White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/224230/

Version: Preprint

Preprint:
Cohn, A.G. and Blackwell, R.E. (2024) Can Large Language Models Reason about the Region Connection Calculus? [Preprint - arXiv]

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

arXiv:2411.19589v1 [cs.CL] 29 Nov 2024

Can Large Language Models Reason about the Region Connection Calculus?

Anthony G Cohn[1,2] **and Robert E Blackwell**[2]

[1]School of Computing, University of Leeds, UK
[2]The Alan Turing Institute, UK
a.g.cohn@leeds.ac.uk, rblackwell@turing.ac.uk

Abstract

Qualitative Spatial Reasoning is a well explored area of Knowledge Representation and Reasoning and has multiple applications ranging from Geographical Information Systems to Robotics and Computer Vision. Recently, many claims have been made for the reasoning capabilities of Large Language Models (LLMs). Here, we investigate the extent to which a set of representative LLMs can perform classical qualitative spatial reasoning tasks on the mereotopological Region Connection Calculus, RCC-8. We conduct three pairs of experiments (reconstruction of composition tables, alignment to human composition preferences, conceptual neighbourhood reconstruction) using state-of-the-art LLMs; in each pair one experiment uses eponymous relations and one, anonymous relations (to test the extent to which the LLM relies on knowledge about the relation names obtained during training). All instances are repeated 30 times to measure the stochasticity of the LLMs. No LLM tested performs particularly well, though rather better than chance.

Introduction

Qualitative Spatial Reasoning (QSR[1]) (Cohn and Renz 2008; Chen et al. 2015; Cohn and Hazarika 2001) is a well developed field which is concerned with the representation of qualitative spatial information and reasoning with it. In natural language, spatial information is usually represented qualitatively (using prepositions such as *on, in, left of, part of, under, touching, ...)* and many calculi have been developed to represent such information. There are calculi for mereological relations e.g. RCC-5[2]. (Jonsson and Drakengren 1997), mereotopological relations (such as RCC-8 (Randell, Cui, and Cohn 1992; Cohn et al. 1997), the 9-intersection model (Egenhofer and Franzosa 1991)), directions (such as OPRA (Moratz 2006)), size (Gerevini and Renz 2002) for example, as well as calculi combining two different aspects of spatial information, such as the Rectangle Algebra (Guesgen 1989; Mukerjee and Joe 1990) which can represent both mereotopological information as well as directional. What is common to all these calculi is that they consist of a set of *jointly exhaustive and pairwise disjoint* (JEPD) *base* relations. E.g., RCC-8 contains eight JEPD *base* relations, illustrated in 2D in Fig. 1.

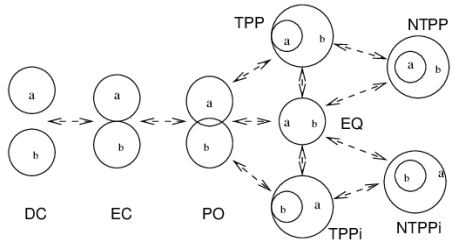

Figure 1: The eight relations of the RCC-8 calculus illustrated in 2D (Cohn et al. 1997): DC (Disconnected), EC (Externally Connected), PO (Partially Overlapping), TPP (Tangential Proper Part), NTPP (Nontangential Proper Part) and EQ (Equals); TPPi and NTPPi are the inverses of TPP and NTPP respectively since they are asymmetric.

Large Language Models (LLMs) (Devlin et al. 2019; Brown et al. 2020), such as GPT-3, LLAMA and GPT-4 are examples of so called *Foundation Models* (Bommasani et al. 2021) which have been trained on very large textual corpora in order to generate text in response to a prompt. This is not the place to survey this burgeoning field, but we note that many claims have been made for the power and apparent intelligent behaviour that these models can display. In particular their performance on some benchmarks may lead one to believe that they possess well developed reasoning capabilities. So the question arises as to whether LLMs can perform the reasoning commonly associated with qualitative spatial calculi. This is the question that we address here. There are several motivations for this, including whether an LLM could be a replacement for a symbolic spatial reasoner (with the added advantage of being able to pose queries in natural language) and also whether an LLM could be used to construct the reasoning mechanisms for a QSR – typically these are manually derived (though sometimes reasoning (Randell, Cohn, and Cui 1992; Hazarika 2005; Bennett 1996) or machine learning (Clementini and Cohn 2024) approaches have been taken), and automating the process could ease the production of new QSRs; here we take an existing calculus (RCC-8) as a test vehicle for answering this question.

[1]We may use QSR as shorthand for both Qualitative Spatial Reasoning and Qualitative Spatial Representation; context should usually make clear which is intended.

[2]RCC is an acronym for the Region Connection Calculus, which comes in various forms, including RCC-5 and RCC-8

Related Work

If we are interested in evaluating the abilities of LLMs to perform spatial reasoning, there are different kinds of spatial reasoning tasks which can be considered. Relational composition is one of the most studied from a theoretical point of view. R3(x, z) is the composition of R1(x, y) and R2(y, z) if it is implied by the latter two relations. In general R3(x, z) is a disjunction of relations. A *composition table* (CT) records the results for all combinations of relations in a particular QSR such as RCC. A second form of reasoning commonly associated with QSR involves what is called a *conceptual neighbourhood* (CN) – also called a *continuity network* in (Randell and Cohn 1989). Fig. 1 illustrates the CN for RCC-8 – there is an edge between two relations iff one relation can be directly followed by the second one, assuming continuous translation or morphological change.

Despite the rapidly growing amount of research into LLMs and their capabilities there has been relatively little devoted specifically to spatial reasoning (some exceptions being, e.g. Cohn and Blackwell (2024); Topsakal and Harper (2024); Yamada et al. (2024)), and almost none exhaustively investigating their abilities to reason about composition in any particular QSR, with the exception of Cohn (2023), which looked at the ability of ChatGPT4 to correctly compute RCC-8's CT, whether it was able to correctly predict human preferred relations in the case of ambiguity; it also investigated whether ChatGPT4 could construct the conceptual neighbourhood diagram of RCC-8. However it only investigated one LLM (ChatGPT-4) and did not investigate the stochasticity of the LLM via repeated experiments. There are a few works which selectively test compositional reasoning, e.g. Cohn and Hernandez-Orallo (2023) which investigated a number of spatial reasoning problems include some limited instances of relational composition, but not exhaustively. Other work investigating the spatial reasoning abilities of LLMs typically which revolved around especially constructed benchmarks such as StepGame (Li, Hogg, and Cohn 2024; Shi, Zhang, and Lipani 2022) can also be regarded as testing compositional reasoning, but not in a methodical or exhaustive manner. StepGame aims to test an LLM's ability to correctly determine the qualitative direction relationship between two objects, given a set direction relations between a larger set of objects, and between 1 and 10 reasoning steps are required to correctly determine the result. Not surprisingly, performance deteriorates as the required number of steps increases. Performance increases markedly when the LLM is used to translate from the English specification to a logical representation and symbolic reasoning is used to compute the relationship. The SpartQA dataset (Mirzaee et al. 2021) is also focused on assessing spatial reasoning, but does not test composition or conceptual neighbourhoods. The bAbI dataset (Weston et al. 2016) also has some tasks which test spatial reasoning and compositional reasoning, in particular about directions to a limited extent. Other work has investigated whether LLMs can acquire an understanding of a spatial environment from a turn-by-turn description of a route, with landmarks named at each turn; whilst the LLMs did perform reasonably well, the experiment did not involve any mereotopological relations, only left/right and up/down (Yamada et al. 2024).

Weaknesses in the reasoning powers of LLMs have previously been noted (e.g. Cai, Chang, and Han (2023)) so one might not expect LLMs to perform well in this regard. But on the other hand, there are a large number of papers about QSR in the literature and these are likely to have formed part of the training corpus of an LLM, and thus might facilitate correctly responding to prompts – though the information concerning the actual reasoning steps are often given in tables (in particular *CT*s – see below) and thus might be hard for LLM training procedures to process effectively.

Experimental Design

There are now many LLMs in the literature, with new ones released frequently. Some of these are open source and are explicit about the training corpus; others are closed and give no specific information about the training, or the precise corpus, such as the GPT family of LLMs. Previous work suggests that models with less than about 40B parameters perform poorly at reasoning (e.g. Leyton-Brown (2024)). We therefore favour larger models, testing those in Table 1.

