When LLMs Benchmark Themselves: Deconstructing Self-Bias in Automated Evaluation

Anonymous ACL submission

Abstract

As LLMs rapidly saturate existing benchmarks, automated benchmark creation using LLMs (LLM-as-a-benchmark)—where a model generates test inputs (LLM-as-a-testset) and evaluates outputs (LLM-as-an-evaluator)—has gained traction as a cheap alternative to human curation. We show that this paradigm has a fundamental problem: *LLM-generated benchmarks systematically favor the model that created them*. Using machine translation as our primary testbed, we find that self-bias arises from two additive sources, LLM-as-a-testset and LLM-as-an-evaluator, and their combination amplifies the effect. Crucially, even when test data is generated with explicit diversity controls, each model's implicit stylistic tendencies produce homogeneous, model-specific outputs that inflate its own scores. Increasing source text diversity, using our proposed diversity metric, partially mitigates this bias. Self-bias is strong enough to cause each model to rank itself first, overriding the peer-consensus ordering. We confirm that the phenomenon extends to open-ended generation on the Chatbot Arena task.

1 Introduction

Keeping NLP benchmarks current is harder than it used to be. LLMs saturate leaderboards within months of a benchmark's release, yet building a new human-curated test set is expensive and slow—and for languages or domains where labeled data is scarce, it may simply not be feasible. Automated benchmark creation using LLMs (LLM-as-a-benchmark) (Farchi et al., 2024; Maheshwari et al., 2024; Pombal et al., 2025a) has emerged as a practical response: a single model generates test inputs (LLM-as-a-testset) and scores the outputs (LLM-as-an-evaluator), producing benchmarks on demand at low cost.

The problem is that this convenience comes with a hidden cost. Prior work has shown that LLM-based evaluators tend to favor their own outputs (Panickssery et al., 2025; Chen et al., 2025; Xu et al., 2024), but the bias introduced by the test set generation step—and how it interacts with evaluator bias—has not been carefully studied. If the source texts a model generates are already tailored to its own strengths, performance scores will reflect the model's idiosyncrasies rather than its actual quality.

We study this problem and make three contributions. **(1)** We formally define self-bias using the estimator-bias framework (Xu et al., 2024), decomposing it into additive LLM-as-a-testset and LLM-as-an-evaluator components and showing that their combination amplifies the effect. Self-bias is strong enough that every model ranks itself first, overriding the peer-consensus ordering. **(2)** We trace self-bias to its mechanistic cause: even under high-quality prompts with explicit diversity controls (Pombal et al., 2025b), each model generates repetitive, model-specific source texts—text it can then translate and score more easily. This effect is strongest in low-resource languages, where limited generation capacity causes outputs to converge into distinctive degeneration patterns. We also verify that self-bias appears in LLM-as-a-benchmark on the Chatbot Arena task, showing the problem is not specific to translation. **(3)** We provide practical guidance for reducing self-bias: filtering source texts using our proposed diversity metric substantially reduces bias; generating benchmark data in languages where models have high proficiency minimizes model-specific signatures; and using frontier models to benchmark open-source models remains reliable when the quality gap is large.

2 Preliminaries and Self-Bias Definition

LLM-as-a-testset We automatically generate a test set using a generator model, M_{test}. For a given instruction s, M_{test} produces a source text x and an

optional reference text y'. The source text x is then used as a prompt for the target model under evaluation, M_{target}, which generates an output y. Finally, an evaluation metric computes a quality score for y by comparing it against source text x. We investigate testset generation self-bias: a generated test set may systematically favor its generator model (M_{test}), inflating that model's scores.

LLM-as-an-evaluator An evaluator model, $M_{\text{evaluator}}$, assesses the quality of an output y generated by a target model M_{target} in response to a prompt x , assigning it a score based on predefined criteria. In our experimental setup, the prompts x are sourced from human-authored benchmark datasets. We investigate evaluator self-bias: an LLM-based evaluator may assign higher scores to its own model's outputs than warranted.

LLM-as-a-benchmark combines LLM-as-a-testset and LLM-as-an-evaluator: a single model generates source text x, prompts a target model M_{target} to produce output y, and then evaluates y. By unifying both roles in one model, LLM-as-a-benchmark inherits—and potentially compounds—the biases of each component.

We operationalize self-bias as follows:

$$\text{bias}_{M_i} = \underbrace{\theta_{M_i,M_i}}_{\text{self-ranking}} - \underbrace{\frac{1}{|M|-1}\sum_{M_o \neq M_i} \theta_{M_i,M_o}}_{\text{ranking by other models}} \tag{1}$$

where M is a set of models, M_i is the model being evaluated, and M_o are peer models. θ_{M_i,M_o} is the *ranking* of model M_i as determined by model M_o, averaged across segments. We use rankings rather than raw scores so that self-bias is directly comparable across settings where quality metrics operate on different scales. θ_{M_i,M_i} is the self-ranking and $\frac{1}{|M|-1}\sum_{M_o \neq M_i} \theta_{M_i,M_o}$ is the peer-consensus ranking, which we treat as a proxy for true performance. A negative bias_{M_i} indicates that M_i ranks itself more favorably than its peers do; a more negative value reflects stronger self-preference.

In our study, M_o plays the role of test set generator (LLM-as-a-testset), evaluator (LLM-as-an-evaluator), or both (LLM-as-a-benchmark).

3 Experimental Setup

We study self-bias through machine translation, which offers established quality metrics (Juraska et al., 2024; Kocmi and Federmann, 2023; Xu et al., 2023; Lavie et al., 2025), annually updated test sets with official human rankings for our models (Kocmi et al., 2025) as ground truth, and the ability to study self-bias across diverse language pairs. We validate the generality of our findings on open-ended generation via the Chatbot Arena task (Section 6.2). For translation, our experiments encompass six low-to-medium resource language directions: Bemba−English, Kurdish−English, Aymara−English, Luo−English, English−Bemba, and English−Aymara. We focus on low-to-medium resource languages where model quality differences are larger; on high-resource pairs, all models score near-perfectly and self-bias becomes difficult to measure (See Appendix Section E for detailed discussion). For the Chatbot Arena task, we contrast high-resource (English, Korean) and low-resource (Bemba, Aymara) prompt languages.

Model selection. We select Gemini 2.5 Pro, GPT-4.1, and Claude Opus 4 for two reasons. First, they are the top-performing systems at WMT25 (Kocmi et al., 2025), for which official human rankings can still distinguish their translation quality. Second, all three have been used in the zero-shot benchmarking framework of Pombal et al. (2025a), demonstrating their capability as LLM-as-a-benchmark generators. In preliminary experiments, multilingual open-source models (Gemma3-27B, Mistral-Large-2411, Qwen3-32B) failed to generate high-quality benchmark data for low-resource languages.

