A Cross-Linguistic Evaluation of Cardinal Direction Reasoning in LLMs

Panagiota Gyftou
Dept. of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

Robert E. Blackwell
School of Computer Science, University of Leeds, UK

Manolis Koubarakis
Dept. of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

Anthony G. Cohn
School of Computer Science, University of Leeds, UK
Alan Turing Institute, London, UK

Abstract

We present a cross-linguistic evaluation of cardinal direction reasoning in large language models. We extend an earlier cardinal direction benchmark, translating the questions into Greek and French and compare 10 LLMs on the new benchmarks.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning

Keywords and phrases Large Language Models, Spatial Reasoning, Cardinal Directions

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.18

Category Short Paper

Supplementary Material *Dataset (Dataset)*: `https://github.com/panagiotagyft/cosit2026`
archived at `swh:1:dir:9ef27812d94d1315e5c45dc3ae396cde58141d0a`

Funding This work was partially supported by the Marie Skłodowska-Curie Actions Staff Exchanges fund (project ThirdWave, GA no. 101236394), the Fundamental Research priority area of The Alan Turing Institute, the Economic and Social Research Council (ESRC) under grant ES/W003473/1, and the EPSRC under grant EP/Z003512/1.

Acknowledgements This work was supported by computational time granted from the National Infrastructure for Research and Technology S.A. (GRNET S.A.) in the National HPC facility – ARIS – under project ID pa260202-A Cross-Linguistic Evaluation of Cardinal Direction Reasoning in LLMs.

1 Introduction

Many claims have been made about the reasoning capabilities of Large Language Models (LLMs) [10] and many benchmarks have been developed to evaluate LLM reasoning performance e.g., Grade School Math 8K (GSM8K) [11]. The majority of widely used LLM evaluation benchmarks remain English-centric, with multilingual benchmarks representing a minority (roughly 20-30%) of well-known benchmarks (as an example, MGSM [8] extends GSM8K to 10 typologically diverse languages). As a result, the reasoning capabilities of LLMs in other languages remain comparatively underexplored. The reasoning problem on which we concentrate in this paper is *spatial reasoning about cardinal directions (CDs)* [1, 2]. CDs are used to describe how regions of space are placed relative to one another (e.g., region

a is north of region b). They are an important case of qualitative spatial relations which are used to model space similarly to the way humans do. They have been found to be useful in GIS, autonomous navigation and robotics and computer vision.

In order to explore the reasoning capabilities of modern LLMs in languages other than English, this paper concentrates on two European languages: French and (Modern) Greek. French is a high-resource language but not as dominant as English. As an example, there are only 4.6052% documents with primary language French in the Common Crawl latest version CC-MAIN-2026-12[1], which is one of the large corpora used to train LLMs, compared to 41.0588% English documents. Greek, on the other hand, is substantially under-represented (0.5494% of the total number of documents).

Despite being major European languages, French and Greek present unique hurdles for LLMs: (i) morphological complexity: French and Greek are highly inflected (words change form based on case, gender and number). This requires more data for a model to learn the rules of the language compared to English. (ii) script heterogeneity (in the case of Greek): Because Greek uses its own alphabet, it does not benefit from lexical sharing the way French does (e.g., a model learns the word "information" in English and instantly understands "information" in French). (iii) tokenization inefficiency: many state-of-the-art LLMs break French and Greek words into more tokens than English words because they use tokenizers optimized for English [3, 15, 6]. This makes processing French and Greek more expensive and slightly slower for the model.

The contributions of this paper are the following:

