Do LLMs exhibit the same commonsense capabilities across languages?

Anonymous ACL submission

Abstract

This paper explores the multilingual commonsense generation abilities of Large Language Models (LLMs). To facilitate this investigation, we introduce MULTICOM, a novel benchmark that extends the COCOTEROS dataset to four languages: English, Spanish, Dutch, and Valencian. The task involves generating a commonsensical sentence that includes a given triplet of words. We evaluate a range of open-source LLMs with a variety of methods, including automatic metrics, human and LLM-as-judge evaluations. Results consistently show superior performance in English, with significantly lower performance in less-resourced languages. While contextual support yields mixed results, it tends to benefit under-represented languages. These findings underscore the current limitations of LLMs in multilingual commonsense generation.

1 Introduction

Commonsense reasoning is a fundamental human ability that allows individuals to make inferences without needing everything to be explicitly stated (Do et al., 2024). For example, people intuitively know that a dog can bark but not meow, or that a pencil is typically used for writing or drawing. Transferring this type of cognitive skill to Artificial Intelligence (AI) systems is challenging, as it involves enabling models to make inferences beyond what is directly expressed. Large Language Models (LLMs) have become the state of the art in Natural Language Generation (NLG) due to their training on vast amounts of text that capture a wide range of world knowledge. As a result, they are able to exhibit at least a superficial level of commonsense, thanks to the information encoded in their parameters. While LLMs are trained on multilingual data, the majority of that data is in English. This imbalance can lead to disparities in performance across languages (Blasi et al., 2022; Xu et al., 2024; Qin et al., 2025). In this paper, we investigate whether LLMs demonstrate the same level of ability to generate text which conforms to commonsense intuitions across different languages. To explore this topic, we use a constrained commonsense generation task, using MULTICOM, a novel dataset inspired by the COCOTEROS benchmark (Maestre et al., 2024): given a triplet of concepts, a model must generate a coherent and plausible sentence that connects them. Additionally, we examine whether providing a supporting context in addition to the concept triple improves the model's ability to produce meaningful outputs. We evaluate the outcomes across four different languages using automatic metrics, LLM-as-judge evaluations, and human judgments. Our results show a significant disparity across different languages in the extent to which LLMs are able to generate texts compatible with commonsense.

In summary, our contributions are as follows: (i) We present MULTICOM, a novel multilingual benchmark dataset (English, Dutch, Spanish and Valencian) for commonsense generation; (ii) We introduce and study the role of supporting linguistic context as a way to assist the model during generation to produce more accurate texts; (iii) We perform a comprehensive evaluation of commonsense generation capabilities across five LLM families, varying in both size and configuration.

2 Related Work

Commonsense Generation Commonsense knowledge refers to the information that people typically infer in everyday situations, even when it is not explicitly stated (Bhargava and Ng, 2022). Incorporating this type of knowledge into NLG systems has been recognised as both important and challenging (Wang et al., 2021), motivating several benchmarks and shared tasks. The Avicenna task (Aghahadi and Talebpour, 2022) focuses on

generating a conclusion that logically connects two sentences with a syllogistic relationship. Winograd Schemas test models' ability to resolve referential ambiguities which require commonense reasoning to arrive at the most likely interpretation in a given context (Levesque et al., 2012). Other research has concentrated on more constrained generation tasks, such as keyword-to-text. For instance, SituatedGen (Zhang and Wan, 2024) involves generating a pair of contrasting sentences based on a set of concepts, including temporal or geographical entities. Similarly, CommonGen (Lin et al., 2020) and C^2Gen (Carlsson et al., 2022) require generating a coherent and plausible sentence that describes an everyday scenario using a given set of keywords. C^2Gen also incorporates contextual information to improve the relevance and accuracy of the generated text. Most of this work has focused on English. One exception is COCOTEROS (Maestre et al., 2024), which extends the principles underlying CommonGen and C^2Gen to Spanish. The present paper takes this line of work further, developing a multilingual dataset based on COCOTEROS and evaluating the multilingual commonsense generation abilities of several LLMs.

Multilinguality in LLMs Because they are trained on massive datasets in multiple languages, LLMs often acquire multilingual capabilities. Nevertheless, recent research suggests that LLM capabilities are not uniform across languages and many languages remain under-represented in training data (Blasi et al., 2022; Xu et al., 2024; Qin et al., 2025). This impacts multilingual performance. For example, Zhang et al. (2023) examine ChatGPT's performance on three NLP tasks and find that it shows a strong bias toward English. Xuan et al. (2025) introduce MMLU-ProX, a multilingual benchmark derived from English, and report significant performance disparities across languages. Thellmann et al. (2024) propose a benchmark based on translated datasets covering several European languages. In the specific case of constrained, commonsense generation tasks, there has been very limited cross-lingual evaluation, in part due to the paucity of resources. This motivates our development of MULTICOM benchmark.

3 The MULTICOM benchmark

MULTICOM is conceived as a new benchmark to test the capabilities of generative models in commonsense generation tasks. Similarly to CommonGen (Lin et al., 2020), we frame the task in terms of generating text from a triple of keywords or concepts. However, MULTICOM further incorporates contextual support, and covers multiple languages (English, Spanish, Dutch and Valencian), thus including low-resourced languages.

As our starting point, we use the COCOTEROS dataset (Maestre et al., 2024), which was specifically created to assess commonsense generation in Spanish. Each instance in the COCOTEROS dataset includes three components:

- **Keywords:** a set of three words that should appear in a generated sentence.
- **Target sentence:** a grammatically correct and commonsense-consistent sentence that includes all three keywords or suitable morphological variations.
- **Context:** a short passage aligned with the target sentence that provides semantic or thematic background, helping to naturally incorporate the keywords. This supports the inclusion of an experimental condition to test the impact of contextual support on commonsense generation.

MULTICOM extends COCOTEROS by adding two new elements to each sample:

- **Counterfactual sentence:** a sentence that includes all the keywords and is grammatically correct and fluent in Spanish, but violates commonsense. Counterfactuals were generated using the LLaMA 3.2 model (Grattafiori et al., 2024), with prompts designed to ensure the inclusion of all keywords, fluency, and the intentional absence of commonsense reasoning. All generated sentences were manually reviewed by the authors to confirm they matched these criteria.
- **Unrelated sentence:** a sentence that is grammatically and semantically correct but does not contain any of the keywords for the given instance. These were created by randomly selecting target sentences from other examples in the dataset, making sure there was no overlap with the current instance's keywords.

Following the augmentation of the Spanish data with the new sentence types, we further translated it into three additional languages:

- **English:** We included English based on the hypothesis that LLMs tend to perform better at generating commonsense text in English compared to other languages. We translated the COCOTEROS dataset into English using the OPUS-MT-ES-EN machine translation model (Tiedemann and Thottingal, 2020)[1].

- **Dutch:** Dutch was selected as a medium-resource language, which is somewhat less represented in NLP than English or Spanish. The Dutch version was produced by translating the English subset using the OPUS-MT-EN-NL model (Tiedemann and Thottingal, 2020).[2]

- **Valencian:** To explore LLM performance in a low-resource setting, we included Valencian, a regional variety of Catalan spoken in Spain's Valencian Community. The dataset was translated from Spanish to Valencian using the Salt translation tool provided by the regional government of Valencia (Valenciana, 2025).

After completing the translations, we used the Grok model (Grok) to verify that all target sentences in each language contained the intended keywords, and, if they did not, to align and correct the keyword usage accordingly. This step was necessary due to potential polysemy and translation variation, where a keyword might be rendered differently in different contexts. The prompt we used within this step is shown in Appendix A.

MULTICOM consists of a training set of 3875 unique input triples and corresponding context sentences, with target, counterfactual and unrelated sentences in 4 languages (total: $15,500$); and a test set with $3,876$ instances (969 inputs in four languages). The inclusion of the three different types of sentences allows us to conduct a detailed analysis of the commonsense capabilities of NLG models, based on the hypothesis that generated sentences should be more similar to the target sentences than to the counterfactual or unrelated ones. Examples from MULTICOM are shown in Table 2 of Appendix B.