LLMs are stochastic in nature and show considerable variability in their answers. Vendors provide various API options (e.g., *seed*, *temperature*, and *top_p*) to try to make sampling more consistent. However, no settings that we have yet tried (including setting *temperature* to 0) result in fully deterministic answers (e.g. see Appendix Fig. 8).

We therefore accept all model defaults and repeat each chat completion multiple times. We choose $n = 30$ repeats being the sample size at which, according to the Central Limit Theorem, the sampling distribution of a mean approximates normality well enough for practical purposes. To measure accuracy, we use the mean plus or minus the 95% confidence interval of the mean of answer scores. In all of the experiments each question was posed in a separate conversation. We switched off the guard rails for models hosted on Azure OpenAI. We set *max_tokens* to 512 in the Anthropic API, it being a required parameter.

Although there are many different prompt engineering strategies (see Schulhoff et al. 2024), our purpose here is to test, rather than optimise, model performance. We therefore prefer simple, natural prompting. Our only concession to prompt engineering is to add *Answer the question and provide the final answer in the form: "### Answer:"* to facilitate pattern matching and automation of answer assessment using regular expressions.

In the experiments below, since multiple relations may be possible as an answer, we use the Jaccard Index (JIx, Jaccard (1901)) to compute the accuracy of the predicted response compared to the expected, ground truth, answer. The JIx is calculated by counting the size of the intersection of the predicted set of relations and the ground truth set, divided that by the number of relations in the union of the two sets. When only a single relation is given and the expected answer is just a single relation, this reduces to a binary 1 or 0 measure of accuracy.

Details of the computing infrastructure used to run the experiments can be found in the appendix. The results of all

runs, and the programs used to do the experiments are in the uploaded supplementary material.

Compositional Reasoning

The most researched form of QSR reasoning is that of composition: i.e. given two relations R1(x, y), and R2(y, z), then which relations are possible between x and z? In general, the answer may be a disjunction of relations, as can be seen in the RCC-8 CT in Fig. 2 (which also uses colouring to show the results of Experiment 1 below). Note that this table omits the rows and columns for the EQ relation (we do not include compositions involving EQ in our experiment since they should be trivial). Given the ubiquity of RCC-8 in the QSR literature, in this paper we focus on the abilities of an LLM to reason with RCC-8.

Experiment 1 : Compositional Reasoning in RCC-8

In this experiment the standard names of the eight relations were given with a sentence stating that the relations were part of the RCC-8 calculus, as well as English language descriptions of their semantics. Each prompt concluded with a question asking for the relations involved in one particular composition. We generated 7x7=49 prompt questions based on the RCC-8 CT (Fig. 2). For example[3]:

system: *You are a helpful assistant. Consider the following set of eight pairwise disjoint and mutually exhaustive binary spatial relations. These relations form part of the well known RCC-8 qualitative spatial reasoning calculus. DC(x,y) means that x and y are disconnected and share no spatial parts. EC(x,y) means that x and y touch at a boundary but do not share any interior parts. PO(x,y) means that x and y share a spatial part, but neither is part of the other. TPP(x,y) means that x is part of y and touches y's boundary. NTPP(x,y) means that x is part of y but does not touch y's boundary. TPPi(x,y) is the same as TPP(y,x). NTPPi(x,y) is the same as NTPP(y,x). Finally, EQ(x,y) means that x and y are coincident. I will now ask you a question about these relations. There may be more than one possible relation, in which case name all of the possible answers. Answer the question and provide the final answer in the form: "### Answer:"*

user: *If DC(x,y) and DC(y,z) then what are the possible relationships between x and z?*

The best performing model for experiment 1 is Claude-3.5S, followed by Gemini-1.5P; the worst performing is GPT-3.5T (Table 4, column "1epon"). All models performed better than guessing ($>\approx 0.31$). More detailed statistics can be found in Fig. 3. Claude-3.5S's first answer (Appendix Example 1) is correct, well justified, and scores 1.0. However, repeat 13 from Claude-3.5S was:

assistant: *To answer this question, we need to consider all possible relationships between x and z given that x is disconnected from y, and y is disconnected from z. ≀ ≀ ≀ Given: ≀ - DC(x,y): x and y are disconnected and share no spatial parts ≀ - DC(y,z): y and z are disconnected and share no spatial parts ≀ ≀ ≀ Possible relationships between x and z: ≀ ≀ ≀ 1. DC(x,z): x and z could be disconnected from each other, just as they are both disconnected from y. ≀ ≀ ≀ 2. EC(x,z): x and z could be externally connected, touching at their boundaries without sharing interior parts. ≀ ≀ ≀ 3. PO(x,z): x and z could partially overlap, sharing some spatial parts but not being fully contained within each other. ≀ ≀ ≀ 4. TPP(x,z) or TPPi(z,x): x could be a tangential proper part of z, or vice versa. ≀ ≀ ≀ 5. NTPP(x,z) or NTPPi(z,x): x could be a non-tangential proper part of z, or vice versa. ≀ ≀ ≀ 6. EQ(x,z): x and z could be equal or coincident. ≀ ≀ ≀ In this scenario, all possible RCC-8 relations between x and z are possible because the given information about their relationships to y does not constrain their relationship to each other in any way. ≀ ≀ ≀ ### Answer: ≀ DC(x,z), EC(x,z), PO(x,z), TPP(x,z), NTPP(x,z), TPPi(z,x), NTPPi(z,x), EQ(x,z)*

This answer is only partially correct (TPPi(z,x) is equivalent to TPP(x,z) but should be TPPi(x,z) and NTPPi(z,x) is equivalent to NTPP(x,z) but should be NTPPi(x,z)) so this answer scores 0.75.

Although we asked for possible relations between x and z, models sometimes reply with relationships between z and x (e.g., Claude-3.5S gave 288/1470 (19.59%) answers that included a relation R(z,x)). Generously, we therefore interpret TPPi(z,x) as TPP(x,z) and NTPPi(z,x) as NTPP(x,z) and vice versa. We also allow the symmetric relationships DC(z,x), EC(z,x), PO(z,x) and EQ(z,x).

Claude-3.5S got 513/1470 (34.90%) answers completely correct, Gemini-1.5P 307/1470 (20.88%) and GPT-3.5T only 47/1470 (3.20%) (the lowest).

Claude-3.5S predicted DC most frequently 944/4937 (19.12%) and EQ least frequently 212/4937 (4.29%). Claude-3.5S gave 14/4937 (0.28%) invalid relations (i.e. involving y, rather than only x and z), EC(x,y), EC(y,x), PO(y,z) and PO(x,y). GPT-3.5T predicted DC most frequently 1100/3925 (28.03%) and TPPi least frequently 124/3925 (3.16%). GPT-3.5T gave 3/3925 (0.08%) invalid relations, EC(y,z), NTPP(x,y), and TPP(y,x). Although few, these invalid relations suggest a fundamental lack of understanding of what composition is.

Claude-3.5S predicted TPP more frequently 574/4937 (11.63%) than TPPi 404/4937 (8.18%) and NTPP more frequently 563/4937 (11.40%) than NTPPi 424/4937 (8.59%). Similarly, GPT-3.5T predicted TPP more frequently 466/3925 (11.87%) than TPPi 124/3925 (3.16%) and NTPP more frequently 865/3925 (22.04%) than NTPPi 157/3925 (4.00%). These statistics suggest a bias away from the inverse in these asymmetrical relations. Indeed these four relations all tend to have fewer correct answers in comparison to other relations (Fig. 3).

Whilst the original CT in Fig. 1 has a certain symmetry around the leading diagonal (after substituting the inverse relations for TPP and NTPPi where appropriate and switching to the relevant rows – e.g. (TPP∘EC) has the same predicted relations as (EC∘TPPi)), the LLM produced ones are not, and this is in no small part due to their reluctance to

[3]To save space, in the **system:** prompt, the text in blue is standard across all the system prompts in the experiments in this paper and will not be repeated subsequently, only the text in black, the missing text indicated by ellipsis. Also in the **assistant:** responses, we remove all the blank lines generated by the LLM, replacing each blank line by a ≀ symbol.

Vendor	Model	Abbreviation	Released	API	Number of Parameters	Context
OpenAI	GPT-3.5 Turbo 0125	GPT-3.5T	Jan 2024	Azure OpenAI	Undisclosed	16K
Meta	Llama 3 70B Instruct	Llama-3 70B	Apr 2024	Azure AI	70B	8K
Google	Gemini 1.5 Pro preview-0409	Gemini-1.5P	Apr 2024	Google Vertex	Undisclosed, est. $\gg$3.5T	128K
OpenAI	GPT-4 Turbo 2024-04-09	GPT-4T	Apr 2024	Azure OpenAI	Undisclosed	128K
OpenAI	GPT-4o 2024-05-13	GPT-4o	May 2024	Azure OpenAI	Undisclosed	128K
Athropic	Claude 3.5 Sonnet 20240620	Claude-3.5S	Jun 2024	Anthropic	Undisclosed	200K

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

include the inverse relations.

