Can Large Reasoning Models Reason about Spatial Relations?

Eirinaios Odysseas Gardelakos*
Vasileios Kyriakopoulos*
cs2210008@di.uoa.gr
cs22200017@di.uoa.gr
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece

Despina-Athanasia Pantazi
dpantazi@di.uoa.gr
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens and
Archimedes/Athena Research Center
Athens, Greece

Orestis-Minas Kapopoulos
sdi2000066@di.uoa.gr
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece

Maria Tsourma
mtsourma@iti.gr
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens and
Information Technologies Institute
Centre for Research and Technology Hellas
Athens, Thessaloniki, Greece

Manolis Koubarakis
koubarak@di.uoa.gr
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens and
Archimedes/Athena Research Center
Athens, Greece

Abstract

Spatial reasoning has been an established area of research since around 1990 with the proposal of the 9-intersection model and RCC-8. Since then, a lot of interesting related work has been carried out by researchers in Geography, AI and Databases. With the arrival of large language models, there has also been research on evaluating their geospatial knowledge and spatial reasoning capabilities, and extending them so that they can be used to solve spatial reasoning problems. In this paper, we contribute to this emerging area of research and evaluate 7 large reasoning models (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 tasks of reasoning with topological and cardinal direction relations.

CCS Concepts

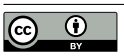
• **Computing methodologies** → **Spatial and physical reasoning**.

Keywords

spatial reasoning, topological relations, RCC-8, cardinal direction relations

ACM Reference Format:
Eirinaios Odysseas Gardelakos, Vasileios Kyriakopoulos, Despina-Athanasia Pantazi, Orestis-Minas Kapopoulos, Maria Tsourma, and Manolis Koubarakis. 2025. Can Large Reasoning Models Reason about Spatial Relations?. In *The 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3764912.3770841

*Both authors contributed equally to this research.



GeoAI '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2179-3/2025/11
https://doi.org/10.1145/3764912.3770841

1 Introduction

Spatial reasoning is a critical element of human cognition and has many real-world applications, for example, in Geographic Information Systems [10] (used to define the concepts of the common sense geographic world upon which such systems can be developed), Image and Vision systems (video analysis and spatial representation of moving objects) [7], natural language understanding (used for the interpretation of spatial issues included in natural language text such as the fitting of objects into containers) [8] and Robotics (used for the enhancement of the robot perception) [44]. Spatial reasoning has been an established area of research since around 1990 with the proposal of the 9-intersection model by Egenhofer [9] and the Region Connection Calculus [55] by Randell, Cohn and Cui. Since then a lot of interesting related work has been carried out by researchers in Geography, AI and Databases (see survey papers [6, 56]).

With the arrival of large language models (LLMs), there has been recent interesting work on evaluating their geospatial knowledge [1, 15, 16, 21, 24, 26, 28, 29, 31, 32, 34, 36–39, 51–53, 57] and spatial reasoning capabilities [4, 5], and extending them so that they are able to solve spatial reasoning problems for various proposed benchmarks [49, 50]. The questions studied in these works are important since LLMs have been advertised for their powerful reasoning capabilities by many and even as one possible avenue for achieving Artificial General Intelligence (AGI) by some. Another related avenue of research is the development of foundation models

for multimodal geospatial data [3, 17–19, 23, 33, 36, 40–42]. Collectively, these recent studies may be regarded as contributing to the emerging area of Geospatial Artificial Intelligence (GeoAI) [20].

In this paper, we follow the lead of [4, 5] and evaluate the capabilities of seven large reasoning models on two spatial reasoning tasks. *Large reasoning models (LRMs)* [35] are the most recent breed of LLMs that have shown strong capabilities in coding tasks and in solving logical and arithmetic problems (see e.g., OpenAI's o1 model [14] and Deep Mind's Gemini Advanced 2.5 model[1]). In this paper we use six commercial models (the OpenAI models o1, o3, and o4-mini, grok-3-mini, claude-3.7-sonnet with extended thinking mode and gemini-2.5-flash) and one open-weight model (deepseek-r1). We choose these models since they represent the state of the art in LRMs and have not been evaluated in prior research on the spatial reasoning tasks of interest. We mainly chose commercial LRMs as opposed to open-weight LRMs (e.g., the Llama 3 family of models [11]) since they are known to exhibit top performance on benchmarks for arithmetic, logical reasoning and coding (see for example [13] and [43]).

The first spatial reasoning task we investigate is reasoning with cardinal directions based on the calculus CDC [45, 46, 59] and, in particular, computing the transitivity (or composition) table for CDC. We perform 3 experiments for this task. In the first experiment, we prompt the models with a complete definition of the CDC calculus. In the second experiment, following the methodology of [4], 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 second task that we study is reasoning based on the calculus RCC-8 [55] following the ideas of [4]. Again, we investigate whether the chosen LRMs can correctly compute the composition table. We perform 2 experiments; in the first one we prompt the models with the definition of RCC-8, while in the second one we use a similar prompt with the first experiment but now with the context provided to the model anonymized.

Our results on the composition table of CDC are original; this problem has not been studied in the literature so far. Our results on RCC-8 complement the results of [4]; in this study we focus on more recent and more powerful LRMs and we demonstrate that these perform better than the LLMs tested in [4].

In more detail, our experimental results show the following contributions:

- For both tasks introduced above, the tested LRMs significantly outperformed the LLMs tested previously in the literature (where, for example, for reconstructing the transitivity table of RCC-8 [4] claude-3-5-sonnet was the best performing model). In our experiments, we used the updated version claude-3-7-sonnet, which was outperformed in most cases by the model o3 which achieves a mean Jaccard index of 99.41% and 99.66% for the CDC and the RCC-8 tasks respectively when the prompts included context. o3 outperforms the other 6 models in most of the other experiments as well.
- In the second experiment of both tasks, where the context of the prompt was anonymized, the evaluated LRMs could provide answers but not as accurate as in the first experiment. The observed performance drop suggests that the LRMs struggle to accurately interpret anonymized relations, likely due to the absence of context information in the prompt. If no context is provided in the prompt and the models are expected to rely on knowledge encoded in their parameters, the models perform poorly (we conducted this experiment only for the CDC task).
- For both tasks, we examined the presence of inductive bias in the LRMs when predicting specific relations. We conclude that no bias is present in the process of predicting any specific spatial relation, as the correct and incorrect predictions are shown to be almost equally distributed.

The remainder of the paper is organized as follows. Section 2 discusses the related work, section 3 analyzes the technical details of our experimental evaluation, section 4 presents our results on the calculus CDC, while section 5 presents our results for the calculus RCC-8. Finally, in section 6 we report our conclusions and discuss future work.

Our code and supplementary materials are available as open source at https://github.com/AI-team-UoA/LRMs-spatial-reasoning.

2 Related Work

The introduction section set the stage for our work and introduced the main papers in the relevant research areas. In this section we discuss in some detail only papers [4, 5, 49, 50] that are most closely related to our work. The reader can also view the AAAI 2025 invited plenary talk by Tony Cohn[2].

[5] 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 the small dataset, 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 [5] 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.

[4] investigates whether LLMs can reason about mereotopological relations in RCC-8. They concentrate on three problems: getting

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

[2] https://underline.io/events/473/sessions/19422/lecture/115523-can-large-language-models-reason-about-spatial-informationquestion

an LLM to reconstruct 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. [4] 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 performing this task.

