GENERATIVE JUDGE FOR EVALUATING ALIGNMENT

Junlong Li[1,6] **Shichao Sun**[3,6] **Weizhe Yuan**[4] **Run-Ze Fan**[5,6] **Hai Zhao**[1]
Pengfei Liu[1,2,6*]
[1]Shanghai Jiao Tong University [2]Shanghai Artificial Intelligence Laboratory
[3]Hong Kong Polytechnic University [4]New York University [5]Chinese Academy of Sciences
[6]Generative AI Research Lab (GAIR)

ABSTRACT

The rapid development of Large Language Models (LLMs) has substantially expanded the range of tasks they can address. In the field of Natural Language Processing (NLP), researchers have shifted their focus from conventional NLP tasks (e.g., sequence tagging and parsing) towards tasks that revolve around aligning with human needs (e.g., brainstorming and email writing). This shift in task distribution imposes new requirements on evaluating these aligned models regarding *generality* (i.e., assessing performance across diverse scenarios), *flexibility* (i.e., examining under different protocols), and *interpretability* (i.e., scrutinizing models with explanations). In this paper, we propose a generative judge with 13B parameters, AUTO-J, designed to address these challenges. Our model is trained on user queries and LLM-generated responses under massive real-world scenarios and accommodates diverse evaluation protocols (e.g., pairwise response comparison and single-response evaluation) with well-structured natural language critiques. To demonstrate the efficacy of our approach, we construct a new testbed covering 58 different scenarios. Experimentally, AUTO-J outperforms a series of strong competitors, including both open-source and closed-source models, by a large margin. We also provide detailed analysis and case studies to further reveal the potential of our method and make a variety of resources public at `https://github.com/GAIR-NLP/auto-j`

1 INTRODUCTION

In natural language processing, the *evaluation methodology* for generation tasks is continually updating with the advancement of *modeling techniques*, ranging from ROUGE (Lin, 2004) to ROUGE-WE (Ng & Abrecht, 2015) (a metric enhanced with word embedding (Mikolov et al., 2013)) and then to BERTScore (Zhang et al., 2019), BARTScore (Yuan et al., 2021), and GPTScore (Fu et al., 2023) (metrics enhanced by pre-trained language models (Peters et al., 2018; Devlin et al., 2019; Lewis et al., 2020)), aiming for a more reliable evaluation for ever-growing modeling techniques. Recently, the advent of large language models (Brown et al., 2020; Touvron et al., 2023a;b) has not only reshaped the implementation for modeling techniques (i.e., *paradigm shift* from "pre-train, fine-tuning" to "pre-train, supervised fine-tune, and reward model-based tune" (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022)) but also broadened the spectrum of tasks that modeling techniques seek to address (i.e., *task distribution shift* from traditional NLP tasks towards those more aligned with human needs (Bai et al., 2022a; OpenAI, 2023; Zhou et al., 2023; Taori et al., 2023)).

Given the evolving modeling techniques, the evaluation methods are in urgent need of upgrading and improvement to adapt to new challenges and requirements, particularly in the following aspects: (i) *generality*: the evaluation method should support massive real-world scenarios where gold references are usually unavailable. Traditional approaches frequently require human references and apply a single evaluation metric to constrained tasks (e.g., ROUGE (Lin, 2004) for text summarization, BLEU (Papineni et al., 2002) for machine translation) are struggling to keep pace with the current demands for evaluation. (ii) *flexibility*: the evaluation method should accommodate different protocols with desirable performance. The current LLM-based modeling paradigm requires methodological support of the evaluation in various aspects, and the evaluation protocols they demand also exhibit variations. For instance, when learning a reward model, it is necessary to compare two responses,

* Corresponding author