Stylistic Transfer from Annotator Communities to Large Language Models

Anonymous ACL submission

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 remain poorly understood. We investigate the stylistic transfer from Nigerian annotators to LLaMa 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 generate corpora from both model families and measure linguistic style by computing the difference-in-difference of Jensen-Shannon distance of bigram frequencies between model outputs and corpora of Nigerian English and USA English. We found that while pre-trained model variants from both LLaMa 2 and LLaMa 3.1 families exhibit similar proximity to both English variants, the LLaMA 2 model post-trained before the Nigerian shutdown moved towards Nigerian English, while the LLaMa 3.1 model post-trained after the Nigerian shutdown moved away from Nigerian English. Qualitatively, we found that post-trained LlaMa 2 models use significantly less contractions, in line with Nigerian English speakers opting to use a formal register due to its role as an index of knowledgeability. Our findings suggests that annotator communities can imprint linguistic style on large language models, with potential implications including a disproportionally high rate of false positives by AI plagiarism checkers on users who share similar linguistic styles with annotator communities.

1 Introduction

Large language models (LLM) undergo two phase 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, uses interactions with humans to make the model useful as an assistant. Specifically, a human *annotator*, often paid, will choose between two model outputs given the same instruction, or prompt. The language model will then be shaped towards one whose outputs are preferred by a human, and thus this process is known as reinforcement learning from human feedback (Ouyang et al., 2022). Models that underwent the second phase are known as *chat* or *instruct* models.

LLM companies hire annotators to train their models to be useful for humans. However, aside from transferring the ideal human preference of an assistant-like behaviour, 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 cost is cheap, since annotation work could be done remotely and with the assumption that the ideal model preferred by these annotators are similar to the ideal model preferred by the user base. Countries with documented annotator communities include Nigeria, Kenya, and the Philippines (Wang et al., 2022). This sociolinguistic imbalance of providing annotator communities that speak a certain a 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 compared this difference to annotator communities.

In this paper, we study the stylistic transfer from Nigerian annotator communities to LLaMa models. We compared the model outputs of LLaMa 2 models (Touvron et al., 2023) and LLaMa 3.1 models (Grattafiori et al., 2024), the latter of which was released after data annotation services 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 USA English using the Jensen-Shannon distance (JSD). We found that although both LLaMa 2 and

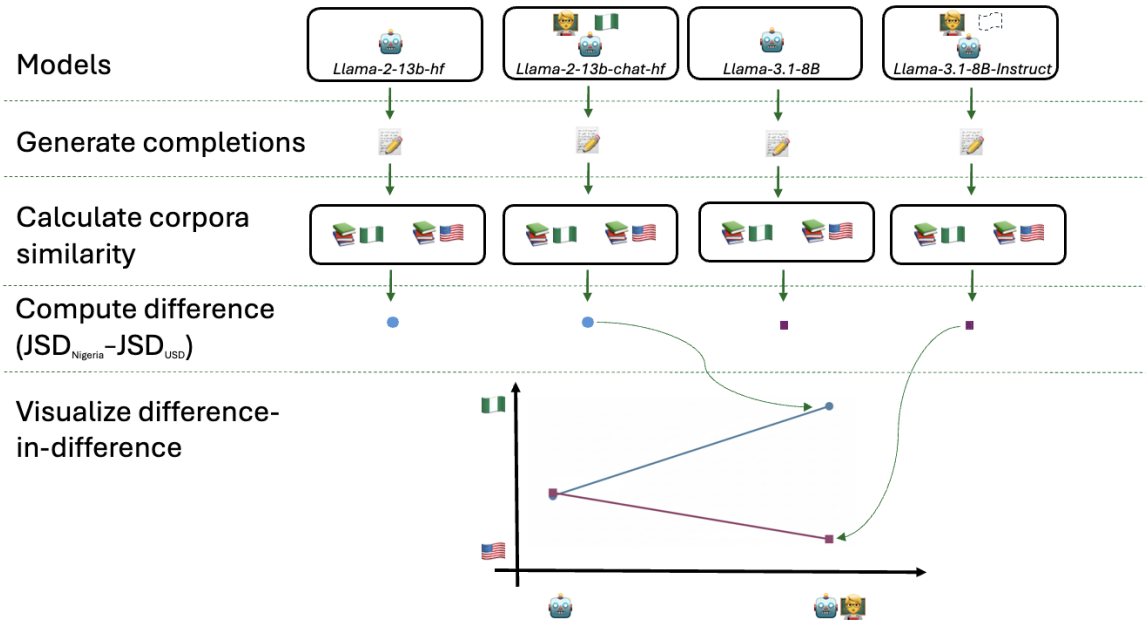

Figure 1: Our setup. (First row) Two similar model families with base and post-trained variants, with Nigerian annotators present in one but not in the other. (Second row) All four models are prompted on AlpacaEval and MT-Bench to generate model-specific corpora. (Third row) The similarity between the model corpora and the corpora in Nigerian English and USA 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.

