

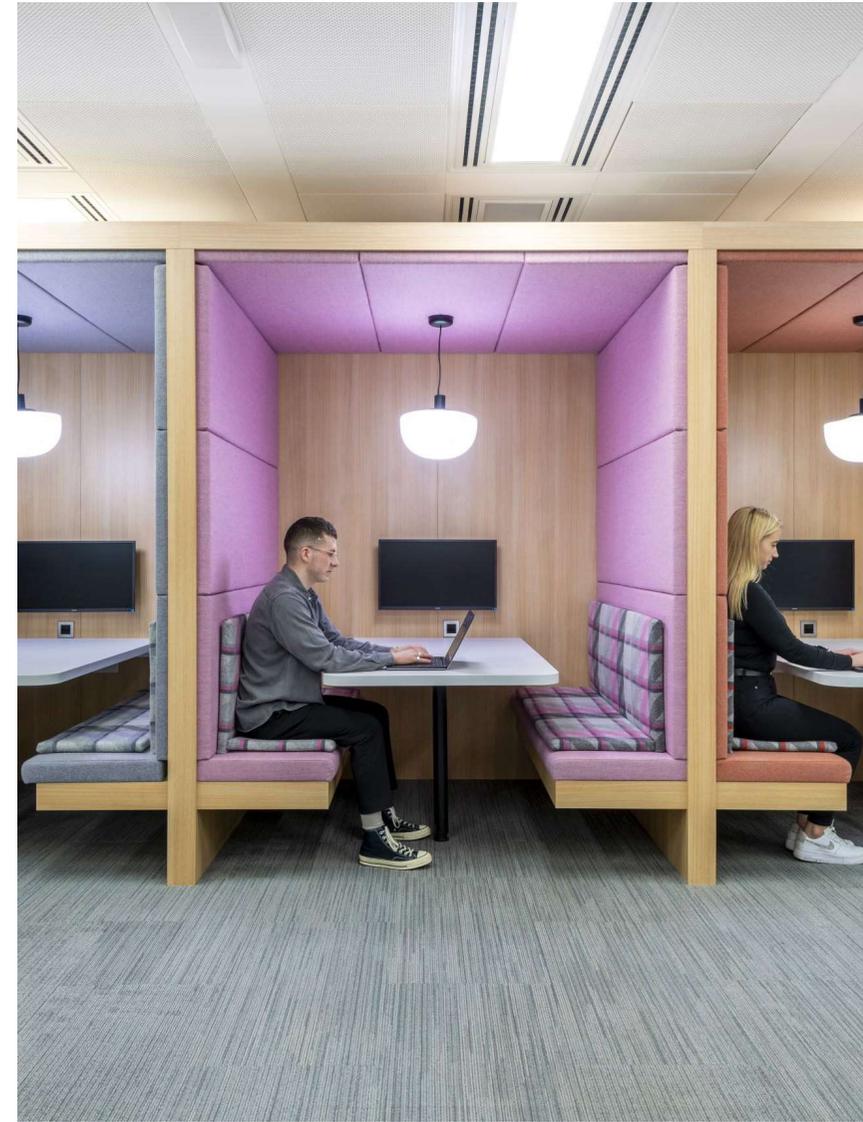
Can AI models reason about qualitative spatial and temporal information?

Robert Blackwell

With thanks to my collaborators:

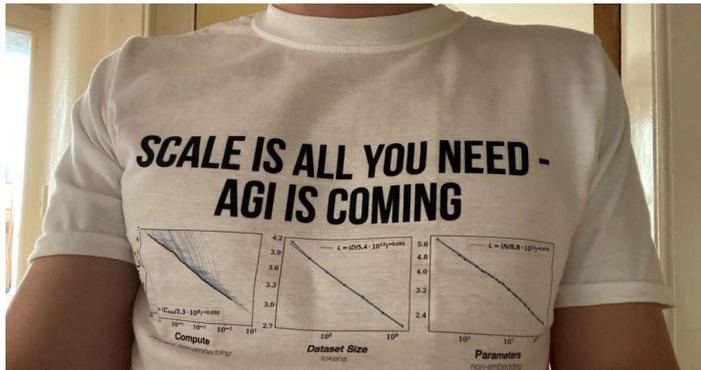
Anthony Cohn (Leeds),
Pranava Madhyastha (City),
Navdeep Kaur (Turing) and
Lachlan McPheat (Turing).

**The
Alan Turing
Institute**



The rise of Generative AI / LLMs

“Reasoning models think before they answer” – OpenAI documentation



Twitter 2021

OpenAI CEO Sam Altman says that the company is confident that it knows “how to build AGI as we have traditionally understood it” Jan 2025

“My guess is that we’ll have AI that is smarter than any one human probably around the end of next year.” Elon Musk, April 2024

“We are going to expand intelligence a millionfold by 2045” Ray Kurzweil

“Deep learning is indeed finally hitting a wall, in the sense of reaching a point of diminishing results.” - Gary Marcus



Give me a picture of a wall clock showing that the time is twenty past three.

ChatGPT 4o 3 April 2025



Grok 3 **beta** ▾

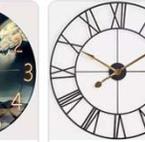
Give me a picture of a wall clock showing that the time is twenty past three.

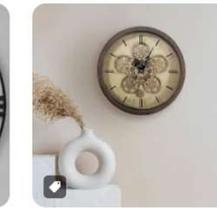
Image



- Living room
- Modern
- Decorative
- Ajanta
- Quartz
- Contemporary
- Ikea
- Art
- Handmade
- Rustic
- Luxury
- Amazon
- Oversized
- Seiko

Sponsored

 <p>Silent large bird modern wall... £89.00 Amy's Deco Free delivery By Google</p>	 <p>Giant living room art deco squar... £89.00 Amy's Deco Free delivery By Google</p>	 <p>Weather Clock With Baromet... £445.00 Bramwell Brow... Free delivery ★★★★★ (392) By Google</p>	 <p>Silent extra large kitchen... £89.00 Amy's Deco Free delivery By Google</p>	 <p>670mm LED Silent Wall Clo... £119.99 Homary UK Free delivery ★★★★★ (30) By Google</p>	 <p>Silent big kitchen... £69.00 Amy's Deco Free delivery By Google</p>	 <p>Silent small living room... £69.00 Amy's Deco Free delivery By Google</p>	 <p>Silent Big kitchen deco... £79.00 Amy's Deco Free delivery By Google</p>	 <p>TimberTime - Scandinavian... £49.95 tryglasskin Free delivery By Producthero</p>	 <p>Northern wooden wall... £39.95 Falkner-Home +£295.00 deli... By Google</p>	 <p>Newgate Battersby Larg... £139.99 Newgate World Free delivery By Google</p>	 <p>Newgate Putney Large Retro... £124.99 Newgate World Free delivery By Google</p>	 <p>12-Inch Silent Glass Wall Clo... £9.38 Temu Free delivery By Google</p>	 <p>John Lewis Large Roman... £80.00 John Lewis Free delivery ★★★★★ (66) By Kelkoo</p>
--	--	--	--	---	--	---	---	---	---	--	--	--	--

 <p>Amazon UK · In stock Modern Silent Pendulu...</p>	 <p>Etsy · In stock Large Wall Clock / 18" ...</p>	 <p>Amazon UK · In stock Large Modern Wall Clo...</p>	 <p>Decorexi · In stock Wall Clocks Chiswick ...</p>	 <p>Picture Perfect Home · I... Moving Gears Round ...</p>	 <p>Curvissa · In stock Jones Clocks 'The...</p>	 <p>Gracie Jaynes · In stock Rustic Wall Clock - Gra...</p>	 <p>Home Store + More Pocket Watch Wall Clo...</p>	 <p>ManoMano · In stock Wall Clock 30cm Diam...</p>	 <p>Next UK · In stock Buy Jones Clocks ...</p>
 <p>eBay · In stock</p>	 <p>Dunelm.com · In stock</p>	 <p>Closs & Hamblin · In stock</p>	 <p>Ubuy UK · In stock</p>	 <p>Housing Units · In stock</p>	 <p>Ruma</p>	 <p>Fusion Living · In stock</p>	 <p>HomesDirect365 · In sto...</p>	 <p>Barnardo's · In stock</p>	 <p>Next UK · In stock</p>

Give me a picture of a cup of hot tea turned upside down.



ChatGPT 4o January 2025



ChatGPT 4o 3 April 2025

Bob is standing on the Northwest side of a mountain. Give me a picture of the scene from the Southeast side?

Nice picture, but the spatial configuration is wrong.

Bob ought to be behind the mountain and so unseen.



Here's the image of the mountain viewed from the Southeast side, with Bob standing on the Northwest side. Let me know if you need any adjustments!

I am walking Southeast. I look down on my compass. Please give me a picture of the compass.



ChatGPT 4o January 2025



ChatGPT 4o April 2025

My sandwich cost £2.73 and I paid with a £5 note. Give me a picture of my change.

Your change from a £5 note after buying a £2.73 sandwich is:

- £2.27 total change
 - £2 coin
 - 20p coin
 - 5p coin
 - 2p coin

I'll generate an image of this combination of UK coins for you.

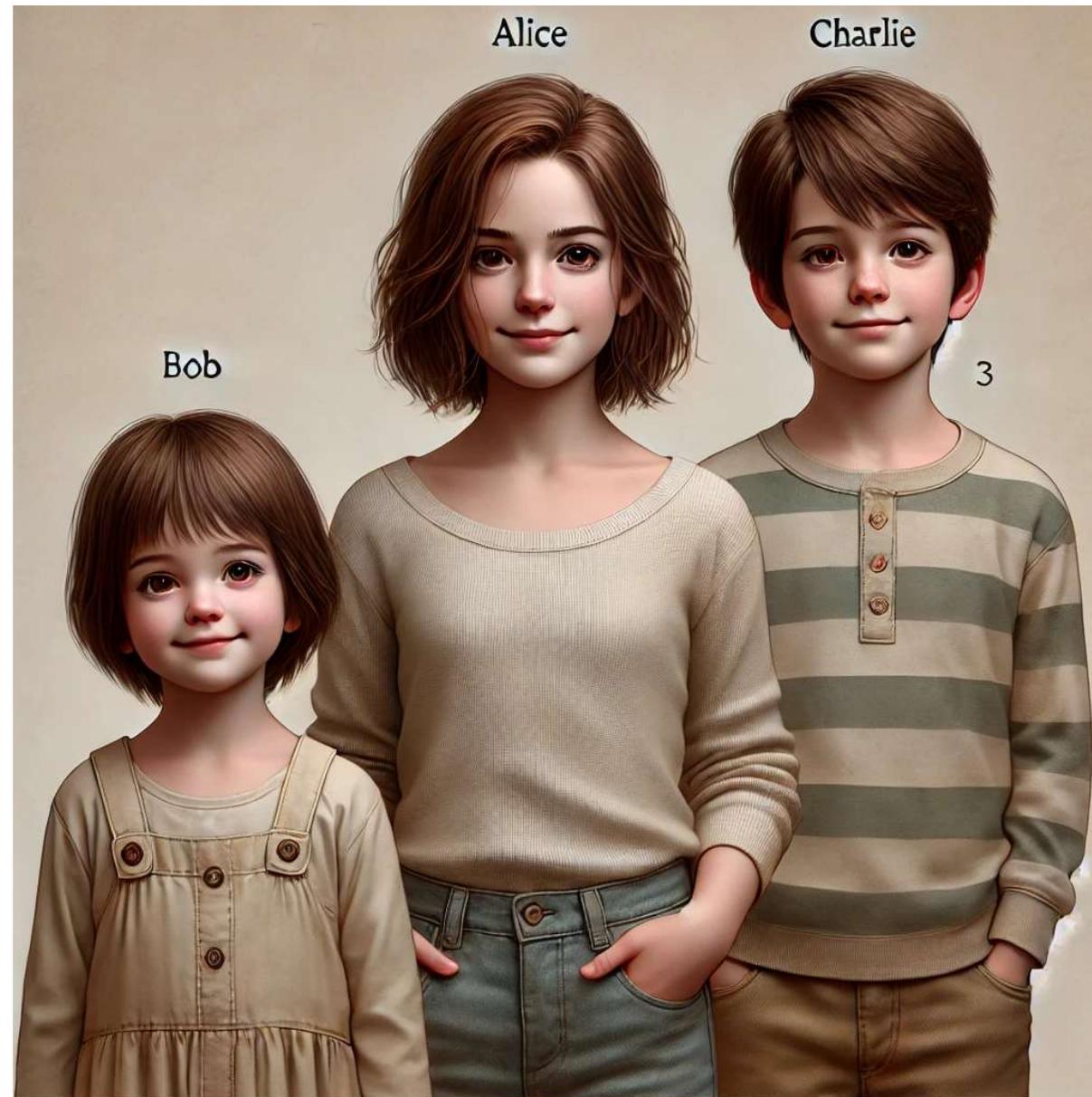


Here's an image of your change, showing £2.27 in British coins. Let me know if you need any adjustments!

