

Style Over Substance: Evaluation Biases for Large Language Models

Minghao Wu^{1,2*} Alham Fikri Aji¹

¹Mohamed bin Zayed University of Artificial Intelligence ²Monash University

{minghao.wu, alham.fikri}@mbzuai.ac.ae

Abstract

As large language models (LLMs) continue to advance, accurately and comprehensively evaluating their performance becomes increasingly challenging. Ranking the relative performance of LLMs based on Elo ratings, according to human or LLM judgment, is becoming more popular. However, the extent to which humans and LLMs are capable of serving as reliable evaluators remains uncertain. This study investigates the behavior of crowd-sourced and expert annotators, as well as LLMs, when comparing outputs from different models. To achieve this, we curate a dataset of intentionally flawed, machine-generated answers. Our findings reveal a concerning bias in the evaluation process: answers with factual errors are rated more favorably than answers that are too short or contain grammatical errors. To address this issue, we propose independently evaluating machine-generated text across multiple dimensions, rather than merging all evaluation aspects into a single score. We instantiate this idea with the Elo rating system, resulting in the Multi-Elo Rating System (MERS). Empirical results from our study reveal that this proposed approach significantly enhances the quality of LLM-based evaluations, particularly in terms of factual accuracy. However, there is no significant improvement in crowd-sourced evaluations, indicating the need for further investigation.

1 Introduction

Recent advancements in the field of natural language processing (NLP) have demonstrated that the utilization of supervised instruction fine-tuning and reinforcement learning from human feedback (RLHF) can yield substantial improvements in the performance of large language models (LLMs) with respect to their ability to comprehend and execute instructions (Ouyang et al., 2022; Wei et al.,

2022; Sanh et al., 2022; Chung et al., 2022; OpenAI, 2023; Wu et al., 2023; Li et al., 2023a; Lyu et al., 2023). This progress signifies a significant stride in the domain of language model development. However, the assessment of these enhanced LLMs presents a notable challenge, particularly when confronted with more generalized instructions that entail open-ended responses. Such instructions often lack a definitive metric for evaluation within the confines of traditional natural language processing benchmarks.

In response to this challenge, recent studies commonly adopt the *Elo rating system* (Elo, 1967) for evaluation. This approach involves enlisting either human or LLM evaluators to compare and rate the outputs generated by two different models (Askeel et al., 2021; Bai et al., 2022a; Srivastava et al., 2022; Dettmers et al., 2023; Zheng et al., 2023). This evaluation method enables the computation of an Elo-based leaderboard to rank the relative performance of LLMs (Chiang et al., 2023).¹ Nonetheless, an important question arises concerning the qualifications of human and LLM judges to serve as effective evaluators in this context. Evaluating model outputs encompasses a multifaceted decision-making process, and it remains an open question whether these judges possess the expertise to accurately determine the superior model output. Further research is needed to address this inquiry comprehensively and refine the evaluation procedures for enhanced LLMs.

In this study, we systematically generate a set of responses, considering factors such as language proficiency, factual accuracy, and response length, and conduct pairwise comparisons for compute Elo ratings.² We employ 40 general-purpose questions

¹<https://arena.lmsys.org/>

²In this work, we refer to various settings used for answering questions as different “models”, even though all answers are generated by GPT-4. Furthermore, we describe the pairwise comparison as a game/battle played between two models.

* work done while visiting MBZUAI

	Answer Features			Elo Ratings							
	# of words	Language Errors	# of Factual Errors	Human		GPT-4	Claude-1				
				Crowd	Expert						
Correct	≈ 100	N.A.	0	1091		1162		1482		1320	
+ Short	≈ 50	N.A.	0	970		1029		1096		1052	
One Minor Factual Error	≈ 100	N.A.	1, minor	1074		1137		1415		1265	
+ Short	≈ 50	N.A.	1, minor	1002		964		988		997	
Several Minor Factual Errors	≈ 100	N.A.	≈ 3, minor	1032		1024		1206		1182	
+ Short	≈ 50	N.A.	≈ 3, minor	952		873		851		891	
Several Major Factual Errors	≈ 100	N.A.	≈ 3, major	1025		892		861		979	
+ Short	≈ 50	N.A.	≈ 3, major	937		832		710		782	
Advanced Learner	≈ 100	Spelling	0	1041		1138		1213		1126	
+ Short	≈ 50	Spelling	0	941		986		824		841	
Intermediate Learner	≈ 100	Grammatical	0	1015		1108		771		904	
+ Short	≈ 50	Grammatical	0	921		855		582		662	

Table 1: Elo ratings for answers in different settings based on the annotations given by crowd-sourced annotators, expert annotators, GPT-4, and Claude-1.

sourced from Chiang et al. (2023) that do not require specialized expertise to ensure the generalization of our study and reduce annotation difficulty. The answers for these questions are generated by GPT-4 with specific instructions. To probe the potential impact of language proficiency towards human and LLM judgments, we instruct GPT-4 to emulate an advanced English learner, occasionally incorporating spelling errors, or an intermediate English learner, occasionally introducing grammatical mistakes during the response generation process. To probe factual accuracy, we direct GPT-4 to include varying degrees of factual errors in the responses. Lastly, we explore the influence of response length by instructing GPT-4 to generate answers of differing lengths. To ensure that the generated responses conformed to the desired criteria, we conduct manual reviews and carry out post-editing as necessary. After obtaining the collection of responses, we conduct annotations with a diverse pool of annotators, including crowd-sourced annotators, expert annotators, GPT-4, and Claude-1. We calculate the Elo ratings for each “model” based on their judgments to understand their preferences when deciding on high-quality model outputs.

