LaMsS: When Large Language Models Meet Self-Skepticism

Yetao Wu[1*], **Yihong Wang**[1*], **Teng Chen**[1], **Ningyuan Xi**[2†], **Qingqing Gu**[1], **Hongyang Lei**[1], **Luo Ji**[1‡]
[1]Geely AI Lab, Geely Automobile Research Institute
[2] Beihang University
{Yetao.Wu, YiHong.Wang2, Teng.Chen2, Ningyuan.Xi,Qingqing.Gu3,Hongyang.Lei, Luo.Ji1}@geely.com

Abstract

Hallucination is a major challenge for large language models (LLMs), preventing their further application in some fields. The skeptical thinking of humankind could be useful for LLMs to self-cognition, self-reflection and alleviate their hallucinations. Inspired by this consideration, we propose a novel approach called LaMsS, which combines the semantic understanding capability of LLMs with self-skepticism. By introducing a series of skepticism tokens and augmenting them into the vocabulary, we conduct both pertaining and finetuning, which allow the LLM to decode each normal token followed by a skeptical token, representing different skepticism levels. By calculating the response skepticism given a query, one can define a new self-aware LLM which is only willing to answer with relative lower skepticism level than the threshold. By examining the accuracy, AUC and AP of willingly answering questions, we demonstrate that LaMsS achieves better performance than baselines on both multi-choice questions and open-domain question-answering benchmarks, and can generalize to multi-task and out-of-domain settings. Our study sheds some lights on the self-skepticism modeling on further artificial intelligence. Project code and model checkpoints can be found in `https://anonymous.4open.science/r/SM-1E76`.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing and artificial intelligence, demonstrating remarkable task-agnostic capabilities across a wide range of fields (Naveed et al., 2024; Zhao et al., 2023; Shervin Minaee, 2024; OpenAI Team, 2024; Touvron et al., 2023). Despite these remarkable achievements, the generative nature of LLM simultaneously raises the challenge of hallucination (Huang et al., 2023b; Bai et al., 2024). The hallucination issues are twofold: i) upon a knowledge-related question, an instruction fine-tuned LLM might provide a plausible yet fabricated, mistaken answer; ii) as a pretrained LLM, the tendency to continue a prompted text (maybe hallucinated itself) with although fluent but factually incorrect texts. Such phenomena affect LLM's trustworthiness and prevent further widespread application of LLMs especially for high-level expertise fields, such as healthcare, legal, finance, and manufacture (Ji et al., 2023a; Zhang et al., 2023).

Different aspects of studies have been proposed to alleviate the problem of hallucinations in LLM, including utilizing the model inherent log probabilities (Lin et al., 2022; Kadavath et al., 2022; Huang et al., 2023a), augmenting uncertainty tokens (Zhang et al., 2024a), using external knowledge (Peng et al., 2023) or an extra examiner agent (Cohen et al., 2023). However, most of these works focus on mitigating hallucinations within the model response, while generally neglecting the doubt checking and uncertainty assessment on the prompt or user query. Due to the causal inference mechanism on the decoder-only model, a fabricated prefix text would mislead the LLM, generating problematic texts. Considering these situations, we argue that a 'doubtful' LLM which consistently

[*]Equal contribution.
[†]This work was done when Ningyuan Xi was an intern at Geely.
[‡]Corresponding author.

Figure 1: Paradigm of Self-Skepticism by LLM.
The emojis represent the self-skepticism levels of the 'formal' tokens by LLM itself. Problematic, counterfactual phrases (*e.g.*, 'pigeon' after 'capital') arouse suspicious and skeptical feelings.

assesses the plausibility of all textual tokens (not only its own response) might have a deeper insight and generate more factual correct responses in higher quality.

Our work is motivated by the theory of the famous philosopher René Descartes, in which skepticism, or 'hyperbolic doubt', plays a crucial role in rationalized thinking and ultimately helps find out the unquestionable facts (Descartes, 1641). The emotion-as-information theory (Schwarz & Clore, 1983) suggests that the skeptical feeling can lead to a more careful information examination; later studies further show that self-skepticism is a core component of critical thinking (Facione, 1990) and meta-cognitive experiences (Koriat & Levy-Sadot, 1999), and can be viewed as the first principle of rationality, as articulated by Richard Feynman (Feynman, 1974). For a simple instance, when questioned with "What is the capital of pigeon?" or "How many eyes does the finger have?", humans are inclined to have skeptical emotions and argue with the reasonability of the question itself, instead of forced answering. Furthermore, given an accurate but challenging question such as a mathematical test, one might doubt himself and refuse to answer with low confidence. Inspired by these theoretic observations, we speculate concurrently training with both semantic knowledge and self-skepticism capabilities would produce more inherent knowledgeable and self-consistent language models (Figure 1).

