From Route to Survey: Probing Spatial Perspective Transformation in Large Language Models

Haotong Zhang
Department of Computer Science
The University of Manchester
haotong.zhang@manchester.ac.uk

Abstract

Spatial perspective transformation is a fundamental component of human spatial cognition, yet it remains unclear whether large language models (LLMs) can exhibit robust spatial reasoning abilities that support such transformations. We introduce a controlled dataset inspired by cognitive science studies, combining route-view and survey-view descriptions across multiple environmental topologies. Experiments on both open-source and closed-source LLMs show that models readily learn within-perspective reasoning but continue to struggle with cross-perspective transformation, especially from route-view to survey-view reasoning. The learnt capabilities remain strongly topology-dependent and generalise asymmetrically between within-perspective and cross-perspective settings. Scaling substantially improves allocentric and within-perspective reasoning, but does not uniformly resolve cross-perspective transformation difficulty, while chain-of-thought prompting provides little benefit. Overall, our results suggest that current LLMs do not consistently demonstrate perspective-invariant spatial reasoning across perspectives and environments. Code and datasets are available at this link.

1 Introduction

Humans routinely reason about space when navigating environments, interpreting maps, or following spatial instructions. Such abilities have long been studied in cognitive psychology as fundamental components of human cognition (Waller and Nadel, 2013; Denis, 2017). At the same time, recent large language models (LLMs) have demonstrated increasingly strong reasoning abilities across a wide range of tasks (Grattafiori et al., 2024; Yang et al., 2025; Guo et al., 2025; OpenAI, 2025), raising the question of whether such models can also acquire robust spatial reasoning capabilities.

One important form of spatial reasoning concerns transformations between route-view and survey-view representations. Route-view 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-view descriptions, by contrast, describe the environment from a global bird's-eye perspective: *At the east end of the Mall stands the Capitol and at the west end* ... These two perspectives correspond broadly to egocentric and allocentric spatial representations, respectively. The ability to switch between them—commonly referred to as *perspective transformation*—has been widely studied in cognitive psychology (Tversky, 1991; Taylor and Tversky, 1992; Tversky, 1993). In this paper, we investigate whether LLMs can perform such transformations and whether their behaviours show qualitative similarities to those reported in human studies.

Recent work has increasingly examined perspective taking in LLMs and multimodal language models. However, most prior work focuses on viewpoint shifts between egocentric observers (Patania et al., 2025), while transformations between route-view and survey-view representations remain underexplored. This raises a question: *Do LLMs exhibit behaviour consistent with stable spatial representations that generalise across perspectives and environments, or do they instead rely on more local and context-dependent reasoning strategies?*

To investigate these questions, we construct a controlled dataset for spatial perspective transformation inspired by cognitive psychology experiments. The dataset combines route-view and survey-view descriptions and questions across multiple topologies, allowing us to disentangle perspective consistency, perspective transformation, and topology-dependent effects. We evaluate both open-source and closed-source LLMs in finetuning and zero-shot settings, and further study the robustness of perspective transformation through generalisation, scaling, and prompt engineering.

Our results suggest that current LLMs do not yet

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21368–21386
October 24-29, 2026 ©2026 Association for Computational Linguistics

exhibit robust perspective transformation behaviour that generalises across perspectives and environmental topologies. Although models can learn certain forms of within-perspective reasoning through finetuning and scaling, these capabilities remain highly dependent on environmental topology and degrade substantially under cross-perspective transformation. This pattern suggests that current LLMs may rely on context-sensitive reasoning behaviours that do not consistently generalise across spatial perspectives and environments. More broadly, our findings position spatial perspective transformation as a useful behavioural probe of whether LLMs can generalise spatial reasoning beyond surface-level task patterns.

2 Related Work

Spatial reasoning has long been studied in cognitive psychology, where researchers investigated how humans comprehend spatial information, construct mental models, and reason about spatial environments (Byrne and Johnson-Laird, 1989; Ragni and Knauff, 2013). Prior work suggests that humans construct mental representations through structured reasoning processes while often overlooking alternative spatial interpretations. Another major line of research focuses on perspective taking, particularly between route (egocentric) and survey (allocentric) representations. Early work by Siegel and White (Siegel and White, 1975) distinguished between route and survey knowledge in the development of spatial representations, while subsequent studies further examined the roles of landmark, route, and survey knowledge in navigation (Werner et al., 1997). Empirical studies have shown that route and survey representations differ not only in representational format but also in the cognitive processes required to construct and manipulate them (Tversky, 1991, 1993). Additional work further demonstrated that original orientation can significantly affect spatial reasoning accuracy (Shelton and McNamara, 2004; Shelton and Pippitt, 2007).

In recent years, textual spatial reasoning has attracted growing attention in NLP, leading to benchmarks such as SpartQA (Mirzaee et al., 2021), SpaRTUN (Mirzaee and Kordjamshidi, 2022), and StepGame (Shi et al., 2022). With the emergence of LLMs, increasing work has examined their spatial reasoning abilities in few-shot and zero-shot settings, reporting persistent limitations and human-like biases in spatial inference (Li et al., 2024; Fulman et al., 2024; Cohn and Blackwell, 2025). However, most existing work focuses on local spatial reasoning rather than perspective transformation across spatial reference frames.

More 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 observers. Empirical analyses report systematic limitations in spatial inference in MLLMs (Shiri et al., 2024). Visual perspective-taking has been evaluated through tasks requiring models to adopt alternative observer viewpoints, revealing persistent difficulties in shifting away from egocentric representations (Góral et al., 2025). Other studies document strong egocentric biases in VLMs (Wang et al., 2026a), while cognitively inspired intermediate representations improve cross-perspective reasoning (Leonard and Murray, 2026). Some work also investigates collaborative spatial integration across partial viewpoints (Sikarwar et al., 2026) and spatial mental modeling from limited visual observations (Wang et al., 2026b), highlighting the difficulty of constructing global spatial representations.

While related, these studies primarily investigate observer-based viewpoint shifts in visual environments. The most similar work to ours is SpatialText (Jiang et al., 2026). This benchmark broadly evaluates text-based spatial cognition using human-annotated indoor scenes and code-generated spatial relations, with perspective transformation included as one component of a wider task taxonomy. In contrast, our work isolates route–survey transformation as the central experimental variable through paired route-view and survey-view descriptions and questions in linguistic spatial reasoning, a setting central to human spatial cognition research but still underexplored in language models.

3 Task Definition and Dataset Construction

We study spatial perspective transformation, which requires reasoning across different spatial reference frames. In humans, this ability has been extensively studied in cognitive science as a fundamental component of spatial reasoning. We are interested in whether LLMs exhibit such abilities in zero-shot settings or can acquire them through finetuning. Following prior work in human spatial cognition, we consider two fundamental perspectives: *route*

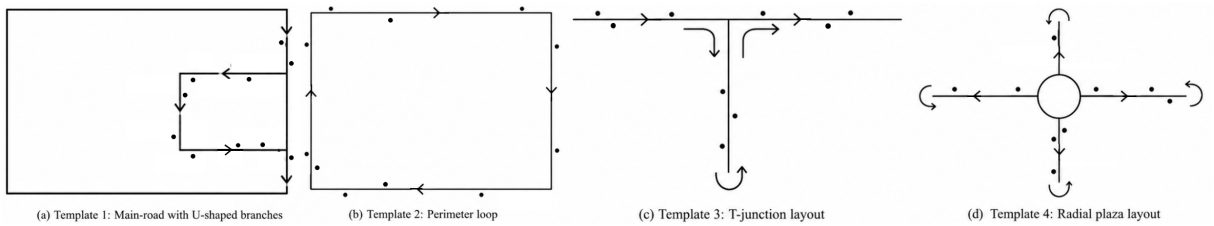


Figure 1: Draft of four templates. Lines indicate linear landmarks. Dots indicate point landmarks. The circle in (d) indicates the plaza. Arrows indicate the traversal order.

and *survey*. The route perspective corresponds to an egocentric viewpoint, where spatial relations are described relative to the observer's orientation (*i.e.*, left, right, front, and back) as the observer moves through an environment. In contrast, the survey perspective corresponds to an allocentric, bird's-eye representation, where spatial relations are expressed using global cardinal and intercardinal directions such as north, east, and northeast[1].

In prior cognitive psychology experiments, human participants are given a textual description of an area containing multiple landmarks and are asked to determine the spatial relation between two landmarks in either the same or a different perspective. We adopt a similar setup in this work. The area contains both linear landmarks (*e.g.*, roads, rivers, beaches) and point landmarks (*e.g.*, hotels, museums, theaters). The model must infer the target spatial relation from the description and answer the question. Both the description and the question can be presented in either route-view or survey-view form, resulting in four descripton-question combinations: route-to-route (**RR**), route-to-survey (**RS**), survey-to-route (**SR**), and survey-to-survey (**SS**). RR and SS require reasoning within the same spatial reference frame, while RS and SR additionally require perspective transformation between egocentric and allocentric representations. For RS, we additionally specify the observer's initial facing direction to ground the egocentric route description within the global coordinate system.