Wrong number of coins, incorrect denominations and inconsistent with UK coinage.

Give me a picture of three children stood side by side in order of height. Alice is 7 years old, Bob is 16 years old and Charlie is only 3 years old.

ChatGPT 4o January 2025



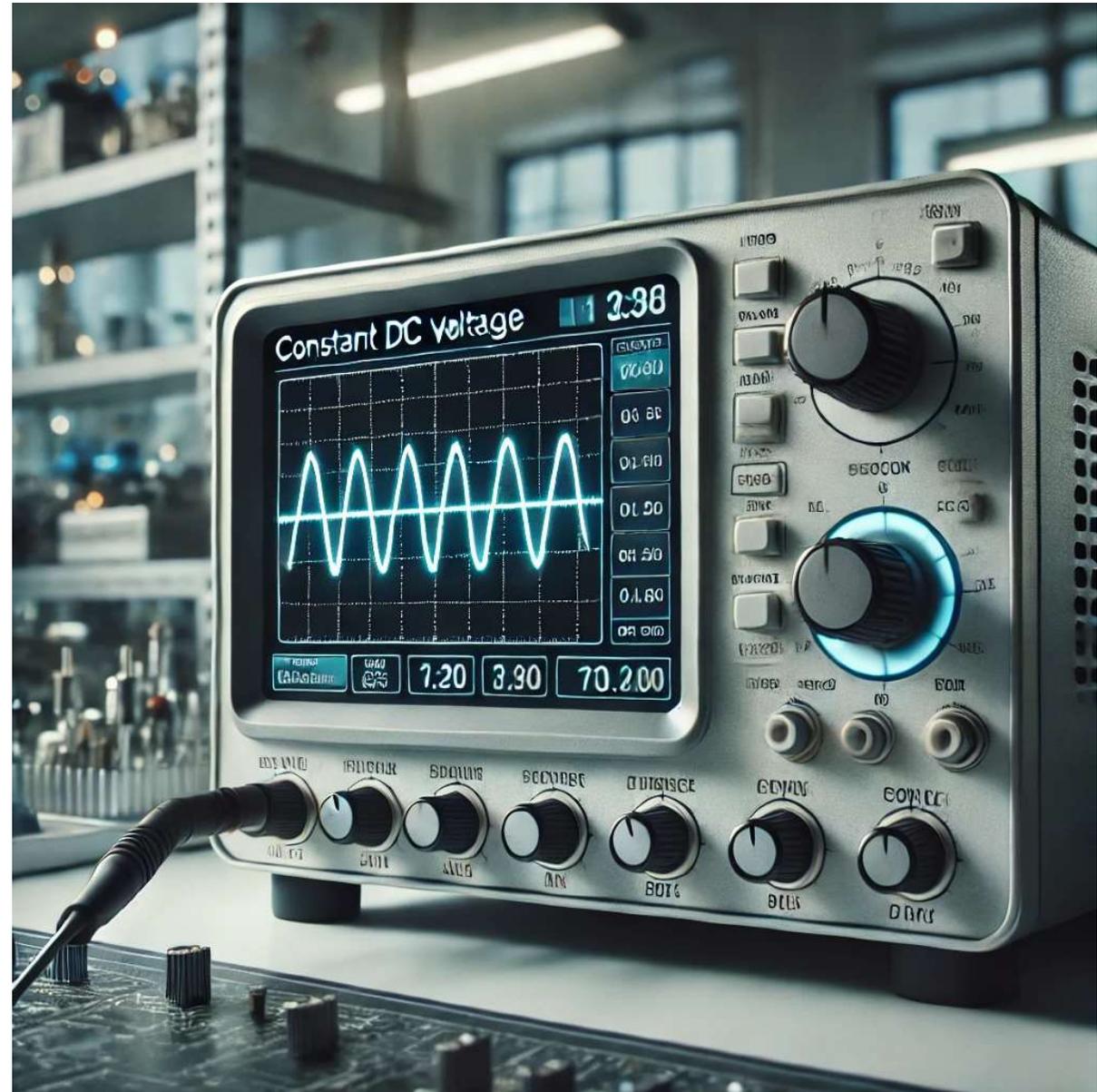
Now order them
alphabetically by
name

(Good but Bob has
apparently changed
gender and has gone from
shortest to tallest).

ChatGPT 4o January 2025



Give me a picture of a digital oscilloscope measuring a constant DC voltage.



ChatGPT 4o January 2025

What do we mean by commonsense reasoning?

Benchmarks for Automated Commonsense Reasoning: a Survey

(E. Davis, arXiv:2302.04752, ACM Comp. Surv. 23)

- is distinguished from common knowledge, encyclopaedic & expert knowledge
- Is concerned with generalities rather than individuals
- is integrated into other cognitive abilities (language, vision...)
 - One never observes it directly, only how it is manifested through language, action...
- is independent of any modality or task i.e. can be (re)used in multiple situations.
- has broad scope (time, space, physical, biological, social realities; meta-knowledge)
- Is not book learning or explicitly taught in schools

Davis gives desiderata for CS benchmarks and notes many benchmarks depart from these <http://cs.nyu.edu/~davise/Benchmarks/>

touching
above
left of
on
part of
under
in
north of
below
over

What do we mean by qualitative spatial and temporal reasoning?

before

during

after

Reasoning about Cardinal Directions

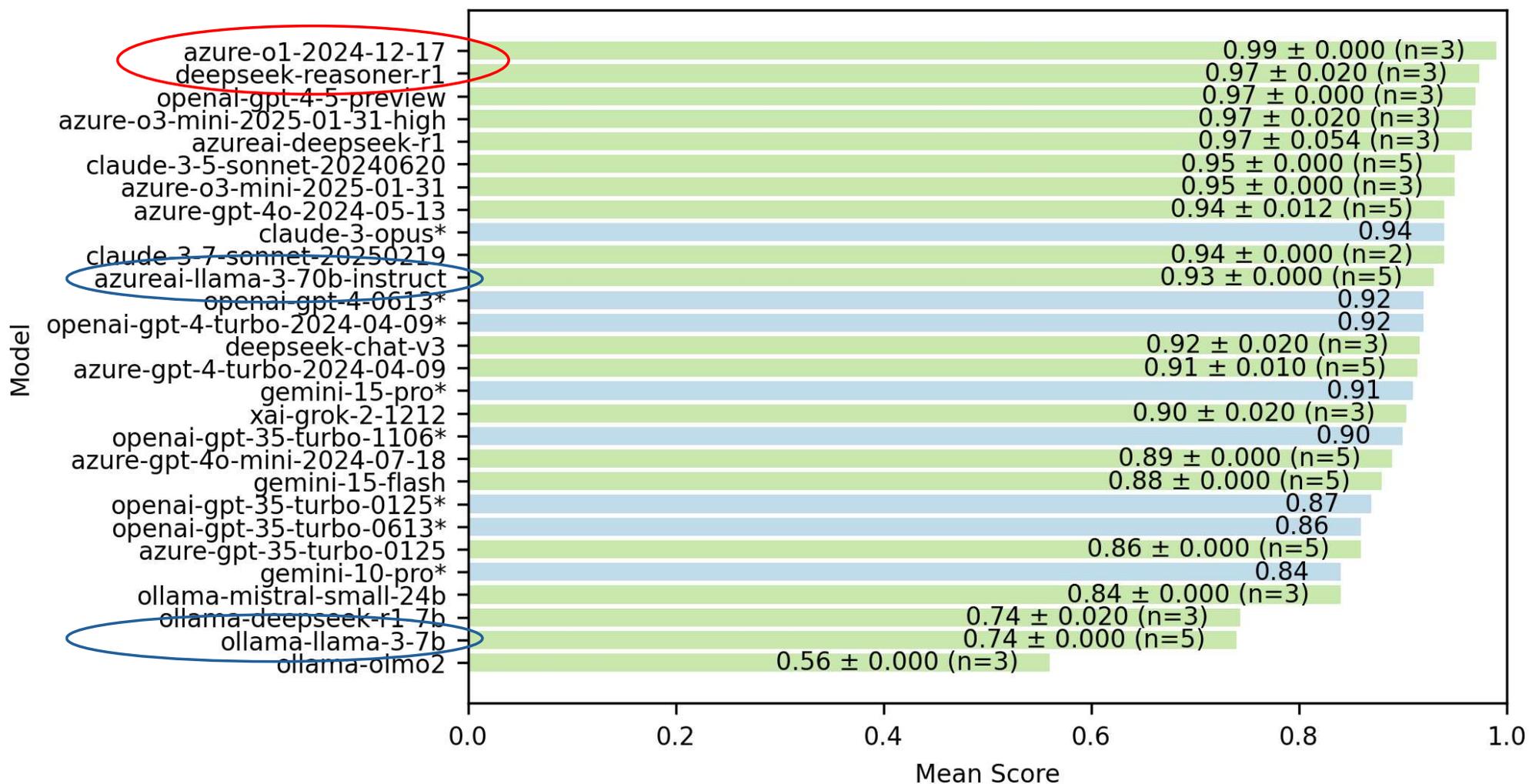
(Cohn & Blackwell, COSIT-24)