4 Experimental setup

We test models by prompting them with the input triple from a MULTICOM test instance, together with an instruction to generate a sentence using all the words in the triple. We compare the generated output to the reference (commonsense-preserving) sentence, as well as to the counterfactual (non-commonsensical) sentence in MULTICOM. We further compare two different settings: (a) With a supporting context sentence included in the prompt (**+Con** condition; cf. Section 3); (b) with no supporting context. The rationale for these conditions is as follows: our evaluation method relies on a comparison between a generated output and a reference (resp. counterfactual) sentence. It is, however, possible that one or more words in an input triple are ambiguous or polysemous, raising the possibility that an acceptable, commonsensical output nevertheless deviates from the reference. Adding a supporting context allows us to guide the model's interpretation of triples. We therefore expect models to perform better in the +Con condition. A further goal is to evaluate to what extent the benefits of further contextual support differ across target languages. See Appendix E for details of the prompt templates used.

4.1 Models

We evaluated the commonsense generation capabilities of LLMs across two primary dimensions: model size and instruction tuning (i.e., base vs. instructed versions). To this end, we selected five recent LLM families: LLaMA (Grattafiori et al., 2024), Qwen (Team, 2025), Gemma (Team et al., 2024), EuroLLM (Martins et al., 2025), and Salamandra (Gonzalez-Agirre et al., 2025). LLaMA, Qwen and Gemma are widely-used and highly performant LLMs. The rationale for including EuroLLM and Salamandra is their additional focus on multilingual capabilities, through training data curation: EuroLLM is trained on data from all the official languages of the EU, including Spanish and Dutch. Salamandra was similarly developed with a focus on multilinguality during the data curation process, and further emphasises the co-official languages in Spain, including Valencian.

For each model family, we chose the most recent available models in small and medium sizes, testing both their base and instruction-tuned variants[3] for each language. The specific model configurations evaluated in MULTICOM benchmark are as follows:

- **LLaMA**: LLaMA-3.2 in 1B and 3B sizes, includ-

[1]Model available at https://huggingface.co/Helsinki-NLP/opus-mt-es-en

[2]Model available at https://huggingface.co/Helsinki-NLP/opus-mt-en-nl

[3]Due to compatibility issues, we used version LLaMA 3.2 for the LLaMA family.

ing both base and instruct versions (1B-Instruct and 3B-Instruct).

- **Gemma**: Gemma-2 3B, tested in both base and instruct variants. In our experiments, we also included Gemmma-9B. However, we observed a strong tendency for the model to misunderstand the task, frequently generating outputs in the form of (pseudo) code. We therefore excluded Gemma-9B from further consideration.
- **Qwen**: Qwen3 in 4B and 8B sizes, with corresponding base and instruct versions.
- **Salamandra**: Models in 3B and 7B sizes, tested in both base and instructed forms.
- **EuroLLM**: EuroLLM models in 1.7B and 9B sizes, evaluated in both base and instructed versions.

Experiments ran on one NVIDIA A100 GPU, averaging 30 minutes.

4.2 Evaluation methods

The evaluation was conducted using three complementary approaches. First, we applied widely-used automatic evaluation metrics to assess model performance quantitatively. Second, we employed LLM-as-a-judge to provide ratings with explicit feedback. Third, a human evaluation was performed on a subset of the dataset to verify the alignment between automated assessments and human judgments.

Automatic metrics We compared the generated outputs against both the target and counterfactual sentences in the test set using the following metrics:

- **BERTScore (Zhang et al., 2020)**: This metric assesses the similarity between a candidate and reference sentence using contextual embeddings from BERT. Unlike traditional approaches, BERTScore captures semantic similarity by computing token-level cosine similarity in the embedding space. To compute BERTScore across all languages in this study, we used the *bert-base-multilingual-cased* model, as its training data includes English, Dutch and Spanish, as well as Catalan, of which Valencian is a variety.
- **Universal Sentence Encoder (USE) (Cer et al., 2018) and Cosine Similarity (Steck et al., 2024)**: To measure semantic similarity, we encode each sentence using the Universal Sentence Encoder to obtain vector representations. Cosine similarity is then computed between these vectors to assess their semantic closeness.
- **Dependency Parsing (Jurafsky and Martin, 2009) and Levenshtein Distance (Miller et al., 2009)**: To evaluate syntactic similarity, we extract dependency triplets (head, relation, dependent) from each sentence using the Stanza (Qi et al., 2020) pipeline and compare them using symbolic Levenshtein distance.[4] Each dependency triple captures a syntactic relation between a head and a dependent word, labelled with the type of grammatical relation. Then, we compare them using symbolic Levenshtein distance. For each triplet, we compute the average distance across its components and consider it a match if this average falls below a predefined threshold. The final score reflects the proportion of matched triplets relative to the larger set.
- **Dependency Parsing (Jurafsky and Martin, 2009) with Vector Representations and Cosine Similarity (Steck et al., 2024)**: We also assess semantic similarity at the syntactic level. Dependency triples are extracted using the Stanza dependency parser (Qi et al., 2020), and each triple, comprising a head, relation, and dependent, is converted into a vector by averaging the embeddings of its components. For this conversion, we used SpaCy.[5] Cosine similarity is then computed between vectors from different sentences, and a match is recorded when the similarity exceeds a predefined threshold. The final score is calculated as the proportion of matched triplets relative to the larger set, capturing both semantic content and syntactic structure.

While the first two metrics above focus on semantic comparison, the metrics relying on dependency triples provide an indication of the extent to which a generated and a reference sentence incorporate the concepts in the input triple in the same or similar syntactic dependency relations.

The idea is that when we evaluate generated sentences using these metrics, better sentences should

[4]We process sentences using the Stanza pipeline configured with the *"tokenize, PoS, lemma, and depparse"* processors. This ensures consistent preprocessing across languages and enables the extraction of universal dependency representations.

[5]The models used were: *"en_core_web_md"* for English, *"es_core_news_md"* for Spanish, *"nl_core_news_md"* for Dutch, and *"ca_core_news_md"* for Valencian, due to its close typological proximity to Catalan.

resemble the reference, commonsense sentence more closely than they resemble the counterfactual sentence. As a result, the score difference when comparing to the reference and when comparing to the counterfactual should be larger for better outputs, and smaller for worse ones.

LLM-as-Judge To complement the automatic metrics, we also employed two LLM-as-a-judge models to assess the commonsense quality of the generated sentences:

- **Prometheus-V2.0** (Kim et al., 2024): Prometheus is an LLM specifically designed for evaluating the outputs of other models through both direct scoring and pairwise ranking. It demonstrates strong performance across evaluation formats, approaching the reliability of proprietary models like GPT-4. Prometheus 2 is trained on the preference collection, a comprehensive dataset containing over 1,000 fine-grained evaluation criteria. These criteria go beyond general categories such as helpfulness or harmlessness to include specific skills like commonsense reasoning, making it particularly well-suited for our task. We used Prometheus to assign a score from 1 to 5 based on the rubric shown in Appendix C.

- **JudgeLM (Zhu et al., 2025)**: JudgeLM is a fine-tuned LLM trained on GPT-4-labelled data to efficiently and reliably evaluate the outputs of other language models, with measures taken during finetuning to support a variety of evaluation formats, and to overcome known biases in LLM-based judgments, such as an excessive reliance on the order in which items are presented. JudgeLM has shown over 90% agreement with GPT-4, making it a strong proxy evaluator. In our case, JudgeLM provided a score from 0 to 10, based solely on the level of commonsense reasoning demonstrated by each generated sentence. This choice of scale mirrored the scale used for direct assessment during the finetuning process of the model (Zhu et al., 2025). The final prompt used to evaluate the sentences can be seen on Appendix D. To facilitate comparison to Prometheus-2, we normalised the scores returned by JudgeLM to a scale from 1 to 5.

Human evaluation For the human evaluation, we selected a random subset of 20 instances from the test set per language, and collected annotations using the Prolific platform[6] (Prolific). Participants were paid at an hourly rate of £9.20, as recommended by the platform. The median time of completion of the study was 15 minutes. For each of the target languages in MULTICOM, three native or fluent speakers were selected via the platform.[7] Annotators were asked to rate the commonsensical quality of generated texts, given the input triple. The instructions mirrored the evaluation rubric used for Prometheus-2 (see Appendix C).

5 Results

In this section, we present the results obtained with the different evaluation methods.

5.1 Automatic Metrics

We conducted automatic evaluations using the metrics described in Section 4.2 across all models. However, due to the extensive volume of results, this section presents only the outcomes for a single model, LLaMA-3.2-3B, which consistently demonstrated strong and stable performance. These results are shown in Table 1. Comprehensive results for all evaluated models are provided in Appendix F for LLaMA, Appendix G for Qwen, Appendix H for EuroLLM, Appendix I for Salamandra, and Appendix J for Gemma.

