Evaluating the Ability of Large Language Models to Reason about Cardinal Directions

Anthony G Cohn ⓘ
School of Computing, University of Leeds, UK

Robert E Blackwell ⓘ
Alan Turing Institute, UK

─── **Abstract** ───────────────────────────────────

We investigate the abilities of a representative set of Large language Models (LLMs) to reason about cardinal directions (CDs). To do so, we create two datasets: the first, co-created with ChatGPT, focuses largely on recall of world knowledge about CDs; the second is generated from a set of templates, comprehensively testing an LLM's ability to determine the correct CD given a particular scenario. The templates allow for a number of degrees of variation such as means of locomotion of the agent involved, and whether set in the first , second or third person. Even with a temperature setting of zero, Our experiments show that although LLMs are able to perform well in the simpler dataset, in the second more complex dataset no LLM is able to reliably determine the correct CD, even with a temperature setting of zero.

1 Introduction

Many claims (e.g. [3, 7, 8]) have been made since the emergence of Large Language Models (LLMs) as to their ability to reason. Spatial reasoning is of particular interest since not only does it underlie a human's ability to operate in the physical world, but also because LLMs are not embodied; so the question arises, have they nonetheless acquired an ability to reason about situations which might occur in the real world? This is the question we address here. Spatial reasoning in general encompasses many aspects including topology, distance, and direction. Here, we restrict ourselves to reasoning about cardinal directions (CDs). CDs are important for many reasons, e.g.: (i) successful navigation and route finding/following usually requires a fundamental understanding and ability to reason about CDs: CDs are crucial to success when using a map. Equally, giving directions often relies, at least in part, on the use of CDs. (ii) Understanding the geography of an area depends on understanding the relative CD of one location to another – migration patterns, climate differences and economic variations are often underpinned by CDs. (iii) Weather patterns are often heavily influenced by the direction of the prevailing wind. (iv) CDs often play a critical role in

cultural and historical contexts, e.g. the alignment of the pyramids to the four CDs, or in certain languages – e.g. the aboriginal language *Guugu Yimithirr* has no words for left or right, and spatial information is mainly conveyed using CDs [5].

In this paper we therefore analyse how well LLMs can reason about cardinal and inter-cardinal directions. We do this by automatically constructing a large set of questions based on templates, for which the correct answer has been pre-determined, and testing each LLM's ability to answer the questions correctly. We also tested the LLMs on a small set of simpler questions, co-created with ChatGPT, in which recall of world knowledge is more prevalent.

2   Related Work

Despite the rapidly growing amount of research into LLMs and their capabilities there has been relatively little devoted specifically to spatial and/or geographic reasoning, and none which has tested their ability to reason about CDs in the way we do here. Of the existing work we note benchmarks such as StepGame [10, 16] which aim to test an LLM's ability to correctly determine the spatial relationship between two objects, given the spatial relations between a larger set of objects, and between 1 and 10 reasoning steps are required to correctly determine the result; the direction relationships are not exclusively CDs, but also include, for example "clock face directions" (B is in the three o'clock direction from C). 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 [13] is also focused on assessing spatial reasoning, but does not contain any CDs. The bAbI dataset [17] has one task which tests CDs understanding, task 19, which contains 1000 training and 1000 test questions: each instance contains 5 facts stating CDs between two objects, and then a question asking about the relation between two of them. Other work [19] 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 CDs, only left/right and up/down.

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. A *composition table* records the results for all combinations of relations in a particular spatial representation such as RCC; an investigation [2] into ChatGPT's abilities to compute all RCC compositions found reasonable accuracy levels (reduced when relations are anonymised); however RCC is a purely mereotopological calculus with no notion of direction embedded in its semantics.

Some LLMs have been built specifically for geo-applications, but these do not focus on reasoning about directions but rather aspects such as toponym recognition, e.g. [12].

