Native Design Bias: Studying the Impact of English Nativeness on Language Model Performance

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel at providing information acquired during pretraining on large-scale corpora and following instructions through user prompts. However, recent studies suggest that LLMs exhibit biases favoring Western native English speakers over non-Western native speakers. Given English's role as a global lingua franca and the diversity of its dialects, we extend this analysis to examine whether non-native English speakers also receive lower-quality or factually incorrect responses more frequently. We compare three groups—Western native, non-Western native, and non-native English speakers—across classification and generation tasks. Our results show that performance discrepancies occur when LLMs are prompted by the different groups for the classification tasks. Generative tasks, in contrast, are largely robust to nativeness bias, likely due to their longer context length and optimization for open-ended responses. Additionally, we find a strong anchoring effect when the model is made aware of the user's nativeness for objective classification tasks, regardless of the correctness of this information. Our analysis is based on a newly collected dataset with over 12,000 unique annotations from 124 annotators, including information on their native language and English proficiency.

1 Introduction

English, as the global lingua franca, is predominant in large-scale text corpora used to train Large Language Models (LLMs) (Ziems et al., 2023; Zhang et al., 2023), including widely used datasets like CommonCrawl. These datasets are primarily tailored to an English-speaking audience located in the United States, and are mainly composed of privileged English dialects from wealthier educated urban zones (Talat et al., 2022; Ziems et al., 2023; Ryan et al., 2024; Gururangan et al., 2022). This biased training dataset composition permeates the

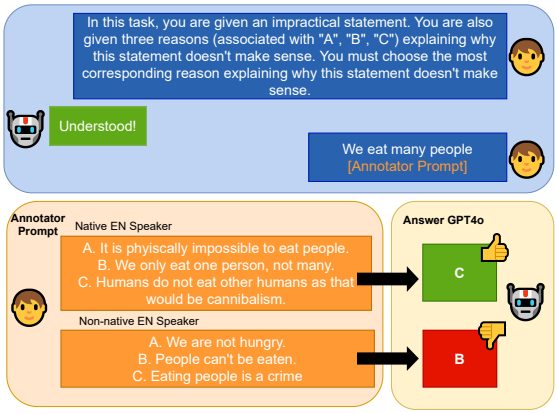


Figure 1: An example prompt of a native and non-native English speaker and the corresponding output given by GPT4o, where *Annotator Prompt* represents the placeholder for the annotations. The desired output is C. The model selects the wrong answer for the non-native English speaker, while semantically the same message was conveyed. Sentence B is a direct translation from the non-native speaker's native language and expresses the same idea as Sentence A from the native speaker.

LLM, resulting in models tailored to these English dialects (Santy et al., 2023; Hall et al., 2022). This highlights underlying design biases in LLMs, a phenomenon where design choices result in improved downstream performance for specific subpopulations (Santy et al., 2023). Consequently, their effectiveness considerably decreases when prompted in other languages or in underrepresented English dialects (Lai et al., 2023; Zhang et al., 2023; Bang et al., 2023; Ziems et al., 2023; Ryan et al., 2024).

LLMs are highly sensitive to prompt formulations (Beck et al., 2024; Chakraborty et al., 2023). Ryan et al. (2024) show how models' responses are tailored to Western English dialects, with prompt selection impacting LLMs' preference tuning. Therefore, prompting models in other dialects can result in performance differences due to these design biases. Ziems et al. (2023) even provide a dataset covering multiple English dialects. How-

ever, unlike those studies focusing only on English dialects from native English-speaking countries, our research also incorporates participants from countries where English is not an official language. We assess if word sensitivity in prompts disproportionately benefits native English speakers, leading to better model performance. In this case, the model has an inherent native language bias.

In this paper, we examine performance differences when LLMs are prompted by speakers from three groups: Western native (WN), non-Western native (NWN), and non-native (NN) English speakers. We find performance differences when LLMs are prompted by both NWN and NN versus WN speakers. More specifically, some models generate inaccurate responses for non-native speakers and rate the WN prompts more positively than intended. We also highlight how LLMs are more robust against this native bias on generative tasks. Moreover, we uncover deeply embedded bias within models towards native speakers for the classification tasks, as explicitly stating that a prompt writer is non-native leads to lower model performance compared to stating that the writer is native regardless of the correctness of this information. We collect a dataset comprising over 12,000 unique prompts from native and non-native English speakers worldwide and demonstrate how different prompt formulations can lead to worse performance despite conveying the same message. An example prompt from our dataset is shown in Figure 1.

Our contributions are as follows: 1) We quantitatively and qualitatively analyze how LLM performance differs between native — both Western and non-Western— and non-native English speakers on objective and subjective classification tasks[1], as well as generative tasks. 2) We investigate the impact of explicitly stating user nativeness. 3) We publish our multilingual instruction-tuning dataset and code used for the experiments[2] containing over 12,000 unique prompts from diverse native and non-native English speakers, with translations into eight languages.

2 Related work

Model Positionality and Design Bias. Model positionality, coined by Cambo and Gergle (2022), refers to the social and cultural position of a model, influenced by the stakeholders involved in its development, such as annotators and developers. This positionality affects the inclusivity of LLMs, as they evolve with certain biases that may disadvantage specific populations (Cambo and Gergle, 2022; Santy et al., 2023). Design biases arise when researchers make choices that improve model performance for specific sub-populations (Santy et al., 2023). A notable example is the overrepresentation of English pretraining corpora, which leads to disproportionate performance improvements in English compared to other languages (Qin et al., 2023; Blasi et al., 2022; Joshi et al., 2020).

Effect of demographic background on LLM performance. Recent literature suggests that LLM performance on subjective tasks is influenced by the demographic attributes of the user (Beck et al., 2024; Santy et al., 2023). Moreover, when assigned a persona, LLMs reveal deep inherent stereotypes against various socio-demographic groups (Cheng et al., 2023; Gupta et al., 2023; Deshpande et al., 2023). For example, Gupta et al. (2023) show how ChatGPT3.5, when asked to solve a math question while adopting the identity of a physically disabled person, generates that it cannot answer the question, as a physically disabled person. Furthermore, Barikeri et al. (2021) demonstrate that LLMs can infer demographic attributes from dialog interactions. Additionally, research shows biases in favor of Western populations (Santy et al., 2023; Durmus et al., 2023). In model alignment literature, Ryan et al. (2024) show this similar bias within preference models and Gururangan et al. (2022) illustrate that even within a Western country like the US, GPT3 prefers the more privileged dialects. Furthermore, Hofmann et al. (2024) illustrate how models show covert biases towards African American English speakers. Additionally, Kantharuban et al. (2024) show how LLMs express racially stereotypical recommendations regardless of whether the user explicitly or implicitly revealed their identity. Finally, Ziems et al. (2023) have provided a cross-dialectal English dataset for countries with English as an official language. Building on these findings, we extend the research to include non-native English speakers, who use English dialects influenced by their native languages. Furthermore, while Gupta et al. (2023) assign a persona to the model, we analyze performance differences of LLMs both with and without explicitly informing the model about the user's native language and thus

[1] By subjective tasks, we mean classification tasks where the correct answer depends on the subjective interpretation as explained in Beck et al. (2024)

[2] https://anonymous.4open.science/r/native_en_bias-EDC5/README.md

with and without assigning a persona to the prompt writer. However, note that models providing different answers based on demographic background is not always problematic as noted in Jin et al. (2024).

3 Methodology

Given the sensitivity of LLMs to prompt formulation (Beck et al., 2024; Chakraborty et al., 2023), the diversity of English dialects (Ziems et al., 2023; Ryan et al., 2024), and alignment of models towards Western native English speakers (Ryan et al., 2024; Santy et al., 2023; Gururangan et al., 2022), we investigate whether LLMs exhibit bias in favor of native English speakers over non-native speakers. More specifically, we aim to answer the following research questions:

1. Do LLMs perform differently when prompted by native vs. non-native English speakers? And is there a performance difference for different groups of native English speakers?
2. Are certain tasks more prone to performance disparities between native and non-native speakers?
3. Which tasks, if any, remain robust to these differences?
4. Are these trends consistent across models, or do they vary by architecture?
5. Does explicitly providing information about a speaker's nativeness amplify performance gaps?

