Evaluating Uncertainty Quantification Methods in Argumentative Large Language Models

Kevin Zhou[1] **Adam Dejl**[1] **Gabriel Freedman**[1]
Lihu Chen[1] **Antonio Rago**[1,2] **Francesca Toni**[1]
[1]Imperial College London [2]King's College London
{kevin.zhou24,adam.dejl18,g.freedman22,lihu.chen,ft}@imperial.ac.uk
antonio.rago@kcl.ac.uk

Abstract

Research in uncertainty quantification (UQ) for large language models (LLMs) is increasingly important towards guaranteeing the reliability of this groundbreaking technology. We explore the integration of LLM UQ methods in argumentative LLMs (ArgLLMs), an explainable LLM framework for decision-making based on computational argumentation in which UQ plays a critical role. We conduct experiments to evaluate ArgLLMs' performance on claim verification tasks when using different LLM UQ methods, inherently performing an assessment of the UQ methods' effectiveness. Moreover, the experimental procedure itself is a novel way of evaluating the effectiveness of UQ methods, especially when intricate and potentially contentious statements are present. Our results demonstrate that, despite its simplicity, direct prompting is an effective UQ strategy in ArgLLMs, outperforming considerably more complex approaches.

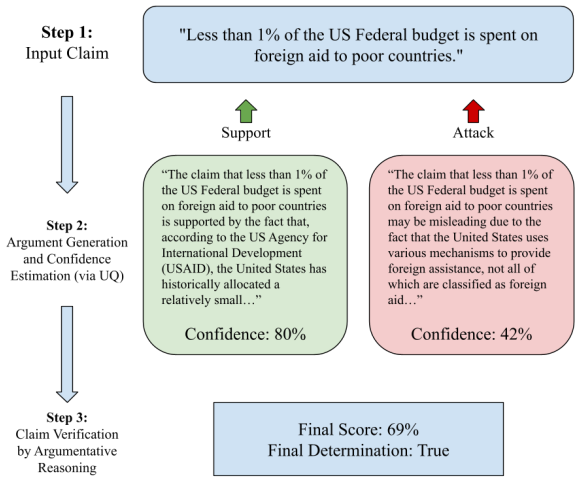

Figure 1: Example of argumentative LLMs, where UQ plays a crucial role in estimating confidence in the generated arguments and thus in the claim verification itself. Here, the input claim is derived from the TruthfulQA dataset (Lin et al., 2022), arguments are generated by Llama 3.1 (Grattafiori et al., 2024), and a default base score 0.50 is used for the input claim.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a range of tasks, such as coding, reasoning, and speech recognition (OpenAI, 2023; Grattafiori et al., 2024). However, they also often generate hallucinated answers (Sahoo et al., 2024), with no clear indication of the uncertainty which caused them. Still, many users are prone to blindly trusting LLMs' responses (Klingbeil et al., 2024), which is especially risky in areas such as healthcare where LLMs are being applied (He et al., 2025). In these settings, the ability to reliably retrieve an LLM's uncertainty would be immensely impactful, highlighting the importance of uncertainty quantification (UQ) research in the development of trustworthy AI systems.

Recent research has shown that LLMs exhibit strong performance in automated decision-making (Ouyang and Li, 2023), but that they also face challenges such as the inability to faithfully explain their decisions (Turpin et al., 2023; Chen et al., 2025) and reliably correct mistakes based on user feedback (Freedman et al., 2025). Towards addressing these challenges, Freedman et al. (2025) introduce argumentative LLMs (ArgLLMs), which leverage computational argumentation to improve explainability and contestability for decision-making tasks such as claim verification.

For a given statement, ArgLLMs generate an argumentation framework of supporting and attacking arguments, quantify argument uncertainty, and aggregate these components using gradual semantics (Baroni et al., 2019), a form of formal reasoning, to determine whether a statement is true (see Figure 1 for an example of an ArgLLM evaluating a claim). ArgLLMs demonstrate comparable performance in claim verification against prompt-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21700–21711

ing methods such as direct questioning and chain-of-thought (Wei et al., 2022) while also providing faithful explanations through the argumentative structure.

It can be seen from Figure 1 that in ArgLLMs, the estimated confidence in the generated arguments, and thus the chosen UQ method, is pivotal in the claim verification task itself. For example, if (unlike in Figure 1) the confidence in the attacking argument(s) was stronger than that in the supporting argument(s), we would expect the final determination to be false.[1] Thus, ArgLLMs are an appealing evaluation setting for UQ methods.

Since the confidence scores from the used UQ method directly feed into the final predictions, ArgLLMs enable evaluation of these scores without requiring access to the ground-truth labels for the arguments, as they only require the label for the top-level claim. For instance, arguments in favor of an incorrect prediction can generally be expected to be less convincing, and should thus be less certain. In this setting, UQ methods can be seen as a proxy for downstream factuality and output reliability, which is a common use-case. Additionally, ArgLLMs involve long and potentially contentious arguments, and the generation of supporting and attacking arguments for each claim ensures a strong diversity of statements to evaluate, making for a challenging, realistic, and wide-ranging setting.