1. We develop the first two benchmarks for reasoning about CDs in French and Greek by translating the (English) benchmark of [1, 2] into these two languages. The new benchmarks consist of 7680 questions in each language.
2. We evaluate the reasoning capabilities of 10 LLMs on the new benchmarks, in order to explore how language affects spatial reasoning performance in state-of-the-art LLMs. We have selected a variety of models: one commercial multilingual LLM with strong reasoning capabilities (*GPT-5.2*), three smaller multilingual LLMs with enhanced reasoning capabilities but not a dedicated reasoning architecture (*Kimi-k2-0905*, *Deepseek V3.2* and *Llama-3.3-70B*), the European multilingual models *EuroLLM-22B* and *EuroLLM-9B*, the state of the art models for the Greek language *Meltemi-7B* and *Krikri-8B*, the model *Lucie-7B* for French, and the small multilingual model *Mistral Nemo*. Our experiments show that, in general, spatial reasoning in English yields better accuracy for the majority of examined LLMs, given that they have been trained mostly on English corpora. French and Greek showed consistently strong performance, even outperforming English in the *GPT-5.2* reasoning model. Generally, the models struggle with intercardinal directions, suggesting that compound words increase reasoning complexity. Additionally, sentence clarity plays a crucial role in the quality of responses, as complexity increases, accuracy tends to decrease. Finally, variations in locomotion type and grammatical person do not seem to affect the responses.

2 Related work

Cohn and Blackwell tested LLM reasoning using a CD benchmark [1, 2] showing evidence of systematic errors and bias in LLM reasoning. The benchmark consists of 5760 questions and answers generated from a set of templates, testing an LLM's ability to determine the correct CD given a particular scenario (for more details see Section 3). However, they only

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

test English language questions and they do not test European LLMs, except for Mistral. It remains unclear whether their findings generalize across languages with different scripts, morphologies, and spatial expressions. Another recent work which considers reasoning about CDs is [7] where the authors evaluate 7 large reasoning models on the task of computing the transitivity table of the CD calculus of [19] which deals with regions and allows relations North, South, East, West and their combinations (e.g., Northwest), but only in English.

The question of whether language influences reasoning in LLMs has been studied by other researchers too. [5] analyzes the internal workings of Qwen1.5-1.8B-Chat in cross-lingual causal reasoning from both syntactic and semantic perspectives. [9] introduce the cross-lingual-thought prompting method to systematically improve the multilingual capability of LLMs. [13] introduced a novel methodology, an open-source framework for generating correct-by-construction natural language spatial reasoning tasks that enable decomposed evaluation of compositional abilities beyond mere accuracy, and DecompSR, a large-scale, customisable benchmark dataset (over 5 Million samples) for multi-hop compositional spatial reasoning. They also evaluated the performance of contemporary LLMs demonstrating that while LLMs show some linguistic resilience, they largely struggle with systematic and productive generalisation in spatial tasks and exhibit overgeneralisation tendencies. Finally, [4] introduces the multilingual dataset xSTREET that covers four tasks across six languages. The authors show that xSTREET demonstrates a gap in base LLM performance between English and non-English reasoning tasks. Then, they propose two methods (one for training time and one for inference time) to remedy this gap, building on the idea that LLMs trained on code are better reasoners.

In Europe, the most important language technologies activity at the moment is the Alliance for Language Technologies (ALT-EDIC). ALT-EDIC is a "European Digital Infrastructure Consortium supporting the development of language technologies[2]. It promotes excellence in language technologies and contributes to the preservation of Europe's linguistic diversity and cultural richness." In addition, it aims to strengthen Europe's technological and economic sovereignty and competitiveness, through the development of transparent and compliant open-source multilingual models. These models are tailored to all official European languages, as well as several languages of strategic international importance. In the context of ALT-EDIC, OpenEuroLLM is a family of LLMs developed in collaboration with the High Performance Language Technologies initiative[3] which has produced 38 monolingual models. These cover all European languages excluding Irish and Maltese.

Another important activity in Europe has been the development of EuroLLM[4] supporting all 24 official EU languages. A significant milestone is the collaboration between OpenEuroLLM and EuroLLM on MultiSynt, an open-source multilingual synthetic dataset for LLM pre-training. As the first effort of its kind, democratizes access to high quality training data, particularly for languages that remain underrepresented in the current LLM space[5].

For the French and Greek languages considered in this paper, there is also considerable recent LLM development activity. The most important LLMs with an emphasis on the French language are Lucie [15] and Cedille [18]. Lucie-7B is trained on equal amounts of data in French and English (roughly 33% each). Cedille is trained only on French data. We also note that the French company Mistral AI is a leading European player in the LLM space, with models that are multilingual rather than exclusively focused on French.

