QSTRBench: A New Benchmark to Evaluate the Ability of Language Models to Reason with Qualitative Spatial and Temporal Calculi – a Summary

A.G. Cohn[1,2] , **R.E. Blackwell**[1]

[1]School of Computer Science, The University of Leeds, Leeds LS2 9JT, UK
[2]The Alan Turing Institute, London NW1 2DB, UK
{a.g.cohn, r.e.blackwell}@leeds.ac.uk

Abstract

We introduce an extensive qualitative spatial and temporal reasoning (QSTR) benchmark for evaluating large language models (LLMs) and report results for contemporary frontier models. All models tested perform better than guessing but none can consistently answer all questions correctly. Performance varies sharply by calculus, with PA being the most straightforward, and RCC-22 the most difficult. We release the benchmark, and our results under an open licence to facilitate further assessment of qualitative spatio/temporal reasoning in LLMs.

1 Introduction

Qualitative Spatial and Temporal Reasoning (QSTR[1]), see e.g. [Cohn and Renz, 2007], is a well developed field which is concerned with the representation of qualitative spatial and temporal information and reasoning with it. In natural language, spatial and temporal information is usually represented qualitatively (using prepositions such as *on, in, left of, part of, under, touching, before, after, ...)* and many calculi have been developed to represent such information [Dylla and others, 2017].

Large Language Models (LLMs) [Devlin *et al.*, 2019; Brown and others, 2020], such as GPT, Gemini, Grok, Claude and Llama are examples of so called *Foundation Models* [Bommasani and others, 2021] which have been trained on very large textual corpora in order to generate text in response to a prompt. Many claims have been made about 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.

We create a benchmark, *QSTRBench*, which is designed to test three kinds of reasoning tasks commonly associated with QSTR calculi: what is a relation's converse, what is the composition of two relations, and what are the conceptual neighbours of a relation. The benchmark covers the three best known temporal qualitative calculi, and six of the best known qualitative spatial calculi. Having created the benchmark we then test the performance on LLMs ranging from small open weights models to large commercial frontier models. This paper is a summary of a paper currently under review for a journal (arXiv:2605.18380).

[1]We may use QSTR as shorthand for both Qualitative Spatial and Temporal Reasoning and Qualitative Spatial and Temporal Representation(s); context should make clear which is intended.

2 Related Work

A QSTR calculus is a formal system comprising a finite set of base relation symbols forming a jointly exhaustive and pairwise disjoint (JEPD) set. For example, the Region Connection Calculus, RCC-8 [Cohn and others, 1997] consists of eight relations: DC (Disconnected), EC (Externally Connected), PO (Partially Overlapping), TPP (Tangential Proper Part), NTPP (Nontangential Proper Part) and EQ (Equals); TPPi and NTPPi are the converses of TPP and NTPP respectively since they are asymmetric. A *composition table* (CT) records the results of relational composition R3(x, z) = R1$(x, y) \circ$ R2(y, z) for all base relation pairs. The *conceptual neighbourhood* (CN) [Freksa, 1992] captures which relations can immediately follow one another under continuous change. The *converse* of R(x, y) is the relation obtained by reversing the arguments.

The ability of LLMs to reason about RCC-8 was first explored by Cohn [2023] using ChatGPT-4, and extended to six models in [Cohn and Blackwell, 2024a]. Gardelakos and others [2025] tested RCC-8 and CDC, finding that larger reasoning models and additional context improve accuracy but that models show inductive bias. Bellodi and others [2025] found LLMs unreliable for temporal interval reasoning on unseen problems. Fatemi and others [2024] showed temporal reasoning is sensitive to graph structure and fact ordering. Yamada and others [2024] concluded that spatial knowledge is only partially learned from text. DecompSR [McPheat and others, 2025] tests compositional direction reasoning, finding that performance degrades with increasing reasoning steps; neurosymbolic approaches help but translation errors accumulate. The SpartQA dataset [Mirzaee and others, 2021] and bAbI [Weston and others, 2015] assess spatial reasoning but not CT or CN tasks. LLM reasoning weaknesses are well documented [Cai and others, 2023].