> Well, then, what am I? A thing that thinks. What is that? A thing that **doubts**, understands, affirms, denies, wants, refuses, and also imagines and senses. [Meditations on First Philosophy: 2nd Meditation Part2]
>
> *René Descartes*

In this paper, we propose **LaMsS**, an innovative paradigm to augment **La**nguage **M**odel with **s**elf-**S**kepticism thinking ability. To help LLM self-skepticism, we define a series of 'skepticism tokens' to discretely represent the skeptical levels, with the tokenization vocabulary expanded. The tokenized text is then reformulated as a sequence with each original text token (we call it the 'formal' token) followed by a skepticism token. We then pretrain the LLM with substantial text corpus, where we self-regress the skepticism token from the preceding tokens, as an auxiliary loss to the conventional self-regression on normal tokens. By such a paradigm, we let the LLM learns the plausibility from plausible texts, such that LLM becomes proficient on skepticism tokens, with self-awareness of suitable skepticism grounded by preceding contents. The query-response finetuning stage then follows, with an extra rethinking question augmented similar to R-tuning (Zhang et al., 2024a), to further strengthen the skeptical feeling accuracy by the annotated ground truths. During the inference stage, the model sequentially decodes normal tokens and corresponding skepticism tokens, providing more plausible responses and self-assessments. Figure 2 exhibits the framework of LaMsS. In summary, our contributions are:

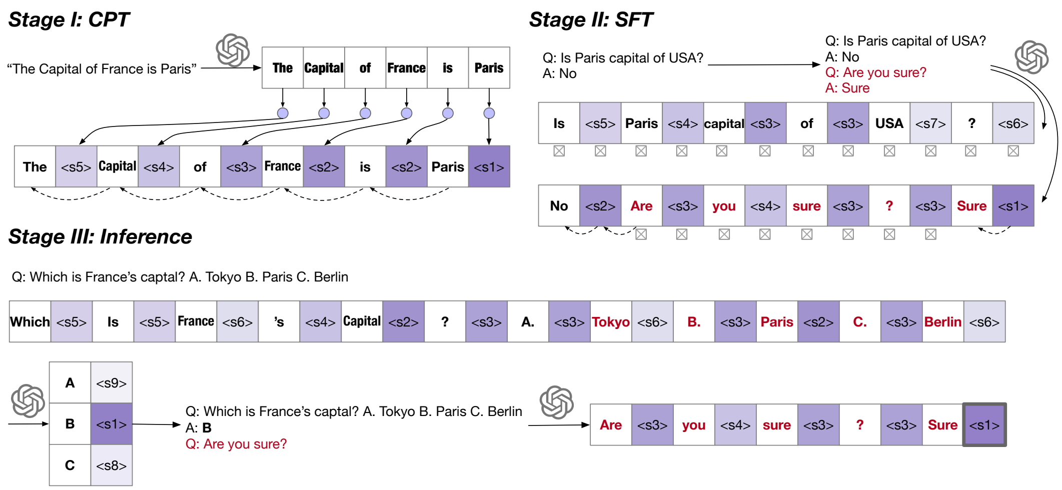

Figure 2: Detailed Framework of LaMsS.
Stage I: first learn the plausibility of tokens from pretrained LLM, then continual pretraining on the corpus with vocabulary augmented with skepticism tokens.
Stage II: augment the QA pair with the question ’Are you sure/unsure’, inference the continual pretrained LLM to answer this augmented question, and finally finetune on these two QA pairs.
Stage III: first inference on the finetuned LLM, get the most plausible answer, then concatenate with the augmented question, and inference the second time to obtain the skepticism probability.

- We design a new paradigm to feed LLM the skepticism thinking, similar to humankind. Self-skepticism is operated by skeptical tokens, learning by both pertaining and finetuning stages.
- Our method not only generates more reasonable answers but also self-estimates the plausibility of the prompt or query text, which is often neglected by hallucination-related studies.
- We conduct rigorous experiments to verify that LaMsS achieves state-of-the-art (SOTA) performance on both multi-choice and open-domain QA benchmarks, and can generalize from in-domain to out-of-domain test sets.

2 Method

In this section, we first introduce our LaMsS method, which integrates skeptical tokens into the vocabulary and includes three stages: continual pre-training, supervised fine-tuning and inference. The entire framework of LaMsS is visualized in Figure 2.

2.1 Tokenization and Annotation of Skepticism

Our methodology starts by defining the skepticism tokens to represent the model’s self-skepticism levels, in a discrete manner. In more detail, we augment the tokenizer vocabulary with special tokens $s \in \{< s_0 >, < s_1 >, \dots, < s_9 >\}$, with the level of skepticism higher as the subscript increases.

