GPCR: A Greek physical commonsense reasoning dataset

Anna Vacalopoulou, Sokratis Sofianopoulos and Prokopis Prokopidis
Institute for Speech and Language Processing, Athena Research Center
Artemidos 6 & Epidavrou, Athens, Greece
`{avacalop, soksof, prokopis}@athenarc.gr`

Abstract

GPCR is a novel dataset of 208 examples, manually created and annotated for the purpose of physical commonsense reasoning in the Greek language. The dataset was submitted for inclusion in the Multilingual Representation Learning 2025 Shared Task on Multilingual Physical Reasoning Datasets.

1 Introduction

Many languages lack evaluation datasets for physical commonsense reasoning, similar to PIQA (Bisk et al., 2020). In this paper we introduce the Greek physical commonsense reasoning (GPCR) dataset that has been submitted for inclusion in the Multilingual Representation Learning 2025 Shared Task on Multilingual Physical Reasoning Datasets[1].

2 Dataset construction

GPCR was constructed and annotated manually by two annotators, both native speakers of Greek, who consulted a variety of online material, including government and non-governmental organization (NGO) publications, academic theses, course presentations, and commercial product brochures. Other sources included Wikipedia and Wikisource articles, as well as the annotators' own knowledge.

Annotators selected paragraphs from these sources and created prompt and solution tuples from text that was relevant to physical reasoning. Paragraphs were adjusted for clarity and brevity and adapted to the dataset format. In most cases, this adaptation involved removing irrelevant and redundant sentences and phrases from a source paragraph, splitting the resulting paragraph or sentence into two parts (a prompt and a "solution") and then modifying the solution by one or two words to create an alternative "solution" that makes no sense given the prompt. Each example was also assigned a domain string and a binary flag to indicate if it was culturally specific. A domain string was chosen from a set that expanded as new domains emerged during the annotation process. Examples were classified as culturally specific if they were not readily translatable into English or required regional and/or cultural commonsense, in accordance with the MRL 2025 Shared Task guidelines.

After an initial set of 268 examples was collected and annotated, the annotators collaborated to double-check each item against a set of inclusion criteria: (a) physical properties, (b) commonsense knowledge, and (c) cultural specificity. We filtered out 60 items that did not meet these criteria.

3 Dataset description

The final dataset contains 208 examples, each with the following columns: `id`, `prompt`, `solution0`, `solution1`, `label` (0 or 1 as the correct solution), `culturally specific`, and `domain`. The distribution of the dataset examples across the final 20 domains is presented in Figure 2, which also indicates the cultural specificity of the examples within each domain. Approximately 40% of the final examples were annotated as culturally specific.

We show below two examples from the dataset. Examples 1 and 2 were assigned to the `local foods` and `driving` domains, respectively. The former was additionally classified as culturally specific.

1. prompt: Τι πρέπει να έχει γίνει στο ταψί πριν τοποθετηθεί πάνω του το φύλλο για πίτα; "How should the baking sheet be prepared before placing the filo dough on it?"
   solution0: Να έχει αλευρωθεί. "It should be lightly floured."
   solution1: Να έχει λαδωθεί. "It should be greased."

[1] `https://sigtyp.github.io/st2025-mrl.html`

label: 1

2. prompt: Ποια είναι η σωστή θέση των χεριών πάνω στο τιμόνι σε αντιστοιχία με την ώρα που δείχνει ένα ρολόι; “What is the correct position of the driver’s hands if you compare the steering wheel with the numbered screen of a clock with hands?”

   solution0: Εννέα και τέταρτο. “A quarter past nine.”

   solution1: Έντεκα και πέντε. “Five past eleven.”

   label: 0

We process prompts and solutions with an NLP toolkit for Greek (Prokopidis and Piperidis, 2020). Prompts average 13.65 tokens and correct and trick solutions average 12.69 and 12.68 tokens (cf. Figure 3, where we observe similar distributions for correct and trick solution lengths). We also see in Figure 4 that the most frequent adjectives and nouns in the GPCR examples focus on physical properties and regional knowledge.

