Overview of CCL25-Eval Task 1: The Fifth Spatial Cognition Evaluation (SpaCE2025)

Yuhang Qin[1] Liming Xiao[1] Nan Hu[1] Sirui Deng[1] Jingyuan Ma[2] Hyang Cui[1]

Zihan Zhang[1] Chihsu Tsai[1] Jingkun Ding[1] Sumin Kang[1] Zhifang Sui[2,3] Weidong Zhan[1,3,4,*]

[1]Department of Chinese Language and Literature, Peking University
[2]School of Computer Science, Peking University
[3]National Key Laboratory for Multimedia Information Processing, Peking University
[4]Center for Chinese Linguistics, Peking University
hezonglianheng@stu.pku.edu.cn

Abstract

The Fifth Spatial Cognition Evaluation (SpaCE2025) presents a benchmark aimed at evaluating the spatial semantic understanding and reasoning capabilities of Large Language Models (LLMs), primarily in Chinese. It consists of five subtasks: (1) Retrieving Spatial Referents (RSR), (2) Detecting Spatial Semantic Anomalies (DSA), (3) Recognizing Synonymous Spatial Expression (RSE), (4) Spatial Position Reasoning (SPR) in Chinese, and (5) SPR in English. The fourth and fifth subtask share the same content and structure, differing only in language, and are designed to assess the cross-linguistic spatial reasoning capability of LLMs. A total of 12 teams submitted their final results, and the best-performing team achieved an accuracy of 0.7931. The results suggest that while LLMs are capable of handling basic spatial semantic understanding tasks such as RSR, their performance on more complex tasks, such as DSA and RSE, still requires improvement. Additionally, finetuning methods that effectively activate LLMs' reasoning ability are essential to improve their performance.

Keywords: evaluation benchmark , Large Language Models , spatial semantic understanding , spatial reasoning , synthetic data

1 Introduction

Spatial expressions are not only common phenomena in natural languages, but also reflect how humans conceptualize the world (Talmy, 1983). Thus, understanding spatial semantics, an important topic in cognitive linguistics, is an essential component of language ability. To evaluate machines' spatial semantic understanding capability, several tasks and benchmarks have been proposed, including tasks on spatial role labeling (Pustejovsky et al., 2015; Kordjamshidi et al., 2017) and spatial reasoning (Weston et al., 2015; Mirzaee et al., 2021; Mirzaee and Kordjamshidi, 2022; Shi et al., 2022; Li et al., 2024). In Chinese, Spatial Cognition Evaluation (SpaCE), which has been held for 4 years, offers a series of datasets designed to comprehensively evaluate models' performance in multiple NLU and reasoning tasks (詹卫东et al., 2022; Xiao et al., 2023a; Xiao et al., 2023b; Xiao et al., 2024). According to recent studies, machine performance on spatial semantic understanding capability lags significantly behind humans capabilities. Thus, spatial semantic understanding continues to be challenging for NLP systems, including large language models (LLMs).

In this work, we introduce a multitask benchmark to assess spatial understanding ability of LLMs. This benchmark consists of five subtasks: (1) **R**etrieving **S**patial **R**eferents (RSR), (2) **D**etecting **S**patial semantic **A**nomalies (DSA), (3) **R**ecognizing **S**ynonymous spatial **E**xpression with different forms (RSE), (4) **S**patial **P**osition **R**easoning (SPR) in Chinese, and (5) Spatial Position Reasoning in English. Subtasks (1), (2) and (3) mainly focus on LLMs' spatial language ability, and (4) and (5) primarily concerntrate on their spatial reasoning ability of LLMs. Our dataset contains 18,423 questions, which are available at

Subtask	Demo set	Train set	Dev set	Test set	Total
DSA	20	0	0	3,500	3,520
RSR	20	0	0	1,763	1,783
RSE	20	0	0	1,100	1,120
SPR(Chinese)	0	2,000	500	3,500	6,000
SPR(English)	0	2,000	500	3,500	6,000
Total	60	4,000	1,000	13,363	18,423

Table 1: Scale of Dataset of each subtask. Demo set is a set that provides some examples of the subtask, and is used to help participants understand the subtask.

https://github.com/PKU-SpaCE/SpaCE2025/tree/main/data. The scale and distribution of data across subtasks is presented in Table 1, and more detailed statistics are presented in Appendix A. In addition to maintaining the features of SpaCE2024 (Xiao et al., 2024), some novel characteristics are introduced in this benchmark.

- **The reasoning benchmark is bilingual.** Compared to existing reasoning benchmarks, Chinese as well as English questions are provided, which can assess the spatial reasoning ability of models in different languages.
- **The benchmark focuses on high-difficulty tasks.** Previous research demonstrated that LLM performance on tasks relying heavily on textual patterns, such as spatial role labeling, is relatively close to the human level. However, their performance on tasks requiring higher cognitive ability is still below the human level. Thus, this benchmark mainly focuses on complicated tasks that go beyond forms on surface, including referent resolution, scene construction, scene comparison, and spatial reasoning.
- **The data are diverse and statistically balanced.** Our benchmark contains spatial expressions that have not appeared in previous SpaCE benchmarks, with the aim of evaluating the spatial cognitive ability of models comprehensively and precisely. In addition, to ensure that the result is statistically significant, the scale of our dataset is increased, and the number of each kind of question is balanced.