[2] https://www.alt-edic.eu/
[3] https://huggingface.co/collections/HPLT/hplt-20-monolingual-reference-models
[4] https://eurollm.io/
[5] https://openeurollm.eu/blog/multisynt-synthetic-training-data

Recent Greek LLMs include Meltemi-7B [12] and Llama-Krikri-8B [6]. The latter model is currently the state of the art for the Greek language. It has been trained on a corpus which includes 56.7 billion monolingual Greek tokens (62.3%), 21 billion monolingual English tokens (23.1%), 5.5 billion parallel data tokens (6.0%), and 7.8 billion math and code tokens (8.6%). It demonstrates substantial improvements for Greek (+10.8%) compared to Llama-3.1-8B [6]. Moreover, Llama-Krikri-8B-Base surpasses Meltemi-7B-v1.5 with a notable +11.6% average improvement across all benchmarks tested.

3 Experimental design

The benchmark questions from [1, 2] are generated from six templates each modulated with ten forms of locomotion (cycling, driving, hiking, jogging, perambulating, racing, riding, running, unicycling and walking), eight directions (north, south, east, west, north-east, north-west, south-east, south-west) and each of the six English person forms: first-person (I am), first-person plural (We are), second-person (You are), third-person singular (He is and She is), and third-person plural (They are). This gives us 6 templates $\times$ 10 locomotion forms $\times$ 6 person forms $\times$ 8 directions $\times$ 2 direction-variations = 5760 questions.

We take these templates and translate them into both Greek and French, which both have two more person forms. Moreover, verbs have to agree in person and number, and both Greek and French exhibit a richer inflectional system for person and number compared to English, where these differences are implicit. Specifically, English uses the single pronoun "You" for both the second person singular and plural. In contrast, both Greek and French employ distinct forms, "Εσύ" (Greek) and "Tu" (French), for the singular, and "Εσείς" and "Vous" for the plural. Additionally, the second person in both languages differs in formal and informal contexts. Specifically, the second person singular is used informally for people one knows or for children, whereas the second person plural can be used in the singular to express politeness. Similarly, although all three languages distinguish gender in the third person singular ("He"/"She", "Αυτός"/"Αυτή", "Il"/"Elle"), in the plural, English uses only "They", whereas Greek and French distinguish by gender: "Αυτοί"/"Ils" are used for masculine or mixed groups, and "'Αυτές'/"Elles" are used for groups of women. So our benchmarks for French and Greek consist of 7680 questions ($6 \times 10 \times 8 \times 8 \times 2$).

We adopt the same experimental setup as [1, 2] setting temperature to 0 with a fixed random seed (where available). We perform three runs per prompt to verify the consistency of the results across iterations. Each prompt is presented as a new LLM conversation so that models do not retain context from earlier prompts.

Table 1 Example question templates for each language.