Various UQ methods exist in the literature, from directly prompting the model for an uncertainty estimate (also known as verbalized UQ), which have proven effective and at times well-calibrated (Yang et al., 2024; Tian et al., 2023), to more complex methods involving token logits or semantic similarity (Geng et al., 2024). It is thus unclear which UQ method would be ideal in the context of ArgLLMs, which use direct prompting in Freedman et al. (2025). This is especially true since verbalized UQ can sometimes have issues arising from LLMs being overconfident and biased when evaluating their own answers (Sun et al., 2025). In this paper, we aim to shed light on this problem by experimenting with a set of LLM UQ methods and examining their effect on the performance of ArgLLMs in the claim verification task. Concretely, our contributions are as follows:

- We integrate several LLM UQ methods that have performed well in long-form generation benchmarks (Vashurin et al., 2025; Zhang et al., 2024) into the explainable framework of ArgLLMs, presenting a novel method for evaluating their effectiveness via the resulting accuracy in downstream claim verification.
- We conduct experiments with three claim verification datasets, three LLMs, and four ArgLLM settings, resulting in 36 different configurations where we evaluate three LLM UQ methods as well as the direct prompting baseline. The results show direct prompting outperforms the other UQ methods.

2 Background and Preliminaries

2.1 Uncertainty Quantification

In addition to direct prompting, we use three LLM UQ methods: Semantic Entropy (Kuhn et al., 2023), Eccentricity (Lin et al., 2024), and LUQ (Zhang et al., 2024). For these additional methods, we focus on the version of each method implemented in the MIT-licensed LM-Polygraph library (Fadeeva et al., 2023). Further details of the methods are given in Appendix A.

Direct Prompting involves the model directly providing a confidence score for the inputted text. The prompt used to obtain confidence scores for direct prompting is given in Appendix B.

Semantic Entropy Each time the model is prompted, it generates (a hyperparameter) M different samples. Then, samples with similar meaning are clustered and the entropy over the meaning-distribution determined by the clustering is computed using token logits.

Eccentricity has multiple variations: in this paper we use the natural language inference (NLI) Entailment version. After generating M samples for an input, an NLI classifier model computes entailment logits between the generations, uses them to calculate similarity scores, and then constructs a graph Laplacian with the similarity scores as edge weights. The uncertainty is then computed as the average distance from the center of the eigenvectors, with the intuition being that a lower uncertainty would result in more similar samples and thus a lower average distance.

LUQ (Long-text Uncertainty Quantification) also involves multiple generated samples. In Zhang et al. (2024), the generations are split into component sentences, but LM-Polygraph considers a simplified version which leaves the generations in their full form. An NLI model is used to obtain logit val-

[1] Note that the argumentative reasoning generated by ArgLLMs can be significantly more complex and deep than the example shown in Figure 1.

ues of "entailment" and "contradiction" between each generated response, and then the uncertainty is computed as a function of both the "entailment" and "contradiction" logits.

Similar to Eccentricity and LUQ, other methods in the literature also make use of entailment and NLI-based scoring, highlighting the capabilities of NLI scores in uncertainty estimation processes. Examples of such methods include SelfCheckGPT (Manakul et al., 2023) and HaLoCheck (Elaraby et al., 2023), which are designed primarily for the task of hallucination detection.

2.2 Argumentative LLMs

ArgLLMs construct quantitative bipolar argumentation frameworks (QBAFs), which consist of arguments connected through support and attack relations where each argument also has a base score representing its intrinsic strength (Baroni et al., 2019). In the context of ArgLLMs, the supporting and attacking arguments have their base score set as the confidence score outputted by the UQ method.[2] QBAFs can then be evaluated via a gradual semantics (Baroni et al., 2019), which determines the final strength of each argument, taking into account its intrinsic strength as well as the strengths of its attackers and supporters. In ArgLLMs, each input claim can have supporting and attacking arguments generated for it, and each of those arguments can have its own supports and attacks (Freedman et al., 2025). From the scores and connections of these QBAF components, the DF-QuAD (Rago et al., 2016) gradual semantics is used to compute the original claim's strength. If the final score of the claim is greater than 0.5, it is predicted to be true; otherwise it is predicted to be false.

The structure of ArgLLMs provides a unique and also realistic setting for LLM UQ method evaluation where the confidence scores of generated arguments are integral to downstream claim verification, which to the best of our knowledge has not yet been studied in the literature.

3 Experiments

We evaluate the performance of ArgLLMs on the claim verification task when using different LLM UQ methods as the uncertainty estimator for the arguments. For the experiments, we build upon the publicly released code provided by Freedman et al. (2025).[3] We also employ the same prompts as Freedman et al. (2025) except for a minor change to argument generation (see Appendix B).

3.1 UQ Integration

We use the LM-Polygraph implementations for Semantic Entropy, Eccentricity, and a simplified version of LUQ without sentence splitting, including LM-Polygraph's default value of 10 for the number of samples generated per input. The performance of ArgLLMs with these methods is compared with the baseline UQ method of direct prompting.

