Evaluating the Capabilities of Large Reasoning Models to Reason about Cardinal Direction Relations

Orestis-Minas Kapopoulos[1] , **Vasileios Kyriakopoulos** [2] , **Maria Tsourma**[2] , **Despina-Athanasia Pantazi**[2] , **Manolis Koubarakis**[2,3]

[1]Dept. of Informatics, King's College London, London, UK

[2]Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens

[3]Archimedes, Athena Research Center

orestis.kapopoulos@kcl.ac.uk, cs22200017@di.uoa.gr, mtsourma@iti.gr, dpantazi@di.uoa.gr, koubarak@di.uoa.gr

Abstract

Spatial reasoning has been an established area of research since around 1990 and a lot of interesting related work has been done by researchers in Geography, AI and Databases. With the arrival of large language models, there has also been research on evaluating their spatial reasoning capabilities. In this paper, we contribute to this emerging research area and evaluate 5 large reasoning models (GPT-5.4, gemini-3-flash-preview, phi-4, llama-4-maverick and magistral-small) on the task of computing the transitivity table of the well-known cardinal direction calculus.

1 Introduction

We follow the lead of [Cohn and Blackwell, 2024b; Cohn and Blackwell, 2024a] and our own work [Gardelakos *et al.*, 2025] and evaluate the capabilities of 5 large reasoning models on the task of computing the transitivity table for the cardinal directions calculus CDC of [Goyal and Egenhofer, 1997; Goyal and Egenhofer, 2001; Skiadopoulos and Koubarakis, 2004]. CDC is based on 9 relations: the cardinal directions North, South, East and West, the intercardinal directions Northeast, Southeast, Northwest and Southwest, and the relation Bounding Box which captures that the reference object falls inside the bounding box of the target object.[1]

Large reasoning models (LRMs) [Liu *et al.*, 2025] are the most recent breed of large language models (LLMs) that have shown strong capabilities in coding tasks and in solving logical and arithmetic problems (see e.g., OpenAI's o1 model [Jaech *et al.*, 2024] and Deep Mind's Gemini Advanced 2.5 model[2]). In this paper we use the commercial models GPT-5.4, Gemini and Claude, and the open-weight models Phi-4, Llama-4 and Magistral-small. We chose these models since they are all state of the art in the commercial and open weight LRMs arena, and have not been tried in our previous study [Gardelakos *et al.*, 2025].

We perform 3 experiments for the task at hand. In the first experiment, we prompt the models with a complete definition of the CDC calculus. In the second experiment, following the methodology of [Cohn and Blackwell, 2024a], we use a similar prompt with the first experiment but now with anonymized context. The last experiment evaluates how the models perform without providing any context regarding the CDC calculus in the prompt (i.e., how they perform based only on the knowledge of cardinal direction relations obtained during their pretraining and stored in their parameters).

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents our experiments. Finally, 4 concludes the paper.

2 Related Work

[Cohn and Blackwell, 2024b] investigates whether LLMs can reason about cardinal direction relations. Their methodology is to create two question and answer datasets which they call *small* and *large*. To create small, they have used ChatGPT to co-create 100 simple questions where the answer is a cardinal direction (North, South, East and West). Two example questions are: *"You are watching the sun set. Which direction are you facing?"* and *"If the South Pole is behind you, which direction are you facing?"*. The second dataset is generated from a set of templates and it is meant to test comprehensively an LLM's ability to determine the correct cardinal direction given a particular scenario. An example question template from this dataset is *"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?"*. Even with a temperature setting of 0, the experiments of [Cohn and Blackwell, 2024b] demonstrate that although LLMs are able to perform well in the small (simpler) dataset, in the second more complex dataset, no LLM is able to reliably determine the correct cardinal direction.

[Cohn and Blackwell, 2024a] investigates whether LLMs can reason about mereotopological relations in RCC-8. They concentrate on three problems: getting an LLM to reconstruct

[1]The definitions of these relations are given in Section 3.1 of [Skiadopoulos and Koubarakis, 2004] so we do not repeat them here. In that paper they are called single-tile relations to distinguish them from multi-tile relations. The 9 relations are defined for regions that are homeomorphic to the closed unit disk in $\mathbb{R}^2$ and regions that are formed by finite unions of these regions.

[2]https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/

the transitivity table of RCC-8, evaluating the alignment of LLM answers to human composition preferences, and reconstructing the conceptual neighbourhood of RCC-8 relations. [Cohn and Blackwell, 2024a] show that LLMs do not perform well on the above tasks; this is not surprising given that, e.g., for the computing the transitivity table, even humans might have difficulty doing it.

[Gardelakos *et al.*, 2025] evaluates the capabilities of 7 commercial LRMs (the OpenAI models o1, o3, and o4-mini, deepseek-r1, grok-3-mini, claude-3.7-sonnet with extended thinking mode and gemini-2.5-flash) on the reasoning tasks of computing the transitivity table of RCC-8 and CDC. [Gardelakos *et al.*, 2025] shows that the model o3 outperforms the other 6 models in most of the experiments and achieves mean Jaccard index 99.66% and 99.57% for the CDC and the RCC-8 tasks respectively when the prompts included context.