For each model and language, we assessed two types of generated sentences: one without context injection, and another with context injected into the prompt to provide a more constrained input. These generated sentences were then evaluated against both the target sentence and the counterfactual sentence. The primary objective was to compare the difference in scores for generated sentences when compared against the target and counterfactual. A larger gap implies that the generated sentence aligns more closely with the reference than with the counterfactual, providing some evidence of the degree of commonsense generation abilities in a given language, based on similarity to a reference.

Consistent with the results observed for LLaMA (Table 1), all models generally follow the same trend: the differences in similarity scores, when comparing generated outputs to both target and counterfactual sentences, are consistently larger in

[6] Platform available at https://www.prolific.com/

[7] No speakers of Valencian are available on Prolific. However, for the purposes of this study, we selected speakers of Catalan, since Valencian is a variety of Catalan. Nevertheless, we urge caution in the interpretation of results for Valencian.

	ES		EN		CA-VA		NL	
	Ref. -Con (Diff Count.)	Ref. +Con (Diff Count.)	Ref. -Con (Diff Count.)	Ref. +Con (Diff Count.)	Ref. -Con (Diff Count.)	Ref. +Con (Diff Count.)	Ref. -Con (Diff Count.)	Ref. +Con (Diff Count.)
BERTScore | Base	0.771 (0.022)	0.754 (0.013)	0.903 (0.024)	0.902 (0.022)	0.759 (0.034)	0.805 (0.087)	0.761 (0.045)	0.742 (0.028)
BERTScore | Instruct	0.777 (0.023)	0.743 (0.016)	0.908 (0.027)	0.902 (0.026)	0.765 (0.04)	0.756 (0.056)	0.771 (0.049)	0.746 (0.039)
USE | Base	0.632 (0.036)	0.604 (-0.004)	0.555 (0.194)	0.516 (0.126)	0.420 (0.098)	0.533 (0.19)	0.637 (0.038)	0.641 (-0.002)
USE | Instruct	0.650 (0.031)	0.549 (0.011)	0.558 (0.194)	0.509 (0.149)	0.440 (0.093)	0.422 (0.118)	0.672 (0.04)	0.639 (0.016)
Dep.Par + Lev. | Base	0.666 (0.117)	0.535 (-0.013)	0.705 (0.212)	0.598 (0.091)	0.671 (0.155)	0.684 (0.135)	0.596 (0.184)	0.478 (0.063)
Dep.Par + Lev. | Instruct	0.680 (0.103)	0.508 (0.006)	0.718 (0.193)	0.621 (0.115)	0.693 (0.139)	0.576 (0.077)	0.621 (0.185)	0.492 (0.087)
Dep.Par + Cos.Sim | Base	0.472 (0.072)	0.423 (-0.005)	0.529 (0.181)	0.496 (0.123)	0.630 (0.097)	0.673 (0.098)	0.448 (0.125)	0.387 (0.051)
Dep.Par + Cos.Sim | Instruct	0.492 (0.064)	0.399 (0.016)	0.556 (0.187)	0.498 (0.129)	0.658 (0.087)	0.552 (0.051)	0.472 (0.136)	0.402 (0.084)

Table 1: Results obtained for the LLaMA-3.2-3B models evaluated with automatic metrics in English (EN), Spanish (ES), Valencian (CA-VA), and Dutch (NL). Each system output is compared against the reference (Ref.) and how the score differs regarding to the evaluation against counterfactual (Diff Count.) sentences, considering two generation settings: with context injection (+Con) and without context injection (-Con). Results for the base and instruction-tuned models are reported in separate rows.

English than in other languages. The only exception to this pattern is BERTScore, for which the difference in score between reference and counterfactual in English is comparable to, or slightly smaller than, that observed in other languages. We attribute this to the fact that although models typically perform better in English, BERTScore may not effectively distinguish between commonsensical and non-commonsensical outputs, as it primarily captures token-level semantic similarity rather than deeper commonsense reasoning.

The impact of contextual information on generation is mixed across models and languages. While certain models – particularly LLaMA – demonstrate improved performance when context is provided, especially for under-represented languages such as Valencian and Dutch, other model families, such as EuroLLM (Figure 6 in Appendix H), show only modest gains across all languages. In contrast, for some models like the Gemma family (Figure 8 in Appendix J), the inclusion of context appears to reduce performance.

5.2 LLMs as Judges

For the LLMs-as-a-judge evaluation (Figures 1 and 2), we assessed the results of the instruction-tuned versions of each tested model family/size combination.

Differences among languages Consistent with our observations on automatic metrics, the results for LLM-as-judge evaluation suggest a disparity among languages, but this depends on the target models. Broadly speaking, we see a division between models for which the score distribution varies substantially (LLaMA, Qwen, Gemma) across languages, with better scores for English; and models for which the distribution is less obviously skewed towards English (EuroLLM, Salamandra). We attribute this to the nature of the pretraining data for the latter models, where greater attention was paid to linguistic diversity. For LLaMA and Qwen models, which are among the best-performing on the automatic metric evaluation, output in English is judged as superior by both models, albeit with some variation between the two LLM judges on the impact of context (see below). Surprisingly, JudgeLM scores outputs in Valencian in the +Con condition on a par with English. On the other hand, the larger version of Salamandra obtains a more even score distribution across languages with both LLM judges.

Impact of context sentence The injection of a supporting context sentence does not appear to exert a significant impact on results, with two exceptions: JudgeLM scores LLaMA-3B outputs for English, and Qwen-4B outputs for Valencian, significantly higher when context is included. Nevertheless, this tendency is not consistent across the two LLM judges.

Effect of model size Larger models can be seen to perform somewhat better than smaller ones, though not by a wide margin. However, larger models do tend to be judged better when generating in under-resourced languages like Valencian, with the possible exception of EuroLLM.

LLM judge disparities Beyond the observations made above, it is worth noting that the two LLM-as-judge models display some inconsistencies even for the same model configurations and languages, suggesting that model-based judgments may depend on the specific fine-tuning regime employed for a specific judge model.

5.3 Human Evaluation

For the human evaluation, we focused on responses generated by the LLaMA-3.2-3B-Instruct model, as it showed strong performance across the automatic metrics in our experiments. In this case, we evaluated only one model across multiple lan-

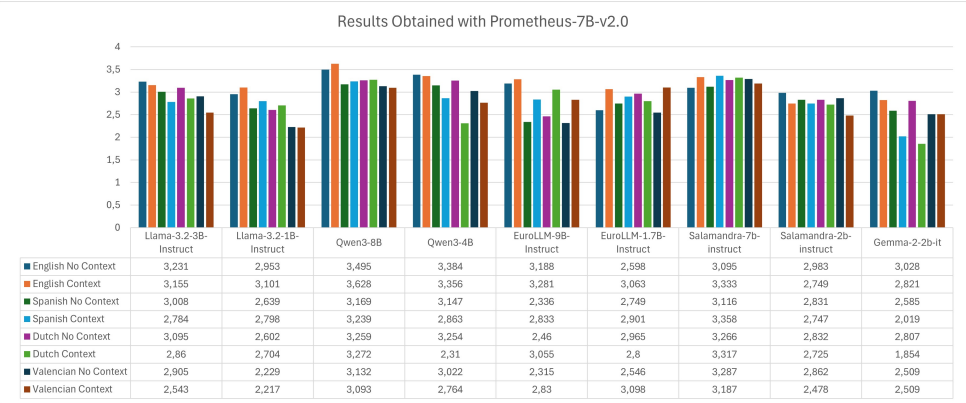


Figure 1: LLM-as-Judge results with Prometheus-2

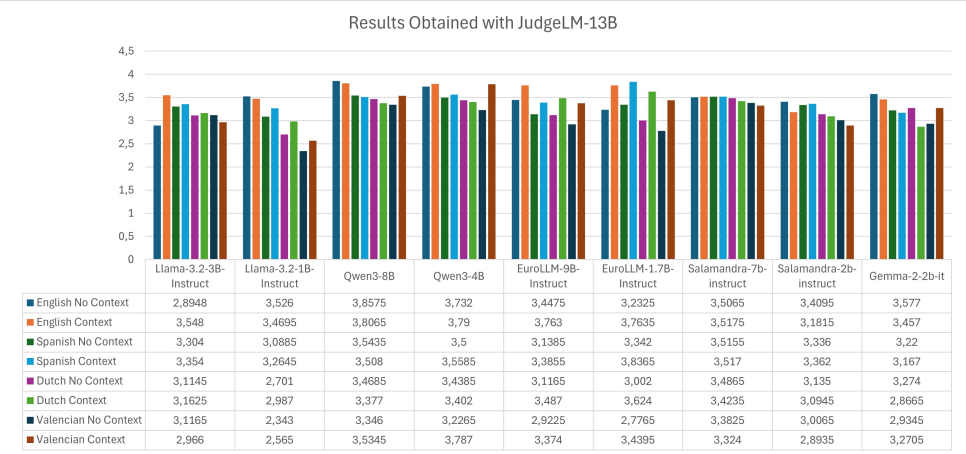


Figure 2: LLM-as-Judge results with JudgeLM, normalised to a 1-5 scale

guages without context. In what follows, we compare the results of the human judgments with those of Prometheus-2 and JudgeLM, for the same sample of 20 instances in the MULTICOM test set.

