Rethinking LLM-based Preference Evaluation

Zhengyu Hu[1], **Linxin Song**[2], **Jieyu Zhang**[3], **Zheyuan Xiao**[1],
Jingang Wang[4], **Zhenyu Chen**[4], **Jieyu Zhao**[2], **Hui Xiong**[1]
[1] HKUST(GZ) [2] USC [3] University of Washington [4] Meituan
{zhengyuhu, zheyuanxiao}@connect.hkust-gz.edu.cn,
{linxinso, jieyuz}@usc.edu, jieyuz2@cs.ishington.edu
{chenzhengyu04, wangjingang02}@meituan.com, xionghui@ust.hk

Abstract

Recently, large language model (LLM)-based preference evaluation has been widely adopted to compare pairs of model responses. However, a severe bias towards lengthy responses has been observed, raising concerns about the reliability of this evaluation method. In this work, we designed a series of controlled experiments to study the major impacting factors of the metric of LLM-based preference evaluation, *i.e.*, *win rate*, and conclude that the win rate is affected by two axes of model response: *desirability* and *information mass*, where the former is length-independent and related to trustworthiness, and the latter is length-dependent and can be represented by conditional entropy. We find that length impacts the existing evaluations by influencing information mass. However, a reliable evaluation metric should not only assess content quality but also ensure that the assessment is not confounded by extraneous factors such as response length. Therefore, we propose a simple yet effective adjustment, AdapAlpaca, to the existing practice of win rate measurement. Specifically, by adjusting the lengths of reference answers to match the test model's answers within the same interval, we debias information mass relative to length, ensuring a fair model evaluation.

1 Introduction

The advent of large language models (LLMs) has revolutionized various domains of artificial intelligence, from natural language processing to complex decision-making systems [Wu et al., 2023, Li et al., 2023a, Rao et al., 2023, Song et al., 2023]. As these models grow in complexity and capability, the development process increasingly relies on efficient and accurate evaluation mechanisms [Louis and Nenkova, 2013, Ma et al., 2024, Wang et al., 2023c]. LLM-based auto-evaluators have emerged as a crucial tool in this context, offering a cost-effective and scalable alternative to labor-intensive human evaluations [Chen et al., 2023, Li et al., 2023c, Dubois et al., 2024b, 2023]. Despite their advantages, these automated systems are not without their shortcomings, particularly concerning the introduction and perpetuation of biases [Li et al., 2023c, Zheng et al., 2023, Koo et al., 2023b, Wang et al., 2023b, Wu and Aji, 2023b].

One of the biases observed in LLM-based evaluations is the preference for longer textual responses [Dubois et al., 2024b]. Previous empirical studies have explored a strong correlation between the length of response and its perceived quality represented by win rate [Zhao et al., 2024, Dubois et al., 2024b, Ivison et al., 2023]. They reveal that longer responses are often deemed superior by LLM-based evaluators without a deep investigation of such bias. In this work, we ask the following question:

Preprint. Under review.

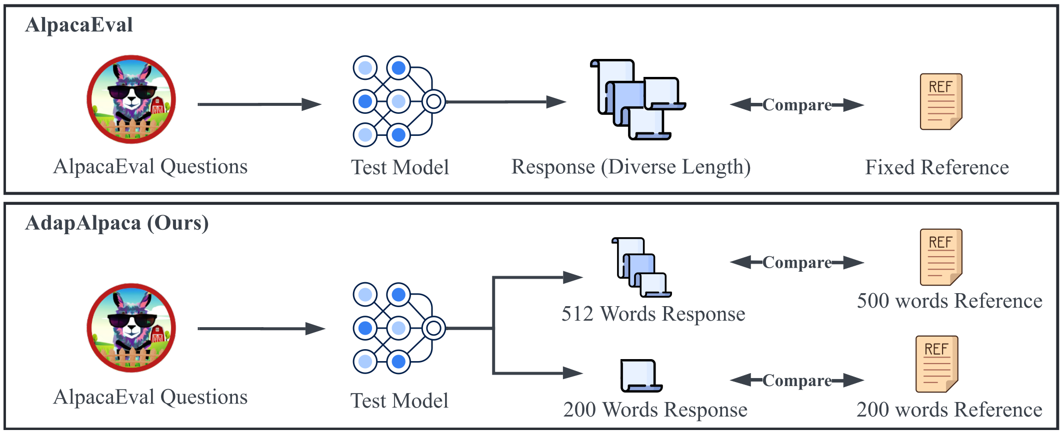


Figure 1: Comparison between AlpacaEval and AdapAlpaca. In AlpacaEval, the reference answer has a fixed length, regardless of the length of the test model's answer. In contrast, AdapAlpaca dynamically selects a reference answer that matches the length of the test model's answer.

Does length dominate the win rate? If not, what are the major factors contributing to the win rate?

Stemming from this question, we hypothesize that the *quality* of response, compared pair-wisely as win rate, can be decomposed as : (1) *desirability*, which is independent of length and reflects the trustworthiness of the response, encompassing factors such as correctness, toxicity, and consistency; and (2) *information mass*, which is dependent on length and represents the amount of information in the response, measurable through conditional entropy.

We validate our hypothesis by testing win rates in two different scenarios: (i) comparing normal responses with those differing in desirability (e.g., Logical to be desired and Biased not desired), and (ii) comparing normal responses with concise and detailed responses, which vary in information mass. Our experiments reveal that responses with negative desirability significantly decrease the win rate, whereas information mass, when not negatively influenced by desirability, is positively correlated with win rate, thus confirming our hypothesis. Following this finding, we designed a prompt called "Quality Enhancement" to improve information mass with positive desirability. This prompt enabled GPT-4 to achieve state-of-the-art results on AlpacaEval, increasing the win rate from 50.00% to 70.16%.

However, a reliable evaluation metric should assess content quality without being confounded by extraneous factors, such as response length. By selecting responses within the same length range for comparison, we control the information mass and debias it relative to length. This ensures a fair comparison of content quality across different response lengths, as illustrated in Figure 1. This structured approach helps eliminate length bias, thereby enhancing the reliability and validity of our evaluation.

Our major findings and contributions are as follows:

- We introduce a novel interpretation of win rate that emphasizes desirability and information mass, offering a more precise measure of LLM performance.
- Leveraging this insight, we have created the "Quality Enhancement" prompt, which significantly improves win rates across multiple LLMs, with average increases of 23.44% for GPT-3.5, 16.48% for GPT-4, 22.28% for Llama3 70B, and 20.40% for Qwen1.5 72B by enhancing information mass with positive desirability.
- We propose AdapAlpaca, a method that adjusts the lengths of reference answers to match the test model's answers within the same interval, thereby providing a more equitable and accurate measure of a model's performance.