[49] studies the StepGame benchmark [58] as a means to test the spatial reasoning abilities of LLMs. The StepGame benchmark contains story-question pairs where the question refers to spatial relations between two specified entities. For example, the story can be *"J is diagonally above B to the right at a 45-degree angle."* and the question can be *"What is the spatial relation of agent J to agent B?"*. This is called 1-hop spatial reasoning in the benchmark but, as one can imagine, there can be stories and questions that require multi-hop reasoning to be answered. The contributions of [49] are the following. First, they point out some inconsistencies in the original StepGame benchmark and they correct them. Secondly, they adopt the approach of [47] and combine GPT-3 with a spatial reasoner implemented using Answer Set Programming (ASP) to solve instances of the StepGame benchmark. In this approach, GPT-3 is employed to parse spatial expressions into symbolic representations, which are then passed to the ASP reasoner to obtain an answer. This neuro-symbolic approach results in *perfect accuracy* in the benchmark. Thirdly, they customize the Tree-of-Thoughts approach of [30] to deal with the important subproblem of object-linking chain building for reasoning in StepGame. With this approach, which uses the LLM as a native spatial reasoner, [49] are able to demonstrate up to 90% accuracy even in complex tasks when the LLM used is GPT-4.

[50] first provides a detailed analysis of existing spatial reasoning benchmarks (bAbI [22], StepGame [58], SpartQA [27] and SpaRTUN [54]) and their limitations. Then, it develops a more realistic benchmark by developing scenarios derived from 3D simulation data and moving away from emphasizing logical expressions and going towards stories that mirror everyday communication. [50] also presents a consistency checking tool, developed using Constraint Programming, for evaluating whether a spatial relation predicted by an LLM is feasible, given a set constraints. The paper also tests the capabilities of various LLMs on the benchmark and demonstrates that GPT-4 achieves the best performance. Finally, the paper shows that all tested models face difficulties in multi-hop spatial reasoning scenarios and that their performance improves as the story's constraint graph becomes more complete.

Finally, the recent paper [48] 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 is passed to LLMs. [48] 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 Setting

This section presents the experimental setting of our work. We assume that the reader is familiar with the concepts of the *cardinal direction calculus CDC* [59] and the *region connection calculus RCC-8* [55].

The LRMs that we evaluated are shown in Table 1. Like their LLM predecessors, LRMs are also stochastic. Each LRM API provides various options for reducing stochasticity. Despite following the suggested actions to reduce the stochastically (using a fixed seed and setting the temperate to 0), some models produced varied outputs. This behaviour is not surprising since some APIs may add variability due to caching, load balancing, or dynamic model updates. In the following analysis, experiments in which neither the seed was set to a random number nor the temperature was set to 0 will be referred to as the *default* experiments. That is, these experiments were conducted using the default configuration specified in the documentation provided by each API vendor.

Regarding the configuration of the experiments with reduced stochasticity, the seed was set to a random number (same for all repetitions of each model/experiment) for every model tested except gemini-2.5-flash and claude-3.7-sonnet, where it was not supported. Furthermore, the temperature was set to 0 for all models, except o1, o3 and o4-mini, where it was not applicable. It should be mentioned that in order to set the temperature to 0 for claude-3.7-sonnet, we needed to disable the thinking mode of the model. In the following analysis, the experiments with reduced stochasticity will be referred to as the *fixed seed* experiments. In Sections 4 and 5 we quantify the uncertainty of the LRMs answers using prediction intervals as proposed in [2].

The experiments described in the following two chapters evaluate the *compositional reasoning* capabilities of the selected LRMs. To compute each cell of a composition table, the LRMs under evaluation were prompted using appropriate prompt templates. Each prompt template consists of two parts; the first part is the background information (context) which defines the relevant calculus and its basic spatial relations, and the second part is the question posed by the user to compute the composition table. Due to space constraints, the precise definitions of all the prompt templates constructed are provided in our GitHub repository.

Similarly to [4], since the answers of the prompts which reconstruct the cells can contain more than one spatial relation, we use the *Jaccard index* to compute the accuracy of the predicted response compared to the expected answer. The Jaccard index is computed by counting the size of the intersection of the predicted set of relations with the expected set of relations and dividing it 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.

We repeat each experiment 3 times for all models evaluated in Section 4, and at least 3 times in all the experiments conducted in Section 5. This approach allows us to calculate and employ the average prediction accuracy of each model as a comparative metric. All experiments were conducted using the APIs provided by the vendors of the evaluated models.

To quantify the variability introduced by repeated runs, we follow the statistical methodology of [2], which involves computing

Vendor	Model	Abbreviation	Released	API	Number of Parameters	Context
OpenAI	o1-2024-12-17-high	o1	Dec 2024	OpenAI	Undisclosed	200K
OpenAI	o3-2025-04-16-high	o3	Apr 2025	OpenAI	Undisclosed	200K
OpenAI	o4-mini-2025-04-16-high	o4-mini	Apr 2025	OpenAI	Undisclosed	200K
DeepSeek	deepseek-reasoner	deepseek-r1	May 2025	DeepSeek API	671B (≈ 37B activated)	128K
xAI	grok-3-mini	grok-3-mini	Feb 2025	xAI API	Undisclosed	131072
Anthropic	claude-3-7-sonnet-20250219-thinking	claude-3-7-sonnet	Feb 2025	Anthropic	Undisclosed	200K
Google	gemini-2.5-flash-preview-04-17	gemini-2.5-flash	Apr 2025	Gemini API	Undisclosed	1M In - 65536 Out

Table 1: The evaluated LRMs referred to by their abbreviation henceforth. Context is the context window size in tokens.

prediction intervals rather than relying solely on mean scores. The methodology of [2] can be described briefly as follows (we take definitions almost verbatim from [2]). Let q be the number of benchmark questions, and let $X_{i,j}$ denote the score for question i in repeat j (0 for incorrect answer and 1 for correct answer). The mean score for the j-th repeat is $\bar{x}_j = \frac{1}{q}\sum_{i=1}^{q} X_{i,j}$.

Let n be the number of repeats. The overall mean score across all repeats is $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} \bar{x}_j$.

To estimate the range in which a future repeat's mean score might fall, we use the prediction interval $\bar{x} \pm t_{\alpha/2,n-1} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{1}{n'}}$ where s is the sample standard deviation of the repeat means $\bar{x}_j$, $t_{\alpha/2,n-1}$ is the critical value from the Student's t-distribution with $n-1$ degrees of freedom, n' is the number of future repeats (typically set to 1), and α is the significance level (e.g., 0.05 for a 95% prediction interval).

As explained in [2], *a prediction interval is wider than its corresponding confidence interval.* A prediction interval provides an estimated range that future observations or their averages are expected to fall into, while a confidence interval indicates the range that is likely to contain the true population parameter derived from the sample data. Since we aim the reproducibility of benchmark scores, we use prediction intervals where $n' = n$ as in [2].

4 Reconstructing the Transitivity Table of the Calculus CDC using LRMs

In this section we evaluate whether the tested LRMs can reconstruct the transitivity table of the calculus CDC by performing three experiments. In the first experiment, we prompt the models with the definition of the CDC calculus and we evaluate their accuracy in reconstructing the transitivity table. In the second experiment, following the methodology of [4], we employ a prompt similar to that used in the first experiment, but with the contextual information provided to the model anonymized. In the final experiment, we examine the models' performance on the same task without providing contextual information in the prompts, relying solely on their pre-existing knowledge of cardinal direction relations.

