Revisiting Epistemic Markers in Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty?

Jiayu Liu, Qing Zong, Weiqi Wang, Yangqiu Song
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
jliufv@connect.ust.hk; {qzong, wwangbw, yqsong}@cse.ust.hk

Abstract

As Large Language Models (LLMs) are increasingly used in high-stakes domains, accurately assessing their confidence is crucial. Humans typically express confidence through epistemic markers (e.g., "fairly confident") instead of numerical values. However, it remains unclear whether LLMs reliably use these markers to reflect their intrinsic confidence due to the difficulty of quantifying uncertainty associated with various markers. To address this gap, we first define *marker confidence* as the observed accuracy when a model employs an epistemic marker. We evaluate its stability across multiple question-answering datasets in both in-distribution and out-of-distribution settings for open-source and proprietary LLMs. Our results show that while markers generalize well within the same distribution, their confidence is inconsistent in out-of-distribution scenarios. These findings raise significant concerns about the reliability of epistemic markers for confidence estimation, underscoring the need for improved alignment between marker based confidence and actual model uncertainty. Our code is available at https://github.com/HKUST-KnowComp/MarCon.

1 Introduction

LLMs have grown increasingly powerful, yet their application in mission-critical tasks is still hindered by reliability issues (Zhang et al., 2024; Maynez et al., 2020). Therefore, accurately measuring output confidence is crucial for their reliable deployment (Li et al., 2024a; Pedapati et al., 2024; Beigi et al., 2024). Traditionally, black-box confidence estimation in LLMs primarily relies on direct numerical outputs (e.g., "30% confidence") or response consistency (Xiong et al., 2024; Chen and Mueller, 2024; Li et al., 2024b), while white-box methods mainly utilize logits, internal states as information source (Geng et al., 2024). However, natural language is the primary interface for human-LLM interaction. Instead of relying solely on abstract numerical measures, humans often use epistemic markers, such as "I am not sure" or "it is unlikely that," to convey uncertainty (Wallsten, 1986; Erev and Cohen, 1990; Juanchich et al., 2017; Kadavath et al., 2022). This similar recognition of uncertainty markers is essential for effective communication (Willems et al., 2019; Belém et al., 2024), which potentially makes it valuable for LLMs to adopt a similar practice (Yona et al., 2024).

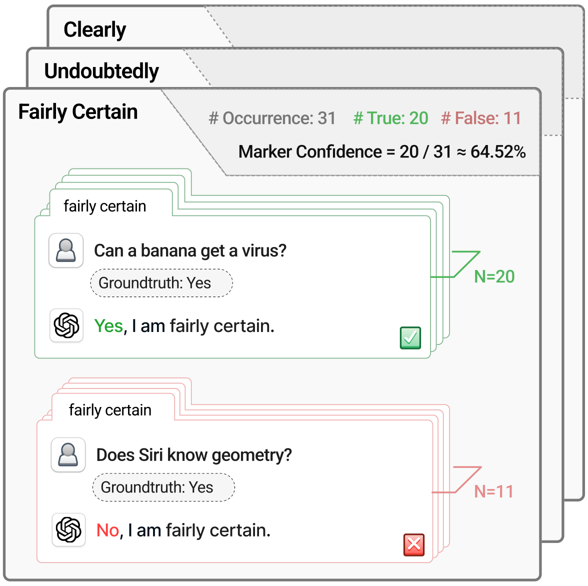

Figure 1: An example of our framework calculating the marker confidence of "fairly certain" for *GPT-4o* on StrategyQA. We calculate the confidence for all the markers across seven models and seven datasets.

However, it remains unexplored whether LLMs are capable of incorporating these epistemic markers in their responses to express their intrinsic confidence stably and consistently. Previous works have primarily concentrated on the misalignment between human and LLM recognition of epistemic markers (Zhou et al., 2024; Tang et al., 2024; Belém et al., 2024), concluding that models always fail to accurately convey confidence in words (Yona

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 206–221
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024). In fact, human interpretations of markers are not completely identical (Pennekamp et al., 2024), so even if these markers may not align well with human reasoning, they can still be useful if the model maintains a consistent internal mapping between markers and their actual accuracy. Thus, previous studies questioning the reliability of markers may be insufficient, as they have not examined whether LLMs can consistently apply their own confidence framework.

To address this gap, we investigate whether epistemic markers produced by LLMs reliably reflect their confidence in question-answering tasks. By defining marker confidence as the accuracy of responses when a model uses a specific marker to convey confidence, we calculate the marker confidence with various models and datasets. Additionally, we propose seven evaluation metrics to systematically assess the stability of these markers in both in-distribution and out-of-distribution contexts. Our findings show that while markers perform well within similar distributions, their stability declines in out-of-distribution contexts. Additionally, we compare a range of widely used models and conclude that the more powerful ones demonstrate a better understanding of epistemic markers.

2 Related Work

Our work primarily intersects with confidence estimation in LLMs and studies about epistemic markers. Please find related works in Appendix A.

3 Study Design

3.1 Formalization

Confidence of Epistemic Markers. Let W denote an epistemic marker, $D = \{Q_1, Q_2, \ldots, Q_n\}$ a labeled dataset, and M a model. We define the confidence associated with each epistemic marker as $\mathit{Conf}(W, D, M)$, computed as the accuracy of the answers that explicitly include marker W when the model provides responses. It is important to note that our definition of marker confidence deviates from the conventional interpretation of verbal confidence, which pertains to the semantic uncertainty associated with epistemic markers. To compute the marker confidence for a specific epistemic marker W_i, we use model M_k to generate answers for all questions in the training set of dataset D_j and then extract the subset $Q_{W_i} \subseteq D_j$ consisting of questions whose generated answers contain W_i. The marker confidence is defined as:

$$\mathit{Conf}(W_i, D_j, M_k) = \frac{1}{|Q_{W_i}|} \sum_{q \in Q_{W_i}} \mathbb{I}\Big(M_k(q)\Big),$$

where Q_{W_i} is the set of questions whose generated answers contain the epistemic marker W_i and $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the answer generated by M_k for question q is correct, and 0 otherwise. An example is provided in Figure 1.

