# From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge

**Dawei Li♠, Bohan Jiang♠, Liangjie Huang♣, Alimohammad Beigi♠, Chengshuai Zhao♠, Zhen Tan♠, Amrita Bhattacharjee♠, Yuxuan Jiang♦, Canyu Chen❦, Tianhao Wu❧, Kai Shu❥, Lu Cheng♣, Huan Liu♠**

♠Arizona State University, ♣University of Illinois Chicago,
♦University of Maryland, Baltimore County, ❦Northwestern University,
❧University of California, Berkeley, ❥Emory University

## Abstract

Assessment and evaluation have long been critical challenges in artificial intelligence (AI) and natural language processing (NLP). Traditional methods, usually matching-based or small model-based, often fall short in open-ended and dynamic scenarios. Recent advancements in Large Language Models (LLMs) inspire the "LLM-as-a-judge" paradigm, where LLMs are leveraged to perform scoring, ranking, or selection for various machine learning evaluation scenarios. This paper presents a comprehensive survey of LLM-based judgment and assessment, offering an in-depth overview to review this evolving field. We first provide the definition from both input and output perspectives. Then we introduce a systematic taxonomy to explore LLM-as-a-judge along three dimensions: *what* to judge, *how* to judge, and *how* to benchmark. Finally, we also highlight key challenges and promising future directions for this emerging area[1][2].

## 1 Introduction

Automatic model assessment and evaluation have long been essential yet challenging tasks in machine learning (ML) and natural language processing (NLP) (Sai et al., 2022; Chang et al., 2024). Traditional static metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure quality by calculating lexical overlap between output and reference texts. While computationally efficient, these metrics perform poorly in dynamic and open-ended scenarios (Liu et al., 2016; Reiter, 2018). With the rise of deep learning, small language model-based metrics like BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) have emerged. However, these metrics still face challenges in capturing nuanced attributes like fairness (Sun et al., 2022) and helpfulness (Zhu et al., 2024a).

Recently, the advancements of large language models (LLMs) such as GPT-4 (Achiam et al., 2023) and o1 (Jaech et al., 2024), have led to striking improvements in various applications, leveraging substantial prior knowledge in vast training corpora. This progress has motivated researchers to propose the concept of "LLM-as-a-judge" (Zheng et al., 2023; Wang et al., 2023c; Liu et al., 2023b; Chiang and Lee, 2023b), where LLMs are used to assess the candidate outputs by assigning scores, producing rankings, or selecting the best options, based on various input formats (e.g., point- and pair-wise), given context and instruction. The strong capability of LLMs combined with well-designed assessment pipelines (Li et al., 2023b; Bai et al., 2023a) leads to fine-grained and human-like judgment for various evaluation applications, addressing the previous limitations.

Beyond evaluation, LLMs-as-a-judge has been adopted across the lifecycle for next generations of LLM developments and applications. LLMs-as-a-judge is often used as a scalable way to provide supervisions for key development steps like alignment (Lee et al., 2023), retrieval (Li et al., 2024c), and reasoning (Liang et al., 2023). LLM-as-a-judge also empowers LLMs with a series of advanced capabilities such as self-evolution (Sun et al., 2024), active retrieval (Li et al., 2024c), and decision-making (Yang et al., 2023), driving their elevations from generative models to intelligent agents (Zhuge et al., 2024). However, as the field develops rapidly, challenges like bias and vulnerability (Koo et al., 2023; Park et al., 2024; Fu et al., 2024; Huang et al., 2024a) are emerging. Therefore, a systematic review of both techniques and limitations is crucial for facilitating this field.

This survey delves into the details of LLM-as-a-judge, aiming to provide a systematic overview of LLM-based judgment systems. We start by for-

[1] More resources on **LLM-as-a-judge** are on the website: https://llm-as-a-judge.github.io

[2] We have released and will maintain a paper list about **LLM-as-a-judge** at: https://github.com/llm-as-a-judge/Awesome-LLM-as-a-judge



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2792
November 4-9, 2025 ©2025 Association for Computational Linguistics

mally defining LLM-as-a-judge with its diverse input and output formats (Section 2). Next, we propose an in-depth and comprehensive taxonomy to address the three key questions (Section 3, 4 6):

- **Attribute: What to judge?** We outline six subtle attributes that are uniquely assessed by LLM-as-a-judge, including helpfulness, safety & security, reliability, relevance, logical, and overall quality.
- **Methodology: How to judge?** We explore ten tuning and prompting methods for LLM-as-a-judge, including manual labeling, synthetic feedback, supervised fine-tuning, preference learning, swapping operation, rule augmentation, multi-agent collaboration, demonstration, multi-turn interaction, and comparison acceleration.
- **Benchmark: How to evaluate LLM-as-a-judge?** We categorize existing benchmarks for LLM-as-a-judge into four types: for general performance, bias quantification, challenging tasks, and domain-specific performance.

Finally, we discuss challenges and potential future directions for LLM-as-a-judge in Section 7.

**Differences from Existing Surveys.** Existing concurrent surveys investigate LLM for the evaluation of natural language generation (NLG) (Gao et al., 2024; Li et al., 2024n; Gu et al., 2024). However, LLM-as-a-judge has been applied across a broader range of scenarios beyond evaluation, as we discussed, necessitating a systematic survey to categorize and summarize its various applications.

## 2 Preliminary

In this section, we provide a detailed definition of LLM-as-a-judge, discussing the various input and output formats as shown in Figure 1.

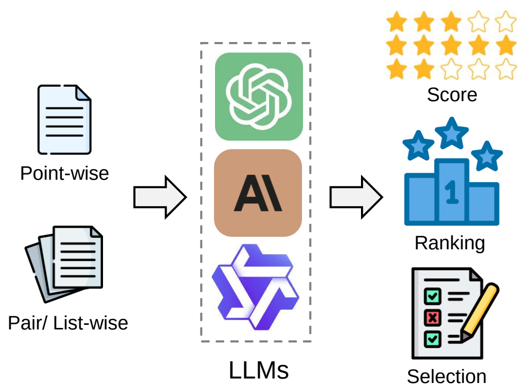


Figure 1: Overview of I/O formats of LLM-as-a-judge.

### 2.1 Input

Given a judge LLM $J$, the assessment process can be formulated as: $R = J(C_1, ...C_n)$. Here $C_i$ is the $i_{th}$ candidate to be judged and $R$ is the judging result. We categorize two input formats based on the different candidate numbers $n$.

**Point-Wise:** When $n = 1$, it becomes a point-wise judgment where the LLMs judges will solely focus on one candidate sample (Gao et al., 2023).

**Pair/ List-Wise:** When $n \geq 2$, it becomes a pair-wise ($n = 2$) or list-wise ($n > 2$) judgment where multiple candidate samples are provided together for the LLM judges to compare and make a comprehensive assessment (Zheng et al., 2023).

### 2.2 Output

In this section, we discuss three kinds of output of the judgment based on the different formats of $R$.

**Score:** When each candidate sample is assigned a continuous or discrete score: $R = \{C_1 : S_1, ..., C_n : S_n\}$, it becomes a score-based judgment. This is the most widely adopted protocol, leveraging LLMs to generate scores for quantitative comparisons (Li et al., 2024a) or attribute detection (Xie et al., 2024a).

**Ranking:** In ranking-based judgment, the output is a ranking of each candidate sample, represented as $R = \{C_i > ... > C_j\}$. This comparative approach is useful in scenarios where establishing a rank order among candidates is required (Li et al., 2023b; Liu et al., 2024b).

**Selection:** In selection-based judgment, the output involves selecting one or more optimal candidates, represented as $R = \{C_i, ..., C_j\} > \{C_1, ...C_n\}$. This method is particularly crucial in decision-making (Yao et al., 2023a) or content-filtering (Li et al., 2024c) contexts.

## 3 Attribute

In this section, we categorize current research in LLM-as-a-judge from attribute perspectives. Figure 2 gives an overview summarization of what aspects can be assessed by the LLM judges.

### 3.1 Helpfulness

Helpfulness is a critical criterion to measure the utility and informativeness of a generated response. Due to the high cost of manually assessing helpfulness in training data, recent studies have explored leveraging LLMs to label helpfulness and to generate or filter alignment data (Bai et al., 2022; Lee


Figure 2: Overview of different judging aspects.

et al., 2023; Guo et al., 2024; Zhang et al., 2025d). Beyond alignment tuning, helpfulness assessment using LLM-as-a-judge also plays a vital role in automatic model evaluation (Zheng et al., 2023; Lin et al., 2023; Li et al., 2024e; Zhang et al., 2025a).

### 3.2 Safety & Security

Safety and security are essential to ensure that models do not generate harmful content or respond inappropriately to malicious inputs. Current studies have validated that LLMs can be effectively used for model safety assessment, either as off-the-shelf models guided by policy instructions (Bai et al., 2022; Phute et al., 2023; Li et al.; Ye et al., 2024b; Wang et al., 2024l; Eiras et al., 2025; Chen and Goldfarb-Tarrant, 2025; Rodriguez et al., 2025; Hengle et al., 2025), or as lightweight models fine-tuned on safety-specific datasets (Inan et al., 2023; Zhang et al., 2024f; Xie et al., 2024a). Besides, LLM-as-a-judge has been widely adopted to detect and purify adversarial and toxic prompts designed with malicious intent (Cantini et al., 2025; Mu et al., 2025; Armstrong et al., 2025).

### 3.3 Reliability

Reliability is a crucial attribute for LLMs, enabling them to generate faithful content while presenting uncertainty or acknowledging missing knowledge about certain topics. Regarding sentence-level faithfulness assessment, existing researches leverage LLM-as-a-judge to either instruct the powerful LLMs (e.g., GPT-4) directly (Cheng et al., 2023; Gekhman et al., 2023; Luo et al., 2024a; Hsu et al., 2024) or train specific reliability judges (Wang et al., 2024a). Several works adopt LLM judges for long-form and fine-grained faithfulness evaluation (Tan et al., 2024a; Bai et al., 2024; Wu et al., 2025), using external retrieval bases (Min et al., 2023; Cao et al., 2025b; Loru et al., 2025) or search engines (Wei et al., 2024b). Jing et al. (2024); Pu et al. (2025) further expand this assessment to the multimodal area. Besides evaluation, there are also many works that adopt LLM-as-a-judge to improve the reliability of the generated content, either by external verifiers (Xie et al., 2024b) or synthetic alignment datasets (Zhang et al., 2024g; Wen et al., 2024). For uncertainty judgment, Xu et al. (2024d) propose SaySelf, a training framework that teaches LLMs to express more fine-grained confidence estimates with self-consistency prompting and group-based calibration training.

### 3.4 Relevance

Relevance assessment with LLM-as-a-judge has been explored and validated to be a more refined and effective manner across various tasks (Chiang and Lee, 2023a; Arabzadeh and Clarke, 2025a). In conversation evaluation, both Lin and Chen (2023a) and Abbasiantaeb et al. (2024) propose to replace expensive human annotation with LLM judgment in relevance assessment. In retrieval-augmented generation (RAG) scenarios, there are also many works that utilize LLMs to determine which demonstrations (Li and Qiu, 2023a) or documents (Li et al., 2024c) are most relevant for solving the current problem. Recently, LLM-as-a-judge has also been used in multimodal applications for cross-modality relevance judgment (Lee et al., 2024b; Chen et al., 2024g; Yang and Lin, 2024; Chen et al., 2024a; Lu et al., 2024b; Luo et al., 2024b; Lin et al., 2025). Additionally, LLM-as-a-judge has also been explored in many traditional retrieval applications for relevance assessment (Zhao et al., 2023a; Alaofi et al., 2024; Dietz et al., 2025; Arabzadeh and Clarke, 2025b; Balog et al., 2025), such as search (Thomas et al., 2024; Sebastian and Hoppe, 2025), retrieval (Ma et al., 2024; Dey et al., 2025), recommendation (Hou et al., 2024; Zhang et al., 2024h).

### 3.5 Logic

In agentic LLMs, assessing the logical correctness of candidate actions or steps is crucial for LLMs' planning, reasoning and decision-making, which further releases their great potential at inference-time. While some works leverage metrics or external tools for this feasibility assessment (Huang et al., 2023a; Yuan et al.), many others leverage LLMs' feedback as the signal (Lightman et al.; Kawabata and Sugawara, 2024) to perform planning and searching in complex reasoning spaces (Hao et al., 2023; Yao et al., 2023a; Besta et al., 2024). In multi-agent collaboration systems, both Liang et al. (2023) and Li et al. (2024b) propose to leverage the judge LLM to select the most feasible solutions among multiple candidates' responses. Besides, other works adopt LLM judges to perform logical assessment in API selection (Zhao et al., 2024b), tool using (Yang et al., 2023) and LLM routing (Ong et al., 2024).

