Overview of CCL24-Eval Task 3: The Fourth Evaluation on Chinese Spatial Cognition

Liming Xiao[‡] **Nan Hu**[‡] **Weidong Zhan**[†,*] **Yuhang Qin**[‡] **Sirui Deng**[‡]

Chunhui Sun[†] **Qixu Cai**[‡] **Nan Li**[†]

Department of Chinese Language and Literature, Peking University

Center for Chinese Linguistics, Peking University

[†]{zwd,sch,linan2017}@pku.edu.cn

[‡]{lmxiao,hunan,hezonglianheng,d2020sr,cqx}@stu.pku.edu.cn

Abstract

The Fourth Chinese Spatial Cognition Evaluation Task (SpaCE 2024) presents the first comprehensive Chinese benchmark to assess spatial semantic understanding and reasoning capabilities of Large Language Models (LLMs). It comprises five subtasks in the form of multiple-choice questions: (1) identifying spatial semantic roles; (2) retrieving spatial referents; (3) detecting spatial semantic anomalies; (4) recognizing synonymous spatial expression with different forms; (5) conducting spatial position reasoning. In addition to proposing new tasks, SpaCE 2024 applied a rule-based method to generate high-quality synthetic data with difficulty levels for the reasoning task. 12 teams submitted their models and results, and the top-performing team attained an accuracy of 60.24%, suggesting that there is still significant room for current LLMs to improve, especially in tasks requiring high spatial cognitive processing.

1 Introduction

Spatial expressions have long been a focus of cognitive linguists because they are not only a high-frequency phenomenon in human language but also embody the fundamental mechanisms of how humans perceive the world (Talmy, 1983). In recent years, some NLP evaluation tasks have also sought to explore machines' cognitive semantic understanding capabilities through the lens of spatial expressions by a single task such as spatial role labeling (Pustejovsky et al., 2015; Kordjamshidi et al., 2017) or spatial reasoning (Mirzaee et al., 2021). As for evaluation in Chinese, the **Spatial Cognition Evaluation Task (SpaCE)** has been held for 3 years since 2021 with the aim of comprehensively evaluating machine's capabilities with regard to Chinese spatial semantic understanding by multi-task learning (詹卫东et al., 2022; Xiao et al., 2023a; Xiao et al., 2023b). The latest results indicate that machines' spatial semantic understanding capabilities lags significantly behind the average level of humans, especially in tasks requiring high cognitive processing. Spatial semantic understanding remains a challenging task for NLP systems, even for large language models (LLMs).

In this work, we construct a benchmark that comprehensively assesses LLMs' performance on understanding spatial expressions and conducting spatial reasoning in the following five subtasks: (1) Identifying Spatial semantic Roles (ISR); (2) Retrieving Spatial Referents (RSR); (3) Detecting Spatial semantic Anomalies (DSA); (4) Recognizing Synonymous spatial Expression with different forms (RSE); (5) Spatial Position Reasoning (SPR). SpaCE 2024 has released 10,373 multiple-choice questions (MCQs) of the benchmark, which is available at https://github.com/2030NLP/SpaCE2024. Summarized statistics are shown in Table 1. In summary, the characteristics of the benchmark can be outlined as follows:

- The subtasks are interconnected. They not only share a common evaluation vision about spatial semantics but also emphasize the relevance of data to each other, reflected in the shared contextual information across some tasks.
- Structured synthetic data are used in the reasoning task. These rule-based generated data are 100% accurate and do not require quality auditing.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

Sub-Task	Answer Type	#Train	#Dev	#Test		#Total
				Non-Disturb.	Disturb.	
ISR	single	1,074	186	746	30	2,083
	multi	19	4	24		
RSR	single	937	226	489	30	1,948
	multi	161	24	81		
DSA	single	1,077	40	500	30	1,647
RSE	single	4	44	517	30	740
	multi	1	11	133		
SPR	single	909	468	1,509	30	3,955
	multi	301	207	531		
#Total	-	4,483	1,210	4,530	150	10,373

Table 1: Statistics of SpaCE 2024 benchmark.

- The consistency of LLMs is concerned. We set some disturbance data in the test set, by repeating, rephrasing questions, and adjusting the order of options. If an LLM cannot perform well under different conditions of disturbance, it's hard to confidently assert that this LLM has some kind of capability.

2 Task Definition

The five subtasks of SpaCE 2024 have different difficulty levels of cognitive processing, which can be roughly divided into three levels: (1) Basic: ISR. This level requires understanding the lexical meaning, grammatical structure, and semantic structure of a sentence, focusing on local information processing. (2) Moderate: RSR and DSA. This level requires understanding discourse semantics with context and commonsense, focusing on global information processing. (3) Advanced: RSE and SPR. This level requires employing advanced cognitive capabilities based on semantic understanding, such as constructing spatial scenes and conducting spatial reasoning. Details about the five subtasks are introduced in the following.

2.1 Identifying Spatial semantic Roles (ISR)

Identifying semantic roles is a fundamental task in semantic understanding. We assume that a machine that can understand underlying semantics should perform well in identifying semantic roles. SpaCE 2022 proposed an annotation scheme called **STEP** to formalize the semantics of spatial expression. S, T, E and P stand for Spatial entity, Time, Event and Place, respectively. The meaning of a spatial expression can be represented by STEP annotation as "an entity is at a certain place at a certain time via an event". Appendix A presents 14 spatial roles of STEP.