3.2 Methods

We calculate $\mathit{Conf}(W_i, D_j, M_k)$ for all combinations of generated markers, datasets and models (W_i, D_j, M_k) in Appendix B.1 to provide an all-rounded insight into the marker distributions. Specifically, we propose seven metrics to systematically evaluate the stability and consistency of LLM generated epistemic markers:

(1) I-AvgECE *In-domain Average ECE* reflects how well the marker confidence of the model aligns with its actual accuracy in a consistent setting within the same distribution.

(2) C-AvgECE *Cross-domain Average ECE* assesses the calibration error of the marker confidence and the actual accuracy, further reflecting the robustness of the model's marker confidence in out-of-domain scenarios.

(3) NumECE *Numerical* ECE measures the overall calibration of the model's numerical confidence outputs across all datasets. All ECE-related metrics are desired with a lower value, indicating better calibration performance on the target dataset.

(4) MAC *Marker Accuracy Correlation* reflects the correlation between marker confidence and the model accuracy on different datasets. The metric is calculated based on Pearson coefficient, so 0 in this value represents no linear correlation and 1 indicates direct propotional relationship between the marker confidence and model's accuracy.

(5) MRC *Marker Ranking Correlation* measures the model's ability to maintain a consistent marker confidence ranking across different datasets. The metric is calculated based on Spearman coefficient (de Winter et al., 2016), so 0 in this value represents no correlation and 1 indicates totally identical between markers' confidence rankings.

(6) I-AvgCV *In-domain Average CV* captures the dispersion of the model-generated confidence scores within each dataset. A relatively higher I-AvgCV value indicates a more decentralized distribution of markers within the dataset, demonstrating

Model	Marker Confidence					Rank	Density
	I-AvgECE ↓	C-AvgECE ↓	NumECE ↓	C-AvgCV ↓	MAC	MRC ↑	I-AvgCV
Llama-3.1-8B-Instruct	10.09	15.95	22.70	20.80	60.91	11.37	20.48
Qwen2.5-7B-Instruct	7.85	23.60	21.84	31.29	68.06	11.85	22.39
Qwen2.5-14B-Instruct	7.66	20.38	17.98	26.44	73.95	34.60	23.83
Qwen2.5-32B-Instruct	**4.78**	10.40	8.86	19.24	78.20	**36.97**	16.26
Mistral-7B-Instruct-v0.3	10.58	24.81	24.46	28.52	84.57	10.54	21.01
GPT-4o	8.55	**11.84**	**7.56**	**15.72**	76.44	27.54	14.30
GPT-4o-mini	7.65	17.15	12.79	21.98	87.68	16.48	20.61
Average	8.17	17.73	16.60	23.43	75.69	21.34	19.84

Table 1: Model performance across seven metrics. For each metric, the data for the best performing model is bolded. For analytical experiments about markers, we only consider those appear no less than 10 times to eliminate the effect of randomness (see Section 4 for details). All values listed in the table are expressed as percentage (%).

a stronger ability to distinguish between different markers.

(7) C-AvgCV *Cross-domain Average CV* measures the consistency of the model's marker-based confidence across different datasets. A higher C-AvgCV value indicates a greater dispersion of marker confidence across various datasets, suggesting the model's instability regarding marker confidence.

More details about the design and implementation of the metrics can be found in Appendix B.4.1.

4 Experiments and Analysis

Models and Datasets. We experiment with two mainstream open-source and five proprietary LLMs over seven datasets from various domains. More introduction can be found in Appendix B.1.

Baseline. Inspired by previous comparison about using numerical values and uncertainty expression in words to express confidence level (Jaffe-Katz et al., 1989; Knapp et al., 2016), we apply the method of directly prompting the model to express a numerical confidence as baseline for comparison. The prompt designed to elicit epistemic markers and numerical confidence is in Appendix B.1. Our main experiment results are in Table 1.

Marker Filtering Strategies We conduct all marker analysis experiments (namely C-AvgCV, MAC, MRC, and I-AvgCV) by filtering markers that occur fewer than 10 times in the training set in Table 1. The filtering threshold is eventually a tradeoff between the completeness and reliability of the data. On one hand, if the filtering threshold (10 in the main table) is too small, the confidence interval of the results would be large. On the other hand, we cannot include diverse markers for the metrics and get limited insights. The confidence intervals and more details are reported in Appendix B.4.2.

4.1 Main Observations

While the in-distribution marker confidence is relatively stable, it lacks robustness across different datasets. This conclusion is supported by the observation that I-AvgECE values consistently remain lower than NumECE and C-AvgECE values are notably higher than NumECE for 6 out of 7 models, indicating that models exhibit shortcomings in generalizing marker confidence to different distributions and datasets.

More direct evidence to support the conclusion may be inferred from the C-AvgCV of the models. The average C-AvgCV of 0.2343 highlights that marker confidence is highly sensitive to distribution shifts, aligning with the observed high value in C-AvgECE. Notably, we observed that stronger models (e.g., *GPT-4o, Qwen2.5-32B-Instruct*) might exhibit smaller C-AvgCV values.

We further examine the relationship between marker confidence and model accuracy across different datasets using MAC. For 5 out of 7 of the models, the MAC value is over 0.7, which indicates that marker confidence are positively related to the model's accuracy in a strict manner. This also suggests that marker confidence is fragile under distribution shifts, highlighting models' lack of robust understanding of epistemic markers.

