Analyzing LLM Evaluations

Max Kuhn Simon Couch

2025-09-05

Frameworks exist for automating the evaluations of LLMs so that queries can be executed and assessed over many experimental factors: LLM models, prompts, replicates, etc. The resulting designs are often factorial in nature but can have a variety of hierarchical structures, such as replicates within queries, scores within raters, and so on.

This talk describes how experimental results can be analyzed and reported for a variety of designs and outcome types (percentage correct, correct/incorrect, ordinal scales, etc.). It also shows how off-the-shelf tools for Frequentist and Bayesian inferential analysis can be utilized. The methods are illustrated with an example evaluation experiment.