3 Experimental Design

There are many QSTR calculi [Dylla and others, 2017] and so we choose well-known examples with binary relations, where the citation count is high and the CT, CN, and converse relations are readily available. We prefer calculi available in open source tooling such as GQR [Gantner and others, 2008] or SparQ [Wolter, 2009] to avoid transcription errors.

We evaluate nine qualitative calculi: Point Algebra (PA; [Vilain and Kautz, 1986]; 3 relations), Allen's Interval Algebra (IA; [Allen, 1983]; 13), Interval and Duration Algebra (INDU; [Pujari and others, 1999]; 25), RCC-5 ([Bennett, 1994; Jonsson and Drakengren, 1997]; 5), RCC-8 ([Randell and others, 1992; Cohn and others, 1997]; 8), RCC-22 ([Cui *et al.*, 1993]; 22), Egenhofer's 9-Intersection Model (9IM; [Egenhofer, 1994]; 8), the Cardinal Direction Calculus (CDC; [Frank, 1996]; 9), and Revised STAR ([Renz *et al.*, 2004]; 9). PA, IA, and INDU are representative temporal calculi, RCC-8 is the best-known qualitative spatial calculus, 9IM provides an alternative formulation of RCC-8 relations, and CDC and STAR are included because of our interest in cardinal directions [Cohn and Blackwell, 2024b; Cohn and Blackwell, 2025].

For each calculus, we test all combinations of relations R1 and R2 from the CT, using a question of the form *If R1(x,y) and R2(y,z) then what are the possible relations between x and z?*. We test the converse of each relation R using a question of the form: *If R(x,y) then what is the relation between y and x?*. We test the CN of a relation R using a question of the form: *What are the conceptual neighbours of R(x,y)?* (except for 9IM[2] when we use the following wording: *What are the closest topological relationships to R(x,y)?*).

We have a total of 102 base relations across all the tested calculi. So there are 102 converse questions, 102 conceptual neighbourhood questions and 1602 composition table questions[3], a total of 1806 questions in our *QSTRBench* data set.

We present each question to the LLMs in a separate chat session. Where an LLM supports setting temperature explicitly, we set temperature to 0.0[4] with a fixed random seed, otherwise we accept the model defaults.

Burnell and others [2023] call for the routine release of instance-by-instance evaluation results and more granular reporting practices to improve transparency and robustness in AI evaluations. Accordingly, we publish all our experimental data including all API request / response pairs for all experiments, providing full traceability of all model runs and the opportunity for future re-analysis[5]. The data include the timing of experiments, prompts, metadata, API endpoints, token counts and reasoning traces where available. We provide the scripts used to generate questions and number questions and answers consistently, to facilitate cross model comparison.

Models are stochastic in nature and so experimental repeats are essential for robust statistical comparison, however we must balance costs. We therefore report experimental repeats for the best performing model (n=3) and in other cases, where affordable.

4 Results

None of the 36 model configurations tested[6] gets all the benchmark questions correct (Fig. 1). *Gemini 3.1 Pro* with high reasoning (1725:1734:1735/1806)[7] is the best performing model. All of the top ten models are frontier models provided by U.S. technology companies. The top four models all use high reasoning effort. The best performing Chinese open weights model is *DeepSeek R1*. The two worst performing models (*Gemma 3:1B* and *Llama 3:8B*) are small edge models, running on local hardware using Ollama[8]. There has been extraordinary improvement in model accuracy over time, e.g. OpenAI models improved from 0.06 *GPT-3.5 Turbo* to 0.96 (*GPT-5.5* with high reasoning effort) in just over three years.

Overall, LLMs were most accurate with the PA calculus, and least accurate with RCC-22. For all calculi except STAR, the LLMs were more accurate with CT than CN (but the CN for STAR is perhaps obvious because of the numeric ordering of relations).