Models fail to maintain a consistent ordering of epistemic markers across different domains. The overall low values of MRC suggest that models do not preserve a consistent ranking of markers when applied to datasets with different distributions. We notice that larger models appear to have a better grasp at maintaining a stable ordering of markers. However, both the maximum and average MRC indicates low consistency performance, suggesting a lack of robust understanding of marker

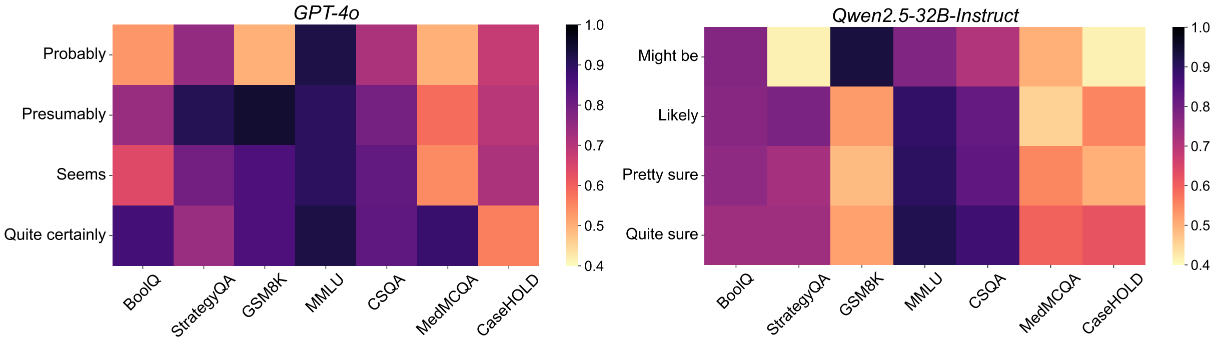

Figure 2: **Model's marker confidence varies greatly across different datasets.** We plot the heatmap of the marker confidence of *GPT-4o* and *Qwen2.5-32B-Instruct* across different datasets, illustrating that even the best models exhibit substantially different confidence levels in various contexts. The markers in the graph are randomly selected from the shared markers of all datasets.

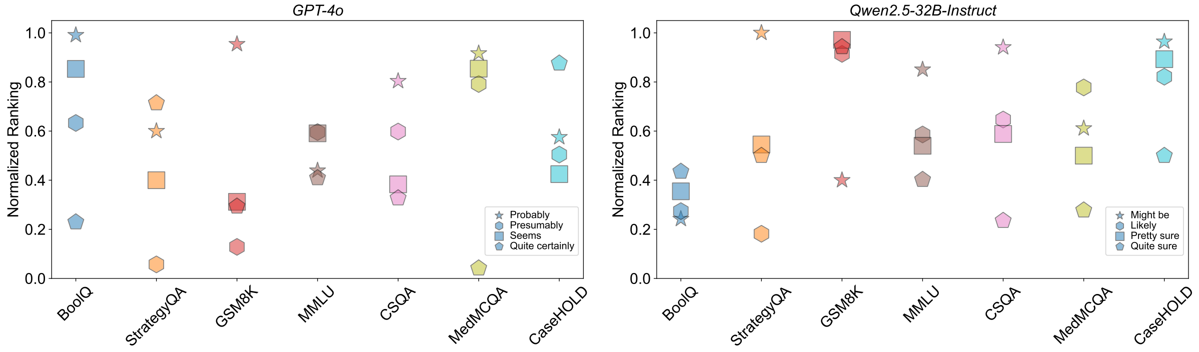

Figure 3: **The rankings of the model's marker confidence fluctuates significantly across different datasets.** We plot the scatter diagram for the marker confidence rankings of the best performing models, but still discovered that the rankings are extremely unstable. The markers in the graph are randomly selected from the shared markers across all datasets.

relative confidence.

The values of the marker confidence are highly concentrated. Since models are expected to express a wide range of confidence including extreme values which is necessary in mission-critical senarios (Alam et al., 2017; Bhise et al., 2018), We expect the models to clearly differentiate the epistemic markers by obtaining a relatively uniform distribution and containing markers with a confidence near 0% or 100%. However, we found the I-AvgCV values typically range from approximately 0.14 and 0.24, demonstrating a concentrated distribution with minor difference. Additionally, only 4 out of 49 settings (dataset, model pair) include markers with confidence under 10% when only those occur no less than 10 times are counted, indicating significant failure in expressing uncertainty.

4.2 Correlation between Performance and Marker Consistency

In Section 4.1, we notice that the statistics in Table 1 suggest that larger models demonstrate a better understanding of epistemic markers, as evidenced by lower C-AvgCV values and higher MRC values. In this section, we quantitatively evaluate the relationship between a model's accuracy and its corresponding C-AvgCV and MRC values to gain a deeper insight into the relationship of model ability and mastery of epistemic markers.

Specifically, for a given model M_k, we use its average accuracy across all datasets as a comprehensive measure of its performance. We then compute the Pearson Correlation Coefficient between each model's overall accuracy and both its C-AvgCV and MRC. The results show a correlation coefficient of -0.88 between model accuracy and C-AvgCV, and a correlation coefficient of 0.75 between model accuracy and MRC. These findings indicate a strong negative relationship between model accuracy and C-AvgCV and a strong positive relationship between model accuracy and MRC. This suggests that more powerful models exhibit greater stability in marker confidence across datasets, as well as a more consistent ordering of markers.

Model	C-AvgCV ↓		MAC		MRC ↑		I-AvgCV	
Threshold	**50**	**100**	**50**	**100**	**50**	**100**	**50**	**100**
Llama-3.1-8B-Instruct	25.50	23.39	77.54	65.98	-8.57	-10.32	11.44	8.59
Qwen2.5-7B-Instruct	30.12	29.30	77.96	69.03	6.01	2.25	18.38	15.41
Qwen2.5-14B-Instruct	26.49	27.66	92.03	94.25	36.56	36.11	20.59	21.13
Qwen2.5-32B-Instruct	20.03	21.27	87.25	86.85	34.91	23.95	13.51	12.18
Mistral-7B-Instruct-v0.3	26.70	26.34	94.80	84.71	30.68	30.60	12.81	9.67
GPT-4o	**15.86**	**16.92**	89.91	90.68	**37.38**	**39.40**	7.52	7.29
GPT-4o-mini	22.15	22.41	86.82	87.45	24.16	24.36	11.98	11.39
Average	23.84	23.90	86.62	82.71	23.02	20.91	13.75	12.24

Table 2: This table presents the results (in %) of the marker analysis experiments, organized by different filtering thresholds. We observed that the conclusion obtained from Table 1 still holds when the threshold increases.