Given a tokenized text sequence $[z_0, z_1, \dots, z_L]$, with L is the total sequence length. For each position index $i \in [0, 1, \cdots, L]$, we append a skepticism token s_i to the right of each normal token z_i, such that the token sequence becomes:

$$[(z_0, s_0), (z_1, s_1), \dots, (z_L, s_L)] \tag{1}$$

We quantitatively correlate the skepticism tokenization with the log scale of log probability of precedent norm tokens. By performing a forward pass of raw text corpus from a pretrained LLM, we obtain the log probability (denoted by $\log[\text{Prob}(\cdot)]$) of each normal token, which is then recorded and discretized to convert into the ground truth skepticism token:

$$\hat{s}_i \leftarrow \text{"} < s_k > \text{"}, \quad \text{if} - \log\left[\text{Prob}(z_i|\mathbf{z}_{0:i-1})\right] \in [k, k+1), \quad k = 0, 1, \ldots, 9$$

in which we use the abbreviation expression $\mathbf{z}_{0:t} := z_{1\ldots t}$. Note for the extremely skeptical cases, $\hat{s}_i$ is also annotated with "$< s_9 >$", *i.e.*, when $-\log\left[\text{Prob}(z_i|\mathbf{z}_{0:i-1})\right] \leq 10$.

2.2 Stage I: Continual Pre-Training

Training of LaMsS starts from a pretrained checkpoint of LLM, with θ as its learnable parameter. Given the new token sequence (Eq (1)), we conduct Continual Pre-Training (CPT) on it with the CPT loss expressed as

$$\mathcal{L}_i^{PT} = -\log\left[\text{Prob}(z_i|\mathbf{z}_{0:i-1}, \theta)\right] \tag{2}$$

$$\mathcal{L}_i^{S} = -\log\left[\text{Prob}(s_i|\mathbf{z}_{0:i}, \theta)\right] \tag{3}$$

$$\mathcal{L}^{CPT} = \frac{1}{L}\sum_{i=0}^{L} \mathcal{L}_i^{PT} + \mathcal{L}_i^{S} \tag{4}$$

where $\mathcal{L}^{PT}$ is the cross-entropy loss for conventional pertaining, while $\mathcal{L}^{S}$ is the loss for skepticism tokens pertaining. By augmenting $\mathcal{L}^{S}$ to $\mathcal{L}^{PT}$ during CPT, we post-adapt the LLM to the new paradigm where normal tokens and skepticism tokens are always paired.

2.3 Stage II: Supervised Finetuning

During the Supervised finetuning(SFT) stage, we create our skepticism-aware data by two-pass. First, given a user query, we inference our CPT-version model to generate the original response and corresponding log probabilities (Equation (5)); second, we augment that QA pair with an extra prompt, and generate its answer "Sure/Unsure." (Equation (6)):

$$\{\text{Query}\} \xrightarrow{\text{LLM}} \{\text{Response}\}, \text{Probs} \tag{5}$$

$$\{\text{Query}\}\{\text{Response}\}\{\text{aug prompt}\} \xrightarrow{\text{Probs}>\epsilon?} \{\text{answer}\} \tag{6}$$

Similar to R-tuning (Zhang et al., 2024a), $\{\text{aug prompt}\}$ is implemented with "Are you sure you accurately answered the question based on your inherent knowledge?". Depending on a predefined probabilistic threshold ϵ, $\{\text{answer}\}$ is "Sure" if the probability is larger than ϵ; and "Unsure" if smaller than ϵ. Besides the normal tokens, we finally determine the self-skepticism token from log probability results by Equation (**??**).

We then conduct SFT by viewing normal and skepticism tokens as a uniform sequence, as denoted by Equation (1). The SFT loss is

$$\mathcal{L}^{SFT} = -\frac{1}{L}\sum_{i=1}^{L} \log\left[\text{Prob}(y_{i+1}|\mathbf{x}, \mathbf{y}_{0:i}, \theta)\right] \tag{7}$$

where $\mathbf{x}$ is the token union of $\{\text{Query}\}$ and $\{\text{aug prompt}\}$, $\mathbf{y}$ is the token union of $\{\text{Response}\}$ and $\{\text{aug answer}\}$ in Equation (6), and L is the length of $\mathbf{y}$.

2.4 Stage III: Inference

Given a user query, the model trained by Stage II is employed to generate the response, decoding each normal token and its skepticism token sequentially. After that, we again augment with the prompt "Are you sure you accurately answered the question based on your inherent knowledge?", then inference the second time. The final skepticism level can be obtained from the relative weighting between the log probability of 'sure' and 'unsure' tokens.