Finally, the recent paper [Ji *et al.*, 2025] investigates whether LLMs can understand the widely-used Well-Known-Text representation of geometries and whether spatial relations (e.g., topological) are preserved during spatial reasoning when the corresponding vector data are passed to LLMs. [Ji *et al.*, 2025] experiments with GPT-3.5-turbo, GPT-4, and DeepSeek-R1-14B and reports their accuracy in the identification of spatial relations when using geometry embedding-based, prompt engineering-based, and everyday language-based approaches. In particular, the GPT-based reasoner studied can understand inverse topological spatial relations. In addition, GPT-4 can translate certain vernacular descriptions about places into formal topological relations, and adding the geometry-type or place-type context in prompts may improve inference accuracy, although this varies by instance.

3 Experimental Setup

We perform our experiments on five models summarized in Table 1. The open-weight models phi-4-reasoning-plus, llama-4-maverick and magistral-small-2506 are accessed via OpenRouter. All models are evaluated using their default hyper-parameters. Henceforth, the models will be referred to by their abbreviations as listed in Table 1. The experiments examine the ability of the chosen LRMs to perform compositional reasoning. This is to reconstruct the transitivity table of CDC. Each experiment is conducted 3 times per model. The prompts used are presented in Appendix A.

To evaluate the answers the models give, we employ Jaccard Similarity since each composition may result in more than one possible relation. The Jaccard index is computed as the size of the intersection of the predicted set and the expected set of relations, divided by the number of relations in the union of the two sets. When only a single relation is predicted and the expected answer is also a single relation, the Jaccard index reduces to a binary 1 or 0 measure of accuracy. Since each experiment is conducted 3 times per model, we report in Table 2 the mean and standard deviation of the Jaccard Similarity averaged over all compositions and all 3 runs.

To visualize the performance across the full transitivity table, we present heatmaps (Figures 1, 3, 5) where each cell represents a composition and is shaded according to its mean Jaccard Index across the 3 runs ranging from 0 (dark purple) to 1 (bright yellow). For example, in the Figure 1a the composition $B \circ B$ has an average Jaccard Index of 1 and is in color yellow, while the composition $N \circ NE$ achieves 0.1667 and has a dark blue color. The relations in the cells of the figures are the ground truth relations. For the sake of readability, we use the letter U to denote the case where the ground truth result includes all relations.

To further analyze the behavior of the models, we decompose the prediction into binary classification counts for each relation r (Figures 2, 4, 6). We count correctly-predicted (TP) meaning the model correctly predicted the relation r as part of the result, correctly-not-predicted (TN) meaning the model correctly excluded relation r from the result, incorrectly-predicted (FP) meaning the model predicted the relation r as part of the result but it should not have, and incorrectly-not-predicted (FN) meaning the model did not predict the relation r but it should have. For example, in Figure 2a the model predicts the relation North as part of the result correctly 45 times (TP), it excludes it from its prediction correctly 154 times (TN), incorrectly includes it 14 times (FP) and incorrectly excludes it 30 times (FN) in all 3 runs of the experiment, summing to $45 + 154 + 14 + 30 = 243 = 81 \times 3$ total decisions across all $9 \times 9 = 81$ compositions and 3 runs.

Model	With context	Without context	Anon. context
phi-4-reasoning-plus	37.54± 2.92	14.58±2.44	-
llama-4-maverick	60.04±0.74	20.08±1.5	18.17±1.01
magistral-small	40.07±2.5	16.23±1.5	-
GPT-5.4	99.25± 1.3	21.71±0.79	31.73 ±0.62
gemini-3-flash-preview	100±0	20.28±0.04	32.21±0.24

Table 2: Comparison of the mean and the standard deviation of the Jaccard index achieved for the CDC experiment in three settings: with context included in the prompt, without context and with anonymized context.

3.1 Open-Weight Models

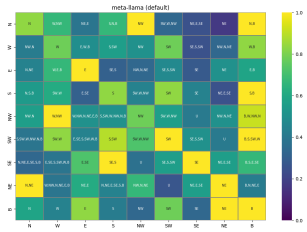
First Experiment. We begin by evaluating the ability of the open-weight models to reconstruct the transitivity table of CDC when provided with the definition of each relation as context. We present results for the open-weight models, phi-4-reasoning-plus, llama-4-maverick, and magistral-small using their default hyperparameters. As shown in Table 2, the models display varying degrees of success. Llama-4-maverick achieves the highest mean Jaccard index of 60.04%, followed by magistral-small at 40.07% and phi-4-reasoning-plus at 37.54%.

By inspecting Figure 1, llama-4-maverick shows many high-scoring cells, several in the diagonal $B \circ B$, $NE \circ NE$, $SE \circ SE$, $SW \circ SW$ and $E \circ E$, as well as some non-diagonal entries such as $NE \circ N$, $N \circ NW$, $N \circ B$. Magistral-small and phi-4-reasoning-plus produce largely dark heatmaps depicting the gap of approximately 20% of Jaccard index from the llama-4-maverick model. In particular, magistral-small achieves high scores only on $B \circ N$, $B \circ B$, $B \circ SW$ and phi-4-reasoning-plus on $NW \circ W$, $W \circ W$, $E \circ B$.

