Can LLMs Learn Uncertainty on Their Own?
Expressing Uncertainty Effectively in A Self-Training Manner

Shudong Liu[♠] **Zhaocong Li**[♠] **Xuebo Liu**[♣*] **Runzhe Zhan**[♠]
Derek F. Wong[♠*] **Lidia S. Chao**[♠] **Min Zhang**[♣]
[♠]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
[♣]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
nlp2ct.{shudong,zhaocong,runzhe}@gmail.com, {liuxuebo,zhangmin2021}@hit.edu.cn
{derekfw,lidiasc}@um.edu.mo

Abstract

Large language models (LLMs) often exhibit excessive, random, and uninformative uncertainty, rendering them unsuitable for decision-making in human-computer interactions. In this paper, we aim to instigate a heightened awareness of self-uncertainty in LLMs, enabling them to express uncertainty more effectively. To accomplish this, we propose an uncertainty-aware instruction tuning (UaIT) method, aligning LLMs' perception with the probabilistic uncertainty of the generation. We conducted experiments using LLaMA2 and Mistral on multiple free-form QA tasks. Experimental results revealed a surprising 45.2% improvement in the effectiveness of uncertainty expression by LLMs, accompanied by reasonably good out-of-domain generalization capabilities. Moreover, this uncertainty expression can serve as a valuable real-time basis for human decision-making, e.g., retrieving external documents and incorporating stronger LLMs[1].

1 Introduction

Large language models (LLMs), such as ChatGPT and GPT-4, are capable of generating fluent and realistic responses tailored to diverse user requirements (Ouyang et al., 2022; OpenAI, 2023). However, LLMs do not consistently exhibit optimal performance, as they can also generate unreliable responses characterized by hallucinations or factual errors. Effective uncertainty estimation is widely recognized as a crucial step in establishing reliable AI systems, as it provides a foundation for decision-making in human-machine interactions.

Unlike previously examined models with distinct labels (e.g. classification), uncertainty estimation for free-form LLM poses a significant challenge due to the inherent flexibility in generation and the unbounded nature of solution domains (Kadavath et al., 2022; Duan et al., 2023; Kuhn et al., 2023). Nevertheless, these methods mainly rely on model probability and multi-sampling to derive uncertainty, which entails substantial time and resources, rendering them impractical for real-time interactions. Moreover, natural language has emerged as the predominant interface for human interaction with AI systems encompassing various tasks (Zhou et al., 2024). Recent research has been dedicated to prompting LLMs to express verbalized confidence (Tian et al., 2023; Xiong et al., 2024). However, LLMs, especially smaller ones, consistently

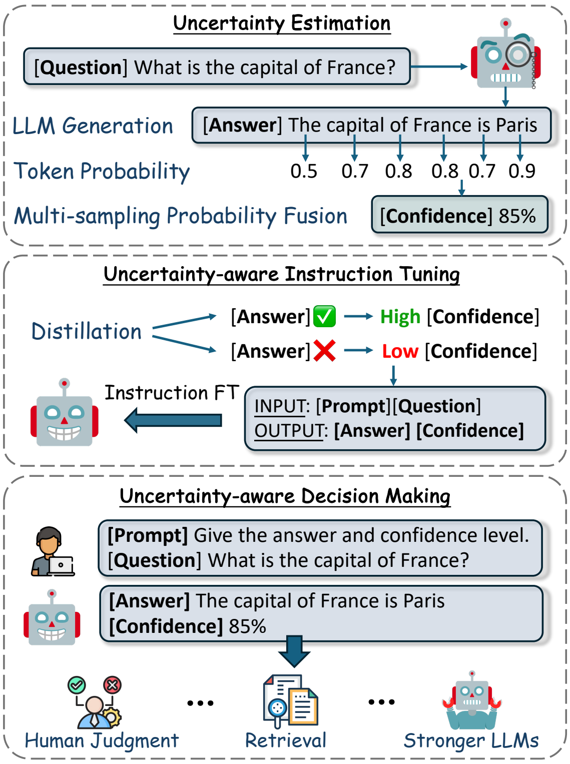
Uncertainty Estimation
[Question] What is the capital of France?
LLM Generation
[Answer] The capital of France is Paris
Token Probability
0.5 0.7 0.8 0.8 0.7 0.9
Multi-sampling Probability Fusion
[Confidence] 85%
Uncertainty-aware Instruction Tuning
Distillation
[Answer]✅ → High [Confidence]
[Answer]❌ → Low [Confidence]
Instruction FT
INPUT: [Prompt][Question]
OUTPUT: [Answer] [Confidence]
Uncertainty-aware Decision Making
[Prompt] Give the answer and confidence level.
[Question] What is the capital of France?
[Answer] The capital of France is Paris
[Confidence] 85%
Human Judgment
...
Retrieval
...
Stronger LLMs