3   Experimental Design

Whilst testing compositional reasoning with CDs would be of interest, here we restrict ourselves to testing simpler reasoning abilities. We created two question and answer sets which we refer to as *small* and *large*. For *small* we used ChatGPT to co-create 100 simple questions where the answer is a CD $\{north, south, east, west\}$. We edited the questions and corrected the answers where necessary. We changed the questions to ensure equal class representation amongst the four answers. Example questions are:

- *You are watching the sun set. Which direction are you facing?*

- *If the South Pole is behind you, which direction are you facing?*

We use this dataset to give an overall assessment of LLM performance in real world scenarios requiring directional common sense spatial reasoning and common sense spatial knowledge.

It would be impractical to generate a substantial question set manually and so we used an automated, template driven approach for *large*. We wanted to investigate the ability of LLMs to reason about CDs in the context of a simple scenario involving locomotion along or around a geographical feature as this is a test of an LLM in a realistic situation. Based on some informal experimentation using GPT and ChatGPT with a selection of questions and noting a lack of accuracy, we chose six question templates to test LLM performance more comprehensively:

- *You are walking [south] along the [east] shore of a lake; in which direction is the lake?* (Template T1).

- *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?* (Template T2).

- *You are walking [south] along the middle of the [east] side of a park; in which direction is the bandstand located in the centre of the park?* (Template T3).

- *You are walking [east] along the [south] side of a road which runs [east to west]. In which direction is the road?* (Template T4).

- *You are walking [south] along the [east] shore of the island. In which direction is the sea?* (Template T5).

- *You are walking [south] along the [east] shore of an island and then turn around to head back in the direction you came from, in which direction is the sea?* (Template T6).

We then exhaustively generated all forms of these questions for all cardinal and inter-cardinal directions and ten different locomotion types *{cycling, driving, hiking, jogging, perambulating, racing, riding, running, unicycling, walking}*. Note that in each template, once one of the directions (between "[ ]") is fixed, then there are only two possibilities for the second direction. Following earlier evidence [9] that an LLM's performance can vary depending on which person a question is phrased as using, we also generated questions in the first-person (*I am*), first-person plural (*We are*), second-person (*You are*), third-person singular (*He is* and *She is*), and third-person plural (*They are*) forms. This gives us 6 questions × 10 forms of locomotion × 6 person forms × 8 directions × 2 directions-variations = 5760 questions.

Previous work suggests that models with less than about 40B parameters perform poorly at reasoning [9]. We therefore favour larger models, testing those listed in Table 1.

API	Model	Released	Num. params	Window
Anthropic Claude	claude-3-opus-20240229	Feb 2024	137B	200,000
Google Vertex	gemini-10-pro	Dec 2023	1.6T	32,000
	gemini-15-pro-preview-0409	Apr 2024	≫3.5T	128,000
OpenAI	gpt-3.5-turbo-0613 [4]	Jun 2023	175B	4,096
	gpt-35-turbo-1106	Nov 2023	175B	16,385
	gpt-3.5-turbo-0125	Jan 2024	175B	16,385
	gpt-4-0613 [14]	Jun 2023	1.76T	8,192
	gpt-4-turbo-2024-04-09	Apr 2024	1.76T	128,000

Table 1 LLMs tested. in our experiments. Window is the context window size in tokens.

3.1   Prompting

Zero-shot prompting is when a model is given a question without any examples to help guide the answer. The model must then attempt to answer the question based solely on its general pre-training. As a system prompt, we use "You are a helpful assistant. I will give you a question about directions. The answer is either north, south, east, west, north-east, north-west, south-east or south-west. Please only reply with the answer. No yapping.". We then present each question in a new chat. We include "No yapping." since that has been reported (`https://tinyurl.com/no-yapping` as being beneficial in persuading an LLM to be brief in its response.