### 3.6 Overall Quality

As previously mentioned, LLM-as-a-judge can be employed to perform multi-aspect and fine-grained assessments. However, in many cases, a general assessment is still required to represent the candidates' overall quality. One straightforward approach to obtain this overall score is based on the aspect-specific scores, either by averaging them (Lin et al., 2023) or referring them to generate an overall judgment (Yu et al., 2024c). Moreover, in many traditional NLP tasks (Lu et al., 2024a; Jiang et al., 2024; Ho et al., 2025; Shibata and Miyamura, 2025; Kartáč et al., 2025) like summarization (Gao et al., 2023; Jain et al., 2023a; Chen et al., 2023; Kumar et al., 2024a; Qi et al., 2025; Barnes et al., 2025; Altemeyer et al., 2025; Jeong et al., 2025; Calderon et al., 2025) and machine translation (Kocmi and Federmann, 2023; Huang et al., 2024b; Piergentili et al., 2025; Wang et al., 2025d), the evaluation dimensions are less diverse compared to more open-ended, long-form generation tasks. As a result, LLM-as-a-judge is often prompted directly to produce an overall judgment in these tasks.

## 4 Methodology

In this section, we present commonly adopted methods and tricks to improve LLMs' judging capabilities, splitting them into tuning (Section 4.1) and prompting strategies (Section 4.2).

### 4.1 Tuning

To enhance the judging capabilities of a general LLM, various tuning techniques have been employed in different studies. In this section, we discuss these tuning approaches for LLM-as-a-judge from two perspectives: data sources (Section 4.1.1) and training techniques (Section 4.1.2).

#### 4.1.1 Data Source

**Manually-labeled Data:** To train a LLM judge with human-like criteria, one intuitive method is to collect manually-labeled judgments. Previous works have leveraged and integrated existing sources annotated by humans, including instruction tuning datasets (Lee et al., 2024a; Wang et al., 2024k) and traditional NLP datasets (Vu et al., 2024), for tuning LLM judges. Other works collect manually-labeled datasets with fine-grained judgment feedback. These fine-grained feedbacks can be rationales behind judgment results (Xu et al., 2023a), multi-aspect judgment formats (Liu et al., 2024a) and fine-grained judgment labels (Yue et al., 2023), all of which facilitate the LLM judges to produce more detailed and context-rich judging results. Notably, Ke et al. (2024) first prompt GPT-4 to generate judgment and then manually verify and revise the outputs to ensure high-quality annotations.

**Synthetic Feedback:** While manually labeled feedback is high-quality and accurately reflects human judgment preferences, it is limited in both scale and coverage. To address it, researchers have also explored synthetic feedback as a data source for LLM judges' tuning. Some rely on the LLM judges themselves to generate the synthetic feedback. It involves instructing the LLM to self-evaluate and improve its judgments (Wu et al., 2024a), or by generating corrupted instructions and corresponding responses as negative samples for Directed Preference Optimization (DPO) training (Wang et al., 2024h). Besides, other powerful and stronger LLMs are also introduced for feedback synthesis. For example, GPT-4 has been widely leveraged to synthesize judging evidence (Wang et al., 2024a), erroneous responses (Park et al., 2024), rationale and feedback (Li et al., 2024e; Kim et al., 2024b; Xiong et al., 2024), and judgment labels (Zhu et al., 2023; Xie et al., 2024a).

#### 4.1.2 Tuning Techniques

**Supervised Fine-tuning:** Supervised fine-tuning (SFT) is the most widely used approach for training LLM judges (Hu et al., 2025a), enabling them to learn from pairwise (Li et al., 2024e; Wang et al., 2023b; Zhu et al., 2023; Wang et al., 2024k; Pombal et al., 2025b; Salinas et al., 2025) or pointwise (Wang et al., 2023b; Yue et al., 2023; Xie et al., 2024a; Lee et al., 2024a; Chiang et al., 2025) judgment data. Among many tricks applied in SFT, multi-task training and weight merging are introduced to enhance the robustness and generalization of LLM judges (Kim et al., 2024b; Vu et al., 2024; Saad-Falcon et al., 2024b). Other works try to enrich the original training set with augmented or self-generated samples. Ke et al. (2024) augment pairwise training data by swapping the order of two generated texts and exchanging the corresponding content in critiques. Xu et al. (2023a) further fine-tune their INSTRUCTSCORE model on self-generated outputs to align diagnostic reports better with human judgment. Additionally, Liu et al. (2024a) propose a two-stage SFT process: an initial phase of vanilla instruction tuning for evaluation diversity, followed by additional tuning with auxiliary aspects to enrich the model's evaluative depth.

**Reinforcement Learning:** Reinforcement learning from human preference is closely tied to judgment and evaluation tasks, particularly those involving comparison and ranking. Rather than directly adopt or augment preference learning datasets for SFT, several studies apply preference learning techniques to enhance LLMs' judging capabilities. One straightforward way is to treat the off-topic responses as inferior samples and apply DPO (Wang et al., 2024a; Yu et al., 2025; Rad et al., 2025). Besides, Wu et al. (2024a) propose meta-rewarding, which leverages the policy LLMs to judge the quality of their own judgment and produce pairwise signals for enhancing the LLMs' judging capability. This concept is also adopted by Wang et al. (2024h), who propose self-taught evaluators that use corrupted instructions to generate suboptimal responses as inferior examples for preference learning. Moreover, Hu et al. (2024b) introduce rating-guided DPO, in which the rating difference between two responses is considered in preferences modeling. Different from RLHF- and DPO-based approaches, several recent works leverage reinforcement learning with verifiable reward (RLVR) (Guo et al., 2025) to train LLM judges by rewarding reasoning trajectories that lead to correct judgments (Saha et al., 2025; Liu et al., 2025e; Zhou et al., 2025).

### 4.2 Prompting

Designing appropriate prompting strategies and pipelines at the inference stage could improve judgment accuracy and mitigate bias. We summarize existing prompting strategies for LLM-as-a-judge into six categories (see Figure 3).

#### 4.2.1 Swapping Operation

Previous studies have demonstrated that LLM-based judges are sensitive to the positions of candidates, and the ranking results of candidate responses can be easily manipulated by merely altering their order in the context (Wang et al., 2023d). To mitigate this positional bias and establish a more fair LLM judging system, (Zheng et al., 2023) propose a swapping operation, which involves invoking the judge LLM twice, swapping the order of the two candidates in each instance. If the two results are inconsistent, it is labeled a "tie", indicating that the LLM is unable to confidently distinguish the quality of the candidates. This swapping operation technique has also been widely adopted in pairwise feedback synthesis to produce more accurate reward signals (Lee et al., 2023; Sun et al., 2024; Lee et al., 2024a).

#### 4.2.2 Rule Augmentation

Rule-augmented prompting involves embedding a set of principles, references, and evaluation rubrics directly within the prompt for LLM judges. This approach is commonly employed in LLM-based evaluations, where LLM judges are guided to assess specific aspects (Lahoti et al., 2023; Li et al., 2024e; Bai et al., 2023a; Yu et al., 2024c; Qian et al., 2024; Dong et al., 2024; Wei et al., 2025; Xie et al., 2025b) and provided with detailed rubrics and criteria (Gao et al., 2023; Kim et al.; Wang et al., 2024g; Murugadoss et al., 2024; Li et al., 2024l,h; Hu et al., 2024a; Liu et al., 2024d; Li et al., 2025b; Fan et al., 2025) to ensure accurate judgments. Following this concept, studies in alignment (Bai et al., 2022; Lee et al., 2023, 2024a; Guo et al., 2024; Sun et al., 2024; Beigi et al., 2024) enhance this principle-driven prompting by incorporating more detailed explanations for each aspect of the principle or rubric. Apart from these human-written rules, some works (Liu et al., 2024c; Zhang et al., 2024f; Xu et al., 2025b; Wen et al., 2025; Zhou et al., 2024a) embed the self-generated or automaticaly-searched scoring criteria and principles as a part of their instruction.

#### 4.2.3 Multi-agent Collaboration

Accessing results from a single LLM judge may not be reliable due to inherent biases in LLMs (Wang et al., 2023d; Liusie et al., 2024; Ohi et al., 2024). To address this limitation, Li et al. (2023b); Chen

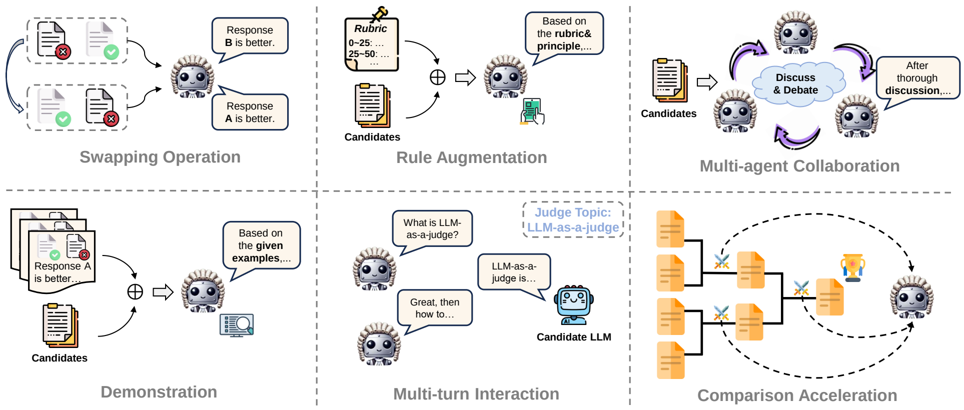


Figure 3: Overview of prompting strategies for LLM-as-a-judge.

et al. (2024c); Ning et al. (2024) introduce the Peer Rank (PR) algorithm, which produces the final ranking based on each LLM judge's output. Building on this, several architectures and techniques for multi-agent LLMs emerge, including mixture-of-agent (Zhang et al., 2023; Xu et al., 2023b; Beigi et al., 2024; Cao et al., 2025a), role play (Wu et al., 2023; Li et al., 2024m; Patel et al., 2024), debating (Chan et al., 2023; Zhang et al., 2024e; Bandi and Harrasse, 2024; Kenton et al., 2024), voting & aggregation (Zhu et al., 2024c; Verga et al., 2024; Li et al., 2025c; Guerdan et al., 2025; Rahmani et al., 2024) and cascaded selection Jung et al. (2024); Badshah and Sajjad (2025). Additionally, others apply multi-agent collaboration for alignment data synthesis, leveraging multiple LLM judges to refine responses (Arif et al., 2024) or provide more accurate feedback (Li et al., 2024i).

#### 4.2.4 Demonstration

In-context samples or demonstrations (Brown et al., 2020; Dong et al., 2023; Agarwal et al.) provide concrete examples for LLMs to follow and have been shown to be a crucial factor in the success of in-context learning for LLMs. Several studies have introduced human assessment results as demonstrations for LLMs-as-judges, aiming to help LLMs learn evaluation standards from a few illustrative examples (Jain et al., 2023b; Kotonya et al., 2023). To improve the robustness of LLM evaluations, Hasanbeig et al. (2023) propose ALLURE, an approach that iteratively incorporates demonstrations of significant deviations to enhance the evaluator's robustness. Additionally, Song et al. (2024b) borrow the insights from many-shot in-context learning and apply it in LLM-as-a-judge applications.

#### 4.2.5 Multi-turn Interaction

A single response may not provide enough information for an LLM judge to thoroughly and fairly assess each candidate. To address this limitation, multi-turn interactions are proposed to offer a more comprehensive evaluation. Typically, the process begins with an initial query or topic, followed by dynamically interacting between the LLM judge and candidate models (Bai et al., 2023b; Yu et al., 2024c; Pombal et al., 2025a). Besides, some approaches facilitate debates among candidates in a multi-round manner, allowing their true knowledge and performance to be fully revealed and evaluated (Zhao et al., 2024c; Moniri et al., 2024).

#### 4.2.6 Comparison Acceleration

Among various input formats in LLM-as-a-judge, pair-wise comparison is the most common approach for model comparison in evaluation or producing pair-wise feedback for training. However, when multiple candidates need to be ranked, this method can be quite time-consuming (Zhai et al., 2024). To mitigate the computational overhead, Zhai et al. (2024) propose a ranked pairing method in which all candidates are compared against an intermediate baseline response. In addition, Lee et al. (2024a); Liu et al. (2025d) utilize a tournament-based approach (Liu et al., 2023a; Zhao et al., 2023b) for rejection sampling during inference to speed up the pair-wise comparison.

## 5 Application

We introduce four applications which LLM-as-a-judge can be applied: evaluation (Section 5.1), alignment (Section 5.2), retrieval (Section 5.3), and reasoning (Section ??). Due to the space limitation, we provide a more detailed version in Appendix C.

### 5.1 Evaluation

LLM judges are initially proposed for and widely adopted in various evaluation scenarios. For open-ended generation, LLM judges assess the quality of outputs like dialogues, summaries, and creative writing, ensuring contextual relevance, coherence, and safety (Badshah and Sajjad, 2024; Kumar et al., 2024b; Zeng et al.; Jones et al., 2024). For reasoning tasks, they judge intermediate steps and final answers (He et al., 2023; Parmar et al., 2024; Xia et al., 2024) in areas such as math (Xia et al., 2024), logic (Parmar et al., 2024), and temporal reasoning (Fatemi et al., 2024). There are also some emerging areas where LLM judges are applied to domains once dominated by humans, including social intelligence (Zhou et al., 2023), multimodal tasks (Chen et al.) and multilingual generation (Fu and Liu, 2025).