Figure 1: Our objective is to align the LLMs' self-generated probabilistic uncertainty estimation and express it. This uncertainty expression can then be applied in real-time human decision-making, guiding judgment, retrieval documents, and leveraging stronger LLMs.

[*] Co-corresponding Author

[1] Code and scripts can be found at: https://github.com/NLP2CT/UaIT

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21635–21645
November 12-16, 2024 ©2024 Association for Computational Linguistics

exhibit a high and unvarying pattern of verbalized confidence, indicating a poor level of competence in uncertainty expression.

In this paper, we seek to elicit the capacity of LLMs to effectively and accurately express uncertainty. We employ advanced method (Duan et al., 2023), based on probability and multi-sampling, to assess the model's uncertainty of its free-form generation. Subsequently, we utilize these uncertainty estimates as labels to construct instructions and train LLMs to align with their own uncertainty. The expressed uncertainty is applied in practical decision-making scenarios, including determining when to retrieve external documents and incorporate more powerful LLMs. We conduct experiments using the LLaMA-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023a) models on a range of free-form question-answering tasks, spanning domains such as reading comprehension, factual, scientific, and medical. We make a remarkable discovery that this simple method has led to a 45.2% improvement in the ability of LLMs to express uncertainty, while also demonstrating commendable cross-domain generalization capabilities. The expressed uncertainty also provides a strong foundation for downstream decision-making processes.

2 Improving Self-Uncertainty Expression

2.1 Uncertainty Estimation

We employ SAR (Duan et al., 2023), an advanced approach based on multi-sampling and probability fusion to estimate the uncertainty of free-form generation. Given $\boldsymbol{x}$ as the input query, LLM generates a response $\boldsymbol{y}$ with the probability distribution $p_\theta\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}\right)$. Then the predictive entropy is:

$$\mathrm{PE}(\boldsymbol{y}, \boldsymbol{x}) = \sum_t - \log p_\theta\left(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}\right). \quad (1)$$

SAR claims that tokens are not equivalent in expressing sentence semantics and should be given different attention in uncertainty estimation. Therefore, SAR quantifies the relevance score of each token by comparing the semantic change upon its removal from the generation. The token-level shifted predictive entropy can be computed as:

$$\mathrm{TokenSAR}\left(\boldsymbol{y}, \boldsymbol{x}\right) = \sum_t - \log p_\theta\left(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}\right) R_T\left(y_t\right), \quad (2)$$

where $R_T(y_t)$ is the relevance weight for the token y_t. Subsequently, this relevance score is also extended to the sentence-level predictive entropy under a multi-sampling setup:

$$\mathrm{SentSAR}(Y, \boldsymbol{x}) = \frac{1}{K} \sum_k \mathrm{PE}(\boldsymbol{y}, \boldsymbol{x}) R_S\left(\boldsymbol{y}\right), \quad (3)$$

where $R_S\left(\boldsymbol{y}\right)$ is the relevance weight for sentence $\boldsymbol{y} \in Y (1 \leq k \leq K)$. Ultimately, SAR combines token-shifted and sentence-shifted predictive entropy to obtain uncertainty scores. Actually, other effective methods for quantifying uncertainty can be employed as substitutes to obtain a fine-grained uncertainty score in our method.

