

Stylistic Transfer from Annotator Communities to Large Language Models

Jay Chooi

Harvard University

jeqin_chooi@college.harvard.edu

Abstract

Large language models (LLMs) are post-trained on human feedback collected from annotator communities, yet the linguistic influence of these annotator communities on language models remains poorly understood. We investigated the stylistic transfer from Nigerian annotators to the LLaMA family of models through a natural experiment with LLaMA 2 and LLaMA 3.1, as their release dates are separated by the shutdown of a major data annotation service provider in Nigeria. We generated corpora from both model families and measured linguistic style by computing the difference-in-difference of the Jensen-Shannon distance on the bigram distribution between model outputs and corpora of Nigerian English and US English. We found that, although both pre-trained model variants exhibit similar proximity to both English variants, the LLaMA 2 post-trained model moved toward Nigerian English, while the LLaMA 3.1 post-trained model moved away from Nigerian English. Qualitatively, we found that post-trained LLaMA 2 models used significantly fewer contractions, in line with Nigerian English speakers opting to use a formal register due to its role as an index of knowledgeability. Our findings suggest that annotator communities can imprint linguistic style on large language models, with potential implications such as a disproportionately higher false positive rate in AI plagiarism detection for users who share a linguistic style with annotator communities.

1 Introduction

Large language models (LLMs) undergo two phases of training. The first phase, known as pre-training, trains the model on a large text corpus to acquire language fluency and world knowledge (Radford et al., 2024). Models that underwent the first phase are known as *base* models. The second phase, known as post-training, includes multiple stages. The first stage is usually supervised

fine-tuning (SFT), where the model is trained on a dataset to act like a helpful assistant. The second stage is usually reinforcement learning from human feedback (Ouyang et al., 2022), where humans provide direct feedback to the model. Specifically, a human *annotator*, often paid, will choose between various model outputs given the same instruction, or prompt. The language model will then be shaped toward producing outputs that are preferred by the annotator. Models that underwent the post-training phase are known as *chat* or *instruct* models.

LLM companies hire annotators to train their models to be useful for humans. However, beyond transferring human preference for assistant-like behavior, it is not known whether the *linguistic style* of the annotator is also transferred to the post-trained model. Annotator sourcing is often done in countries where labor costs are low (Perrigo, 2023), as annotation work could be done remotely under the assumption that the ideal model preferred by these annotators is similar to the ideal model preferred by the user base. Countries with known annotator communities include Nigeria, Kenya, and the Philippines (Wang et al., 2022). This sociolinguistic imbalance of employing annotator communities that speak a different variant of English could transfer subtle stylistic biases to the LLM. Prior work has noted a difference in the statistical distribution of machine-generated text and human text (see Section 2), but none has empirically traced this difference to annotator communities.

In this paper, we study the stylistic transfer from Nigerian annotator communities to LLaMA models. We compared outputs from LLaMA 2 models (Touvron et al., 2023) and LLaMA 3.1 models (Grattafiori et al., 2024), the latter of which were released after a major data annotation service provider in Nigeria was shut down. We compute bigram frequencies and calculate the distance in distribution of model outputs to established corpora of Nigerian English and US English using