2 SpaCE2025 Task Overview

2.1 Retrieving Spatial Referents (RSR)

In Modern Chinese, spatial expressions typically anchor an entity's location using a localizer relative to a referent, such as 树下面(*the area under the tree*) and 剧院里面(*the area inside the theater*). However, the referent noun can be omitted if it is contextually evident. Consider Figure 1a, where the localizer 里面(*inside*) appears without an explicit referent. The context makes it clear that the wooden partition serves as the referent, dividing the floor into an "inside" and an "outside". Resolving such cases requires LLMs to engage in contextual understanding, commonsense knowledge, and spatial cognition. This makes the RSR task effective for evaluating the spatial semantic understanding of LLMs.

In each RSR question, LLMs are given a text containing an omitted referent and a proposed interpretation that specifies a potential referent. LLMs must then judge the correctness of this interpretation. For example, based on our analysis above, Interpretation 1 in Figure 1a is correct, while Interpretation 2 is incorrect, even though the spatial relationship "the teacher's dormitory is inside the Tujia building" is factually true.

2.2 Detecting Spatial Semantic Anomalies (DSA)

The ability to detect semantic anomalies within a text, especially those involving spatial concepts such as *up*, *down*, *left*, *right*, *come*, and *go*, is a critical indicator of deep language comprehension. To evaluate

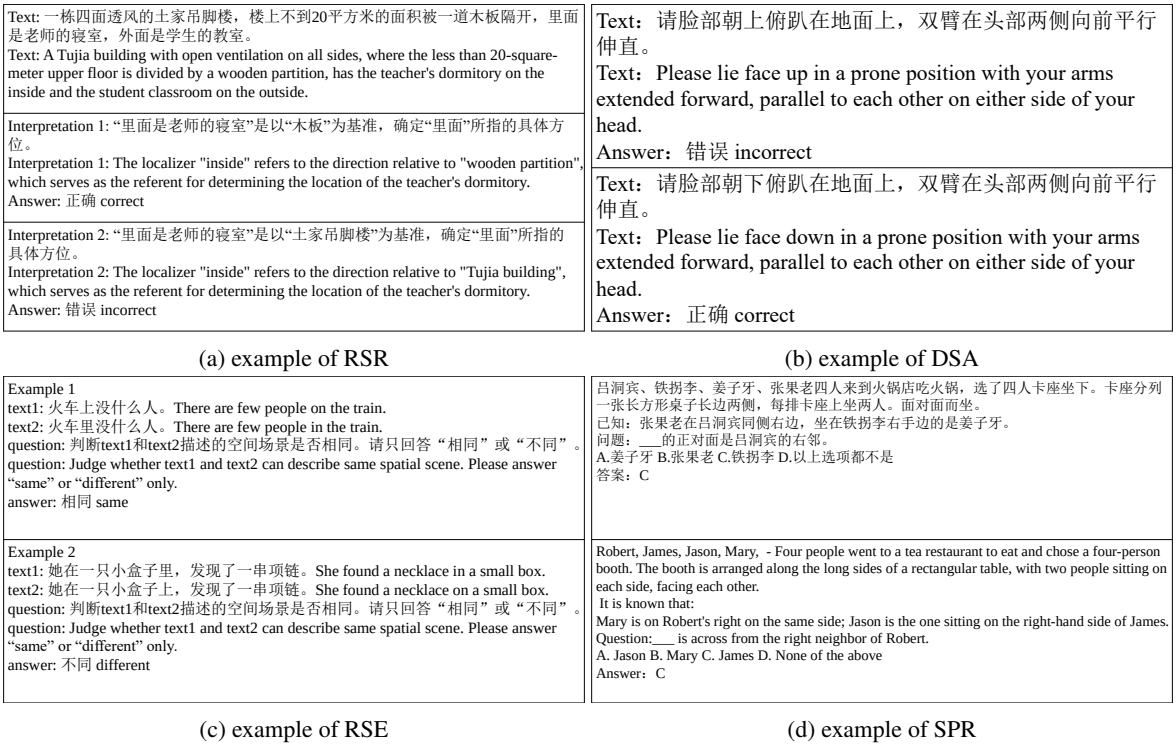


Figure 1: Illustrative examples of the RSR, DSA, RSE, and SPR subtasks. For RSR, DSA, and RSE, tasks are presented to the models exclusively in Chinese; the accompanying English translations are included solely for readability. The SPR task is bilingual, with questions formulated in both Chinese and English. To ensure linguistic fluency, entity names are idiomatic to each language and are therefore not literal translations of one another.

the spatial semantic understanding of LLMs, we introduce the DSA task. In this diagnostic task, LLMs are given a sentence and must determine if its spatial expressions are semantically coherent. Consider the example in Figure 1b, the sentence *Please lie face up in a prone position* contains a semantic anomaly, because the phrase *face up* contradicts the physical requirement of *a prone position*. A coherent phrasing would be *lie face down in a prone position*.

2.3 Recognizing Synonymous Spatial Expression (RSE)

In Modern Chinese, different localizers typically represent distinct meanings, as seen in 他站在桥上(*He stands on the bridge*) versus 他站在桥下(*He stands under the bridge*). However, specific contexts can render expressions with different localizers spatially synonymous, which means that they describe the same spatial scene (詹卫东et al., 2024). For example, with the verb 倚(*lean*), the localizers 上(*on*) and 下(*under*) can cease to be antonymous, with 他倚在树上 and 他倚在树下 both meaning *He leans against the tree*, and the spatial scenes depicted are considered equivalent.

Discerning this synonymy requires nuanced comprehension and comparison of the underlying spatial scenes, drawing upon both commonsense and spatial knowledge. To test this advanced cognitive ability, we introduce RSE, a natural language understanding task where LLMs must determine if a pair of sentences represents the same spatial scene. For instance, in Example 1 of Figure 1c, both sentences describe people inside a train, so the answer is “same”. In contrast, in Example 2, the necklaces are in different locations (inside vs. on the box), so the judgement is “different”.