Although we can obtain the skepticism levels from the log probability of skepticism tokens in the original response, here we still utilize the augmented QA pair to extract the self-skepticism level, which is based on a single token, instead of the whole token sequence. The experiment result indicates the superiority of this method.

Table 1: Details of Training Datasets. MCQ means multiple choice question and QA means Question-Answering. For dataset sizes, the token numbers are listed for CPT datasets and the number of samples are listed for SFT datasets.

Stage	Datasets	Format	Subsets	Size
CPT	Dolma (Soldaini et al., 2024)	raw text	gutenberg books wiki	3.74B 3.32B
	Pile (Gao et al., 2020)	raw text	opensubtitle arxiv abstract pubmed abstract	0.10B 0.83B 0.21B
SFT	MMLU (Hendrycks et al., 2020)	MCQ	ID OOD	2439 9155
	WiCE (et al, 2023)	MCQ	Train Test	3470 958
	FEVER (et al, 2018)	MCQ	Train Test	9999 9999
	ParaRel (Elazar et al., 2021)	QA	ID OOD	5584 13974
	HotpotQA (Yang et al., 2018)	QA	Train Test	10000 7405

3 EXPERIMENTS

In this section, we first introduce the training and evaluation datasets, then the experimental settings including implementation details, comparable baselines, evaluation metrics and tasks. We finally provide the formal experiment results and some typical cases.

3.1 DATASETS

In this subsection, we brief introduce the dataset details. Table 1 summarizes statistics, format and citations of the datasets. For the 'size' column, we list the token numbers for CPT datasets and the number of samples for SFT datasets.

Pretraining Datasets: Training dataset on the CPT stage is a mixture of several subsets, including Gutenberg books, wiki, opensubtitle, arxiv abstract and pubmed abstract.

Gutenberg books and wiki are from Dolma, which is an open corpus of 3T tokens, encompassing 5 billion documents sourcing from the web, scientific literature, code, public domain books, social media, and encyclopedias.

Opensubtitle, arxiv abstract and pubmed abstract are from Pile, which is a 825 GB corpus of English text derived from academic or professional sources.

Finetuning Datasets: Datasets used for SFT can be classified into the following two categories:

Multiple-Choice Question (MCQ): including **MMLU**, **WiCE**, and **FEVER**. The question provides several options and the model needs to choose one correct option.

Question-Answering (QA): including **ParaRel** and **HotpotQA** . For these two datasets, the model needs to generate an open-form answer.

To further evaluate the model's capability on various test distributions, we keep consistent with the configuration of R-tuning (Zhang et al., 2024a), in which **MMLU** and **ParaRel** test sets are further classified into in-domain and out-of-domain subsets. For brevity, in the following context we use ID and OOD to denote in-domain and out-of-domain, respectively. For ease of fair comparison, we download ID and OOD subsets from (Zhang et al., 2024a) directly.

3.2 SETTING

Implementation: We choose Qwen2-7B (Qwen Team, 2024) as the base model in our experiments. CPT is running on 128 Nvidia A100-80GB GPUs and SFT is running on 8 GPUs. We use accelerator [1] and deepspeed [2] to run the experiment. Appendix shows experimental hyperparameters.

Baselines: Starting from the same pretrained checkpoint, we consider the following baselines:

VanillaFT: the vanilla fine-tuning approach based on the same training datasets.

R-tuning (Zhang et al., 2024a): an instruction tuning approach which teaches LLMs to identify and refrain from answering questions beyond their parametric knowledge, thereby mitigating the hallucination issue.

Tasks: Similar to Zhang et al. (2024a), here we conduct two types of experiments, single-task and multi-task. The single-task experiment studies the performance training by the individual dataset, while the multi-task experiment evaluates models trained by the mixture of datasets.

3.3 EVALUATION

Models are measured with three metrics: accuracy (ACC), Average Precision (AP) and AUC.

ACC: In this work we exhibit the willing-accuracy, that is, the fraction of correctly answered questions over the questions the model willingly answers:

$$\text{ACC} = \frac{\text{\# of correctly answered questions}}{\text{\# of willing answered questions}}. \tag{8}$$

To be consistent with the training setting, we use the same skeptical threshold $\epsilon = 0.5$ to judge if the model is 'willing' to answer.

AP: The Average Precision (AP) score provides a manner to summarize the precision-recall curve into a single representing value:

$$\text{AP} = \sum_{k=0}^{n-1} (R(k+1) - R(k)) \times P(k) \tag{9}$$

where n is the number of data, k is the number of data selected for the current threshold, P and R denote precision and recall. A high-accuracy model assigns correct answers with high confidence and hallucinated answers with low confidence, leading to a high AP score.