In our investigation (Table 1 and Section 4), we uncover intriguing findings related to the assessment of answer quality. Firstly, we observe uncertainty in determining answer quality among humans, including both crowd-sourced and expert annotators, whereas LLMs exhibit greater certainty in their evaluations. Furthermore, we notice a substantial difference in fact-checking capabilities between humans and LLMs. Humans generally do not thor-

oughly fact-check answers unless the factual error is glaringly evident, whereas LLMs demonstrate some degree of fact-checking ability, albeit with imperfections. Another significant finding is that both humans and LLMs favor longer answers. Interestingly, crowd-sourced annotators appear to be (almost) unbiased in their evaluation of answer ordering, while LLMs lean toward the first answer and experts prefer the second answer.

We note that a unified measure is insufficient for evaluating LLMs. Furthermore, some aspects, such as factuality, are arguably more crucial than others. Yet, we observe judges preferring factually incorrect models over those that are grammatically incorrect or short. With these identified issues, we present the Multi-Elo Rating System (MERS) designed to assess machine-generated text from multiple dimensions. Our proposed approach requires human judges and LLMs to independently evaluate the machine-generated text from three aspects: “Accuracy”, “Helpfulness”, and “Language”. This allows us to achieve a more comprehensive and transparent understanding of the quality of model outputs. Note that our approach is not limited to these three dimensions and can be easily extended to more dimensions. Our empirical findings demonstrate a significant improvement in the evaluation quality of GPT-4, particularly regarding factual accuracy. However, we observe that humans still exhibit uncertainty in their assessments.

Based on our findings, we highly recommend that practitioners evaluate machine-generated text from various perspectives rather than depending solely on a single unified measure. Additionally,

we advise practitioners to exercise caution when using crowd-sourced annotators.

2 Related Work

Large Language Models Large Language Models (LLMs) commonly refer to Transformer-based language models with billions of parameters (Vaswani et al., 2017). Examples of these models include GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), BLOOM (Scao et al., 2022), and LLaMA (Touvron et al., 2023). These models demonstrate impressive abilities in understanding natural language and handling complex tasks. It is found that supervised fine-tuning and reinforcement learning from human feedback can further enhance their performance following general language instructions (Mishra et al., 2022; Wang et al., 2022b; Wei et al., 2022; Ouyang et al., 2022; Chung et al., 2022; Muennighoff et al., 2022; Wu et al., 2023). Therefore, accurately and comprehensively assessing the performance of these LLMs remains an unsolved challenge.

Evaluation NLP models are commonly assessed using standardized benchmark test suites. Several recent studies suggest using a diverse set of NLP benchmarks for LLM evaluation (Hendrycks et al., 2021; Liang et al., 2022; Li et al., 2023b). As pointed out by Gudibande et al. (2023) and Zheng et al. (2023) that there is a gap between users’ perception and standardized evaluation suites, recent LLM studies often incorporate human evaluation for a more nuanced understanding of model performance (Wang et al., 2022a; Chiang et al., 2023). As human evaluations can be costly, some recent works utilize state-of-the-art LLMs to evaluate model outputs (Wang et al., 2023b; Kim et al., 2023). More recently, several works employ the Elo rating system from chess games to gauge the LLMs’ capabilities (Bai et al., 2022a; Srivastava et al., 2022; Dettmers et al., 2023; Zheng et al., 2023). However, these previous works operate under the assumption that human evaluations serve as the gold standard. In contrast, Clark et al. (2021) demonstrate that humans are not able to distinguish between human-written and machine-generated text. We investigate the judgments of humans and LLMs in evaluating text across multiple aspects.

3 Evaluation Method

In this section, we cover the Elo rating system, methods for generating incorrect answers for mod-

els, and evaluations for humans and LLMs.

3.1 Elo Rating System

The Elo rating system is a method used to calculate the relative skill levels of players in two-player games, such as chess. Given two players \mathcal{A} and \mathcal{B} whose Elo ratings are \mathcal{R}_A and \mathcal{R}_B respectively, the expected scores (winning probabilities) for these two players are:

$$\mathcal{E}_A = \frac{1}{1 + 10^{\frac{\mathcal{R}_B - \mathcal{R}_A}{400}}}, \text{ and } \mathcal{E}_B = \frac{1}{1 + 10^{\frac{\mathcal{R}_A - \mathcal{R}_B}{400}}}. \quad (1)$$

Suppose the player \mathcal{A} is expected to obtain \mathcal{E}_A scores from this game but actually get \mathcal{S}_A scores, the updated Elo rating of player \mathcal{A} is:

$$\mathcal{R}'_A = \mathcal{R}_A + \mathcal{K} \cdot (\mathcal{S}_A - \mathcal{E}_A), \quad (2)$$

where \mathcal{K} is adjustment parameter, called the \mathcal{K} -factor, and \mathcal{S}_A is the actual battle result. Equation 2 is also used for obtaining \mathcal{R}'_B . Following Chiang et al. (2023), we set $\mathcal{K} = 32$. Additionally, if \mathcal{A} is better than \mathcal{B} , we set $\mathcal{S}_A = 1$ and $\mathcal{S}_B = 0$. If \mathcal{B} is better than \mathcal{A} , we set $\mathcal{S}_A = 0$ and $\mathcal{S}_B = 1$. We set both $\mathcal{S}_A = 0.5$ and $\mathcal{S}_B = 0.5$ if both players are equally good. For instance, in a game featuring the player \mathcal{A} with an Elo rating of 1200 and the player \mathcal{B} with a rating of 900, the player \mathcal{A} is expected to win with a probability of 84.9%. Moreover, to minimize the influence of the game orders, the Elo rating calculation is performed $10K$ times with random orders following Dettmers et al. (2023).