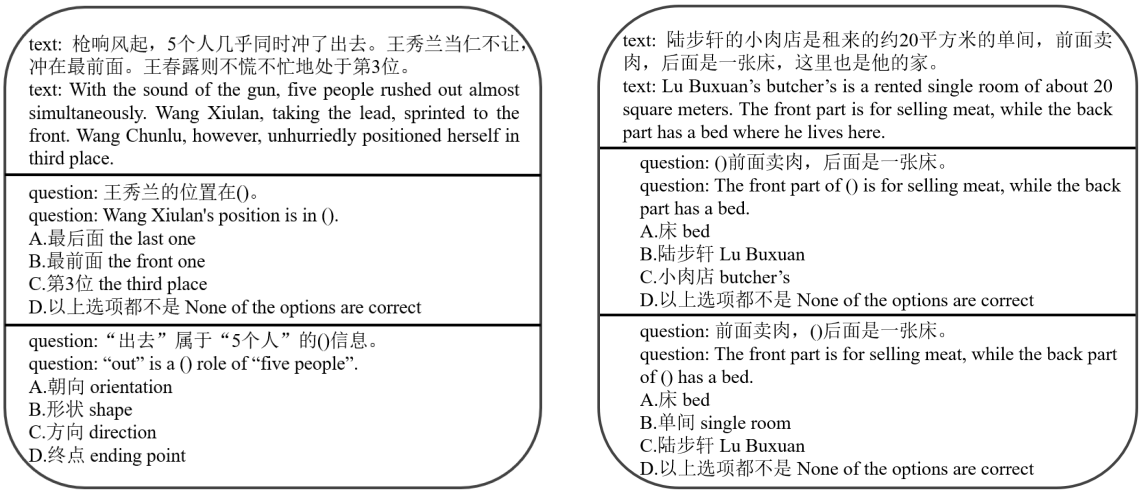
In ISR, machines are required to select the spatial role corresponding to a spatial expression or, conversely, select the spatial expression corresponding to a spatial role, as exemplified in Figure 1(a). The answers of two questions in example are (B) 最前面 and (C) 方向.

2.2 Retrieving Spatial Referents (RSR)

In Modern Chinese, a spatial entity's position is commonly expressed through the combination of a noun as the referent and a following localizer, such as 树下面*(the area under the tree)* and 剧院里面*(the area inside the theatre)*. However, referents can sometimes be omitted and instead be found in a more distant position in the text. In Figure 1(b), there are no referents preceding the localizers 前面*(front)* and 后面*(back)*, yet we can still understand that the meat is sold at the front part of the butcher's and the bed is at the back part of the butcher's.

Since referents play a crucial role in accurately mapping localizers to real-world spatial scenes, the ability to retrieve omitted referents reflects the capability of LLMs in text understanding. In RSR, the

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations



(a) ISR example (b) RSR example

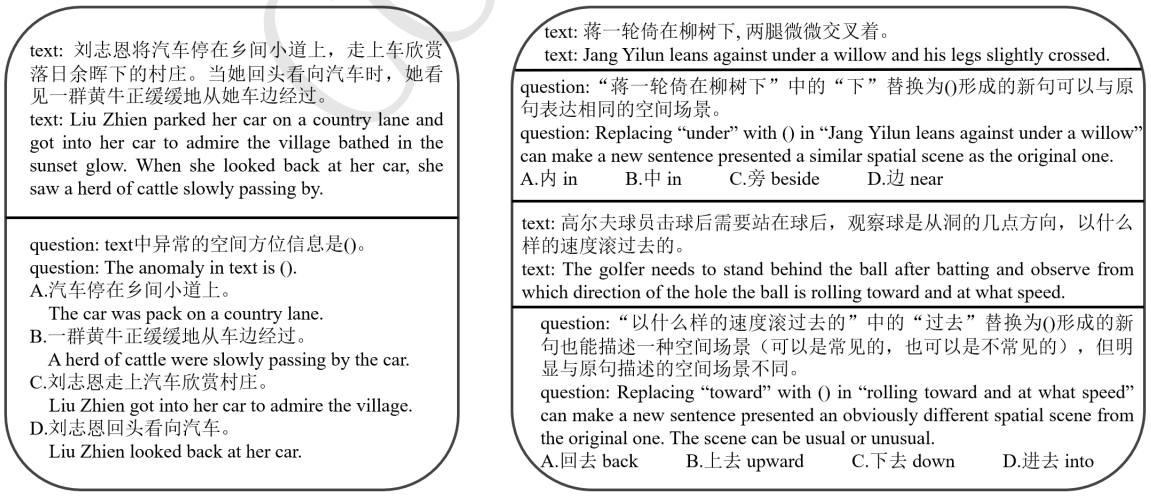
Figure 1: Examples from ISR and RSR training set. English translations are shown for better readability.

machine needs to select one or more entities as the referent(s) of the localizer in question. The answers to two questions in the example are (C) 小肉店 and (B) 单间.