English	T1	You are walking *[south]* along the *[east]* shore of a lake; in which direction is the lake?
	T2	You are walking *[south]* along the *[east]* shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake?
	T3	You are walking *[south]* along the middle of the *[east]* side of a park; in which direction is the bandstand located in the centre of the park?
	T4	You are walking *[east]* along the *[south]* side of a road which runs *[east to west]*. In which direction is the road?
	T5	You are walking *[south]* along the *[east]* shore of the island. In which direction is the sea?
	T6	You are walking *[south]* along the *[east]* shore of an island and then turn around to head back in the direction you came from, in which direction is the sea?
French	T1	Tu marches *[au sud]* le long de la rive *[est]* d'un lac. Dans quelle direction se trouve le lac?
	T2	Tu marches *[au sud]* le long de la rive *[est]* d'un lac et puis tu retournes pour revenir dans la direction d'où tu es venu(e). Dans quelle direction se trouve le lac?
	T3	Tu marches *[au sud]* au milieu du côté *[est]* d'un parc. Dans quelle direction se trouve la banderole située au centre du parc?
	T4	Tu marches *[à l'est]* le long du côté *[sud]* d'une route qui s'étend de *[l'est à l'ouest]*. Dans quelle direction se trouve la route?
	T5	Tu marches *[au sud]* le long de la côte *[est]* d'une île. Dans quelle direction se trouve la mer?
	T6	Tu marches *[au sud]* le long de la côte *[est]* d'une île et puis tu retournes pour revenir dans la direction d'où tu es venu(e). Dans quelle direction se trouve la mer?
Greek	T1	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* όχθης μιας λίμνης. Προς ποια κατεύθυνση βρίσκεται η λίμνη;
	T2	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* όχθης μιας λίμνης και στη συνέχεια γυρίζεις για να επιστρέψεις προς την κατεύθυνση από την οποία ήρθες. Προς ποια κατεύθυνση βρίσκεται η λίμνη;
	T3	Περπατάς *[νότια]* στη μέση της *[ανατολικής]* πλευράς ενός πάρκου. Προς ποια κατεύθυνση βρίσκεται η εξέδρα της μπάντας στο κέντρο του πάρκου;
	T4	Περπατάς *[ανατολικά]* κατά μήκος της *[νότιας]* πλευράς ενός δρόμου που εκτείνεται από τα *[ανατολικά προς τα δυτικά]*. Προς ποια κατεύθυνση βρίσκεται ο δρόμος;
	T5	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* ακτής ενός νησιού. Προς ποια κατεύθυνση βρίσκεται η θάλασσα;
	T6	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* ακτής ενός νησιού και στη συνέχεια γυρίζεις για να επιστρέψεις προς την κατεύθυνση από την οποία ήρθες. Προς ποια κατεύθυνση βρίσκεται η θάλασσα;

4 Results

4.1 Accuracy of models across all languages

To compute accuracy we adopt the formula [17] which calculates the mean accuracy for a model while simultaneously determining a prediction interval to quantify accuracy variance across repeated evaluations. Therefore, in the heatmaps appearing in this section, we present accuracy values as a prediction interval (i.e., mean value ± a margin of error).

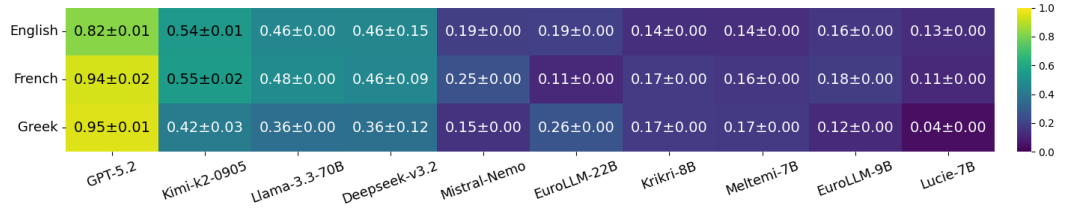

Figure 1 Comparison of model performance across three languages.

As we see in Figure 1, the best performing model across all languages is *GPT-5.2*, showing the highest accuracy in Greek (95%), closely followed by French (94%) and finally English (82%). It is surprising that the English language performance of GPT-5.2 reported here (0.82) is worse than that of o1 (0.92) reported by Cohn and Blackwell [2], despite GPT-5.2 being a newer OpenAI model. This discrepancy is particularly pronounced for template T4 questions: GPT-5.2 achieves a score of only 0.08, compared to 0.55 for o1. The superiority of French and Greek might be considered a surprising and counterintuitive result, given that *GPT-5.2* has been trained predominantly on English corpora. We conjecture that this result may be due to one or more of the following factors: (i) the fact that the translated questions may be less ambiguous compared to the English ones, e.g., with respect to person form, given that French and Greek are highly inflected languages. (ii) data contamination and overfitting given that the English dataset has been published on the Web since 2024 and it is may be part of the training data of *GPT-5.2* which has a knowledge cutoff of late 2025 (though the answers are in a separate file); in contrast, the French and Greek datasets are new so the model has to reason about them from first principles (unless the model were to translate into English for the reasoning), (iii) denoising: English training data is noisy in comparison with Greek and French which might be of higher quality (e.g., from Wikipedia) in a higher percentage, and (iv) more tokens per cardinal direction for French and Greek which could make processing less efficient but enable more detailed reasoning steps. (v) *GPT-5.2* uses a dynamic reasoning approach – i.e. by default, it auto-decides how much reasoning effort to devote to a problem; it might be that the presence of a language other than English makes it more likely to trigger a higher reasoning effort which may in turn lead to more accurate answers. Regarding the English dataset, excluding *GPT-5.2*, the leading models are *Kimi-k2-0905* with an accuracy of 54%, *Deepseek V3.2* with 46% and *Llama-3.3-70B* with the same accuracy of 46%. The reason for this lower performance is that these models have significantly fewer parameters than *GPT-5.2* and are not truly reasoning models with a specialized reasoning architecture; however, according to their developers, they have enhanced reasoning capabilities as opposed to pure LLMs. The leading models for Greek follow the same hierarchy as in English but with lower accuracy scores. *Kimi-k2-0905* achieved 42%, while *Deepseek V3.2* and *Llama-3.3-70B* scored 36%. On the other hand, for French, the models behave similarly to the English dataset: *Kimi-k2-0905* achieved