In prior cognitive psychology experiments, spatial descriptions and questions are written manually. However, this process is both labour-intensive and difficult to control, and may unintentionally introduce commonsense or linguistic biases. To address this issue, we design an automatic generation and annotation pipeline for constructing synthetic area descriptions and question-answer pairs. To study how different area layouts affect spatial reasoning performance, we define four templates with distinct environmental topologies, as shown in Figure 1.

Template 1 (Main-road with U-shaped branches). The area is bounded on all four sides, with one boundary serving as a main road. Three secondary roads connect to the main road in a U-shaped structure. The description starts at one end of the main road and proceeds along it. Whenever a secondary road is encountered, the observer traverses the secondary-road loop before returning to the main road and continuing forward. Landmarks are placed along both sides of all traversed roads.

Template 2 (Perimeter loop). The area is enclosed by four boundaries forming a rectangular loop. The description starts at one corner and follows the perimeter before returning to the starting point. Landmarks are placed along both sides of each boundary segment.

Template 3 (T-junction layout). The area consists of two intersecting roads forming a T-junction. The description begins at one end of the main road. After reaching the intersection, the observer traverses the secondary road to its end, returns to the intersection, and then continues along the main road until reaching the opposite end. Landmarks are placed along both sides of both roads.

Template 4 (Radial plaza layout). The area contains a central plaza from which four roads extend radially in the cardinal directions. The description starts at the plaza and traverses each road in turn: the observer walks to the end of a road, returns to the plaza, and then continues with the next road until all four roads have been visited. Landmarks are placed along both sides of each road.

To construct the dataset, we first define a symbolic spatial layout for each template, specifying the environmental topology and traversal structure of the area. For each instance, a set of landmarks is randomly sampled and assigned to valid loca-

[1]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.

tions within the layout. Each point landmark is attached to exactly one linear landmark and associated with a relative position along that landmark (*e.g.*, one quarter along the main road). Based on this symbolic representation, we automatically generate both route-view and survey-view textual descriptions using predefined description templates.

To generate question-answer pairs, we manually define a set of question templates corresponding to spatial relations that are linguistically natural and unambiguous. For each question instance, valid landmark pairs satisfying the constraints of the selected question template are randomly sampled. The correct answer is then computed automatically from the underlying symbolic spatial layout.

Although every instance from the same template shares the same underlying topology, we vary the global orientation, landmark identities, the number of point landmarks attached to each linear landmark, their side assignments, and their relative positions along the attached linear landmarks. These sources of variation make individual areas highly diverse and substantially reduce the possibility of memorising specific instances.

Each generated area therefore yields both route-view and survey-view descriptions, together with corresponding route-view and survey-view question-answer pairs. For each template, we generate 200 areas for training and 20 areas for testing. Each area contains 20 route-view and 20 survey-view question-answer pairs. In other words, training/test sets are split at the **area** level, rather than at the question-answer pair level. All question-answer pairs generated from the same area are assigned to the same split. The route-view and survey-view descriptions of the same area are also kept together. Example instances for all templates are provided in Appendix B. We also provide ground-truth label distributions per template in Appendix C.

4 Experimental Setup

To examine LLMs' perspective transformation ability, we consider two open-source LLMs: LLaMA-3.1 8B (Grattafiori et al., 2024) and Qwen-3.5 9B (Qwen Team, 2026), and two closed-source LLMs: GPT-5.2 (OpenAI, 2025) and Gemini-3 (Google, 2025). The input consists of a textual description of an area together with a question asking about the spatial relation between two landmarks. The model is required to answer the question based on the description. We train separate models for the four description-question combinations (SS, SR, RS, and RR) to match the different output spaces and reasoning settings.

For supervised finetuning, we cast the task as a classification problem over a fixed set of directional labels. Open-source LLMs are obtained from Hugging Face with a classification head added on top of the language model. The classification head takes the final-layer hidden state of the last input token as input and outputs an N-dimensional vector, where $N = 4$ for route-view questions and $N = 8$ for survey-view questions. Inputs are truncated to the model's maximum sequence length. We finetune LLMs via Low-Rank Adaptation (LoRA). We follow the standard PEFT implementation with rank 8 and scaling factor $\alpha = 16$. LoRA adapters are applied to all attention projection layers (*i.e.*, query, key, value, and output projections). During finetuning, the objective is to reduce the cross-entropy loss between target labels and predicted labels. We use the AdamW optimizer with default settings (Loshchilov and Hutter, 2017). The learning rate is linearly warmed up from 0 to 1e-4, then linearly decayed to 0. The batch size is set to 8. We finetune each model for 3 epochs. We do not perform extensive hyperparameter tuning, as our goal is to probe the general behaviour of LLMs rather than achieve the best possible performance. To ensure fair comparison, we use the same training configuration across all models. During inference, predictions are made by selecting the label with the highest logit. We use three random seeds 42, 123 and 418 and report the mean and standard deviation across the three runs where appropriate.

For zero-shot evaluation of open-source models, we use the original causal language-model checkpoints from Hugging Face. Each input is inserted into a natural-language prompt asking the model to generate an answer. The model is required to choose answer from the four (for route-view questions) or eight (for survey-view questions) valid relations. Generated answers are normalised using case folding, token splitting and simple regular-expression rules before evaluation. Since the prompt explicitly restricts outputs to the provided choices, outputs that cannot be matched to a valid label are counted as incorrect. We use greedy decoding to ensure that the generated outputs are deterministic. Therefore, we report results from a single run.

For closed-source models, we likewise use generative prompting. We query the official APIs and

set temperature to 0 to minimise decoding variability. We evaluate performance in a zero-shot setting with different prompts. Results are reported from a single run due to API cost constraints. Prompt templates are provided in Appendix A.

5 Experiments and Results

In this section, we investigate whether LLMs can acquire robust spatial perspective transformation abilities through supervision, and whether the learnt capabilities generalise across environments, perspectives, and model scales. Rather than focusing solely on overall task accuracy, we aim to understand what types of spatial reasoning can be learnt, how these abilities transfer across description-question combinations, and whether current LLMs exhibit spatial reasoning behaviour that generalises consistently across perspectives.

To this end, we conduct a series of experiments designed to answer the following research questions: (1) Can LLMs acquire spatial perspective transformation abilities through supervised finetuning? (2) Do the learnt spatial reasoning capabilities generalise across different environmental topologies and description-question combinations? (3) How do different perspective transformation abilities scale with model size? (4) Can frontier closed-source LLMs solve the task through prompting alone, and does chain-of-thought prompting improve spatial reasoning performance?

We first evaluate finetuned open-source LLMs in different environment templates and analyse their behaviour across spatial directions. We then investigate generalisation across templates and description-question combinations, followed by a scaling analysis across multiple model sizes. Finally, we compare these findings with zero-shot learning on different prompts on frontier closed-source models.

5.1 Overall Performance

We first investigate whether supervised finetuning enables LLMs to acquire spatial perspective transformation abilities. We finetune LLaMA and Qwen models on all four templates and evaluate them under the four description-question combinations: SS, SR, RS, RR. We further compare the finetuned models with their zero-shot counterparts. Results are shown in Table 1. Since route-view and survey-view questions have different answer spaces, we additionally report chance-normalised accuracy for cross-combination comparison in Appendix D.1.

Across all templates and both models, the most consistent pattern is that within-perspective combinations (*i.e.*, SS, RR) substantially outperform cross-perspective combinations (*i.e.*, SR, RS). In particular, SS achieves the highest and most stable performance, with accuracies approaching perfect performance across nearly all templates after finetuning. RR also shows strong performance, especially on Template 3 and 4 where accuracies are close to 1.0, and outperforms SR and RS on Template 1 and 2. These results suggest that finetuned LLMs readily acquire within-perspective spatial reasoning abilities. The advantage of SS over the other combinations further suggests that survey-view reasoning is easier to learn than route-view reasoning. Unlike route descriptions, which require sequential traversal tracking and continuous orientation updating, survey descriptions provide a global allocentric representation of the environment, making spatial relations more directly accessible to the models. Moreover, the strong RR performance indicates that route-view description itself is not inherently difficult for the models to process. Instead, the primary challenge lies in transforming spatial information between perspectives.