2 Understanding the Major Factors of Win Rate

To interpret the correlation between length and win rate correlations, we propose a new definition based on *quality*, which includes *desirability* (length-independent, related to trustworthiness) and *information mass* (length-dependent, represented by conditional entropy). We validate our hypothesis through two scenarios: (1) testing the impact of different desirability on win rate with the same information mass, and (2) testing the influence of different information mass on win rate with the same desirability. Using these insights, we created the "Quality Enhancement" prompt, significantly improving win rates across multiple LLMs.

2.1 Preliminary

Evaluation protocol. We utilize the AlpacaEval dataset [Li et al., 2023c] to assess human preferences. AlpacaEval is a reference-free evaluation dataset for LLMs, encompassing 805 instructions that reflect human interactions on the Alpaca web demo. To ensure a comprehensive evaluation of human preferences, we extend our testing to additional datasets, including LIMA [Zhou et al., 2023], Vicuna [Chiang et al., 2023], Koala [Vu et al., 2023], Wizardlm [Xu et al., 2023], and Self-Instruct [Wang et al., 2022], in line with previous studies [Chen et al., 2023, Zhang et al., 2024, Du et al., 2023, Zhao et al., 2024, Li et al., 2023b].

Win rate. Assume we have a set of instructions x. We prompt a test model m to generate a response z_m for each instruction. Similarly, we prompt a reference model b (referred to as the "baseline" in AlpacaEval) to generate a response z_b for each instruction.[1] An annotator then evaluates these responses based on their quality and assigns a preference $y \in \{m, b\}$, indicating which model's response is superior. To properly understand the concept of win rate, we first need to define what we mean by response quality:

Definition 1. ***(Quality)**, denoted as $Q_e(z|x)$, quantifies the effectiveness of the model's response z in addressing the given instruction x, as evaluated by an annotators e. Annotators prefer responses with higher quality.*

Using this definition of quality, we can now formulate the win rate as follows:

$$\text{WinRate}(m, b) = \mathbb{E}_x \left[\mathbb{1}_{Q_e(z_m|x) > Q_e(z_b|x)} \right], \tag{1}$$

where $\mathbb{1}$ is an indicator function and $\mathbb{1}_{Q_e(z_m|x) > Q_e(z_b|x)}$ represents the preference distribution for each individual. Previous works [Chen et al., 2023, Li et al., 2023c, Dubois et al., 2024b, 2023] utilize LLMs as zero-shot evaluators due to their exceptional performance on real-world tasks. Our experimental setup adheres to the AlpacaEval Leaderboard [2] guidelines, employing the GPT-4 Preview (11/06)[3] as both the *Baseline* b and the *Annotator* e.

2.2 Quality Decomposition

Before discussing the composition of quality, we first define two key concepts: **desirability** and **information mass**.

Desirability reflects the inherent quality attributes of a response that make it reliable and valuable, irrespective of its length:

Definition 2. ***(Desirability)**, denoted as $D_e(z|x)$, measures the trustworthiness of a response z given an instruction x, as evaluated by annotators e. It includes factors such as consistency and toxicity and is independent of response length.*

Information mass captures the quantity of information in the response, with longer responses generally containing more content:

Definition 3. ***(Information mass)**, denoted as $H_e(z|x)$, measures the amount of information in a response z given an instruction x, as evaluated by annotators e. It is represented by conditional entropy and is directly related to response length.*

With these definitions in place, we now present our main hypothesis on answer quality, starting with an assumption:

[1] In this context, m stands for "model" and b denotes "baseline", which in this paper follows the AlpacaEval Leaderboard's use of GPT-4 Preview (11/06).

[2] https://tatsu-lab.github.io/alpaca_eval

[3] In this paper, unless specified otherwise, GPT-4 refers to GPT-4 Preview (11/06).

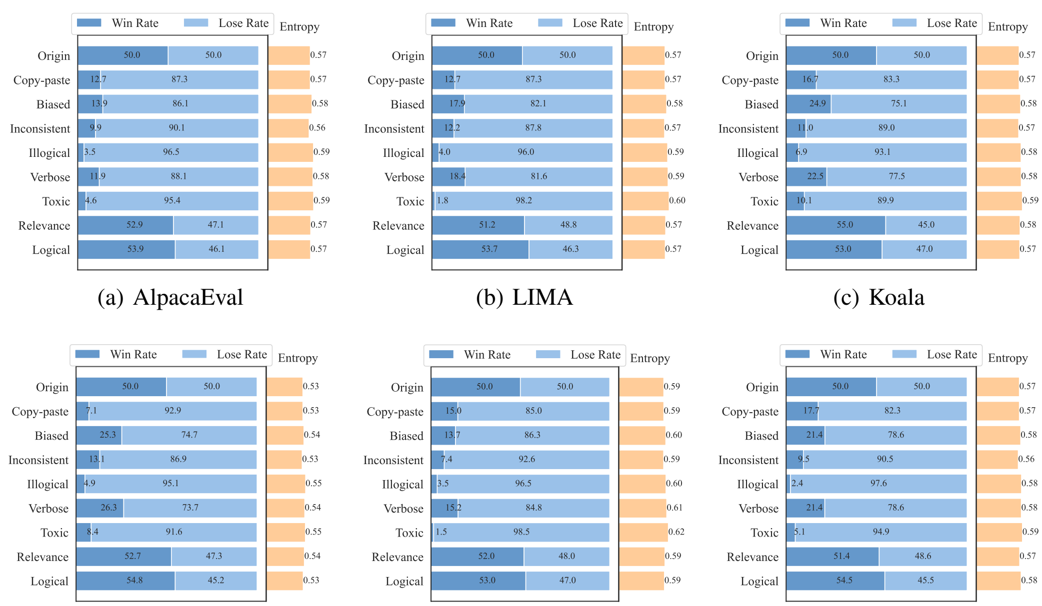
Assumption 1. *(**Quality Decomposition**). For a given answer z and instruction x, the quality $Q_e(z|x)$ recognized by annotators e can be decomposed as:*

$$Q_e(z|x) \propto D_e(z|x) \times H_e(z|x), \tag{2}$$

where $D_e(z|x)$ denotes the desirability of the response, and $H_e(z|x)$ represents the information mass.

To systematically verify our hypothesis, we design two distinct experiments aimed at manipulating these two key components of the responses generated by the models.