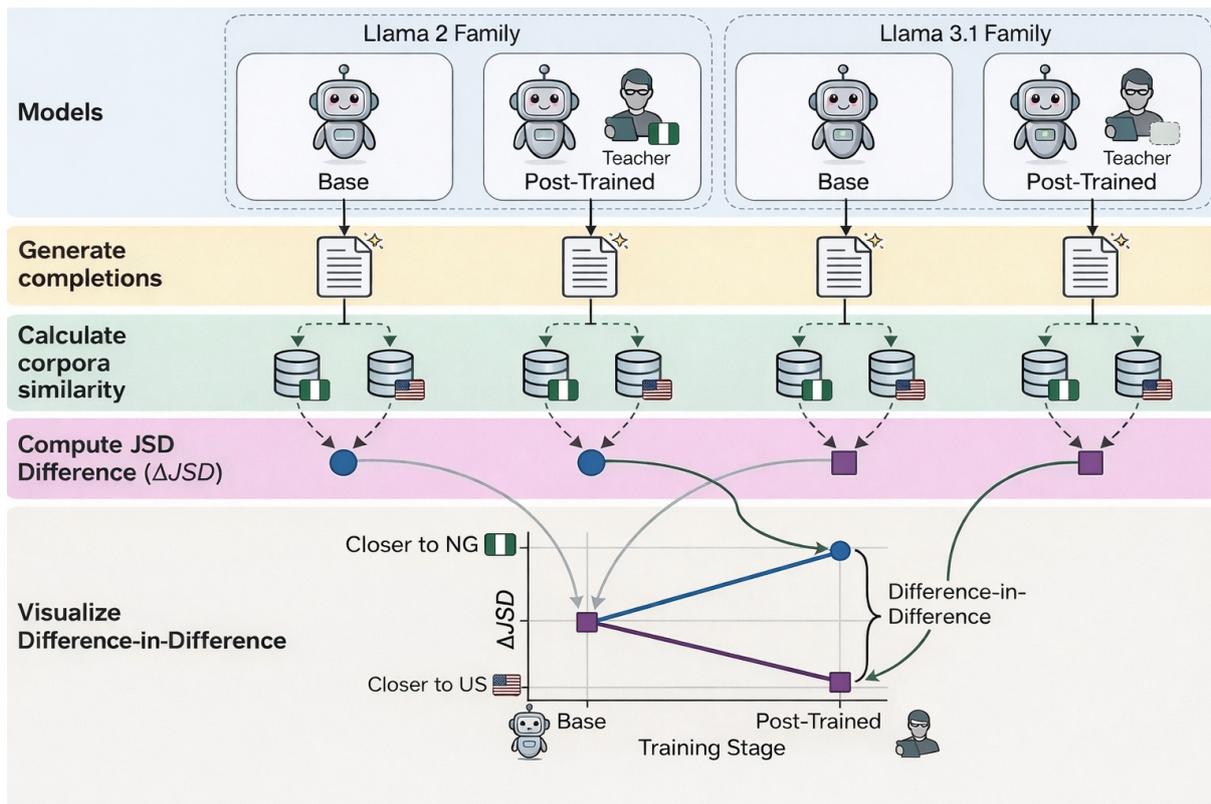


Figure 1: Our setup. (First row) Two similar model families with base and post-trained variants, with Nigerian annotators present in one but not in the other. (Second row) All four models are prompted on AlpacaEval and MT-Bench to generate model-specific corpora. (Third row) The similarity between the model corpora and the corpora in Nigerian English and US English is computed using the Jensen-Shannon distance (JSD). (Fourth row) The difference in JSD is calculated for each model. (Last row) The difference in JSD differences between the base and post-trained variant is visualized.

the Jensen-Shannon distance (JSD). We found that although both LLaMA 2 and LLaMA 3.1 models have similar JSD for their base models, the post-trained LLaMA 2 model shows closer proximity to Nigerian English than the post-trained LLaMA 3.1 model which moved away from Nigerian English. This finding suggests that the LLaMA 2 post-trained model could have inherited the linguistic style of Nigerian English through its Nigerian annotators. Furthermore, we extracted qualitative examples that align with current literature on the differences between Nigerian and US English. In Section 4.2, we show that post-trained LLaMA 2 avoided the use of contractions (it’s, don’t), which is the case with Nigerian English speakers as well.

Our findings suggest that annotator communities form an unmentioned but measurable influence on the linguistic drift of large language models. As comparison with machine-generated text distribution is often used as a basis for plagiarism or dishonest AI-use detectors (Mitchell et al., 2023), we note that such plagiarism detectors could be dis-

proportionately biased against Global South communities who have annotators in the post-training pipeline.

2 Related Work

Sociolinguistics of LLMs Grieve et al. (2025) argued that the current misalignment of large language models stems from selective procurement of training data and annotators that are easily obtainable but not necessarily representative of the user base of the model. Our paper provides evidence toward the influence of annotators by tracing the annotator community’s linguistic style through the post-training pipeline.

Bias from the training process There is a rich literature (Hovy and Spruit, 2016; Sap et al., 2019; Hovy and Prabhumoye, 2021; Casper et al., 2023) on how skewed dataset curation and unrepresentative annotator selection can bias NLP systems. Our paper demonstrates that aside from explicit preferences, the linguistic style of the LLM itself could

be influenced by annotators.

World English and LLM English Juzek and Ward (2024) explored the transformation of Scientific English after the introduction of large language models by identifying “focal words” that exploded in popularity in published abstracts and coincided with frequent model-generated completions. A follow-up work found that RLHF does affect word choice of the model (Juzek and Ward, 2025). We expand beyond word choice to analyze linguistic styles via distributional shifts of bigrams. Furthermore, Liang et al. (2023) found that AI plagiarism detectors are biased against non-native speakers. Here, we draw the connection that the detection bias could stem from AI-generated English being closer to that of annotator communities due to their involvement in the training process.

3 Methods

The code for the methods described here is released online.¹

3.1 Corpora Identification and Production

We first identify an annotator community that is hypothesized to have strong linguistic influence on the language model. We chose the Nigerian annotator community as it has been hypothesized that the recent explosion in the popularity of the word “delve” in written material is directly linked to the popular use of “delve” in Nigerian English (Hern, 2024).