### 5.2 Alignment

Model alignment also benefits from the automatic LLM-as-a-judge to produce and filter data at scale. Typically, larger and powerful LLMs are usually used as judges to align smaller models, providing synthetic preference data. This includes methods like multi-agent collaboration (Arif et al., 2024) and specialized tasks such as code alignment (Weyssow et al., 2024). Additionally, self-judging methods have LLMs rank or critique their own outputs to generate preference data without external teachers. To improve the judging capability of the policy model, techniques such as meta-rewarding (Wu et al., 2024a), Judge Augmented Supervised Fine-Tuning (JSFT) (Lee et al., 2024a), and self-evaluation (Zhang et al., 2024g) have been proposed. Apart from pairwise data, some other studies also use LLM-as-a-judge to judge and filter synthetic SFT data for instruction tuning (Liang et al., 2024c; Yasunaga et al., 2024).

### 5.3 Retrieval

LLM judges can assist with both traditional retrieval tasks and retrieval-augmented generation (RAG). For traditional retrieval, LLM-as-a-judge ranks documents by relevance (Zhuang et al., 2024a) without task-specific data (Ma et al., 2023), using permutation-based (Sun et al., 2023), pairwise (Qin et al., 2024), and listwise (Zhuang et al., 2024b) approaches to improve reranking for complex queries and domain-specific search tasks. For RAG, LLM judges guide how external knowledge is fetched and used during generation, ensuring coherence, accuracy, and relevance. This includes frameworks like Memory-of-Thought (Li and Qiu, 2023b), Self-Retrieval (Tang et al., 2024a), and Self-RAG (Asai et al.), where the judge selects or filters retrieved content, particularly in specialized fields such as biomedicine (Li et al., 2024c).

### 5.4 Reasoning

Reasoning is a critical capability of LLMs for complex and dynamic problem-solving. LLM judges can aid reasoning tasks by improving reasoning path selection and external tool use. Reasoning path selection involves identifying the correct trajectory for the LLM's reasoning process, where LLM-as-a-judge are adopted to evaluate intermediate reasoning steps (Lahoti et al., 2023), perform trajectory-level selection (Musolesi, 2024), and act as a process reward model for reasoning state scoring (Lightman et al., 2023) or a fine-grained critic to provide verbal feedback (Ankner et al., 2024). For external tool use, LLM judges help AI systems decide which external tools, modules, or agents to activate at each step of reasoning, acting as controllers that coordinate tool choice (Sha et al., 2023), agent communication (Ong et al., 2024), and message flow management (Liang et al., 2023) to ensure accurate and coherent problem solving.

## 6 Benchmark: Judging LLM-as-a-judge

We categorize benchmarks for evaluating LLMs-as-judges into four groups: general performance (Section 6.1), bias quantification (Section 6.2), challenging task performance (Section 6.3), and domain-specific performance (Section 6.4).

### 6.1 General Performance

Benchmarks focusing on general performance aim to evaluate the overall competence of LLMs in various tasks. One direct way to benchmark LLM judges' performance is to calculate the alignment between LLM prediction and the manual judgment result, using various metrics like Cohen's kappa, Discernment Score, and normalized accuracy (Li et al., 2023a; Tan et al., 2024b; Wang et al., 2024j; Lambert et al., 2024; Penfever et al., 2024; Qu et al., 2025; Xu et al., 2025a; Chang et al., 2025; Hu et al., 2025b; Calderon et al., 2025; Elangovan et al., 2024; Schroeder and Wood-Doughty, 2024; Gera et al., 2024). Moreover, several studies build LLM leaderboards using LLM-as-a-judge and assess their validity by comparing model rankings with those from established benchmarks and leaderboards, such as Chatbot Arena (Zheng et al., 2023)) (Zheng et al., 2023; Dubois et al., 2024; Li et al., 2024k; Zhao et al., 2024c; Chi et al., 2025).

### 6.2 Bias Quantification

Quantifying and mitigating bias in LLM judgments is critical to ensuring fairness and reliability (Xie et al., 2025a). Typical benchmarks include EvalBiasBench (Park et al., 2024) and CALM (Ye et al., 2024a), focus explicitly on quantifying biases, including those emerging from alignment and robustness under adversarial conditions. Besides, Shi et al. (2024) adopt metrics such as position bias and percent agreement in question-answering tasks. Recently, (Tripathi et al., 2025) examine the influence of protocol choice (pairwise and pointwise) on the bias degree of LLM judges.

### 6.3 Challenging Task Performance

Benchmarks designed for difficult tasks push the boundaries of LLM evaluation. For example, Arena-Hard (Li et al., 2024k) and JudgeBench (Tan et al., 2024b) select harder questions based on LLMs' performance for conversational QA and various reasoning tasks, respectively. CALM (Ye et al., 2024a) explores alignment and challenging scenarios, using metrics like separability and agreement to evaluate performance in manually identified hard datasets.

### 6.4 Domain-Specific Performance

Domain-specific benchmarks provide task-focused evaluations to assess LLMs' effectiveness in specialized contexts. Concretely, Raju et al. (2024) measure separability and agreement across tasks in specific domains such as coding, medical, finance, law and mathematics. CodeJudge-Eval (Zhao et al., 2024a) specifically evaluates LLMs for judging code generation with execution-focused metrics such as accuracy and F1 score. This idea has also been adopted by several following works in code summarization and generation evaluation (Wu et al., 2024b; Yang et al., 2024; Tong and Zhang, 2024). Besides, there are also domain-specific benchmarks focusing on LLMs' assessing capabilities in multimodal (Chen et al., 2024a), multilingual (Son et al., 2024b,a), instruction following (Murugadoss et al., 2024) and LLM agent (Lù et al., 2025).

## 7 Challenges & Future Works

### 7.1 Bias & Vulnerability

The use of LLMs-as-a-judge inherently introduces significant challenges related to bias and vulnerability, which significantly compromise fairness and reliability when LLMs are deployed for diverse judging tasks. Among the various types of bias, some are consistent across all LLM judges, for example, a tendency to prefer longer (Koo et al., 2023; Dubois et al., 2024; Domhan and Zhu, 2025; Yuan et al., 2024a), authoritative-looking (Stephan et al., 2024; Chen et al., 2024b) and well-formatted (Chen et al., 2024b) responses. In addition, other biases stem from individual judges' own preferences or knowledge, such as egocentric bias (Liu et al., 2023c; Wataoka et al., 2024; Panickssery et al., 2024; Chen et al., 2025c) and preference leakage (Li et al., 2025a; Goel et al., 2025; Naseh and Mireshghallah, 2025). LLM judges are also susceptible to adversarial manipulations. Techniques like prompt injection attacks (Shi et al., 2024; BENCHMARK; Banerjee et al., 2024; Tong et al., 2025) and adversarial phrases (Liusie et al., 2023; Raina et al., 2024; Doddapaneni et al., 2024b) can drastically influence LLMs' judgment, thus raising concerns about the reliability of LLM judges in high-stakes scenarios (Shi et al., 2024; Raina et al., 2024).

**Future Direction.** Existing studies have already explored approaches, such as providing more detailed evaluation principles (Zheng et al., 2023; Zhu et al., 2023; Liusie et al., 2023; Krumdick et al., 2025) and eliminating spurious features through calibration (Li et al., 2024d; Raina et al., 2024; Zhou et al., 2024b; Liu et al., 2024c; Chen et al., 2025a; Wang et al., 2025c; van den Burg et al., 2025), to mitigate LLM judges' bias. Future work could focus more on analyzing and understanding the **root causes** of these biases. For example, why would LLMs prefer their own generation (Panickssery et al., 2024)?

### 7.2 Scaling Judgment at Inference Time.

Motivated by recent inference-time scaling (ITS) studies in LLMs (Snell et al., 2024; Zhang et al., 2025b), several works have begun to explore how to scale LLMs' judgment capabilities at inference time (Saha et al., 2025; Liu et al., 2025e; Zhou et al., 2025). By expanding the reasoning process in judgment tasks and incorporating advanced behaviors such as reflection and exploration, both the accuracy and fairness (Chen et al., 2025c; Wang et al., 2025a) of judge LLMs have seen significant improvements. A straightforward approach to scaling judge LLMs is to employ Large Reasoning Models (LRMs) that generate judgments via long CoT reasoning (Chen et al., 2025b). Additionally, traditional sampling and search strategies, such as self-consistency, best-of-N, and Monte Carlo Tree Search (MCTS), have been used to more thoroughly explore the space of possible judgment trajectories (Wang et al., 2025f; Kalra and Tang, 2025). Other methods leverage golden labels as supervision, applying rule-based reinforcement learning (Chen et al., 2025b; Liu et al., 2025e; Whitehouse et al., 2025; Chen et al., 2025d; Shi and Jin, 2025), DPO (Saha et al., 2025) or distillation (Zhao et al., 2025) to train LLMs to serve as more effective judges.

**Future Directions.** While LLM-as-a-judge approaches benefit from ITS techniques, it is also important to recognize the associated challenges. These include efficiency bottlenecks (Sui et al., 2025), performance degradation from overthinking (Chen et al., 2024e), and increased vulnerability of long CoTs to adversarial attack (Jiang et al., 2025). Future research could investigate these limitations and develop mitigation strategies, paving the way for more efficient, accurate, and robust judge LLMs enhanced by ITS.

### 7.3 Dynamic & Complex Judging Strategy

Compared with earlier static and straightforward approaches that directly prompt LLMs for judgment (Zheng et al., 2023), more dynamic and complex judgment pipelines have been proposed recently to address various limitations, improving the robustness and effectiveness of LLM-as-a-judge. One approach is to follow the concept of "**LLM-as-a-examiner**", where the system dynamically and interactively generates both questions and judgments based on the candidate LLMs' performance (Yu et al., 2024c; Bai et al., 2023a; Pombal et al., 2025a; Dammu et al., 2025; Khalili and Smyth, 2025; Wang et al., 2024i; Kim et al., 2024a; Zhang et al., 2025e). Other works focus on making judgments based on multiple candidate LLMs' **battling and debating** (Moniri et al., 2024; Zhao et al., 2024c). Additionally, building complex judgment agents is another popular research area (Li et al., 2023b; Chan et al., 2023; Zhuge et al., 2024), which typically involves multi-agent collaboration with well-designed planning systems.

**Future Direction.** One promising direction for future research is to equip LLMs with human-like and agentic judgment capabilities (Yuan et al., 2024a; Liang et al., 2024b; Li et al., 2024o; Saha et al., 2024; Zhang et al., 2024b; Wang et al., 2025e; Song et al., 2025), such as anchoring, comparing, and meta-judgment. Another intriguing avenue would be to develop an **adaptive difficulty assessment system** (Hu, 2024; Patel et al., 2025), dynamically adjusting problems' difficulties based on candidates' performance.

### 7.4 Human-LLMs Co-judgement

As mentioned earlier, the biases and vulnerabilities in LLM-as-a-judge can be addressed through human-in-the-loop for further intervention and proofreading. However, only a few studies have focused on this direction (Wang et al., 2023d; Faggioli et al., 2023; Pradeep et al., 2025).

**Future Direction.** As **data selection** (Xie et al., 2023; Albalak et al., 2024) becomes an increasingly popular research area for improving the efficiency of LLMs' training and inference, it also holds the potential for enhancing LLMs-based evaluation. LLM-as-a-judge can draw insights from data selection to enable judge LLMs to serve as a critical sample selector, choosing a small subset of samples based on specific criteria (e.g., difficulty) for human annotators to conduct evaluation.

Due to the space limitation, we put the application of LLM-as-a-judge, paper collection for our taxonomy, tuning techniques and benchmark for LLM-as-a-judge in Appendix 5, D, E and F.

## 8 Conclusion

This survey explores the intricacies of LLM-as-a-judge. We begin by categorizing existing LLM-based judgment methods based on input and output formats. Then, we propose a comprehensive taxonomy for LLM-as-a-judge, encompassing judging attributes, methodologies and benchmarks. After this, a detailed and thoughtful analysis of current challenges and future directions of LLM-as-a-judge is proposed, aiming to provide more resources and insights for future works in this emerging area.

## Limitations

This work aims to provide a comprehensive survey of the LLM-as-a-judge paradigm. Due to space constraints, we focus on three core aspects in the main paper: judging attributes, methods, and benchmarks. Applications of LLM-as-a-judge and a detailed list of related papers are included in the appendix. Additionally, as discussed in Section 7.1, LLM-as-a-judge carries inherent limitations and biases. The substantial computational resources required for deploying LLMs may also pose challenges in resource-constrained scenarios.

## Acknowledgment

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-08-00. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. Lu Cheng is supported by the National Science Foundation (NSF) Grant #2312862, NSF CAREER #2440542, NSF-Simons SkAI Institute, National Institutes of Health (NIH) #R01AG091762, Google Research Scholar Award, and a Cisco gift grant.

## References

Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can we use large language models to fill relevance judgment holes? *ArXiv preprint*, abs/2405.05600.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie CY Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. i-srt: Aligning large multimodal models for videos by iterative self-retrospective judgment. *ArXiv preprint*, abs/2406.11280.

Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*, pages 135–159. Springer.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, and 1 others. 2024. A survey on data selection for language models. *ArXiv preprint*, abs/2402.16827.

Moritz Altemeyer, Steffen Eger, Johannes Daxenberger, Tim Altendorf, Philipp Cimiano, and Benjamin Schiller. 2025. Argument summarization and its evaluation in the era of large language models. *arXiv preprint arXiv:2503.00847*.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*.

