

DO YOU SPEAK BASQUENGLISH? ASSESSING LOW-RESOURCE MULTILINGUAL PROFICIENCY OF PRETRAINED LANGUAGE MODELS

Iñigo Parra¹

¹Department of Modern Languages and Classics, University of Alabama,
Tuscaloosa, Alabama

ABSTRACT

Multilingual language models' have democratized access to information and artificial intelligence (AI). Still, low-resource languages (LRL) remain underrepresented. This study compares the performance of GPT-4, LLaMa (7B), and PaLM 2 when asked to reproduce English-Basque code-switched outputs. The study uses code-switching as a test to argue for the multilingual capabilities of each model and compares and studies their cross-lingual understanding. All models were tested using 84 prompts (N = 252), with their responses subjected to qualitative and quantitative analysis. This study compares the naturalness of the outputs, code-switching competence (CSness), and the frequency of hallucinations. Results of pairwise comparisons show statistically significant differences in naturalness and the ability to produce grammatical code-switched output across models. This study underscores the critical role of linguistic representation in large language models (LLMs) and the necessity for improvement in handling LRLs.

KEYWORDS

Basque, code-switching, low-resource languages, multilingual models

1. INTRODUCTION

Code-switching (CS) is the linguistic phenomenon that refers to alternating between two or more languages in a conversation. It usually arises within multilingual environments, where speakers are exposed to multiple languages. Several factors, such as the topic of conversation, the relationship between the interlocutors, and the interaction context, can influence language selection. This linguistic ability is a reflection of the speakers' cultural competence and cognitive flexibility. In this way, code-switching represents a complex, rule-governed use of language that provides rich insight into issues of identity, culture, and power.

Studies on bilingualism and code-switching have identified two related but distinct phenomena, code-switching (CS) and code-mixing (CM). Code-mixing involves the practice of mixing languages in a single sentence, whereas code-switching can occur either within or across sentence boundaries within a single discourse or constituent. Although these two phenomena are often used interchangeably, there are some subtle differences in their usage. Previous literature [1] [2] has differentiated code-switching from code-mixing, emphasizing that the former involves a more conscious and deliberate choice of language, while the latter is often more spontaneous and less planned. However, this study refers to code-switching and code-mixing

synonymously, taking both to refer to the practice of using more than one language within a single discourse or conversation. Understanding the nature of code-switching and code-mixing is important for researchers in the field of bilingualism, as well as for educators and policymakers who work with bilingual communities. By examining the patterns and contexts in which these practices occur, we can gain insight into the linguistic and cognitive processes that underlie bilingualism, as well as the social and cultural factors that shape language use.

Code-switching behavior has been studied differently depending on the age of the group looked at. Code-switching studies conducted in bilingual adults have analyzed grammatical and communicative functions [3] [4]. The complexity found in bilingual adults' code-switching attitudes revealed a sophisticated knowledge of the grammatical schemes of both languages and reflected the individuals' aptitude in using them appropriately. However, early research argued that code-switching revealed grammatical disorders caused by bilingual or multilingual language learning. From this point, language switching is considered a linguistic failure. Most studies pointing in this direction analyze what children's code-switching behavior suggests about their linguistic competency. Studies on kids' language alternation have postulated that bilingual children's mixing or switching of languages is provoked either by confusion or linguistic incompetency.

With multilingualism becoming common in today's world, there has been increasing interest in code-switching within natural language processing (NLP). This study focuses on the case of Basque and English code-switching to test the multilingual capabilities of models when dealing with low-resource languages (LRLs). The Basque language is a linguistically unique and low-resource language primarily spoken in the Basque Country, an autonomous region located in northern Spain, and to a lesser extent, in Navarre and some areas of southern France (Figure 1). These regions are the stronghold of the Basque language, which is noted for its distinct linguistic features, not related to any other known language. Despite its low number of native speakers (750,000) compared to global languages such as English, Basque has exhibited resilience and adaptability, largely due to active language revitalization efforts [5].

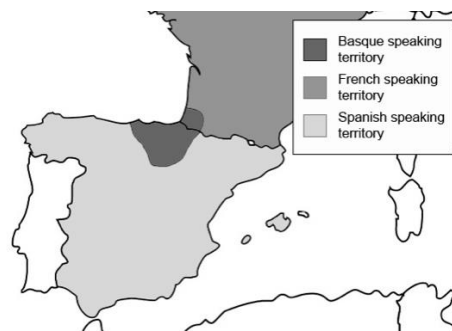


Figure 1. Map of Basque-speaking territory

In the evolution of language models, GPT-4 [6], LLaMa (7B) [7], and PaLM 2 [8], signify notable milestones. GPT-4, developed by OpenAI, is an enhanced version of its predecessor GPT-3.5, demonstrating remarkable capability in generating human-like text based on provided prompts. It harnesses the power of a transformer-based architecture and is trained on a diverse range of internet text. However, it also shares the limitations of its earlier iterations such as potential biases in the training data and lack of an understanding of factual correctness. LLaMa (7B) is the most lightweight model of LLaMa, developed by Meta's Fundamental AI Research (FAIR) team. This model distinguishes itself through its capacity to understand, learn, and adapt to different

