge it.

A.4 Illustration to the State-of-the-Art (SOTA) Methods

Current SOTA research primarily addresses rejection in noncontextual QA tasks, leaving contextual QA underexplored. We categorise SOTA methodologies into post-generation, prompt-based, and tuning methods. Notable tuning approaches for rejecting unknown questions include Honesty-Alignment (Yang et al., 2023) and Calibration-Tuning (Kapoor et al., 2024). They focus on noncontextual QA tasks while tuning for rejecting answering in contextual QA tasks remains unaddressed. C-DPO (Bi et al., 2024) emphasizes model faithfulness to context rather than rejecting unknown questions. Context-Faithful-Prompting (CFP) (Zhou et al., 2023) aims to enhance model fidelity to context through third-person paraphrasing in prompts. Post-generation methods for uncertainty detection include Multi-Sampling (Cole et al., 2023) and LM-Validation (Kadavath et al.,

2022). The sampling method generates three outputs at a temperature of 0.6, selecting the most frequent response, while LM-Validation allows for further refinement of the generation. This study compares these methodologies with our proposed US-Tuning.

Model	Category	Contexutal	Task
Validation	Post-Gen.	Both	Faithfulness
Sampling	Post-Gen.	Both	Faithfulness
CFP	Prompt	Contextual	Faithfulness
Calibration	Tuning	Noncontextual	Rejection
Honesty	Tuning	Noncontextual	Rejection
C-DPO	Tuning	Contextual	Faithfulness
US-Tuning	Tuning	Contextual	Rejection

Table 8: Categories and targeted tasks for the SOTAs.

A.5 Further Ablation Study on the SOTA Method with Sensitivity-Aware Tuning

In Section 4.5, we evaluate the performance of the SOTA methods on our QA uncertainty-recognition benchmark. We attribute the low performance of Honesty-Alignment (Yang et al., 2023) to the instruction-reduction problem, evidenced by an $Acc_{unknown}$ of only 61.1%, despite it being tuned on unknown noncontextual QA samples. In contrast, our US-Tuned Llama2 achieves 93.0%. This section further elucidates the instruction-sensitivity reduction problem by implementing our Sensitivity-Aware Tuning (ST), which aims to enhance the model's sensitivity to constraint instructions alongside the Honesty-Alignment approach.

Method	Acc_{known}	$Acc_{unknown}$	$F1$
Vanilla Llama2	79.3	58.3	67.2
US-Tuning	**79.7**	**93.0**	**85.8**
Honesty	74.7	61.1	67.2
Honesty + ST	**80.4**	**80.8**	**80.6**

Table 9: Results of sensitivity-aware tuned Honesty-Alignment on QA uncertainty-recognition benchmark.

Table 9 yields several key conclusions. First, our proposed ST effectively mitigates the instruction-sensitivity reduction problem, improving the $Acc_{unknown}$ of Honesty-Alignment by 19.7%, resulting in a 13.4% enhancement in overall performance. Second, our initial stage, focused on assessing the sufficiency of the given context relative to the question, outperforms other methods, as demonstrated by a 5.2% improvement of our US-Tuned Llama2 over the Sensitivity-Aware Tuned Honesty-Alignment. This advancement is attributable to both the quality and quantity of the dataset used for ST, enabling the model to recognize knowledge gaps as a metacognitive capacity, as discussed in Section 4.7. Finally, the samples utilized for Sensitivity-Aware Tuned Honesty-Alignment are strictly non-overlapping with our benchmark, further validating the effectiveness and robustness of our proposed US-Tuning approach.

A.6 Figure of Ablation Study

In Section 4.6, we present a comparative analysis of each model configuration. Table 11 details the specific names associated with each setting.

Model	UT	HP	CI
Llama2			
UT	✓		
HotpotQA		✓	
ST		✓	✓
UT & HotpotQA	✓	✓	
US-Tuning	✓	✓	✓

Table 11: Corresponding model name to each setting in the ablation study.

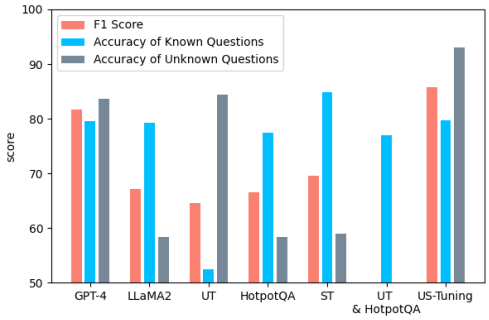

Figure 3: Comparison of different models in the ablation study. A higher score means better performance.