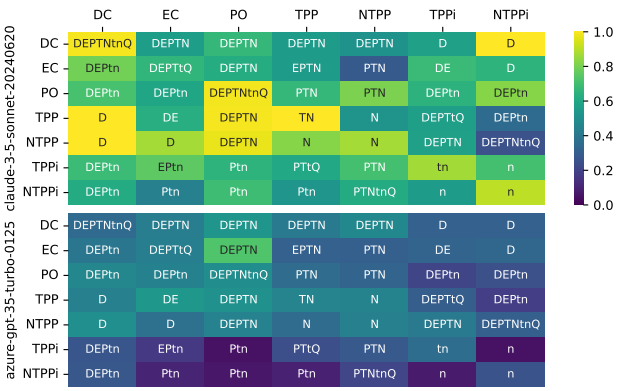

Figure 2: RCC8 CT shaded by mean Jaccard Index (n=30 repeats) for the best performing model, Claude-3.5S and the worst performing model GPT-3.5T. The entry in each cell uses the following coding: D (DC), E (EC), P (PO), T (TPP), N (NTPP), t (TPPi), n (NTPPi), Q (EQ). The full results are in the appendix, Table 5.

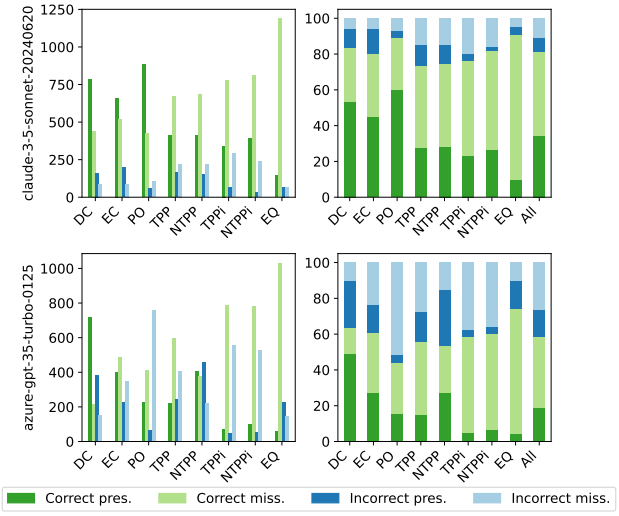

Figure 3: Relation statistics for the CT for RCC-8. The left hand chart shows the absolute number of relations, and the right hand the relative percentage for each relation across all thirty repeats. *All* is the aggregate of all relations.

In order to test whether the result was influenced by prior knowledge of RCC-8 gained during training, we also performed the same experiment, but with all the relation names anonymised[4] to disguise the connection to RCC-8. The prompt was the same as above except for the change of relation names and the omission of the third sentence. All models showed lower performance with anonymised relations (see Fig. 4), and the GPT-3.5T performance was worse than guessing (Table 4, column 1anon). These results support the view that models have difficulty generalising the information contained in the literature and cannot reason well about spatial information from first principles.

Experiment 2 : Preferred Compositions in RCC-8

As noted above, in general a composition of two relations will yield more than one possible base relation, but it turns out that humans tend to have a "preferred" relation. For example, Ragni et al's (2007) report on experiments performed on native German and native Mongolian speakers (20 of each) for RCC-8. The relations were described, but the subjects were not allowed to draw possible configurations, so the setting is essentially equivalent to an LLM setting.

Given that humans may struggle to see all the possible relations[5], determining whether there is agreement about the most preferred is a good question to ask. It turns out that there is good agreement in general across and within the two cultures, with the the percentage of people agreeing with the same preferred relation ranging from 30% to 87.5% (a random choice would yield 12.5% on average since there are eight relations to choose from). (They did not query cases where the composition yields a unique relation, nor did they consider EQ as one of the two relations as this should be a trivial task.) This agreement is perhaps surprising since the two languages are linguistically very different. Ragni et al (2007) do report some differences though – for example although both language speakers preferred DC whenever

[4]We asked Chat-GPT to make up eight words not in the dictionary and then mapped DC to *fablon*, EC to *narkil*, PO to *quonty*, NTPPi to *piflox*, TPPi to *dregly*, NTPP to *lufrex*, TPP to *zorpin*, and EQ to *womfer*. After running the experiments we discovered that fablon is actually a kind of sticky-backed fabric.

[5]The fact that some humans may struggle to compute the CT does not stop it being a valid question to see if an LLM can determine the correct entries. Indeed, given that it has already been remarked upon that determining the correct CT is challenging (Randell, Cohn, and Cui 1992), if LLMs could automatically compute CTs for new sets of relations, this would be very helpful.

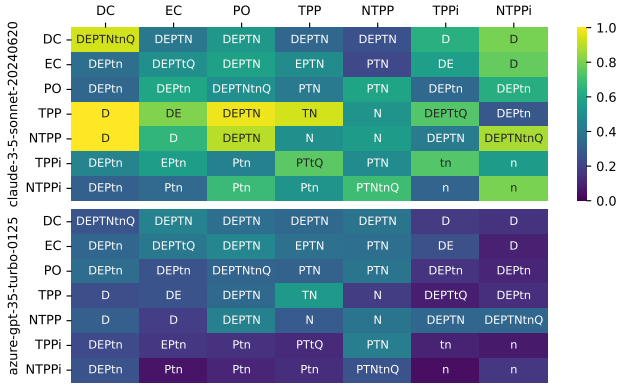

Figure 4: Same as for Figure 2 but for anonymous.

it was consistent, Mongolians preferred PO over NTPPi whereas for Germans the converse was true. Both cultures only chose EQ when composing a relation with its inverse (e.g. TPP with TPPi).

We found an apparent error when transcribing the preferred relations from Table 2 in Ragni et al (2007) . The preferred relation for NTPPi followed by TPP is given as PO but the mean score is given as 45% rather than 35% and so we use the NTPPi (40%) relation instead.

The theory of *preferred mental models* (Knauff, Rauh, and Schlieder 1995) states that people construct the simplest (computationally cheapest) model consistent with the premises. Their experiments showed that humans prefer models with the smallest overlapping complexity which explains the preference for DC noted above.

Given the difficulty reported in Experiment 1 that LLMs have in correctly inferring all possible relations in a composition, asking this more specific query seems a natural task to try; this gives some measure of whether the cognitive preferences of humans correspond at all to LLM choices.

Fig. 5 shows the JIx for each cell for the best performing model, Claude-3.5S and the worst performing model Llama-3 70B, with actual predictions made by each model in the appendix, Table 6.

Some responses contained a clear preference with a good argument (E.g. Appendix Example 2 where Claude-3.5S's response of DC is indeed the preference reported by Ragni et al (2007). A more problematic example, also from Claude-3.5S (Appendix Example 3), yields a confused explanation and an incorrect answer of EC, when only PO, TPP or NTPP are possible.

In some cases, models justify their choice by saying it was "cautious", 22/8820 (0.25%), or "safest" 117/8820 (1.33%). The humans in the Ragni et al (2007) experiment were not asked to justify their choices, but the claim made by the authors of that paper is that choices were based on computational/simplicity considerations – which differs from the reasons claimed here.

As can be seen in Appendix Table 6, Claude-3.5S agreed with the average human preference in 21/37 (56.76%) of cases and only chose two impossible relations. GPT-4o agreed with the average human preference in 19/37 (51.35%), but chose six impossible relations. GPT-3.5T agreed with the average human preference in 18/37 (48.65%) but chose 17 impossible relations. Whilst better than guessing ($\approx 13\%$, these results do not show that LLM preferences are aligned with human preferences.