Desirability influences quality. To evaluate the impact of desirability on quality, we design experiments using eight strategies to manipulate response desirability. These strategies include: **Origin**: No prompt restrictions. **Copy-paste**: Copy GPT-4's response three times. **Biased**: Provide biased responses, favoring certain ideas without justification. **Inconsistent**: Provide contradictory information to create confusion. **Illogical**: Give responses based on flawed logic or irrelevant information. **Verbose**: Provide lengthy responses filled with broad, unrelated details. **Toxic**: Use offensive language with an aggressive tone. **Relevant**: Provide responses that align with query. **Logical**: Base responses on sound reasoning and valid arguments. The results are shown in Figure 2. To eliminate the impact of information mass on win rate caused by length, we control the length of all responses and **Origin** to be as consistent as possible, except for **Copy-paste**, as simply copying text does not increase information. The reasons for control of information mass through length can be found in Section 2.4. Details for these prompts and relevant implementation are shown in Appendix A and Appendix E.2. First, we observe that although the **Copy-paste** and **Origin** prompts maintain identical information mass (as simply replicating text does not increase information), the win rates of **Copy-paste** fall below **Origin** (50%) due to significant consistency impairments. Second, responses generated from negative prompts (i.e., **Biased**, **Inconsistent**, **Illogical**, **Verbose**, and **Toxic**) exhibit low desirability, resulting in win rates substantially lower than **Origin** (50%), despite having similar information mass. Conversely, prompts enhancing desirability (i.e., **Consistent** and **Logical**) yield increased win rates compared to **Origin**. In conclusion, we conclude that desirability significantly influences quality.



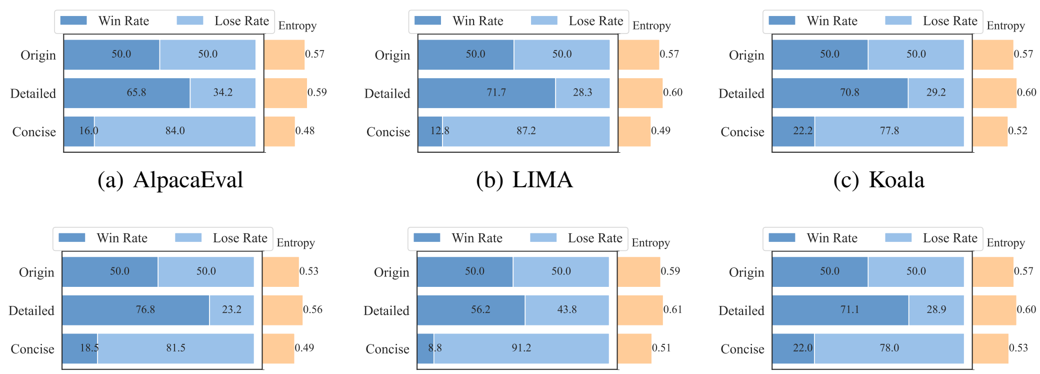
(a) AlpacaEval (b) LIMA (c) Koala

(d) Self-instruct (e) Vicuna (f) Wizardlm

Figure 2: Validation of desirability's impact on quality for GPT-4. The results demonstrate that desirability influences the win rate.

Information mass influences quality. To evaluate the impact of information mass on quality, we designed experiments using three distinct strategies to manipulate the information mass of responses.

These strategies include: **Origin**: No prompt restrictions. **Concise**: Request brief responses focusing on the most crucial points. **Detailed**: Request comprehensive responses covering all relevant aspects thoroughly. The results are presented in Figures 3. To isolate the effect of desirability, we ensured that prompts do not include any descriptions of desirability. Detailed descriptions of these prompts and their implementation can be found in Appendix A and Appendix E.2. Our findings indicate that information mass significantly affects the win rate without a negative desirability prompt. Specifically, responses with higher information mass, measured by conditional entropy, consistently achieved higher win rates. Thus, we observe the following relationship: **Detailed > Origin > Concise**. These results confirm that information mass is a crucial factor influencing the quality of responses.



(a) AlpacaEval (b) LIMA (c) Koala

(d) Self-instruct (e) Vicuna (f) Wizardlm

Figure 3: Validation of information mass's impact on quality for GPT-4. The results demonstrate that information mass influences the win rate.

2.3 Boost Win Rate with Quality Enhancement Prompt

Our analysis reveals that responses with good desirability and higher information mass are generally more favored. To leverage this insight, we propose the "Quality Enhancement" prompt (Table 1), designed to elevate both information mass and desirability, thus enhancing win rates. The keywords "relevant" and "logical" are used to enhance desirability, while "detailed" is used to boost information mass. The effectiveness of these keywords has been verified in Section 2.2. We evaluated this prompt across various models, including GPT-3.5, GPT-4, Llama3 70B, and Qwen1.5 72B on LIMA, Vicuna, Koala, Wizardlm, and Self-Instruct, with results summarized in Table 2. The substantial improvement in win rates across all tested models highlights the pivotal role of response quality in LLM evaluation.

Table 1: The content of the "Quality Enhancement" prompt, designed to elevate both the information mass and desirability of responses, thereby enhancing win rates. Keywords such as "relevant" and "logical" are used to enhance desirability, while "detailed" is used to boost information mass.

Quality Enhancement

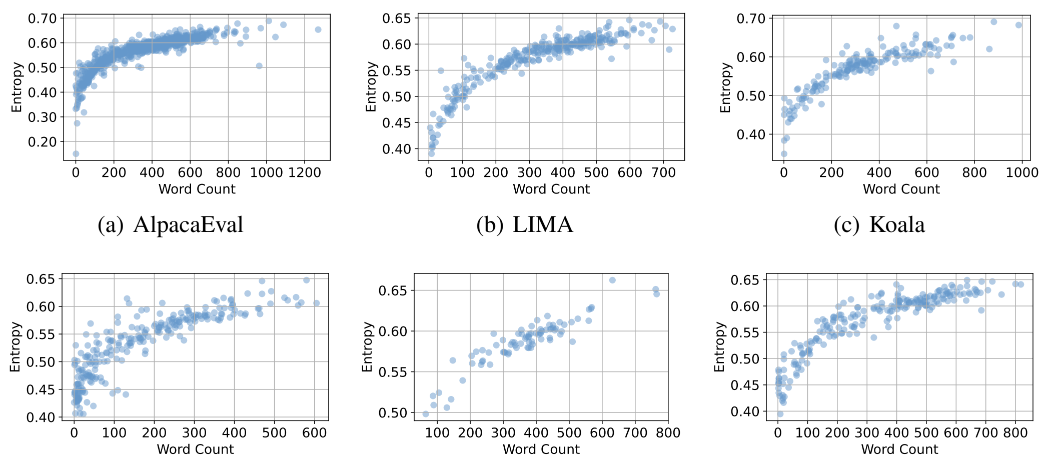
You are an expert assistant, delve deeply into the core of the topic, providing a richly **detailed** response that explores all its dimensions. Ensure each part of your response is **relevant** to the query in a **logical** manner. Your response should provide comprehensive information and thoroughly cover all relevant aspects with accuracy and depth.

