Evaluating Large Language Models at Evaluating Instruction Following

Zhiyuan Zeng1, Jiatong Yu2, Tianyu Gao2, Yu Meng3, Tanya Goyal2, Danqi Chen2
1Tsinghua University 2Princeton University 3University of Illinois Urbana-Champaign

Background
- LLM evaluator = [base LLM] + [prompting strategy]
- Given one instruction and corresponding outputs from two models, an LLM evaluator is asked to choose a preferred one.
- meta-evaluation benchmark = [instruction] + [2 outputs] + [gold preference]
- Previous benchmarks: randomly-sampled output pairs & crowdsourcing
- Previous benchmarks suffer from inherent subjectivity of human preferences - One piece of evidence: Very low human agreement on existing benchmarks

Prompting Strategy
- LLMBar has the following qualities
  - All the instances are manually examined by authors
  - Only focuses on instruction-following and enforces objectivity
  - Includes two sets
    - Natural set: filtered and modified based on existing datasets
    - Adversarial set: Multiple ways of generating challenging candidate instances

Experiments
- We propose novel prompting for LLM evaluators (left) and use existing ones
- CoT: The LLM first generates a concise reasoning before its preference.
- ChatEval: Multiple LLMs with different prompts evoke a discussion.

Human Agreement Experiment
- We sample 80 instances randomly from LLMBar and assign each to two paper authors. LLMBar has a very high expert agreement rate of
  - 90% on the Natural set
  - 95% on the Adversarial set
  - 94% overall

A New Meta-Evaluation Benchmark: LLMBar
- LLMBar has the following qualities
  - All the instances are manually examined by authors
  - Only focuses on instruction-following and enforces objectivity
  - Includes two sets
    - Natural set: filtered and modified based on existing datasets
    - Adversarial set: Multiple ways of generating challenging candidate instances

Examples
- Average accuracies of 8 representative LLM evaluators on LLMBar:
  - Note that humans achieve an accuracy of 95% on (sampled) Adversarial set.
- Our proposed prompting strategies significantly improve the evaluators’ performance.

Results of GPT-4-based evaluators on LLMBar. Note that Random guess would achieve a 50% accuracy.

LLMBar can be used also for evaluating reward models (RMs) and preference models. Existing open-source RMs and preference models struggle to identify instruction-following outputs.

Ways to construct adversarial candidate instances.