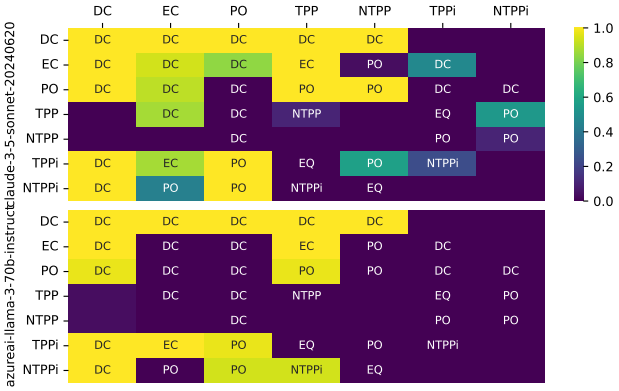

Figure 5: RCC8 preferred relations shaded by mean Jaccard Index (n=30 repeats) for the best performing model, Claude-3.5S and the worst performing model Llama-3 70B. Labels show preferred relations as reported by Ragni et al (2007).

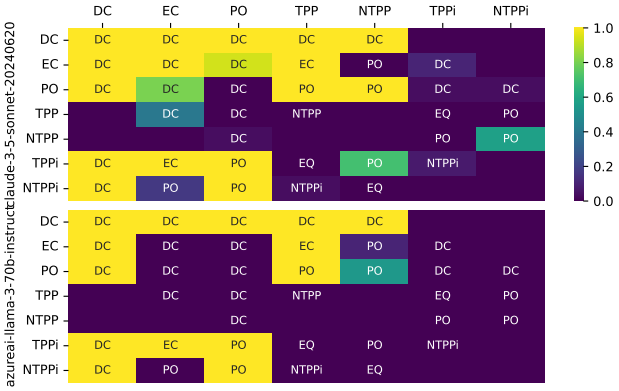

Figure 6: Same as Figure 5 but for anonymous.

In the anonymous version of this experiment GPT-4o and Llama-3 70B had similar performance with both the eponymous and anonymous relations, but all other models had worse performance with anonymous relations (Table 4).

Experiment 3 : Spatial Continuity

Continuity networks were introduced by Randell and Cohn (1989) to represent the set of possible 'next' relations that might apply, assuming that motion is continuous and any transformations in the shape and/or size of an object are also continuous. Subsequently these have been termed "conceptual neighbourhoods" (Freksa 1992). Fortunately, the abbreviation for both of these terms is CN, which we shall use henceforth. The CN for RCC-8 is depicted in Fig. 1. Each

prompt emphasises the need for the relation to be an immediate next relation. Examples of Claude-3.5S correct and incorrect prompt/response pairs can be found in the Appendix (correct: Example 4 for DC; incorrect Example 5, where Claude-3.5S correctly predicted PO, and EQ as neighbours of TPPi but incorrectly predicted EC, DC, NTPP and TPP and missed NTPPi).

GPT-4T performed best, see Table 4 column 3epon for all the LLM results. It is immediately noticeable that the generated tables are not symmetric, although they should be (since all the arrows in Fig. 1 are bi-directional).

azure-gpt-4-turbo-2024-04-09	DC	EC	PO	TPP	NTPP	EQ	TPPi	NTPPi
DC		x						
EC	x		x					
PO	x	x				x		
TPP		x	x		x			
NTPP				x		x		
EQ				x	x		x	x
TPPi				x		x		x
NTPPi						x	x	

azure-gpt-35-turbo-0125	DC	EC	PO	TPP	NTPP	EQ	TPPi	NTPPi
DC		x						
EC	x					x		
PO				x			x	
TPP	x	x			x	x		
NTPP				x		x		
EQ	x	x						
TPPi					x			
NTPPi						x	x	

Table 2: The Continuity Table for RCC-8 produced by our best performing LLM (GPT-4T) and the worst performing LLM (GPT-3.5T). An 'x' means that the relation in that column is predicted as an immediate neighbour of the relation in that row. An empty box means that the relation is not predicted as an immediate neighbour. Green means that the prediction was correct and red that it was incorrect. For each model and question, we take the most commonly occurring answer across thirty repetitions.

In the anonymous version of this experiment Claude-3.5S performed best, Table 4, column 3anon). The results are given in Table 3. It is striking how much worse all models except Claude-3.5S do in the anonymous case compared to the eponymous case, suggesting that they are more of a "parrot" (Bender et al. 2021) than a reasoner, though we suspect that the RCC-8 neighbours are rarely, if ever explicitly stated in textual form in the literature.

Summary

The best performing models overall were Claude-3.5S and GPT-4T (Table 4) with Llama-3 70B and GPT-3.5T noticeably less performant. LLMs are non-deterministic and experiment repeats are essential to quantify the variability of answers.

Although Claude-3.5S is the newest and best performing model overall and GPT 4 series models outperform GPT-3.5T, GPT-4o was released one month after GPT-4T and has worse performance in experiments 2 and 3 suggesting that newer models are not necessarily better than older models in the same family. Although GPT-4T and GPT-4o are both GPT 4 models, it is evident that GPT-4o has had different alignment training and changes to its underlying engineering[6]. Our results are similar to those from the MATH benchmark (Hendrycks et al. 2021) (Fig. 7), except that GPT-4o performs noticeably better in MATH than in our experiments.

azure-gpt-4-turbo-2024-04-09	DC	EC	PO	TPP	NTPP	EQ	TPPi	NTPPi
DC		x						
EC	x		x					
PO	x	x		x	x	x	x	x
TPP			x		x			
NTPP				x		x		
EQ		x	x	x	x		x	x
TPPi		x		x		x		
NTPPi							x	

azure-gpt-35-turbo-0125	DC	EC	PO	TPP	NTPP	EQ	TPPi	NTPPi
DC		x	x					
EC	x		x					
PO		x						
TPP	x	x						
NTPP								x
EQ	x	x						
TPPi								x
NTPPi	x	x						

Table 3: The Continuity Table for RCC-8 produced by two LLMs using anonymised relation names. The meaning of the colouring is the same as in Table 2

It is noticeable that models seem to rarely miss predicting DC in Experiment 1 and DC is often a human preferred relation in Experiment 2 which is similarly well predicted by the LLMs. This suggests that DC is perhaps the easiest relation for LLMs to reason about. It can also be seen, from the bottom row of Table 4, that Experiment 3 was the easiest reasoning task for the LLMs on average with both the eponymous and anonymous average scores being better than the corresponding score for the other two experiments.

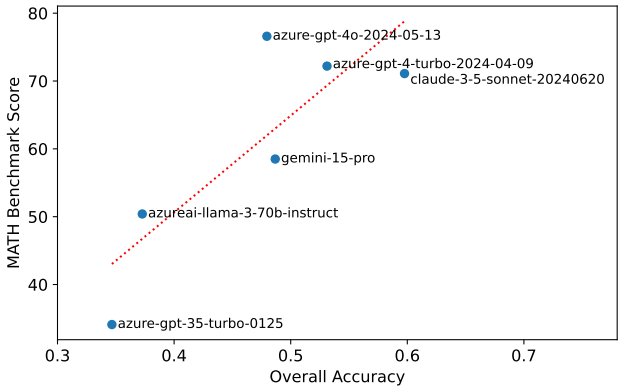

Figure 7: Overall accuracy compared to the MATH benchmark (Hendrycks et al. 2021). MATH data courtesy vellum.ai, retrieved 8-Aug-2024.

All models except Claude-3.5S and Llama-3 70B perform better with spatial continuity than composition (Table 4).

Concluding Remarks and Future Work

The results support the widely-held view that LLMs can struggle to do reasoning tasks[7]. In Experiment 1, where

[6]https://openai.com/index/hello-gpt-4o/ accessed 10-Aug-24

[7]Bender et al (2021) have observed that LLMs might be regarded just as "stochastic parrots" and thus it is not surprising that

Model	1epon	1anon	2epon	2anon	3epon	3anon	Overall
claude-3-5-sonnet-20240620	$\mathbf{0.69 \pm 0.008}$	$\mathbf{0.58 \pm 0.012}$	$\mathbf{0.56 \pm 0.011}$	$\mathbf{0.54 \pm 0.008}$	0.61 ± 0.023	$\mathbf{0.61 \pm 0.026}$	$\mathbf{0.60 \pm 0.016}$
azure-gpt-4-turbo-2024-04-09	0.50 ± 0.014	0.45 ± 0.013	0.50 ± 0.014	0.49 ± 0.014	$\mathbf{0.67 \pm 0.029}$	0.57 ± 0.036	0.53 ± 0.022
gemini-15-pro	0.51 ± 0.013	0.49 ± 0.011	0.51 ± 0.010	0.46 ± 0.007	0.52 ± 0.021	0.42 ± 0.021	0.49 ± 0.015
azure-gpt-4o-2024-05-13	0.47 ± 0.013	0.43 ± 0.012	0.50 ± 0.012	0.50 ± 0.014	0.57 ± 0.039	0.41 ± 0.033	0.48 ± 0.024
azureai-llama-3-70b-instruct	0.43 ± 0.006	0.40 ± 0.010	0.40 ± 0.004	0.40 ± 0.005	0.43 ± 0.018	0.18 ± 0.021	0.37 ± 0.013
azure-gpt-35-turbo-0125	0.33 ± 0.009	0.25 ± 0.011	0.44 ± 0.015	0.36 ± 0.011	0.39 ± 0.034	0.30 ± 0.030	0.35 ± 0.021
Overall	0.49 ± 0.011	0.43 ± 0.012	0.49 ± 0.012	0.46 ± 0.011	0.53 ± 0.028	0.42 ± 0.028	0.47 ± 0.057
Guess rate	0.31 ± 0.012	0.31 ± 0.012	0.13 ± 0.017	0.13 ± 0.017	0.26 ± 0.025	0.26 ± 0.025	0.23 ± 0.033

Table 4: Comparison of model performance based on the mean JIx for each question and answer. Entries in bold are the best performing model in each experiment. Uncertainty is expressed as the 95% confidence interval, n=30 repeats.