2.4 Spatial Position Reasoning (SPR)

Spatial reasoning is a cognitive process that involves the construction of mental representations of spatial entities, relations, and transformations, which is fundamental to spatial semantic understanding

(Clements and Battista, 1992). Our SPR task evaluates this skill using automatically generated texts based on preset spatial scenarios. Appendix B shows four types of predefined spatial scenario schema, each containing a certain number of entities and the spatial relations between them. In each SPR question, some of the spatial relations are provided by the text, leaving the others to be inferred through deduction. These scenarios allow for a robust evaluation of LLMs' ability to comprehend diverse spatial relations, including synonymous expressions for the same relation, within a unified structure.

The example in Figure 1d instantiates the Four-people Booth scenario. After deducing the complete seating arrangement of the four people based on provided clues, the people opposite to the one to the right of 吕洞宾(Robert), (C) 铁拐李(James), can be found.

3 Dataset Construction

Due to the disparate characteristics of subtasks, the datasets were constructed using different strategies. Datasets for evaluating spatial language ability relies mainly on human annotation given the high cognitive demands. However, manually generating evaluation questions for spatial reasoning ability poses significant challenges. To ensure both the quantity and quality of these questions, an automatic method was proposed in the generation procedure. The detailed data construction pipelines for each type of subtask are outlined below.

3.1 Crowdsourced Spatial Language Datasets with Lexical Variation

Each item of the spatial language ability evaluation data was formulated as a binary judgment question. DSA questions includes 4 fields: id, instruction, text, and answer. In RSR questions, another field, interpretation, is included. In RSE, where comparison of meanings requires sentence pairs, the text field is divided into text1 and text2. The instruction for each subtask is fixed (see Appendix C). The contents of other fields are discribed in the following sections on question construction.

3.1.1 Lexical Replacement for Variant Generation

Data for DSA and RSE were generated by placing one or more words in the original sentence. In Chinese, spatial information is usually conveyed by localizers or directional verbs. Thus, we first constructed a spatial vocabulary comprising words expressing spatial relations. Words sharing similar syntactic functions and syllable counts were grouped together, and spatial terms in the texts were systematically replaced with alternatives from the same group to generate candidate sentences.

Then, human annotators verified whether the new sentences were grammatically and semantically correct. Only data that received consistent annotations from at least two annotators were considered valid. The sentence was used as the text field of DSA if it contains any errors or conflicts. Otherwise, we regarded the original and the modified sentence were treated as an RSE data, and assigned to the text1 and text2 field respectively.

In addition, to expand the RSE dataset, annotators were asked to construct new sentence pairs based on given word pairs.

3.1.2 Multi-Rater Annotation and Consensus Filtering

Since 2022, about 30 students majoring in linguistics have been recruited each year to annotate the data. These annotators were instructed not only to check grammatical and semantical correctness when replacing the words or creating new RSE materials, but also to annotate additional labels. Only data that received consistent annotations from at least two annotators were considered valid.

In RSR, an interpretation comprises three elements: a specific spatial expression extracted from the text, an entity from the context, and a localizer within the expression. For example, in the Interpretation 1 shown in Figure 1a, "里面是老师的寝室" is the spatial expression, "木板"(*wooden partition*) is the entity, and "里面"(*inside*) is the localizer. Annotators were asked to judge whether an interpretation accurately describes the spatial information in the text. If so, the answer is marked as 正确(correct); otherwise, it is 不正确(incorrect). Annotators could also propose alternative entities from the context that they considered more appropriate as the correct or incorrect referents for the given spatial expression.

For RSE questions, annotators were asked to judge whether a sentence pair has the same meaning in terms of spatial expressions by their intuition. Then, 相同(same) or 不同(different) was recorded as the answer to the questions depending ot the annotators' judgement.

3.2 Knowledge-Driven Generation of Spatial Reasoning Questions

Each item of data from subtasks of spatial reasoning ability evaluation datasets is a four-choice question, consisting of following fields: id, instruction, text, question, and options. The instruction for each subtask is fixed (see Appendix C). The datasets are generated using an automated framework that leverages a structured knowledge base, and consists of three core modules. This pipeline is used to generate synthetic data for the SPR task. Details of the pipeline are discussed in Appendix D.

4 Evaluation Metrics and Overall Results

In our benchmark, accuracy is adopted as the primary metric for evaluation and ranking. The spatial language ability evaluation score $S_{language}$ and the spatial reasoning ability evaluation score $S_{reasoning}$ are computed independently, and the overall score S is obtained by averaging the two. The score of each part is the mean accuracy of each subtask. The scoring fomulae are shown below.

$$S = \frac{S_{language} + S_{reasoning}}{2} \tag{1}$$

$$S_{language} = \frac{1}{3}\sum_{i=1}^{3} Acc_i \tag{2}$$

$$S_{reasoning} = \frac{1}{2}\sum_{i=1}^{2} Acc_i \tag{3}$$

$$Acc = \frac{\#correct}{\#total} \tag{4}$$

38 teams participated in the evaluation, among which 12 teams submitted their final results. In our evaluation, participants were required to either use or finetune LLMs with fewer than 7B parameters for the spatial language ability evaluation. For the spatial reasoning ability evaluation they were only required to finetune DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025). In addition, we developed a baseline using DeepSeek-R1-Distill-Qwen-7B in zero-shot mode across all subtasks. The scores are presented in Table 2. Moreover, for each subtask, a subset of test set was extracted to assess human performance. Appoximately 10 human participants completed the test for each subtask. Table 3 reports the average human accuracy alongside the performance of the top six teams on the human-evaluated subset.

