Preprint:
Cohn, A.G. and Blackwell, R.E. (2025) Evaluating the Ability of Large Language Models to Reason about Cardinal Directions, Revisited. [Preprint - arXiv]

UNIVERSITY OF LEEDS

University of Sheffield

UNIVERSITY of York

Evaluating the Ability of Large Language Models to Reason about Cardinal Directions, Revisited[*]

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

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

Abstract

We investigate the abilities of 28 Large language Models (LLMs) to reason about cardinal directions (CDs) using a benchmark generated from a set of templates, extensively 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 the newer Large Reasoning Models are unable to reliably determine the correct CD for all questions. This paper summarises and extends earlier work presented at COSIT-24.

1 Introduction

This paper summarises and extends an earlier paper which investigated the ability of Large Language Models (LLMs) to reason about cardinal directions [Cohn and Blackwell, 2024b]. That paper introduced two benchmarks: a small dataset of 100 questions designed to test simple factual understanding of cardinal directions, and a more complex one, with 5760 questions, generated from a set of six templates which requires some spatial reasoning to answer correctly. Already in the earlier paper several LLMs were obtaining very good performance on the small dataset, and newer models do even better. Therefore this update concentrates solely on the larger dataset and documents the performance of newer LLMs on it, conducting a more thorough analysis overall.

Many claims (e.g. [Creswell and Shanahan, 2022; Huang and Chang, 2023; Kojima *et al.*, 2022]) 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 [Haviland, 1998].

In this paper we therefore analyse how well LLMs can reason about cardinal and intercardinal 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.

In the original paper, Large Reasoning Models (LRMs), in which the models have been trained to perform inference-time reasoning, had not yet emerged, though chain-of-thought prompting (CoT) [Wei *et al.*, 2022] was already popular[1]. In the intervening months, LRMs have generally become the most performant LLMs, and this paper provides an update on how LRMs perform on our dataset, as well as some other new LLMs, such as GPT-4.5 preview. Since many LLMs now claim to have reasoning capability, we use the strict definition that an LRM is one which reports the use of reasoning tokens in its response to prompts in our benchmark. In this paper we investigate the following LRMs: azure-o1-2024-12-17, azure-o3-mini-2025-01-31, azure-o3-mini-2025-01-31-high, azure-o4-mini-2025-04-16, and deepseek-reasoner. For futher information on LRMs see Xu *et al.* [2025] who provide a survey covering the emergence and architectures of LRMs.

---

[1]We did not explore CoT as we wanted to measure raw LLM performance, independent of any particular prompting strategy; we do not explore CoT prompting in this paper either.

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 [Li *et al.*, 2024a; Shi *et al.*, 2022] which aims 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. However, performance increases markedly when an LLM is used to translate from the English specification to a logical representation and symbolic reasoning is used to compute the relationship. It should also be noted that the setting in StepGame is rather artificial compared to the naturalistic settings in our dataset described below. The SpartQA dataset [Mirzaee *et al.*, 2021] is also focused on assessing spatial reasoning, but does not contain any CDs. The bAbI dataset [Weston *et al.*, 2016] 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 [Yamada *et al.*, 2024] 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; investigations [Cohn, 2023; Cohn and Blackwell, 2024a] into various LLMs' 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. Li *et al.* [2023].

3 Experimental Design

Whilst testing compositional reasoning with CDs would be of interest, here we restrict ourselves to testing simpler reasoning abilities. It would be impractical to generate a substantial question set manually and so we used an automated, template driven approach. 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 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, six question templates (T1-T6, shown in Figure 5) were chosen to test LLM performance more comprehensively. Note that five of these concern areal objects (lake, island, park), one concerns a linear object (a road, T4); of the areal objects, the island differs from the other two, in that when on the, say, east shore of an island the sea is to one's east, differently to being on the east shore of lake, where the lake is to the west.

We then exhaustively generated all forms of these questions for all cardinal and intercardinal 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[2]. Following earlier evidence [Raman *et al.*, 2024] that an LLM's performance can vary depending on whether a question is phrased as using the first or second person, 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 templates × 10 locomotion forms × 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 [Raman *et al.*, 2024]. We test state-of-the-art models, favouring larger offerings. However, we did test four smaller models (OLMO 2, DeepSeek R1 7B, LLama 3 7B, and Mistral Small 24B) which were able to run locally on an Apple M2 Pro MacBook using Ollama[3] for comparison. All LLM experiments were conducted using the Golem software[4]. An advantage of Golem is that all HTTP API requests and responses are logged, providing a comprehensive record of experiments and results in JSONL[5] format. Unlike some other benchmark studies, all model versions, settings, hyper-parameters, and results are preserved for subsequent analyses, facilitating transparent and reproducible science, as advocated by Burnell *et al.* [2023].