2.2 Uncertainty-aware Instruction Tuning

To construct the training set for uncertainty-aware instruction tuning, we input the question to the LLMs and obtain a confidence score in percentage form using the above uncertainty estimation approach. Given the free-form nature of LLM outputs, current uncertainty estimation methods still demonstrate limited effectiveness. To enhance the quality of the training set, as illustrated in Figure 1, we filter samples that exhibit consistency between accuracy and confidence scores. Specifically, we distill samples with both correct answers and confidence scores above a specific threshold, as well as samples with incorrect answers and confidence scores below the threshold. The distilled dataset $\mathcal{D}$ can be defined as $\mathcal{D} = \{(p_i, q_i, a_i, c_i)\}_{i=1}^n$, where p_i, q_i, a_i, and c_i represent the user's prompt, question, answer, and confidence level associated with the answer respectively, and n is the dataset size. Then the process of instruction tuning is represented as:

$$\underset{\triangle\theta}{\operatorname{argmin}} \sum_{i=1}^{n} - \log\left(p\left(a_i, c_i \mid q_i, p_i; \theta + \triangle\theta\right)\right), \quad (4)$$

where θ and $\triangle\theta$ are the original weights and updated weights. We demonstrate that such a simple fine-tuning approach effectively stimulates uncertainty perception in LLMs. It is worth emphasizing that our objective is to cultivate self-awareness in LLMs rather than modifying their beliefs, as we input the answers they themselves generate.

2.3 Uncertainty-aware Decision Making

To further validate the effectiveness of uncertainty expressed by LLMs in practical interaction, we leverage uncertainty as a basis for human decision-making. Specifically, we demonstrate its effectiveness in downstream tasks through three scenarios: uncertainty-based human judgment (evaluated for correlation with accuracy), retrieval of external documents, and leveraging more powerful LLMs for

Model	Method	In-domain	Out-of-domain	
		TriviaQA	SciQA	MedQA
Mistral	Verbalized	0.644	0.579	0.503
	PE	0.705	0.585	0.569
	SAR	0.762	0.672	0.564
	UaIT	**0.846**	**0.775**	**0.582**
LLaMA2	Verbalized	0.536	0.507	0.499
	PE	0.726	0.583	0.530
	SAR	0.759	0.637	0.530
	UaIT	**0.867**	**0.730**	**0.574**

Table 1: The AUROC scores on three QA datasets.

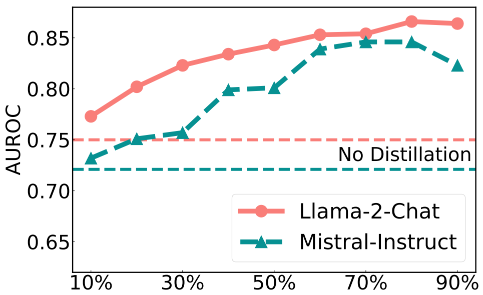

Figure 2: Effect of different thresholds on AUROC during data distillation.

assistance. Since well-calibrated LMs tend to lack knowledge when exhibiting low confidence/high uncertainty (Kadavath et al., 2022; Jiang et al., 2023b), we proactively trigger retrieval/stronger LLM when the LLM's confidence falls below a specified threshold. Taking retrieval as an example, decision-making can be formalized as:

$$\boldsymbol{y}_t = \begin{cases} \text{LLM}\left([\boldsymbol{x}, \boldsymbol{y}_{<t}]\right) & \text{if Conf } \geq \alpha, \\ \text{LLM}\left([\mathcal{D}_{\boldsymbol{x}}, \boldsymbol{x}, \boldsymbol{y}_{<t}]\right) & \text{otherwise,} \end{cases} \quad (5)$$

where $\mathcal{D}_{\boldsymbol{x}}$ is the retrieval document and α is the threshold. We demonstrate that LLMs, through such simple fine-tuning, are capable of effectively expressing meaningful uncertainty and can serve as a real-time basis for human decision-making.

3 Experiments

3.1 Setup

Datasets and Metric In our experiments, we consider TriviaQA (Joshi et al., 2017), SciQA (Welbl et al., 2017), and MedQA (Jin et al., 2020), which respectively represent fact-based, science-related, and medical-related question-answering tasks. We utilize RougeL (Lin, 2004) to measure the accuracy of generation and AUROC to assess the effectiveness of uncertainty. More details of the datasets and metrics can be found in Appendix A.1 and A.2.

Baseline We compare our method with the following Uncertainty Expression/Estimation methods: (1) **Verbalized** (Tian et al., 2023; Xiong et al., 2024) refers to directly querying the verbalized confidence of LLMs, which has recently been demonstrated as effective, particularly for large-scale RLHF-LMs. (2) **PE** is the predictive entropy of the model, as shown in Equation 1. It is the most fundamental method of measuring uncertainty based on probability. (3) **SAR** (Duan et al., 2023) (Shifting Attention to Relevance) is one of the latest uncertainty estimation methods based on probability, sampling and attention allocation.

