GreekCommonGen: A Benchmark for Evaluating Generative Commonsense Reasoning in Greek

Aristotelis Stamopoulos[1] and Dimitrios Galanis[2]

[1]SmartRep.ai, Greece
[2] "Athena" R. C., Institute for Language and Speech Processing (ILSP), Greece

Abstract

This paper introduces GreekCommonGen, the first benchmark designed for generative commonsense reasoning in Greek. The dataset is created by automatically translating the original English CommonGen corpus and subsequently refining the outputs through manual post-editing to ensure linguistic and semantic quality. We conduct a comprehensive evaluation of a range of approaches/models on this benchmark, exploring the impact of different prompting strategies, decoding methods, and model sizes/architectures. Our findings provide valuable insights into the challenges of commonsense generation in Greek, paving the way for future research in the field.

Keywords: Large Language Models, Greek NLP, Generative Commonsense Reasoning

1. Introduction

Large Language Models (LLMs) have demonstrated strong performance across a wide range of natural language tasks. For English there are a plethora of large-scale evaluation benchmarks, however, for Greek, as noted by Papantoniou and Tzitzikas (2020); Zhang et al. (2026), there is a lack of such resources.

A popular benchmark for testing how well models can perform Generative Commonsense Reasoning (GCSR) for English is CommonGen, introduced by Lin et al. (2020). In CommomGen, given a set of common concepts (e.g., {dog, frisbee, catch, throw}) the task is to generate a coherent sentence describing an everyday scenario using these concepts; e.g., "a man throws a frisbee and his dog catches it". The task is considered challenging mainly because it requires 1) relational reasoning based on commonsense knowledge, and 2) compositional generalization for unseen concept combinations. Several methods/approaches for the specific task/dataset in English have been proposed.

The contributions of the paper are the following: a) we make freely available GreekCommomGen[1], the first Greek benchmark for evaluating Generative Commonsense Reasoning b) we evaluate well-known Greek LLMs on the dataset using different models and decoding/prompting strategies. The results provide valuable insights for the difficulty of the task in Greek and for future research.

The remainder of this paper is structured as follows: Section 2 presents the related work while Section 3 describes the construction of GreekCommonGen. Section 4 presents the evaluation methods used for GCSR. Section 5 analyzes the experiments that were conducted on GreekCommonGen, and finally Section 6 presents the concluding remarks and outlines future research directions, as well as the main limitations of this study.

2. Related work

A large number of LLMs have been released in the last few years; e.g., GPT, Mistral, Llama, Qwen, Gemini, Gemma, etc. However, most of them have been trained predominantly on data for widely spoken languages, particularly English. To address this imbalance, several initiatives in recent years have focused on adapting existing LLMs to low or medium resource languages through continual pretraining. Examples of such initiatives are the BIELIK models (Ociepa et al., 2024) for Polish based on Mistral as well as LeoLM[2], a German model derived from Llama. For Greek, two models have been made available the last years, Meltemi (Voukoutis et al., 2024) and Krikri (Roussis et al., 2025), the former is based on Mistral (Jiang et al., 2023) and the latter on Llama 3.1 (Grattafiori et al., 2024). Experimental results have shown that the two Greek LLMs surpass in several Greek NLP tasks the models on which they were based on.

The evaluation of these two Greek LLMs (Voukoutis et al., 2024; Roussis et al., 2025) has been based on a) machine-translated datasets that are widely used for that purpose in English; e.g., MMLU, ARC-Challenge, HellaSwag, etc. and b) native Greek benchmarks such as Belebele (Bandarkar et al., 2024) and Medical MCQA. Recently, more native Greek datasets were made available such

[1] https://github.com/aristotelisStam/CommonGen_Greek/

[2] https://laion.ai/blog/leo-lm/

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 94–100
11–12 May 2026.

as GreekMMLU (Zhang et al., 2026) and GREEK-BARBENCH (Chlapanis et al., 2025); also, there is the Greek portion of Global PIQA (Chang et al., 2025).

Regarding Generative Commonsense Reasoning evaluation there is COMMONGEN (Lin et al., 2020) for English and the recently released COCOTEROS (Maestre et al., 2024) and MULTICOM (Martínez-Murillo et al., 2025). COCOTEROS is for Spanish while MULTICOM extends it and covers English, Spanish, Dutch, and Valencian. To the best of our knowledge, however, there are no available datasets in Greek for Generative Commonsense Reasoning (GCSR).