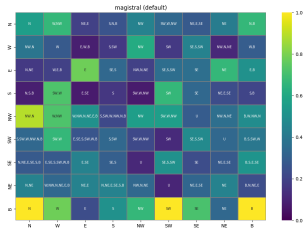
Figure 2 details each model's prediction behaviour across

Vendor	Model	Abbreviation	Released	API	Number of Parameters	Context
Microsoft	phi-4-reasoning-plus	phi-4-reasoning-plus	Apr 2025	OpenAI	14B	32K
Meta	llama-4-maverick-17B-128E-Instruct	llama-4-maverick	Apr 2025	OpenAI	400B(≈17B activated)	1M
Mistral	magistral-Small-2506	magistral-small	Jun 2025	OpenAI	24B	127K
OpenAI	GPT-5.4 Thinking	GPT-5.4	March 2026	OpenAI	Undisclosed	1M
Google	gemini-3-flash-preview	gemini-3-flash-preview	December 2025	Gemini API	Undisclosed	1M

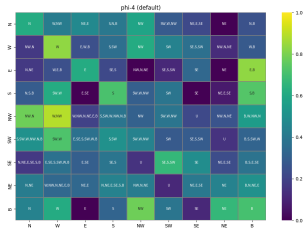
Table 1: LRMs tested. Context is the context window size in tokens. Models will be referred to by their abbreviation henceforth.



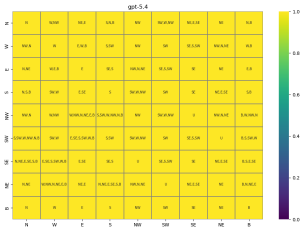
(a) llama-4-maverick



(b) magistral-small



(c) phi-4-reasoning-plus



(d) GPT-5.4

(e) gemini-3-flash-preview

Figure 1: CDC transitivity table shaded by mean Jaccard index for the experiment with context.

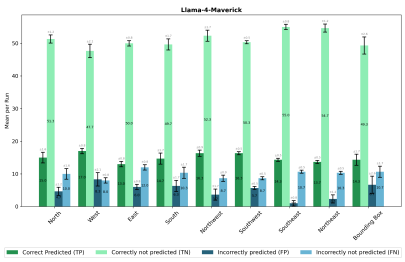
the nine spatial relations. For each relation, performance is decomposed into four classification outcomes: correctly-predicted (True Positives, TP), correctly-not-predicted (True Negatives, TN), incorrectly-predicted (False Positive, FP) and incorectly-not-predicted (False Negative, FN). Each bar reports the mean count per runs alongside its standard deviation (± std) to quantify variability across runs. This reporting format and metric layout are used consistently across all such figures.

All three models exhibit a pattern of high correctly-not-predicted (TN) counts relative to correctly-predicted (TP) counts, indicating that models are more confident in excluding a relation than in identifying one. For phi-4-reasoning-plus, the mean correctly-predicted (TP) values range from 5.7 (Bounding Box) to 18 (Southwest) against the mean correctly-not-predicted (TN) values of range 37.7 (West) and 51.3 (Southeast). Similarly, magistral-small, exhibits mean correctly-predicted (TP) values between 5.7 (Bounding Box) and 14.3 (Northwest) and correctly-not-predicted (TN) between 44 (North) and 49 (Southeast), while llama-4-maverick exhibits correctly-predicted (TP) values between 13 (East) and 17 (West) and correctly-not-predicted (TN) between 47.7 (West) and 55 (Southeast). Finally, the low standard deviation across runs ranging from $\pm$ 0.5 to $\pm$ 2.5 demonstrates that all three models exhibit stable and consistent prediction behaviors across iterations.

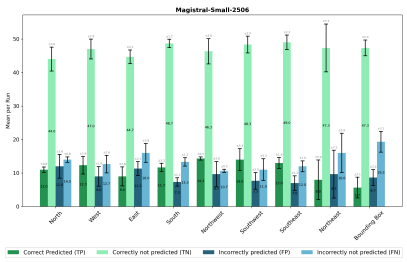
Second Experiment. Moving on we investigate whether the models can reconstruct the transitivity table without any contextual description of the cardinal direction relations and relying solely on their internal knowledge. As shown in Table 2, all three models exhibit a sharp drop in performance compared to the first experiment. Specifically, phi-4-reasoning-plus, llama-4-maverick, and magistral-small achieve mean Jaccard similarities of 14.58%, 20.08%, and 16.23% respectively. This drop indicates that models lack sufficient internal knowledge about cardinal direction composition to perform well without explicit guidance.

Figure 3 shows that all three models exhibit largely dark heatmaps reflecting the poor Jaccard indices they achieve. Llama-4-maverick exhibits high scores in cells $N \circ N$, $E \circ S$ and $S \circ S$ while magistral-small and phi-4-reasoning-plus only in one cell $N \circ N$ and $B \circ B$ respectively.

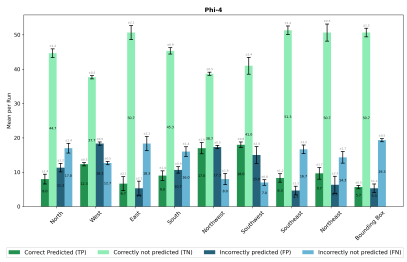
Figure 4 shows the same pattern we saw in the first experiment i.e., of the correctly-not-predicted (TN) counts remaining high compared to correctly-predicted (TP). For example, llama-4-maverick records 5 correctly-predicted (TP) in the Northeast direction, while correctly-not-predicted (TN) reach 38.7 in the same direction. In addition, the incorrectly-not-predicted (FN) and incorrectly-predicted (FP) counts rise, reflecting poor performance since the models frequently