Evaluation setup. We evaluate models under the three conditions defined in Section 2. For LLM-as-a-testset, translation quality is measured by MetricX-QE (Juraska et al., 2024), a SOTA reference-free translation metric where lower scores indicate higher quality. For LLM-as-an-evaluator, source texts are drawn from the FLORES-200 benchmark (Team et al., 2022) to ensure human-authored test set inputs, so that the LLM acts solely as an evaluator. For LLM-as-a-benchmark, models generate source texts and evaluate translations; a fully reference-free generation variant is analyzed in Section 5. We additionally use chrF (Popović, 2015) to measure textual similarity across source texts generated by different models. As described in Section 2, all raw scores are normalized into per-segment system rankings to ensure comparability across settings.

For LLM-as-a-testset and LLM-as-a-benchmark, we generate 200 source texts per language direction; for LLM-as-an-evaluator, we sample 200

FLORES examples per direction. In all cases, translations are obtained from all three LLMs. **All evaluations are reference-free.** All prompts are provided in Appendix Section C. Our sample size of N=200 is well justified: bootstrap resampling (B=10,000) shows that confidence intervals for self-rankings and peer-rankings do not overlap, confirming statistically significant self-bias at this sample size (Appendix Section A).

4 Does the use of LLM-as-a-benchmark introduce self-bias?

We investigate self-bias using machine translation, where source and target languages can be precisely controlled. To generate our benchmark data, we employ the high-quality prompt proposed by Pombal et al. (2025a) in their framework. This prompt explicitly controls for diversity and quality by instructing models to generate source-reference text pairs covering diverse topics, varying in length and complexity, and targeting different styles (see Appendix Section C for details). Despite these explicit controls, we find that self-bias persists due to implicit model-specific features in the generated texts.

Table 1 quantifies self-bias for Gemini-2.5-pro, GPT-4.1, and Claude-Opus-4 across XX→En directions. In our framework, a negative bias score indicates that a model ranks itself more favorably than its peers do. The consistently negative diagonal scores confirm that **these models systematically favor their own outputs** across four language directions—except Gemini on Kurdish→English. We hypothesize that this exception arises because Kurdish is a relatively higher-resource language compared to Bemba, Aymara, and Luo, leading to more diverse and less model-specific source text generation (further analyzed in Section 5.2).

Self-bias causes ranking flips. To demonstrate the practical impact of self-bias, we compare the rankings produced by each model's own LLM-as-a-benchmark. Under LLM-as-a-benchmark, **each model ranks itself first**: as shown in Table 2, the diagonal entries (bolded) are always the lowest average rankings in each column, meaning each benchmark ranks its own translations as best. The per-language bias scores in Table 1 confirm this pattern: the diagonal entry is consistently the most negative in each column. For example, on Bemba→English, Claude-as-benchmark ranks Claude first (-0.617) even though its peer models disagree. This rank-flip effect—where self-bias overrides the peer-consensus ordering—is the most direct evidence that LLM-as-a-benchmark distorts practical model selection decisions.

Bemba→En.		LLM-as-a-benchmark		
		Gemini	**GPT**	**Claude**
Translator	**Gemini**	-0.591	0.161	0.430
	GPT	-0.145	-0.202	0.347
	Claude	0.102	0.515	-0.617
Aymara→En.		**LLM-as-a-benchmark**		
		Gemini	**GPT**	**Claude**
Translator	**Gemini**	-0.157	0.180	-0.022
	GPT	0.232	-0.315	0.083
	Claude	0.253	0.365	-0.617
Luo→En.		**LLM-as-a-benchmark**		
		Gemini	**GPT**	**Claude**
Translator	**Gemini**	-0.300	0.120	0.180
	GPT	0.385	-0.508	0.123
	Claude	0.188	0.352	-0.540
Kurdish→En.		**LLM-as-a-benchmark**		
		Gemini	**GPT**	**Claude**
Translator	**Gemini**	0.005	0.215	-0.219
	GPT	0.173	-0.150	-0.023
	Claude	-0.095	0.520	-0.425

Table 1: Self-bias (LLM-as-a-benchmark) across four XX→En directions. Diagonal entries are always most negative: each model ranks itself first, overriding the peer-consensus ordering.

LLM as MT system	LLM-as-a-benchmark		
	Gemini	**GPT**	**Claude**
Gemini	**1.222**	1.509	1.457
GPT	1.525	**1.222**	1.506
Claude	1.564	1.782	**1.123**

Table 2: Rankings converted from LLM-as-an-evaluator scores (lower=better), averaged over four XX→En directions. Each column represents a different LLM acting as the benchmark. The diagonal (bolded) is always the lowest in each column: every model ranks itself as the best MT system on its own benchmark.

Decomposition of self-bias. To understand what drives self-bias, we decompose LLM-as-a-benchmark into its two constituent components (Table 3): LLM-as-a-testset, where models generate test sets and translation quality is scored by MetricX-QE, and LLM-as-an-evaluator, where models act as judges for translations of human-authored FLORES test data. **Both components independently exhibit measurable self-bias**. Crucially, LLM-as-a-benchmark consistently shows larger bias than either component alone, indicating that testset generation bias and evaluator bias are additive—the combination amplifies the distortion

XX→En	Gemini	GPT	Claude
LLM-as-a-testset	-0.124	-0.239	-0.093
LLM-as-a-benchmark	-0.261	-0.294	-0.550
LLM-as-an-evaluator	-0.302	-0.443	-0.303

Table 3: Self-bias under LLM-as-a-testset, LLM-as-a-benchmark, and LLM-as-an-evaluator (averaged XX→En). All models self-favor in every condition; LLM-as-a-benchmark compounds both sources of bias.

Lang	Setup	Gemini	GPT	Claude
Bemba→En	Src-only	8.49	5.33	4.07
	Src+Ref	2.73	4.76	3.89
Aymara→En	Src-only	12.9	10.4	8.29
	Src+Ref	7.54	9.60	8.13

Table 4: Src+Ref generation yields better translation quality than src-only (MetricX; lower is better), indicating co-generation biases sources toward the model's translation strengths.

beyond what either source produces in isolation.

