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. This study investigates whether the quality of LLM responses varies depending on the demographic profile of users. Considering English as the global lingua franca, along with the diversity of its dialects among speakers of different native languages, we explore whether non-native English speakers receive lower-quality or even factually incorrect responses from LLMs more frequently. Our results show that performance discrepancies occur when LLMs are prompted by native versus non-native English speakers and persist when comparing native speakers from Western countries with others. Additionally, we find a strong anchoring effect when the model recognizes or is made aware of the user's nativeness, which further degrades the response quality when interacting with non-native speakers. 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 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 certain design choices result

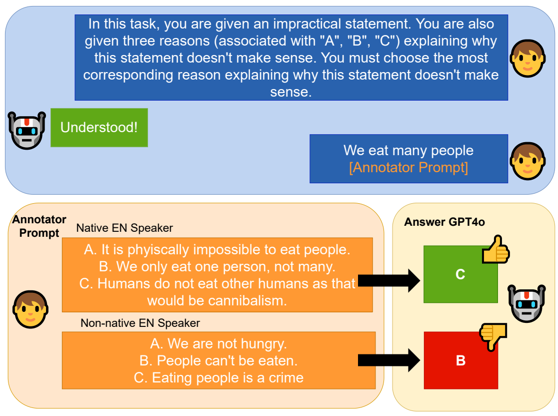


Figure 1: An example prompt of a native and non-native English speaker and the corresponding output given by GPT4o. The desired output is C. The model selects an incorrect answer choice for the non-native English speaker, although semantically the same message was conveyed.

in improved downstream performance for specific sub-populations (Santy et al., 2023). Consequently, their effectiveness considerably decreases when prompted in other languages or even 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. However, unlike those studies focusing only on English dialects from English-speaking countries, our research also incorporates participants from countries where English is not an official language.

In this paper, we find performance differences when LLMs are prompted by native versus non-native English speakers. More specifically, models often generate inaccurate responses for non-native speakers and rate the native prompts more positively than intended. We collect a dataset comprising over 12,000 unique prompts from native and non-native English speakers around the world, and demonstrate how different prompt formulations can lead to worse performance despite conveying the same message. An example of a prompt in our dataset is shown in Figure 1. Moreover, we find that these performance differences increase when comparing native English speakers from Western countries (US, UK, Canada) with other native and non-native English speakers. Furthermore, when the model recognizes or is informed about the user's nativeness, a strong anchoring effect occurs, where the added information substantially affects model performance, leading to increased bias towards native English speakers.

Our contributions are as follows:

- We quantitatively and qualitatively analyze the performance of LLMs on objective and subjective classification tasks, as well as generative tasks, when prompted by native and non-native English speakers
- We investigate the impact on the LLM performance when the nativeness of the user is explicitly stated or inferred by the model.
- We publish our multilingual instruction-tuning dataset[1] containing over 12,000 unique prompts from a diverse group of native and non-native English speakers worldwide, including translations of the prompts into eight different native 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. 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 with and without assigning a persona to the prompt writer.

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 hypothesize that these design choices affect LLM performance when interacting with native versus non-native English speakers. Furthermore, we hypothesize that the specific English dialect learned may influence the responses from LLMs,

[1] https://anonymous.4open.science/r/native_en_bias-EDC5

particularly if the LLMs are optimized for English variants from the US, UK, and other predominantly English-speaking countries, as shown by i.e. Santy et al. (2023). We also anticipate that performance differences will increase when a model is explicitly informed about the user's nativeness. Considering the LLM's ability to infer demographic annotator information from text (Staab et al., 2023), we also expect the LLM to infer a user's nativeness and that this correct inference leads to greater performance differences.

To test these hypotheses, 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

We selected ten diverse datasets from various natural language instruction tasks[2] (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, and multiple-choice answering. From each original dataset, we randomly selected 100 examples, ensuring they were correctly annotated and free of offensive language. Additionally, we included one extra example per dataset to serve 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 the annotators to have a minimum level of English equivalent to a high school or university degree to ensure English proficiency. Each annotator worked on 20 to 240 examples, and 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 rate of at least 12.11 euros per hour.

In addition to gathering self-reported linguistic information, i.e. the respective native language, level of English, and frequency of English use, we also collected information from native English speakers about how they learned English. To investigate our hypotheses, we categorize English speakers based on their nativeness and based on their learning contexts. The term *Western native* refers to native English speakers who learned English from native speakers from countries like the UK, US, Australia, or Canada. The term *not Western native* refers to all other annotators in our dataset.

Annotators performed different tasks depending on the assigned datasets. An example annotation is shown in Figure 2, 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. More details about the annotation setup can be found in Appendix C.

Before including the annotations in our final dataset, they were validated. An annotation was deemed invalid if it met any of the following criteria: 1) 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 for each validation criterion are included in Appendix B.

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[3].

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[4], GPT4o[5], Haiku (Anthropic, 2024), Sonnet (Anthropic, 2024),using the appropriate APIs, and Qwen1.5 7B[6] (Bai et al., 2023) in line with the provided

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

[3] 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.

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

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

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

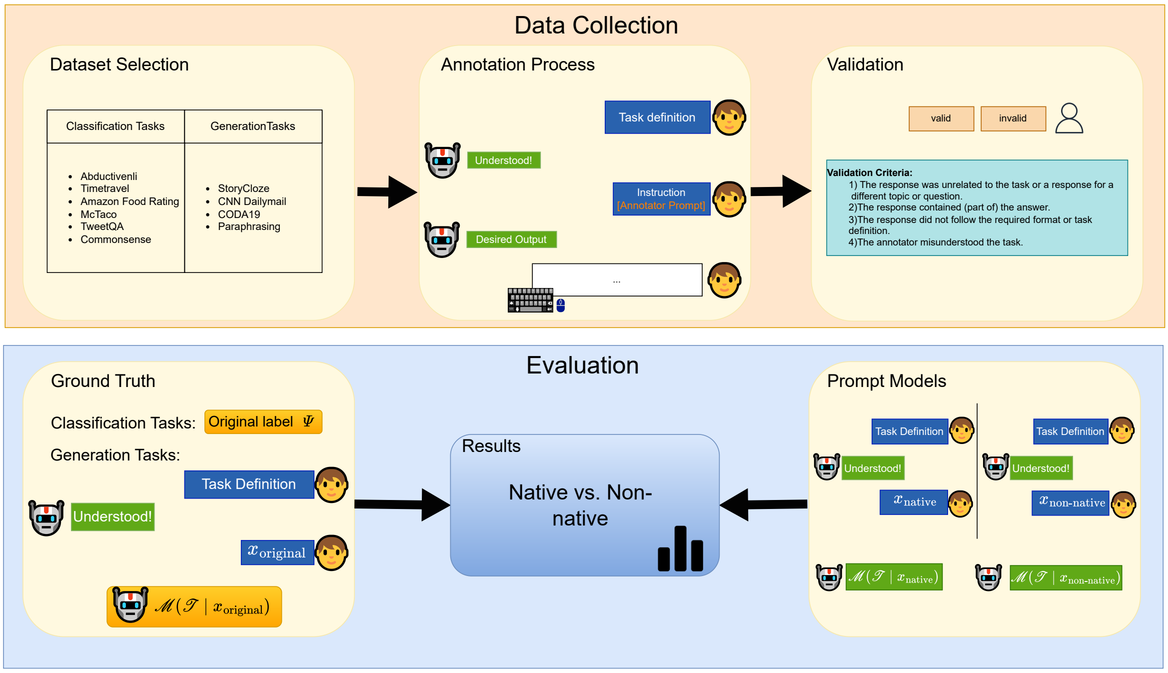


Figure 2: Methodology and experimental setup. The upper part of the figure shows the data collection steps. After gathering the different datasets, study participants annotated the examples. Then we validated them and used them as input to generate LLM responses. The lower part of the figure shows the evaluation phase. Before gathering the results, we got a new ground truth value for the dataset based on the original example for the generation tasks.

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.