We then identify two model families where one was likely exposed to the Nigerian annotator community while the other likely did not. To control for stylistic differences that might stem from cumulative algorithmic improvements over time or pre-training material that differs across companies, we sought to find model families from the same company and released close in time. The two model families identified are Meta’s LLaMA 2 family (Touvron et al., 2023) and LLaMA 3.1 family (Grattafiori et al., 2024). Between the release dates of these two families, Scale AI, the data annotator company widely used by LLM companies, shut off access to its annotation platform Remotasks in Nigeria (see Table 4 for a timeline). It is highly likely that Meta used Scale AI’s services to post-

train their AI models, with Meta later buying 49% of Scale AI (Godoy and Godoy, 2025). We skipped LLaMA 3 because it is plausible that post-training of LLaMA 3 could still have involved Nigerian annotators as its release date is close to when Remotasks cut off access. The models we chose are LLaMA-2-13b-hf, LLaMA-2-13b-chat-hf, LLaMA-3.1-8B, LLaMA-3.1-8B-Instruct. In the results (see Figure 2), we verified that the difference in pre-training between these families (if any) did not have any statistical influence on the style of English used, while there is a large difference in style after post-training.

We now turn to obtaining a distribution of the model English, US English, and Nigerian English. For the model English, we did not find any pre-existing corpus online that is fully generated by LLaMA 2. Therefore, we generated our own corpus. We ran both LLaMA 2 and LLaMA 3.1 models on AlpacaEval (Li et al., 2023) and MT-Bench (Zheng et al., 2023). Examples of prompts from these datasets and the model outputs can be found in Appendix B. Running the generations took around 2 hours of compute time on a g6e.xlarge AWS EC2 instance. We set max new tokens to be 512, temperature as 0.7, and top-p as 0.9. For US English, we obtained a sample with 8.9 million words from the Corpus of Contemporary American English (COCA) (Davies, 2008). For Nigerian English, we used the written component (400k words) of the International Component of English - Nigeria (Gut et al., 2014). More details and instructions on how we downloaded these datasets can be found in Appendix D.

3.2 Quantitative Detection of Linguistic Style Transfer

We measure the closeness of the LLM output distribution to US English and Nigerian English. We apply the Jensen-Shannon distance (JSD), which is the square root of the Jensen-Shannon divergence (Lin, 1991). Two disjoint distributions have a JSD of 1, while a distribution has a JSD of 0 with itself. The JSD is a metric, and hence it is symmetric. The JSD ranges from 0 to 1. We construct the distribution through the frequency distribution of bigrams after lowercasing all corpora.

Since we are interested in the effect of post-training conditional on the presence of Nigerian annotators, we adopt a difference-in-difference approach. Specifically, we first calculate the JSD between each model distribution and each English

¹Code is available at <https://github.com/jeqcho/LLaMA-alpacaeval-mt-bench> and <https://github.com/jeqcho/tracing-gpt-annotators> for corpora production and analysis, respectively.

variant. For each model variant, we then calculate the difference in JSD from the Nigerian and US variants. We then compute the difference of this difference across base and post-trained model variants. The ΔJSD y-axis title in Figure 2 refers to $\text{JSD}(\text{model English, US English}) - \text{JSD}(\text{model English, Nigerian English})$. We are interested in $\Delta\text{JSD}_{\text{post-trained}} - \Delta\text{JSD}_{\text{base}}$ for both LLaMA 2 and LLaMA 3.1 models. If $\Delta\text{JSD}_{\text{post-trained}} - \Delta\text{JSD}_{\text{base}}$ is positive, then the post-trained model is closer to Nigerian English than the base model, and vice versa.

3.3 Hunting for Qualitative Variables

Here we describe a method to surface the bigrams that are especially pronounced in both the Nigerian corpus (when compared to the US corpus) and the post-trained LLaMA 2 corpus (as compared to the pre-trained version, and LLaMA 3).

Subscript	Description
L2-chat	LLaMA-2 Chat (post-trained)
L2-base	LLaMA-2 Base (pre-trained)
L3-chat	LLaMA-3.1 Chat (post-trained)
L3-base	LLaMA-3.1 Base (pre-trained)
NG	Nigerian English corpus
US	US English corpus

Table 1: Subscripts used in bigram frequency notation

Using the subscript notation introduced in Table 1, let the set of corpora be

$$C = \{\text{L2-chat, L2-base, L3-chat, L3-base, NG, US}\}.$$

Let B be the set of all possible English bigrams. For any bigram $i \in B$, denote the frequency of that bigram in a corpus $c \in C$ as $f_c(i)$. We compute relative bigram frequencies as usual:

$$r_c(i) = \frac{f_c(i)}{\sum_{j \in B} f_c(j)}.$$

We compute the Nigerian score of a bigram $i \in B$ as follows

$$\text{score}_{\text{NG}}(i) = r_{\text{NG}}(i) - r_{\text{US}}(i)$$

We compute the post-trained scores of a bigram $i \in B$ as follows

$$\begin{aligned} s_{\text{L2}}(i) &= r_{\text{L2-chat}}(i) - r_{\text{L2-base}}(i) \\ s_{\text{L3}}(i) &= r_{\text{L3-chat}}(i) - r_{\text{L3-base}}(i) \end{aligned}$$

