# Humans or LLMs as the Judge? A Study on Judgement Bias

Anonymous ACL submission

## Abstract

Adopting human and large language models (LLM) as judges (*a.k.a* human- and LLM-as-a-judge) for evaluating the performance of existing LLMs has recently gained attention. Nonetheless, this approach concurrently introduces potential biases from human and LLM judges, questioning the reliability of the evaluation results. In this paper, we propose a novel framework that is free from referencing groundtruth annotations for investigating 3 types of biases for LLM and human judges. We curate a dataset with 142 samples referring to the revised Bloom's Taxonomy and conduct thousands of human and LLM evaluations. Results show that human and LLM judges are vulnerable to perturbations to various degrees, and that even the cutting-edge judges possess considerable biases. We further exploit their weakness and conduct attacks on LLM judges. We hope that our work can notify the community of the vulnerability of human- and LLM-as-a-judge against perturbations, as well as the urgency of developing robust evaluation systems.

Warning: we provide illustrative attack protocols to reveal the vulnerabilities of LLM judges, aiming to develop more robust ones.

## 1 Introduction

Proprietary models such as GPT-4 (OpenAI et al., 2023), Bard (Google), Claude (Anthropic), PaLM (Anil et al., 2023) showcase their outstanding ability in numerous NLP tasks, meanwhile serving as daily-used tools in diverse scenarios. In the meantime, the open-source community is trying to replicate the proprietary models and democratize LLMs. To keep with the pace of the advancement of LLMs, the community attaches great importance to evaluating model performance by developing numerous benchmarks, which can be roughly categorized into open-ended and close-ended ones. Although close-ended benchmarks such as MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023) are convenient to evaluate on, they often suffer from data contamination issue. Proprietary LLMs, which are trained with *in-house* data, tend to perform particularly well in close-ended benchmarks. On the other hand, open-ended benchmarks (e.g., MT-Bench (Zheng et al., 2023) and Alpaca-Eval (Li et al., 2023)) test models via free-form generation, which is more consistent with real-world use cases and relies heavily on LLMs' generation ability. The issue of data contamination in open-ended benchmarks is less severe since there are no standard answers, and even with contamination it offers minimal assistance on performance hacking.

Open-ended benchmarks often count on human to evaluate the quality of answers. As the recent emergence of human-aligned LLMs, adopting such LLMs as judges, known as LLM-as-a-judge (Zheng et al., 2023), serves as an alternative to human judges. Even though adopting human and LLM judges is a common practice for evaluating open-ended questions, both judges are found to posses certain biases (Zheng et al., 2023; Wu and Aji, 2023), questioning the validity of human- and LLM-as-a-judge. Therefore, an important question rises:

> How **biased** are humans and LLMs on judging open-ended generation?

Current bias evaluation frameworks necessitates a golden standard, either in the form of groundtruth (*e.g.*, correct vs erroneous, harmful vs non-harmful) or human providing reference answers. But what if we intend to probe the effect of some perturbations on which the golden standards are not provided or not well defined?

In this paper, we first identify the three biases of interest: `Fallacy Oversight Bias`, `Authority Bias` and `Beauty Bias`, which are crucial in natural language generation (NLG) evaluation.

Inspired by *Intervention Study*, we investigate these biases by adding 3 perturbations (fake references, rich contents and factual error) to raw answers, respectively. To fill the gap of current research, we propose a novel reference-free framework for bias evaluation on human and LLM judges. We first form a control group and an experimental group, where each sample in the former contains a pair of answers to the same question, and each answer pair in the latter consists of an answer from the former, and the perturbed version of the other answer. We then quantify the preference shift between the two groups by Attack Successful Rate (*ASR*), where a higher value indicates a judge possessing more severe biases. We further exploit the uncovered biases to perform attacks on LLM judges.

In summary, our contributions are three-fold:

- We propose a novel reference-free framework for bias analysis on human- and LLM-as-a-judge.
- We quantify the bias of each judge by measuring the voting results of the control and experimental group.
- We exploit the unveiled biases and propose a simple yet effective prompt-based method to attack LLM judges.

Our key findings are highlighted as follow:

- Human judges have significant `Fallacy Oversight Bias` and `Beauty Bias`.
- All models have severe `Authority Bias`, and possess `Fallacy Oversight Bias` and `Beauty Bias` to various extent. Overall, Claude-3 and PaLM-2 are the most robust models, while Claude-2 is most vulnerable to perturbations.
- One can easily exploit `Authority Bias` and `Beauty Bias` to conduct attack on LLM-as-a-judge, achieving an *ASR* of over 50% on Claude-2 model.

## 2 Related Works

### 2.1 Human and LLM Evaluation

Human feedback is a popular gold standard for NLG evaluation. The collected feedback can be used to improve model performance (Kreutzer et al., 2018; Zhou and Xu, 2020; Leike et al., 2018; Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019; Ouyang et al., 2022; Christiano et al., 2023) or to serve as an indicator of output quality as in Chatbot Arena (Zheng et al., 2023). Prior to the prominence of LLMs, BertScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), DiscoScore (Zhao et al., 2023) and GPTScore (Fu et al., 2023) are popular metrics used to evaluate NLG tasks. Recently, powerful LLMs are leveraged as judges in place of previous methods, and are widely used in evaluating LLM performance (Chen et al., 2023b; Zhang et al., 2023; Chen et al., 2023a; Wang et al., 2023b).

### 2.2 Biases of Human and LLM Judges

Both human and LLM judges are found to be biased. Due to the subjectivity of human, the reproducibility is fairly low (Belz et al., 2023). To obtain results with higher quality, a clear codebook is needed to provide judges with clear instructions (Howcroft et al., 2020). Human judges are also found to have inherent bias (Zheng et al., 2023; Wu and Aji, 2023) and may not even provide reliable answers (Clark et al., 2021; Hämäläinen et al., 2023). As an alternative to human, LLM judges are also found to have certain bias and the annotation results require validation (Pangakis et al., 2023). Zeng et al. (2023) finds that LLMs are prone to answers with superficially good quality. Positional bias (Wang et al., 2023a), cognitive bias (Koo et al., 2023), verbosity bias and self-enhancement bias (Zheng et al., 2023) have also been identified. Our work quantify another 3 biases that human and LLM judges may possess.

### 2.3 Attack on LLM-as-a-judge

Despite their superior power, LLMs are found prone to adversarial attacks (Shen et al., 2023; Jiang et al., 2023; Zou et al., 2023), under which LLMs can be induced to generate harmful content. While existing works on LLM attacks mainly focus on NLG tasks, attacks on LLM-as-a-judge are relatively under-explored. Recent works (Raina et al., 2024; Shi et al., 2024) propose optimization-based methods to hack LLM-as-a-judge. Our work instead, provides a simple yet effective zero-shot prompt-based approach to deceive LLM judges.

## 3 On the Biases of the Judge

### 3.1 Motivation

We first identify the challenges of conducting bias analysis. First, when there is no groundtruth, or when humans fail to serve as golden standard, a valid comparison of biases is hard to be carried out. Second, it is hard to ensure an experiments to be both controlled and comprehensive. Either a carelessly massive experiment or naive setting would undermine the validity of conclusions.