2.3 Detecting Spatial semantic Anomalies (DSA)

Having capability of semantic understanding means being able to detect an expression with semantic anomalies. The text in Figure 2(a) has a spatial semantic anomaly: the protagonist should get out of her car to enjoy the scenery after parking instead of getting in, because she was already inside the car. The following context ”she looked back at her car” also indicates that the protagonist was outside the car.

In DSA, the machine needs to select an expression corresponding to the anomaly in text. The answer of the question in example is (C) 刘志恩走上汽车欣赏村庄. Actually, the expression in option generally don’t have semantic anomalies, but it always conflicts with contextual information in text that makes the anomaly stick out.



(a) DSA example (b) RSE example

Figure 2: Examples from DSA and RSE training set. English translations are shown for better readability.

2.4 Recognizing Synonymous spatial Expression (RSE)

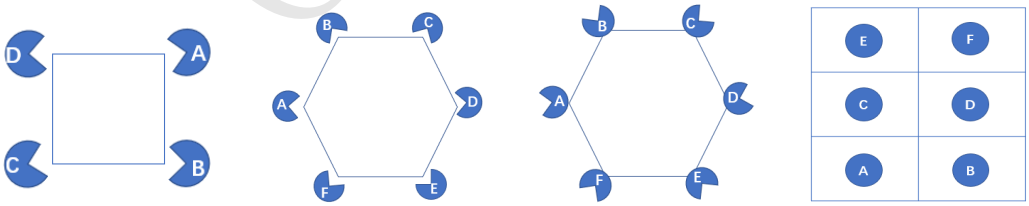
In Modern Chinese, different localizer generally causes to different meaning of the expression. For instance, if we change the localizer 上*(on)* of the sentence 把筷子放在碗上*(Putting the chopsticks on the bowl)* into 边*(beside)*, then the position of the chopsticks will change from the top of the bowl to the area near the bowl. However, there are certain situations where two spatial expressions can have similar meanings even though the localizer is different. For instance, the meaning of both sentence 蒋一轮倚在柳树下 and sentence 蒋一轮倚在柳树上 is *Jiang Yilun leans against a tree*, while the former uses localizer 下*(under)* and the latter uses localizer 上*(on)*. The spatial scenes pictured by these sentences are considered equivalent.

Comprehending such sentence groups needs to utilize commonsense and knowledge to compare the spatial areas of the entity activated by different localizer in the real world. RSE has two kinds of questions. One requires selecting all localizers or directional verbs which ensure that the new sentence after replacement can describe a spatial scene similar to the original one, as exemplified by the top question in Figure 2(b), where the answers are (C) 旁 and (D) 边. The incorrect options are so because the replaced sentence implies "the person leans inside the tree", which is an anomaly. The other type requires to ensure that the two sentences describe different spatial scenes, as seen in the bottom question of Figure 2(b), where the answer is (D) 进去 and the others are incorrect: the scenario described by 滚上去*(rolling upward)* or 滚下去*(rolling down)* is similar to 滚过去*(rolling toward)*, because the trend of the ball's movement are similar and the former is just more detailed in the description of direction; 滚回去*(rolling back)* makes the sentence anomalous because 回去*(back)* implies a return to a previous position but the context suggests batting the ball to a new position.

2.5 Spatial Position Reasoning (SPR)

Spatial Reasoning is a cognitive process based on the construction of mental representations for spatial entities, relations, and transformations, which is necessary for spatial semantic understanding (Clements and Battista, 1992). In SPR, the context is automatically generated based on a preset spatial layout, where entities within the text can form a spatial layout, but their positions must be deduced using the spatial relations provided by the text. Figure 3 shows 4 types of spatial layout schema used in SPR. These layouts contribute to evaluating the comprehension of different spatial relations, as well as the same relation expressed in different ways, in an unified structure. Moreover, a larger number of entities and more advanced layouts indicate a more complex spatial relation, which is conducive to building an evaluation system with varying levels of difficulty.

The example in Figure 4 is an instance of the concentric hexagon layout. First, you need to use the 4 given conditions to infer the position of six statues, as the resolution pictured. Then, find the statues that are not adjacent to Zhang Tianshi: (B) 曹国舅 and (C) 张果老.



(a) square (b) concentric hexagon (c) eccentric hexagon (d) 3×2 matrix

Figure 3: Schema of 4 kinds of spatial layout. Blue circles with letters are spatial entities, with gaps indicating their orientation

3 Dataset

Each data is a four-choice question containing five fields: *qid, text, question, option and answer*. The *qid* field is the unique identity label of the data. For example, *qid* as "3-test-s-1660" points to the 1660^{th}