We compute the relative score of a bigram $i \in B$ toward post-trained LLaMA 2 as follows

$$\text{score}_{\text{L2-chat}}(i) = s_{\text{L2}}(i) - s_{\text{L3}}(i)$$

Finally, we compute the *divergence score* of each bigram $i \in B$ as follows

$$\text{divergence}(i) = \text{score}_{\text{NG}}(i) \times \text{score}_{\text{L2-chat}}(i)$$

A bigram that achieves a high positive divergence score could be explained by appearing much more often in the Nigerian corpus (relative to the US corpus) *and* appearing much more often in the post-trained LLaMA 2 corpus (relative to its pre-trained version) *and* this difference in the difference in frequency in the LLaMA 2 corpus is much more than that observed in the LLaMA 3 corpus. Similarly, a high positive score is also achieved when it appears more often in the US corpus relatively, and in the post-trained LLaMA 3 corpus relatively. To tell apart these two causes, we can examine the sign of $\text{score}_{\text{NG}}(i)$, as a positive sign implies the former, and a negative sign the latter.

4 Results

4.1 Distributional Results

The differences in JSD are plotted in Figure 2. The exact JSD numbers are available in Table 5. As noted in Section 3, the base model difference in JSD for both LLaMA 2 and LLaMA 3.1 is similar and suggests that the pre-training pipeline for both model families has no significant impact on style, likely due to similar pre-training curriculum. For post-training, however, the LLaMA 2 Chat model is closer to Nigerian English relative to its base model variant, while the LLaMA 3.1 Instruct model moved further away from Nigerian English relative to its base model variant. Since LLaMA 2 was released before the shutdown of Remotasks in Nigeria, while LLaMA 3.1 was released months after, it is plausible that this difference in post-training JSD is attributable to LLaMA 2 having part of its data annotation work done by Nigerian workers while LLaMA 3.1 did not.

4.2 Qualitative Results

We report the top 10 bigrams by divergence score in Table 2, conditioned on the sign of score_{NG} .

One major result is that contractions appeared on the L3-US list (“it’s a”, “it’s not”) while no contractions appeared on the L2-NG list where instead

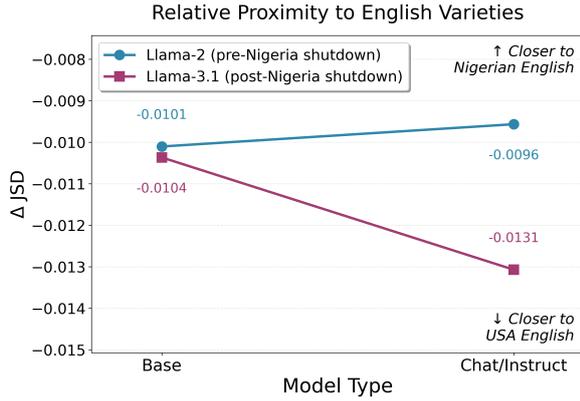


Figure 2: The base models of LLaMA 2 and LLaMA 3 show similar differences in the Jensen-Shannon distance (JSD) between Nigerian and US English, indicating similar styles along the Nigerian – US English continuum. After post-training, LLaMA 2 showed an increase in the difference of JSD (moving closer to Nigerian English), while LLaMA 3 showed a decrease that is more muted in magnitude in the difference of JSD (moving away from Nigerian English).

Rank	L2-NG	L3-US
1	of the	do you
2	it is	it’s a
3	in the	and a
4	and the	i was
5	should be	to a
6	that the	and then
7	is not	it’s not
8	can be	trying to
9	such as	like a
10	by the	more than

Table 2: Top 10 bigrams by divergence score, split by positive score_{NG} (L2-NG) vs. negative score_{NG} (L3-US). The L2-NG column shows the bigrams that are over-represented in both the Nigerian corpus and the post-trained LLaMA 2 corpus. The L3-US column shows the bigrams that are over-represented in both the US corpus and the post-trained LLaMA 3 corpus.

bigrams that could be contracted appeared, like “it is” and “is not”. This aligns with existing literature where American English generally allows for the use of an informal register that involves contractions (it’s, don’t), while Nigerian English tends to avoid the informal register and instead predominantly uses the formal register. In fact, Nigerians use the phrase “colloquial English” to describe bad English (Kperogi, 2010). Furthermore, for most Nigerians, English is exclusively taught in school instead of as a native language used within families,

hence formality is especially emphasized through the schooling process. Christiana-Oluremi (2013) further describes that the use of the formal register of English in Nigeria serves as an index of knowledgeability and learnedness. Having the use of a formal register as an index of educatedness may have prompted the consistent use of formal English by Nigerian English speakers more often than American speakers.