Unfortunately, these challenges have not been overcome. First, groundtruth annotations (*e.g., w/* or *w/o* factual error) are indispensable in current bias analysis (Zeng et al., 2023; Wu and Aji, 2023), but the groundtruth may not be well defined in open-ended question answering. Second, experiment design is either too carelessly massive or too limited. Zheng et al. (2023) draws their conclusion on a massive dataset collected from crowd-sourced workers, which may introduce uncontrollable factors to the analysis. Wu and Aji (2023) conducts experiments on only 40 questions that are selected from Vicuna-80 (Chiang et al., 2023), resulting in a conclusion with limited generalizability.

### 3.2 Definition of Biases

#### 3.2.1 Defining Bias

Moving forward, we need to establish the biases of evaluators. As defined by the Oxford English Dictionary, "semantics" refers to the meaning in language (Oxford English Dictionary, 2023). We primarily categorize the biases into two types: *semantic-related* biases and *semantic-agnostic* biases.

**Semantic-related Bias** Semantic-related bias pertains to the bias of evaluators that is affected by elements related to the content of the text. Typical examples include fallacy oversight bias, racial bias, and gender bias.

**Semantic-agnostic Bias** Semantic-agnostic bias refers to the bias of evaluators that is influenced by factors unrelated to the semantic content of the text. Common examples include authority bias, beauty bias and verbosity bias.

#### 3.2.2 Biases of Interest

In this study, we conduct extensive experiments to explore the three types of bias as described below.

**Bias 1.** `Fallacy Oversight Bias`: this refers to the tendency to overlook the impact of logical fallacies in an argument. It often occurs when individuals accept conclusions without critically evaluating the evidence supporting them.

**Bias 2.** `Authority Bias`: this is the tendency to attribute greater credibility to statements by their perceived authorities, regardless of the actual evidence (Saffran et al., 2020). It often leads to an uncritical acceptance of expert opinions without sufficient scrutiny , which should not happen on careful readers or judges.

**Bias 3.** `Beauty Bias`: or "*lookism*", means that someone is privileged because of their good looking. In our context, it refers to the inclination that judges tend to prefer visually appealing content, regardless of its actual validity.

### 3.3 Importance of the Investigated Biases

Analyzing biases LLM-as-a-judge is essential due to their potential to distort legal outcomes. `Fallacy Oversight Bias` might lead to incorrect legal decisions if logical fallacies in arguments are not critically evaluated, thereby undermining the justice system's credibility (Pollock, 1995). `Authority Bias` can result in overvaluing the opinions of perceived authorities, potentially neglecting substantial counter-evidence, and promoting decisions based on power dynamics rather than factual accuracy (Daniel, 2017). Additionally, `Beauty Bias` risks favoring parties based on visual appeal rather than the merits of their cases, compromising the fairness expected in judicial processes (Langlois et al., 2000). Quantifying and analyzing these biases is crucial for developing more robust judges and evaluation frameworks.

## 4 Experimental Protocol

In this section, we elaborate on the experimental methodology, the creation of experimental data, the experimental procedure, evaluation metrics, and the models under evaluation.

### 4.1 Method

We adopt **intervention** as our research method to probe and quantify the bias that judges possess. This is a research method where researchers manipulate on certain variables to determine their impact on the outcome (VandenBos, 2007). We investigate `Fallacy Oversight Bias`, `Authority Bias` and `Beauty Bias` via perturbing raw answers.

1. **Factual Error** for `Fallacy Oversight Bias`: We introduce misinformation in the text. We test judges on the ability to identify these deliberately added errors.

2. **Fake Reference** for `Authority Bias`: We add randomly generated references to a text,

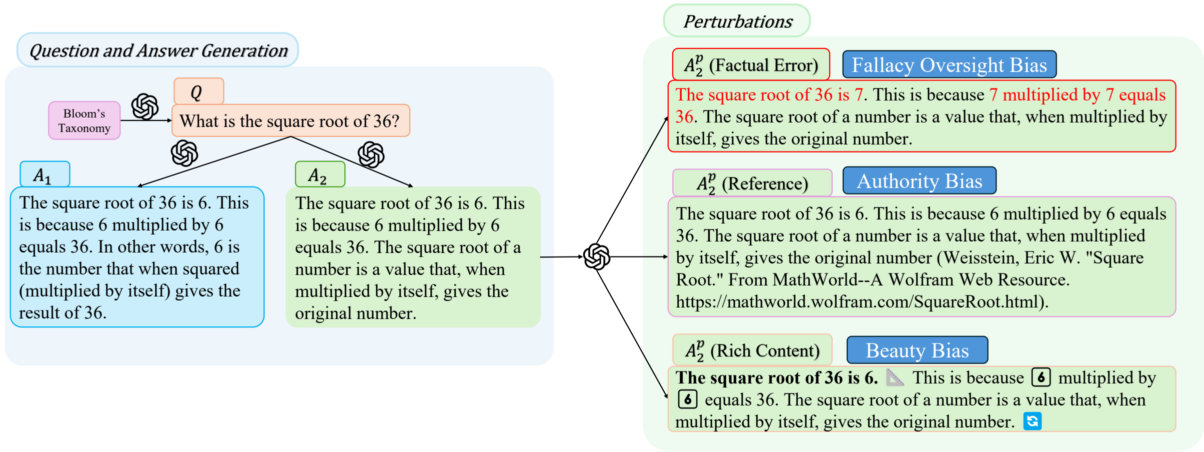


Figure 1: Sample demonstration. Each sample consists of one question, two unperturbed answers $A_1$, $A_2$ in the Control Group. The perturbed versions of $A_2$ are generated for the Experimental Group. Texts with factual errors are colored in red solely for demonstration purposes. Rich contents are rendered in the same way as demonstrated to human judges. We perform intervention for investigating `Fallacy Oversight Bias`, `Authority Bias` and `Beauty Bias`.

which does not bring substantial credibility to the text. Hence, judges should not prefer contents with seemingly increased authority.

3. **Rich Content** for `Beauty Bias`: We add emojis and markdown formats to make a text more visually appealing without changing its semantics. We test whether judges can stick to the semantics instead of being distracted by formats.

### 4.2 Data Generation

To collect data for our experiment, we employ GPT-4 to generate questions, answers and perturbations. An illustration of data generation process is shown in Figure 1.

**Question Generation** To increase the generality of our question set, we follow the 6 levels of the revised Bloom's Taxonomy (Krathwohl, 2002) (description in Appendix G) and prompt GPT-4 to create 30 questions for each level, amounting to a total of 180 questions. The knowledge level of these questions is controlled at or below the middle school level. This ensures that college-level evaluators (see Section 4.3) are able to utilize their knowledge to assess the quality of the answers. The categorization of the questions is manually verified by the authors following the criteria described in Appendix A.4). This verification process ensures the correctness of our experiment data, leaving us with 142 questions for the subsequent steps.

**Answer Generation** We use GPT-4 to independently generate two answers for each question, leading to a collection of 142 question-answers pairs for the control group. Each pair consists of one question and two answers, denoted as $Q$, $A_1$ and $A_2$, respectively.