3.2 Answer Generation

Chiang et al. (2023) release a set of 80 questions categorized into 8 categories to evaluate the generation capability of LLMs.³ This dataset is widely used by recent studies (Dettmers et al., 2023; Li et al., 2023a), due to its high quality. However, due to the requirement for specialized expertise to answer some of these questions, we exclude the “fermi”, “coding”, and “math” questions, as they typically demand extra efforts to evaluate the answers. This step aims to reduce the potential impact of varying human raters’ capabilities on the evaluation process. Additionally, we also remove the “roleplay” and “writing” questions, as they involve creative writing and are prone to subjectivity in human assessment. As a result, our

³<https://github.com/lm-sys/vicuna-blog-eval/blob/main/eval/table/question.jsonl>

Model type	Output
Correct	Effective ways to deal with stress include regular exercise, maintaining a healthy diet, getting enough sleep, and taking time for relaxation activities like meditation. It’s also important to have a support system in place, whether that’s friends, family, or a mental health professional.
Minor Errors	Effective ways to deal with stress include physical activities like playing tennis, doing yoga, or running a marathon . Also, adopt healthy habits like sleeping for 10 hours , eating three full meals a day, and avoiding caffeine. Lastly, meditation, guided by a professional like John Zen , can be very soothing.
Major Errors	Some effective ways to deal with stress include drinking 10 cups of coffee a day , binge-watching Netflix for at least 5 hours , and attending Dr. Monty Farquhar’s stress-relief seminars. Additionally, practicing the art of Floga, a combination of Flute-playing and Yoga , can significantly reduce stress levels.
Advanced Learner	Whenevr feeling stressed, always remember to take deep breathes . This can calm nerves. Regular physical exercise also reduce stress hormonal levels. A balance diet promoting overral well-being can lead to up healthy stress response. Finally, finding hobby or activity you enjoy can help distract you from stressful situations.
Intermediate Learner	Deal with stress , you can do exercise regularly, practicing a meditation , get with plenty sleep , and eat healthy foods also . You can too to connect with others so express your feelings, and avoiding caffeine, alcohol, and nicotine, and take time to relax and have fun.

Table 2: Examples of different error types for the question “What are the most effective ways to deal with stress?”. Factual errors are **highlighted in red** and language errors are **highlighted in orange**.

final question set consists of 40 questions, focusing on the “generic”, “knowledge”, “common sense”, and “counterfactual” categories. We believe these retained questions can be easily understood and answered by the general public.

Once we have the set of questions, we require GPT-4 to generate answers with specific error types in addition to providing the correct answers. We provide some examples of these error types in [Table 2](#). Regarding language quality errors, we ask GPT-4 to respond as either an advanced English learner or an intermediate English learner. The answers generated by an advanced English learner may occasionally contain spelling errors, while those from an intermediate English learner commonly include grammatical mistakes. In terms of factual accuracy, we expect GPT-4 to produce answers with either minor or major errors. Minor errors primarily involve fabricated names or incorrect numbers, while major errors contain incorrect facts and suggestions. Furthermore, we utilize GPT-4 to generate both long (approximately 100 words) and short (approximately 50 words) answers for each question to investigate the preference of both humans and LLMs regarding answer length. Hence, there are 12 models (settings) in total in this study. We present the prompts used for answer generation in [Appendix A](#). To ensure unbiased evaluation results regarding answer ordering, all evaluators, including crowd-sourced human annotators, expert annotators, and LLMs, evaluate all answer pairs from both forward and reversed directions. **In the end, we have 5280 unique pairwise comparisons across all 12 models and 40 questions.**⁴ During

⁴The total number of pairwise comparisons (5280) is derived by multiplying three components: First, with 12 models,

the generation process, we notice that GPT-4 may not always fulfill our requirements. For instance, the output sometimes may not contain any factual or language errors. Therefore, we manually review all the answers and make necessary edits to ensure they align with our requirements.

3.3 Crowd-Sourced Evaluation

Crowd-sourced human annotators are commonly used to assess the quality of machine-generated text. For this study, we utilize Amazon Mechanical Turk (AMT) to collect text evaluations from human raters in NLG evaluations. To minimize the potential impact of annotator demographics, we only recruit crowd-sourced human annotators from the United States. We also ensure the quality of annotations by exclusively offering tasks to annotators who have completed over 5,000 assignments and maintain a success rate of 99% or higher. To ensure the diversity of the annotators, we limit each annotator to provide no more than 20 annotations. As a result, we engage 377 unique annotators, each contributing an average of 14 annotations.

During annotation, we provide a simple interface that mainly follows the guidelines of [Dettmers et al. \(2023\)](#), as shown in [Appendix C](#). Annotators are presented with a question and two model-generated responses placed side-by-side. Their task is to select the better output or indicate a tie between them. To ensure the annotators’ attentiveness and thorough reading of the responses, we incorporate a mandatory 20-second delay before they can submit their answers. Furthermore, we anonymize the

there are $\binom{12}{2} = 66$ possible unique pairs. Each pair is compared in both directions (A vs B and B vs A), so we multiply by 2. Finally, this is done for each of the 40 questions, resulting in $66 \times 2 \times 40 = 5280$ total unique pairwise comparisons.

```

[Question]
What are the most effective ways to deal with stress?

[The Start of Assistant 1's Answer]
Effective ways to deal with stress include ...
[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]
Whenever feeling stressed, ...
[The End of Assistant 2's Answer]

[System]
We would like to request your feedback on the performance of two AI
↔ assistants ...

```

Figure 1: A simplified example for the evaluation prompt employed by GPT-4 and Claude-1 is designed to assess the quality of responses. A complete example can be found in Figure 10.

model name from our human annotators.

3.4 Expert Evaluation

To address concerns about the reliability of crowd-sourced annotators, we have implemented a parallel system involving a team of 20 experts, each holding at least a master’s degree from English-speaking countries. We organize two tutorials on the annotation criteria to ensure a comprehensive understanding of the language and proficiency in the nuances of annotation. It is important to note that employing expert annotators comes at a higher cost than using crowd-sourced alternatives. Therefore, we sample 200 games out of a total pool of 5280 games, which is statistically sufficient to yield meaningful insights into the model’s performance.⁵ To ensure a fair comparison, we provide the same annotation instructions to both the expert team and the crowd-sourced participants. Each expert is assigned to annotate 10 games, which takes approximately 20 minutes. This approach enables us to accurately evaluate the effectiveness and reliability of each annotation method.