Models typically perform best on converse questions, then CT questions, with CN questions being the most challenging. All models tested achieve higher accuracy than the guess rate for the combined Converse, CT and CN benchmark.

For RCC, overall accuracy declines with the number of base relations (RCC-5 is more accurate than RCC-8, which is more accurate than RCC-22) but in general, accuracy is not proportional to the number of base relations. For example, INDU has higher overall accuracy than RCC-22, despite it having three more base relations and having the highest number of base relations of all calculi tested. Also, STAR has a higher accuracy than 9IM (9 vs 8 relations) and CDC and IA have a higher accuracy than RCC-8 and 9IM (9 and 13 relations respectively compared to 8 and 8).

Temporal calculi (PA, IA, INDU). PA is the easiest calculus: 19/36 models, including *Gemini 3.1 Pro*, answer all 15 questions correctly. IA and INDU are harder; with only *Gemini 3.1 Pro* getting these completely correct. Converse is better handled for PA (33/36 perfect) than IA (25/36) and INDU (18/36).

Region Connection calculi (RCC-5, RCC-8, RCC-22). Accuracy declines with the number of relations across the RCC family. Only three models achieve perfect RCC-5 accuracy; no model is perfect on RCC-8 or RCC-22. Converse accuracy falls sharply from RCC-5 (34/36) through RCC-8

[2]The 9IM description does not talk about conceptual neighbourhoods but closest topological distances, measured in terms of the similarity of the matrices defining two relations; closest topological distance is related to but not identical to the notion of two relations being conceptual neighbours.

[3]We test only compositions within a calculus, not across calculi, so the total number is not 102^2.

[4]Previous work [Blackwell and others, 2025] found better and more predictable performance at zero temperature.

[5]See github.com/RobBlackwell/QSTRBench, accessed July 2026

[6]For *GPT-5.2* we tested with both default and high reasoning effort.

[7]Where we had the resources to do three experimental repeats, we use the notation (a:b:c/d) to mean a/d, b/d and c/d.