4.3 Conclusion Robustness under Different Filtering Thresholds

Our primary conclusions regarding marker consistency and model understanding of epistemic markers remain robust across various filtering thresholds. As detailed in Table 2, even when increasing the filtering threshold to 50 or 100 occurrences, the observed trends persist: 1) The C-AvgCV and MRC values consistently remain low, while the I-AvgCV values remain high. This pattern continues to suggest that models struggl to maintain consistent confidence across different datasets and to differentiate effectively between markers. 2) The MAC values consistently remain high. This reinforces the strong positive relationship between model performance and its marker usage ability. Moreover, **the observed shortcomings extend even to frequently occurring markers** (with over 100 instances), indicating a severe challenge in the model's ability to reliably utilize epistemic expressions. This pervasive issue further substantiates our claims regarding the difficulties in ensuring reliable marker usage.

4.4 Discussions

This section aims to explain the observed differential in the consistent deployment of epistemic markers within in-domain versus out-of-domain contexts. Furthermore, we aim to investigate the distributional characteristics that potentially modulate the confidence associated with these markers.

Within similar distributions, models tend to maintain a stable "preference" for employing these markers to express uncertainty. For instance, in in-domain scenarios, responses on the test set leverage a consistent pattern of epistemic marker "preference" due to the similarity of the distribution. However, this consistency breaks down when the data distribution shifts. As the distribution changes, the model's preference usage for epistemic markers also changes, hindering the transfer of marker-based confidence to test sets from different datasets.

To further investigate this phenomenon, we conducted an in-depth analysis by calculating the average ECE values for each dataset to quantify the generalizability of the marker consistency from the training set to the test set. This involved averaging the ECE values across seven models in the in-domain scenario. Our analysis revealed that **datasets with greater component diversity generally exhibit a higher ECE**. This suggests that more complex or multi-sourced distributions make it challenging to transfer marker confidence from the training to the test set, as exemplified by datasets like MMLU and StrategyQA, which are not domain-specific. This observation further supports our claim that models perform well in in-domain scenarios primarily due to the consistency of the underlying data distribution. The details about the experiments are reported in Appendix C.

5 Conclusion

Our study evaluates whether LLMs can reliably express confidence using epistemic markers. We define marker confidence as the observed accuracy of responses containing specific markers, conduct extensive experiments and evaluate the results with several metrics. The results show that the marker confidence shifts significantly under distribution changes, following the trend of model accuracy, which highlights poor stability. Additionally, models struggle to effectively differentiate between markers and maintain consistent marker rankings across datasets. These findings suggest that the LLM generated markers to express their confidence is unreliable and requires improved alignment between verbal confidence and actual performance. Our work contributes to more consistent confidence estimation frameworks, ultimately facilitating reliable and trustworthy LLM responses.

Limitation

Human language is remarkable for its complexity, variability, and rich connotations, particularly when expressing uncertainty. Within a complete linguistic system, factors like sentence structure can significantly influence confidence, which is often difficult to quantify with epistemic markers alone. Moreover, in the context of long-form communication, it is clear that confidence cannot be simply measured by the confidence values of epistemic markers. To facilitate simplicity in evaluation and to focus on the study of epistemic markers, we adopt a relatively idealized approach: using epistemic markers generated by LLMs in closed-source QA tasks to represent the confidence of the responses while keeping them relatively brief. Additionally, epistemic markers may carry different meanings across cultures and languages. However, we only consider epistemic markers in English.

Despite our idealized conditions and using state-of-the-art models, LLMs still fail to consistently align epistemic markers with their true confidence levels, revealing that the issue lies not only with our approach but also with the models themselves. While they perform well in question-answering tasks, they do not truly understand epistemic markers (Zhou et al., 2023), struggle to express consistent confidence in these markers, and have difficulty aligning their confidence expressions with human expectations (Belém et al., 2024). This points to a deeper challenge in model behavior, suggesting that future research should focus on addressing fundamental gaps in model linguistic alignment.

Ethics Statement

Our work contains no offensive contents and have no potential risks.

Licenses. We will share our code under the MIT license, allowing other researchers free access to our resources for research purposes. The datasets used in this paper, including BoolQ (Clark et al., 2019), StrategyQA (Geva et al., 2021), GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021), CSQA (Talmor et al., 2019), MedMCQA (Pal et al., 2022), and CaseHOLD (Zheng et al., 2021), are shared under either the CC BY-SA license, Apache License Version 2.0, or the MIT License, all of which permit their use for research purposes. As for language models, we access all open-source LMs via the Huggingface Hub (Wolf et al., 2020). Our use of *GPT-4o* (OpenAI, 2024b) and *GPT-4o-mini* (OpenAI, 2024a) is conducted through OpenAI's official website[1]. All associated licenses permit user access for research purposes, and we have agreed to adhere to all terms of use.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong,SAR, China.

References

Rahul Alam, Sudeh Cheraghi-Sohi, Maria Panagioti, Aneez Esmail, Stephen Campbell, and Efharis Panagopoulou. 2017. Managing diagnostic uncertainty in primary care: a systematic critical review. *BMC Family Practice*, 18:1–13.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. Internalinspector i^2: Robust confidence estimation in llms through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 12847–12865. Association for Computational Linguistics.

Catarina G. Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8467–8502. Association for Computational Linguistics.

Viraj Bhise, Suja S Rajan, Dean F Sittig, Robert O Morgan, Pooja Chaudhary, and Hardeep Singh. 2018. Defining and measuring diagnostic uncertainty in medicine: a systematic review. *Journal of general internal medicine*, 33:103–115.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

[1] https://platform.openai.com/docs/api-reference/introduction

Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8301–8327. Association for Computational Linguistics.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5186–5200. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Joost C. F. de Winter, Samuel D. Gosling, and Jeff Potter. 2016. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273–290.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5050–5063. Association for Computational Linguistics.

Ido Erev and Brent L Cohen. 1990. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45(1):1–18.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:539–555.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *CoRR*, abs/2306.04459.

Amanda Jaffe-Katz, David V Budescu, and Thomas S Wallsten. 1989. Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition*, 17(3):249–264.

Zahra Jalilibal, Amirhossein Amiri, Philippe Castagliola, and Michael B. C. Khoo. 2021. Monitoring the coefficient of variation: A literature review. *Comput. Ind. Eng.*, 161:107600.

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. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Marie Juanchich, Amélie Gourdon-Kanhukamwe, and Miroslav Sirota. 2017. "i am uncertain" vs "it is uncertain". how linguistic markers of the uncertainty source affect uncertainty communication. *Judgment and Decision Making*, 12(5):445–465.