Implementation Details We use the LLaMa-2-Chat (Touvron et al., 2023) and Mistral-7b-Instruct (Jiang et al., 2023a) as the backbone model. We use greedy search for all the generations and set the temperature as 0.5. The max length of each generation is 64 tokens for all the datasets. More implementation Details about inference and instruction tuning are shown in Appendix A.3. All the experiments are run on NVIDIA H800 GPU.

3.2 Effective Uncertainty Expression

Effectiveness The results presented in Table 1 demonstrate that our approach significantly enhances LLMs' awareness of self-uncertainty, expressing reliable and effective confidence within the domain. In comparison to directly querying verbalized confidence in vanilla LLMs, our method achieves a notable 45.2% AUROC increase on average. In contrast to the SAR method based on probability calculation and multiple sampling, our approach consistently outperforms it by 12.6%, surpassing its own "teacher". Moreover, such confidence expression incurs negligible time and computational costs during inference, as it is solely dedicated to generating a few tokens.

Generalizability To demonstrate the efficacy of our instruction tuning method beyond mere data set distribution fitting, we extended our evaluation to additional domains. Our findings indicate that the confidence expression capability inspired in LLMs exhibits a certain level of generalization and proves effective in the other two domains as well. LLMs demonstrate relatively better generalization on SciQA compared to MedQA, which may be attributed primarily to the high domain specificity of the medical field. Furthermore, the questions in MedQA were adapted from multiple-choice questions and accompanied by longer dis-

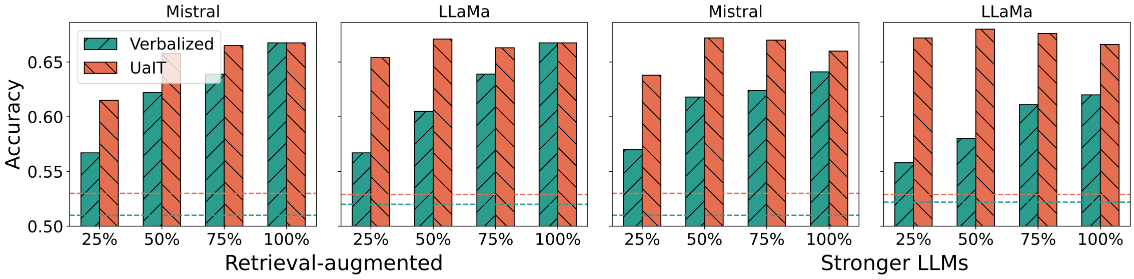

Figure 3: Accuracy of different proportions of low confidence samples with the assistance of retrieved evidence and stronger LLMs. The dashed line represents the accuracy without retrieval or powerful LLM assistance.

Model	Type	In-domain	Out-of-domain	
		TriviaQA	SciQA	MedQA
Mistral	Q	0.798	0.734	0.529
	Q+R	**0.846**	**0.775**	**0.582**
LLaMA2	Q	0.777	0.702	0.567
	Q+R	**0.867**	**0.730**	**0.574**

Table 2: AUROC for uncertainty expression. “Q” represents Query and “R” represents Response.

Model	Method	Accuracy	AUROC
Mistral	SAR	0.510	0.762
	SAR w/ft	0.530	0.778
	UaIT w/ft	0.530	**0.846**
LLaMA2	SAR	0.522	0.759
	SAR w/ft	0.529	0.780
	UaIT w/ft	0.529	**0.867**

Table 3: Effect of fine-tuning on accuracy and AUROC.

ease descriptions, imposing a significant challenge for the model’s comprehension abilities.

Effect of Thresholds during Distillation We explore the effect of distillation and different thresholds on uncertainty expression, as shown in Figure 2. It can be observed that fine-tuning on distilled data at all thresholds significantly improves performance, thereby demonstrating the effectiveness, robustness, and efficiency of this distillation process. Notably, using thresholds above 50% often yields more significant performance improvements.