We set temperature $= 0$ for each model to try to achieve deterministic answers. Temperature is a parameter that affects the randomness or variability of the responses generated by a language model and helps to control the predictability of the the model's output but even a 0 temperature does not guarantee reproducible, deterministic behaviour. To explore the effect of temperature on accuracy, we take our best performing model on the *large* dataset and vary temperature t , $0 \leq t \leq 2$.
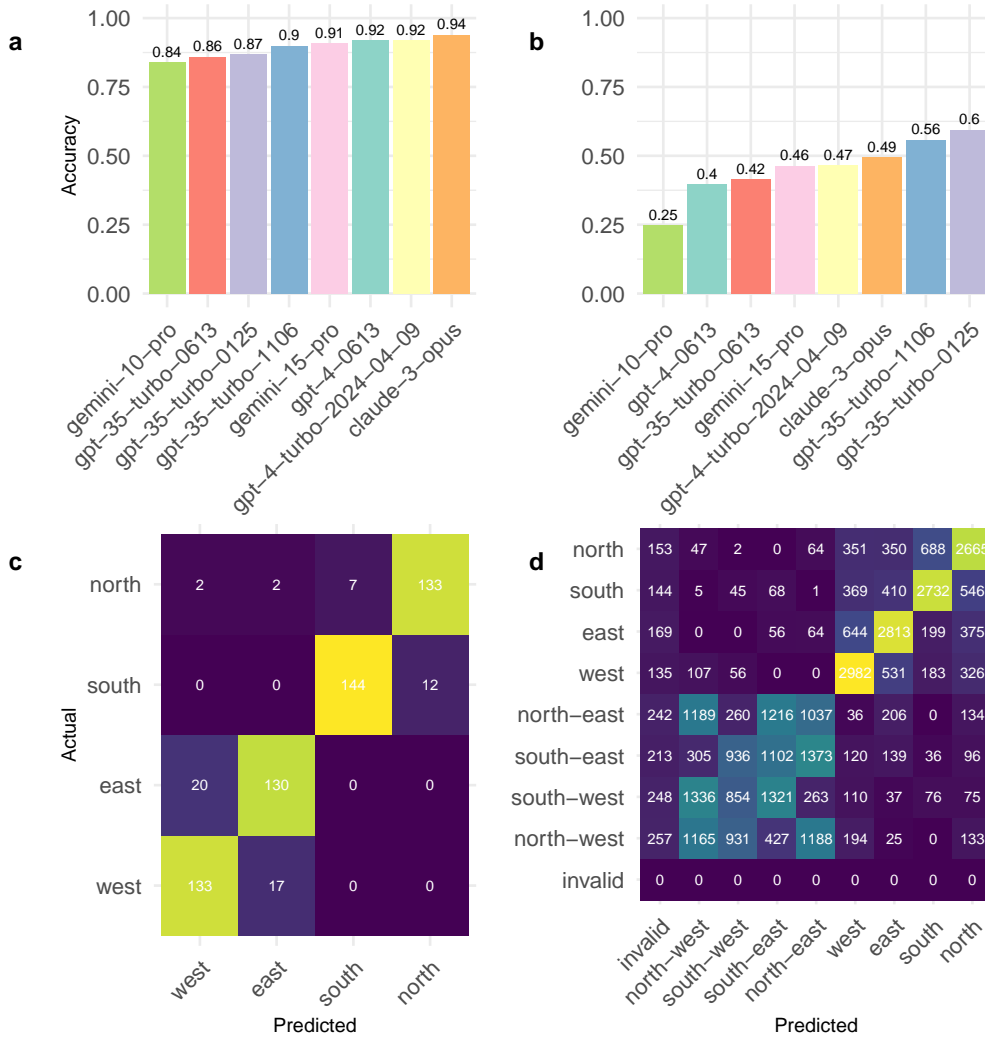
We use case-insensitive string comparison and remove spurious punctuation and white space before comparing answers; e.g. we regard "'North East'." and "north-east" as equivalent. Our prompts ask for cardinal or inter-cardinal direction answers only: we count answers such as "The lake is to the west" as correct if the intended answer is "west", but we note instances where answers do not strictly meet the rubric. We assess performance using accuracy, and report variability using the standard error of the mean.

4   Results

All models tested showed an accuracy of > 0.8 for *small* (Fig. 1a). Where confusion occurred, it was mostly *north* confused with *south*, and *east* confused with *west* (Fig. 1c). In one case a model ignored the rubric: gemini-1.0-pro answered the question *"On a hike, a duck pond is to your north and the nearest town is south. What direction is the pond from the town?"* with *"The pond is north of the town"*, which is correct but not a one word answer. Of the 100 questions, 77 were always correctly answered. Only one question was always answered incorrectly: *In a stadium with a north-facing entrance, if the VIP section is on the left side, which direction would it be in? (east)*; all answered *west*.