55% accuracy, while *Llama-3.3-70B* reached 48% and *Deepseek V3.2* 46%. It is notable that *Llama-3.3-70B* has fewer parameters than *Deepseek V3.2* yet it achieved the same or sometimes better accuracy on the French dataset.

The smaller models in Figure 1 exhibit different behaviors across the three languages. Specifically, for the European instruct models, we see that *EuroLLM-9B* performed better on the French dataset with 18% accuracy, compared to 16% and 12% for the English and Greek datasets, respectively (chance is 12.5%). In addition, for *EuroLLM-22B*, we observed better performance in the Greek language with 26%, while English showed a lower accuracy score of 19% and French 11%. *EuroLLM-9B* was trained initially on 50% English, with this percentage being reduced gradually during the next two phases of its training. The French dataset accounted for $\sim 5-6\%$ across all three phases, whereas Greek was only $\sim 0.5-1\%$; this explains these results [16]. *EuroLLM-22B* was also trained with 60% English, $\sim$9% French and $\sim$1% Greek, but in the last phases the authors of the EuroLLM models used the EuroFilter to choose the best quality data [14]. We conjecture that in the case of *EuroLLM-22B*, the Greek data was of better quality than the French. Furthermore, regarding the Greek models, *Meltemi-7B* and *Krikri-8B*, as anticipated, both showed superior performance in the Greek dataset, each reaching 17% accuracy. On the English dataset, however, both models achieved lower accuracy, 14%. On the French dataset, *Meltemi-7B* scored 16%, while *Krikri-8B* scored 17%. On the other hand, the French model *Lucie-7B* showed its best score on the English dataset with 13%, followed by the French dataset with 11%. The model performs poorly on the Greek dataset due to a lack of prior training in that language. Finally, *Mistral Nemo* shows fair performance on the French dataset with an accuracy of 25%, whereas in English it achieved 19%. Despite the fact that it has not been trained for Greek, the model scored 15%; we hypothesize that its multilingual Tekken tokenizer provides better support for Greek than other tokenizers.

4.2 Performance by cardinal and intercardinal directions

In this section, we investigate whether the type of direction – cardinal (e.g., north) versus intercardinal (e.g., northwest) – affects the LLM's responses to benchmark questions. Our results indicate that reasoning involving cardinal directions is more accurate than reasoning involving intercardinal directions; e.g. even in the highest scoring overall, *GPT-5.2*, the accuracies are 84% and 81% respectively[6]. As illustrated in Figure 2 in the Appendix, high accuracy (above 43%) is also observed across all languages for the larger models (*GPT-5.2*, *Kimi-k2-0905*, *Llama-3.3-70B*, *Deepseek V3.2*) with cardinal directions. With intercardinal directions, accuracy is high only for the *GPT-5.2* model as already observed; for the remaining large models, the maximum score achieved is 40% (*Kimi-k2-0905*). This agrees with the findings of [1, 2] for the English dataset. The *Mistral Nemo* model shows the worst results for intercardinal directions; its accuracy across all languages is close to zero (0.01-0.02%). Additionally, *EuroLLM-22B* achieved a 0% score on the French intercardinal directions, while the smaller corresponding model (9B) achieved 16%. Regarding the Greek models, *Meltemi-7B* showed better performance in Greek for intercardinal directions than for cardinal ones (19% vs. 14%). The larger Greek model, *Krikri-8B* showed a score of 16% in intercardinal directions for the French dataset, outperforming its results in Greek (12%) and English (9%). Finally, the French model *Lucie-7B* doesn't perform well with cardinal directions, achieving

[6] A similar result has been seen reported for the non-reasoning model GPT-4.5 in [2].

only 8-10% across all languages. However, it reasons about intercardinal directions more accurately, although it achieves 0% for the Greek dataset, which is expected as the model hasn't been trained on it.