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.

Peter Knapp, Peter H Gardner, and Elizabeth Woolf. 2016. Combined verbal and numerical expressions increase perceived risk of medicine side-effects: a randomized controlled trial of ema recommendations. *Health Expectations*, 19(2):264–274.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2024. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *CoRR*, abs/2410.20774.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024a. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *CoRR*, abs/2403.09972.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024b. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11858–11875. Association for Computational Linguistics.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang, and Yangqiu Song. 2024. Gprooft: A multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 118–129.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. *OpenAI*.

OpenAI. 2024b. Hello gpt-4o. *OpenAI*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Tejaswini Pedapati, Amit Dhurandhar, Soumya Ghosh, Soham Dan, and Prasanna Sattigeri. 2024. Large language model confidence estimation via black-box access. *CoRR*, abs/2406.04370.

Pia Pennekamp, Jamal K. Mansour, and Rhiannon J. Batstone. 2024. Variability in verbal eyewitness confidence. *Applied Cognitive Psychology*.

Limin Su and Huimin Li. 2021. Project procurement method decision-making with spearman rank correlation coefficient under uncertainty circumstances. *Int. J. Decis. Support Syst. Technol.*, 13(2):16–44.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Zhisheng Tang, Ke Shen, and Mayank Kejriwal. 2024. An evaluation of estimative uncertainty in large language models. *arXiv preprint arXiv:2405.15185*.

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

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Lyudmila Rvanova, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *CoRR*, abs/2406.15627.

Thomas S Wallsten. 1986. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4):348–365.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Chen Luo, Sheikh Muhammad Sarwar, Yang Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. 2025. Ecomscriptbench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. *Preprint*, arXiv:2505.15196.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. 2024a. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2351–2374. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, and Yangqiu Song. 2024b. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. *CoRR*, abs/2406.10885.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. MARS: benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Sanne J. W. Willems, Casper J. Albers, and Ionica Smeets. 2019. Variability in the interpretation of dutch probability phrases - a risk for miscommunication. *Preprint*, arXiv:1901.09686.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Hui Xiong, Shashi Shekhar, Pang-Ning Tan, and Vipin Kumar. 2004. Exploiting a support-based upper bound of pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 334–343. ACM.

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 Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *CoRR*, abs/2410.02736.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7752–7764. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1946–1965. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings. In *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, pages 159–168. ACM.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3623–3643. Association for Computational Linguistics.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5506–5524. Association for Computational Linguistics.

Qing Zong, Zhaowei Wang, Tianshi Zheng, Xiyu Ren, and Yangqiu Song. 2024. Comparisonqa: Evaluating factuality robustness of llms through knowledge frequency control and uncertainty. *CoRR*, abs/2412.20251.

Appendices

A Related Work

Confidence Estimation in LLMs. Confidence estimation in LLMs refers to the process of assessing a model's confidence in its output. Previous research on this topic can be categorized into white-box and black-box methods, distinguished by whether they utilize the model's internal information. White-box methods leverage the internal states of LLMs, with key approaches including information-based methods that analyze these inner states (Vashurin et al., 2024; Geng et al., 2024; Burns et al., 2023), such as perplexity (Fomicheva et al., 2020; Zong et al., 2024), the negative log probability of generated tokens (Duan et al., 2024), and others. In the field of black-box methods, Lin et al. (2022) first introduces the concept *verbal confidence* that prompts LLM to output its confidence directly. Most subsequent methods are based on either directly prompting the model to generate an output or consistency sampling (Lin et al., 2024; Xiong et al., 2024; Chen and Mueller, 2024; Liu et al., 2024). However, previous methods primarily focus on processing numerical values to estimate the LLM's confidence, leading to a research gap in exploring LLM confidence expression through linguistic patterns, especially epistemic markers.

Studies on Epistemic Markers. Epistemic markers are essential for expressing confidence in conversation, playing a key role in human-LLM interactions (Hu et al., 2023). Recent studies have examined the reliability of LLMs in mastering epistemic markers, primarily investigating their interpretation of various uncertainty expressions. For instance, Tang et al. (2024) and Belém et al. (2024) use sentence templates with uncertainty expressions to prompt the model to assess overall confidence, though this approach is limited by the fixed nature of the templates, restricting generalizability. Zhou et al. (2023) and Zhou et al. (2024) argue that LLMs often mimic marker distributions from training data rather than truly understanding them, with the latter highlighting a tendency for overconfidence. Lee et al. (2024) also examine epistemic markers but focus on robustness and biases in model interpretations. While these studies investigate LLM generation of epistemic markers, our work aligns most closely with Yona et al. (2024), which directly challenges the ability of LLMs to accurately convey confidence using epistemic markers. However, they employ LLM-as-a-judge and few-shot prompting to assess the numerical confidence of uncertainty expressions, introducing potential bias (Chen et al., 2024; Ye et al., 2024; Ma et al., 2023). Additionally, they use human judges to assess the quality of LLM judges, which essentially aligns the model's interpretation of markers with human understanding. This approach is similar to other works in the field, which also focus on aligning human and LLM recognition of epistemic markers. However, as long as LLMs maintain a consistent framework for incorporating epistemic markers to express confidence, we can learn from its intrinsic mapping of marker usage and align it with human expectations through external means.

B Technical Details

B.1 Experiment Setup

Models We incorporate a range of commonly used LLMs of different scales, including Llama-3.1-8B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct (Yang et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), GPT-4o (OpenAI, 2024b), GPT-4o-mini (OpenAI, 2024a). For all models, we use a temperature of 0.5 to balance between logical consistency and creativity. All the open-source models are run on 4 NVIDIA A6000 (40G) GPUs with BF16.

We observed that instruction-tuned models exhibit greater variation and demonstrate better linguistic diversity when using epistemic markers. Although we also tested some base models, we found that for the same dataset, they emitted significantly fewer markers compared to the instruction-tuned models (See Figure 4) . As a result, we chose to focus our experiments on instruction-tuned models instead.