Model accuracy was worse for *large*, (the more complex dataset), with the best performing model (gpt-35-turbo-0125) achieving only 0.595. Of the 5760 questions, only 294 (5.10%) were correctly answered by all the models. 628 questions (10.90%) were not answered correctly by any of the models. Of those questions not answered correctly by any of the models, 368 (58.60%) were T4 questions (suggesting roads running from one direction to another are a cause of confusion), 129 (20.54%) were T6 and 98 (15.61%) were T2 (suggesting that turning backwards is a difficulty). The rubric was not followed for 1762 of the 46080 answers (3.82%) from the eight models. gpt-4-turbo-2024-04-09 failed to follow the rubric on 618 questions (10.73%). gpt-4-0613 failed to the follow the rubric on 608 questions (10.56%) claude-3-opus failed to follow the rubric on 207 questions (3.59%). gemini-10-pro failed to follow the rubric on 190 questions (3.30%). All other models followed the rubric on more than 98% of questions, with gpt-35-turbo-0125 failing to follow the rubric on only one question. Only gemini-10-pro gave correct answers when not following the rubric (33 such answers, 17%).

1595 (90.52%) of the answers where the rubric was not followed were answers to T4 questions, and the answer given was one of 'north-south', 'east-west','south-east to north-west','north-east to south-west','south-west to north-east','north-west to south-east','south
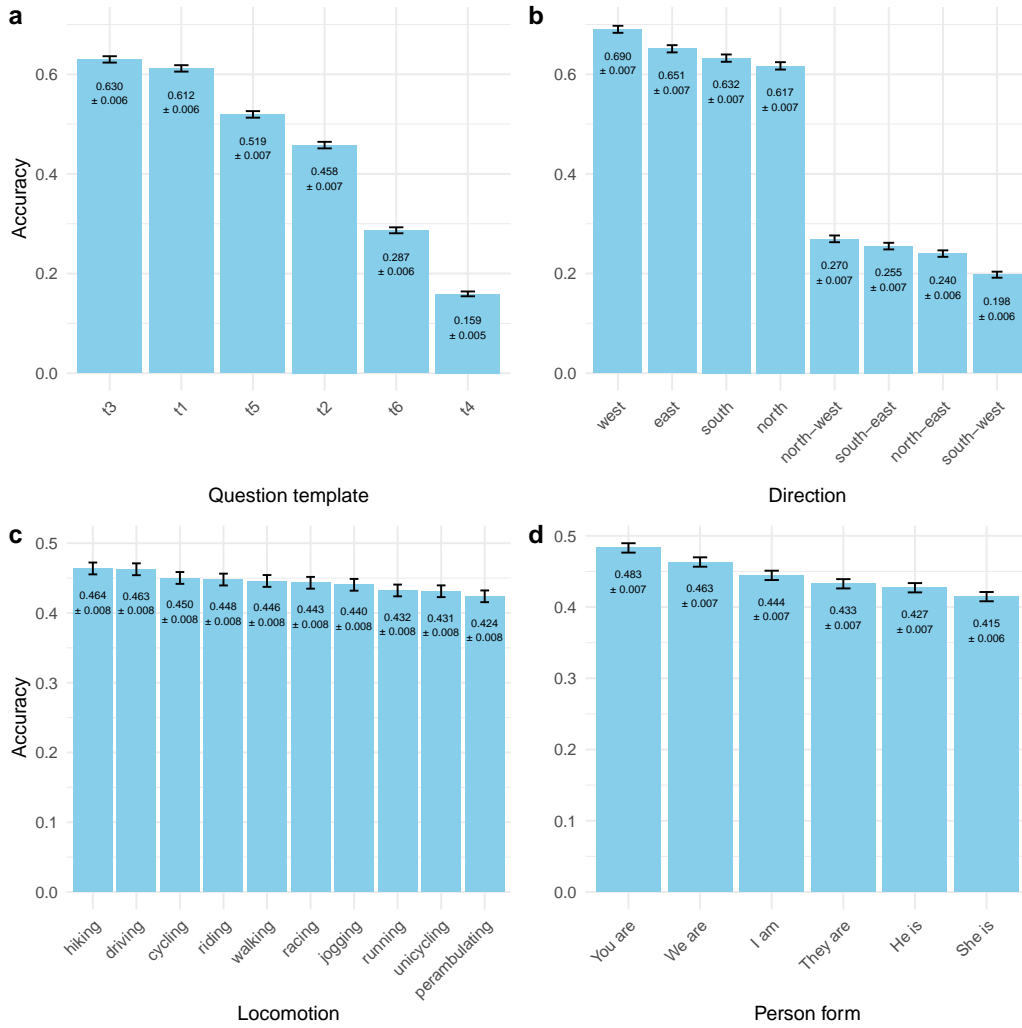
Figure 1 (a) and (c) show accuracy by LLM and confusion matrix respectively, for *small* question set; (b) and (d) for *large*. Answers that cannot be interpreted as a CD or an inter-CD are considered invalid. To avoid bias from three gpt-35-turbo models, the confusion matrices exclude gpt-35-turbo-0613 and gpt-35-turbo-1106 but include gpt-35-turbo-0125.