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 USA English. In Section 4.2, we show that both 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 checkers (Mitchell et al., 2023), we note that such plagiarism checkers could be disproportionately biased against Global South communities who have annotators in the post-training pipeline.

2 Related Work

Grieve et al. (2025) explored the sociolinguistics of large-language models and argued that the current misalignment of large language models stem from selective procurement of training data and annotators that are easily obtainable but not necessarily representative of the user base of the model. Juzek and Ward (2024) explored the transformation of Scientific English after the introduction of large language models by identifying 21 "focal words" that exploded in popularity in published abstracts and model-generated completions. Juzek and Ward (2025) found that post-training in language models affects word choice when given the task to complete half of an abstract. (Lu et al., 2025) used n-grams to characterize creativity and for language model output detection.

3 Methods

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

[1] Code is available at REDACTED and REDACTED for corpora production and analysis, respectively.

3.1 Corpora Identification and Production

To control for the presence of a specific annotator community on the training of a large language model, we first identify the Nigerian annotator community, as it has been hypothesized that the recent explosion in popularity in the word "delve" in written material is directly linked to the popular use of "delve" in Nigerian English (Hern, 2024). Nigeria is known to be a popular

We then identify two model families where one was exposed to the Nigerian annotator community while the other 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 that are 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 popular among LLM companies, had shutdown access in Nigeria to its annotation platform Remotasks (see Table 4 for a timeline). It is highly likely that Meta used the services from Scale AI 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 Remostasks closed 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 big difference in style after post-training.

We now turn to obtaining a distribution of the model English, USA 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 USA 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). 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 USA 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 lower-casing 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 English variant. For each model variant, we then calculate the difference in JSD from the Nigeria and USA variants. We then compute the difference in this difference across base and post-trained model variants. The ΔJSD y-axis title in Figure 2 refers to $\text{JSD}(\text{model English}, \text{USA English}) - \text{JSD}(\text{model English}, \text{Nigeria English})$. 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 the USA corpus) and the post-trained LlaMa 2 corpus (as compared to the pre-trained version, and LlaMa 3).

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

$$C = \{\text{L2-chat}, \text{L2-base}, \text{L3-chat}, \text{L3-base}, \text{Nig}, \text{USA}\}.$$

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 $f_c(i)$. For any corpus

Subscript	Description
L2-chat	Llama-2 Chat (post-trained)
L2-base	Llama-2 Base (pre-trained)
L3-chat	Llama-3 Chat (post-trained)
L3-base	Llama-3 Base (pre-trained)
Nig	Nigerian English corpus
USA	USA English corpus

Table 1: Subscripts used in bigram frequency notation

$c \in C$, denote the set of bigrams present in that corpus as

$$B_c = \{\, i \in B \mid f_c(i) > 0 \,\}.$$

We compute relative bigram frequencies as usual:

$$r_c(i) = \frac{f_c(i)}{|B_c|}.$$

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

$$\text{score}_{\text{Nig}}(i) = r_{\text{Nig}}(i) - r_{\text{USA}}(i)$$

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

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

We compute the relative score of a bigram $i \in B$ towards 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

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

A bigram that achieves a high positive divergence score could be explained by appearing much more often in the Nigerian corpus (relative to the USA 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, it is also achieved when it appears more often in the USA corpus relatively, and in the post-trained Llama 3 corpus relatively. To tell apart these two causes, we report the sign of $\text{score}_{\text{Nig}}(i)$, where a positive sign implies the former, and the negative sign the latter.

4 Results

4.1 Distributional Results

The results are plotted in Figure 2. The exact 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 verifies that the pretraining pipeline for both model families have no significant impact on style, likely due to similar pretraining curriculum. For post-training however, the LLaMa 2 Chat model is closer to Nigerian English than its base model variant, while the LLaMa 3.1 Instruct model moved further away from Nigerian English than 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 attribution work done by Nigerian workers while LLaMa 3.1 did not.