Small – 100 simple questions with cardinal directions (N, S, E, W) as answers

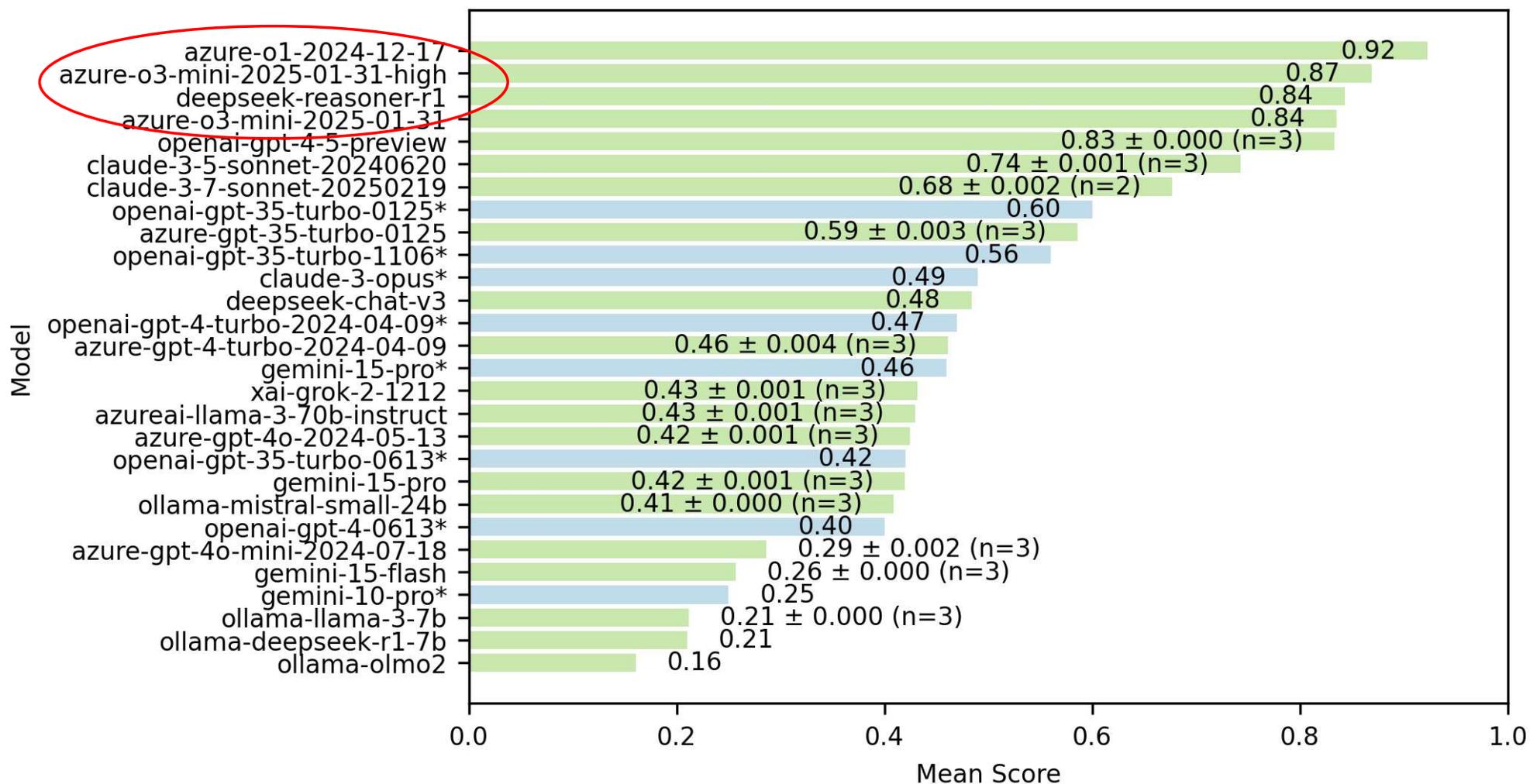
1. *You are watching the sun set. Which direction are you facing?*
2. *If the South Pole is behind you, which direction are you facing?*
3. ...

Large – 6 templates * 10 locomotion * 6 person forms * 8 directions (cardinal and inter-cardinal) * 2 direction variations = 5760 questions

1. *You are walking south along the east shore of a lake; in which direction is the lake?*
2. *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?*
3. *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?*
4. *You are walking east along the south side of a road which runs east to west. In which direction is the road?*
5. *You are walking south along the east shore of the island. In which direction is the sea?*
6. *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?*

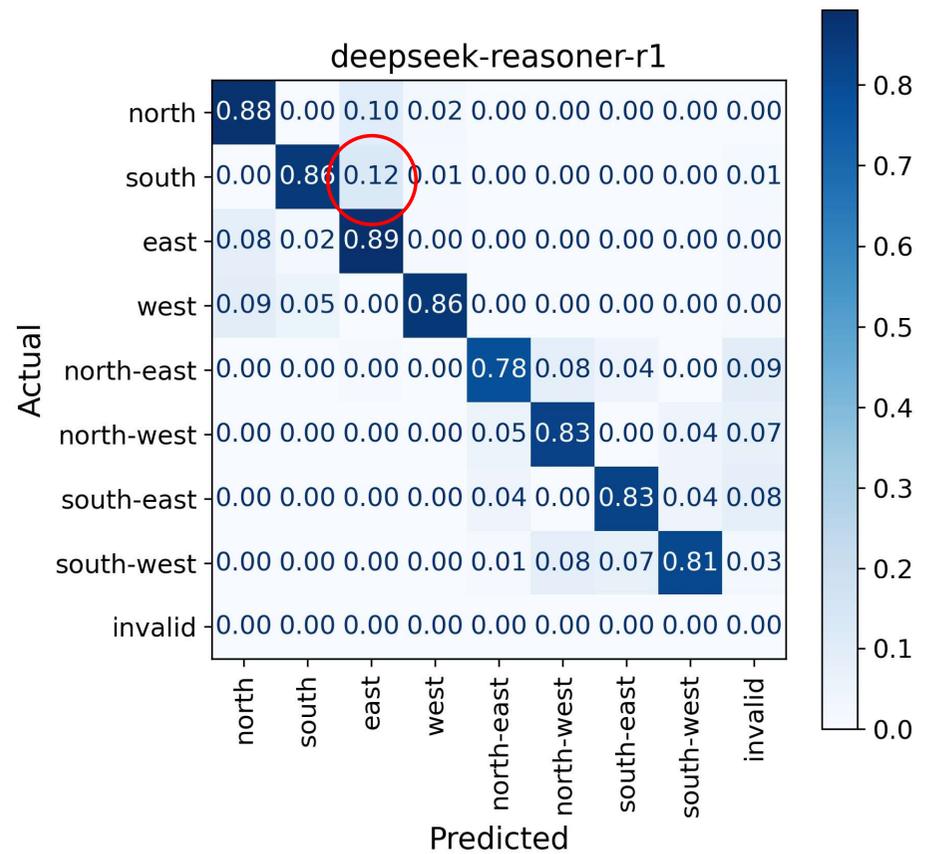
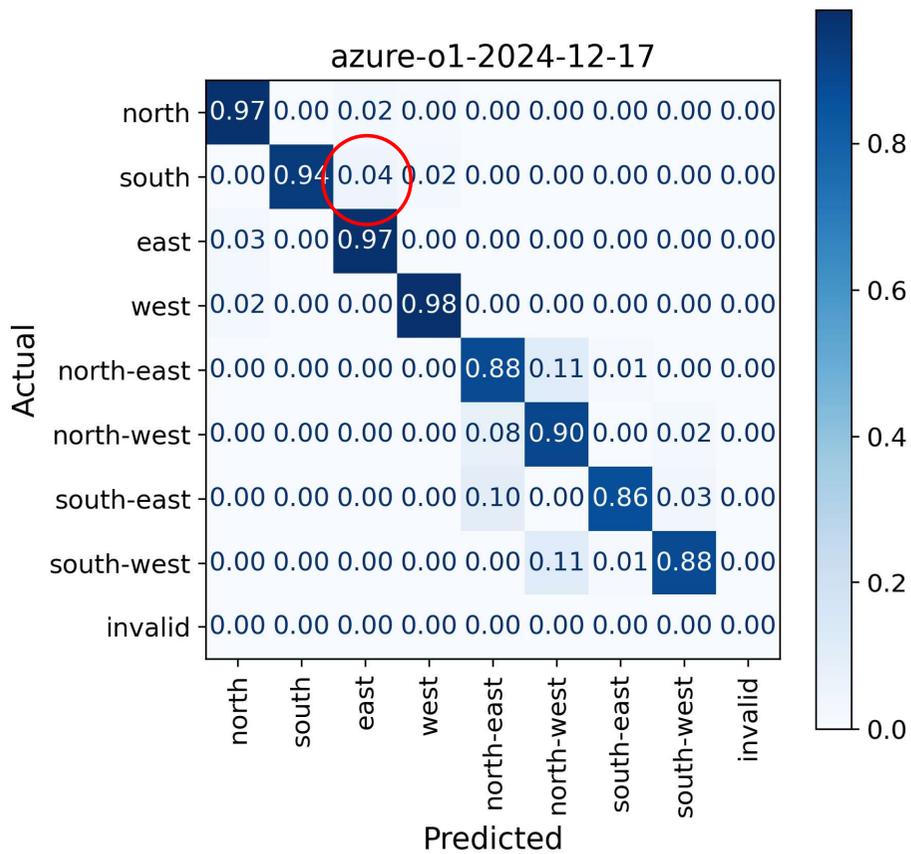


Accuracy by model for our **small** data set (100 questions, temperature = 0.0, fixed seed where available). Blue : results from the paper. Green: new mean ± prediction interval for n repeats.

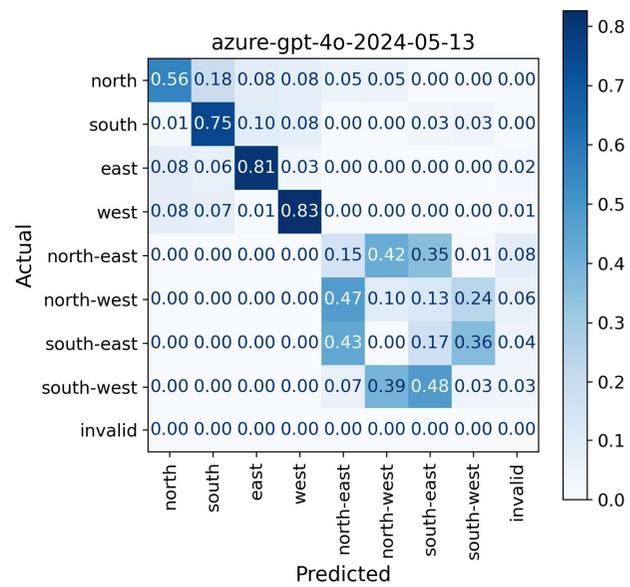
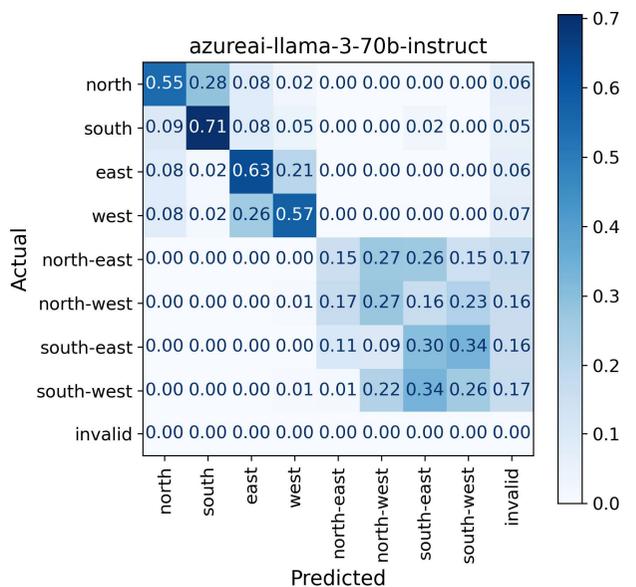
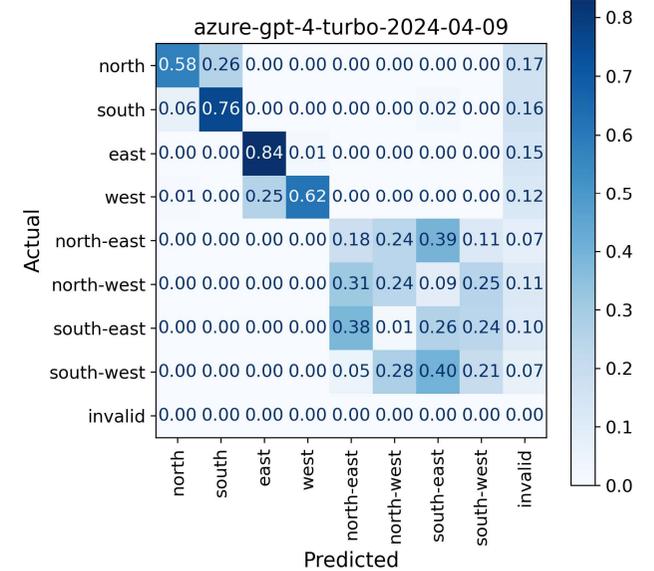
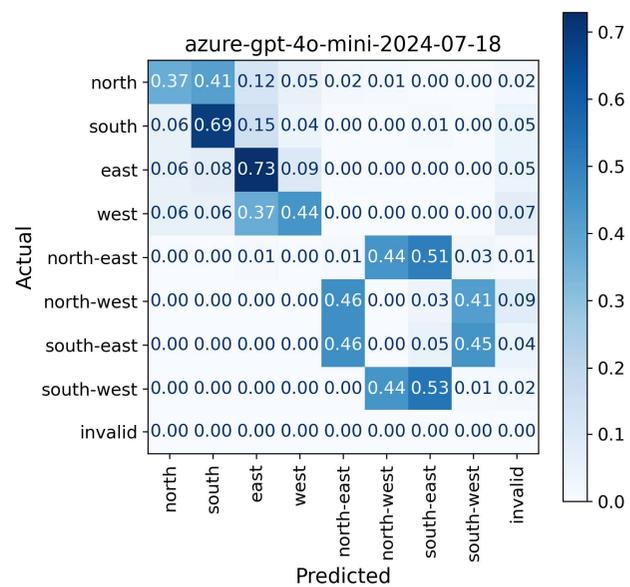
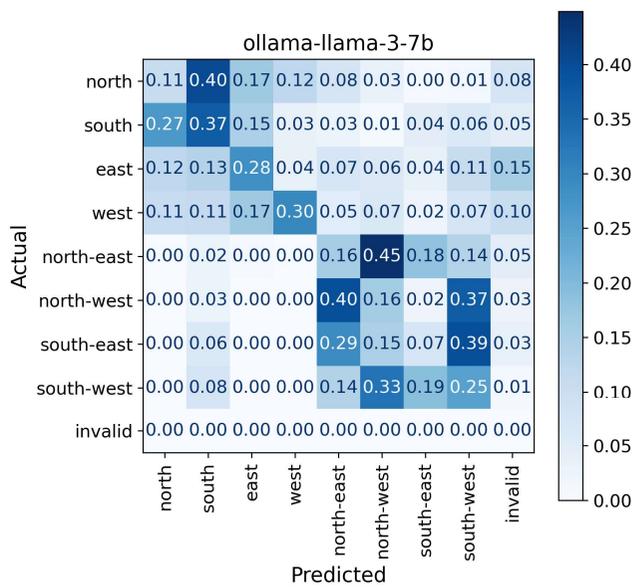