Where does the self-bias stem from? Co-generating source texts and translation references, as in (Pombal et al., 2025b), can introduce two biases: (1) a *translatability bias*, where the model avoids source texts it cannot translate well, and (2) a *dialect bias*, where each model generates text in its own stylistic "dialect." We confirm translatability bias by comparing MetricX-QE scores for src-only vs. src+ref generation (Table 4): src+ref consistently yields better scores, indicating that **co-generation tailors sources to the model's translation strengths**.

Our primary interest is the more fundamental dialect bias. Each LLM generates text in its own implicit "dialect"—characteristic n-gram distributions and stylistic patterns. This dialect advantages the same model in two ways: (1) the model translates text in its own dialect more accurately, and (2) since LLMs can recognize their own outputs (Panickssery et al., 2025), they score them more favorably as evaluators. In high-resource languages, all models produce diverse text and dialect differences are small. In low-resource languages, each model falls back on a narrow set of patterns, making its outputs distinctive and amplifying self-bias. To isolate dialect bias from the translatability confound, we study source-only generation in subsequent sections.

Bemba→En	LLM-as-a-benchmark		
Translator	**Gemini**	**GPT**	**Claude**
Gemini	-0.515	0.055	0.460
GPT	0.235	-0.253	0.018
Claude	0.220	0.543	-0.763

Aymara→En	LLM-as-a-benchmark		
Translator	**Gemini**	**GPT**	**Claude**
Gemini	-0.293	0.285	0.008
GPT	0.308	-0.330	0.022
Claude	0.015	0.473	-0.488

Table 5: LLM-as-a-benchmark self-bias with src-only generation for Bemba→En and Aymara→En. Self-bias persists on the diagonal for all three models.

5 How does source text generation impact self-bias?

This section traces self-bias to the properties of generated source texts. We show that src-only generation still exhibits self-bias (Section 5.1), that each model's implicit linguistic fingerprint produces homogeneous outputs despite explicit diversity controls (Section 5.2), that this lack of diversity drives self-bias (Section 5.3), and that improving diversity partially mitigates it (Section 5.4). We focus on Aymara→English and Bemba→English.

5.1 Source text generation leads to self-bias

As Table 5 shows, a measurable self-bias persists even when the data generation pipeline is constrained to produce only source texts without co-generated references. The negative bias scores appear exclusively on the diagonal, confirming that self-bias is inherent to each model's generated texts. Source-only generation thus does not eliminate self-bias; the model's inherent stylistic tendencies in monolingual generation are sufficient to create evaluation advantages for its own translation system.

5.2 What makes source text generation exhibit bias?

We investigate how each LLM's source text characteristics contribute to self-bias. Each model's source texts are more similar to its own other generated texts—even those for entirely different topics—than to texts from other models. To quantify this within-model versus cross-model similarity, we define a chrF@K metric.

Let $M = \{M_1, M_2, M_3\}$ be the set of LLMs. Each model $M_k \in M$ generates N source texts $S_k = \{s_{k,1}, \ldots, s_{k,N}\}$, where the index j in $s_{k,j}$ denotes the topic ID. For a source text $s_{A,i} \in S_A$, we compute pairwise chrF scores against all $s_{B,j} \in$

Bemba→En	chrF across prompts + models		
Data from	**Gemini**	**GPT**	**Claude**
Gemini	35.76	32.70	32.03
GPT	32.28	37.78	31.63
Claude	36.42	35.99	38.64

Aymara→En	chrF across prompts + models		
Data from	**Gemini**	**GPT**	**Claude**
Gemini	37.78	32.19	32.58
GPT	33.52	40.79	32.44
Claude	39.56	37.04	42.28

Table 6: Average chrF@K similarity scores. Diagonal (within-model) entries are consistently higher than off-diagonal (cross-model) entries, showing each model's outputs are more similar to itself than to other models.

S_B (excluding self-matches when $M_A{=}M_B$), select the $K{=}5$ highest scores, and average them:

$$\text{chrF@K}(s_{A,i}, M_B) = \frac{1}{K}\sum_{k=1}^{K}\text{chrF}_k \qquad (2)$$

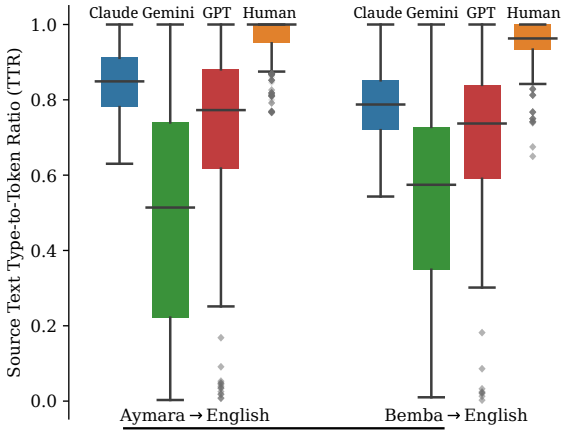
Averaging over all $s_{A,i} \in S_A$ yields the model-level similarity:

$$\text{chrF@K}(M_A, M_B) = \frac{1}{N}\sum_{i=1}^{N}\text{chrF@K}(s_{A,i}, M_B) \qquad (3)$$

This process yields three scores for each model M_A: one for its average within-model similarity (diagonal, $M_A{=}M_B$), and two for its cross-model similarities to the other LLMs (off-diagonal). Table 6 presents the results. The diagonal entries—representing within-model similarity—are consistently higher than off-diagonal entries across all models and language pairs. This demonstrates that each model exhibits significantly higher textual similarity to its own generated content, even across different topics, compared to texts generated by other models. Rather than producing truly diverse source texts following the prompt instructions, the models tend to reproduce content and style patterns from their limited knowledge, especially for low-resource languages like Bemba and Aymara.

5.3 Limited Diversity in Low-Resource Source Texts

We hypothesize that the elevated within-model similarity stems from limited lexical diversity in low-resource language generation. We measure two complementary metrics: the Type-Token Ratio (TTR), which captures the proportion of unique words in a text and serves as a standard measure of lexical diversity, and the degeneration ratio, which quantifies the frequency of pathological repetition in generated outputs.



Model	Degeneration (%)
Gemini	22.3
GPT4.1	5.8
Claude	0

Figure 1: TTR of LLM-generated vs. human (FLORES) source texts. All LLMs are less lexically diverse than humans, and each model exhibits a distinct TTR profile. Degeneration (% texts with ≥10 repeating 4-grams) is worst for Gemini (22.3%).

Figure 1 illustrates the TTR distributions and degeneration ratios across the three models. Each model exhibits a visually distinct TTR profile, confirming that their lexical diversity patterns diverge substantially. Notably, all three LLMs—including Claude, which shows the highest diversity—are less lexically diverse than human-written benchmarks[1], supporting prior work (Yu et al., 2023) on data diversity's role in mitigating systematic bias.