To assess consistency among the three annotators, we calculated the majority agreement percentage, which reflects the proportion of items where at least two annotators assigned the same label. Given the inherently subjective nature of commonsense evaluation, achieving perfect consensus among all annotators is often unrealistic. Therefore, majority agreement serves as a practical and meaningful measure of reliability in this context. The observed scores -— 0.75 for English and Dutch, 0.80 for Spanish, and 0.95 for Valencian -— indicate a generally strong level of agreement. These values suggest that despite some variability in individual judgments, the annotators largely concur on the commonsense quality of the responses, supporting the validity of the human evaluation results.

The results of the human annotation process are shown in Figure 3. Based on the results from the human evaluation, we confirm that human judgments generally align with those of the LLM-as-a-judge evaluators. In particular, the English-generated sentences tend to demonstrate stronger commonsense reasoning compared to other languages. Interestingly, although Spanish is generally considered the second-best language, Prometheus shows a slight preference for Dutch, which it ranks above Spanish as the third-best language overall. In contrast, the other evaluations consistently rank Spanish as the

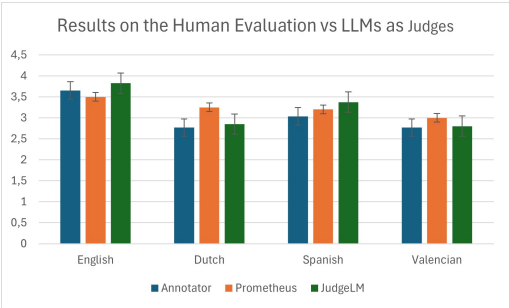


Figure 3: Results obtained for the tested models in the human evaluation.

second-best language, while Valencian and Dutch yield comparable results. This discrepancy may be due to Prometheus having been trained on very limited data in Dutch, which could impair its ability to accurately evaluate texts in that language. We further note that the annotators for Valencian are native speakers of Catalan; hence their knowledge of Valencian may vary, since it is a dialect of Catalan spoken primarily in specific regions of Spain.

6 Discussion and Conclusions

This study provides key insights into LLMs' multilingual commonsense generation abilities using the MULTICOM benchmark.

English commonsense predominance Our findings indicate a clear imbalance in LLMs' ability to generate commonsensical sentences across languages. Although these models are trained on multilingual data and can produce text in several languages, they demonstrate a significantly stronger performance in English. This is because, while LLMs tend to generate commonsensical outputs in English, they often struggle to do so in other languages. This finding aligns with the hypothesis of this paper, and was supported through three distinct evaluation approaches: (1) automatic evaluation using standard metrics, (2) LLM-based evaluation with LLMs as judges, and (3) human evaluation on a small subset of data.

Context Disparity During the experimentation, we designed two main setups for commonsense generation. In the first, the models were prompted to generate a commonsensical sentence based on a given triplet of keywords. In the second, we included in the prompt a linguistic context intended to support the model in producing a more plausible output. However, the results obtained from the context-enhanced setup were not as expected. Instead of consistently improving the performance, the inclusion of context resulted in mixed outcomes. While some models seem not to benefit from the context, in a few cases it even had a detrimental impact. In contrast, there were some cases where models demonstrated improved performance when enhanced with the context. This was observed with EuroLLM and Salamandra. This variation suggest that the effectiveness of context injection is model dependent, potentially influenced by the model's training or instruction tuning process. Another important observation concerns language dependence. For under-represented languages such as Valencian and Dutch, contextual information generally had a more positive impact compared to major languages like English and Spanish. This may be because pre-trained models already possess richer representations for high-resource languages, and adding external context could potentially introduce noise rather than improve understanding.

LLMs versus human judges To evaluate the outputs of our experiments, we leveraged the capabilities of LLMs-as-judges to assess a challenging task such as commonsense generation. Additionally, we conducted a human evaluation via the Prolific platform on a subset of generated sentences, aiming to verify the alignment between human judgments and those produced by the automated judges. The results obtained from Prometheus and JudgeLM were largely consistent, with minor variations. Both judges (human and LLMs) supported the conclusions drawn from automatic metrics, English being the language in which models most reliably generated commonsensical sentences. Similarly, both judges consistently assigned the lowest scores to outputs in Valencian, suggesting that LLMs struggled the most in this under-represented language. These trends were further validated by human evaluation. We compared the human-assigned scores with those from Prometheus and JudgeLM on the same set of sentences. The analysis revealed that humans and LLM judges are broadly compatible, particularly in ranking English as the top-performing language by a substantial margin. This alignment reinforces the reliability of LLM-as-a-judge approaches for commonsense evaluation, while also highlighting persistent performance disparities across languages.

Future Work Building on this study's findings, we outline several promising directions for future research to advance multilingual commonsense evaluation and generation. First, future work could further fine-tune an LLM-as-judge model to better align its assessments with the nuanced demands of multilingual commonsense reasoning. Second, while this study focused on open-weight models, comparing their performance with proprietary systems such as GPT-4, Grok, and Gemini would provide valuable context. Finally, improving commonsense reasoning and generation in low-resource languages remains an important challenge—one that could be addressed through continued training and integration of external knowledge sources.

Limitations

This study presents some limitations that may have influenced the evaluation results.

First, while we employed both automatic metrics and LLM-as-a-judge approaches, the automatic evaluation relied on language-specific encoders and parsers. These tools are typically developed and benchmarked primarily on English, which may lead to more favourable performance scores for English texts. Consequently, the evaluation of non-English outputs might not fully reflect their true quality, and could under-represent the commonsense reasoning abilities of models in less-resourced languages.

Second, for the LLM-as-a-judge evaluations, both prompts and scoring rubrics were written in English, even when evaluating outputs in other languages. This decision was made to ensure consistency and comparability across evaluations. Writing prompts in multiple target languages could introduce variability due to differences in how judges understood the prompt, which might inadvertently bias the results. Using a single, English-language prompt ensures that observed differences in model performance across languages are more likely to reflect the model's multilingual capabilities rather than inconsistencies in the prompts themselves. While this approach ensures consistency across evaluations, it may introduce subtle biases, particularly if the judging LLM has varying levels of exposure to the target language. This could affect the model's ability to fully capture the nuances of commonsense reasoning in certain languages, potentially influencing the reliability of cross-linguistic comparisons.

Third, our evaluation was limited to open-source models. While proprietary models such as GPT-4, Grok, or Gemini may exhibit different behavioural patterns or performance characteristics, they were excluded from this study due to the significant costs and access restrictions associated with their use. By focusing on open-source models, we aimed to ensure full transparency, reproducibility, and accessibility of our experimental setup, which are essential for fostering open research and facilitating further comparative studies.

Fourth, for the evaluation of Valencian, we were restricted by the availability of models and participants. Due to the paucity of resources for this language, we relied on models trained on Catalan for some parts of the study (such as the computation of BERTScore), and on Catalan speakers for the human evaluation. While this means that results for Valencian should be treated with caution, we also hope that its inclusion in MULTICOM will motivate further work on this language.

Finally, to our knowledge, the data in MULTICOM does not include any explicitly harmful text. We therefore have no reason to believe that its public release will invite harmful (dual) uses. At the same time, since the data was based on an existing Spanish dataset (COCOTEROS; Maestre et al., 2024), which was initially sourced from human annotators, we cannot exclude the possibility of cultural or linguistic biases having crept into the data creation process.