We experimented with various prompt schemata, structuring our methodological setup as follows. Firstly, we analyzed the results on our dataset without any modifications, comparing responses for native and non-native English speakers. Additionally, we distinguished between Western native English speaking and not Western native English speaking. Secondly, we hypothesize that the bias becomes more pronounced when adding explicit information on whether the annotator is a native or non-native English speaker, thereby analyzing whether an anchoring effect occurs. Anchoring is a term 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. Finally, we analyze the performance difference when letting the model first guess the nativeness of the annotator.

4.2 Evaluation

To measure the bias within the models, we look into the performance difference between two groups, i.e. native versus non-native and Western native versus not Western native. 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:

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

$$\begin{aligned}\Delta_{\text{non-native}} = &\phi(\mathcal{M}\left(\mathcal{T} \mid x_{\text{non-native}}\right), \\ &\mathcal{M}\left(\mathcal{T} \mid x_{\text{original}}\right))\end{aligned}$$

with native bias generative = $\Delta_{\text{native}} - \Delta_{\text{non-native}}$, template $\mathcal{T}$, user prompt x, model $\mathcal{M}$, performance metric ϕ, and generated ground truth from the original prompt $\mathcal{M}\left(\mathcal{T} \mid x_{\text{original}}\right)$. The Western native bias can be similarly inferred.

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: BLEU, ROUGE, BERT-score, and BART-score. These metrics are explained in detail in Appendix G.

The performance metrics for generative tasks are calculated using the LLM-generated gold answer as reference. This gold answer was generated using the original prompt in the dataset as displayed in Figure 2. We used this instead of the original gold answer in the dataset, given that optimal answers are model and context-dependent. By employing the LLM-generated gold answers, we align responses for each model. We include the results, using the original gold output in Appendix.

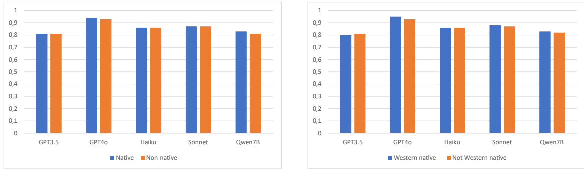
5 Results

In line with our hypotheses, the results indicate performance differences between native and non-native English speakers and Western native and not Western native speakers for GPT4o, Sonnet, and Qwen for the classification tasks, as shown in Figures 3 and 4. However, the generative tasks indicate no or even opposite preferences as displayed in Figure 5. Furthermore, informing the models about the user's nativeness further reduces answer quality for non-native English speakers, also resulting in a strong anchoring effect.

5.1 Classification tasks

Do LLMs perform better for certain groups on objective classification tasks? Figure 3 shows the overall average performance for the different models and groups on the objective classification tasks. These contain all classification tasks, except the Amazon Food review task.

Figure 3a shows that **only GPT4o and Qwen 7B perform better for the native group of annotators**. All other models provide on-par results for both groups. Interestingly, Qwen, which is predominantly trained on Chinese text, also prefers native English speakers over non-native English speakers.



(a) Native vs. non-native

(b) Western native vs. not

Figure 3: Overall average accuracy of the objective classification tasks per model and group.

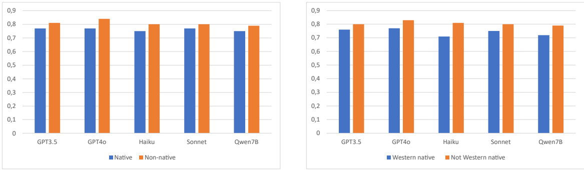
Given that GPT4o is the best-performing model from our set, the results indicate that its superior performance may come at the cost of being more specifically tailored to native English speakers.

Figure 3b indicates that **beside GPT4o and Qwen 7B, also Sonnet shows a higher performance for the Western native English speakers**. Interestingly, the performance difference between the two groups for Qwen has decreased, compared to Figure 3a, indicating that the model is less tailored towards Western native English.

Similarly, GPT 3.5 performs best for the not Western English speakers. The results suggest that GPT4o, a larger and generally more effective model than GPT3.5, is more sensitive to prompts from native and non-native speakers, leading to the observed performance disparity. Additionally, this behavior is not exclusive to the two GPT models. A comparison between Haiku and Sonnet, both part of the Claude3 family, reveals that Sonnet also exhibits greater sensitivity to prompts from Western native English speakers.

Can we make similar conclusions for the subjective classification task?
Surprisingly, Figure 4 shows the opposite effect in comparison to the objective classification tasks, preferring both the non-native and not Western native English speakers. This finding is remarkable, given the results in the subjective classification literature (Santy et al., 2023; Durmus et al., 2023).

To understand this contrasting phenomenon, we first analyzed the results for both the native and non-native English speakers, and the Western native and not Western native English speakers, in depth. **We find that for the native English-speaking group, the models often predict the rating more positively than actually intended, and this effect is even more pronounced for the Western native English speakers.** While for the non-native English group and not Western native English speaking group, GPT4o predicted 50% of all wrongly predicted annotations to be more positive than in-



(a) Native vs. non-native (b) Western native vs. not

Figure 4: Overall average accuracy of the subjective classification task per model and group.

model	correct information		wrong information	
	native	non-native	native	non-native
GPT3.5	**0.84**	0.82	**0.84**	**0.84**
GPT4o	**0.94**	0.93	**0.94**	0.93
Haiku	**0.86**	0.84	0.85	**0.86**
Sonnet	**0.86**	0.66	0.66	**0.86**
qwen 7B	**0.83**	0.82	**0.83**	**0.83**

Table 1: Overall average accuracy of all classification tasks per model when adding correct/wrong information about whether or not the prompt was written by a native English speaker.

tended, this was 60% for the native English speaking group and even 70% for the Western native English speaking group for GPT4o indicating cultural differences. More details about these different distributions for the different models can be found in Appendix H.

Although we find that GPT4o seems to be most tailored towards the native English speakers, it tends to estimate native or Western native responses more positive than non-native responses. This implies cultural differences in this subjective classification, indicating that these models do not always align with native or Western native opinions as previously shown in the literature.

What is the effect on the performance when informing the model about the nativeness of the annotator? For these experiments, we added information about the user's nativeness to the system prompt. More specifically, we asked the model to respond as if it is interacting with a native or non-native English speaker respectively.

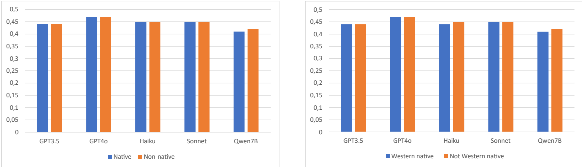
Table 1 shows that when adding information about the nativeness of the annotator to the system prompt when generating the predictions, the bias towards the native English-speaking group becomes more pronounced. **All models show better performance for the native English speaking group in comparison to the non-native English speakers**. The biggest performance difference is seen for Sonnet. When looking into these results, we find that Sonnet started answering several questions languages other than English, such as Spanish, French or Indonesian. This is remarkable and demonstrates a clear anchoring effect, considering we only instructed the model to respond as if interacting with a non-native English speaker without specifying another language.

When we inaccurately labeled native speakers as non-native and vice versa, we observed that only the Claude models adhered to this incorrect information, showing a preference toward the non-native group. GPT4o still performs better on the native prompts, and both Qwen and GPT3.5 now show equal performance, compared to the slight preference towards the native group in our previous experiment. This indicates that GPT4o is most robust against the additional information on the user's nativeness, concentrating primarily on the prompt itself. It thus seems to be better hedged against direct bias being added. The same applies to GPT3.5 and Qwen, however to a lesser extent given that adding information that mislabels the non-native English speakers as native English speakers does enhance performance.

model	guess native	guess non-native	guess native correctly	guess non-native correctly
GPT3.5	**0.78**	0.77	**0.76**	0.73
GPT4o	**0.83**	**0.83**	**0.83**	0.81
Haiku	**0.65**	0.64	0.53	**0.64**
Sonnet	**0.65**	**0.65**	0.53	**0.56**
qwen 7B	0.69	**0.70**	0.61	**0.76**

Table 2: Overall average accuracy of all classification tasks per model and group when guessed by the model whether the person is a native or non-native speaker, and when it is guessed correctly.