Negar Arabzadeh and Charles LA Clarke. 2025a. Benchmarking llm-based relevance judgment methods. *arXiv preprint arXiv:2504.12558*.

Negar Arabzadeh and Charles LA Clarke. 2025b. A human-ai comparative analysis of prompt sensitivity in llm-based relevance judgment. *arXiv preprint arXiv:2504.12408*.

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2024. The fellowship of the llms: Multi-agent workflows for synthetic preference optimization dataset generation. *ArXiv preprint*, abs/2408.08688.

Stuart Armstrong, Matija Franklin, Connor Stevens, and Rebecca Gorman. 2025. Defense against the dark prompts: Mitigating best-of-n jailbreaking with prompt evaluation. *arXiv preprint arXiv:2502.00580*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *ArXiv preprint*, abs/2408.09235.

Sher Badshah and Hassan Sajjad. 2025. Dafe: Llm-based evaluation through dynamic arbitration for free-form question-answering. *arXiv preprint arXiv:2503.08542*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023a. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023b. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation. *arXiv preprint arXiv:2503.19092*.

Chaithanya Bandi and Abir Harrasse. 2024. Adversarial multi-agent evaluation of large language models through iterative debates. *arXiv preprint arXiv:2410.04663*.

Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The vulnerability of language model benchmarks: Do they accurately reflect true llm performance?

Jeremy Barnes, Naiara Perez, Alba Bonet-Jover, and Begoña Altuna. 2025. Summarization metrics for spanish and basque: Do automatic scores and llm-judges correlate with humans? *arXiv preprint arXiv:2503.17039*.

Alimohammad Beigi, Bohan Jiang, Dawei Li, Tharindu Kumarage, Zhen Tan, Pouya Shaeri, and Huan Liu. 2024. Lrq-fact: Llm-generated relevant questions for multimodal fact-checking. *ArXiv preprint*, abs/2410.04616.

JUDGE BENCHMARK. Jailjudge: Acomprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023. Oceangpt: A large language model for ocean science tasks. *ArXiv preprint*, abs/2310.02031.

Nathan Brake and Thomas Schaaf. 2024. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 352–363.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*.

Riccardo Cantini, Alessio Orsino, Massimo Ruggiero, and Domenico Talia. 2025. Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge. *arXiv preprint arXiv:2504.07887*.

Hongliu Cao, Ilias Driouich, Robin Singh, and Eoin Thomas. 2025a. Multi-agent llm judge: automatic personalized llm judge design for evaluating natural language generation applications. *arXiv preprint arXiv:2504.02867*.

Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Jiahua Dong, Wangbo Yu, Ge Zhang, Ian Reid, and 1 others. 2025b. Video simpleqa: Towards factuality evaluation in large video language models. *arXiv preprint arXiv:2503.18923*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. Exploring the multilingual nlg evaluation abilities of llm-based evaluators. *arXiv preprint arXiv:2503.04360*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge:

Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or llms as the judge? a study on judgement biases. *ArXiv preprint*, abs/2402.10669.

Hongyu Chen and Seraphina Goldfarb-Tarrant. 2025. Safer or luckier? llms as safety evaluators are not robust to artifacts. *arXiv preprint arXiv:2503.09347*.

Junjie Chen, Weihang Su, Zhumin Chu, Haitao Li, Qinyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024c. An automatic and cost-efficient peer-review framework for language generation evaluation. *arXiv preprint arXiv:2410.12265*.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, and 1 others. 2024d. Automated evaluation of large vision-language models on self-driving corner cases. *ArXiv preprint*, abs/2404.10595.

Meilin Chen, Jian Tian, Liang Ma, Di Xie, Weijie Chen, and Jiang Zhu. 2025a. Unbiased evaluation of large language models from a causal perspective. *arXiv preprint arXiv:2502.06655*.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025b. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*.

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. 2025c. Do llm evaluators prefer themselves for a reason? *arXiv preprint arXiv:2504.03846*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024e. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025d. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.

Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. 2024f. Llms are biased evaluators but not biased for retrieval augmented generation. *ArXiv preprint*, abs/2410.20833.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, and 1 others. 2024g. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *ArXiv preprint*, abs/2407.04842.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and 1 others. 2023. Evaluating hallucinations in chinese large language models. *ArXiv preprint*, abs/2310.03368.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot arena: A platform for code llm evaluation in the wild. *arXiv preprint arXiv:2502.09328*.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Cheng-Han Chiang, Hung-yi Lee, and Michal Lukasik. 2025. Tract: Regression-aware fine-tuning meets chain-of-thought reasoning for llm-as-a-judge. *arXiv preprint arXiv:2503.04381*.

Marianne Chuang, Gabriel Chuang, Cheryl Chuang, and John Chuang. 2025. Judging it, washing it: Scoring and greenwashing corporate climate disclosures using large language models. *arXiv preprint arXiv:2502.15094*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, and Chirag Shah. 2025. Dynamic-kgqa: A scalable framework for generating adaptive question answering datasets. *arXiv preprint arXiv:2503.05049*.

Soumik Dey, Hansi Wu, and Binbin Li. 2025. To judge or not to judge: Using llm judgements for advertiser keyphrase relevance at ebay. *arXiv preprint arXiv:2505.04209*.

Kaustubh D. Dhole, Kai Shu, and Eugene Agichtein. 2024. Conqret: Benchmarking fine-grained evaluation of retrieval augmented argumentation with llm judges.

Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Llm-evaluation tropes: Perspectives on the validity of llm-evaluations. *arXiv preprint arXiv:2504.19076*.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, and Mitesh M Khapra. 2024a. Cross-lingual auto evaluation for assessing multilingual llms. *arXiv preprint arXiv:2410.13394*.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M Khapra. 2024b. Finding blind spots in evaluator llms with interpretable checklists. *ArXiv preprint*, abs/2406.13439.

Tobias Domhan and Dawei Zhu. 2025. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. *arXiv preprint arXiv:2505.01761*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey on in-context learning. *ArXiv preprint*, abs/2301.00234.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *ArXiv preprint*, abs/2406.11657.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Francisco Eiras, Eliott Zemour, Eric Lin, and Vaikkunth Mugunthan. 2025. Know thy judge: On the robustness meta-evaluation of llm safety judges. *arXiv preprint arXiv:2503.04474*.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Bodapati, and Dan Roth. 2024. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. *arXiv preprint arXiv:2410.03775*.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and 1 others. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Zhiyuan Fan, Weinong Wang, Xing Wu, and Debing Zhang. 2025. Sedareval: Automated evaluation using self-adaptive rubrics. *arXiv preprint arXiv:2501.15595*.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *ArXiv preprint*, abs/2406.09170.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *ArXiv preprint*, abs/2309.16289.

Tao Feng, Yihang Sun, and Jiaxuan You. 2025. Grapheval: A lightweight graph-based llm framework for idea evaluation. *arXiv preprint arXiv:2503.12600*.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2024. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. *arXiv preprint arXiv:2412.20127*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual llm-as-a-judge? *arXiv preprint arXiv:2505.12201*.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *ArXiv preprint*, abs/2402.01383.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *ArXiv preprint*, abs/2304.02554.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. 2024. Justrank: Benchmarking llm judges for system ranking. *arXiv preprint arXiv:2412.09569*.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. 2025. Validating llm-as-a-judge systems in the absence of gold labels. *arXiv preprint arXiv:2503.05965*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, and 1 others. 2024. Direct language model alignment from online ai feedback. *ArXiv preprint*, abs/2402.04792.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. 2023. Allure: auditing and improving llm-based evaluation of text using iterative in-context-learning. *arXiv e-prints*, pages arXiv–2309.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Socreval: Large language models with the socratic method for reference-free reasoning evaluation. *arXiv preprint arXiv:2310.00074*.

Junda He, Jieke Shi, Terry Yue Zhuo, Christoph Treude, Jiamou Sun, Zhenchang Xing, Xiaoning Du, and David Lo. 2025. From code to courtroom: Llms as the new software judges. *arXiv preprint arXiv:2503.02246*.

Amey Hengle, Aswini Kumar, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. Cseval: Towards automated, multi-dimensional, and reference-free counterspeech evaluation using auto-calibrated llms. *arXiv preprint arXiv:2501.17581*.

Xanh Ho, Jiahao Huang, Florian Boudin, and Akiko Aizawa. 2025. Llm-as-a-judge: Reassessing the performance of llms in extractive qa. *arXiv preprint arXiv:2504.11972*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381.

Aliyah R Hsu, James Zhu, Zhichao Wang, Bin Bi, Shubham Mehrotra, Shiva K Pentyala, Katherine Tan, Xiang-Bo Mao, Roshanak Omrani, Sougata Chaudhuri, and 1 others. 2024. Rate, explain and cite (rec): Enhanced explanation and attribution in automatic evaluation by large language models. *arXiv preprint arXiv:2411.02448*.

Aaron Hu. 2024. Developing an ai-based psychometric system for assessing learning difficulties and adaptive system to overcome: A qualitative and conceptual framework. *ArXiv preprint*, abs/2403.06284.

Renjun Hu, Yi Cheng, Libin Meng, Jiaxin Xia, Yi Zong, Xing Shi, and Wei Lin. 2025a. Training an llm-as-a-judge model: Pipeline, insights, and practical lessons. *arXiv preprint arXiv:2502.02988*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024a. Are llm-based evaluators confusing nlg quality criteria? *arXiv preprint arXiv:2402.12055*.

Xinyu Hu, Mingqi Gao, Li Lin, Zhenghan Yu, and Xiaojun Wan. 2025b. A dual-perspective nlg meta-evaluation framework with automatic benchmark and better interpretability. *arXiv preprint arXiv:2502.12052*.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024b. Themis: A reference-free nlg evaluation language model with flexibility and interpretability. *ArXiv preprint*, abs/2406.18365.

Hui Huang, Yingqi Qu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024a. On the limitations of fine-tuned judge models for llm evaluation. *arXiv preprint arXiv:2403.02839*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024b. Lost in the source language: How large language models evaluate the quality of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Yue Huang, Qihui Zhang, Lichao Sun, and 1 others. 2023b. Trustgpt: A benchmark for trustworthy and responsible large language models. *ArXiv preprint*, abs/2306.11507.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv preprint*, abs/2312.06674.

Andrés Isaza-Giraldo, Paulo Bala, Pedro F Campos, and Lucas Pereira. 2024. Prompt-gaming: A pilot study on llm-evaluating agent in a meaningful energy game. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023a. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023b. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. *arXiv preprint arXiv:2501.09993*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2820–2837.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment. *arXiv preprint arXiv:2412.13746*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement. *ArXiv preprint*, abs/2407.18370.

Nimit Kalra and Leonard Tang. 2025. Verdict: A library for scaling judge-time compute. *arXiv preprint arXiv:2502.18018*.

Ivan Kartáč, Mateusz Lango, and Ondřej Dušek. 2025. Openlgauge: An explainable metric for nlg evaluation with open-weights llms. *arXiv preprint arXiv:2503.11858*.

Akira Kawabata and Saku Sugawara. 2024. Rationale-aware answer verification by pairwise self-evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16178–16196.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, and 1 others. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.

Zachary Kenton, Noah Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, and 1 others. 2024. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276.

Boshra Khalili and Andrew W Smyth. 2025. Autodrive-qa-automated generation of multiple-choice questions for autonomous driving datasets using large vision-language models. *arXiv preprint arXiv:2503.15778*.

Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, and Alice Oh. 2024a. Llm-as-an-interviewer: Beyond static testing through dynamic llm evaluation. *arXiv preprint arXiv:2412.10424*.

Heegyu Kim, Taeyang Jeon, Seungtaek Choi, Ji Hoon Hong, Dong Won Jeon, Ga-Yeon Baek, Gyeong-Won Kwak, Dong-Hee Lee, Jisu Bae, Chihoon Lee, and 1 others. 2025. Towards fully-automated materials discovery via large-scale synthesis dataset and expert-level llm-as-a-judge. *arXiv preprint arXiv:2502.16457*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *ArXiv preprint*, abs/2405.01535.

Chhavi Kirtani, Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Murari Mandal, and Dhruv Kumar. 2025. Revieweval: An evaluation framework for ai-generated reviews. *arXiv preprint arXiv:2502.11736*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *ArXiv preprint*, abs/2309.17012.

Neema Kotonya, Saran Krishnasamy, Joel Tetreault, and Alejandro Jaimes. 2023. Little giants: Exploring the potential of small LLMs as evaluation metrics in summarization in the Eval4NLP 2023 shared task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 202–218, Bali, Indonesia. Association for Computational Linguistics.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*.

Abhishek Kumar, Sonia Haiduc, Partha Pratim Das, and Partha Pratim Chakrabarti. 2024a. Llms as evaluators: A novel approach to evaluate bug report summarization. *arXiv preprint arXiv:2409.00630*.