First Experiment. For the first experiment, each model is prompted with background information (the context) that defines the CDC calculus using the default setting. Then, we pose to the models a set of questions which compute the cells of the transitivity table (one question per cell, for a total of $9 \times 9 = 81$ questions). An example of such a prompt is provided in the appendix A. The context information of each prompt concludes with the following sentences: "`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.`". The sentences about possible answers and "No yapping." were appended because we observed that the models' answers exhibited greater precision in this setting (also observed in the methodology of [5]). The questions posed have the following format: "`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}.

In this experiment o3 achieved the best accuracy across all 3 repeats, achieving 99.66 % mean Jaccard index, compared to the worst performing model, grok-3-mini, which achieved 89.13%. Figure 1 shows the mean Jaccard coefficient for 4 models. The models chosen to be further displayed were the best, worst and two models of different vendors, Google and Anthropic. We have detailed results for all the other models of Table 1 as well included in our GitHub repository, but we omit them due to space considerations. We also performed experiments using the 'fixed seed' configuration which is described in Section 3, but were not displayed again due to space limitations. However, all information regarding the results for each model is presented in Figure 4.

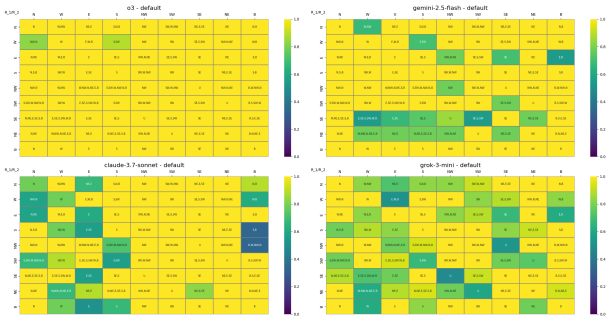


Figure 1: CDC transitivity table shaded by the mean Jaccard Index (n=3 repeats) for o3 (best performing model), gemini-2.5-flash, claude-3.7-sonnet and grok-3-mini (worst performing model), using the default input parameters.

Figure 2 depicts the total counts of all the answers provided by each model per each cardinal direction categorized as: correctly predicted (true positive), correctly not predicted (true negative),

incorrectly predicted (false positive) and incorrectly not predicted (false negative). The aim is to identify possible inductive biases of the models when predicting specific relations. We show the counts for the models o3 (best performing model), gemini-2.5-flash, claude-3.7-sonnet and grok-3-mini (worst performing model), using default input parameters. The model o3, which achieved 99.66% mean Jaccard index, has 1 incorrectly not predicted relation by predicting $W \circ N = NW$ instead of the correct $\{NW, N\}$, and 1 incorrectly predicted relation by predicting $W \circ S = \{S, SW, W\}$ instead of the correct $\{S, SW\}$. Upon closer examination of the first four bars for o3 (relation N), we can observe that the model predicted correctly the relation N 74 times across all answers for all 3 repetitions that were conducted. Thereafter, it correctly did not predict the relation N 168 times. Finally, it failed to predict the relation N 1 time, and it did not predict incorrectly the relation N at all (0 times). The same comments hold for the rest of the models regarding the default CDC experiment. Overall, Figure 2 illustrates the models' good performance on this experiment, as shown in Table 2. We can conclude the lack of bias in any relation, as the correctly and incorrectly predictions are somewhat equally distributed. However, we should mention that the worst model (grok-3-mini) has incorrectly not predicted the relation B 15 times, the most out of all the relations for all models, which indicates the model's incapacity to manage this relation.

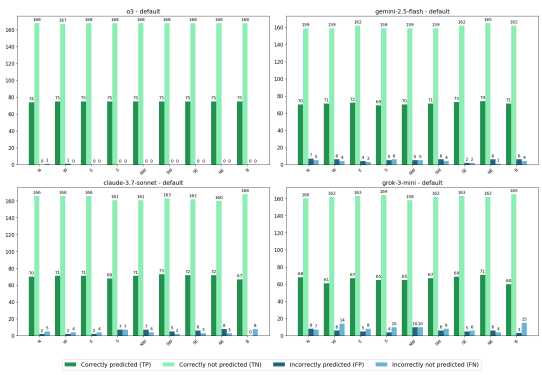


Figure 2: Relation statistics for the CDC transitivity table computation. for the models o3 (best performing model), gemini-2.5-flash, claude-3.7-sonnet and grok-3-mini (worst performing model), using the default input parameters.

An additional point of interest is the duration needed for each model to process a prompt. This is critically important in our setting since most of the models tested are commercial ones, hence, longer execution times incur higher fees. Figure 3 depicts the average time in seconds per prompt for each of the models. We can observe clearly that deepseek-r1 is the slowest, followed by claude-3.7-sonnet. This is primarily due to the large amount of reasoning tokens used by the model for each prompt. More specifically, 8280 reasoning tokens were used per each prompt on the average for deepseek-r1, using default parameters, compared to o3, which used 3535 i.e., less than half of deepseek-r1. The model claude-3.7-sonnet with fixed seed consumed the least amount of time because the thinking mode was disabled in order to be able to set a fixed seed.

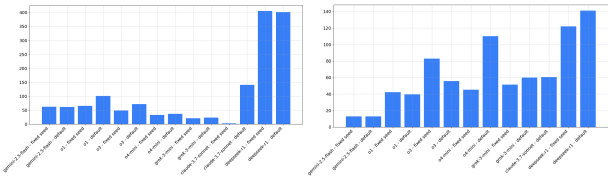


Figure 3: Average time (in seconds) per prompt when computing the CDC table (left) and the RCC-8 table (right).

Finally, Figure 4 shows a precise sense of the stochasticity of the tested models using prediction intervals as explained in Section 3. The figure includes two columns for each model; the left column shows the prediction interval of the mean Jaccard index for each repeat for the default setting, while the right column shows the same values when a fixed random seed and a temperature equal to zero are used for the models that this setting was permitted. For the models o1 and o3, the experiment with the fixed seed was repeated only twice, as our focus lies mainly to models with default values and the monetary cost of each repetition was high.

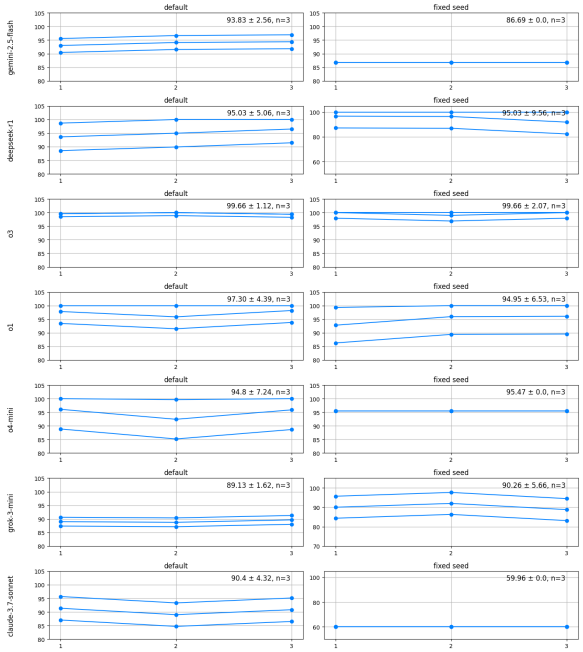


Figure 4: Prediction intervals for the CDC calculus.