5 Analysis

5.1 Performance Overview and LLM-Human Comparison

From the overall results illustrated in Table 2, All 12 teams exceed our baseline in total score. Furthermore,

- Without finetuning, LLM performance on basic spatial language tasks surpasses that on spatial reasoning tasks, which represents a more advanced capability requiring improvement.

- Complexity of the language task significantly impacts LLM performance. The subtasks from the spatial language ability evaluation are on different difficulty levels (see 5.2.3). LLMs handle the simplest task (RSR) effectively, but encounter difficulties with the more complicated ones (DSA and RSE).

Team	Language ability task				Reasoning ability task			Total
	DSA	RSR	RSE	Total	SPR(Chinese)	SPR(English)	Total	
UPC	**0.7089**	**0.8491**	0.6600	**0.7393**	**0.8686**	**0.8251**	**0.8469**	**0.7931**
SHU	0.6454	0.7720	0.7082	0.7085	0.6254	0.5997	0.6126	0.6606
HYQ	0.6911	0.8259	0.6827	0.7332	0.6914	0.3766	0.5340	0.6336
PAIC	0.6766	0.7992	0.6527	0.7095	0.5106	0.5263	0.5184	0.6140
ZZU1	0.6889	0.7856	**0.7200**	0.7315	0.4694	0.4609	0.4651	0.5983
ZZU2	0.6626	0.8100	0.6973	0.7233	0.4446	0.4503	0.4474	0.5854
PKU	0.6637	0.8332	0.6355	0.7108	0.4431	0.4703	0.4567	0.5838
CPIC	0.6994	0.8168	0.6636	0.7266	0.4349	0.4283	0.4316	0.5791
DJT	0.6829	0.8168	0.6336	0.7111	0.4329	0.4374	0.4351	0.5731
BIT	0.6017	0.7079	0.6173	0.6423	0.3151	0.2714	0.2933	0.4678
SAU	0.5177	0.5394	0.5309	0.5293	0.3734	0.3986	0.3860	0.4577
CCNU	0.6003	0.7317	0.6336	0.6552	0.2426	0.2034	0.2230	0.4391
Baseline	0.6274	0.6375	0.5809	0.6153	0.2266	0.2977	0.2622	0.4388

Table 2: Scores of 12 teams who submitted the final results and our baseline. The highest scores of each evaluation part and each subtask are **bolded**.

- In SPR, due to the difference in prompts and strategies, performance varies significantly across different teams. This suggests that LLM performance in SPR is heavily influenced by techniques or prompts capable of activating the reasoning ability of LLMs, which is detailed in Section 6. Additionally, no significant difference is in performance between Chinese and English SPR tasks, indicating that LLMs can handle spatial reasoning tasks in both languages.

Compared to human performance, none of the team exceeds the average human accuracy. While LLM performance is close to human level in RSR, noticeable gaps remains in other subtasks, suggesting that LLMs' cognitive capacity of spatial understanding still require improvement. In addition, high level of human accuracy confirms the high quality of our dataset. Interestingly, human accuracy in the spatial language ability evaluation is lower than that in the spatial reasoning ability evaluation. This is likely due to the greater vagueness in semantic understanding task, making it harder for human participants to be consistent.

5.2 Systematic Errors and Cross-Task Correlation

5.2.1 Bias towards Typical Situations

LLMs exhibit systematic bias when predicting answers. In the spatial language ability evaluation, they tend to give specific answers compared to human participants. Table 4 shows the average accuracy of the top 6 teams on questions with different answers in DSA, RSR, and RSE. The results show that LLMs tend to consider that no error is in the sentence in DSA. And in RSR, LLMs tend to assume that the interpretations provided in the questions are correct. However, in RSE, they tend to assume that the meanings of the sentences are different.

These performance asymmetries are likely relevant to the training of LLMs. Most LLMs are trained using next-token prediction on large-scale natural language corpora, where sentences containing semantic errors or synonymous are uncommon. Consequently, the LLMs tend to develop a bias toward assuming correctness. In addition, instruction tuning and alignment phases rarely contain task-related examples, especially those involving errors or similar meanings, which may further limit their sensitivity to such distinctions. These findings suggests that although LLMs achieve surface-level linguistic fluency, they are still struggled with deeper spatial-semantic understanding. Addressing these limitations will be crucial for deploying LLMs in applications that requires physical reasoning and commonsense validation.

Team	Language ability task			Reasoning ability task	
	DSA	RSR	RSE	SPR(Chinese)	SPR(English)
UPC	0.62	0.85	0.70	0.87	0.93
SHU	0.51	0.72	0.72	0.63	0.53
HYQ	0.63	0.83	0.75	0.73	0.43
PAIC	0.59	0.80	0.71	0.47	0.43
ZZU1	0.64	0.81	0.74	0.43	0.43
ZZU2	0.60	0.79	0.73	0.47	0.43
Average	0.60	0.80	0.73	0.60	0.53
Human	**0.82**	**0.85**	**0.89**	**0.97**	**0.93**

Table 3: Accuracy of the top 6 teams and human participants in human-evaluated subset. *Average* refers to average performance of LLMs, and *Human* to the average performance of humans. For subtasks, the highest results are **bolded**.

Label	DSA	RSR	RSE
correct/same	0.88	0.90	0.55
incorrect/different	0.33	0.79	0.82
total	0.68	0.81	0.69

Table 4: Average accuracy of the top 6 teams on questions with different answers in DSA, RSR, and RSE. The results show that LLMs tend to consider that no error is in the sentence in DSA. And in RSR, LLMs tend to assume that the interpretations provided in the questions are correct. However, in RSE, they tend to assume that the meanings of the sentences are different.