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024b. Decoding biases: Automated methods and llm judges for gender bias detection in language models. *ArXiv preprint*, abs/2408.03907.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-Ling Mao. Criticeval: Evaluating large-scale language model as critic. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*, abs/2309.00267.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024a. Aligning large language models by on-policy self-judgment. *ArXiv preprint*, abs/2402.11253.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024b. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *arXiv preprint arXiv:2406.06004*.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025a. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Dawei Li, Zhen Tan, and Huan Liu. 2024a. Exploring large language models for feature selection: A data-centric perspective. *ArXiv preprint*, abs/2408.12025.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu, and Jiayi Shen. 2024b. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. *ArXiv preprint*, abs/2411.03284.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sunkwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, and 1 others. 2024c. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *ArXiv preprint*, abs/2405.04819.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2024d. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. *ArXiv preprint*, abs/2410.15393.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024e. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2024f. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024g. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *ArXiv preprint*, abs/2402.05044.

Mingxuan Li, Hanchen Li, and Chenhao Tan. 2025b. Hypoeval: Hypothesis-guided evaluation for natural language generation. *arXiv preprint arXiv:2504.07174*.

Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy F Chen, and Min-Yen Kan. 2024h. Decompose and aggregate: A step-by-step interpretable evaluation framework. *arXiv preprint arXiv:2405.15329*.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023a. Exploring the reliability of large language models as customized evaluators for diverse nlp tasks. *arXiv preprint arXiv:2310.19740*.

Renhao Li, Minghuan Tan, Derek F Wong, and Min Yang. 2024i. Coevol: Constructing better responses for instruction finetuning through multi-agent cooperation. *ArXiv preprint*, abs/2406.07054.

Ruosen Li, Ruochen Li, Barry Wang, and Xinya Du. 2024j. Iqa-eval: Automatic evaluation of human-model interactive question answering. *Advances in Neural Information Processing Systems*, 37:109894–109921.

Ruosen Li, Teerth Patel, and Xinya Du. 2023b. Prd: Peer rank and discussion improve large language model based evaluations. *ArXiv preprint*, abs/2307.02762.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024k. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *ArXiv preprint*, abs/2406.11939.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. 2024l. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*.

Xiaonan Li and Xipeng Qiu. 2023a. MoT: Memory-of-thought enables ChatGPT to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, Singapore. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023b. Mot: Memory-of-thought enables chatgpt to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374.

Yu Li, Shenyu Zhang, Rui Wu, Xiutian Huang, Yongrui Chen, Wenhao Xu, Guilin Qi, and Dehai Min. 2024m. Mateval: A multi-agent discussion framework for advancing open-ended text evaluation. In *International Conference on Database Systems for Advanced Applications*, pages 415–426. Springer.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations*.

Yuran Li, Jama Hussein Mohamud, Chongren Sun, Di Wu, and Benoit Boulet. 2025c. Leveraging llms as meta-judges: A multi-agent framework for evaluating llm judgments. *arXiv preprint arXiv:2504.17087*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024n. Leveraging large language models for nlg evaluation: Advances and challenges.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024o. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. 2024a. Debatrix: Multi-dimensinal debate judge with iterative chronological analysis based on llm. *arXiv preprint arXiv:2403.08010*.

Sirui Liang, Baoli Zhang, Jun Zhao, and Kang Liu. 2024b. Abseval: An agent-based framework for script evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12418–12434.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *ArXiv preprint*, abs/2305.19118.

Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, and 1 others. 2024c. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. *ArXiv preprint*, abs/2408.08072.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023a. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023b. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Beiming Liu, Zhizhuo Cui, Siteng Hu, Xiaohua Li, Haifeng Lin, and Zhengxin Zhang. 2025a. Llm evaluation based on aerospace manufacturing expertise: Automated generation and multi-model question answering. *arXiv preprint arXiv:2501.17183*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8560–8579, Mexico City, Mexico. Association for Computational Linguistics.

Rundong Liu, Andre Frade, Amal Vaidya, Maxime Labonne, Marcus Kaiser, Bismayan Chakrabarti, Jonathan Budd, and Sean Moran. 2025b. On iterative evaluation and enhancement of code quality using gpt-4o. *arXiv preprint arXiv:2502.07399*.

Shuliang Liu, Xinze Li, Zhenghao Liu, Yukun Yan, Cheng Yang, Zheni Zeng, Zhiyuan Liu, Maosong Sun, and Ge Yu. 2025c. Judge as a judge: Improving the evaluation of retrieval-augmented generation through the judge-consistency of large language models. *arXiv preprint arXiv:2502.18817*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023a. Statistical rejection sampling improves preference optimization. *ArXiv preprint*, abs/2309.06657.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025d. Pairwise rm: Perform best-of-n sampling with knockout tournament. *arXiv preprint arXiv:2501.13007*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023c. Llms as narcissistic evaluators: When ego inflates evaluation scores. *ArXiv preprint*, abs/2311.09766.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024c. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024d. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. *arXiv preprint arXiv:2402.15754*.

Zijun Liu, Boqun Kou, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024e. Meta ranking: Less capable language models are capable for single response judgement. *ArXiv preprint*, abs/2402.12146.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025e. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023. Zero-shot nlg evaluation through pairware comparisons with llms. *ArXiv preprint*, abs/2307.07889.

Edoardo Loru, Jacopo Nudo, Niccolò Di Marco, Matteo Cinelli, and Walter Quattrociocchi. 2025. Decoding ai judgment: How llms assess news credibility and bias. *arXiv preprint arXiv:2502.04426*.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J Pal, and Siva Reddy. 2025. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. *arXiv preprint arXiv:2504.08942*.

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Heyan Huang, Chen Xu, and Xiaoyan Gao. 2024a. Beyond exact match: Semantically reassessing event extraction by large language models. *arXiv preprint arXiv:2410.09418*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024b. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024a. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *ArXiv preprint*, abs/2406.07070.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. 2024b. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. *arXiv preprint arXiv:2411.13281*.

Shengjie Ma, Chong Chen, Qi Chu, and Jiaxin Mao. 2024. Leveraging large language models for relevance judgments in legal case retrieval. *ArXiv preprint*, abs/2403.18405.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *ArXiv preprint*, abs/2305.02156.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *ArXiv preprint*, abs/2404.01869.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. 2024. Evaluating the performance of large language models via debates. *ArXiv preprint*, abs/2406.11044.

Wenhan Mu, Ling Xu, Shuren Pei, Le Mi, and Huichi Zhou. 2025. Evaluate-and-purify: Fortifying code language models against adversarial attacks using llm-as-a-judge. *arXiv preprint arXiv:2504.19730*.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. *ArXiv preprint*, abs/2408.08781.

Mirco Musolesi. 2024. Creative beam search: Llm-as-a-judge for improving response generation. ICCC.

Linyong Nan, Ellen Zhang, Weijin Zou, Yilun Zhao, Wenfei Zhou, and Arman Cohan. 2024. On evaluating the integration of reasoning and action in LLM agents with database question answering. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4556–4579, Mexico City, Mexico. Association for Computational Linguistics.

Ali Naseh and Niloofar Mireshghallah. 2025. Synthetic data can mislead evaluations: Membership inference as machine text detection. *arXiv preprint arXiv:2501.11786*.

Kun-Peng Ning, Shuo Yang, Yuyang Liu, Jia-Yu Yao, Zhenhui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Pico: Peer review in llms based on the consistency optimization.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias in large language models. *arXiv preprint arXiv:2402.15987*.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Shao-yen Tseng, and Vasudev Lal. 2024. Steering large language models to evaluate and amplify creativity. *arXiv preprint arXiv:2412.06060*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *ArXiv preprint*, abs/2406.18665.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *ArXiv preprint*, abs/2401.12086.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *ArXiv preprint*, abs/2407.06551.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Arkil Patel, Siva Reddy, and Dzmitry Bahdanau. 2025. How to get your llm to generate challenging problems for evaluation. *arXiv preprint arXiv:2502.14678*.

Bhrij Patel, Souradip Chakraborty, Wesley A Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. Aime: Ai system optimization via multiple llm evaluators. *arXiv preprint arXiv:2410.03131*.

John Penfever and 1 others. 2024. Style over substance: Failure modes of llm judges in alignment benchmarking. *ArXiv preprint*, abs/2410.17578.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2023. Llm self defense: By self examination, llms know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024*.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An llm-as-a-judge approach for scalable gender-neutral translation evaluation. *arXiv preprint arXiv:2504.11934*.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025a. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models. *arXiv preprint arXiv:2504.01001*.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André FT Martins. 2025b. M-prometheus: A suite of open multilingual llm judges. *arXiv preprint arXiv:2504.04953*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. *arXiv preprint arXiv:2504.15068*.

Archiki Prasad, Elias Stengel-Eskin, Justin Chih-Yao Chen, Zaid Khan, and Mohit Bansal. 2025. Learning to generate unit tests for automated debugging. *arXiv preprint arXiv:2502.01619*.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.

Siya Qi, Rui Cao, Yulan He, and Zheng Yuan. 2025. Evaluating llms' assessment of mixed-context hallucination through the lens of summarization. *arXiv preprint arXiv:2503.01670*.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orašan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Huaizhi Qu, Inyoung Choi, Zhen Tan, Song Wang, Sukwon Yun, Qi Long, Faizan Siddiqui, Kwonjoon Lee, and Tianlong Chen. 2025. Efficient map estimation of llm judgment performance with prior transfer. *arXiv preprint arXiv:2504.12589*.

Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. *ArXiv preprint*, abs/2406.14783.

Melissa Kazemi Rad, Huy Nghiem, Andy Luo, Sahil Wadhwa, Mohammad Sorower, and Stephen Rawls. 2025. Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment. *arXiv preprint arXiv:2501.13080*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hossein A Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2024. Judgeblender: Ensembling judgments for automatic relevance assessment. *arXiv preprint arXiv:2412.13268*.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *ArXiv preprint*, abs/2402.14016.

Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakkar. 2024. Constructing domain-specific evaluation sets for llm-as-a-judge. *ArXiv preprint*, abs/2408.08808.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

David Rodriguez, William Seymour, Jose M Del Alamo, and Jose Such. 2025. Towards safer chatbots: A framework for policy compliance evaluation of custom gpts. *arXiv preprint arXiv:2502.01436*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024a. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024b. Lmunit: Fine-grained evaluation with natural language unit tests. *arXiv preprint arXiv:2412.13091*.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8345–8363.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

David Salinas, Omar Swelam, and Frank Hutter. 2025. Tuning llm judge design decisions for 1/1000 of the cost. *arXiv e-prints*, pages arXiv–2501.

Piotr Sawicki, Marek Grześ, Dan Brown, and Fabrício Góes. 2025. Can large language models outperform non-experts in poetry evaluation? a comparative study using the consensual assessment technique. *arXiv preprint arXiv:2502.19064*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Ratan J Sebastian and Anett Hoppe. 2025. Validating llm-generated relevance labels for educational resource search. *arXiv preprint arXiv:2504.12732*.

Saptarshi Sengupta, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification.

Kwangwook Seo, Donguk Kwon, and Dongha Lee. 2025. Mt-raig: Novel benchmark and evaluation framework for retrieval-augmented insight generation over multiple tables. *arXiv preprint arXiv:2502.11735*.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. 2023. Languagempc: Large language models as decision makers for autonomous driving. *ArXiv preprint*, abs/2310.03026.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2024. Optimization-based prompt injection attack to llm-as-a-judge. *ArXiv preprint*, abs/2403.17710.

Wenlei Shi and Xing Jin. 2025. Heimdall: test-time scaling on the generative verification. *arXiv preprint arXiv:2504.10337*.

Takumi Shibata and Yuichi Miyamura. 2025. Lces: Zero-shot automated essay scoring via pairwise comparisons using large language models. *arXiv preprint arXiv:2505.08498*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024a. Llm-as-a-judge & reward model: What they can and cannot do. *ArXiv preprint*, abs/2409.11239.

Guijin Son and 1 others. 2024b. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *ArXiv preprint*, abs/2410.17578.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024a. Finesure: Fine-grained summarization evaluation using llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922.

Mingyang Song, Mao Zheng, and Xuan Luo. 2024b. Can many-shot in-context learning help long-context llm judges? see more, judge better! *ArXiv preprint*, abs/2406.11629.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025. Grp: Goal-reversed prompting for zero-shot evaluation with llms. *arXiv preprint arXiv:2503.06139*.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2024. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. *ArXiv preprint*, abs/2409.04168.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuan-Jing Huang. 2022. Bertscore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2024. Salmon: Self-alignment with instructable reward models. In *The Twelfth International Conference on Learning Representations*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and 1 others. 2024a. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042*.

Sijun Tan and 1 others. 2024b. Judgebench: A benchmark for evaluating llm-based judges. *ArXiv preprint*, abs/2410.12784.

Qiaoyu Tang, Jiawei Chen, Bowen Yu, Yaojie Lu, Cheng Fu, Haiyang Yu, Hongyu Lin, Fei Huang, Ben He, Xianpei Han, and 1 others. 2024a. Self-retrieval: Building an information retrieval system with one large language model. *ArXiv preprint*, abs/2403.00801.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024b. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *ArXiv preprint*, abs/2406.12624.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences, 2023. *ArXiv preprint*, abs/2309.10621.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Terry Tong, Fei Wang, Zhe Zhao, and Muhao Chen. 2025. Badjudge: Backdoor vulnerabilities of llm-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20032–20051.

Yongqi Tong, Sizhe Wang, Dawei Li, Yifan Wang, Simeng Han, Zi Lin, Chengsong Huang, Jiaxin Huang, and Jingbo Shang. 2024. Optimizing language model's reasoning abilities with weak supervision. *ArXiv preprint*, abs/2405.04086.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*.