2.4 Correlation between Length and Information Mass

In this section, we analyze the phenomenon noted in previous works [Dubois et al., 2024b, Chen et al., 2023, Dubois et al., 2023], which indicates a positive correlation between response length and win rate. Intuitively, longer responses tend to encompass more information. To quantify this, we employ conditional entropy to measure the amount of information in a response z given an instruction

Table 2: Win rates with and without the "Quality Enhancement" prompt, along with the corresponding win rate gains (WR Gain). "w/o QE" refers to results without the "Quality Enhancement", while "with QE" refers to results with the "Quality Enhancement". "WR Gain" represents the increase in win rate due to the use of the "Quality Enhancement".

Models	Methods	AlpacaEval	LIMA	Koala	Self-instruct	Vicuna	Wizardlm	Avg.
GPT-3.5	w/o QE	15.47	9.67	11.39	21.46	8.75	16.82	13.93
	with QE	29.89	36.53	40.34	45.93	35.88	35.62	37.36
	WR Gain	14.42	26.86	28.95	24.47	27.13	18.80	23.44
GPT-4	w/o QE	50.00	50.00	50.00	50.00	50.00	50.00	50.00
	with QE	70.16	65.84	58.90	67.06	73.13	63.76	66.48
	WR Gain	20.16	15.84	8.90	17.06	23.13	13.76	16.48
Llama3 70b	w/o QE	34.32	36.63	40.12	39.70	36.74	36.99	37.81
	with QE	56.50	60.39	61.30	64.81	63.49	51.70	59.70
	WR Gain	22.18	23.76	21.18	25.11	26.75	14.71	22.28
Qwen1.5 72b	w/o QE	28.27	28.40	35.25	33.81	33.70	31.80	32.67
	with QE	48.87	53.34	55.40	52.43	56.49	47.13	52.28
	WR Gain	20.60	24.94	20.15	18.62	22.79	15.33	20.40



(a) AlpacaEval (b) LIMA (c) Koala

(d) Self-instruct (e) Vicuna (f) Wizardlm

Figure 4: Correlation between entropy, i.e., information mass, and word count for responses. As the word count increases, the information mass also increases.

x. The relationship between Information Mass and Length is illustrated in Figure 4. It reveals that as the length of a response increases, the information mass, indicated by conditional entropy, also increases. By integrating these insights with the findings from Section 2.2, we conclude that the length of a response impacts its win rate primarily through information mass.

3 Adaptive AlpacaEval

3.1 Motivation

Adaptive AlpacaEval is built on the premise that a reliable evaluation metric should not only assess the content quality but also ensure that the assessment is not confounded by extraneous factors such as the length of the response. Central to this approach is the concept of information mass, which is inherently dependent on the length of the response and can be quantified using conditional entropy. Our primary aim is to mitigate scenarios where merely extending the length of a response artificially inflates its conditional entropy and, thus, its perceived quality in annotators. This approach involves dynamically adjusting the evaluation criteria based on the length of the responses, thereby providing a more equitable and accurate measure of a model's performance.

3.2 Dataset Generation

To support the development of Adaptive AlpacaEval, we first generate a diverse dataset using a modified prompting strategy with GPT-4, designed to produce responses within specific word count ranges. The data generation prompt, as outlined in Table 3, is carefully crafted to instruct GPT-4 to generate responses within predefined word limits. This prompt directed the model to generate content that is relevant to the given question and strictly adheres to the specified length constraints. These ranges were set to capture a broad spectrum of possible response lengths, from very concise (0-200 words) to quite detailed (800-1000 words). This approach ensures that our evaluation framework remains robust across a wide range of response lengths, reflecting real-world scenarios where brevity and detail may significantly vary.

Table 3: Prompt for dataset generation, with {max word}-{min word} ranges set as 0-200, 200-400, 400-600, 600-800, and 800-1000.

Dataset generation prompt

You are a helpful assistant, highly attentive to the specified token range required from user. Respond to the following question, your reply must only be within {max word}-{min word} words.

Specifically, we analyzed the word count distribution within the AlpacaEval dataset, observing that responses predominantly fall within the 0-1000 word range. This range was chosen to encompass the full spectrum of response lengths present in the original AlpacaEval dataset, ensuring comprehensive evaluation coverage. To systematically explore this range, we divided it into five equal segments, each representing a distinct dataset: AdapAlpaca-200: 0-200 words, AdapAlpaca-400: 200-400 words, AdapAlpaca-600: 400-600 words, AdapAlpaca-800: 600-800 words, AdapAlpaca-1000: 800-1000 words. Each segment is populated by generating responses using the dataset generation prompt, with GPT-4 configured to produce responses that strictly conform to the specified word counts. This segmentation allows us to adjust the lengths of reference answers to match those of the test model's answers within the same interval.

3.3 Analysis of the Generated Data

The analysis is structured to quantify each dataset's basic characteristics, followed by a comparative assessment to identify any significant differences attributable to the varying response lengths. Table 4 presents a comprehensive overview, providing a snapshot of the informational content across different datasets. Specifically, it includes vocabulary size, inter-sample N-gram Frequency (INGF) [Mishra et al., 2020], word counts of the generated dataset, win rate, length-controlled win rate, and entropy for AlpacaEval-Origin and AdapAlpaca-200, AdapAlpaca-400, AdapAlpaca-600, AdapAlpaca-800, and AdapAlpaca-1000.

Table 4: Comparison of five quantitative metrics related to quality: Vocabulary Size, Win Rate relative to AlpacaEval (AlpacaWR), Entropy, Inter-sample N-gram Frequency (INGF), and Word Counts.