3.1 Prompting

We used zero-shot prompting with a system prompt of "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 presented each question in a new chat. We included "No yapping." since that has been reported as being beneficial in persuading an LLM to be brief in its response [Stifter, 2024].

Where models allow the temperature to be explicitly set (all models except o1, o3-mini, o4-mini and GPT4.5), we set

---

[2]Note that for T4, there could potentially be twice the number of choices if one were to consider, e.g. "[west-east]" as well as "[east-west]". This could conceivably make a difference to an LLM since in one case the agent is walking in the direction specified, and in the other case in the opposite direction.

[3]https://ollama.com, accessed May 2025.

[4]https://github.com/RobBlackwell/golem, accessed May 2025.

[5]https://jsonlines.org, accessed May 2025.

temperature = 0 to try to achieve deterministic answers, although we show [Blackwell *et al.*, 2024] (and in the results below) that this cannot in general be achieved on models run in the cloud.

We use case-insensitive string comparison and remove spurious punctuation and white space before comparing answers. Our prompts ask for cardinal or intercardinal direction answers only, but allow answers such as "The lake is to the west" if the intended answer is "west", recording such instances where answers do not strictly meet the rubric. We assess performance using accuracy, using prediction intervals [Blackwell *et al.*, 2024] to report variability (where practical).

4 Results

None of the models tested are accurate for all questions (Fig. 1), although all models have higher accuracy than the guess rate. The best performing model in Cohn and Blackwell [2024b] was GPT3.5 Turbo (0.60), but the best performing model now is o1 (0.92), an increase in accuracy of more than 50% in a year. The top 22 results are all commercially-hosted Large Reasoning Models. The best performing locally-hosted model is mistral-small-24b (0.41). In the original paper we only ran models once, and have not conducted further runs here; new models were run more than once to obtain error bars (even with temperature = 0 there can be non determinism [Blackwell *et al.*, 2024]), except when cost precluded this – many LRMs are expensive, e.g. o1 cost $458 for one run.

We also see slightly different accuracy when accessing the same OpenAI model via the Azure API versus OpenAI API (GPT3.5-Turbo and GPT4-Turbo) and further analysis suggests that these are statistically different [Blackwell *et al.*, 2024].

Where confusion occurred, it was mostly with intercardinal directions rather than cardinal directions (e.g. Fig. 2). It can be seen that confusions are not symmetric in general which is perhaps revealing about the reasoning process, as a purely logical reasoner would be expected to produce symmetric computations.

Reasoning tokens consumed: LRMs report the number of completion tokens used for reasoning about a question, and we investigate the distribution of the count of reasoning tokens by correct direction (See Fig. 3 for o1). More completion tokens are used for questions involving intercardinal directions than cardinal directions. We see a similar pattern for the other LRMs tested. We also note that the interquartile range is much larger for the intercardinal directions. Finally, we observe that, for o1, north is the easiest cardinal direction and east the most difficult, whilst north-west appears to be the easiest intercardinal direction, and southwest the most difficult.

We also investigated how many reasoning tokens were consumed when the LRMs got the answer right compared to when they got the answer wrong (see (Figure 4). Three of the LRMs (o1, o3-mini High, and Deepseek R1) have a noticeably higher median count of reasoning tokens used for incorrect answers than correct answers. o3-mini uses more reasoning tokens when reasoning_effort[6] is set to "High" rather