References

Zeinab Aghahadi and Alireza Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, 32(1):55–71.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.

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.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and 1 others. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Quyet V. Do, Junze Li, Tung-Duong Vuong, Zhaowei Wang, Yangqiu Song, and Xiaojuan Ma. 2024. What really is commonsense knowledge? *CoRR*, abs/2411.03964.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza,

Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grok. 2025. Grok 3 beta — the age of reasoning agents. Technical report, x.ai. [Accessed 01-07-2025].

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

María Miró Maestre, Iván Martínez-Murillo, Elena Lloret, Paloma Moreda, and Armando Suárez Cueto. 2024. Cocoteros: A spanish corpus with contextual knowledge for natural language generation. In *40th Annual Conference of the Spanish Association for Natural Language Processing*, page 2024.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press.

Prolific. 2025. Quality data. from real people. for faster breakthroughs. [Accessed 09-07-2025].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1). Publisher: Elsevier.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards multilingual llm evaluation for european languages. *CoRR*, abs/2410.08928.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Generalitat Valenciana. 2025. Salt.usu. `https://salt.gva.es/es/traductor`. [Accessed 01-07-2025].

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *arXiv preprint*. ArXiv:2404.00929 version: 1.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke

Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *CoRR*, abs/2503.10497.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

Yunxiang Zhang and Xiaojun Wan. 2024. SituatedGen: Incorporating geographical and temporal contexts into generative commonsense reasoning. *Advances in Neural Information Processing Systems*, 36.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

A Keyword-Sentence Alignment with Grok

As explained in Section 3, during the translation of the corpus into other languages, there were instances where the translated keywords were not present in the corresponding target sentences due to word polysemy. To address this issue, we applied a post-processing step using Grok to ensure proper alignment between keywords and sentences throughout the whole dataset. The prompt used for this task is shown below. In this example, we illustrate the alignment process for the Valencian translation; for other languages, we simply replaced "Valencian" with the corresponding language name.

```
I have a TSV file with a column named
    keywords, which contains Valencian
    keywords. These keywords are
    intended to be translations of the
    Spanish keywords in the keywords_es
    column. However, the translations in
     the keywords column are often
    incorrect or do not appear in the
    corresponding reference sentence (
    written in Valencian).

Your task is to revise the keywords
    column so that:
1. Each keyword is a correct Valencian
    translation of the corresponding
    word in the keywords_es column.
2. Each translated keyword must appear
    exactly in the associated reference
    sentence.
3. If a correct translation of a keyword
     is not present in the reference
    sentence, omit it from the final
    list of keywords, and include
    another word from the sentence.
```

B Examples from MULTICOM

Examples from the MULTICOM dataset are provided in Table 2.

C Rubric Used for Prometheus Evaluation

To evaluate the generated outputs using Prometheus as an evaluator, it is essential to carefully define the evaluation rubrics that guide the model in assigning a final score to each sentence. The rubrics used for this evaluation are as follows:

```
RUBRICS = {
"criteria": "Does the model's answer
    demonstrate commonsense capability
    ?",
"score1_description": "No Commonsense.
    The answer is illogical, implausible
    , or nonsensical from a commonsense
    perspective. It may reflect a
    fundamental misunderstanding of how
    the world works.",
"score2_description": "Poor Commonsense.
     The answer is grammatically
    readable but lacks commonsense
    reasoning. It may include
    unrealistic, contradictory, or
    illogical actions or relationships
    .",
"score3_description": "Limited
    Commonsense. The answer contains a
    plausible idea, but it may be vague,
     confusing, or reflect partial
    misunderstandings of how the world
    works. The logic is somewhat
    strained, but the meaning is
    recoverable.",
"score4_description": "Good Commonsense.
     The answer makes sense and reflects
     commonsense understanding, though
    it may be slightly awkward or overly
     simplistic. Any issues present do
    not significantly affect its
    plausibility.",
"score5_description": "Excellent
    Commonsense. The answer demonstrates
     strong commonsense reasoning. It
    reflects how people typically think
    and behave in the real world."
}
```

D Prompt used for JudgeLM Evaluation

The JudgeLM model is trained to process an instruction provided to the generation models, the

Keyword	Context	Target Sentence	Counterfactual Sentence	Unrelated Sentence
learn, piano, dad	They gave me a musical instrument to develop my talent.	My dad encouraged me to learn the piano.	My dad's piano makes me learn to dance on the moon with his plastic shoes.	Please give me something to eat.
open, man, door	The noise of the lock woke the woman. She looked at the door and saw a dark figure walking out the hall.	The door opened and a man came out.	The woman opened the door with an invisible man who danced tango.	I have to remove a lot of files from my computer.
read, book, night	After a long day's work, Maria decided to spend the night at home. She didn't feel like going out, so she made herself a cup of tea and plunged into a new book.	Tonight I'd rather read a book than go out.	Mary's night became a night of reading without a book.	There is so much beauty in ambiguity.
have, negotiation, come	The lack of consensus prevented agreement from being reached.	We had a lot of negotiations but they didn't come to anything.	The negotiation of having a cup of coffee reaches the limit of denial.	Thank you very much for your letter of 7 January.

Table 2: Samples from the MULTICOM dataset. For readability, only English samples are shown, but they are aligned across languages.

corresponding target response, the generated response, and the specific evaluation aspect to focus on, along with the desired scoring template. Based on this framework, we designed the following prompt to guide the evaluation process using JudgeLM:

```
### Instruction:
Your task is to construct a
    commonsensical and grammatically
    correct sentence in {lang}, using
    all of the given words. The sentence
     should sound natural and make
    logical sense within everyday usage.
### Keywords:
{keywords}
### Response 1: {target_sentence}
### Response 2: {generated_sentence}
###Evaluation: Evaluate the two answers
    based on their level of commonsense
    reasoning. Assign a score from 0 to
    10 to each answer, where 0 indicates
     no commonsense reasoning and 10
    indicates excellent commonsense
    reasoning. For each score, include a
     brief explanation justifying your
    evaluation. Format your response
    exactly as shown below, replacing [
    Score] with your assigned score and
    [Explanation] with your
    justification:
### Answer 1: [Score]/10 - [Explanation]
### Answer 2: [Score]/10 - [Explanation]
```

E Prompts for Generating the Sentences

To generate the sentences, we designed a descriptive prompt using a one-shot configuration, in which a single example demonstrated the task to be performed. This prompt was identical across all models and written in English for all target languages, as our preliminary analysis showed that models better understood the task when instructions were provided in English. However, the example within the prompt was written in the target language to guide the model appropriately. For readability, we present the example in English here. Additionally, we used a placeholder lang to indicate the target language for generation. The prompt used in the configuration without context is as follows:

```
Your task is to construct a
    commonsensical and grammatically
    correct sentence in {lang}, using
    all of the given words. The sentence
     should sound natural and make
    logical sense within everyday usage.
     I will provide an example to show
    you what is expected:

### Keywords:
cat, owner, pet
### Sentence:
The owner pets the cat.

Your turn:
### Keywords:
{example["keywords"]}:
### Sentence:
}
```

For the configuration with context injection, the contextual information was incorporated directly into the prompt as follows:

```
Your task is to construct a
    commonsensical and grammatically
    correct sentence in {lang}, using
    all of the given words, and adhering
     to the given context. The sentence
    should sound natural and make
    logical sense within everyday usage.
     I will provide an example to show
    you what is expected:

### Keywords:
cat, owner, pet
### Context:
The person arrived home after work.
### Sentence:
The owner pets the cat.

Your turn:
### Keywords:
{example["keywords"]}
### Context:
{example["context"]}
### Sentence:
}
```

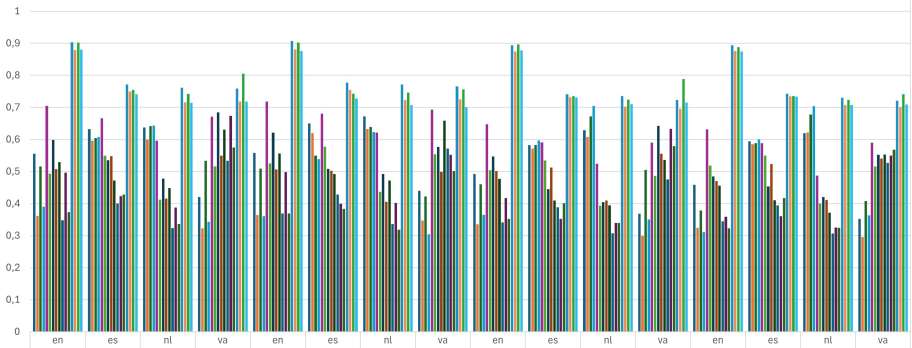
F Results with LLaMA models

Regarding the results obtained for the LLaMA models, we can see them in Figure 4.