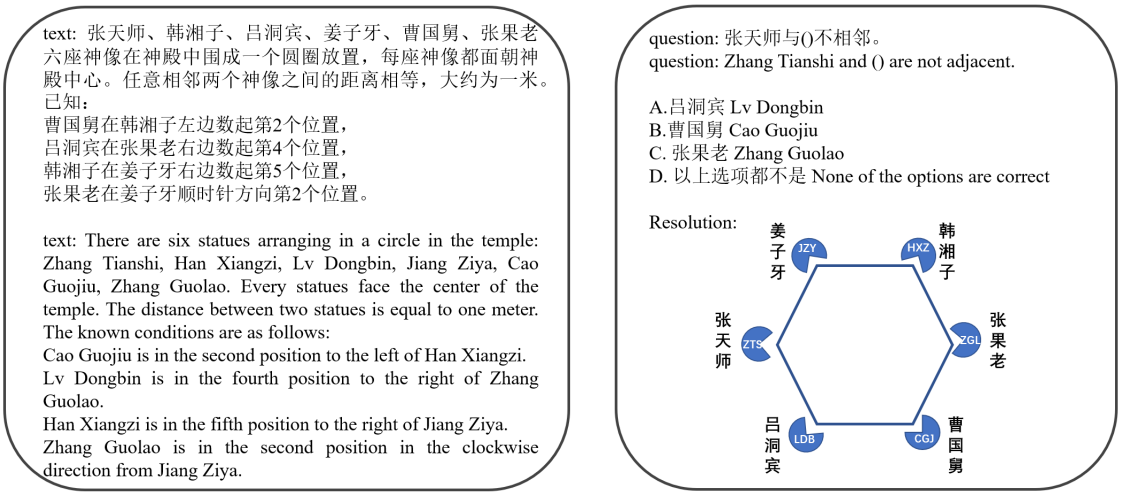


Figure 4: Example from SPR training set. English translations and resolution schema are shown for better readability.

data in the test set, with only one answer, belonging to DSA. We use numbers 1 to 5 to present the five subtasks and use ”s” or ”m” for a single answer or multiple answers, at least 2 up to 4. The *text* and *question* field are string shown above in Figures. Notably, the genre of the text covers general domains such as textbooks and newspapers, as well as specific domains including text of traffic accidents, body movements and geographical encyclopedias. The *option* field is a dictionary with the keys from the letters A to D. The *answer* field is an array containing the letter corresponding to the correct option.

To ensure data quality, data in ISR, RSR, DSA and RSE were generated based on manual annotation and data auditing with the rules that data could only be accepted if two people agreed or one person audited it twice. The SPR data were automatically generated by a spatial knowledge base and an interpretable program, without the need for data auditing.

3.1 Word Replacement for Text

Texts in DSA and RSE were generated by word replacement. We first grouped localizers and directional verbs based on their morphological characteristics and the degree to which they can be grammatically substituted. For example, 上边,下边,前边,后边*(up,down,front,back)* were in one group. The groups also served as options of RSE. Then, using the results of part-of-speech tagging, we replaced the original words in the text with words from the same group. Finally, we manually annotated the replaced sentences to determine if there were any anomalies in the sentences. Data with consistent annotations by two annotators were considered valid. If there were any anomalies, they belonged to DSA, otherwise became a resource of RSE for further annotation about whether the two sentences can picture a similar scene. To generate more sentence pairs for RSE, the members of the SpaCE team needed to construct texts that satisfy the requirement of the given word pair.

3.2 Annotation Extraction for Option

Since 2022, we have recruited nearly 30 students majoring in linguistics each year to annotate STEP information, achieving about 5000 high-quality annotations as a data resource for subtasks. To construct appropriate options for ISR, we extracted annotations that include at least three target roles and then filled in question templates designed for different semantic roles to generate data. RSR took advantage of the discontinuity between the omitted referents and the localizers. We selected discontinuous texts from the annotations and, using the results of part-of-speech tagging, chose nouns as candidate options. For DSA, before word replacement, the text had been annotated with STEP information, making it easy to locate the annotation of the original word and perform the replacement to create a wrong annotation. Together with three other correct annotations of this text, they were sent to GPT-4 to generate natural

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

sentences, thereby forming options related to this text.

3.3 Synthesis Data for SPR

It's too challenging for humans to entirely manually construct a spatial reasoning data, so we try to use synthetic data techniques. Current technology employs LLMs to generate more data, such as data augmentation or to generate complex prompt and narrative texts.(Mukherjee et al., 2023; Josifoski et al., 2023; Eldan and Li, 2023; Dai et al., 2023; Yehudai et al., 2024) For reasoning tasks, LLMs are still unable to produce high-quality data, especially in cases where the reasoning chain is long(Liu et al., 2023). Since the time required for human to audit a reasoning question is much longer than the other tasks, we have decided to use a rule-based method to generate data with transparent intermediate processes. SpartQA is a representative work of the rule-based method, but it only considered 7 types of spatial relations which might no longer be that challenging for LLMs(Mirzaee et al., 2021). To address this, we proposed a program that can automatically generate reasoning questions driven by expert knowledge. Figure 5 illustrates the entire process.

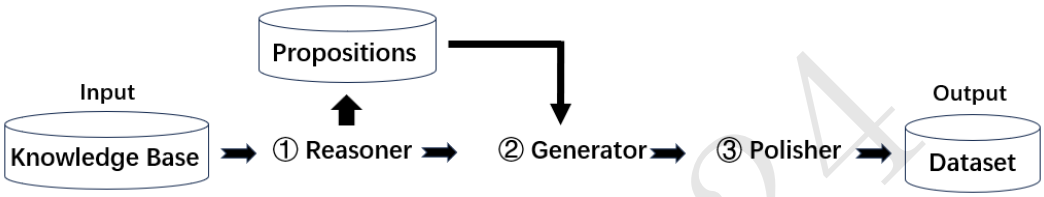


Figure 5: A flowchart for SPR data generation program.