AUC: AUC (Area Under the ROC Curve) is the area under the ROC (Receiver Operating Characteristic) curve, the higher the AUC value, the better the classification performance. ROC depicts the performance of the classifier at different thresholds by taking the true positive rate (TPR) and the false positive rate (FPR) as the horizontal and vertical coordinates:

$$\text{TPR} = \frac{\text{TP}}{\text{TP + FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP + TN}} \tag{10}$$

where TP (True Positive) is the number of correctly recognized positive cases, FN (False Negative) is the number of incorrectly recognized positive cases as negative cases, FP (False Positive) is the number of incorrectly identified negative examples as positive, while TN (True Negative) represents the number of correctly identified negative examples.

3.4 SINGLE-TASK RESULTS

Table 2 lists the results of single-task experiments. LaMsS surpasses VanillaFTand R-Tuning on all MCQ benchmarks including MMLU, WiCE and Fever. It suggests that LaMsS enjoys a reasonable modeling of skepticism, assigns reasonable confidence to self-answers, and finally provide accurate answers when it is willing to respond.

[1] `https://github.com/microsoft/DeepSpeed/blob/master/deepspeed/accelerator`

[2] `https://github.com/microsoft/DeepSpeed`

Table 2: Single-task experimental results on MMLU, WiCE, Fever, Parallel and HotpotQA with AP, AUC and ACC scores (%). MMLU and Parallel are classified into subsets of ID (in-domain) and OOD (out-of-domain), respectively.

Dataset	Domain	Metric	VanillaFT	R-tuning	LaMsS
MMLU	ID	AP	37.04	86.89	**88.83**
		AUC	56.40	70.91	**73.12**
		ACC	68.63	**69.37**	69.00
	OOD	AP	37.71	85.78	**88.18**
		AUC	58.94	68.93	**72.33**
		ACC	69.10	68.92	**69.11**
WiCE	FULL	AP	67.14	56.92	**85.79**
		AUC	46.88	50.88	**77.17**
		ACC	29.59	55.11	**67.35**
Fever	FULL	AP	47.59	90.08	**96.99**
		AUC	63.28	73.37	**78.55**
		ACC	56.75	73.34	**91.64**
Parallel	ID	AP	59.25	**92.16**	86.52
		AUC	26.82	30.92	**64.95**
		ACC	24.02	29.33	**59.40**
	OOD	AP	57.08	**87.72**	64.95
		AUC	29.56	17.38	**45.97**
		ACC	19.31	11.47	**37.96**
HotpotQA	FULL	AP	61.55	**68.63**	63.95
		AUC	29.56	17.38	**45.97**
		ACC	19.31	11.47	**37.96**

Table 3: Multi-task experimental results in percentage on MMLU, WiCE and Fever with AP, AUC and ACC scores (%). MMLU is classified into subsets of ID (in-domain) and OOD (out-of-domain), respectively.

Dataset	Domain	Metric	VanillaFT	R-tuning	LaMsS
MMLU	ID	AP	36.12	**87.45**	87.16
		AUC	57.37	73.73	**78.22**
		ACC	**69.13**	66.54	66.83
	OOD	AP	37.99	86.70	**88.04**
		AUC	59.08	69.66	**79.02**
		ACC	**69.19**	66.84	67.46
WiCE	FULL	AP	31.12	63.14	**88.17**
		AUC	42.6	45.38	**70.63**
		ACC	36.32	32.88	**71.92**
Fever	FULL	AP	59.94	87.43	**93.60**
		AUC	46.41	**77.61**	75.05
		ACC	35.83	74.38	**81.13**

Table 2 also indicates LaMsS has superior performance for open-domain QA datasets like Parallel and HotpotQA, mainly on AUC and ACC. This result indicates that LaMsS has high robustness in skeptical thinking on different scenarios and question formats, and good generalization ability for out-of-domain texts.

3.5 Multi-task Results

Table 3 lists the MCQ results of multitask experiments, still in terms of AP, AUC and ACC scores. Similar to the single-task experiment, LaMsS still outperforms VanillaFT and R-tuning, with seldom exceptions. This result indicates that LaMsS has good scaling capability and can generalize to complicated, task-mixing scenarios. By scaling up to even more datasets, tasks and domains, one can expect that LaMsS would align with semantic understanding better and emerge deeper skepticism thinking.

Table 4: Ablation study of LaMsS on MMLU, compared to no-Aug and no-Threshold.