languages and multimodal inputs, emphasizing linguistic diversity and inclusivity. It addresses some of the critiques of earlier monolingual-centric models and works towards reducing the digital language divide. Finally, PaLM 2 represents a paradigm shift towards procedural language models. It goes beyond the standard language model capabilities, demonstrating an understanding of procedural and sequential tasks. This ability to follow a sequence of instructions and generate logical, step-by-step explanations represents a significant advancement in the practical applications of language models [9].

2. PREVIOUS WORK

Literature on code-switching language modeling has underpinned the challenges that arise from the limited availability of large-scale code-switched data for training [10] [11]. Other code-switching language modeling studies have explored the use of recurrent neural network language models (RNNLM) and factored language models (FLM) for language modeling in code-switched speech [12]. These techniques integrated part-of-speech tags (POS) and language information (LID), significantly improving perplexity scores. Synthetic data augmentation has also been proposed as a solution to these limitations. [13] offer a sequence-to-sequence model that generates code-switching data by leveraging parallel mono-lingual translations from limited sources of code-switched data using a copy mechanism.

Some studies have focused on this problem in the context of automatic speech recognition (ASR). New methodologies to confront these limitations have consisted in introducing ASR-motivated evaluation setups that demonstrated to outperform generative language modeling [11]. Within the ASR framework, researchers have also proposed end-to-end (E2E) models to deal with intra-sentential CS [14], new datasets for ASR systems capable of identifying code-switching [15], and named entity recognition (NER) improvements in ASR through speech editing data augmentation [16].

This study proposes a different focus; it explores the utility of code-switching to evaluate the multilingual capabilities and cross-lingual understanding of language models. Drawing inspiration from [17], it analyzes the performance of three major models in handling low-resource-high-resource code-switching. Unlike previous studies that primarily focus on monolingual contexts, this research contributes to the field by simulating real-world language behaviour.

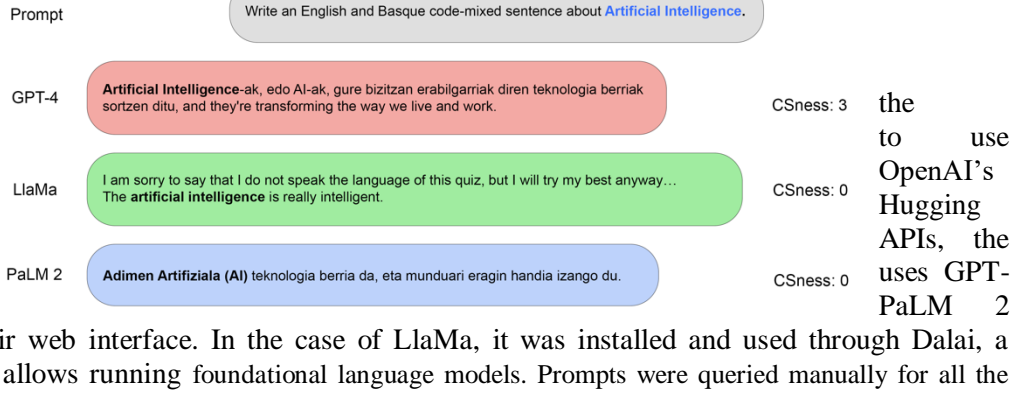
3. METHODOLOGY

3.1. Design

This study compares three of the most widely used text-generation models: GPT-4, LLaMa (7B), and PaLM 2. GPT-4 and PaLM 2 were selected due to their power and extensive capabilities; LLaMa (7B) was selected to compare the performance of models with lesser parameters. New currents in artificial intelligence and natural language processing may benefit from developing and deploying computationally light models.

The experiment consisted of 252 queries ($N = 252$), 84 for each model ($n = 84$). These were divided into groups of 6 ($n = 6$) by topic. Topics include artificial intelligence, family, sports, language, weather, food, politics, news, economy, education, video games, music, traveling, and photography. Topics were selected with the objective of having a holistic view of the linguistic (or code-switching) capabilities of each model. Figures 2, 3, and 4 are some of the examples of the queries.

Due to inability or Face's study 4 and



through their web interface. In the case of LlaMa, it was installed and used through Dalai, a library that allows running foundational language models. Prompts were queried manually for all the models.