In contrast, cross-perspective reasoning remains substantially more difficult after finetuning. Although SR and RS improve over zero-shot performance, their accuracies remain below SS and RR. Interestingly, unlike the strong asymmetry observed in zero-shot settings, where RS is consistently the worst-performing combination, finetuning largely reduces this gap, with SR slightly higher in most cases. This suggests that supervision partially helps models acquire perspective transformation abilities. Nevertheless, performance remains highly dependent on both model and template. In particular, LLaMA generally exhibits stronger or at least comparable cross-perspective performance compared with Qwen, especially in templates where RS reasoning becomes substantially more robust.

In the zero-shot setting, RS is consistently the most difficult combination across all templates. This provides a qualitative parallel with the difficulty of route-to-survey transformation reported in prior human studies (Tversky, 1991).

Finally, substantial performance differences persist across templates in both finetuned and zero-shot settings. For example, Template 3 and 4 yield stronger SR performance than the other two tem-

			SS	SR	RS	RR
Template 1	Finetune	LLaMA	0.917±0.142	0.535±0.015	0.531±0.024	0.626±0.064
		Qwen	0.983±0.022	0.523±0.009	0.491±0.013	0.544±0.012
	Zero-shot	LLaMA	0.205	0.323	0.075	0.405
		Qwen	0.426	0.422	0.185	0.415
Template 2	Finetune	LLaMA	0.948±0.007	0.574±0.004	0.470±0.168	0.674±0.003
		Qwen	0.953±0.013	0.537±0.032	0.325±0.059	0.671±0.008
	Zero-shot	LLaMA	0.233	0.350	0.072	0.310
		Qwen	0.387	0.305	0.234	0.328
Template 3	Finetune	LLaMA	0.998±0.003	0.627±0.013	0.484±0.038	0.992±0.002
		Qwen	0.998±0.001	0.629±0.013	0.552±0.123	0.998±0.004
	Zero-shot	LLaMA	0.328	0.140	0.090	0.355
		Qwen	0.376	0.280	0.189	0.347
Template 4	Finetune	LLaMA	0.999±0.001	0.598±0.014	0.626±0.099	1.000±0.000
		Qwen	0.997±0.006	0.612±0.010	0.284±0.031	0.997±0.004
	Zero-shot	LLaMA	0.273	0.305	0.043	0.307
		Qwen	0.434	0.263	0.173	0.270

Table 1: Performance comparison between finetuned and zero-shot models across different description-question combinations and templates. In the finetune setting, values are averaged over three runs and reported as mean ± standard deviation.

plates, indicating that the learnt spatial reasoning remains strongly topology-dependent rather than being consistent across environmental structures.

To further investigate whether finetuned models exhibit stable direction-specific biases, we analyse accuracy breakdowns across different directions for each description-question combination (see Appendix G for detailed results). Overall, we do not observe consistent direction-specific patterns across templates or models. This suggests that the learnt spatial reasoning strategies remain highly context-dependent and topology-dependent.

5.2 Generalisation Ability

5.2.1 Cross-Template Generalisation

We next investigate whether the learnt spatial reasoning abilities generalise across environmental topologies. For each description-question combination, we finetune Qwen on one template and evaluate it on the remaining templates. Results are shown in Table 2. Overall, we observe limited but non-trivial cross-template transfer, with patterns strongly dependent on both description-question combination and environmental topology. In particular, SS exhibits substantially stronger generalisation than other combinations. For example, models finetuned on Template 1 retain relatively high performance on the remaining templates under SS. This suggests that survey-view reasoning captures more abstract allocentric relations that transfer more effectively across environments. In contrast, combinations involving route-view reasoning remain less transferable; these tasks additionally require sequential traversal and orientation tracking, which vary with topology.

We also observe partial transfer between structurally similar templates. In particular, Templates 1 and 2 exhibit relatively stronger mutual transfer in SS, while Templates 3 and 4 show stronger mutual transferability in almost all combinations. One possible explanation is that the models learn topology-dependent traversal strategies that partially transfer across environments with similar structures.

Cross-template generalisation in SR and RS remains weak overall, especially in RS. Since Qwen itself achieves only moderate in-domain performance under these cross-perspective combinations, the learnt perspective transformation abilities may not generalise reliably across unseen environments. Similar generalisation patterns are also observed with LLaMA (Appendix D.2).

5.2.2 Cross-Combination Generalisation

We further investigate whether spatial reasoning abilities learnt under one description-question combination transfer to other combinations. Since survey-view and route-view questions require different output spaces, we only evaluate transfer between combinations that share the same question type. Specifically, we evaluate transfers between RR and SR, and between RS and SS. Results on Qwen are shown in Table 3.

Overall, we observe a clear asymmetry in

	finetune \ test	Template 1	Template 2	Template 3	Template 4
SS	**Template 1**	0.983±0.022	0.831±0.045	0.757±0.092	0.643±0.155
	Template 2	0.695±0.038	0.953±0.013	0.464±0.105	0.598±0.154
	Template 3	0.482±0.005	0.523±0.077	0.998±0.001	0.692±0.144
	Template 4	0.491±0.098	0.422±0.168	0.622±0.085	0.997±0.006
SR	**Template 1**	0.523±0.009	0.345±0.007	0.400±0.000	0.291±0.025
	Template 2	0.354±0.024	0.537±0.032	0.393±0.065	0.478±0.024
	Template 3	0.443±0.009	0.423±0.017	0.629±0.013	0.527±0.036
	Template 4	0.345±0.019	0.442±0.020	0.529±0.034	0.612±0.010
RS	**Template 1**	0.491±0.013	0.236±0.003	0.361±0.030	0.253±0.009
	Template 2	0.224±0.019	0.325±0.059	0.251±0.034	0.236±0.019
	Template 3	0.257±0.003	0.218±0.028	0.552±0.123	0.231±0.016
	Template 4	0.203±0.028	0.215±0.003	0.191±0.014	0.284±0.031
RR	**Template 1**	0.544±0.012	0.352±0.014	0.438±0.069	0.282±0.011
	Template 2	0.417±0.011	0.671±0.008	0.555±0.021	0.524±0.062
	Template 3	0.499±0.001	0.515±0.039	0.998±0.004	0.785±0.041
	Template 4	0.420±0.040	0.525±0.014	0.843±0.004	0.997±0.004

Table 2: Cross-template generalisation results on Qwen. Each row indicates the template used for finetuning, while each column indicates the template used for evaluation. Values are averaged over three runs and reported as mean ± standard deviation.

	RR-to-SR	**SR-to-RR**	**RS-to-SS**	**SS-to-RS**
Template 1	0.487 (-0.057)	0.501 (-0.022)	0.263 (-0.228)	0.322 (-0.661)
Template 2	0.467 (-0.204)	0.486 (-0.051)	0.256 (-0.069)	0.213 (-0.740)
Template 3	0.655 (-0.343)	0.631 (+0.002)	0.336 (-0.216)	0.306 (-0.692)
Template 4	0.610 (-0.387)	0.623 (+0.011)	0.246 (-0.038)	0.308 (-0.689)

Table 3: Cross-combination generalisation results on Qwen. Each column "A-to-B" denotes finetuning on combination A and evaluating on combination B. The values in parentheses indicate the performance difference between the accuracy on B and the original in-domain accuracy on A. All values are averaged over three runs.

cross-combination transfer. Transfer from within-perspective combinations to cross-perspective combinations remains substantially weaker than the reverse direction. In particular, models finetuned on SS experience dramatic performance drops when evaluated on RS. This suggests that finetuning on SS primarily teaches the model allocentric spatial relations, without requiring the model to process route-view representations or perform sequential orientation tracking. As a result, the learnt capabilities transfer poorly to route-view perspective transformation tasks. In contrast, transfer from RR to SR is more robust. One possible explanation is that survey-view reasoning may already partially exist in pretrained language representations, whereas route-view reasoning requires more explicit learning during finetuning. Consequently, models trained on route-view tasks may implicitly acquire some of the capabilities required for survey-view reasoning.

Interestingly, transfer from cross-perspective combinations to within-perspective combinations remains relatively strong, especially from SR to RR, where performance sometimes remains nearly unchanged or even increases. This asymmetry is consistent with cross-perspective tasks imposing additional transformation requirements beyond those needed for the corresponding within-perspective tasks. In other words, perspective transformation may behave as a higher-level compositional spatial skill that partially subsumes within-perspective reasoning. We also observe consistent patterns with LLaMA, as shown in Appendix D.2, indicating that the behaviours are robust across examined model families.

5.3 Scalability

We then examine how spatial perspective reasoning scales with model size using the Qwen-3.5 family, including 0.8B, 4B, 9B, and 27B dense models. Overall, performance improves with model size in most cases, but the improvement is highly uneven across description-question combinations and templates. Results are shown in Figure 2.