3.5 LLM Evaluation

As human evaluation can be costly and inefficient, there is an increase in the use of advanced LLMs, such as GPT-4, to evaluate model outputs. In our work, we also use LLMs as judges to assess answer quality. However, previous studies rely solely on GPT-4 as the LLM judge (Chiang et al., 2023; Li et al., 2023a; Zheng et al., 2023), which may not be appropriate for our work as our answers are

⁵The US Chess Federation believes that 25 games are enough to assess a player’s ability, and in our sample of 200 games, each model participated in at least 28 games. Source: <https://new.uschess.org/frequently-asked-questions-member-services-area>

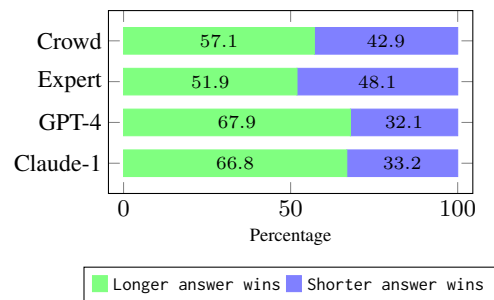


Figure 2: The percentage distribution of decision choices made by humans and LLMs in terms of answer length (in words).

generated by GPT-4. This raises concerns about potential biases that GPT-4 may have towards its own outputs, which could skew the evaluation process. To ensure fair comparisons and mitigate any bias, we introduce Claude-1 from Anthropic (Bai et al., 2022b) as an additional LLM judge, in addition to GPT-4.⁶ By incorporating multiple LLM judges, we can establish a more comprehensive and unbiased assessment of the generated answers in our study. We utilize the evaluation prompt from Dettmers et al. (2023), as presented in Figure 1. The prompt assesses the answers based on their helpfulness, relevance, accuracy, and level of detail, while also aiming to avoid bias related to answer ordering.

4 Analysis

In this study, we assess the performance of 12 models using Elo ratings, as evaluated by crowd-sourced annotators, expert annotators, GPT-4, and Claude-1. The Elo ratings are presented in Table 1. In this section, we delve into a comprehensive analysis of these distinctions.

All the judges exhibit a bias toward longer texts.

Text evaluation by both human and LLM judges often exhibits a bias towards longer responses, where GPT-4 demonstrates the most bias and the expert annotators demonstrate the least bias, as depicted in Figure 2. This inclination is expected since one of the scoring criteria requested is “the level of detail”, which often correlates with the length of the output. However, such an inclination is arguably undesirable. For example, we observe that GPT-4 considers “Several Minor Factual Errors” (1206 Elo) to be better than “Correct + Short” (1096 Elo). When manually reviewing the justifications

⁶GPT-4 signature: gpt-4-0613 and Claude-1 signature: Claude-1.3

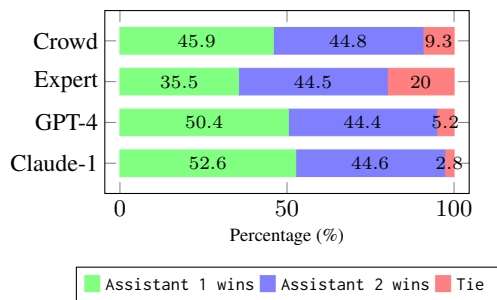


Figure 3: The distribution of decision choices made by humans and LLMs.

provided by GPT-4 for its evaluation decisions, we observe that GPT-4 sometimes considers the response of “Several Major Factual Errors” as superior to that of “Correct + Short”, even when factual errors are detected. We present an example in [Appendix D](#).

Humans are indecisive. Unlike LLM judges, both expert and crowd-sourced annotators demonstrate indecisiveness, leading to Elo scores that remain relatively close to the initial value of 1000. The Elo scores assigned by crowd-sourced annotators range from 926 to 1086, while those by expert annotators range from 832 to 1162. These ranges are significantly narrower than those observed for GPT-4 and Claude-1. Interestingly, human judges are more inclined to assign “tie” judgments compared to LLMs, as illustrated in [Figure 3](#). These findings raise important questions about the level of reading comprehension and attention to detail demonstrated by humans, especially when they are crowd-sourced workers evaluating answers.

The order of answers affects the judges’ decisions. Our analysis also reveals an interesting bias toward the order of answers, which is evident in the assessments made by both human judges and LLM judges. As illustrated in [Figure 3](#), it can be observed that the crowd-sourced annotators, GPT-4, and Claude-1 exhibit a distinct and notable preference for Assistant 1. This intriguing finding is consistent with prior research ([Wang et al., 2023a](#)), further underscoring the significance of this bias. Interestingly, the expert annotators favor the answer given by Assistant 2 and are more likely to assign “tie” judgments. To ensure fairness and comprehensive evaluation, we strongly recommend considering both answer orders when analyzing the performance of the systems. We leave the further study of the behavior of humans for future work.

	Crowd	Expert	GPT-4	Claude-1
Crowd	—	0.08	0.11	0.10
Expert	0.08	—	0.09	0.14
GPT-4	0.11	0.09	—	0.51
Claude-1	0.10	0.14	0.51	—

Table 3: Pairwise inter-annotator agreement measured by Cohen’s kappa coefficient κ . The pairs involving the experts only cover 200 games.

