Controllable Epistemic Sensitivity in Large Language Models: Probing, Benchmarking, and Adaptive Reasoning

Srivarshinee S

Vellore Institute of Technology, Chennai

Abstract

This proposal aims to investigate epistemic uncertainty - uncertainty about knowledge or truth, often conveyed by modals like might or probably in Large Language Models (LLMs). By probing how such cues affect reasoning, we seek to achieve controllable epistemic sensitivity: enabling models to interpret and adapt to uncertainty. Using activation-level analyses and multilingual benchmarks, this work advances transparent, context-aware, and trustworthy reasoning in uncertainty-critical domains.

Introduction

Large Language Models (LLMs) are increasingly integrated into high-stakes domains such as healthcare, law, and public policy-settings where accurate reasoning under uncertainty is essential. While prior research has focused on calibrating output confidence and evaluating probability estimates, far less is known about how LLMs internally represent input-side uncertainty, particularly linguistic uncertainty expressed through epistemic modality (e.g., might, could, probably). Figure 1 explains this with an example - minimal shifts in modality, such as replacing should with could, might yield markedly different generations despite identical contexts, revealing a consistent and systematic sensitivity to uncertainty cues rather than random sampling variance.

Much like temperature settings, epistemic sensitivity could serve as a user- or context-specific hyperparameter, allowing adaptive control of how models interpret and respond to uncertain inputs. Epistemic modality, unlike other semantic features, directly reflects a speaker's confidence in a claim, a dimension particularly consequential in reasoning-driven applications such as medical diagnostics or legal reasoning.

Linguistic uncertainty is inherent in all natural language communication. Prior work shows that introducing epistemic uncertainty often degrades LLM reasoning (Holliday, Mandelkern, and Zhang 2024), as models treat modals like "must" and "might" as surface patterns rather than operators encoding necessity and possibility. Addressing this gap can enable models to reason more faithfully under ambiguity, enhancing interpretability and reliability.

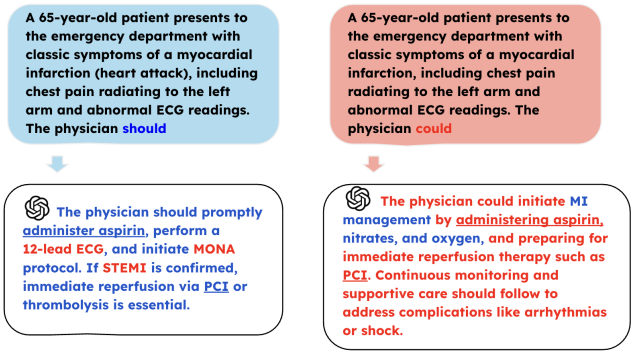

Figure 1: Although the prompt pairs differ only in epistemic modality (*should vs could*), the responses must vary: those prompted with *could* may offer a broader range of medical possibilities and more open-ended compared to those ending with *should*. The miscolors indicate minor differences in the output, while the underlines highlight the common elements.

This research aims to make epistemic modality a controllable aspect of LLM behavior. By enabling models to interpret and adapt to input uncertainty, it can reduce misinformation and enhance user trust in reasoning-intensive applications, from medical diagnostics to legal assistance.

Background

Uncertainty in LLM Outputs

Recent work on LLM uncertainty has focused on outputs via calibration (Desai and Durrett 2020), truthfulness (Lin, Hilton, and Evans 2022), or confidence alignment (Ghafouri et al. 2024) but does not examine how input uncertainty is internally represented or whether models differentiate certain from uncertain prompts.

Epistemics and Modal Reasoning

Several recent studies have examined the role of modal verbs in model reasoning. (Holliday, Mandelkern, and Zhang 2024) show that LLMs often struggle with logical tasks involving modal operators, suggesting a lack of systematic reasoning with modality.

(Zhou, Jurafsky, and Hashimoto 2023) show that epistemic markers in LLM outputs significantly affect accuracy, but do not examine their internal neural representations. Similarly, (Lee et al. 2025) find that LLM evaluators are systematically biased against responses expressing uncertainty.

This work aims to complement the above works by reading into the effects of input-side epistemic modality during generation.

Approach

This work approaches the question of epistemic modality in LLMs as both a problem in and for AI. Similar to interpretability studies, the objective is to understand how subtle variations in epistemic cues (e.g., might, could, probably) influence reasoning in LLMs. The focus lies in identifying where and how such cues are internally represented, and quantifying their effect on downstream behavior.

This work begins by systematically varying epistemic modality within controlled input contexts while holding all other semantics constant. Differences in output generations are then analyzed to isolate model sensitivity to input-side uncertainty. Probing and intervention methods are employed to localize and quantify where epistemic information is encoded across layers and components.