Datasets We benchmark the LLMs' responses with confidence expression using epistemic markers using the following seven datasets requiring knowledge in different domains: 1) Factual and commonsense knowledge: BoolQ (Clark et al., 2019), StrategyQA (Geva et al., 2021), CSQA (Talmor et al., 2019). 2) Mathematical reasoning: GSM8K (Cobbe et al., 2021). 3) Medical reasoning: MedMCQA (Pal et al., 2022). 4) Law reasoning: CaseHOLD (Zheng et al., 2021). 5) Mixed factual datasets: MMLU (Hendrycks et al., 2021).

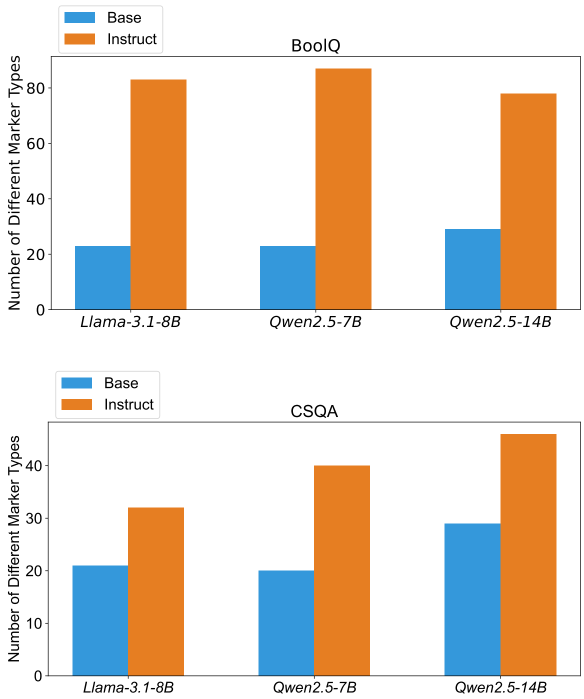

Figure 4: The number of epistemic markers that six different models generated in BoolQ and CSQA dataset. The results indicate that the instruct-tuned models exhibit much better linguistic diversity than base models in expressing confidence, which is desired by our experiment.

Prompts The prompts used in our experiments are shown in Table 4. We performed five permutations on the prompt with *Qwen2.5-7B-Instruct* and *Mistral-7B-Instruct-v0.3* on BoolQ, StrategyQA, CSQA, MedMCQA and found that both accuracy and C-AvgCV, MAC, MRC and I-AvgCV values varied only slightly. Consequently, we randomly selected one version and conducted all subsequent experiments using it.

B.2 Model Sources

This section clarifies the sources of the models used in our study. For methods involving LLMs, we utilize their instruction fine-tuned versions (see Appendix B.1 for more details) accessed via the Hugging Face Hub (Wolf et al., 2020). Specifically, for *Llama-3.1-8B-Instruct*, we employ the version *meta-llama/Llama-3.1-8B-Instruct*. The models related to Qwen include *Qwen/Qwen2.5-7B-Instruct*, *Qwen/Qwen2.5-14B-Instruct*, and *Qwen/Qwen2.5-32B-Instruct*. For *Mistral-7B-Instruct-v0.3*, we use *mistralai/Mistral-7B-Instruct-v0.3*.

B.3 Detailed Implementation

Data Preprocessing We applied data preprocessing methods to the datasets. For GSM8K, we transformed it into a binary-question dataset for convenience. Specifically, for half of the questions Q_i with even index in GSM8K, we first extracted the correct answer A_i, then used a question template to create a binary question by incorporating (Q_i, A_i) and setting the correct binary answer to "yes." For the remaining half of the questions Q_j, we randomly selected an answer A_j different from the correct answer, and used the same question template to create a binary question by incorporating (Q_j, A_j) and setting the correct binary answer to "no." The question template and two examples are given in Table 5. For the training set of the MMLU, due to its massive size, we randomly sampled a subset of 20000 QA-pairs for the MMLU training set. For MedMCQA dataset, for the convenience of evaluation, we pick the subset of the answer with only one correct answer. We then randomly sample 9686 QA-pairs for the MedMCQA training set and 2422 for its test set. For BoolQ, we did not expose the model to the "passage" part and treated it as a closed-book question-answering dataset in our experiment. Since CaseHOLD isn't explicit split into training set and test set, we divide the former 80% as training set and the rest as test set. A detailed statistics for our dataset usage is on Table 3.

Dataset	Train Size	Test Size
BoolQ	9427	3270
StrategyQA	2061	229
GSM8K	7473	1319
MMLU	20000	14041
CSQA	9741	1221
MedMCQA	9686	2422
CaseHOLD	8396	2099

Table 3: A detailed statistcs of our dataset usage.

Epistemic Marker Extraction For each model, we extract the epistemic marker from each response by few-shot prompting (Brown et al., 2020) the same model to recognize the epistemic markers emitted by itself. We manually examined a subset of each dataset and find out most of them are able to recognize the epistemic markers. For the unrecognized ones or the one that didn't match the desired format, we uniformly use *GPT-4o-mini* to extract its epistemic markers. According to Zhou et al. (2024), models are relunctant to express confidence in words, so we also provided responses that did not include any epistemic markers as few-shot samples, and those with no markers are grouped together as a special epistemic marker.

B.4 Evaluation Metrics

This section introduces the evaluation metrics in detail and explain our experiment settings.

B.4.1 Detailed Implementation and Formulas

Our evaluation metrics are categorized into three kinds: ECE-based (Xiong et al., 2024), CV-based (Jalilibal et al., 2021) and Spearman/Pearson coefficient based (Su and Li, 2021; Xiong et al., 2004), aiming to reflect the calibration error, the degree of dispersion and correlation respectively.

NumECE, I-AvgECE, and C-AvgECE To evaluate the calibration of model-generated confidence, we introduce three metrics: NumECE, I-AvgECE, and C-AvgECE. Since the latter two are based on generalization of marker confidence, these metrics assess the model's stability on marker confidence, considering both within-domain (I-AvgECE) and cross-domain (C-AvgECE) scenarios and comparing with number-based methods.

NumECE measures the overall calibration of the model's outputted numerical confidence. For each dataset D_j, we compute the expected calibration error (ECE-num) based on the model's numerical confidence on the test set. The final NumECE is the average of these ECE values across all datasets, providing a comprehensive evaluation of the model's confidence calibration.