Figure 2. Prompt example on artificial intelligence AI. Models' CSness score is shown on the right



Figure 3. Prompt example on news

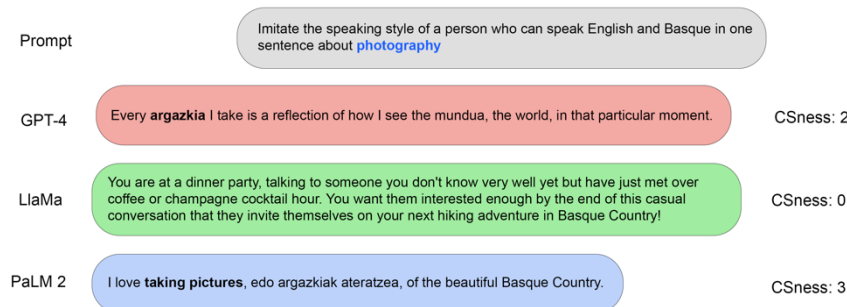


Figure 4. Prompt example on photography

3.2. Evaluation

Results were evaluated from a qualitative and quantitative perspective. The qualitative analysis provided tools for judging the naturalness, code-switching ability, and hallucinations from a human perspective. After annotation, quantitative analysis provided the necessary tools to judge the statistical significance of the results obtained.

Through naturalness, the study measured the coherence, grammatically, and fluency of the outputs. Every answer was given a naturalness score ranging from 0 (not natural) to 10 (very natural). Even if naturalness judged three aspects, coherence (i.e., responding to the prompt) was the minimum requirement to obtain the minimum score of 1.

The models' ability to produce code-switching was measured by code-switchness (CSness). Each code-mixed answer was given a score from 0 to 3. 0 score meant no code-switching –or code-mixing–; answers were given just in English. Those outputs with loanwords from Basque (e.g., *sirimiri*, *aupa*) were given a score of 1. Cases in which switching consisted of single, isolated tokens, such as cases in context L1_L1, were also given 1. The study provided a score of 2 to the generated code-switched text that showed the ability to go beyond the loanwords but that stayed in the domain of the topic queried (e.g., photography is great, *argazkiak ateratzea gustatzen zait*). Finally, a score of 3 was provided to the answers that showed code-switching abilities that went beyond the domain of the topic queried about (e.g., Sports are a good way of staying healthy *baita lagunekin momentu onak partekatzeke era bat*).

In artificial intelligence, hallucination refers to the creation of outputs that seem plausible but are either untrue or irrelevant to the given context. These outputs can arise from the AI model's inherent biases, lack of real-world knowledge, or training data constraints. Hallucinations were evaluated across groups to compare the linguistic performance of the three chosen models. This study classified hallucinations according to qualitative criteria as follows:

- **Out-of-Switch Language (OOS):** Out-of- speech (OOS) hallucinations happened when the output included keywords or phrases in a language that was not mentioned in the query. Most OOS hallucinations included languages that are geographically close to Basque (Spanish and French).
- **Linguistic Hallucination:** Linguistic hallucinations are ungrammatical responses. Some of the outputs included odd semantics, redundant syntax, or made-up lexicon.
- **Unexpected Output:** Unexpected output hallucinations included the outputs in which the model provided information that was not included in the prompt. This category also included those responses that did not answer the original query or that were produced by the comprehensive failures of the model.
- **Translation:** Translation hallucinations occurred when the model conceived code-switching as a direct translation of statements. This category also included outputs in which the model inserted translations in the form of clarification.

3.3. Exclusion Criteria

Exclusion criteria were applied to the outputs of the prompts. It was found that when models misinterpreted queries, they provided answers that followed similar patterns. These patterns included explanatory add-ons that were excluded to evaluate the code-switchness (CSness) of the output. Some examples are shown in Table 1.

Other analysis criteria included the evaluation of non-sensical outputs. To keep the analysis consistent, those outputs were considered hallucinations (classified as non-expected output) and got 0 score on code-switchness. There was not any case that showed high code-switchness and non-sensical outputs. Similarly, structures connoting clarification (often between quotation marks

or dashes) were not included in code-switch or code-mixing for this study, thus they got 0 code-switchness.

In outputs where two or more degrees of code-switchness (CSness) were found the output got the highest-ranked degree. This criterion was used mainly in answers to queries consisting of recreating a code-switched conversation. These outputs usually showed a variety of code-switches, which required a consistent evaluation across queries and models. In the same line, outputs that showed direct translation were given a score of 1 in CSness due to their lack of code-switching complexity.