LLMs were asked to compute the entire CT for RCC-8, this is a non-trivial task even for humans, so it is perhaps not surprising that LLMs fell well short of 100% accuracy.

A detailed analysis of the actual conversations in the supplementary material show that sometimes LLMs do appear able to do some interesting (qualitative) spatial reasoning, but often fail, sometimes making elementary mistakes, and showing confusion about what the task was. They also show inconsistency in being able to reason correctly about a relation but not its inverse. It is possible that fine tuning, explicit chain-of-thought prompting, or more carefully engineered prompts might improve performance; however, given the stochastic nature of LLMs it seems unlikely that the results would be as good as logical reasoning (Experiment 2 on preferred relations is of course not strictly a logical reasoning exercise, except for the requirement not to predict spatially impossible relations).

There are a variety of avenues for further work which present themselves. Other calculi could be investigated – for example the coarser calculus RCC-5[8], or calculi for reasoning about direction or size (Cohn and Renz 2008). Other LLMs could be evaluated – though since new LLMs and new LLM versions are continually being released, this is a challenge with no definite stopping point. Tracking the change in performance of a particular LLM across releases would also be of interest – though in the case of closed LLMs such as the GPT series where the owners have the right to harvest user conversations and use them for future training, it will not be clear if any improvement is the result of leakage from the previous conversation or more general performance improvement[9]. Determining which LLMs are better at which spatial reasoning tasks would also be worth of future investigation. The overall conclusion that LLMs in general struggle with more complex spatial reasoning tasks is likely to remain the case, at least for the foreseeable future. As already noted above, even with temperature set to zero, LLMs are not deterministic, different temperatures in the LLM APIs could be tried to see if settings other than the default are reliably better.

Experiment 2 above already investigated how LLM performance compared to human performance to a limited extent but further investigation would be worthwhile, including a head-to-head comparison rather than simply taking a result from the literature originally intended to investigate a different question. Another interesting avenue for further work will be to explore the use of multimodal models – when humans perform spatial reasoning tasks including the challenge of building a CT, it is natural to use pencil and paper to sketch diagrams and possible scenarios – investigating whether a multi-modal model with such abilities (including the ability to analyse its own drawings) would be of great interest to the spatial reasoning community. The LLMs failed to produce symmetric CTs, partly as a result of apparent bias against the inverse relations. The extent and reasons for this warrants further investigation and analysis.

As mentioned above, another possible avenue of research is to investigate different prompting strategies, including k-shot (Dang et al. 2022), chain-of-thought (Wei et al. 2022) and tree-of-thought (Yao et al. 2023). Not doing so was deliberate in this paper as we wanted to establish a set of baseline results before considering other prompting strategies. In this paper we use the mean JIx across 30 repeats of each answer to measure answer accuracy and take the mean of the means across a set of answers to measure model accuracy. There are other methods which could be investigated, though it is already clear the LLMs are far from good reasoners for RCC-8. one alternative is to compute the modal JIx for each RCC-8 relation individually and average those for each cell or one could use an alternative error metric such as the Levenshtein (1965) or Edit distance. One could also take account of the distance between relations in the CN and treat deviations which are further from the correct response more harshly (e.g. if the single correct response was DC and a model predicted EC then that would have a smaller penalty than if it predicted PO – cf (Bruns and Egenhofer 1996)).

Although the results presented here are comparable to the MATH benchmark, there are some notable differences with the GPT 4 series of models, and we suggest that a separate comprehensive QSR benchmark be established.

Data Access

The data associated with this paper are available from the following GitHub repository: https://github.com/RobBlackwell/can-llms-reason-about-the-rcc Note that the data are encrypted to avoid the questions

logically correct deductive reasoning is challenging for an LLM.

[8]Preliminary results for the eponymous RCC-5 CT in the appendix suggest that LLMs find RCC-5 easier than RCC-8.

[9]However, note that no feedback was given to any of the LLMs investigated as to whether the response was correct or not.

and answers being automatically harvested to become LLM training data. However, the decryption procedure and password are included in the README file.

Funding

This work was supported by the Fundamental Research priority area of The Alan Turing Institute.

AGC thanks the Turing's Defence and Security programme through a partnership with the UK government in accordance with the framework agreement between GCHQ and The Alan Turing Institute, and for support provided by the Economic and Social Research Council (ESRC) under grant ES/W003473/1.

We thank Microsoft Research - Accelerating Foundation Models Research program, for the provision of Azure resources to access OpenAI models.

Contribution Statement

AC conceived the original idea. RB implemented the benchmark in consultation with AC and performed the evaluations. AC wrote the first draft and both authors contributed to subsequent drafts.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bennett, B. 1996. Modal logics for qualitative spatial reasoning. *Logic Journal of the IGPL*, 4(1): 23–45.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Bruns, H. T.; and Egenhofer, M. 1996. Similarity of spatial scenes. In *Seventh international symposium on spatial data handling*, 31–42. Delft, The Netherlands.

Cai, Z.; Chang, B.; and Han, W. 2023. Human-in-the-Loop through Chain-of-Thought. *arXiv preprint arXiv:2306.07932*.

Chen, J.; Cohn, A. G.; Liu, D.; Wang, S.; Ouyang, J.; and Yu, Q. 2015. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(1): 106–136.

Clementini, E.; and Cohn, A. G. 2024. Extension of RCC*-9 to Complex and Three-Dimensional Features and Its Reasoning System. *ISPRS International Journal of Geo-Information*, 13(1): 25.

Cohn, A. G. 2023. An Evaluation of ChatGPT-4's Qualitative Spatial Reasoning Capabilities in RCC-8. *arXiv preprint arXiv:2309.15577, Working notes of QR-23*.

Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica*, 1(3): 275–316.

Cohn, A. G.; and Blackwell, R. E. 2024. Evaluating the Ability of Large Language Models to Reason about Cardinal Directions. In *arXiv:2406.16528, To appear in Proceedings of COSIT-24*.

Cohn, A. G.; and Hazarika, S. M. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2): 1–29.

Cohn, A. G.; and Hernandez-Orallo, J. 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs. *arXiv preprint arXiv:2304.11164*.

Cohn, A. G.; and Renz, J. 2008. Qualitative Spatial Representation and Reasoning. In *Handbook of Knowledge Representation*. Elsevier.

Dang, H.; Mecke, L.; Lehmann, F.; Goller, S.; and Buschek, D. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv:2209.01390.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Egenhofer, M. J.; and Franzosa, R. D. 1991. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2): 161–174.

Freksa, C. 1992. Temporal reasoning based on semi-intervals. *Artificial intelligence*, 54(1-2): 199–227.

Gerevini, A.; and Renz, J. 2002. Combining topological and size information for spatial reasoning. *Artificial Intelligence*, 137(1-2): 1–42.

Guesgen, H. W. 1989. Spatial reasoning based on Allen's temporal logic, Technical Report TR-89-049. Technical report, International Computer Science Institute Berkeley.

Hazarika, S. M. 2005. *Qualitative spatial change: space-time histories and continuity*. Ph.D. thesis, PhD Thesis, University of Leeds.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37: 547–579.

Jonsson, P.; and Drakengren, T. 1997. A Complete Classification of Tractability in RCC-5. *J. Artif. Intell. Res.*, 6: 211–221.