I-AvgECE focuses on the model's performance within the same domain, where the training and test datasets are identical. For each dataset D_p, we calculate the marker-based expected calibration error (ECE-mar) by using the marker's confidence $\textit{Conf}(W_i, D_p, M_k)$ obtained from the training set of D_p on the test set of it when the model also emits W_i as confidence expression. The final I-AvgECE is obtained by averaging these values across all datasets as well.

C-AvgECE evaluates the model's ability to generalize its marker confidence across different datasets. For each pair of distinct datasets (D_p, D_q), where $D_p \neq D_q$, we calculate the marker-based expected calibration error ECE-mar(D_p, D_q, M_k). This is done by using the

	Binary Question	**Multiple Choice Question**
Eliciting epistemic markers	*User*: The following is a binary question. When responding, answer with a binary answer from **[choices]** and incorporate only one epistemic marker to reflect your confidence level. You must include your binary answer at the beginning of your response then respond with the epistemic markers in a concise and brief manner. The question is: **[Question]** And your answer is:	*User*: The following is a multiple choice question. When responding, answer with a letter from **[choices]** and incorporate only one epistemic marker to reflect your confidence level. You must include your choice of letter at the beginning of your response then respond with the epistemic markers in a concise and brief manner. The question is: **[Question]** The options are: **[Options]** And your answer is:
Eliciting numerical values	*User*: The following is a binary question. When responding, answer with a binary answer from **[choices]** and incorporate a number between 0 and 100 to reflect your confidence level. You must include your binary answer at the beginning of your response then respond with the confidence score in a concise and brief manner. The question is: **[Question]** And your answer is:	*User*: The following is a multiple choice question. When responding, answer with a letter from **[choices]** and incorporate a number between 0 and 100 to reflect your confidence level. You must include your choice of letter at the beginning of your response then respond with the confidence score in a concise and brief manner. The question is: **[Question]** The options are: **[Options]** And your answer is:

Table 4: Our prompt to elicit epistemic markers and numerical confidence values. The text inside the square brackets is filled with actual content in the dataset. Specifically, choices are capital letters that represent the options (e.g., "A, B, C, D" or "A, B, C, D, E") for multiple choice questions and "yes or no" for binary questions.

model's confidence $\mathit{Conf}(W_i, D_p, M_k)$ on the training dataset D_p to estimate the model's marker confidence on the test set of D_q, thereby measuring the model's ability to transfer its marker confidence to a new dataset. The final C-AvgECE is computed by averaging the ECE-mar values across all dataset pairs, providing insight into the model's robustness in handling cross-distribution variations. The mathematical formulas of three ECE-based metrics is given by:

$$\text{NumECE} = \frac{1}{|D|} \sum_{D_j \in D} \text{ECE-num}(D_j, M_k),$$

$$\text{I-AvgECE} = \frac{1}{|D|} \sum_{D_p \in D} \text{ECE-mar}(D_p, D_p, M_k),$$

$$\text{C-AvgECE} = \frac{\sum_{\substack{D_p, D_q \in D \\ D_p \neq D_q}} \text{ECE-mar}(D_p, D_q, M_k)}{|D| * (|D| - 1)}$$

where $|D|$ is the total number of datasets $(D_1, D_2, \ldots, D_n)$. Note that for all ECE values, we use a ECE bin number of *N*, where *N* is the number of confidence predictions.

I-AvgCV and C-AvgCV To quantify the concentration and variation of marker confidence, we propose I-AvgCV and C-AvgCV. These metrics assess how dispersed and consistent marker confidence is within individual datasets and across datasets, respectively.

I-AvgCV measures the concentration of marker confidence within a single dataset. For each dataset D_j, we calculate the coefficient of variation (CV) of the confidence of different markers $\mathit{Conf}(W_i, D_j, M_k)$. The final I-AvgCV is the average CV value across all datasets, providing an overall measure of confidence concentration.

It is important to note that while we expect the

Question Template

For the question **[original question]**, is the answer **[original answer]** its correct answer?

Sample 1

Q_i: Each bird eats 12 beetles per day, each snake eats 3 birds per day, and each jaguar eats 5 snakes per day. If there are 6 jaguars in a forest, how many beetles are eaten each day?

A_i: 1080 (The correct answer is 1080)

Integrated binary question: For the question ‘Each bird eats 12 beetles per day, each snake eats 3 birds per day, and each jaguar eats 5 snakes per day. If there are 6 jaguars in a forest, how many beetles are eaten each day?”, is the answer 1080 its correct answer?

Binary answer: Yes.

Sample 2

Q_j: James writes a 3 - page letter to 2 different friends twice a week. How many pages does he write a year?

A_j: 223 (Randomly generated answer, the correct answer is 624)

Integrated binary question: For the question ‘James writes a 3 - page letter to 2 different friends twice a week. How many pages does he write a year?”, is the answer 223 its correct answer?

Binary answer: No.

Table 5: Data pre-processing method used in GSM8K. The text inside the square brackets is replaced by actual content in the dataset. Sample 1 keeps the original correct answer and incorporate it into the binary answer while setting the binary answer to “Yes.” Sample 2 randomly generates a wrong answer and set the binary answer to “No.”

distribution of marker confidence to be more dispersed, we are not claiming that greater dispersion is inherently better. Our desired result for models is to cover a relatively wide range of confidence values across all markers, while also clearly differentiating between different markers. This would facilitate more effective confidence expression in a variety of scenarios. However, as shown in Table 1, the average I-AvgCV value marker is lower than 0.2, which indicates that the marker confidence is highly concentrated, leading us to conclude that the model fails to clearly differentiate between the markers.

C-AvgCV evaluates the consistency of marker confidence across datasets. For each shared marker (markers that appear in each dataset) W_i, we compute the CV of the marker confidence across different datasets and then average these values over all shared markers. The final C-AvgCV quantifies the consistency of model-generated confidence across multiple datasets.