---

[6]https://platform.openai.com/docs/guides/reasoning?api-mode=
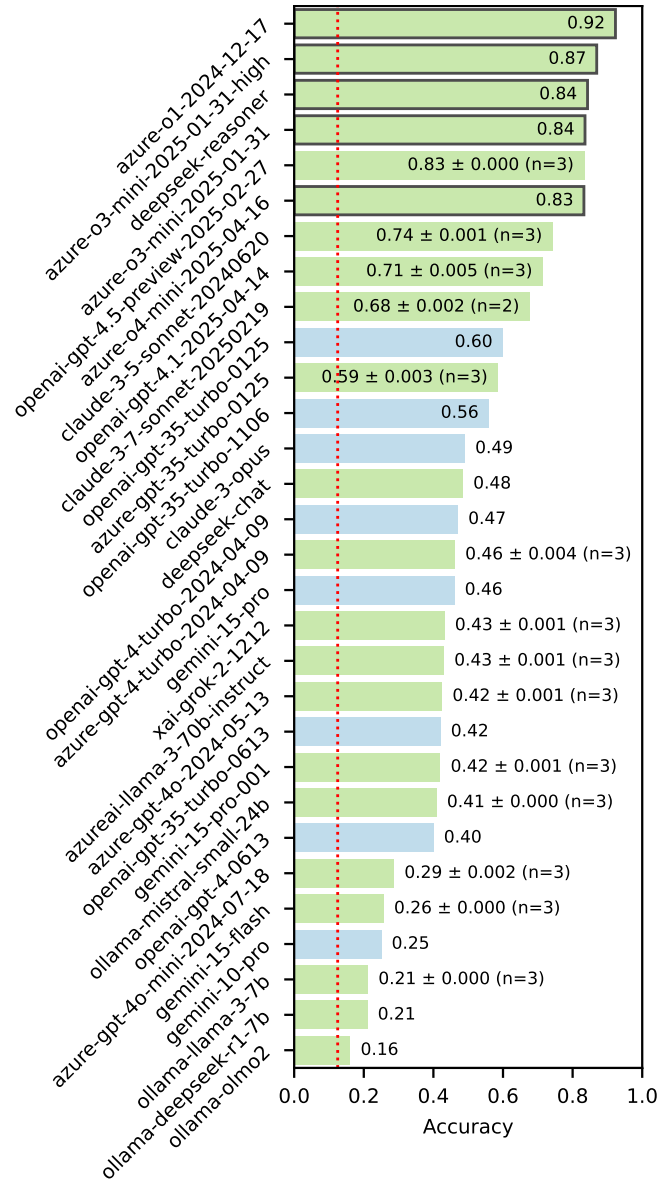

Figure 1: Accuracy by LLM. Results shaded in blue are from Cohn and Blackwell [2024b]. Results with a black border are LRMs. The red dotted line is the guess rate (0.125, since there are eight possible answers). Where possible we use the model names and versions reported in the LLM response, prefixed by the API provider (e.g. openai or azure). Models run locally are prefixed with ollama.

than the default "Medium". o4-mini has a slightly lower median for correct compared to incorrect answers but has a very long but small tail of correct answers requiring a very large number of reasoning tokens.

Fig. 5 analyses seven models' performance on CDs, person, locomotion and template – we choose just seven models since looking at all the models in Fig. 1 would be overwhelming. We chose the best model overall (o1), the best non LRM (GTP4.5 preview), the best model in the
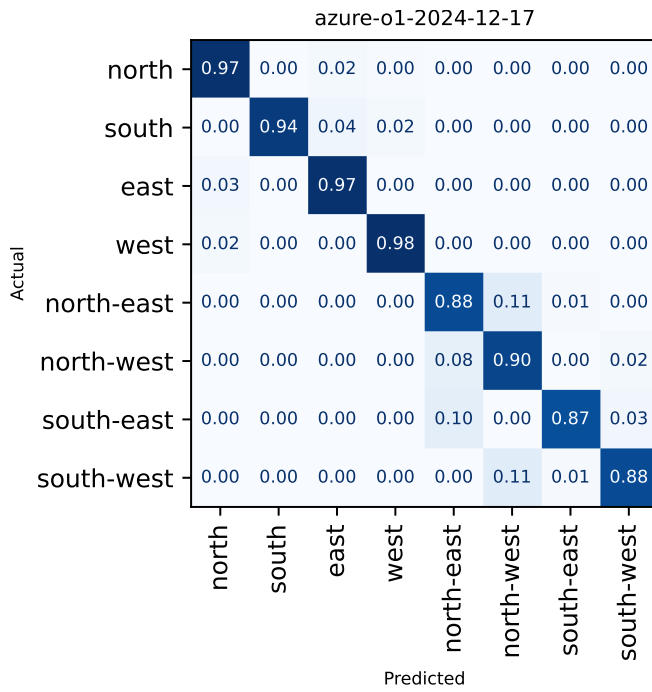
chat