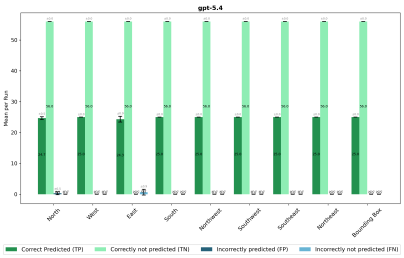
(a) llama-4-maverick



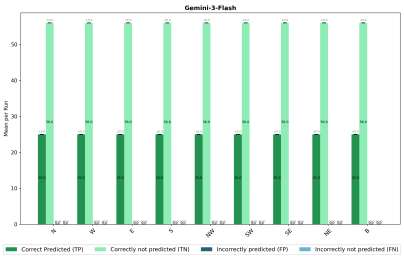
(b) magistral-small



(c) phi-4-reasoning-plus



(d) GPT-5.4



(e) gemini-3-flash-preview

Figure 2: Relation statistics for the CDC transitivity table computation for the experiment with context.

fail to identify correct compositions. For instance, phi-4-reasoning-plus records 26 incorrectly-predicted (FP) relations in the Southeast direction and 21.7 incorrectly-not-predicted (FN) in East, and llama-4-maverick reaches 22.7 incorrectly-predicted (FP) relations in Northwest and 21.3 incorrectly-not-predicted (FN) in Southeast. Similarly to the first experiment, the low standard deviation ranging from ± 0.5 to $\pm$ 2.9 indicate the models show stable and consistent behavior acroos the 3 runs.

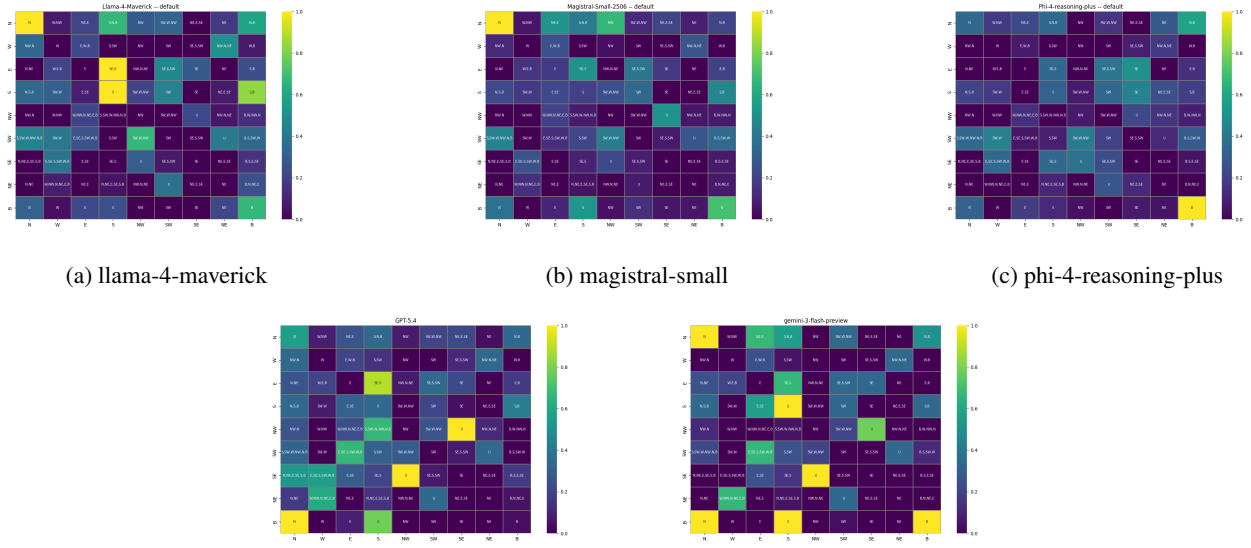
Third Experiment. Finally, we evaluate the ability of the models to reconstruct the transitivity table of CDC when provided with the definition of relations, but now using anonymous labels R1 through R9 instead of the cardinal directions names. Due to the discontinuation of support on OpenRouter for the models phi4-reasoning-plus and magistral-small, this experiment is done only for llama-4-maverick. As shown in Table 2, the Jaccard index drops even further in this experiment for llama-4-maverick to 18.17%. Notably, the Jaccard index is near the one of the second experiment, suggesting that good accuracy is achieved when a model is provided by both by the definition of cardinal directions and the recognition of named cardinal directions.

Figure 5a illustrates the above behaviour. While the first experiment (Figure 1) yielded bright cells at multiple positions, Figure 5a shows only a small number of high scoring entries such as $R9 \circ R4$, $R9 \circ R9$, with the vast majority remaining dark. Figure 6 shows the same pattern identified in Figures 2 and 4. The correctly-not-predicted (TN) count, which ranges from 33 ($R6$) to 55 ($R1$), is substantially greater than the correctly-predicted count (TP), which ranges from 1 ($R7$) to 12 ($R6$). Interestingly, the standard deviation indicators reveal severe instability in the model's incorrect predictions across runs. While successful predictions correcly-predicted (TP) and correctly-not-predicted (TN) remain relatively stable ($\pm$ 0 to $\pm$ 2.9), both incorrectly-predicted (FP) and incorrectly-not-predicted (FN) mean values exhibit massive variance such as incorrectly-not-predicted standard deviation of $\pm$ 10.6 in R7 and incorrectly-predicted (FP) standard deviation of $\pm$ 8.0 in R1. This suggests substantial run-to-run volatility specifically when the model makes errors.

3.2 Closed-Weight Models

First Experiment. We present results for the closed-weight models GPT-5.4 and gemini-3-flash-preview using their default hyperparameters. As shown in Figure 1, both models achieve near-perfect performance where gemini-3-flash-preview exhibits 100% and GPT-5.4 99.25% and their respective heatmaps appear visually identical. For both models, the heatmap shows a highly consistent pattern, with all cells appearing at the maximum similarity level. The diagonal cells correctly preserve the expected relations, as well as the off-diagonal cells which contain the expected composed relation sets. However, some cells contain broader relation sets in GPT-5.4, especially in compositions involving NW, SW, SE, and NE explaining the small standard deviation of $\pm 1.3\%$.