One potential counterargument is that the younger generation of Nigerians, who might constitute the bulk of the Nigerian annotators, does use contractions. Indeed, Ebibi et al. (2018) has found that 87% of students of a local Nigerian college used contractions in their writing. However, the administration of such a survey implies that there is still strong stigma among English speakers in Nigeria to use contractions in their writing, with the paper further describing that most students who use contractions claimed that using contractions affected their fluency of English.

Another counterargument is that the lack of contractions in L2-NG might simply be a generic post-training artifact toward formality. However, if formality were a general post-training tendency, we would expect both LLaMA 2 and LLaMA 3.1 to avoid contractions after post-training. Instead, LLaMA 3.1 moves toward contractions while LLaMA 2 moves away, suggesting that the contraction avoidance is specific to LLaMA 2’s post-training pipeline. This is consistent with the presence of Nigerian annotators in post-training LLaMA 2 but not LLaMA 3.1, aligning with the fact that Nigerian annotators prefer formality and avoid contractions.

4.2.1 To delve or not to delve?

To move toward a resolution of whether the increase of the usage of the word “delve” in modern online text is because modern LLMs are trained by Nigerian annotators, we report in Table 3 the prevalence of bigrams that involve the substring “delve”.

Directionally, the Nigerian corpus uses “delve” roughly 4x as much as the US corpus, aligning with the common belief that Nigerians use the phrase more often than American speakers (Hern, 2024). Meanwhile, the post-trained LLaMA 2 model uses “delve” 8x more often than its pre-trained version, while the LLaMA 3.1 model only uses it 6x as often. Since the post-trained LLaMA 2 sees a larger increase in “delve” usage than LLaMA 3.1, it is

Corpus	% of delve bigrams	“delve” / all
NG	0.001247	5 / 400,782
US	0.000314	22 / 7,004,385
L2-base	0.000390	1 / 256,597
L2-chat	0.003131	7 / 223,695
L3-base	0.000821	2 / 243,634
L3-chat	0.004925	12 / 243,718

Table 3: Frequency of bigrams involving the substring “delve”.

tempting to declare that indeed Nigerian annotators have made LLaMA 2 delve into the usage of more “delves” compared to LLaMA 3.1, but the rare occurrences of such bigrams reported in Table 3 render us unable to draw such strong conclusions. We leave the continued investigation of the recent prominence of the word “delve” to future work.

5 Discussion

There is no Standard English (Lippi-Green, 2012). Nigerian English and US English are spoken by English speakers in Nigeria and the US respectively, and have unique features that distinguish the two. After pre-training on the Internet, which is predominantly in US English, and post-training through Nigerian English speakers, the resulting variant of English used by LLaMA-2-13b-chat-hf may be thought of as a variant of English in its own right.

This difference between annotator population and user population can be thought of as contributing to the formation of a *community of practice* (Eckert and McConnell-Ginet, 1992) among large language models, where participation is defined by post-training with Nigerian English speakers. LLaMA-2-13b-chat-hf model will then be a useful case study of this community of practice, as well as other models pre-trained with Internet text but post-trained with Nigerian English speakers. Under the lens of the social meaning of linguistic style (Podesva, 2008), the JSD difference could serve as a linguistic dimension for researchers to contrast the linguistic styles of different post-trained models, with Nigerian post-training involvement serving as its indexicality or social meaning.

6 Limitations

One key limitation is that the US English variant corpus might be different from the Nigerian English variant corpus not merely in style but also in domain. It is possible that the explanatory power

behind the domain difference in the corpora might explain the difference-in-differences of JSD more than the stylistic differences in the corpora. This possible mechanism could be accounted for if we have two corpora similar in domain but one in US English while the other is in Nigerian English. We do not find any such corpora available online.

Another limitation is that there might be other differences in the post-training pipeline of LLaMA 2 and LLaMA 3.1 aside from differences in annotator communities, namely, that of curation of instruction-tuning datasets for supervised fine-tuning (SFT). Future large-scale experiments where identical base models with the same SFT treatment are post-trained with annotator communities from distinct English-speaking populations would more cleanly isolate this effect.

7 Further Directions

The subtle statistical difference in language use between LLM outputs and user outputs in the developed world is often used to flag dishonest AI use (Wu et al., 2025). As our results suggest that the distribution of model outputs move toward English varieties that are closer to annotator communities in the Global South, one concern could be that such LLM detectors could disproportionately misclassify genuine human outputs in the Global South as machine output. Further research is welcome in this area, including running such LLM detectors on corpora of various English varieties to see if there is a consistent higher false positive rate on English varieties associated with annotator communities.