Further inspection reveals that the notably low diversity in Gemini-2.5-pro and GPT-4.1 outputs often stems from text degeneration. **We quantify degeneration by counting repeating n-grams, marking texts with ≥10 repeating 4-grams as degenerated.** As shown in Figure 1, both models exhibit substantial degeneration levels (22.3% for Gemini, 5.8% for GPT-4.1), while Claude shows none. Crucially, degeneration can be a direct driver of self-bias: when a model translates its own degenerated source texts, translation quality improves compared to when other models translate the same inputs (see Appendix Section D). This asymmetry—where the generating model handles its own degeneration artifacts better than its competitors—

[1]We use 200 source texts from FLORES-200 Aymara and Bemba as the human baseline.

Subset	Self-bias Estimation Gemini	GPT	Claude
Max chrF	-0.400	-0.280	-0.685
Random	-0.342	**-0.256**	-0.616
Min chrF	**-0.250**	-0.265	**-0.600**

Table 7: Self-bias of LLM-as-a-benchmark for subsets of LLM-generated source texts: those with the highest within-model chrF similarity (lowest lexical diversity), those with the lowest within-model chrF similarity (highest lexical diversity), and a randomly selected baseline. Diverse source texts mitigate self-bias. Self-bias is averaged for Aymara to English and Bemba to English.

Lang	Direction	Gemini	GPT	Claude
Bemba	En→Bemba	-0.15	0.10	0.14
	Bemba→En	-0.59	-0.20	-0.62
Aymara	En→Aymara	0.02	-0.09	0.07
	Aymara→En	-0.16	-0.32	-0.62

Table 8: LLM-as-a-benchmark self-bias by translation direction. XX→En directions show substantially more self-bias (< -0.15) than En→XX (consistently < 0.15).

Setting	Direction	Gemini	GPT	Claude
As-a-testset	En→XX	0.02	0.05	-0.08
	XX→En	-0.17	-0.24	-0.09
As-an-evaluator	En→XX	-0.11	0.05	0.10
	XX→En	-0.30	-0.44	-0.30

Table 9: Self-bias decomposed into LLM-as-a-testset and LLM-as-an-evaluator by direction. Both components show stronger bias for XX→En than En→XX.

inflates its translation scores on its own test set, which contributes to the self-bias we observe.

5.4 Improving diversity reduces self-bias

We hypothesize that the lack of diversity in generated source texts is a key contributor to self-bias, and test this by conducting an ablation study using our within-model similarity metric. As defined in Equation (3) (with $M_A{=}M_B$), a high within-model similarity score indicates a model's tendency to repeat its own content patterns. For each of the three LLMs, we selected three subsets of 50 source texts from the total 200: those with the highest within-model chrF@K similarity (representing the least lexically diverse texts), those with the lowest similarity (representing the most diverse texts), and a randomly selected subset as a control.

Table 7 demonstrates that source texts with high within-model similarity (Max chrF) consistently exhibit the highest self-bias across all three LLMs, surpassing the bias observed in the random 50-sample baseline. Conversely, source texts with low within-model similarity (Min chrF) consistently lead to reduced or comparable self-bias relative to the random control. This pattern confirms that **generating more diverse source texts can mitigate self-bias**, pointing toward a practical avenue for improving LLM-as-a-benchmark reliability.

6 Self-Bias is Driven by Language Proficiency

Self-bias stems from limited diversity in model-generated texts—but *when does this diversity collapse, and what governs its severity?* We show that language proficiency is the key factor. When models generate text in languages where they are highly proficient (e.g., English), their outputs are lexically varied and stylistically diverse across models, substantially reducing model-specific signatures and, consequently, self-bias. Conversely, in low-proficiency languages, models default to a narrow set of patterns, amplifying self-bias. We demonstrate this through translation direction asymmetry (Section 6.1) and cross-task validation on open-ended generation via Chatbot Arena (Section 6.2).

6.1 Translation Direction Asymmetry

Table 8 shows a striking asymmetry: En→XX directions exhibit substantially lower self-bias than XX→En. The magnitude of self-bias for English→Aymara and English→Bemba is consistently less than 0.15 in absolute value, whereas the XX→En directions show much larger effects. Table 9 decomposes LLM-as-a-benchmark into its LLM-as-a-testset and LLM-as-an-evaluator components by direction. Both components consistently exhibit more pronounced self-bias in the XX→En direction than in En→XX. This suggests that when generating source texts in low-resource languages, the LLM-as-a-testset produces texts with stronger model-specific linguistic features that advantage its own translation system. Simultaneously, the LLM-as-an-evaluator appears more attuned to these self-generated patterns when judging XX→En outputs, leading to systematically inflated scores.

Why does translation asymmetry exist for self-bias? Table 10 reports pairwise Cohen's D (Cohen, 1988) between models' TTR distributions, measuring how much each pair of models diverges in their lexical diversity profiles ($D{>}0.5$ indicates

	Gem.×Claude	Gem.×GPT	Claude×GPT
English→XX	0.076	0.190	0.111
XX→English	**1.487**	**0.777**	**0.663**

Table 10: Lexical diversity divergence across models (pairwise TTR Cohen's D). English source texts show near-zero divergence, while low-resource source texts diverge substantially (D>0.5).

substantial divergence; see Appendix Section B).

En→XX (low self-bias): When generating English source texts, all three models are highly proficient and follow the diversity prompt effectively. As a result, all three produce lexically diverse texts—their TTR distributions are independently high and look similar across models, yielding low pairwise Cohen's D (D<0.2 for all pairs). Importantly, this does *not* mean the models produce the same texts; it means that none of the models leave distinctive low-diversity fingerprints. Without model-specific signatures, translators and evaluators cannot favor their own sources over competitors', suppressing self-bias.

XX→En (high self-bias): When generating low-resource language texts, limited proficiency causes each model to fall back on its own narrow, repetitive patterns—model-specific artifacts. Each model's outputs become low in TTR, but the specific degeneration patterns differ across models (e.g., Gemini degenerates 22.3% of texts, Claude 0%). This creates *divergent* low-diversity TTR distributions, reflected in high pairwise Cohen's D (D>0.5 for all pairs). These distinctive fingerprints give the generating model a systematic advantage: its translator handles its own artifacts better than other models' translators do, and its evaluator recognizes its own stylistic patterns more favorably. This two-sided advantage compounds into the large self-bias we observe for XX→En. Correspondingly, Appendix Table 15 shows that within-model chrF similarity is not elevated for English sources, whereas XX→En diagonal entries are consistently higher, confirming the presence of model-specific fingerprints only in the low-resource direction.