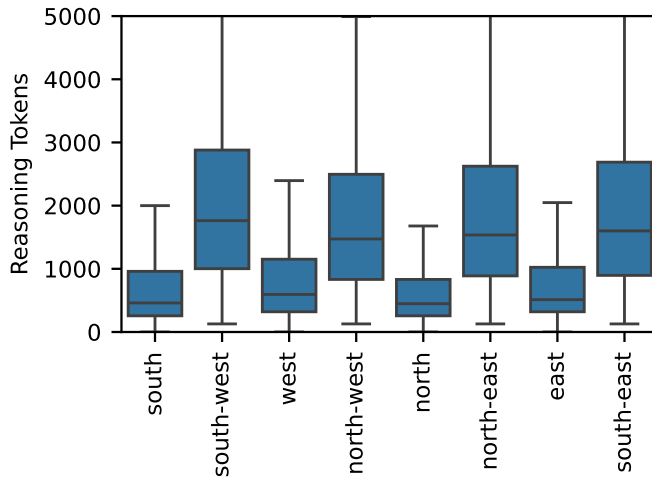Figure 2: Confusion matrix for the best performing model, o1.

Figure 3: Median and inter-quartile range of o1 reasoning tokens by ground truth direction.

earlier paper (GPT3.5T), the best LRM not from OpenAI (Deepseek-reasoner), the deepseek model from which this was built (Deepseek-chat), the largest open weights model tested (LLama 3 70B instruct) and the best model which can be run on a laptop (mistral-small-24b). The key point about varying the person form and the locomotion type is that such changes should not make any difference to the accuracy as the result is clearly (at least to a human) independent of such variations; but as can be seen in the top right and lower left charts in Fig. 5, there can be considerable variations in accuracy. For cardinal directions, one would have expected that if a model can solve for one choice of directions in the prompt,
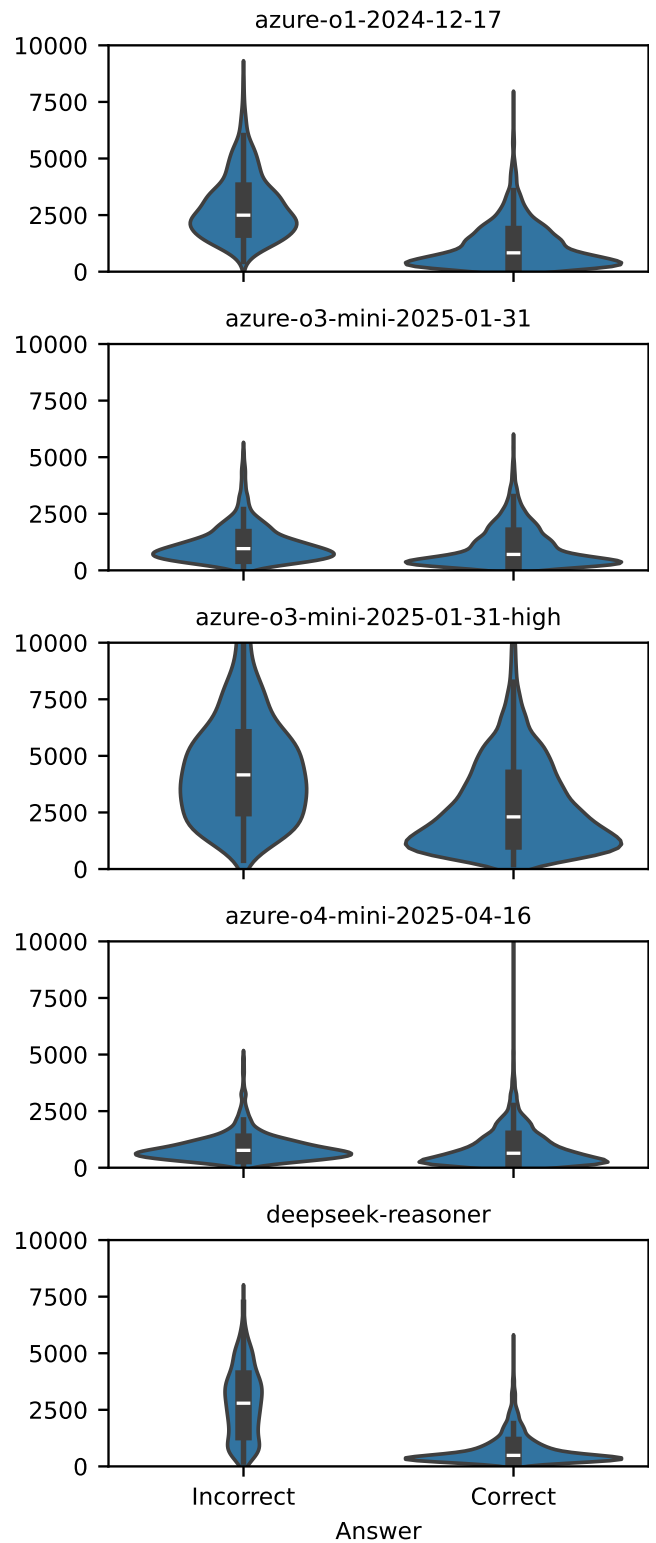