A.7 Experiment Setting for Hallucination Benchmarks

R-Tuning (Zhang et al., 2024) divides the ParaRel (Elazar et al., 2021) and MMLU (Hendrycks et al., 2020) into in-domain and out-of-domain datasets, treating the entire HaluEval (Li et al., 2023b) as an out-of-domain evaluation set. Models are initially fine-tuned on in-domain datasets and subsequently evaluated on out-of-domain datasets. Our fine-tuning is based on Llama2-7B-Chat (Touvron et al., 2023) US-Tuned on ASQA (Stelmakh et al., 2022) and HotpotQA (Yang et al., 2018), as detailed in Section 3. We further fine-tune it on in-domain datasets with LoRA, adhering to the settings in R-Tuning: 1 epoch, a learning rate of 2e-5,

Context: This is a passage about Harry Potter: An epilogue "Nineteen Years Later" (set on 1 September 2017) describes the lives of the surviving characters and the effects of Voldemort's death on the Wizarding World. In the epilogue, Harry and Ginny are married with three children, and Ron and Hermione are married with two children.
Unknown Question: What year were the Harry Potter books first released?
Vanilla Llama2: Sure, I'd be happy to help! Based on the context you provided, the Harry Potter books were first released in the year 1997. The passage states that the epilogue is set on September1, 2017, which is 19 years after the events of the final book in the series, "Harry Potter and the Deathly Hallows," which was published in 1997. Therefore, the Harry Potter books were first released in 1997.
US-Tuned Llama2: Based on the given context, the answer to the question "What year were the Harry Potter books first released?" is: Not Provided. The context does not provide sufficient information to determine the year the Harry Potter books were first released. The passage only mentions the epilogue set in 2017, which is 19 years after the events of the books. Therefore, the year of the books' release is not mentioned or implied in the given context.

Table 10: An illustrative case from our proposed benchmark, along with the corresponding model responses before and after the implementation of US-Tuning.

and a batch size of 4. Similarly to R-Tuning, our evaluation spans several out-of-domain datasets.

CoCoNot (Brahman et al., 2024) offers a comprehensive dataset for benchmarking the performance in rejecting answering. In this study, we utilise a subset containing 302 unknown QA pairs in a non-contextual setting and evaluate the effectiveness of US-Tuning on it. Following the methodology outlined in that paper, we employ GPT-3.5 (OpenAI, 2023) to assess the compliance with the response.

A.8 Zero-Shot Effectiveness on RealtimeQA

RealtimeQA (Kasai et al., 2024) is a dataset designed for high-stakes scenarios that necessitate timely responses, thereby challenging the faithfulness of LLMs to contextual information. Our study utilizes 113 contextual QA pairs from RealtimeQA, of which 50 are unknown pairs. Our benchmark is distinguished by a larger sample size compared to RealtimeQA. We directly implement our pre-trained model without further tuning on RealtimeQA. As shown in Table 12, our model demonstrates significant improvements in addressing unknown questions, underscoring the effectiveness and robustness of our approach.

Method	Acc_{known}	$Acc_{unknown}$	$F1$
Vanilla Llama2	**88.7**	36.7	51.9
US-Tuning	71.8	**56.0**	**62.9**

Table 12: Accuracies (%) on RealtimeQA.

A.9 Case Study

The case provided in Table 10 addresses a key challenge regarding uncertain information. The vanilla Llama2 incorrectly claims that the Harry Potter books were released in 1997, despite the context only referencing an epilogue set in 2017.

In contrast, our US-Tuned Llama2 effectively mitigates this issue by prioritising uncertainty detection. Rather than offering an uncertain answer, it appropriately responds with "Not Provided." This approach not only rejects uncertain responses but also clarifies the source of uncertainty, thereby enhancing the model's reliability. The implementation of US-Tuning is particularly vital in high-stakes fields, such as medicine, where a low wrong answer rate is essential. By refining LLMs' ability to recognise and communicate uncertainty, US-Tuning promotes responsible and trustworthy interactions, ensuring users receive reliable information.

A.10 Example of Constructing Benchmark

Table 13 presents an example that illustrates the construction of our uncertainty-recognition benchmark, as detailed in Section 4.1. In this process, we shuffle the questions and their corresponding contexts to introduce uncertainty, thereby challenging the model's ability to respond to uncertain queries.

B Instructions

In this section, we present an overview of all the prompt templates utilized in this study. Key descriptions are highlighted in red, while blue descriptions are designated for performance adjustments.

B.1 Question Answering

Question Answering Task: You need to do the Question Answering for the following query.

Question: Who discovered and developed an explanationfor the photoelectric effect in 1887?
Positive Context: This is a passage about Photoelectric effect: Light, and especially ultra-violet light, discharges negatively electrified bodies with the production of rays of the samenature as cathode rays. Under certain circumstances it candirectly ionize gases. The first of these phenomena was discovered by Heinrich Hertz and Wilhelm Hallwachs in 1887.The second was announced first by Philipp Lenard in 1900.
Negative Context: This is a passage about Photoelectric effect: In 1905, Einstein proposed an explanation of the photoelectriceffect using a concept first put forward by Max Planck that light waves consist of tiny bundles or packets of energy knownas photons or quanta.

Table 13: An example from ASQA (Stelmakh et al., 2022), where the positive context adequately supports the question, whereas the negative is insufficient.