Model	Query	Original	After Exclusion Criteria
GPT-4	Code-mixing refers to a phenomenon of combining two or more languages in a single sentence. Can you generate a code-mixed English-Basque sentence about food?	Sure, here's a code-mixed sentence: "This paella is delicious, oso gozoa da!"	This paella is delicious, oso gozoa da!
LlaMa (7B)	Write an English and Basque code-mixed sentence about sports.	In Euskara bat (basketball) is called gorri da or jaiusko da, depending on where it's being played/who plays in it...	In Euskara bat is called gorri da or jaiusko da, depending on where it's being played/who plays in it...
PaLM 2	Code-mixing refers to a phenomenon of combining two or more languages in a single sentence. Can you generate a code-mixed English-Basque sentence about politics?	I'm so excited to vote for the new Basque president! Eta irabazi behar du!	I'm so excited to vote for the new Basque president! Eta irabazi behar du!

Table 1. Examples of the exclusion criteria per model

4. RESULTS

4.1. General Overview

The study analyzed three categories to judge the multilingual capabilities of each model in the case of English-Basque code-switching. After annotation, the descriptive features of the results from the prompts are described in Tables 2, 3, and 4.

GPT-4													
Hallucinations					Code-switching					Naturalness			
Total	T	LH	OOS	UO	Switches	Total (CSness)	<i>M</i>	<i>mo</i>	<i>stdev</i>	Total	<i>M</i>	<i>mo</i>	<i>stdev</i>
36	15	14	3	4	69	127	1.51	1	1.07	518	6.17	6	2.01

Table 2. Descriptive analysis of the results for the data obtained from GPT-4

LlaMa (7B)													
Hallucinations					Code-switching					Naturalness			
Total	T	LH	OOS	UO	Switches	Total (CSness)	<i>M</i>	<i>mo</i>	<i>stdev</i>	Total	<i>M</i>	<i>mo</i>	<i>stdev</i>
46	1	8	5	30	2	4	0.05	0	0.34	393	4.68	0	3.03

Table 3. Descriptive analysis of the results for the data obtained from LlaMa (7B)

PaLM 2													
Hallucinations					Code-switching					Naturalness			
Total	T	LH	OOS	UO	Switches	Total (CSness)	<i>M</i>	<i>mo</i>	<i>stdev</i>	Total	<i>M</i>	<i>mo</i>	<i>stdev</i>
42	36	2	2	2	11	21	0.25	0	0.73	513	6.11	5	2.32

Table 4. Descriptive analysis of the results for the data obtained from PaLM 2

Figure 5 provides a general overview of the metrics under analysis. Total scores were summed and normalized to compare the three models equally. The initial landscape pointed out that GPT-4 was the model with the most code-switches produced (CS total) in Basque-English or vice-versa (0.69). PaLM 2 was significantly behind its direct opponent (0.11) and LLaMa (7B) was the model that least code-switches produced (0.02). This meant that GPT-4 was able to produce code-switches for the 82% of the prompts, PaLM 2 for the 11%, and LLaMa for the 2%. GPT-4 was also the model that showed fewer total hallucinations (0.36), followed by PaLM 2 (0.42), while LLaMa (7B) showed the highest number of hallucinations (0.46). As for naturalness, both PaLM 2 and GPT-4 were paired (0.518 and 0.513, respectively); LLaMa (7B) obtained 0.39.

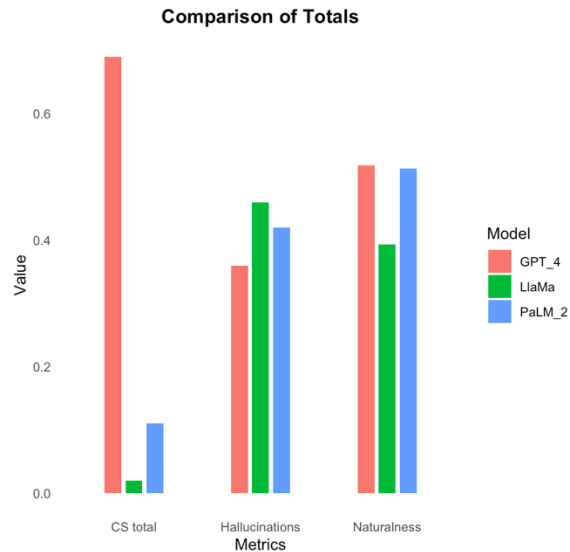


Figure 5. Totals of code-switched answers (CS total), hallucinations, and naturalness

4.2. Code Switchness

Code-switchness (CSness) measured the quality of the switches produced. It was hypothesized that the data did not follow a normal distribution. To have a closer insight, the Shapiro-Wilk test was used to assess the normality of the data distribution for the CSness of the models. The study assumed a significance level of $p = 0.05$. The results indicated a significant deviation from normality for all three models: GPT-4 ($W = 0.858$, $p = 0.0001$), LLaMa (7B) ($W = 0.123$, $p = 0.0002$), and PaLM 2 ($W = 0.386$, $p = 0.0002$). To test the homogeneity of variances among the three models a Brown-Forsythe test was performed. The F statistic was 7.126 (numerator $df = 2$, denominator $df = 138.16$), with a p -value of 0.0002, indicating a significant difference in variances among the three models. The assumptions of normality and homogeneity of variance, which are prerequisites for parametric statistical tests, were violated in this case. This was taken into account at the time of considering further statistical analyses that compare the CSness of the models.