Crowd-sourced annotators lack fact-checking, while experts and LLMs can fact-check, albeit imperfectly. The close Elo scores of those models with factual errors from crowd-sourced annotators in [Table 1](#) suggest inadequate fact-checking capability of humans, raising concerns about false information and malicious intent in LLM-generated outputs. People are vulnerable to believing such outputs, particularly when they appear convincing. Relatedly, instances are found where humans overlooked basic verification tasks, like identifying fake URLs or citations ([Nakov et al., 2021](#)),⁷. Although experts are more diligent in fact-checking, the general public’s behavior towards LLM output tends to be closer to that of crowd-sourced annotators, posing significant safety risks. On the other hand, LLMs do notice errors, but not consistently. When an LLM fails to detect inaccuracies, it often favors flawed outputs over shorter or grammatically incorrect responses.

LLMs only reach a moderate consensus, while humans embrace more diversity in thought. We assess inter-annotator agreement using Cohen’s kappa coefficient κ ([Cohen, 1960](#)) and present the results in [Table 3](#). Our analysis, following the interpretation of κ by [McHugh \(2012\)](#), shows that only GPT-4 and Claude-1 achieve moderate agreement, while other comparisons demonstrate only slight agreement. This confirms that human annotators exhibit inconsistencies during annotation. [Zheng et al. \(2023\)](#) define the agreement between two types of judges as the probability of non-identical individuals of each type agreeing on a randomly selected question and shows an approximately 80% agreement between LLMs and crowd-sourced humans, which seems to contradict our findings. However, this discrepancy arises because they remove ties and inconsistent annotations, while we consider all annotations. When ties and inconsisten-

⁷<https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>

	Accuracy			Helpfulness			Language		
	Crowd	Expert	GPT-4	Crowd	Expert	GPT-4	Crowd	Expert	GPT-4
C.	1056	1180	1200	1045	1208	1384	1036	1109	1415
+ S.	963	1040	1158	983	979	1009	1007	1068	1199
OMin.FE	1026	1090	1120	1048	1153	1378	1019	1114	1334
+ S.	978	898	1016	993	941	965	990	1012	1109
SMin.F	1036	1044	993	1051	1069	1248	1029	1096	1200
+ S.	978	931	857	956	865	845	996	935	988
SMaj.FE	1030	963	794	1037	1015	926	1023	1010	995
+ S.	955	787	746	940	766	726	982	879	871
AL	1028	1121	1139	1032	1146	1196	1004	1039	1051
+ S.	979	971	1051	969	891	804	994	863	814
IL	1015	1076	1018	1002	1095	908	991	992	560
+ S.	956	898	908	945	872	612	930	884	465

Table 4: Elo ratings for different models with regard to “Accuracy”, “Helpfulness”, and “Language” given by crowd-sourced annotators, expert annotators, and GPT-4. C. stands for Correct. OMin.FE stands for One Minor Factual Error. SMin.F stands for Several Minor Factual Errors. SMaj.FE stands for Several Major Factual Errors. AL stands for Advanced Learner. IL stands for Intermediate Learner. S. stands for Short.

cies are retained, Zheng et al. (2023) report an approximately 60% agreement between LLMs and crowd-sourced humans, which is slightly better than random guessing and aligns with our findings.

5 Multi-Elo Rating System

As discussed in Section 4, we identify the limitations in the evaluation of machine-generated answers based on humans and LLMs. We observe that the current widely used evaluation approach fails to yield satisfactory results and lacks a comprehensive understanding of the factors influencing the outcomes. Collapsing multiple decision components (e.g., accuracy, level of details, relevance, language, helpfulness, etc.) into a single score undermines the importance of individual components and introduces subjectivity regarding their relative significance in the final decision.

To overcome these limitations, we propose a novel multi-dimensional evaluation approach for assessing the outputs of LLMs, which we refer to as the Multi-Elo Rating System (MERS). This approach is inspired by machine translation research, where evaluations are often divided into at least two parts: fluency and adequacy. Recent advancements in MT also advocate breaking down the evaluation into multiple aspects (Burchardt, 2013). In this section, we present a detailed description of our approach and the results.

5.1 Methodology

The Multidimensional Quality Metrics (MQM) framework provides a comprehensive approach for evaluating and establishing standards for transla-

tion quality (Burchardt, 2013). Drawing inspiration from this framework, we propose a similar approach to evaluate the outputs of LLMs from multiple dimensions. Our evaluation focuses on three dimensions of the generated text, as follows:

- **Accuracy:** The accuracy of text involves considering factual correctness and logical consistency.
- **Helpfulness:** The helpfulness of the text involves considering its relevance of the information and whether it addresses the question given, taking into account the depth of the response given.
- **Language:** The language of the text involves considering its clarity, coherence, grammar, syntax, and tone.

Note that our approach is not limited to these dimensions and can be easily extended.

The quality of an answer is dependent on its specific context. For instance, if a model gives a detailed but complicated explanation of black holes to an 8-year-old, the answer may be accurate but not useful. Conversely, if a model is asked to compose an email and produces a message with incorrect information, the response may lack accuracy but still have some value. By taking this multi-dimensional approach, we can gain a clearer understanding of model performance and prioritize different aspects based on our individual requirements.

To facilitate the multi-dimensional evaluation by humans, we introduce a simple modification to the interface, asking them to rate the quality across the three different aspects, as shown in Appendix C. Additionally, we experiment with two approaches for GPT-4 evaluation: asking three independent queries versus a single query that requests judg-

	Accuracy		Helpfulness		Language	
	Separate	Compound	Separate	Compound	Separate	Compound
Correct	1200	1284	1384	1429	1415	1429
+ Short	1158	1146	1009	1054	1199	1178
One Minor Factual Error	1120	1221	1378	1399	1334	1346
+ Short	1016	1045	965	993	1109	1090
Several Minor Factual Errors	993	1054	1248	1208	1200	1187
+ Short	857	895	845	833	988	956
Several Major Factual Errors	794	805	926	884	995	968
+ Short	746	730	726	711	871	842
Advanced Learner	1139	1178	1196	1210	1051	1093
+ Short	1051	969	804	810	814	839
Intermediate Learner	1018	911	908	853	560	565
+ Short	908	761	612	615	465	506

Table 5: The Elo ratings for different models with regard to “Accuracy”, “Helpfulness”, and “Language” given by GPT-4. “Separate” means that GPT-4 assesses the factual accuracy of the models using a separate prompt, while “Compound” implies that GPT-4 evaluates all three dimensions simultaneously using a compound prompt.

ments for all three aspects together. In this paper, we report the results given by asking three independent queries.