4 Evaluation

We use GPCR to evaluate three open models: Llama-Krikri-8B (Roussis et al., 2025), Llama 3.1-8B (Grattafiori et al., 2024) and Gemma3-27B (Kamath et al., 2025). Figure 1 summarizes the results.[2] Among the tested models, Gemma3-27B demonstrates the highest overall accuracy (71%) and consistently strong performance on both culturally specific (70%) and non-culturally specific (71%) examples. Krikri-8B, a model based on Llama 3 and trained on a data mix with a Greek focus, achieves an overall accuracy of 59% and shows a relative advantage in handling culturally specific cases (63%) compared to non-culturally specific ones (57%). Llama-3.1-8b scores lower (53%) and struggles particularly with culturally specific examples (51%). Based on these initial results, the dataset seems to be challenging for the two relatively small models, as suggested by their lower accuracy scores compared to Gemma3-27B. Krikri-8B, which is built on Llama 3.1-8B and is trained on a data mix with a Greek cultural focus, outperformed its parent model—especially on culturally specific examples. This suggests that targeted, culturally relevant training data can enhance a model’s ability to handle nuanced language tasks and improve overall accuracy. We submit the dataset for inclusion in the Shared Task on Multilingual Physical Reasoning Datasets and also make it available from https://huggingface.co/datasets/ilsp/greek_pcr.

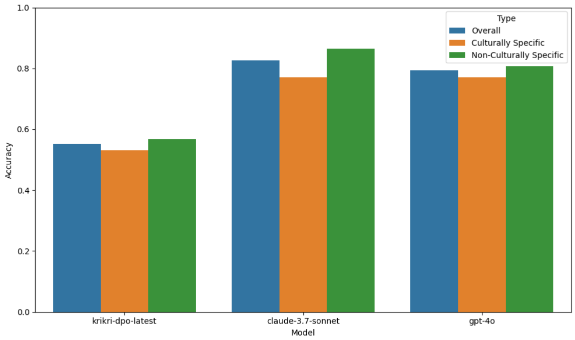

Figure 1: Accuracy of three LLMs on GPCR

[2] We provide evaluation code, predictions and results on this repo.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Prokopis Prokopidis and Stelios Piperidis. 2020. A Neural NLP toolkit for Greek. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 125–128, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2505.13772.