What is the result of first asking the model to guess about the nativeness of the prompt writer? In these experiments, the model first guesses the user's nativeness from their annotation, and then generates a task response within the same chat. Table 2 shows that for Haiku and GPT3.5, first explicitly letting the model guess about the nativeness of the prompt writer, results in a better score for the prompts that were indicated to be native than the ones that were seen as non-native. GPT4o and Sonnet seem on par for both predicted groups. However, when analyzing the performance of the prompts that were guessed correctly, we do see a large performance difference for both groups for GPT3.5 and GPT4o. Haiku, Sonnet and Qwen, on the other hand, show a preference for non-native English speakers. The performance difference for Haiku between the two correctly guessed groups is



(a) Native vs. non-native (b) Western native vs. not

Figure 5: Overall average BLEU score of the generation tasks per model and group.

model	correct information		wrong information	
	native	non-native	native	non-native
GPT3.5	**0.45**	0.39	0.39	**0.42**
GPT4o	**0.46**	**0.46**	0.45	**0.47**
Haiku	**0.45**	**0.45**	0.44	**0.45**
Sonnet	**0.45**	**0.45**	**0.45**	**0.45**
qwen 7B	0.41	**0.42**	0.41	**0.42**

Table 3: Overall average BLEU score of generative tasks per model when adding correct/wrong information about whether or not the prompt was written by a native English speaker.

remarkable. It seems that the model has predicted many well-performing non-native annotations to be actually native. Thus, despite this opposite performance difference, it does show the underlying bias within Haiku to associate better-performing prompts with native speakers. Interestingly, Qwen performs best for the non-native group. For Sonnet, the overall accuracy has dropped significantly in this experimental set-up. Moreover, both Qwen and Sonnet have correctly guessed a very low amount of annotations. This is mainly due to the models' indecisiveness.

5.2 Generative tasks

Do we find similar performance differences for the generative tasks?
On average, we do not see a clear performance difference for all generative tasks between the native English speakers and the non-native English speakers. Figure 5 shows a slightly higher performance for Qwen for non-native English speakers and both Qwen and Haiku perform better for Western native English speakers. All figures display only the BLEU score, but similar trends are shown by the other metrics. All performance metrics are included in Appendix L.

When analyzing these results per dataset, we find that the best-performing group depends on the specific task at hand. In general, the generative tasks can be divided into two groups: group one with CODA19, which includes a medical research article generation task, and the paraphrasing dataset, group two with StoryCloze and CNN Dailymail. We find that all models prefer the non-native English speakers or not Western native English speakers for group one. Most models prefer (Western) native English speakers for group two. Detailed results are included in Table 17 in Appendix.

In interpreting these results, it is important to consider the specific nature of the generation tasks involved. CODA19 comprises medical articles that utilize specialized medical terminology. Given that most annotators were unfamiliar with this vocabulary, native English speakers did not have a specific advantage over non-native speakers. Additionally, research articles are commonly written in English by authors from various backgrounds. Therefore, it might be that this specific task is robust against the native and non-native preference. Finally, for the paraphrasing task, we found that native and non-native English annotators handled this task differently. While native English speakers often paraphrased the sentences more freely, non-native English speakers closely followed the original structure, mainly providing synonyms and maintaining the original sentence's format. An example is added in Appendix J.

How does informing the model about the user's nativeness impact performance?
Interestingly, only for GPT3.5 adding correct information about the user's nativeness results in a better performance for the native group versus the non-native group, as shown in Table 3. Qwen even displays the opposite effect, resulting in a lower performance for the native group. The other models are equally preferring both groups. However, when looking into the specific datasets, we again find that the result differs depending on the task. For both GPT3.5 and GPT4o, a better performance is found for native English speakers for StoryCloze and CNN Dailymail, while the other datasets are preferred for non-native speakers. Qwen and Haiku on the other hand, perform better for the non-native English speakers on all datasets, even when given the information that the prompt writer is a non-native speaker. Sonnet performs best for the native group for StoryCloze, all other datasets prefer the non-native group or are on par.

Additionally, when incorrect information about the user's nativeness is introduced, we find that all models show improved performance for the non-native group, which is presented to the model as native speakers. This is a remarkable finding, as it emphasizes the importance of the prompt over

model	guess native	guess non-native	guess native correctly	guess non-native correctly
GPT3.5	**0.42**	**0.42**	**0.45**	0.42
GPT4o	**0.46**	**0.46**	**0.47**	0.46
Haiku	0.44	**0.45**	0.34	**0.43**
Sonnet	0.44	**0.45**	0.39	**0.46**
Qwen 7B	**0.39**	**0.39**	**0.43**	0.33

Table 4: Overall average BLEU score for all generative tasks per model and group when guessed by the model whether the person is a native or non-native speaker, and when it is guessed correctly.

the user's nativeness, unlike in classification tasks. GPT3.5 is particularly sensitive to the inclusion of incorrect information regarding the user's nativeness, with metrics across all datasets increasingly favoring the non-native English speakers and diminishing the performance of the native group when compared to the addition of correct information. GPT4o, still prefers the native group for StoryCloze and CNN Dailymail. The model thus shows similar robustness as for the classification tasks. In this experiment, Haiku and Qwen perform even better for the non-native group. Sonnet, however, is a bit more robust than Haiku.

What is the result of first asking the model to guess about the nativeness of the prompt writer? Table 4 shows the average BLEU scores when the model (correctly) guessed whether or not the writer of the prompt was a native English speaker. Overall, there is no distinct performance variation based on the model's general guesses. However, when looking at the guesses that were made correctly, differences emerge. All models now perform better for the native group on average, except Haiku and Sonnet. For the individual datasets, we find that CNN Dailymail performs better for the native group, but again CODA19 and Paraphrasing perform better for the non-native group. Similar to the classification tasks, GPT3.5 is more influenced by this additional information than GPT4o.

6 Discussion

In our experiments, we define native bias, or western native bias, as the model's performance disparity when prompted by native versus non-native English speakers, or Western versus not Western natives respectively. **We find that there are performance differences when the model is prompted by people from different backgrounds.**

In fact, we find that for the objective classification tasks, the larger models perform better for the native or Western native groups compared to the non-native and not Western native groups respectively. For the subjective classification task, however, we find an opposite effect, where the model interprets prompts of (Western) native English speakers to be more positive than intended. The generative tasks show no clear performance differences in general, but the bias varies per dataset.

Moreover, **a strong anchoring effect occurs when the model knows whether or not the prompt writer is a native English speaker, either by correctly guessing, or by inserted information in the system prompt**. The bias is so deeply engraved that informing the models of the wrong native and non-native groups, results in a preference towards the group that was indicated as native. The model is largely led by this added information, more 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 seems to be less present for the generative tasks than for the classification tasks.

Interestingly, the results also suggest that **the more advanced versions of models, such as GPT4o compared to GPT3.5 and Sonnet compared to Haiku, tend to be more sensitive to the subtleties of native English**. On the other hand, we also notice that they are less sensitive to added explicit information of the nativeness or non-nativeness of the annotator, and thus seem to be better hedged against explicitly added bias.

7 Conclusion

In this work, we analyze bias in LLMs towards native English speakers. More specifically, we analyze whether models perform better for native compared to non-native English speakers. We also analyze 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 become more inaccurate for the latter and this effect is even more pronounced for the Western native English versus not Western native English comparison. Furthermore, we find a strong anchoring effect when information about the user's nativeness is added or inferred by the model. 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. Finally, the main limitation of using LLMs and especially the closed-source variant thereof, is the lack of reproducibility of the results.

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.

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*.

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.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

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*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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 5 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.

B Annotation validation

Examples for each of the criteria of an invalid annotation are shown in Table 6.

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.

C 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, 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 6 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 7. We have anonymized all annotations by only providing the self-reported linguistic information in the dataset along with the user id number.

