Iter	Gemini	GPT3.5	GPT4	LLaMA2	Mixtral	DeepS
1st	93%	98%	100%	100%	100%	99%
5th	93%	97%	100%	100%	100%	98%

Table 2: We report human evaluation of format accuracy at six LLM's outputs. We observed that all LLMs have either perfect or nearly perfect format at first and fifth iteration of self-feedback at Yor-En translation. Mixtral stands for MixtralMOE and DeepS stands for DeepSeek-MoE that we used in the experiment.

is incorrect as it provides an alternative translation that does not match the source text." We conclude that this is due to their intrinsic instability of their instruction following capabilities. Gemini model contains surprisingly low format accuracy compared to other LLMs. This is due to the Gemini model refusing to generate any content that involves sensitive topics. There are 7 sentences in our testing set, Gemini refuses to provide responses. However, since our study focuses on self-bias amplification at iterations, this will not impact our experimental conclusions (The effects canceled out when comparing 1st and 5th iteration).

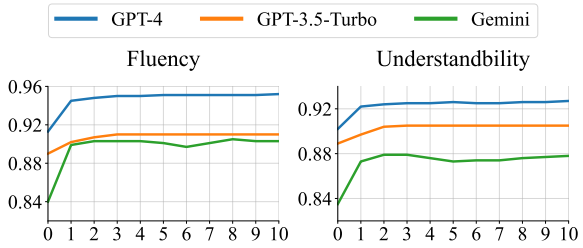

Figure 7: We measure the fluency and understandability aspects of GPT-4, GPT-3.5-Turbo, and Gemini's generated texts at Yor-En through self-refine steps. Despite no gains in quality, all LLMs have consistent performance improvements in fluency and understandability.

4.3 What improves after self-refinement?

Self-refinement can improve fluency and understandability but not quality. We demonstrate that LLM with biased feedback can impede the model's self-refine process. This raises a natural question: if an LLM does not improve its generation quality, does it improve in any other aspects throughout the iterative refine phase? To investigate this, we utilize the learned metric UniEval (Zhong et al., 2022) to measure the LLM's improvement beyond quality metrics. UniEval, a multidimensional learned metric, estimates various evaluation dimensions, including fluency, understandability, engagement and more. We focus on two dimensions, fluency and understandability, which UniEval is not trained on task-specific data. Our results, illustrated in Figure 6, show that GPT-4, GPT-3.5-Turbo, and Gemini consistently exhibit improvements in both fluency and understandability. This suggests an alternative perspective on the self-refine pipeline, indicating that while an LLM may not strictly adhere to instruction-following in terms of quality improvements, it can still improve certain intrinsic text qualities, such as fluency and understandability.

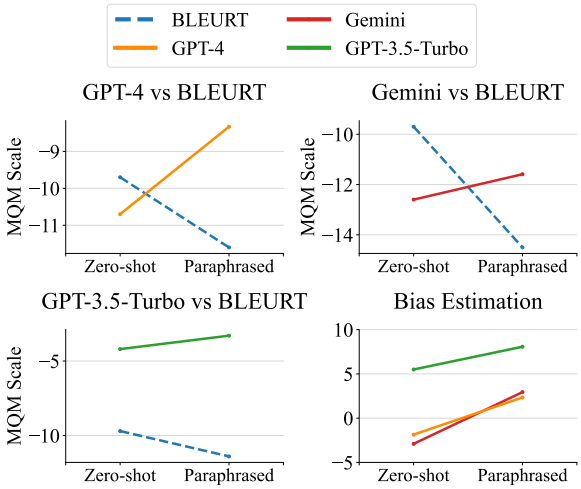

Figure 8: We used Madlad400-10b to translate 100 Yor-En translations and asked GPT-4, GPT-3.5-Turbo, and Gemini to paraphrase 100 translations. We show the BLEURT and LLM scores before and after paraphrasing. In the lower right of the figure, we show the bias estimation before and after paraphrasing. GPT-4 and Gemini have negative self-bias before paraphrasing. After paraphrasing, all LLMs increase their bias against their paraphrased outputs.