The goal is to demonstrate that, for each language, the difference in similarity scores between the generated sentence and the reference sentence versus the generated sentence and the counterfactual sentence (e.g., using the USE evaluation method) is greater in English than in other languages. A larger difference indicates that the generated sentence is more similar to the reference and less similar to the counterfactual, suggesting stronger commonsense reasoning.

We analyse three key aspects:

Differences among languages In the case of LLaMA models, a consistent trend is observed: the difference in similarity scores between the generated sentence and the reference versus the counterfactual is substantially larger in English than in other languages, where this gap is notably smaller. This suggests that the models perform better at commonsense reasoning in English.



	en	es	nl	va	en	es	nl	va	en	es	nl	va	en	es	nl	va
	meta-llama/Llama-3.2-3B				meta-llama/Llama-3.2-3B-Instruct				meta-llama/Llama-3.2-1B				meta-llama/Llama-3.2-1B-Instruct			
USE Reference	0,555	0,632	0,637	0,42	0,558	0,65	0,672	0,44	0,492	0,582	0,629	0,368	0,458	0,594	0,619	0,352
USE Counterfactual	0,361	0,596	0,599	0,322	0,364	0,619	0,632	0,347	0,335	0,571	0,608	0,299	0,324	0,585	0,622	0,295
USE Reference Context	0,516	0,604	0,641	0,533	0,509	0,549	0,639	0,422	0,46	0,582	0,672	0,505	0,378	0,588	0,678	0,408
USE Counterfactual Context	0,39	0,608	0,643	0,343	0,36	0,538	0,623	0,304	0,365	0,597	0,705	0,35	0,311	0,6	0,704	0,363
Levensthein Reference	0,705	0,666	0,596	0,671	0,718	0,68	0,621	0,693	0,647	0,591	0,524	0,59	0,631	0,588	0,487	0,59
Levensthein Counterfactual	0,493	0,549	0,412	0,516	0,525	0,577	0,436	0,554	0,504	0,534	0,393	0,486	0,518	0,549	0,4	0,516
Levensthein Reference Context	0,598	0,535	0,478	0,684	0,621	0,508	0,492	0,576	0,547	0,445	0,404	0,642	0,484	0,453	0,42	0,552
Levensthein Counterfactual Context	0,507	0,548	0,415	0,549	0,506	0,502	0,405	0,499	0,501	0,512	0,409	0,555	0,47	0,523	0,411	0,54
Vector similarity Reference	0,529	0,472	0,448	0,63	0,556	0,492	0,472	0,658	0,477	0,409	0,394	0,536	0,456	0,41	0,371	0,553
Vector similarity Counterfactual	0,348	0,4	0,323	0,533	0,369	0,428	0,336	0,571	0,341	0,388	0,307	0,475	0,344	0,394	0,306	0,527
Vector similarity Reference Context	0,496	0,423	0,387	0,673	0,498	0,399	0,402	0,552	0,417	0,352	0,339	0,633	0,359	0,36	0,325	0,549
Vector similarity Counterfactual Context	0,373	0,428	0,336	0,575	0,369	0,383	0,318	0,501	0,352	0,401	0,339	0,579	0,322	0,417	0,324	0,568
BertScore Reference	0,903	0,771	0,761	0,759	0,907	0,777	0,771	0,765	0,894	0,741	0,735	0,723	0,894	0,743	0,73	0,721
BertScore Counterfactual	0,879	0,749	0,716	0,718	0,881	0,754	0,722	0,725	0,874	0,732	0,702	0,695	0,876	0,734	0,707	0,7
BertScore Reference Context	0,902	0,754	0,742	0,805	0,902	0,743	0,746	0,756	0,896	0,735	0,724	0,788	0,888	0,735	0,723	0,741
BertScore Counterfactual Context	0,88	0,741	0,714	0,718	0,876	0,727	0,707	0,7	0,878	0,731	0,71	0,715	0,874	0,733	0,707	0,709

Figure 4: Results obtained for the LLaMA models with the automatic metrics.

Impact of context sentence The role of context in sentence generation appears ambiguous. For high-resource languages (English and Spanish), results show that adding context does not lead to significant improvements. However, in lower-resource languages (Dutch and Valencian), context seems to help produce more commonsensical outputs. This benefit is particularly evident in the base versions of the models, where outputs generated with context obtain higher similarity scores compared to those without.

Effect of model size Larger models consistently achieve slightly higher scores across all languages. This is expected, as larger models have been trained on more extensive multilingual data, enhancing their ability to generate sentences that align better with commonsense knowledge.

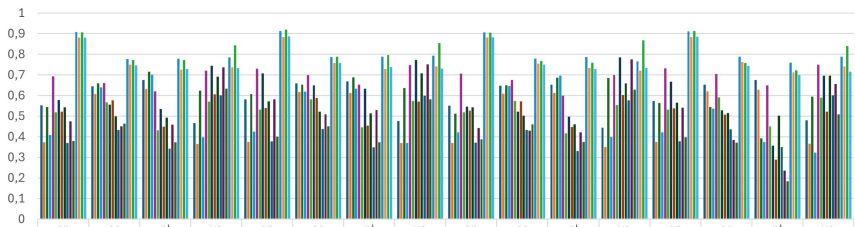
G Results with Qwen models

The results obtained by the Qwen models are shown in Figure 5. These models follow a similar trend to those observed with the LLaMA models.

Differences among languages The largest differences between the reference and counterfactual comparisons are observed in English, indicating stronger commonsense reasoning capabilities in that language. In contrast, the differences are minimal for Spanish and Dutch, suggesting that the generated sentences in these languages exhibit limited commonsense reasoning. Interestingly, Valencian shows a greater difference than Spanish and Dutch. This unexpected result may be attributed to limitations in the sentence encoders used for Valencian, which may fail to accurately capture the nuances of the language. A notable observation for the Qwen models is that, when evaluated using Levenshtein distance, the differences remain relatively high across all languages.

Impact of context sentence Another key finding relates to the impact of contextual information. Overall, providing additional context appears to have a limited positive effect on performance. However, in less-represented languages, such as Valencian and Dutch, the inclusion of context tends to yield better results. This suggests that context injection may help compensate for the models' weaker prior knowledge in these languages, leading to more commonsensical outputs.

Effect of model size With respect to model size, there appears to be no significant impact on the commonsensical quality of the generated sentences. The scores obtained using automatic evaluation metrics are similar across different model sizes, suggesting that increasing model size does not necessarily lead to better commonsense reasoning in this context.