Accuracy by model for our **large** data set (5760 template-driven questions, temperature = 0.0, fixed see where available). Blue : results from the paper. Green: new mean \pm prediction interval for n repeats.

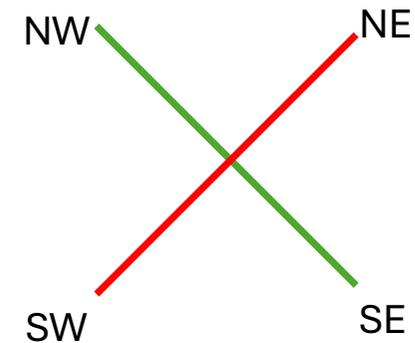
Confusion matrices for Large (5760 questions)



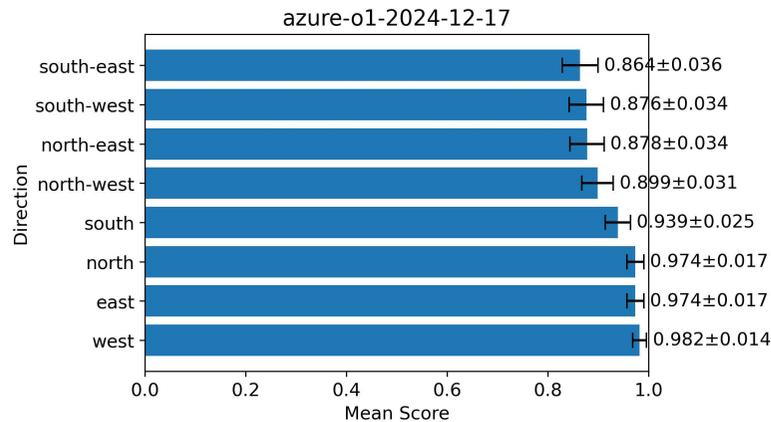
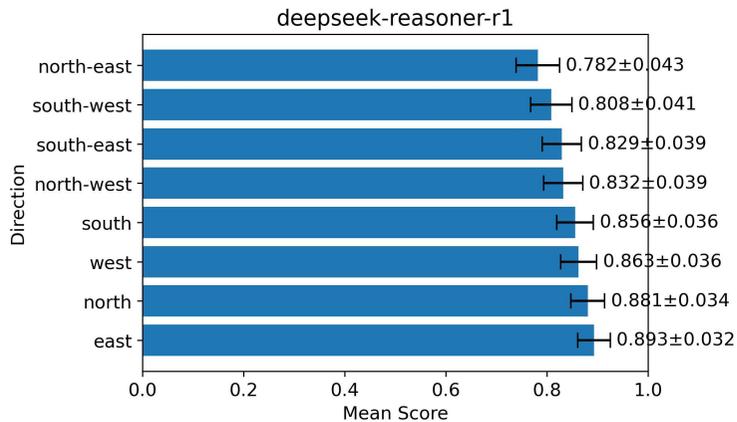
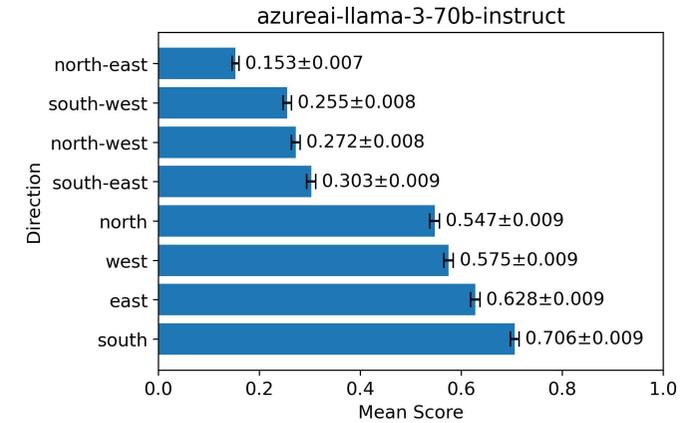
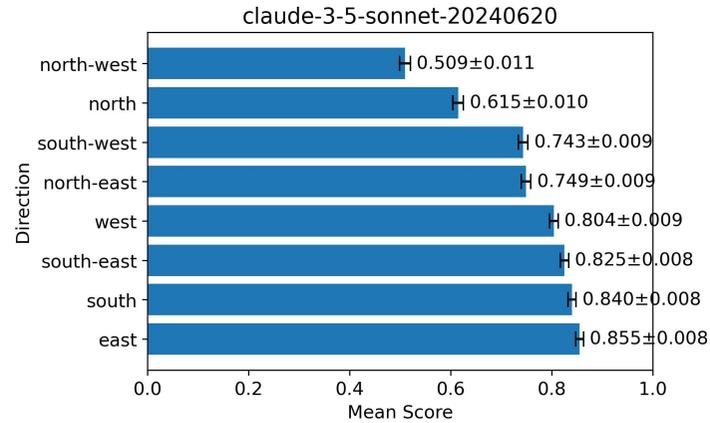
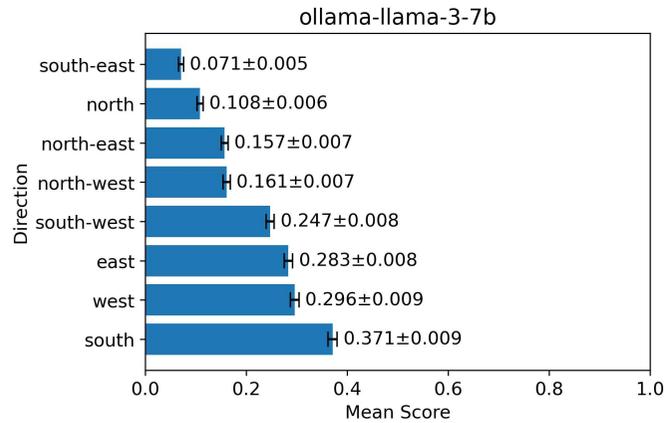
Note: confusions not always symmetric.



- Large question set
- Confusion matrices for other tested models.
- Note “squares” in lower right