3.2 Datasets

We use the TruthfulClaim, StrategyClaim and MedClaim datasets (Freedman et al., 2025), which are tailored versions of TruthfulQA (Lin et al., 2022), StrategyQA (Geva et al., 2021), and MedQA (Jin et al., 2021). Details for each dataset are discussed further in Appendix C.

An important consideration is the risk of data contamination, especially since these claim-based datasets are derived from popular QA datasets. We believe that in this case, the associated risk is mitigated by the nature of ArgLLMs. Specifically, using ArgLLMs substantially changes the original task and data distribution, as the model is generating attacking and supporting arguments and providing subsequent confidence scores for these arguments rather than directly answering the questions from the original QA datasets.

3.3 Models

The LLMs we use are Google's gemma-2-9b-it (Mesnard et al., 2024), Meta's Llama-3.1-8B (Grattafiori et al., 2024), and OpenAI's GPT-4o-mini (OpenAI, 2024). We choose these models because they have demonstrated strong performance on model benchmarks and fit within our compute resources. The gemma-2-9b-it and Llama-3.1-8B models are open-source[4], while GPT-4o-mini is closed-source. Notably, since Semantic Entropy requires direct access to model logits in its computations, the LM-Polygraph implementation is not compatible with GPT-4o-mini. All open-source models are quantized to 4 bits to lower the computational cost (Dettmers et al., 2023).

[2] In this work, the confidence score is considered the antonym of uncertainty.

[3] Our code is available at https://github.com/CLArg-group/argumentative-llms-uq. All experiments are run on a system with RTX 4090 24GB GPUs with the seed set to 42 for reproducibility.

[4] We adopt a broad notion of "open-source", not necessarily implying licenses approved by the Open Source Initiative.

Model	UQ Method	TruthfulClaim				StrategyClaim				MedClaim			
		D=1		D=2		D=1		D=2		D=1		D=2	
		0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS
Llama 3.1	Direct Prompting	**0.626**	0.664	**0.652**	**0.658**	0.594	0.600	0.592	0.590	**0.594**	**0.606**	**0.574**	0.604
	Semantic Entropy	0.604	0.650	0.592	0.650	0.574	0.592	0.552	0.588	0.548	0.596	0.556	0.566
	Eccentricity	0.514	0.606	0.508	0.638	0.530	0.566	0.534	0.566	0.484	0.518	0.490	0.548
	LUQ	0.556	**0.668**	0.552	0.654	**0.622**	**0.614**	**0.614**	**0.600**	0.578	0.594	**0.574**	**0.608**
Gemma 2	Direct Prompting	**0.682**	0.732	**0.674**	0.732	**0.656**	**0.708**	**0.652**	**0.702**	**0.596**	0.578	**0.576**	0.582
	Semantic Entropy	0.516	**0.746**	0.558	0.734	0.466	0.666	0.490	0.696	0.518	0.580	0.546	0.578
	Eccentricity	0.504	0.714	0.500	**0.740**	0.464	0.634	0.450	0.676	0.534	0.580	0.496	0.582
	LUQ	0.560	0.714	0.584	0.712	0.526	0.672	0.538	0.686	0.566	**0.590**	0.542	**0.584**
GPT-4o-mini	Direct Prompting	**0.748**	**0.816**	**0.764**	**0.822**	**0.646**	**0.742**	**0.690**	0.736	**0.638**	**0.718**	**0.644**	**0.710**
	Semantic Entropy	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Eccentricity	0.512	0.722	0.496	0.760	0.548	0.680	0.534	0.724	0.528	0.656	0.516	0.686
	LUQ	0.610	0.780	0.618	0.796	0.610	0.722	0.632	**0.742**	0.546	0.662	0.510	0.704

Table 1: Accuracy (↑, best in bold) of ArgLLMs in all experiments. Values other than the best that are not statistically significantly worse than the best accuracy are underlined. Semantic Entropy results for GPT-4o-mini are marked as N/A for the reasons discussed in Section 3.3. In the "0.5 BS" setting, the claim's base score is set at 0.5, while in "Est. BS" it is estimated through prompting.

3.4 Experimental Procedure

Each experiment is defined by the dataset, LLM, UQ method, the method of determining the claim's base score, and the depth. A depth of 1 (i.e. D=1) means that for each claim, one supporter and one attacker will be generated, and a depth of 2 (i.e. D=2) means that each of those arguments will also have a supporter and attacker. For the claim's base score, we adopt the two methods from Freedman et al. (2025) of either setting it to 0.5 or estimating it by prompting the LLM. We are not able to use the other LLM UQ methods with these claims because they are pre-existing text from the datasets and are not generated by the LLM itself.

Thus, for each claim, the LLM generates supporting and attacking argument(s), and uses the UQ method to obtain confidence scores for each argument. Importantly, the UQ methods besides the direct prompting baseline produce a raw score that is not necessarily in $[0, 1]$, which ArgLLMs require. To address this, we employ binned normalization by grouping the outputs into 20 quantile bins linearly mapped to a score in $[0, 1]$, which is more robust to skewed distributions than a strict linear normalization. The DF-QuAD semantics is then used to compute a final confidence measure, which determines the prediction. We adopt the accuracy of ArgLLMs on claim verification as the primary downstream performance metric to observe the impact of integrating the different UQ methods. All experiments use 500 data samples, which are identical to those used by Freedman et al. (2025).