Query vs. Query+Responce To investigate the sources of uncertainty, we also employ the input query as a basis for uncertainty assessment, training LLM to express uncertainty solely based on the query. Table 2 demonstrates that individual queries alone enable LLMs to express reasonable levels of uncertainty, possibly due to LLMs assessing uncertainty based on the similarity between the query and their pre-trained knowledge. However, incorporating both the query and response to determine uncertainty provides a more accurate assessment.

Accuracy vs. AUROC AUROC measures the correlation between accuracy and uncertainty. Our method fine-tunes the model on TriviaQA, utilizing answers generated by the model itself. To minimize the potentially significant impact of fine-tuning on accuracy, we show the accuracy and AUROC in Table 3. For fine-tuned models with equivalent accuracy, SAR results in only a slight improvement on AUROC, whereas significant progress is achieved through UaIT due to its superior calibration.

3.3 Uncertainty-aware Decision Making

To validate the effectiveness of uncertainty expression in practical human decision-making, we conducted experiments in two scenarios: knowledge retrieval (Liu et al., 2023; Wang et al., 2024) and stronger LLM assistance (Chen et al., 2023). We divide all samples into four equal parts based on their confidence levels and set corresponding thresholds, to trigger retrieval or employ more powerful LLM when LLM’s confidence falls below the thresholds. Figure 3 presents the accuracy of incorporating retrieval document and LLaMa2-13b at different proportions of low confidence levels. UaIT achieves significant improvements by incorporating additional knowledge at the lowest 25% confidence level, and relatively saturated performance is obtained by incorporating additional knowledge at the 50% confidence level. Compared to the Verbalized Confidence of vanilla model, UaIT better reflects knowledge gaps in uncertainty expression. More details and examples are in Appendix A.3 and C.

4 Related Work

Uncertainty estimation constitutes an essential step in developing reliable AI systems, which are instrumental in detecting unreliable responses characterized by hallucinations (Zhang et al., 2023; Agrawal et al., 2024) or factual errors (Bian et al., 2023; Karpinska and Iyyer, 2023) generated by LLMs.

Traditional uncertainty estimation methods have mainly focused on text classification (Vazhentsev et al., 2022; Ulmer et al., 2022; Jiang et al., 2021; Desai and Durrett, 2020) or regression (Wang et al., 2022; Glushkova et al., 2021; Zhan et al., 2023) tasks with clear and distinct labels. However, for free-form LLMs, multiple different but semantically equivalent generations can be considered correct. Recent research transformed the free-form questions into multiple-choice form to align with traditional categorical uncertainty estimation methods (Lin et al., 2022b; Shrivastava et al., 2023; Ye et al., 2024). Some recent works estimated the uncertainty by quantifying the consistency of multiple generations, computing predictive entropies with generations, or incorporating paraphrase detection (Geng et al., 2023; Malinin and Gales, 2021; Kadavath et al., 2022; Manakul et al., 2023; Sai et al., 2023; Bakman et al., 2024). Semantic Entropy (SE) (Kuhn et al., 2023) proposes the notion of "semantic equivalence" to aggregate generations with similar semantics. SAR (Duan et al., 2023) advocates assigning more attention to tokens and sentences with higher relevance.

The research on estimating uncertainty with the expression of LLM is still in its early stages. Recent research has explored various prompt strategies to enhance the uncertainty expression (Kadavath et al., 2022; Zhou et al., 2023; Tian et al., 2023). Lin et al. (2022a) group examples based on the mathematical computation type and fine-tune LLMs with the empirical accuracy of each group to predict the correctness of problem-solving. However, this group-based method, where the answer comprises solely a single numerical token, lacks generality in applications. Xiong et al. (2024) further combine direct expression and multi-sampling methods to achieve more accurate assessment. Kumar et al. (2024) analyze the correlation between internal model probability and the verbalized uncertainty expression. Another concurrent work develops a comprehensive framework that incorporates sampling, clustering, and the use of external LLMs (GPT-4) to generate rationales, to enhance the uncertainty expression (Xu et al., 2024). Our work focuses on enhancing uncertainty awareness in LLMs by simply aligning powerful probabilistic uncertainty estimation and utilizing output uncertainty as a basis for real-time human decision-making. We highlight that such a simplified approach that avoids extensive multi-sampling and reliance on external commercial LLMs (e.g. GPT4 or ChatGPT), is capable of demonstrating robust and immediate uncertainty expression in real-time interactions.