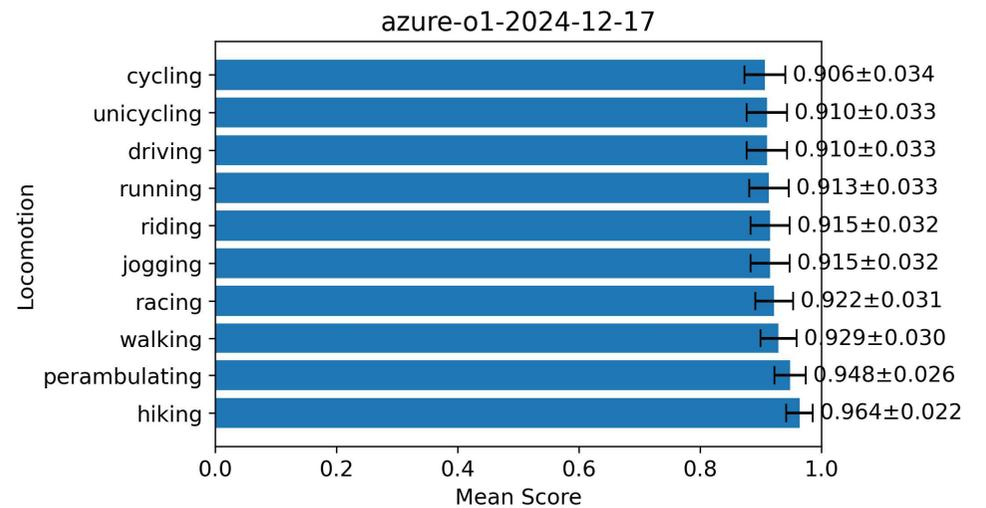
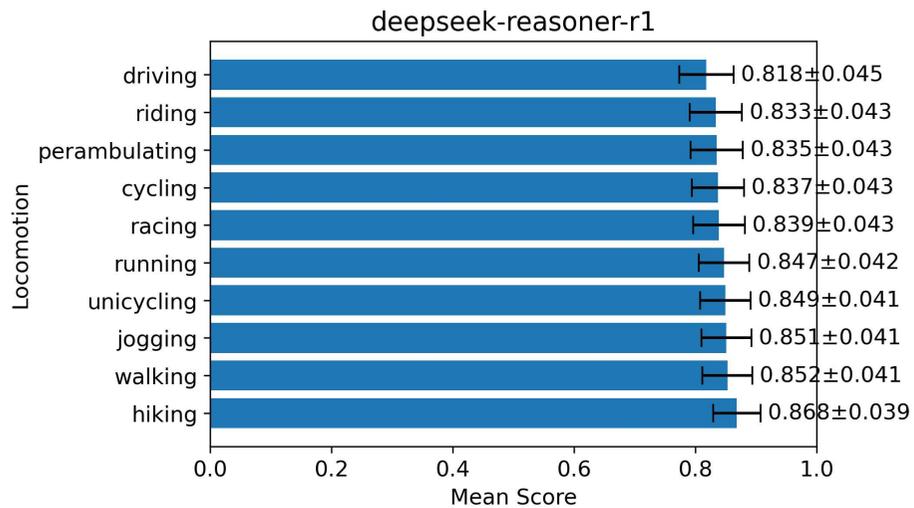
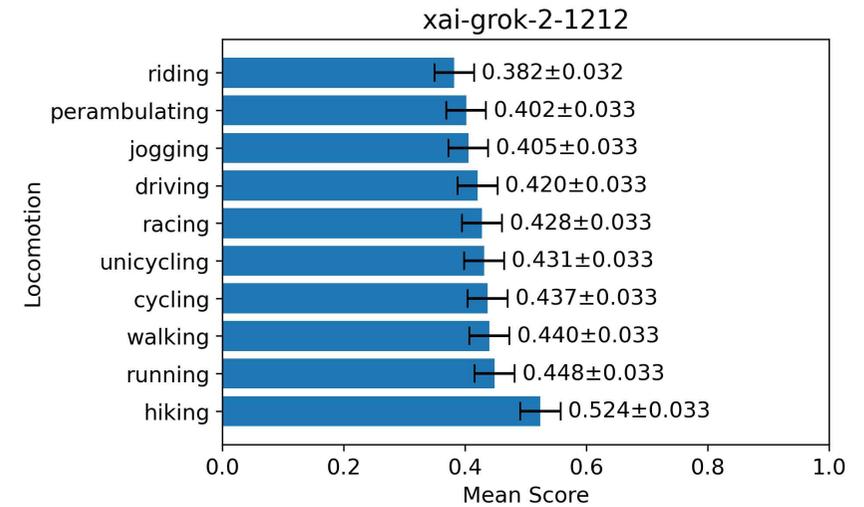
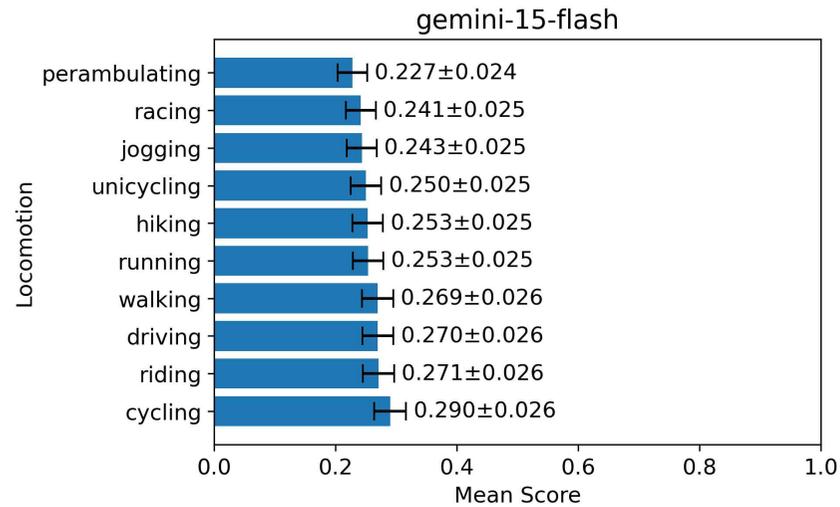
Mean score by correct answer direction for selected tested models.



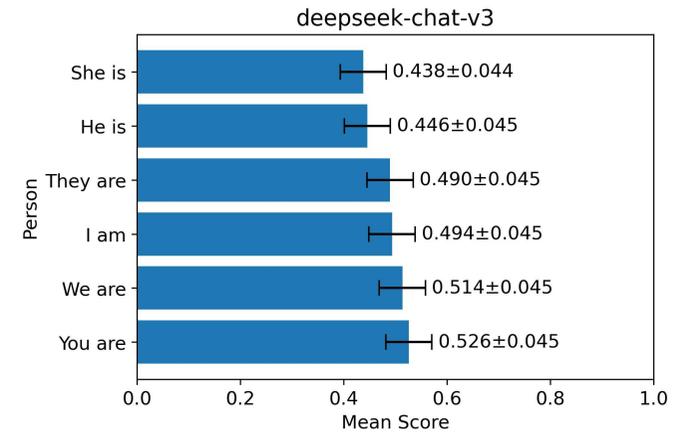
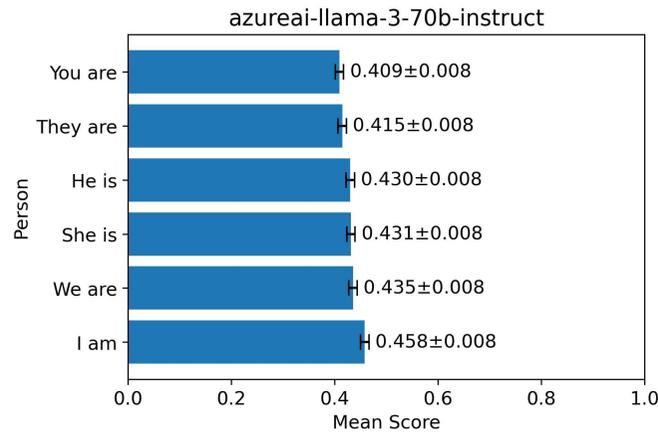
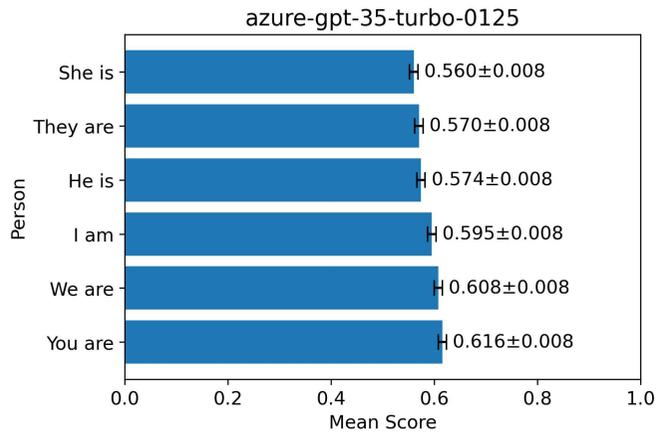
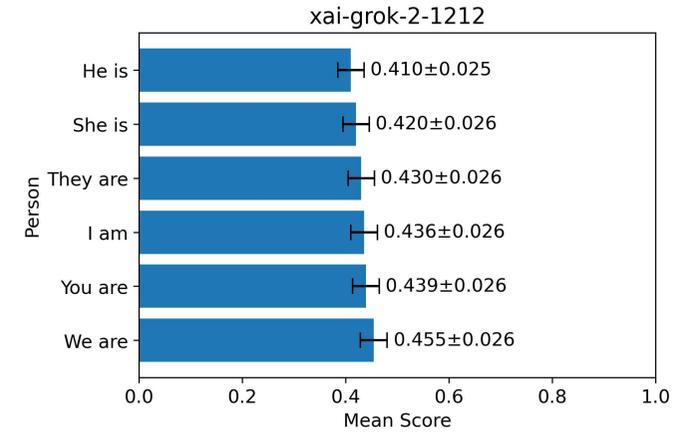
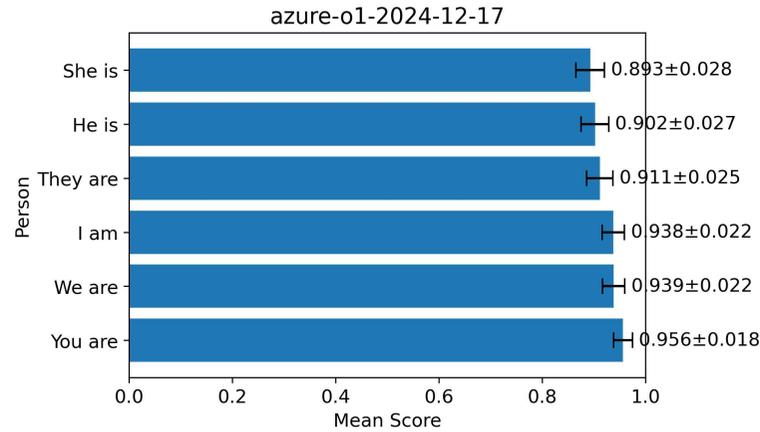
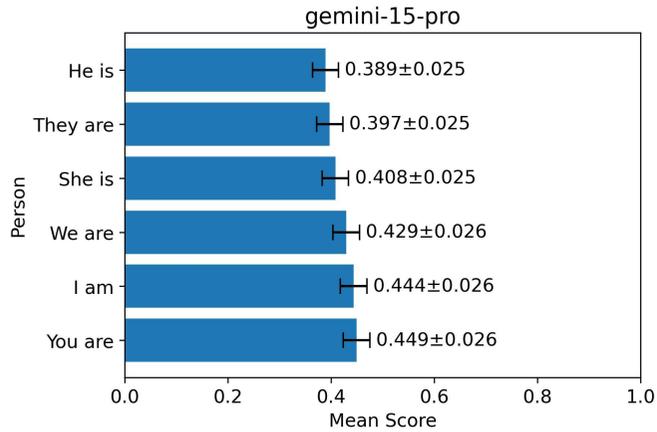
All except llama-3-7b and Claude 3.5 have better CDs than inter-CDs

East always in top 2 except for llama-3-7b

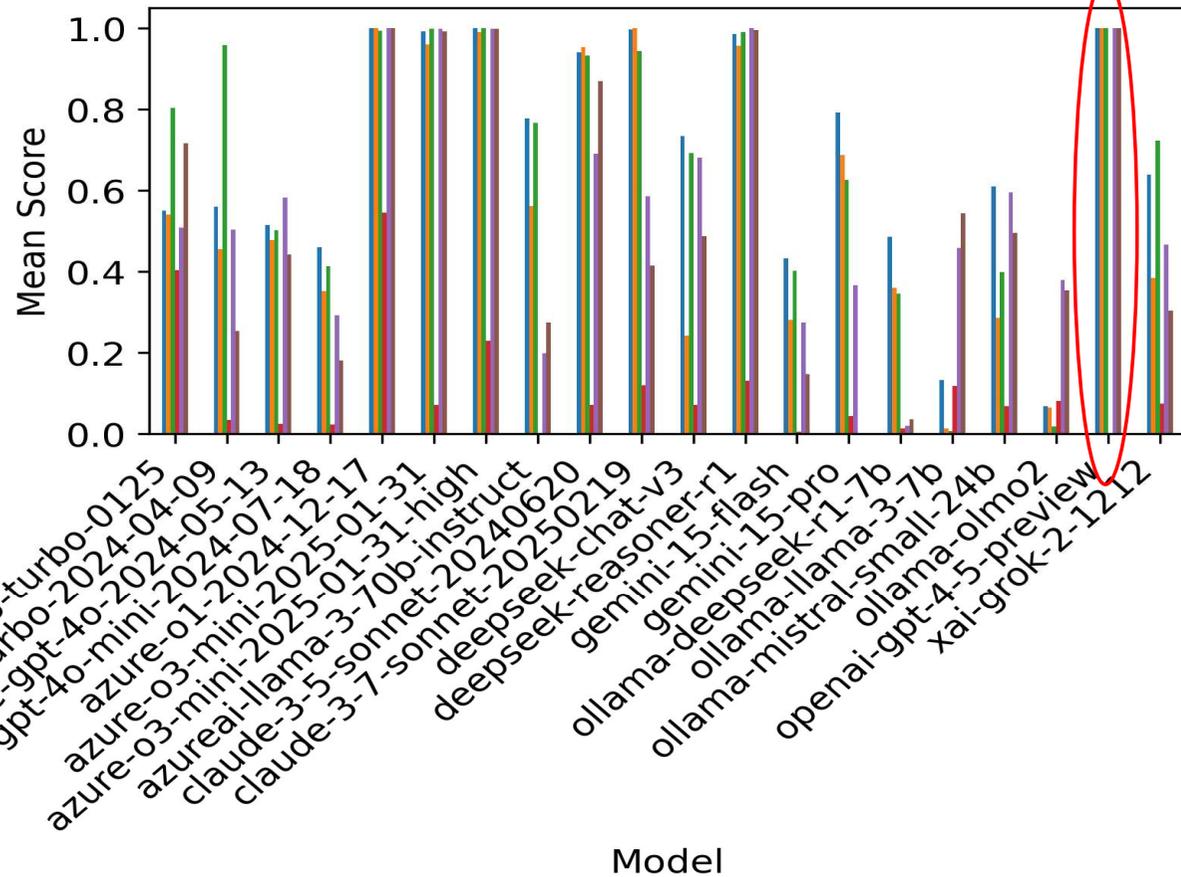
Mean score by locomotion for selected tested models.