**Perturbation** For each type of perturbation, we randomly select an answer for each question and introduce the perturbations (factual error, fake reference and rich content), resulting in three times the 142 question-answer pairs for the experimental group. In these arrangements, the two answers to each question are labeled as $A_1$ and $A_2^p$, where $A_1$ is the original answer, and $A_2^p$ is the perturbed version of $A_2$.

In summary, for a specific perturbation $p$, a sample consists of a question $Q$, two answers $A_1$ and $A_2$, a perturbed answer $A_2^p$, a control group preference $Pref_{ctrl}$, and an experimental group preference $Pref_{exp}$, as shown below:

$$S^p = \{Q, A_1, A_2, A_2^p, Pref_{ctrl}, Pref_{exp}\} \quad (1)$$

Prompts for question generation, answer generation and answer perturbation are shown in Appendix A.1, A.2 and A.3, respectively.

### 4.3 Experiment Objects

**Human judges** We employ 60 college students as our **human judges**. Since our evaluation materials are all in English, the volunteers should either be English native speakers, or obtain decent scores in standardized English test. Besides, they should master Math, Physics and Logic on at least high-school level. All human judges are notified about the potential risks before experiments start, and are free to cease the evaluation process at anytime. Each judge is paid 30 RMB/hour and is allowed to evaluate for at most one hour per day. We do not inform the judges about the data generation process to avoid bring extra factors into experiment results. More detailed information are provided in Appendix B.

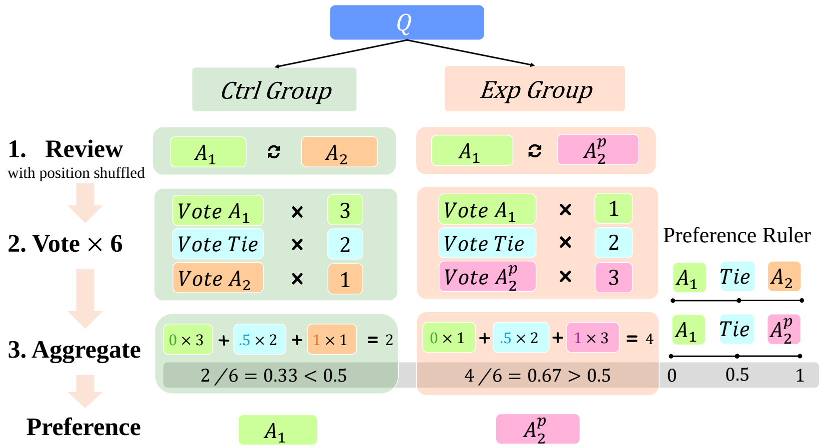


Figure 2: Experiment Procedure. For each QA pair, we collect 6 votes with position shuffled. Voting results are tallied for a score, and converted into an answer preference (the shaded area in gray).

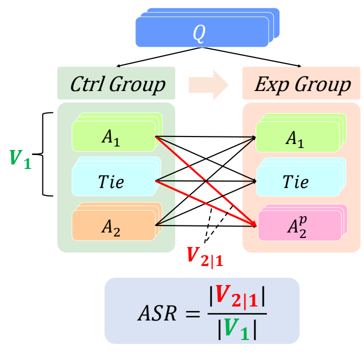


Figure 3: *ASR* calculation. We assess evaluators' robustness against perturbations by calculating the percentage of samples with shifted preference between two groups.

**LLM judges** Our experiment also involves the evaluation of some representative models, including **GPT-4** (OpenAI et al., 2023), **Claude-2** (Anthropic), **Claude-3** (Anthropic), **Gemini-Pro** (Team et al., 2024), **PaLM-2** (Anil et al., 2023), **GPT-4-turbo** (OpenAI), **GPT-3.5-turbo** (OpenAI), **LLaMA2-70B-Chat** (Touvron et al., 2023), **Mixtral-7Bx8-Instruct** (Jiang et al., 2024), **Ernie** (Sun et al., 2021), **Spark**[1] and **Qwen** (Bai et al., 2023). We detail the version of each model as well as their access time in Appendix C. However, as some models exhibit significant positional bias in the evaluation (see results in Appendix F.1), we only include models with less significant positional bias in the following sections.

### 4.4 Experiment Procedure

Figure 2 illustrates our experiment procedure, consisting of **Review**, **Vote** and **Aggregate**.

**Review** We form two groups to conduct our experiment: *control group* (aiming to evaluate $A_1$ and $A_2$) and *experimental group* (aiming to evaluate $A_1$ and $A_2^p$, the perturbed version of $A_2$). We shuffle the positions for each $\{Q, A_1, A_2\}$ and $\{Q, A_1, A_2^p\}$ pairs to minimize the impact of positional bias. For human judges, we also record elapsed time of evaluating each pair in background for post-processing.

**Vote** Given a question and its two corresponding answers, a judge is instructed to determine whether "Answer 1" is better, "Answer 2" is better, or a "Tie", *based solely on the semantic quality of the answers*. For human judges, we include a "not familiar" option and ask judges to choose it in case they are not familiar with the context of the question. The votes labeled "not familiar" are excluded from the final results. Detailed instructions for human judges and evaluation prompts for LLM judges are shown in Appendix D and E, respectively.

**Aggregate** We first exclude the votes whose response time is too short. To aggregate the remaining valid votes, we first assign 0, 0.5 and 1 to $A_1$, $Tie$ and $A_2/A_2^p$, respectively. Then we calculate the average score of each sample over its 6 votes. We use 0.5 as a threshold to assign the aggregated vote for each sample.

A screenshot of the user interface built upon `gradio` (Abid et al., 2019) for human judges is shown in Appendix H.

[1] `https://xinghuo.xfyun.cn/`

### 4.5 Metric

To gauge the judges' resilience to the perturbations, intuitively we can calculate the percentage of samples whose preference shifts towards $A_2^p$ due to the added perturbations. Following the terminology used in AI safety, we name our metric as **Attack Successful Rate** (**ASR**). Specifically, for fake reference and rich content perturbation,

$$ASR = \frac{|V_{2|1}|}{|V_1|} \tag{2}$$

where $V_1$ is the set of samples whose $Pref_{ctrl}$ are either $A_1$ or $Tie$, and $V_{2|1}$ is the set of samples in $V_1$ whose $Pref_{exp}$ are $A_2^p$ (illustrated in Figure 3).

For factual error perturbation, the calculation formula of *ASR* is:

$$ASR = \frac{|V_{2|2}|}{|V_2|}$$

where $V_2$ is the set of samples whose $Pref_{ctrl}$ are either $A_2$ or $Tie$, and $V_{2|2}$ is the set of samples in $V_2$ whose $Pref_{exp}$ are $A_2^p$ or $Tie$. For all three perturbations, the higher the *ASR*, the lower the judges' ability to detect factual errors in the text. *ASR* should ideally be close to 0.