5 Conclusion

Expressing uncertainty by LLMs poses a significant challenge that has not been thoroughly explored. In this paper, we address this challenge by training the model to align the probabilistic uncertainty of its own generation, thereby enhancing the model's ability to perceive and express uncertainty. Experimental results demonstrate that the model not only exhibits strong uncertainty expression capabilities within the domain but also showcases promising generalization capabilities.

Limitations

Our study provides preliminary evidence of the effectiveness of uncertainty-aware instruction tuning. In the future, we aim to investigate how uncertainty perception is learned by incorporating different prompts and analyzing the interplay of the model's probability and attention distributions. Additionally, our fine-tuning process was conducted using a limited amount of data from a single domain. Exploring the optimal data balancing across different domains and scenarios, designing improved training strategies, incorporating more diverse prompts, and utilizing full-scale fine-tuning to achieve reliable and robust uncertainty-aware LLM remains an important avenue for further exploration.

It is also challenging and valuable to extend our method to more general scenarios and tasks, e.g. long-form QA and summarization, although this kind of exploration is still in its nascent stages (Huang et al., 2024). Most of the existing uncertainty estimation studies primarily focused on short-form generations. Applying our method to long-form generation also requires obtaining probabilistic uncertainty, e.g. assessing the uncertainty using token probabilities of reasoning text, which we leave as future work.

Acknowledgment

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF), the National Natural Science Foundation of China (Grant No. 62261160648), the Research Program of

Guangdong Province (Grant Nos. 2220004002576, EF2023-00090-FST), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011491), and the Tencent AI Lab Rhino-Bird Gift Fund (Grant No. EF2023-00151-FST). We appreciate the insightful suggestions from the anonymous reviewers and meta-reviewer.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 912–928. Association for Computational Linguistics.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. MARS: meaning-aware response scoring for uncertainty estimation in generative llms. *CoRR*, abs/2402.11756.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink makes a million think: The spread of false information in large language models. *CoRR*, abs/2305.04812.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *CoRR*, abs/2302.01318.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *CoRR*, abs/2307.01379.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *CoRR*, abs/2311.08298.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. *CoRR*, abs/2402.06544.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

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. *CoRR*, abs/2207.05221.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 419–451. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022b. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4160–4173, Dublin, Ireland. Association for Computational Linguistics.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. kNN-TL: k-nearest-neighbor transfer learning for low-resource neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1891, Toronto, Canada. Association for Computational Linguistics.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):26:1–26:39.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation. *CoRR*, abs/2311.08877.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Zhexuan Wang, Shudong Liu, Xuebo Liu, Miao Zhang, Derek Wong, and Min Zhang. 2024. Domain-aware k-nearest-neighbor knowledge distillation for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9458–9469, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Nineth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *CoRR*, abs/2405.20974.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *CoRR*, abs/2401.12794.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–820, Toronto, Canada. Association for Computational Linguistics.

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. *CoRR*, abs/2309.01219.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *CoRR*, abs/2401.06730.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Experiment Details

A.1 Datasets

We follow Duan et al. (2023) to select 2000 questions from the validation set of TriviaQA for evaluation. For SciQA and MedQA, we utilize the complete validation sets.

We only employ TriviaQA for instruction tuning, evaluate the model's in-domain performance on TriviaQA, and assess its cross-domain generalization abilities on SciQA and MedQA. We utilize the refined version of the TriviaQA dataset curated by Wang et al. (2023), which consists of 78,785 question-answers. After distillation (2.2), we ultimately train the model with 31,391 and 25,362 samples on LLaMa-2 and Mistral, respectively.

A.2 Metric

We employ Rouge-L (Lin, 2004) to measure the accuracy of the response generated by LLMs. Rouge-L calculates the longest common subsequence between the generated content and reference answers and considers it correct if it exceeds a predefined threshold. We set the threshold at 0.5. For MedQA with longer answers, we consider predictions that contain the complete golden answer to be correct.