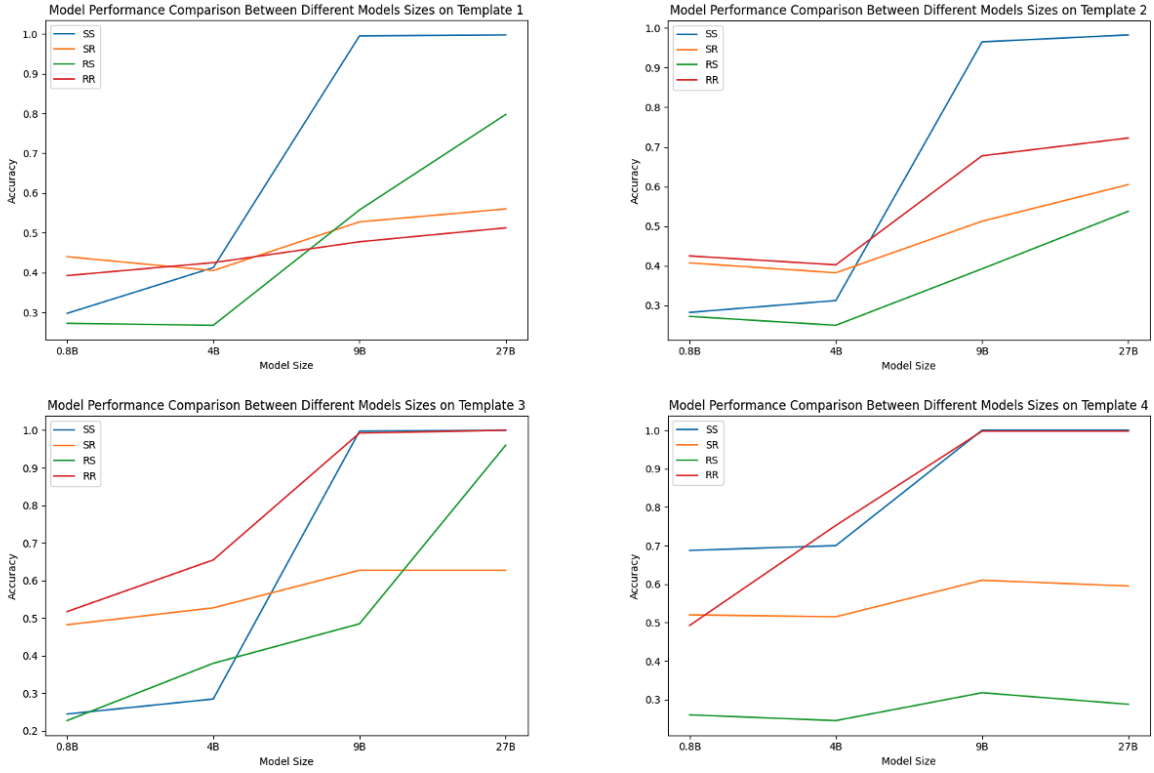


Figure 2: Model performance comparison between different model size from Qwen-3.5 family on all templates and all description-question combinations.

The clearest scaling trend appears in the SS. In all templates, SS shows the steepest improvement: smaller models perform poorly, while performance approaches saturation by 9B, leaving little additional gain for the 27B model. This suggests that survey-view reasoning is highly learnable with scale. One possible explanation is that SS mainly requires reasoning over global allocentric relations, which are more directly encoded in text and do not require sequential orientation tracking. RR also benefits from scale in most templates, though the effect is weaker and less consistent than in SS. In particular, RR improves substantially in Templates 2–4, but remains relatively flat in Template 1.

The two cross-perspective combinations show divergent scaling behaviour. SR exhibits relatively flat scaling across templates, indicating that increasing model size alone does not substantially improve this transformation. In contrast, RS is highly topology-dependent: it improves sharply in Templates 1 and 3, with the 27B model reaching high accuracy, but remains weak in Templates 2 and 4. This suggests that scaling can alleviate route-to-survey difficulty when the topology supports relatively stable global reconstruction, but does not uniformly solve perspective transformation.

Taken together, these results show that model scale improves spatial reasoning selectively. Within-perspective reasoning generally scales better than cross-perspective reasoning. Larger models are better at learning within-perspective and allocentric reasoning, but cross-perspective transformation remains sensitive to environmental topology. Thus, scaling alone does not consistently produce spatial reasoning behaviour that generalises across topologies and perspectives.

5.4 Reasoning in Closed-Source LLMs

We finally evaluate frontier closed-source LLMs under standard zero-shot, chain-of-thought (CoT) prompting and combination-specific structured CoT prompting, using GPT-5.2 and Gemini-3. CoT prompting is constructed by appending *Let's think step by step by tracking the location and direction changes described in the description* to the standard zero-shot prompt. The structured CoT prompting further gives step-by-step instructions to guide the model to track current position and orientation. We refer readers to Appendix A for the full prompt. Due to API cost constraints, we evaluate CoT and structured CoT only on Templates 2 and 3 using GPT. Results are shown in Table 4.

		SS	SR	RS	RR
Template 1	GPT-zs	0.625	0.492	0.290	0.480
	Gemini-zs	0.900	0.963	0.777	0.887
Template 2	GPT-zs	0.552	0.405	0.223	0.492
	GPT-cot	0.557 (+0.005)	0.405 (0)	0.242 (+0.019)	0.470 (-0.022)
	GPT-scot	0.580 (+0.028)	0.383 (-0.022)	0.195 (-0.028)	0.490 (-0.002)
	Gemini-zs	0.672	0.958	0.583	0.905
Template 3	GPT-zs	0.795	0.520	0.435	0.667
	GPT-cot	0.775 (-0.020)	0.540 (+0.020)	0.465 (+0.030)	0.640 (-0.027)
	GPT-scot	0.775 (-0.020)	0.495 (-0.025)	0.440 (+0.005)	0.637 (-0.030)
	Gemini-zs	0.870	0.963	0.922	0.968
Template 4	GPT-zs	0.625	0.422	0.253	0.475
	Gemini-zs	0.588	0.993	0.605	0.995

Table 4: Zero-shot performance of GPT and Gemini in different templates and description-question combinations. "zs" denotes standard zero-shot prompting, "cot" denotes CoT prompting and "scot" denotes structured CoT prompting. The values in parentheses indicate the performance difference between (structured) CoT prompting and standard zero-shot prompting.

The two models exhibit substantially different spatial reasoning profiles. Overall, GPT shows a behavioural pattern that shares some qualitative similarities with findings reported in human spatial cognition studies: SS consistently achieves the highest performance, followed by RR and SR, while RS remains the most difficult combination. This pattern indicates that GPT performs more consistently on survey-view reasoning, while route-to-survey transformation remains challenging. Across templates, GPT achieves its strongest overall performance on Template 3, again suggesting topological variance.

In contrast, Gemini exhibits a markedly different pattern. The model performs strongly on SR and RR across templates, while SS and RS vary substantially with the environment. Interestingly, SS performance is sometimes even lower than RS in Templates 3 and 4. These results suggest that Gemini may rely more on local directional relations and route-view reasoning strategies, rather than stable global survey representations. Consistent with this interpretation, route-view questions (*i.e.*, SR and RR) remain relatively stable across templates, while survey-view questions (*i.e.*, RS and SS) show much stronger topology dependence.

Finally, CoT and structured CoT prompting provide almost no improvement over standard zero-shot prompting, even occasionally slightly decreases performance. This suggests that more explicit textual decomposition alone is insufficient to reliably improve spatial perspective transformation. Instead, the difficulty may involve maintaining and updating spatial information during reasoning. This interpretation is also consistent with prior MLLM studies showing that explicitly visualising intermediate spatial states can improve spatial reasoning performance, whereas purely textual reasoning chains often fail to provide comparable gains (Wu et al., 2024; Li et al., 2025).

6 Conclusion

We investigate spatial perspective transformation in LLMs using a controlled dataset inspired by human spatial cognition studies. Our experiments show that finetuning and scaling substantially improve within-perspective reasoning, particularly under survey-view settings, while cross-perspective transformation remains considerably more difficult. The learnt capabilities generalize only partially across environment templates and description-question combinations, and remain strongly topology-dependent. Experiments on closed-source LLMs reveal qualitatively different reasoning profiles, while CoT and structured CoT prompting provide little improvement. Together, these findings suggest that current LLMs acquire useful but highly context-dependent spatial reasoning behaviours, which do not yet generalise robustly across perspectives and environments.

Limitations

Our work has several limitations. First, although the proposed templates cover multiple environmental topologies, future work could investigate a broader range of spatial organisations and more complex layouts. Second, our dataset is synthetically generated to ensure controllability and automatic annotation, but the resulting descriptions may

still differ linguistically from naturally occurring spatial descriptions such as navigation instructions or tourist guides. Third, while our task is inspired by classic cognitive psychology experiments, it is not an exact replication of the original human studies. Future work could therefore evaluate human participants directly on our dataset to enable better calibration of task difficulty and provide a more direct behavioural comparison between humans and LLMs. Finally, our dataset focuses on symbolic spatial reasoning from textual descriptions, rather than embodied or interactive spatial grounding, which remains an important direction for future research.