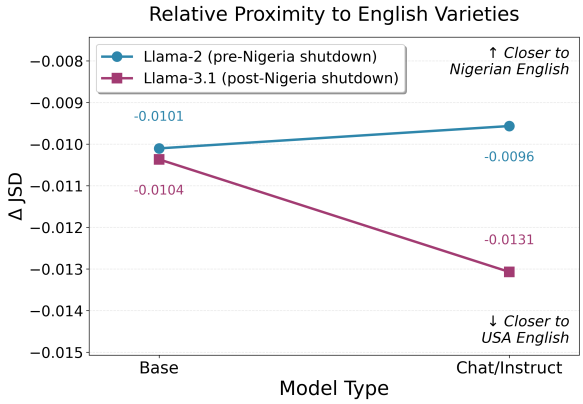

Figure 2: The base models of LLaMa 2 and LLaMa 3 show similar differences in the Jensen-Shannon distance (JSD) between Nigerian and USA English, indicating similar styles in the continuum of Nigeria English – USA English. After post-training, LLaMa 2 showed an increase in the difference of JSD (moving 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).

4.2 Qualitative Results

We report the top 10 bigrams by divergence score in Table 2, conditioned on the sign on $\text{score}_{\text{Nig}}$.

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-Nig list where instead bigrams that could be contracted appeaered, like "it is" and "is not". This aligns with existing literature where American English generally allows for the

Rank	L2-Nig	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 in divergence scores with positive $\text{score}_{\text{Nig}}$ (L2-Nig) and negative $\text{score}_{\text{Nig}}$ L3-US. The middle column shows the bigrams that are over-represented in both the Nigeria corpus and the post-trained Llama 2 corpus. The right column shows the bigrams for over-representation in both the USA corpus and the post-trained Llama 3 corpus.

use of an informal register that involves contractions (it's, don't), while Nigerian English shuns away from informal register and instead predominantly uses the formal register. In fact, Nigerians use the phrase "colloquial English" to describe bad English (Kperogi, 2010). Furthermore, English in Nigeria is exclusively being taught in school instead of as a native language used within families for most Nigerians, hence formality is especially emphasized through the schooling process and is thus possessed by Nigerian English speakers. 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, command a form of English that has 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 within English speakers in Nigeria to use contractions in their writing, with the paper further describing that most students who use contractions claimed that its usage has affected their fluency of English.

4.2.1 To delve or not to delve?

To move towards 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".

Corpus	"delve"	all bigrams	Percent (%)
Nig	5	400,782	0.001247
USA	22	7,004,385	0.000314
L2-base	1	256,597	0.000390
L2-chat	7	223,695	0.003131
L3-base	2	243,634	0.000821
L3-chat	12	243,718	0.004925

Table 3: Frequency of bigrams involving the substring "delve".

Directionally, the Nigerian corpus uses "delve" roughly 4x as much as the US corpus, as commonly believed that Nigerians use the phrase more often than American speakers (Hern, 2024). Meanwhile, the post-trained Llama-2 model uses "delve" 8x times more often than its pre-trained version, while the Llama-3 model only uses it 6x times as often. It is tempting to declare that indeed Nigerian annotators has made Llama 2 delve into the usage of more "delves" compared to Llama 3, but the rare occurrences of such bigrams reported in Table 3 render us unable to draw strong conclusions.

5 Discussion

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

This difference between annotator population and user population can be thought as contributing towards a formation of a *community of practice* (Eckert and McConnell-Ginet, 1992) among large language models, where the act is defined through the post-training with Nigerian English speakers. `LLaMa-2-13b-chat-hf` model will then be a model organism 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 serves 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 USA 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 JSD differences 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 USA English while the other is in Nigerian English. We do not find any such corpora available online.

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 we have shown that the distribution of model outputs move towards English varieties that are closer to annotator communities in the Global South, one concern could be that such LLM detector could disproportionately misidentify genuine human outputs in the Global South as machine output. Further research is welcome in this area, including running such LLM detectors on corpus of various English varieties and see if there is a consistent higher false positive rate on English varieties associated with annotator communities.

Replicating the experiment using LLaMa 2 7B instead of LLaMa 2 13B might be a fairer comparison to LLaMa 3 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 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).

References

Russell Brandom. 2024. Scale AI's Remotasks platform is dropping whole countries without explanation.

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. *AFREV 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. Publisher: Frontiers.

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*.

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.

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.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025. AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text. *arXiv preprint*. ArXiv:2410.04265 [cs].

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. ISSN: 2640-3498.

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].

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 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. Publisher: MIT Press.

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.

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,
```

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)

```
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:

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'uhonua 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
usa	llama2-base	0.7062
usa	llama2-chat	0.7128
nigeria	llama3-base	0.6915
nigeria	llama3-chat	0.7091
usa	llama3-base	0.7064
usa	llama3-chat	0.7224
nigeria	usa	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 divergence (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 downloaded ICE Nigeria from https://sourceforge.net/projects/ice-nigeria/.