Mean score by person form for selected tested models.



Not all models show gender bias



Mean score by template by model.

Models seem to be confounded by the road in question template **T4**.

Template

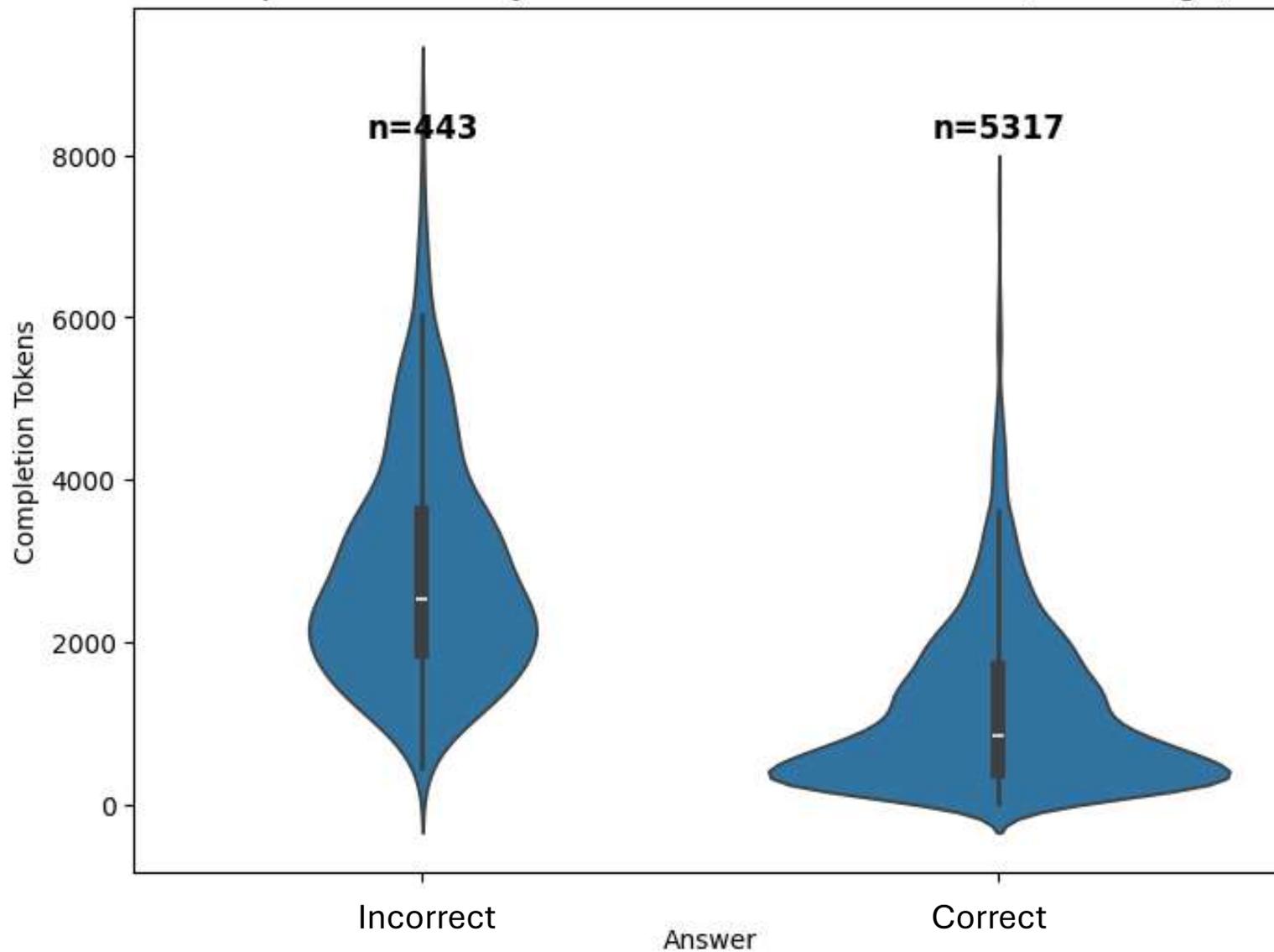
- T1 You are walking [south] along the [east] shore of a lake; in which direction is the lake?
- T2 You are walking [south] along the [east] shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake?
- T3 You are walking [south] along the middle 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?

OpenAI GPT4.5 preview (launched 27 Feb 2025)

Not claimed to be a reasoning model?

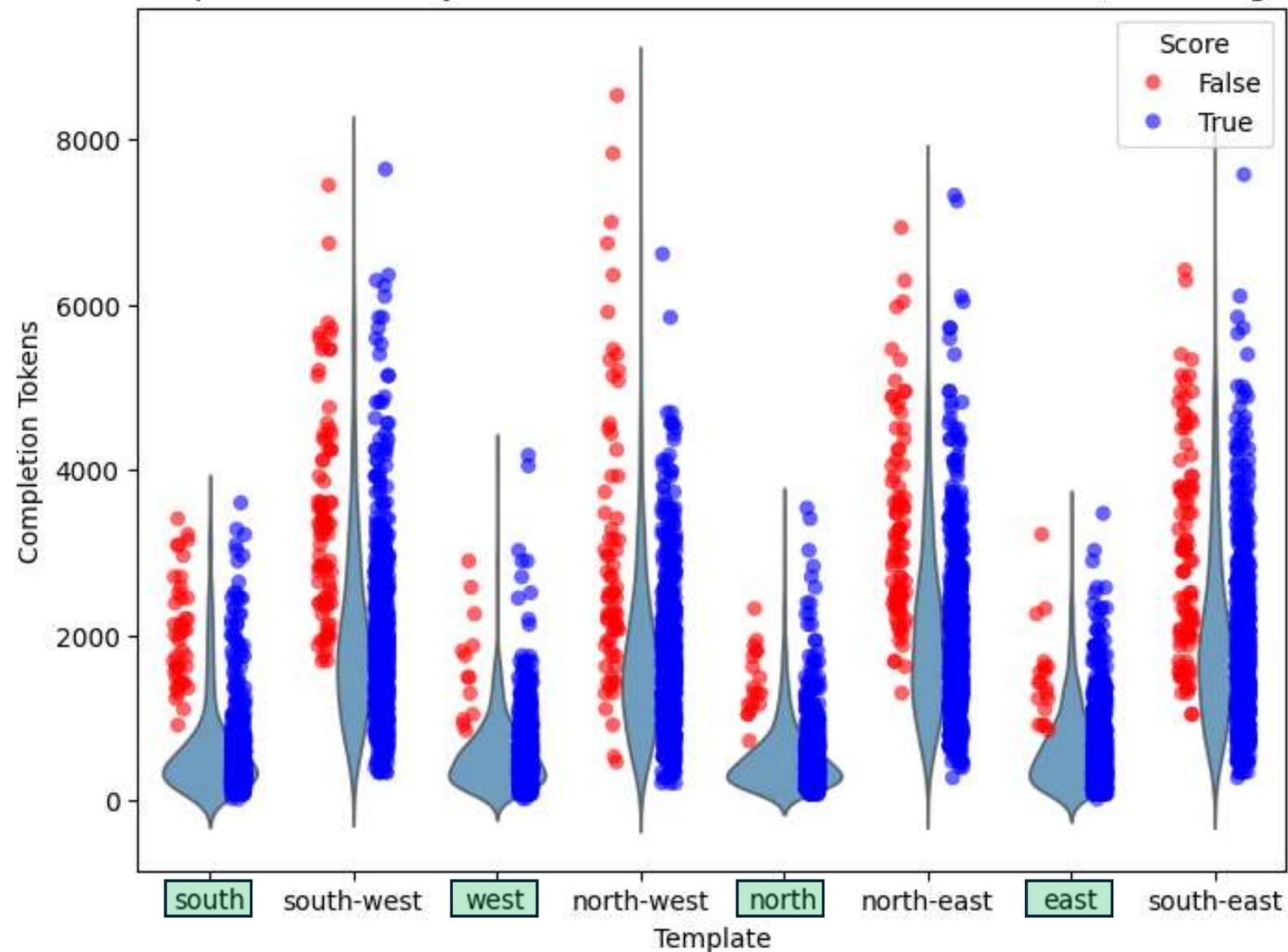
- Suspiciously, it gets all COSIT large questions correct **except for Template 4 questions, where it gets those all incorrect!**
- Hence accuracy = 0.83 (5/6)
- 5th place in the ranking
- Game over for this benchmark?
- Has GPT4.5 been trained on our benchmark???

Completion Tokens by Score for azure-o1-2024-12-17 (COSIT large)



Correct answers generally use fewer completion tokens (~ less reasoning effort).

Completion tokens by correct answer for azure-o1-2024-12-17 (COSIT large)



More “reasoning effort” required for intercardinal directions than cardinal directions.

Reasoning LLMs may never terminate!

Using Ollama deepseek-r1:7b (a distilled version), one question from COSIT-SMALL never terminates (temperature 0):

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

- We have tried it several times and even let it run for three hours, but it just keeps “thinking”.
 - The same sentence recurs after a while and a new cycle starts...
- However, if we use the default temperature it soon returns, with the correct answer.

Okay, so I'm trying to figure out this geography question about directions. Let me read it again: "On a hike, a duck pond is to your north and the nearest town is south. What direction is the pond from the town?"

Hmm, okay.

Alright, let's break this down. The is on a hike, so they're somewhere in the wilderness or maybe at a trailhead. They have two landmarks: a duck pond and a town. The pond is to their north, which means if you imagine looking towards the sky from where they are standing, the pond is above them. Then, the nearest town is south of them.

Wait, so if the town is south, that's directly opposite the direction of the pond. So the pond is north of the person, and the town is south of the same person. Now, the question is asking what direction the pond is from the town. That means I need to figure out where the town is relative to the pond.

Let me visualize this. Imagine a map with the person at point A. The pond is at point B, which is north of A. The town is at point C, south of A. So if you're standing at A looking towards the town (C), it's directly south. But from the perspective of the town (C), where is the pond? Since the pond is north of A and A is somewhere between B and C, but closer to one or the other?

Wait, no, actually, directionally, if you're at point A, and the pond is north, then the town being south means that from the town's perspective, the person would be north. But wait, I'm getting confused.