6.2 Cross-Task Validation: Chatbot Arena

To confirm generality, we apply LLM-as-a-benchmark to the Chatbot Arena task (Chiang et al., 2024) using prompts from Pombal et al. (2025b) (details in Appendix Section C), comparing high-resource (English, Korean) versus low-resource (Bemba, Aymara) settings.

English	LLM-as-a-benchmark		
Generator	**Gemini**	**GPT**	**Claude**
Gemini	-0.018	-0.122	0.140
GPT	0.035	-0.205	0.170
Claude	0.103	-0.107	0.005

Korean	LLM-as-a-benchmark		
Generator	**Gemini**	**GPT**	**Claude**
Gemini	0.068	-0.060	-0.007
GPT	0.107	-0.208	0.100
Claude	0.105	-0.068	-0.038

Bemba	LLM-as-a-benchmark		
Generator	**Gemini**	**GPT**	**Claude**
Gemini	-0.385	-0.010	0.396
GPT	0.052	-0.218	0.167
Claude	0.334	-0.154	-0.180

Aymara	LLM-as-a-benchmark		
Generator	**Gemini**	**GPT**	**Claude**
Gemini	-0.216	0.346	-0.130
GPT	0.247	-0.413	0.165
Claude	0.177	-0.108	-0.069

Table 11: LLM-as-a-benchmark self-bias on Chatbot Arena. Low-resource prompts (Bemba, Aymara) elicit much stronger self-bias than high-resource prompts (English, Korean).

As shown in Table 11, results closely mirror our translation findings. Self-bias is near-zero in high-resource settings (Gemini: 0.025, Claude: −0.017) but increases sharply for low-resource prompts (−0.301 and −0.125 respectively). **This confirms that self-bias driven by low LLM proficiency generalizes beyond translation.**

7 Merits of LLM-as-a-benchmark

Despite self-bias concerns, LLM-as-a-benchmark retains practical value in two settings.

LLM-as-a-benchmark benefits open-source model evaluation. Table 12 shows that frontier LLMs consistently rank open-source models (Gemma3-27B, Mistral-large-2411, Qwen3-32B) with low bias—except GPT-4.1's specific bias toward Qwen3-32B. When the quality gap between generator and evaluated models is large, self-bias does not distort rankings. However, bias remains substantial when frontier models rank each other.

High-resource languages reduce self-bias. Self-bias is substantially reduced for high-resource language pairs (Section 6), and applying diversity filtering to source texts mitigates it further (Section 5.4). Practitioners should exercise particular caution when using LLM-as-a-benchmark for low-resource languages or when comparing models of

Translator	Bemba	LLM-as-a-benchmark (Rank)		
		Gemini	**GPT**	**Claude**
	Gemma3	1.125	1.110	1.160
	Mistral	1.675	1.623	1.630
	Qwen3	2.335	2.558	2.285

Translator	Bemba	LLM-as-a-benchmark (Bias)		
		Gemini	**GPT**	**Claude**
	Gemma3	-0.01	0.04	-0.03
	Mistral	0.05	-0.02	-0.03
	Qwen3	-0.09	-0.16	0.25

Table 12: Frontier models rank open-source models (Gemma3-27B, Mistral-large-2411, Qwen3-32B) consistently and with low bias (Pombal et al., 2025b), except GPT-4.1's bias toward Qwen3-32B. Self-bias remains significant when frontier models rank each other.

similar capability.

8 Related Work

Automatic Benchmark Creation. As existing benchmarks become increasingly saturated by rapid LLM advancements (Glazer et al., 2025), automatic benchmark construction using LLMs has gained significant traction (Pombal et al., 2025a). Methods range from synthesizing test sets via prompt workflows (Sprague et al., 2024; Zouhar et al., 2025) to constructing prompts from existing data (Li et al., 2024; Vergopoulos et al., 2025). Efficacy is assessed via benchmark agreement (Perlitz et al., 2024) or ranking correlation with human benchmarks (Pombal et al., 2025a), using accuracy for verifiable tasks or LLM judges for open-ended outputs (Xu et al., 2023; Pombal et al., 2025a). A key advantage is the ability to generate evaluation data for arbitrary domains and language pairs on demand, which is particularly valuable for settings where human data curation is expensive or difficult.

Self-bias in LLM Evaluation. The LLM-as-a-benchmark paradigm is susceptible to self-bias from two sources: the LLM acting as an evaluator and the LLM generating the test set. A well-documented issue is the tendency of LLM judges to systematically favor their own outputs (Xu et al., 2024; Panickssery et al., 2025), attributed to familiarity with their own stylistic patterns or a preference for low-perplexity text (Wataoka et al., 2025). While this preference may sometimes reflect genuine quality (Chen et al., 2025), judge reliability is questionable for problems beyond the model's own capability (Krumdick et al., 2025). Our work extends this line of research by investigating the overlooked self-bias introduced by testset generation and, crucially, the additive effects when both sources of bias combine. Unlike Yuan et al. (2025), whose mitigation strategies target verifiable tasks like math reasoning, we address generative settings where evaluation itself requires subjective judgment.

9 Conclusion

We formally define and quantify self-bias in LLM-as-a-benchmark, decomposing it into additive LLM-as-a-testset and LLM-as-an-evaluator components. Using machine translation as our primary testbed, we demonstrate that self-bias causes each model to rank itself first, overriding the peer-consensus ordering, and validate this on the Chatbot Arena task. We trace the mechanistic cause to limited model proficiency: in low-resource settings, each model's implicit linguistic fingerprint produces homogeneous, model-specific outputs whose degeneration patterns the generating model handles better than its competitors. Based on these findings, we offer actionable guidance: increasing source text diversity using our proposed diversity metric partially mitigates self-bias; generating benchmark data in high-resource languages where models exhibit similarly high diversity minimizes model-specific signatures; and using frontier models to benchmark open-source models remains reliable when the quality gap is large. We hope this work alerts practitioners to self-bias as a validity threat in automated benchmarking and motivates the development of frameworks robust to models' generative idiosyncrasies.

Limitations

Our conclusions are based on three frontier closed-source LLMs and two tasks (machine translation and Chatbot Arena); broader coverage of models, tasks, and languages would further validate generalizability. We use N=200 samples per language direction; larger sample sizes may reveal more nuanced patterns. We used AI assistance (Claude Opus 4.6) to help revise grammar and sentence structures in this paper.