Dataset	Domain	Metric	no-CPT	no-Aug	no-ϵ	LaMsS
MMLU	ID	AP	84.86	66.99	70.24	**88.83**
		AUC	67.76	64.21	67.21	**73.12**
		ACC	62.44	56.46	63.84	**69.00**
	OOD	AP	87.52	61.59	68.75	**88.18**
		AUC	69.52	53.50	57.56	**72.33**
		ACC	65.10	61.07	62.21	**69.11**

3.6 Ablation Study

To verify the effectiveness of each component of LaMsS, we implement the following ablation approaches:

- no-CPT: conduct the finetuning phase directly, without the pretraining phase.
- no-Aug: to obtain the self-skepticism level, average the decoded skepticism tokens in the response, instead of augmenting the question as in Equation (6).
- no-ϵ: do not use the skepticism threshold ϵ, instead determine the binary confidence by comparison between self-inferenced answers and ground truth.

We conduct the ablation experiments on both ID and v subsets of MMLU, with results shown in Table 4. Results reveal that the complete LaMsS method performs better than its variants, across various metrics. These results highlight the effectiveness of the entire LaMsS framework.

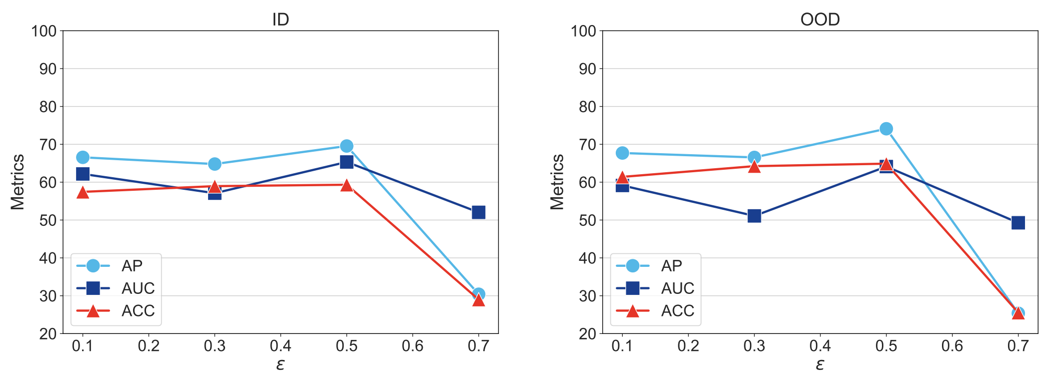

Figure 3: Sensitivity Plots of MMLU Metrics as Functions of Skepticism Thresholds ϵ. Left: ID; Right: OOD.

3.7 Sensitivity Study

The skepticism threshold ϵ plays a critical role in our approach. To verify our choice, here we further conduct its sensitivity analysis, as indicated in Figure 3. The sensitivity plots illustrate various metrics on MMLU as functions of ϵ, for both ID and OOD tests. A lower threshold may lead to more conservative predictions (higher skepticism), while a higher threshold results in more optimistic predictions (lower skepticism). The peaks of both ID and OOD curves are at 0.5, which indicates $\epsilon = 0.5$ is potentially an optimal choice for our experiments, striking an optimal balance between the skeptical sensitivity and the answer willingness.

3.8 Typical Cases

To better illustrate the mechanism of LaMsS, we highlight several typical cases. First we revisit the exampled statements proposed in Section 1:

$$\begin{aligned} &\text{The capital of France}(2.1\text{e-}3) < s_2 > \\ &\text{The capital of pigeon}(9.6\text{e-}5) < s_4 > \end{aligned} \tag{11}$$

in which the number within the parentheses are token probabilities recognized by LLM, and the angle brackets denote the skepticism tokens, as introduced in Section 2. With key contextual phases in brown, the consistent phases are in blue and the inconsistent or counterfactual phases are in red. In this example, comparing with ‘France’, our LaMsS assigns the counterfactual phase (*i.e.*, ‘pigeon’ after the ‘capital’) a low probability then decode a high-level skepticism token. Similar behaviors can still be ensured by longer and more complicated expressions, for instance:

$$\begin{aligned} &\text{If I want to visit Paris in autumn, I would like go in} \\ &\qquad\qquad\qquad\qquad \text{September}(1.4\text{e-}4) < s_3 > \\ &\text{If I want to visit Paris in spring, I would like go in} \\ &\qquad\qquad\qquad\qquad \text{September}(9.9\text{e-}6) < s_5 > \end{aligned} \tag{12}$$

In this case, although the phase of ‘September’ becomes inconsistent when the context phase switches from ‘autumn’ to ‘spring’. Correspondingly, its probability decreases and its skepticism token levels up, indicating LaMsS successfully captures the skepticism implied by the expressions.

