Verbosity Bias in Preference Labeling
by Large Language Models

Keita Saito[*]
University of Tsukuba & RIKEN AIP
Tsukuba, Ibaraki 305-8573, Japan
keita.saito@bbo.cs.tsukuba.ac.jp

Akifumi Wachi
LY Corporation
Chiyoda-ku, Tokyo 102-8282, Japan
akifumi.wachi@lycorp.co.jp

Koki Wataoka
LY Corporation
Chiyoda-ku, Tokyo 102-8282, Japan
koki.wataoka@lycorp.co.jp

Youhei Akimoto
University of Tsukuba & RIKEN AIP
Tsukuba, Ibaraki 305-8573, Japan
akimoto@cs.tsukuba.ac.jp

Abstract

In recent years, Large Language Models (LLMs) have witnessed a remarkable surge in prevalence, altering the landscape of natural language processing and machine learning. One key factor in improving the performance of LLMs is alignment with humans achieved with Reinforcement Learning from Human Feedback (RLHF), as for many LLMs such as GPT-4, Bard, etc. In addition, recent studies are investigating the replacement of human feedback with feedback from other LLMs named Reinforcement Learning from AI Feedback (RLAIF). We examine the biases that come along with evaluating LLMs with other LLMs and take a closer look into verbosity bias – a bias where LLMs sometimes prefer more verbose answers even if they have similar qualities. We see that in our problem setting, GPT-4 prefers longer answers more than humans. We also propose a metric to measure this bias.

1 Introduction

Large Language Models (LLMs) have made tremendous strides in recent years and continue to gain popularity (Zhao et al., 2023). With its growing size in network parameters, its wide application ranges from conventional natural language processing tasks such as chat-bots, summarization, and translation, to other applications beyond its original intended use such as search engines, programming assistance, and foundation models (Zhao et al., 2023; Brants et al., 2007; Katz, 1987).

After pretraining for general purposes, LLMs are fine-tuned to further better their performance for specific tasks with supervised learning and RLHF – reinforcement learning from preference labeling feedback from humans (Stiennon et al., 2020; Ouyang et al., 2022). However, issues arise with RLHF where human feedback can become costly. To work around this problem, Reinforcement Learning from AI Feedback (RLAIF) was proposed (Bai et al., 2022b; Lee et al., 2023), which replaces human feedback with inexpensive feedback from other LLMs.

In many cases, the question lack a clear-cut "correct answer" and require creativity and imagination. As evident in an example of feedback by an LLM provided in Figure 1, when LLMs are tasked to assess responses to such prompts, the evaluation process can become arbitrary and introduce various biases. One prominent bias is the verbosity bias, which occurs when LLMs are influenced

---

[*]Research done during internship at LY Corp.

Figure 1: An example of a prompt to an LLM to judge two texts and the verdict. There is no one correct answer, and a comprehensive judgement is required.

by verbosity, favoring longer and more verbose texts, even if they appear wordy or of lower quality. Without accounting for this bias, LLM agents may learn to generate unnecessarily long texts. This may result in failures in downstream tasks such as lengthy summarizations or chatbots that return verbose responses to simple questions.

While previous studies have explored the concept of verbosity bias, they have tended to focus on specific cases. Zheng et al. (2023) limits their problem setting to questions answered with lists in their experiment on verbosity bias, and Huang et al. (2023) conducted experiments on summarization tasks. Moreover, these do not compare the preferences of LLMs to those of humans. We believe that such a comparison is crucial in challenging the conjecture that longer answers are inherently better and that LLMs are actually correct in their preferences.

Our contributions.   In this paper, we conduct experiments on verbosity bias and saw that 1) LLMs exhibit a preference for longer answers in creative writing tasks, and 2) there is a discrepancy between of LLMs and those of humans in verbosity preference. Additionally, we formulate a quantification for measuring verbosity bias based on accuracy parity. This can be used to compare LLMs on their degree of verbosity bias.