To answer these research questions, we collected a new dataset containing both classification and generation tasks, along with information about the native languages of the annotators, as this is lacking in existing literature. An overview of our methodology and experimental setup is shown in Figure 2.

3.1 Dataset

Our dataset was constructed including samples from ten diverse task datasets from various natural language instruction tasks[3] (Mishra et al., 2022; Wang et al., 2022), covering classification (subjective and objective) and generation tasks. These tasks, representing typical LLM interactions, follow a standard instruction pattern and should not inherently favor native speakers. The tasks include paraphrasing, article generation based on a summary or title, sentiment analysis, natural language understanding, multiple-choice answering, and review writing. This last task is the subjective classification task in our experiments.
From each original dataset, we randomly selected 100 examples, manually ensuring they were correctly annotated and free of offensive language. Additionally, one extra example per dataset served as a tutorial for the annotator to get used to the task.

More information about the different tasks included in our dataset can be found in Appendix A.

3.2 Annotations

We required all annotators to have a minimum English proficiency level equivalent to a high school or university-level proficiency to establish a baseline, ensuring that performance differences stem primarily from dialectal variation rather than overall language proficiency. Each annotator worked on 20 to 240 examples. We gathered them through direct recruitment, opting for an open annotation process rather than an existing annotation platform to ensure high-quality annotations. All annotators were reimbursed at a minimum rate of 12.11 euros per hour.

In addition to gathering self-reported linguistic data—such as native language, English proficiency, and frequency of English use—we also collected information from native English speakers about how they acquired the language. This allows us to compare three groups: the non-native speakers (NN), Western native speakers (WN), and non-Western native speakers (NWN). The term *Western native* here refers to native English speakers who learned English from native speakers from countries like the UK, US, Australia, or Canada.

Annotators performed different tasks depending on the assigned datasets. An example annotation is shown in Figure 1, where a task definition is provided together with an impractical statement. The annotator has to provide the `[Annotator PROMPT]` based on the task definition and the desired output, which is *C* in this example. We identified the `[Annotator PROMPT]` per example depending on the dataset. More details about the annotation setup including information about the annotator prompts per dataset can be found in Appendix B.

The authors manually validated the annotations before including them in the final dataset, deeming one invalid if it met any of the following criteria: 1)

[3] https://github.com/allenai/natural-instructions

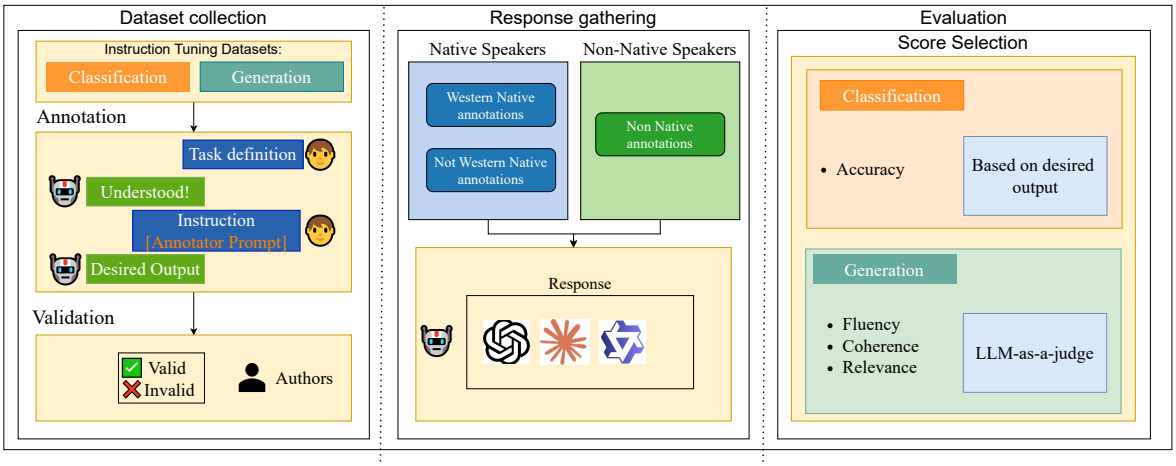


Figure 2: Methodology and experimental setup. The left part shows the data collection steps. After gathering the different datasets, study participants annotated the examples. Then we validated them and used them as input to generate LLM responses. The right part of the figure shows the evaluation phase, where we gathered the respective scores depending on the task.

The response was unrelated to the task, i.e. *"I don't know / understand"*, or a response for a different topic or question. 2) The response contained (part of) the answer. 3) The response did not follow the required format or task definition. 4) The annotator misunderstood the task. Examples per validation criterion are included in Appendix C.

After validation, we removed instances with more than 50% rejected annotations to ensure the quality of the dataset. In total, we removed 12 examples entirely and a total of 162 individual annotations. Our final dataset contains 12,519 annotations from 124 annotators. More information on the dataset statistics can be found in Appendix D[4].

4 Experimental setup

4.1 Gathering LLM responses

Using gathered annotations, we conducted experiments with the chat-versions of well-established LLMs, as these are used in daily life. An overview of the checkpoints per model is shown in Appendix F. We included GPT3.5[5], GPT4o[6], Haiku (Anthropic, 2024), Sonnet (Anthropic, 2024), using the appropriate APIs, and Qwen1.5 7B[7] (Bai et al., 2023) in line with the provided licenses and all consistent with the intended use. This set includes models of varying sizes, different performances, and from different developers, ensuring a diverse representation. Moreover, Qwen, developed by Chinese researchers, provides an interesting comparison in terms of design bias.

To answer our predefined research questions mentioned in Section 3, we first ran our experiments for all models without any additional information. Next, to answer the last research question, we provided information about the nativeness of the prompt writer to the LLM. To see whether the LLM entails an inherent bias against native speakers, we included both correct and incorrect information.

4.2 Evaluation

To measure the bias within the models, we look into the performance difference between the the native and non-native speaking groups. We measure these performance differences across classification tasks and generative tasks. Concretely, native bias measured for the classification tasks is defined as follows:

$$\Delta_{\text{native}} = \phi\left(\mathcal{M}\left(\mathcal{T} \mid x_{\text{native}}\right), \psi\right)$$
$$\Delta_{\text{non-native}} = \phi\left(\mathcal{M}\left(\mathcal{T} \mid x_{\text{non-native}}\right), \psi\right)$$

with native bias discriminative = $\Delta_{\text{native}} - \Delta_{\text{non-native}}$, template $\mathcal{T}$, user prompt x, model $\mathcal{M}$, accuracy ϕ, and original ground truth ψ. The native generative bias is defined as follows:

$$\Delta_{\text{native}} = \phi\left(\mathcal{M}\left(\mathcal{T} \mid x_{\text{native}}\right)\right)$$
$$\Delta_{\text{non-native}} = \phi\left(\mathcal{M}\left(\mathcal{T} \mid x_{\text{non-native}}\right)\right)$$

with native bias generative = $\Delta_{\text{native}} - \Delta_{\text{non-native}}$, template $\mathcal{T}$, user prompt x, model $\mathcal{M}$, and performance metric ϕ. The Western native bias can be

[4]Due to the nature of the tasks, we did not calculate inter-annotator agreement scores, as annotators were providing prompts, and invalid prompts were filtered out.

[5]https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

[6]https://openai.com/index/hello-gpt-4o/

[7]We ran the experiments for Qwen using A100 GPUs.

similarly inferred by splitting the native group into a Western native and non-Western native group.

Classification tasks. When assessing classification tasks, we focus on the accuracy of the predictions. We only consider classifications as correct if they follow the instructions correctly or if the correct classification can be determined automatically.

Generative tasks. In assessing the generative tasks, we include the following metrics: fluency, coherence, and relevance (Bavaresco et al., 2024). All metrics were evaluated using a Likert scale: fluency was rated on a 3-point scale. Coherence and relevance were scored on a 5-point scale. Fluency is defined as the quality of the generated text in terms of grammar, spelling, etc. Coherence assesses the collective quality of the sentences. Finally, relevance refers to the inclusion of important content in the generated text. These definitions are based on the ones used in Bavaresco et al. (2024). The prompt templates used are shown in Appendix G. All results were rescaled to a range of 0 to 1 to ensure clarity.