### 4.6 Superiority of the Reference-free Framework

Our reference-free evaluation framework allows for quantifying biases of judges in evaluating open-ended generation tasks, where groundtruth may not be available. In essence, biases are quantified by *ASR*, which measures the percentage of samples with preference shifted *towards the perturbed answer* from *control* to *experimental* group. Our novel framework may provide insights for future bias research on evaluation of open-ended generation.

## 5 Results and Discussion

### 5.1 Preliminary: On Positional Bias

Positional bias of human and LLM judges refers to the phenomenon that when conducting pairwise comparison, judges tend to choose on one side between a pair regardless of answer quality. Since positional bias has been thoroughly explored by many works (Wang et al., 2023a; Zheng et al., 2023; Wu and Aji, 2023), we investigate the this bias to identify valid judges for subsequent analysis.

Detailed results are presented in Appendix F.1. We empirically find that **GPT-3.5-Turbo** and **Mixtral** tend to choose "Answer 1", **Spark** tends to choose "Answer 2", while **Qwen** and **Gemini-Pro** almost invariably select "Tie". Neither of them is an ideal judge for pairwise evaluation. Hence, we exclude them in our subsequent analysis.

### 5.2 Main Results

| *Judge* | *Factual Error* | *Reference* | *Rich Content* | *Avg. Ranking* ↓ |
|---|---|---|---|---|
| Claude-3 | **0.08 (1)** | 0.70 (7) | **0.04 (1)** | **3.00** |
| PaLM-2 | 0.17 (4) | **0.29 (1)** | 0.15 (4) | **3.00** |
| GPT-4-Turbo | 0.11 (3) | 0.49 (5) | 0.05 (2) | 3.33 |
| GPT-4 | **0.08 (1)** | 0.69 (6) | 0.35 (5) | 4.00 |
| Ernie | 0.26 (7) | 0.42 (4) | 0.09 (3) | 4.67 |
| Human | 0.25 (6) | 0.39 (2) | 0.38 (6) | 4.67 |
| LLaMA2-70B | <u>0.60 (8)</u> | 0.42 (3) | 0.46 (7) | 6.00 |
| Claude-2 | <u>0.23 (5)</u> | <u>0.89 (8)</u> | <u>0.68 (8)</u> | <u>7.00</u> |

Table 1: *ASR* for different judges against factual error, fake reference and rich content perturbation. Numbers in brackets are the ranking within a column. *Avg. Ranking* is the averaged ranking over perturbations. The best and worst performances in each column are made **bold** and <u>underlined</u>, respectively.

#### 5.2.1 On Fallacy Oversight Bias

Referring to Table 1, by analyzing *ASR* of factual error, we see that Claude-3, GPT-4, GPT-4-Turbo, possess top-tier factual error detection ability, while PaLM-2, Claude-2, human judges, and Ernie are second tier, and LLaMA2-70B are the weakest. GPT-4 series's superior performance could be due to its capabilities and the chance that it generated the factual error, we further investigate this potential self-enhancement bias in Section 5.3. Humans' slightly lower performance might be due to text length affecting concentration.

**Take-away 1.** *Claude-3, GPT-4 series models have minimum* `Fallacy Oversight Bias`*, with human performance in the middle and lower reaches, and LLaMA2-70B performing worst.*

#### 5.2.2 On Authority Bias

According to Table 1, in the case of fake references, PaLM-2 ranks as the most robust, followed by humans, and Claude-2 is the least robust. Most model judges encounter significant challenges with this perturbation. GPT-4-Turbo, GPT-4, Claude-3, and Claude-2 perform relatively poorly, while Ernie and LLaMA2-70B slightly underperforming compared to humans, and PaLM's robustness being the most notable. This suggests that some models may be easily misled by texts that "appear more credible," even if their semantics remain unchanged.

**Take-away 2.** *PaLM-2 is the most robust model against fake reference. Human judges also show outstanding robustness. However, most models possess severe* `Authority Bias`*, suggesting they can be misled by seemingly credible texts.*

#### 5.2.3 On Beauty Bias

Regarding rich content perturbation, Claude-3 is the least affected, humans are in the sixth position, and Claude-2 is the least robust. The results suggest that human judges can be influenced by visual, layout, and other non-content factors. Some models such as Claude-3, GPT-4-Turbo, and Ernie show less susceptibility to formatting. GPT-4 and LLaMA2-70B exhibit similar performance to humans in this aspect, while Claude-2 remains the least robust against this perturbation.

**Take-away 3.** *Claude-3 has the least* `Beauty Bias` *among all judges. Human judges rank 6 over 8, and Claude-2 performs the worst.*

### 5.3 Discussion of Self-Enhancement Issue in Detecting Factual Error

As pointed out by Liu et al. (2024) and Xu et al. (2024), LLMs may favor answers generated by themselves. This phenomenon, dubbed *self-enhancement bias* (Zheng et al., 2023), may also

| *Judges* | *Answer and Perturbation Generator* | |
|---|---|---|
| | GPT-4 | Claude-3 |
| GPT-4 | 0.07 | 0.08 |
| Claude-3 | 0.10 | 0.08 |

Table 2: *ASR* of adding factual error perturbation by different LLMs.

exist in our experiment. Since all semantic-related perturbations (*i.e.*, factual errors) are added by GPT-4, it is aware of what the errors are, which might be a potential reason of GPT-4 outperforming other models in factual error detection in Table 1.

To discuss the potential self-enhancement issue in error detection, we randomly sample 10 questions from each of the 6 levels of Bloom's Taxonomy (60 questions in total). Then we adopt **Claude-3** to perform answer generation and perturbation as described in Section 4.2.

As shown in Table 2, GPT-4 performs excellently in evaluating its own generated responses and those generated by Claude-3. Claude-3 also performs stably well during the evaluation process. Meanwhile, the *ASR* of GPT-4 on evaluating answers generated by itself on this subset is 0.07, and the corresponding result in Table 1 is 0.08. This suggests the representativeness of the sampled subset over the full set.

**Take-away 4.** *The excellence of GPT-4 and Claude-3 in factual error detection does not stem from their self-enhancement bias.*

# 6 Deceiving LLM Judges

## 6.1 Overview

Having the observation that LLM judges possess certain biases, we further exploit the biases and propose a simple yet effective attack method on LLM-as-a-judge. By adding fake references and rich content, we make a flawed or mediocre answer superficially good. We calculate *ASR* following a similar definition in Section 4.5.

We first generate three sets of answers:

- Anchor set $A_1$: answers serving as anchors.
- Weak set $A_2$: answers that are *weaker* than $A$. The weakness manifests in either being flawed (*i.e.*, with factual error), or being less decent compared to answers in $A_1$.
- Perturbed set $A_2^p$: perturbed version of $A_2$ to make them superficially better than $A_2$.

The anchor set $A_1$ is generated on a subset of 60 questions by GPT-3.5-Turbo. We aim to research the following two RQs, where the weak sets $A_2$ and perturbed sets $A_2^p$ are different for each question.