5.2 Stop Using Crowd-Sourced Annotators!

In this section, we compare the annotation outcomes provided by both crowd-sourced and experts, as presented in Table 4. Regarding “Accuracy”, we find that experts are proficient in identifying factual errors in answers, although not entirely flawless. However, crowd-sourced annotators exhibit indecisiveness in their evaluations. Notably, the crowd-sourced annotators perceive Several Major Factual Errors + Short (955 Elo) and Correct + Short (963 Elo) as nearly equally good. Regarding “Helpfulness”, the experts display a stronger preference for longer answers, while the crowd-sourced annotators only slightly favor them, as evidenced by their Elo scores. In terms of “Language”, both expert and crowd-sourced annotators face challenges in recognizing spelling or grammatical errors, suggesting that humans may be less sensitive to language errors. Overall, the experts outperform the crowd-sourced annotators in the evaluation, despite not being entirely error-free themselves. These results serve as a warning against over-reliance on crowd-sourced judgments and highlight concerns regarding the general audience’s ability to critically interpret LLM’s output.

5.3 Experts versus GPT-4

In this section, we discuss the difference between the experts and GPT-4 in evaluation from multiple dimensions and present the results in Table 4. Regarding the “Accuracy” dimension, it is noteworthy

that the Elo scores for factual accuracy closely align with the single Elo scores presented in Table 1, suggesting that experts indeed prioritize factual accuracy during evaluation. GPT-4 can also effectively rank models based on the severity of the errors. Regarding the “Helpfulness” dimension, both experts and GPT-4 consistently consider longer answers to be more helpful. Similar to the discussion in Section 5.2, we believe that this preference stems from the strong correlation between “helpfulness” and “the level of detail”, as longer answers tend to convey more information, making them perceived as more helpful. Regarding the “Language” dimension, recognizing spelling or grammatical errors in text is challenging for experts, while GPT-4 effectively distinguishes between answers based on their language quality and appears to penalize grammatical errors more heavily during assessment. Overall, this comprehensive analysis sheds light on the evaluation process and reveals the differing perspectives of experts and GPT-4 in assessing various aspects of model outputs.

5.4 Separate versus Compound Prompt

In this section, we explore two evaluation strategies: assessing each dimension separately or evaluating all dimensions simultaneously using a compound prompt. The results obtained from GPT-4 using these two strategies are presented in Table 5.

Regarding the “Accuracy” dimension, our findings indicate that GPT-4 performs better when assessing the factual accuracy of models independently. However, when using the compound prompt, GPT-4 ranks “One Minor Factual Error” and “Advanced Learner” higher than “Correct

+ Short”. This observation leads us to hypothesize that evaluation dimensions can mutually influence each other when evaluated concurrently, even when explicitly instructing GPT-4 to evaluate each dimension independently. For the “Helpfulness” dimension, GPT-4, when using the separate prompt, ranks “Several Minor Factual Errors” higher than “Advanced Learner”. However, when using the compound prompt, GPT-4 ranks “Advanced Learner” higher than “Several Minor Factual Errors”. Interestingly, the “Language” dimension is the most consistent, as GPT-4 produces the same rankings using both evaluation strategies.

Based on our findings, we choose to use the separate prompts for each dimension in this work, as this strategy yields better results in terms of factual accuracy.

6 Discussion

When evaluating the quality of a Language Model (LLM), it is essential to recognize that a single, aggregated score may not adequately capture the model’s multifaceted performance. The absence of consensus on which specific factors should take precedence suggests the need for a more nuanced approach. We propose examining three distinct factors individually to achieve a comprehensive assessment. While this categorization enhances our understanding, it is important to acknowledge that it might not encompass all elements necessary for an ideal response. This highlights the necessity for ongoing research, as the field continues to evolve and refine its evaluation standards.

Another critical aspect of LLM evaluation is the incorporation of human judgments. Crowd-sourced feedback can offer valuable insights into the general audience’s perception of the model’s output. On the other hand, expert evaluators are typically more adept at distinguishing between accurate and inaccurate information. Nevertheless, employing experts presents challenges, particularly in terms of scalability and resource allocation. Furthermore, both crowd-sourced and expert human annotators are subject to various biases. Such biases can skew evaluations and do not always reflect the true quality of the LLM’s output. To mitigate these biases and achieve a more balanced assessment, we encourage the practitioners to integrate multiple evaluation strategies.

7 Conclusion

This study investigates the limitations of humans and large language models (LLMs) when evaluating machine-generated text. We introduce intentional factual and grammatical errors into these texts and analyze the responses from crowd-sourced annotators, experts, and LLMs. Our findings reveal that both humans and LLMs exhibit various biases during evaluation. To address these issues, we propose evaluating text across multiple dimensions using the Elo rating system, leading to the development of the Multi-Elo Rating System. This method significantly enhances GPT-4’s evaluation quality, especially in terms of factual accuracy. Despite these improvements, crowd judges often remain indecisive, tend to favor longer responses, and have limited fact-checking capabilities. Therefore, we recommend a multi-dimensional evaluation approach for assessing machine-generated text rather than relying on a single metric, and advise caution when using crowd annotators for LLM evaluations.

8 Limitations

Question Coverage We select only 40 questions from Chiang et al. (2023), mainly due to explosive annotation cost. We acknowledge that this limited selection may not capture the full spectrum of question types and variations. Consequently, there is a potential risk that some aspects of the research question may not receive sufficient representation or exploration.