Figure 2 shows that both closed-weight models perform almost perfectly across the nine spatial relations. For GPT-5.4, the correctly-predicted (TP) mean counts are very high and stable, since the counts mean ranges from 24.3 for East to 25 for West, for example and standard deviation is at most $\pm$ 0.5. Gemini-3-flash-preview shows an even more uniform pattern, with 25 mean TP and 56 mean TN for every spatial relation. It produces no FP or FN errors across all categories, suggesting that in this setting the model reconstructs the expected relation sets with complete consistency. Overall, both models demonstrate very strong performance, but gemini-3-flash-preview achieves a perfectly stable result across all relations, while GPT-5.4 shows only minimal errors in North and



(a) llama-4-maverick (b) magistral-small (c) phi-4-reasoning-plus

(d) GPT-5.4 (e) gemini-3-flash-preview

Figure 3: CDC transitivity table shaded by mean Jaccard index for the experiment without context.

East.

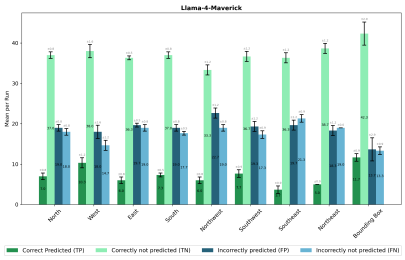
Second Experiment. As shown in Table 2, all models exhibit a sharp drop in performance compared to the first experiment, and the same occurs for gemini-3-flash-preview and GPT-5.4. This drop indicates that the models lack sufficient internal knowledge about cardinal direction composition to perform without explicit guidance.

According to Figure 3, gemini-3-flash-preview performs best on cardinal relations, especially N, S, and B, which appear with very high Jaccard scores on the diagonal. However, its performance is weaker and more inconsistent for intercardinal relations such as NW, SW, SE, and NE. The model often produces broad sets of relations instead of a single precise relation, creating thus several off-diagonal cells with moderate or high Jaccard scores, because part of the predicted set overlaps with the expected relation; suggesting overgeneralization and confusion between neighboring or opposite cardinal directions. On the other hand, the results of GPT-5.4 show that the diagonal is not consistently strong: while some expected relations such as $N \circ N$ and $B \circ B$ achieve higher Jaccard similarity, other diagonal cells appear darker, indicating weaker overlap between the expected and predicted relation sets. There are also several bright off-diagonal cells, such as combinations involving South, Northwest, and Southeast, suggesting that the model sometimes assigns high similarity to relation pairs that should not be the main match, indicating that GPT-5.4 does not reconstruct the spatial transitivity structure perfectly in this setting; instead, it shows partial understanding, with noticeable confusion between intercardinal directions and broader relation sets.

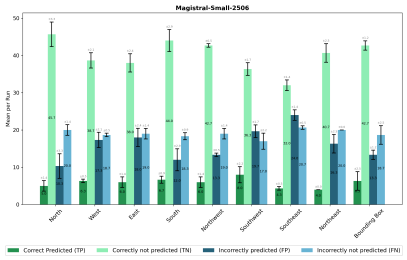
Figure 4 presents the prediction behaviour of GPT-5.4 and gemini-3-flash-preview across the nine spatial relations. GPT-5.4 (Figure 4d) achieves higher correctly-predicted (TP) mean counts than gemini-3-flash-preview for every relation, with its strongest results observed for South (17), North (16.3), Northwest (12.7), and East (12). This indicates that GPT-5.4 is better at identifying relations that should be included in the expected answer set. However, the model also produces a high number of incorrectly-predicted (FP) cases, especially for Southwest (29.7), Northeast (29.3), West (27.7), Northwest (26.3), and Southeast (26.3). This suggests a tendency to overpredict, meaning that GPT-5.4 often includes additional relations that are not part of the correct output.

In contrast, gemini-3-flash-preview has lower TP mean counts overall, ranging from 5.7 for Northeast to 10.7 for South, showing that it identifies fewer expected relations correctly (Figure 4e). It achieves much higher correctly-not-predicted (TN) counts, particularly for Bounding Box (52), South (47), North (46), and East (45), indicating that gemini-3-flash-preview is more conservative and better at excluding relations that should not be predicted. However, this leads to many incorrectly-not-predicted (FN) cases, especially for West and Northeast (18.3), East (17.7), and Southeast and Bounding Box (17.3). Overall, GPT-5.4 shows stronger recall of the expected relations but tends to overgenerate, whereas gemini-3-flash-preview is more restrictive, producing fewer false positives but missing many correct relations.