2   Preliminaries

After undergoing pretraining for general purposes, LLMs are fine-tuned to further improve their performance in specific tasks. Pretraining is accomplished through self-supervised learning, where the model is trained to predict the next token in a sentence. Once the LLM is able to generate cohesive sentences, we proceed to fine-tune the model to solve specific tasks. One approach to fine-tuning involves supervised learning using expert data. This method relies on examples where experts have solved the task at hand. An example of a conversational LLM trained solely using this approach is Vicuna (Chiang et al., 2023). Vicuna achieved performance comparable to ChatGPT by utilizing user-shared conversations with ChatGPT as expert data. However, it is worth noting that obtaining expert data is often challenging.

RLHF addresses the challenge of limited training data in supervised learning by leveraging human feedback (Stiennon et al., 2020; Ouyang et al., 2022). This approach not only mitigates data scarcity but also significantly enhances alignment with human preferences, a critical factor in applications such as question answering. In RLHF, a reward model is trained to closely match human feedback data, which acts as the reward signal in the subsequent RL phase. Prominent LLMs like ChatGPT and Bard adopt a hybrid approach, combining both supervised learning and RLHF techniques to further refine their alignment with human preferences.
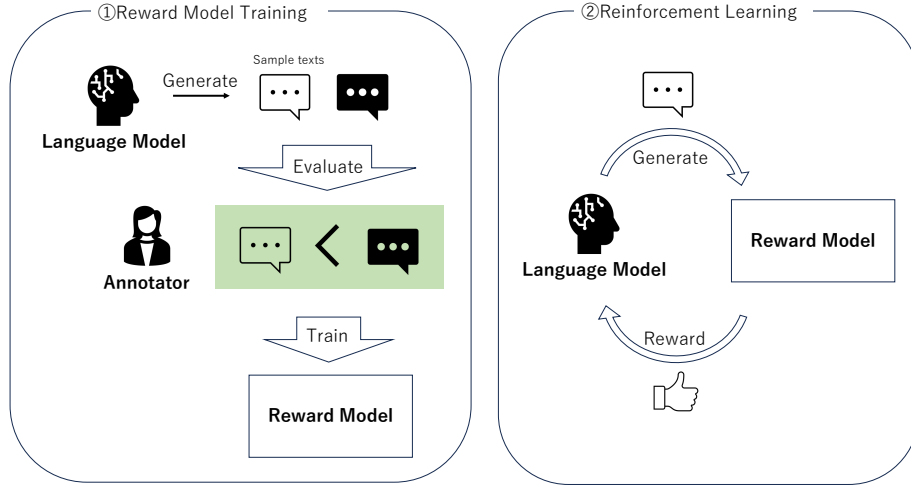
Figure 2: Two phases in RLHF. First the reward model is trained to align with human preference by matching with human feedback. In the second RL phase, the trained reward model provides reward signals to the language model.

2.1 RLHF

The first step of RLHF is to fit the reward function to align with human feedback. RL directly from human feedback as reward signal is unstable and requires volume. Therefore, a reward model that acts as the reward signal later in the process is trained to be consistant with human preference. Given a dataset $\mathcal{D}$ consisting of the original question, a pair of generated text, and the human preference label on which is chosen and rejected, the reward model is trained by minimizing

$$\mathcal{L}(\phi) = - \mathbb{E}_{(x, y_{\text{chosen}}, y_{\text{rejected}}) \sim \mathcal{D}} \left[\log \sigma (r_\phi(y_{\text{chosen}} \mid x) - r_\phi(y_{\text{rejected}} \mid x)) \right], \qquad (1)$$

where x is the prompt to the LLM, y_{chosen} is the preferred text, y_{rejected} is the rejected text, and r_ϕ parametrised with ϕ is the reward model that takes text as input and outputs the rating score.

In the second step of RLHF, now that we have a reward model to evaluate a generated text without human interaction, regular RL can take place. In this context, the state is the question and the generated text so far, action is the next token to generate, and the reward is $r_\phi(y)$ given after the full text is generated. This is equivalent to a task where a sparse reward is given only at episode termination. The LLM maximizes the signal from the reward model with the following:

$$\max_\theta \mathbb{E}[\, r_\phi(\pi_\theta(x) \mid x)\,], \qquad (2)$$

where π_θ is a policy parameterized by θ. An optional KL divergence term is added to penalize the policy from deviating from the original policy.