[8]https://ollama.com, accessed July 2026

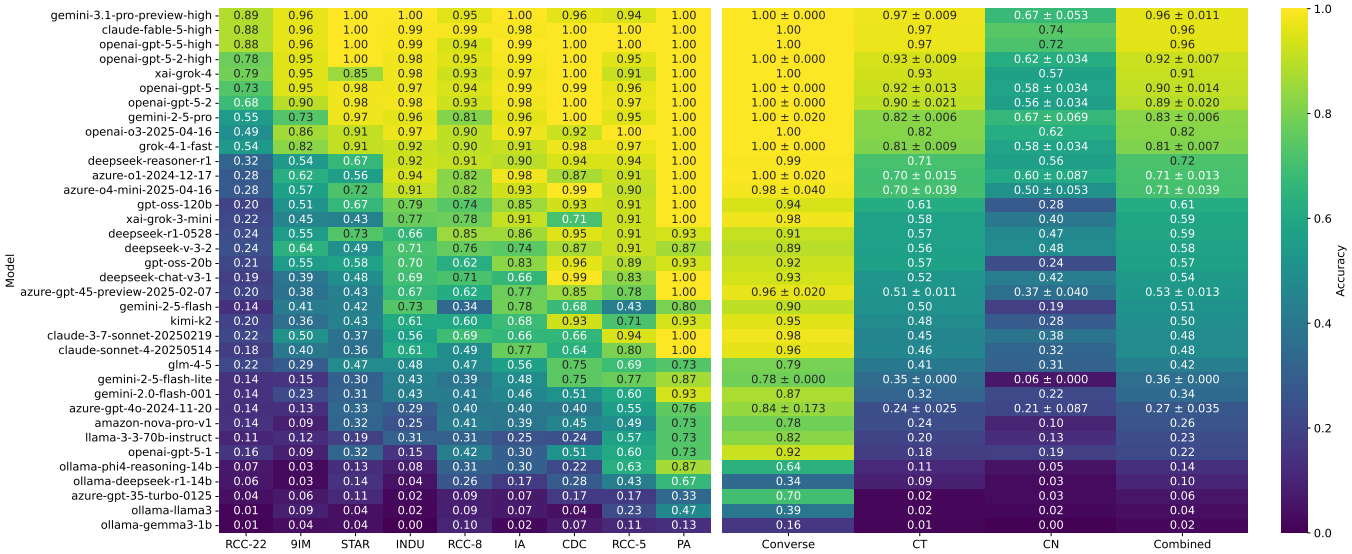

Figure 1: Accuracy by model, by calculus, and by question type. Values are mean ± prediction interval (n=3) where repeats are available.

(28/36) to RCC-22 (12/36). Some models hallucinate invalid RCC-8 relation names (e.g. *TPPPi*), suggesting tokenisation issues with multi-character symbols. RCC-22 is the hardest calculus; even *GPT-5.2* reverts to RCC-8 names on 11 occasions. Collapsing *GPT-5.2*'s RCC-22 answers to RCC-8 granularity improves accuracy, but the analogous narrowing from RCC-8 to RCC-5 does not.

Nine intersection model (9IM). Despite sharing the same underlying topological structure as RCC-8, models show substantially lower converse accuracy with 9IM (only 12/36 perfect models compared to 28/36), likely because the matrix representation obscures converseness. None of the models gets the RCC-8 CN or 9IM CN completely correct, even though the 9IM CN is in principle computable purely syntactically from the matrix entries, rather than involving actual spatial reasoning.

Direction calculi (CDC, STAR). Only six models answer all 99 CDC questions correctly; most (33/36) handle CDC converse correctly. Inter-cardinal directions are harder than cardinal directions, consistent with our earlier work [Cohn and Blackwell, 2025]. For STAR, only four models achieve perfect accuracy – all have the high reasoning effort parameter set. *Kimi K2* failed to produce any valid STAR answer on 7/99 questions, attempting to apply trigonometry to the numeric relation labels.

5 Conclusions

All models perform better than guessing, but none correctly answers all questions. LLM reasoning is non-deterministic – the same question can yield different answers across repeats even at temperature 0. Performance varies sharply by calculus: PA is easiest, RCC-22 hardest. Accuracy improves with model size; small edge models perform very poorly. There has been an extraordinary improvement over time but even the best contemporary frontier models are still inconsistent. RCC-22 is harder than INDU despite having fewer base relations, consistent with its limited documentation, and also arguably its more complex definitions. Although STAR and CDC are similar domains, and RCC-8 and 9IM share structure, results differ across these pairs, showing that models cannot generalise across representations. Models also misunderstand identity (e.g. EQ) and parthood relationships, which ought to be among the easiest to reason about. Although the QSTR literature likely appears in LLM training data, the key inference rules are encoded in tables and diagrams that may be difficult for training procedures to process effectively.

Acknowledgements

This work was supported by The Alan Turing Institute, ESRC grant ES/W003473/1, EPSRC grant EP/Z003512/1, and Tongji University. We thank Microsoft Research (Azure credits) and Google DeepMind (Gemini credits) for compute support, and David Randell for discussions on the RCC-22 CN.

References

[Allen, 1983] J F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[Bellodi and others, 2025] P Bellodi et al. Assessing the (in) ability of LLMs to reason in interval temporal logic. In *32nd International Symposium on Temporal Representation and Reasoning (TIME 2025)*, pages 4–1, 2025.

[Bennett, 1994] B Bennett. Spatial reasoning with propositional logics. In *Principles of Knowledge Representation and Reasoning*, pages 51–62, 1994.

[Blackwell and others, 2025] R E Blackwell et al. Towards reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores. *arXiv preprint arXiv:2410.03492*, 2025.