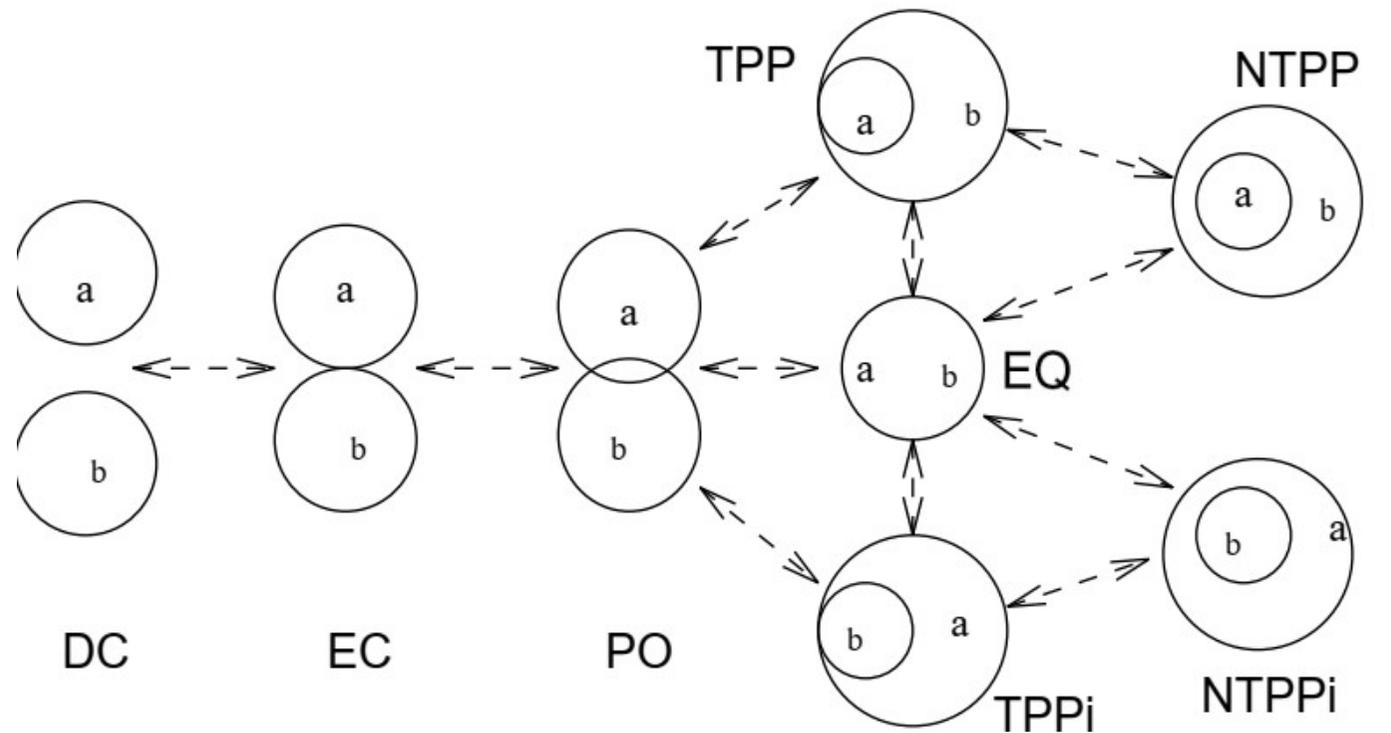
Let me think again. If the pond is north of where you are (the hiker), and the town is south of where you are, then relative to each other, the town is south of the pond? Or is it the other way around?

Wait no, ...

Can LLMs reason about the Region Connection Calculus, RCC-8?

(Cohn and Blackwell, arXiv:2411.19589)

Eight jointly exhaustive and pairwise disjoint relations:



RCC-8 Conceptual neighbourhood

Compositional reasoning

Given $R1(x,y)$ and $R2(y,z)$ what relations are possible between x and z ?

	DC	EC	PO	TPP	NTPP	TPPi	NTPPi	=
DC	*	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC	DC	DC
EC	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPP, TPPi, =	DC, EC, PO, TPP, NTPP	EC, PO, TPP, NTPP	PO, TPP, NTPP	DC, EC	DC	EC
PO	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPPi, NTPPi	*	PO, TPP, NTPP	PO, TPP, NTPP	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPPi, NTPPi	PO
TPP	DC	DC, EC	DC, EC, PO, TPP, NTPP	TPP, NTPP	NTPP	DC, EC, PO, TPP, TPPi, =	DC, EC, PO, TPPi, NTPPi	TPP
NTPP	DC	DC	DC, EC, PO, TPP, NTPP	NTPP	NTPP	DC, EC, PO, TPP, NTPP	*	NTPP
TPPi	DC, EC, PO, TPPi, NTPPi	EC, PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPP, TPPi, =	PO, TPP, NTPP	TPPi, NTPPi	NTPPi	TPPi
NTPPi	DC, EC, PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPP, NTPP, TPPi, NTPPi, =	NTPPi	NTPPi	NTPPi
=	DC	EC	PO	TPP	NTPP	TPPi	NTPPi	=

The prompt

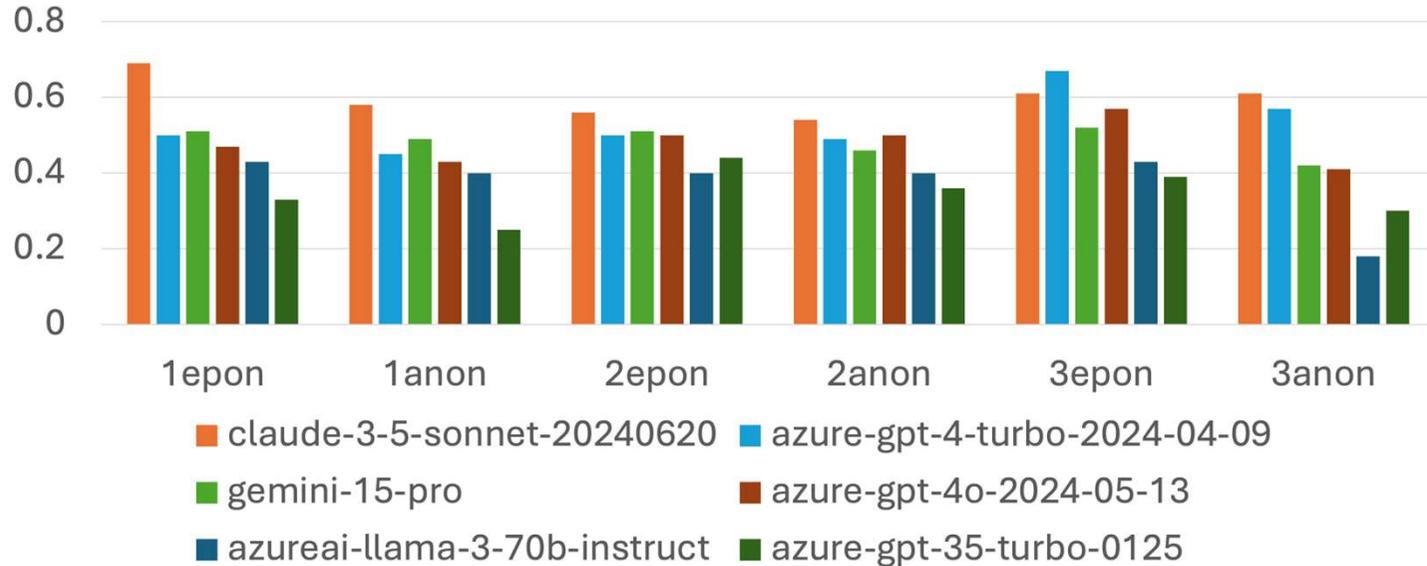
System: Consider the following set of eight pairwise disjoint and mutually exhaustive binary spatial relations. These relations form part of the well known RCC-8 qualitative spatial reasoning calculus. $DC(x,y)$ means that x and y are disconnected and share no spatial parts. $EC(x,y)$ means that x and 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. If all eight relations are possible, just say ALL. 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 ?

Results across three experiments

Model	1epon	1anon	2epon	2anon	3epon	3anon	Overall
claude-3-5-sonnet-20240620	0.69 ± 0.008	0.58 ± 0.012	0.56 ± 0.011	0.54 ± 0.008	0.61 ± 0.023	0.61 ± 0.026	0.60 ± 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	0.67 ± 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

Exp 1 – Composition table
 Exp 2 – Single Preferred composition
 Exp 3 – Spatial Continuity