Acknowledgements

The author would like to acknowledge the use of the Computational Shared Facility at The University of Manchester and the Department of Computer Science at the University for funding this research. We also thank the anonymous reviewers from the ARR May 2026 cycle for their valuable feedback.

References

Ruth M. J. Byrne and P. N. Johnson-Laird. 1989. Spatial Reasoning. *Journal of Memory and language*, 28:564–575.

Anthony G. Cohn and Robert E. Blackwell. 2025. Evaluating the ability of large language models to reason about cardinal directions, revisited. *Preprint*, arXiv:2507.12059.

Michel Denis, editor. 2017. *Space and spatial cognition: A multidisciplinary perspective*. Routledge, New York.

Nir Fulman, Abdulkadir Memduhoglu, and Alexander Zipf. 2024. 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. Springer Cham.

Google. 2025. Gemini 3 flash preview.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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

Peiyao Jiang, Zequn Qin, and Xi Li. 2026. Spatial-text: A pure-text cognitive benchmark for spatial understanding in large language models. *Preprint*, arXiv:2603.03002.

Bridget Leonard and Scott O. Murray. 2026. Cognitively-inspired tokens overcome egocentric bias in multimodal models. *Preprint*, arXiv:2601.16378.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *Preprint*, arXiv:2501.07542.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024. 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. AAAI Press, Washington, DC, USA.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *Preprint*, arXiv:2104.05832.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. *Preprint*, arXiv:2210.16952.

OpenAI. 2025. Introducing gpt-5.2.

Sabrina Patania, Luca Annese, Anita Pellegrini, Silvia Serino, Anna Lambiase, Luca Pallonetto, Silvia Rossi, Simone Colombani, Tom Foulsham, Azzurra Ruggeri, and Dimitri Ognibene. 2025. Perspact: Enhancing llm situated collaboration skills through perspective taking and active vision. *Preprint*, arXiv:2511.08098.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

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

Amy L. Shelton and Timothy P. McNamara. 2004. Orientation and Perspective Dependence in Route and Survey Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30:158–170.

Amy L. Shelton and Holly A. Pippitt. 2007. Fixed versus dynamic orientations in environmental learning from ground-level and aerial perspectives. *Psychological Research*, 71:333–346.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. 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. AAAI Press, Palo Alto, California USA.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. *Preprint*, arXiv:2411.06048.

Alexander W. Siegel and Sheldon H. White. 1975. The Development of Spatial Representations of Large-Scale Environments. *Advances in Child Development and Behavior*, 10:9–55.

Ankur Sikarwar, Debangan Mishra, Sudarshan Nikhil, Ponnurangam Kumaraguru, and Aishwarya Agrawal. 2026. Communicating about space: Language-mediated spatial integration across partial views. *Preprint*, arXiv:2603.27183.

Holly A. Taylor and Barbara Tversky. 1992. Spatial mental models derived from survey and route descriptions. *Journal of Memory and language*, 31:261–292.

Barbara Tversky. 1991. Spatial Mental Models. *Psychology of learning and motivation*, 27:109–145.

Barbara Tversky. 1993. Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In *International Conference on Spatial Information Theory (COSIT 1993)*, page 14–24, Marciana Marina, Elba Island, Italy. Springer Berlin, Heidelberg.

David Ed Waller and Lynn Ed Nadel, editors. 2013. *Handbook of spatial cognition*. American Psychological Association.

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

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. 2026b. Mindcube: Spatial mental modeling from limited views. *Preprint*, arXiv:2506.21458.

Steffen Werner, Bernd Krieg-Brückner, Hanspeter A. Mallot, Karin Schweizer, and Christian Freksa. 1997. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation1. In *Informatik '97 Informatik als Innovationsmotor*, pages 41–50, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Prompting Templates

In the following prompt templates, all italicised text denotes placeholders that should be replaced with the corresponding instance-specific content. *area_textual_description* and *question* represents the area description and question in the input, respectively. *additional_information* will be replaced by *Assume that at the start of the description you are facing xxx*[2]*. Use this information to answer the following question.* if the description is in route view and the question is in survey view. Otherwise, no additional information is provided. If the question is in route view, *choices* will be replaced by *Choose answer from left, right, front and back. Only give your answer.* If the question is in survey view, *choices* will be replaced by *Choose answer from north, south, east, west, northeast, northwest, southeast, southwest. Only give your answer.*

Standard Zero-shot Prompt

You will read a description of an area and answer a question.
Description: *area_textual_description*
Question: *additional_information*
question
choices
Answer:

CoT Prompt

You will read a description of an area and answer a question.
Description: *area_textual_description*

[2]One of the four cardinal directions: north, south, east, west.

Question: *additional_information*
question
Let's think step by step by tracking the location and direction changes described in the description.
choices
Answer:

Structured CoT Prompt - SS

You will read a description of an area and answer a question.
Description: *area_textual_description*
Question: *additional_information*
question
Let's think step by step. 1. Extract the global locations and directions of all landmarks from the description. 2. Determine the locations of the two landmarks mentioned in the question in the global coordinate system. 3. Compute the direction between the two landmarks. 4. Check that the answer is expressed using a cardinal or intercardinal direction.
choices
Answer:

Structured CoT Prompt - SR

You will read a description of an area and answer a question.
Description: *area_textual_description*
Question: *additional_information*
question
Let's think step by step. 1. Extract the global locations and directions of all landmarks from the description. 2. Determine the locations of the landmarks mentioned in the question in the global coordinate system. 3. Identify the facing direction specified in the question. 4. Convert the global direction into the corresponding egocentric direction. 5. Check that the answer is expressed relative to the observer.
choices
Answer:

Structured CoT Prompt - RS

You will read a description of an area and answer a question.
Description: *area_textual_description*
Question: *additional_information*
question
Let's think step by step. 1. Begin with the initial facing direction. 2. Track every movement and turn described in the route description, updating the facing direction after each step. 3. Maintain the global locations of all landmarks throughout the traversal. 4. Determine the locations of the two landmarks mentioned in the question in the global coordinate system. 5. Compute the direction between the two landmarks. 6. Check that the answer is expressed using a cardinal or intercardinal direction.
choices
Answer:

Structured CoT Prompt - RR

You will read a description of an area and answer a question.
Description: *area_textual_description*
Question: *additional_information*
question
Let's think step by step. 1. Begin with the initial facing direction. 2. Track every movement and turn described in the route description, updating the facing direction after each step. 3. Identify the facing direction specified in the question. 4. Compute the relative direction of the queried landmark. 5. Check that the answer is expressed relative to the observer.
choices
Answer:

B Data Example for each Template

B.1 Template 1

Route description: You enter Old City Town through Black Avenue. At this point, you see Maple Street on your left, intersecting with Black Avenue. Drive along Black Avenue. Drive for 0.38 kilometre and you will see cafe on your left. Drive for 0.06 kilometre and you will see bank on your right. Drive for 0.06 kilometre and you will see supermarket on your left. You continue driving on Black Avenue and come to, on your left, Cedar Highway. Turning left onto Cedar Highway. Drive for 2.64 kilometre and you will see school on your right. Drive for 1.21 kilometre and you will see cinema on your left. Continue driving on Cedar Highway until you are forced to make a right turn. Turning right onto Madison Road. Drive for 0.40 kilometre and you will see library on your left. Drive for 1.54 kilometre and you will see hospital on your left. Drive for 1.02 kilometre and you will see restaurant on your left. Continue driving along Madison Road until you are again forced to make a right turn. Turning right onto Willow Highway. Drive for 0.29 kilometre and you will see grocery on your right. Drive for 1.65 kilometre and you will see bookstore on your left. Drive for 2.46 kilometre and you will see fountain on your left. Continue driving along Willow Highway until you meet a dead-end into Black Avenue. Turn left

onto Black Avenue. Drive for 0.26 kilometre and you will see statue on your right. Drive for 0.02 kilometre and you will see flower shop on your right. You drive along Black Avenue until you see Silver Bay on your left, intersecting with the one you are on. Greenwood Beach runs parallel to your far left, on the other side of the town.