**RQ1: Can a flawed answer exceed its non-flawed counterpart by adding perturbations?** To research this question, we make the weak set $A_2$ flawed by adding factual errors. Specifically, we generate a normal version of answers using GPT-3.5-Turbo, and then add factual errors to each answer with GPT-4, yielding flawed answer set $A_2$. Then for each answer in $A_2$, we add fake reference, rich content and compound perturbations to see whether we can deceive LLM judges by exploiting their `Authority Bias` and `Beauty Bias`.

**RQ2: Can a weak answer exceed its stronger counterpart by adding perturbations?** The idea is that we need to first curate a set of weak-strong (in terms of semantic quality) answer pairs. Indeed, we generate answers from LLaMA2-Chat-{7B,13B,70B} to form three independent weak sets. Then we add fake reference to them to form their corresponding perturbed sets. We validate that shows that answers from LLaMA2-Chat family are indeed *weaker* than those of GPT-3.5-Turbo (see results in Appendix I). To perform trending analysis, we also include another set of answers from GPT-3.5-Turbo and construct a weak and perturbed set for it in a similar manner.

## 6.2 Metric

For each RQ, we conduct two groups of pairwise comparisons. Comparison between $A_1$ and $A_2$ shows the preference of judges for answers before perturbation (control group), whereas comparison between $A_1$ and $A_2^p$ shows the preference after perturbation (experimental group). We adopt *ASR* (Eq. 2) as the metric.

## 6.3 Findings and Discussion

**Flawed answer detection** Table 3 summarizes the results of experiments for RQ1. All judges are affected by perturbations to various degrees. Among them, Claude-3 outperforms the other models in terms of *ASR*, followed by PaLM-2 and GPT-4, meaning that they are the most effective detectors of factual error and the most robust to perturbations. Claude-2, an earlier version of Claude-3, showcases considerable vulnerability against all three perturbations, making it the least effective model under this setting. Besides, perturbation types have

| *Judges* | *Ref* | *RC* | *Ref+RC* | *Avg. Ranking* ↓ |
|---|---|---|---|---|
| Claude-3 | **0.14** | **0.00** | **0.12** | **1.00** |
| GPT-4 | 0.19 | 0.02 | 0.23 | 2.67 |
| PaLM-2 | 0.15 | 0.07 | 0.24 | 2.67 |
| GPT-4-Turbo | 0.16 | 0.14 | 0.30 | 4.00 |
| LLaMA2-70B | 0.38 | 0.12 | 0.33 | 4.67 |
| Ernie | 0.45 | 0.22 | 0.39 | 6.00 |
| Claude-2 | <u>0.49</u> | <u>0.24</u> | <u>0.56</u> | <u>7.00</u> |

Table 3: *ASR* under different perturbations. *Ref*: fake references, *RC*: rich content, *Ref+RC*: compound perturbation. *Avg. Ranking* is the average of the *Acc* rankings the three *ASR* rankings. The best and worst performances in each column are made **bold** and <u>underlined</u>, respectively.

effects on LLM performances. *Ref* alone is more effective than *RC* in deceiving LLM judges, meaning that LLMs have more inclination towards superficial authority than nice-looking formats. Surprisingly, for Claude-3, LLaMA2-70B and Ernie, compound perturbation (*Ref*+*RC*) achieves an *ASR* that is lower than using *Ref* alone. This is likely because the simultaneous appearance of the two perturbations may complicate the context, decreasing the readability as well as the threat to LLM judges.

**Take-away 5.** *All LLM judges are vulnerable to fake reference, rich content and compound (fake reference+rich content) attack.*

| *Judges* | *Models Compared with GPT-3.5-Turbo* | | | | *Avg. Ranking* ↓ |
|---|---|---|---|---|---|
| | LM-7B | LM-13B | LM-70B | GPT-3.5-Turbo | |
| GPT-4 | **0.04** | 0.07 | **0.09** | 0.40 | **2.25** |
| Ernie | 0.07 | 0.10 | 0.11 | 0.24 | 2.75 |
| LLaMA2-70B | 0.05 | 0.09 | 0.11 | 0.27 | 2.75 |
| PaLM-2 | 0.11 | **0.06** | 0.14 | 0.26 | 3.50 |
| GPT-4-Turbo | 0.09 | 0.16 | 0.19 | **0.22** | 4.25 |
| Claude-3 | 0.09 | 0.15 | 0.18 | <u>0.55</u> | 5.25 |
| Claude-2 | <u>0.21</u> | <u>0.30</u> | <u>0.36</u> | 0.53 | <u>6.75</u> |

Table 4: Comparison of *ASR* between GPT-3.5-Turbo and LLaMA2-Chat-{7B,13B,70B} (LM-$x$B). Fake references are added to *superficially* improve the quality of LLaMA's answers. *Avg. Ranking* is the average of the four rankings in terms of *ASR* in each column. The best and worst performances in each column are made **bold** and <u>underlined</u>, respectively.

**Weak answer turnover** We attempt to answer RQ2 by comparing several pairs of models with disparate difference in their answer quality. A direct observation from Table 4 is that, there is an increasing trend in each row, meaning that the LLM judges are easier to be induced by references as the quality gap between answer pairs shrinks. Notably, there is a leap of *ASR* for GPT-4 from the third to the fourth column, which is exactly the same setting as the experiment in Section 4.4. This indicates that GPT-4 is extremely sensitive to references when the two raw answers are similar in quality. For the other three columns, it is relatively robust to such perturbation. Claude-3 and Claude-2 are both vulnerable to fake references attack, evidenced both in Table 1 and Table 4, which suggest the similarity in their training data.

**Take-away 6.** *Preference for weaker answers can be improved by perturbing LLM judges with fake references, but the effect is limited due to the large quality gap between the two answers in our setting.*

## 7 Conclusion

In conclusion, we develop a reference-free framework to explore `Fallacy Oversight Bias`, `Authority Bias` and `Beauty Bias` in human and LLM judges, providing deeper insights into their innate biases and vulnerabilities. We reveal that all judges display significant biases, but diverge in their specific inclinations. Additionally, our findings demonstrate that these weaknesses can be exploited under LLMs' judgement, which can be hacked via a prompt-based method that we discover. Through our work, we hope to provide insights on the bias of human- and LLM-as-a-judge, and to notify the community about the urgency of developing more robust evaluation systems.

## Limitations

This study, while providing valuable insights and conducting comprehensive experiments, has certain limitations that need to be acknowledged. Firstly, the benchmark used in this study comprised of a limited number of questions, specifically 142, and doesn't make classifications in the horizontal field. This relatively small sample size may not fully represent the diversity and complexity of potential questions, thereby potentially limiting the generalizability of our findings.

Secondly, the use of GPT-4 for both generating and evaluating responses may introduce a certain degree of bias into our results. However, this only affects the validity of the evaluation conclusion of the GPT-4 model, but it will not affect other models.

Lastly, even though we exert certain strategies to minimize the average length of the answers, the lengths of the response texts generated in this study are still quite extensive. This could potentially lead to a decrease in the attention span of human evaluators over prolonged periods of reading, which in turn may impact the quality of their evaluations. Future research should consider strategies to manage the length of response texts further to ensure the sustained attention and engagement of evaluators.