We evaluated the performance of the generative tasks using an LLM-as-a-judge approach, specifically leveraging Llama-3.3-70B-Instruct to assess each prompt's output. To ensure reliability of the LLM-generated responses, we manually annotated 100 examples and observed a correlation of 81.3% with the model's evaluations.

5 Results

Below, we analyze the results from our experiments answering each of the research questions. Throughout the next paragraphs, we analyze the performance of the native speakers— consisting of Western native speakers (WN) and native speakers that are non-Western (NWN)— and non-native English speakers (NN).

The WN group performs best for the objective classification tasks, outperforming both NWN and NN. This is shown in Figure 3, where the average performance per group on the objective classification tasks is displayed on the left. WN speakers achieve the highest overall performance in objective classification tasks, reinforcing findings from previous research (Hofmann et al., 2024; Ryan et al., 2024) that models favor Western privileged dialects. In contrast, NWN and NN English speakers perform similarly, with the NN group slightly outperforming NWN speakers. However, this difference is minimal and not substantial

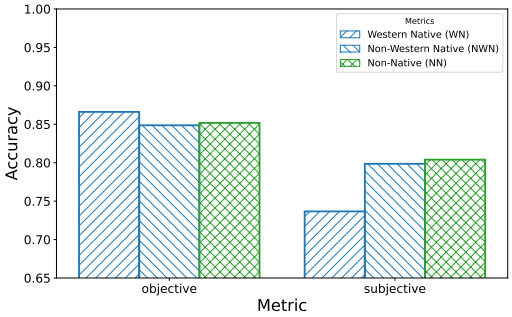


Figure 3: This figure shows the average performance per group and task type. You can see how for the objective tasks, the (western) native group performs slightly better then the not western or non-native group. For the subjective classification task, this distinction becomes more pronounced, with the non-native and western native groups outperforming the other groups. These results are the average results for all models and runs. We adjusted the y-axis to range from 0.65 to 1 for clarity.

enough to draw strong conclusions. The performance gap between WN and the other groups, however, suggests the advantage of Western dialects.

The WN group performs worst for the subjective classification task as models predict their rating more positively than actually intended. The right part of Figure 3 shows this opposite effect for the subjective classification tasks. For these tasks, both the NN and NWN show again similar performance and are now outperforming the WN group. This finding is remarkable, as it contradicts the results in the subjective classification literature (Santy et al., 2023; Durmus et al., 2023). When further analyzing the results, we find that for the Western native English-speaking group, we find that the models often predict the rating more positively than actually intended. While for the NN and NWN groups, GPT4o predicted around 50% of all wrongly predicted annotations to be more positive than intended, this was around 70% for the WN English-speaking group for GPT4o indicating cultural differences. Appendix H includes more information on the different answer distributions per model.

The generation tasks are more robust against (Western) native bias. Figure 4 shows the average performance scores for all models and groups. The figure shows that no clear performance difference exists among de groups compared to the classification results. A slight performance difference favoring the WN group is found for coherence, with the NWN and NN groups performing simi-

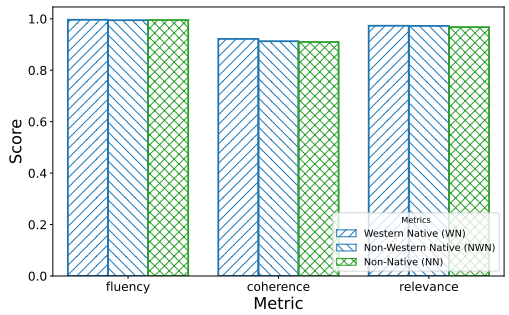


Figure 4: This figure shows the overall performance across the three groups: (western) native speakers and non-native speakers. However, when looking into the coherence metric, we do see a preference for the western native group. The results show how there is no difference regarding fluency and only a slight performance difference when comparing the native categories with the non-native category for relevance. We rescaled the results so that they range from 0 to 1.

larly. Nevertheless, the performance differences are not substantial. Therefore, we conclude that generation tasks are rather robust against (Western) native bias. Nevertheless, when zooming in on the results, we find discrepancies depending on the specific task at hand. These are shown in Appendix L. For two of the datasets, namely Story Cloze and Paraphrase, we find differences in terms of the coherence scores. More specifically, the WN group is here outperforming both the NWN and NN groups. Interestingly, these two tasks also include the smallest written annotations by the prompt writer and generated text by the model. Additionally, the CODA19 dataset comprises medical articles that utilize specialized medical terminology. Given that most annotators were unfamiliar with this vocabulary, native English speakers (WN and NWN) did not have a specific advantage over non-native speakers. Additionally, research articles are commonly written in English by authors from various backgrounds. Therefore, this specific task might be robust against the native versus non-native preference.

(Western) Native bias is model-dependent for the classification tasks. Figure 5 illustrates that the preference for WN speakers over NWN speakers in objective classification tasks varies by model. Notably, this trend is pronounced in GPT-3.5 and GPT-4o, while Qwen and Claude models show little to no performance difference between WN and NWN speakers. Interestingly, OpenAI's models even appear to favor NN speakers over NWN speakers. Moreover, it is interesting to see how the Qwen model, developed by Chinese researchers shows almost on par results between both native groups. Additionally, within a model family, the performance disparity increases with model size and overall capability. This aligns with prior research showing a positive correlation between model size and biases, such as gender bias (Tal et al., 2022). Furthermore, Sclar et al. (2023) demonstrate that prompt sensitivity does not decrease as models scale, suggesting that larger models may reinforce rather than mitigate biases. Also for the subjective classification tasks, the results are strongly model-dependent. However, all models do provide the lowest performance for the WN group. For the generative results, on the other hand, all models show similar trends as is shown in Appendix L.

Objective classification tasks are largely affected by adding information about the nativeness of the prompt writer. Figure 6 shows the effect of providing the model with (in)correct information about the nativeness of the annotator on model performance. This figure clearly shows how the additional information of the nativeness highly affects the results. Adding correct information about the nativeness results in a clear performance preference for the native group, while adding incorrect information results in a preference for the non-native group. Moreover, it not only shows how the performance is influenced by this information, but it also reveals deeply embedded bias towards non-native speakers. Adding this information results in a different performance, where the model focuses more on the initial given information than on the prompt itself. This phenomenon is called anchoring.This term is used for human cognitive bias indicating that a person might insufficiently change its estimates away from an initially provided value (Jones and Steinhardt, 2022; Tversky and Kahneman, 1974). This effect is demonstrated in LLMs by Jones and Steinhardt (2022), who found that code generation models modify their outputs to align with related solutions included in the prompt. Moreover, also Nguyen (2024) shows how LLM responses are highly influenced by previously given information. Our results reveal a similar anchoring effect, where the model focuses on the additional information about the nativeness of the prompt writer, regardless of whether or not this information is correct. This anchoring effect

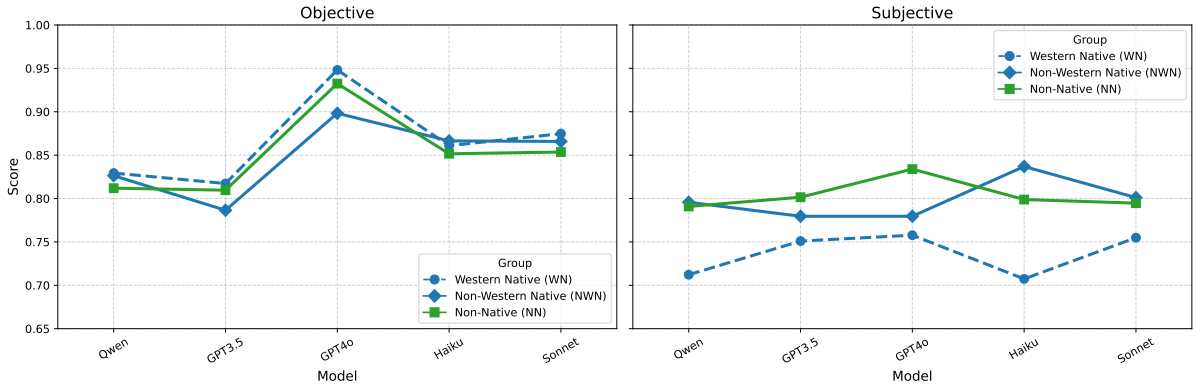


Figure 5: This figure shows the average performance for the different classification tasks per model and group. We see how both GPT models clearly prefer the Western native group, while the other models show similar preference for both native groups for the objective classification task. For the subjective classification tasks, the Western native group is the worst performing group for all models. We adjusted the y-axis to range from 0.65 to 1 for clarity.