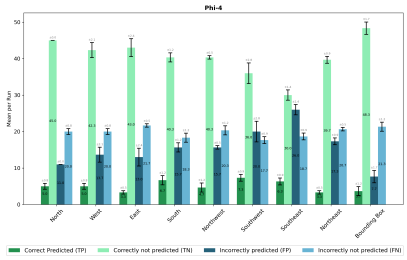
Morever, gemini-3-flash-preview exhibits stable and consistent prediction behavior shown by low standard deviation ranging from $\pm$ 0.5 to $\pm$1.4. In contrast, gpt-5.4 displays more variable behavior since standard deviation ranges from $\pm$ 0.5 to $\pm$ 2.4. Specifically, it exhibits low standard deviation on the correctly-predicted (TP) and incorrectly-not-predicted (FN) but high standard deviation on correctly-not-predicted (TN) and incorrectly-predicted (FP) which indicates that GPT-5.4 maintains stable behaviour when identifying ground-truth relations, but highly variable behavior re-



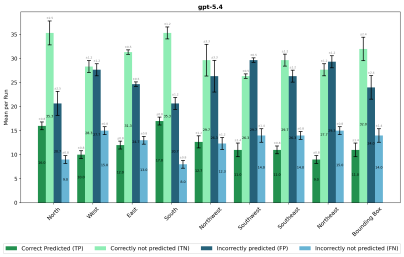
(a) llama-4-maverick



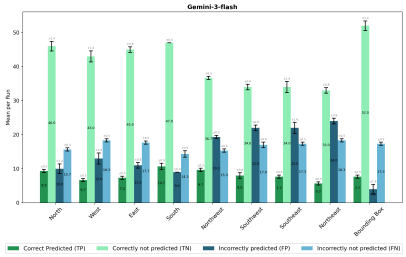
(b) magistral-small



(c) phi-4-reasoning-plus

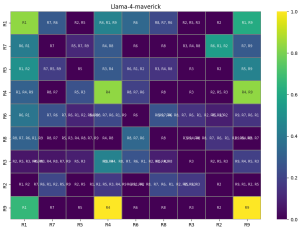


(d) GPT-5.4

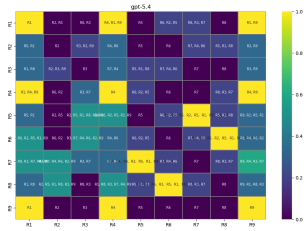


(e) gemini-3-flash-preview

Figure 4: Relation statistics for the CDC transitivity table computation for the experiment without context.



(a) llama-4-maverick



(b) GPT-5.4

(c) gemini-3-flash-review

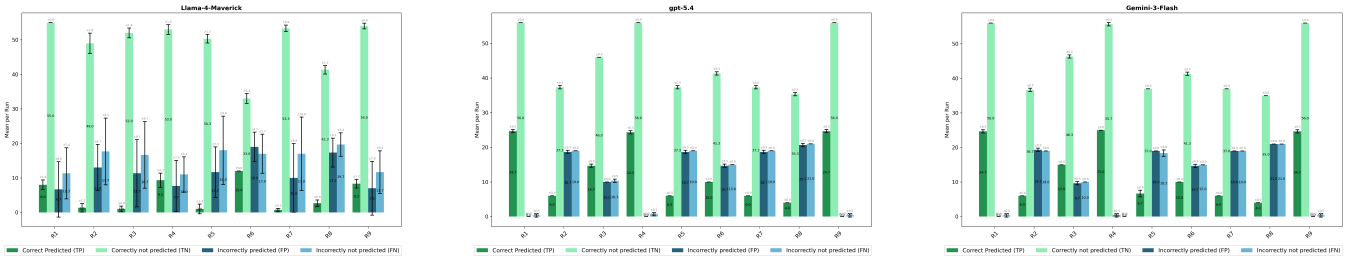
Figure 5: CDC transitivity table shaded by mean Jaccard index for the experiment with anonymized relations.

garding how many incorrect relations it falsely predicts across different samples.

Third Experiment. Finally, we evaluate the ability of GPT-5.4 to reconstruct the transitivity table of cardinal directions provided with a textual description but using anonymous labels R1 through R9, for the cardinal directions names. As shown in Table 1, the Jaccard Index increases slightly in this case, compared to the second experiment. A likely explanation is that the anonymization in the second experiment removes the semantic cues that the model relies on. In this experiment the models may already have learned during pretraining that combinations such as $N \circ E$ result in relations involving NE, or that cardinal directions compose in certain intuitive ways. Therefore, it can use its general spatial knowledge and memorized/learned associations with cardinal directions to produce answer sets that overlap more with the expected CDC transitivity table cells.

Based on the results illustrated in Figure 5b, several cells reach a Jaccard index of 1.0, especially in cases involving identity-like or highly intuitive relations, such as $R1 \circ R1$, $R4 \circ R4$, $R9 \circ R9$, and some combinations involving $R1, R4, R7$, and $R8$. This indicates that the model can recover parts of the transitivity table even without being given the CDC definitions in the prompt. However, the heatmap also contains many low-similarity cells, especially around relations such as $R2, R3, R5, R6$, and $R8$, where the predicted sets only partially overlap with the expected sets or fail to overlap at all. This suggests that the model does not reconstruct the full algebra systematically, while instead, it seems to recover some relations through pretraining-based spatial intuition, while struggling with less direct or more abstract compositions.

Gemini-3-flash-preview is still able to recover parts of the transitivity structure, but its performance is uneven (Figure 5c). Several cells reach a Jaccard index of 1.0, while some off-diagonal cells also obtain very high similarity, suggesting that the model captures some repeated structural patterns among these anonymous relations, even though their original semantic meaning is hidden. Several cells contain long relation sets showing uncertainty: instead of predicting a precise



(a) llama-4-maverick (b) GPT-5.4 (c) gemini-3-flash-review

Figure 6: CDC transitivity table shaded by mean Jaccard index for the experiment with anonymized relations.

(a) Full Context (b) Anonymized Context (c) No Context

Figure 7: CDC transitivity table shaded by Jaccard index standard deviation for all experiments for llama4-maverick model

composition result, the model sometimes includes many possible relations. Such broad predictions can increase the Jaccard index when part of the expected answer is included, but they also reveal that the model is not confidently identifying the exact transitivity relation.