Figure 4: Reasoning token counts for correct and incorrect answers by LRM. The white bar shows the median. Note that all LRMs tested have more correct than incorrect answers and so incorrect sample sizes are small.

then swapping to a different set is mathematically/formally equivalent, but again we see this is not the case.

Cardinal Directions: all models except GPT4.5[7] had higher overall accuracy with cardinal than intercardinal directions[8] though there were some notably strange results such as with Claude Sonnet where for example, South East had higher accuracy than North (Claude Sonnet 3.5: 0.825 / 0.614, Claude Sonnet 3.7: 0.801 / 0.607). The previously best model, GPT-3.5T still performs very competitively on the cardinal directions but is let down on its performance on the intercardinals, suggesting that the reasoning available in LRMs really helps for intercardinals (though GPT-4.5 which is not a reasoning model has similar performance across both CDs and intercardinals). Deepseek-reasoner and GPT-4.5 have a very similar performance, though with GPT-4.5 slightly outperforming Deepseek-reasoner on the intercardinals, whilst the situation is reversed on the cardinal directions. It is striking how much better Deepseek-reasoner is compared to the base model, Deepseek-chat, from which it was trained. The smallest model, Mistral, which was run on a laptop, is the bottom of the pack, except for north, northeast, and east, yielding an unexplained asymmetry.

Person form: It can be easily observed that the top five models form "concentric hexagons", showing reducing performance from o1 through to Deepseek-chat, with roughly similar performance for each person-case; the bottom two LLMs, Llama-3-70b and Mistral-24B jostle for position with Llama outperforming Mistral on "He is", "They are", and "She is", but slightly worse on "You are". GPT-3.5T is worse at "she is" versus "he is" suggesting a possible slight gender bias (possibly as a result of biased training data) – this is also true of some of the other models not selected for this figure but in Fig. 1.

Locomotion form: The best four models along with Llama-3-70b form concentric decagons, showing strictly reducing accuracies. However Deepseek-chat and Mistral-small-24b show more variation with each outperforming their nearest rival for unicycling, and the latter outperforming Llama-3-70b for hiking. o1 does best overall on hiking, which maybe surprising when according to our investigations walking is likely to appear more frequently in the training data – but perhaps hiking appears more frequently in contexts where cardinal directions are used (such as trail maps/instructions). The original model, GPT-3.5T is best for walking and hiking, and worst for unicycling, racing and driving.

Template variation: If we look at the performance of all the seven selected models across the six different templates

---

[7]Please see our conclusions about GPT4.5 which we suspect may have been trained on our benchmark.

[8]It is interesting to note that humans have been reported to be worse at "oblique" directions, rather than the two orthogonal axes [Meng and Qian, 2005; Appelle, 1972] – the so called "oblique effect". It is suggested to have a neuropsychological basis, though we are not suggesting the effect here is for the same reason. A more likely explanation, in the world of large *language* models, might be that the intercardinal relation names are composed of two words (in English and many other, but not all languages), and this might create confusion in an LLM.