The Kolmogorov-Smirnov test measures the extent to which two data samples are drawn from the same distribution. In this context, it was used to determine whether the distributions of CSness scores from each pair of models were similar. GPT-4 and LLaMa ($p < 0.05$), LLaMa and PaLM 2 ($p = 0.017$), and GPT-4 and PaLM ($p < 0.05$) indicated that the pairs did not have the same distribution. The Wilcoxon rank sum test with continuity correction pointed out in the same direction: GPT-4 and LLaMa ($p < 0.05$), LLaMa and PaLM 2 ($p = 0.009$), and GPT-4 and PaLM 2 ($p < 0.05$) statistically differed in the distributions of their CSness scores.

The results of the code-switching ability in English-Basque were tested to gain insight into the statistical significance of the differences between models. In line with the previous data exploration, a pairwise comparison through the Wilcoxon rank sum test with continuity correction was performed. In addition, to control Type I errors, the Bonferroni p -value adjustment method was applied. Results are shown in Figure 6.

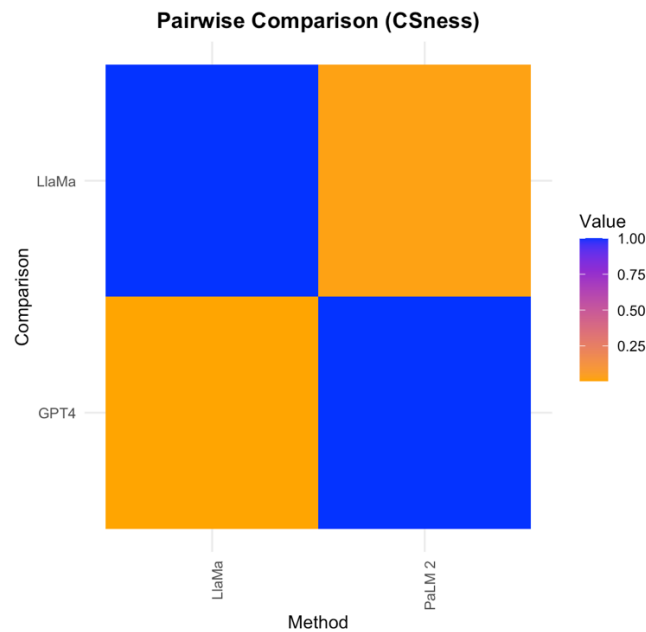


Figure 6. Comparison of CSness shown across models

The comparison between GPT-4 and LLaMa indicated a statistically significant difference in the code-switching ability of both models ($p = 0.0047$). For GPT-4 and PaLM 2 the p -value was 1, therefore, their code-switching abilities in the case under study did not show a meaningful statistical difference. Finally, LLaMa and PaLM 2 indicated a statistically significant difference in English-Basque code-switching performance ($p = 0.0215$). In conclusion, GPT-4 differed significantly in terms of code-switching ability from LLaMa but not from PaLM 2, its direct competitor. LLaMa and PaLM 2 also showed significantly different code-switching abilities.

4.3. Naturalness

For naturalness, the same procedure as for CSness was followed. The Shapiro-Wilk test indicated a significant deviation from normality for all three models: GPT-4 ($W = 0.949$, $p = 0.002$), LLaMa (7B) ($W = 0.924$, $p = 0.0001$), and PaLM 2 ($W = 0.943$, $p = 0.001$). To test the homogeneity of variances among the three models a Brown-Forsythe test was performed. The F statistic was 7.823 (numerator $df = 2$, denominator $df = 161.99$), with a p -value of 0.0005, indicating a significant difference in variances among the three models. The assumptions of normality and homogeneity of variance were also violated in this case.

The Kolmogorov-Smirnov indicated that GPT-4 and LLaMa ($p = 0.001$) and LLaMa and PaLM 2 ($p = 0.029$) did not have the same distribution. In contrast, GPT-4 and PaLM ($p = 0.329$), suggested that their distributions of naturalness scores were similar. The Wilcoxon rank sum test with continuity correction pointed out in the same direction: GPT-4 and LLaMa ($p = 0.0015$) and

LLaMa and PaLM 2 ($p = 0.007$) statistically differed in the distributions of their naturalness scores. By contrast, GPT-4 and PaLM 2 ($p = 0.522$) indicated no significant difference.