2.2 RLAIF

While RLHF brings down the cost of human labor compared to generating an expert data from scratch, human feedback is still costly. For example, in Wang et al. (2023), it cost around 3 minutes ($0.75 if $15.00 per hour) per evaluation. In one occasion, OpenAI worked around this by employing people in Kenya on less than $2 per hour pay in the process of labeling violent or innappropriate texts. They were under scrutiny for the unideal working conditions.

To combat these promblems, RLAIF was proposed. This method replaces human feedback with feedback from other LLMs. This brings down the cost significantly; in our case, evaluation cost around $0.05 each, which is 1/15th compared to human feedback in the previously cited paper (Wang et al., 2023).

2.3 Biases in Automated LLM Evaluation

When LLMs evaluate generated texts, various biases are introduced. We provide below a list of biases discussed in various papers (Zheng et al., 2023; Bai et al., 2022a; Wang et al., 2023).

Position Bias: Position bias occurs when, in comparing generated texts, LLMs prefer the answer given in certain positions. If we define the ground truth probability of a (the first parameter) preferred over b (the second parameter) to be $P(a, b)$, it should be that $P(y_0, y_1) = 1 - P(y_1, y_0)$, meaning the position should have no affect on the judgement. Position bias is when the comparison by model is $\hat{P}(y_0, y_1) \neq (1 - \hat{P}(y_1, y_0))$. For example, GPT-4 tends to prefer the first option given to it, while ChatGPT prefers the second option (Wang et al., 2023). To account for this bias, we can simply swap the positions and evaluate the options twice. If the model gives contradicting results between permutations, we count it as a draw.

Wang et al. (2023) has proposed several methods to calibrate this bias further: Multiple Evidence Calibration asks the LLM to provide evidence before making judgement, and Human-in-the-Loop Calibration involves human adjustment when deemed neccessary.

Self-enhancement Bias: LLMs tend to prefer answers generated by itself compared to answers generated by other models. This becomes a problem when benchmarking LLMs by evaluating them with LLMs (Zheng et al., 2023), but not so much in the context of RLAIF, as the comparisons are always between answers generated by the same model.

Verbosity Bias: Verbosity bias refers to the bias where LLMs prefer longer, more verbose answers even if there are no difference in quality. Training with RLAIF with verbosity bias present can lead to LLMs generating excessively long responses, when in reality a much more concise response would suffice. In tasks such as question answering, a verbose response can be critical to its usefulness, but there aren't enough researches that look into this. For these reasons we take a closer look into this.

There are several proposed methods to mitigate the effect of biases.

Chain-of-thought Prompting is a prompting technique where the LLM is asked to provide the thought process before generating the actual evaluation. This way, at the time when the LLM generates the actual evaluation, it has its chain-of-thought to base its evaluation from. This encourages human alignment and more accurate evaluations, rather than arbitrary evaluations without thought.

One-shot/Few-shot Prompting is another prompting technique which gives one example/several examples of a prompt and its corresponding correct answer when promting the LLM. When generating the response, the LLM can continue the pattern from the examples to better align with the intended response.

3 Related Works

3.1 RLAIF Advancements

There have been several recent advancements in the field of RLAIF. Bai et al. (2022b) trained an LLM via RLAIF with limited human feedback. In this work, they claim that helpfulness and harmfulness have a trade-off relationship, and aim to train an LLM that keeps a balance between those two. Their method only requires human feedback in the helpfulness aspect, and harmless behavior is achieved purely from RLAIF. The LLMs trained in the work by Lee et al. (2023) achieved near-human performance in summarization tasks with RLAIF without any human feedback. While not a study on RLAIF itself, Zheng et al. (2023) evaluate LLMs with other LLMs as a judge and show that GPT-4 has a high human alignment and agrees with humans on over 80% of evaluations.

