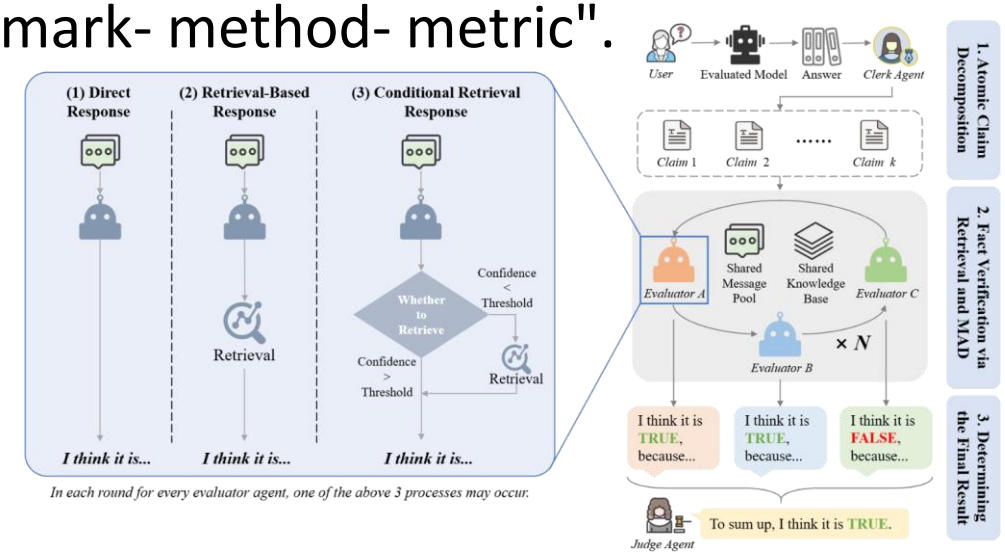
MAD-Fact: A Multi-Agent Debate Framework for Long-Form Factuality Evaluation in LLMs

Yucheng NING, Xixun LIN, Fang FANG, Yanan CAO

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51369-x

Problems & Ideas

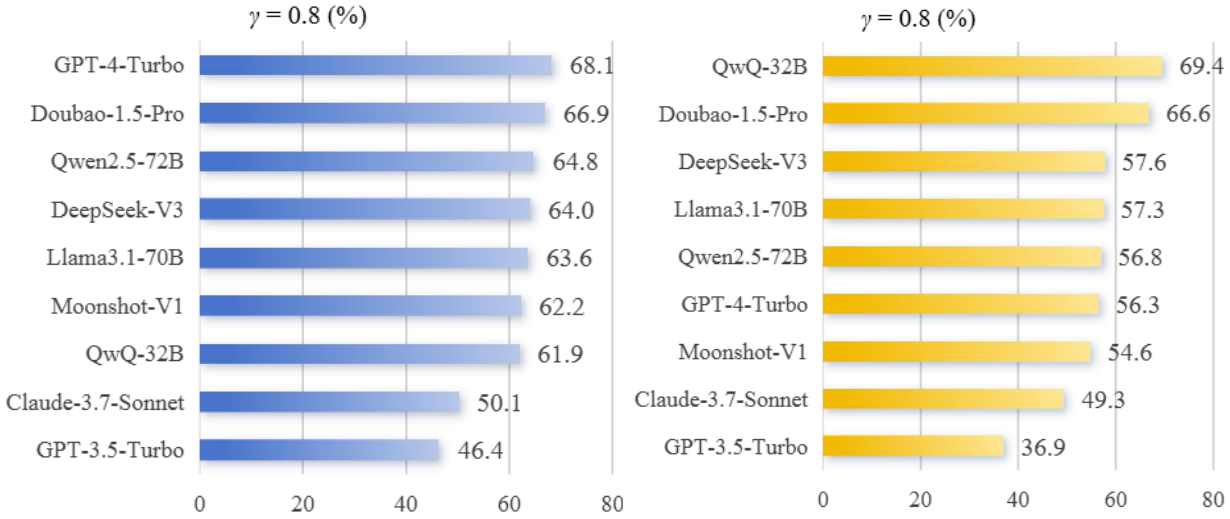
- Problems of current long-form factuality evaluation:
  - Most current benchmarks are designed for English, while resources for Chinese long-form evaluation are severely lacking.
  - Single-model evaluation architectures are prone to systematic biases, potentially misrepresenting the factuality of generated content.
  - Existing metrics typically treat all claims equally, without considering differences in their relative importance.
- Ideas: A innovative long-form factuality evaluation framework of "benchmark- method- metric".



The overall framework of the MAD-Fact system. The MAD-Fact system consists of three types of agents: the Clerk Agent decomposes the long-form response into multiple atomic claims; the Jury assesses the factuality of each atomic claim through external retrieval and multi-agent debate; The Judge Agent produce the final prediction and calculates score.

Main Contributions

- Contributions:
  - A Chinese long-form factuality benchmark, LongHalluQA, containing 2,746 high-quality samples across 7 topics;
  - A multi-agent debate framework, MAD-Fact, that enhances factual verification through structured interactions among Clerk, Jury, and Judge modules, outperforming strong baselines such as SAFE;
  - A fact importance hierarchy model that weights claims by significance, enabling weighted metrics strongly correlated with human judgments;
  - A benchmarking study of 9 major LLMs from 7 model families.



The long-form factuality evaluation results of the selected models. Left: the evaluation performance of the selected models on LongFact (English); Right: the evaluation performance of the selected models on LongHalluQA (Chinese).