[Bommasani and others, 2021] R Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brown and others, 2020] T Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[Burnell and others, 2023] R Burnell et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

[Cai and others, 2023] Z Cai et al. Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*, 2023.

[Cohn and Blackwell, 2024a] A G Cohn and R E Blackwell. Can large language models reason about the Region Connection Calculus? *arXiv preprint arXiv:2411.19589*, 2024.

[Cohn and Blackwell, 2024b] A G Cohn and R E Blackwell. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 28:1–28:9, 2024.

[Cohn and Blackwell, 2025] A G Cohn and R E Blackwell. Evaluating the ability of large language models to reason about cardinal directions, revisited. *arXiv preprint arXiv:2507.12059*, 2025. Accepted at the 38th International Workshop on Qualitative Reasoning (QR 2025), co-located with IJCAI.

[Cohn and others, 1997] A G Cohn et al. Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1(3):275–316, 1997.

[Cohn and Renz, 2007] A G Cohn and J Renz. Qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation, 1*, pages 551–596. Elsevier, 2007.

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

[Cui *et al.*, 1993] Z Cui, A G Cohn, and D A Randell. Qualitative and topological relationships in spatial databases. In *Advances in Spatial Databases*, pages 296–315. Springer, 1993.

[Devlin *et al.*, 2019] J Devlin, M Chang, K Lee, and K Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, 2019.

[Dylla and others, 2017] F Dylla et al. A survey of qualitative spatial and temporal calculi: Algebraic and computational properties. *ACM Comput. Surv.*, 50(1), 2017.

[Egenhofer, 1994] M J Egenhofer. Deriving the composition of binary topological relations. *Journal of Visual Languages & Computing*, 5(2):133–149, 1994.

[Fatemi and others, 2024] B Fatemi et al. Test of time: A benchmark for evaluating LLMs on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024.

[Frank, 1996] A U Frank. Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3):269–290, 1996.

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

[Gantner and others, 2008] Z Gantner et al. GQR – a fast reasoner for binary qualitative constraint calculi. In *AAAI Workshop on Spatial and Temporal Reasoning*, page 6, 2008.

[Gardelakos and others, 2025] EO Gardelakos et al. Can large reasoning models reason about spatial relations? In *Proc. 8th ACM SIGSPATIAL Workshop on AI for Geographic Knowledge Discovery*, pages 81–91, 2025.

[Jonsson and Drakengren, 1997] P Jonsson and T Drakengren. A complete classification of tractability in RCC-5. *Journal of Artificial Intelligence Research*, 6:211–221, 1997.

[McPheat and others, 2025] L McPheat et al. DecompSR: A dataset for decomposed analyses of compositional multihop spatial reasoning. *arXiv preprint arXiv:2511.02627*, 2025.

[Mirzaee and others, 2021] R Mirzaee et al. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proc. NAACL*, pages 4582–4598, 2021.

[Pujari and others, 1999] A K Pujari et al. INDU: An interval & duration network. In *Australasian Joint Conference on Artificial Intelligence*, pages 291–303, 1999.

[Randell and others, 1992] D Randell et al. A spatial logic based on regions and connection. In *3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176, 1992.

[Renz *et al.*, 2004] J Renz, D Mitra, et al. Qualitative direction calculi with arbitrary granularity. In *PRICAI*, volume 3157, pages 65–74, 2004.

[Vilain and Kautz, 1986] M B Vilain and H A Kautz. Constraint propagation algorithms for temporal reasoning. In *AAAI*, volume 86, pages 377–382, 1986.

[Weston and others, 2015] J Weston et al. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[Wolter, 2009] D Wolter. SparQ – A Spatial Reasoning Toolbox. In *AAAI Spring Symposium: Benchmarking of Qualitative Spatial and Temporal Reasoning Systems*, page 53, 2009.

[Yamada and others, 2024] Y Yamada et al. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*, 2024.