Second Experiment. Our second experiment investigates whether the tested LRMs can reconstruct the transitivity table of the calculus CDC when prompted with anonymized CDC-specific definitions (e.g., the relation names). For comparison purposes, we present the results of the four LRMs discussed in the first experiment.

Figure 5 presents the transitivity table shaded by mean Jaccard index for these four LRMs. By comparing these results with the ones presented in Figure 1, we observe that all LRMs perform better when the complete context of the first experiment is provided. Specifically, gemini-2.5-flash achieved the highest mean Jaccard

index (86.14%) having achieved 74.15%, 92.11%, and 92.18% in each experiment, while the lowest was achieved by claude-3.7-sonnet (30.25%) as presented in Table 2. claude-3.7-sonnet achieved 30.12%, 35.16%, 25.47% in each of the experiments conducted. The grok-3-mini and o3 LRMs achieved mean Jaccard index between 68.00% and 62.56%, respectively. Generally, the evaluated LRMs could provide answers but not as accurately as in the first experiment. The performance drop observed suggests that the LRMs cannot interpret accurately anonymized relations without semantics due to the lack of background information on the prompt. Moreover, it should be highlighted that claude-3.7-sonnet failed to answer in some questions. This might be due to the lack of associations created between its prior knowledge and the given information.

As in the first experiment, Figure 6 examines the presence of inductive bias in the LRMs when predicting specific relations. The figure shows that the highest percentages correspond to the correctly not predicted label. Meanwhile, each LRM exhibits different distributions across the other three labels. Specifically, o3 has a higher percentage of incorrectly predicted in three out of the nine relations, while gemini-2.5-flash-default records the lowest false positive counts, with 18 instances in two relations. Also, gemini-2.5-flash shows better generalization by identifying more actual relations compared to claude-3.7-sonnet which has achieved very low correctly predicted counts (18-29), indicating difficulty in detecting positive compositions. From the incorrectly not predicted relations, we can conclude that this model misses several valid relations. We do not show graphs regarding the time consumption of the above experiments due to space considerations.

Third Experiment. We proceed to investigate if the models can reconstruct the transitivity table of CDC relying solely on the knowledge obtained during their pretraining. Comparing the results depicted in Figure 8 with the ones in Figure 1, we observe a significant drop in accuracy for all models. o3, gemini-2.5-flash, claude-3.7-sonnet and grok-3-mini achieved a mean Jaccard index of 21.27%, 21.4%, 19.9% and 16.6% respectively. This performance drop suggests that the tested models do not include prior knowledge about CDC, so if they are not prompted with the relevant context, they cannot provide correct answers. Finally, it is interesting to note that 3 out of 4 models, achieve 100% accuracy when computing $B \circ N$, $B \circ B$, $SE \circ NW$ and $NW \circ SE$ (presumably by chance).

Model	Context	No context	Anonymized
o3	99.66	21.27	62.56
gemini-2.5-flash	93.83	21.4	86.14
claude-3.7-sonnet	90.4	19.9	30.25
grok-3-mini	89.13	16.6	68.00

Table 2: Comparison of mean Jaccard index achieved for the CDC experiment in the three settings.

Similarly to Figure 2, Figure 8 examines the possibility of inductive bias when the models predict specific relations. We observe that all of the models exhibit relatively high percentages of correctly not predicted relations compared to correctly predicted ones. For instance, grok-3-mini returns 155 correctly not predicted relations versus only 6 correctly predicted relations in direction S,

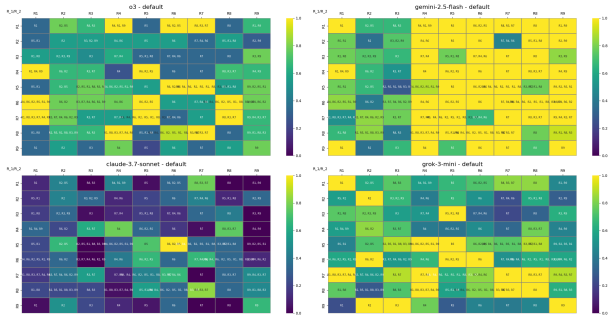


Figure 5: CDC transitivity table shaded by mean Jaccard index for the experiment with anonymized relations.

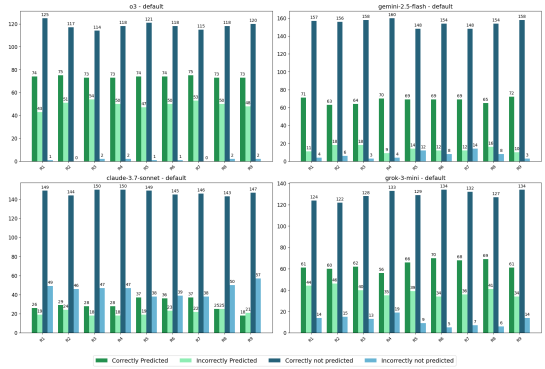


Figure 6: Relation statistics for the CDC transitivity table computation for the experiment with anonymized context.

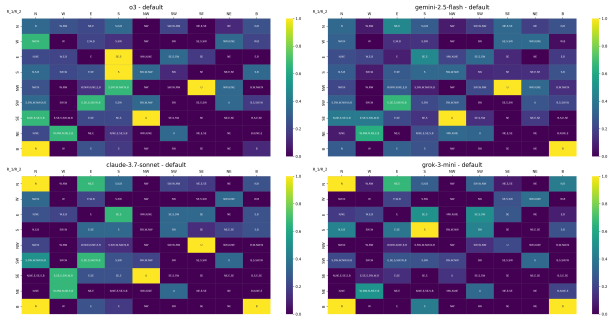


Figure 7: CDC transitivity table shaded by mean Jaccard index for the experiment without context.

while claude-3.7-sonnet concludes 47 correctly not predicted relations but only 7 correctly predicted relations for the direction B. In addition, we observe high percentages of incorrectly predicted relations, which indicates that the models make numerous incorrect predictions. In particular, gemini-2.5-flash shows 110 incorrectly predicted relations in the NE direction and 99 in the SE direction, while o3 reaches 32 incorrectly predicted relations in the SW direction and 21 in SW. These high incorrectly predicted values imply that the positive predictions issued are not always reliable.

o3	N	W	E	S	NW	SW	SE	NE	B
N	N	W, NW	NE, E	S, N, B	NW	SW, W, NW	NE, E, SE	NE	N, B
W	NW, N	W	E, W, B	W, S, SW	NW	SW	SE, S, SW	NW, N, NE	W, B
E	N, NE	W, E, B	E	SE, S	NW, N, NE	SE, S, SW	SE	NE	E, B
S	N, S, B	SW, W	E, SE	S	SW, W, NW	SW	SE	NE, E, SE	S, B
NW	NW, N	W, NW	W, NW, N, NE, E, B	S, SW, W, NW, N, B	NW	SW, W, NW	U	NW, N, NE	B, W, NW, N
SW	S, SW, W, NW, N, B	SW, W	E, SE, S, SW, W, B	S, SW	SW, W, NW	SW	SE, S, SW	U	B, S, SW, W
SE	N, NE, E, SE, S, B	E, SE, S, SW, W, B	E, SE	SE, S	U	SE, S, SW	SE	NE, E, SE	B, S, E, SE
NE	N, NE	W, NW, N, NE, E, B	NE, E	N, NE, E, SE, S, B	NW, N, NE	U	NE, E, SE	NE	B, N, NE, E
B	N	W	E	S	NW	SW	SE	NE	B