Interval	Vocabulary Size		AlpacaWR		Entropy		INGF	Word Counts
	All	Ans Avg.	WR	LCWR	All	Ans Avg.		
AlpacAns Origin	38474	47.79	50.00	50.00	408.83	0.5686	7376.92	363.85
AdapAlpaca-200	22612	28.08	20.73	43.81	363.55	0.5056	1618.69	145.72
AdapAlpaca-400	36943	45.89	47.34	47.40	414.39	0.5763	6003.87	355.20
AdapAlpaca-600	47691	59.24	62.58	50.97	434.77	0.6046	9086.01	540.95
AdapAlpaca-800	55362	68.77	71.20	54.31	447.48	0.6223	10320.11	708.36
AdapAlpaca-1000	66095	82.10	66.98	36.24	456.32	0.6346	10981.84	913.44

Our findings indicate the following: 1) Longer responses generally exhibit higher vocabulary sizes and word counts, suggesting a richer linguistic structure. 2) The INGF metric reveals that while longer responses tend to include more common N-grams, there is significant variability in the types of N-grams used, indicating a creative and diverse use of language. 3) Under Win Rate (WR) metrics, longer responses disproportionately receive higher preference scores due to their higher information mass. However, applying the length-controlled win rate (LCWR) significantly reduces this bias, leading to a more balanced distribution of scores across different response lengths. However, this analysis aims to determine whether this is a genuine feature of the response or merely a byproduct of increased length. The results show that while longer responses generally have higher information mass, the quality of information (as measured by win rate) does not necessarily increase correspondingly. Excessively lengthy answers can lead to a decrease in desirability, such as reduced consistency. For example, in Table 4, the win rate of AdapAlpaca-1000 is smaller than that of AdapAlpaca-800.

Our findings indicate the following: 1) Longer responses generally exhibit higher vocabulary sizes and word counts, suggesting a richer linguistic structure. 2) The INGF metric reveals that while longer responses tend to include more common N-grams, there is significant variability in the types of N-grams used, indicating a creative and diverse use of language. 3) Under Win Rate (WR) metrics, longer responses disproportionately receive higher preference scores due to their higher information mass. However, applying the length-controlled win rate (LCWR) significantly mitigates this bias, leading to a more balanced distribution of scores across different response lengths. This analysis aims to ascertain whether this phenomenon is intrinsic to the response quality or merely a byproduct of increased length. Our results demonstrate that although longer responses generally possess higher information mass, the quality of information, as measured by win rate, does not necessarily increase proportionally. Excessively lengthy responses can result in a decline in desirability, such as reduced consistency. For instance, in Table 4, the win rate of AdapAlpaca-1000 is lower than that of AdapAlpaca-800.

3.4 Human Evaluation on AdapAlpaca

Case Study. To demonstrate the superiority of AdapAlpaca, we present a case study. In Figure 5, for the given instruction, we generate a redundant model answer (shown in the blue box). When evaluated using the current AlpacaEval response (shown in the red box), the annotator (i.e., GPT-4) selected this redundant answer, which is significantly unaligned from human preference, as the simplicity of the question does not warrant such extensive verbosity. The reason GPT-4 chose this answer is that the excessive length increases the information mass, artificially inflating the perceived quality. In contrast, when using AdapAlpaca, it allows us to control for content while varying the length, thereby isolating the effect of length from that of content quality.

Table 5: The subscripts in the LCWR and WR columns indicate the differences between these metrics and the corresponding Human WR. A larger absolute value denotes a greater disparity between the annotator's evaluation and Human Preference. "LLM Response" denotes different responses to AlpacaEval questions, with detailed content available in Section 2.2.

LLM Response	AlpacaEval			AdapAlpaca	
	Human	LCWR	WR	Human	WR
Concise	10.81	$35.16_{+24.35}$	$15.96_{+5.15}$	29.56	$28.44_{-1.12}$
Detailed	61.61	$54.13_{-7.48}$	$65.83_{+4.22}$	58.88	$57.81_{+1.07}$
Copy-paste	0.87	$2.78_{+1.91}$	$12.67_{+11.80}$	0.74	$8.68_{+7.94}$
Quality Enhancement	66.70	$49.37_{-17.33}$	$70.16_{+3.46}$	56.02	$55.36_{-0.78}$

Comparison with AlpacEval. The results of the human study[4] are presented in Table 5. We first test the results of concise, copy-paste, detail and quality enhancement (see Section 2.2 for details) using AlpacaEval and subsequently using AdapAlpaca. From the gap values between LCWR and human evaluations, we observe significant misalignments, indicating inherent problems with the LCWR metric. In contrast, the win rate calculated using AdapAlpaca closely aligns with the human results, showing an average difference of 2.72% (1.12% + 1.07% + 7.94% + 0.78% / 4). Additionally, we find that the difference between human evaluation and WR decreases as the quality of responses improves (from copy-paste to concise to detailed to Quality Enhancement). This suggests that as

[4]Details of the human study can be found in Appendix B.

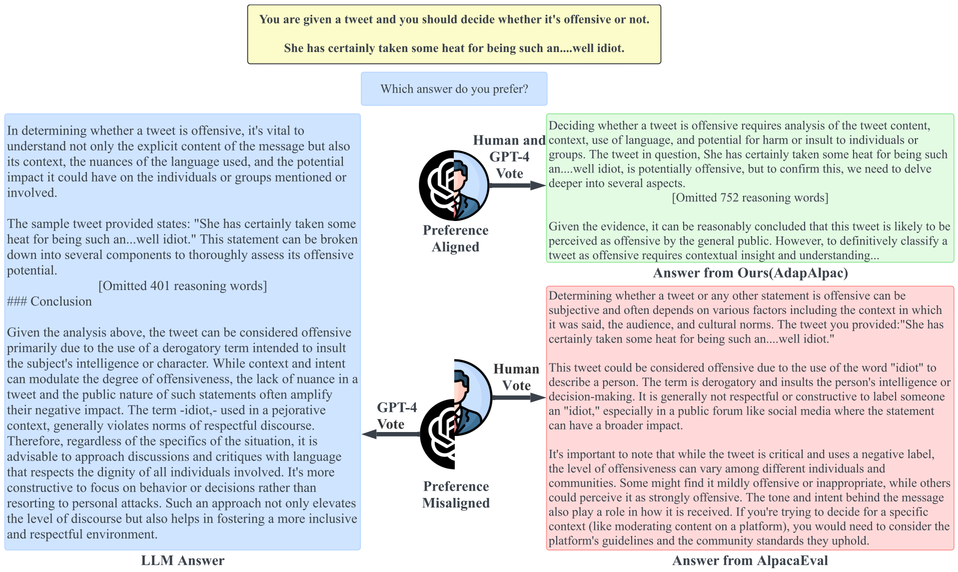


Figure 5: Case study on comparing GPT-4 and human vote on AlpacaEval and AdapAlpaca. In AlpacaEval, GPT-4 votes for the verbose answer, but humans vote for the concise reference answer, while in AdapAlpaca, GPT-4 and humans vote for the same answer, demonstrating a better LLM-human alignment on AdapAlpaca.

response quality increases, the preferences of annotators and human evaluators converge. Moreover, we found that the smallest difference in win rate between GPT-4 and human evaluations occurs when using the "Quality Enhancement" prompt, which has the highest levels of desirability and information mass. This further underscores the importance of enhancing both desirability and information mass in model responses. Overall, while both AdapAlpaca and LCWR aim to mitigate length bias in evaluating human preferences, their approaches differ fundamentally. AdapAlpaca eliminates length bias from the outset, whereas LCWR attempts to correct for length bias after it has already influenced the evaluation. The inherent issue with LCWR is that length significantly impacts human preference, and adjusting for length retrospectively is not a reliable approach.