In SPR, each question in the datasets is a four-choice question and between one and four of them is correct. LLM performance is influenced by numbers of answers, as shown in Table 5. Noteably, questions with three or more answers is challenging for LLMs. It implies that the structural design of the questions plays an significant role on the LLM performance.

5.2.2 Low Cross-Task and Cross-Language Correlation

To comprehensively evaluate LLMs' capability to understand spatial semantics, texts of some questions are designed to be related questions. In the spatial language ability evaluation, some texts were shared across subtasks. And in SPR, Chinese and English questions are paired since their texts describe the same scenario, sharing the same relations.

The distribution of questions sharing the same texts in the subtasks of spatial language ability evaluation are shown in Table 6. These questions are grouped by the texts they share. We calculated the Spearman correlation between the accuracy of these questions in different subtasks, shown in Table 7. The results show that the accuracy of questions sharing the same texts in different subtasks is appoximately not correlated. This suggests that the abilities used by LLMs to solve these questions are different, and there may not be a holistic spatial language ability of LLMs.

The Spearman correlation of questions describing the same spatial scenario in SPR is shown in Table 7. Similar to the results in spatial language ability evaluation, the accuracy of Chinese and English questions with the same spatial scenario is weakly correlated. This suggests that though LLMs performed well in SPR, the ability to understand spatial semantics in Chinese and English is not the same, and the ability in a certain language can not be transplanted to another simply. However, due to the method adopted by ZZU1 team which is more focused on activating the reasoning ability of LLM, the accuracy has a relatively strong correlation.

Team	SPR(Chinese)				SPR(English)			
	1	2	3	4	1	2	3	4
UPC	0.87	0.91	0.67	0.17	0.81	0.88	0.70	0.12
SHU	0.62	0.67	0.42	0.44	0.59	0.65	0.35	0.27
HYQ	0.70	0.71	0.56	0.02	0.68	0.00	0.00	0.00
PAIC	0.54	0.54	0.10	0.00	0.56	0.54	0.27	0.05
ZZU1	0.49	0.51	0.09	0.17	0.49	0.49	0.02	0.39
ZZU2	0.45	0.48	0.10	0.56	0.47	0.47	0.10	0.66

Table 5: Performances of the top 6 team on SPR questions with different number of answers. The result suggests that it is hard for LLMs to deal with questions with multiple answers.

Related Subtasks	DSA	RSR	RSE
DSA-RSE	1779	–	973
RSR-RSE	–	452	112

Table 6: Numbers of questions sharing the same texts in the subtasks of spatial language ability evaluation. DSA-RSE means that the texts of DSA questions also appeared in RSE, and RSR-RSE means that the texts of RSR questions also appeared in RSE.

Team	DSA-RSE		RSR-RSE		SPR(Chinese)-SPR(English)	
	ρ	p	ρ	p	ρ	p
UPC	0.046	0.193	-0.009	0.922	0.467	*
SHU	0.037	0.291	-0.077	0.423	0.505	*
HYQ	0.045	0.203	-0.029	0.760	0.257	*
PAIC	0.013	0.716	0.180	0.058	0.301	*
ZZU1	0.009	0.804	0.143	0.133	0.641	*
ZZU2	0.127	*	0.089	0.352	0.577	*
Average	0.090	0.011	0.132	0.164	0.480	*

Table 7: Spearman correlation between the accuracy of the related questions. ρ is the Spearman correlation coefficient, and p is the p-value. * means that the absolute value is between 0.001 and -0.001 for the Spearman correlation, and the p-value is less than 0.001.

5.2.3 A Difficulty Gradient of Subtasks

According to the data illustrated in Table 2, the accuracy scores of the subtasks reveal the following trend: RSR>RSE>DSA>SPR. This difficulty trend is distinct with the trend shown in SpaCE2024.

Difficulties of the subtasks are influenced by the length of contexts that need to be concerntrated on. In SpaCE2024, RSE is thought harder than DSA due to the more advanced cognitive capability related to the task. However, in RSR, LLMs only need to focus on a certain spatial expression of the sentences. In RSE, the expression along with related contexts need to be considered. And in DSA, multiple spatial expressions in the sentence need to be compared, for which DSA is the most difficult.

5.3 Subtask-Specific Analysis

5.3.1 DSA - Lexical Errors Easy and Discourse Confilct Hard

To better understand the performance of the LLMs on spatial descriptions which is labeled as incorrect. To further investigate model behavior, we divided the incorrect-labeled samples into two semantic categories:

1. **Spatial Language Expression Errors**: These include incorrect or awkward spatial terms that are semantically inappropriate or physically implausible in context, such as 月亮里的村庄(**a village inside the moon**). The model performed relatively well on this category, achieving an accuracy of 0.71.

2. **Spatial Logical Reasoning Errors**: These involve inconsistencies or contradictions in spatial progression or transitions, requiring higher-level discourse reasoning. The model performed poorly in this category, with an accuracy of only 0.29.

We speculate that the stark contrast in accuracy between these categories stems from the model's limited capacity for discourse reasoning. While it may recognize abnormal spatial word usage based on learned patterns (e.g., "vomit along the stomach" being implausible), it struggles to infer inconsistencies that span across sentences and involve narrative coherence or event sequencing.