4 Results and Discussion

Table 1 presents the accuracy results from all experiments, indicating that direct prompting clearly achieves the best performance. Table 2 shows that direct prompting is either the best UQ method or not statistically significantly worse than the best in all 36 configurations, and its 25 instances of being the best UQ method are by far the most. An important caveat for our results and subsequent conclusions is that the models used are relatively small compared with the leading LLMs, and each experiment is only run once.

UQ Method	Best	Not Significantly Worse than Best
Direct Prompting	25 (0.69)	11 (0.31)
Semantic Entropy	1 (0.04)	15 (0.63)
Eccentricity	1 (0.03)	8 (0.22)
LUQ	10 (0.28)	13 (0.36)

Table 2: Summary of the accuracy results from Table 1, counting the number of experiments in which each method performed best or did not perform statistically significantly worse than the best method. The values in the parentheses are the counts divided by the number of experiments the UQ method is used in. The "Best" column counts add up to 37 since LUQ and Direct Prompting tied for best with MedClaim, Llama 3.1, 0.5 BS, and D=2.

In some cases, the advantage of direct prompting was substantial, such as in the StrategyClaim Gemma-2 0.5 BS (D=1) setting where it results in a 0.130 higher accuracy than the next best method. In contrast, the most direct prompting is ever outperformed is 0.028 by LUQ in the StrategyClaim, Llama 3.1, 0.5 BS, and D=1 setting.

We assessed the statistical significance of the accuracy results by conducting bootstrap tests with 5000 resamples to obtain 95% confidence intervals for the pairwise accuracy differences between the methods. These confidence intervals are used to

determine the statistical significance of the best performances per configuration in Table 1. Furthermore, across all 180 confidence intervals comparing UQ method performances, 74 indicate statistically significant differences. Of these, 44 involve direct prompting having a statistically significant advantage, followed by 24 for LUQ, 6 for Semantic Entropy, and 0 for Eccentricity. The table of confidence intervals and further details are included in Appendix E.

In addition to accuracy, we also measured the Brier scores for all experiments, which computes the mean squared difference between the predicted probability and the true label. In our case, the final ArgLLM score is used as the predicted probability, and the label is 1 if the topic claim is true and 0 if it is false. The full table of Brier scores is presented in Appendix E. In summary, direct prompting scored the best in 18 instances, the most of any method, followed by LUQ with 9, Semantic Entropy with 7, and Eccentricity with 2.

Overall, these results support the notion that verbalized confidence scores from direct prompting can be well-calibrated (Tian et al., 2023) and represent the model's internal knowledge well with effective prompting (Yang et al., 2024).

Moreover, direct prompting likely outperforms the other methods due to the nature of long-form contentious generations in argumentation. Sampling-based methods such as Semantic Entropy require the capture of semantic consistency among multiple arguments, which can lead to degraded performance as the length of texts grows. As shown in Figure 1, ArgLLMs have long generations which can sometimes be contentious, unlike more definitive true or false statements. In this situation, direct prompting is often better suited to estimate a reasonable uncertainty level based on the LLM's self-knowledge. Also, it does not rely on an additional normalization step to map the UQ outputs to suitable confidence scores, which can be prone to noise and introduces further estimation compounded with the existing estimation task.

On top of its superior performance, the advantage of direct prompting is further amplified by its lower resource requirements. Many of the other high-performing LLM UQ methods, including all three tested in this paper, require a separate NLI model and multiple generations per instance, increasing both memory and time complexity.

While not outperforming direct prompting overall, LUQ's relatively strong performance and high frequency of performing the best are intriguing. For example, when using Llama 3.1 with Strategy-Claim, LUQ is the best in all four settings, illustrating the potential for a UQ method to perform the best in a specific setup. In total, as seen in Table 2, LUQ performs the best 10 times, which is much greater than the 1 time for each of Semantic Entropy and Eccentricity. However, it is worth noting that Semantic Entropy often closely trails behind the best performing method with no statistically significant difference, even if it is rarely the best method.

LUQ's capabilities in these experiments are also consistent with its strong performance in factuality tasks (Zhang et al., 2024). While sentence splitting is not included in the LUQ implementation we use, computing uncertainty more directly from the entailment and contradiction logits between responses is another potential advantage which could have contributed to the stronger performance. All in all, the experiments demonstrate that ArgLLMs offer a compelling benchmark for LLM UQ methods. The QBAF structure of supporting and attacking arguments poses challenges distinct from those in existing benchmarks. At the same time, certain features that enhance performance on other tasks can also improve ArgLLM claim verification performance as in the case of LUQ, lending further credibility to ArgLLMs as an evaluation environment.

5 Conclusion

Our work integrates commonly used and high-performing LLM UQ methods into ArgLLMs and assesses their performance on claim verification. We find that direct prompting leads to notably better performance than the other UQ methods. Among the latter, the LM-Polygraph implementation of LUQ performs better than Semantic Entropy and Eccentricity, reflecting LUQ's advantages seen in other tasks. Overall, the experiments affirm the role of verbalized confidence prompting in eliciting confidence scores in ArgLLMs and suggest that prompt-based methods offer benefits for LLM UQ with long-form and potentially contentious statements. Our research also presents and highlights the value of evaluating LLM UQ methods in argumentative settings and faithfully explainable frameworks such as ArgLLMs.