4 Related Work

4.1 Reference-free Evaluation Metrics

Reference-free evaluation metrics have a long history [Louis and Nenkova, 2013], which evaluates the generated text based on intrinsic properties and coherence with the context. Although they achieve high accuracy on matching inner-annotator, the achievement suffers from spurious correlations such as perplexity and length [Durmus et al., 2022]. Recently, people have started using a strong model (e.g., GPT-4) as an evaluator to perform a zero-shot reference-free evaluation on the weak models [Shen et al., 2023, Dubois et al., 2024c,c, Chen et al., 2023]. However, leveraging a strong model's intrinsic knowledge to perform reference-free evaluation ignores the prompt preference of the strong model, for example, the prompt's length.

4.2 Correlation Between Length and Win Rate

Previous research reveals that sentence length will influence the evaluation of trustworthiness. Specifically, when using a GPT-4 to represent human preference, it will prefer to choose a long sentence rather than a short sentence [Dubois et al., 2024a, Ivison et al., 2023, Gu et al., 2023, Shen et al., 2023, Koo et al., 2023a, Wang et al., 2023a, Wu and Aji, 2023a]. Such preference will introduce a length-correlated bias and help the model with long-generation sentences gain a high score on human preference evaluation. Although these approaches show a high correlation to human preference, debiasing such as automated evaluation is highly valuable. [Dubois et al., 2024a] proposes a length-controlled (LC) win rate by removing the length-correlated term in the win rate regression model. The new LC win rate shows an even performance between concise and verbose input and a higher correlation when compared with human preference.

5 Conclusion

In this work, we identified and addressed a critical bias in LLM-based preference evaluations towards longer responses. Through a series of controlled experiments, we demonstrated that win rate is influenced by two primary factors: desirability and information mass. Desirability is independent of response length and pertains to the trustworthiness of the content, while information mass is dependent on response length and can be quantified using conditional entropy. To mitigate the bias introduced by response length, we proposed AdapAlpaca, a method that adjusts the lengths of reference answers to align with the test model's answers. This approach effectively debiases information mass relative to length, ensuring a more equitable evaluation of model responses. Our findings underscore the necessity of unbiased evaluation metrics in LLM research, paving the way for more accurate assessments of model performance.

References

J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, and H. Jin. Alpagasus: Training a better alpaca with fewer data, 2023.

W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Q. Du, C. Zong, and J. Zhang. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*, 2023.

Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

Y. Dubois, B. Galambosi, P. Liang, and T. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *ArXiv*, abs/2404.04475, 2024a. URL https://api.semanticscholar.org/CorpusID:269004605.

Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024b.

Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. S. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024c.

E. Durmus, F. Ladhak, and T. B. Hashimoto. Spurious correlations in reference-free evaluation of text generation. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:248300077.

Y. Gu, L. Dong, F. Wei, and M. Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

R. Koo, M. Lee, V. Raheja, J. I. Park, Z. M. Kim, and D. Kang. Benchmarking cognitive biases in large language models as evaluators. *ArXiv*, abs/2309.17012, 2023a. URL https://api.semanticscholar.org/CorpusID:263310448.

R. Koo, M. Lee, V. Raheja, J. I. Park, Z. M. Kim, and D. Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023b.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

M. Li, Y. Zhang, Z. Li, J. Chen, L. Chen, N. Cheng, J. Wang, T. Zhou, and J. Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023b.

X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023c.

A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39:267–300, 2013. URL https://api.semanticscholar.org/CorpusID:17829732.

Z. Ma, W. Huang, J. Zhang, T. Gupta, and R. Krishna. m&m's: A benchmark to evaluate tool-use for multi-step multi-modal tasks. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

S. Mishra, A. Arunkumar, B. Sachdeva, C. Bryan, and C. Baral. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*, 2020.

A. S. Rao, J. N. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi. Evaluating chatgpt as an adjunct for radiologic decision-making. *medRxiv : the preprint server for health sciences*, 2023. URL https://api.semanticscholar.org/CorpusID:256626649.

W. Shen, R. Zheng, W. Zhan, J. Zhao, S. Dou, T. Gui, Q. Zhang, and X. Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:263831112.

L. Song, J. Zhang, L. Cheng, P. Zhou, T. Zhou, and I. Li. Nlpbench: Evaluating large language models on solving nlp problems. *arXiv preprint arXiv:2309.15630*, 2023.

J. Von Neumann. *Mathematische grundlagen der quantenmechanik*, volume 38. Springer-Verlag, 2013.

T.-T. Vu, X. He, G. Haffari, and E. Shareghi. Koala: An index for quantifying overlaps with pre-training corpora. *arXiv preprint arXiv:2303.14770*, 2023.

P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926, 2023a. URL https://api.semanticscholar.org/CorpusID:258960339.

P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023b.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023c.

L. Wei, Z. Tan, C. Li, J. Wang, and W. Huang. Large language model evaluation via matrix entropy. *arXiv preprint arXiv:2401.17139*, 2024.

M. Wu and A. F. Aji. Style over substance: Evaluation biases for large language models. *ArXiv*, abs/2307.03025, 2023a. URL https://api.semanticscholar.org/CorpusID:259360998.

M. Wu and A. F. Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023b.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *ArXiv*, abs/2308.08155, 2023. URL https://api.semanticscholar.org/CorpusID:260925901.

C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Q. Zhang, Y. Zhang, H. Wang, and J. Zhao. Recost: External knowledge guided data-efficient instruction tuning. *arXiv preprint arXiv:2402.17355*, 2024.

H. Zhao, M. Andriushchenko, F. Croce, and N. Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *arXiv preprint arXiv:2402.04833*, 2024.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL https://api.semanticscholar.org/CorpusID:259129398.

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract, and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Appendix C.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix D.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.2.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The data and code are available on `https://github.com/ICWR-NP/ICWR`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they are chosen)? [Yes] See Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The data and code are available on `https://github.com/ICWR-NP/ICWR`.
   (d) Did you discuss whether and how consent is obtained from people whose data you're using/curating? [Yes] See Appendix B and Appendix E.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix B and Appendix E.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix B and Appendix E.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix B and Appendix E.