5.3.2 RSR - Textual Structures and Frames of Reference Influence

LLMs exhibit relatively strong performance in the RSR subtask, with the top-performing team approaching human-level accuracy. However, several limitations remain evident. First, LLMs are sensitive to textual structures, such as syntactic patterns and lexical distributions. This leads LLMs to prioritize referents with closer textual proximity and higher semantic similarity to the localizer rather than accurately interpreting the entire context. Performance drops notably on tasks demanding full-text comprehension and spatial scene construction. For example, in RSE texts describing a three-tiered scene, "上面…中间…再下…" (upper…middle…further below…) and "上面…下面…再下…" (upper…lower…further below…) depict identical layouts, but most models misinterpret the latter, probably confused by the local adjacent localizers. Second, LLMs show uneven accuracy across Frames of Reference (FoR). Their performance tends to be better on Intrinsic FoR, where direction is anchored to inherently oriented entities. In contrast, accuracy declines on Relative FoR, which requires viewpoint-based interpretation, especially concerning front/back/left/right terms.

5.3.3 RSE - Referent Divergence and Reasoning Challenges

Drawing on a previous study (詹卫东et al., 2024) and the data collected this year, four main reasons why two sentences have equivalent meaning is identified:

(1) The schemas of the localizers in the two sentences are similar;

(2) The schemas of the directional verbs in the two sentences are similar;

(3) The omitted referents are different in the two sentences;

(4) The sentences describe the spatial relations of an entity and its projection respectively.

The average accuracy of the top six teams across the four categories were 0.7044, 0.7355, 0.4123, and 0.7292, respectively. These result suggest that though LLMs can realize some spatial meaning equivalence, they still face challenges in integrating different spatial cognitive capabilities, such as identifying referents and reasoning to comprehensively understand spatial scenes.

5.3.4 SPR - Scenario Complexity and Absolute-Direction Effects

Different spatial scenarios preset in SPR contain various spatial relations. In addition, to evaluate LLMs' capability of connecting spatial relations with the absolute directions, entities in some questions are set in a certain cardinal direction, such as east, west, north, or south. We compare LLM performance on different scenarios with entities in different cardinal directions. The results are shown in Appendix E. The results show the following trend of the difficulty of the scenarios: Hexagon>Three-level Two-column>Four-person Booth. This trend suggests that the complexity of scenarios increases when the entity increases and spatial relations become irregular. LLMs can handle simple scenarios but struggle with the more complicated ones. Additionally, when the scenarios combine with absolute directions, a

LLM performance drop can be seen. This suggests that it is difficult for LLMs to link absolute directions with relative spatial relations.

Interestingly, Chinese accuracy is consistently higher than English, possibly due to better alignment between model training data and Chinese spatial concepts or prompt design in Chinese.

6 Participant Strategies

The top six teams submitted their models and technical reports. While most teams adopted several mainstream methods, some also uses unique strategies. The approaches applied in each part are briefly introduced as follows.

For the spatial language ability evaluation, due to the limited scale of the dataset and the specific requirement, prompt engineering is approached by almost all teams. The team from China University of Petroleum (UPC) wrote step-by-step prompts for each subtask. The team from Shanghai University (SHU) used Supervised Instruction Fine-tuning with In-Context Learning (ICL) to generate prompts and finetuned Qwen2.5-7B-instruct. The team HYQ and the team from Ping An Insurance (Group) Company of China (PAIC) used Qwen3-4B, and the team PAIC used Deepseek-Chat to obtain thinking rules for dynamic prompts. The 1^{st} team of Zhengzhou University (ZZU1) used Chain-of-Thought (CoT) prompts to improve model's performance. In addition, the 2^{nd} team of Zhengzhou University (ZZU2) used the spatial vocabulary and the demo set provided to generate additional data to finetune Qwen2.5-7B and Qwen3-4B.

For the spatial reasoning ability evaluation, teams were required to finetune DeepSeek-R1-Distill-Qwen-7B to participate in the evaluation. According to the proposed reports, Low-Rank Adaptation (LoRA) was generally used in the finetuning procedure. The team UPC proposed a mathematical framework to represent spatial relationships precisely, enabling LLMs to reason mathematically. The team SHU extracted constraints from questions using GPT-o3-mini-high, then finetuned the model by the questions with their constraints. The team HYQ and the team PAIC generated Chain of Thought (CoT) prompts for finetuning - the team HYQ using DeepSeek-R1, and the team PAIC using DeepSeek-Chat. For team ZZU1, Chinese data and English data were merged together, and multi-task learning was adopted. In addition, structured prompts were designed to guide model to reason step by step. The team ZZU2 considered the similarity of data in the spatial language ability evaluation and the spatial reasoning ability evaluation, and therefore the data from the spatial language ability evaluation were also used when finetuning LLM for the spatial reasoning ability evaluation.

In answer generation and extraction, voting is a widely used method, adopted by most teams to choose answers from multiple responses. Details of approaches can be found in their respective technical reports.

7 Conclusion

In this paper, we introduce the The Fifth Spatial Cognition Evaluation (SpaCE2025) benchmark. This benchmark consists of five subtasks to assess two core dimensions of spatial comprehension, semantic understanding and reasoning. To ensure the correctness of the questions, different approaches were adopted for different part. For SPR, an automatic pipeline was introduced to generate high-quality large-scale synthesis data. And for the spatial language ability evaluation, the consistency of human-annotated data was confirmed. In addition, we expand our reasoning dataset to English, enabling to evaluate multilingual spatial reasoning ability.

Based on fine-grained analysis, the following insights into spatial semantic understanding can be drawn:

- LLMs' ability of analyzing simple spatial semantics is comparable to human levels. However, in more complex understanding tasks such as RSE and DSA, there remains a noticeable gap between LLMs and humans. This indicates that LLMs' capability in comprehending complex spatial semantics still needs further improvement.