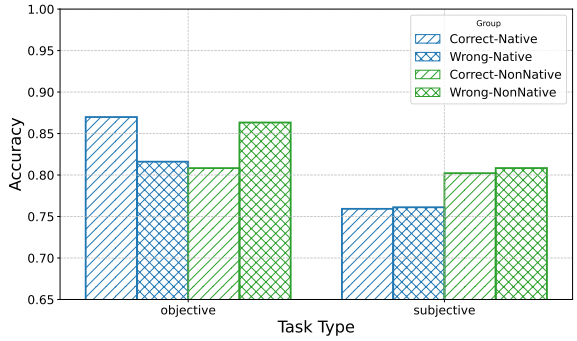


Figure 6: This figure shows the average performance per group and task type when the model is (in)correctly informed about the nativeness of the annotator. You can see how the information about the nativeness plays a significant role in the performance, whether the information was correct or not for the objective classification tasks. For the subjective classification tasks, however, this does not play a role. These results are the average results for all models and runs. We adjusted the y-axis to range from 0.65 to 1 for clarity.

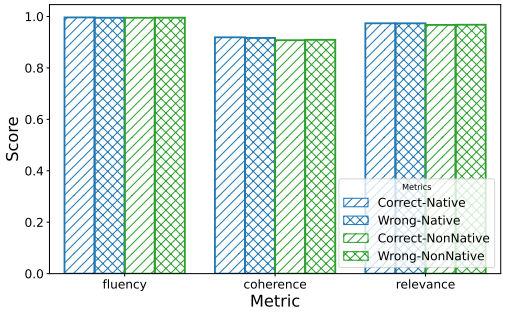


Figure 7: This figure shows a similar trend to the original experiments where no information was added. However, we do see that when adding information that the prompt writer was a native English speaker, the performance is slightly higher than when adding that the prompt writer is a non-native English speaker. We rescaled the results so that they range from 0 to 1.

was most clearly present for Sonnet. We find that Sonnet answered several questions in languages other than English, such as Spanish, French, or Indonesian, when responding as if interacting with non-native speakers. This resulted in a clear drop in performance as is also shown in Figure 11 in Appendix K. Note that this occurred both for native and non-native speakers. From the other models, we see that Qwen and GPT4o seem to be most robust against this added information. GPT3.5 and Haiku did show performance differences, however, not as pronounced as Sonnet. We manually analyzed examples for GPT3.5 and Haiku to gather more insight into the performance difference. GPT3.5 makes more mistakes when informed about the prompt writer being non-native, due to repetition of the instructions, rather than answering the question. Haiku explains the answers, arguing why one option is better than another, thereby failing to follow the instructions. If both answers are mentioned, we classify the response as inaccurate.

The subjective classification tasks and generation tasks are more robust against this additional information about the prompt writer's nativeness. In the subjective classification tasks, we observed only slight performance differences, with the non-native group consistently outperforming the native group. These experiments appear largely unaffected by the addition of information, as the non-native group remains the best-performing regardless of whether accurate or inaccurate details

about nativeness are introduced. Also for the generative tasks, the addition of information about the prompt writer's nativeness does not impact performance ranking, as shown in Figure 7. All different groups continue to perform similarly, regardless of the additional information provided.These findings suggest that models are more robust to nativeness cues in generative and subjective classification tasks than in objective classification tasks. This is likely due to their primary optimization for generation rather than classification, particularly given that we use the chat-based versions. Additionally, the longer context in both the initial prompt and generated output may reduce the impact of the anchoring effect.

6 Discussion

In our experiments, we define native bias as the model's performance disparity when prompted by native versus non-native English speakers. Additionally, we also further split the native speakers into two groups: Western Natives (WN) and Non-Western Natives (NWN). **In general, we find that there are performance differences when the model is prompted by people from different backgrounds.**

More specifically, we find an interesting overall preference towards the WN group, where the NWN group is performing similarly as the NN group. This aligns with literature showing how models are tuned towards western native English dialects. The subjective classification tasks on the other hand, favor the western native group the least, across all different models, contradicting findings from (Santy et al., 2023; Durmus et al., 2023). This is explained by the models interpreting the (western) native results more positively than intended.

When analyzing the generative results, we find that this task type is more robust and performing similarly for all evaluated groups. This is probably due to the longer context length in both input and generated output, which seems to help the models to perform similarly across different groups. Additionally, the model checkpoints used were also optimized for these generative tasks rather than classification tasks.

We show how the performance ranking of the three groups are also model-dependent for the classification tasks. More specifically, the two GPT models are even preferring the NN group over the NWN group on objective classification tasks. The other models show similar performance for both native groups, for the objective classification tasks. Also for the subjective tasks, we see how group preference depends on the model. Nevertheless, all models perform worst for the WN group. The generative tasks on the other hand seem to perform similarly across all models.

Finally, we show how a strong anchoring effect occurs when the model is made aware of the nativeness of the prompt writer for the objective classification tasks. The bias is so deeply engraved that informing the models about the nativeness of both groups results in a preference towards the group that was indicated as native, regardless of the correctness of this information, being led by this additional information rather than by the prompt itself. However, we find differences between the models. GPT4o appears most resistant to this anchoring effect, while Sonnet on the other hand even changes the language of the response based on this anchor. Furthermore, this anchoring effect is not clearly present for the generative results, probably due to the optimization of these models towards generative tasks compared to classification tasks, given that we used chat-versions.

7 Conclusion

In this work, we analyze bias in LLMs towards native English speakers. We analyze if models perform better for native compared to non-native English speakers and whether the models are even further tuned towards Western native English speakers. We find that there are performance differences between native and non-native prompts. More specifically, models are most accurate for the Western-native English speakers on objective classification tasks. A slightly lower performance is shown for the NWN group compared to the NN, nevertheless, we show that this is mostly model-dependent. Both GPT models seem even to prefer NN over NWN, while the other models in our analysis show similar performance for both native groups. Furthermore, we find a strong anchoring effect when information about the user's nativeness is added for objective classification tasks. Generative tasks seem to be in general more robust against this native bias, probably due to the longer context length and the optimization of the used models towards these generative tasks. For our experiments, we used a newly collected dataset consisting of over 12,000 unique prompts from a diverse set of annotators.

8 Limitations

Our dataset contained a very diverse set of annotators. Nevertheless, it would be interesting to have more study participants for every sub-population, such that general findings at sub-population level could be made as well. Furthermore, our experiments contained mostly annotators having a self-reported level of English of C1 and C2. It would be very interesting to analyze the effects on the performance of LLMs when prompted by people having different levels of English as this will probably also be impactful. Additionally, our results were only gathered for five different models. It would be insightful to extend this analysis to more models, as every model is trained differently and therefore these design choices might lead to different biases within the model. An important limitation of using LLMs and especially the closed-source variant thereof, is the lack of reproducibility of the results. We make available a multilingual dataset, however, have only analyzed the English answers. We leave the analysis of bias in the multilingual dataset for future research. Finally, we acknowledge how the LLM-as-a-judge implementation for gathering generative results might be suboptimal to human annotators due to model-specific biases. Therefore, we chose a different LLM than the ones we will evaluate to serve as a judge to avoid self-preference bias and we manually validated a sample. Given the high correlation between the manual annotations and the LLM annotations, we assume that the LLM annotations are representative.