References

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. 2025. Do llm evaluators prefer themselves for a reason? *Preprint*, arXiv:2504.03846.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edition. Lawrence Erlbaum Associates.

Eitan Farchi, Shmulik Froimovich, Rami Katan, and Orna Raz. 2024. Automatic generation of benchmarks and reliable llm judgment for code tasks. *Preprint*, arXiv:2410.21071.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, and 5 others. 2025. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *Preprint*, arXiv:2411.04872.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *Preprint*, arXiv:2503.05061.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi kiu Lo, Vilém Zouhar, Frédéric Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *Preprint*, arXiv:2409.11968.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2025. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Do these llm benchmarks agree? fixing benchmark evaluation with benchbench. *Preprint*, arXiv:2407.13696.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025a. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models. *Preprint*, arXiv:2504.01001.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025b. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *Preprint*, arXiv:2310.16049.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Konstantinos Vergopoulos, Luca Di Petrillo, Giancarlo Pellegrino, Luca Salucci, Sebastian Biedermann, Fabio Grasso, Julian S. S., Evgeny Khramtsov, and Markus Wagner. 2025. Automated benchmark generation for repository-level coding tasks. *Preprint*, arXiv:2503.07701.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-preference bias in llm-as-a-judge. *Preprint*, arXiv:2410.21819.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Preprint*, arXiv:2306.15895.

Peiwen Yuan, Yiwei Li, Shaoxiong Feng, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. Silencer: From discovery to mitigation of self-bias in llm-as-benchmark-generator. *Preprint*, arXiv:2505.20738.

Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2025. Generating difficult-to-translate texts.

A Statistical Significance and Sample Size Justification

To justify our sample size of N=200, we conduct a bootstrap analysis on the Bemba→English translation task under the LLM-as-a-testset and LLM-as-an-evaluator setup. For each of the three models (Gemini 2.5 Pro, GPT-4.1, Claude Opus 4), we construct the 3×3 ranking table and perform bootstrap resampling (B=10,000) to compute 95% confidence intervals.

Table 13 reports the mean rankings with standard deviations, and Table 14 reports the corresponding 95% confidence intervals. The diagonal entries—representing each model's ranking of itself—show no overlap with the off-diagonal entries in the same column, confirming that each model ranks itself significantly higher than the other two. Based on this analysis and budget constraints, we use N=200 for all experiments.

Table 13: Mean rankings (± std) from bootstrap resampling on Bemba→English. Diagonal entries (self-rankings) are consistently lower (better) than off-diagonal entries.

Model	Gemini	GPT-4.1	Claude
Gemini	1.00 ± 0.03	1.52 ± 0.09	1.70 ± 0.10
GPT-4.1	1.21 ± 0.07	1.18 ± 0.05	1.54 ± 0.09
Claude	1.64 ± 0.11	1.93 ± 0.11	1.17 ± 0.06

Table 14: 95% bootstrap confidence intervals for rankings on Bemba→English. Diagonal intervals do not overlap with off-diagonal intervals in each column, confirming statistical significance of self-bias.

Model	Gemini	GPT-4.1	Claude
Gemini	$[0.97, 1.02]$	$[1.43, 1.61]$	$[1.59, 1.80]$
GPT-4.1	$[1.15, 1.28]$	$[1.12, 1.23]$	$[1.45, 1.64]$
Claude	$[1.52, 1.75]$	$[1.82, 2.04]$	$[1.11, 1.23]$

B Cohen's d: Effect Size for Mean Differences

Cohen's d (Cohen, 1988) is a widely used standardized effect size measure to quantify the difference between two means, expressed in standard deviation units. It is particularly useful when comparing the central tendency of two groups (e.g., two sets of scores, two distributions) and interpreting the practical significance of their difference, independent of sample size.

Given two groups, Group 1 and Group 2, with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations s_1 and s_2 respectively, Cohen's d is typically calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

n_1 and n_2 are the sample sizes for Group 1 and Group 2.

Interpretation Guidelines (Cohen's conventions):

- $d = 0.2$: Small effect
- $d = 0.5$: Medium effect
- $d = 0.8$: Large effect

C Prompts used for LLM-as-a-benchmark

All prompts for both LLM-as-a-testset and LLM-as-an-evaluator are included. To ensure fair comparisons, these prompts are adapted from the zero-shot benchmark paper by (Pombal et al., 2025b). The LLM-as-a-testset prompts incorporate a few randomly chosen seed variables—such as length, topic, subtopic, and style—to guarantee diverse generated source texts. A complete list of options for each seed variable is provided below.

Prompt for LLM-as-a-Testset (source + reference generation) for translation.

```
You are a multilingual content creator
and translation expert. Your task is
to generate a comprehensive translation
exercise based on the given attributes.
Follow these instructions carefully:
1. Review the following input variables:
- Source language: SOURCE LANGUAGE
- Target language: TARGET LANGUAGE
- Topic: topic
- Subtopic: subtopic
- Source length: length
- Style: style
2. Generate a source text: Create an
original text in the source language,
adhering to the specified topic, subtopic,
and length. The text should be coherent,
informative, and suitable for translation.
3. Generate a reference translation:
Produce a high-quality, fluent translation
of the source text in the target language.
This translation should serve as a reference
for evaluating other translations. IT IS
CRUCIAL THAT THE REFERENCE TRANSLATION
SOUNDS NATURAL IN THE TARGET LANGUAGE.
Format your output as follows:
<START OF SOURCE>
INSERT THE SOURCE TEXT HERE
<END OF SOURCE>
<START OF REFERENCE TRANSLATION>
INSERT THE REFERENCE TRANSLATION HERE
<END OF REFERENCE TRANSLATION>
Ensure that your response is comprehensive,
coherent, and follows all the instructions
provided above. Abide strictly by the
requested format and generated until the
end of the requested output. Only generate
source and reference translation. Do not
generate any other text such as reasoning or
explanations.
<START OF SOURCE>
```