- Prompt-base methods only achieve limited improvement in LLMs' ability in spatial semantic understanding. However, by finetuning, best-performing LLM's accuracy in spatial reasoning increase

from 0.2622 to 0.8469, which is relatively high. This indicates that it is challenging to improve LLMs' spatial ability only by train-free methods, and finetuning is essential. Therefore, constructing high-quality data in quantity is significant for improving LLMs' language and reasoning capability.

In future works, we aim to further develop our benchmark. For the spatial language ability evaluation, more data need to be collected for both testing and finetuning. We also plan to increase the proportion of questions that shared texts. For the spatial reasoning ability evaluation, additional spatial relations and scenarios will be introduced. In addition, integrating other modals is another significant aspect that needs to be taken into consideration.

Acknowledgements

We would like to acknowledge the contributions of 10 human annotators and 40 human participants who participated in the evaluation from Peking University. This work was supported by the Major Program of the Key Research Center of the Ministry of Education of Humanities and Social (Grant No.22JJD740004).

References

Douglas H. Clements and Michael T. Battista. 1992. Geometry and spatial reasoning. *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*, pages 420–464.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 367–376, Cham. Springer International Publishing.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024. Reframing spatial reasoning evaluation in language models: a real-world simulation benchmark for qualitative reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, June. Association for Computational Linguistics.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. SemEval-2015 task 8: SpaceEval. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June. Association for Computational Linguistics.

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*, volume 36, pages 11321–11329.

Leonard Talmy, 1983. *How Language Structures Space*, pages 225–282. Springer US, Boston, MA.

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

Liming Xiao, Chunhui Sun, Weidong Zhan, Dan Xing, Nan Li, Chengwen Wang, and Fangwei Zhu. 2023a. SpaCE2022中文空间语义理解评测任务数据集分析报告(a quality assessment report of the Chinese spatial cognition evaluation benchmark). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 547–558, Harbin, China, August. Chinese Information Processing Society of China.

Liming Xiao, Weidong Zhan, Zhifang Sui, Yuhang Qin, Chunhui Sun, Dan Xing, Nan Li, Fangwei Zhu, and Peiyi Wang. 2023b. CCL23-eval任务4总结报告:第三届中文空间语义理解评测(overview of CCL23-eval task 4:the 3rd Chinese spatial cognition evaluation). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 150–158, Harbin, China, August. Chinese Information Processing Society of China.

Liming Xiao, Nan Hu, Weidong Zhan, Yuhang Qin, Sirui Deng, Chunhui Sun, Qixu Cai, and Nan Li. 2024. The fourth evaluation on Chinese spatial cognition. In Hongfei Lin, Hongye Tan, and Bin Li, editors, *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 122–134, Taiyuan, China, July. Chinese Information Processing Society of China.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—space2021数据集的研制. 语言文字应用, (02):99–110.

詹卫东, 秦宇航, and 肖力铭. 2024. 基于异形同义现象的机器空间语义理解能力评测研究. 辞书研究, (05):1–19+125.

Appendix A Text-Length Statistics

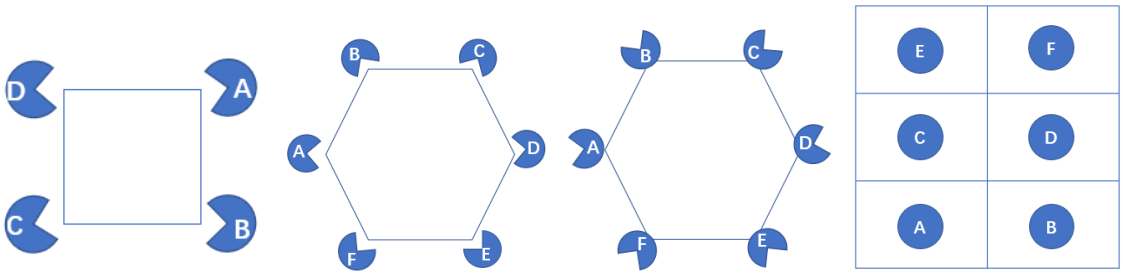
Table 8 shows the statistics of the length of the text in each subtask of the spatial language ability evaluation. The length is measured by the number of Chinese characters in the text. The lengths of the texts in all datasets are on the same level, and the text lengths varies from short to long.

Subtask	Max length	Min length	Average length
RSR	251	25	104.8
DSA	429	24	133
RSE	264	9	88.9

Table 8: Statistics of the length of the text in each subtask of the spatial language ability evaluation. The length is measured by the number of Chinese characters in the text.

Appendix B Scenarios Used in SPR

Figure 2 shows the four spatial scenarios preset in SPR. The scenarios are abstracted from everyday life scenes, and each scenario contains several spatial entities with different orientations. The blue circles with letters are spatial entities, with gaps indicating their orientation.



(a) Four-people Booth (b) Centripetal Hexagon (c) Centricfugal Hexagon (d) Three-level Two-column

Figure 2: Schema of 4 kinds of spatial scenario preset in SPR. Blue circles with letters are spatial entities, with gaps indicating their orientation.

Appendix C Instructions for All Subtask

Table 9 shows the instructions of each subtask in our benchmark. The instructions are fixed and provided in each question.