Survey description: Old City Town is surrounded by four major landmarks: Black Avenue, Maple Street, Silver Bay and Greenwood Beach. The east is made up of the Black Avenue, running north-south. The south is made up of the Maple Street, running west-east. The north is made up of the Silver Bay, running west-east. The west is made up of the Greenwood Beach, running north-south. You enter the town through Black Avenue. Drive north along Black Avenue. Drive for 0.38 kilometre and you will see cafe on the west. Drive for 0.06 kilometre and you will see bank on the east. Drive for 0.06 kilometre and you will see supermarket on the west. You continue driving on Black Avenue and come to, on the west, Cedar Highway. Turning west onto Cedar Highway. Drive for 2.64 kilometre and you will see school on the north. Drive for 1.21 kilometre and you will see cinema on the south. Continue driving on Cedar Highway until you are forced to make a turn to the north. Turning north onto Madison Road. Drive for 0.40 kilometre and you will see library on the west. Drive for 1.54 kilometre and you will see hospital on the west. Drive for 1.02 kilometre and you will see restaurant on the west. Continue driving along Madison Road until you are again forced to make a turn to the east. Turning east onto Willow Highway. Drive for 0.29 kilometre and you will see grocery on the south. Drive for 1.65 kilometre and you will see bookstore on the north. Drive for 2.46 kilometre and you will see fountain on the north. Continue driving along Willow Highway until you meet a dead-end into Black Avenue. Turn north onto Black Avenue. Drive for 0.26 kilometre and you will see statue on the east. Drive for 0.02 kilometre and you will see flower shop on the east. Continue driving north along Black Avenue and you leave the town.

Route question: If you go along Black Avenue with Madison Road on your right, what side of Black Avenue is bank on?

Survey question: In which direction is fountain from Greenwood Beach?

B.2 Template 2

Route description: You enter Highland Town through Cedar Highway. Drive for 0.71 mile and you will see hospital on your left. Drive for 2.14 mile and you will see flower shop on your right. Drive for 0.58 mile and you will see cafe on your right. Drive for 0.90 mile and you will see fountain on your left. Drive for 0.04 mile and you will see town hall on your left. Make a left turn onto Madison Highway. Drive for 2.17 mile and you will see supermarket on your right. Drive for 0.22 mile and you will see restaurant on your right. Drive for 0.23 mile and you will see barbershop on your right. Drive for 0.77 mile and you will see pharmacy on your left. Drive for 1.22 mile and you will see gazebo on your left. Make a left turn onto Lincoln Road. Drive for 1.86 mile and you will see hotel on your right. Drive for 0.76 mile and you will see playground on your right. Drive for 0.10 mile and you will see theatre on your right. Drive for 0.79 mile and you will see statue on your right. Make a left turn onto Madison Avenue. Drive for 0.58 mile and you will see clock tower on your right. Drive for 1.05 mile and you will see toy store on your right. Drive for 1.78 mile and you will see school on your left. You arrive back at your starting point on Cedar Highway, completing the tour.

Survey description: Highland Town is formed by four boundary routes: Lincoln Road, Madison Highway, Cedar Highway, Madison Avenue. The north side is made up of the Lincoln Road, running from west to east. The south side is made up of the Cedar Highway, running from west to east. The west side is made up of the Madison Avenue, running from north to south. The east side is made up of the Madison Highway, running from north to south. You enter the town through Cedar Highway. Drive east along the Cedar Highway. Drive for 0.71 mile and you will see hospital on the north. Drive for 2.14 mile and you will see flower shop on the south. Drive for 0.58 mile and you will see cafe on the south. Drive for 0.90 mile and you will see fountain on the north. Drive for 0.04 mile and you will see town hall on the north. Make a turn to the north onto Madison Highway. Drive for 2.17 mile and you will see supermarket on the east. Drive for 0.22 mile and you will see restaurant on the east. Drive for 0.23 mile and you will see barbershop on the east. Drive for 0.77 mile and you will see pharmacy on the west. Drive for 1.22 mile and you will see gazebo on the west. Make a turn to the

west onto Lincoln Road. Drive for 1.86 mile and you will see hotel on the north. Drive for 0.76 mile and you will see playground on the north. Drive for 0.10 mile and you will see theatre on the north. Drive for 0.79 mile and you will see statue on the north. Make a turn to the south onto Madison Avenue. Drive for 0.58 mile and you will see clock tower on the west. Drive for 1.05 mile and you will see toy store on the west. Drive for 1.78 mile and you will see school on the east. Continue driving until you complete the tour and return to your starting point on Cedar Highway.
Route question: If you go along Madison Avenue with Lincoln Road ahead, what side of Madison Avenue is school on?
Survey question: In which direction is playground from Cedar Highway?

B.3 Template 3

Route description: You enter Silver Town through Lincoln Avenue. Drive along Lincoln Avenue. Drive for 0.30 mile and you will see hotel on your left. Drive for 0.59 mile and you will see clock tower on your left. You continue driving along Lincoln Avenue and come to Oak Highway on your right. Turn right onto Oak Highway. Drive for 0.55 mile and you will see fountain on your left. Drive for 0.37 mile and you will see playground on your right. Drive for 0.60 mile and you will see town hall on your right. Drive for 0.26 mile and you will see supermarket on your left. You reach the end of Oak Highway. Go back to where you leave Lincoln Avenue and turn right back onto Lincoln Avenue. Drive for 0.50 mile and you will see barbershop on your left. You continue along Lincoln Avenue and leave the town.
Survey description: Silver Town consists of two main roads: Lincoln Avenue and Oak Highway. You enter the town via Lincoln Avenue. Drive east along Lincoln Avenue. Drive for 0.30 mile and you will see hotel on the north. Drive for 0.59 mile and you will see clock tower on the north. You come to Oak Highway on the south. Turn south onto Oak Highway. Drive for 0.55 mile and you will see fountain on the east. Drive for 0.37 mile and you will see playground on the west. Drive for 0.60 mile and you will see town hall on the west. Drive for 0.26 mile and you will see supermarket on the east. You reach the end of Oak Highway. Go back to where you leave Lincoln Avenue and turn east back onto Lincoln Avenue. Drive for 0.50 mile and you will see barbershop on the north. Continue driving east and you leave the town via Lincoln Avenue.
Route question: If you go along Lincoln Avenue with Oak Highway on your right, what side of Lincoln Avenue is clock tower on?
Survey question: In which direction is supermarket from fountain?

B.4 Template 4

Route description: You arrive at Black Square. Standing at the centre of Black Square and facing Black Avenue, Willow Highway will be on your left, Oak Avenue will be on your right and Lincoln Road will be on your back. Walk onto Black Avenue. Walk for 0.23 kilometre and you will see pharmacy on your right. Walk for 0.44 kilometre and you will see post office on your left. Walk for 0.00 kilometre and you will see supermarket on your right. Walk for 0.05 kilometre and you will see bookstore on your right. Walk for 0.23 kilometre and you will see parking lot on your right. Return back along Black Avenue to Black Square. Walk onto Lincoln Road. Walk for 0.13 kilometre and you will see theatre on your left. Walk for 0.46 kilometre and you will see gazebo on your right. Walk for 0.22 kilometre and you will see hospital on your left. Walk for 0.05 kilometre and you will see cinema on your left. Return back along Lincoln Road to Black Square. Walk onto Willow Highway. Walk for 0.06 kilometre and you will see cafe on your right. Walk for 0.08 kilometre and you will see bus stop on your right. Walk for 0.62 kilometre and you will see museum on your left. Return back along Willow Highway to Black Square. Walk onto Oak Avenue. Walk for 0.15 kilometre and you will see grocery on your left. Walk for 0.46 kilometre and you will see restaurant on your left. Walk for 0.22 kilometre and you will see bank on your right. Return back along Oak Avenue to Black Square. You finish your exploration at Black Square area.
Survey description: You arrive at Black Square. Extending from it are four roads: Black Avenue to the north, Lincoln Road to the south, Willow Highway to the west, and Oak Avenue to the east. From Black Square, you head north along Black Avenue. Walk for 0.23 kilometre and you will see pharmacy on the east. Walk for 0.44 kilometre and you will see post office on the west. Walk for 0.00 kilometre and you will see supermarket on the east. Walk for 0.05 kilometre and you will see bookstore on the east. Walk for 0.23 kilometre and you will see parking lot on the east. You then

	left	right	front	back
Template 1	0.41/0.38	0.41/0.44	0.08/0.10	0.10/0.08
Template 2	0.30/0.32	0.30/0.30	0.20/0.17	0.20/0.21
Template 3	0.32/0.30	0.32/0.31	0.17/0.19	0.18/0.19
Template 4	0.24/0.22	0.24/0.27	0.26/0.26	0.26/0.25

Table 5: Answer label distributions for route-view questions across templates. Each entry shows the proportion in the training/test set.

return along Black Avenue to Black Square. From Black Square, you head south along Lincoln Road. Walk for 0.13 kilometre and you will see theatre on the east. Walk for 0.46 kilometre and you will see gazebo on the west. Walk for 0.22 kilometre and you will see hospital on the east. Walk for 0.05 kilometre and you will see cinema on the east. You then return along Lincoln Road to Black Square. From Black Square, you head west along Willow Highway. Walk for 0.06 kilometre and you will see cafe on the north. Walk for 0.08 kilometre and you will see bus stop on the north. Walk for 0.62 kilometre and you will see museum on the south. You then return along Willow Highway to Black Square. From Black Square, you head east along Oak Avenue. Walk for 0.15 kilometre and you will see grocery on the north. Walk for 0.46 kilometre and you will see restaurant on the north. Walk for 0.22 kilometre and you will see bank on the south. You then return along Oak Avenue to Black Square. You finish your exploration at Black Square area.
Route question: If supermarket is on your left as you go along Black Avenue, what side of Black Avenue is bookstore on?
Survey question: In which direction is restaurant from Black Avenue?