to north','north to south','west to east','west-east','east to west' (further suggesting roads running from one direction to another are a cause of confusion, Fig. 2a).

Models achieve higher accuracy on cardinal directions than inter-cardinal directions (Fig. 2b). We found little difference in accuracy amongst the various forms of locomotion (Fig. 2c). These minor differences may be due to the incidence of the words in the training data of the models. Second person prompts (*You are*) have the highest accuracy, followed by first and then third person (Fig. 2d).

Fig. 3 gives a breakdown of confusion for each LLM. The upper right quadrant of each matrix gives the performance for the 4 main CDs and it can be seen that in general (except for gemini-10-pro and to a lesser extent gpt-35-turbo-0613) all models perform well here – it is the inter-CD relations which cause most confusion. Surprisingly, there is asymmetry in the north/south and east/west confusion. For example, gpt-35-turbo-0125 predicted north

Figure 2 Accuracy by (a) question template, (b) direction, (c) locomotion and (d) person form for *large*. To avoid bias from using three gpt-35-turbo models, we exclude gpt-35-turbo-0613 and gpt-35-turbo-1106 but include gpt-35-turbo-0125.

when the answer was south on 138 occasions but predicted south when the answer was north on only ten occasions. gpt-4-turbo-2024-4-09 predicted east when the answer was west on 212 occasions but never predicted west when the answer was east. But this bias towards north and east is not universal – some of the models have a reverse bias. We do not have a good explanation for this unexpected asymmetry.