	en	es	nl	va	en	es	nl	va	en	es	nl	va	en	es	nl	va
	Qwen/Qwen3-8B-Base				Qwen/Qwen3-8B				Qwen/Qwen3-4B-Base				Qwen/Qwen3-4B			
USE Reference	0,553	0,645	0,675	0,467	0,581	0,659	0,669	0,476	0,551	0,647	0,653	0,444	0,574	0,653	0,675	0,479
USE Counterfactual	0,373	0,607	0,631	0,365	0,375	0,617	0,613	0,37	0,369	0,609	0,612	0,351	0,375	0,62	0,629	0,367
USE Reference Context	0,544	0,659	0,715	0,623	0,608	0,652	0,689	0,637	0,512	0,649	0,687	0,685	0,564	0,545	0,392	0,595
USE Counterfactual Context	0,409	0,64	0,701	0,398	0,425	0,619	0,634	0,369	0,422	0,646	0,696	0,399	0,422	0,536	0,375	0,323
Levensthein Reference	0,693	0,66	0,62	0,721	0,731	0,699	0,652	0,748	0,706	0,676	0,6	0,7	0,732	0,704	0,649	0,75
Levensthein Counterfactual	0,518	0,567	0,431	0,57	0,532	0,582	0,445	0,574	0,519	0,574	0,416	0,554	0,531	0,592	0,45	0,589
Levensthein Reference Context	0,579	0,556	0,535	0,744	0,708	0,649	0,634	0,773	0,546	0,521	0,498	0,786	0,668	0,528	0,357	0,697
Levensthein Counterfactual Context	0,521	0,576	0,449	0,606	0,539	0,588	0,454	0,57	0,526	0,572	0,447	0,603	0,538	0,507	0,289	0,522
Vector similarity Reference	0,543	0,499	0,492	0,691	0,572	0,521	0,514	0,707	0,542	0,503	0,461	0,659	0,566	0,516	0,502	0,696
Vector similarity Counterfactual	0,37	0,433	0,342	0,601	0,378	0,438	0,348	0,599	0,372	0,432	0,331	0,577	0,378	0,436	0,351	0,601
Vector similarity Reference Context	0,475	0,451	0,459	0,736	0,582	0,508	0,529	0,751	0,443	0,43	0,421	0,776	0,541	0,385	0,236	0,656
Vector similarity Counterfactual Context	0,38	0,463	0,373	0,634	0,4	0,451	0,373	0,582	0,388	0,461	0,375	0,628	0,398	0,372	0,184	0,509
BertScore Reference	0,908	0,777	0,779	0,786	0,913	0,787	0,788	0,793	0,907	0,779	0,787	0,766	0,912	0,788	0,759	0,788
BertScore Counterfactual	0,881	0,75	0,725	0,736	0,884	0,757	0,729	0,741	0,882	0,754	0,733	0,721	0,884	0,763	0,713	0,741
BertScore Reference Context	0,907	0,772	0,773	0,843	0,92	0,788	0,796	0,854	0,905	0,768	0,76	0,867	0,913	0,757	0,722	0,84
BertScore Counterfactual Context	0,881	0,747	0,728	0,734	0,887	0,757	0,738	0,731	0,882	0,75	0,729	0,733	0,885	0,743	0,7	0,715

Figure 5: Results obtained for the Qwen models with the automatic metrics.

H Results with EuroLLM models

In the case of the EuroLLM models, the results are presented in Figure 6.

Differences among languages These models exhibit a better ability to generate commonsensical sentences in English compared to other languages. This is reflected in the larger gap between the similarity scores of the generated sentence when compared to the reference versus the counterfactual. Notably, for some non-English languages, the generated sentence is, in fact, more similar to the counterfactual than to the reference. This suggests that the EuroLLM models are more prone to generating nonsensical outputs relative to the previously discussed models. As well as for Qwen, is also observed a bigger difference for the Valencian, that can be due the use of the encoder, which may be not working well.

Impact of context sentence In the EuroLLM model family, context appears to have an overall positive impact. Most models benefit from the inclusion of contextual information, consistently showing improved results across languages. The only exception is the larger variant of the base model, where the addition of context does not provide the same level of improvement observed in the other models.

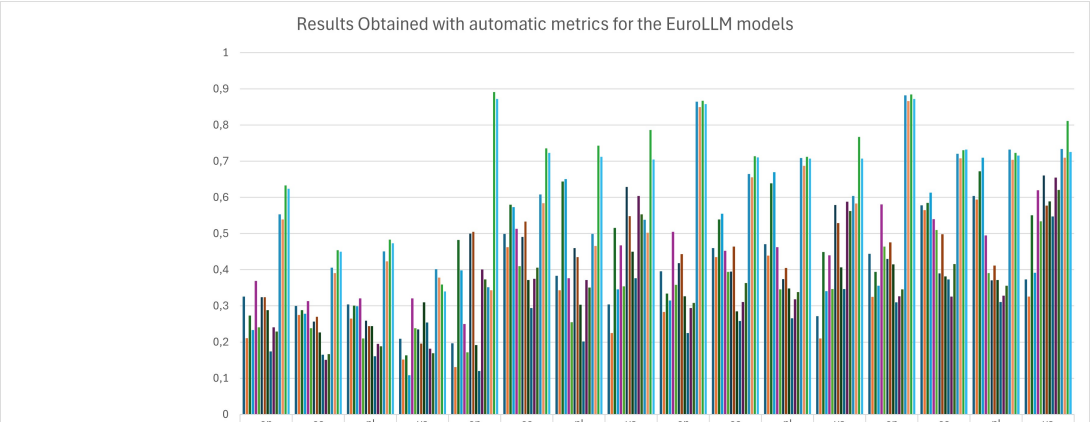
Effect of model size Curiously, within the EuroLLM family, the smaller model versions consistently yield better results across all languages. This may be attributed to differences in training data exposure, as larger models may have been exposed to more diverse or noisier data, which could hinder their ability to generate consistently commonsensical outputs across languages.

I Results with Salamandra models

The results for the Salamandra models are presented in Figure 7.

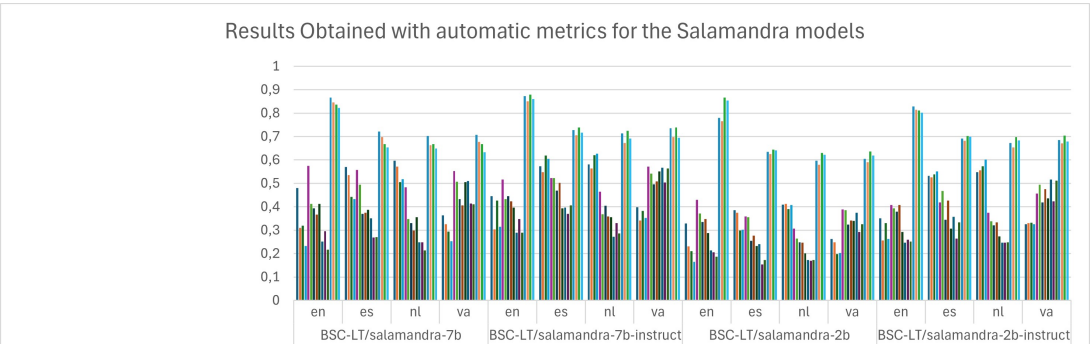
Differences among languages These results show that English continues to exhibit the largest differences in similarity scores when comparing the generated outputs against the reference and counterfactual sentences. This further supports the observation that, for English, models are more likely to generate commonsensical responses. Notably, this model family was trained on extensive data in Catalan, a variant of Valencian. However, the scores showed in the automatic metrics does not shows as much improvement, as the results for other languages shows a bigger distance.

Impact of context sentence In the case of Salamandra models, adding context provides a slight improvement in the instructed versions. However, for the base versions, the inclusion of context appears to actually lower the performance, indicating



	en	es	nl	va	en	es	nl	va	en	es	nl	va	en	es	nl	va
	utter-project/EuroLLM-9B				utter-project/EuroLLM-9B-Instruct				utter-project/EuroLLM-1.7B				utter-project/EuroLLM-1.7B-Instruct			
USE Reference	0,326	0,3	0,304	0,209	0,197	0,499	0,383	0,304	0,396	0,46	0,471	0,272	0,444	0,578	0,604	0,373
USE Counterfactual	0,211	0,275	0,265	0,152	0,131	0,462	0,343	0,225	0,283	0,435	0,439	0,21	0,325	0,565	0,594	0,326
USE Reference Context	0,273	0,288	0,301	0,163	0,482	0,58	0,644	0,516	0,334	0,539	0,639	0,449	0,394	0,585	0,672	0,551
USE Counterfactual Context	0,233	0,278	0,299	0,108	0,398	0,573	0,651	0,346	0,315	0,555	0,67	0,341	0,356	0,613	0,71	0,392
Levensthein Reference	0,369	0,313	0,321	0,321	0,25	0,513	0,377	0,467	0,505	0,452	0,462	0,44	0,581	0,54	0,495	0,62
Levensthein Counterfactual	0,241	0,238	0,21	0,238	0,172	0,41	0,255	0,354	0,358	0,394	0,346	0,347	0,464	0,51	0,391	0,534
Levensthein Reference Context	0,324	0,257	0,259	0,235	0,5	0,491	0,46	0,629	0,418	0,395	0,374	0,579	0,43	0,39	0,371	0,661
Levensthein Counterfactual Context	0,324	0,27	0,244	0,196	0,505	0,533	0,435	0,548	0,443	0,464	0,405	0,529	0,476	0,498	0,412	0,577
Vector similarity Reference	0,288	0,227	0,244	0,31	0,192	0,372	0,303	0,45	0,327	0,285	0,348	0,407	0,415	0,382	0,372	0,589
Vector similarity Counterfactual	0,174	0,165	0,161	0,254	0,12	0,294	0,202	0,377	0,225	0,258	0,266	0,347	0,31	0,373	0,311	0,547
Vector similarity Reference Context	0,241	0,151	0,195	0,182	0,401	0,375	0,372	0,604	0,294	0,311	0,318	0,588	0,327	0,326	0,328	0,655
Vector similarity Counterfactual Context	0,229	0,167	0,188	0,169	0,373	0,406	0,351	0,553	0,308	0,363	0,338	0,562	0,346	0,416	0,356	0,621
BertScore Reference	0,553	0,406	0,451	0,402	0,352	0,608	0,499	0,538	0,865	0,665	0,709	0,604	0,882	0,721	0,732	0,734
BertScore Counterfactual	0,539	0,391	0,423	0,378	0,343	0,584	0,466	0,502	0,85	0,656	0,687	0,583	0,866	0,708	0,704	0,71
BertScore Reference Context	0,633	0,454	0,483	0,359	0,891	0,736	0,743	0,786	0,867	0,714	0,712	0,767	0,885	0,731	0,723	0,811
BertScore Counterfactual Context	0,624	0,45	0,473	0,34	0,872	0,723	0,712	0,705	0,858	0,711	0,707	0,707	0,872	0,732	0,716	0,726