(see Fig. 5, bottom right) then the most striking aspect is that T4 (the only template concerning a linear rather than an areal object) has the worst accuracy, by a substantial margin (this is also the case for the other LLMs in Fig. 1. We do acknowledge that the wording of T4 could be improved but this alone would not seem to account for this effect. Worst of these was GPT-4.5 preview which got all of the T4 questions wrong (91% of the time it gives an answer 90 degrees off (i.e. consistent with the road direction). The remainder of the time it gives an invalid answer echoing the road description, e.g. north-south) – and all the other questions on the other templates completely correct; given the probabilistic nature of LLMs this seems statistically unlikely, but we do not have a good explanation for it.

o1 does best on T4 (0.55), followed by GPT-3.5T (0.41); apart from its performance on T4, GPT-4.5 is pretty similar to o1. T2 is a slightly more complicated variant of T1 (since it involves "head[ing] back in the direction you came from" and it can be seen that this results in a slight performance dip for each of the models. T3 is very analogous to T1, except for a change of scenario from lake to a park; many models achieve similar accuracy, though Deepseek-chat is worse, as is Mistral-small-24b, whilst GPT-3.5T has better accuracy. T5 and T6 are similar to T1 and T2 respectively, but the answer is the opposite in in each case (a lake is west of its east shore but the sea is east of the east shore of an island); o1 and GPT-4.5 have very similar performances in each case, whilst GPT-3.5T is about the same on T1/T5 but better on T6 compared to T2. We might have expected LLMs to confuse the direction of the lake (T1,T2) and the sea (T5,T6), and 180 degrees is indeed the most common confusion for five models (o3-mini, o4-mini, Claude 3.5 Sonnet, Claude 3.7 sonnet, and gemini-15-pro). However in most cases (11 models) the most common confusion is 90 degrees from the correct answer suggesting that confusion is not systematic.

5 Discussion and Conclusions

None of the models tested is able to reason completely reliably about cardinal direction; however all models show some capacity for spatial reasoning, and o1 in particular. A model that randomly selects cardinal or intercardinal direction answers to an MCQ would have an accuracy of 0.125, but even the worst performing model OLMO2 achieved an accuracy of 0.16.

None of the models is able to fully generalise person form or locomotion. If a model was truly reasoning we would not expect the mode of locomotion to have any impact on accuracy, nor would the gender of the agent.

Differences in results amongst models of the same name but with different versions and different vendors (e.g. GPT3.5T versions 0613, 1106, and 0125 hosted on OpenAI and Azure) highlight the importance of recording precise experimental conditions in benchmark studies.