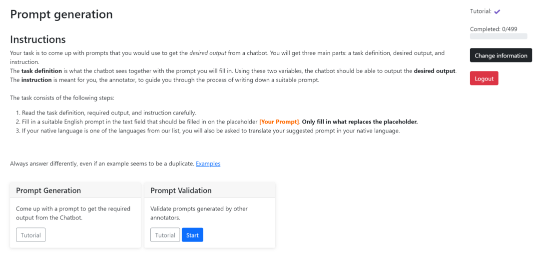


Figure 6: Screenshot of the landing page of the annotation platform.

Figure 7: An annotation example of the Abductivenli dataset.

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 7. 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 8 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.

D.1 Prompt length

Table 9 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

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 1̈ör 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) ÿesïf the given answer is right for question, and (2) n̈oïf 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 5: Overview of the different datasets used for the experiments in this paper.

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 6: 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, ET, 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 7: 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 8: Overview of the number of annotators per group and set.

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 9: Average prompt length per group and dataset.

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 10 shows an overview of the average duration of annotating one example per group in minutes. Table 11, on the other hand, shows the average time for annotating the given set in hours.

E Prompt templates

General

system prompt:*You are a helpful assistant.*

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

Table 10: 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 11: Average overall duration of annotating all annotations in set per group, in hours

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]*

Guess Native
system prompt: *You are a helpful assistant.*
chat history1:
User: *Guess whether the writer of the following prompt is a native or non-native English speaker: [ANNOTATOR PROMPT]*,

chat history2:

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

User: *Guess whether the writer of the following prompt is a native or non-native English speaker: [ANNOTATOR PROMPT]*,
Assistant:*[RESPONSE]*,
User: *Next, execute the following task taking this information into account. [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 12 shows the instructions that were added to the task definition for the different datasets.

F Checkpoints models

We used the following checkpoints of the different models:

GPT 3.5 was made by OpenAI[7]. We used *gpt-3.5-turbo-0125*.

GPT 4o was made by OpenAI[8]. 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*

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

G Evaluation metrics

BLEU.[9] This established performance measure relies on counting the n-gram overlap between candidate and reference sentence(Papineni et al., 2002). **ROUGE.**[10] This metric is commonly used for summarization tasks and calculates the n-gram recall between a reference and candidate summary. We employ ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004). **BERT-score.**[11] This evaluation metric calculates and aggregates a similarity score between the BERT input embeddings for both the candidate and reference sentences. We employed *microsoft/deberta-xlarge-mnli* which was mentioned to be the best-performing model at the moment of writing the paper (Zhang et al., 2019). **BART-score.**[12] This evaluation measure approaches the evaluation as a text-generation task, including all parameters learned during the pre-training phase (Yuan et al., 2021). All metrics were implemented using the default metrics of the used packages.

H Distribution Amazon food reviews

Figures 8 and 9 show an overview of the wrong predictions of the AmazonFood review dataset for the different groups and models. 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 13. 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.

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

Table 13: Occurrences of different languages in Sonnet

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.

K Prompt guess native

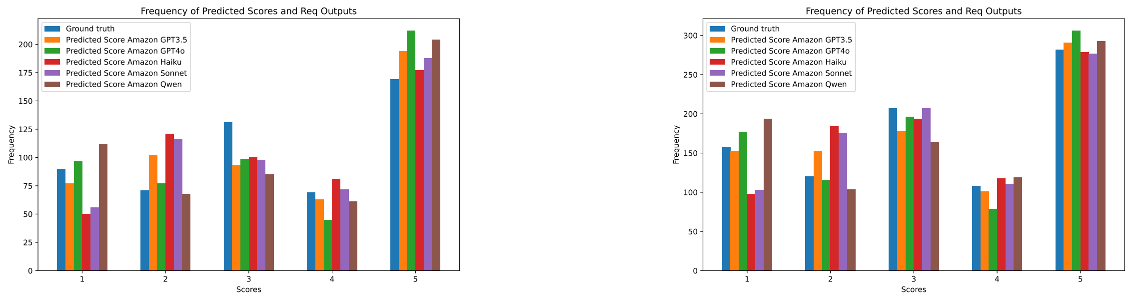
For this prompt, the number of identified correct predictions by the models was gathered automatically. Table 14 shows an overview of the correctly identified native and non native prompts for the entire dataset. In total there are 4264 native prompts and 8255 non-native prompts in our dataset. Many of the incorrectly predicted prompts, where in fact indecisive guesses or when the model was in doubt.

L Generative results

Tables 15 and 16 show the full performance metrics for the different experiments throughout the paper.

Table 17 shows the overall average BLEU scores per dataset and model. All models perform better or equally well for the non-native group for the

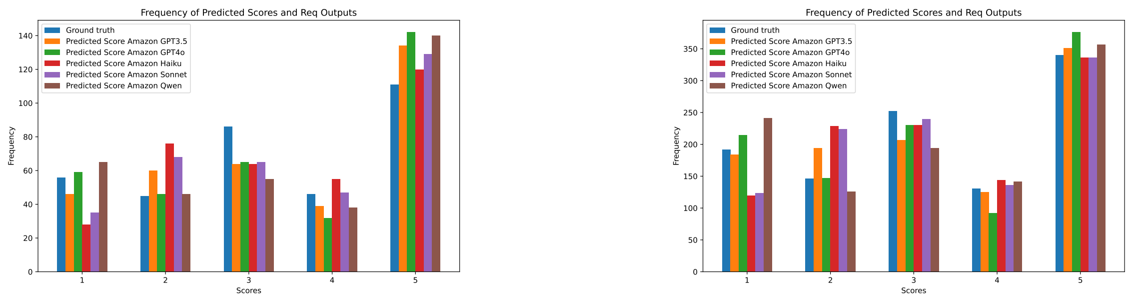
[9] https://www.nltk.org/api/nltk.translate.bleu_score.html
[10] https://huggingface.co/spaces/evaluate-metric/rouge
[11] https://huggingface.co/spaces/evaluate-metric/bertscore
[12] github.com/neulab/BARTScore



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

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

Figure 8: Overall misclassification for native and non-native English speakers



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

(b) Overview of the wrong predictions for the not Western native English speakers.

Figure 9: Overall misclassification for Western native English speakers and not Western native English speakers

model	amount guess native correctly	amount guess non-native correctly
GPT3.5	350	3080
GPT4o	114	436
Haiku	164	914
Sonnet	87	304
qwen 7B	95	152

Table 14: Amount of times the user's nativeness was correctly guessed by the different models.

CODA19, which includes a medical research article generation task, and Paraphrase dataset. Interestingly, for the other tasks, only Qwen and GPT3.5 show better performance for the native group on the Story Cloze dataset and GPT3.5 also on the CNN article generation. For the Western native and not Western native groups, all models perform better on the CODA19 and Paraphrasing datasets for the not Western group, and all models prefer the Western group for the Story Cloze dataset, except Haiku. For the CNN dailymail dataset, we find some conflicting performance differences, with GPT4o performing better for the Western group, while Qwen prefers the non-Western group.

Tables 18, 19, 20, 21, and 22 show per model all different performance measures for the native groups.

M Overall results per prompt

Tables 23, 24, 25, 26, and 27 show an overview of the overall average results per group, model, and prompt.