Subtask	Instruction
DSA	判断text的空间语言表达是否正确。请只回答“正确”或“错误”。(Determine whether the spatial language expression in the text is correct. Please respond with “Correct” or “Incorrect” only.)
RSR	判断interpretation是否正确。请只回答“正确”或“错误”。(Determine whether the interpretation is correct. Please respond with “Correct” or “Incorrect” only.)
RSE	判断text1和text2描述的空间场景是否相同。请只回答“相同”或“不同”。(Determine whether the spatial scenes described in text1 and text2 are the same. Please respond with “Same” or “Different” only.)
SPR(Chinese)	题目是多选题，有两个或两个以上的正确答案。答案选项必须与标准答案完全一致才能得分。请逐步思考，并最终输出答案选项。
SPR(English)	The question is multiple-choice with more than one correct answers. Answer choices must exactly match the gold answer to be considered correct. Please think step by step and finally output the answer choices.

Table 9: Instructions of subtasks in our benchmark. The instruction of each task is fixed. English translations are provided for DSA, RSR and RSE only for readability, and the instructions of SPR(Chinese) and SPR(English) are aligned.

Appendix D Knowledge-Based Pipeline for SPR Question Generation

The dataset generation for SPR is based on an automated framework that leverages a structured knowledge base, and consists of three core modules:

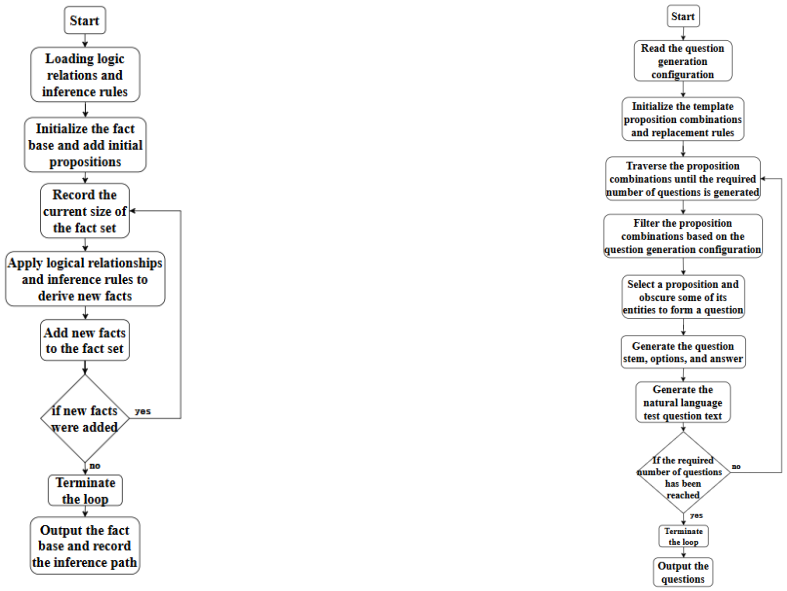
1. a commonsense-based spatial orientation relation reasoning knowledge base, which structurally describes spatial schemas and their constraint rules;
2. an automatic natural language question generation algorithm based on the knowledge base, enabling the automatic conversion from knowledge base to question bank using logical rules;
3. multidimensional evaluation metrics based on the knowledge labels of the questions, facilitating dynamic and fine-grained evaluation of the spatial reasoning abilities of large language models.

The spatial orientation relation reasoning knowledge base includes spatial scenarios abstracted from everyday life scenes, spatial reasoning templates (e.g., “X is to the right of Y”), logical relationships and reasoning rules between the templates (e.g., “X is to the right of Y” is equivalent to “Y is to the left of X”), and initial propositions that describe the overall layout of the scenario. The spatial reasoning knowledge base covers four spatial schemas, ten specific spatial scenes, and thirty types of spatial relationships.

The automated generation program primarily consists of the inference module and the question generation module. The inference module functions by reading the knowledge base based on user configurations, including initial facts, spatial reasoning templates, logical relationships between templates, and reasoning rules. It then iteratively applies all the logical rules to deduce all spatial relations between entities from the initial facts, generating a complete fact base.

The question generation module filters proposition sets that describe the overall spatial schema from the fact base to form question stems. From these, one proposition is selected to generate the question and options. Additionally, the abstract proposition in the question is transformed into a natural language text description by adding scene descriptions and entity names, ultimately outputting reasoning questions that are suitable for general readers. The workflows of the inference module is shown in Figure 3a, and

the question generation module is shown in Figure 3b. Other details of the pipeline will be discussed in future works.



(a) Inference Module of SPR

(b) Question Generation Module of SPR

Figure 3: Flowcharts of the inference module and the question generation module in the SPR pipeline. The inference module generates a fact base from the knowledge base, and the question generation module generates questions based on the fact base.

Appendix E Scenario-Wise SPR Accuracy and Orientation Effects

Table 10 shows the average accuracy of the top six teams on SPR questions with different spatial scenarios and different entity orientations. The results show that LLMs can handle simple scenarios but struggle with the more complicated ones. Additionally, when the scenarios combine with absolute directions, a LLM performance drop can be seen. This suggests that it is difficult for LLMs to link absolute directions with relative spatial relations.

Scenario	Orientation/Direction	Accuracy (Chinese)	Accuracy (English)
Four-person Booth	No cardinal directions	0.815	0.741
	East-West	0.797	0.716
	North-South	0.835	0.709
Centrifugal Hexagon	No cardinal directions	0.625	0.571
	East-West	0.447	0.407
	North-South	0.530	0.470
Centripetal Hexagon	No cardinal directions	0.609	0.554
	East-West	0.357	0.313
	North-South	0.425	0.367
Three-level Two-column	East-West	0.626	0.569

Table 10: Average accuracy of the top six teams on SPR questions with different spatial scenarios and different entity orientations. *East-West* means that at least one entity in the scenario is set in the east or west direction, and *North-South* means that at least one entity in the scenario is set in the north or south direction. *No cardinal directions* means that no entity in the scenario is set in a certain cardinal direction.