6 Limitations

We now discuss some limitations of our work, particularly with regard to the experiments:

- We did not conduct multiple runs for each configuration due to the high computational and time cost of each run, which limits the robustness of statistical analysis on the experiments.
- We had to choose models which together with the entire experimental pipeline could fit within our compute resources. This limited the possible size of LLMs in our experiments.
- As discussed in Section 3.4 and Section 4, the use of binned normalization in the UQ process is a limitation and likely negatively impacts the calibration and performance of the additional UQ methods. Future work could benefit from a more robust and separately evaluated calibration procedure.
- We only evaluate the task of claim verification with true or false as the possible labels in this paper; future experiments with other types of datasets and tasks would further enrich our understanding of LLM UQ integration with ArgLLMs.

7 Ethical Considerations

One potential risk of UQ with ArgLLMs in general is that malicious actors could theoretically devise a bad-faith UQ method to output confidence scores in line with an agenda, and then integrate it into the background of ArgLLMs and present the ArgLLM outputs for means of persuasion or demonstration. As a result, it is paramount that any user presenting ArgLLM outputs is also transparent and truthful about the UQ method used. Additionally, some sample claims in the datasets may contain untrue stereotypes or beliefs. We do not endorse any of the statements or opinions in the datasets, and their only purpose in the experiments would be to serve as sample claims that the ArgLLM evaluates.

References

Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.*, 105:252–286.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *Preprint*, arXiv:2505.05410.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. 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*.

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

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 446–461. Association for Computational Linguistics.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative large language models for explainable and contestable claim verification. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 14930–14939. AAAI Press.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. 2024. The llama 3 herd of models. *CoRR*, arXiv:2407.21783.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf. Fusion*, 118:102963.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI - an experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.*, 160:108352.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

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

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.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

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

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Siqi Ouyang and Lei Li. 2023. Autoplan: Automatic planning of interactive decision-making tasks with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3114–3128. Association for Computational Linguistics.

Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, pages 63–73. AAAI Press.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11709–11724. Association for Computational Linguistics.

Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. Large language models are overconfident and amplify human bias. *Preprint*, arXiv:2505.02151.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 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*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Trans. Assoc. Comput. Linguistics*, 13:220–248.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. 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*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *CoRR*, abs/2412.14737.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5244–5262. Association for Computational Linguistics.

A Uncertainty Quantification Methods

Semantic Entropy The clustering in Semantic Entropy is performed based on the concept of bi-directional entailment; given two inputs, a natural language inference (NLI) model such as DeBERTa-large (He et al., 2021) is used to determine if one entails the other and vice versa, and the generations are clustered together if both directions are true. The likelihoods of each sample are summed together within a cluster using the token logits, and then the entropy is computed over the meaning-distribution to determine the semantic entropy for the input text. Formally, Kuhn et al. (2023) express the semantic entropy as:

$$\begin{aligned} SE(x) &= -\sum_{c} p(c|x) \log p(c|x) \\ &= -\sum_{c} ((\sum_{s \in c} p(s|x)) \log [\sum_{s \in c} p(s|x)]) \end{aligned} \quad (1)$$

where x is the input text, c represents a semantic equivalence class, and s is a sequence. However, in practice, only the distribution generated by the model is accessible for the calculations. Thus, the semantic entropy is estimated through Monte Carlo integration as such:

$$SE(x) \approx -|C|^{-1} \sum_{i=1}^{|C|} \log p(C_i|x) \quad (2)$$

where C is the set of semantic equivalence classes induced by the model.

Eccentricity The steps for the Eccentricity NLI Entailment algorithm as described in Lin et al. (2024) are as follows, with more detailed explanations afterwards:

1. Generate M samples from the model.
2. Use an NLI classifier to obtain predicted entailment scores between generations. Then, apply a softmax function to these logit values to obtain probabilities of entailment which are used as the similarity measure.
3. Using the graph Laplacian and its corresponding embedding, calculate the average distance from the center as an eccentricity estimate. This serves as the input's uncertainty score.

In step 2), the NLI classifier Lin et al. (2024) use is the DeBERTa-large model, which Kuhn et al.

(2023) also use for Semantic Entropy. Instead of using the DeBERTa model to check for sequences entailing each other or not, the entailment logits are retrieved with a softmax applied afterwards to obtain a probability measure in $[0, 1]$ of entailment as a proxy for similarity.

In step 3), the embedding of a sequence can be represented through the eigenvectors of L, where L is the symmetric normalized graph Laplacian. Specifically, if there are M responses and $u_1, ..., u_k \in \mathbb{R}^M$ are the k eigenvectors of L with the smallest eigenvalues, then the embedding v_j of sequence s_j is $[u_{1,j}, ..., u_{k,j}]$ (von Luxburg, 2007). From this embedding, the eccentricity uncertainty is calculated by the following:

$$U(x) = ||[v_1'^\top, ..., v_M'^\top]||_2 \quad (3)$$

where $v_j' = v_j - \frac{1}{M}\sum_{j'=1}^{M} v_{j'}$ is the difference between v_j and the average embedding.

