On Spatial Reasoning and Perspective Transformation in Language Models

Haotong Zhang[1]
Department of Computer Science, University of Manchester, UK

Ian Pratt-Hartmann
Department of Computer Science, University of Manchester, UK
Instytut Informatyki, Uniwersytet Opolski, Opole, Poland

Abstract

In this paper, we investigate the performance of language models on spatial reasoning with textually presented data, with particular focus on the ability to switch between different perspectives (route vs. survey) – an important component of human spatial reasoning. Our methodology involves the construction of large, automatically labelled corpora, as opposed to crowd-sourced, human-annotated datasets; this approach focuses attention on the language models' command of underlying geometrical principles, rather than their access to background knowledge and commonsense rules-of-thumb. Results reveal that language models can acquire basic spatial reasoning ability after finetuning on targeted tasks, though they do not fully capture the underlying rules. Furthermore, language models show significant differences from humans in the perspective-transformation task, exhibiting distinct patterns of performance across perspective conditions.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Spatial Reasoning, Language Models, Perspective Transformation, Spatial Cognition

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.6

Supplementary Material *Software (Source code and datasets)*: `https://github.com/zhanghaotong1/cosit-2026-spatial-reasoning-and-perspective-transformation-in-LMs`
archived at `swh:1:dir:c9382b808e5a17a051732b9787ceacd5e16c3c2c`

Funding *Haotong Zhang*: Supported by the University of Manchester.
Ian Pratt-Hartmann: Supported by the Polish NCN, grant number 2025/57/B/ST6/04170.

Acknowledgements We would like to acknowledge the use of the Computational Shared Facility at the University of Manchester. We also thank the anonymous reviewers for their helpful feedback.

1 Introduction

By *spatial reasoning*, we understand the faculty of making inferences mediated by the reasoner's knowledge – implicit or explicit – of the space in which it is operating. This definition comprises all types of spatial knowledge from the banal (if you are heading due north, objects to the east will appear on your right) to the recherché (rotating an object through 180° about each of three orthogonal axes will return it to its original orientation). Spatial reasoning underlies the ability to execute numerous everyday (as well as some not-so-everyday) tasks. Such is its centrality in human – and indeed animal – intelligence, that it has generated a great quantity of psychological research. How do humans (or other animals) solve spatial reasoning problems? What representations do they employ to describe spatial

[1] Corresponding author

arrangements? What algorithms do they use to manipulate them [4]? At the same time, the astonishing development of transformer-based language models (LMs) and their more recent evolution into large *reasoning* models [64] have demonstrated at least an apparent ability on the part of these architectures to engage in humanlike reasoning, suggesting some grasp of rudimentary logical principles. The question therefore arises as to how well such models can engage in *spatial* reasoning, and, more specifically, what their characteristics are [6]. Can LMs reason about space? If so, do they reason in ways that resemble the patterns observed in humans? That is the subject of this paper.

When considering research into the reasoning abilities of LMs, we must distinguish two very different paradigms. The first understands inference as a matter of arriving at *probably correct* judgments mediated by the sort of commonsense, background knowledge expected of every normal person [3, 44, 63]. For example, from *The cat is in the garden*, it may be inferred that *The cat is not in the kitchen*, using the background knowledge that gardens are seldom found in kitchens (and that, quantum effects aside, cats cannot be in two places at once). The second paradigm examines entailment *sensu stricto* [48, 46, 56, 17]: a collection of premises Φ is said to entail a conclusion ψ if all logically possible situations making all the premises Φ true also make ψ true. For example, from the two premises *All artists are creative* and *Some beekeepers are artists*, we can infer *Some beekeepers are creative* as a matter of logic. These two paradigms differ sharply in the datasets typically employed. Whereas the former pradigm is largely reliant on human-annotated, crowd-sourced data, the latter is free to exploit automatically generated corpora of essentially unlimited size. For example, in training systems to recognise logically valid inferences couched in natural language, we may automatically construct large collections of premises and (putative) conclusions in fragments of language specified by some grammar, and label them (valid or invalid) using an automated theorem prover. As for reasoning in general, so for spatial reasoning in particular: corpora of reasoning problems – even those couched in natural language – can be constructed and labelled using fully automated techniques [7]. It is this second paradigm that we adopt here.

The problem of determining logical entailments mediated by knowledge of the operative space is, of course, the domain of geometry, broadly construed. Where considerations of the language employed and the algorithms required to process them are to the fore, we are in the more specific realm of *spatial logic*. By a spatial logic, we understand a formal language whose variables range over some prescribed domain of geometrical entities – *e.g.*, points, lines, regions of space – and whose non-logical primitives are interpreted as relations defined on those geometrical entities – *e.g.* incidence, parallelism, inclusion. The relations considered may be purely qualitative (as in topological calculi, or indeed as in traditional presentations of Euclidean geometry) or quantitative (as in real algebraic geometry). For a survey of spatial logic, see [1]. When investigating the performance of LMs on a particular class of spatial reasoning problems, the mathematical and computational aspects of the class of problems in question must be thoroughly understood. Thus, we find ourselves at the confluence of three research programmes – psychological studies of human spatial reasoning, computational investigations of spatial logic, and the use of LMs from artificial intelligence. What spatial inferences can LMs perform effectively? How does this performance correlate with the intrinsic complexity of the problems in question? Do these models exhibit similar behaviour to humans in respect of spatial reasoning?

The present paper tackles these issues from the perspective of a particular problem which has attracted the attention of cognitive psychologists, namely that of shifting between route-based and survey-based views. Route-based spatial descriptions adopt a first-person perspective: *As you leave the Capitol going along the Mall, the first building you pass on your*

right is ... Survey-based descriptions, by contrast, take the view from nowhere: *At the east end of the Mall stands the Capitol and at the west end ...* The (obviously useful) ability to switch between these sorts of descriptions – *perspective transformation* – has been the object of a certain amount of psychological study, notably by B. Tversky [57, 55, 58]. How, we ask, do LMs perfom on such tasks? To help focus our investigations, we proceed step-by-step, considering first a range of reasoning problems where the underlying spatial logic is well understood, and investigate the extent to which LMs can learn to solve these problems. Specifically, we define three spatial reasoning tasks – which we call the one-dimensional (**1D**) task, the two-dimensional (**2D**) task and the oriented matroid realisability (**OMR**) task – and construct datasets of sets of statements, couched in natural language – representing descriptions of spatial arrangements. Broadly speaking, we may regard the **2D**-task as involving reasoning within the *survey view*, while the **OMR**-task involves reasoning within the *route view*. The datasets in question are automatically labelled as either *satisfiable* (*i.e.* geometrically realisable) or *unsatisfiable* (geometrically unrealisable), and used to finetune pre-trained LMs to perform this classification. From an algorithmic perspective, the **1D**- and **2D**-tasks are relatively simple in their underlying logical structure, whereas the **OMR**-task involves more intricate reasoning. In each case, datasets are crafted to eliminate data artefacts. The difficulty of each problem instance is also measured to make sure that our datasets are non-trivial. With a preliminary understanding of the spatial reasoning abilities of LMs in place, we turn our attention to the issue of perspective transformation. Specifically, we define two further spatial reasoning tasks – which we call the **2D2OMR**-task, and the **OMR22D**-task. These tasks are intended – very approximately – to replicate perspective-transformation experiments reported in the psychological literature.