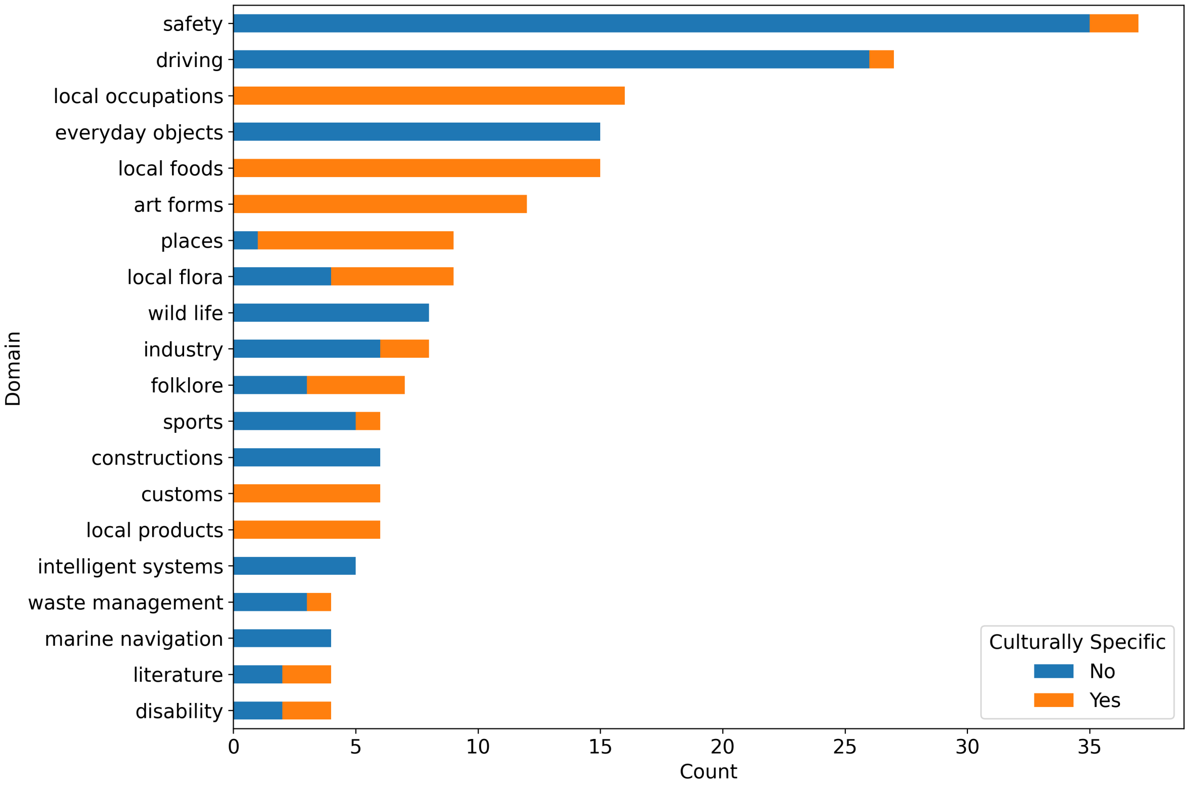

Figure 2: Distribution of (culturally specific) examples across domains in GPCR

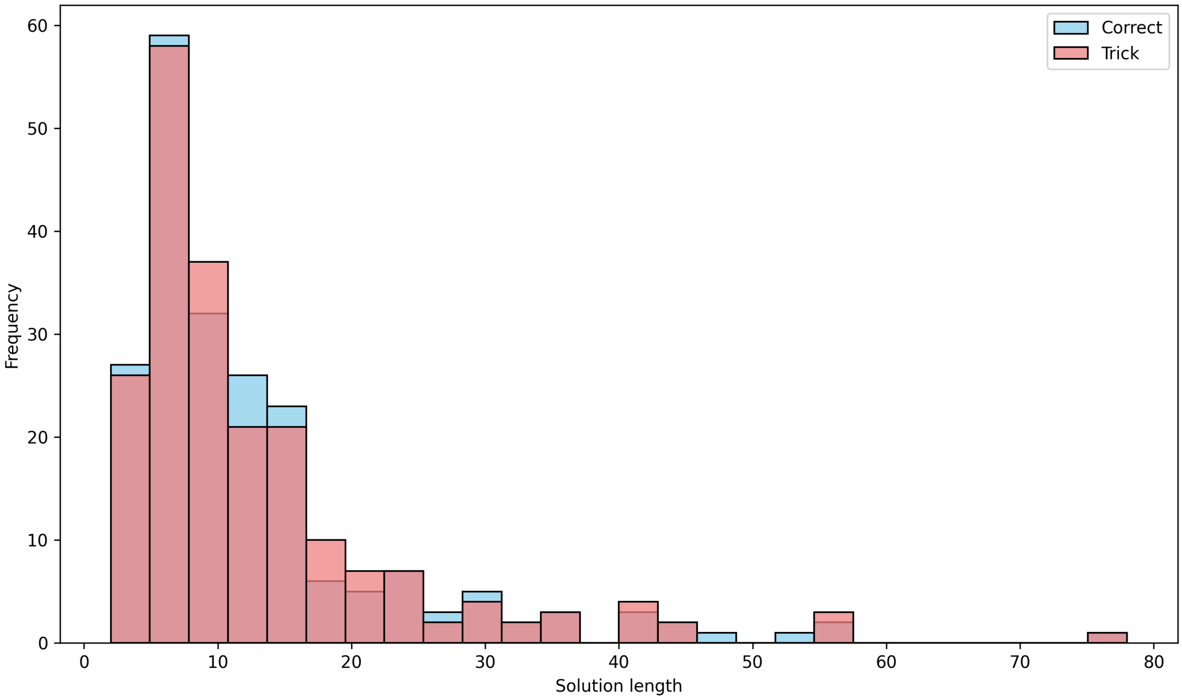

Figure 3: Length distributions in tokens for correct and trick solutions in GPCR

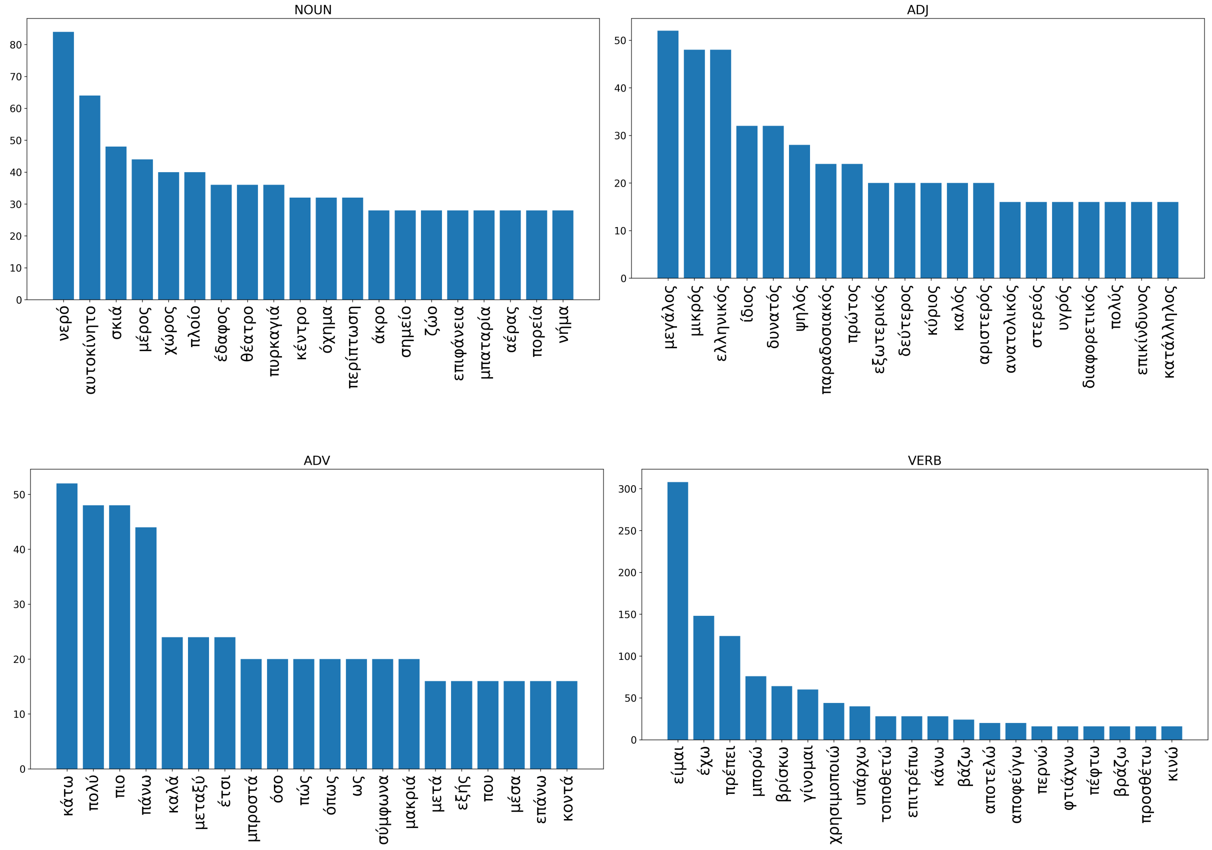

Figure 4: Frequency distributions for the top 20 words tagged as nouns, adjectives, adverbs and verbs in the prompt and correct solution texts of GPCR. We observe both a focus on physical properties (e.g. NOUN: νερό “water”, έδαφος “terrain”, επιφάνεια “surface”; ADJ: μεγάλος “large”, μικρός “small”, εξωτερικός “external”, υγρός “wet/liquid”; ADV: κάτω “down”, πάνω “up”; VERB: τοποθετώ “place”, πέφτω “fall”) and often on regional and cultural knowledge (e.g. ADJ: ελληνικός “Greek”, παραδοσιακός “traditional”; NOUN: θέατρο “theater”).