Knauff, M.; Rauh, R.; and Schlieder, C. 1995. Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. In *Proceedings of the seventeenth annual conference of the Cognitive Science Society*, 200–205.

Levenshtein, V. I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10: 707–710.

Leyton-Brown, K. 2024. *Rationality Report Cards – slides presented at a AAAI-24 workshop. https://tinyurl.com/Leyton-Brown-AAAI24.*

Li, F.; Hogg, D. C.; and Cohn, A. G. 2024. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18500–18507.

Mirzaee, R.; Rajaby Faghihi, H.; Ning, Q.; and Kordjamshidi, P. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *Proc. NAACL*, 4582–4598.

Moratz, R. 2006. Representing relative direction as a binary relation of oriented points. In *ECAI*, volume 6, 407–411.

Mukerjee, A.; and Joe, G. 1990. A Qualitative Model for Space. In Shrobe, H. E.; Dietterich, T. G.; and Swartout, W. R., eds., *Proceedings of the 8th National Conference on Artificial Intelligence. Boston, Massachusetts, USA, July 29 - August 3, 1990, 2 Volumes*, 721–727. AAAI Press / The MIT Press.

Ragni, M.; Tseden, B.; and Knauff, M. 2007. Cross-cultural similarities in topological reasoning. In *Spatial Information Theory: 8th International Conference, COSIT 2007*, 32–46. Springer.

Randell, D.; and Cohn, A. G. 1989. Modelling topological and metrical properties in physical processes. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, 357–368.

Randell, D.; Cui, Z.; and Cohn, A. G. 1992. A Spatial Logic Based on Regions and Connection. In *3rd International Conference on Knowledge Representation and Reasoning, 1992*, 165–176.

Randell, D. A.; Cohn, A. G.; and Cui, Z. 1992. Computing transitivity tables: A challenge for automated theorem provers. In Kapur, D., ed., *Automated Deduction—CADE-11*, 786–790. Berlin, Heidelberg: Springer Berlin Heidelberg.

Schulhoff, S.; Ilie, M.; Balepur, N.; Kahadze, K.; Liu, A.; Si, C.; Li, Y.; Gupta, A.; Han, H.; Schulhoff, S.; et al. 2024. The Prompt Report: A Systematic Survey of Prompting Techniques. *arXiv preprint arXiv:2406.06608*.

Shi, Z.; Zhang, Q.; and Lipani, A. 2022. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proc. AAAI*, volume 36, 11321–11329.

Topsakal, O.; and Harper, J. B. 2024. Benchmarking Large Language Model (LLM) Performance for Game Playing via Tic-Tac-Toe. *Electronics*, 13(8).

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; Van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *ICLR*.

Yamada, Y.; Bao, Y.; Lampinen, A. K.; Kasai, J.; and Yildirim, I. 2024. Evaluating Spatial Understanding of Large Language Models. arXiv:2310.14540.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Appendix

This appendix provides supplementary information and gives examples of interactions which could not fit into the main paper – instead the the main paper refers to this appendix. The appendix also contains additional figures and tables which would not fit into the main paper, giving and comparing the performance of the different LLMs. Benchmark questions and results are available in a GitHub repository at https://github.com/RobBlackwell/can-llms-reason-about-the-rcc.

Example answers referred to in the paper

Example 1. Claude-3.5S's first answer to DC◦DC

assistant: *To answer this question, let's consider the given information and the possible relationships between x and z: ≀ ≀ ≀ 1. We know that DC(x,y) - x and y are disconnected ≀ 2. We also know that DC(y,z) - y and z are disconnected ≀ ≀ ≀ Given these conditions, x and z could have any of the following relationships: ≀ ≀ ≀ 1. DC(x,z) - x and z could be disconnected ≀ 2. EC(x,z) - x and z could be externally*

connected ≀ 3. PO(x,z) - x and z could partially overlap ≀ 4. TPP(x,z) - x could be a tangential proper part of z ≀ 5. NTPP(x,z) - x could be a non-tangential proper part of z ≀ 6. TPPi(x,z) - z could be a tangential proper part of x ≀ 7. NTPPi(x,z) - z could be a non-tangential proper part of x ≀ 8. EQ(x,z) - x and z could be equal ≀ ≀ ≀ The reason all relationships are possible is that the disconnection of both x and y from z does not constrain the relationship between x and z in any way. They could be completely separate or in any other spatial configuration. ≀ ≀ ≀ ### Answer: ≀ DC(x,z), EC(x,z), PO(x,z), TPP(x,z), NTPP(x,z), TPPi(x,z), NTPPi(x,z), EQ(x,z)

Example 2. One instance of Claude-3.5S's preferred answer to DC∘DC

system: *... It is possible that in some cases there could be more than one relation that might hold; in these cases just give me your single most preferred relationship. ...*

user: *If DC(x,y) and DC(y,z) then what is your preferred relationship between x and z?*

assistant: *To answer this question, we need to consider the implications of the given information and the possible relationships between x and z. ≀ ≀ ≀ Given: ≀ - DC(x,y): x and y are disconnected and share no spatial parts ≀ - DC(y,z): y and z are disconnected and share no spatial parts ≀ ≀ ≀ When we consider the relationship between x and z, we can deduce that: ≀ ≀ ≀ 1. x and z could be completely separate from each other, just as they are both separate from y. ≀ 2. x and z could be touching at their boundaries. ≀ 3. x and z could be overlapping or even one could be part of the other. ≀ ≀ ≀ However, since we are asked for the preferred relationship, we should consider the most likely scenario based on the given information. ≀ ≀ ≀ The fact that both x and z are disconnected from y suggests that they are likely to be separate from each other as well. While it's possible for them to have a different relationship, the most conservative and likely assumption based on the given information is that x and z are also disconnected. ≀ ≀ ≀ Therefore, the preferred relationship between x and z is DC(x,z). ≀ ≀ ≀ ### Answer: DC(x,z)*

Example 3. An answer from Claude-3.5S to EC∘NTPP

user: *If EC(x,y) and NTPP(y,z) then what is your preferred relationship between x and z?*

assistant: *To answer this question, let's consider the relationships given: ≀ ≀ ≀ 1. EC(x,y): x and y are externally connected, meaning they touch at their boundaries but don't share interior parts. ≀ 2. NTPP(y,z): y is a non-tangential proper part of z, meaning y is completely inside z without touching z's boundary. ≀ ≀ ≀ Given these relationships, we need to determine the most likely relationship between x and z. ≀ ≀ ≀ Since y is completely inside z (NTPP), and x is only touching y at its boundary (EC), x must be outside of z. However, there are two possibilities for how x and z might relate: ≀ ≀ ≀ 1. DC(x,z): x and z could be disconnected, with some space between them. ≀ 2. EC(x,z): x could be touching the boundary of z. ≀ ≀ ≀ Of these two options, EC(x,z) is more likely because we know that x is touching y, which is inside z. For x to be disconnected from z, there would need to be some space between y and z's boundary, which is not guaranteed by the given information. ≀ ≀ ≀ Therefore, the preferred relationship between x and z is EC(x,z). ≀ ≀ ≀ ### Answer: EC(x,z)*

Example 4. An example correct response from Claude-3.5S when asked about the conceptual neighbours of DC.

system: *... In the question, assume that all motion is continuous, and any deformations of the spatial regions are also continuous. Also please describe the motion or changes of size or shape that would be required for the change of relation to take place. ...*

user:*If DC(a,b) holds, then what are the next possible relations that might hold as the two regions translate or change their shape or size? I am interested in the immediate next relation, so no intermediate relation must hold.*

assistant: *To answer this question, let's consider the possible changes that could occur to regions a and b starting from the DC (disconnected) relation: ≀ ≀ ≀ 1. Translation: If either region a or b (or both) move towards each other, the next possible relation would be EC (externally connected). This occurs when the regions just touch at their boundaries. ≀ ≀ ≀ 2. Change in shape: If either region a or b (or both) change shape in a way that extends towards the other region, they could also reach the EC relation when their boundaries just touch. ≀ ≀ ≀ 3. Change in size: If either region a or b (or both) grow in size towards each other, they could again reach the EC relation when their boundaries just touch. ≀ ≀ ≀ It's important to note that no other relation can be reached directly from DC without first passing through EC. This is because: ≀ ≀ ≀ - To reach PO (partially overlapping), the regions would first need to touch (EC) before overlapping. ≀ - To reach TPP or NTPP (tangential or non-tangential proper part), one region would need to first touch the other (EC) before becoming part of it. ≀ - TPPi and NTPPi are inverse relations of TPP and NTPP, so they also cannot be reached directly from DC. ≀ - EQ (equal) cannot be reached directly from DC as the regions would need to touch and then perfectly align, which would involve intermediate relations. ≀ ≀ ≀ Therefore, the only immediate next relation possible from DC is EC. ≀ ≀ ≀ ### Answer: ≀ The only next possible relation that can immediately follow DC(a,b) is EC(a,b). This can occur through translation of one or both regions towards each other, or through growth or shape change of one or both regions until their boundaries just touch.*

