

Peer Review

Review of: "DAFE: LLM-Based Evaluation Through Dynamic Arbitration for Free-Form Question-Answering"

Hyeoncheol Kim¹

1. Korea University, Seoul, South Korea

This paper introduces **DAFE**, a lightweight framework for evaluating free-form QA outputs using LLMs. It employs two primary LLM judges and selectively invokes a third arbitrator only in cases of disagreement, aiming to improve evaluation reliability while reducing computational cost. The approach is simple but well-motivated, mimicking human adjudication processes. Experiments across five QA datasets demonstrate that DAFE consistently aligns more closely with human judgments than EM, BERTScore, or single LLM evaluators. The framework is practical, easy to implement, and the writing is clear and well-organized.

The framework heavily depends on the quality of reference answers, which can lead to misjudgments when the references are incomplete or ambiguous. The paper also provides only a brief comparison with recent reference-free evaluation methods (e.g., MT-Bench, Chatbot Arena, G-Eval), leaving DAFE's positioning somewhat underexplored. Finally, while the selective arbitration structure is practical, the idea of involving a third-party adjudicator is a well-established practice in human evaluation and does not represent a methodologically novel contribution on its own.

This is a well-executed paper that offers a practical solution to improving the reliability of factual evaluations in free-form QA while minimizing cost. Although the structural novelty is limited, the overall contribution is meaningful and timely for LLM evaluation research.

Accept with Minor Revision

Declarations

Potential competing interests: No potential competing interests to declare.