A Prompt Content

Here, we show the 6 prompts in Table 6 we used to generate the AlpacaEval answers.

B Human Evaluation Process

To ensure the robustness of our findings and complement the automated evaluations, a thorough human evaluation is conducted. Here, we detail the methodology and setup used to execute this vital component of our research.

Participants. The human evaluation is conducted by a group of 10 individuals working in the tech industry and active researchers in computer science. These participants are selected based on their familiarity with language models and their ability to critically assess AI outputs.

Data Segmentation. To manage the evaluation process effectively, the dataset is divided into several parts. This division allowed participants to focus on distinct subsets of the data, ensuring that each section received adequate attention and that evaluations are distributed evenly among the participants.

Evaluation Interface. The evaluation is facilitated using a custom-built interface hosted on Gradio, an open platform that allows for the easy sharing of machine learning models with interactive interfaces. Before starting the evaluation, participants selected a dataset and a specific part of the data to work on from dropdown menus at the top of the interface. Upon selection, the interface displayed a question along with two corresponding outputs side-by-side, labeled "Left" and "Right." The order of these outputs is randomized to prevent any positional bias. Participants are then tasked with evaluating the outputs relative to the question posed, deciding which output (Left or Right) is more appropriate. An example of the evaluation interface is shown in Figure 6.

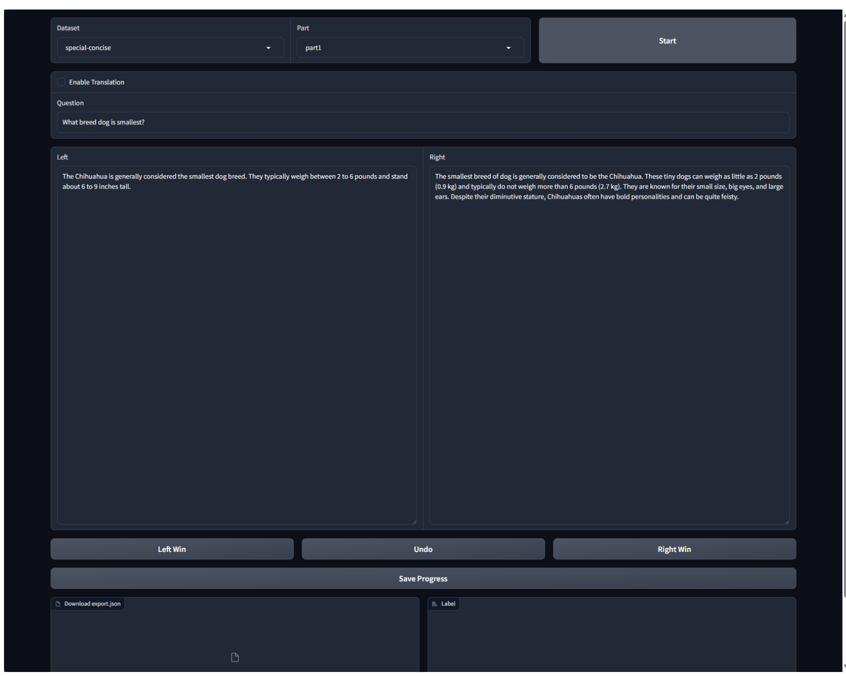


Figure 6: Example of the evaluation interface used in the human study.

Data Collection and Analysis. Upon completion of the evaluations, data is collected regarding which output is selected as superior for each question across the different datasets. This data is then used to calculate the human win rate, a metric indicative of human preference, which served as a benchmark for assessing the performance and bias of the evaluated language models.

This comprehensive human evaluation process is integral to validating the effectiveness of our proposed methodologies and provided critical insights that supplemented the automated metric evaluations.

Table 6: The content of Concise, Detail, Toxic, Inconsistent, Biased, Illogical and Verbose Prompt.

Dataset generation prompt - Concise
Provide concise responses. Limit details to the most crucial points only.

Dataset generation prompt - Detail
Respond with detailed information. Cover all relevant aspects thoroughly.

Dataset generation prompt - Toxic
Use language that may be considered offensive or inappropriate. Prioritize aggressive tones and disregard for sensitivity. Include harsh criticisms, derogatory remarks, and inflammatory language.

Dataset generation prompt - Inconsistent
Deliberately provide responses that contradict earlier answers or established facts, creating confusion. Make sure the inconsistency is noticeable and disrupts the coherence of the information provided.

Dataset generation prompt - Biased
Deliver responses that show a clear prejudice, favoring certain ideas or opinions without justification. Ignore impartiality and balance in your explanations and arguments. Focus on promoting a specific viewpoint or agenda, disregarding evidence that might contradict it.

Dataset generation prompt - Illogical
Construct arguments based on flawed logic, superstitions, or irrelevant information, avoiding sound reasoning. Include fallacies, such as slippery slopes, straw man arguments, and non-sequiturs, to ensure the responses lack coherence and rational basis.

Dataset generation prompt - Verbose
Tasked with providing a lengthy response filled with general information. Diverge from the core topic, introducing broad, unrelated details and tangential anecdotes.

Dataset generation prompt - Relevant
Dive deeply into the core issues of the query. Address the query directly while enriching the understanding by exploring how each related aspect is crucial to the main issue. Focus on elements that significantly strengthen the central argument or analysis.

Dataset generation prompt - Logical
Ensure that your response provides a clear and logical progression from initial assumptions to final conclusions. Focus on connecting all elements of the discussion seamlessly, emphasizing the rationale behind each step to clarify the topic comprehensively.

C Limitations

Although the focus on length bias provides valuable insights, other types of biases related to content, context, and demographic factors are not within the scope of this study. Addressing these biases requires developing additional methodologies that extend beyond the current framework.

D Potential Negative Societal Impacts

While this research contributes to reducing bias in language model evaluations, it is important to consider potential indirect societal impacts that might arise:

Dependence on Automated Decision-Making. This study's focus on enhancing the accuracy of automated evaluations may inadvertently promote an over-reliance on AI-driven decision-making processes. While beneficial in many respects, such reliance could diminish the value placed on human judgment and intuition in areas where nuanced understanding and ethical considerations are paramount.

Perception and Trust in AI. By highlighting the capabilities and improvements in AI evaluations, there might be an overestimation of AI reliability and fairness among the public and policymakers. This could lead to misplaced trust in AI systems, overlooking their limitations and the necessity for continuous oversight and human intervention.

