Research Project in the Data Science Profile

Assessing the Uncertainty of LLM Responses

Situation

The Institute for Data Science FHNW has been working on Natural Language Processing for several years, more recently focusing on Large language model-based applications. LLMs have become part of our daily lives and are used for countless decisions. Answers generated by LLMs are not necessarily correct, even though these models usually appear confident and their answers very convincing. Incorrect statements can go undetected and have far-reaching consequences in many applications critical to humans, which is why methods for assessing the uncertainty of LLM answers are urgently needed.

The aim of this project is to investigate how the uncertainty of the generated answers can be determined. In a previous student project LLM uncertainty was analyzed for a first time showing the problem of overconfidence of LLM models. New research approaches have been deviced and there are different strategies or learning methods (unsupervised vs supervised) to determine LLM uncertainty.

Goals / Methodology / Tasks

The goal of this project is to analyze the problem of uncertainty estimation and calibration for LLMs. To this end you will investigate unsupervised and supervised approaches and apply them to a variety of state-of-the-art commercial/black-box and open-source/white-box models. These investigations will include an assessment of the impact of prompt engineering applied to a variety of tasks like e.g., Machine Translation, Question Answering or Multiple Choice for existing benchmark data. In view of industry applications the project will also critically analyze LLM uncertainty for real-life data from different domains like e.g., energy and health care.

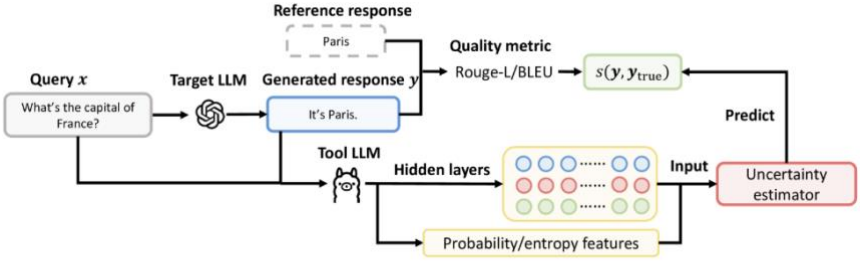

Figure 2: Illustration of our proposed supervised method. The tool LLM is an open-source LLM and can be different from the target LLM. In the training phase, where the reference response is available, we train the uncertainty estimator using the quality of the response as the label. In the test phase, the uncertainty estimator predicts the quality of the generated response to obtain an uncertainty score.

Required Skills

Good programming skills, interest in deep learning and especially in Natural Language Processing (NLP), willingness to work closely in a team, knowledge of Swiss German would be an advantage.

Tasks for the Master Student

The project can cover several aspects in 2 to 3 sub-projects (IP7, IP8, IP9).

Full/Parttime: [X] Full time study
[X] Part time study

Location: Windisch

Advisor: Prof. Dr. Daniel Perruchoud