CS772: Project Proposal

4-Armed Bandits

February 13, 2025

1 Introduction

Large language models (LLMs) are commonly used in academics, medicine, finance, and countless other structures. LLM outputs are uncertain by their generative nature, and the quantification for this uncertainty plays a key role in improving their reliability and accuracy. In particular, uncertainty can be used to enhance the trustworthiness of LLMs by detecting factually incorrect model responses, commonly called hallucinations. Critically, one should seek to capture the model's semantic uncertainty, i.e., the uncertainty over the meanings of LLM outputs, rather than uncertainty over lexical or syntactic variations that do not affect answer correctness.

To address this problem, a novel method for uncertainty estimation in white-box and black-box LLMs is proposed in Nikitin et al. (2024) called Kernel Language Entropy (KLE). KLE defines positive semidefinite unit trace kernels to encode the semantic similarities of LLM outputs and quantifies uncertainty using the von Neumann entropy. They theoretically prove that. KLE generalizes the previous state-of-the-art method called semantic entropy and empirically demonstrates that it improves uncertainty quantification performance across multiple natural language generation datasets and LLM architectures.

Our project aims to understand and re-implement the methodology and the experiments explained in the paper and introduce the concept of evaluation of the Gaussian process for text classification (Jayashree and Srijith (2020)) to compare the LLM outputs. We aim to make the model more robust and introduce a probabilistic mechanism to the underlying methodology.

2 Proposed Contributions

2.1 Integration of Probabilistic Latent Variable Models into KLE

Extend the KLE framework by incorporating probabilistic latent variable models such as Variational Autoencoders (VAEs) and Gaussian Processes (GPs). We can implement a Gaussian Process Latent Variable Model (Titsias and Lawrence (2010)) to capture uncertainty in the semantic space and compare the KLE scores derived from these representations to those obtained from observable token-level behaviour.

2.2 Multi-LLM Latent Truth Inference

We propose inferring a latent truth from multiple model outputs, treating each LLM's response as a noisy observation weighted by its reliability. Using latent variable models (e.g., Bayesian truth inference) and Kernel Language Entropy (KLE), we can combine responses to better estimate uncertainty and achieve a consensus. This approach leverages multiple models' collective insight and improves accuracy and robustness.

2.3 Benchmarking Across Diverse LLMs

Systematically benchmark a broader range of LLMs on multiple datasets to assess whether these models are implicitly trained for specific tasks. This evaluation will determine the generalizability of the proposed uncertainty measures and contribute valuable insights into the strengths and limitations of current LLM architectures.

3 Relevant Research

Recent research has laid a solid foundation for our proposed approach. Deep Gaussian Processes have been applied to text classification (Jayashree and Srijith (2020)), showing that Bayesian non-parametric models can capture complex language patterns. Probabilistic relation graphs have also been used to analyze word similarities in short texts (Alnahas et al. (2023)), providing structured insights into relational semantics. Classical methods, such as the EM algorithm for estimating observer error rates (Dawid and Skene (1979)), demonstrate how latent truth can be inferred from noisy data. The Learning From Crowds framework (Raykar et al. (2010)) further refines this idea by jointly estimating annotator reliability and true labels. In contrast, Truth Inference at Scale (Li et al. (2019)) offers scalable Bayesian models for handling redundant annotations. Building on these contributions, our work aims to infer a latent truth from multiple LLM outputs instead of selecting a single answer. We combine semantic uncertainty measures with probabilistic truth inference to capture consensus across models. This unified approach leverages deep learning and classical probabilistic methods to improve accuracy and reliability in language tasks.

4 Team Information

Following is the team information :

Name	**Roll No.**	**Email ID**
Aditi Khandelia	220061	aditikh22@iitk.ac.in
Arush Upadhyaya	220213	arushu22@iitk.ac.in
Kushagra Srivastava	220573	skushagra22@iitk.ac.in
Zehaan Naik	221238	zehaan22@iitk.ac.in

Table 1: Student Information

References

Alnahas, D., Ateş, A., Aydın, A. A., and Alagöz, B. B. (2023). Revisiting probabilistic relation analysis: Using probabilistic relation graphs for relational similarity analysis of words in short texts. *Turkish Journal of Mathematics and Computer Science*, 15(2):334–354.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jayashree, P. and Srijith, P. K. (2020). Evaluation of deep Gaussian processes for text classification. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1485–1491, Marseille, France. European Language Resources Association.

Li, Y., I. P. Rubinstein, B., and Cohn, T. (2019). Truth inference at scale: A bayesian model for adjudicating highly redundant crowd annotations. In *The World Wide Web Conference*, WWW '19, page 1028–1038, New York, NY, USA. Association for Computing Machinery.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of machine learning research*, 11(4).

Titsias, M. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy. PMLR.