Uncertainty Quantification for Large Language Models

Artem Shelmanov, Maxim Panov, Roman Vashurin,
Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin

Website: http://uncertainty-for-llm.nlpresearch.group

Large language models (LLMs) are widely used in NLP applications, but their tendency to produce hallucinations poses significant challenges to the reliability and safety, ultimately undermining user trust. This tutorial offers the first systematic introduction to uncertainty quantification (UQ) for LLMs in text generation tasks – a conceptual and methodological framework that provides tools for communicating the reliability of a model answer. This additional output could be leveraged for a range of downstream tasks, including hallucination detection and selective generation. We begin with the theoretical foundations of uncertainty, highlighting why techniques developed for classification might fall short in text generation. Building on this grounding, we survey state-of-the-art white-box and black-box UQ methods, from simple entropy-based scores to supervised probes over hidden states and attention weights, and show how they enable selective generation and hallucination detection. Additionally, we discuss the calibration of uncertainty scores for better interpretability. A key feature of the tutorial is practical examples using LM-Polygraph, an open-source framework that unifies more than a dozen recent UQ and calibration algorithms and provides a large-scale benchmark, allowing participants to implement UQ in their applications, as well as reproduce and extend experimental results with only a few lines of code. By the end of the session, researchers and practitioners will be equipped to (i) evaluate and compare existing UQ techniques, (ii) develop new methods, and (iii) implement UQ in their code for deploying safer, more trustworthy LLM-based systems.

Artem Shelmanov, Senior Research Scientist at MBZUAI, UAE.
Email: artem.shelmanov@mbzuai.ac.ae
Website: https://iinemo.github.io

Dr. Artem Shelmanov leads a research team focused on debiasing, hallucination detection, and uncertainty quantification methods for LLMs. His team has developed a series of robust UQ techniques for text classification models and generative LLMs, as well as a series of active learning algorithms for various NLP tasks. Dr. Artem leads the development of LM-Polygraph – one of the most comprehensive Python libraries for uncertainty quantification and hallucination detection in LLMs. He is also one of the organizers of the UncertainNLP workshop at EMNLP-2025.

Maxim Panov, Assistant Professor at MBZUAI, UAE.
Email: maxim.panov@mbzuai.ac.ae
Website: https://mbzuai.ac.ae/study/faculty/maxim-panov/

Maxim Panov's research is focused on uncertainty quantification for machine learning model predictions and Bayesian approaches to machine learning. In particular, Maxim's research group works on hallucination detection for LLMs. Maxim also co-leads the development of the LM-Polygraph library.

Roman Vashurin, Senior Research Engineer at MBZUAI, UAE.
Email: roman.vashurin@mbzuai.ac.ae

Roman Vashurin conducts research on uncertainty quantification in LLMs with a focus on unsupervised and semi-supervised approaches, developing new and extensively benchmarking existing methods. Roman is one of the core developers of the LM-Polygraph library.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 3–4

Artem Vazhentsev, PhD student at Skoltech and a researcher at AIRI.
Email: vazhentsev@airi.net
Artem Vazhentsev works on novel uncertainty quantification methods for LLMs and other NLP models. Artem developed several density-based and attention-based supervised UQ methods for text generation and classification models. He is one of the core developers of the LM-Polygraph library.

Ekaterina Fadeeva, PhD student at the LRE Lab, Department of Computer Science, ETH Zürich, Switzerland.
Email: efadeeva@ethz.ch
Ekaterina Fadeeva's research focuses on uncertainty quantification for LLMs, with applications to claim-level hallucination detection and reasoning models. She is one of the core developers of the LM-Polygraph library.

Timothy Baldwin, Provost, Professor at MBZUAI, UAE.
Email: timothy.baldwin@mbzuai.ac.ae
Website: https://mbzuai.ac.ae/study/faculty/timothy-baldwin/
Tim Baldwin is a full Professor in the Natural Language Processing department and Provost of MBZUAI, as well as a Melbourne Laureate Professor at The University of Melbourne. He was also the President of the Association for Computational Linguistics in 2022. His research interests encompass natural language processing, algorithmic fairness, AI safety, and computational social science. Tim has authored a series of publications related to uncertainty quantification and the interaction between debiasing and UQ techniques. He leads several projects focused on enhancing the safety and trustworthiness of LLMs.