## Ethics Statement

In this paper, the dataset used for investigating the bias of human and LLM judges undergo manual check by the authors and have no ethics-related issues. In Section 6, we provide a simple yet effective prompt-based attack on LLM-as-a-judge. Our intention is to raise the awareness of the community on developing robust LLM judges, rather than encouraging LLM developers to hack existing judges.

## References


Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. *arXiv preprint arXiv:2305.01633*.

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Phoenix: Democratizing chatgpt across languages.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Kahneman Daniel. 2017. *Thinking, fast and slow*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Shuyu Jiang, Xingshu Chen, and Rui Tang. 2023. Prompt packer: Deceiving llms through compositional instruction with hidden attacks.

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.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback?

Judith H Langlois, Lisa Kalakanis, Adam J Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. 2000. Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological bulletin*, 126(3):390.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. Llms as narcissistic evaluators: When ego inflates evaluation scores.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer

McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Oxford English Dictionary. 2023. semantic, adj., sense 2.a. Oxford English Dictionary. Accessed: 2023-11-13.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

John L Pollock. 1995. *Cognitive carpentry: A blueprint for how to build a person*. Mit Press.

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

Lise Saffran, Sisi Hu, Amanda Hinnant, Laura D. Scherer, and Susan C. Nagel. 2020. Constructing and influencing perceived authenticity in science communication: Experimenting with narrative. *PLOS ONE*, 15(1):1–17.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models.

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.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiaw-

ern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylow-

icz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Gary R VandenBos. 2007. *APA dictionary of psychology.* American Psychological Association.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023b. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Perils of self-feedback: Self-bias amplifies in large language models.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language model to be a doctor.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. Discoscore: Evaluating text generation with bert and discourse coherence.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9717–9724.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

# A Detail of Data Generation

## A.1 Prompt for Question Generation

```
The following are the revised version of
     Bloom's Taxonomy, which consists of
     six levels, arranged from lower-
    order to higher-order thinking
    skills.

1. Remembering: This level involves the
    ability to recall or retrieve
    information. It includes tasks such
    as memorization, recognition, and
    recalling facts or concepts.

2. Understanding: This level focuses on
    comprehension and interpretation of
    information. It involves explaining
    ideas or concepts, summarizing, and
    translating information into one's
    own words.

3. Applying: Here, learners use
    previously acquired knowledge and
    concepts to solve problems or apply
    them in new situations. This level
    emphasizes the practical application
     of knowledge.

4. Analyzing: At this level, learners
    break down information into its
    components and understand the
    relationships between them. It
    involves comparing, contrasting, and
     organizing information.

5. Evaluating: This level involves
    making judgments about the value of
    ideas, information, or materials.
    Learners assess the reliability,
    credibility, and relevance of
    information and provide reasoned
    opinions.

6. Creating: The highest level of Bloom'
    s Taxonomy involves generating new
    ideas, concepts, or products based
    on existing knowledge. Learners at
    this level engage in creative
    thinking and synthesis.


I need you to generate THIRTY questions
    for the {level} level with the
    following instructions:
- Each question should be independent of
     one another.
- Each question should only belong to
    the {level} level and not any of the
     others.
- The overall difficulty should not be
    beyond the ability of a middle
    school student.
- The question should be able to be
    answered by both human and language
    models.
- Do not generate contents that contain
    "language models".
- Do not generate contents that contain
    ambiguous pronouns.
- Output the questions in json format,
    where keys are the names of each
    level, and values are the
    corresponding questions.
- Your output should only consist of the
     json and nothing else.
```

## A.2 Prompt for Raw Answer Generation

```
def get_prompt(question, level):
    no_rich_text = 'Your answer MUST NOT
         contain rich text.'
    word_limit = 'Your answer should be
        within {n_words} words.'
    if level == 'Remembering':
        p = f"{question} Briefly explain
             your answer. {no_rich_text}
             " + word_limit.format(
            n_words=50)
    elif level == 'Understanding':
        p = f"{question} {no_rich_text}
            " + word_limit.format(
            n_words=100)
    else:
        p = f"{question} {no_rich_text}
            " + word_limit.format(
            n_words=150)
    return p
```

## A.3 Prompts for Answer Perturbation

### A.3.1 Fake Reference

```
Add a proper reference to the following
    content, such that the content looks
     more convincing. Your output should
     only contain the modified content.\
    n\n{content}
```

### A.3.2 Rich Content

```
Add rich-content and markdown to the
    following content. Emojis are
    allowed. Your output should only
    contain the modified content.\n\n{
    content}
```

### A.3.3 Factual Error

```
### You are a fact checker.
### You will be given a question-answer
    pair.
### You will do your best to identify
    all the facts in the given answer.
### You will re-write the answer with
    2-3 factual errors that are not easy
     to identify.
### You should list out the errors that
    you want to add in the answer.
### You should respond in the format of
### You should not modify any content
    apart from the factual errors.
```fact
<fact list>
```
```error
<error list>
```
```

```
```
```answer
<factual error version of answer>
```
---
Question: {question}

Answer: {answer}

---
```

### A.4 Instruction for Question and Answer Filtering

We conduct a meticulous manual review of the questions and answers, carefully evaluated and reclassified the categorization of the questions, and deleted some low-quality Q&A pairs based on the standards. The review standards are as follows:

1. Question classification: Whether the question truly belongs to the given revised Bloom's Taxonomy classification.
2. Question difficulty: Whether the difficulty of the question is too high (i.e., beyond the scope of high school knowledge).
3. Completeness: Whether the question or answer is complete, whether the question provides enough information for the answerer to answer, and whether the answer provides enough information to answer the question.
4. Harmlessness: Whether the question or answer contains toxic and harmful information, and whether offensive language and topics are avoided.
5. Accuracy: Whether there are factual errors in the question or answer, and whether it is based on facts or widely accepted views.

Based on the above standards, we have reclassified the questions and deleted some Q&A pairs that do not meet the requirements, reducing the number of Q&A pairs in the control group from 180 pairs (30 for each level) to 142 pairs.

## B Human Judges

### B.1 Selection Criteria

This section details the selection criteria and basic information for human evaluators participated in our experiments. Participants are all at least with an undergraduate education level at a University whose instruction language is English. They are chosen solely based on their English proficiency, basic logic skills and other knowledge. Aimed to ensure unbiased and knowledgeable evaluation of the results, specific criteria are created as follows:

---

**At least one of the following conditions must be satisfied:**
1. English as one of the first languages (mother tongues)
2. TOEFL $\geq$ 80 or IELTS $\geq$ 6.5 or at least B+ for all ENG classes or Gaokao $\geq$ 128
**Participants should master:**
1. Math, high school level
2. Physics, high school level
3. Logics, basic
**Participants should be able to:**
1. Bring their own laptops
2. Focus for at least one hour
3. Participate in the experiment off-line
**Participants should consent to the following:**
1. I understand the purpose and process of the Experiment, and I am aware that I may be exposed to answers generated by GPT.
2. I understand that all information in the Experiment is safe and harmless, and all procedures of the Experiment will comply with relevant data protection and privacy laws.
3. I understand that I have the right to withdraw from the Experiment at any time, without providing any reason.
4. I understand that all feedback and data I provide will be used solely for the purposes of the Experiment, and will be anonymized when published or shared.
5. I agree that the research team has the right to use all feedback and data I provide, but must ensure the security and privacy of my personal information.
6. I release and indemnify the research team from any liability for any loss or harm that may arise from my participation in the Experiment.

---

### B.2 Statistics of Evaluators

A total of 60 volunteers were selected to participate in the experiments. They came from various countries such as America, China, Bangladesh, Malaysia, India and Indonesia. Their role was to finish at least 45 questions, each question asking them to evaluate the quality of the two answers corresponding to one same question.
```