En-Qi Tseng, Pei-Cing Huang, Chan Hsu, Peng-Yi Wu, Chan-Tung Ku, and Yihuang Kang. 2024. Codev: An automated grading framework leveraging large language models for consistent and constructive feedback. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5442–5449. IEEE.

Gerrit JJ van den Burg, Gen Suzuki, Wei Liu, and Murat Sensoy. 2025. Aligning black-box language models with human judgments. *arXiv preprint arXiv:2502.04997*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *ArXiv preprint*, abs/2407.10817.

Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. 2024a. Halu-j: Critique-based hallucination judge. *ArXiv preprint*, abs/2407.12943.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Chengrui Wang, Qingqing Long, Xiao Meng, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024b. Biorag: A rag-llm framework for biological question reasoning. *ArXiv preprint*, abs/2408.01107.

Chihang Wang, Yuxin Dong, Zhenhong Zhang, Ruotong Wang, Shuo Wang, and Jiajing Chen. 2024c. Automated genre-aware article scoring and feedback using large language models. *arXiv preprint arXiv:2410.14165*.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2023b. Learning personalized story evaluation. *ArXiv preprint*, abs/2310.03304.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023c. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024d. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2024e. Direct judgement preference optimization. *ArXiv preprint*, abs/2409.14664.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023d. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. 2025a. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025b. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *arXiv preprint arXiv:2502.06193*.

Sizhe Wang, Yongqi Tong, Hengyuan Zhang, Dawei Li, Xin Zhang, and Tianlong Chen. 2024f. Bpo: Towards balanced preference optimization between knowledge breadth and depth in alignment. *arXiv preprint arXiv:2411.10914*.

Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2024g. Ceb: Compositional evaluation benchmark for fairness in large language models. *ArXiv preprint*, abs/2407.02408.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024h. Self-taught evaluators. *ArXiv preprint*, abs/2408.02666.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. 2025c. Improving llm-as-a-judge inference with the judgment distribution. *arXiv preprint arXiv:2503.03064*.

Wanying Wang, Zeyu Ma, Pengfei Liu, and Mingang Chen. 2024i. Revisiting benchmark and assessment: An agent-based exploratory dynamic evaluation framework for llms. *arXiv preprint arXiv:2410.11507*.

Xiao Wang, Daniil Larionov, Siwei Wu, Yiqi Liu, Steffen Eger, Nafise Sadat Moosavi, and Chenghua Lin. 2025d. Contrastscore: Towards higher quality, less biased, more efficient evaluation metrics with contrastive evaluation. *arXiv preprint arXiv:2504.02106*.

Xinchen Wang, Pengfei Gao, Chao Peng, Ruida Hu, and Cuiyun Gao. 2025e. Codevisionary: An agent-based framework for evaluating large language models in code generation. *arXiv preprint arXiv:2504.13472*.

Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, Mark Cusick, Param Kulkarni, Zhengping Ji, Yasser Ibrahim, and Xia Hu. 2024j. Dhp benchmark: Are llms good nlg evaluators? *ArXiv preprint*, abs/2408.13704.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024k. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Yutong Wang, Pengliang Ji, Chaoqun Yang, Kaixin Li, Ming Hu, Jiaoyang Li, and Guillaume Sartoretti. 2025f. Mcts-judge: Test-time scaling in llm-as-a-judge for code correctness evaluation. *arXiv preprint arXiv:2502.12468*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024l. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Ziyu Wang, Hao Li, Di Huang, and Amir M Rahmani. 2024m. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *ArXiv preprint*, abs/2409.19487.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *ArXiv preprint*, abs/2410.21819.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024a. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *ArXiv preprint*, abs/2408.13006.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and 1 others. 2024b. Long-form factuality in large language models. *ArXiv preprint*, abs/2403.18802.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. Rocketeval: Efficient automated llm evaluation via grading checklist. *arXiv preprint arXiv:2503.05142*.

Bosi Wen, Pei Ke, Yufei Sun, Cunxiang Wang, Xiaotao Gu, Jinfeng Zhou, Jie Tang, Hongning Wang, and Minlie Huang. 2025. Hpss: Heuristic prompting strategy search for llm evaluators. *arXiv preprint arXiv:2502.13031*.

Xueru Wen, Xinyu Lu, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2024. On-policy fine-grained knowledge feedback for hallucination mitigation. *arXiv preprint arXiv:2406.12221*.

Martin Weyssow, Aton Kamanda, and Houari Sahraoui. 2024. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *arXiv preprint arXiv:2403.09032*.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. 2025. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *Preprint*, arXiv:2505.10320.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 695–707. Springer.

Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. 2025. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm. *arXiv preprint arXiv:2502.19103*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024a. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *ArXiv preprint*, abs/2407.19594.

Yang Wu, Yao Wan, Zhaoyang Chu, Wenting Zhao, Ye Liu, Hongyu Zhang, Xuanhua Shi, and Philip S. Yu. 2024b. Can large language models serve as evaluators for code summarization?

Yang Wu, Yao Wan, Zhaoyang Chu, Wenting Zhao, Ye Liu, Hongyu Zhang, Xuanhua Shi, and Philip S Yu. 2024c. Can large language models serve as evaluators for code summarization? *arXiv preprint arXiv:2412.01333*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2024. Evaluating mathematical reasoning beyond accuracy. *ArXiv preprint*, abs/2404.05692.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. 2025a. An empirical analysis of uncertainty in large language model evaluations. *arXiv preprint arXiv:2502.10709*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, and 1 others. 2024a. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *ArXiv preprint*, abs/2406.14598.

Wenwen Xie, Gray Gwizdz, and Dongji Feng. 2025b. Prompting a weighting mechanism into llm-as-a-judge in two-step: A case study. *arXiv preprint arXiv:2502.13396*.

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rose, and 1 others. 2024b. Improving model factuality with fine-grained critique-based evaluator. *arXiv preprint arXiv:2410.18359*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2024c. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *ArXiv preprint*, abs/2410.02712.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025a. Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings. *arXiv preprint arXiv:2503.15620*.

Kaishuai Xu, Tiezheng Yu, Wenjun Hou, Yi Cheng, Liangyou Li, Xin Jiang, Lifeng Shang, Qun Liu, and Wenjie Li. 2025b. Learning to align multi-faceted evaluation: A unified and robust framework. *arXiv preprint arXiv:2502.18874*.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024a. Academically intelligent llms are not necessarily socially intelligent. *ArXiv preprint*, abs/2403.06591.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. 2024b. Benchmarking llms' judgments with no gold standard.

Shuying Xu, Junjie Hu, and Ming Jiang. 2024c. Large language models are active critics in nlg evaluation. *arXiv preprint arXiv:2410.10724*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024d. Sayself: Teaching llms to express confidence with self-reflective rationales. *ArXiv preprint*, abs/2405.20974.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023a. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023b. Towards reasoning in large language models via multiagent peer review collaboration. *arXiv preprint arXiv:2311.08152*.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Autogpt for online decision making: Benchmarks and additional opinions. *ArXiv preprint*, abs/2306.02224.

Jheng-Hong Yang and Jimmy Lin. 2024. Toward automatic relevance judgment using vision–language models for image–text retrieval evaluation. *ArXiv preprint*, abs/2408.01363.

Jian Yang, Jiaxi Yang, Ke Jin, Yibo Miao, Lei Zhang, Liqun Yang, Zeyu Cui, Yichang Zhang, Binyuan Hui, and Junyang Lin. 2024. Evaluating and aligning codellms on human preference.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Michihiro Yasunaga, Leonid Shamis, Chunting Zhou, Andrew Cohen, Jason Weston, Luke Zettlemoyer, and Marjan Ghazvininejad. 2024. Alma: Alignment with minimal annotation. *arXiv preprint arXiv:2412.04305*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024a. Justice or prejudice? quantifying biases in llm-as-a-judge. *ArXiv preprint*, abs/2410.02736.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024b. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Zihuiwen Ye, Luckeciano Carvalho Melo, Younesse Kaddar, Phil Blunsom, Sam Staton, and Yarin Gal. 2025. Uncertainty-aware step-wise verification with generative reward models. *arXiv preprint arXiv:2502.11250*.

Seungjun Yi, Jaeyoung Lim, and Juyong Yoon. 2024. Protocollm: Automatic evaluation framework of llms on domain-specific scientific protocol formulation tasks. *arXiv preprint arXiv:2410.04601*.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025. Improve llm-as-a-judge ability as a general ability. *arXiv preprint arXiv:2502.11689*.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. 2024a. xfinder: Robust and pinpoint answer extraction for large language models. *arXiv preprint arXiv:2405.11874*.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, and 1 others. 2024b. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024c. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *ArXiv preprint*, abs/2402.15043.

Zhuohao Yu, Weizheng Gu, Yidong Wang, Xingru Jiang, Zhengran Zeng, Jindong Wang, Wei Ye, and Shikun Zhang. Reasoning through execution: Unifying process and outcome rewards for code generation.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, and 1 others. Advancing llm reasoning generalists with preference trees. In *AI for Math Workshop@ ICML 2024*.

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, and Kan Li. 2024a. Batcheval: Towards human-like text evaluation. *ArXiv preprint*, abs/2401.00437.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, and 1 others. 2024b. R-judge: Benchmarking safety risk awareness for llm agents. *ArXiv preprint*, abs/2401.10019.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024c. Self-rewarding language models. *ArXiv preprint*, abs/2401.10020.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Yuwei Zeng, Yao Mu, and Lin Shao. 2024. Learning reward for robot skills using large language models via self-alignment. *ArXiv preprint*, abs/2405.07162.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Yuanzhao Zhai, Zhuo Zhang, Kele Xu, Hanyang Peng, Yue Yu, Dawei Feng, Cheng Yang, Bo Ding, and Huaimin Wang. 2024. Online self-preferring language models. *ArXiv preprint*, abs/2405.14103.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, and 1 others. 2025a. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19515–19524. AAAI Press.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. 2024b. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024c. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1941–1951.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024d. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Mingqing Zhang, Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024e. Breaking event rumor detection via stance-separated multi-agent debate.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025b. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*.

Qiyuan Zhang, Yufei Wang, Tiezheng Yu, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and 1 others. 2024f. Reviseval: Improving llm-as-a-judge via response-adapted references. *ArXiv preprint*, abs/2410.05193.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025c. Process-based self-rewarding language models. *arXiv preprint arXiv:2503.03746*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025d. Persona-judge: Personalized alignment of large language models via token-level self-judgment. *arXiv preprint arXiv:2504.12663*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024g. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *ArXiv preprint*, abs/2402.09267.

Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024h. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 33–42.

Xiechi Zhang, Shunfan Zheng, Linlin Wang, Gerard De Melo, Zhu Cao, Xiaoling Wang, and Liang He. 2024i. Ace-m3: Automatic capability evaluator for multimodal medical models. *arXiv preprint arXiv:2412.11453*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *ArXiv preprint*, abs/2308.01862.

Yueheng Zhang, Xiaoyuan Liu, Yiyou Sun, Atheer Alharbi, Hend Alzahrani, Basel Alomair, and Dawn Song. 2025e. Can llms design good questions based on context? *arXiv preprint arXiv:2501.03491*.

Fuheng Zhao, Lawrence Lim, Ishtiyaque Ahmad, Divyakant Agrawal, and Amr El Abbadi. 2023a. Llm-sql-solver: Can llms determine sql equivalence? *arXiv preprint arXiv:2312.10321*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and 1 others. 2025. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

John Zhao and 1 others. 2024a. Codejudge-eval: A benchmark for evaluating code generation. *ArXiv preprint*, abs/2401.10019.

Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. 2024b. Diffagent: Fast and accurate text-to-image api selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6390–6399.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Deli Zhao, and Lidong Bing. 2024c. Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions. *ArXiv preprint*, abs/2405.20267.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023b. Slic-hf: Sequence likelihood calibration with human feedback. *ArXiv preprint*, abs/2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunfan Zheng, Xiechi Zhang, Gerard de Melo, Xiaoling Wang, and Linlin Wang. 2025. Hierarchical divide-and-conquer for fine-grained alignment in llm-based medical evaluation. *arXiv preprint arXiv:2501.06741*.

Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. 2024a. Fairer preferences elicit improved human-aligned large language model judgments. *arXiv preprint arXiv:2406.11370*.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. 2024b. Mitigating the bias of large language model evaluation. *ArXiv preprint*, abs/2409.16788.

Lexin Zhou, Youmna Farag, and Andreas Vlachos. 2024c. An llm feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5389–5409.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024d. Self-discover: Large language models self-compose reasoning structures. *ArXiv preprint*, abs/2402.03620.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024e. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *ArXiv preprint*, abs/2310.11667.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. 2025. Evaluating judges as evaluators: The jetts benchmark of llm-as-judges as test-time scaling evaluators. *arXiv preprint arXiv:2504.15253*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024a. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. 2024b. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024c. Dynamic evaluation of large language models by meta probing agents. In *Forty-first International Conference on Machine Learning*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *ArXiv preprint*, abs/2310.17631.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024b. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, and 1 others. 2024. Agent-as-a-judge: Evaluate agents with agents. *ArXiv preprint*, abs/2410.10934.