Throughout this paper, we use the term *spatial reasoning* in an operational sense. That is, we evaluate whether a model can correctly solve formally defined spatial inference tasks expressed in natural language. Our use of the term does not presuppose that language models employ reasoning processes identical to those used by humans, nor does successful task performance necessarily imply human-like spatial cognition. Rather, the tasks provide a behavioural framework for comparing the spatial inference capabilities of different systems.

2 Related Work

Spatial reasoning first caught the attention of psychologists in the last century, who carried out experiments to explore how humans comprehend spatial information, how they store such information as mental models and how they reason with mental models. Experiments revealed that humans construct mental models following certain processes and tend to overlook alternative possibilities [4, 43]. Another line of research focuses on perspective taking, especially between route (egocentric) and survey (allocentric) perspectives. Early work by Siegel and White [53] distinguished between route view and survey view in the development of spatial representations. Subsequent work further examined the respective roles of landmark, route, and survey knowledge in navigation [61]. Empirical studies have demonstrated that these two forms of representation differ not only in format but also in the cognitive processes required to construct and manipulate them [57, 55, 58]. Further studies indicate that original orientation also affects reasoning accuracy [49, 50].

It is instructive to consider many of these reasoning problems from a purely logical perspective. By a *spatial logic*, we understand a formal language interpreted over a class of structures featuring geometrical entities and relations [54]. A central line of research in spatial logic concerns qualitative spatial reasoning (QSR), which develops formal calculi for representing and reasoning about spatial relations without relying on precise metric

information [9]. Foundational work in this area includes topological calculi for regions [45], directional calculi for points [27, 28], and east and west calculi for intervals [26]. Another set of approaches is often motivated by observer-dependent spatial descriptions and their relationship to human spatial cognition. Examples include acceptance regions for projective spatial expressions, which have been argued to better reflect human usage of relative directional concepts [36]. Related approaches include the Dipole Calculus [37], which represents spatial entities as oriented line segments and supports reasoning about intrinsic orientation, and the OPRA family of calculi [38, 39], which models relative directions between oriented points at configurable granularity levels. These formalisms provide cognitively motivated representations of relative direction and have been applied to navigation and qualitative spatial reasoning tasks. A salient issue in QSR involves determining the computational complexity of particular spatial logics, especially with regard to the satisfiability-checking problem [21, 13, 11]. Of particular interest here are relatively expressive spatial logics whose satisfiability problems are decidable, ideally with low complexity.

In the last decade, textual spatial reasoning has been introduced to the field of NLP. BAbI was the first dataset to include spatial reasoning [62]. It was relatively simple and has since been largely solved by neural models. Later, three more complex datasets with more relations, entities and longer reasoning paths were introduced: SpartQA [35], SpaRTUN [34] and StepGame [51]. StepGame has been influential as a synthetic benchmark for multi-step textual spatial reasoning, although later work has identified template errors and inconsistencies in its definition of reasoning hops [25, 24]. These datasets nevertheless proved to be difficult for LMs such as BERT [10], ALBERT [22] and XLNet [68]. Premsri and Kordjamshidi claimed that the unsatisfactory performance of LMs on spatial reasoning is partly due to their poor generalisation ability on unseen data samples [41]. Therefore, they proposed training LMs with neuro-symbolic techniques. Although proven to be effective, the approach did not address how well LMs can perform on purely textual spatial reasoning.

With the emergence of large language models (LLMs), their performance on spatial reasoning in few-shot or zero-shot settings (*i.e.*, without finetuning) has attracted much attention [20, 5, 29, 8]. New benchmarks have been constructed and tested on LLMs [66, 47]. Models are reported to exhibit shortcomings on a variety of spatial reasoning benchmarks [2, 25, 65, 67]. Nevertheless, Li, Hogg, and Cohn [25] showed that, on a corrected version of the StepGame benchmark, LLMs can accurately extract spatial relations from natural-language descriptions, and that their performance can be improved through Chain-of-Thought prompting [60] and Tree-of-Thoughts prompting [69]. More recently, neuro-symbolic approaches have also been explored for spatial reasoning benchmarks similar to StepGame [33]. Furthermore, LLMs have been reported to exhibit certain performance patterns and biases similar to those observed in humans [6, 14, 7].

Recent work has examined spatial reasoning and perspective-taking in vision-language models (VLMs) and multimodal large language models (MLLMs), typically focusing on viewpoint transformations between different observers. Empirical analyses report systematic limitations in spatial inference in MLLMs [52]. Visual perspective-taking has been evaluated through tasks requiring models to adopt alternative observer viewpoints, revealing persistent difficulties in shifting away from egocentric representations [16]. More recent studies further document strong egocentric biases in VLMs [59], and suggest that incorporating cognitively inspired mechanisms can improve performance on cross-perspective reasoning tasks [23]. While related, these studies primarily address observer-based perspective shifts. In contrast, our work focuses on transformations between route and survey perspectives in linguistic spatial reasoning, which have been central in research on human spatial cognition and remain comparatively underexplored in studies of LMs.

In summary, our study links human perspective transformation, qualitative spatial logic, and natural language processing through a logical framework for spatial reasoning.

3 Reasoning Tasks

In this section, we define our spatial reasoning tasks. We first begin with a very simple task, which deals with points in one-dimensional space. A statement in this task takes the form $A\ r\ B$, interpreted as *A is on the r of B*. A and B represent points and r represents the spatial relation between A and B: left, right or overlap. Left and right follow our daily convention, while overlap means that two points are at exactly the same position. The task is formatted as a satisfiability-checking problem. From the geometric point of view, a set of statements is *satisfiable* just in case a set of points can be arranged on a line in accordance with the statements. We call this task the **1D**-task.

Next, we define a similar task which is only slightly more complicated than the **1D**-task: the **2D**-task. As indicated by its name, we now consider points in two-dimensional space. Statements are again of the form $A\ r\ B$, interpreted as *A is to the r of B*, except that r represents cardinal and intercardinal directions: north, south, east, west, northeast, southeast, northwest, southwest and overlap. This set of directional relations is closely related to the projection-based notion of cardinal directions introduced by Frank [12], and more broadly to the Cardinal Direction Calculus (CDC) literature. When we refer to "north", we specifically mean due north (*i.e.*, 0° on the compass). And "northeast" means any direction strictly between due north and due east. The same convention applies to other (inter)cardinal directions. We say a set of statements is satisfiable if there is a set of points in the two-dimensional space satisfying all statements at the same time.

The third task is the oriented matroid realisability problem (or **OMR** for simplicity). In this task, locations are described from an egocentric point of view. A problem instance consists of a set of statements having either of the forms $\mathit{left}(A, B, C)$ or $\mathit{right}(A, B, C)$, interpreted as: *standing at A and facing B, C is on the left/right*. Here A, B, C are objects that may regarded as unextended (*i.e.*, single points). We do not differentiate between front and back. The term "left" encompasses both front-left and back-left regions, and "right" is defined in an analogous manner. Note that "left" and "right" are interpreted in the strict sense. Thus, $\mathit{left}(A, B, C)$ or $\mathit{right}(A, B, C)$ is satisfied only when C lies strictly in the corresponding half-plane. Collinear configurations, where A, B, and C lie on the same line, are excluded and satisfy neither relation. Furthermore, the points A, B, C must be distinct; coincident points are not permitted. Despite the statements' simple form, this task is more complicated than the above two tasks, as each statement involves the relative orientation among three points. A set of statements is satisfiable if we can find a set of points on the two-dimensional Euclidean plane making all statements true. For example, $\mathit{left}(A, B, C), \mathit{left}(A, D, B), \mathit{left}(A, C, D), \mathit{left}(B, C, D)$ together is satisfiable (see Figure 1 for a realising arrangement). However, a closer look reveals that changing the last statement to $\mathit{left}(B, D, C)$ leads to unsatisfiability, no matter how we arrange A, B, C and D.