Based on both the Kolmogorov-Smirnov test and Wilcoxon rank sum test, GPT-4 and LLaMa (7B), as well as LLaMa and PaLM 2, generated text with different degrees of naturalness when code-switching in English and Basque. However, GPT-4 and PaLM 2 produced text with comparable naturalness in the same context of code-switching. GPT-4 is the model that most robustness showed, as indicated by its range (Figure 7).

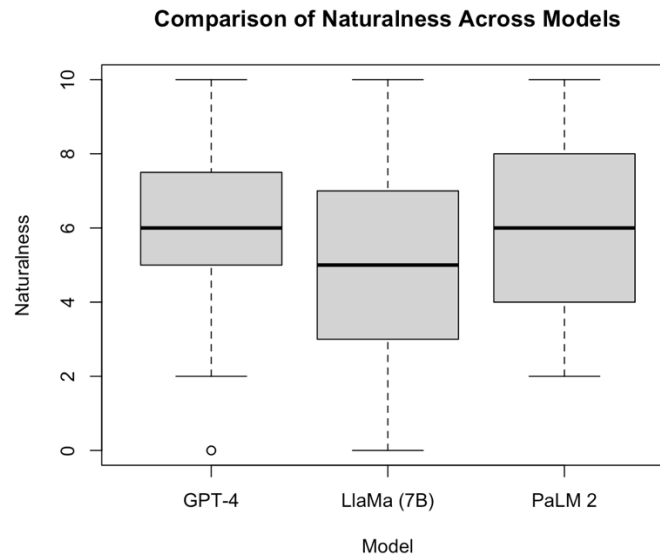


Figure 7. Comparison of naturalness shown across models

A pairwise comparison through the Wilcoxon rank sum test with continuity correction was performed (with Bonferroni p -value adjustment method). The comparison between GPT-4 and LLaMa yielded a p -value that was effectively zero ($< 2e - 16$). It suggested a statistically significant difference in the naturalness of the text produced by these two models. Similarly, the comparison between GPT-4 and PaLM 2 was also close to zero ($4.1e - 16$), indicating a statistically significant difference. The comparison between LLaMa and PaLM 2 yielded $p = 0.029$, which was less than the significance value ($p = 0.05$), again suggesting a statistically significant difference in the naturalness of the text produced. In conclusion, each pair of language models had significantly different levels of naturalness. The results are shown in Figure 8.

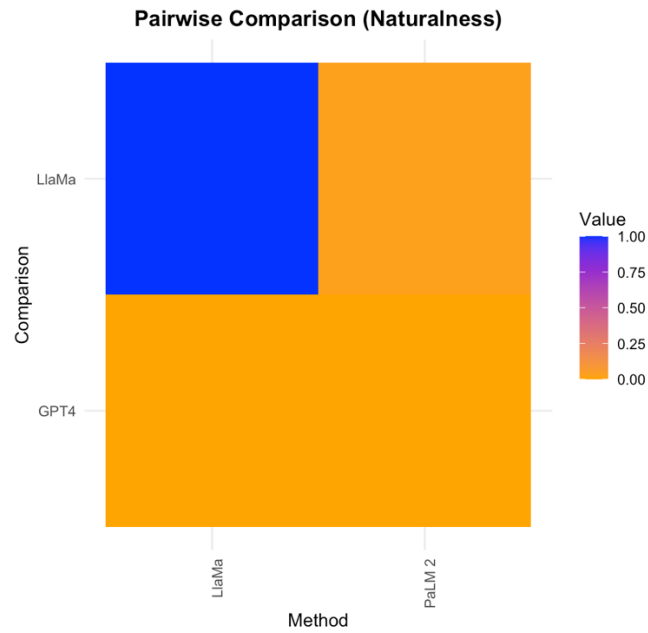


Figure 8. Comparison of naturalness shown across models

4.4. Hallucinations

Models varied when analyzing the types of hallucinations shown. This study analyzed the types of hallucination prominent in each of the models. This study sought to have an insight into the cross-lingual understanding based on the amount and type of hallucinations shown by each of the models. Results are shown in Figure 9.