9 Ethical considerations

We included human annotators in this study. All annotators were paid for the provided annotations and the annotations were done on a voluntary base. Moreover, our paper shows some of the consequences of unfair design choices when developing models. We think this work is important to highlight the necessity of taking into account multiple English dialects, as these models should work equally well for everyone. In this paper, we focus on the English language. We wanted to point out that even in English, this problem of not having enough diversified training data might also result in performance differences among certain populations. However, this does not mean that other languages do not require the same attention.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, E. Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's

languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Scott Allen Cambo and Darren Gergle. 2022. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. Zero-shot approach to overcome perturbation sensitivity of prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5698–5711, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. 2022. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, Miami, Florida, USA. Association for Computational Linguistics.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anjali Kantharuban, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. 2024. Stereotype or personalization? user identity biases chatbot recommendations. *arXiv preprint arXiv:2410.05613*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Jeremy K. Nguyen. 2024. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Dataset overview

We used the datasets as they were assembled by Mishra et al. (2022) and Wang et al. (2022). Table 1 shows an overview of the selected datasets, together with their task ID in the original instructions dataset. The task definition given in the table is the one we used when prompting the models. For CNN Dailymail and CODA19, this differs from the original task definition in the dataset because we flipped the task. Instead of letting our annotators write the article, we asked them to write the summary or title respectively. Datasets Abductivenli, Timetravel, Amazonfood, McTaco, TweetQA, and Commonsense are thus classification tasks, while datasets StoryCloze, CNN Dailymail, CODA19, and Paraphrase are generation tasks.

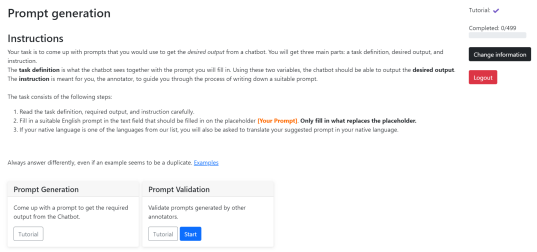


Figure 8: Screenshot of the landing page of the annotation platform.

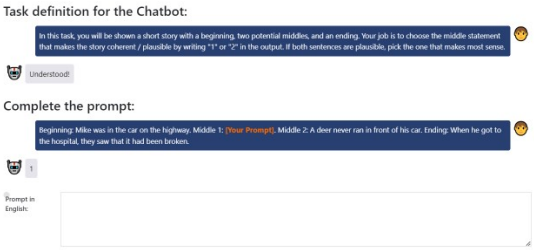


Figure 9: An annotation example of the Abductivenli dataset.

B Annotation set-up

We have set up an annotation platform to gather the annotations. The annotators first get information about the task. They will get a task definition, a prompt where part of the answer is marked out with the placeholder [YOUR PROMPT], and the desired output of the LLM. The annotators should complete the prompt such that the desired output would be generated by the LLMs. Figure 8 shows a screenshot of the landing page of the annotation platform together with annotation instructions. An example of an annotation that had to be annotated is shown in Figure 9. An example of the different `[Annotator PROMPT]` per dataset is shown in Table 2. We have anonymized all annotations by only providing the self-reported linguistic information in the dataset along with the user ID number.

C Annotation validation

Examples for each of the criteria of an invalid annotation are shown in Table 3.

Task ID	Name	Task Definition
task069	Abductivenli	In this task, you will be shown a short story with a beginning, two potential middles, and an ending. Your job is to choose the middle statement that makes the story coherent / plausible by writing 1̈or 2̈in the output. If both sentences are plausible, pick the one that makes most sense.
task105	Story Cloze	In this task, you're given four sentences of a story written in natural language. Your job is to complete the end part of the story by predicting the appropriate last sentence which is coherent with the given sentences.
task065	Timetravel	In this task, you are given a short story consisting of exactly 5 sentences where the second sentence is missing. You are given two options and you need to select the one that best connects the first sentence with the rest of the story. Indicate your answer by 'Option 1' if the first option is correct, otherwise 'Option 2'. The incorrect option will change the subsequent storyline, so that at least one of the three subsequent sentences is no longer consistent with the story.
task588	Amazonfood rating	In this task, you're given a review from Amazon's food products. Your task is to generate a rating for the product on a scale of 1-5 based on the review. The rating means 1: extremely poor, 2: poor, 3: neutral or mixed, 4: good, 5: extremely good.
task020	Mctaco	The answer will be 'yes' if the provided sentence contains an explicit mention that answers the given question. Otherwise, the answer should be 'no'. Instances where the answer is implied from the sentence using ïnstinctör c̈ommon sense(̈as opposed to being written explicitly in the sentence) should be labeled as 'no'.
task241	TweetQA	In this task, you are given a context tweet, a question and the corresponding answer of the given question. Your task is to classify this question-answer pair into two categories: (1) "yes" if the given answer is right for question, and (2) "no" if the given answer is wrong for question.
task1553	CNN Dailymail	In this task, you are given highlights ,i.e., a short summary, in a couple of sentences, of news articles and you need to generate the news article with a maximum length of 2 paragraphs.
task1161	CODA19	In this task, you're given a title from a research paper and your task is to generate a paragraph for the research paper based on the given title. Under 10 lines is a good paragraph length.
task177	Paraphrase	This is a paraphrasing task. In this task, you're given a sentence and your task is to generate another sentence which express same meaning as the input using different words.
task295	Commonsense	In this task, you are given an impractical statement. You are also given three reasons (associated with Ä, B̈, C̈) explaining why this statement doesn't make sense. You must choose the most corresponding reason explaining why this statement doesn't make sense.

Table 1: Overview of the different datasets used for the experiments in this paper.

Dataset	Example Prompt
Abductivenli	Beginning: Mike was in the car on the highway. Middle 1 : `[Annotator Prompt]`. Middle 2: A deer never ran in front of his car. Ending: When he got to the hospital, they saw that it had been broken
Story Cloze	Sentence1: `[Annotator Prompt]` Sentence2: Suddenly, there was an announcement. Sentence3: The school was on a lockdown. Sentence4: The kids sat quietly, and waited.
Timetravel	Sentence 1: Little Charlie and his dad were painting the garage. Sentence 3: His dad turned around and started to laugh Sentence 4: Charlie had paint on him from head to toe Sentence 5: His dad rinsed him off with water from the hose Option 1: `[Annotator Prompt]` Option 2: Charlie had some trouble controlling the brush.
AmazonFood rating	This is `[Annotator Prompt]`
McTaco	Sentence: The legitimization of gambling led to its increased legalization across the US. Question: `[Annotator Prompt]`
TweetQA	Context: Praying for everyone here in Vegas. I witnessed the most unimaginable event tonight. We are okay. Others arent. Please pray. –Jake Owen (@jakeowen) October 2, 2017 Question: `[Annotator Prompt]` Answer: people were not okay
CNN DailyMail	`[Annotator Prompt]`
CODA19	`[Annotator Prompt]`
Paraphrase	`[Annotator Prompt]`
Commonsense	I walk under the park. `[Annotator Prompt]`

Table 2: Example of a prompt to annotate per dataset. `[Annotator Prompt]` indicates where the prompt of that the annotator should come up with, should fit in the text.

Criteria	Dataset	Example	Desired Answer
The response is unrelated to the task or it includes a response for a different topic or question	TweetQA	Context: I lost the role in 50 Shades of Grey so you won't be hearing from me for awhile— Lena Dunham (@lenadunham) September 2, 2013 Question: which countries are next to France? Answer: liverpool and everybody.	no
The response contains (part of) the answer.	Amazonfood	These are Amazon fish fingers, 5 stars from me - extremely good!	5
The response does not follow the required format or task definition.	TweetQA	Context: Kasich's daughter on his dance moves: "You're not going to go on 'Dancing with the Stars'" #KasichFamily CNN Politics (@CNNPolitics) April 12, 2016 Question: no, as he is terrible at dancing Answer: dozen	no
The person misunderstood the task.	Commonsense	He is wearing a green car choose an alphabet rating for this sentence, "A" for unreasonable meaning, otherwise "B"	A

Table 3: Examples for the criteria of an invalid annotation.

Native language	Number of annotators	Languages	Validation rate
Other	36	BG, SL, RU, SW, ML, HU, FA, VI, BE, EL, TN, ID,PL, MR, TR, PT, T, RO, FIL, UR, SQ	0.83
NL	23		0.80
EN	28		0.83
ZH	11		0.82
EN, other	9	PA, JA, SW, UR, VI, MR, EL	0.86
EN, ZH	1		0.88
ES	5		0.77
FR	4		0.94
IT	3		0.94
HI	2		0.93
AR	1		0.94
ES, Other	1	CA	0.84

Table 4: Overview of the native languages of the annotators and the validation rate per native language.