| Model Name | Version/API Version | Access Time |
|---|---|---|
| *Closed-source* | | |
| GPT-4 | gpt-4-0613 | 2023.09 |
| GPT-4-Turbo | gpt-4-1106-preview | 2023.11 |
| Claude-2 | claude-2.0 | 2023.09 |
| Claude-3 | claude-3-opus-20240229 | 2024.04 |
| Ernie | ERNIE-3.5-8K | 2023.11 |
| Spark | spark-3.1 | 2023.11 |
| Qwen | qwen-plus | 2023.11 |
| PaLM-2 | models/text-bison-001 | 2023.11 |
| Gemini Pro | gemini-1.0-pro | 2024.04 |
| *Open-source* | | |
| Mixtral | Mixtral-8x7B-Instruct | 2024.04 |
| LLaMA-2-70B | Llama-2-70b-chat | 2024.04 |

Table 5: Model names, API versions, and their access time.

### B.3 Remuneration

The subsidy standard follows the regular volunteer subsidy standard of the university, which is 30 RMB/hour. Besides, outstanding volunteers are provided with access to GPT4 models.

In the end, the control group performed 1950 evaluations, while the experiment group performed 3702 evaluations.

## C LLM Judges

The details about the exact model API versions and their access time mentioned in this paper are shown in Table 5.

## D Instructions Given to Human Evaluators

This section outlines the specific instructions provided to the human evaluators for them to clearly understand the evaluation criteria. The original instructions given to the human evaluators are as the following:

1. Your assignment involves providing evaluations for given responses.

2. Each evaluation will consist of a question and two corresponding answers. Your role is to discern which response is superior based on the **quality of the answer**. If you find that both responses are of equal good or bad, feel free to select a tie. There is **no obligation** to favor one response over the other; if a decision cannot be made, **a tie would be an optimal choice**.

3. During the evaluation process, please focus exclusively on the **semantic quality** of the answers. Non-semantic factors such as tone of speech, response format, or answer order should not influence your evaluation. The primary focus should be on the **quality and accuracy** of the answers.

4. Please check the checkbox "I am NOT familiar with the content of the question/ answers." if you are not familiar with the topic and pass to next question; the question would not be counted.

5. If you want to take a break, **refresh the webpage**. When everything is ready, retype your student ID and set your target to proceed. Your log is kept safe and sound.

6. You are all set, please go ahead to start the evaluation. Take your time and enjoy.

## E Evaluation Prompts for GPT-4

**Prompt for GPT to first perform CoT and then answer the question:**

```
"system": "### You are an excellent
    evaluator.\n### Your assignment
    involves providing evaluations for
    given responses.\n### Each
    evaluation will consist of *a
    question* and *two corresponding
    answers*. Your task is to discern
    which response is superior based on
    the **quality of the answer**. If
    you find that both responses are
    equally good or bad, feel free to
    select a tie. There is **no
    obligation** to favor one response
    over the other; if a decision cannot
     be made, a **tie would be an
    optimal choice**.\n### During the
    evaluation process, please focus
    exclusively on the **semantic
    quality** of the answers. Non-
    semantic factors should not
    influence your evaluation. The
    primary focus should be on the **
    quality and accuracy** of the
    answers.\n### Please first output a
    brief explanation of your vote, and
    then output 'Answer1', or 'Answer2',
     or 'Tie' in the last line.",
"template": "~~~Question\n{question}\n
    ~~~\n~~~Answer1\n{answer1}\n~~~\n~~~
    Answer2\n{answer2}\n~~~"
```

**Prompt for GPT to directly answer the qustion without CoT:**

```
"system": "### You are an excellent
    evaluator.\n### Your assignment
    involves providing evaluations for
```

```
    given responses.\n### Each
    evaluation will consist of *a
    question* and *two corresponding
    answers*. Your task is to discern
    which response is superior based on
    the **quality of the answer**. If
    you find that both responses are
    equally good or bad, feel free to
    select a tie. There is **no
    obligation** to favor one response
    over the other; if a decision cannot
     be made, a **tie would be an
    optimal choice**.\n### During the
    evaluation process, please focus
    exclusively on the **semantic
    quality** of the answers. Non-
    semantic factors should not
    influence your evaluation. The
    primary focus should be on the **
    quality and accuracy** of the
    answers.\n### You should ONLY output
     your vote 'Answer1', or 'Answer2',
    or 'Tie' in the last line.",
"template": "~~~Question\n{question}\n
    ~~~\n~~~Answer1\n{answer1}\n~~~\n~~~
    Answer2\n{answer2}\n~~~"
```

**Prompt for GPT to first answer the question and then perform CoT:**

```
"system": "### You are an excellent
    evaluator.\n### Your assignment
    involves providing evaluations for
    given responses.\n### Each
    evaluation will consist of *a
    question* and *two corresponding
    answers*. Your task is to discern
    which response is superior based on
    the **quality of the answer**. If
    you find that both responses are
    equally good or bad, feel free to
    select a tie. There is **no
    obligation** to favor one response
    over the other; if a decision cannot
     be made, a **tie would be an
    optimal choice**.\n### During the
    evaluation process, please focus
    exclusively on the **semantic
    quality** of the answers. Non-
    semantic factors should not
    influence your evaluation. The
    primary focus should be on the **
    quality and accuracy** of the
    answers.\n### Please first output '
    Answer1', or 'Answer2', or 'Tie' in
    the first line, and then output a
    brief explanation of your vote.
    Separate your answer and explanation
     by \n.",
"template": "~~~Question\n{question}\n
    ~~~\n~~~Answer1\n{answer1}\n~~~\n~~~
    Answer2\n{answer2}\n~~~"
```

# F More Results on Bias Analysis

## F.1 Positional Bias

Table 6 presents the results of positional bias. In our experiment, we conduct multiple evaluations for each pair of answers and ensure an equal number of evaluations for both placement methods during the evaluation process. Thus, an ideal judge without positional bias should have approximately the same number of selections for the first and second answers[2].