Based on the spatial layout shown in Figure 3, the knowledge base uses templates and reasoning rules to describe the positions of entities within the layout. Each template is a spatial expression that contains slots for entities, like X在Y左边数起第2个位置*(X is in the second position to the left of Y)*, where "X" and "Y" are the slots. Reasoning rules between two templates are logical relations, including equivalence, implication, inclusion, conflict, and reversibility, ensuring that the program has the ability to generate a large number of varying expressions based on a small number of premises. For example, X在Y左边数起第2个位置can deduce X在Y顺时针方向第2个位置*(X is in the second position in the clockwise from Y)* because they are equivalent. We also set up multiple templates as conditions and one template as a conclusion in reasoning rules to expand the program's reasoning ability. For instance, X在Y左边数起第2个位置groups with Z在Y左边数起第1个位置*(Z is in the first position to the left of Y)* can deduce X在Z左边数起第1个位置*(X is in the first position to the left of Z)*. The current knowledge base contains a total of 592 templates written by linguistic experts, expressing more than 30 types of spatial relations.

After loading logical relations and condition-conclusion pairs from knowledge base, the Reasoner needs to input some initial propositions that can form a complete spatial layout and automatically generate more prepositions. Then the Generator will randomly take several prepositions as known conditions and one preposition as a question until these prepositions can form a spatial layout. The entity will be removed from the question as an answer and the other entities will be the options. These extracted propositions are then fed back into the Reasoner to output the intermediate reasoning process. Finally, the Polisher will replace the slots in template like X as a predefined entity like 张天师 *(Zhang Tianshi)* and add a lead in the text to make it conform to natural language.

Given that the knowledge base is constructed correctly, the data generated by this method is guaranteed to be 100 % accurate. It can generate a vast amount of diverse data without limitations if the spatial templates and inference rules in the knowledge base are sufficiently comprehensive. Futhermore, compared to using LLMs for data synthesis, the rule-based method offers strong interpretability, transparency in intermediate processes, and traceable reasoning steps. This is extremely beneficial in assessing the quality and difficulty of spatial reasoning data.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

Team	Total	Single	Multi	ISR	RSR	DSA	RSE	SPR
TeleAI	**60.24**	**64.90**	37.45	93.64	**89.47**	**84.80**	56.31	34.71
SHU	59.69	64.34	36.93	91.43	84.91	81.00	54.31	**37.16**
PKU	59.49	63.55	**39.66**	**94.29**	77.19	78.00	**58.77**	37.11
ZUT	56.60	61.71	31.60	91.43	85.26	79.80	49.23	32.11
GDUT	56.47	61.26	33.03	92.86	73.68	76.20	58.62	32.40
BNU	56.20	61.79	28.87	**94.29**	79.65	74.20	56.92	30.64
Knowdee	54.48	59.29	30.95	91.17	75.09	75.40	52.31	30.44
DataGround	53.55	61.69	13.78	90.13	73.33	79.60	51.23	28.58
SU	51.99	56.16	31.60	90.00	80.53	70.20	24.15	34.07
CPIC	48.65	52.91	27.83	86.10	76.67	62.20	40.31	26.03
Insgeek	47.24	51.53	26.27	81.82	67.19	70.00	33.69	27.35
XTU	33.64	36.80	18.21	67.27	49.47	21.60	26.62	21.72
Baseline1 (Fine-tuning)	47.92	54.37	16.38	88.18	75.09	68.60	42.00	21.96
Baseline2 (GPT-4-1106-preview)	46.29	52.94	13.78	85.19	54.39	65.60	45.08	25.00

Table 2: Accuracy of 12 teams (%)

4 Evaluation and Results

We use accuracy as the ranking metric. The higher the accuracy, the better the LLM's capability of comprehensive spatial semantic understanding.

$$Accuracy = \frac{\#Correct}{\#NonDisturbDatas} \tag{1}$$

There are 150 disturbance data in test set that use to evaluate the consistency of LLMs. A group of disturbance data contains an original data, a repeated version, a version with different order of options, and a version with different expression of question. A LLM is considered stable only if all the answers in a group are the same. The higher the consistency score, the less the LLM is affected by the data format, making the evaluation results more reliable.

$$Consistency = \frac{\#StableGroups}{\#DisturbGroups} \tag{2}$$

Totally 30 teams enrolled and 12 teams stick to the end. The top 6 teams won the prize. They all employed LLMs either fine-tuned or utilized prompt engineering. The accuracy results are shown in Table 2. We develop two baselines, one is using prompt engineering in GPT-4-1106-preview[0], the other one is fine-tuning in Qwen1.5-7B-Chat model[1]. Both baselines were tested in zero-shot mode. Most of the teams surpassed our baselines, yet the best total accuracy only slightly exceeded 0.6, suggesting that SpaCE 2024 is quite challenging for current LLMs, especially in RSE and SPR. In contrast, LLMs performed well on tasks from basic level and moderate level, notably the top accuracy was close to 0.9

[0]https://openai.com/index/gpt-4
[1]https://github.com/QwenLM/Qwen1.5

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

in ISR. We also analyzed the performance of questions with single answers versus multiple answers and found that LLMs are much better at handling single-answer questions.