Example 5. An example incorrect response from Claude-3.5S when asked about the conceptual neighbours of TPPi.

user:*If TPPi(a,b) holds, then what are the next possible relations that might hold as the two regions translate or change their shape or size? I am interested in the immediate next relation, so no intermediate relation must hold.*

assistant: *To answer this question, let's consider the initial state where TPPi(a,b) holds, meaning that b is a tangential proper part of a. This means that b is completely contained within a, and touches a's boundary at least at one point. From this state, we can consider the possible immediate next relations based on continuous motion, translation,*

or changes in shape or size. ≀ ≀ ≀ The possible next relations are: ≀ ≀ ≀ 1. PO(a,b): If b grows or a shrinks in such a way that part of b extends outside of a while still maintaining some overlap, the relation would change to partial overlap. ≀ ≀ ≀ 2. EC(a,b): If a shrinks precisely along its boundary with b, or if b grows exactly to match a's boundary, they would become externally connected. ≀ ≀ ≀ 3. DC(a,b): If a rapidly shrinks away from b or b rapidly shrinks away from a's boundary, they could become disconnected. This is less likely in a continuous motion scenario but technically possible. ≀ ≀ ≀ 4. EQ(a,b): If a shrinks or b grows so that their boundaries exactly match, they would become equal. ≀ ≀ ≀ 5. NTPP(a,b): If b grows to completely engulf a, including a's boundary, the relation would flip to a being a non-tangential proper part of b. ≀ ≀ ≀ 6. TPP(a,b): If b grows to completely contain a but still touches a's boundary at least at one point, the relation would flip to a being a tangential proper part of b. ≀ ≀ ≀ It's important to note that the relation cannot directly transition to NTPPi(a,b) from TPPi(a,b) without passing through an intermediate state, as this would require b to first separate from a's boundary before becoming properly contained within a. ≀ ≀ ≀ ### Answer: The next possible relations that might immediately hold after TPPi(a,b) are PO(a,b), EC(a,b), DC(a,b), EQ(a,b), NTPP(a,b), or TPP(a,b).

Experiment 1 Supplementary Information

Table 5 is provided in addition to Figure 2 to show specific errors of omission and commission for the most commonly occurring answers.

Experiment 2 Supplementary Information

Table 6 is provided in addition to Figure 5 to show preferred relations for the most commonly occurring answers.

Variability

To explore the effect of LLM parameters, we fixed the *seed* and varied *temperature* for GPT-3.5T applied to our first Experiment 1 question set, and then fixed the *seed* and varied *top_p* (Fig. 8). Results show considerable variability.

Anonymised Experiments

The following tables and figures show results of the anonymised experiments for comparison with the eponymous experiments included in the main paper (Table 7, Figure 9 and Table 8).

Preliminary Results for RCC-5

Preliminary results for the RCC-5 CT with eponymous relations are Claude-3.5S 0.84±0.011, GPT-4T 0.72±0.012, Gemini-1.5P 0.72±0.013, GPT-4o 0.66±0.014, Llama-3 70B 0.60±0.013, and GPT-3.5T 0.47±0.016. All models performed better with RCC-5 than with RCC-8, consistent with RCC-5 being a coarser calculus that is easier to reason about.

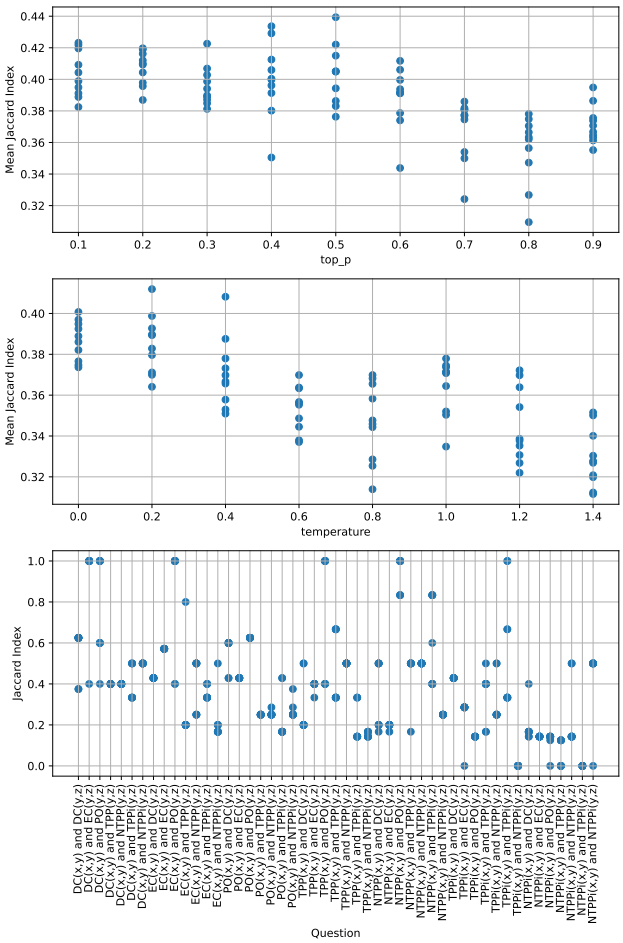

Figure 8: Variability of scores when using GPT-3.5T i) varying *top_p*, ii) varying *temperature* and iii) when fixing *temperature* $= 0$. Ten repeats. In all cases we fix the *seed* $= 123$.

Computing Infrastructure

All experiments were conducted on an Apple MacBook Pro with an Apple M2 Pro processor and 32GB of memory. The Large Languages Models themselves were hosted by suppliers (Google Vertex, Microsoft Azure, Anthropic) and accessed using HTTP APIs as listed in in Table 1. A `Makefile` is provided with each experiment to aid reproducibility. All interactions with LLM APIs were recorded in `answers.jsonl` files including prompts, responses, date and time stamps, API and model versions.

claude-3-5-sonnet-20240620	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	D,E,P,T,N,t,n,Q	D,E,P,+T,+N	D,E,P,+T,+N	D,E,P,+T,+N	D,E,P,+T,+N	D,-E	D
EC	D,E,P,t,n	D,E,P,+T,+t,+Q	D,E,P,+T,+N	-D,E,P,T,+N	-D,-E,P,T,+N	D,E,-P	D
PO	D,E,P,t,n	D,E,P,+t,+n	D,E,P,T,N,t,n,Q	-D,-E,P,T,N	P,T,N	D,E,P,-T,-N,+t,+n	D,E,P,t,n
TPP	D	D,E,-P,-T	D,E,P,T,N	T,N	-T,N	D,E,P,T,-N,+t,Q	+D,+E,P,-T,-N,+t,+n
NTPP	D	D	D,E,P,T,N	N	N	+D,+E,+P,T,N	+D,+E,+P,+T,N,+t,+n,+Q
TPPi	D,E,P,+t,+n	-D,E,P,t,n	-D,-E,P,t,n	P,T,-N,+t,Q	P,T,N,-Q	t,n	-t,n
NTPPi	D,E,P,+t,+n	-E,P,+t,+n	P,t,n,-Q	P,t,n,-Q	P,T,N,t,n,Q	-t,n	n

azure-gpt-35-turbo-0125	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	D,+E,+P,+T,+N,+t,+n,+Q	D,E,+P,+T,+N	D,E,P,+T,+N	D,+E,+P,+T,N	D,+E,+P,+T,N	D,-N	D,-N
EC	D,E,+P,+t,+n	D,E,+P,+T,+t,+Q	D,E,P,T,N	-D,+E,+P,+T,N	-D,+P,+T,N	D,+E,-N	D,-N
PO	D,E,P,+t,+n	D,E,+P,+t,+n	D,E,P,T,N,+t,+n,+Q	+P,T,N	-D,+P,+T,N	D,+E,+P,-N,+t,+n	D,+E,+P,-N,+t,+n
TPP	D,-N	D,+E,-N	D,+E,+P,+T,N	T,+N,-Q	-D,N	+D,+E,+P,T,+t,Q	D,+E,+P,-N,+t,+n
NTPP	D,-N	D,-N	D,E,+P,+T,+N	-D,N	-D,N	+D,+E,+P,T,N,-Q	+D,+E,+P,+T,N,+t,n,+Q
TPPi	D,+E,+P,-N,+t,+n	-D,+E,+P,-N,+t,+n	-D,-E,+P,+t,+n	-D,+P,+T,-N,+t,+Q	-D,+P,+T,N	t,+n,-Q	-D,-N,+n
NTPPi	D,+E,+P,-N,+t,+n	-D,-E,+P,+t,+n	-D,+P,-N,+t,+n	-D,+P,-N,+t,+n	-D,+P,+T,N,+t,+n,+Q	-T,-N,+n	-D,n