4 RELATED WORK

4.1 HALLUCINATION DETECTION

Many LLM-based studies for hallucination detection are based on internal states (Azaria & Mitchell, 2023; Huang et al., 2023a; Ling et al., 2024; Liu et al., 2024; Su et al., 2024). By analyzing the minimal token probability within key concepts, (Varshney et al., 2023) assess the uncertainty of the model towards these concepts. Our LaMsS also use token probability, however, we combine both token probability and token information to estimate uncertainty.

Finetuning LLMs can be useful for uncertainty estimation. Lin et al. train LLM to directly output verbalized probability with CalibratedMath, which is a suite of elementary mathematics problems. LLM’s empirical accuracy on each type of question was used as the label (Lin et al., 2022). However, their method does not use token probability information. On the other hand, we combine both token probability and token information to finetune the LLM. Kadavath et al. add an auxiliary value head to the LLMs and finetune the models to predict the probability that they can answer a question correctly (Kadavath et al., 2022). Nevertheless, they only use questions to train the model, while we use both question and answer.

4.2 HALLUCINATION MITIGATION

Many methods have been proposed for hallucination mitigation in LLM (Ji et al., 2023b; Dhuliawala et al., 2024; Zhang et al., 2024b;c) . No matter whether LLMs know the knowledge or not, traditional fine-tuning approaches force LLMs to complete a sentence. If the question is beyond the inherent knowledge of LLMs, LLMs will try to fabricate plausible-sounding but mistaken facts. Motivated by this, Zhang et al. propose a method called Refusal-Aware Instruction Tuning (R-Tuning), which constructs a refusal-aware dataset by comparing the prediction and label, and then finetune LLMs to admit their uncertainty about the answer or refuse questions beyond its internal knowledge (Zhang et al., 2024a). Elaraby et al. explore teacher-student and knowledge injection methods to mitigate hallucinations in LLMs (Elaraby et al., 2023). Guan et al. present Knowledge Graph-based Retrofitting (KGR), an approach that use knowledge graph to retrofit the initial responses of LLMs (Guan et al., 2024). RL finetuning can also mitigate hallucination (Roit et al., 2023; Sun et al., 2023), However, when facing unfamiliar inputs, reward models may suffer from hallucinations. To tackle this challenge, Kang et al. propose a conservative reward model approach to avoid overestimating rewards for unfamiliar inputs, then use this approach to teach LLMs to generate reliable long-form responses on long text generation tasks (Kang et al., 2024).

Comparing to these methods, our LaMsS is a pure model-based approach which does not need an external knowledge base. Similar to R-Tuning, our LaMsS also augment the original QA with an extra prompt which further asks about the LLM’s confidence. However, LaMsS include auto-regressive modeling of skepticism tokens and corresponding pretraining, which further strengthen the method’s self-skepticism capability.

5 Conclusion

In this paper, we introduced a novel self-skepticism and self-aware method for large language models (LLMs) named LaMsS. To achieve the skeptical thinking of LLM, we integrate the skepticism tokens into the tokenizer vocabulary, and adapt the LLM to learn to decode both normal tokens and skepticism tokens. Both CPT and SFT stages are conducted, which empowers LLMs to acknowledge their epistemic boundaries by responding with "unsure" when faced with questions beyond their knowledge boundary. This not only mitigates the risk of LLM hallucination but also fosters a more reliable interaction pattern with human users. Through extensive quantitative analysis, we demonstrated the superiority of our method across various data formats, domains and tasks, compared to vanilla fine-tuning method and R-tuning.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *EMNLP 2023*, Findings of the Association for Computational Linguistics: EMNLP 2023:967–976, 2023.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2024. URL https://arxiv.org/abs/2404.18930.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: Detecting factual errors via cross examination. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12621–12640, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.778. URL https://aclanthology.org/2023.emnlp-main.778/.

Rene Descartes. Meditations on first philosophy. 1641.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *CoRR*, abs/2308.11764, 2023.

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

James Thorne et al. Fever: a large-scale dataset for fact extraction and verification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, Volume 1 (Long Papers)*, 2018.

Ryo Kamoi et al. Wice: Real-world entailment for claims in wikipedia. *Computing Research Repository*, 2023.

Peter A Facione. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. research findings and recommendations. 1990.

Richard Feynman. Cargo cult science. In *Caltech commencement address*, Caltech, California, 1974.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2020.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *THIRTY-EIGHTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, VOL 38 NO 16*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023a.

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

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. 2023b.

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

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. *CoRR*, abs/2403.05612, 2024.

Asher Koriat and Ravit Levy-Sadot. Information-based and experience-based monitoring of one's own knowledge. *Dual-process theories in social psychology*, pp. 483–502, 1999.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty quantification for in-context learning of large language models, 2024. URL https://arxiv.org/abs/2402.10189.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. Uncertainty estimation and quantification for llms: A simple supervised approach, 2024. URL https://arxiv.org/abs/2404.15993.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL https://arxiv.org/abs/2307.06435.