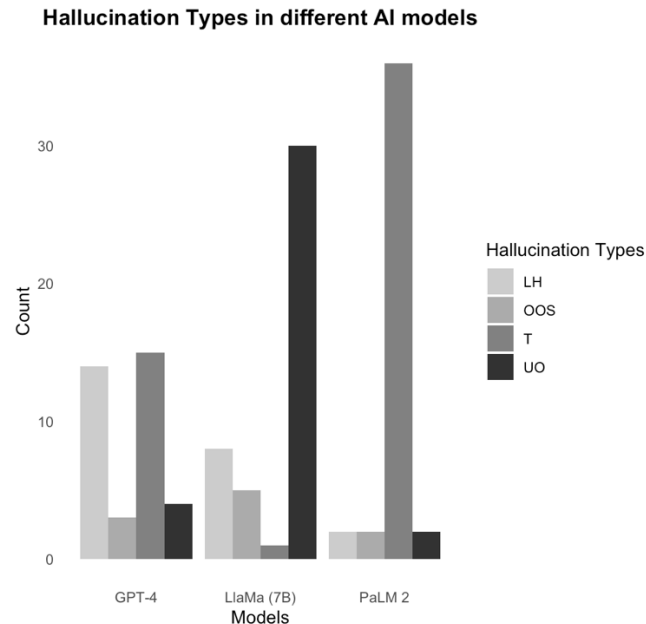


Figure 9. Types of hallucination by model. The study differentiates linguistic hallucination (LH), out-of-switch (OOS), translations (T), and unexpected output (UO). These are defined in Evaluation (3.2)

GPT-4 showed a tendency to produce hallucinations that mix the languages prompted (Basque and English) with others that keep a geographical and political relation (Spanish and French). It was observed that for prompts such as ‘Imitate the speaking style of a person who can speak English and Basque in one sentence about sports’, GPT-4 provided answers such as ‘*kirolak* (Basque) are really great for *gure osasuna* (Basque), and they help build *camaradería* (Spanish)’. Furthermore, it was observed that the model amplified some stereotypes through the Basque cuisine (e.g., excellent idea! *Sagardoa eta pintxos*), or that it mixed it with typical Spanish dishes (e.g., this *paella* is delicious, *oso gozoa da!*).

In GPT-4 linguistic hallucinations were also prominent. Many of them were vocabulary related, such as incorrect lexicon (e.g., *Gaurko eguraldia oso beroa da*, I think I will go for a **txotxongilo / izozki!*). Still, the most interesting linguistic hallucination was that GPT-4 generated odd syntactic constructions that merged Basque syntax and the English lexicon. Basque is characterized by having the determiner connected to the noun: the *-a* in *etxea* is equivalent to the English determiner *the*. GPT-4 generated instances such as *familya*, where it used the Basque determiner *-a* with the English word *family*.

LlaMa (7B) showed the highest number of hallucinations. Of those, unexpected outputs were the most significant. Most of the time, LlaMa was not capable of generating coherent responses that answered the prompts. When able to produce a switch, while the English text was correct, the part in Basque was rarely grammatical or coherent (e.g.,[?] **Euskara eta engelialan irektatu gaurken duzuko diren ondoren dira*. Language is a mixture of Spanish and[?] **Euskarian*.). This pointed out in the direction of a lack of sufficient training in low-resource languages such as Basque. It has to be said that LlaMa (7B) is a lightweight model; its lack of parameters could be one of the reasons to its poor performance.

PaLM 2 showed a tendency to produce switches that were direct translations of parts that were mentioned previously in the same sentence (e.g., *Hau da nire kirol gogokoena: eskubaloia*. This is my favorite sport: handball.). These cases were not considered code-switching in this study. Most of the translation hallucinations followed a similar scheme, where the translated part was between parenthesis or after a comma.

5. DISCUSSION

The evaluation of the Basque and English code-switching capabilities of GPT-4, LlaMa (7B), and PaLM 2 has shed light on the significant disparities among these models, particularly in terms of language-specific adaptations and their capability to generate answers in low-resourced languages. The experiments underline the broad-spectrum challenges of deploying large language models (LLMs) and the constraints that come into play, particularly in low-resourced languages and code-mixing scenarios.

The study shows that multilingual training does not necessarily lead to code-switching ability. The inability to create code-switched outputs of models such as PaLM 2 shows that, in many multilingual models, code-mixing is not recognized as an essential component. In fact, both PaLM 2 and LlaMa (7B) seem to misunderstand code-switching, interpreting it as simply a translation or random combination of two or more languages. Researchers must take code-mixing into account as an integral part of many linguistic repertoires around the world. By building LLMs that include code-mixing, NLP researchers can capture the dynamic elements of many languages more accurately while also helping to improve users’ understanding of cultural aspects embedded in dialogues. Through these considerations, new models will be able to provide a true multilingualism that goes beyond syntax and semantics.