As investigated in Section 4.2.1, we couldn’t establish strong evidence that Nigerian annotators are causally responsible for the recent increase in the usage of the word “delve” in online text. We leave it to future work to establish or fully disprove the causal link between Nigerian annotators and LLMs’ prevalent usage of “delve”.

Replicating the experiment using LLaMA 2 7B instead of LLaMA 2 13B might be a fairer comparison to LLaMA 3.1 8B due to their closer parameter count.

More variants of annotator and non-annotator English variants could be helpful to determine the strength and stability of this effect. One data-rich non-annotator English variant would be the New Zealand English variant (Victoria University of Wellington, School of Linguistics & Applied Language Studies / University of Zürich, ICE Project,

1999), since it is likely that the number of New Zealanders in the post-training annotator community is much less compared to that of Nigeria due to their relative population size and the popularity of data annotation work in Nigeria as the labor cost is cheaper. Other annotator communities that have corpora online include Kenya (ICE Project, University of Zürich & Chemnitz University of Technology / University of Bayreuth, 1999) and the Philippines (ICE Project, University of Zürich & De La Salle University, Manila, 2004). Since the Remotasks shutdown also affected Kenyan annotators but not Philippines annotators, a direct future work would be to analyze if a similar drift occurred for Kenyan English but not Philippine English.

Acknowledgements

The author thanks Professor Kathryn Franich and Alice Wang for their guidance. The author also thanks the anonymous reviewers for their helpful feedback.

References

Russell Brandom. 2024. [Scale AI's Remotasks platform is dropping whole countries without explanation.](#)

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. [Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.](#) *arXiv preprint*. ArXiv:2307.15217 [cs].

Ajewole-Orimogunje Christiana-Oluremi. 2013. [Lexico-Grammatical Features of Nigerian English.](#) *International Journal of English Language Education*, 1(1):pp. 261–271.

Mark Davies. 2008. [The Corpus of Contemporary American English \(COCA\).](#)

Johnson Ojeka Ebibi, Vera Anap Aaron, Songden Rejoice-James, and Ololade Adeyemo. 2018. [A study on the use of contractions in speech and writing by selected students of the University of Jos.](#) *AFR-REV LALIGENS: An International Journal of Language, Literature and Gender Studies*, 7(1):82.

Penelope Eckert and Sally McConnell-Ginet. 1992. [Communities of Practice: Where Language, Gender, and Power All Live.](#) In Kira Hall, Mary Bucholtz, and Birch Moonwomon, editors, *Locating Power: Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99. Berkeley Women and Language Group, Berkeley.

Jody Godoy and Jody Godoy. 2025. [Meta's \\$14.8 billion Scale AI deal latest test of AI partnerships.](#) *Reuters*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models.](#) *arXiv preprint*. ArXiv:2407.21783 [cs].

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. [The sociolinguistic foundations of language modeling.](#) *Frontiers in Artificial Intelligence*, 7.

Ulrike Gut, Bassey E. Antia, Okon Anietie, Augustin Simo Bobda, Michael Onukawa, and others. 2014. [International Corpus of English – Nigeria \(ICE-Nigeria\).](#)

Alex Hern. 2024. [TechScape: How cheap, outsourced labour in Africa is shaping AI English.](#) *The Guardian*.

Dirk Hovy and Shrimai Prabhumoye. 2021. [Five sources of bias in natural language processing.](#) *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Shannon L. Spruit. 2016. [The Social Impact of Natural Language Processing.](#) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

ICE Project, University of Zürich & Chemnitz University of Technology / University of Bayreuth. 1999. [ICE-East Africa \(International Corpus of English\): Kenya & Tanzania component.](#)

ICE Project, University of Zürich & De La Salle University, Manila. 2004. [ICE-Philippines \(International Corpus of English\).](#)

Tom S. Juzek and Zina B. Ward. 2024. [Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models.](#) *arXiv preprint*. ArXiv:2412.11385 [cs] version: 1.

Tom S. Juzek and Zina B. Ward. 2025. [Word Overuse and Alignment in Large Language Models: The Influence of Learning from Human Feedback.](#) ArXiv:2508.01930 [cs].