3.2 On Verbosity Bias in Evaluations by LLMs

Zheng et al. (2023) also provides lists of biases and methods to overcome them. Alongside their experiment on position bias, they experimented on verbosity bias by attempting a "repetitive list attack" on several LLMs. This attack pertains to "listing" tasks, in which the prompt asks to list several items (e.g. "What are examples of fruits that are round?"). The "repetitive list attack" is done by making the answers verbose by repeating items multiple times, and then asking the LLMs to evaluate these augmented answers. If the LLM evaluates these "repetitive lists" to be better than the

original, the attack is considered a success. Their results show GPT-4 is significantly less prone to this attack with below 10% success rate, while GPT-3.5 and Claude-v1 both suffer over 90% success rate. Compared to this research, we expand the problem setting to general question-answering tasks. Huang et al. (2023) tackle verbosity bias in summarization tasks. They found that GPT-4 actually prefers short responses in faithfulness and coverage when it comes to summarization, although this is seen strongly only in single-answer grading, and not in comparison grading. This suggests that verbosity bias can be different between different tasks.

Compared to these studies, our problem setting is more general and we compare the verbosity preference between humans and LLMS. The experiments conducted in these papers measure the difference in evaluations when the texts are artificially made verbose while maintaining the same content. The assumption is that elongating the texts would have no effect on a true evaluator, so the difference in evaluation indicates verbosity bias. In our attempt to broaden the problem setting, we make use of human feedback as the oracle instead of making this assumption.

4   Verbosity Preference of LLMs

First, we experiment to see how much LLMs actually prefer longer answers. We ask GPT-4 to choose between pairs of responses and examine if it prefers longer responses or not. We did not limit our scope to prompts answered with a specific format (like lists in Zheng et al. (2023)) in order to observe the LLMs' general tendency to prefer longer answers.

We generated 100 sample answers each to 3 prompts, all from the same model (Vicuna-7b-v1.5) generated with the temperature parameter set to 0.7. One of the questions and two examples of the answers are as follows.

- Question: *Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions*
  1. I recently had the opportunity to travel to Hawaii, and it was an experience that I will never forget. From stunning beaches to vibrant culture, there was so much to explore and enjoy during my time on the islands...
  2. Hawaii is a tropical paradise, and I recently had the opportunity to experience it for myself. This archipelago is made up of eight main islands, each with its unique culture, traditions, and breathtaking landscapes. During my trip, I had the chance to visit several cultural sites, such as the Polynesian Cultural Center on Oahu...
  3. ...

The prompts are taken from the library introduced by Zheng et al. (2023), all from the "creative" category because 1) answers generated to other categories didn't vary in word count enough to see verbosity bias and 2) GPT-4 was not good at judging answers in those categories. We then take answers from these generated samples and insert them into the template shown in Figure 1. With the template complete, we asked GPT-4 to evaluate preferences between pairs of answers with the template. The outcome is either the first option selected, a draw, or the second option selected. In order to account for position bias, GPT-4 evaluated the pair twice with the position swapped the second time. It was considered a draw unless it gave the same result on both permutations.

The results are shown in Figure 3a for the overall result, and Figures 3b to 3d for results from each prompt. Both in the overall result and the individual results, there is a tendency for GPT-4 to prefer longer answers. When the word count difference is large enough, GPT-4 almost always prefers the longer answer. For question 1 and 2, the preference is smooth and clear, while for question 3, when the word count difference is small, there is high variance in evaluation. As we can see the shape varies between questions, and therefore we can deduce that verbosity does not rely entirely on word count and is different for each question. This makes adjusting for verbosity post-evaluation hard unless we know the verbosity preference shape for the prompt in question.

From this experiment, we can draw the conclusion that GPT-4 generally prefers longer answers among those that are generated by the same LLM with the same prompt. However, this experiment by itself does not indicate that GPT-4 suffers from verbosity bias; it could be that the longer answers generated by vicuna are actually higher in quality and helpfulness. In order to truly measure verbosity bias, we would need the ground truth of each comparison which we do not have. Instead, we next utilize a dataset of human evaluations as the baseline.

(a) Results with all questions combined.