N Generative results original label

Tables 28, 29, 30, 31, and 32 show the results for the overall average results for the generative tasks per model on the original ground truth label in the dataset.

group	GPT3.5 native	GPT3.5 non-native	GPT4o native	GPT4o non-native	Haiku native	Haiku non-native	Sonnet native	Sonnet non-native	Qwen7B native	Qwen7B non-native
BLEU	**0.44**	**0.44**	**0.47**	**0.47**	**0.45**	**0.45**	**0.45**	**0.45**	0.41	**0.42**
ROUGE-1	**0.39**	**0.39**	0.40	**0.41**	0.40	**0.41**	**0.38**	**0.38**	**0.38**	**0.38**
ROUGE-2	**0.15**	**0.15**	0.15	**0.16**	**0.16**	**0.16**	0.13	**0.14**	**0.14**	**0.14**
ROUGE-L	**0.29**	**0.29**	**0.29**	**0.29**	0.29	**0.3**	**0.27**	**0.27**	**0.27**	**0.27**
BERTscore precision	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BERTscore recall	**0.72**	**0.72**	**0.72**	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BERTscore F1	**0.72**	**0.72**	**0.72**	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BARTscore faithful	**-2.8**	-2.81	**-2.95**	-2.97	**-2.88**	-2.9	**-3.12**	-3.13	-3.13	**-3.1**
BARTscore precision	**-2.73**	**-2.73**	**-2.74**	-2.75	**-2.7**	-2.71	**-2.86**	**-2.86**	-2.91	**-2.9**
BARTscore recall	-2.73	**-2.70**	-2.78	**-2.76**	-2.68	**-2.66**	-2.88	**-2.87**	-2.93	**-2.92**
BARTscore F1	-2.70	**-2.69**	-2.74	**-2.73**	**-2.66**	**-2.66**	**-2.84**	**-2.84**	-2.89	**-2.88**

Table 15: Overall average generative performance metrics for all generative tasks per model and group.

group	GPT3.5 western native	GPT3.5 not western native	GPT4o western native	GPT4o not western native	Haiku western native	Haiku not western native	Sonnet western native	Sonnet not western native	Qwen7B western native	Qwen7B not western native
BLEU	**0.44**	**0.44**	**0.47**	**0.47**	0.44	**0.45**	**0.45**	**0.45**	0.41	**0.42**
ROUGE-1	**0.39**	**0.39**	0.4	**0.41**	0.4	**0.41**	**0.38**	**0.38**	**0.38**	**0.38**
ROUGE-2	**0.15**	**0.15**	**0.16**	0.15	0.15	**0.16**	**0.14**	**0.14**	**0.14**	**0.14**
ROUGE-L	**0.29**	**0.29**	**0.29**	**0.29**	0.29	**0.3**	**0.27**	**0.27**	**0.27**	**0.27**
BERTscore precision	**0.72**	**0.72**	**0.72**	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BERTscore recall	**0.72**	**0.72**	**0.72**	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BERTscore F1	**0.72**	**0.72**	**0.72**	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**
BARTscore faithful	**-2.8**	-2.81	**-2.96**	**-2.96**	**-2.89**	**-2.89**	-3.13	**-3.12**	-3.12	**-3.1**
BARTscore precision	**-2.72**	-2.73	**-2.73**	-2.75	**-2.7**	**-2.7**	**-2.85**	-2.87	-2.92	**-2.9**
BARTscore recall	-2.73	**-2.71**	-2.77	**-2.76**	-2.69	**-2.66**	-2.88	**-2.87**	-2.95	**-2.91**
BARTscore F1	-2.7	**-2.69**	**-2.73**	**-2.73**	-2.67	**-2.66**	**-2.84**	**-2.84**	-2.9	**-2.87**

Table 16: Overall average generative performance metrics for all generative tasks per model and group.

model	**GPT3.5**		**GPT4o**		**Haiku**		**Sonnet**		**Qwen 7B**	
dataset	native	non-native	native	non-native	native	non-native	native	non-native	native	non-native
Story Cloze	**0.39**	0.38	**0.41**	**0.41**	**0.37**	0.36	**0.39**	**0.39**	**0.39**	0.38
CNN dailymail	**0.46**	0.45	**0.53**	**0.53**	**0.49**	**0.49**	**0.5**	**0.5**	**0.43**	**0.43**
CODA19	0.55	**0.56**	0.57	**0.58**	0.57	**0.58**	0.57	**0.58**	0.52	**0.53**
Paraphrasing	**0.36**	**0.36**	0.35	**0.37**	0.36	**0.38**	0.32	**0.33**	0.31	**0.33**

model	**GPT3.5**		**GPT4o**		**Haiku**		**Sonnet**		**Qwen 7B**	
dataset	Western	non Western	Western	non Western	Western	non Western	Western	non Western	Western	non Western
Story Cloze	**0.4**	0.38	**0.43**	0.41	0.36	**0.37**	**0.4**	0.39	**0.39**	0.38
CNN dailymail	0.46	0.46	**0.54**	0.53	**0.49**	**0.49**	**0.50**	**0.50**	0.42	**0.43**
CODA19	0.55	**0.56**	0.58	**0.58**	0.57	**0.58**	0.58	**0.58**	0.52	**0.53**
Paraphrasing	0.35	**0.36**	0.35	**0.37**	0.35	**0.38**	0.32	**0.33**	0.3	**0.33**

Table 17: Average BLEU scores per dataset and model for the native-non-native group and the Western non Western group.

dataset id	**StoryCloze**		**CNN Dailymail**		**CODA19**		**Paraphrase**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	**0.39**	0.38	**0.46**	0.45	0.55	**0.56**	**0.36**	**0.36**
ROUGE-1	**0.39**	0.38	**0.39**	0.38	0.43	**0.45**	0.35	**0.36**
ROUGE-2	**0.19**	**0.19**	**0.11**	**0.11**	0.14	**0.15**	**0.15**	0.14
ROUGE-L	**0.34**	**0.34**	**0.22**	**0.22**	0.26	**0.27**	**0.32**	**0.32**
BERTscore precision	**0.74**	**0.74**	**0.66**	**0.66**	0.70	**0.71**	0.77	**0.78**
BERTscore recall	**0.74**	**0.74**	**0.67**	**0.67**	0.70	**0.71**	**0.76**	**0.76**
BERTscore F1	**0.74**	**0.74**	**0.66**	**0.66**	0.70	**0.71**	**0.77**	**0.77**
BARTscore faithful	**-3.32**	-3.35	-2.3	**-2.29**	**-2.87**	**-2.87**	**-2.70**	-2.71
BARTscore precision	**-2.45**	-2.48	**-2.83**	-2.87	-2.77	**-2.71**	-2.86	**-2.85**
BARTscore recall	**-2.47**	-2.49	-2.68	**-2.66**	-2.8	**-2.74**	-2.96	**-2.94**
BARTscore F1	**-2.43**	-2.45	**-2.74**	-2.75	-2.78	**-2.72**	-2.87	**-2.86**

Table 18: Average Performance metrics per dataset for the native-non-native group for GPT3.5.

dataset id	StoryCloze		CNN Dailymail		CODA19		Paraphrase	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	**0.41**	**0.41**	**0.53**	**0.53**	0.57	**0.58**	0.35	**0.37**
ROUGE-1	**0.41**	**0.41**	**0.42**	**0.42**	0.41	**0.43**	0.36	**0.37**
ROUGE-2	**0.25**	0.23	0.11	**0.12**	0.12	**0.13**	**0.14**	**0.14**
ROUGE-L	**0.38**	**0.38**	**0.21**	**0.21**	0.23	**0.24**	0.32	**0.34**
BERTscore precision	**0.76**	0.75	**0.65**	**0.66**	0.69	**0.7**	0.78	**0.79**
BERTscore recall	**0.76**	0.75	**0.67**	**0.67**	**0.69**	**0.69**	**0.76**	**0.76**
BERTscore F1	**0.76**	0.75	**0.66**	**0.66**	0.69	**0.7**	**0.77**	**0.77**
BARTscore faithful	**-3.52**	-3.53	-2.65	**-2.62**	**-3.15**	-3.18	**-2.48**	-2.52
BARTscore precision	**-2.42**	**-2.42**	**-2.89**	-2.92	-2.9	**-2.88**	-2.77	**-2.76**
BARTscore recall	**-2.38**	-2.4	-2.82	**-2.81**	-2.96	**-2.92**	-2.95	**-2.9**
BARTscore F1	**-2.36**	-2.38	**-2.85**	-2.86	-2.92	**-2.9**	-2.82	**-2.79**

Table 19: Average Performance metrics per dataset for the native-non-native group for GPT4o.