Prompt for LLM-as-a-Testset (source only) for translation.

```
You are a multilingual content
creator and translation expert. Your task
is to generate a comprehensive translation
exercise based on the given attributes.
Follow these instructions carefully:
1. Review the following input variables:
- Source language: SOURCE LANGUAGE
- Topic: topic
- Subtopic: subtopic
- Source length: length
- Style: style
2. Generate a source text: Create an
original text in the source language,
adhering to the specified topic, subtopic,
and length. The text should be coherent,
informative, and suitable for translation.
Format your output as follows:
<START OF SOURCE>
INSERT THE SOURCE TEXT HERE
<END OF SOURCE>
Ensure that your response is comprehensive,
coherent, and follows all the instructions
provided above.
Abide strictly by the requested format and
generated until the end of the requested
output. Only generate source text. Do not
generate any other text such as reasoning or
```

explanations.
<START OF SOURCE>

Topics. "Tech Innovation", "Global Markets", "Environmental Policy", "Public Health", "Urban Development", "International Relations", "Education Reform", "Cultural Trends", "Scientific Discoveries", "Economic Policy", "Sports Industry", "Media & Entertainment", "Workplace Transformation", "Transportation & Mobility", "Food & Agriculture", "Medical & Healthcare", "Legal & Compliance", "E-commerce & Retail", "Financial Services", "Gaming & Software", "Marketing & Advertising", "Government Documentation", "Academic Research", "Patents & Intellectual Property", "Manufacturing & Safety", "Tourism & Hospitality", "Religious & Cultural Studies", "Insurance & Risk Management", "Consumer Electronics", "Pharmaceutical Industry", "Fashion & Apparel", "Beauty & Cosmetics", "Home & Living", "Automotive Industry", "Social Media", "Dating & Relationships", "Parenting & Family", "Arts & Culture", "Music Industry", "Film & Cinema", "Books & Literature", "Food & Cuisine", "Sports & Recreation", "Fitness & Wellness", "Mental Health", "Architecture & Design", "Real Estate", "Telecommunications", "Renewable Energy", "Space Exploration", "Wildlife & Nature", "Weather & Climate", "History & Heritage", "Politics & Governance", "NGOs & Nonprofits", "New York City", "London", "Tokyo", "Paris", "Berlin", "Singapore", "Dubai", "São Paulo", "Sydney", "Mumbai", "Madrid", "Lisbon", "Stockholm", "Amsterdam", "Seoul", "Japan", "France", "Germany", "Brazil", "India", "Italy", "Spain", "China", "United Kingdom", "Portugal", "Poetry"

Subtopics. "Poetry": ["Modernism", "Contemporary", "Modernism", "Haiku", "European Poetry", "Asian Poetry", "Theme identification",], "Tech Innovation": ["Artificial Intelligence", "Quantum Computing", "Robotics", "5G/6G Networks", "Biotechnology", "Green Tech", "Edge Computing", "Cybersecurity",], "Global Markets": ["Stock Exchanges", "Cryptocurrency", "International Trade", "Foreign Investment", "Commodity Markets", "Emerging Markets", "Foreign Exchange", "Market Regulations",], "Environmental Policy": ["Carbon Trading", "Renewable Energy Initiatives", "Wildlife Protection", "Urban Planning", "Waste Management", "Climate Agreements", "Marine Conservation",], "Public Health": ["Disease Prevention", "Healthcare Systems", "Vaccination Programs", "Mental Health Services", "Maternal Health", "Epidemiology", "Health Technology",], ...

Styles. "creative", "concise", "technical", "formal", "informal", "narrative", "persuasive", "descriptive", "analytical", "humorous", "poetic", "casual", "academic", "journalistic", "neutral", "elaborate", "minimalist", "rushed"

Lengths. "short", "medium"

Prompt for LLM-as-an-Evaluator for translation. You are an expert judge evaluating translation quality. You will be presented with: - An original text - A translation to evaluate

Rate the translation on a scale of 1-6 based on these key criteria: - Accuracy and fidelity to source - Grammar and language correctness - Natural flow and readability - Terminology consistency - Completeness of translation - Technical precision

Scoring Rubric:

6 - Outstanding - Perfect accuracy with source meaning - Flawless grammar and language use - Reads completely naturally in target language - Consistent and precise terminology - Complete translation with no omissions - Excellent technical accuracy

5 - Excellent - Very accurate rendering of source - Strong grammar with minimal issues - Natural-sounding translation - Good terminology consistency - Nearly complete coverage - Strong technical accuracy

4 - Good - Generally accurate translation - Mostly correct grammar - Readable with some awkward passages - Generally consistent terminology - Minor omissions only - Adequate technical accuracy

3 - Fair - Some accuracy issues - Notable grammar problems - Often unnatural phrasing - Inconsistent terminology - Several omissions - Technical inaccuracies present

2 - Poor - Significant accuracy issues - Frequent grammar errors - Unnatural throughout - Poor terminology consistency - Major omissions - Many technical errors

1 - Inadequate - Fails to convey source meaning - Severe grammar issues - Incomprehensible in target language - No terminology consistency - Incomplete translation - Technical meaning lost

Format your output as follows: Put detailed explanation between <START OF FEEDBACK> and </END OF FEEDBACK> Put result between <START OF RESULT> and </END OF RESULT> Don't provide any other text

<START OF FEEDBACK> Put detailed explanation of the score based on the criteria here </END OF FEEDBACK>

<START OF RESULT> Put only a number from 1 to 6 here </END OF RESULT> <START OF SOURCE

TEXT> prompt </END OF SOURCE TEXT>

<START OF TRANSLATION> answer </END OF TRANSLATION>

Prompt for Translation. You are a professional translator. You are given a source text in SOURCE LANGUAGE. You need to translate the source text to TARGET LANGUAGE. Don't include any other text except the translation. Please output the translation between <START OF TRANSLATION> and <END OF TRANSLATION>. Source text: SOURCE TEXT

Prompt for LLM-as-a-Testset for Chatbot Arena. You are tasked with creating a diverse and engaging prompt for a chatbot arena. This prompt will be used to test and compare the capabilities of different language models. Your goal is to generate a question or prompt that will challenge these models and showcase their strengths or weaknesses.

Here are the input variables you will use to craft your prompt:
- Language: language
- Topic: topic
- Subtopic: subtopic
- Difficulty: difficulty
- Style: style
- Writer: writer
- Writing proficiency: writing_proficiency
- Prompt length: length

Guidelines for creating the prompt:
1. Abide strictly by the input variables provided.
2. Ensure that your prompt is open-ended enough to allow for varied and interesting responses from different language models.
3. Avoid prompts that are overly specific to a particular AI model's capabilities or training data.
4. Create a prompt that is engaging and thought-provoking, encouraging creative or analytical thinking.

Generate output in the following format:
<START OF PROMPT> [Your generated prompt here. IMPORTANT: include only the prompt.] <END OF PROMPT>

Ensure the generated prompt is in the requested language. Remember to abide strictly by the provided input variables and the requested format.

An example question for Chatbot Arena (English). In Albion's ledger, where fortunes entwine, what unseen current shapes the City's final line?

An example question for Chatbot Arena (Bemba). Finshi fingacitika mu fashion design mu Zambia mu myaka ikumi iileisa? Ndiye, muli busy sana, lelo ndefwaya ukwishiba ifyo mwingalanda pali ifi.