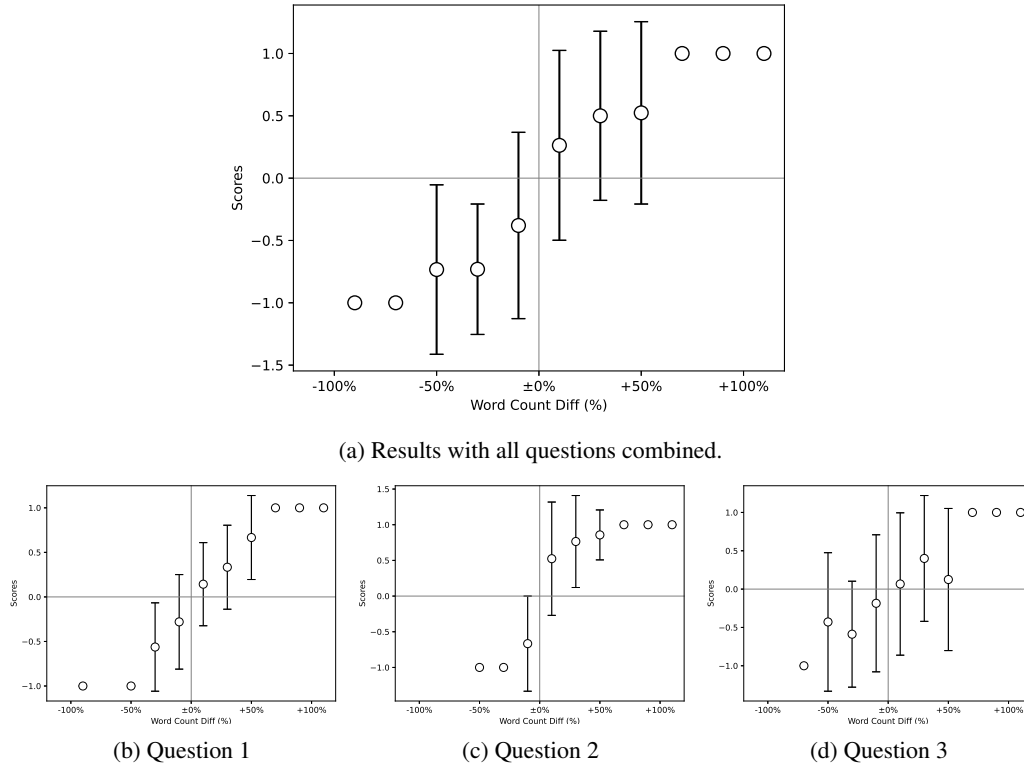(b) Question 1

(c) Question 2

(d) Question 3

Figure 3: Example experimental results for three questions (Question 1: Blog about trip to Hawaii, Question 2: Email to professor about paper, Question 3: Blog comparing smartphones). Figure 3a combines results from all three questions and Figures 3b to 3d show results for each indivisual quesion. X-axis is the percentage of the difference between the first and the second option compared to the length of the second option. Y-axis is the actual score, with 1.0 meaning the first option was selected, 0 meaning it was a draw, and -1.0 meaning the second option was selected. There is an positive correlation between word count difference between the two options and the resulting evaluation. The data points are binned with each ranging 20%. The circles represent the average in each range and the errorbars show the standard deviation.

5 Is There A Difference in Verbosity Preference Between LLMs and Humans?

Considering that LLMs replace humans as annotators in RLAIF, it is sufficient if LLMs could replicate human feedback and it does not necessarily have to be aligned with the ground truth. As seen in Figure 4 which plots verbosity preference of humans in the HH-RLHF dataset described later, humans seem to prefer longer answers as well. It is important to keep in mind that the aim of RLAIF is to replace human feedback with feedback from LLMs. Thus, the goal here is human alignemnt, and whether or not the longer answers are actually helpful is irrelevant as long as the LLM and the human come to the same conclusion. In light of this, we compare the difference in verbosity preference between LLMs and humans. We can view this as verbosity bias since the aim of LLM judgment in RLAIF is human alignment and not the eradication of biased preference in verbosity.