LLMs favor texts that follow their style. To explore this propensity, we conducted experiments to investigate if LLMs display a preference for outputs that align with their generation style. We asked the GPT4, GPT-3.5-Turbo, and Gemini model to paraphrase external translation outputs. In this prompt, LLMs aimed not to improve the quality of translations but rather to rewrite sentences in their corresponding styles. Using the multilingual translation system Madlad400-10b (Kudugunta et al., 2023), we produced 100 Yoruba-to-English translations. Subsequently, each LLM was instructed to paraphrase the generated sentences. Our findings, shown in Figure 8, reveal that GPT-4 and Gemini have negative self-bias before paraphrasing. However, after paraphrasing, all LLMs showed an in-

	DeepSeekMOE		MixtralMOE		LLaMA2-7B	
Sample Size	Bias	Dskew	Bias	Dskew	Bias	Dskew
1	14.8	0.735	12.4	0.483	8.75	0.491
4	16.1	0.795	10.1	0.490	14.1	0.580
8	16.7	0.800	13.0	0.610	19.8	0.810
16	18.0	0.830	16.9	0.730	20.7	0.840
32	18.5	0.840	18.5	0.790	20.9	0.850

Table 3: We report Bias and Dskew on Deepseek-MOE, MixtralMOE and LLaMA2-7B's self-feedback with varying sample size at Yor-En. Our results indicate that both bias and distance skewness tend to increase as the sample size grows larger.

creased bias against their paraphrased outputs. This is mainly attributed to a decline in quality performance post-paraphrasing, with LLMs erroneously perceiving these paraphrased outputs as indicative of improvements.

4.4 Self-Bias is Amplified at Self-Rewarding Pipeline

In this section, we will explore the concept of self-bias in the self-rewarding pipeline, as outlined in (Yuan et al., 2024). The pipeline begins with an instruction fine-tuned large language model (LLM). Initially, we generate k candidate responses for each input provided to the LLM. Next, the same LLM is used as a reward model to identify the best-performing candidate or to rank pairs within the collection of samples. Finally, various training objectives are applied to further train the LLM using the top-performing samples.

To illustrate the potential drawbacks of this pipeline, we carried out experiments on Yoruba to English translation task using three open-source LLMs: Deepseek-MOE, MixtralMOE, and LLaMA2-7B. For each source input, we sampled k candidate responses from each model. Subsequently, we obtained self-feedback scores on these candidates employing the prompt detailed in Section 4.1 and computed the corresponding self-bias. We varied k across $1, 4, 8, 16$, and 32 to examine the influence of sample size on the self-bias within the self-rewarding pipeline.

As shown in Table 3, we observed that all LLMs displayed an increase in bias and distance skewness as the sample size increased. This occurs when the LLM has a biased estimation of its self-feedback, and this bias can be amplified when the sample size is increased to find the top-performing candidate according to the self-feedback. Notably, selecting samples from a larger pool, e.g. a sample size of 32, significantly increases this bias compared to selections from a smaller pool, such as a sample size of 4. When the LLM optimizes over these samples, it can further increase its self-bias and generate samples that are biased by its self-feedback.

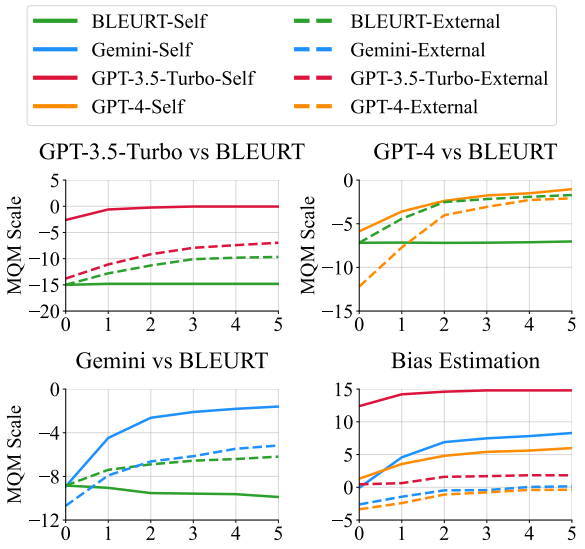

Figure 9: Using an external feedback model, we provide external feedback for GPT-4, GPT-3.5-Turbo, and Gemini in Yoruba-to-English translation task, across 5 refinement steps. We compare the models' true performance (measured by BLEURT) against external feedback-evaluated performance and self-feedback evaluated performance. Additionally, we plot the bias estimation for the three LLMs, considering both feedback types over 5 iterative refinement steps.