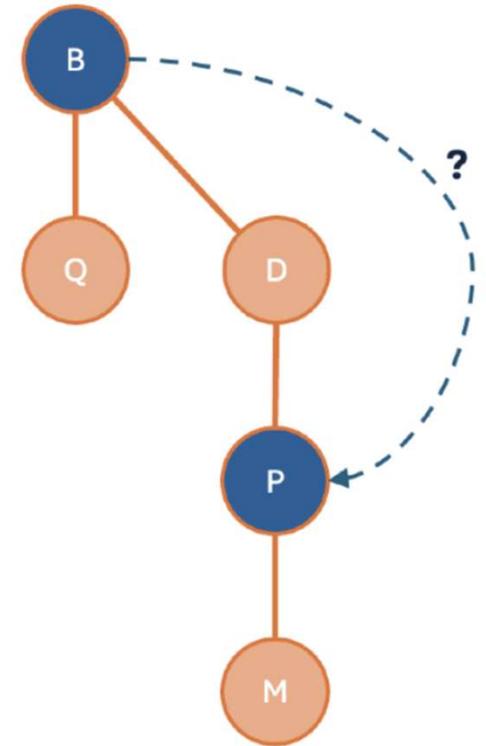
RCC-8 Conclusions

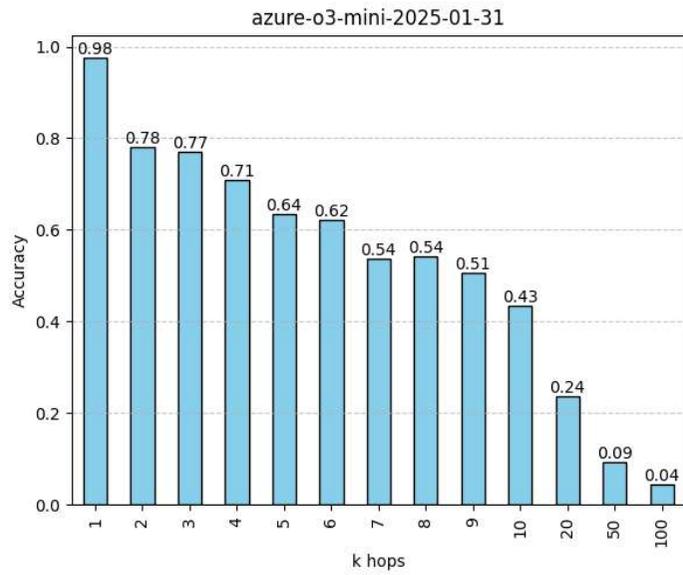
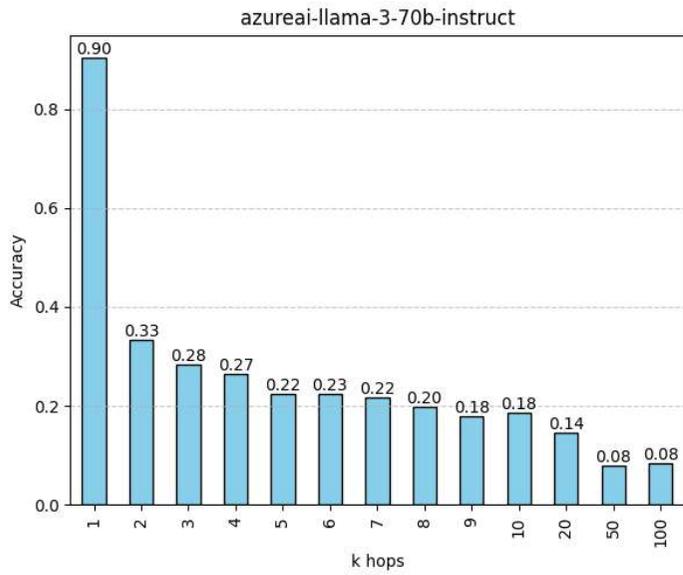
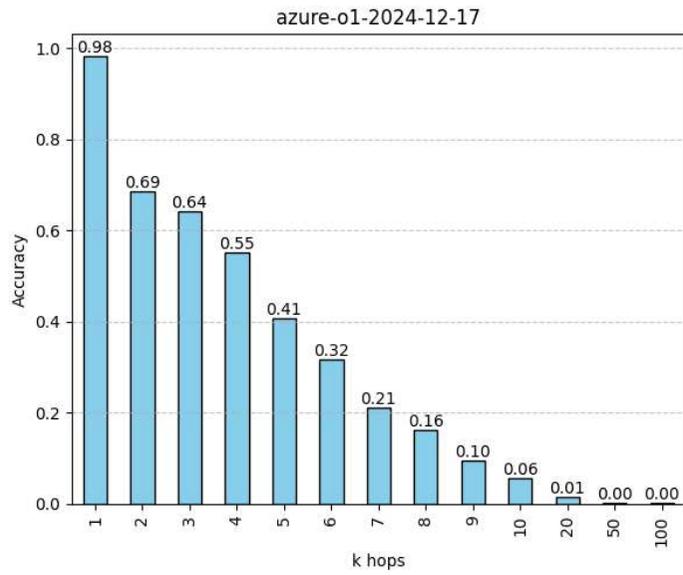
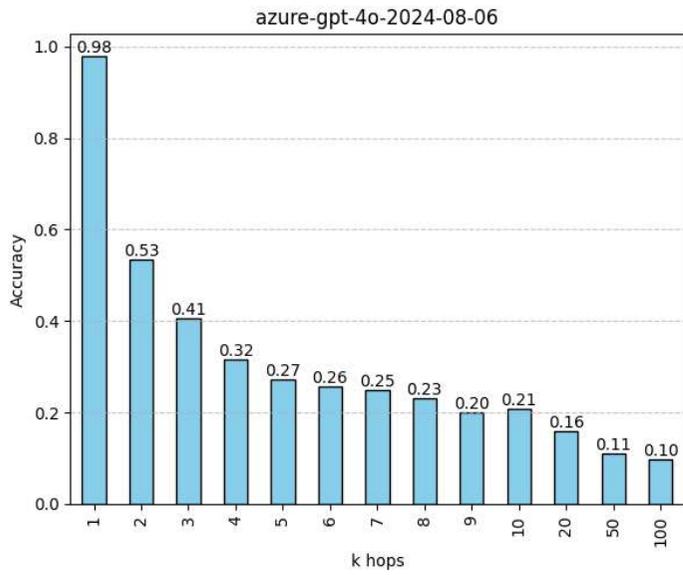
- None of the LLMs tested could reliably perform compositional reasoning or reason about spatial continuity with RCC-8.
- The LLM “preferences” did not align with human preferences.
- Reasoning with anonymised relation names was notably worse than with well-known relation names. (Stochastic Parrots [1] versus generalised reasoners?)

[1] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the dangers of stochastic parrots: Can language models be too big?  .

k hop reasoning

- Rise of chain of thought prompting and then explicit Large *Reasoning* Models (e.g. o1, DeepSeek-R1 etc.)
- Inferences seem to consist of a number of reasoning steps or “hops”
- Revised StepGame (Shi et al., 2022) – Questions can be generated from a spatial graph, e.g. above, below, left, right, upper-left, upper-right, lower-left, lower right





Accuracy by k, by model for 0 shot clean experiment.

Azure o3-mini refusal example

- 5-shot prompt for

Story:

C is on top of O.

AJ is sitting in the right direction of AF.

O is positioned up and to the right of Q.

C is below K with a small gap between them.

Q is over there with AF below.

What is the relation of the agent K to the agent AJ?

- **Consistently trips the Microsoft “Sexual” content filter!**

Chessboard problems trip guard rails

System: You are a helpful assistant. Consider a standard chessboard using algebraic notation ...

User: If the black queen is on square d4 and the white king is on e4, in which direction is the queen from the king?

This trips the Microsoft Azure OpenAI Hate and Fairness Filter (July 2024)

“Bake off” Benchmarks

- Design a set of q questions and answers for some domain (math's, statistics, geometry, ..)
 - Modulate aspects of the questions (gender, tense, reasoning steps, complexity..)
 - E.g. {He, she} is {walking, running} {north, south} along the east shore of a {lake, island}, in which direction is the water?
- Test against a set of state-of-the-art LLMs, scoring each question 1 for a correct answer and 0 for incorrect.
- Analyse results
 - Create a leaderboard comparing LLM performance.
 - Draw conclusions from the modulations.
- But are the results reproducible?
- What combination of temperature, top_p and seed should I use?

How stochastic are LLMs?

Number of unique responses by model to

- i) Tell me a rhyming poem about a tortoise and a friendly dinosaur.
- ii) What is two multiplied by two?

Temperature and nucleus (top_p) sampling are compared. Orange bars show results where the seed was fixed; blue bars, where no seed was specified.

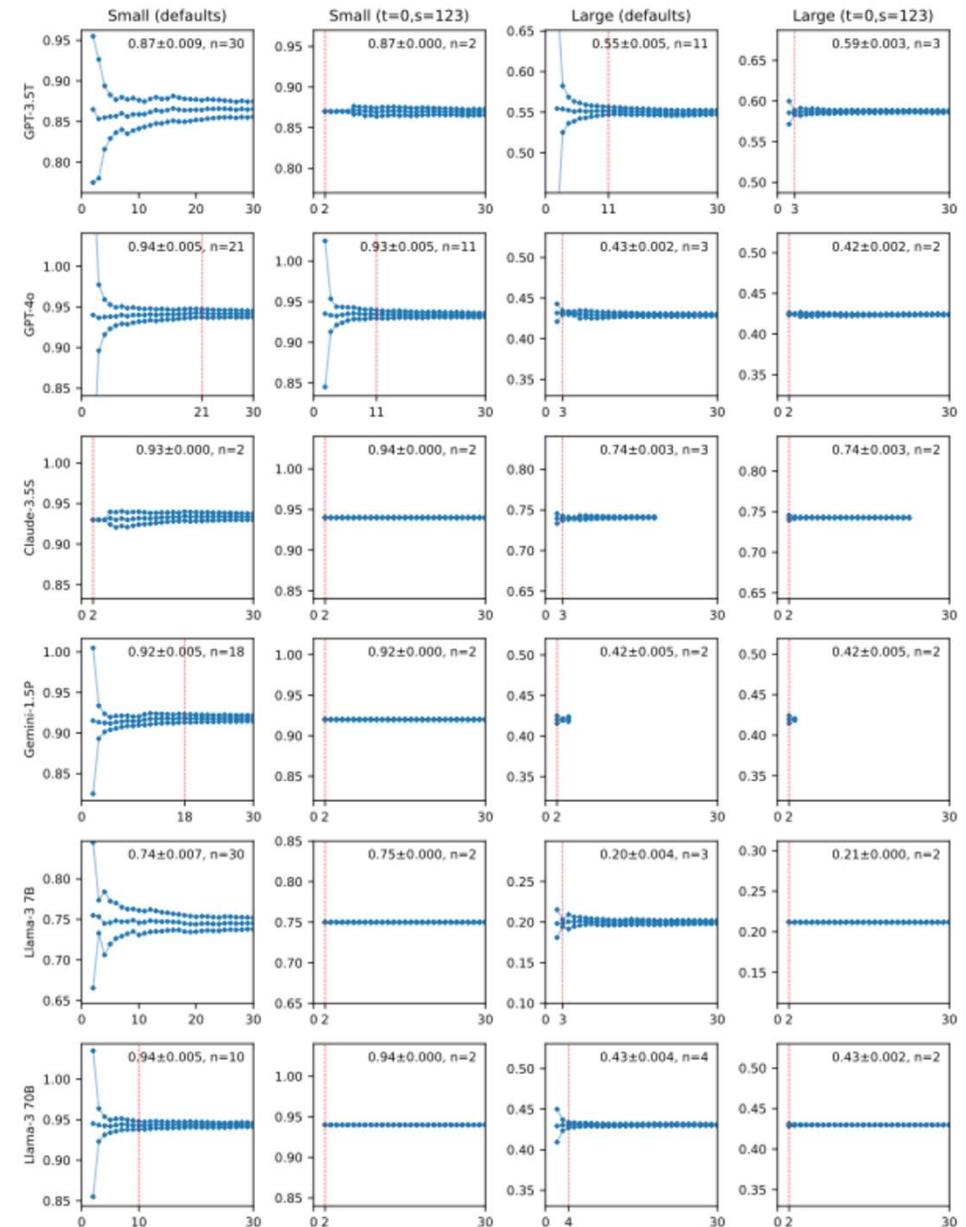


How many experimental repeats?

Now I can run a small number of repeats, see if the prediction interval is below a threshold e.g. ± 0.01 and decide whether to repeat.

In practice, the larger benchmarks converge within three repeats if I fix the temperature and seed.

Blackwell, Barry, Cohn, Arxiv 2410.03492, 2024



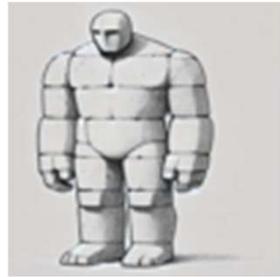
Lessons learned

- Llama-3 7B always gave deterministic results when setting the seed to a fixed value.
 - Only model that was run locally
 - All others run on distributed architectures in the cloud
- Fixing temperature and seed can increase accuracy
 - E.g. GPT-3.5T applied to Large increased its score from 0.55 to 0.59
- With temperature = 0.0 & fixed seed, rarely need > 3 repeats to achieve a prediction interval width of ≤ 0.01 .
- Not all models allow setting seed or temperature
- Fair model comparison remains challenging!

Comparing OpenAI GPT 3.5 Turbo and Microsoft Azure OpenAI GPT 3.5 Turbo.

- Same model, different APIs
- Gave statistically significantly different results for the Small benchmark (n=90, two-sample t-test, $t = 2.51$, $p = 0.013$).

Golem – simple, unified command line interface to LLM APIs



- OpenAI, Azure OpenAI, Azure AI, Google, Anthropic and Ollama.
- Full logging and traceability of request/response pairs.
- Fault tolerance.
- Plays well with jq and JSONL.

```
% golem --provider azure --model gpt-35-turbo --temperature "0.0,0.5,1.0" --  
repeats "1,2,3" "Why is the sky blue?"
```

```
% golem --provider google -f prompts.jsonl > answers.jsonl
```

Blackwell, R.E. and Cohn, A.G., 2024. Golem - a command line interface for benchmarking Large Language Models. Zenodo. <https://doi.org/10.5281/zenodo.14035711>. Available at <https://github.com/RobBlackwell/golem>

Conclusions

- Impressive that AI models demonstrate reasoning at all. Often much better than guessing (especially LRMs).
- Models are improving over time for our benchmarks.
- However:
 - Reasoning is not always reliable.
 - Reasoning is not always convincing.
 - Benchmarks don't always capture reasoning flaws [1].
 - Reasoning often appears to be more like a stochastic parrot than generalised.
 - AI reasoning seems to have a different character to human reasoning.

[1] Burnell, R., et al. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641), pp.136-138.

Thanks!

Robert Blackwell

With thanks to my collaborators:
Anthony Cohn (Leeds),
Pranava Madhyastha (City),
Navdeep Kaur (Turing) and
Lachlan McPheat (Turing).

**The
Alan Turing
Institute**



roblackwell.com