Figure 6: Results obtained for the EuroLLM models with the automatic metrics.



	en	es	nl	va	en	es	nl	va	en	es	nl	va	en	es	nl	va
	BSC-LT/salamandra-7b				BSC-LT/salamandra-7b-instruct				BSC-LT/salamandra-2b				BSC-LT/salamandra-2b-instruct			
USE Reference	0,48	0,57	0,597	0,364	0,446	0,573	0,581	0,398	0,328	0,385	0,409	0,262	0,35	0,532	0,548	0,325
USE Counterfactual	0,31	0,535	0,572	0,326	0,304	0,548	0,563	0,341	0,231	0,374	0,412	0,248	0,256	0,526	0,555	0,33
USE Reference Context	0,319	0,442	0,506	0,294	0,426	0,619	0,621	0,383	0,211	0,299	0,39	0,198	0,33	0,538	0,573	0,332
USE Counterfactual Context	0,233	0,432	0,518	0,253	0,315	0,605	0,626	0,353	0,165	0,302	0,408	0,202	0,262	0,551	0,602	0,326
Levensthein Reference	0,574	0,558	0,483	0,553	0,516	0,522	0,464	0,572	0,43	0,358	0,307	0,388	0,408	0,418	0,374	0,457
Levensthein Counterfactual	0,412	0,494	0,347	0,507	0,433	0,523	0,368	0,541	0,372	0,355	0,264	0,385	0,394	0,467	0,338	0,494
Levensthein Reference Context	0,394	0,37	0,331	0,432	0,445	0,469	0,405	0,495	0,335	0,255	0,248	0,324	0,379	0,344	0,32	0,418
Levensthein Counterfactual Context	0,366	0,374	0,298	0,406	0,424	0,502	0,358	0,509	0,348	0,276	0,247	0,342	0,408	0,427	0,334	0,475
Vector similarity Reference	0,412	0,387	0,355	0,506	0,396	0,394	0,355	0,551	0,288	0,233	0,201	0,34	0,292	0,306	0,274	0,436
Vector similarity Counterfactual	0,252	0,351	0,249	0,51	0,289	0,396	0,272	0,567	0,214	0,24	0,173	0,375	0,247	0,357	0,247	0,516
Vector similarity Reference Context	0,296	0,269	0,248	0,414	0,348	0,37	0,331	0,503	0,206	0,153	0,17	0,293	0,26	0,264	0,247	0,423
Vector similarity Counterfactual Context	0,216	0,271	0,213	0,41	0,289	0,406	0,286	0,563	0,187	0,173	0,173	0,326	0,252	0,334	0,248	0,512
BERTSCORE Reference	0,867	0,722	0,703	0,707	0,872	0,727	0,714	0,735	0,779	0,635	0,596	0,605	0,829	0,691	0,673	0,685
BERTSCORE Counterfactual	0,845	0,697	0,663	0,677	0,851	0,706	0,673	0,698	0,766	0,625	0,579	0,591	0,814	0,682	0,653	0,67
BERTSCORE Reference Context	0,837	0,668	0,667	0,668	0,879	0,739	0,725	0,738	0,867	0,644	0,63	0,636	0,811	0,703	0,697	0,704
BERTSCORE Counterfactual Context	0,822	0,654	0,648	0,633	0,86	0,717	0,691	0,694	0,854	0,641	0,622	0,619	0,802	0,699	0,683	0,678

Figure 7: Results obtained for the Salamandra models with the automatic metrics.

that context may sometimes hinder rather than help the generation process in these models.

Effect of model size Regarding commonsense capability and model size, as expected, larger models demonstrate a better ability to generate commonsensical sentences. This is likely because they

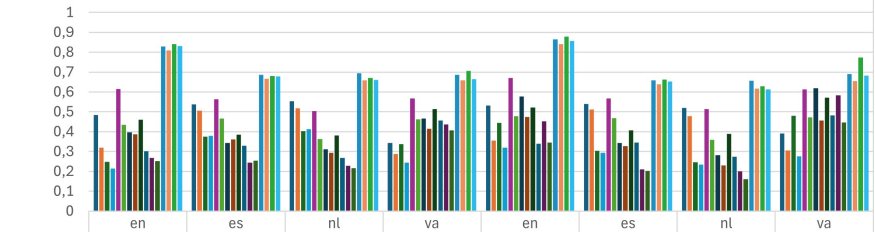
have been trained on larger volumes of text, allowing them to learn and reproduce commonsense knowledge more effectively.

J Results with Gemma models

The results from evaluating the Gemma models are shown in Figure 8.

Differences Among Languages For this model, the difference in similarity scores is notably higher in English than in other languages, reflecting stronger commonsense reasoning capabilities in English. In contrast, the results for Valencian reveal an inverse pattern: the generated sentences are more similar to the counterfactual than to the reference, suggesting a lack of commonsensical understanding. This may be attributed to the model's limited or nonexistent exposure to training data in Valencian.

Impact of Context Sentence Regarding contexts, these models do not seem to benefit from their inclusion. In fact, the scores obtained when context is injected are often lower than those without context, indicating that adding context may sometimes hinder performance.



	en	es	nl	va	en	es	nl	va
	google/gemma-2-2b				google/gemma-2-2b-it			
USE Reference	0,484	0,537	0,553	0,344	0,531	0,54	0,519	0,39
USE Counterfactual	0,32	0,505	0,517	0,288	0,355	0,511	0,478	0,305
USE Reference Context	0,249	0,375	0,403	0,337	0,445	0,304	0,246	0,48
USE Counterfactual Context	0,214	0,378	0,413	0,245	0,32	0,294	0,234	0,276
Levensthein Reference	0,615	0,564	0,504	0,568	0,671	0,568	0,513	0,613
Levensthein Counterfactual	0,435	0,467	0,363	0,463	0,479	0,469	0,359	0,473
Levensthein Reference Context	0,396	0,343	0,311	0,466	0,578	0,344	0,282	0,619
Levensthein Counterfactual Context	0,387	0,362	0,294	0,415	0,474	0,328	0,23	0,456
Vector similarity Reference	0,461	0,384	0,38	0,514	0,522	0,407	0,389	0,571
Vector similarity Counterfactual	0,302	0,329	0,268	0,456	0,34	0,345	0,274	0,482
Vector similarity Reference Context	0,268	0,244	0,228	0,437	0,453	0,21	0,2	0,583
Vector similarity Counterfactual Context	0,253	0,255	0,217	0,407	0,346	0,203	0,16	0,447
BERTSCORE Reference	0,829	0,686	0,695	0,687	0,865	0,658	0,656	0,69
BERTSCORE Counterfactual	0,81	0,667	0,659	0,659	0,841	0,638	0,616	0,654
BERTSCORE Reference Context	0,841	0,681	0,671	0,707	0,878	0,662	0,628	0,773
BERTSCORE Counterfactual Context	0,832	0,679	0,661	0,664	0,856	0,653	0,613	0,683

Figure 8: Results obtained for the Gemma models with the automatic metrics.