3. GREEKCOMMONGEN dataset

We created GREEKCOMMONGEN, a Greek counterpart of COMMONGEN. COMMONGEN consists of concept sets (3–5 words), each set is paired with multiple reference sentences describing plausible scenarios integrating these concepts. Our objective was GREEKCOMMONGEN to preserve the commonsense reasoning nature of the original benchmark while ensuring linguistic naturalness in Greek. The process consisted of two stages: (i) automatic translation and (ii) manual revision. In the first phase, reference sentences were translated using MADLAD-3B[3](Kudugunta et al., 2023), a multilingual machine translation model based on T5. Concept sets were not translated at this stage, as their interpretation depends on the contextual alignment with the reference sentences.

Due to the rich morphology of Greek, flexible order of words and pro-drop properties etc., automatic translations frequently required a revision to ensure grammatical agreement, idiomatic fluency, and semantic fidelity. Literal renderings often produced unnatural expressions or semantic shifts. For example, in the concept set ['bow', 'ribbon', 'tie'], automatic translation produced "τόξο" (the weapon used to shoot arrows) for 'bow' instead of "φιόγκος" (the decorative knot). Revisions were restricted to minimal interventions with the aim of restoring grammaticality, preserve the original scenario, cover all concepts as well as to ensure the dataset's linguistic naturalness in Greek. Both manual revisions of automatic translations and concept set translations were carried out by a native Greek speaker to ensure the quality of the translations. Each translated instance was evaluated using the following principals: (i) grammatical correctness, including agreement in gender, number, and case, (ii) The best possible semantic fidelity to the original English sentence while maintaining linguistic naturalness in Greek and (iii) coverage of all concepts within the set. Sentences that failed to meet these criteria were either minimally corrected or discarded if the original meaning could not be preserved without substantial modification. We ensured that the final dataset preserves the structural properties of the original benchmark and kept the concept sets to three concepts each. As a result, GREEKCOMMONGEN constitutes a linguistically natural yet structurally faithful adaptation of COMMONGEN

Concept sets were translated only after finalizing the corrected Greek reference sentences, to ensure lexical and semantic alignment. Each concept set is paired with one to three reference sentences. All concepts were translated according to the following principles: (i) Verbs were expressed in first-person singular present tense, reducing morphological ambiguity in Greek (e.g., distinctions of person, tense, voice, and gender). Furthermore, unlike English, the Modern Greek language lacks an infinitive mood, making the first person indicative a natural and grammatically sound base form for sentence generation. (ii) Passive forms were explicitly preserved where required. (iii) Nouns were rendered in nominative singular. (iv) Participial or derived verbal forms were retained when necessary to reflect the structural diversity of the original dataset.

Overall, the curation of the dataset, the manual translations and revisions required approximately 3 months of careful manual work.

Statistics	Training	Test
References	3,868	1,004
Concept Sets	1,843	342

Table 1: Statistics for GREEKCOMMONGEN

The produced GREEKCOMMONGEN dataset (Table 1) consists of two parts: A test set used for evaluation purposes and a training set that can be used as a parameter tuning dataset. The test set contains 1,004 reference sentences with 342 unique concept sets while the training set contains 3,868 references with 1,843 unique concept sets. Figure 1 showcases the part-of-speech distribution analysis of GREEKCOMMONGEN. The PoS distribution shows that nouns and verbs constitute the majority of the words in both parts, which is justified by the concept-focused nature of the dataset. Determiners and prepositions also form a substantial portion. Due to the nature of the GreekCommonGen task, it is also characterized by noun-heavy sentences and rich verbal morphology.