Finally we introduce two perspective transformation tasks, which are the main focus of this paper. Notice that our **2D**-task is similar to the survey view, in the sense that each statement describes the absolute direction between two points. On the other hand, our **OMR**-task can be regarded as the route view. Previous cognitive psychology experiments conducted on humans are formulated as question-answering tasks, where subjects are given a textual description of an area in one perspective and asked to determine whether a question from the other perspective is true or false. However, to maintain consistency with above tasks,

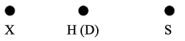
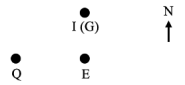
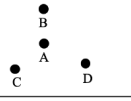
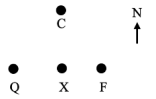
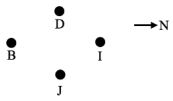
Task	Statements	Natural Language Form	Realising Arrangement
1D	$H\ left\ S$	H is on the left of S.	
	$H\ overlap\ D$	H overlaps with D.	
	$X\ left\ H$	X is on the left of H.	
	$D\ right\ X$	D is on the right of X.	
2D	$I\ overlap\ G$	I overlaps with G.	
	$E\ south\ I$	E is to the south of I.	
	$I\ northeast\ Q$	I is to the northeast of Q.	
	$E\ east\ Q$	E is to the east of Q.	
OMR	$left(A,B,C)$	Standing at A and facing B, C is on the left.	
	$left(A,D,B)$	Standing at A and facing D, B is on the left.	
	$left(A,C,D)$	Standing at A and facing C, D is on the left.	
	$left(B,C,D)$	Standing at B and facing C, D is on the left.	
2D2OMR	$C\ north\ X$	C is to the north of X.	
	$C\ northeast\ Q$	C is to the northeast of Q.	
	$F\ southeast\ C$	F is to the southeast of C.	
	$X\ east\ Q$	X is to the east of Q.	
	$left(C,X,F)$	Standing at C and facing X, F is on the left.	
OMR22D	$right(J,D,I)$	Standing at J and facing D, I is on the right.	
	$right(J,B,D)$	Standing at J and facing B, D is on the right.	
	$right(I,J,D)$	Standing at I and facing J, D is on the right.	
	$D\ northwest\ B \rightarrow I\ northeast\ D$	If D is to the northwest of B, then I is to the northeast of D.	

Figure 1 A satisfiable example in each task. The *Natural Language Form* column shows how to interpret logical statements into natural language and is the actual input to the models. We also give a realising arrangement for each example.

we again formulated two tasks as the satisfiability-checking problem. Specifically, in our **2D2OMR**-task, we append a single statement from the **OMR**-task (*i.e.* $left/right(A,B,C)$) to the end of a set of statements from the **2D**-task (*i.e.*, $A\ r\ B$), switching from the survey view to the route view. This setting is roughly similar to the psychological experiments in the sense that LMs first see a set of statements in the survey view (textual description), then a statement in the route view (question). Checking whether the question statement is consistent with the textual description amounts to checking whether they are jointly satisfiable. Similarly, in our **OMR22D**-task, we append a single statement from the **2D**-task to the end of a set of statements from the **OMR**-task[2]. This task requires the ability to switch from the route view to the survey view. In both tasks, a set of statements is satisfiable if a set of points satisfied all spatial relations described by the statements. While our setup is not a strict replication of previous psychology experiments, it captures the core feature of cross-perspective transformation between survey and route perspectives.

Figure 1 depicts a satisfiable example for each task. We also provide a realising arrangement alongside, which, as can be checked, satisfies every statement.

4 Data Construction

As mentioned in Section 1, datasets in our tasks can be labelled by automated algorithms, due to their purely logical character. An *instance* in the dataset comprises a set of statements. For the **1D**-task, it is straightforward that unsatisfiability can only rise from a *strict directed cycle*, namely a subset of statements of the form $A_1\ r_1\ A_2, A_2\ r_2\ A_3, ..., A_n\ r_n\ A_1$, where

[2] We also need to assign a random (inter)cardinal direction between two points since a pure route perspective does not contain any absolute direction information.

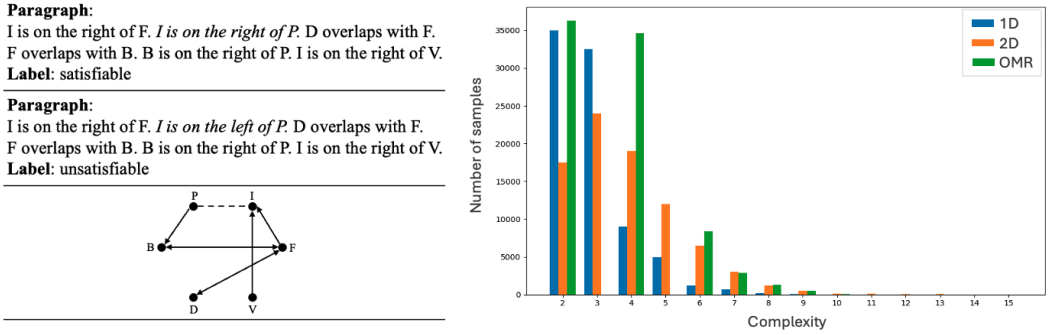


Figure 2 (a) A minimally distinguished pair from the **1D**-task. The second sentence (in *Italics*) is different in the two instances. To give a clearer illustration, we visualise the instances in a "graph-like" style where each point is a node in the graph, instead of plotting an one-dimensional arrangement. Directed edges (with one arrow) denote the left or reversed right relation. Bidirectional edges denote the overlap relation. The dashed edge between P and I corresponds to the different statement. If it goes from I to P (*i.e.*, I *left* P, as in the unsatisfiable instance), then there is a strict directed cycle $F - I - P - B - F$; (b) Complexity distribution in the **1D**-, **2D**- and **OMR**-task. We omit the complexity distribution for the **2D2OMR**- and **OMR22D**-task due to computational infeasibility.

either: (i) each r_i is *left* or *overlap*, with at least one of them being *left*; or (ii) each r_i is *right* or *overlap*, with at least one of them being *right*. (Here, of course, we are implicitly taking A *right* B to be equivalent to B *left* A.) Labelling data in the **1D**-task is equivalent to detecting strict directed cycles in the statements.

We can label data in the **2D**-task in a similar way, except that now we need to check for two independent axes: north-south and west-east. For example, when evaluating the north-south axis, the east and west direction is ignored: A is considered "latitudinally north" of B if it lies to the north, northeast or northwest of B. The same convention applies when evaluating the east–west axis: A is considered "longitudinally east" of B if it lies to the east, northeast or southeast of B. We remark that, here, we imagine points located in a small region, and disallow circumnavigation of the globe: thus, Berlin is north*east* of Paris, not north*west*.) A set of statements is easily seen to be satisfiable if there are no strict directed cycles along either axis.