We selected 100 questions from each of the four tasks, excluding SPR, to create a human test set. Each task had 6 human participated. Table 3 shows the performance of the top 6 teams and the average of human accuracy on the human test set. None of a team surpassed human performance. Although accuracy in ISR and RSR is close to human, there is still a gap between LLMs and human on DSA and RSE. For SPR, we believe that spatial reasoning questions tend to reflect individual human differences rather than the overall human level. Additionally, the quality of our synthesis data does not need to be reflected by human scores, as our members have sampled on every type of questions, ensuring no error.

Team	ISR	RSR	DSA	RSE
TeleAI	94.00	92.00	81.00	54.00
SHU	82.00	85.00	72.00	58.00
PKU	91.00	81.00	75.00	60.00
ZUT	86.00	90.00	77.00	49.00
GDUT	86.00	79.00	77.00	54.00
BNU	93.00	83.00	71.00	58.00
Human	**96.00**	**92.33**	**91.50**	**86.33**

Table 3: Accuracy of top 6 teams in Human test set (%)

Table 4 shows the top 6 teams' consistency scores. The results indicate that the performance of LLMs is affected by the question format, which is unrelated to the test content, particularly the order of the options. The 1st was not only the most accurate but also the most stable.

Team	Consistency	Repeated	Reform Question	Switch Option
TeleAI	**84.85**	98.46	85.00	**89.23**
SHU	81.82	95.38	90.00	87.69
PKU	69.70	81.54	85.00	78.46
ZUT	71.21	96.92	**95.00**	72.31
GDUT	77.27	90.77	85.00	84.62
BNU	72.73	98.46	80.00	76.92

Table 4: Consistency scores of top 6 teams (%)

5 Overview of Approaches

Only the top 6 teams submitted their models. Model ensemble with majority voting is the main strategy.

The team from TeleAI company manually selected 5 examples for each subtask to construct in-context learning (ICL) training data, then fine-tuned Qwen1.5-7B-Chat 3 times to do the majority voting. TeleAI set weights on various models and tasks, thereby performing a more nuanced and context-sensitive aggregation to enhance the accuracy and consistency score of the final predictions.

The team from Shanghai University (SHU) recategorized the dataset into four types of questions based on the training size and the difficulty of the tasks. For the questions having large training data with low difficulty, SHU fine-tuned Qwen1.5-7B-Chat, Yi-6B-Chat, and intern2-chat-7B, then voted for the majority as answer. For the structured data in SPR, SHU assumed that it was too difficult and not suitable for training with the easy question, so they fine-tuned the above models again and voted. The remaining questions did not have enough training data to fine-tune well, so SHU ran GPT-4o with Chain of Thought (CoT) and Qwen1.5-110B-Chat with 5-shot to solve them.

The team from Peking University (PKU) ran GPT-4o three times in ISR, RSR, DSA, SPR and conducted a majority vote. For RSE, it ran GPT-4o and Deepseek-chat, meanwhile set another Deepseek

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

as the judge to decide which one was better when two answers were not the same. PKU also set an error correction mechanism to rerun the question in some situations like only outputted one answer in a multi-answer question.

The team from Zhongyuan University of Technology (ZUT) used data augmentation by generating pseudo-labeled data. They used training data to fine-tune Qwen1.5-7B-Chat in different hyperparameter and chose high-score models to predict the unlabeled test data. After majority voting, nearly 3k reliable pseudo-labeled data were added to the training set to fine-tune again until no further improvement in the accuracy of the test set. ZUT also ran GLM-4 in the test set and made model ensemble.

The team from Guangdong University of Technology (GDUT) employed 5-fold cross-validation and majority voting to produce the final prediction by fine-tuning ChatGLM3-6B-base. Another model for ensemble was Ernie-4.0-8k-0329, whose overall score was higher than ChatGLM3 in ISR, DSA and RSE. For ICL, GDUT dynamically selected 5-shot examples by computing the similarity of options between the target test data and the training data in same subtask. The top 5 of similarity would be chosen as examples.

The team from Beijing Normal University (BNU) constructed prompt with one-shot. To select the most representative example for each task, they used Sentence-BERT(Reimers and Gurevych, 2019) to encode all training data into embedding and averaged all embedding to get the centroid of the task. The data closest to the centroid in terms of embedding distance was the most representative.

In general, the models of the top 6 teams were primarily based on ICL and the majority voting strategy. Fine-tuning can bring about a more stable model, and it can improve the performance of a 7B model close to that of GPT-4o, but this improvement is limited to simple tasks. For the two most complex tasks, fine-tuning might not bridge the gap because it's fundamentally related to the reasoning capabilities of the base LLMs.

6 Additional Analysis

Following an analysis based on the results of the top 6 teams, We studied each subtask, the shared text across tasks, and conducted a comparison between SpaCE 2024 and SpaCE 2023.

