

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

5. LLMs Should Eat Their Words: Unveiling the Contradictions of LLM-as-Judges

Authors: Céline Leuzinger¹

Affiliations: 1. Department of Swedish, Multilingualism and Language Technology, University of Gothenburg, Sweden

Abstract: Evaluating the natural language outputs of generative AI systems represents an ongoing challenge in NLP, particularly in emerging fields such as Feedback Generation. Possible evaluation methods include automatic metrics such as BLEU or ROUGE, or manual checks with human evaluators. However, both approaches have their drawbacks: automatic metrics show little correlation with human judgements, while human evaluation is costly and time-consuming. In recent years, studies have turned to "LLM-as-judges" to evaluate the quality of AI-generated texts. This technique consists of prompting a Large Language Model (LLM) to give a score to a piece of textual data, as well as a short natural language justification. This promising approach could tackle the challenge of natural language evaluation if proven effective. However, a potential issue is the misalignment between the scores and natural language justifications provided. Preliminary manual checks on the outputs of an LLM-as-judge pipeline for feedback evaluation reveal that LLMs tend to attribute high score even when the justification is unfavourable, raising concerns regarding the reliability of such systems. In this experiment, we use few-shot learning to classify LLM-as-judge justifications for sentiment (Negative, Neutral, Positive) and unveil whether they align with the scores provided. We classify the outputs of various open-source and open-weight LLM-as-judges, prompted to generate a score and a justification for both AI-generated and human-crafted feedback.