Next we turn to the **OMR**-task. Checking satisfiability is equivalent to the problem of solving a set of quadratic inequalities in the real domain. Specifically, a statement $left(A, B, C)$ holds if and only if $(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1) > 0$, where $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ are coordinates of A, B, C in the two-dimensional Euclidean plane, respectively. We refer readers to, for example, [40], for a detailed proof. In principle, the feasibility of a quadratic inequality system can be determined exactly using symbolic solvers such as Mathematica[3]. However, exact solving becomes computationally expensive when the number of points grows. In constructing the dataset, we therefore rely on iterative quasi-Newton methods implemented in SciPy library[4] in Python to obtain approximate solutions. These methods provide solutions that satisfy the constraints up to a numerical tolerance (we use 1e-6) rather than guaranteeing exact symbolic correctness, but offer substantially improved efficiency and scalability for large numbers of instances. We randomly check some instances using Mathematica to make sure that SciPy is highly accurate.

[3] `https://www.wolfram.com/mathematica`
[4] `https://scipy.org`

We also use SciPy to label data in our perspective transformation tasks. As mentioned in Section 3, a single statement from one task is appended to the end of a set of statements from the other task (*i.e.*, perspective). We deliberately make sure that statements from the same perspective are satisfiable, to prevent intrinsic unsatisfiability from arising within one perspective. In other words, unsatisfiability can only come from the inconsistency during perspective transformation.

To eliminate potential shortcuts in instances that are easily detected by LMs and reduce task difficulty, we generate satisfiable and unsatisfiable instances in pairs for all tasks. Specifically, we first randomly generate sets of statements until we find an unsatisfiable one. Then we randomly reverse a statement in the set and check if the new set of statements is satisfiable. By "reverse", we mean to replace the spatial relation with its inverse relation[5] (*e.g.*, replace "left" with "right", replace "north" with "south"). If so, we call these two instances a *minimally distinguished pair*, and say that they are *contra-instances* of each other. We generate datasets in this way so that satisfiable and unsatisfiable instances cannot be easily distinguished by any superficial characteristics. In fact, two instances in a minimally distinguished pair only differ by one word. Figure 2 (a) shows an example of such pair in the **1D**-task and embodies the strict directed cycle in the unsatisfiable instance.

For an instance, both the number of statements (*instance size*) and the number of points involved (*domain size*) influence the probability of satisfiability of a random instance. Through some pilot experiments, we empirically show that setting both to be equal makes it easy to generate unsatisfiable instances while allowing satisfiable contra-instances to be found for most of them. The dataset construction process is as follows: (1) Generate a set of statements with determined instance size and domain size. (2) If the generated instance is unsatisfiable and there exists a satisfiable contra-instance, we add both to the dataset. (3) Repeat (1) and (2) until we have enough instances. Notice that the dataset constructed in this way is automatically balanced. We obtain a dataset for each of our tasks and randomly split each into a training set with around 74K instances and a test set with 10K instances. We make sure that each minimally distinguished pair is in the same set to avoid potential data leakage. Domain size varies from 3 to 26 (inclusive), with an equal number of instances for each domain size. Spatial relations are sampled uniformly during dataset generation. As a result, all relation types occur with approximately equal frequency in the datasets.

We additionally checked for duplicate instances and ensured that no identical instance appears in both the training and test sets. However, because points are represented by arbitrary uppercase letters, there may still exist pairs of instances that become identical after a permutation of uppercase letters. For example, (1) *T is on the left of K. K is on the left of B. B is on the left of T.* and (2) *U is on the left of I. I is on the left of Z. Z is on the left of U.* are essentially the same under uppercase letter permutation. Such instances are not treated as duplicates by our generation procedure. Although this does not affect the correctness of the dataset, it may reduce the effective diversity of the data and could potentially be exploited by models that learn to ignore specific point names. We leave a more systematic analysis of this issue for future work.

Finally, we provide a complexity measurement for an instance, to help use assess the difficulty of our datasets. Intuitively, a contradiction in an unsatisfiable instance is generally harder to detect if it is derived from more statements. Following this idea, we define the *complexity* of an unsatisfiable instance by the size of the *minimal unsatisfiable subset*, which is the smallest subset of statements that leads to a contradiction. For example, the instance

[5] For instances in the **2D2OMR**- or **OMR22D**-task, we only reverse the last (appended) statement.

A *left* B, B *left* C, C *left* A, A *right* D has complexity 3 since the first three statements form a strict directed cycle and any two statements do not contradict each other. Similarly, the instance $\mathit{left}(A, B, C), \mathit{left}(A, D, B), \mathit{left}(A, C, D), \mathit{left}(B, D, C)$ has complexity 4 because four statements together are unsatisfiable while any 3 statements are satisfiable. The complexity of a satisfiable instance is defined as the complexity of its contra-instance for simplicity. Figure 2 (b) shows the complexity distribution for the **1D**-, **2D**- and **OMR**-task in our constructed datasets, with domain size ranging from 3 to 26.

5 Experimental Setup

5.1 Language Models

To explore how well LMs can solve the satisfiability-checking problem in our tasks, we finetune three (small) language models, namely, RoBERTa large, DeBERTa-v3 large and T5 base[6]. This choice allows us to study spatial reasoning under relatively controlled conditions and to more clearly separate task learning from the complex and often opaque effects of large-scale pretraining. Our goal is therefore not to benchmark state-of-the-art large language models (LLMs), but to investigate how spatial reasoning can be acquired and represented in language models. To provide a point of comparison with modern LLMs, we additionally evaluate Gemini-3.5-Flash [15] in a zero-shot setting for the main experimental comparisons.

RoBERTa. The Robustly optimized BERT approach (RoBERTa) is trained in a more optimised manner based on the BERT architecture [30]. It removes the next sentence prediction objective, trains on longer sequences over more data and changes the masking pattern applied to the training data. It outperforms BERT on GLUE, SQuAD and RACE. We use RoBERTa large with 355M parameters.

DeBERTa-v3. The new architecture DeBERTa improves the BERT and RoBERTa models using a disentangled attention mechanism and an enhanced mask decoder [19]. The DeBERTa model consistently performs better than BERT and RoBERTa. DeBERTa-v3 further improves the efficiency of DeBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing [18]. We use DeBERTa-v3 large with 304M parameters.

T5. The Text-to-Text Transfer Transformer (T5) reframes all NLP tasks into a unified text-to-text-format where the input and output are text strings [42]. It is one of the state-of-the-art encoder-decoder language models. We employ T5 base with 220M parameters.

For all models, we utilise the Hugging Face[7] implementation with a classification layer on the top.

5.2 Implementation Details

Formally, we define the satisfiability-checking problem as a binary classification task. The input to the model is an instance (*i.e.*, a set of statements) that describes spatial relations between points, couched in natural language. The model needs to predict satisfiability of the input. Since the dataset is balanced, we use accuracy as the evaluation metric. For Gemini,

[6] We also considered T5 large, but it was unstable on our tasks, with several runs failing to converge.
[7] `https://huggingface.co`

Table 1 Accuracy of four models on our reasoning tasks. The first three models were finetuned on our task. Cells in the first three rows report post-finetuning (mean accuracy and standard deviation over three runs with random seeds 42, 418, and 2048) and pre-finetuning accuracies, shown before and after the slash (/), respectively. Gemini was evaluated without task-specific finetuning; we only report its zero-shot performance. *Model failed to converge even with smaller learning rates.