LUQ Here, we discuss the calculation of similarity scores and the subsequent step to obtain the uncertainty measure in LUQ. As discussed in the paper, we are considering a simplified version of LUQ which does not include sentence splitting. Formally, if the "entailment" value is l_{entail} and the "contradiction" value is $l_{contradict}$:

$$similarity = \frac{e^{l_{entail}}}{e^{l_{entail}} + e^{l_{contradict}}} \quad (4)$$

Then, the uncertainty is computed via an average of the pairwise similarity scores, where higher similarity between responses indicates lower uncertainty.

B Prompts

As mentioned in the paper, we employ the same prompts that Freedman et al. (2025) use, except for the prompt used to generate the supporting and attacking arguments. For this prompt, we make a slight modification. In the original prompt, shown in the top box in Figure 3, the section that tells the LLM to generate "N/A" if

> "there is a non zero probability that this claim is true"

does not distinguish between the supporting and attacking case. At times, this could lead to an "N/A" generation when prompted for an attacking argument with a topic claim that the LLM perceives to not be true, when in fact a strong attacking argument should be generated in this situation. Our new prompting accounts for this by conditioning this portion of the prompt on whether it is an attacking or supporting situation, as illustrated in Figure 3.

For reference, we also provide the prompt used by Freedman et al. (2025) and us to obtain confidence scores for the generated supporting and attacking arguments through direct prompting, in Figure 2.

You are an analyst evaluating the validity and relevance of arguments. For the argument:

Argument: "{argument}"

please give your confidence that the argument presents a compelling case {'in favour of'/'against'} the statement:

Statement: "{parent argument}"

Your assessment should be based on how well the argument {'supports'/'refutes'} the considered statement as well as the correctness, accuracy and truthfulness of the given argument. Your response should be between 0% and 100% with 0% indicating that the considered argument is definitely invalid, 100% indicating that the considered argument is definitely valid and values in between indicating various levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment. Please respond in the following form:

Likelihood: The predicted likelihood that the considered argument is valid
Likelihood:

Figure 2: The prompt used in the direct prompting method to obtain confidence scores for the generated supporting and attacking arguments (reproduced from Freedman et al. (2025)).

C Datasets

The original QA datasets consist of question and answer pairs which are transformed into claims by an LLM and then manually checked and edited if needed to ensure faithfulness to the original data in Freedman et al. (2025). We use 500 samples per experiment in order to have a sample size consistent with prior work on these datasets (Freedman et al., 2025) as well as provide a representative number of samples while keeping the experiments

Please provide a single short argument {“supporting" if support else “attacking”} the following claim. Construct the argument so it refers to the truthfulness of the claim. Only provide an argument if you think there is a valid and convincing {“support" if support else “attack”} for this claim (there is a non-zero probability that the claim is **true**), otherwise return: N/A.
Claim: {statement}
Now take a deep breath and come up with an argument.
Argument:

↓

Please provide a single short argument {“supporting" if support else “attacking”} the following claim. Construct the argument so it refers to the truthfulness of the claim. Only provide an argument if you think there is a valid and convincing {“support” if support else “attack”} for this claim (there is a non-zero probability that the claim is **{“true” if support else “false”}**), otherwise return: N/A.
Claim: {statement}
Now take a deep breath and come up with an argument.
Argument:

Figure 3: Prompt modification for the generation of supporting and attacking arguments, with the prompt from Freedman et al. (2025) in the top box and the new prompt we use in the bottom box. The changed portion is shown in bold.

computationally tractable.

D Experiment Parameters

A key parameter for the LLMs in these experiments, especially considering the importance of semantic consistency in some of the UQ methods, is the temperature. We use a temperature of 0.7 for the main LLM, which is used for direct prompting and the generation of supporting and attacking arguments. In addition, we use $p = 0.95$ top-p sampling and set the repetition penalty to 1.0. For the generation of samples in the UQ procedures of Semantic Entropy, Eccentricity, and LUQ with LM-Polygraph, we use the default value of 1.0 for the temperature, p, and repetition penalty.

E Results

E.1 Brier Scores

Table 3 shows the table of Brier scores for all experiments.

E.2 Confidence Intervals

Table 4, Table 5, and Table 6 show the confidence intervals from the bootstrapping procedure for the TruthfulClaim, StrategyClaim, and MedClaim experiments respectively. Confidence intervals where the values are either both negative or both positive indicate statistical significance. In these tables, if the values of the confidence interval are both negative, that means the first UQ method in the UQ Pair column performed statistically significantly worse than the second UQ method in this configuration, and if the values are both positive, then the first UQ method performed statistically significantly better than the second UQ method.