I will give a question and several contexts. Based on the given contexts, give an answer to the question. Your answer must not using any additional knowledge that is not mentioned in the contexts. If the context is not sufficient to answer the question, please answer it with 'Not Provided'

QUERY: q

CONTEXT: c

ANSWER:

B.2 Uncertainty-Aware Tuning

Cognition Assessment Task: You need to do the Cognition Assessment for the following query.

I will give a query and a related context about the query. Your task is to judge whether the context is sufficient to answer the query.

Assessment: You must append either '<Sufficient>' or '<Insufficient>' after your answer.

Finetuning: You must only answer either 'Sufficient' or 'Insufficient' without any other output.

Here is the example.

QUERY: What happened to Jay when he got old?

CONTEXT: Jay Chou was the most famous singer in China when he was young, releasing many nostalgic albums and songs that are memorable to middle-aged people today.

ANSWER:

Assessment: Jay Chou was the most famous singer in China.<Insufficient>

Finetuning: Insufficient

Here is the provided information that you need to accomplish follow the provided example:

QUERY: q

CONTEXT: c

ANSWER:

B.3 Additional Casual Instruction Generation

Your task is to provide various instructions for the questions answering task.

The questions answering task provides a context and a query. e.g. "Context: XXX Query: XXX Answer:". And your task is to add some specific requirement to the answer. e.g. "The answer must be all in upper case", "There should be no punctuation in the answer". The added instruction should be general to the query. You should generate hundreds of such instructions.

B.4 Sensitivity-Aware Tuning

You should check whether your answer aligned the requirement by generating a Checking part, checking each sentence of the above instruction, with either <fulfilled> or <not fulfilled> mark behind the sentence, indicating whether the requirement is fulfilled or not. If there is <not fulfilled> mark behind the sentence, you must modify your answer again to fulfill the requirement, by appending a new ANSWER and CHECKING part.

Here is an example for this task:

e.g. Question Answering Task Requirements: You need to do the Task Prompt for the following query and context. Ensure the response is written in the past tense.

QUESTION: Who is Jack Chen?

CONTEXTS: People saying that Jack Chen is a famous singer in China.

ANSWER: Jack Chen is a famous singer in China.

CHECKING: Question Answering Task: You need to do the Task Prompt for the following query and context.<fulfilled>Ensure the response is written in the past tense.<not fulfilled>

ANSWER: Jack Chen is a famous singer in China.

CHECKING: Question Answering Task: You need to do the Task Prompt for the following query and context.<fulfilled>Ensure the response is written in the past tense.<fulfilled>

Here is the information of your task:

{Question Answering Instruction}

B.5 Trustworthy Question Answering for Benchmark

Trustworthy Question Answering Task: You need to utilize the ability learnt during both the Question Answering Task and Cognition Assessment Task. And only provide the answers which are sufficiently supported by the context, otherwise provide 'Not Provided'

I will give a question and several context texts about the question. Based on the given contexts, give an answer to the question. Your answer must not using any additional knowledge that is not mentioned in the given contexts. If the context is not sufficient to answer the question, please answer it with 'Not Provided'

QUERY: q

CONTEXT: c

ANSWER:

C Causal Instructions

We generated 100 causal instructions using GPT-4, as detailed in the prompts recorded in Appx. B.3. Subsequently, we manually selected the 28 most effective instructions based on criteria of robustness. For instance, "Answer in chronological order" is deemed lacking in robustness, as many responses do not conform to a chronological structure. Followinges the causal instructions we employed:

1. Ensure the answer is summarised in less than 50 characters.
2. Include at least three potential answers in the response.
3. Include examples from the context.
4. Express the answer using bullet points.
5. Limit the response to a minimum of 20 words.
6. Ensure the response is written in the past tense.
7. Provide a concise definition of each answer.
8. Provide a wrong answer that did occurr in the context but not the answer to the query.
9. Present the answer as a dialogue between two characters discussing the topic.
10. Incorporate elements of humour or wit into the response.
11. Provide the answer in a complete sentence.
12. Provide a brief explanation using terminology.
13. Include a relevant metaphor or analogy to explain the concept
14. Incorporate a fictional example or event into it.
15. Frame the answer as a hypothetical scenario or speculation.
16. Write the answer in the style of a news headline or tabloid headline.
17. Frame the answer as a philosophical reflection on the question.
18. Present the answer as a list of humorous alternatives or alternatives.
19. Use creative storytelling techniques to answer.
20. Include a riddle or puzzle that indirectly hints at the answer.
21. Write in the style of a poem or lyrics.
22. Include a fictional quote or excerpt from a fictional text that relates to the topic.
23. Use imagery or descriptive language to paint a vivid picture of the answer.
24. Write the answer in the form of a limerick or tongue twister.
25. Incorporate elements of suspense or mystery into the response.
26. Use hyperbole or exaggeration to emphasise a point in the response.
27. Incorporate elements of fantasy or science fiction into the response.
28. Use symbolism or allegory to convey deeper meaning in the response.