Terry Yue Zhuo. 2024. Ice-score: Instructing large language models to evaluate code. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2232–2242.

# A Attribute Definition

We provide a detailed definition for each judgment attribute in Table 1.

# B Prompting Methods Categories

Based on each prompting strategy's target, we categorize them into following four group: (1) bias reduction, which involves reducing bias caused by candidate output position or reliance on a single LLM judge (swapping operations, multi-agent collaboration); (2) boosting instruction-following, which helps the LLM judge learn clear judging criteria and principles from rules or demonstrations (rule augmentation, in-context demonstration); (3) enhancing evaluation depth, which enables a better understanding of model capabilities (multi-turn interaction); and (4) improving evaluation efficiency, which refers to reducing the computational budget required during judgment (comparison acceleration).

# C Application with More Details

## C.1 Evaluation

LLM-as-a-judge is first proposed for evaluation. It enables human-like evaluations rather than overlap-based matching (Post, 2018; Lin and Chen, 2023b). We discuss how LLM-as-a-judge has been utilized to evaluate open-ended generation (Section C.1.1), reasoning (Section C.1.2), and emerging NLP tasks (Section C.1.3).

### C.1.1 Open-ended Generation Tasks

Open-ended generation includes tasks like dialog response, text summarization, and creative writing, where outputs must be safe, accurate, and contextually relevant with multiple "correct" answers (Badshah and Sajjad, 2024; Kumar et al., 2024b; Zeng et al.; Song et al., 2024a; Jones et al., 2024). Unlike traditional metrics, LLM-as-a-judge enables nuanced and adaptable evaluation (Zheng et al., 2023). This approach has been used for single-model evaluations and competitive comparisons (Gao et al., 2023; Wu et al., 2023). While LLMs-as-judges demonstrate human-like judgments, longer outputs risk hallucinations (Wang et al., 2024a; Cheng et al., 2023). Another concern is biased and unsafe judgements (Yu et al., 2024a; Li et al., 2024g; Ye et al., 2024a), though excessive caution may cause overly refusal (Xie et al., 2024a). To address these, researchers have proposed conversational frameworks like self-reflection (Ji et al., 2023) and debating (Moniri et al., 2024). Besides, multilingual LLM-as-a-judge research has advanced with various methods and benchmarks that address cross-lingual evaluation challenges. Approaches include scoring non-English answers against English references (Doddapaneni et al., 2024a), using multi-agent debate frameworks for fine-grained evaluation (Feng et al., 2024), and developing open-source multilingual judges that outperform English-centric evaluators across 20+ languages (Pombal et al., 2025b). Benchmarks like MM-Eval and PARIKSHA test the consistency and fairness of multilingual LLM judges, showing that evaluators tuned in English often underperform on low-resource languages (Son et al., 2024b).

However, key challenges still remain in LLM-based multilingual judgment. Studies highlight cross-lingual inconsistency, where judges show low agreement across languages, especially for low-resource settings (Fu and Liu, 2025). Evaluators may also suffer from factual errors, cultural misrepresentations, and toxic content (Hada et al., 2024). Additionally, dialectal variation further complicates the bias, with weaker alignment between LLM and human toxicity ratings in regional varieties [8]. These issues underscore the need for more culturally sensitive and robust multilingual evaluation methods.

### C.1.2 Reasoning Tasks

The reasoning abilities of LLMs can be assessed through their intermediate thinking processes and final answers (He et al., 2023; Parmar et al., 2024; Mondorf and Plank, 2024). For mathematical reasoning, Xia et al. (2024) introduce a framework using judge LLMs to assess the quality of reasoning steps. Similarly, for temporal reasoning, Fatemi et al. (2024) create synthetic datasets to evaluate models' ability to reason about event sequences, causality, and dependencies. To distinguish genuine reasoning ability from pattern memorization, Wang et al. (2023a) propose a human-in-the-loop framework where LLMs and users adopt opposing positions to reach correct decisions. Nan et al. (2024) develop a multi-agent framework simulating peer review, leveraging LLMs-as-judges to collaboratively assess reasoning capabilities in data-driven tasks.

### C.1.3 Emerging Tasks

LLM-as-a-judge is also applied to tasks once exclusive to humans, particularly in context-specific

| Attribute | Definition |
|---|---|
| Helpfulness | Helpfulness is a critical criterion to measure the utility and informativeness of a generated response. |
| Safety & Security | Safety & security refer to whether the model avoids generating and is not affected by harmful, toxic, biased, or adversarial content. |
| Reliability | Reliability is the degree to which a response is faithful to verifiable sources and appropriately calibrated in expressing uncertainty. |
| Relevance | Relevance is a metric to measure how well a response aligns with the user query, topic, or task context. |
| Logic | Logic refers to the internal coherence and correctness of reasoning steps within a response, independent of factual accuracy. |
| Overall Quality | Overall quality is a holistic assessment of a response's merit, typically integrating multiple dimensions into one comprehensive score. |

Table 1: Common judgment attributes and their definitions.

areas. A prominent task is in social intelligence, where models are presented with complex social scenarios requiring the understanding of cultural values, ethical principles, and potential social impacts (Xu et al., 2024a; Zhou et al., 2023). Research has also extended to evaluating Large Multimodal Models (LMMs) and Large Vision-Language Models (LVLMs) (Zhu et al., 2024b). For example, Xiong et al. (2024) use LMM-as-a-judge to provide transparent evaluations with rationales, while Chen et al. (2024d) propose a benchmark for LVLMs in self-driving scenarios, showing that LLM-based evaluations align better with human preferences than LVLM-based ones. Recently, we have seen more customized utilization of LLM-as-a-judge to evaluate emerging tasks such as code understanding and generation (Zhao et al., 2024a; Zhuo, 2024; Tseng et al., 2024; Wu et al., 2024c; He et al., 2025; Yu et al.; Wang et al., 2025b; Prasad et al., 2025; Liu et al., 2025b; Chi et al., 2025), legal knowledge (Fei et al., 2023), game development (Isaza-Giraldo et al., 2024), nature science (Bi et al., 2023; Chuang et al., 2025; Kim et al., 2025), manufacture engineering (Liu et al., 2025a), healthcare conversations (Wang et al., 2024m; Zhang et al., 2024a; Zhou et al., 2024c), debating judgment (Liang et al., 2024a), RAG (Dhole et al., 2024; Saad-Falcon et al., 2024a; Jin et al., 2024; Liu et al., 2025c; Seo et al., 2025), biomedical application (Brake and Schaaf, 2024; Zheng et al., 2025; Zhang et al., 2024i), paper review (Zhou et al., 2024e; Wang et al., 2024c; Zhu et al., 2025; Kirtani et al., 2025), novelty & creativity evaluation (Olson et al., 2024; Feng et al., 2025; Sawicki et al., 2025), and human-computer interaction (Li et al., 2024j).

## C.2 Alignment

Alignment tuning is a vital technique to align LLMs with human preferences and values (Wei et al., 2022a; Ouyang et al., 2022; Rafailov et al., 2023). In this section, we discuss the use of larger LLMs as judges (Section C.2.1) and self-judging (Section C.2.2) for alignment.

### C.2.1 Larger Models as Judges

Recently, alignment tuning leverages feedback from larger LLMs to guide smaller models. Bai et al. (2022) first propose to train reward models with synthetic preferences from pre-trained LLMs. Following this, there are also some works explore online learning (Guo et al., 2024) and direct preference optimization (Lee et al., 2023) with larger models as judges. To prevent reward hacking, Sun et al. (2024) develop an instructable reward model enabling real-time human interventions for alignment. Moreover, multi-agent collaborations employ diverse workflows and LLM debates to improve judgments in alignment tuning (Arif et al., 2024; Sengupta et al., 2024; Li et al., 2024i). For code alignment, Weyssow et al. (2024) create CodeUltraFeedback, a dataset using LLM judges to align smaller code models. Wang et al. (2024f) introduce BPO, employing GPT-4 as a judge to augment pairwise feedback.

### C.2.2 Self-Judging

Self-judging utilizes LLMs' own preference signals for self-improvement. Some focus on directly judging the preference ranking with the policy LLMs. Yuan et al. (2024c); Zhang et al. (2025c) first introduce self-rewarding, where LLMs judge their outputs to construct pairwise data. Following works adopt various methods to improve the judg-

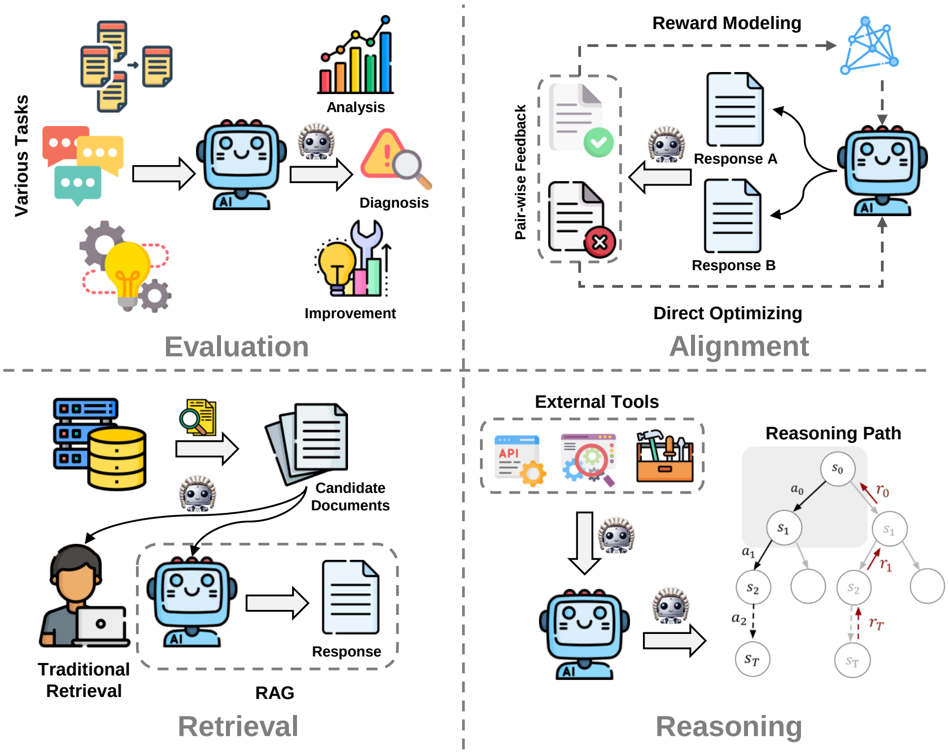


Figure 4: Overview of application and scenario for LLM-as-a-judge.

ing capabilities, including meta-rewarding (Wu et al., 2024a), Judge-Augmented Supervised Fine-Tuning (JSFT) (Lee et al., 2024a) and self-evaluation (Zhang et al., 2024g). To guarantee the quality of synthetic pairwise data, Pace et al. (2024) introduce West-of-N approach while Tong et al. (2024) apply self-filtering to produce high-quality synthetic data pairs for reasoning tasks. To reduce computational overhead, Zhai et al. (2024) propose ranked pairing for self-preferring models. Liu et al. (2024e) introduce meta-ranking, enabling smaller LLMs to act as judges and combining this method with Kahneman-Tversky optimization for post-SFT alignment. Besides pairwise data, (Liang et al., 2024c) and (Yasunaga et al., 2024) leverage LLM-as-a-judge to filter synthetic instruction tuning data. Other works adopt self-assessment and self-judgment in specific domains, such as robotics (Zeng et al., 2024; Yi et al., 2024) and multimodal (Ahn et al., 2024).

## C.3 Retrieval

In traditional retrieval, LLM-as-a-judge ranks documents by relevance with minimal labeled data (Section C.3.1). LLM judges can also enhance the RAG system by dynamically integrating retrieved knowledge into the final response (Section C.3.2).

### C.3.1 Traditional Retrieval

LLMs enhance document ranking by employing methods like permutation-based ranking (Sun et al., 2023), fine-grained relevance labeling (Zhuang et al., 2024a), and listwise reranking without task-specific training (Ma et al., 2023). Moreover, Setwise (Zhuang et al., 2024b) and Pairwise Ranking Prompting (PRP) (Qin et al., 2024) offer a cost-efficient alternative for complex tasks. Tang et al. (2024b) introduce a permutation self-consistency technique that averages across multiple orders to obtain order-independent rankings. Domain-specific knowledge retrieval with LLM-as-a-judge includes legal information, recommender systems and searching (Ma et al., 2024; Hou et al., 2024; Thomas et al., 2023).

### C.3.2 Retrieval-Augmented Generation (RAG)

Li and Qiu (2023a) propose the Memory-of-Thought (MoT) framework, where LLMs store and recall reasoning to enhance response relevance. Tang et al. (2024a) introduce Self-Retrieval, an architecture integrating retrieval into document generation, enabling end-to-end IR within a single LLM. Similarly, Asai et al. (2024) develop SELF-RAG, combining retrieval with self-reflection to enhance response quality. In the do-

| Benchmark | Definition |
|---|---|
| General Performance | Benchmarks that assess the general accuracy performance of LLM judges (e.g., MT-Bench) |
| Bias Quantification | Benchmarks focused on measuring and analyzing biases in LLM judgments (e.g., CALM) |
| Challenging Performance | Benchmarks that test LLM judges on difficult or adversarial tasks designed to probe the limits of their evaluation capabilities (e.g., Arena-Hard) |
| Domain-Specific Performance | Benchmarks that measure LLM judges' effectiveness in specific domains, such as biomedical, legal, and coding evaluation (e.g., Raju et al. (2024)) |

Table 2: Categories of benchmarks for evaluating LLM judges.

main of Q&A, Rackauckas et al. (2024) present an LLM-based evaluation framework using synthetic queries to judge RAG agent performance. Zhang et al. (2024c) study LLMs' ability to assess relevance versus utility. In the biomedical area, several studies explore the usage of LLM-as-a-judge for active and dynamic retrival (Wang et al., 2024b) or retrieved knowledge filtering (Jeong et al., 2024; Li et al., 2024c).

