To study this, we introduce two benchmarks with human ratings as the primary evaluation metric: MT-bench and Chatbot Arena. MT-bench is a series of open-ended questions that evaluate a chatbot’s multi-turn conversational and instruction-following ability – two critical elements for human preference. MT-bench is also carefully constructed to differentiate chatbots based on their core capabilities, such as reasoning and math. In addition, we develop Chatbot Arena, a crowdsourced platform featuring anonymous battles between chatbots in real-world scenarios – Users engage in conversations with two chatbots at the same time and rate their responses based on personal preferences.

Existing benchmarks mostly fall into the following three categories.

• Core-knowledge benchmarks, including MMLU [19], HellaSwag [50], ARC [9], WinoGrande [36], HumanEval [6], GSM-8K [10], and AGIEval [51], evaluate the core capabilities of pre-trained LLMs using zero-shot and few-shot benchmark sets. They typically require LLMs to generate a short, specific answer to benchmark questions that can be automatically validated.

• Instruction-following benchmarks, such as Flan [27, 46], Self-instruct [44], NaturalInstructions [28], Super-NaturalInstructions [45], expand to slightly more open-ended questions and more diverse tasks and are used to evaluate LLMs after instruction fine-tuning.

• Conversational benchmarks, like CoQA [35], MMDialog [15] and OpenAssistant [23], are closest to our intended use cases. However, the diversity and complexity of their questions often fall short in challenging the capabilities of the latest chatbots.

Types of LLM-as-a-Judge We propose 3 LLM-as-a-judge variations. They can be implemented independently or in combination:

• Pairwise comparison. An LLM judge is presented with a question and two answers, and tasked to determine which one is better or declare a tie. The prompt used is given in Figure 5 (Appendix).

• Single answer grading. Alternatively, an LLM judge is asked to directly assign a score to a single answer. The prompt used for this scenario is in Figure 6 (Appendix)

. • Reference-guided grading. In certain cases, it may be beneficial to provide a reference solution if applicable. An example prompt we use for grading math problems is in Figure 8 (Appendix).

LLM-as-a-judge offers two **key benefits: scalability and explainability.**

We identify certain biases and limitations of LLM judges. However, we will also present solutions later and show the agreement between LLM judges and humans is high despite these limitations. **Position bias** is when an LLM exhibits a propensity to favor certain positions over others.

Most LLM judges favor the first position

**Verbosity bias** is when an LLM judge favors longer, verbose responses, even if they are not as clear, high-quality, or accurate as shorter alternatives. As a calibration, we find LLM judges are able to correctly judge identical answers (i.e., they always return a tie for two identical answers) but cannot pass the more advanced “repetitive list” attack

**Self-enhancement bias.** We adopt the term “self-enhancement bias” from social cognition literature [4] to describe the effect that LLM judges may favor the answers generated by themselves

**Addressing limitations**

We present a few methods to address position bias and the limited grading ability for math questions.

**Swapping positions.** The position bias can be addressed by simple solutions. A conservative

approach is to call a judge twice by swapping the order of two answers and only declare a win when

an answer is preferred in both orders. If the results are inconsistent after swapping, we can call it a

tie. Another more aggressive approach is to assign positions randomly, which can be effective at a

large scale with the correct expectations. In the following experiments, we use the conservative one.

**Few-shot judge.** We assess whether few-shot examples can improve consistency in the position bias

benchmark. We select three good judgment examples using MT-bench-like questions, GPT-3.5 and

Vicuna for generating answers, and GPT-4 for generating judgments. The examples cover three cases:

A is better, B is better, and tie. As shown in Table 12 (Appendix), the few-shot judge can significantly

increase the consistency of GPT-4 from 65.0% to 77.5%. However, high consistency may not imply

high accuracy and we are not sure whether the few-shot examples will introduce new biases. Besides,

**the longer prompts make API calls 4× more expensive**. We use the zero-shot prompt by default in

our following experiments but leave an additional study in Appendix D.2.

**Chain-of-thought and reference-guided judge.** In Section 3.3, we have shown LLM’s limited

capability in grading math and reasoning questions. We propose two simple methods to mitigate

this issue: chain-of-thought judge and reference-guided judge. Chain-of-thought is a widely used

technique to improve LLM’s reasoning capability [47]. We propose a similar technique to prompt

an LLM judge to begin with answering the question independently and then start grading. Detailed

prompt in Figure 7 (Appendix). However, even with the CoT prompt, we find that in many cases

LLM makes exactly the same mistake as the given answers in its problem-solving process (See

example in Figure 15 (Appendix), suggesting that LLM judge may still be misled by the context.

Hence, we propose a reference-guided method, in which we first generate LLM judge’s answer

independently, and then display it as a reference answer in the judge prompt. In Table 4, we see a

significant improvement in failure rate (from 70% to 15%) over the default prompt.

**Fine-tuning a judge model.** We try fine-tuning a Vicuna-13B on arena data to act as a judge and

show some promising preliminary results in Appendix F.

**High agreement between GPT-4 and humans**

We compute agreement on MT-bench data. In Table 5, GPT-4 with both pairwise comparison and single answer grading show very high agreements with human experts. The agreement under setup S2 (w/o tie) between GPT-4 and humans reaches 85%, which is even higher than the agreement among humans (81%). This means GPT-4’s judgments closely align with the majority of humans.