	1D	2D	OMR
RoBERTa large	99.03 ± 0.39 / 50.43	76.79 ± 2.78 / 50.12	50.03 ± 0.06* / 50.69
DeBERTa-v3 large	99.99 ± 0.01 / 51.05	99.42 ± 0.16 / 51.40	91.99 ± 0.02 / 50.06
T5 base	93.06 ± 3.22 / 50.26	67.69 ± 2.77 / 50.28	50.20 ± 0.16* / 50.05
Gemini-3.5	85.17	92.20	84.80

performance is evaluated in a zero-shot setting with standard prompts (Appendix A). We use the official API with temperature set to 0 to ensure deterministic outputs. Results are reported from a single run due to API cost constraints.

During finetuning, we optimise the standard cross-entropy loss for binary classification. We use the AdamW optimizer [31] with default parameters except for the learning rate. The learning rate is linearly warmed up from 0 to 1e-6 (for RoBERTa large and DeBERTa-v3 large) or 1e-5 (for T5 base) during the first epoch, followed by a linear decay to 0. We use a batch size of 24 and finetune each model for 10 epochs. These hyperparameters may not be optimal for every model; however, since the models already achieve strong and consistent performance, we do not conduct extensive hyperparameter tuning, accepting a trade-off between model-specific optimisation and experimental comparability. During inference, we pick the highest logit as the model's prediction.

6 Results and Discussion

6.1 Overall Model Performance

All investigated LMs yielded satisfactory performance on the **1D** and **2D** tasks after finetuning. By contrast, before finetuning the models performed at approximately chance level (50% accuracy), as shown in Table 1. This demonstrates that finetuning enables LMs to distinguish between minimally contrasting pairs. Accuracy decreases from the **1D** task to the **2D** task and further to the **OMR** task, which accords with our intuition that these tasks exhibit progressively greater difficulty. The performance of DeBERTa-v3 on the computationally demanding OMR task is particularly impressive.

The difference between models' performance increases from the **1D**-task to the **2D**-task to the **OMR**-task. All models performed remarkably well on the **1D**-task, while some of them dropped significantly on the other two tasks, or even failed to converge. Among them, DeBERTa-v3 always performed the best, even better than RoBERTa, which has more parameters. This might be due to its more effective architecture.

As a point of comparison with a powerful general-purpose language model, we additionally evaluated Gemini-3.5-Flash in a zero-shot setting. Gemini-3.5 achieved performance comparable to that of the finetuned language models across all three tasks, although its accuracy remained below that of DeBERTa-v3, the best-performing model. The strong zero-shot performance indicates that contemporary state-of-the-art LLMs already capture a considerable amount of spatial reasoning knowledge during pretraining, allowing them to generalise to our tasks without additional task-specific supervision.

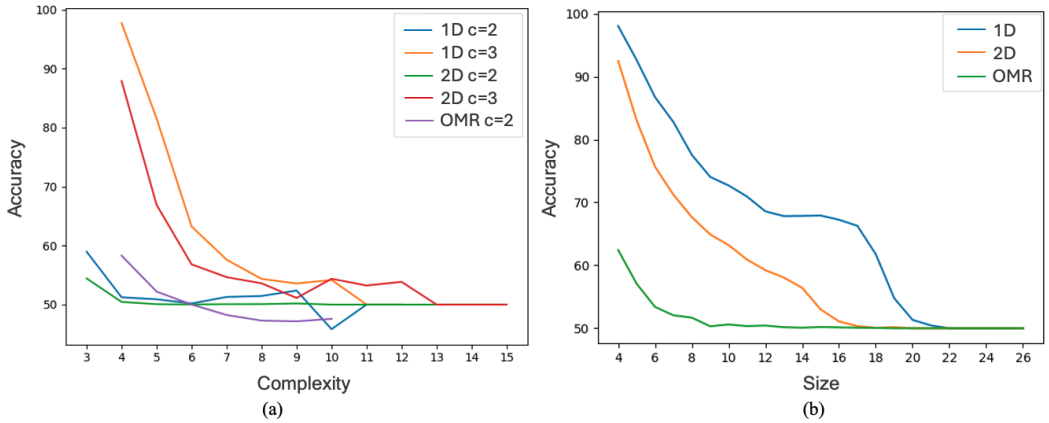


Figure 3 (a) Finetune DeBERTa-v3 on instances with complexity 2 or 3 and test on instances with higher complexity. c=2/3 indicates which complexity is the model finetuned on; (b) Finetune DeBERTa-v3 on instances with size 3 and test on instances with larger sizes.

6.2 Generalisation Ability

In the following experiments, we will use DeBERTa-v3 since it performed the best on all tasks. Figure 3 shows that DeBERTa-v3 exhibits different generalisation behaviour according to instance complexity or instance size. When the model is finetuned on instances with complexity 2, it shows poor generalisation ability on higher complexity, with accuracy only slightly better than a random guess (*i.e.*, 50%). Since complexity 2 means two statements directly contradict each other, only learning this type of contradiction barely helps the model to detect more difficult cases. By contrast, when the model is finetuned on instances with complexity 3, it is exposed to more complicated situations.[8] Therefore, it acquires the ability to handle more difficult instances to some extent. Nevertheless, accuracy quickly drops to below 70% above complexity 6. Confusion matrices show that the model tends to predict unsatisfiable instances as satisfiable, which reveals that the model fails to detect much longer strict directed cycles than what it has seen during finetuning.

When the model is finetuned on instances with instance size 3, and tested on larger instance sizes, it demonstrates better generalisation ability than when it is finetuned on instances with small complexity. A possible reason is that, although the instance size increases, the complexity distribution remains relatively similar, which is seen by the model during finetuning. Therefore, the model can utilise the learnt knowledge within a reasonable scope. Despite that, accuracy falls back to random guess when the instance size is very large. This is anticipated as a larger instance size means it is more difficult to find the statements that lead to contradiction. In other words, the model is not robust to "distractors" (*i.e.*, statements that are not in the minimal unsatisfiable subset).

In both cases, it is obvious that the model shows best generalisation ability on the **1D**-task, with the worst generalisation ability observed on the **OMR**-task. As observed in [32], LMs can learn to understand logically significant words and grammatical constructs more easily from simpler fragments than harder ones.

[8] No instance has complexity 3 in the **OMR**-task. In other words, any three statements are satisfiable whenever every pair of them is satisfiable.

Table 2 Accuracy of DeBERTa-v3 when finetuned on one task and tested on the other two tasks. The column headings denote which dataset is the model finetuned on. The row headings denote which dataset is the model tested on.