[3] https://huggingface.co/google/madlad400-3b-mt

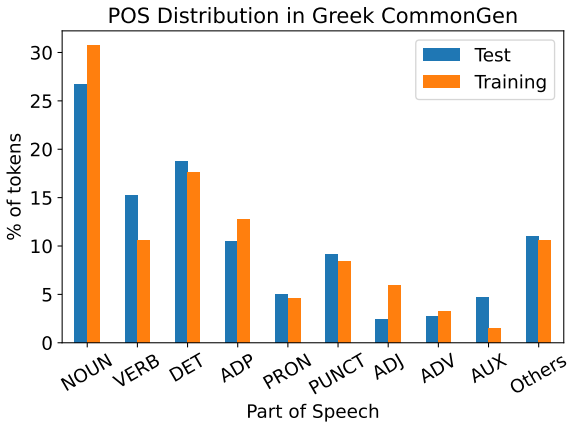

Figure 1: POS Distribution

4. Evaluation measures

Systems and methods developed for the CommonGen task were usually evaluated on the whole test set (1497 concept sets) based on widely-used automatic metrics that measure the overlap of machine-generated between human-generated reference sentences. Such metrics are BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016) and METEOR (Banerjee and Lavie, 2005). The respective leaderboard that reports these scores is available online[4]. In recent years CommonGen-Lite[5], a subset of 400 concept sets from the CommonGen test is used for evaluation. The respective leaderboard is also available online on the same web page as the initial leaderboard. The overall CommonGen-Lite scores are based on a) Length b) Coverage: the percentage of examples where all given concepts are covered by model outputs c) PoS: the percentage of examples where the part-of-speech (PoS) of all given concepts are correct in the outputs. d) Win+Tie Rate: the percentage of examples where GPT-4-turbo prefers the model outputs over the human-written references. Win+Tie Rate can be thought of as an LLM-as-a-judge approach (Zheng et al., 2023).

In the experiments that we present for Greek-CommonGen in the following section we use automatic scores and human scores for evaluation:

- **Automatic measures**: We adapted the provided implementation of **BLEU-4** to Greek's rich morphology. Specifically, we replaced the default tokenizer with spaCy's Greek model, el_core_news_sm[6], and instead of raw tokens, we computed n-grams over lemmatized forms. In addition, we applied Unicode normalization and accent removal, followed by lowercasing. BLEU-4 computation on the preprocessed tokens followed the original approach and was based on precision and brevity penalty. For **CIDEr**, we modified the preprocessing and n-gram representation following the same steps as with BLEU-4. Minor adjustments to the TF-IDF computation were made, including logarithmic smoothing of document frequencies, to prevent several issues caused by rare n-grams in the Greek reference corpus. The code for **SPICE** for the task was not adapted in Greek because of its design; i.e., it would require significant re-engineering.
- **Human-assigned scores**: Following standard practices in NLG evaluation, annotators rated each generated sentence on a 5-point Likert scale (1=very poor, 5=excellent) along three distinct dimensions: a) **Grammaticality**: the degree to which the sentence is fluent, well-formed, and syntactically correct b) **Commonsense Plausibility** how realistic and plausible the described event is with respect to commonsense and everyday knowledge c) **Coverage of Concepts**: how effectively the generated sentence incorporates and relates all the given concept words as instructed. The final scores were averaged per sentence.

Automatic measures BLEU-4 and CIDEr were calculated for the whole test set. For human scores, a subset of 50 concept sets was used.

5. Experiments and Results

Our experimental setup consisted of a variety of model configurations, including instruct models, quantized versions of them, and models fine-tuned using QLoRA (Dettmers et al., 2023). Additionally, we explored different decoding and prompting strategies to examine how these influence model performance.

The evaluated models are a) **Meltemi-7B-Instruct-v1.5 (5-bit quantized)**[7]: A 7B instruction-tuned model optimized for Greek. We use its 5-bit quantized GGUF version. This model serves as a baseline instruction-following system b) **Llama-Krikri-8B-Instruct** [8]: An 8B LLaMA-based model instruction-tuned for Greek. This model represents a higher-capacity instruction-following system. c) **Llama-Krikri-8B-Instruct (5-bit quantized)**[9]: A 5-

[4] https://inklab.usc.edu/CommonGen/leaderboard.html
[5] https://huggingface.co/datasets/allenai/commongen_lite
[6] https://spacy.io/models/el
[7] https://huggingface.co/tensorblock/Meltemi-7B-Instruct-v1.5-GGUF
[8] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct
[9] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

bit GGUF quantized variant of the above model d) **Llama-Krikri-8B-Instruct + QLoRA**: A 4-bit quantized version of Llama-Krikri-8B-Instruct fine-tuned on the training split of GREEKCOMMONGEN using QLoRA. This variant allows us to measure the effect of task-specific adaptation.