C Dataset Analysis

The ground-truth label distributions for every topology and train/test split are reported in Table 5 and Table 6 for route-view questions and survey-view questions, respectively. The distributions are not perfectly uniform because QA pairs are generated only when the spatial relation is linguistically natural and unambiguous. Consequently, route-view questions contain more left/right than front/back answers, while survey-view questions contain more cardinal than intercardinal answers. However, no individual answer label is overwhelmingly dominant: left and right are approximately balanced with each other, and the four cardinal directions are also approximately balanced with one another. There is no obvious label distribution shift between training sets and test sets.

D More Experiment Results

D.1 Chance-Normalised Overall Accuracy

The chance-normalised metric $(accuracy - chance)/(1 - chance)$ provides a fairer comparison across the four- and eight-way classification settings, where $chance$ means the accuracy of random guessing (0.25 for route-view questions and 0.125 for survey-view questions). For conciseness, we recompute the finetuned results in Table 1 and report them in Table 7. Compared with the original accuracy, the main relative performance patterns remain largely unchanged.

D.2 Generalisation Results with LLaMA

we replicated the cross-template (Section 5.2.1) and cross-combination (Section 5.2.2) experiments with LLaMA using the same setup as Qwen. Table 8 shows the cross-template result, and Table 9 shows the cross-combination result. Results are largely consistent between two models, indicating that the observed patterns are robust across examined model families.

E Model Details

The open-source models are loaded from Hugging Face using the following checkpoints: LLaMA-3.1 8B (`meta-llama/Llama-3.1-8B-Instruct`) and Qwen-3.5 9B (`Qwen/Qwen3.5-9B`). For the scalability experiments, we use Qwen-3.5 0.8B (`Qwen/Qwen3.5-0.8B`), Qwen-3.5 4B (`Qwen/Qwen3.5-4B`) and Qwen-3.5 27B (`Qwen/Qwen3.5-27B`). GPT-5.2 and Gemini-3 are accessed through their official APIs in March 2026.

F Artifact Usage

The open-source models used in this work are subject to their respective licenses and intended research usage policies. Closed-source models are accessed through official APIs in accordance with their terms of service. All datasets used in this work are synthetically generated and contain no personal information, identifiable individuals, or offensive content. The datasets and code will be released for research purposes only.

	north	northeast	east	southeast	south	southwest	west	northwest
Template 1	0.17/0.15	0.07/0.07	0.17/0.21	0.08/0.09	0.17/0.19	0.07/0.07	0.18/0.15	0.09/0.07
Template 2	0.19/0.20	0.05/0.05	0.19/0.20	0.06/0.05	0.20/0.17	0.06/0.07	0.20/0.23	0.05/0.04
Template 3	0.17/0.15	0.10/0.07	0.16/0.17	0.08/0.10	0.14/0.18	0.10/0.08	0.16/0.14	0.09/0.11
Template 4	0.20/0.17	0.05/0.07	0.19/0.20	0.05/0.07	0.19/0.20	0.06/0.04	0.20/0.20	0.05/0.05

Table 6: Answer label distributions for survey-view questions across templates. Each entry shows the proportion in the training/test set.

		SS	**SR**	**RS**	**RR**
Template 1	LLaMA	0.905±0.162	0.380±0.020	0.464±0.027	0.501±0.070
	Qwen	0.980±0.025	0.364±0.013	0.418±0.016	0.392±0.014
Template 2	LLaMA	0.940±0.008	0.432±0.005	0.394±0.191	0.566±0.004
	Qwen	0.947±0.014	0.382±0.042	0.229±0.067	0.561±0.010
Template 3	LLaMA	0.997±0.003	0.502±0.017	0.410±0.044	0.989±0.004
	Qwen	0.998±0.002	0.506±0.017	0.489±0.141	0.997±0.006
Template 4	LLaMA	0.999±0.002	0.464±0.018	0.572±0.113	1.000±0.000
	Qwen	0.996±0.007	0.482±0.013	0.182±0.036	0.996±0.005

Table 7: Chance-normalised accuracy of finetuned open-source models. Values are averaged over three runs and reported as mean ± standard deviation.

	test / finetune	**Template 1**	**Template 2**	**Template 3**	**Template 4**
SS	**Template 1**	0.917±0.142	0.770±0.083	0.622±0.156	0.723±0.134
	Template 2	0.689±0.009	0.948±0.007	0.557±0.012	0.703±0.087
	Template 3	0.438±0.047	0.447±0.041	0.998±0.003	0.487±0.173
	Template 4	0.419±0.040	0.367±0.040	0.572±0.078	0.999±0.001
SR	**Template 1**	0.535±0.015	0.347±0.013	0.478±0.026	0.286±0.018
	Template 2	0.408±0.034	0.574±0.004	0.441±0.012	0.435±0.089
	Template 3	0.439±0.009	0.422±0.013	0.627±0.013	0.527±0.038
	Template 4	0.345±0.007	0.431±0.008	0.497±0.013	0.598±0.014
RS	**Template 1**	0.531±0.024	0.198±0.063	0.379±0.051	0.187±0.014
	Template 2	0.273±0.052	0.470±0.168	0.282±0.078	0.224±0.031
	Template 3	0.264±0.006	0.236±0.071	0.484±0.038	0.208±0.028
	Template 4	0.210±0.040	0.218±0.031	0.200±0.050	0.626±0.099
RR	**Template 1**	0.626±0.064	0.357±0.013	0.437±0.021	0.278±0.004
	Template 2	0.419±0.013	0.674±0.003	0.445±0.011	0.512±0.019
	Template 3	0.469±0.033	0.432±0.031	0.992±0.002	0.652±0.170
	Template 4	0.416±0.018	0.506±0.017	0.826±0.053	1.000±0.000

Table 8: Cross-template generalisation results on LLaMA. Each row indicates the template used for finetuning, while each column indicates the template used for evaluation. Values are averaged over three runs and reported as mean ± standard deviation.

	RR-to-SR	**SR-to-RR**	**RS-to-SS**	**SS-to-RS**
Template 1	0.502 (-0.124)	0.505 (-0.030)	0.263 (-0.268)	0.283 (-0.634)
Template 2	0.518 (-0.156)	0.507 (-0.067)	0.244 (-0.226)	0.264 (-0.684)
Template 3	0.611 (-0.381)	0.616 (-0.011)	0.364 (-0.120)	0.334 (-0.664)
Template 4	0.601 (-0.399)	0.587 (-0.011)	0.221 (-0.405)	0.315 (-0.684)

Table 9: Cross-combination generalisation results on LLaMA. Each column "A-to-B" denotes finetuning on combination A and evaluating on combination B. The values in parentheses indicate the performance difference between the accuracy on B and the original in-domain accuracy on A. All values are averaged over three runs.

G Accuracy Breakdown Based on Directions

We compute accuracy on each direction for each description-question combination in all templates on both LLaMA and Qwen. Different patterns are observed between different templates and models. Results are shown in Figure 3-6.