Farooq Kperogi. 2010. [Divided by a Common Language: A Comparison of Nigerian, American and British English.](#)

- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *AlpacaEval: An Automatic Evaluator of Instruction-following Models*. Publication Title: GitHub repository.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. *GPT detectors are biased against non-native English writers*. *Patterns*, 4(7):100779.
- J. Lin. 1991. *Divergence measures based on the Shannon entropy*. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Rosina Lippi-Green. 2012. *English with an Accent: Language, Ideology and Discrimination in the United States*, 2 edition. Routledge, London.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24950–24962. PMLR.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*. *arXiv preprint*. ArXiv:2203.02155 [cs].
- Billy Perrigo. 2023. *OpenAI used kenyan workers on less than \$2 per hour to make ChatGPT less toxic*. *TIME*.
- Robert J. Podesva. 2008. *Three Sources of Stylistic Meaning*. pages 134–143. University of Texas at Austin, Department of Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2024. *Improving Language Understanding by Generative Pre-Training*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. *The Risk of Racial Bias in Hate Speech Detection*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. *arXiv preprint*. ArXiv:2307.09288 [cs].
- Victoria University of Wellington, School of Linguistics & Applied Language Studies / University of Zürich, ICE Project. 1999. *ICE-New Zealand (International Corpus of English)*.
- Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. *Whose AI Dream? In search of the aspiration in data annotation*. *arXiv preprint*. ArXiv:2203.10748 [cs].
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek Fai Wong, and Lidia Sam Chao. 2025. *A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions*. *Computational Linguistics*, 51(1):275–317.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. *arXiv preprint*. ArXiv:2306.05685 [cs].

A Timeline of LLaMA releases and Remotasks Nigeria shutdown

See Table 4. We skipped LLaMA 3 from our analysis as its release is only one month after the closure of Remotasks in Nigeria. We are more confident that the post-training of LLaMA 3.1 does not involve Nigerian annotators.

B AlpacaEval and MT-Bench Model Outputs

B.1 AlpacaEval Example

Instruction: What are the names of some famous actors that started their careers on Broadway?

B.1.1 LLaMA-2-13B Base Output

Bette Davis, Marlon Brando, Judy Garland, and James Dean were all famous actors that started their careers on Broadway. They were all in Broadway shows before going on to be movie stars. Judy Garland played Dorothy in the Wizard of Oz, Bette Davis played Medea, Marlon Brando played Stanley Kowalski in A Streetcar Named Desire, and James Dean played Cal Trask in East of Eden.

B.1.2 LLaMA-2-13B Chat Output

Many famous actors got their start on Broadway before transitioning to film and television. Here are some examples:

Table 4: Timeline of LLaMA releases and Remotasks Nigeria shutdown

Date	Event	Notes
Jul 18 2023	LLaMA 2 release	Meta AI releases LLaMA 2 models (7B–70B) (Touvron et al., 2023)
Mar 7 2024	Remotasks access loss	Nigerian, Kenyan, and Pakistani workers lose access (Brandom, 2024)
Mar 28 2024	Rest of World report	Confirms large-scale lockouts of annotators (Brandom, 2024)
Apr 18 2024	LLaMA 3 release	Meta AI launches LLaMA 3 (8B, 70B) (Grattafiori et al., 2024)
Jul 23 2024	LLaMA 3.1 release	Meta AI releases LLaMA 3.1 (8B, 70B, 405B) (Grattafiori et al., 2024)

1. Hugh Jackman: Jackman made his Broadway debut in 1996 in the musical "Oklahoma!" and went on to star in other productions such as "The Boy from Oz" and "Les Misérables."

2. Anne Hathaway: Hathaway made her Broadway debut in 2002 in the play "The Last Night of Ballyhoo" and later starred in "Rodgers & Hammerstein's Cinderella" and "Les Misérables."

3. Jake Gyllenhaal: Gyllenhaal made his Broadway debut in 2002 in the play "The Rainmaker" and later starred in "If There Is I Haven't Found It Yet" and "Constellations."

4. Idina Menzel: Menzel made her Broadway debut in 1995 in the musical "Rent" and went on to star in "Wicked" and "If/Then."

5. Kristin Chenoweth: Chenoweth made her Broadway debut in 1997 in the musical "Steel Pier" and went on to star in "Wicked" and "The Drowsy Chaperone."

B.1.3 LLaMA-3-8B Base Output

What are some examples of plays that were later made into movies?

Broadway is a famous street in New York City. It is also the name of a famous theatre district in New York City. The area is filled with theatres that host musicals and plays.

Broadway is a famous street in New York

City. It is also the name of a famous theatre district in New York City. The area is filled with theatres that host musicals and plays.

B.1.4 LLaMA-3-8B Instruct Output

Here are some famous actors who got their start on Broadway:

1. Audrey Hepburn: Began her career on Broadway in the 1950s, performing in shows like "High Button Shoes" and "Ondine."

2. Julia Roberts: Made her Broadway debut in the 1987 revival of "Steel Magnolias" and later starred in "Six Degrees of Separation."

3. Neil Patrick Harris: Got his start on Broadway in the 1990s, performing in shows like "Rent" and "Hedwig and the Angry Inch."