We use the HH-RLHF dataset (Bai et al., 2022a) which contains human feedback data comparing pairs of answers to a prompt. It only has one feedback data per prompt, so we cannot plot the verbosity preference of humans like in the experiment in the previous chapter. Instead we can see the dissimilarity between LLMs and humans in verbosity preference in general across various questions. Precisely, this experiment looks into the relationship between the difference in a number of words in the pair of responses and the human alignment of LLMs, meaning how often LLMs give the same judgment as humans.
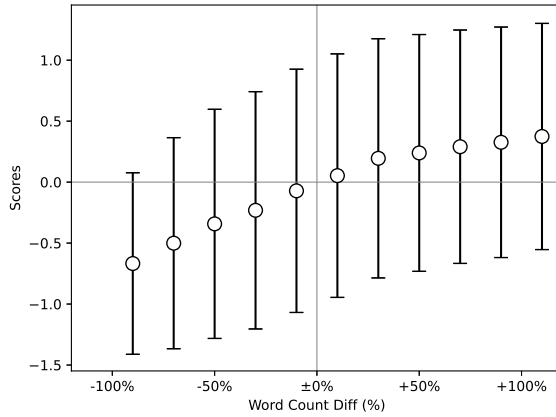
Figure 4: X-axis is the percentage of the difference between the first and the second option compared to the length of the second option. Y-axis is the resulting score by human judgement.
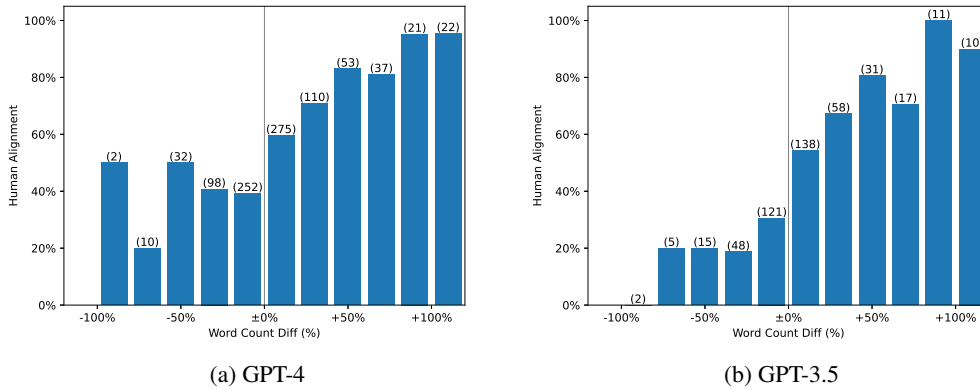

(a) GPT-4

(b) GPT-3.5

Figure 5: X-axis is the percentage of the difference between the chosen and the rejected option, compared to the length of the rejected option. Y-axis is the human alignment measured by the rate of LLM's decision agreeing with humans. The numbers in brackets indicate the sample size in each bracket.

We used the same prompt template as the previous experiment but asked GPT-4 to evaluate the whole conversation. Unlike the previous experiment which evaluated answers to a single question, HH-RLHF contains conversations between a human and an assistant. Therefore we asked GPT-4 to evaluate the pair of whole conversations and answer which assistant was more helpful.

In cases where human feedback preferred the longer answer, human alignment was high for the LLMs, meaning the LLMs preferred the longer answers as well. However, when human feedback chose the answer with fewer words, human alignment was low, because the LLMs still chose the longer answers regardless of the helpfulness of the shorter answer.

One possible explanation for this is that LLMs learned to mimic human behavior heuristically by choosing longer answers – in this dataset, human feedback did tend to favor longer responses as seen in Figure 4, and it is possible the dataset used to train GPT-3.5/GPT-4 had the same tendency. Nevertheless, a closer look into the cause is up for debate.

7

Table 1: Verbosity bias values calculated with (6) for GPT-4 and GPT-3.5 with data from experiment.

Model	GPT-4	GPT-3.5
Verbosity Bias	0.328	0.428

6 Formulation of Verbosity Bias

In the second experiment, we observed the tendency of LLMs to have low human alignment for cases where human feedback preferred shorter answers. In this section, we formulate verbosity bias to allow for quantative comparison between models.

In our problem setting, we define the given pair of text inputs as y_0 and y_1, the LLM outputs decision as $Y' \in \{0, 1\}$, and the more helpful option labeled by humans as $Y \in \{0, 1\}$. We define the sensitive attribute $S \in \{0, 1\}$ which equals 0 when y_0 has more words than y_1, and 1 when y_1 has more words than y_0.