6.1 Performance of Per-Task

Questions in ISR were labeled as STEP roles. The good performance on most types may be due to the powerful ability of LLMs to capture the correspondence between form and meaning, such as the negation marker of *Factivity* and the preposition marker "到" of *Ending Point*. There were two types of questions where performance needed to be improved: 1) Questions asking for selecting a spatial role. For example, only two teams correctly answered the question "车辆后部"属于"外卖箱"的()信息*("the back of the car" is a () role of "the delivery box")*. Some LLMs may not have a good grasp of STEP, leading to confusion between *Internal Place*, the correct answer, and *Part*. This issue was more serious in the *Direction* type, where LLMs often misidentified directional verbs as other roles. It was possible that the machine's internal knowledge did not classify directional verbs under semantic of *Direction* like STEP did. Therefore, the prompt must include a detailed introduction of STEP to achieve better performance. 2) Questions requiring reasoning ability, including event sequencing questions in *Time*, qualitative distance reasoning questions in type *Distance*, and position reasoning questions similar to SPR in *External Place*.

In RSR, we found that the difficulty of referent retrieval is related to the type of localizers. LLMs performed better on questions where localizers describe up-down, inside-outside, and front-back relations than on those describing left-right and cardinal directions (north, south, east, west). We also observed that LLMs tended to choose entities closer to the target localizer. For example, in the question "()后边架着一架摄像机记录整个授课过程"*("Behind ( ) is a camera recording the entire teaching process"*, only one team correctly chose "classroom" as the answer. The remaining five teams chose the closer option, "blackboard", which is illogical because blackboards in classrooms are commonly mounted on walls, making it impossible to place a camera behind them. Another observation is that LLMs tended to avoid using the subject as the referent. For example, in the question "袁格兵将绳子系在()左边的门柱

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

上，把另一头抛给司机”*”Yuan Gebing tied a rope to the left doorpost of () and threw the other end to the driver”*, the correct answer is the subject ”Yuan Gebing”, but all teams chose ”streetlight” or ”car”, which are inconsistent with the context of the flood rescue situation described in the text.

The DSA's questions were annotated with the amount of textual information needed to infer anomalies. Statistics show that the more information required, the harder the question, and the poorer the LLMs' performance. In texts describing body movements and traffic accidents, embodied experience and traffic knowledge were necessary to infer the anomalies but not included in the text, making it more difficult for LLMs. For example, in texts describing body movements, LLMs performed poorly on questions related to the orientation of the chest in a prone position. Similarly, in texts describing traffic accidents, without common knowledge, such as north-south roads have east-west lanes, it was hard to judge whether two cars will collide. In terms of understanding spatial relations, we observed the same phenomenon as in RSR that LLMs performed worse with left-right and cardinal directions compared to other spatial relations.

RSE had two types of questions. LLMs performed better with synonymous spatial expressions than with those have different meanings, which was consistent with human performance. However, unlike humans, the pairs of words with which LLMs performed better in synonymous spatial expressions were quite different from another type. The former mainly included synonyms, such as ”旁-边” *(beside-next to)* and ”中-内” *(within-inside)*, while the latter mainly included antonyms, such as ”内-外” *(inside-outside)* and ”上面-下面” *(above-below)*. Humans did not show such a significant difference. For example, with common antonyms like ”上-下” (up-down), humans performed well on both types of questions, although antonyms were rarely used in synonymous spatial expressions. It indicates that LLMs' understanding of spatial semantics remains at surface semantics, lacking the capability of modeling spatial scenes case by case.

The difficulty of SPR questions was measured based on the number of actual reasoning conditions used, which were automatically outputted by the reasoning program. The more reasoning conditions required, the longer the reasoning chain, the harder the question, and the worse the performance. For instance, the score of the questions with using 2 conditions was 40 % less than those with only 1 condition. The complexity of spatial layouts and the number of entities also affected LLMs' performance: 1) The more complex the layout, the worse the LLM performs, shown as Square layout >hexagonal layout >matrix layout. 2) LLMs performed better on texts involving 4 entities compared to those with 5 or 6 entities. 3) If a cardinal direction was added to a hexagonal layout, LLMs' performance decreased by approximately 8 %. In terms of spatial relation, we labeled each question with spatial relation based on the localizer used. Statistics show that LLMs perform better in reasoning about opposite and adjacent relation compared to left-right and cardinal directions. Reasoning about up-down relation is even worse.

6.2 Discussion on the Shared Text

20% of the test set questions shared textual information between subtasks. Taking the most basic task, ISR, as a clue, we constructed 801 question pairs in the form of (ISR, RSR/DSA/RSE/SPR). The conditional probability that ISR is correct when the other subtask is correct reached 92.26%, suggesting that the LLMs' ability to complete advanced tasks is based on a fundamental understanding of semantics. On the other hand, the number of pairs where both tasks are incorrect is much fewer than the number of pairs where ISR is incorrect and another one is correct. Other fundamental capabilities of LLMs like reasoning may compensate for this deficiency when the LLMs' semantic understanding is flawed.