Test \ Finetune	1D	2D	OMR
1D	100.00	99.39	58.22
2D	59.89	99.77	55.84
OMR	48.70	49.02	95.67

Finally – essentially as a sanity check – we test how well the model can generalise to out-of-domain tasks. Specifically, DeBERTa-v3 is finetuned on one task, then tested on the other two tasks. Results in Table 2 support that the **1D**- and **2D**-task are similar to each other, but are different from the **OMR**-task. The model is robust to the change in relation names, as the model finetuned on the **2D**-task achieved very high accuracy when tested on the **1D**-task. However, the model finetuned on the **1D**-task only attained an adequate result on the **2D**-task, probably because it did not learn to check for both axes (*i.e.*, north-south and west-east) during finetuning. The model finetuned on the **OMR**-task presents generalisation ability on the **1D**- and **2D**-task to some extent. This indicates that the model finetuned on the hardest task still manages to gain some general spatial reasoning ability to employ on simpler tasks, despite the discrepancy in underlying rules.

Analysing confusion matrices gives us additional insights. We observed that if the model is finetuned on a simpler task and tested on a harder task, then it tends to predict unsatisfiable instances as satisfiable. By contrast, if the model is finetuned on a harder task and tested on a simpler task, then more satisfiable instances are predicted as unsatisfiable. This asymmetric generalisation likely stems from the difference in data structure among different tasks. Unsatisfiable instances in the simpler task tend to involve more explicit contradictions, whereas those in the harder task often depend on more complicated interactions. As a result, training on the simpler task may lead to a conservative unsatisfiability criterion that fails to generalise to the harder task, while training on the harder task may induce a more aggressive criterion that overgeneralises to the simpler task.

6.3 Perspective Transformation

The above results have demonstrated that LMs can gain basic spatial reasoning ability to some extent after finetuning on targeted tasks. We then move to perspective transformation. Switching between route-based view and survey-based view is an important ability which is – sometimes implicitly – involved in everyday life. There have long been studies in psychology that try to understand whether different perspectives lead to different mental representations [57, 55, 49, 50]. These studies have revealed that humans display asymmetrical behaviour when switching between the two perspectives. We are interested in whether LMs exhibit similar bias as humans.

We finetune DeBERTa-v3 on the two training sets corresponding to the **2D2OMR**- and **OMR22D**-task, respectively, and test on the test sets. The model achieved 94.44% accuracy on the **OMR22D**-task, while only 81.92% accuracy on the **2D2OMR**-task. The model appears to be more proficient at transforming from route perspective to survey perspective, in contrast to the tendency observed in human subjects, who performed better when transforming from survey perspective to route perspective. If we regard the **2D**- and the **OMR**-task as "single-perspective tasks" (*i.e.*, no perspective transformation is needed

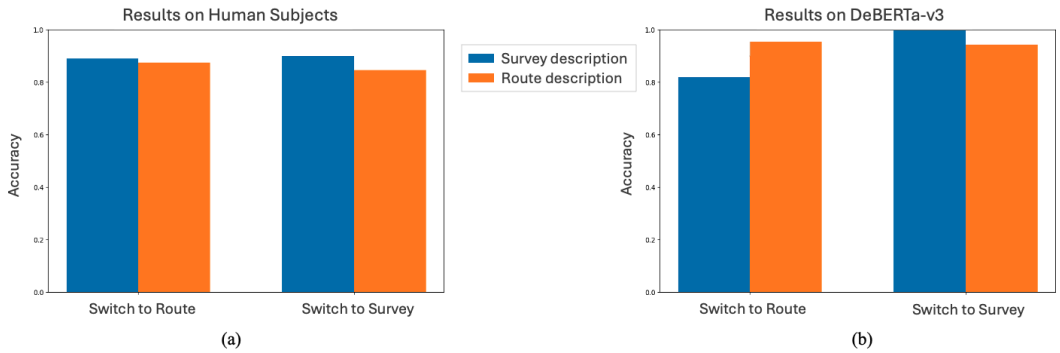


Figure 4 Performance of perspective transformation of human subjects (a) and DeBERTa-v3 (b). Survey/route description means the perspective taken in the textual description. x-axis indicates the perspective required to do inference. For humans, this is the perspective in which the question is asked. For the model, this is the perspective of the appended statement. Figure (a) is redrawn from Figure 5 in [57] based on the results reported in the paper. The figure is intended to illustrate qualitative performance patterns rather than provide a direct comparison of absolute accuracy.

during inference), comparing between these two tasks and the two "cross-perspective task" (*i.e.*, **2D2OMR**- and **OMR22D**-task) uncovers more differences between LMs and humans. Results on human subjects and DeBERTa-v3 are shown in Figure 4. Human results are drawn from a related experiment reported by Tversky [57] rather than being collected on our dataset. Moreover, our experimental setup is only inspired by, rather than a direct replication of, the original human study. In particular, the human experiment involved additional cognitive processes such as memory retention, delayed recall, and information retrieval, which cannot be straightforwardly mirrored in the language model setting. Accordingly, the comparison is intended to highlight qualitative similarities and differences in performance patterns across conditions, rather than to support a direct comparison of absolute accuracy.

Human performance appears to be more influenced by the perspective of the textual description itself: being given survey descriptions yields higher accuracy overall. In contrast, model performance seems more sensitive to whether perspective transformation is required, with single-perspective tasks generally outperforming cross-perspective tasks. As noted above, these comparisons should be interpreted cautiously, given the differences in datasets and experimental protocols. Nevertheless, the results suggest that humans and language models may be affected differently by perspective-related manipulations, leading to distinct patterns of performance across conditions.

Interestingly, Gemini-3.5 does not exhibit the same asymmetry. In the zero-shot setting, it achieved 85.77% accuracy on the **OMR22D**-task and 87.58% accuracy on the **2D2OMR**-task, comparable to the performance of finetuned DeBERTa-v3. This observation suggests that the asymmetry observed for DeBERTa-v3 may not generalise to all language models and warrants further investigation. At the same time, the strong performance of Gemini-3.5 suggests that contemporary LLMs can handle perspective transformations reasonably well even without task-specific supervision.

Finally, we investigate how the ratio of the number of statements from two perspectives influences model performance. We fix the total number of statements (equal to the number of points) in an instance to n, with the number of statements from the **2D**-task, $\#survey$, varying from 0 to n, and the number of statements from the **OMR**-task varying from n to 0. We choose n to be 5, 10, 15 and construct three balanced datasets. Each dataset is split into a training set with approximately 72K instances and a test set with approximately 10K instances. The number of instances with different ratios is the same within each dataset.

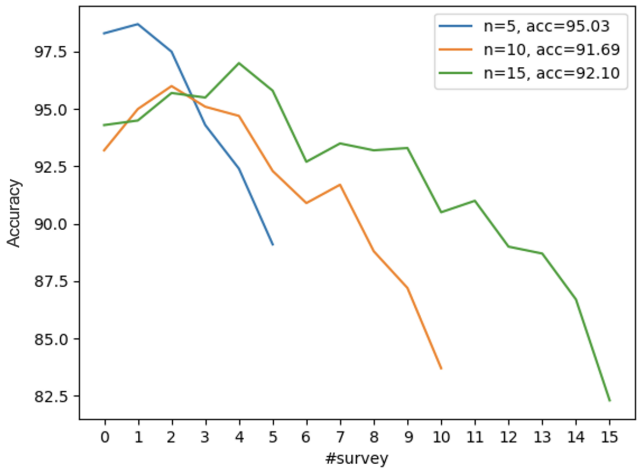


Figure 5 Accuracy breakdown based on the number of survey-perspective sentences on DeBERTa-v3. "acc" means the overall accuracy on the test set after finetuning.