The mathematical formulations for I-AvgCV and C-AvgCV are given by:

$$CV(D_j, M_k) = \frac{\sigma(D_j, M_k)}{\mu(D_j, M_k)},$$

$$\text{I-AvgCV}(M_k) = \frac{1}{|D|}\sum_{j=1}^{|D|} CV(D_j, M_k),$$

$$CV(W_i, M_k) = \frac{\sigma(W_i, M_k)}{\mu(W_i, M_k)},$$

$$\text{C-AvgCV}(M_k) = \frac{1}{|W|}\sum_{i=1}^{|W|} CV(W_i, M_k)$$

where $|W|$ is the number of shared markers across every datasets for model M_k, $\sigma(D_j, M_k)$ is the standard deviation of the confidence scores for all markers in dataset D_j for model M_k, $\mu(D_j, M_k)$ is the mean of the confidence scores for all markers in dataset D_j for model M_k, $\sigma(W_i, M_k)$ is the standard deviation of the confidence scores for the marker W_i across different datasets for model M_k and $\mu(W_i, M_k)$ is the mean of the confidence scores for the marker W_i across different datasets for model M_k.

MRC To assess the alignment of marker rankings across datasets, we introduce a metric based on the Spearman rank correlation coefficient: *Marker Rank Correlation* (MRC). This metric capture the degree of consistency in marker confidence rankings across datasets.

Specifically, For each pair of datasets (D_p, D_q), we compute the Spearman rank correlation coefficient between the confidence rankings of shared markers. The final MRC for the model is the average correlation across all dataset pairs. The mathematical formulations for MRC are given by:

$$\mathrm{MRC} = \frac{1}{|P|} \sum_{\substack{(D_p, D_q) \in P \\ D_p \neq D_q}} \rho(D_p, D_q)$$

where $W_1, \ldots, W_i$ are all the epistemic markers that shared by D_p and D_q, $\mathrm{S}(X, Y)$ denotes the Spearman rank correlation coefficient between the rankings of X and Y and $\mathrm{Conf}(W_i, D_j, M_k)$ represents the confidence of marker W_i for model M_k on dataset D_j.

MAC To analyze whether the confidence of markers and the accuracy of the model are positively correlated, we propose the *Marker Accuracy Correlation* (MAC) based on the Pearson correlation coefficient.

Specifically, for a given model M_k, we consider the confidence of a specific shared marker W_i, which is present across all datasets associated with M_k. We then compute the Pearson correlation coefficient between the set of marker confidences across these datasets and the model's overall accuracies on the same datasets. Finally, we compute the average of the correlation coefficients $\rho(W_i, M_k)$ across all shared markers W_i to obtain the overall correlation coefficient for the model, denoted as $\mathrm{MAC}(M_k)$. It's mathematical formula is given by:

$$\mathrm{MAC}(M_k) = \frac{1}{|W|} \sum_{W_i \in W} \rho(W_i, M_k),$$

where W is the set of all shared markers W_i, $|W|$ is the number of all shared markers and $\rho(W_i, M_k)$ is the Pearson correlation coefficient between the confidence of marker W_i and the model's accuracy on all the datasets.

These metrics provide a quantitative assessment of the consistency and concentration of model-generated confidence values across different datasets.

Threshold	Confidence Interval
10	∼23%
50	∼10%
200	∼5%

Table 6: The exact confidence interval of different filtering thresholds.

B.4.2 Details on Marker Filtering Strategies

The filtering is necessary because our method of quantifying marker confidence is based on accuracy, which is affected by the confidence interval. If the sample size for a particular marker is too small, its corresponding confidence values can be heavily influenced by random variations. For instance, if the marker "unsure" appears only once in the training set and the response happens to be correct, the marker confidence for "unsure" would be 100%, which may not accurately reflect the model's true intent. Previous works also show that confidence expression could be more reflective for humans when the it is determined in a crowd-source manner (Pennekamp et al., 2024), which supports our setting.

Furthermore, we have enough shared markers (more than ten for each model) after implementing the filtering, which ensures the reliability of our experiment.

On the other hand, all epistemic markers obtained from the training set are used for the experiment related to ECE values. Since the estimated confidence for each question in the test set is derived from the marker confidence in the training set, it is essential to ensure that the vast majority of markers in the test set can be mapped to corresponding markers in the training set. This approach is reasonable since the low frequency of marker occurrences results in minimal impact on the overall calibration performance, which ensures both completeness and reliability.

C Distribution Analysis

Our investigation primarily focused on the influence of **distribution diversity** and **difficulty** on the generalizability of marker confidence. We measured this generalizability by calculating the Average ECE value across various models for each specific dataset.

Table 7 illustrates that multi-domain datasets, specifically StrategyQA (14.42%) and MMLU (15.97%), exhibit significantly higher Average ECE values compared to single-domain datasets such as GSM8K (Math, 6.68%), MedMCQA (Medical, 6.25%), and CaseHOLD (Law, 9.53%). This finding leads to the conclusion that a more diverse data distribution impedes the transfer of models' marker confidence preferences from the training to the test set (Wang et al., 2023b,a, 2024a,b, 2025; Wang and Song, 2024).

Dataset	**Domain**	**Average ECE**
StrategyQA	Multi-domain	14.42
MMLU	Multi-domain	15.97
GSM8K	Math	6.68
MedMCQA	Medical	6.25
CaseHOLD	Law	9.53

Table 7: The Average ECE values across different models for certain dataset. We found that multi-domain datasets exhibit higher average ECE values than single-domain ones.

Additionally, we observed that a substantial **difficulty gap** between the training and test sets of the marker confidence also compromises its transferability. As shown in Table 8, we utilized a challenging math dataset, MATH-500, and a simpler math dataset, GSM8K, to compute the MRC value, representing the relevance of marker rankings across these two datasets. For both *Qwen2.5-7B-Instruct* (0.38) and *Mistral-7B-Instruct-v0.3* (0.26), the models demonstrated low relevance in marker rankings obtained from datasets with differing difficulties. This indicates that the **difficulty of the data distribution** can significantly affect the generalizability of marker confidence preferences.

Model	**MRC**
Qwen2.5-7B-Instruct	0.38
Mistral-7B-Instruct-v0.3	0.26

Table 8: The MRC value for two models across MATH-500 and GSM8K. The results shows that marker rankings obtained from the two datasets are relevant to little extent.