Evaluation Dimension Coverage In our proposed Multi-Elo Rating System, we only explore three crucial dimensions: “Accuracy”, “Helpfulness”, and “Language”. We acknowledge that while these dimensions provide valuable insights, they may not encompass the entirety of the multifaceted nature of text evaluation. Furthermore, it is important to recognize that our definitions for the three dimensions we have chosen are not infallible. Different stakeholders may have diverse perspectives on these dimensions.

We leave the investigation of addressing these limitations to future work.

References

Amanda Askill, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson

- Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. [A general language assistant as a laboratory for alignment](#). *CoRR*, abs/2112.00861.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *CoRR*, abs/2204.05862.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. [Constitutional AI: harmfulness from AI feedback](#). *CoRR*, abs/2212.08073.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Aljoscha Burchardt. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- 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](#).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *CoRR*, abs/2204.02311.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *CoRR*, abs/2210.11416.
- 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](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *CoRR*, abs/2305.14314.
- Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247.
- Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. [The false promise of imitating proprietary llms](#). *CoRR*, abs/2305.15717.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. [Prometheus: Inducing fine-grained evaluation capability in language models](#). *CoRR*, abs/2310.08491.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. [Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation](#). *CoRR*, abs/2305.15011.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. [CMMLU: measuring massive multitask language understanding in chinese](#). *CoRR*, abs/2306.09212.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. [Holistic evaluation of language models](#). *CoRR*, abs/2211.09110.
- Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. [Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration](#). *CoRR*, abs/2306.09093.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Swaroop Mishra, Daniel Khoshabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. [Crosslingual generalization through multitask finetuning](#). *CoRR*, abs/2211.01786.
- Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. [Automated fact-checking for assisting human fact-checkers](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, 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](#). In *Advances in Neural Information Processing Systems*.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. [Multi-task prompted training enables zero-shot task generalization](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. [BLOOM: A 176b-parameter open-access multilingual language model](#). *CoRR*, abs/2211.05100.

- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokkandov, Ashish Sabharwal, Austin Herick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.](#) *CoRR*, abs/2206.04615.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models.](#) *CoRR*, abs/2302.13971.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need.](#) In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators.](#) *CoRR*, abs/2305.17926.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. [Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization.](#) *CoRR*, abs/2306.05087.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022a. [Self-instruct: Aligning language model with self generated instructions.](#) *CoRR*, abs/2212.10560.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners.](#) In *International Conference on Learning Representations*.
- Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. [Lamini-lm: A diverse herd of distilled models from large-scale instructions.](#) *CoRR*, abs/2304.14402.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#) *CoRR*, abs/2306.05685.

```

Question/Instruction:
$instruction

Answer the question/instruction.
The answer should be roughly 100 words long.

```

Figure 4: The prompt employed by GPT-4 for generating answers of the “Correct” model.

```

Question/Instruction:
...

Answer the question/instruction.
The answer should be roughly 100 words long.
The answer must contain one minor factual error.
The factual error can be made-up names, wrong
↪ numbers, incorrect facts, or incorrect
↪ suggestions.
List the error and its corresponding
↪ justification separately.
Enclose your answer within <answer> and
↪ </answer> tags.
Enclose the error and justification within
↪ <error> and </error> tags.

```

Figure 5: The prompt employed by GPT-4 for generating answers of the “One Minor Factual Error” model.

A Answer Generation Prompts

We present the answer generation prompt for “Correct” (Figure 4), “One Minor Factual Error” (Figure 5), “Several Minor Factual Errors” (Figure 6), “Several Major Factual Errors” (Figure 7), “Advanced Learner” (Figure 8), and “Intermediate Learner” (Figure 9) in this section. For those prompts generating short answers, we simply require GPT-4 to ensure that “The answer should be roughly 50 words long”.

B LLM Evaluation Prompt

We utilize the evaluation prompt for LLMs from Dettmers et al. (2023), as presented in Figure 10.

C Human Evaluation Interface

The interface used for crowd-source evaluation is shown in Figure 12 and Figure 13.

D LLM Evaluation Case Study

We present a concrete example query for GPT-4 and Claude-1 in Figure 10. For such a query, the answer given by Assistant 1 is detailed but contains minor factual errors (Several Minor Factual Errors), while the answer given by Assistant 2 is

```

Question/Instruction:
...

Answer the question/instruction.
The answer should be roughly 100 words long.
The answer must contain several minor factual
↪ errors.
The factual errors can be made-up names, wrong
↪ numbers, incorrect facts, or incorrect
↪ suggestions.
List the errors and their corresponding
↪ justifications separately.
Enclose your answer within <answer> and
↪ </answer> tags.
Enclose the errors and justifications within
↪ <error> and </error> tags.

```

Figure 6: The prompt employed by GPT-4 for generating answers of the “Several Minor Factual Errors” model.

```

Question/Instruction:
...

Answer the question/instruction.
The answer should be roughly 100 words long.
The answer must contain several major factual
↪ errors.
The factual errors can be made-up names, wrong
↪ numbers, incorrect facts, or incorrect
↪ suggestions.
List the errors and their corresponding
↪ justifications separately.
Enclose your answer within <answer> and
↪ </answer> tags.
Enclose the errors and justifications within
↪ <error> and </error> tags.

```

Figure 7: The prompt employed by GPT-4 for generating answers of the “Several Major Factual Errors” model.

```

Question/Instruction:
...

Answer the question/instruction.
The answer must be written as if you're an
↪ advanced-level English learner.
The answer must contain 2 or 3 minor
↪ grammatical and spelling errors.
The answer should be roughly 100 words long.
List the errors and their corresponding
↪ justifications separately.
Enclose your answer within <answer> and
↪ </answer> tags.
Enclose the errors and justifications within
↪ <error> and </error> tags.