grok-3-mini	N	W	E	S	NW	SW	SE	NE	B
N	N	W, NW	NE, E	S, N, B	W, NW	SW, W, NW	NE, E, SE	NE	N, B
W	NW, N	W	W, SW, NW, B, E	SW, S	NW	SW	SW, S, SE	NW, N, NE	W, SW, NW, B
E	N, NE	W, E, B	E	SE, S	NW, N, NE	SW, SE, S	SE, E	NE	B, NE, SE, N, S, E
S	N, S, B	SW, W	E, SE	S	SW, W, NW	SW	S, SE	NE, E, SE	S, B
NW	NW, N	W, NW	N, W, NW, B, E, NE	NW, B, SW, N, S, W	NW	SW, W, NW	U	NW, N, NE	B, W, NW, N
SW	W, SW, S, B, N, NW	SW, W	B, W, SW, SE, S, E	W, SW, S, B	SW, W, NW	SW	SE, S, SW	U	B, S, SW, W
SE	N, NE, E, SE, S, B	SE, S, E, W, SW, B	SE, S, E	SE, S	U	SW, S, B, SE	SE	NE, E, SE	B, S, E, SE
NE	N, NE	NW, B, NE, N, E, W	N, E, NE	B, NE, SE, N, E, S	NE, N, NW	N, E, NE, NW, B, W, SW, SE, S	NE, E, SE	NE	B, NE, SE, N, E
B	N	W	E	S	NW	SW	SE	NE	B

Table 3: The CDC transitivity table for the best (o3) and the worst (grok-3-mini) performing models when the context is used.

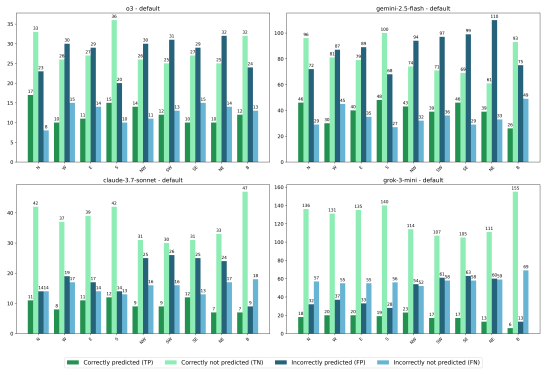


Figure 8: Relation statistics for the CDC transitivity table computation for the experiment without context.

Finally, the transitivity table created by the best (o3) and worst (grok-3-mini) performing models for the first repetition of the experiment with context used and default input parameters is depicted in Table 3. Each answer provided is highlighted with black for correctly predicted, red for incorrectly predicted and blue for incorrectly not predicted answers.

5 Reconstructing the Composition Table of RCC-8 using LRMs

In this section we perform two experiments to evaluate whether the tested LRMs of Table 1 can reconstruct the RCC-8 composition table. In the first experiment, following the methodology of [4], we prompt the models with the definition of each one of the 8 RCC-8 spatial relationships, and then we evaluate their accuracy in reconstructing the composition table. In the second experiment, we employ a prompt similar to that used in the first experiment, but with the contextual information anonymized. Since the third experiment of the CDC calculus provided poor results, we decided not to perform the corresponding experiment for the RCC-8 calculus.

First Experiment. As in [4], each model is prompted with background information that defines the RCC-8 calculus. Then, we pose to the models a set of questions which compute the cells of the composition table (one question per cell, for a total of $8 \times 8 = 64$ questions). The prompt used for the experiment follows the same template used at [4] for a fair comparison. The prompt starts with defining each RCC-8 binary relation and then questions are posed to compute the transitivity table. The questions posed have the following format: `If` R_1`(x,y) and` R_2`(y,z) then what are the possible relationships between` x `and` z`?`". In the above question, variables R_1 and R_2 are relations from the set {DC, EC, PO, TPP, NTPP, TPPi, NTPPi, EQ}.

In this experiment, we show clearly that the LRMs tested outperform the older LLMs tested in the same task by [4] (GPT-3.5 Turbo 0125, Llama 3 70B Instruct, Gemini 1.5 Pro preview-0409, GPT-4 Turbo 2024-04-09, GPT-4o 2024-05-13 and Claude 3.5 Sonnet 20240620). In our case, o3 achieved the best accuracy across all repeats, with a mean Jaccard index of 99.57%, compared to the worst performing model, gemini-2.5-flash, which achieved 88.55%. Figure 9 depicts the mean Jaccard coefficient for 4 models (best, worst and two models of different vendors, xAI and DeepSeek). In addition, we observe that gemini-2.5-flash provided an empty response due to exceeding the 65536-token context window (as also indicated by the heatmap), which constitutes a significant factor affecting the results. We computed the results for all the other models of Table 1 as well but we omit them due to space considerations. We also performed experiments using the fixed seed configuration, as described in Section 3, but these are not shown again due to space limitations. However, all the models' results can be seen in Figure 10 and in our GitHub repository.

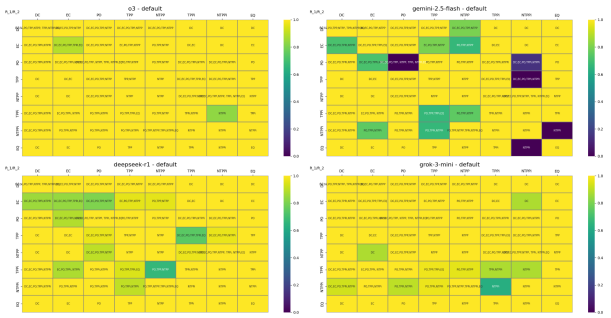


Figure 9: The RCC8 Composition Table shaded by the mean Jaccard index for the models o3 (best performing model), grok-3-mini, deepseek-r1 and gemini-2.5-flash (worst performing model), using the default input parameters.

Figure 11 depicts the total counts of all the answers provided by each model per each topological relation categorized as: correctly predicted (true positive), correctly not predicted (true negative), incorrectly predicted (false positive) and incorrectly not predicted (false negative). The best performing model is o3 and achieved a mean Jaccard index of 99.57%. Upon closer examination of the first four bars of o3 (relation *DC*), we can observe that the model predicted correctly the relation *DC* 93 times across all answers for all 3 repetitions that were done. Thereafter, it correctly did not predict the relation *DC* 1443 times. Finally, it avoided to predict the relation incorrectly and it did not predicted it incorrectly at all (0 times). The same comments hold for the rest of the models and the RCC-8 relations. Overall, Figure 11 illustrates the models' good performance on this experiment, as shown in Table 4, since they all demonstrated high correctly predicted and correctly not predicted scores over the incorrectly predicted and incorrectly not

