TINY: SEMANTIC-BASED UNCERTAINTY QUANTIFICATION IN LLMS: A CASE STUDY ON MEDICAL EXPLANATION GENERATION TASK.

Nicholas K. B. Tan
National University of Singapore
Singapore, 119077
nicholas.kb.tan@u.nus.edu

Mehul Motani
Department of Electrical and Computer Engineering
Institute of Data Science, Institute for Digital Medicine
N.1 Institute for Health, National University of Singapore
Singapore, 119077
motani@nus.edu.sg

ABSTRACT

Given the often sensible and sometimes nonsensical outputs that modern Large Language Models (LLMs) generate, how should we interpret confident claims such as 'Strawberry has two "r"s'? One tool that can be applied to such overconfident and hallucinatory claims is uncertainty quantification. In particular, this paper investigates a semantic density method to quantify uncertainty in LLM-generated medical explanations. Semantic density makes use of semantic similarity comparisons instead of lexical matching, and delivers per-response estimates of uncertainty. The results demonstrate that the semantic density framework remains performant when applied in specialized domains, and raises additional considerations around the utility of the ROUGE metric for semantic evaluations.

1 INTRODUCTION

As the performance of Large Language Models (LLMs) improves, their usage has increasingly shifted from natural language tasks to generalized tasks such as agentic execution, code execution, and robot learning. By utilizing them as part of a wider system, especially one that actively interacts with reality, the need for output evaluation and sanitation has become more pressing than ever. In particular, task-agnostic and data-independent metrics that can be applied on both open-sourced and close-sourced LLMs can be regarded as the holy grail for AI safety. Given that LLMs are probabilistic models, uncertainty quantification (UQ) methods are well-suited to this role. However, a consequence of the increasing size and complexity of modern LLMs is that traditional UQ methods are too computationally intensive for application to LLMs outside of research applications.

Given a generative model with parameters $\theta \sim P(\theta)$, the most direct method to quantify uncertainty given an input $x \in X$ is to measure the characteristics of the posterior distribution $P(Y|\theta, X = x)$. Unfortunately, it is intractable to compute the exact marginal likelihood to obtain the posterior distribution by applying Bayes Theorem to LLMs, due to the sheer number of parameters in θ. Therefore, most methods focus on estimating the posterior distribution via sampling-based or optimization-based methods, while other methods retrain base models with additional modules that output UQ estimates directly (Sankararaman et al., 2022). With modern LLMs, this is largely infeasible and highly unscalable due to the computational resources required to retrain and fine-tune each model.

Thus, we choose to look at test-time methods that attempt to extract UQ estimates from sampling LLM responses without retraining. Notable methods include those that directly estimate uncertainty based on output sequences (Lin et al., 2024) or direct elicitation (Xiong et al., 2024), as well as those that sample output logits to characterize the output logits distribution as a proxy to measure uncertainty (Malinin & Gales, 2021; Mielke et al., 2022; Kadavath et al., 2022; Qiu & Miikkulainen, 2024). For methods that utilize output logits, a key assumption made is that the logits are calibrated, with high confidence corresponding to a correct response (whether measured via factuality, relevance, or other metrics). While there are methods that focus on calibrating LLM outputs (Mielke et al., 2022), this work assumes that calibration is not of concern.

With a logits-based approach, information regarding the lexical sequence of outputs is readily available. However, a lexically valid answer may not necessarily be a semantically correct answer. Therefore, the set of generated outputs in this work is evaluated based on semantic consistency in addition to utilizing per-token logits in the output sequence. This allows for the quantification of uncertainty based on the semantic similarity between model outputs, which is more relevant in encouraging lexically diverse and varied generation outputs and a better proxy for performance on complex natural language processing tasks such as summarization and long-form question-answering.

In this work, the semantic density framework proposed by Qiu & Miikkulainen (2024), which extends semantic entropy as introduced by Kuhn et al. (2023) to generate per-response uncertainty estimates instead of per-prompt estimates, is evaluated. As the preceding works are evaluated on general long-form question answering datasets only, this work explores the performance of semantic density in specialized datasets of interest, namely explanation generation based on PubMedQA (Jin et al., 2019) and MedExQA datasets (Kim et al., 2024). This work[1] thus explores whether the general semantic space utilized in previous works is applicable in specialized domains.

2 METHODOLOGY

Semantic density utilizes the concept of a semantic space to measure the semantic similarity between different responses to a given input prompt. Given its causal model architecture and semantically-rich representation, a pre-trained natural language inference (NLI) model (specifically, Deberta-large-mnli (He et al., 2021)) is used by Kuhn et al. (2023) as a reference semantic space from which to compare the semantic similarity of outputs. Qiu & Miikkulainen (2024) then modifies an early variant of KDE to compute the semantic distance of a specific response y_* to the other responses based on the NLI model's output logits together with the output logits from the generation LLM.

The same experimental method and evaluation method used by Qiu & Miikkulainen (2024) and Kuhn et al. (2023) is utilized in our experiments. The following uncertainty metrics are computed for comparison against semantic density (SD): semantic entropy (SE) (Kuhn et al., 2023), degree (Deg) (Lin et al., 2024), length-normalized likelihood (NL) (Murray & Chiang, 2018), length-normalized entropy (NE) (Malinin & Gales, 2021) and P(True) and predictive entropy (PE) from Kadavath et al. (2022). Generation LLMs tested were the following: Mistral-7B-v0.1 (Jiang et al., 2023), Meta-Llama-3-8B (AI@Meta, 2024), and Llama-2-7b-hf (Touvron et al., 2023).

After the sample of responses is collected from each generation LLM, a binary label correct/incorrect is then computed based on the ROUGE-L score (Lin & Och, 2004) thresholded at 0.3 for each response. The uncertainty quantification methods below are subsequently evaluated against this binary label based on their Area Under the Receiver Operating Characteristic (AUROC) score.