5 Alleviating Self-Bias

External Feedback Reduces Self-Bias. We demonstrated that self-feedback from a large language model can self-amplify bias with iterative refinement. We aim to answer if external feedback with low bias estimation can improve the model's generation performance and elicit self-correction capability. We leverage a reference-based feedback model, InstructScore (Xu et al., 2023), to provide external feedback. InstructScore will take in both reference and candidate text and output fine-grained feedback, including error location, severity label, and error type. To ensure a fair comparison, we parse all outputs with the same format as self-feedback. Since InstructScore can access reference text to provide feedback, we recognize this external feedback as oracle feedback. However, models will only receive information about error location, error type, and severity labels. Therefore, refinement

External Feedback Example at GPT-4

Yoruba text: Ní bayii a ni àwon eku oloshu merin ti ko ni dayabetesi telele to ti ni ayabetesi," o she afikun.
Reference English text: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.
(Red span indicates a major error and blue span indicates a minor error annotated by GPT-4.)

GPT-4's 1st generation [Human: -11, InstructScore: -10, Bias: 1]: "At this point, we have four rats without diabetes that have developed diabetes," he added.

GPT-4's 1st refinement [Human: -2, InstructScore: -6, Bias: -4]: "At this point, we have four mice without diabetes that were diabetic," he added.

GPT-4's 2nd refinement [Human: -1, InstructScore: -1, Bias: 0]: "We now have 4-month-old mice that are non-diabetic that were diabetic," he added.

Table 4: This case study demonstrates that external feedback (oracle) from InstructScore (Xu et al., 2023) can remain low self-bias during iterative self-refine. By providing accurate error type, error location, and severity labels, InstructScore effectively elicits GPT-4's self-correction capability and improves its translation quality. Despite InstructScore's oracle-like role (which it can access reference text to make error annotations), it does not provide explicit corrections, requiring GPT-4 to rely on its internal knowledge for corrections.

still relies on LLM's self-correction capability.

In Figure 9, we demonstrate that external feedback with accurate assessment can significantly lower the model's bias at iterative refinement (shown at the lower right of the figure. All dotted curves are below solid curves with corresponding colors). Interestingly, both Gemini and GPT-4's bias estimation is improved throughout the refinement process, as the external feedback model can over-penalize low-quality outputs. As refinement proceeds, the external feedback model converges to BLEURT quality assessment that samples achieve improved quality. Most importantly, we demonstrate that all LLMs with external feedback can elicit their self-correction ability with consistent BLEURT improvements at self-refine iterations. We include a case study example in Table 4. Our finding of model improvement is consistent with prior study (Xu et al., 2024) and we further demonstrate that external feedback can significantly reduce self-bias.

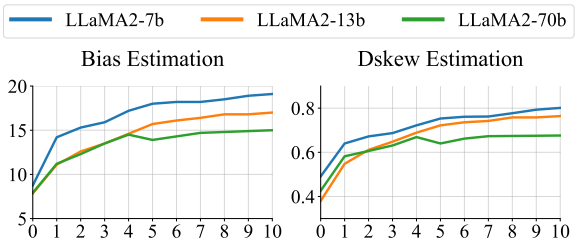

Figure 10: We show that bias and distance skewness estimation on LLaMA-2 7B, 13B, and 70B models at Yor-En translation across self-refinement steps. LLM with larger parameter size can have less self-bias.

Larger Model Reduces Self-Bias. In Figure 10, we demonstrate that LLMs with larger parameter size can have less self-bias throughout self-refinement steps. Specifically, we tested the LLaMA2 models with 7B, 13B, and 70B parameters on Yoruba-to-English (Yor-En) translation tasks. Our findings indicate that while the LLaMA2-70B model exhibits self-bias in the earlier iterations, its self-bias begins to plateau after the 5th iteration. In contrast, the 7B and 13B models continue to amplify their self-bias in later iterations. This observation aligns with prior work (Huang et al., 2023a), which posited that larger LLMs possess better self-refinement capabilities. Our study contributes to this discussion from the perspective of self-bias, proposing that larger LLMs are more resilient to self-bias. Consequently, they can assess their own outputs more accurately and possess a greater capacity for self-correction.

6 Conclusion

In this study, we define and quantify self-bias in LLMs with two principled estimated statistics. Our experiments across six LLM families, four languages, and three tasks reveal that self-bias is prevalent in self-refine or self-rewarding pipelines. This biased self-feedback leads to false positive objectives, hindering performance improvements during iterative refinement. Further analysis reveals that while LLM improves fluency and understanding of its generated text, they do not necessarily progress in the intended direction, such as improving quality in machine translation or expanding coverage in concept-to-word generation. Instead, LLMs tend to favor texts that adhere to their inherent styles. Finally, our research suggests that larger models are more resistant to self-bias, and incorporating external feedback significantly reduces bias, leading to performance improvements in LLMs.