```

Figure 8: The prompt employed by GPT-4 for generating answers of the “Advanced Learner” model.

```
Question/Instruction:
...

Answer the question/instruction.
The answer must be written as if you're an
↪ intermediate-level English learner.
The answer must contain 5 or more major
↪ grammatical and fluency errors.
The answer must be roughly 100 words long.
List the errors and their corresponding
↪ justifications separately.
Enclose your answer within <answer> and
↪ </answer> tags.
Enclose the errors and justifications within
↪ <error> and </error> tags.
```

Figure 9: The prompt employed by GPT-4 for generating answers of the “Intermediate Learner” model.

correct and concise yet contains a few spelling errors (Advanced Learner+ Short). We present the response of GPT-4 to this query in [Figure 11](#). Even though GPT-4 identifies the concept “Ollivander Method” as hallucinating, it still favors Assistant 1’s answer.

[Question]
 What are the most effective strategies for conflict resolution in the workplace?

[The Start of Assistant 1's Answer]
 Some effective strategies for conflict resolution in the workplace include diligent communication,
 ↳ effective engagement, practicing empathy, implementing the **Ollivander Method**, and applying the
 ↳ Watson 5-step resolution technique. Effective communication is vital, which means actively
 ↳ listening to others, simplifying the complex issues, and frequently updating everyone involved.
 ↳ Engage all parties constructively and aim for consensus-based solutions. Empathy is key, try to
 ↳ understand the positions of the others. **The Ollivander Method encourages finding a magical
 ↳ solution by performing a trust-building ritual.** Lastly, Watson 5-step resolution technique
 ↳ preaches recognizing the issue, understanding different perspectives, finding solutions,
 ↳ executing the solutions and reviewing of these actions.

[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]
 Effectives strategies for resolving conflicts in work place includes: Active listening, this
 ↳ encourages the feeling of worth. One can also differing emotions to minimize encounters.
 ↳ Maintaining a calm demeanor help defuse tension. Finally, finding common grounds and using them
 ↳ to build agreements is very important.

[The End of Assistant 2's Answer]

[System]
 We would like to request your feedback on the performance of two AI assistants in response to the
 ↳ user question displayed above.
 Please rate the helpfulness, relevance, accuracy, level of details of their responses. First,
 ↳ provide your evaluation of the assistant's helpfulness, relevance, accuracy, and level of detail.
 ↳ Please provide a comprehensive explanation of your evaluation, avoiding any potential bias and
 ↳ ensuring that the order in which the responses were presented does not affect your judgment.
 Once you have carefully reviewed both submissions, in a new line, choose between the answers of
 ↳ Assistant 1 and Assistant 2 by outputting the number 1 or 2 respectively, or choose 3 if the two
 ↳ assistants are equivalent. Do not output anything else other than the number in this last line.

Figure 10: A complete example for the evaluation prompt employed by GPT-4 and Claude-1 is designed to assess the quality of responses. These language models evaluate answers based on criteria such as helpfulness, relevance, accuracy, and level of detail. The first Assistant is based on long yet factually incorrect model, whereas the latter is correct but concise with some grammatical errors.

Assistant 1's answer was quite helpful, relevant, accurate, and detailed. The strategies they've
 ↳ mentioned - communication, engagement, empathy, and a 5-step resolution technique, are all
 ↳ practical methods used in conflict resolution. **However, the reference to the "Ollivander Method"
 ↳ was a bit confusing** and didn't seem to be immediately applicable or clearly explained which
 ↳ mildly affects the accuracy. **It might be a misunderstanding or a typographical error due to it
 ↳ being associated with 'Harry Potter' and not usually in context of conflict resolution.**

On the other hand, Assistant 2's response was also helpful and relevant, but not as detailed as
 ↳ Assistant 1. They did mention key strategies like active listening, managing emotions,
 ↳ maintaining calm demeanor, and finding common ground, which are indeed effective strategies.
 ↳ However, these points could have been elaborated a bit more, making it more helpful for someone
 ↳ seeking advice on this topic.

**Accuracy was maintained by both in terms of the topic, though the level of detail varied. Assistant
 ↳ 1 provided more in-depth strategies while Assistant 2 kept it short yet relevant.**

1

Figure 11: The response given by GPT-4 to the query presented in Figure 10 where it prefers longer yet incorrect answer.

Task

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed below.

Please consider the helpfulness, relevance, accuracy, level of details of their responses.

Please choose the best response, avoiding any potential bias, ensuring that the order in which the responses were presented does not affect your judgment.

User Question

QUESTION

Response A

ANSWER_A

Response B

ANSWER_B

Response Comparison

- Response A is better than Response B.
- Response B is better than Response A.
- The responses are EXACTLY the same in quality.

Figure 12: Annotation interface for single Elo score.

Task

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed below. We would like to evaluate them across three aspects **independently**: accuracy, helpfulness, and language.

- The accuracy of the text involves considering factual correctness and logical consistency.
- The helpfulness of the text involves considering its relevance of the information and whether it addresses the question given, taking into account the depth of the response given.
- The language of the text involves considering its clarity, coherence, grammar, syntax, and tone.

Please choose the best response, avoiding any potential bias, ensuring that the order in which the responses were presented does not affect your judgment.

User Question

QUESTION

Response A

ANSWER_A

Response B

ANSWER_B

Response Comparison

Accuracy

The accuracy of the text involves considering factual correctness and logical consistency.

- Response A is more accurate than Response B.
- Response B is more accurate than Response A.
- The responses are EXACTLY the same in accuracy.

Helpfulness

The helpfulness of the text involves considering its relevance of the information and whether it addresses the question given, taking into account the depth of the response given.

- Response A is more helpful than Response B.
- Response B is more helpful than Response A.
- The responses are EXACTLY the same in helpfulness.

Language

The language of the text involves considering its clarity, coherence, grammar, syntax, and tone.

- Response A's language is better than Response B.
- Response B's language is better than Response A.
- The responses' language are EXACTLY the same in quality.

Figure 13: Annotation interface for multiple Elo scores.