We finetune DeBERTa-v3 on the training sets and test on the test sets. Figure 5 shows a trend of first increasing then decreasing with $\#survey$ increasing. The peak occurs when the proportion of survey-perspective statements (*i.e.*, $A\ r\ B$) is around 20%. We observe similar results as above that the model performed better when route-perspective statements (*i.e.*, $left/right(A, B, C)$) accounts for a larger proportion.

At present, we do not have a clear account of the observed results. While this pattern is robust across our experiments and is theoretically intriguing, we currently lack sufficient evidence to determine the underlying cause, and any explanation would therefore be highly speculative. We therefore wish to highlight this phenomenon as an important open question raised by the present study. Understanding why this specific mixture yields optimal performance may provide new insights into how language models integrate route-based and survey-based spatial information.

7 Discussion

Our findings have several implications for both the study of language models and the broader literature on spatial cognition. First, the results suggest that spatial reasoning should not be viewed as a single, homogeneous capability. Across our experiments, model performance varied substantially among the **1D**-, **2D**-, and **OMR**-tasks, despite all three being formulated as satisfiability-checking problems involving spatial relations. This observation indicates that different forms of spatial reasoning place qualitatively different computational demands on language models. In particular, the comparatively strong performance of DeBERTa-v3 on the **OMR**-task suggests that contemporary pretrained models can learn non-trivial forms of spatial-logical reasoning when provided with sufficient task-specific supervision.

Second, our perspective-transformation experiments suggest that reasoning within a perspective and transforming between perspective may constitute distinct challenges. Although DeBERTa-v3 achieved high accuracy on both the **2D**- and **OMR**-tasks individually, performance decreased when information had to be integrated across route and survey perspectives. This finding echoes long-standing observations in spatial cognition that different representational formats can place different demands on reasoning processes. More generally, it suggests that success on spatial reasoning tasks does not necessarily imply equal proficiency in perspective transformation.

A further implication concerns the relationship between language models and human spatial cognition. While our comparison with human results should be interpreted cautiously due to differences in datasets and experimental protocols, the observed performance patterns do not appear to align well with those reported in the psychological literature. Rather than supporting strong conclusions about underlying cognitive mechanisms, these findings highlight the value of perspective transformation as a diagnostic setting for comparing human and model behaviour. At present, our analysis is limited to behavioural outcomes, namely performance across different perspective conditions. Future work could therefore move beyond accuracy-based comparisons and investigate intermediate reasoning traces, such as chain-of-thought explanations, reasoning trajectories, or other forms of model-generated rationales. Such analyses may help clarify whether similarities and differences in task performance reflect comparable representational strategies or fundamentally different computational processes.

Finally, our results may have implications for GeoAI and language-based spatial reasoning systems. Practical applications such as spatial question answering, navigation assistance, and wayfinding support often require information to be translated between route-based and survey-based descriptions. The reduced performance observed in cross-perspective tasks suggests that such transformations remain a challenge even when basic spatial reasoning is largely mastered. This observation motivates future work on architectures that explicitly represent perspective information, as well as hybrid approaches that combine language models with established qualitative spatial reasoning techniques.

8 Conclusion

We begin with three basic spatial reasoning tasks that can be viewed as abstract versions of everyday inferential tasks: the **1D**-, **2D**- and **OMR**-tasks. All language models we investigated achieved satisfactory performance on the **1D**- and **2D**-task, indicating that they can learn to solve these problems. Notably, DeBERTa-v3 copes well with the computationally demanding **OMR**-task. It also demonstrates a certain degree of both in-domain and out-of-domain generalisation. Concentrating on DeBERTa-v3, given its superior performance, we then turn our attention to the issue of perspective transformation, which has long been studied in cognitive psychology on humans. We abstract the perspective transformation task into two tasks: the **2D2OMR**- and **OMR22D**-tasks, switching from survey view (2D) to route view (OMR), or from route view to survey view, respectively. We find that the model is better at switching from route view to survey view, in contrast to the phenomenon observed on humans in psychological experiments. These findings suggest that language models and humans may exhibit different patterns of performance in perspective-dependent spatial reasoning, although the differences in experimental setup preclude strong conclusions regarding the underlying cognitive processes.

References

1 Marco Aiello, Ian Pratt-Hartmann, and Johan Van Benthem, editors. *Handbook of spatial logics*. Dordrecht: Springer, 2007. doi:10.1007/978-1-4020-5587-4.

2 Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.ijcnlp-main.45.

3 Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1075.

4 Ruth M. J. Byrne and P. N. Johnson-Laird. Spatial Reasoning. *Journal of Memory and language*, 28:564–575, 1989. doi:10.1016/0749-596X(89)90013-2.

5 Anthony G. Cohn. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8, 2023. doi:10.48550/arXiv.2309.15577.

6 Anthony G Cohn and Robert E Blackwell. Evaluating the ability of large language models to reason about cardinal directions. In *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. doi:10.4230/LIPIcs.COSIT.2024.28.

7 Anthony G. Cohn and Robert E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions, revisited. In *QR 2025: 38th International Workshop on Qualitative Reasoning at IJCAI*, Montreal, Canada, 2025. doi:10.48550/arXiv.2507.12059.

8 Anthony G Cohn and Jose Hernandez-Orallo. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs, 2023. doi:10.48550/arXiv.2304.11164.

9 Anthony G. Cohn and Jochen Renz. Chapter 13 Qualitative spatial representation and reasoning. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 551–596. Elsevier, 2008. doi:10.1016/S1574-6526(07)03013-1.

10 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

11 Heshan Du, Natasha Alechina, and Anthony G. Cohn. A logic of directions. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1695–1702, Yokohama, Japan, 2020. International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2020/235.

12 Andrew U. Frank. Qualitative spatial reasoning: cardinal directions as an example. *International journal of geographical information systems*, 10(3):269–290, 1996. doi:10.1080/02693799608902079.

13 Christian Freksa. Using orientation information for qualitative spatial reasoning. In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space: International Conference GIS—From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning*, pages 162–178, Pisa, Italy, 1992. Springer Berlin Heidelberg. doi:10.1007/3-540-55966-3_10.

14 Nir Fulman, Abdulkadir Memduhoglu, and Alexander Zipf. Evidence for systematic bias in the spatial memory of large language models. In *GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024*, pages 57–62, Glasgow, Scotland, 2024. Springer Cham. URL: https://ceur-ws.org/Vol-3683/paper8.pdf.

15 Google. Gemini 3.5 flash, 2025. URL: https://deepmind.google/models/model-cards/gemini-3-5-flash/.

16 Gracjan Góral, Alicja Ziarko, Piotr Miłoś, Michał Nauman, Maciej Wołczyk, and Michał Kosiński. Beyond recognition: Evaluating visual perspective taking in vision language models, 2025. doi:10.48550/arXiv.2505.03821.

17 Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou,

Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.1229.

18 Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

19 Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

20 Majid Hojati and Rob Feick. Large language models: Testing their capabilities to understand and explain spatial concepts. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 31:1–31:9, Québec City, Canada, 2024. doi:10.4230/LIPIcs.COSIT.2024.31.