dataset id	StoryCloze		CNN Dailymail		CODA19		Paraphrase	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	**0.37**	0.36	**0.49**	**0.49**	0.57	**0.58**	0.36	**0.38**
ROUGE-1	**0.39**	0.38	0.39	**0.4**	0.45	**0.46**	0.38	**0.4**
ROUGE-2	**0.2**	0.19	**0.12**	**0.12**	0.15	**0.16**	0.16	**0.18**
ROUGE-L	**0.36**	**0.36**	**0.22**	**0.22**	0.26	**0.27**	0.34	**0.36**
BERTscore precision	**0.75**	**0.75**	**0.63**	**0.63**	**0.7**	**0.7**	0.78	**0.79**
BERTscore recall	**0.75**	**0.75**	**0.65**	**0.65**	0.69	**0.7**	**0.77**	**0.77**
BERTscore F1	**0.75**	**0.75**	**0.64**	**0.64**	**0.7**	**0.7**	0.77	**0.78**
BARTscore faithful	**-3.28**	-3.31	-2.59	**-2.57**	**-2.85**	-2.88	**-2.78**	-2.82
BARTscore precision	**-2.39**	-2.42	**-2.97**	-2.99	**-2.66**	**-2.66**	-2.78	**-2.76**
BARTscore recall	**-2.39**	-2.4	-2.82	**-2.8**	-2.71	**-2.66**	-2.82	**-2.79**
BARTscore F1	**-2.35**	-2.37	**-2.88**	**-2.88**	-2.68	**-2.65**	-2.76	**-2.73**

Table 20: Average Performance metrics per dataset for the native-non-native group for Haiku.

dataset id	StoryCloze		CNN Dailymail		CODA19		Paraphrase	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	**0.39**	**0.39**	**0.5**	**0.5**	0.57	**0.58**	0.32	**0.33**
ROUGE-1	**0.38**	**0.38**	**0.41**	**0.41**	0.42	**0.43**	0.3	**0.31**
ROUGE-2	**0.18**	**0.18**	**0.12**	**0.12**	0.13	**0.14**	0.1	**0.11**
ROUGE-L	0.33	**0.34**	**0.22**	**0.22**	0.24	**0.25**	0.27	**0.28**
BERTscore precision	**0.74**	**0.74**	**0.65**	0.64	**0.7**	**0.7**	**0.76**	**0.76**
BERTscore recall	**0.74**	**0.74**	**0.66**	**0.66**	0.69	**0.7**	0.74	**0.75**
BERTscore F1	**0.74**	**0.74**	**0.65**	**0.65**	0.69	**0.7**	**0.75**	**0.75**
BARTscore faithful	**-3.29**	**-3.29**	-2.88	**-2.86**	**-3.07**	-3.08	**-3.24**	-3.26
BARTscore precision	-2.5	**-2.49**	**-3**	-3.03	**-2.78**	-2.79	**-3.16**	**-3.16**
BARTscore recall	-2.48	**-2.47**	**-2.91**	**-2.91**	-2.83	**-2.8**	-3.31	**-3.29**
BARTscore F1	-2.46	**-2.45**	**-2.94**	-2.95	-2.8	**-2.78**	-3.18	**-3.17**

Table 21: Average Performance metrics per dataset for the native-non-native group for Sonnet.

dataset id	StoryCloze		CNN Dailymail		CODA19		Paraphrase	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	**0.39**	0.38	**0.43**	**0.43**	0.52	**0.53**	0.31	**0.33**
ROUGE-1	**0.4**	**0.4**	0.38	**0.39**	0.4	**0.41**	0.33	**0.34**
ROUGE-2	**0.24**	0.23	**0.1**	**0.1**	0.1	**0.11**	**0.12**	**0.12**
ROUGE-L	**0.38**	0.37	**0.19**	**0.19**	**0.22**	**0.22**	0.29	**0.3**
BERTscore precision	**0.75**	**0.75**	**0.63**	**0.63**	**0.68**	**0.68**	0.74	**0.75**
BERTscore recall	**0.75**	**0.75**	**0.64**	**0.64**	0.68	**0.69**	0.72	**0.73**
BERTscore F1	**0.75**	**0.75**	**0.64**	**0.64**	**0.68**	**0.68**	0.73	**0.74**
BARTscore faithful	-3.59	**-3.55**	-2.82	**-2.78**	**-3.21**	-3.25	-2.87	**-2.83**
BARTscore precision	-2.45	**-2.43**	**-3**	**-3**	**-3.02**	**-3.02**	-3.2	**-3.14**
BARTscore recall	**-2.47**	-2.49	-2.92	**-2.91**	-3.07	**-3.02**	-3.28	**-3.24**
BARTscore F1	**-2.41**	-2.42	**-2.94**	-2.95	-3.03	**-3.01**	-3.17	**-3.13**

Table 22: Average Performance metrics per dataset for the native-non-native group for Qwen.

	Standard		Native Prompt		Non-Native Prompt		Guess Native		Guess Native correctly	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
accuracy	0.81	**0.82**	0.83	**0.84**	**0.83**	0.82	**0.78**	0.77	**0.76**	0.73
BLEU	**0.44**	**0.44**	**0.42**	**0.42**	**0.39**	**0.39**	**0.42**	**0.42**	**0.45**	0.42
ROUGE-1	**0.39**	**0.39**	**0.38**	**0.38**	0.36	**0.37**	**0.37**	**0.37**	**0.39**	0.37
ROUGE-2	**0.15**	**0.15**	**0.14**	**0.14**	**0.13**	**0.13**	**0.13**	**0.13**	**0.14**	0.12
ROUGE-L	**0.29**	**0.29**	**0.28**	**0.28**	0.27	**0.28**	**0.27**	**0.27**	**0.27**	**0.27**
BERTscore precision	**0.72**	**0.72**	**0.72**	**0.72**	0.72	**0.73**	**0.72**	**0.72**	**0.72**	0.71
BERTscore recall	**0.72**	**0.72**	**0.71**	**0.71**	**0.7**	**0.7**	**0.71**	**0.71**	**0.71**	**0.71**
BERTscore F1	**0.72**	**0.72**	**0.72**	0.71	**0.71**	**0.71**	**0.71**	**0.71**	**0.72**	0.71
BARTscore faithful	**-2.8**	-2.81	**-2.75**	-2.79	-2.72	**-2.71**	**-2.78**	**-2.78**	**-2.74**	-2.75
BARTscore precision	**-2.73**	**-2.73**	**-2.7**	-2.75	**-2.69**	-2.7	**-2.77**	-2.78	**-2.73**	-2.84
BARTscore recall	-2.73	**-2.7**	**-2.76**	**-2.76**	-2.8	**-2.79**	-2.76	**-2.75**	**-2.76**	-2.77
BARTscore F1	-2.7	**-2.69**	**-2.7**	-2.73	**-2.72**	**-2.72**	**-2.74**	**-2.74**	**-2.72**	-2.78

Table 23: Overall average generative performance metrics for all generative tasks per group and prompt of GPT3.5.

	Standard		**Native Prompt**		**Non-Native Prompt**		**Guess Native**		**Guess Native correctly**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
accuracy	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.83**	**0.83**	**0.83**	0.81
BLEU	**0.47**	**0.47**	0.46	**0.47**	0.45	**0.46**	**0.46**	**0.46**	**0.47**	0.46
ROUGE-1	0.4	**0.41**	**0.4**	**0.4**	0.39	**0.4**	0.39	**0.4**	0.39	**0.4**
ROUGE-2	0.15	**0.16**	**0.15**	**0.15**	0.14	**0.15**	**0.15**	**0.15**	0.12	**0.13**
ROUGE-L	**0.29**	**0.29**	0.28	**0.29**	0.28	**0.29**	0.28	**0.29**	0.24	**0.27**
BERTscore precision	**0.72**	**0.72**	**0.72**	**0.72**	0.72	**0.73**	**0.72**	**0.72**	0.68	**0.71**
BERTscore recall	**0.72**	**0.72**	**0.72**	**0.72**	**0.71**	**0.71**	0.71	**0.72**	0.69	**0.71**
BERTscore f1	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	0.69	**0.71**
BARTscore faithful	**-2.95**	-2.97	-2.95	**-2.94**	**-2.87**	-2.88	-2.92	**-2.91**	**-2.46**	-2.7
BARTscore precision	**-2.74**	-2.75	-2.76	**-2.74**	-2.73	**-2.72**	-2.75	**-2.74**	-2.78	**-2.73**
BARTscore recall	-2.78	**-2.76**	-2.8	**-2.77**	-2.82	**-2.8**	-2.81	-2.79	**-2.71**	-2.83
BARTscore f1	-2.74	**-2.73**	-2.76	**-2.73**	-2.75	**-2.73**	-2.76	**-2.74**	**-2.73**	-2.76

Table 24: Overall average generative performance metrics for all generative tasks per group and prompt of GPT4o.