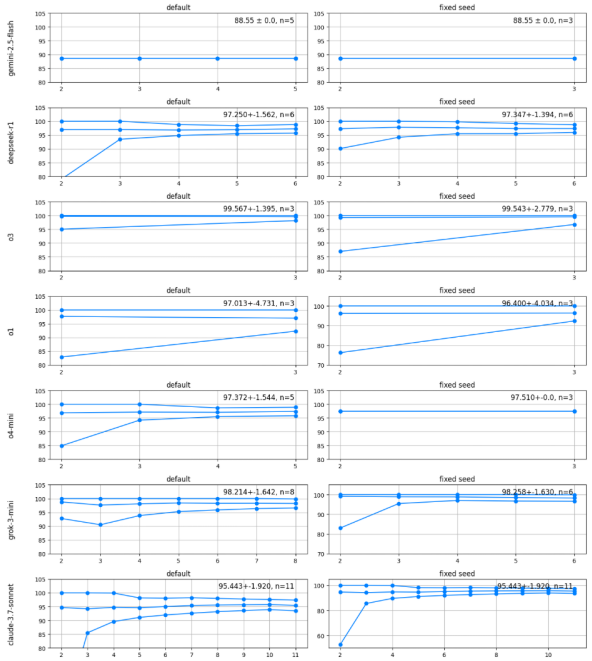


Figure 10: Prediction intervals by repeat for for the RCC-8 calculus. We discontinue further iterations when the prediction interval width approaches values close to 3-6%. We observe that, depending on a model's architecture, some models such as gemini-2.5-flash, o3, o1 need fewer repeats to converge to a narrow prediction interval, while others like claude-3.7-sonnet demand more repeats.

predicted ones. We can conclude the lack of bias in any relation, as the correctly and incorrectly predictions are somewhat equally distributed. However, we should highlight that the worst model (gemini-2.5-flash) has incorrectly predicted the relation *NTPP* 35 times, the most out of all the relations for all models, which shows the model's incapacity to manage this relation.

In relation to prior work, our results shown in Figure 11 indicate a noticeable improvement over those reported in Figure 3 of [4], where there are many incorrectly predicted and incorrectly not predicted relations in their statistics (3 digit numbers), while our highest incorrectly predicted relations are 35 and our highest incorrectly not predicted ones are 25.

Second Experiment. Our second experiment investigates whether the tested LRMs can reconstruct the composition table of the RCC-8 calculus when prompted with anonymized RCC-8 specific definitions (e.g., the relation names). For comparison purposes, we present results for the four LRMs discussed in the first experiment.

Figure 12 presents the transitivity table shaded by mean Jaccard index for these four LRMs. By comparing these results with the ones presented in Figure 9, we observe that in contradiction to [4] and our previous experiment, some LRMs exibit comparable performance, whereas others underperform. Specifically, o3 achieved the highest mean Jaccard index on the anonymized and the non-anonymized experiment by achieving a mean Jaccard index of 99.41% and 99.57%

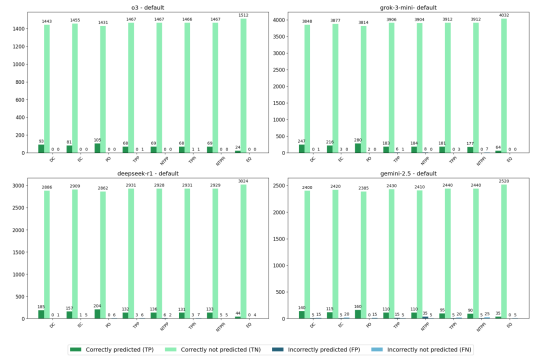


Figure 11: Relation statistics for the RCC-8 composition table for the models o3 (best performing model), grok-3-mini, deepseek-r1 and gemini-2.5-flash (worst performing model), using the default input parameters.

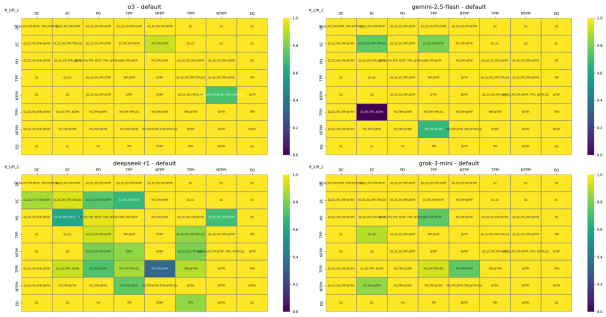


Figure 12: The RCC8 Composition Table shaded by the mean Jaccard index for the anonymized context.

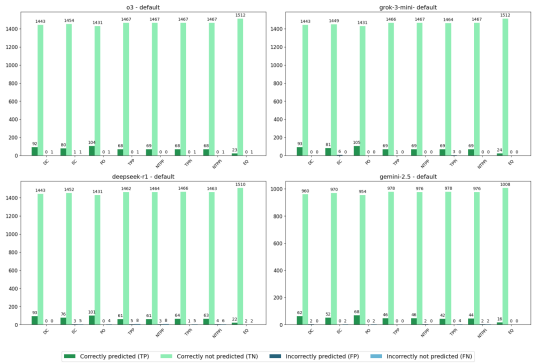


Figure 13: Relation statistics for the composition table for RCC-8 for the anonymized context.

respectively. These comparable results suggest that the LRM that exibit comparable results are trained on the RCC-8 calculus context. An additional noteworthy observation is about the model gemini-2.5-flash which achieved 97.21% on the anonymized experiment and 88.55% on the non-anonymized one. This outcome can be justified if we consider that in the anonymized experiment, the tokens used

(17690) were allowed from the context window of the model, while in the non-anonymized experiment, the model run out of tokens. The exceedance of this limit affected the mean Jaccard index greatly since empty sets were produced as the answer.

As in the first experiment, Figure 13 examines the presence of inductive bias in the LRMs when predicting specific relations. The figure illustrates that the magnitudes for each relation are consistent with those observed in the previous experiment, thereby confirming the similarity of the results between the two experiments.

In Figure 3, we show the average time in seconds per prompt for each model. We can observe clearly that the deepseek-r1 model is the slowest compared to the other models due to the large amount of reasoning tokens used for each prompt. Moreover, in the RCC-8 experiments many models demanded more time to complete the tasks compared to the CDC calculus. We hypothesize that this is the reason for the improved results obtained in these experiments.

Finally, the composition table created by the best (o3) and worst (gemini-2.5-flash) performing models for the first repetition of the experiment with context used and default input parameters is depicted in Table 5. Each answer provided is highlighted with black for correctly predicted, red for incorrectly predicted and blue for incorrectly not predicted answers. From the tables 3 and 5 we observe the significant improved accuracy of the evaluated LRMs compared to the previous LLMs tested by [4].

Model	With context	Anonymized
o3	99.57	99.41
grok-3-mini	98.21	98.68
deepseek-r1	97.25	93.71
gemini-2.5-flash	88.55	97.21

Table 4: The mean Jaccard index for RCC-8.