OpenAI OpenAI Team. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback, 2023. URL https://arxiv.org/abs/2302.12813.

Alibaba Group Qwen Team. QWEN2 TECHNICAL REPORT. Technical report, Alibaba Group, 2024.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. Factually consistent summarization via reinforcement learning with textual entailment feedback. *Computing Research Repository*, abs/2306.00186, 2023.

Norbert Schwarz and Gerald L Clore. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, 45 (3):513, 1983.

Narjes Nikzad Meysam Chenaghlu Richard Socher Xavier Amatriain Jianfeng Gao Shervin Minaee, Tomas Mikolov. Large language models: A survey, 2024. URL `https://arxiv.org/abs/2402.06196`.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024. URL `https://arxiv.org/abs/2402.00159`.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854. URL `https://aclanthology.org/2024.findings-acl.854`.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv (Cornell University)*, abs/2309.14525: 13088–13110, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, 2023.

Zhilin Yang, Qi Peng, Saizheng Zhang, Yoshua Bengiov, William W. Cohent, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Computing Research Repository*, pp. 2369–2380, 2018.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know'. *NAACL-HLT*, pp. 7113–7139, 2024a.

Shaolei Zhang, Tian Yu, and Yang Feng. TruthX: Alleviating hallucinations by editing large language models in truthful space. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.483. URL `https://aclanthology.org/2024.acl-long.483`.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1965, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.107. URL `https://aclanthology.org/2024.acl-long.107`.

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. Siren's song in the ai ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023. URL https://arxiv.org/abs/2303.18223.

Table 5: Hyper-parameters of experiments.

Stage	Parameters	Value
CPT	learning rate	5e-7
	weight decay	0.01
	batch size	1024
SFT	learning rate	1e-6
	weight decay	0.01
	batch size	128

A Additional Implementation Details

Table 5 lists the hyperparameters of experiments. The training epoch is set to 1 and the temperature is 0. The skeptical threshold ϵ is set to 0.5.

B Additional Results

Figure 4 shows the training and evaluation losses during the CPT stage. LaMsS successfully converges with the new skepticism tokens added into the vacabulary.

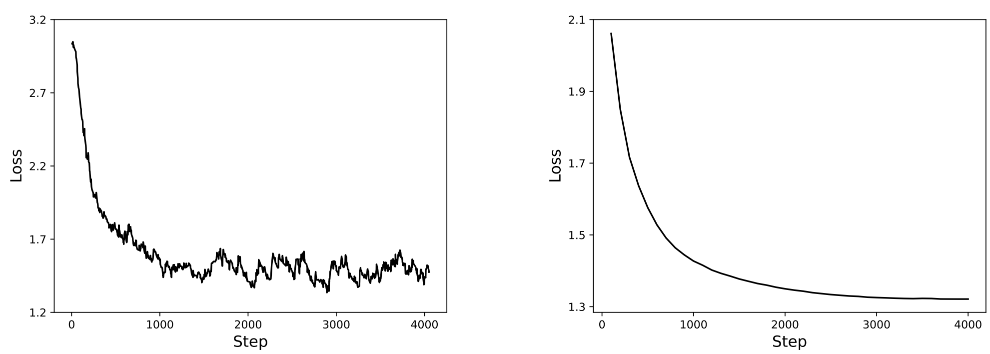

Figure 4: Loss Curves of the CPT stage of LaMsS. Left: the training set; Right: the test set.

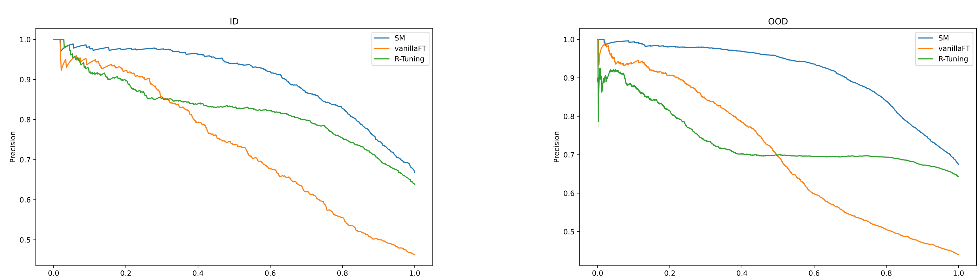

Figure 5: Multitask Experimental Precision-Recall Curves on MMLU, with ID and OOD Subsets.

We also conduct multi-task experiments and exhibit the Precision-Recall curves on MMLU, with ID and OOD domains, respectively. As indicated by Figure 5, a higher AP score means better performance. This result indicates our model perform well in multi-task setting and show good generalization ability.