Table 5: The CTs for RCC-8 produced by our best performing LLM (Claude-3.5S) and the worst performing LLM (GPT-3.5T). The entry in each cell uses the following coding: D (DC), E (EC), P (PO), T (TPP), N (NTPP), t (TPPi), n (NTPPi), Q (EQ). For each model and composition, we take the most commonly occurring answer across 30 repetitions. Black means that relation is correctly predicted, red that relation is incorrectly predicted (also indicated with a $-$), blue that the relation was incorrectly missing (also indicated with a $+$).

claude-3-5-sonnet-20240620	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	DC	DC	EC	DC	EC	DC	DC
EC	DC	PO	PO	PO	DC	PO	PO
PO	DC	DC	PO	TPP	NTPP	TPP	PO
TPP	DC	DC	DC	NTPP	NTPP	NTPP	NTPP
NTPP	NTPP	DC	EC	PO	PO	DC	PO
TPPi	TPPi	NTPPi	DC	EC	PO	PO	PO
NTPPi	NTPPi	NTPPi	DC	DC	DC	DC	DC

azureai-llama-3-70b-instruct	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	DC	EC	EC	EC	EC	EC	DC
EC	EC	PO	PO	NTPP	DC	PO	NTPP
PO	DC	EC	TPP	TPP	NTPP	TPP	NTPP
TPP	DC	DC	NTPP	NTPP	NTPP	NTPP	NTPP
NTPP	NTPP	DC	EC	PO	TPP	DC	NTPP
TPPi	TPPi	NTPPi	DC	EC	PO	NTPPi	NTPPi
NTPPi	NTPPi	NTPPi	DC	DC	DC	DC	EC

Table 6: The table of preferred compositions for RCC-8 produced by our best performing LLM (Claude-3.5S) and the worst performing (Llama-3 70B); each cell indicates the most commonly occurring answer across 30 repetitions. Green text means it agreed with the human most likely relation, and blue a possible but not the most human-preferred relation. Red means it chose an impossible relation. Since the humans in Ragni et al's experiment were not unanimous in their choices, the minority humans also made choices not preferred by the majority.

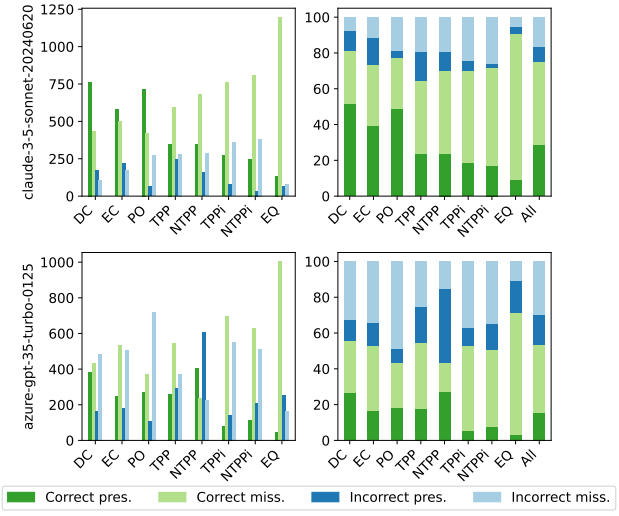

Figure 9: Relation statistics for the CT for RCC-8 with anonymised relations. Layout as for Fig. 3.

claude-3-5-sonnet-20240620	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	D,E,P,T,N,t,n,Q	D,E,+P,+T,+N	D,E,P,+T,+N	D,E,+P,+T,+N	D,+E,+P,+T,+N	D	D
EC	D,E,+P,+t,+n	D,E,P,+T,+t,+Q	D,E,P,+T,+N	-D,E,P,T,+N	-D,-E,+P,+T,+N	D,E,-P	D
PO	D,E,+P,+t,+n	D,E,P,+t,+n	D,E,P,+T,+N,+t,+n,+Q	-D,-E,P,T,N	-D,-E,P,T,N	+D,+E,P,-T,-N,+t,+n,-Q	D,E,P,-T,-N,+t,+n
TPP	D	D,E,-T	D,E,P,T,N	T,N	-T,N	+D,E,P,T,-N,t,-n,Q	D,E,P,-T,-N,+t,+n
NTPP	D	D	D,E,P,T,N	-T,N	-T,N	+D,+E,+P,T,N	+D,E,P,T,N,t,n,Q
TPPi	D,+E,+P,+t,+n	-D,E,P,t,+n	P,+t,+n	P,T,t,Q	P,T,N	t,n	-t,n
NTPPi	D,+E,+P,+t,+n	-D,-E,P,+t,+n	P,t,n	+P,t,n	P,T,N,t,n,Q	-t,n	n

azure-gpt-35-turbo-0125	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	D,+E,+P,+T,+N,+t,+n,+Q	+D,+E,P,T,N,-t,-n,-Q	+D,+E,+P,+T,+N	+D,+E,+P,+T,N,-Q	+D,+E,+P,T,N	+D,-T,-N	+D,-N,-Q
EC	D,E,+P,+t,+n	D,E,+P,+T,+t,+Q	+D,+E,+P,T,N	+E,+P,+T,N,-t	P,T,+N	+D,+E,-T,-N	+D,-T,-N
PO	D,+E,P,+t,+n	+D,+E,+P,-N,+t,n	D,+E,P,+T,+N,+t,+n,+Q	+P,+T,N,-n	+P,T,N	+D,+E,+P,-N,+t,n	+D,+E,+P,-T,-N,+t,+n
TPP	D,-N,-n	+D,+E,-N,-t	D,+E,+P,+T,N	T,N	+N,-t,-n	+D,+E,+P,+T,-N,+t,-n,+Q	+D,+E,+P,-N,t,+n
NTPP	D,-N	+D,-E,-N	+D,+E,P,+T,N	-T,N,-t	N,-Q	+D,+E,+P,T,N	+D,+E,+P,+T,N,+t,n,+Q
TPPi	+D,+E,+P,-N,+t,n	-D,E,P,-T,-N,+t,+n,-Q	+P,-T,-N,+t,+n	+P,+T,-N,+t,-n,+Q	+P,T,N	-T,-N,+t,+n	-T,-N,+n
NTPPi	D,+E,+P,-N,+t,+n	-D,+P,-N,+t,+n	-D,+P,-N,+t,+n	+P,-N,t,+n	+P,T,N,+t,+n,+Q	-T,-N,+n	-N,+n,-Q

Table 7: The CTs for RCC-8 produced by our best performing LLM (Claude-3.5S) and the worst performing LLM (GPT-3.5T) using anonymised relations. The representation used in the entries is the same as in Table 5.

claude-3-5-sonnet-20240620	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	DC	DC	EC	EC	EC		DC
EC	DC	PO	PO	PO	DC	PO	PO
PO		EC	PO	TPP		PO	NTPP
TPP		DC		PO			NTPP
NTPP	PO	DC	EC	PO	PO	DC	PO
TPPi	TPPi		DC	EC	PO	PO	PO
NTPPi			DC			DC	DC

azureai-llama-3-70b-instruct	DC	EC	PO	TPP	NTPP	TPPi	NTPPi
DC	DC	EC	EC	NTPP	EC		DC
EC	EC	PO	PO	PO	DC	TPP	PO
PO		EC	PO	TPP		TPP	TPP
TPP		DC		PO			NTPP
NTPP	NTPP	DC	EC	PO	TPP	DC	NTPP
TPPi	TPPi		DC	EC	PO	TPP	NTPP
NTPPi			DC			DC	EC

Table 8: The table of preferred compositions for RCC-8 produced by our best performing LLM (Claude-3.5S) and the worst performing LLM (Llama-3 70B) LLMs with anonymised relations. The relation names are coloured in the same way as in Table 6.