Figure 6b further clarifies this behaviour across the nine relations where the strongest classes are North, South, and Bounding Box. GPT-5.4 correctly-predicted (TP) mean counts 24.7, 24.3 and 24.7 correspondingly, 56 correctly-not-predicted (TN) cases and 0 FP and 0 FN an these relations, indicating a stable behaviour for these relations. Based on these results, the third experiment suggests that GPT-5.4 has some pretrained knowledge of cardinal direction composition, particularly for the main cardinal directions and bounding-box relation, but this knowledge is uneven.

For gemini-3-flash-preview, the bar chart confirms the mixed behaviour observed in the heatmap (Figure 6c). The model performs accurately on R1, R4, and R9. These results show that gemini-3-flash-preview can reconstruct some anonymized relations almost perfectly, especially the relations that appear to have a stable or identity-like role in the transitivity table. However, the performance drops considerably for the remaining relations, suggesting that the model does not reconstruct the anonymized transitivity table in a fully systematic way, since it can recover some relations with very high accuracy, especially R1, R4, and R9, but it struggles with several other anonymized relations, indicating both overprediction and omission of correct relations.

Our final tables are shown in Figure 7. There we visualize the mean and standard deviation of the Jaccard index for every composition as a heatmap following similar strategy as the generation of mean Jaccard index heatmaps shown earlier. Each cell shows the mean of the jaccard index across all 3 repeats of the corresponding experiment and their standard deviation. Due to space limitations we show the heatmaps only for llama4-maverick model. The heatmaps show that as the context decreases, the standard deviations in the three tables decrease as well. More specifically, the full context experiment has higher standard deviations than the anonymized context experiment which also has higher standard deviations than the experiment without context. This indicates that the more context we give the model about the CDC calculus (or any information we need it to handle in general), the more knowledge it has to handle, which leads to more randomness of the given answers and thus higher standard deviation.

4 Summary

We studied the capabilities of 5 large reasoning models (the closed-weight commercial models GPT-5.4 and gemini-3-flash-preview, and the open-weight models phi-4, llama-4-maverick and magistral-small) on the task of computing the transitivity table of the cardinal direction calculus CDC defined in [Goyal and Egenhofer, 1997; Goyal and Egenhofer, 2001; Skiadopoulos and Koubarakis, 2004]. We performed three experiments. The first experiment evaluated the accuracy of the models when prompted with the exact definition of the cardinal direction relations of CDC. The second experiment evaluated the accuracy of the models based only on the knowledge of cardinal direction relations stored on their parameters. The third experiment was like the first one, but, this time, relations were anonymized. The best results we got were in the first experiment; for the second and third exper-

iment, there were significant drops in accuracy of the tested models. In the first experiment, the closed-weight commercial models gemini-3-flash-preview and GPT-5.4 achieved perfect and almost perfect accuracy respectively, while the open-weight models scored less satisfactorily, with llama-4-maverick scoring better than magistral-small and phi-4-reasoning-plus.

In future work we plan to experiment with other closed-weight and open-weight models and, most importantly, with more qualitative calculi, e.g., the one which combines topological and cardinal direction information [Cohn *et al.*, 2014]. This paper used only the functionality of LLMs to tackle the spatial reasoning problem studied. Another approach is also possible and it is illustrated in the recent paper [Georgiadis *et al.*, 2026]. In that paper, the authors use a graph RAG approach in which the LLM together with an external reasoning mechanism based on a transitivity table is used to answer questions involving qualitative spatial relations.

References

[Cohn and Blackwell, 2024a] A. G. Cohn and R. E. Blackwell. Can large language models reason about the region connection calculus? *CoRR*, abs/2411.19589, 2024.

[Cohn and Blackwell, 2024b] A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions (short paper). In *16th International Conference on Spatial Information Theory, COSIT 2024, September 17-20*, 2024.

[Cohn *et al.*, 2014] A. G. Cohn, S. Li, W. Liu, and J. Renz. Reasoning about topological and cardinal direction relations between 2-dimensional spatial objects. *J. Artif. Intell. Res.*, 51:493–532, 2014.

[Gardelakos *et al.*, 2025] E. Gardelakos, V. Kyriakopoulos, D. Pantazi, O. Kapopoulos, M. Tsourma, and M. Koubarakis. Can large reasoning models reason about spatial relations? In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2025, Minneapolis, MN, USA, November 3-6, 2025*. ACM, 2025.

[Georgiadis *et al.*, 2026] Thanasis Georgiadis, John Pavlopoulos, and Nikos Mamoulis. Sparagraph: Spatial reasoning using retrieval-augmented generation. *ACM Trans. Spatial Algorithms Syst.*, February 2026.

[Goyal and Egenhofer, 1997] R. Goyal and M.J. Egenhofer. Similarity of cardinal directions. In *The Annual Assembly and the Summer Retreat of University Consortium for Geographic Information Systems Science*, 1997.

[Goyal and Egenhofer, 2001] R. K. Goyal and M. J. Egenhofer. Similarity of cardinal directions. In *Advances in Spatial and Temporal Databases, 7th International Symposium, SSTD 2001, Redondo Beach, CA, USA, July 12-15, 2001, Proceedings*, volume 2121 of *Lecture Notes in Computer Science*, pages 36–58. Springer, 2001.

[Jaech *et al.*, 2024] A. Jaech, A. Kalai, A. Lerer, and A. Richardson et al. OpenAI o1 system card. *CoRR*, abs/2412.16720, 2024.