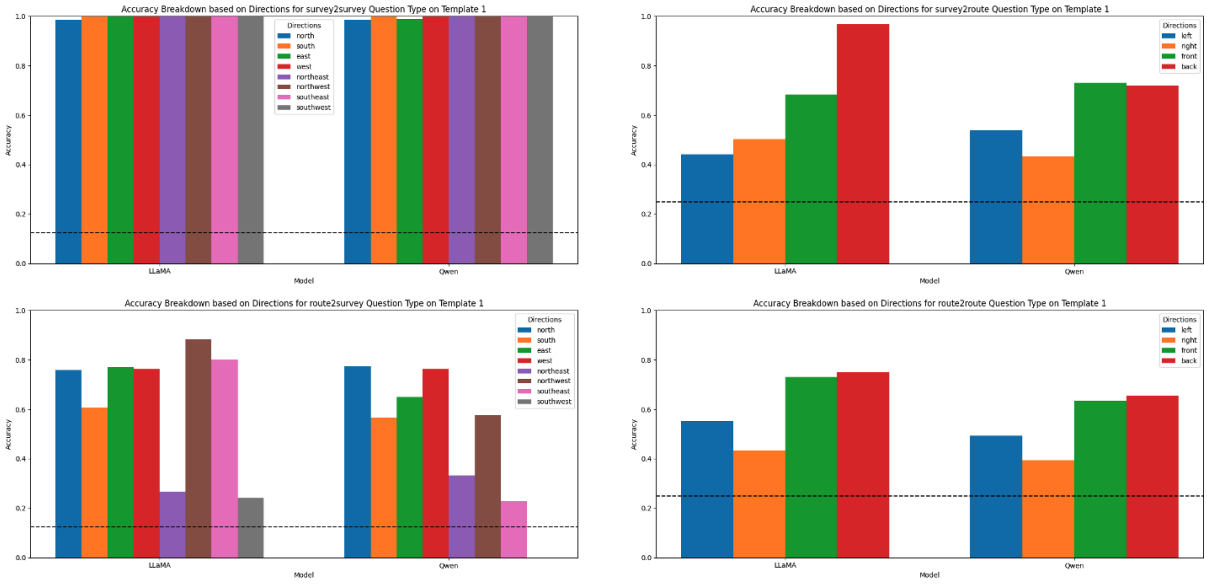


Figure 3: Accuracy breakdown based on directions for different description-question combination in Template 1 on LLaMA and Qwen. The black dashed line indicates the random-guessing accuracy.

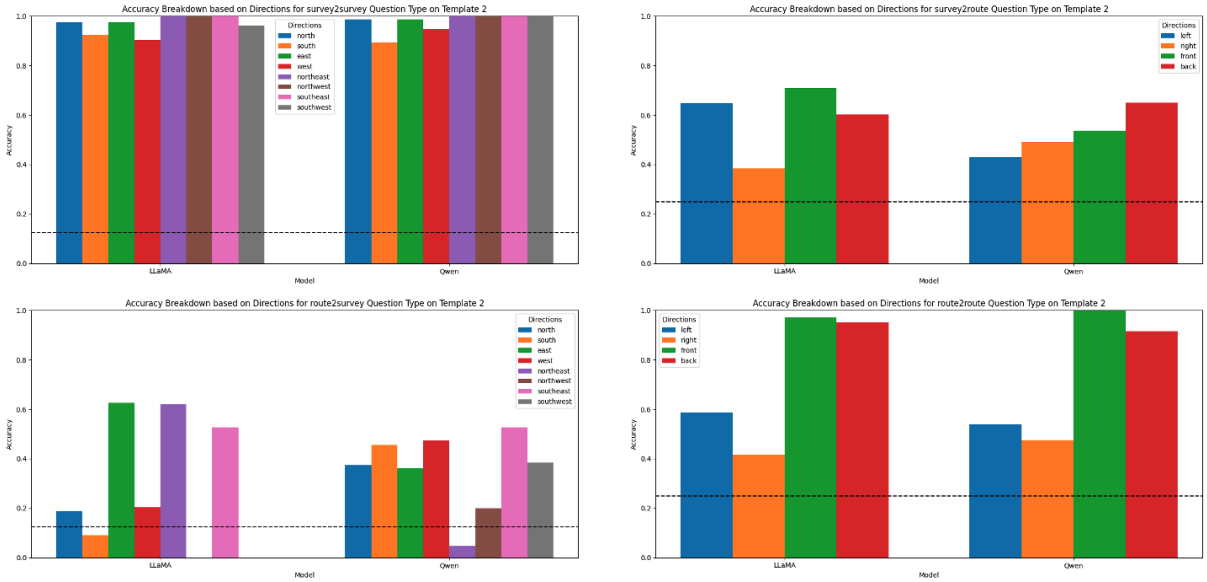


Figure 4: Accuracy breakdown based on directions for different description-question combination in Template 2 on LLaMA and Qwen. The black dashed line indicates the random-guessing accuracy.

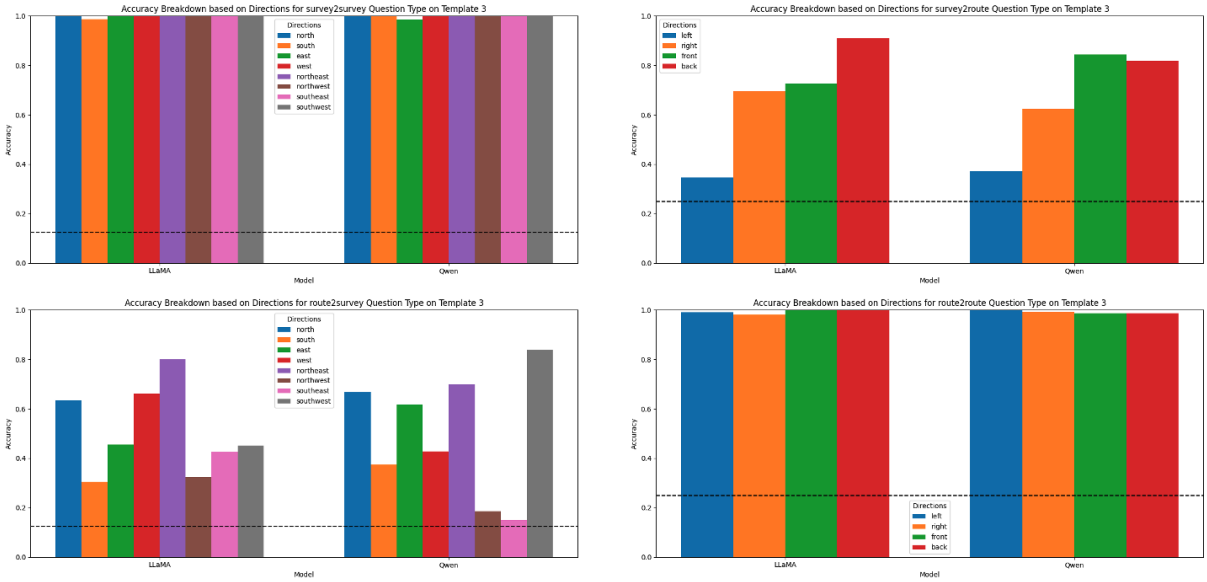


Figure 5: Accuracy breakdown based on directions for different description-question combination in Template 3 on LLaMA and Qwen. The black dashed line indicates the random-guessing accuracy.

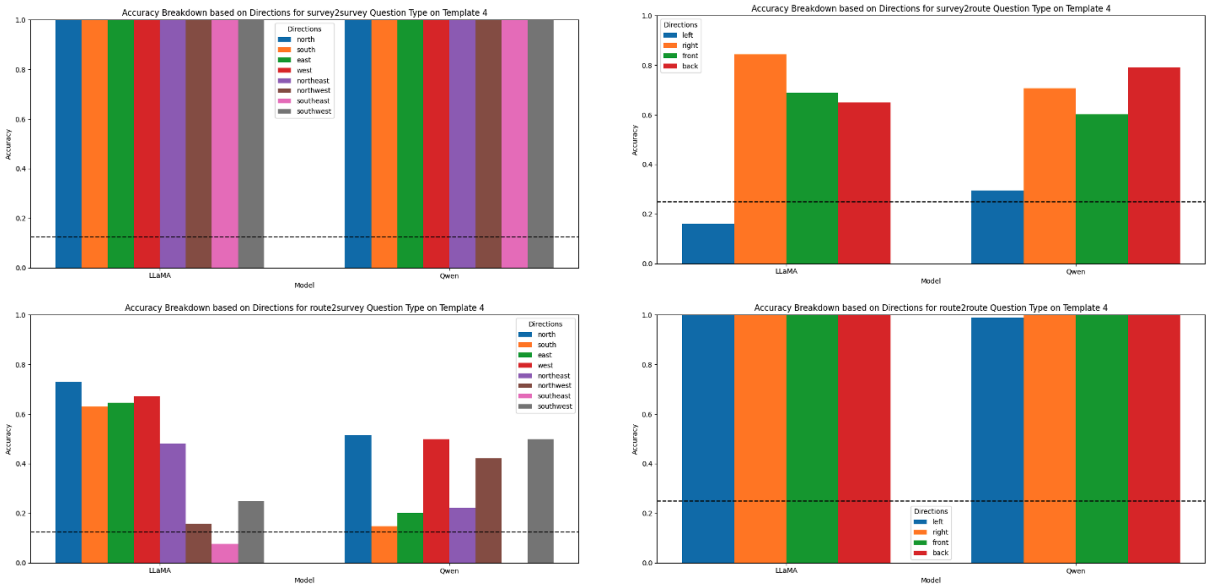


Figure 6: Accuracy breakdown based on directions for different description-question combination in Template 4 on LLaMA and Qwen. The black dashed line indicates the random-guessing accuracy.