Model	UQ Method	TruthfulClaim				StrategyClaim				MedClaim			
		D=1		D=2		D=1		D=2		D=1		D=2	
		0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS	0.5 BS	Est. BS
Llama 3.1	Direct Prompting	0.217	0.253	0.217	0.244	0.239	0.324	0.252	0.324	0.243	0.274	0.251	0.272
	Semantic Entropy	0.245	0.244	0.241	0.241	0.250	0.303	0.246	0.309	0.260	0.284	0.252	0.290
	Eccentricity	0.292	0.282	0.263	0.257	0.282	0.338	0.264	0.331	0.310	0.342	0.271	0.314
	LUQ	0.251	0.248	0.248	0.246	0.236	0.300	0.239	0.311	0.262	0.273	0.245	0.271
Gemma 2	Direct Prompting	0.234	0.209	0.242	0.214	0.239	0.240	0.240	0.244	0.297	0.328	0.328	0.353
	Semantic Entropy	0.263	0.196	0.247	0.191	0.274	0.242	0.259	0.231	0.265	0.305	0.244	0.300
	Eccentricity	0.277	0.210	0.264	0.195	0.298	0.261	0.278	0.241	0.272	0.293	0.261	0.296
	LUQ	0.250	0.207	0.252	0.199	0.265	0.241	0.254	0.237	0.257	0.296	0.254	0.298
GPT-4o-mini	Direct Prompting	0.170	0.135	0.170	0.137	0.204	0.197	0.208	0.201	0.220	0.187	0.225	0.195
	Semantic Entropy	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Eccentricity	0.274	0.187	0.267	0.169	0.269	0.220	0.253	0.202	0.270	0.223	0.258	0.200
	LUQ	0.235	0.166	0.231	0.149	0.232	0.195	0.235	0.186	0.255	0.216	0.253	0.207

Table 3: Brier scores (↓) in all experiments. Semantic Entropy results for GPT-4o-mini are marked as N/A for the reasons discussed in Section 3.3. In the "0.5 BS" setting, the claim's base score is set at 0.5, while in "Est. BS" it is estimated through prompting.

Model	UQ Pair	D=1		D=2	
		0.5 BS	Est. BS	0.5 BS	Est. BS
Llama 3.1	Direct, SE	(-0.0340, 0.0800)	(-0.0200, 0.0480)	(0.0040, 0.1160)	(-0.0240, 0.0380)
	Direct, Ecc	(0.0540, 0.1700)	(0.0140, 0.1020)	(0.0860, 0.2020)	(-0.0120, 0.0520)
	Direct, LUQ	(0.0160, 0.1240)	(-0.0380, 0.0320)	(0.0460, 0.1560)	(-0.0280, 0.0360)
	SE, Ecc	(0.0320, 0.1480)	(0.0000, 0.0860)	(0.0260, 0.1440)	(-0.0160, 0.0380)
	SE, LUQ	(-0.0120, 0.1060)	(-0.0520, 0.0180)	(-0.0180, 0.1020)	(-0.0280, 0.0220)
	Ecc, LUQ	(-0.0980, 0.0140)	(-0.1060, -0.0160)	(-0.1000, 0.0140)	(-0.0440, 0.0120)
Gemma 2	Direct, SE	(0.0940, 0.2180)	(-0.0460, 0.0180)	(0.0540, 0.1780)	(-0.0280, 0.0240)
	Direct, Ecc	(0.1080, 0.2300)	(-0.0200, 0.0560)	(0.1140, 0.2340)	(-0.0380, 0.0220)
	Direct, LUQ	(0.0580, 0.1680)	(-0.0120, 0.0500)	(0.0340, 0.1480)	(-0.0060, 0.0460)
	SE, Ecc	(-0.0380, 0.0640)	(-0.0020, 0.0680)	(0.0100, 0.1060)	(-0.0321, 0.0220)
	SE, LUQ	(-0.1020, 0.0160)	(0.0000, 0.0660)	(-0.0820, 0.0320)	(-0.0040, 0.0480)
	Ecc, LUQ	(-0.1080, -0.0040)	(-0.0360, 0.0360)	(-0.1320, -0.0360)	(0.0000, 0.0560)
GPT-4o-mini	Direct, Ecc	(0.1840, 0.2880)	(0.0600, 0.1280)	(0.2140, 0.3200)	(0.0320, 0.0940)
	Direct, LUQ	(0.0840, 0.1920)	(0.0060, 0.0640)	(0.0920, 0.1980)	(0.0000, 0.0520)
	Ecc, LUQ	(-0.1520, -0.0440)	(-0.0960, -0.0200)	(-0.1780, -0.0660)	(-0.0660, -0.0060)

Table 4: Confidence intervals for the accuracy differences between UQ methods in the TruthfulClaim experiments, with the order of the compared UQ methods given in the UQ Pair column (Direct = direct prompting, SE = Semantic Entropy, Ecc = Eccentricity).