Fig. 4 Shows accuracy by temperature for gpt-35-turbo-0125 applied to *large*. When temperature increases, accuracy decreases. This can be explained by more random next token prediction in the model. As temperature, $t \rightarrow 2.0$ the number of errors from the model (*HTPP 500 - The server had an error while processing your request. Sorry about that!*) also increases, requiring repeated retries before obtaining any answer and making $t = 2.0$ impractical.
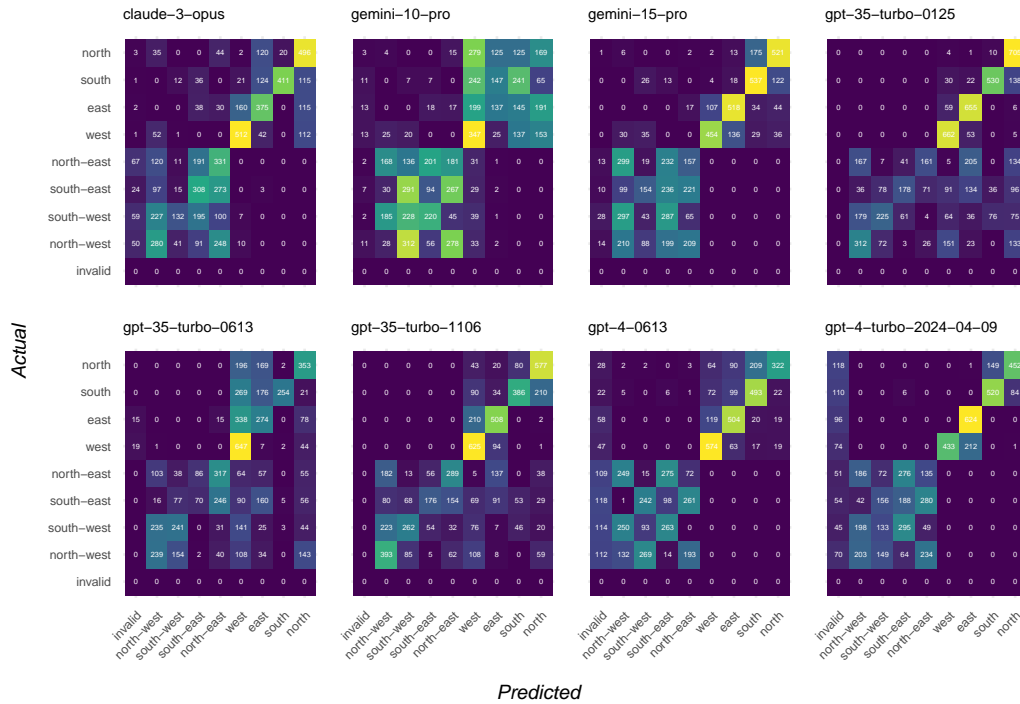
Figure 3 Confusion matrices for each of the LLMs used to test *large*. Answers that cannot be interpreted as CD or inter-CD are considered invalid.
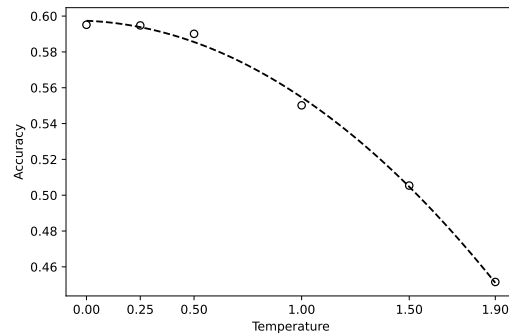

Figure 4 Accuracy by temperature for gpt-35-turbo-0125 applied to *large*.

5   Discussion and Conclusions

None of the models tested is able to reliably reason about cardinal direction, however all models show some capacity for spatial reasoning. A model that randomly selects cardinal or inter-cardinal direction answers to an MCQ would have an accuracy of 0.125, but even the worst performing model (gemini-10-pro) achieved an accuracy of 0.25 on *large*.

All models showed higher accuracy on *small* compared to *large*, though this would be expected given that *small* only has four possible answers whilst *large* has eight. The questions in *large* arguably all require reasoning and a model-based approach to solving, unlike *small*. It is unclear if LLMs produce the answer by reasoning or by recalling memorized information [6] – *large* requires more reasoning, *small* relatively more factual recall.

Given the results here, LLMs appear to perform better at factual recall than spatial reasoning. Of the questions in *small*, 16 can be reasonably regarded as only requiring world knowledge, and 84 as requiring simple reasoning. All the LLMs answered the world knowledge questions correctly. Of all the answers to the simple reasoning questions, 87% were correct.

Unlike results generally reported in the literature (e.g. [15]) comparing GPT-35-Turbo and GPT-4 performance, for *large*, the OpenAI interface to GPT-35-turbo has the highest accuracy. Moreover the latest GPT-35-Turbo beats the latest GPT-4-Turbo and the latter was released more recently than the former. We do not currently have an explanation for this. One author had experimented briefly with T1-T3 on ChatGPT-4, but never provided feedback, so contamination seems unlikely. The same author has also given talks at several venues using T1-T3 as examples, and it is conceivable that this might have reached OpenAI who may have improved their model as a result, but this does not explain why GPT-35-Turbo is better than GPT-4-Turbo, particularly when the release date of the former is before that of the latter.

The development of LLMs is progressing rapidly (though many believe they will never achieve AGI, let alone ever achieve reliable reasoning abilities, at least without a neuro-symbolic component): the Open AI GPT-35-turbo model was updated twice in seven months. Using *large* as a benchmark, we observed a 43% increase in performance between the Open AI GPT-35-turbo 0613 and 0125 versions. However, any evaluation such as this can only ever be a snapshot evaluation, so we hesitate to draw conclusions in general as to which LLM (family) is better than another.