	Standard		**Native Prompt**		**Non-Native Prompt**		**Guess Native**		**Guess Native correctly**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
accuracy	0.84	**0.85**	0.84	**0.85**	**0.83**	**0.83**	**0.65**	0.64	0.53	**0.64**
BLEU	**0.45**	**0.45**	**0.45**	**0.45**	0.44	**0.45**	0.44	**0.45**	0.34	**0.43**
ROUGE-1	0.4	**0.41**	0.4	**0.41**	0.4	**0.41**	**0.39**	**0.39**	0.35	**0.38**
ROUGE-2	**0.16**	**0.16**	0.15	**0.16**	0.15	**0.17**	0.14	**0.15**	**0.15**	0.13
ROUGE-L	0.29	**0.3**	0.29	**0.3**	0.29	**0.3**	**0.28**	**0.28**	**0.31**	0.27
BERTscore precision	**0.72**	**0.72**	0.71	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.73**	0.71
BERTscore recall	0.71	**0.72**	0.71	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.73**	0.71
BERTscore f1	0.71	**0.72**	0.71	**0.72**	0.71	**0.72**	**0.71**	**0.71**	**0.73**	0.71
BARTscore faithful	**-2.88**	-2.9	**-2.9**	**-2.9**	**-2.88**	-2.89	**-2.97**	**-2.97**	-3.31	**-2.95**
BARTscore precision	**-2.7**	-2.71	-2.7	**-2.69**	**-2.7**	**-2.7**	**-2.85**	-2.87	**-2.71**	-2.91
BARTscore recall	-2.68	**-2.66**	-2.69	**-2.66**	-2.7	**-2.66**	-2.75	**-2.74**	**-2.66**	-2.81
BARTscore f1	**-2.66**	**-2.66**	-2.67	**-2.65**	-2.67	**-2.66**	**-2.77**	-2.78	**-2.64**	-2.83

Table 25: Overall average generative performance metrics for all generative tasks per group and prompt of Haiku.

	Standard		**Native Prompt**		**Non-Native Prompt**		**Guess Native**		**Guess Native correctly**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
accuracy	**0.86**	**0.86**	**0.86**	**0.86**	**0.66**	**0.66**	**0.65**	**0.65**	0.53	**0.56**
BLEU	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	0.44	**0.45**	0.39	**0.46**
ROUGE-1	**0.38**	**0.38**	0.38	**0.39**	**0.38**	**0.38**	0.37	**0.38**	0.34	**0.38**
ROUGE-2	0.13	**0.14**	**0.14**	**0.14**	0.13	**0.14**	**0.13**	**0.13**	0.1	**0.13**
ROUGE-L	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.26**	**0.26**	0.24	**0.25**
BERTscore precision	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	0.7
BERTscore recall	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	0.7
BERTscore f1	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**	0.7
BARTscore faithful	**-3.12**	-3.13	**-3.13**	-3.14	**-3.08**	-3.09	**-3.1**	**-3.1**	**-3.04**	-3.06
BARTscore precision	**-2.86**	**-2.86**	**-2.87**	**-2.87**	**-2.87**	**-2.87**	**-2.88**	-2.89	**-2.89**	-2.95
BARTscore recall	-2.88	**-2.87**	-2.87	**-2.85**	-2.89	**-2.88**	**-2.89**	**-2.89**	-3.01	**-2.88**
BARTscore f1	**-2.84**	**-2.84**	-2.85	**-2.83**	-2.85	**-2.84**	**-2.86**	**-2.86**	-2.89	**-2.88**

Table 26: Overall average generative performance metrics for all generative tasks per group and prompt of Sonnet.

	Standard		**Native Prompt**		**Non-Native Prompt**		**Guess Native**		**Guess Native correctly**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
accuracy	**0.82**	0.81	**0.82**	**0.82**	**0.82**	**0.82**	0.69	**0.7**	0.61	**0.76**
BLEU	0.41	**0.42**	0.41	**0.42**	0.41	**0.42**	**0.39**	**0.39**	**0.43**	0.33
ROUGE-1	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**	0.34	**0.35**	**0.38**	0.32
ROUGE-2	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.1**	**0.1**	**0.12**	0.07
ROUGE-L	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.23**	**0.23**	**0.24**	0.21
BERTscore precision	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.65**	**0.65**	**0.63**	**0.63**
BERTscore recall	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.68**	**0.68**	**0.67**	**0.67**
BERTscore f1	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.66**	**0.66**	**0.65**	**0.65**
BARTscore faithful	-3.13	**-3.1**	**-3.11**	**-3.11**	**-3.11**	**-3.11**	**-3.73**	**-3.73**	**-3.28**	-3.64
BARTscore precision	-2.91	**-2.9**	**-2.91**	**-2.91**	**-2.91**	**-2.91**	**-3.74**	-3.76	**-3.46**	-3.89
BARTscore recall	-2.93	**-2.92**	-2.93	**-2.91**	-2.92	**-2.91**	-3.15	**-3.14**	**-3.03**	-3.25
BARTscore f1	-2.89	**-2.88**	-2.89	**-2.88**	**-2.88**	**-2.88**	**-3.38**	**-3.38**	**-3.21**	-3.51

Table 27: Overall average generative performance metrics for all generative tasks per group and prompt of Qwen.

	Standard		**Native Prompt**		**Non-Native Prompt**		**Guess Native**		**Guess Native correctly**	
group	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**	**native**	**non-native**
BLEU	0.21	**0.22**	**0.21**	**0.21**	**0.21**	**0.21**	0.21	**0.22**	0.21	**0.22**
ROUGE-1	0.26	**0.27**	0.26	**0.27**	0.26	**0.27**	0.26	**0.27**	**0.28**	0.27
ROUGE-2	**0.08**	**0.08**	0.08	**0.09**	**0.09**	**0.09**	0.08	**0.09**	**0.09**	**0.09**
ROUGE-L	0.2	**0.21**	0.2	**0.21**	**0.21**	**0.21**	0.2	**0.21**	0.2	**0.21**
BERTscore precision	0.68	**0.69**	**0.69**	**0.69**	0.69	**0.7**	0.68	**0.69**	0.68	**0.69**
BERTscore recall	**0.64**	**0.64**	0.63	**0.64**	0.63	**0.64**	**0.64**	**0.64**	0.61	**0.64**
BERTscore F1	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**	0.64	**0.66**
BARTscore faithful	**-2.8**	-2.81	**-2.75**	-2.79	-2.72	**-2.71**	**-2.78**	**-2.78**	**-2.74**	-2.75
BARTscore precision	-2.84	**-2.82**	**-2.8**	**-2.8**	-2.75	**-2.73**	-2.84	**-2.83**	**-2.72**	-2.85
BARTscore recall	-3.29	**-3.26**	-3.3	**-3.27**	-3.29	**-3.27**	-3.28	**-3.25**	-3.34	**-3.25**
BARTscore F1	-3.01	**-2.99**	-2.99	**-2.98**	-2.96	**-2.94**	-3.0	**-2.99**	**-2.97**	-3.0

Table 28: Overall average generative performance metrics for all generative tasks per group and prompt of GPT3.5 for the original gold answer.

