while evaluating the final system output often involves assessing a single response (Stiennon et al., 2020).[1] (iii) *interpretability:* evaluation results are encouraged to provide more than solely numerical scores. Additional explanations are crucial to enhance the reliability of evaluation outcomes and facilitate humans' involvement in the evaluation loop (Saunders et al., 2022).

In this context, researchers have engaged in some explorations, with the central idea being to conceptualize evaluation as an instruction-following problem (Fu et al., 2023; Liu et al., 2023) based on a high-capacity LLM. For example, Zheng et al. (2023); Zhou et al. (2023); Dubois et al. (2023) employ proprietary LLMs (e.g., ChatGPT, Claude or GPT-4) through API calls to perform various evaluation protocols. Such methods have shown decent agreement with human judgment, but they also face challenges in terms of consistency and reproducibility due to the opacity of API models as well as the high API cost. An alternative is to train a specialized evaluator based on open-source LLMs. PandaLM (Wang et al., 2023c) is able to compare a pair of responses for a given query with a brief explanation of the evaluation process, and Shepherd (Wang et al., 2023b) can provide critiques to a LLM's response to pinpoint its shortcomings. These models have achieved remarkable performance in certain settings; however, they are relatively limited in the following aspects: (a) Some are not optimized to evaluate various deployed LLMs under massive real-world scenarios but are only trained on synthetic data (e.g., the Alpaca dataset Taori et al. (2023) by GPT-3.5), online forums, or traditional NLP datasets, without the consideration of scenario-specific evaluation criteria. (b) Each of these models only supports one evaluation protocol, like pairwise comparison or single-response evaluation, making them less flexible for various evaluation requirements. (c) They only provide brief or no natural language explanation for their evaluation, reducing the reliability of the result.

To address the above challenges, we develop AUTO-J, a generative judge with 13B parameters trained on user queries and model-generated responses from massive real-world scenarios. Methodologically, to train a more generalized judge, we created a new dataset from a large collection of data, encompassing 58 different scenarios, with most samples coming from real-world user queries and LLMs' responses. Based on the dataset, we guide GPT-4 (OpenAI, 2023) with carefully hand-written criteria for each scenario to collect desired evaluation judgments as our supervised training signals and apply heuristic filtering strategies and post-processing to unify output formats and mitigate noise. We also design new testbeds from the above dataset for pairwise comparison and single-response evaluation, with a diverse and balanced scenario distribution (§5.1). Through comprehensive meta-evaluation on its evaluation functionalities, we show that AUTO-J outperforms various strong baselines, including both open-source and closed-source models (§6.1, §6.2, §6.3). We also conduct detailed analysis and case studies (§6.4) to show a series of advantages offered by AUTO-J, from lessened positional bias in pairwise comparison, more specific critiques in single-response evaluation to the potential as a generative reward model to help improve base LLMs. To summarize, our contributions are:

(i) We develop AUTO-J, a new open-source model that can effectively and flexibly evaluate LLMs for both pairwise comparison and single-response assessment, with well-structured natural language critiques to support its evaluation. It establishes a new state-of-the-art performance among open-source models across all 58 scenarios (e.g., 8.9% improvement in pairwise evaluation in §6.1) and surpasses strong proprietary models such as ChatGPT and Claude-2 (e.g., with 12.1% and 12.4% gains in pairwise evaluation in §6.1) (ii) We construct a judgment dataset (§3) that covers 58 real-world scenarios. Each judgment consists of both a numerical rating (or a pairwise comparison result) and a critique generated in accordance with our curated 332 criteria. These data resources serve as a valuable foundation for both training and benchmarking evaluation methodologies under emerging technologies. (iii) We have released a wealth of resources to meet the diverse needs for future research: out-of-the-box models with superior performance; scenario typology and classifier; curated scenario-aware evaluation criterion and prompts; judgments with well-formatted critiques.

2 RELATED WORK

2.1 EVALUATION OF LLMS

It is universally known that the best way to evaluate LLMs is human judgment, but collecting human annotations can be costly, time-consuming, and laborious (Ouyang et al., 2022; Zheng et al., 2023).

[1] Traditional metrics such as BLEU and ROUGE are capable of but not adept at conducting pairwise evaluation due to the worse performance in sample-level evaluation. (Bhandari et al., 2020)