We also tested Microsoft Azure API access to gpt-35-turbo and gpt-4. Although the Microsoft Azure API is designed to include additional guardrails, compliance and data governance certification and enterprise support, we found that accuracy was similar to the Open AI API models gpt-35-0613 and gpt-4-0613 respectively. The Azure documentation does not specify which OpenAI model their models exploit.

Possibilities for future work include: (1) Improving the question design; there are minor flaws in our current questions, e.g. differing punctuation in T1-T6, and a potential ambiguity in T4 (since a road is a linear object it might have been clearer to add "from $\langle agent \rangle$" to make it clear that the question is not relating to the orientation of the road. (2) Exploration and/or development of prompting strategies [1] to improve performance – either using general methods such as *chain of thought* or *tree of thoughts*, or spatial-specific ones such as *Visualization-of-Thought*[18]. (3) Other LLMs could be evaluated, or existing ones fine-tuned for the tasks under consideration. (4) Extend the variety of experiments to create a comprehensive benchmark for evaluating reasoning about directions – in this paper we have deliberately only considered questions whose answer is a CD, but a more comprehensive dataset would also consider other directions (left, right, behind, above...). (5) Building a comprehensive benchmark for other aspects of spatial reasoning (e.g. topological, distance) and combinations of these; ideally these would be generated programmatically (cf [11]). (6) Consider situations with more than two objects of interest, so that, e.g. compositional reasoning can be tested, and also reasoning about trajectories (cf [17]).

Data Access Statement

Contribution Statement

AC conceived the original idea for the *large* dataset and RB subsequently the *small* dataset. RB implemented the benchmark in consultation with AC, and performed the evaluations. Both authors wrote the original draft of the paper and contributed to the subsequent drafts.

References

1 Prabin Bhandari. A survey on prompting techniques in LLMs, 2024. `arXiv:2312.03740`.

2 A G Cohn. An evaluation of ChatGPT-4's Qualitative Spatial Reasoning Capabilities in RCC-8. *arXiv preprint arXiv:2309.15577*, 2023.

3 A Creswell and M Shanahan. Faithful reasoning using large language models, 2022. `arXiv: 2208.14271`.

4 T T Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

5 J B Haviland. Guugu Yimithirr cardinal directions. *Ethos*, 26(1):25–47, 1998.

6 Y Hou, J Li, Y Fei, A Stolfo, W Zhou, G Zeng, A Bosselut, and M Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. *arXiv preprint arXiv:2310.14491*, 2023.

7 J Huang and K C-C Chang. Towards reasoning in large language models: A survey, 2023. `arXiv:2212.10403`.

8 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. `arXiv:2205.11916`.

9 K Leyton-Brown. Rationality Report Cards. Slides presented at a AAAI-24 workshop. URL: `"https://tinyurl.com/Leyton-Brown-AAAI24"`.

10 F Li, D C Hogg, and A G Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the StepGame benchmark. In *Proc. AAAI*, 2024.

11 F Li, D C Hogg, and A G Cohn. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. In *Proc. IJCAI*, 2024.

12 Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. Geolm: Empowering language models for geospatially grounded language understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL: `https://api.semanticscholar.org/CorpusID: 264426067`, `arXiv:2310.14478`.

13 R Mirzaee, H Rajaby Faghihi, Q Ning, and P Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proc. NAACL*, pages 4582–4598, 2021.

14 OpenAI and Josh Achiam et al. GPT-4 technical report, 2024. `arXiv:2303.08774`.

15 Narun Krishnamurthi Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. STEER: Assessing the economic rationality of large language models. In *Forty-first International Conference on Machine Learning*, 2024. `arXiv:2402.09552`.

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

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

18 W Wu, S Mao, Y Zhang, Y Xia, L Dong, L Cui, and F Wei. Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*, 2024.

19 Y Yamada, Y Bao, A K Lampinen, J Kasai, and I Yildirim. Evaluating spatial understanding of large language models, 2024. `arXiv:2310.14540`.