Model	UQ Pair	D=1		D=2	
		0.5 BS	Est. BS	0.5 BS	Est. BS
Llama 3.1	Direct, SE	(-0.0360, 0.0760)	(-0.0220, 0.0380)	(-0.0160, 0.0960)	(-0.0221, 0.0280)
	Direct, Ecc	(0.0080, 0.1160)	(-0.0040, 0.0720)	(0.0040, 0.1120)	(-0.0040, 0.0540)
	Direct, LUQ	(-0.0800, 0.0240)	(-0.0440, 0.0160)	(-0.0780, 0.0340)	(-0.0360, 0.0160)
	SE, Ecc	(-0.0100, 0.0960)	(-0.0100, 0.0620)	(-0.0360, 0.0720)	(-0.0020, 0.0460)
	SE, LUQ	(-0.1060, 0.0100)	(-0.0540, 0.0080)	(-0.1200, -0.0020)	(-0.0340, 0.0100)
	Ecc, LUQ	(-0.1440, -0.0400)	(-0.0860, -0.0100)	(-0.1340, -0.0260)	(-0.0600, -0.0080)
Gemma 2	Direct, SE	(0.1160, 0.2420)	(0.0040, 0.0780)	(0.1020, 0.2220)	(-0.0220, 0.0340)
	Direct, Ecc	(0.1200, 0.2440)	(0.0280, 0.1160)	(0.1420, 0.2620)	(-0.0100, 0.0620)
	Direct, LUQ	(0.0620, 0.1780)	(-0.0020, 0.0700)	(0.0580, 0.1720)	(-0.0120, 0.0440)
	SE, Ecc	(-0.0440, 0.0480)	(-0.0020, 0.0660)	(-0.0100, 0.0880)	(-0.0080, 0.0500)
	SE, LUQ	(-0.1160, -0.0040)	(-0.0440, 0.0320)	(-0.1020, 0.0080)	(-0.0160, 0.0360)
	Ecc, LUQ	(-0.1140, -0.0100)	(-0.0741, -0.0020)	(-0.1380, -0.0360)	(-0.0400, 0.0200)
GPT-4o-mini	Direct, Ecc	(0.0480, 0.1480)	(0.0300, 0.0940)	(0.1020, 0.2100)	(-0.0180, 0.0420)
	Direct, LUQ	(-0.0160, 0.0900)	(-0.0100, 0.0500)	(0.0080, 0.1080)	(-0.0340, 0.0200)
	Ecc, LUQ	(-0.1160, -0.0060)	(-0.0760, -0.0080)	(-0.1520, -0.0440)	(-0.0480, 0.0140)

Table 5: Confidence intervals for the accuracy differences between UQ methods in the StrategyClaim experiments, with the order of the compared UQ methods given in the UQ Pair column (Direct = direct prompting, SE = Semantic Entropy, Ecc = Eccentricity).

Model	UQ Pair	D=1		D=2	
		0.5 BS	Est. BS	0.5 BS	Est. BS
Llama 3.1	Direct, SE	(-0.0080, 0.1020)	(-0.0360, 0.0540)	(-0.0340, 0.0700)	(-0.0060, 0.0800)
	Direct, Ecc	(0.0519, 0.1700)	(0.0360, 0.1400)	(0.0240, 0.1460)	(0.0100, 0.1020)
	Direct, LUQ	(-0.0400, 0.0720)	(-0.0360, 0.0600)	(-0.0520, 0.0540)	(-0.0500, 0.0420)
	SE, Ecc	(0.0060, 0.1200)	(0.0320, 0.1240)	(0.0080, 0.1220)	(-0.0100, 0.0480)
	SE, LUQ	(-0.0840, 0.0220)	(-0.0460, 0.0500)	(-0.0720, 0.0360)	(-0.0740, -0.0100)
	Ecc, LUQ	(-0.1560, -0.0320)	(-0.1240, -0.0280)	(-0.1440, -0.0240)	(-0.0920, -0.0280)
Gemma 2	Direct, SE	(-0.0100, 0.1140)	(-0.0120, 0.0520)	(-0.0300, 0.0880)	(-0.0180, 0.0260)
	Direct, Ecc	(-0.0260, 0.0980)	(-0.0200, 0.0580)	(0.0140, 0.1420)	(-0.0280, 0.0300)
	Direct, LUQ	(-0.0500, 0.0580)	(-0.0160, 0.0380)	(-0.0220, 0.0880)	(-0.0260, 0.0220)
	SE, Ecc	(-0.0720, 0.0420)	(-0.0420, 0.0420)	(-0.0040, 0.1060)	(-0.0300, 0.0240)
	SE, LUQ	(-0.1060, 0.0120)	(-0.0440, 0.0240)	(-0.0540, 0.0620)	(-0.0320, 0.0200)
	Ecc, LUQ	(-0.0840, 0.0220)	(-0.0481, 0.0280)	(-0.0940, 0.0020)	(-0.0280, 0.0240)
GPT-4o-mini	Direct, Ecc	(0.0560, 0.1620)	(0.0280, 0.0960)	(0.0680, 0.1880)	(-0.0060, 0.0540)
	Direct, LUQ	(0.0320, 0.1500)	(0.0220, 0.0900)	(0.0740, 0.1920)	(-0.0180, 0.0320)
	Ecc, LUQ	(-0.0760, 0.0360)	(-0.0460, 0.0340)	(-0.0500, 0.0600)	(-0.0520, 0.0160)

Table 6: Confidence intervals for the accuracy differences between UQ methods in the MedClaim experiments, with the order of the compared UQ methods given in the UQ Pair column (Direct = direct prompting, SE = Semantic Entropy, Ecc = Eccentricity).