Set ids	Native or not		Western native or not		Total
	Native	Non-native	Western	Not Western	
10	7	16	5	18	23
20	7	12	4	15	19
30	7	10	4	13	17
40	4	8	3	9	12
50	4	9	2	11	13
60	5	14	3	16	19
70	5	11	4	12	16
80	3	10	3	10	13
90	4	10	4	10	14
100	6	5	4	7	11

Table 5: Overview of the number of annotators per group and set.

For the annotations that did not follow the required format, we tried to change it into the correct format without changing the content of the prompt, if possible (i.e. removing *Question:*). If this was not possible, the annotation was rejected.

D Dataset Statistics -Annotations

The native-bias dataset consists of 12,519 annotations from 124 annotators. Our dataset initially contained 1,000 different examples. After deleting the examples that were not validly annotated by at least 50 % of annotators, we retained 988 examples for 10 different tasks.

The annotators have varying native languages as shown in Table 4. The languages are shown in isocode format. Moreover, per native language, we have also included the average validation rate, that is the amount of annotations per person that were valid over the total number of annotated examples.

Table 5 shows an overview of the number of annotators per group and set-id. All annotators were given sets of examples that had to be annotated. Every example has a unique set-id.

Furthermore, the annotators have reported their level of English proficiency and the frequency of which English was spoken. We provide this information for the non-native speakers in Tables 6 and 7.

English proficiency level	Number of non-native annotators
C2	31
C1	41
B2	13
B1	1

Table 6: Overview of the self-reported English proficiency of the non-native annotators.

English usage frequency	Number of non-native annotators
Daily	60
A few times per week	21
Once a week	4
Less than once a week	1

Table 7: Overview of the self-reported frequency of English usage of the non-native annotators.

D.1 Prompt length

Table 8 shows the average prompt length per dataset and per group. It is interesting to note the large difference for the CNN dailymail dataset, where the non-native English speakers have provided on average longer summaries. For the Western native English group versus the not Western native English group, the summaries for the latter are on average 10 words longer than for the former.

D.2 Time analysis annotators

Table 9 shows an overview of the average duration of annotating one example per group in minutes. Table 10, on the other hand, shows the average time for annotating the given set in hours.

Dataset ids	Native or not		Western native or not	
	native	non-native	not Western native	Western native
0	11.08	10.17	11.52	10.14
1	9.15	8.94	8.31	9.26
2	9.40	9.71	9.73	9.58
3	14.95	13.00	14.8	13.39
4	7.56	7.57	7.41	7.61
5	7.53	7.74	6.91	7.93
6	59.32	66.14	56.48	66.41
7	12.09	11.74	12.04	11.77
8	11.28	11.38	11.38	11.34
9	25.91	28.30	24.65	28.66

Table 8: Average prompt length per group and dataset.

group	duration (in min)
native	2.07
non native	3.27
Western native	1.87
not Western native	3.25

Table 9: Average duration of annotating 1 example per group, in minutes

Set-ids	Native or not		Western native or not	
	native	non-native	Western native	not Western native
10	2.70	3.06	2.33	3.12
20	3.22	4.01	3.72	3.71
30	3.05	5.95	3.68	5.09
40	3.31	5.12	2.95	5.04
50	4.82	4.68	2.98	5.04
60	1.30	3.37	1.33	3.10
70	1.39	3.66	1.45	3.46
80	2.39	4.54	2.39	4.54
90	4.63	7.14	4.63	7.14
100	2.79	3.39	2.03	3.65

Table 10: Average overall duration of annotating all annotations in set per group, in hours

E Prompt templates

General
system prompt:*You are a helpful assistant.*
chat history:
User: *task definition*,
Assistant:*'Understood'*,
User: *Full English Prompt*

Add Native
system prompt:*You are a helpful assistant. Respond as if you are interacting with a native English speaker*
chat history:
User: *[task definition]*,
Assistant:*Understood*,
User: *[Full English Prompt]*

Add Non-Native
system prompt:*You are a helpful assistant. Respond as if you are interacting with a non-native English speaker*
chat history:
User: *[task definition]*,
Assistant:*Understood*,
User: *[Full English Prompt]*

Since we found that some of the models were not following the task definitions correctly for some of the tasks, we added extra instructions as to how the model should reply. Table 11 shows the instructions that were added to the task definition for the different datasets.

Task	Task definition
StoryCloze	Only respond with the predicted last sentence.
AmazonFood	Only respond with the rating.
McTaco	Only respond with "yes" or "no".
TweetQA	Only respond with "yes" or "no".
CNN Dailymail	Only respond with the news article.
CODA19	Only respond with the paragraph.
Paraphrase	Only respond with the paraphrased sentence.
Commonsense	Only respond with the letter indicating the most corresponding reason.

Table 11: Overview of the added instructions per dataset to ensure consistent answers from the LLMs.

F Checkpoints models

We used the following checkpoints of the different models:
GPT 3.5 was made by OpenAI[8]. We used *gpt-3.5-turbo-0125*.
GPT 4o was made by OpenAI[9]. We used *gpt-4o-2024-05-13*.
Haiku was made by Anthropic (Anthropic, 2024). We used *claude-3-Haiku-20240307*.
Sonnet was made by Anthropic (Anthropic, 2024). We used *claude-3-Sonnet-20240229*.
Qwen 7B is an open source model made by the Alibaba group (Bai et al., 2023). We used *Qwen/Qwen1.5-7B-Chat*

G Evaluation metrics

All generative performance metrics were gathered using Llama-3.3-70B-Instruct. Below, the prompts for the different metrics were given. These are based on the ones used for the Summeval dataset by Bavaresco et al. (2024). Depending on the generative task, the prompts differ slightly. we use the following structure: Final prompt= ""start_prompt + doc_gen: [model generated reply] + metric"" where the start prompt for fluency is:

'You will be given a doc_gen generated based on a doc_comp. Your task is to rate the doc_gen on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

[8] https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/
[9] https://openai.com/index/hello-gpt-4o/

and the start prompt for the other metrics is: *You will be given a doc_gen and a doc_comp. Your task is to rate the doc_gen on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*

doc_gen and doc_comp depend on the dataset. They are a 'closing sentence' and a 'story' for Story Cloze; a 'news article' and a 'summary' for CNN Dailymail; an 'article' and a 'title' for CODA19; and a 'paraphrased sentence' and a 'sentence' for Paraphrase.

The metric is one of the following:

Fluency:

Evaluation Criteria: Fluency (1-3): the quality of the doc_gen in terms of grammar, spelling, punctuation, word choice, and sentence structure. Assign a score on a scale of 1 to 3 where: - 1: Poor. The doc_gen has many errors that make it hard to understand or sound unnatural.- 2: Fair. The doc_gen has some errors that affect the clarity or smoothness of the text, but the main points are still comprehensible.- 3: Good. The doc_gen has few or no errors and is easy to read and follow. Evaluation Form (scores ONLY):- Fluency:

Coherence.