A crucial theme that emerged during the analysis was the need for increased transparency in machine learning research. Specifically, this study emphasizes the need for a thorough and understandable disclosure of the methodologies used to train these models. A comprehensive knowledge of the learning strategies employed by these LLMs, including their dataset and hyperparameter details, would significantly aid researchers in discerning the reasons behind the performance disparities observed. This study shows that GPT-4 outperforms its direct opponent, PaLM 2, and the lightweight model, LLaMa (7B). However, in the absence of transparency, we cannot confidently identify why the models show these performances.

6. CONCLUSION

This study found that GPT-4 was the model that best reflected cross-lingual understanding. It was capable of combining two languages through natural code-mixing utterances, to the point of understanding the grammar of Basque. This was shown in the addition of the Basque determiner *-a* to the English word *family* (*family[a]*).

A salient observation drawn from this study was the unique positioning of GPT-4 in the realm of low-resourced text generation. GPT's adaptability to work with low-resourced languages, such as Basque, was distinctly superior to PaLM 2 and LLaMa (7B). It is hypothesized that this observation could be attributed to its built-in design or its training process. While PaLM 2 and LLaMa (7B) showed reasonable performances in data-rich languages like English, their performance dwindled in the context of Basque, emphasizing the intrinsic issues when it comes to generating low-resource language text.

Low-resource languages are still left behind in natural language processing (NLP). The difficulties of high-parameter models such as PaLM 2 reflect the lack of attention on LRLs. Research on how to improve low-resource data augmentation could bring significant benefits to LLMs to be deployed.

7. LIMITATIONS

7.1. Data Availability

This study provided the first Basque-English AI-generated code-switched dataset. This study showed statistical significance in the experiments performed. However, future studies focusing on Basque code-switching may benefit from increasing the data available.

7.2. Annotators

This study only used a single Basque native annotator, speaker of the neutral and Biscayan variants. No other candidate was proficient in the targeted low-resource language. Future studies may benefit from the annotation of several native speakers to further validate results.

REFERENCES

- [1] J. Paradis, F. Genesee, M.B. Crago, "Dual language development and disorders," A handbook on bilingualism and second language learning, vol. 2, 2011.
- [2] W.Q. Yow, J. Tan, S. Flynn, "Code-switching as a marker of linguistic competence in bilingual children," Bilingualism: Language and Cognition, vol. 21, no. 5, pp. 1075-1090, 2018.
- [3] J. MacSwan, K.T. McAlister, "Naturalistic and elicited data in grammatical studies of codeswitching," Studies in Hispanic and Lusophone linguistics, vol. 3, no. 2, pp. 521--532, 2010.
- [4] J. MacSwan, A minimalist approach to intrasentential code switching. Routledge, 2014.

- [5] K. Colmenero, D. Lasagabaster, "Native and non-native teachers in a minority language: An analysis of stakeholders' opinions," *International Journal of Bilingualism*, 13670069221149671, 2023.
- [6] OpenAI, GPT-4 Technical Report, 2023.
- [7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, L LaMA: Open and Efficient Foundation Language Models, 2023.
- [8] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos... and Y. Wu, PaLM 2 Technical Report, 2023.
- [9] Zhao, Wayne Xin, Zhou, Kun, Li, Junyi, Tang, Tianyi, Wang, Xiaolei, Hou, Yupeng, Min, Yingqian, Zhang, Beichen, Zhang, Junjie, Dong, Zican, others, "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.
- [10] G.I. Winata, Madotto, C.S. Wu and P. Fung, Code-Switching Language Modeling using Syntax-Aware Multi-Task Learning, 2018.
- [11] H. Gonen and Y. Goldberg, Language Modeling for Code-Switching: Evaluation, Integration of Monolingual Data, and Discriminative Training, 2019.
- [12] H. Adel and N. Vu, Thang and Schultz, Tanja, Combination of recurrent neural networks and factored language models for code-switching language modeling, 2013.
- [13] G.I. Winata, Madotto, C.S. Wu and P. Fung, Code-Switched Language Models Using Neural Based Synthetic Data from Parallel Sentences, 2019.
- [14] K. Li, J. Li, G. Ye, R. Zhao, Y. Gong, Towards code-switching ASR for end-to-end CTC models, 2019.
- [15] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, others, "Multilingual and code-switching ASR challenges for low resource Indian languages," arXiv preprint arXiv:2104.00235, 2021.
- [16] Z. Liang, Z. Song, Z. Ma, C. Du, K. Yu, X. Chen, "Improving Code-Switching and Named Entity Recognition in ASR with Speech Editing based Data Augmentation," arXiv preprint arXiv:2306.08588, 2023.
- [17] Z.X. Yong, R. Zhang, J. Zosa Forde, S. Wang, S. Cahyawijaya, H. Lovenia, G.I. Winata, L. Sutawika, J.C. Blaise Cruz, L. Phan, others, "Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages," arXiv e-prints, arXiv:2303, 2023.