Consistent with prior work (Kuhn et al., 2023; Duan et al., 2023), we evaluate the effectiveness of uncertainty by assessing the reliability of the model's generated content, i.e., whether the answer to a question is trustworthy. Specially, we employ the AUROC metric, which is considered a more suitable uncertainty evaluation measure for free-form generations (Kuhn et al., 2023; Xiong et al., 2024). The AUROC metric quantifies the likelihood of a correct answer having a lower uncertainty score compared to an incorrect answer. Higher AUROC indicates superior performance, with perfect uncertainty scoring 1 and random uncertainty measuring 0.5.

A.3 Implementation Details

We follow Duan et al. (2023) to generate 1 most likely generation with greedy search and 5 sentences for each question with multinomial sampling for uncertainty estimation in SAR. For PE, we maintain the same configurations as SAR, with the sole distinction being that the probabilistic averaging in this method is unweighted.

During the data filtering process, we empirically set the thresholds for LLaMa-2 and Mistral, as mentioned in Section 2.2, to 80 and 70 respectively. This means that the confidence scores of all correct and incorrect answers in the distilled training set are set to be higher and lower than these thresholds, respectively. After the distillation process, 47k and 53k samples are filtered out from the total of 79k, leaving less than only 40% remaining for instruction tuning.

Configuration	Value
Model	LLaMa2-7B-Chat Mistral-7B-Instruct
Epochs	4
Batch Size	32 samples
Max Length	512
Optimizer	Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning Rate	2×10^{-5}
LR scheduler	cosine
Warmup Ratio	0.1
LoRA Dropout	0.05
$lora_r$	64
$lora_\alpha$	16
Device	1 Tesla H800 GPU (80GB)

Table 4: Finetuning Details of LLaMa and Mistral.

For instruction tuning, the detailed parameters are presented in Table 4. We finetune the model by low-rank adaptation (LoRA) (Hu et al., 2022). Uncertainty-aware instruction tuning achieves excellent results with training on a single GPU for less than 3 hours.

In Experiment 3.3, our objective is to demonstrate that models with uncertainty-aware instruction tuning (UaIT) are better calibrated. Expressed uncertainty is considered truly reliable when it can be used for decision-making regarding trustworthiness and improving accuracy. Hence, we employ accuracy (rather than AUROC) as our measurement metric in this experiment. All the dashed lines and bars in Figure 3 correspond to the model's accuracy. Wang et al. (2023) provide five pertinent documents for each question in TriviaQA, and we utilize the top-ranked document as an external knowledge source to retrieve. For LLaMa-2-13b, we use its Chat version to ensure high performance.

B Prompt List

B.1 Uncertainty-aware Instruction Tuning

The following is the input prompt and output in uncertainty-aware instruction tuning (Session 2.2).

```
Please directly return the answer to the
 following question without any
explanation and indicate your level of
confidence. Note that the confidence
```

Question	Which journalist first told the world about the My Lai massacre?
Reference	Seymour Hers
Original Answer	Ronald Haeberle
Confidence Level	49.7%
Retrieval Document	Seymour Hersh Seymour Myron "Sy" Hersh (born April 8, 1937) is an American investigative journalist and political writer based in Washington, D.C. He is a longtime contributor to "The New Yorker" magazine on national security matters and has also written for the "London Review of Books" since 2013. Hersh first gained recognition in 1969 for exposing the My Lai Massacre and its cover-up during the Vietnam War, for which he received the 1970 Pulitzer Prize for International Reporting. During the 1970s, Hersh covered Watergate for The New York Times and revealed the clandestine bombing of Cambodia.
Corrected Answer	Seymour Hersh

Table 5: An example of uncertainty expression and retrieval correction.

```
level indicates the degree of certainty
you have about your answer and is
represented as a percentage. For
instance, if your confidence level is
80.0%, it means you are 80.0% certain
that your answer is correct and there is
 a 20.0% chance that it may be incorrect
.
Question: [QUESTION]

Answer and Confidence (1-100): [ANSWER];
 [CONFIDENCE]%
```