The definition depends on the dataset. For CNN Dailynews and CODA19 it is as follows:

Evaluation Criteria: Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby the doc_gen should be well-structured and well-organized. The doc_gen should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic. Evaluation Steps: 1. Read the doc_comp carefully and identify the main topic and key points.2. Read the doc_gen and compare it to the doc_comp. Check if the doc_gen covers the main topic and key points of the doc_gen, and if it presents them in a clear and logical order.3. Assign a score for coherence on a scale of 1 to 5, where 1: Very low coherence ; 2: Low coherence; 3: Mediocre coherence ; 4: High coherence ; 5: Very high coherence. Evaluation Form (scores ONLY):- Coherence:

For Paraphrase it is as follows:

Evaluation Criteria: Coherence (1-5) - The overall quality of the paraphrased sentence in terms of logical flow, structure, and alignment with the original sentence. A coherent paraphrase should preserve the meaning of the original sentence, avoid redundancy, and introduce variation without altering the main idea. The paraphrased sentence should not feel disjointed or incomplete but should read smoothly as a standalone sentence. Evaluation Steps: 1. Read the doc_comp carefully and identify the main topic and key points. 2. Read the doc_gen and compare it to the doc_comp. 3. Assign a score for coherence on a scale of 1 to 5, where 1: Very low coherence ; 2: Low coherence; 3: Mediocre coherence ; 4: High coherence ; 5: Very high coherence. Evaluation Form (scores ONLY): - Coherence:

For Story Cloze it is as follows:

Evaluation Criteria: Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby the sentences should be well-structured and well-organized. The sentences should not just be a heap of related information, but should build from sentence to a coherent story.Evaluation Steps: 1. Read the doc_comp carefully and identify the main topic and key points. 2. Read the doc_gen and compare it to the doc_comp. Check if the sentences are clear and in a logical order. 3. Assign a score for coherence on a scale of 1 to 5, where 1: Very low coherence ; 2: Low coherence; 3: Mediocre coherence ; 4: High coherence ; 5: Very high coherence. Evaluation Form (scores ONLY): - Coherence:

Relevance.

The definition depends on the dataset. For Story Cloze it is as follows:

Evaluation Criteria: Relevance (1-5) - The degree to which the generated doc_gen effectively reflects the main themes and purpose of the doc_comp. A relevant closing sentence should provide a meaningful and appropriate conclusion, aligning with the tone and key points of the narrative. Evaluation Steps: 1. Read the doc_comp and the doc_gen carefully. 2. Compare the doc_gen to the doc_comp and identify the main points of the doc_comp. 3. Assess how well the doc_gen concludes the doc_comp, and how much irrelevant or redundant information it contains. 4. Assign a relevance score from 1 to 5 where 1: Very low relevance ; 2: Low relevance; 3: Mediocre relevance ; 4: High relevance ; 5: Very high relevance. Evaluation Form (scores ONLY): - Relevance:

For all other datasets it is as follows:

Evaluation Criteria: Relevance (1-5) - inclusion of important content from the doc_comp. The doc_gen should include all important information

from the doc_comp. Evaluation Steps: 1. Read the doc_comp and the doc_gen carefully. 2. Compare the doc_gen to the doc_comp and identify the main points of the doc_comp. 3. Assess how well the doc_gen covers the main points of the doc_comp, and how much irrelevant or redundant information it contains. 4. Assign a relevance score from 1 to 5 where 1: Very low relevance ; 2: Low relevance; 3: Mediocre relevance ; 4: High relevance ; 5: Very high relevance. Evaluation Form (scores ONLY): - Relevance:

H Distribution Amazon food reviews

Figure 10 shows an overview of the wrong predictions of the AmazonFood review dataset for the different groups and models for one of the three runs. This shows the distribution between what was predicted and what should be predicted. We only consider here the cases where the model predicted one of the given ratings, and excluded cases where no prediction was given. As shown, for both the native and Western native group, we find a large amount of misclassification for the highest rating. Additionally, neutral is not often predicted for these classes compared to the other groups.

I Results Sonnet different languages

When adding that the model is interacting with a non-native English speaker, we find that Sonnet starts to answer in different languages. We find that for 668 prompts the model answers in Spanish, for 25 sentences in French, and for 5 sentences in Indonesian. There were a couple of other languages that also occurred sporadically. An overview is shown in Table 12. However, these answers were not related to the native language of the prompt writer. This phenomenon was encountered mainly for the Timetravel dataset. Interestingly, this effect was not seen for the other models, not even for Haiku.

J Example Paraphrase

As said, there are differences between native and non-native speakers as to how they perceived the paraphrasing task. For example given this desired output: *At this time of rapid change, those who lag behind fall into irrelevance*. Native speakers came up with very freely paraphrased sentences, such as: *If you are not adapting to the quick changes of the world, you will not succeed.* while non-native speakers stuck to *In this fast changing ages, whoever is lagging becomes irrelevant*. When giving these different sentences to the model to paraphrase, the result for the more freely paraphrased sentences might cause the model to shift away further from the initial sentence or gold answer.

Language	Times Occurring
es	668
fr	25
id	5
it	2
lt	1
sw	1
ru	1

Table 12: Occurrences of different languages in Sonnet

K Classification results

Figure 11 shows the accuracy scores for the objective and subjective classification tasks per model when information about the nativeness of the prompt writer is added. We see how sonnet clearly performs differently than the other models.

L LLM as a judge: Generative results

Figure 12 shows how similar behavior is found across all three performance metrics per model. Moreover, Figure 13 shows the results per dataset for the generative results.

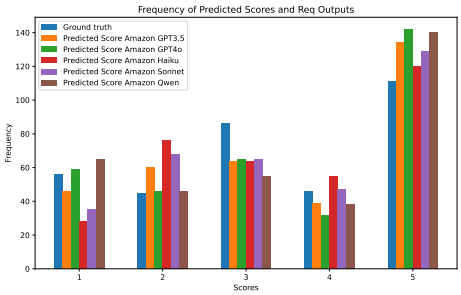
M Additional Analysis

In this section, we include some extra analyses on the performance of the different groups within the non-native English speakers. More specifically, we add the results per level of English proficiency, as well as per frequency of English. We see that there are differences in performance across the different groups.

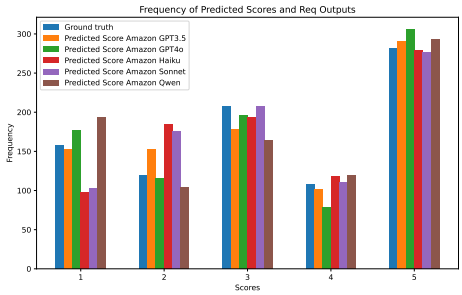
M.1 Classification results

For the classification results, we see a clear connection between performance and level of English, and frequency of usage of English. The groups with the highest levels of English also obtain better results. This is shown in Figures 14 and 15.

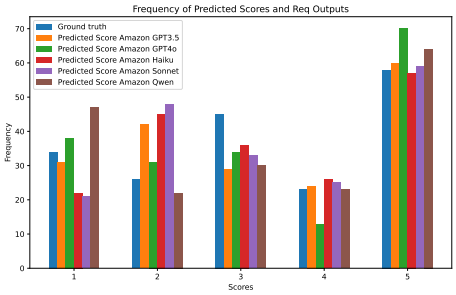
As we saw a performance difference, in terms of levels of English, we also compare the results when only taking into account level C1 and C2 non-native English speakers. The results are shown



(a) Overview of the predictions for the Western native English speakers.



(b) Overview of the predictions for the non-native English speakers.



(c) Overview of the predictions for the native English speakers that are not western native.

Figure 10: Overall classifications for Western native, native that are not Western native, and non-native English speakers

in Figure 16. Here, we still see the same order in performance as in Figure 3 was shown. However, now there is a clearer performance difference between the natives that are not western native and the non-native group.

M.2 Generative Results

For the generative tasks, however, we do not see clear differences in terms of frequency of English usage and performance, as shown in 17 and 18. Only the people with the lowest level of English proficiency perform better in terms of coherence, which is unexpected.

When analyzing the performance differences only for the groups with highest proficiency (C2 and C1), as shown in Figure 19, we see similar findings to Figure 4.

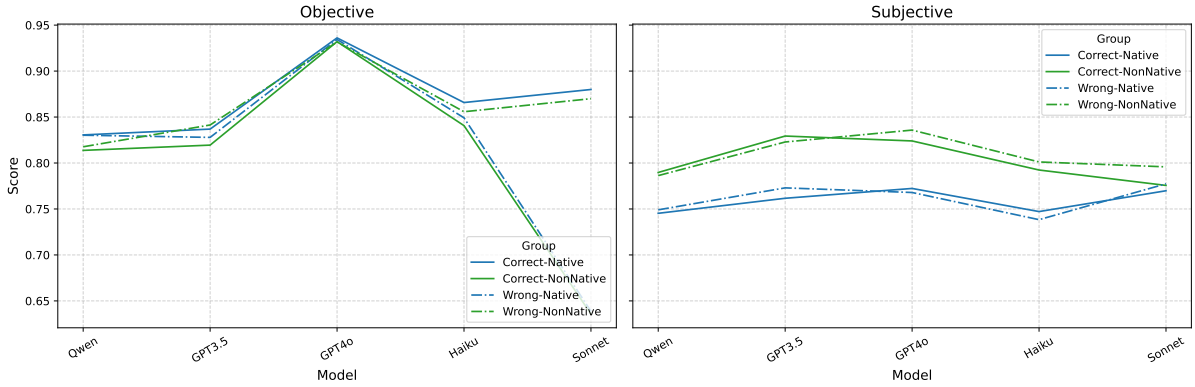


Figure 11: Classification results per model and classification task when information about the nativeness of the prompt writer was added. We clearly see how Sonnet is highly influenced by this additional information.