group	Standard native	Standard non-native	Native Prompt native	Native Prompt non-native	Non-Native Prompt native	Non-Native Prompt non-native	Guess Native native	Guess Native non-native	Guess Native correctly native	Guess Native correctly non-native
BLEU	**0.28**	**0.28**	**0.27**	**0.27**	0.26	**0.27**	**0.27**	**0.27**	0.18	**0.26**
ROUGE-1	0.3	**0.31**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	0.27	**0.31**
ROUGE-2	0.09	**0.1**	0.09	**0.1**	**0.1**	**0.1**	0.09	**0.1**	0.08	**0.11**
ROUGE-L	0.22	**0.23**	**0.22**	**0.22**	0.22	**0.23**	**0.22**	**0.22**	0.18	**0.23**
BERTscore precision	**0.68**	**0.68**	**0.68**	**0.68**	0.68	**0.69**	**0.68**	**0.68**	0.67	**0.69**
BERTscore recall	0.65	**0.66**	0.65	**0.66**	0.65	**0.66**	0.65	**0.66**	0.62	**0.65**
BERTscore F1	0.66	**0.67**	0.66	**0.67**	**0.67**	**0.67**	0.66	**0.67**	0.64	**0.67**
BARTscore faithful	**-2.95**	-2.97	-2.95	**-2.94**	**-2.87**	-2.88	-2.92	**-2.91**	**-2.46**	-2.7
BARTscore precision	-2.89	**-2.88**	-2.89	**-2.86**	-2.85	**-2.83**	-2.88	**-2.85**	-2.79	**-2.74**
BARTscore recall	-3.25	**-3.22**	-3.25	**-3.22**	-3.24	**-3.22**	-3.27	**-3.23**	-3.29	**-3.2**
BARTscore F1	-3.03	**-3.01**	-3.03	**-3.0**	-3.0	**-2.98**	-3.03	**-3.0**	-3.0	**-2.92**

Table 29: Overall average generative performance metrics for all generative tasks per group and prompt of GPT4o for the original gold answer.

group	Standard native	Standard non-native	Native Prompt native	Native Prompt non-native	Non-Native Prompt native	Non-Native Prompt non-native	Guess Native native	Guess Native non-native	Guess Native correctly native	Guess Native correctly non-native
BLEU	0.25	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.25**	**0.25**	0.18	**0.26**
ROUGE-1	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	0.28	**0.29**	0.24	**0.3**
ROUGE-2	**0.08**	**0.08**	0.08	**0.09**	**0.09**	**0.09**	**0.08**	**0.08**	0.05	**0.08**
ROUGE-L	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**	0.2	**0.21**	0.2	**0.22**
BERTscore precision	**0.67**	**0.67**	0.67	**0.68**	0.67	**0.68**	**0.67**	**0.67**	0.66	**0.67**
BERTscore recall	**0.64**	**0.64**	0.64	**0.65**	**0.64**	**0.64**	0.64	**0.65**	**0.7**	0.65
BERTscore F1	0.65	**0.66**	0.65	**0.66**	0.65	**0.66**	**0.65**	**0.65**	**0.67**	0.66
BARTscore faithful	**-2.88**	-2.9	**-2.9**	**-2.9**	**-2.88**	-2.89	**-2.97**	**-2.97**	-3.31	**-2.95**
BARTscore precision	**-2.86**	**-2.86**	**-2.85**	**-2.85**	**-2.85**	-2.86	**-2.98**	**-2.98**	**-2.96**	**-2.96**
BARTscore recall	-3.33	**-3.32**	-3.33	**-3.31**	-3.33	**-3.32**	-3.33	**-3.32**	**-3.31**	**-3.31**
BARTscore F1	**-3.04**	**-3.04**	-3.04	**-3.03**	**-3.04**	**-3.04**	-3.11	**-3.1**	**-3.08**	-3.09

Table 30: Overall average generative performance metrics for all generative tasks per group and prompt of Haiku for the original gold answer.

group	Standard native	Standard non-native	Native Prompt native	Native Prompt non-native	Non-Native Prompt native	Non-Native Prompt non-native	Guess Native native	Guess Native non-native	Guess Native correctly native	Guess Native correctly non-native
BLEU	**0.25**	**0.25**	**0.25**	**0.25**	**0.26**	**0.26**	**0.25**	**0.25**	0.2	**0.24**
ROUGE-1	**0.27**	**0.27**	0.27	**0.28**	**0.28**	**0.28**	**0.27**	**0.27**	0.24	**0.27**
ROUGE-2	**0.07**	**0.07**	**0.07**	**0.07**	0.07	**0.08**	**0.07**	**0.07**	0.06	**0.07**
ROUGE-L	**0.19**	**0.19**	**0.19**	**0.19**	**0.2**	**0.2**	**0.19**	**0.19**	**0.19**	0.18
BERTscore precision	0.66	**0.67**	0.66	**0.67**	**0.67**	**0.67**	**0.66**	**0.66**	**0.66**	**0.66**
BERTscore recall	**0.65**	**0.65**	**0.65**	**0.65**	**0.65**	**0.65**	**0.65**	**0.65**	**0.66**	0.63
BERTscore F1	**0.65**	**0.65**	0.65	**0.66**	**0.66**	**0.66**	**0.65**	**0.65**	**0.66**	0.64
BARTscore faithful	**-3.12**	-3.13	**-3.13**	-3.14	**-3.08**	-3.09	**-3.1**	**-3.1**	**-3.04**	-3.06
BARTscore precision	-3.06	**-3.05**	**-3.08**	**-3.08**	-3.06	**-3.05**	**-3.06**	-3.07	-3.08	**-3.07**
BARTscore recall	-3.31	**-3.3**	**-3.3**	**-3.3**	**-3.29**	**-3.29**	**-3.31**	**-3.31**	**-3.33**	-3.36
BARTscore F1	-3.14	**-3.13**	**-3.15**	**-3.15**	**-3.13**	**-3.13**	**-3.14**	**-3.14**	**-3.17**	**-3.17**

Table 31: Overall average generative performance metrics for all generative tasks per group and prompt of Sonnet for the original gold answer.

group	Standard native	Standard non-native	Native Prompt native	Native Prompt non-native	Non-Native Prompt native	Non-Native Prompt non-native	Guess Native native	Guess Native non-native	Guess Native correctly native	Guess Native correctly non-native
BLEU	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	0.23	**0.24**	0.17	**0.25**
ROUGE-1	**0.29**	**0.29**	**0.29**	**0.29**	0.28	**0.29**	0.26	**0.27**	0.28	**0.3**
ROUGE-2	**0.08**	**0.08**	**0.08**	**0.08**	**0.08**	**0.08**	**0.06**	**0.06**	**0.08**	**0.08**
ROUGE-L	0.2	**0.21**	0.2	**0.21**	0.2	**0.21**	0.18	**0.19**	0.18	**0.23**
BERTscore precision	0.66	**0.67**	0.66	**0.67**	0.66	**0.67**	**0.62**	**0.62**	**0.62**	**0.62**
BERTscore recall	0.64	**0.65**	0.64	**0.65**	**0.65**	**0.65**	**0.63**	**0.63**	0.61	**0.65**
BERTscore F1	0.65	**0.66**	0.65	**0.66**	**0.65**	**0.65**	**0.62**	**0.62**	0.61	**0.64**
BARTscore faithful	-3.13	**-3.1**	**-3.11**	**-3.11**	**-3.11**	**-3.11**	**-3.73**	**-3.73**	**-3.28**	-3.64
BARTscore precision	-3.07	**-3.05**	**-3.06**	**-3.06**	**-3.07**	**-3.07**	**-3.82**	-3.83	**-3.51**	-3.94
BARTscore recall	-3.29	**-3.27**	-3.29	**-3.27**	-3.29	**-3.27**	-3.41	**-3.39**	-3.4	**-3.33**
BARTscore F1	-3.14	**-3.12**	-3.13	**-3.12**	-3.14	**-3.13**	-3.55	**-3.53**	**-3.4**	-3.56

Table 32: Overall average generative performance metrics for all generative tasks per group and prompt of Qwen7B for the original gold answer.