B.2 Direct Query for Verbalized Confidence

The following is the prompt in Verbalized uncertainty expression.

```
Here are some examples.
Question: What type of organism is
commonly used in preparation of foods
such as cheese and yogurt?
Answer and Confidence (0-100):
mesophilic organisms; 90%
Question: What phenomenon makes global
winds blow northeast to southwest or the
 reverse in the northern hemisphere and
northwest to southeast or the reverse in
 the southern hemisphere?
Answer and Confidence (0-100): coriolis
effect; 70%
Question: Changes from a less-ordered
state to a more-ordered state (such as a
 liquid to a solid) are always what?
Answer and Confidence (0-100):
exothermic; 80%

According to the format of the above
examples, directly write the answer with
 one or few words to the following
question without any explanation and
indicate your level of confidence. Note
that the confidence level indicates the
degree of certainty you have about your
answer and is represented as a
percentage. For instance, if your
confidence level is 80%, it means you
are 80% certain that your answer is
correct and there is a 20% chance that
it may be incorrect.
Question: [QUESTION]
Answer and Confidence (0-100):
```

B.3 UaIT based on Query

The following is the input prompt and output in uncertainty-aware instruction tuning based on the query (Experiments 3.2).

```
Please directly give your confidence
level that you can answer the following
question correctly, and then directly
return the answer without any
explanation. Note that the confidence
level indicates the degree of certainty
you have about your answer and is
represented as a percentage. For
instance, if your confidence level is
80%, it means you are 80% certain that
your answer is correct and there is a
20% chance that it may be incorrect.
Question: [QUESTION]

Confidence (1-100) and Answer:[
CONFIDENCE]%; [ANSWER]
```

B.4 Decision Making with Retrieval

The following is the prompt using the retrieval documents.

```
Please directly return the answer to the
 following question without any
explanation and indicate your level of
confidence. Note that the confidence
level indicates the degree of certainty
you have about your answer and is
represented as a percentage. For
instance, if your confidence level is
```

```
80.0%, it means you are 80.0% certain
that your answer is correct and there is
 a 20.0% chance that it may be incorrect
.
Question: "[DOCUMENT]" According to this
 passage, [QUESTION]
Answer and Confidence (0-100):
```

C Case Study

Table 5 illustrates an example of utilizing LLM-expressed uncertainty as the basis for knowledge retrieval. LLM exhibits low confidence in the given question and its own answer, and subsequently corrects the erroneous answer upon incorporating a document containing external knowledge.

D Supplementary Information on the Use of SAR

Given that our method employs SAR (Duan et al., 2023) to provide uncertainty scores, we include additional explanations to elucidate this approach. The core idea of SAR is that tokens are not created equally in presenting semantics and should not be treated equally when estimating uncertainty. For example, consider a given question, "What is the ratio of the mass of an object to its volume?" and a model generation "density of an object." Clearly, "density" is the most relevant token in conveying the semantics than the rest tokens. Therefore, relevance weights (R_T in Equation 2) are proposed to measure the importance of each token by comparing the semantic changes before and after removing it from the generation. Formally, for an input $\boldsymbol{x}$, an output $\boldsymbol{y}$, and a token y_t within $\boldsymbol{y}$, the relevance of y_t can be expressed as:

$$R_T\left(y_t, \boldsymbol{y}, \boldsymbol{x}\right)=1-\left|g\left(\boldsymbol{x} \cup \boldsymbol{y}, \boldsymbol{x} \cup \boldsymbol{y} \backslash\left\{y_t\right\}\right)\right| \quad (6)$$

where $g(\cdot, \cdot)$ represents any semantic similarity measure, providing a similarity score between 0 and 1. A larger semantic change indicates higher relevance weights for that token, and vice versa. Relevance weights are then used to compute a weighted average of log probabilities.

SAR also extends the aforementioned token-level relevance weights to the sentence-level. Previous methods often generate multiple generations for the same question (multi-sampling) and improve performance by averaging the probabilities of these generations. SAR claims that generations (i.e. sentences) are more persuasive when they exhibit semantic similarity with other generations. Therefore, they define sentence-level relevance weights (R_S in Equation 3) as the probability-weighted semantic similarity with other sentences:

$$R_S\left(\boldsymbol{y}_i, Y, \boldsymbol{x}\right)=\sum_{j=1, j \neq i} g\left(\boldsymbol{y}_i, \boldsymbol{y}_j\right) \operatorname{PE}\left(\boldsymbol{y}_j \mid \boldsymbol{x}\right) \quad (7)$$

Here, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ represent two distinct responses from the set of all responses $\boldsymbol{Y}$, and $\operatorname{PE}(\cdot)$ corresponds to Equation (1) in the referenced paper. If a generation is more semantically similar to other generations, its relevance weight is higher.