4. Hugh Jackman: Starred in the 1999 Broadway production of "Sunset Boulevard" and later appeared in "Oklahoma!" and "The Boy from Oz."

5. Kristin Chenoweth: Began her career on Broadway in the 1990s, performing in shows like "Steel Pier" and "You're a Good Man, Charlie Brown."

B.2 MT-Bench Example

Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

B.2.1 LLaMA-2-13B Base Output

This travel blog post should be engaging and informative, providing readers with an inside look at what makes Hawaii a unique and exciting destination. The post should be well-researched and include interesting facts and details about the destination, as well as practical information for travelers. Include a mix of text and images to illustrate the points made in the blog post and make it visually appealing.

B.2.2 LLaMA-2-13B Chat Output

Title: Aloha! Discovering the Heart of Hawaiian Culture and Must-See Attractions

Introduction:

Hawaii, the tropical paradise in the Pacific, is a destination that offers a unique blend of stunning natural beauty, rich cultural heritage, and unparalleled adventure. Recently, I embarked on a journey to explore the islands of Oahu, Maui, and Kauai, and I was left with unforgettable memories and a deeper appreciation for the Hawaiian culture. In this blog post, I'll share my top cultural experiences and must-see attractions that will inspire you to plan your own dream trip to Hawaii.

Cultural Experiences:

1. Attend a traditional Hawaiian luau: A luau is a quintessential Hawaiian experience that offers a glimpse into the island's rich cultural heritage. I attended a luau at the Waikiki Starlight Luau on Oahu, which featured a traditional Hawaiian feast, live music, and authentic hula dancing.
2. Visit a Hawaiian temple: Hawaii is home to numerous temples, or heiaus, that are sacred to the native Hawaiian people. I visited

the Pu'uuhonua o Honaunau National Historical Park on the Big Island, which is a historic temple ground.

3. Learn to hula: Hawaiian dance is an essential part of the island's culture, and I had the opportunity to learn the art of hula from a local dance teacher.

B.2.3 LLaMA-3-8B Base Output

Incorporate SEO strategies, formatting, and a conclusion for a well-rounded piece.

Aloha, fellow travelers! Hawaii is a dream destination, and I'm

here to share my recent trip to the beautiful islands. Join me as we explore the vibrant culture, stunning landscapes, and unforgettable experiences that await in this tropical paradise.

Hawaii is a melting pot of cultures, with influences from Polynesia, Asia, and the United States. As I stepped off the plane, I was greeted by the warm hospitality of the locals and the sweet aroma of freshly caught fish.

B.2.4 LLaMA-3-8B Instruct Output

Island Magic: Unveiling the Hidden Gems of Hawaii

As I stepped off the plane in Honolulu, the warm tropical air enveloped me, and the sweet scent of plumeria flowers filled my senses. My recent trip to the Hawaiian Islands was a dream come true, and I'm excited to share with you the incredible experiences and must-see attractions that left me enchanted.

A Cultural Immersion in Honolulu

My journey began in Honolulu, the capital city of Oahu, where I immersed myself in the rich cultural heritage of the islands. I visited the Bishop Museum, the largest

museum in Hawaii, which showcases the history and traditions of the native Hawaiian people. The museum's collections include an impressive array of artifacts, including ancient Hawaiian tools, traditional clothing, and even a replica of a 19th-century Hawaiian village.

Next, I headed to the Honolulu Fish Auction, where I witnessed the vibrant local fish market in action. The sights, sounds, and smells of the bustling market were a true sensory experience, and I was fascinated by the variety of fresh seafood on display.

C Jensen-Shannon Distance Results

See Table 5.

Corpus 1	Corpus 2	JSD
Nigeria	LLaMA2-base	0.6913
Nigeria	LLaMA2-chat	0.6946
US	LLaMA2-base	0.7062
US	LLaMA2-chat	0.7128
Nigeria	LLaMA3-base	0.6915
Nigeria	LLaMA3-chat	0.7091
US	LLaMA3-base	0.7064
US	LLaMA3-chat	0.7224
Nigeria	US	0.6617
LLaMA2-base	LLaMA2-chat	0.6486
LLaMA3-base	LLaMA3-chat	0.6691
LLaMA2-base	LLaMA3-base	0.6062
LLaMA2-chat	LLaMA3-chat	0.6011

Table 5: Jensen–Shannon distance (JSD) between different corpora. Values rounded to 4 significant figures.

D Downloading English corpora

We downloaded the COCA sample from <https://www.corpusdata.org/formats.asp>. We chose the linear text random sample provided (0.1% of the corpus). We excluded the files on spoken word and TV as their dialogue formatting biases the bigram analysis.

We downloaded ICE Nigeria from <https://sourceforge.net/projects/ice-nigeria/>.