To uncover the internal mechanisms underlying epistemic sensitivity, this work employs and extends methods from mechanistic interpretability and representation analysis:

- Activation patching (Meng et al. 2023) to trace how epistemic information propagates through the network.
- Steering vectors (Panickssery et al. 2024) to evaluate whether latent activations can be manipulated to modulate epistemic sensitivity.
- Latent-space intervention techniques(Gat et al. 2021), vector manipulation (Huang, Chen, and Umrawal 2025) and sparse autoencoders (SAEs)(Gao et al. 2024), to operationalize fine-grained control over epistemic representations.

Findings from these analyses are used to explore the possibility of defining a controllable epistemic sensitivity hyperparameter, analogous to temperature, allowing context-dependent adjustment of reasoning fidelity.

A carefully curated dataset, augmented by AI will be constructed to evaluate epistemic sensitivity across domains and languages. This includes domain-specific datasets in medical, legal, and commonsense reasoning, alongside multilingual corpora that capture the gradient expression of epistemic modality in natural language.

The results from probing and intervention studies will form the basis of a new framework enabling dynamic control of reasoning behavior in LLMs based on input-side uncertainty. *Scoped to a single language and model, this work is expected to take 6-8 months and requires moderate compute, making it feasible.*

Evaluation

Given the novelty of this research area, no existing benchmark currently measures model sensitivity to epistemic modality. As part of this work, a multilingual benchmark spanning across high-, mid- and low-resource languages and a quantitative framework will be developed to evaluate how well model generations align with predefined levels of epistemic sensitivity. The benchmark will include reasoning-intensive tasks such as question answering and classification (e.g., medical diagnosis from natural language input), spanning multiple domains like healthcare and legal decision-making. Inputs will systematically vary in epistemic modality, with each instance annotated by corresponding sensitivity levels.

To quantify performance, a locality-perturbation-based sensitivity metric will be introduced to assess the consistency and directionality of model responses with respect to input uncertainty. This will enable local sensitivity analysis and cross-model comparison.

Discussion

This work is expected to reveal consistent patterns in how LLMs encode and process epistemic cues, identifying specific layers, neuron groups, or latent dimensions responsible for representing linguistic uncertainty. Comparative analyses across languages and resource tiers may further illuminate whether epistemic sensitivity correlates with cross-lingual transfer effects during training.

If successful, this research could offer a new lens for understanding how models interpret fine-grained linguistic semantics and introduce a principled mechanism for controllable reasoning. By enabling user- or context-specific modulation of epistemic sensitivity, it would advance transparency in LLM behavior. The resulting multilingual benchmark, evaluation metric, and control framework would provide valuable resources for the interpretability community.

Societally, such control can improve AI reliability in reasoning-intensive applications like medical diagnostics, legal assistance, and decision support by aligning responses with uncertainty cues, reducing overconfident outputs, and enhancing user trust and safety.

Conclusion

LLMs are increasingly used in reasoning-intensive domains, yet their sensitivity to input-side linguistic uncertainty, particularly epistemic modality, remains unclear. This work investigates how such cues are represented and influence reasoning through mechanistic probing, latent-space interventions, and curated multilingual datasets. A benchmark and evaluation framework will measure alignment between outputs and user-specified sensitivity levels across models, languages, and resource tiers. By revealing and controlling how LLMs interpret epistemic cues, this research enables fine-grained modulation of model outputs, enhancing transparency, reliability, and alignment with user intent. Even partial control over uncertainty can reduce overconfident responses in high-risk domains such as medical diagnostics or legal reasoning, improving AI-driven decision-making, reducing misinformation, and fostering user trust.

References

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295–302. Online: Association for Computational Linguistics.

Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. arXiv:2406.04093.

Gat, I.; Lorberbom, G.; Schwartz, I.; and Hazan, T. 2021. Latent Space Explanation by Intervention. arXiv:2112.04895.

Ghafouri, B.; Mohammadzadeh, S.; Zhou, J.; Nair, P.; Tian, J.-J.; Goel, M.; Rabbany, R.; Godbout, J.-F.; and Pelrine, K. 2024. Epistemic Integrity in Large Language Models. In *Neurips Safe Generative AI Workshop 2024*.

Holliday, W. H.; Mandelkern, M.; and Zhang, C. E. 2024. Conditional and Modal Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3800–3821. Miami, Florida, USA: Association for Computational Linguistics.

Huang, Y.; Chen, D.; and Umrawal, A. K. 2025. JAM: Controllable and Responsible Text Generation via Causal Reasoning and Latent Vector Manipulation. arXiv:2502.20684.

Lee, D.; Hwang, Y.; Kim, Y.; Park, J.; and Jung, K. 2025. Are LLM-Judges Robust to Expressions of Uncertainty? Investigating the effect of Epistemic Markers on LLM-based Evaluation. In *Proceedings of the 2025 NAACL-Long Papers*.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2023. Locating and Editing Factual Associations in GPT. arXiv:2202.05262.

Panickssery, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2024. Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681.

Zhou, K.; Jurafsky, D.; and Hashimoto, T. 2023. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. https://arxiv.org/abs/2302.13439. ArXiv:2302.13439.