As noted, OpenAI GPT4.5 preview (launched 27 Feb 2025) gets all questions correct except for Template 4 questions, where it gets those all incorrect. These results are suspicious and we wonder whether GPT4.5 has been trained on our benchmark.
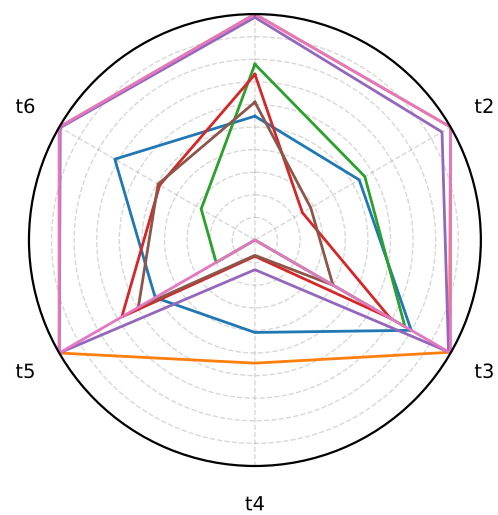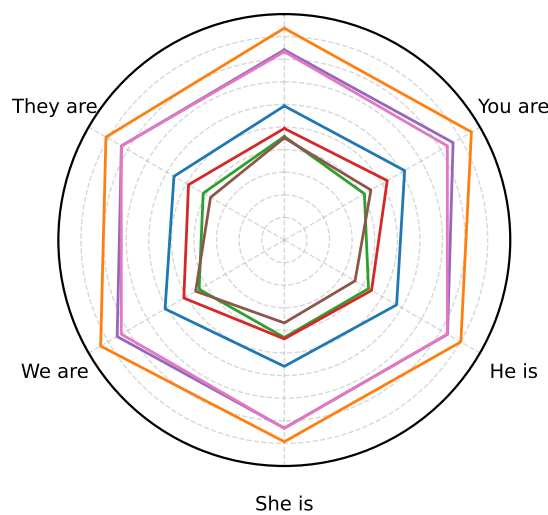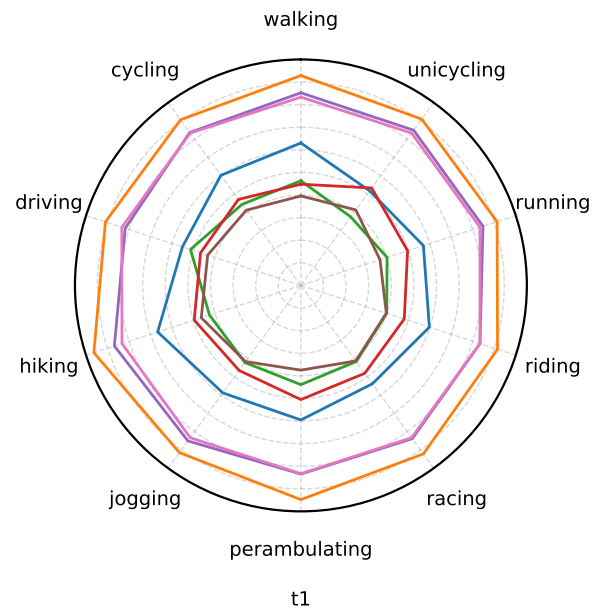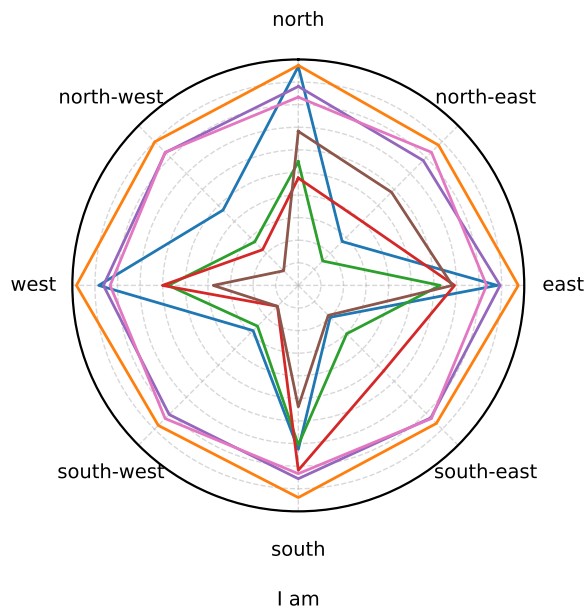
T1. You are walking [south] along the [east] shore of a lake; in which direction is the lake?

T2. You are walking [south] along the [east] shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake?

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

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

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

T6. You are walking [south] along the [east] shore of an island and then turn around to head back in the direction you came from, in which direction is the sea?
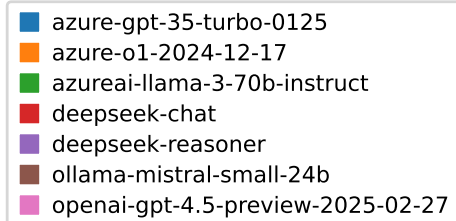
Figure 5: Accuracy by direction, locomotion, person form, and question template for selected models. The grey concentric circles in the background are set at an accuracy of 0.1 apart.

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[9]).

The early success of prompting strategies such as Chain-of-Thought (CoT) and Tree-of-Thought (ToT) have led to models that have been explicitly trained for reasoning at inference time. These so-called Large Reasoning models such OpenAI's o1 and DeepSeek's R1 comprise our four best performing models.

An analysis of reasoning tokens use by o1 shows that more tokens are required for intercardinal rather than cardinal directions. We posit that completion tokens are a proxy for reasoning effort and that intercardinal directions require more reasoning effort, since they are defined in terms of two cardinal directions.

The latest and most sophisticated large reasoning models are expensive to test. Our o1 experiment cost approx. $458 to run with no repeats. By contrast, the OpenAI model, o1 pro[10] costs $150 for 1M input tokens and $600 for 1M output tokens, putting it beyond our budget as it would have cost ∼$4.6k to run a single repeat!

Some LRMs take much longer to respond to prompts than traditional LLMs. In our experiments, Deepseek R1 took a median of 34s to reply (o1 7s, GPT4.5 0.9s, GPT 3.5 Turbo 0.7s). Our 5760 benchmark question set takes more than 48 hours to complete sequentially using Deepseek R1.