6.3 Comparison between 2023 and 2024

SpaCE 2024 inherited ISR, DSA and RSE tasks from SpaCE 2023, providing a basis for comparison. However, the data formats differ between the two years. In 2023, ISR and DSA were sequence labeling tasks, and RSE was a true-false task, while in 2024, they were all multiple-choice questions. To allow a comparison, we selected questions with the same text from both years and used F1 score as a metric for ISR and DSA to be consistent with 2013. For ISR, we only considered the single labeling corresponding to the MCQs as the result of SpaCE 2023.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 122-134, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

Year	ISR		DSA		RSE	
	#Num	F1(%)	#Num	F1(%)	#Num	Acc.(%)
2023	92	45.40	40	62.54	46	68.48
2024		95.47		86.25		52.54

Table 5: Per-task Comparison between 2023 and 2024 of SpaCE

As Table 5 shows, the F1 scores in 2023 are the average of two BERT-based models and the accuracy score is the average of two LLM systems. All scores in 2024 are the averages of six LLM systems. The state-of-the-art improved markedly for ISR and DSA, but went down for RSE. This displays the strengths of LLMs compared to BERT-based models, but also indicates the effect of data format. For ISR and DSA, sequence labeling tasks are overly complex in format, preventing the machine's capabilities from being fully demonstrated. For RSE, MCQs are more challenging than true-false questions because true-false questions only require comparing one pair of words, while MCQs provides four pairs, resulting in a lower score. In general, it was the first time that the SpaCE series of tasks employed multiple-choice questions, helping to improve the validity of the evaluation.

7 Conclusion

This paper introduces the fourth evaluation of the Chinese Spatial Cognition Task. The benchmark consists of five subtasks covering 5 dimensions of spatial semantic understanding to comprehensively assess the machine's language comprehension and reasoning capabilities on spatial semantics. In addition to capability assessment, SpaCE 2024 concerns the consistency of LLMs' performance and sets a disturbance set. As for quality control, a novel rule-based method was used in the spatial reasoning task, successfully generating high-quality synthesis data with transparent intermediate processes and difficulty levels.

The accuracy scores across 5 subtasks reveal the following trend: ISR >RSR >DSA >RSE >SPR. Based on a fine-grained analysis, we have gained the following insights into the spatial semantic understanding capabilities of LLMs:

- Regarding the spatial relation, understanding "left and right" and "cardinal directions" are weaknesses of LLMs, which have been observed among tasks.
- LLMs' ability of semantic analysis is comparable to those of humans. However, the semantics of spatial expressions also encode the interactive experiences between humans and other entities, which currently cannot be simulated by LLMs. This limits LLMs' performance in some advanced tasks.
- Spatial reasoning tasks pose a significant challenge for LLMs. Current LLMs do not possess the ability to solve complex reasoning problems. Data fine-tuning is unlikely to improve their performance in spatial reasoning tasks.

In future work, our aim is to generate a large amount of structured, high-quality, and knowledge-controllable data to train LLMs to enhance their semantic understanding and reasoning capabilities. To this end, we will further investigate data synthetic methods to incorporate more spatial relations such as 里*(inside)* and 外*(outside)* into the knowledge base and to extend this method to four other subtasks, exploring a technical path for data synthesis driven by knowledge.

Acknowledgements

We would like to acknowledge the contributions of the members of the SpaCE 2024 team, Jiajun Wang, Dan Xing, Xihao Wang, Zihan Zhang, Xiang Cui, and 25 annotators from Peking University, Tsinghua University, etc. We thank Professor Zhifang Sui and her team members, Yixin Yang, and Jingyuan Ma, for helpful discussions on topics related to this work. 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.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

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 *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 367–376. Springer.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

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

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Leonard Talmy. 1983. How language structures space. In *Spatial orientation: Theory, research, and application*, pages 225–282. Springer.

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 *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 547–558.

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 *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 150–158.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*.

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.

A STEP scheme

Role	Description	Example
Spatial Entity	the entity in a spatial relation.	桌子上有一杯牛奶。 There is a glass of milk on the table.
External Place	the position of an entity relative to an external referent.	桌子上有一杯牛奶。 There is a glass of milk on the table.
Internal Place	the position of an entity within the referent.	这杯鸡尾酒的上面是红色的。 The top of this cocktail is red.
Starting Point	the initial position of an entity when it moves.	汤姆从宿舍走到了图书馆。 Tome walked from the dormitory to the library.
Ending Point	the final position of an entity when it moves.	汤姆从宿舍走到了图书馆。 Tome walked from the dormitory to the library.
Path	the path taken by an entity when it moves.	汤姆沿着这条街道往南走去。 Tome walked south along the street.
Direction	the position that an entity towards when it moves.	汤姆沿着这条街道往南走去。 Tome walked south along the street.
Orientation	the position that an entity's surface is facing.	汤姆面朝北往南倒着走。 Tome walked backward to the south while facing north.
Part	the piece of a spatial entity.	汤姆把手放在了桌子上。 Tom placed his hand on the table.
Shape	the physical form of a spatial entity.	请大家排成一条直线。 Please line up in a straight line.
Distance	the amount of space between two entities.	两地相隔3公里。 The two places are 3 kilometers apart.
Spatial Event	the event related to a spatial relation.	汤姆把手放在了桌子上。 Tom placed his hand on the table.
Spatial Time	the time that a spatial expression happens.	汤姆在警察到达之前离开了现场。 Tom left the scene before the police arrived.
Spatial Factivity	the state that a spatial expression is a fact.	汤姆不在家。 Tom is not at home.

Table 6: Introduction of STEP scheme. The underline in the example indicates the words corresponding to the roles. English translations are shown for better readability.