### C.4 Reasoning

Reasoning is a critical aspect of LLMs because it directly affects their ability to solve complex problems. Recently, many studies leverage LLM-as-a-judge in reasoning path selection (Section C.4.1) and external source utilization (Section C.4.2).

#### C.4.1 Reasoning Path Selection

While many complex reasoning and cognition structures emerges for LLMs' reasoning (Yao et al., 2023a; Hao et al., 2023), one crucial challenge is how to select a reasonable and reliable reasoning path or trajectory for LLMs to reason. To achieve this, LLM-as-a-judge has been introduced. Some works adopt the reasoner LLMs to perform self-assessment, alternatively executing reasoning and judging steps to achieve the best result (Lahoti et al., 2023; Creswell et al., 2023; Xie et al., 2024c; Kawabata and Sugawara, 2024) or perform sample-level selection among a group of candidates (Musolesi, 2024). Additionally, there are also many work train LLM-based verifiers, leveraging the judge LLM as the process reward model (PRM) to evaluate each state (Lightman et al., 2023; Setlur et al., 2024; Zhang et al., 2024d; Ye et al., 2025). Besides, there are also studies train critique-based LLM judges (Xu et al., 2024c; Ankner et al., 2024; Yu et al., 2024b; Wang et al., 2024e; Lan et al.; Xie et al., 2024b) which provide fine-grained verbal feedback to boost the reasoning process.

#### C.4.2 Reasoning with External Source

Selecting an appropriate external source to use is essential in the success of agentic LLM systems (Xi et al., 2023; Wang et al., 2024d). Auto-GPT (Yang et al., 2023) is the first to benchmark LLMs' performance in real-world decision-making scenarios. Following them, many other works adopt LLM-as-a-judge in various external tool selection applications, including autonomous driving (Sha et al., 2023), reasoning structure selection (Zhou et al., 2024d) and multi-modal area (Zhao et al., 2024b). In addition to selecting among external tools or APIs, LLM-as-a-judge has also been widely adopted as a controller in multi-agent systems, to selectively activate agents for a given problem (Ong et al., 2024) or to assess and manage message flow among a group of agents (Liang et al., 2023; Li et al., 2024b).

### C.5 Definition of each LLM-as-a-judge Benchmark Category

We provide the definition of each LLM-as-a-judge benchmark in Table 2.

# D Taxonomy

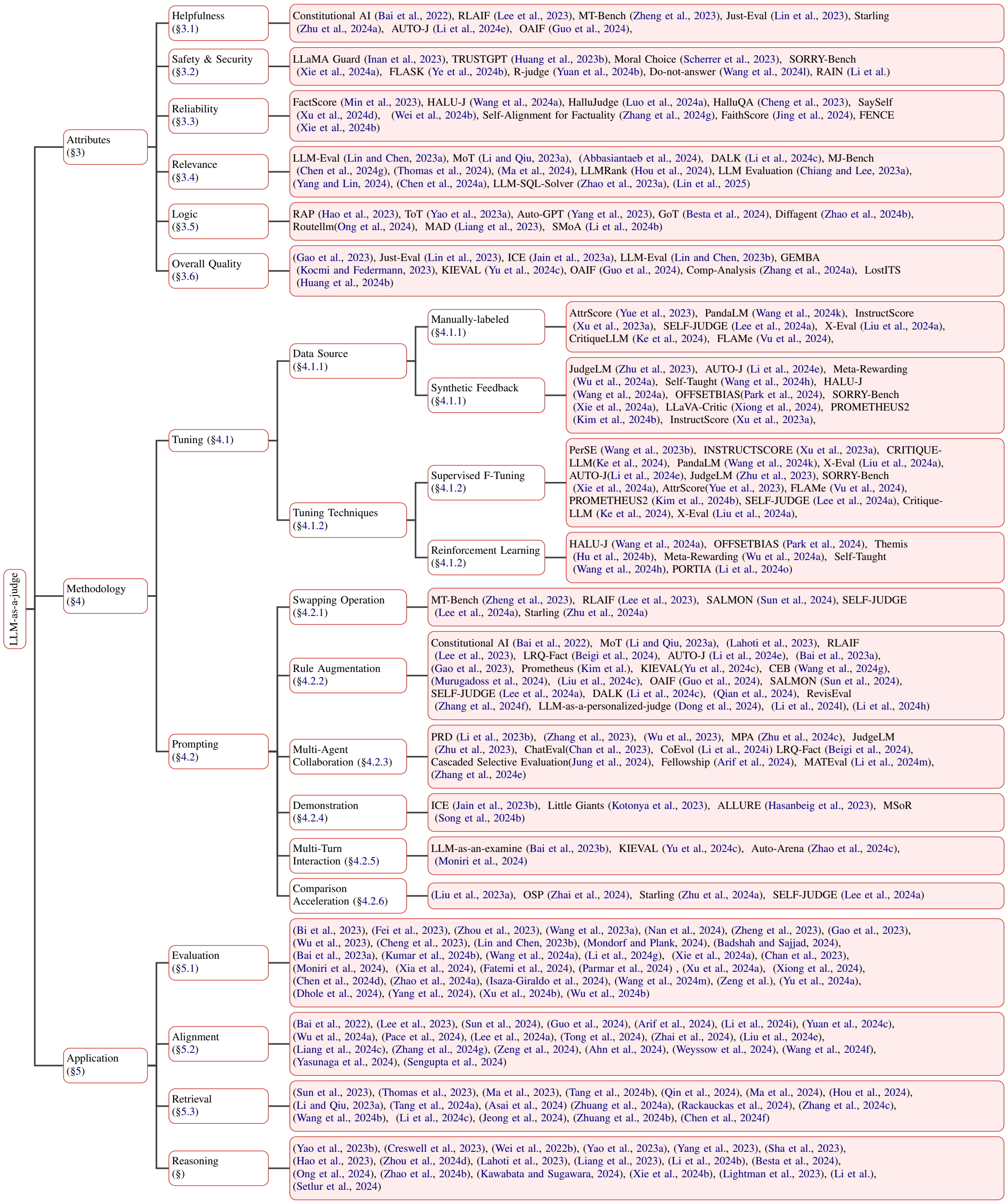


Figure 5: Taxonomy of research in LLM-as-a-judge that consists of judging attribution, methodology and application.

## E Tuning Methods

| Method | Data | | | | Tuning Method | | Base LLM |
|---|---|---|---|---|---|---|---|
| | Source | Annotator | Type | Scale | Technique | Trick | |
| AttrScore (Yue et al., 2023) | Manual | Human | QA, NLI, Fact-Checking, Summarization | 63.8K | SFT | - | Multiple LLMs |
| PandaLM (Wang et al., 2024k) | Manual | Human | Instruction Following | 300K | SFT | - | Multiple LLMs |
| AUTO-J (Li et al., 2024e) | Synthetic | GPT-4 | Real-world Scenarios | 4K | SFT | - | LLaMA-2 |
| JudgeLM (Zhu et al., 2023) | Synthetic | GPT-4 | Instruction Following | 100K | SFT | - | Vicuna |
| Self-Judge (Lee et al., 2024a) | Manual | Human | Preference Learning | 65/57K | SFT | JSFT | LLaMA-2 |
| X-EVAL (Liu et al., 2024a) | Manual | Human | Dialogue, Summarization, Data-to-Text | 55K | SFT | Two-Stage Instruction Tuning | Flan-T5 |
| FLAMe (Vu et al., 2024) | Manual | Human | Various Tasks | 5M+ | SFT | Multi-task Training | PaLM-2 |
| InstructScore (Xu et al., 2023a) | Manual& Synthetic | Human& GPT-4 | Various Tasks | 20K | SFT | Meta-Feedback | LLaMA |
| CritiqueLLM (Ke et al., 2024) | Manual | Human | Instruction Following, real-world scenarios | 5K | SFT | Prompt Simplify, Swapping Augmentation | ChatGLM3 |
| Meta-Rewarding (Wu et al., 2024a) | Synthetic | LLaMA-3 | Preference Learning | 20K | Preference Learning | Meta-Rewarding | LLaMA-3 |
| Self-Taught Evaluator (Wang et al., 2024h) | Synthetic | Mixtral | Various Tasks | 20K | Preference Learning | Self-Taught | LLaMA-3 |
| HALU-J (Wang et al., 2024a) | Synthetic | GPT-4o | Fact Extraction | 2.6K | Preference Learning | DPO | Mistral |
| OffsetBias (Park et al., 2024) | Synthetic | GPT-4, Claude3 | Preference Learning | 8.5K | SFT | Debiasing Augmentation | LLaMA-3 |
| SorryBench (Xie et al., 2024a) | Synthetic | GPT-4 | Safety | 2.7K | SFT | - | Multiple LLMs |
| LLaVA-Critic (Xiong et al., 2024) | Synthetic | GPT-4o | Preference Learning | 113K | Preference Learning | DPO | LLaVA-v.1.5 |
| PROMETHEUS2 (Kim et al., 2024b) | Synthetic | GPT-4 | Preference Learning | 300K | SFT | Joint Training, Weight Merging | Mistral |
| Themis (Hu et al., 2024b) | Manual & Synthetic | Human & GPT-4 | Various Tasks | 67K | Preference Learning | Multi-perspective Consistency Verification, Rating-oriented DPO | LLaMA-3 |

Table 3: Overview of tuning methods in LLM-as-a-judge.

## F Benchmark

## G AI Assistants In Writing

We acknowledge the use of ChatGPT-4o in paper polishing, but not in any direct paper writing or relevant work collections.

| Method | Data Type | Scale | Reference | Metrics | Purpose |
|---|---|---|---|---|---|
| MT-Bench (Zheng et al., 2023) | Multi-turn Conversation | 80 | Human Expert | Consistency, Bias, Error | General Performance, Position/Verbosity/Self-enhancement Bias |
| Chatbot Arena (Zheng et al., 2023) | Single-turn Conversation | 30K | User | Consistency, Bias, Error | General Performance, Position/Verbosity/Self-enhancement Bias |
| CodeJudge-Eval (Zhao et al., 2024a) | Code | 457 | Execution System | Accuracy, F1 | General Performance |
| JudgeBench (Tan et al., 2024b) | Various Tasks | 70K | Human | Cohen's kappa, Correlation | General Performance |
| SOS-BENCH (Penfever et al., 2024) | Various Tasks | 152K | Human | Normalized Accuracy | General Performance |
| LLM-judge-eval (Wei et al., 2024a) | Summarization, Alignment | 1K | Human | Accuracy, Flipping Noise, Position Bias, Length Bias | General Performance |
| DHP (Wang et al., 2024j) | Various Tasks | 400 | Human | Discernment Score | General Performance |
| EvalBiasBench (Park et al., 2024) | Alignment | 80 | Human | Accuracy | Various Bias |
| Raju et al. (2024) | Various Tasks | 1.5K | Human | Separability, Agreement, BrierScore | Domain-specific Performance |
| MLLM-as-a-judge (Chen et al., 2024a) | Various Tasks | 30K | Human | Human Agreement, Analysis Grading, Hallucination Detection | Multimodal |
| MM-EVAL (Son et al., 2024b) | Various Tasks | 5K | Human | Accuracy | Multilingual |
| KUDGE (Son et al., 2024a) | Question Answering | 3.3K | Human & GPT-4o | Accuracy, Correlation | Non-English & Challenging |
| Murugadoss et al. (2024) | Various Tasks | - | Human | Correlation | Evaluation Instruction Following |
| Thakur et al. (2024) | Question Answering | 400 | Human | Scott's $\pi$, Percent Agreement | Vulnerability |
| Rewardbench (Lambert et al., 2024) | Various Tasks | 20K | Human & LLMs | Accuracy | General Performance |
| Arena-Hard Auto (Li et al., 2024k) | Alignment | 500 | GPT-4-Turbo | Separability, Agreement | Challenging |
| R-Judge (Yuan et al., 2024b) | Multi-turn Interaction | 569 | Human | F1, Recall, Spec, Effect | Safety |
| Shi et al. (2024) | Alignment | 100K | Human | Repetition Stability, Position Consistency, Preference Fairness | Position Bias |
| CALM (Ye et al., 2024a) | Various Tasks | 14K | Human | Robustness/Consistency Rate, 0riginal/ Hacked Accuracy | Bias Quantification |
| VL-RewardBench (Li et al., 2024f) | Various Tasks | 1.2K | Human & LLMs | Overall Accuracy, Macro Average Accuracy | Multimodal |

Table 4: Overview of various benchmarks and datasets for LLM-as-a-judge.