We experimented with two prompting approaches: **zero-shot** and **few-shot**[10], in order to assess model performance under varying levels of task guidance. In the zero-shot configurations, models received only a direct instruction to generate a single semantically coherent Greek sentence that includes all provided concept words. In the few-shot configurations, models received multiple {concept-set, sentence} demonstration pairs. Those demonstrations illustrate expected structure, grammaticality, and integration of concepts. The conducted few-shot experiments allowed us to examine whether demonstration-based conditioning improves concept coverage, sentence fluency and task-specific instruction following abilities of the Greek LLMs. The prompt configurations can be found in Appendix A

We included in our experiment setup different decoding strategies since prior work shows that it affects fluency, diversity, and faithfulness (Wan et al., 2023) to required concepts. We therefore evaluate three common decoding methods: 1) **Greedy Decoding**: It deterministically selects the highest-probability token at each step. 2) **Beam Search**: Maintains the top-k partial hypotheses at each step. Beam search is commonly used in the original COMMONGEN benchmark and favors higher likelihood sequences. 3) **Top-p (Nucleus) Sampling**: Samples from the smallest token set whose cumulative probability exceeds threshold p. This introduces stochasticity and improves diversity.

In Table 2, we report automatic evaluation results of our experiments in Tables 3 and 4 and human evaluation results for two annotators; both are Greek native speakers.

Model	Prompt	Decoding	BLEU-4	CIDEr
Meltemi-7B-5bit	0-shot	Greedy	0.0570	0.4282
Meltemi-7B-5bit	0-shot	Top-p	0.0579	0.4353
Krikri-8B-5bit	0-shot	Greedy	0.1603	1.0915
Krikri-8B-5bit	Few-shot	Greedy	0.1771	1.1747
Krikri-8B-5bit	Few-shot	Top-p	0.1451	1.0357
Krikri-8B-Full	0-shot	Beam (5)	0.1578	1.0698
Krikri-8B-Full	Few-shot	Greedy	0.1954	1.2240
Krikri-8B-Full	Few-shot	Beam (5)	0.1818	1.1356
Krikri-8B-4bit-QLoRA	0-shot	Greedy	**0.1977**	**1.3073**
Krikri-8B-4bit-QLoRA	Few-shot	Greedy	0.1969	1.3050

Table 2: Automatic evaluation results on Greek CommonGen.

[10]We adopted an 8-shot configuration

Model	Prompt	Decoding	Gram.	Plaus.	Cov.	Overall
Krikri-8B-5bit	0-shot	Greedy	4.70	3.78	4.98	4.49
Krikri-8B-5bit	Few-shot	Greedy	4.92	4.54	4.98	4.81
Krikri-8B-5bit	Few-shot	Top-p	4.94	4.54	4.88	4.79
Krikri-8B-Full	0-shot	Beam (5)	4.74	4.50	4.96	4.73
Krikri-8B-Full	Few-shot	Greedy	4.96	4.60	4.92	4.82
Krikri-8B-Full	Few-shot	Beam (5)	**4.98**	**4.74**	**5.00**	**4.90**
Krikri-8B-4bit-QLoRA	0-shot	Greedy	**4.98**	4.32	4.98	4.76
Krikri-8B-4bit-QLoRA	Few-shot	Greedy	4.96	4.48	4.98	4.80

Table 3: Human evaluation results (Annotator 1).

Model	Prompt	Decoding	Gram.	Plaus.	Cov.	Overall
Krikri-8B-5bit	0-shot	Greedy	4.64	4.10	4.96	4.56
Krikri-8B-5bit	Few-shot	Greedy	4.90	4.50	4.98	4.79
Krikri-8B-5bit	Few-shot	Top-p	4.82	4.56	4.84	4.74
Krikri-8B-Full	0-shot	Beam (5)	4.62	4.38	4.92	4.64
Krikri-8B-Full	Few-shot	Greedy	4.88	4.68	4.96	4.84
Krikri-8B-Full	Few-shot	Beam (5)	**4.92**	**4.77**	**5.00**	**4.86**
Krikri-8B-4bit-QLoRA	0-shot	Greedy	4.88	4.42	4.96	4.75
Krikri-8B-4bit-QLoRA	Few-shot	Greedy	4.84	4.56	4.98	4.79

Table 4: Human evaluation results (Annotator 2).