[Ji *et al.*, 2025] Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *IJGIS*, page 1–38, June 2025.

[Liu *et al.*, 2025] Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. Efficient inference for large reasoning models: A survey, 2025.

[Skiadopoulos and Koubarakis, 2004] S. Skiadopoulos and M. Koubarakis. Composing cardinal direction relations. *Artif. Intell.*, 152(2):143–171, 2004.

A Prompts used in the Experiments

In this section we present the prompts used in the experiments of the previous section. We start with the prompt used for the experiments for reconstructing the transitivity table. The prompt starts with some background information which essentially defines the CDC calculus and its basic relations and then questions are posed to compute the transitivity table.

Background information (context): `Given a region a, the greatest lower bound or infimum of the projection of a on the x-axis (resp. on the y-axis) is denoted by infx(a) (resp. infy(a)). The least upper bound or the supremum of the projection of a on the x-axis (resp. on the y-axis) is denoted by supx(a) (resp. supy(a)). These bounds define the minimum bounding box of a region a, which is the box formed by the straight lines x=infx(a), x=supx(a), y=infy(a) and y=supy(a). Let us now consider regions that are homeomorphic to the closed unit disk` $\{(x, y)\colon x^2 + y^2 <= 1\}$. `The set of these regions will be denoted by REG. Regions in REG are closed, connected and have connected boundaries. A cardinal direction relation between regions in REG is one of the following relations: B (bounding box), S (South), SW (South West), W (West), NW (North West), N (North), NE (North East), E (East) and SE (South East). These relations are defined as follows: a B b if and only if infx(b) <= infx(a), supx(a) <= supx(b), infy(b) <= infy(a) and supy(a) <= supy(b). a S b if and only if supy(a) <= infy(b), infx(b) <= infx(a) and supx(a) <= supx(b). a SW b if and only if supx(a) <= infx(b) and supy(a) <= infy(b). a W b if and only if supx(a) <= infx(b), infy(b) <= infy(a) and supy(a) <= supy(b). a NW b if and only if supx(a) <= infx(b) and supy(b) <= infy(a). a N b if and only if supy(b) <= infy(a), infx(b) <= infx(a) and supx(a) <= supx(b). a`

NE b if and only if supx(b) <= infx(a) and supy(b) <= infy(a). a E b if and only if supx(b) <= infx(a), infy(b) <= infy(a) and supy(a) <= supy(b). a SE b if and only if supx(b) <= infx(a) and supy(a) <= infy(b). You are a helpful assistant. I will now give you a question regarding the cardinal direction relations I defined above. The possible answer can be one or more of N, NE, SE, S, E, NW, W, SW, B. No yapping.

Question. Let R_1 and R_2 be cardinal direction relations. If region x is R_1 of region y and region y is R_2 of region z, then which could the possible relations between region x and region z be?.

In the above question, variables R_1 and R_2 are relations from the set { N, NE, SE, S, E, NW, W, SW, B }.

Let us now give the prompt for the experiment with anonymized relations.

Backround information (context): Let a be a region on the Cartesian plane. We define functions f, g, h and g that apply to regions and return real values. The straight lines x=f(a), x=g(a), y=h(a) and y=i(a) define a box in the Cartesian plane. Let us now consider regions that are homeomorphic to the closed unit disk $\{(x, y) : x^2 + y^2 <= 1\}$. The set of these regions will be denoted by REG. Regions in REG are closed, connected and have connected boundaries. A relation between regions in REG is one of the following: R1, R2, R3, R4, R5, R6, R7, R8 and R9. These relations are defined as follows (a and b are regions as above). a R9 b if and only if f(b) <= f(a), g(a) <= g(b), h(b) <= h(a) and i(a) <= i(b). a R4 b if and only if i(a) <= h(b), f(b) <= f(a) and g(a) <= g(b). a R6 b if and only if g(a) <= f(b) and i(a) <= h(b). a R2 b if and only if g(a) <= f(b), h(b) <= h(a) and i(a) <= i(b). a R5 b if and only if g(a) <= f(b) and i(b) <= h(a). a R1 b if and only if i(b) <= h(a), f(b) <= f(a) and g(a) <= g(b). a R8 b if and only if g(b) <= f(a) and i(b) <= h(a). a R3 b if and only if g(b) <= f(a), h(b) <= h(a) and i(a) <= i(b). a R7 b if and only if g(b) <= f(a) and i(a) <= h(b). You are a helpful assistant. I will now give you a question regarding cardinal direction relations. The possible answer can be one or more of R1, R2, R3, R4, R5, R6, R7, R8, R9. No yapping.

Question: Let S_1 and S_2 be relations. If region x is in relation S_1 with region y and region y is in relation S_2 with region z, then which could the possible relations between region x and region z be?.

In the above question, variables S_1 and S_2 are relations from the set {R1, R2, R3, R4, R5, R6, R7, R8, R9 }.

Let us now give the prompt for the experiment where we want to test whether LRMs can construct the transitivity table of CDC based only on their parametric knowledge.

Background information (context): You are a helpful assistant. I will give you a question regarding cardinal direction relations. The possible answer can be one or more of these: North, Northeast, Southeast, South, East, Northwest, West, Southwest, Bounding Box. No yapping.

Question: Let R_1 and R_2 be cardinal direction relations. If region x is R_1 of region y and region y is R_2 of region z, then which could the possible relations between region x and region z be?.

In the above question, variables R_1 and R_2 are relations from the set { North, Northeast, Southeast, South, East, Northwest, West, Southwest, Bounding Box }.