4.3 Performance by question template

The syntactic complexity of a question template is an important factor in the accuracy of an LLM's response. As can be seen in Figure 3 in the Appendix, the easiest template for the LLMs is T5, with the strongest model being *GPT-5.2* which achieved 100% across all languages. The hardest template is T4. Only *GPT-5.2* could achieve high scores in T4 for the French and Greek datasets, with 77% and 75%, respectively. However, for the English dataset, it scored only 8%. Our speculation above that the adaptive reasoning mechanism of *GPT-5.2* may not be working in English's favour may again be explanation for this perhaps surprising difference. The difficulty of T4, compared to other templates, stems from conceptual/semantic confusion introduced by the phrase "road which runs from [one direction] to [another]" since the LLM must distinguish between the agent's movement and the additional information regarding the road's orientation. It is also interesting to note that for T4, *Mistral Nemo* scored 14% in French, 12% in Greek and 15% in English – scoring better than models with a significantly higher number of parameters. Templates T1 and T5 have similar questions, the only difference being the environment: T1 uses "lake", whereas T5 uses "island" and "sea". Additionally, T2 and T6 are more complex extensions of T1 and T5, respectively. Our experiments across all languages show two main results: first, models understand simpler questions (T5, T1) better, and second, models understand the terms "sea shore"/"island" better than "lake shore"/"lake". The reason might be that, for the island question, when you are on the east shore, the sea is to the east, whereas for the lake it is to the west; the latter is arguably the more complicated inference.

4.4 Performance by locomotion and person form

The accuracy across all languages does not change drastically for any particular type of locomotion. Overall, the best performance across all languages was achieved for "hiking" while "riding" performed slightly worse. Similarly, the grammatical person form does not significantly impact LLM performance. The highest accuracy was observed with the second person singular "You" while the third person singular "She"/"He" yielded the lowest results. Nevertheless, the performance difference between them remains negligible. This is in agreement with [1].

5 Future work

We plan to develop benchmark datasets for more European languages (e.g., German and Spanish), and to test our models with very low resource languages such as Basque, as well as right-to-left languages (e.g. Arabic or Hebrew), and also logographic systems such as Chinese. We would also like to experiment with a broader range of reasoning models such as Deepseek-R1 and Phi-4. Given the current poor performance of small models, we would like to try enhancing them using fine-tuning or in context learning. Finally, we would like to consider other spatial reasoning tasks.

References

1 A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions (short paper). In *COSIT 2024*, 2024. doi:10.4230/LIPIcs.COSIT.2024.28.

2 A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions, revisited, 2025. Workshop QR@IJCAI 2025. arXiv:2507.12059.

3 A. Petrov et al. Language model tokenizers introduce unfairness between languages, 2023. arXiv:2305.15425.

4 Bryan Li et al. Eliciting better multilingual structured reasoning from LLMs through code, 2024. arXiv:2403.02567.

5 C. Wang et al. Under the shadow of babel: How language shapes reasoning in llms, 2025. arXiv:2506.16151.

6 D. Roussis et al. Krikri: Advancing open large language models for Greek, 2025. arXiv:2505.13772.

7 E. Gardelakos et al. Can large reasoning models reason about spatial relations? In *ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2025.

8 F. Shi et al. Language models are multilingual chain-of-thought reasoners, 2022. arXiv:2210.03057.

9 H. Huang et al. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting, 2023. arXiv:2305.07004.

10 J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models, 2022. arXiv:2201.11903.

11 K. Cobbe et al. Training verifiers to solve math word problems, 2021. arXiv:2110.14168.

12 L. Voukoutis et al. Meltemi: The first open large language model for Greek, 2024. arXiv:2407.20743.

13 Lachlan McPheat et al. Decompsr: A dataset for decomposed analyses of compositional multihop spatial reasoning, 2025. arXiv:2511.02627.

