Title:

Statistical methods for assessing the factual accuracy of large language models

Abstract:

We develop new statistical methods for obtaining validity guarantees on the output of large language models (LLMs). These methods enhance conformal methods to filter out claims (hallucination removal) while providing a finite-sample guarantee on the error rate of what it being presented to the user. This error rate is adaptive in the sense that it depends on the prompt to preserve the utility of the output by not removing too many claims. We shall explain how the theory works and demonstrate performance on real-world examples.

This is joint work with John Cherian and Isacc Gibbs.