Automatic evaluation results (Table 2) show that all Krikri-8B configurations/variants substantially outperform Meltemi-7B on both BLEU-4 and CIDEr. This is due to instruction-following failures; i.e., Meltemi generated (in many cases) multi-sentence outputs despite being instructed to produce a single sentence. For the aforementioned reasons, Meltemi models were excluded from human evaluation. The few-shot and 0-shot QLoRA-tuned Llama-Krikri variants achieve the highest automatic scores, indicating that task-specific training improves alignment with the reference sentences. The few-shot Krikri-8B-Full with greedy and beam search perform strongly, demonstrating that instruction-tuned models can provide competitive baselines without task-specific fine-tuning. Finally, the quantized Krikri models (Krikri-8B-5bit) show minor degradation compared to the full model, unless top-p was used.

Examining the **human evaluation** results, across both annotators (Tables 3 and 4), we can deduce that all Krikri configurations received high scores for grammaticality and coverage, typically 4.6–5.0. Commonsense plausibility shows slightly greater variation between configurations. The few-shot Krikri-8B-Full with beam search configuration receives the highest overall human scores for both annotators, 4.90 and 4.86. The Krikri-8B-4bit-QLoRA models are also very close, especially the ones that use few-shot prompting (4.80 and 4.79). Importantly, quantization does not substantially degrade generation quality, as 5-bit variants (Krikri-8B-5bit) have human-rated scores comparable to full models (Krikri-8B-Full). For example, Krikri-8B-5bit with few-shot prompting and Greedy decoding achieve very high overall scores according to both annotators (4.81, 4.79). We note that although automatic metrics distinguish configurations more clearly, human scores are high and range in a nar-

row interval, suggesting a potential ceiling effect.

Regarding the **prompting strategy**, the few-shot prompts consistently outperformed zero-shot prompting both in automatic and human evaluation. This aligns with studies showing that few-shot examples improve model grounding and guide reasoning by providing structural and semantic cues (Brown et al., 2020).

The QLoRA-tuned models achieve strong performance in both prompting regimes, with minimal difference between zero and few-shot settings. This suggests that task-specific adaptation partially compensates for the absence of in-context demonstrations.

When comparing **decoding algorithms**, Greedy decoding achieves strong and stable performance across configurations, suggesting that its simplicity is well-suited to constrained commonsense generation tasks where faithfulness to required concepts is critical. Beam search performs slightly below greedy decoding in automatic metrics but receives comparable scores in human evaluation for Commonsense Plausibility and Concept Coverage. The broader search appears to improve semantic completeness in some cases. Top-p sampling underperforms and lags behind greedy and beam decoding in this task. The stochasticity introduced by top-p appears less suitable for the GreekCommonGen task.

6. Discussion

6.1. Limitations

The initial CommonGen was only partially translated due to its size. The training and test part can GreekCommonGen can be enhanced with additional human-curated automatic translations. Only Greek LLMs were evaluated on the dataset. Other LLMs could be tested; e.g., the multilingual Teuken 7B (Ali et al., 2024) and EuroLLM 9B (Martins et al., 2025) that were recently released and officially support Greek.

6.2. Conclusions and Future Work

The results of our experiment show that the choice of model and the prompting strategy had the strongest impact on performance. The Krikri models consistently outperformed Meltemi, reflecting its larger-scale and more experience in Greek pretraining. Few-shot prompting led to substantial improvements over zero-shot, confirming that in-context examples provide strong inductive guidance for constrained concept-to-sentence generation. Parameter-efficient fine-tuning (QLoRa) further improved concept coverage and plausibility, while remaining cost efficient and confirming it as a practical strategy for improving Greek LLM performance with limited data. In contrast, quantization caused only minor degradation. This is in contrast with the findings of previous studies that claim that "non-Latin or lower-resource languages suffer more" from the quantization effects (Marchisio et al., 2024). The decoding strategy had the smallest impact overall; deterministic methods greedy and beam search generated similar results but outperformed stochastic sampling (top-p), suggesting that faithful concept realization is more important than diversity in constrained generation tasks.