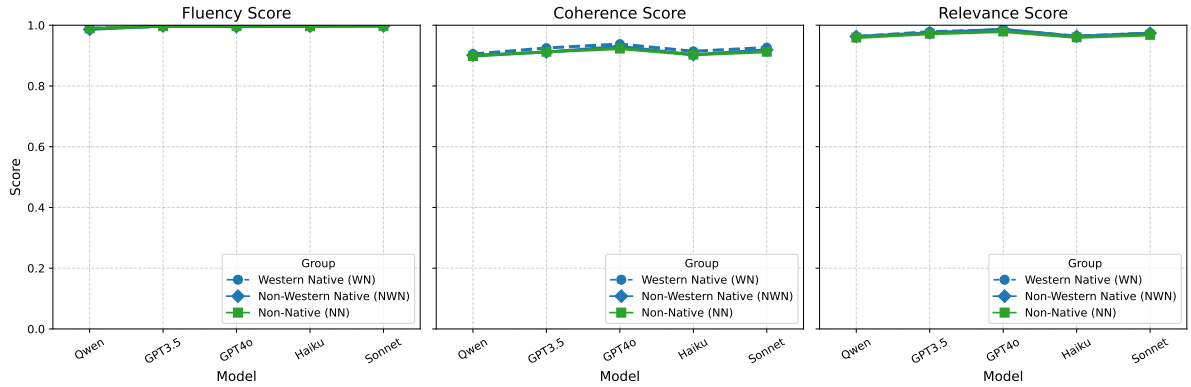


Figure 12: This Figure shows the performance of (western) native speakers and non-native speakers. We see how the highest performance for Coherence and is obtained for the western native group across all different models. The relevance scores show slightly less difference between groups, but the non-native and not western native group performs worse overall. The fluency scores are similar for all groups. We rescaled the results so that they range from 0 to 1.

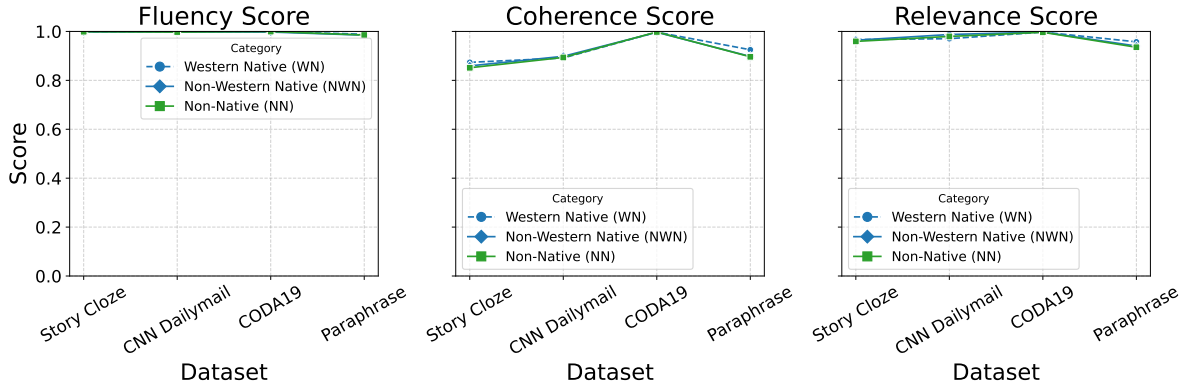


Figure 13: This figure shows the overall performance across the three groups: (western) native speakers and non-native speakers. However, when looking into the coherence metric, we do see a preference for the western native group. The results show how there is no difference regarding fluency and only a slight performance difference when comparing the native categories with the non-native category for relevance.

Figure 14: This figure shows the performance of English non-native speakers per self-reported level of English for the classification tasks.We adjusted the y-axis to range from 0.65 to 1 for clarity.

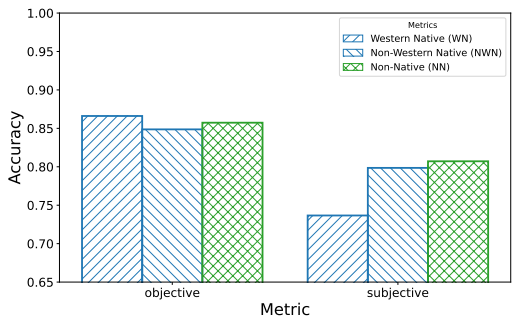


Figure 16: This figure shows the performance of the three groups only including C2 and C1 level English speakers. We adjusted the y-axis to range from 0.65 to 1 for clarity.

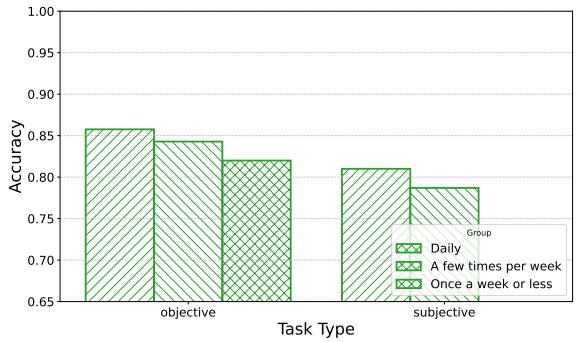


Figure 15: This figure shows the performance of English non-native speakers per self-reported frequency of English usage for the classification tasks. We adjusted the y-axis to range from 0.65 to 1 for clarity.

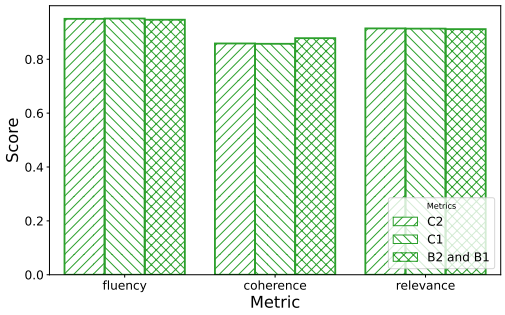


Figure 17: This figure shows the performance of English non-native speakers per self-reported level of English for the generative tasks. We rescaled the results so that they range from 0 to 1.

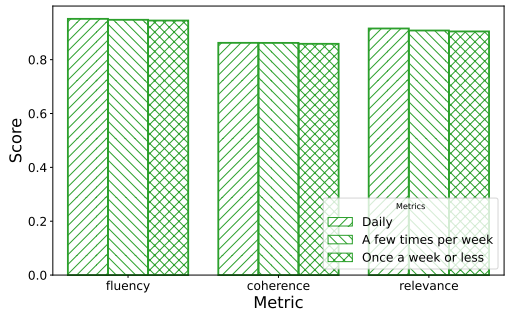


Figure 18: This figure shows the performance of English non-native speakers per self-reported frequency of English usage for the generative tasks. We rescaled the results so that they range from 0 to 1.

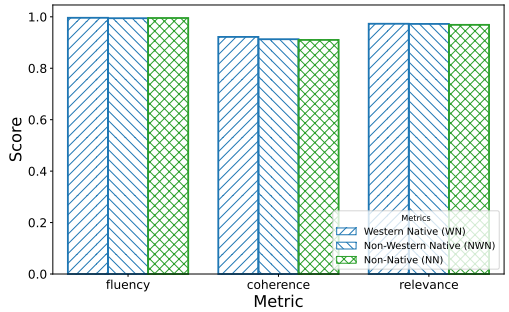


Figure 19: This figure shows the generative results only for the C2 and C1-level speakers per group.We rescaled the results so that they range from 0 to 1.