Although we have not presented further results on newer LLMs on the small dataset in Cohn and Blackwell [2024b] it is worth noting that we did run all the models reported here on the small dataset and most models had an accuracy of 0.9 or greater. One point of interest is that in one case[11], Ollama deepseek-r1:7b (a distilled version of Deepseek's best reasoning model, R1), we found that with a temperature of 0 it did not terminate (or at least we cut it off after 3 hours on a single prompt); however, when given the same prompt and with the default temperature it terminated very quickly and gave the correct answer.

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 [Bhandari, 2024] 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* [Wu *et al.*, 2024].

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 egocentric directions (left, right, behind, above...) with respect to an actor or other objects with an intrinsic front, or the landscape, or architectural artefacts.

5. We noted earlier that there are languages where the intercardinal relation names are not composed of pairs of the cardinal names, for example: Breton, Finnish and Estonian (https://en.wikipedia.org/wiki/Cardinal_direction), and it would be interesting to see if LLMs do not have this effect of reduced accuracy for the intercardinals in such languages. Experimenting with other languages would also be of interest more generally too.

6. 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 Li *et al.* [2024b]).

7. Consider situations with more than two objects of interest, so that, e.g. compositional reasoning can be tested, and also reasoning about trajectories (cf Weston *et al.* [2016]).

8. Look at CD reasoning in a multi-modal setting, an area that has already received some attention, e.g. Yin *et al.* [2025].

9. Investigate LLM performance in reasoning about the cardinal directions between actual geographic entities – e.g. Fulman *et al.* [2024] built a 14 question benchmark to investigate whether LLMs display distortions in estimations of intercardinal directions between cities similarly to the errors made by humans (e.g. estimating whether Reno is east or west of San Diego).

We would not advocate including "no yapping" in the prompt in future because it may constrain chain of thought in non-reasoning models.

Data availability

The data for all experiments (including the data from Cohn and Blackwell [2024b]) are available in a GitHub repository at http://bit.ly/4e069XO.

Ethical Statement

There are no ethical issues.

Acknowledgments

---

[9]Li *et al.* [2024a] shows that the StepGame benchmark can be almost perfectly solved by using an LLM to translate the problem instances into logical form and then using an ASP reasoner to actually solve the problem.

[10]platform.openai.com/docs/models/o1-pro, accessed May 2025

[11]"On a hike, a duck pond is to your north and the nearest town is south. What direction is the pond from the town?"

References

Stuart Appelle. Perception and discrimination as a function of stimulus orientation: the "oblique effect" in man and animals. *Psychological bulletin*, 78(4):266, 1972.

Prabin Bhandari. A survey on prompting techniques in LLMs. *arXiv 2312.03740*, 2024.

Robert E. Blackwell, Jon Barry, and Anthony G. Cohn. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores. *arXiv 2410.03492*, 2024.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

Anthony G Cohn and Robert E Blackwell. Can large language models reason about the region connection calculus? *arXiv 2411.19589*, 2024.

Anthony G Cohn and Robert E Blackwell. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In Benjamin Adams, Amy L. Griffin, Simon Scheider, and Grant McKenzie, editors, *16th International Conference on Spatial Information Theory (COSIT 2024)*, volume 315 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:9, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

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

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

Nir Fulman, Abdulkadir Memduhoğlu, and Alexander Zipf and. Distortions in judged spatial relations in large language models. *The Professional Geographer*, 76(6):703–711, 2024.

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

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

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.

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.

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.

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.

Xin Meng and Ning Qian. The oblique effect depends on perceived, rather than physical, orientation and direction. *Vision Research*, 45(27):3402–3413, 2005.

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.

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 *International Conference on Machine Learning*, pages 42026–42047. PMLR, 2024.

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.

Florian Stifter. Automated unit test generation and improvement with large language models (LLMs). Master's thesis, University of Linz, Linz, 2024. Supervised by Reinhold Plösch. vi, 131 pages. URN urn:nbn:at:at-ubl:1-82793 https://epub.jku.at/obvulihs/content/titleinfo/11324135.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

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.

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.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv 2501.09686*, 2025.

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

Hang Yin, Zhifeng Lin, Xin Liu, Bin Sun, and Kan Li. Do multimodal language models really understand direction? a benchmark for compass direction reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.