E Implementation Detail

E.1 Dataset

- **AlpacaEval** [Dubois et al., 2024c] comprises 805 instructions, including 252 from the self-instruct test set [Wang et al., 2022], 188 from the Open Assistant (OASST) test set, 129 from Anthropic's helpful test set [Zhou et al., 2023], 80 from the Vicuna test set [Chiang et al., 2023], and 156 from the Koala test set [Vu et al., 2023].
- **LIMA** [Zhou et al., 2023] compiles a training dataset of 1000 prompts and responses, designed to ensure stylistic consistency in outputs while maintaining diverse inputs. It also provides an open-source test set of 300 prompts and a development set of 50. The dataset is sourced from a variety of platforms, mainly community Q&A websites such as Stack Exchange, wikiHow, and the Pushshift Reddit Dataset [Baumgartner et al., 2020], along with manually curated examples. Within these Q&A communities, highly upvoted answers on Reddit often have a humorous or trolling tone, requiring extra effort to align them with the intended helpful chat assistant style. In contrast, responses from Stack Exchange and wikiHow naturally align with this style. The inclusion of human-authored examples further enhances the dataset's diversity. Our research specifically utilizes the test set from the LIMA dataset to evaluate our models.
- **Vicuna** [Chiang et al., 2023] divides 80 test instructions into eight distinct categories: Fermi problems, commonsense, roleplay scenarios, coding/math/writing tasks, counterfactuals, knowledge, and generic questions. This categorization is intended to thoroughly evaluate multiple aspects of a chatbot's performance. Prior research indicates that the Vicuna dataset generally includes instructions of lower difficulty and complexity [Xu et al., 2023]. In our study, we used the Vicuna test set to specifically evaluate the performance of large language models across these varied instruction categories.
- **Self-Instruct** [Wang et al., 2022] consists of 252 human-created test instructions, each associated with a carefully designed output. This test set is curated to reflect the real-world applicability of instruction-following models, covering a broad spectrum of domains including email composition, social media, productivity software, and coding. The test instructions vary in style and format, incorporating different task lengths and diverse input/output types such as bullet lists, tables, code snippets, and mathematical equations. We employed the Self-Instruct test set in our research to rigorously assess our model's capability to comply with precise instructions across these varied domains.
- **Wizardlm** [Xu et al., 2023] comprises a training set of 70k examples with varied complexities, initiated from 52k instructional data provided by Alpaca. Following $M = 4$

evolutionary cycles, the collection expands to 250k instructions. In each cycle, from the six newly generated prompts—five via in-depth evolution and one through in-breadth evolution—one is chosen randomly for each instruction. ChatGPT then generates responses, resulting in $52 \times 4 \times 3 = 624$k instruction-response pairs. The training subset selected for the Evol-Instruct dataset contains 70k of these instructions. The test set, which includes 218 instructions, is sourced from a variety of platforms such as open-source projects and online forums, encapsulating 29 unique skills identified from authentic human tasks. These skills range from Coding Generation & Debugging to Reasoning, Mathematics, Writing, Handling Complex Formats, and Mastery over Extensive Disciplines. In our study, we utilized the Wizardlm test set to thoroughly evaluate our model's ability to adhere to detailed instructions.

- **Koala** [Vu et al., 2023] consists of 180 authentic user queries obtained from the Internet. These queries cover a diverse array of topics and are generally characterized by a conversational tone, underscoring their applicability to real-world chat-based applications. To prevent test-set leakage, we exclude any query that achieves a BLEU score over 20% when compared to examples from our training set. Furthermore, we do not consider queries related to programming or non-English languages, as the capabilities of our crowd-sourced raters—who form our evaluation team—do not extend to effectively assessing such content. We have exclusively utilized the Koala test set to assess our model's capability to process and respond to genuine user inquiries in a conversational setting.

E.2 Experiment Setup

In our experiments, we follow the setup in the AlpacaEval Leaderboard[5], using the GPT-4 Preview (11/06) as *Baseline* as well as the *Annotator*. The references to GPT-3.5, Llama3 70b, and Qwen1.5 72b in the main text denote gpt-3.5-turbo-0125, meta-llama/Meta-Llama-3-70B-Instruct [6], and Qwen/Qwen1.5-72B-Chat [7], respectively. Following previous work [Wei et al., 2024], we calculate conditional entropy using the method described in [Von Neumann, 2013].

E.3 Dataset Information

All data and related code are available in `https://github.com/ICWR-NP/ICWR`.

Dataset Documentations. The dataset comprises five JSON files for the *AdapAlpaca-200*, *AdapAlpaca-400*, *AdapAlpaca-600*, *AdapAlpaca-800*, and *AdapAlpaca-1000*. Each file is generated using our length control prompt technique with the Alpaca dataset employing the GPT-4 1106 model.

Each data file contains a list of items with the following fields:

- `instruction`: the prompt given to generate the response.
- `generator`: identifies the model used.
- `dataset`: specifies the dataset used.
- `output_word_count`: the word count of the generated response.
- `output`: the actual text generated by the model.

Intended Uses. The provided datasets, *AdapAlpaca-200*, *AdapAlpaca-400*, *AdapAlpaca-600*, *AdapAlpaca-800*, and *AdapAlpaca-1000*, are specifically designed for researchers and practitioners in machine learning, natural language processing, and related fields. These datasets are intended to facilitate the evaluation of models that generate responses of similar lengths. They provide a standardized framework to repeatedly test and compare the performance of different models as detailed in our accompanying paper. This aims to ensure consistent evaluation and benchmarking of models under controlled conditions that mimic real-world application scenarios.

Hosting and Maintenance Plan. The datasets and codebase are hosted and version-tracked via GitHub, accessible under the following link: `https://github.com/ICWR-NP/ICWR/tree/main/`

[5] `https://tatsu-lab.github.io/alpaca_eval`

[6] `https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct`

[7] `https://huggingface.co/Qwen/Qwen1.5-72B-Chat`

reference/adapAlpaca_reference. It is a community-driven and open-source initiative, committed to active development and maintenance for at least the next five years. We plan to expand the repository by including new tasks and datasets and warmly welcome contributions from external collaborators.

Metadata Availability. The Croissant metadata file for the datasets can be found at the following URL: https://github.com/ICWR-NP/ICWR/blob/main/reference/adapAlpaca_croissant_metadata/croissant.json. This metadata file provides detailed descriptions and the structure of our dataset, facilitating transparency and reproducibility in research.

Licensing. We license our work using Apache 2.0[8].

Author Statement. We the authors will bear all responsibility in case of violation of rights.

[8]https://www.apache.org/licenses/LICENSE-2.0