Prompt for LLM-as-an-Evaluator for Chatbot Arena. You are an expert judge evaluating response quality for prompts on a variety of topics. You will be presented with:

- An original prompt
- A response to evaluate

Rate the response on a scale of 1-6 based on these key criteria:
- Accuracy and factual correctness
- Comprehensiveness of the answer
- Clarity and organization
- Relevance to the prompt
- Supporting evidence/examples (if applicable)
- Objectivity and balance (if applicable)

Scoring Rubric:
6 - Outstanding

- Exceptionally accurate and well-researched information
- Comprehensive coverage with valuable additional context
- Perfectly structured and crystal clear explanation
- Directly addresses all aspects of the prompt
- Strong supporting evidence and examples
- Maintains perfect objectivity where needed

5 - Excellent

- Highly accurate information
- Thorough coverage of the topic
- Very clear and well-organized
- Closely aligned with the prompt
- Good supporting evidence
- Strong objectivity

4 - Good

- Generally accurate information
- Adequate coverage of main points
- Clear organization with minor issues
- Mostly relevant to the prompt
- Basic supporting evidence
- Reasonable objectivity

3 - Fair

- Some inaccuracies present
- Incomplete coverage
- Organization needs improvement
- Partially misses prompt intent
- Limited supporting evidence
- Some bias present

2 - Poor

- Multiple inaccuracies
- Superficial coverage
- Unclear organization
- Largely misses prompt intent
- Minimal supporting evidence
- Clear bias present

1 - Inadequate

- Fundamentally incorrect information
- Severely incomplete
- Disorganized or incomprehensible
- Fails to address prompt
- No supporting evidence
- Extremely biased or misleading

Format your output as follows:
Put detailed explanation between <START OF FEEDBACK> and </END OF FEEDBACK>
Put result between <START OF RESULT> and </END OF RESULT>
Don't provide any other text

<START OF FEEDBACK>
Put detailed explanation of the score based on the criteria here
</END OF FEEDBACK>
<START OF RESULT>
Put only a number from 1 to 6 here
</END OF RESULT>

```
   <START OF PROMPT>
prompt
<END OF PROMPT>

   <START OF ANSWER>
answer
<END OF ANSWER>
```

D Self-repair Degeneration in Translation

In Table 16, we examine Gemini-2.5-pro's self-repair ability when translating its own degenerated source texts. Gemini-2.5-pro more effectively corrects degenerated content it produces during the translation process compared to other models, across both language directions. The observed non-deterministic behavior of GPT-4.1 is likely due to the significantly smaller sample of degenerated source sentences it processed (n=23/400) compared to Gemini-2.5-pro (n=89/400). An example of this self-repair during translation is provided below.

Example of Gemini's self repair during translation.

```
Gemini's degenerate source text:
Jichhürunakanxa, celularanakax wali
wakiskiripuniwa.  Aka tecnología ukax janiw
mayni jaqimpi aruskipt'añatakikiti, jan
ukasti yatiñanak jikxatañataki, anatañataki,
ukat yaqha lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak [repeated
50 times] lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak lurawinak
lurawinak lurawinak lurawinak luraw

   Gemini's translation:  Nowadays, cell
phones are truly necessary.  This technology
is not only for communicating with other
people, but also for finding information,
for entertainment, and for carrying out
various other activities.
```

E High-Resource Language Analysis

We considered high-resource languages for the analysis as well. However, the focus of our study is on medium- and low-resource machine translation, primarily due to the near-perfect performance of frontier models in high-resource settings. As shown in Table 17 for Chinese-to-English translation, the scores from Gemini, Claude, and GPT-4.1 for all translations are nearly perfect (on a 1–6 scale). While models still display self-bias in their own generation—each model assigns itself the highest score in its column—the magnitude is quantitatively much smaller than in low-resource settings.

We argue that the near-perfect translation quality of all models makes studying self-bias on high-resource languages less informative: there is little room for diverse error patterns, and the differences between self-scores and peer-scores are marginal. Consequently, we prioritize medium- and low-resource language directions, where models exhibit distinct error patterns and evaluate them differently due to self-bias. More importantly, the LLM-as-a-benchmark setup is most practically relevant for low-resource translation, as these tasks inherently lack the data to construct high-quality benchmarks through human curation.

En→Bemba	chrF across prompts + models		
Data from	**Gemini**	**GPT**	**Claude**
Gemini	36.07	35.77	34.13
GPT	32.98	34.86	31.50
Claude	36.75	36.47	36.85

En→Aymara	chrF across prompts + models		
Data from	**Gemini**	**GPT**	**Claude**
Gemini	36.15	35.89	34.07
GPT	32.83	35.28	31.27
Claude	36.83	36.62	37.14

Table 15: Average chrF@K similarity scores, differentiating within-model (diagonal) from cross-model (off-diagonal) comparisons. Unlike XX→En translation, a clear diagonal trend is not observed. For instance, Claude-generated outputs exhibit cross-model chrF similarities with GPT4.1 and Gemini-2.5-pro that are comparable to its within model similarity.

	Degeneration ratio after translating on degenerated source texts			
	Aymara→English		**Bemba→English**	
Translator	Gemini-2.5-Pro	GPT4.1	Gemini-2.5-Pro	GPT4.1
Gemini-2.5-Pro	**90.4**	86.7	**81.1**	75.0
GPT4.1	94.2	93.3	86.5	**62.5**
Claude-Opus-4	94.2	**73.3**	94.6	87.5

Table 16: Model Self-Repair Ability during Translation of Self-Generated Degenerated Source Texts. This table presents the percentage of translations that retain degenerated content. Gemini-2.5-pro demonstrates a consistently superior ability to correct degenerated content in its self-generated source texts during the translation process, outperforming other models across both language directions. GPT-4.1's observed non-deterministic behavior is potentially attributable to the significantly smaller sample size of degenerated source sentences it processed (n=23/400), compared to Gemini-2.5-pro (n=89/400).

Translator \ Benchmark	**Gemini**	**Claude**	**GPT-4.1**
Gemini	**6.000**	5.584	5.980
Claude	5.998	**5.633**	5.986
GPT-4.1	5.998	5.542	**5.987**

Table 17: LLM-as-a-benchmark scores for Chinese→English translation (1–6 scale; higher is better). All scores are near-perfect, leaving minimal room for self-bias to manifest. The diagonal entries (bolded) are highest in each column, indicating residual self-preference, but the absolute differences are negligible compared to low-resource settings.