A notable finding is the consistently high human scores, indicating that modern Greek LLMs (i.e., KriKri) handle the GreekCommonGen task with near-perfect performance. This suggests that the benchmark may no longer adequately challenge current models. Future work could expand the GreekCommonGen dataset with larger and more diverse concept sets that require more advanced reasoning skills. The use of more advanced automatic metrics, such as embedding-based approaches (e.g. BERTScore) or scene-graph-based metrics (like SPICE), could provide better assessment than n-gram overlap. In addition, LLM-as-a-judge evaluation framework also serve as a more thorough evaluation strategy. Importantly, the near-perfect performance observed by the models on our task indicates that more challenging versions of the benchmark are needed. A Greek version of Ordered CommonGen (Sakai et al., 2025), requiring specific concept ordering, would better test instruction-following and compositional generalization. Extending the task to multilingual or context-enriched settings, for example, a Greek version of the MULTICOM (Martínez-Murillo et al., 2025) dataset, would further advance research on Greek generative commonsense reasoning.

7. Bibliographical References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, et al. 2024. Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs. *arXiv preprint arXiv:2410.03730*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan,

Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

T. B. Brown et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Tyler A Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, et al. 2025. Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures. *arXiv preprint arXiv:2510.24081*.

Odysseas S Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. GreekBarBench: A Challenging Benchmark for Free-Text Legal Reasoning and Citations. *arXiv preprint arXiv:2505.17267*.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*.

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

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10:3.

Sneha Kudugunta, Isaac Caswell Gupta, Orhan Firat, Joshua Ainslie, Hyung Won Chung, Yi Tay, Xavier Garcia, David Uthus, William Fedus, et al. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *arXiv preprint arXiv:2309.04662*.

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.

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 *SEPLN Posters*, pages 36–47.

K. Marchisio et al. 2024. How Does Quantization Affect Multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Ivan Martínez-Murillo, Elena Lloret, Paloma Moreda, and Albert Gatt. 2025. Do LLMs Exhibit the Same Commonsense Capabilities Across Languages? *arXiv preprint arXiv:2509.06401*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62.

Krzysztof Ociepa, Krzysztof WrĂłbel, Adrian GwoĹşdziej, Remigiusz Kinas, et al. 2024. Bielik 7b v0. 1: A Polish Language Model–Development, Insights, and Evaluation. *arXiv preprint arXiv:2410.18565*.

Katerina Papantoniou and Yannis Tzitzikas. 2020. NLP for the Greek Language: A Brief Survey. In *11th Hellenic Conference on Artificial Intelligence*, pages 101–109.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing Open Large Language Models for Greek. *arXiv preprint arXiv:2505.13772*.

Y. Sakai, H. Kamigaito, and T. Watanabe. 2025. Revisiting Compositional Generalization Capability of Large Language Models Considering Instruction Following Ability. *arXiv preprint arXiv:2506.15629*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The First Open Large Language Model for Greek. *arXiv preprint arXiv:2407.20743*.

D. Wan, M. Liu, K. McKeown, M. Dreyer, and M. Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. *arXiv preprint arXiv:2303.03278*.

Yang Zhang, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang, and Michalis Vazirgiannis. 2026. Greek-MMLU: A Native-Sourced Multitask Benchmark for Evaluating Language Models in Greek. *arXiv preprint arXiv:2602.05150*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A. Prompts used in the experiments

Here we provide the prompt configurations used in the experiments.

0-shot prompt:

messages = ["role": "user", "content": f΄Δημιούργησε μόνο μία νοηματικά σωστή πρόταση στα ελληνικά που να περιέχει τις εξής λέξεις:concepts."]

Few-shot prompt:

messages = ["role": "system", "content": ¨Είσαι ένα βοηθητικό γλωσσικό μοντέλο που δημιουργεί ελληνικές προτάσεις, με βάση λίστες λέξεων.’,"role": "user", "content": "γήπεδο, κοιτάζω, στέκομαι’, "role": "assistant", "content": ¨Ο παίκτης στεκόταν στο γήπεδο κοιτάζοντας το ρόπαλό του.’, "role": "user", "content": ¨αναπηδώ, χρόνος, μπάλα’, "role": "assistant", "content": ¨Χρειάζεται χρόνος για να μάθεις πώς να κάνεις μια μπάλα να αναπηδάει.’, "role": "user", "content": ¨σταματώ, τραβάω, φωτογραφία’,"role": "assistant", "content": ¨Ο άντρας σταμάτησε για να τραβήξει μια φωτογραφία.’, "role": "user", "content": ¨πλέω, μέρα, βάρκα’, "role": "assistant", "content": ¨Ο ψαράς έπλεε όλη την ημέρα με την βάρκα του.’, "role": "user", "content": concepts]