o3	DC	EC	PO	TPP	NTPP	TPPI	NTPPI	EQ
DC	DC,EC,PO,TPP,NTPP,TPPI,NTPPI,EQ	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC	DC	DC
EC	DC,EC,PO,TPP,NTPPI	DC,EC,PO,TPP,TPPI,EQ	DC,EC,PO,TPP,NTPP	EC,PO,TPP,NTPP	PO,TPP,NTPP	DC,EC	DC	EC
PO	DC,EC,PO,TPP,NTPPI	DC,EC,PO,TPPI,NTPPI	DC,EC,PO,TPP,NTPP,TPPI,NTPPI,EQ	PO,TPP,NTPP	PO,TPP,NTPP	DC,EC,PO,TPPI,NTPPI	DC,EC,PO,TPPI,NTPPI	PO
TPP	DC	DC,EC	DC,EC,PO,TPP,NTPP	TPP,NTPP	NTPP	DC,EC,PO,TPP,TPPI,EQ	DC,EC,PO,TPP,NTPPI	TPP
NTPP	DC	DC	DC,EC,PO,TPP,NTPP	NTPP	NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP,TPPI,NTPPI,EQ	NTPP
TPPI	DC,EC,PO,TPPI,NTPPI	EC,PO,TPPI,NTPPI	PO,TPPI,NTPPI	PO,TPP,TPPI,EQ	PO,TPP,NTPP	TPPI,NTPPI	NTPPI	TPPI
NTPPI	DC,EC,PO,TPPI,NTPPI	PO,TPPI,NTPPI	PO,TPPI,NTPPI	PO,TPPI,NTPPI	PO,TPP,NTPP,TPPI,NTPPI,EQ	NTPPI	NTPPI	NTPPI
EQ	DC	EC	PO	TPP	NTPP	TPPI	NTPPI	EQ
gemini-2.5-flash	DC	EC	PO	TPP	NTPP	TPPI	NTPPI	EQ
DC	DC,EC,PO,TPP,NTPP, TPPI,NTPPI,EQ	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP	DC	DC	DC
EC	DC,EC,PO,TPPI,NTPPI,TPPI,NTPP	DC,EC,PO,TPP,TPPI,EQ	DC,EC,PO,TPP,NTPP	EC,PO,TPP,NTPP,DC,EC	PO,TPP,NTPP	DC,EC	DC	EC
PO	DC,EC,PO,TPPI,NTPPI, TPPI,NTPP	DC,EC,PO,TPP, NTPP, TPPI, NTPPI,EQ	PO,TPP,NTPP	PO,TPP,NTPP	DC,EC,PO,TPPI,NTPPI	DC,EC,PO,TPPI,NTPPI,TPP,NTPP	PO	
TPP	DC	DC,EC	DC,EC,PO,TPP,NTPP	TPP,NTPP	NTPP	DC,EC,PO,TPP,TPPI,EQ	DC,EC,PO,TPP,NTPP,TPPI,NTPPI	TPP
NTPP	DC	DC	DC,EC,PO,TPP,NTPP	NTPP	NTPP	DC,EC,PO,TPP,NTPP	DC,EC,PO,TPP,NTPP, TPPI, NTPPI,EQ	NTPP
TPPI	DC,EC,PO,TPPI,NTPPI	EC,PO,TPPI, NTPPI	PO,TPPI,NTPPI	PO,TPP,NTPP, TPPI, NTPPI,NTPP	PO,TPP,NTPP TPPI	TPPI,NTPPI	NTPPI	TPPI
NTPPI	DC,EC,PO,TPPI,NTPPI	PO,TPPI,NTPPI,EC	PO,TPPI,NTPPI	PO, TPPI, NTPPI	PO,TPP,NTPP,TPPI,NTPPI,EQ	NTPPI	NTPPI	TPPI,NTPP
EQ	DC	EC	PO	TPP	NTPP	TPPI	NTPP, NTPPI	EQ

Table 5: The RCC-8 composition table for the best (o3) and worst (gemini-2.5-flash) performing models when the context is provided.

6 Conclusions

In this paper we evaluated the capabilities of 7 state-of-the-art LRMs on the tasks of the cardinal directions transitivity table computation and the RCC-8 composition table computation. Our results show that the LRMs outperform the LLMs previously evaluated in the literature. o3 excels in both tasks, by achieving a mean Jaccard index of 99.41% and 99.66% respectively when the prompts include context, while it outperforms the other models in most of the other experiments as well. Moreover, when evaluating if the LRMs exhibit inductive bias when predicting specific relations, we concluded that there is none since the correct and incorrect predictions were evenly distributed across the relation types.

In future work, we intend to evaluate more open-weight LRMs such as the ones of the Llama 3 family [11], the Phi family [25] and OpenThinker-32B[3] on the two problems studied in this paper. In addition, we would like to explore neurosymbolic approaches and LRMs extensions using Chain-of-Thoughts and Tree-of-Thoughts ideas on our problems in the spirit of the papers [49, 50]. Finally, we aim to experiment with more qualitative calculi, e.g., the one which combines topological and cardinal direction information [12].

7 Acknowledgments

We would like to thank Tony Cohn and Robert Blackwell for sharing their papers with us and giving us crucial advice regarding our experiments.

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] P. Bhandari, A. Anastasopoulos, and D. Pfoser. 2023. Are Large Language Models Geospatially Knowledgeable?. In *Proceedings of the 31st SIGSPATIAL, Hamburg, Germany.* doi:10.1145/3589132.3625625

[2] R. E. Blackwell, J. Barry, and A. G. Cohn. 2024. Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores. https://arxiv.org/abs/2410.03492

[3] V. V. Cepeda, G. K. Nayak, and M. Shah. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In *NeurIPS 2023, New Orleans, LA, USA.*

[4] A. G. Cohn and R. E. Blackwell. 2024. Can Large Language Models Reason about the Region Connection Calculus? *CoRR* abs/2411.19589 (2024). doi:10.48550/ARXIV.2411.19589

[5] A. G. Cohn and R. E. Blackwell. 2024. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions (Short Paper). In *COSIT 2024, Québec City, Canada.* doi:10.4230/LIPICS.COSIT.2024.28

[6] A. G. Cohn and J. Renz. 2008. Qualitative Spatial Representation and Reasoning. In *Handbook of Knowledge Representation.* doi:10.1016/S1574-6526(07)03013-1

[7] A. G. Cohn, J. Renz, and M. Sridhar. 2012. Thinking Inside the Box: A Comprehensive Spatial Representation for Video Analysis. In *KR 2012, Rome, Italy.* http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4563

[8] E. Davis. 2013. Qualitative Spatial Reasoning in Interpreting Text and Narrative. *Spatial Cogn. Comput.* (2013). doi:10.1080/13875868.2013.824976

[9] M. J. Egenhofer. 1991. Reasoning about Binary Topological Relations. In *SSD'91.* doi:10.1007/3-540-54414-3_36

[10] M. J. Egenhofer and D. M. Mark. 1995. Naive Geography. In *COSIT '95, Semmering, Austria.* doi:10.1007/3-540-60392-1_1

[11] A. Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[12] A. G. Cohn et al. 2014. Reasoning about Topological and Cardinal Direction Relations Between 2-Dimensional Spatial Objects. *J. Artif. Intell. Res.* 51 (2014), 493–532. doi:10.1613/JAIR.4513

[13] A. G. Davoodi et al. 2025. LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs. In *NAACL 2025, New Mexico, USA.* https://aclanthology.org/2025.naacl-long.161/

[14] A. Jaech et al. 2024. OpenAI o1 System Card. *CoRR* (2024). doi:10.48550/ARXIV.2412.16720

[15] C. Deng et al. 2024. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. In *WSDM 2024, Merida, Mexico.* doi:10.1145/3616855.3635772

[16] D. Hong et al. 2024. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024). doi:10.1109/TPAMI.2024.3362475

[17] D. Yu et al. 2025. Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Spatial Reasoning Questions. *CoRR* (2025). doi:10.48550/ARXIV.2502.18470

[18] G. Mai et al. 2023. CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. In *ICML 2023, Honolulu, Hawaii, USA.* https://proceedings.mlr.press/v202/mai23a.html