With these definitions, equal opportunity (Hardt et al., 2016) with respect to sensitive attribute S is satisfied if

$$P(Y' = 0|S = 0, Y = 0) = P(Y' = 0|S = 1, Y = 0). \tag{3}$$

This only accounts for cases where human feedback prefers y_0. Although this can be attained by sorting the inputs beforehand, the equation can be generalized with accuracy parity instead of equal opportunity. Accuracy parity is satisfied if the accuracy of prediction is equal among both demographics:

$$P(Y' = Y|S = Y) = P(Y' = Y|S = 1 - Y). \tag{4}$$

The deviance from accuracy parity can be calculated with the following equation:

$$|P(Y' = Y|S = Y) - P(Y' = Y|S = 1 - Y)|. \tag{5}$$

Even though this is how the deviance is calculated in general, we thought it important that the directional information of the bias isn't lost. With the formulation below (6), a positive value indicates that the LLM prefers verbose answers, and a negative value indicates it prefers shorter answers. This distinction is crucial as some tasks may have a negative bias, for example in summarization tasks as shown in Huang et al. (2023). We also opted for the difference in *inaccuracy* between demographics

$$P(Y' = 1 - Y|S = 1 - Y) - P(Y' = 1 - Y|S = Y) \tag{6}$$

because verbosity bias refers to the inacuracy influenced by verbosity.

Table 1 shows the verbosity bias values of GPT-3.5 and GPT-4 calculated with data from Section 5. From these numbers, we can conclude that GPT-4 has improved in verbosity bias. Compared to Wang et al. (2023), which had a limited problem setting and gave the impression that GPT-4 is significantly less prone to verbosity bias, we see that the verbosity bias still exists for GPT-4. A further experiment on other LLMs for comparison is required.

7 Discussion

7.1 Other Metrics of Equality

In the context of our study, we treat the verbosity of the response pair as the sensitive attribute in our formulation of verbosity bias in Section 6. What verbosity differs from sensitive attributes generally discussed in other cases of biases is the fact that verbosity **should** actually be taken into consideration when evaluating the responses, whereas attributes like gender or race **shouldn't** be a factor in the outcome in other cases. This is why employing other metrics of equality like demographic parity doesn't make sense here, and therefore we base the measurement of verbosity bias on equal opportunity and accuracy parity.

7.2 Limitations of Our Experiments

In the experiment in Section 4, we generate the sample responses from the same questions (before concatenation of results from all three questions). However, in our experiment in Section 5, we mix together results from various questions. This has led us to only attain the result across many kinds of questions, not the result on any specific question like in the experiment in Section 4. It is debatable which of these results is preferable.

7.3 Limitation of Our Metric of Verbosity Bias

Our formulation of verbosity bias only accounts for bias between two groups divided by whether y_0 is longer than y_1. What it cannot detect is the bias within each of these groups; it is agnostic to the bias between cases where y_0 is barely longer than y_1 and cases where y_0 is significantly longer than y_0. Hence, if there were to be an instance where the model has high human alignment when there is a large difference in length between the pair of responses – the plot would have a concave shape symmetrical around the vertical line down the middle – our metric would suggest that the model has close to zero verbosity bias. To avoid such a situation, showing the human alignment plot alongside the metric is recommended.

8 Conclusion and Future Work

In this paper, we conducted experiments on the verbosity bias seen in LLMs' judgment by LLMs. In previous works, the problem settings were limited and did not compare the verbosity preference to humans. With our experiments, we saw that 1) LLMs tend to favor longer answers for creative writing tasks, and 2) alignment with humans varies on verbosity with lower human alignment in cases where humans preferred shorter answers. We then formulated verbosity bias based on accuracy parity that can be used to quantitatively compare verbosity biases among models.

Potential future directions of our research includes a deeper look into the possible causes of verbosity bias. This includes experiments on Supervised Fine-Tuning (SFT) models and reward models to determine how the verbosity preference by humans is passed down in the RLAIF learning process. Methods to mitigate the bias during training is also desired.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022a). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022b). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Huang, K.-H., Laban, P., Fabbri, A. R., Choubey, P. K., Joty, S., Xiong, C., and Wu, C.-S. (2023). Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. *arXiv preprint arXiv:2309.09369*.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

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

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.