14 Miguel Moura Ramos et al. Eurollm-22b: Technical report, 2026. arXiv:2602.05879.

15 O. Gouvert et al. The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation, 2025. arXiv:2503.12294.

16 Pedro Henrique Martins et al. Eurollm-9b: Technical report, 2025. arXiv:2506.04079.

17 R. E. Blackwell et al. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores, 2024. arXiv:2410.03492.

18 M. Müller and F. Laurent. Cedille: A large autoregressive french language model, 2022. arXiv:2202.03371.

19 S. Skiadopoulos and M. Koubarakis. Composing cardinal direction relations. *Artif. Intell.*, 152(2), 2004. doi:10.1016/S0004-3702(03)00137-1.

A Visualizations of Experimental Results

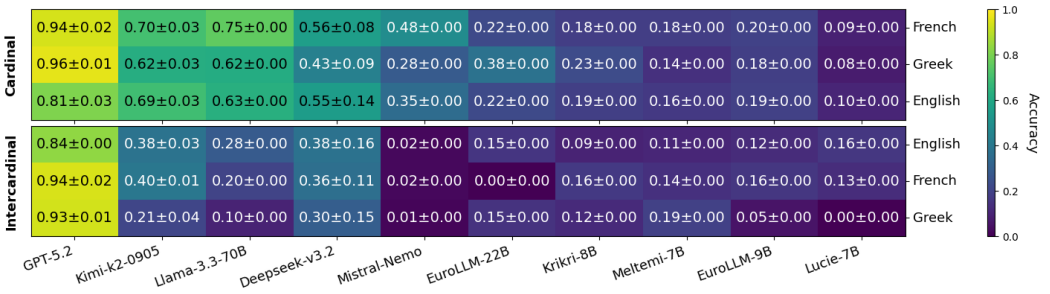

Figure 2 Benchmarking model accuracy: A cross-linguistic comparison of cardinal and intercardinal directions.

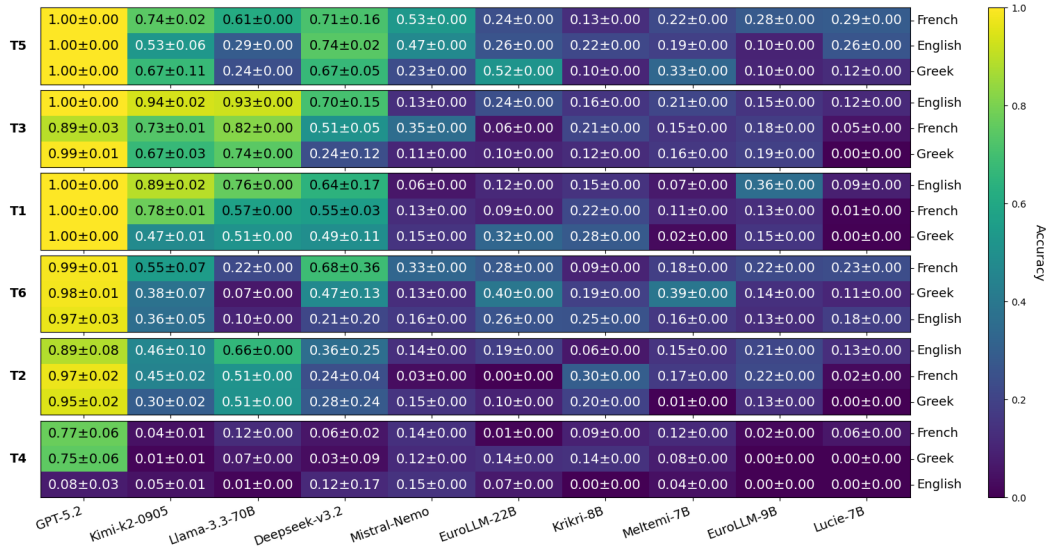

Figure 3 Benchmarking model accuracy: A cross-linguistic comparison by question template.