Semantic density is evaluated on two medical question-answering datasets, PubMedQA and MedExQA, based on the task of free-form explanation generation. PubMedQA provides one reference explanation per sample, while MedExQA has two reference explanations per sample. The test split is used for both datasets, with MedExQA's 5 subject splits merged into a single test set. We note that PubMedQA is usually evaluated against provided 'yes/no/maybe' ground-truth labels; in this case, we evaluate against the accompanying explanations as the ground truth for generation.

3 PERFORMANCE ON MEDICAL EXPLANATION GENERATION

The AUROC score of each method is presented in Table 1. Semantic density outperforms other methods in 4 of the 6 cases, and is within the top three performing metrics for the remaining 2 cases.

3.1 BASELINE PERFORMANCE WITHOUT UNCERTAINTY

The average ROUGE-L performance for each model for the most likely generation (beam 0) is as shown below in Table 2. Our empirical Llama-2-7b-hf performance on the MedExQA dataset is higher than the baseline performance of 0.0492 reported by Kim et al. (2024). The low ROUGE-L performance (especially with many samples having 0 score) means that further experiments should be explored with better performing models more reflective of models in deployment.

[1]The source codes for this paper are provided at: https://github.com/plebianstribe/semantic-density-medexp.

Table 1: Performance of different uncertainty/confidence models across various LLMs and datasets

		PubMedQA					
AUROC	**SD**	SE	P(True)	Deg	NL	NE	PE
Mistral-7B-v0.1	**0.699**	0.599	0.664	0.526	0.653	0.578	0.578
Meta-Llama-3-8B	**0.700**	0.560	0.531	0.581	0.494	0.582	0.582
Llama-2-7b-hf	**0.713**	0.567	0.681	0.626	0.484	0.601	0.601
		MedExQA					
AUROC	**SD**	SE	P(True)	Deg	NL	NE	PE
Mistral-7B-v0.1	0.857	**0.902**	0.568	0.285	0.746	0.859	0.859
Meta-Llama-3-8B	0.805	0.764	**0.821**	0.616	0.608	0.712	0.712
Llama-2-7b-hf	**0.918**	0.600	0.717	0.461	0.352	0.531	0.531

Table 2: Average ROUGE-L scores for Beam 0 across various LLMs and datasets

PubMedQA	**Mistral-7B-v0.1**	**Meta-Llama-3-8B**	**Llama-2-7b-hf**
	0.150	0.173	0.194
MedExQA	**Mistral-7B-v0.1**	**Meta-Llama-3-8B**	**Llama-2-7b-hf**
	0.106	0.118	0.123

The distribution for semantic density against ROUGE-L values for the best performing beam for each model is visualized in Figure 1 for both PubMedQA and MedExQA. Note that the KDE plot is not generated for the positive MedExQA class as there are insufficient samples to plot.

The semantic density values seem to have no clear trend when compared against ROUGE-L values. This may stem from ROUGE being agnostic towards semantic content, instead relying on overlapping subsequences to compute a score. Thus, its possible that additional experiments utilizing a different metric that considers semantic content for evaluating responses against reference answers may result in different AUROC scores.

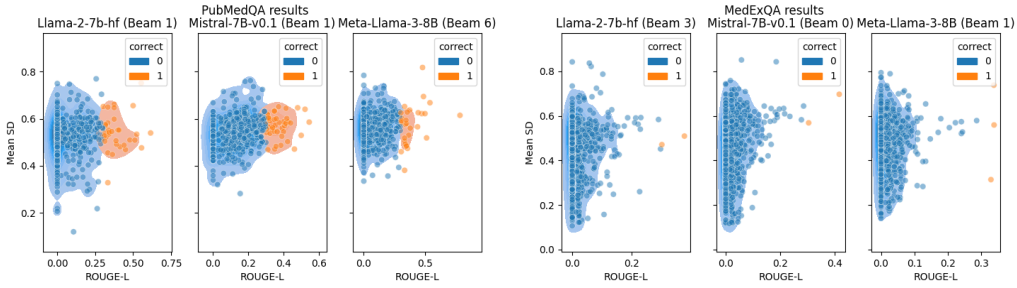

Figure 1: Distributions of semantic density estimate against ROUGE-L score

3.2 Future Work

Given the widespread literature on manual prompting methods such as CoT (Wei et al., 2022; Kojima et al., 2022) and MedPrompt (Nori et al., 2023) that improve the performance of LLMs, the application of UQ methods together with such strategies seems interesting. While there are some recent works in this area (Harsha Tanneru et al., 2024; Ling et al., 2024), it remains relatively understudied. Applying uncertainty quantification to automated prompting methods such as APE (Zhou et al., 2023) and OPRO (Yang et al., 2024) , could also yield additional insights into their underlying principles, and bridge the usage of UQ methods with general-purpose prompting strategies to reduce hallucinations in knowledge-intensive tasks.

REFERENCES

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/harsha-tanneru24a.html.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

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. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

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

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. MedExQA: Medical question answering benchmark with multiple explanations. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 167–181, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.14. URL https://aclanthology.org/2024.bionlp-1.14/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032. URL https://aclanthology.org/P04-1077/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty quantification for in-context learning of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3357–3370, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.184. URL https://aclanthology.org/2024.naacl-long.184/.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50/.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://aclanthology.org/W18-6322/.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. URL https://arxiv.org/abs/2311.16452.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Karthik Abinav Sankararaman, Sinong Wang, and Han Fang. Bayesformer: Transformer with uncertainty estimation. https://arxiv.org/abs/2206.00826, jun 2 2022. [Online; accessed 2024-04-15].

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=92gvk82DE-.