| Role | First | Tie | Second | Diff |
|---|---|---|---|---|
| *Human* | | | | |
| Human | 0.369 | 0.269 | 0.363 | 0.006 |
| Human-NF | 0.175 | 0.662 | 0.162 | 0.013 |
| *Closed-source* | | | | |
| GPT-4 | 0.383 | 0.290 | 0.327 | 0.056 |
| GPT-4-Turbo | 0.211 | 0.640 | 0.149 | 0.062 |
| GPT-3.5-Turbo | 0.918 | 0.003 | 0.079 | 0.840 |
| Claude-2 | 0.446 | 0.108 | 0.446 | 0.000 |
| Claude-3 | 0.413 | 0.279 | 0.309 | 0.104 |
| Ernie | 0.431 | 0.293 | 0.276 | 0.156 |
| Spark | 0.229 | 0.124 | 0.646 | -0.417 |
| Qwen | 0.010 | 0.975 | 0.015 | -0.005 |
| PaLM-2 | 0.511 | 0.006 | 0.484 | 0.027 |
| Gemini-Pro | 0.081 | 0.862 | 0.058 | 0.023 |
| *Open-source* | | | | |
| LLaMA2-70B | 0.517 | 0.182 | 0.302 | 0.215 |
| Mixtral | 0.646 | 0.034 | 0.320 | 0.327 |

Table 6: Preferences (by percentage) of different evaluators for answer positions. Column "Diff" is calculated by subtracting Second from First. Human-NF refers to human preference when the "not familiar" button is chosen. Differences that are smaller than 10% are highlighted by green, differences that are between 10% and 30% are noted as yellow. Results that are more than 30% are marked as red

From Table 6, it is evident that most evaluators exhibit some degree of positional preference, particularly GPT-3.5-Turbo, Spark, Qwen, Gemini-Pro and Mixtral, which demonstrate a strong positional preference in their choices. GPT-3.5-Turbo consistently favors the first answer, similar situations apply to Mixtral. Spark prefers the second answer, while Qwen and Gemini-Pro invariably selects Tie [3]. Additionally, Claude-3, Ernie, and LLaMA2-70B also show some positional bias, but to a less extent than the aforementioned models, with a preference difference of about 10% to 30% between the first and second answers. Human evaluators, human choices in not familiar scenarios, GPT-4, GPT-

[2]For human evaluators, first and second correspond to answers on the left and right, respectively.

[3]Based on this observation, we have excluded these three models from all other experiments.

4-Turbo, Claude-2, and PaLM-2 exhibit a smaller positional bias, with the preference difference between the first and second answers all within 10%.

### F.2 Verbosity Bias

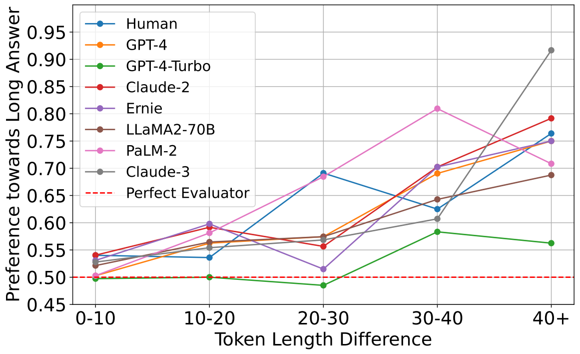


Figure 4: Verbosity Bias of different judges. The X-Axis indicates the absolute length difference between the long answer and the short answer. The Y-Axis indicates the preference towards the long answer. 0 refers to a total favor for the short answer, 0.5 indicates a neutral preference, and 1 indicates a total preference towards the long answer.

We conduct a statistical analysis of judges' verbosity preferences at the vote level [4]. Initially, we assign a value of 0 to votes favoring shorter answers, 0.5 to Tie votes, and 1 to votes favoring longer answers. Subsequently, we calculate the average value of votes based on the difference in answer length. Ideally, as depicted by the Perfect Evaluator in the figure, an evaluator's preference for length should consistently be 0.5.

From Figure 4, it is observable that as the difference in answer length increases, all evaluators exhibit a tendency to prefer longer answers to varying extents. GPT-4-Turbo's judgments are least influenced by length, whereas Claude-3 is most affected by length, and human evaluators also showing significant length bias. In the 0-10 length difference interval, the preferences of all evaluators are near 0.5, suggesting that when the length difference is minimal, the evaluators' length preference is not pronounced. However, as the length difference expands, all evaluators, including humans, demonstrate a preference for longer answers, and this preference intensifies with the growth in length difference. Excluding GPT-4-Turbo, when the length difference exceeds 40, the preference scores of all evaluators approach or surpass 0.7, indicating a pronounced length bias[5].

## G Revised Bloom's Taxonomy

The Revised Bloom's Taxonomy serves as a framework for categorizing educational goals, objectives, and standards. Our study applies this taxonomy to structure the design of questions to evaluate the nuanced bias in human evaluators and LLMs. This taxonomy differentiates cognitive processes into six ascending levels of complexity: remembering, understanding, applying, analyzing, evaluating, and creating. Our research chose this taxonomy as a guidance to create more diverse and cognitive-comprehensive questions.

## H User Interface

We show a screenshot of the user interface in Figure 5.

## I Supplementary Results of Deceiving Models

In Table 7, we show that the answer quality of GPT-3.5-Turbo is much higher than the that of the LLaMA2 family. This proves the validity of using LLaMA2's answers to form the weak set $W$.

| Judges | percentage of votes | |
|---|---|---|
| | LLaMA2-Chat Family | GPT-3.5-Turbo |
| GPT-4 | 0.08 | 0.73 |
| Claude-2 | 0.09 | 0.62 |
| Ernie | 0.07 | 0.70 |
| LLaMA2-Chat-70B | 0.08 | 0.65 |
| PaLM-2 | 0.07 | 0.70 |
| GPT-4-turbo | 0.08 | 0.45 |

Table 7: Percentage of votes of each judge for LLaMA2-Chat family and GPT-3.5-Turbo. Results for LLaMA2-Chat-{7B,13B,70B} are averaged. Tie votes account for the remaining percentages in each row.

[4]Lengths are computed using `tiktoken` library from OpenAI.

[5]To prevent the confounding of length bias with perturbation, we only show statistics on the control group.

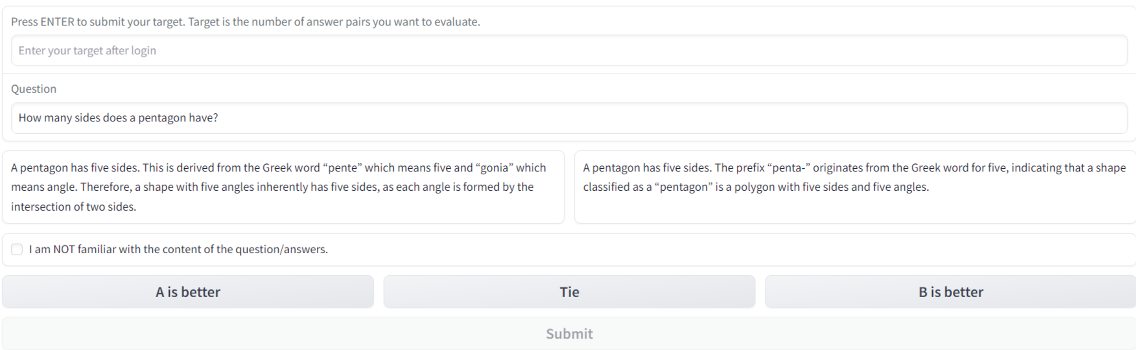


Figure 5: User Interface.