[3] https://www.open-thoughts.ai/blog/scale

[19] G. Mai et al. 2024. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). *ACM Trans. Spatial Algorithms Syst.* (2024). doi:10.1145/3653070
[20] G. Mai et al. 2025. Towards the next generation of Geospatial Artificial Intelligence. *Int. J. Appl. Earth Obs. Geoinformation* (2025). doi:10.1016/J.JAG.2025.104368
[21] J. Roberts et al. 2023. GPT4GEO: How a Language Model Sees the World's Geography. (2023). doi:10.48550/ARXIV.2306.00020
[22] J. Weston et al. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *ICLR 2016, San Juan, Puerto Rico.*
[23] K. Klemmer et al. 2025. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. In *AAAI-25, Philadelphia, PA, USA.* doi:10.1609/AAAI.V39I4.32457
[24] L. Haas et al. [n. d.]. Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocalization. *CoRR* ([n. d.]). doi:10.48550/ARXIV.2302.00275
[25] Marah Abdin et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219
[26] P. Balsebre et al. 2023. CityFM: City Foundation Models to Solve Urban Challenges. *CoRR* abs/2310.00583 (2023). doi:10.48550/ARXIV.2310.00583
[27] R. Mirzaee et al. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *NAACL-HLT 2021, Online.* doi:10.18653/V1/2021.NAACL-MAIN.364
[28] R. Manvi et al. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *ICLR 2024, Vienna, Austria.* https://openreview.net/forum?id=TqL2xBwXP3
[29] S. Khanna et al. 2024. DiffusionSat: A Generative Foundation Model for Satellite Imagery. In *ICLR 2024, Vienna, Austria.* https://openreview.net/forum?id=I5webNFDgQ
[30] S. Yao et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS 2023, New Orleans, LA, USA.*
[31] T. Akinboyewa et al. 2024. GIS Copilot: Towards an Autonomous GIS Agent for Spatial Analysis. *CoRR* (2024). doi:10.48550/ARXIV.2411.03205
[32] T. Nguyen et al. 2023. ClimaX: A foundation model for weather and climate. In *ICML 2023, Honolulu, Hawaii, USA.* https://proceedings.mlr.press/v202/nguyen23a.html
[33] Xin Guo et al. 2024. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In *CVPR 2024, Seattle, WA, USA.* doi:10.1109/CVPR52733.2024.02613
[34] Y. Cong et al. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *NeurIPS 2022, New Orleans, LA, USA.*
[35] Y. Liu et al. 2025. Efficient Inference for Large Reasoning Models: A Survey. arXiv:2503.23077 [cs.CL] https://arxiv.org/abs/2503.23077
[36] Y. Zhang et al. 2024. BB-GeoGPT: A framework for learning a large language model for geographic information science. *Inf. Process. Manag.* (2024). doi:10.1016/J.IPM.2024.103808
[37] Yuanshao Zhu et al. 2024. UniTraj: Learning a Universal Trajectory Foundation Model from Billion-Scale Worldwide Traces. *CoRR* abs/2411.03859 (2024). doi:10.48550/ARXIV.2411.03859 arXiv:2411.03859
[38] Y. Zhang et al. 2025. Geospatial large language model trained with a simulated environment for generating tool-use chains autonomously. *Int. J. Appl. Earth Obs. Geoinformation* (2025). doi:10.1016/J.JAG.2024.104312
[39] Z. Li et al. 2023. GeoLM: Empowering Language Models for Geospatially Grounded Language Understanding. In *EMNLP 2023, Singapore.* doi:10.18653/V1/2023.EMNLP-MAIN.317
[40] Z. Liu et al. 2025. GAIR: Improving Multimodal Geo-Foundation Model with Geo-Aligned Implicit Representations. *CoRR* (2025). doi:10.48550/ARXIV.2503.16683
[41] Z. Yan et al. 2023. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* (2023). doi:10.1109/TGRS.2023.3332219
[42] Z. Zhou et al. 2024. Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. In *SIGIR 2024, Washington DC, USA.* doi:10.1145/3626772.3657673
[43] E. Evstafev. 2025. Token-by-Token Regeneration and Domain Biases: A Benchmark of LLMs on Advanced Mathematical Problem-Solving. *CoRR* (2025). doi:10.48550/ARXIV.2501.17084
[44] Z. Falomir. 2012. Qualitative distances and qualitative description of images for indoor scene description and recognition in robotics. *AI Commun.* (2012). doi:10.3233/AIC-2012-0535
[45] R. Goyal and M.J. Egenhofer. 1997. Similarity of Cardinal Directions. In *The Annual Assembly and the Summer Retreat of University Consortium for Geographic Information Systems Science.*
[46] R. K. Goyal and M. J. Egenhofer. 2001. Similarity of Cardinal Directions. In *SSTD 2001, Redondo Beach, CA, USA, July 12-15.* doi:10.1007/3-540-47724-1_3
[47] A. Ishay, Z. Yang, and J. Lee. 2023. Leveraging Large Language Models to Generate Answer Set Programs. In *KR 2023, Rhodes, Greece.* doi:10.24963/KR.2023/37
[48] Y. et al. Ji. 2025. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science* (June 2025), 1–38. doi:10.1080/13658816.2025.2511227
[49] F. Li, D. C. Hogg, and A. G. Cohn. 2024. Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark. In *AAAI 24 Vancouver, Canada.* doi:10.1609/AAAI.V38I17.29811
[50] F. Li, D. C. Hogg, and A. G. Cohn. 2024. Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning. In *IJCAI 2024, Jeju, South Korea.* https://www.ijcai.org/proceedings/2024/701
[51] Z. Li and H. Ning. 2023. Autonomous GIS: the next-generation AI-powered GIS. *Int. J. Digit. Earth* (2023). doi:10.1080/17538947.2023.2278895
[52] B. Liétard, M. Abdou, and A. Søgaard. 2021. Do Language Models Know the Way to Rome?. In *BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic.* doi:10.18653/V1/2021.BLACKBOXNLP-1.40
[53] J. K. Mbuya, D. Pfoser, and A. Anastasopoulos. 2024. Trajectory Anomaly Detection with Language Models. In *Proceedings of the 32nd SIGSPATIAL, Atlanta, GA, USA.* doi:10.1145/3678717.3691257
[54] R. Mirzaee and P. Kordjamshidi. 2022. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. In *EMNLP 2022, Abu Dhabi, United Arab Emirates.* doi:10.18653/V1/2022.EMNLP-MAIN.413
[55] D. A. Randell, Z. Cui, and A. G. Cohn. 1992. A Spatial Logic based on Regions and Connection. In *Proceedings of KR'92. Cambridge, MA, USA, October 25-29, 1992.*
[56] J. Renz and B. Nebel. 2007. Qualitative Spatial Reasoning Using Constraint Calculi. In *Handbook of Spatial Logics.* doi:10.1007/978-1-4020-5587-4_4
[57] K. Salmas, D.-A. Pantazi, and M. Koubarakis. 2023. Extracting Geographic Knowledge from Large Language Models: An Experiment. In *(KBC-LM) and (LM-KBC) with (ISWC 2023), Athens, Greece.* https://ceur-ws.org/Vol-3577/paper13.pdf
[58] Z. Shi, Q. Zhang, and A. Lipani. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *AAAI 22, IAAI 22, EAAI 22.* doi:10.1609/AAAI.V36I10.21383
[59] S. Skiadopoulos and M. Koubarakis. 2004. Composing cardinal direction relations. *Artif. Intell.* (2004). doi:10.1016/S0004-3702(03)00137-1

A Prompts' template for the CDC experiment

In this appendix we present the template of the prompts for the experiments on the calculus CDC for the default setting. The complete definitions of all the prompt templates constructed are provided in our GitHub repository.

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` $\{(x, y) : x^2 + y^2 <= 1\}$`. 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}.