21 Roman Kontchakov, Ian Pratt-Hartmann, and Michael Zakharyaschev. Spatial reasoning with RCC8 and connectedness constraints in Euclidean spaces. *Artificial Intelligence*, 217:43–75, 2014. doi:10.1016/j.artint.2014.07.012.

22 Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

23 Bridget Leonard and Scott O. Murray. Cognitively-inspired tokens overcome egocentric bias in multimodal models, 2026. doi:10.48550/arXiv.2601.16378.

24 Fangjun Li. *Benchmarking and enhancing spatial reasoning in large language models*. PhD thesis, University of Leeds, 2025.

25 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18500–18507, Vancouver, Canada, 2024. AAAI Press, Washington, DC, USA. doi:10.1609/aaai.v38i17.29811.

26 Zekai Li, Amin Farjudian, and Heshan Du. A logic of east and west for intervals. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 17:1–17:8, Québec City, Canada, 2024. doi:10.4230/LIPIcs.COSIT.2024.17.

27 G E Ligozat. Reasoning about cardinal directions. *Journal of Visual Languages and Computing*, 9:23–44, 1998. doi:10.1006/jvlc.1997.9999.

28 Gérard Ligozat, editor. *Qualitative Spatial and Temporal Reasoning*. John Wiley & Sons, 2013. doi:10.1002/9781118601457.

29 Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods, 2025. doi:10.48550/arXiv.2511.15722.

30 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

31 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

32 Tharindu Madusanka, Riza Theresa Batista-Navarro, and Ian Pratt-Hartmann. Identifying the limits of transformers when performing model-checking with natural language. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3539–3550, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.eacl-main.257.

33 Lachlan McPheat, Navdeep Kaur, Robert Blackwell, Alessandra Russo, Anthony G. Cohn, and Pranava Madhyastha. DecompSR: A dataset for decomposed analyses of compositional multihop spatial reasoning, 2026. doi:10.48550/arXiv.2511.02627.

34 Roshanak Mirzaee and Parisa Kordjamshidi. Transfer learning with synthetic corpora for spatial role labeling and reasoning, 2022. doi:10.48550/arXiv.2210.16952.

35 Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.364.

36 Reinhard Moratz and Marco Ragni. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1):75–98, 2008. Spatial and Image-based Information Systems. doi:10.1016/j.jvlc.2006.11.001.

37 Reinhard Moratz, Jochen Renz, and Diedrich Wolter. Qualitative spatial reasoning about line segments. In *European Conference on Artificial Intelligence*, 2000. URL: https://api.semanticscholar.org/CorpusID:1424882.

38 Till Mossakowski and Reinhard Moratz. Qualitative reasoning about relative direction of oriented points. *Artificial Intelligence*, 180-181:34–45, 2012. doi:10.1016/j.artint.2011.10.003.

39 Till Mossakowski and Reinhard Moratz. Relations between spatial calculi about directions and orientations. *Journal of Artificial Intelligence Research*, 54(1):277–308, September 2015. doi:10.1613/jair.4631.

40 Joseph O'Rourke, editor. *Computational geometry in C*. Cambridge university press, 2012. doi:10.1017/CBO9780511804120.

41 Tanawan Premsri and Parisa Kordjamshidi. Tuning language models with spatial logic for complex reasoning. In *The 4th International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, 2024. URL: https://openreview.net/forum?id=cDm7zGenhZ.

42 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21:1–67, 2020. URL: http://jmlr.org/papers/v21/20-074.html.

43 Marco Ragni and Markus Knauff. A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, 120:561–588, 2013. doi:10.1037/a0032460.

44 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi:10.18653/v1/D16-1264.

45 David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 165–176, Cambridge, United States, 1992. Morgan Kaufmann Publishers Inc.

46 Kyle Richardson and Ashish Sabharwal. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11209–11219, Virtual, 2022. AAAI Press, Palo Alto, California USA. doi:10.1609/aaai.v36i10.21371.

47 Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Seoul, South Korea, 2025. doi:10.48550/arXiv.2507.07644.

48 Felipe Salvatore, Marcelo Finger, and Roberto Hirata Jr. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for*

Low-Resource NLP (DeepLo 2019), pages 22–30, Stroudsburg, PA, USA, May 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-6103.

49 Amy L. Shelton and Timothy P. McNamara. Orientation and perspective dependence in route and survey learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30:158–170, 2004. doi:10.1037/0278-7393.30.1.158.

50 Amy L. Shelton and Holly A. Pippitt. Fixed versus dynamic orientations in environmental learning from ground-level and aerial perspectives. *Psychological Research*, 71:333–346, 2007. doi:10.1007/s00426-006-0088-9.

51 Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11321–11329, Virtual, 2022. AAAI Press, Palo Alto, California USA. doi:10.1609/aaai.v36i10.21383.

52 Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.1195.

53 Alexander W. Siegel and Sheldon H. White. The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, 10:9–55, 1975. doi:10.1016/S0065-2407(08)60007-5.

54 Michael Sioutis and Diedrich Wolter. Qualitative spatial and temporal reasoning: Current status and future challenges. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21): Survey Track*, 2021. doi:10.24963/ijcai.2021/624.

55 Holly A. Taylor and Barbara Tversky. Spatial mental models derived from survey and route descriptions. *Journal of Memory and language*, 31:261–292, 1992. doi:10.1016/0749-596X(92)90014-O.

56 Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.303.

57 Barbara Tversky. Spatial Mental Models. *Psychology of learning and motivation*, 27:109–145, 1991. doi:10.1016/S0079-7421(08)60122-X.

58 Barbara Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *International Conference on Spatial Information Theory (COSIT 1993)*, pages 14–24, Marciana Marina, Elba Island, Italy, 1993. Springer Berlin, Heidelberg. doi:10.1007/3-540-57207-4_2.

59 Maijunxian Wang, Yijiang Li, Bingyang Wang, Tianwei Zhao, Ran Ji, Qingying Gao, Emmy Liu, Hokin Deng, and Dezhi Luo. Egocentric bias in vision-language models, 2026. doi:10.48550/arXiv.2602.15892.

60 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

61 Steffen Werner, Bernd Krieg-Brückner, Hanspeter A. Mallot, Karin Schweizer, and Christian Freksa. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. In *Informatik '97 Informatik als Innovationsmotor*, pages 41–50, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. doi:10.1007/978-3-642-60831-5_8.

62 Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks, 2015. arXiv:1502.05698.

63 Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1101.

64 Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6:1–30, 2025. doi:10.1016/j.patter.2025.101370.

65 Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. Evaluating large language models on spatial tasks: A multi-task benchmarking study, 2025. doi:10.48550/arXiv.2408.14438.

66 Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition, 2025. doi:10.48550/arXiv.2511.21471.

67 Anran Yang, Cheng Fu, Qingren Jia, Weihua Dong, Mengyu Ma, Hao Chen, Fei Yang, and Hui Wu. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, 39(9):2009–2044, 2025. doi:10.1080/13658816.2025.2490701.

68 Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, 2019. Curran Associates, Inc. doi:10.48550/arXiv.1906.08237.

69 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=5Xc1ecxO1h.

A Zero-shot Prompt

Input: *the_original_input*

Question: Each uppercase letter represents a point, and different points cannot overlap. Can all statements in the input be true at the same time? Only answer Yes or No.

Answer: