Spatial Representation of Large Language Models in 2D Scene

Anonymous ACL submission

Abstract

Spatial representations are fundamental to human cognition, as understanding spatial relationships between objects is essential in daily life. Language serves as an indispensable tool for communicating spatial information, creating a close connection between spatial representations and spatial language. Large language models (LLMs), theoretically, possess spatial cognition due to their proficiency in natural language processing. This study examines the spatial representations of LLMs by employing traditional spatial tasks used in human experiments and comparing the models' performance to that of humans. The results indicate that LLMs resemble humans in selecting spatial prepositions to describe spatial relationships and exhibit a preference for vertically oriented spatial terms. However, the human tendency to better represent locations along specific axes is absent in the performance of LLMs. This finding suggests that, although spatial language is closely linked to spatial representations, the two are not entirely equivalent.

1 Introduction

The apparent proficiency of large language models (LLMs) in understanding and generating natural language suggests that they may exhibit cognitive abilities akin to those of humans, such as theory of mind and reasoning (Strachan et al., 2024; Rahimi Moghaddam and Honey, 2023; Lampinen et al., 2024; Webb et al., 2023; Gandhi et al., 2023). Consequently, the evaluation of these models has garnered increasing attention, particularly given their expanding applications across domains like code generation and translation (Hong et al., 2023), where minimizing potential errors in their responses is critical. A promising direction for the LLM industry lies in advancing embodied intelligence, which necessitates a robust capacity for spatial understanding (Fan et al., 2024; Zhang et al., 2024). While spatial reasoning is more prominent in the multi-modal domain, where spatial phenomena are often integrated with visual information, it remains essential to investigate spatial representations grounded in natural language to further enable LLMs to support and enhance various aspects of social life.

Spatial relations, which describe the connections between physical objects, are essential for spatial understanding and play a critical role in spatial reasoning. Humans naturally use language to convey spatial relations in everyday life. Trained on extensive natural language datasets, large language models (LLMs) may encode not only spatial linguistic structures but also develop implicit representations of spatial relations, even without direct sensory inputs. Understanding the interaction between spatial language and spatial representations in LLMs can offer valuable insights into how these models process and "comprehend" spatial concepts. Recent studies suggest that LLMs have achieved acceptable proficiency in representing simple cardinal directions and planning navigation tasks (Cohn and Blackwell, 2024; Zhou et al., 2024). However, their performance remains inconsistent and is influenced by factors such as environmental complexity. LLMs tend to excel in addressing basic spatial questions but struggle with more advanced and intricate spatial concepts (Hojati and Feick, 2024). Considering that spatial representations are vital for achieving embodied intelligence and advancing toward artificial general intelligence (AGI), the sensitivity of LLMs to spatial relations in 2D space warrants more comprehensive exploration.

Building on the *CogEval* protocol recently proposed for the general evaluation of LLMs' cognitive capacities (Momennejad et al., 2023), this study aims to assess the spatial intelligence of LLMs. Specifically, we examine the structure of LLMs' representations of spatial relations between two objects within a 7*7 grid scene and evaluate the similarity of these representations to those of

humans using two spatial tasks: the spatial generation task and the spatial rating task. The central research question is whether LLMs can derive visual-like representations from textual input and coordinate descriptions in a 2D space, and to what extent their representations align with those of humans. We evaluate the spatial sensitivity of five LLMs, including state-of-the-art (SOTA) models such as GPT-4, and compare their performance to human behavior data obtained from a previous related study. The research hypothesis posits that LLMs can partially capture 2D spatial representations and exhibit certain features embedded in human spatial language.

The results reveal both similarities and differences between the spatial representations of LLMs and humans. Similar to humans, LLMs more frequently select vertically oriented spatial prepositions to describe spatial relations, as opposed to horizontally oriented terms. State-of-the-art (SOTA) models, such as GPT-4, demonstrate significant proficiency in judging spatial relations, with the exception of accurately identifying the rightward relationship. However, weaker models, such as Llama3-8B, exhibit lower spatial intelligence. Furthermore, the temperature parameter appears to have minimal impact on the models' performance, suggesting that spatial representations may be fundamental to human cognition. Nonetheless, LLMs show limitations in capturing certain subtle characteristics of human spatial cognition, such as the tendency for more precise representations along specific axes.

In summary, the main contributions of this study are as follows:

1) **Adaptation of a standardized experimental paradigm**: We transferred a well-established experimental paradigm from cognitive psychology, used to examine spatial representations in humans, to the evaluation of LLMs. This approach reveals the models' spatial capacities in a 2D scene, which serves as a foundational aspect of spatial intelligence required in more complex environments.

2) **Comparison of spatial representations**: By comparing the spatial representations of five mainstream LLMs with human behavior based on previous studies, this research provides insights into the spatial capabilities of LLMs while also contributing to an indirect understanding of human spatial cognition.

2 Related Works

2.1 Spatial representations and spatial language

Fundamental to cognition in both humans and other animals, spatial representations play a critical role in encoding the geometric properties of objects and the spatial relationships among them. These representations often encompass cognitive models or mental maps that individuals use to mentally visualize and manipulate spatial information. Spatial representations are typically derived from sensory modalities such as vision, hearing, or touch, and they provide crucial information to motor systems and language processing (Landau and Jackendoff, 1993). As a result, frequent translation occurs between spatial representations and spatial language, which generally consists of spatial words or simple phrases.

Spatial language specifically refers to linguistic expressions used to describe spatial properties such as location, orientation, direction, and distance. These expressions are integral to how individuals communicate their understanding of spatial environments. Three basic elements underpin linguistic descriptions of spatial locations: the figure object (the object being located), the reference object, and the spatial relationship between them. Spatial relationships are often encoded through prepositions such as "above" and "below," while both the figure object and the reference object are typically expressed as noun phrases denoting object names. For example, in the sentence "The apple is on the desk," "the apple" functions as the figure object, "the desk" serves as the reference object, and the preposition "on" reflects the spatial relationship between them.

In cognitive psychology, spatial language and spatial representations are intricately linked. Spatial language serves as a key mechanism through which humans convey and process information about space, while spatial representations act as mental constructs that help organize and navigate spatial relationships. It has been proposed that spatial language is grounded in the geometry of visual scenes represented in spatial cognition (Mirzaee et al., 2021). Furthermore, the articulation of spatial concepts in language may influence how they are mentally represented. Empirical evidence suggests that limited exposure to spatial language impairs individuals' performance on non-linguistic spatial tasks, with deaf children showing weaker

abilities to convey spatial relations (Gentner et al., 2013). Cross-linguistic comparisons reveal that similar spatial properties are encoded in both spatial language and spatial representations, suggesting parallels between these two systems (Munnich et al., 2001). Consequently, spatial language can be viewed as a window into the spatial representations that underlie human cognition.

2.2 Spatial understanding of LLMs

Given that LLMs are trained on vast amounts of natural language data, which inherently contains rich spatial language, it is reasonable to infer that these models may acquire a certain degree of spatial understanding. This inference aligns with the established link between spatial representations and spatial language in human cognition. Although LLMs lack access to visual or sensorimotor information, studies suggest that they can partially derive spatial representations from textual input. For instance, LLMs have shown promise in reasoning about simple cardinal directions (CDs), such as "north," "south," "east," and "west," though their performance declines with more complex CDs, such as "northeast" (Cohn and Blackwell, 2024). Additionally, LLMs demonstrate some ability to perform spatial calculations and apply spatial prepositions correctly (Bhandari et al., 2023). Prompting strategies, including Chain-of-Thought (CoT), one-shot or few-shot prompting, and advanced techniques like Visualization-of-Thought (VoT), have been shown to enhance LLMs' spatial reasoning and path-planning capabilities (Wu et al., 2024; Xu et al., 2024). Breaking complex spatial reasoning tasks into smaller, manageable subtasks also improves performance (Peng and Powers, 2024).

However, challenges remain. LLMs' representations of spatial relations can be distorted, often influenced by the hierarchical structure of the environment (Fulman et al., 2024). In many cases, models identify only the nearest cardinal directions, reflecting an associative learning mechanism rather than a robust understanding of spatial concepts. Furthermore, substantial variability exists in their ability to recognize and represent geometric structures, such as squares or hexagons, leaving significant room for improvement (Yamada et al., 2024). The construction of cognitive maps—representations of relational structures in tasks or environments—has also been explored. While cognitive maps are essential for human spatial planning and navigation, systematic evaluations reveal that LLMs often fail in planning tasks, and there is insufficient evidence to support their competence in cognitive map construction (Momennejad et al., 2023).

In summary, while LLMs have made measurable progress in spatial understanding, further advancements are necessary for practical applications in real-world scenarios. Discrepancies and inconsistent findings regarding their spatial representation capacities may stem from the absence of standardized experimental paradigms. To address this, it is essential to compare LLMs' spatial representations with those of humans, using well-established testing paradigms from cognitive science. This approach could provide critical insights into optimizing LLMs' spatial reasoning capabilities while ensuring the scientific rigor and validity of experimental evaluations.

3 Methods

3.1 Spatial representation tasks and datasets generation

The spatial language capabilities of LLMs were examined by requesting the models to describe spatial relationships between given object pairs. Two tasks, adapted from human psychological experiments (Munnich et al., 2001; Hayward and Tarr, 1995), were employed to assess their spatial abilities: (1) generating spatial terms to capture spatial relationships and (2) rating the appropriateness of given statements about object locations in a 2D scene. The procedures for these tasks are as follows.

Spatial Generation Task. In the spatial generation task, LLMs were required to produce spatial terms that described the relationships between two objects on a 2D 7*7 grid (Figure 1). The two objects in each trial were the reference object and the figure object. The reference object was always positioned at the center of the grid, while the figure object could appear in any of the remaining 48 positions, centered in the corresponding cells. Five reference-figure object pairs—"computer-ring", "apple-fish", "bird-tree", "book-pen", and "desk-sofa"—were used to create a diverse dataset. This design resulted in a total of 240 trials (48 positions * 5 object pairs). For each trial, a query prompt was generated using the following template, where [reference], [figure], and [x1, y1] were replaced with specific values for the trial, and [relation] was to be completed by LLMs.

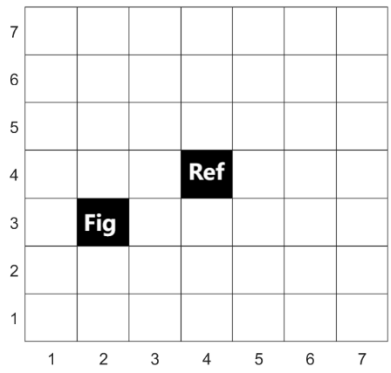


Figure 1: The 7*7 grid plane in spatial representation tasks. The cells noted as 'Ref' and 'Fig' represent the reference object and the figure object respectively, the former of which is always located at the center ([4,4]) while the latter might appear in all the other 48 cells ([2,3] for instance).

Spatial Rating Task. To address the limitation that some LLMs provide only general and coarse terms instead of detailed spatial prepositions in the spatial generation task, a spatial rating task was introduced to further examine their spatial cognition. Unlike the spatial generation task, which required free-form responses, the spatial rating task presented LLMs with predefined statements about the locations of two objects. The models were then required to rate the applicability of these spatial statements on a scale from 1 to 7, where 1 indicated "least appropriate" and 7 indicated "most appropriate." Two reference-figure object pairs—"computer-ring" and "apple-fish"—were selected for this task, combined with four types of spatial relationships: "above," "below," "left," and "right." This design resulted in 384 trials (48 locations * 2 object pairs * 4 relationships). The query prompt for this task followed a specific template, where placeholders were replaced with appropriate values for each trial. The complete set of prompts is available in the supplementary material B.

3.2 LLMs evaluated

The LLMs evaluated in this study include both open-source and closed-source models, incorporating several SOTA models: GPT-3.5-Turbo, GPT-4 (via Azure OpenAI API), Qwen-Turbo, ZhipuAI, and Llama3-8B. To explore the effect of model output variability, experiments were conducted across three temperature settings (0, 0.5, 1) for each LLM. Temperature is a key parameter that controls the uncertainty in the generated content. A higher temperature encourages more diverse and creative responses, but may also reduce reliability and precision. Since this study aims to assess both the creativity and accuracy of LLMs in generating spatial prepositions to describe spatial relationships, varying the temperature allowed for a comprehensive evaluation of the models' ability to balance creativity with precision. Consequently, the spatial representation tasks were repeated across these different temperature settings to account for variability in the models' responses.

3.3 Baseline and evaluation metrics

According to previous studies, most spatial terms used by humans to describe spatial relationships can be categorized into two main types: horizontally oriented and vertically oriented prepositions (Munnich et al., 2001; Hayward and Tarr, 1995). Specifically, horizontally oriented prepositions (e.g., "above" and "below") describe the position of the figure object relative to the reference object in terms of horizontal relations, while vertically oriented prepositions (e.g., "left" and "right") capture vertical relationships between the two objects.

For the spatial generation task, the proportion of horizontally and vertically oriented spatial prepositions used in the LLMs' responses was computed for each cell in the 7*7 grid, with averages taken across different scenarios. Since the concept of 'front' or 'behind' does not apply on a 2D plane, responses involving such prepositions were considered nonsensical or ineffective. Additionally, as neither angles nor compass directions were allowed in the prompts to LLMs, the models' adherence to the instructions was evaluated by examining the proportion of invalid responses. Given that LLMs often use both horizontal and vertical spatial terms simultaneously when describing spatial relationships, the first spatial preposition that appeared in the models' responses was taken as the primary indicator of their axial preference.

In the spatial rating task, LLMs' ratings of statements regarding the spatial relationships between the figure object and the reference object were averaged across all scenarios for each location. To better understand LLMs' basic spatial perception, the 7*7 grid was divided into four 3*7 sub-grids (up, down, left, and right relative to the centrally positioned reference object at [4,4]). The ratings for each sub-grid were then compared to those from the other three sub-grids. This analysis aimed to

Temperature	0	0.5	1
GPT-4	91.25%	94.17%	88.75%
GPT-3.5-Turbo	40.83%	43.33%	39.58%
Qwen-Turbo	71.67%	65.83%	54.17%
ZhipuAI	95.42%	94.17%	93.33%
Llama3-8B	28.75%	25.83%	30.42%

Table 1: Validness of LLMs' responses on the spatial generation task.

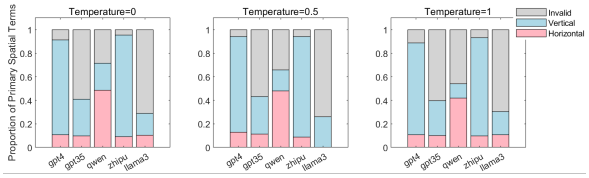


Figure 2: Proportions of different types spatial prepositions shown in models' responses at first. GPT-4 and ZhipuAI show better validness. Most LLMs except Qwen-Turbo tend to prefer vertically oriented spatial terms relative to horizontally oriented spatial terms.

reflect the models' ability to recognize and distinguish primary axial relations.

In both spatial tasks, LLMs' performance was compared to that of humans based on a previous related study (Hayward and Tarr, 1995). Specifically, the Euclidean distance between the rating matrices of LLMs and humans was calculated and normalized to quantify the difference in performance. The relative difference, denoted as $Diff_{\text{norm}}$, is formulated as follows. A smaller value of $Diff_{\text{norm}}$ indicates a closer match between the performance of the models and humans.

$$Diff_{\text{norm}} = \frac{\|\text{LLM}_{\text{matrix}} - \text{Human}_{\text{matrix}}\|_F}{\max\left(\|\text{LLM}_{\text{matrix}}\|_F, \|\text{Human}_{\text{matrix}}\|_F\right)}$$

(F means Frobenius norm; *matrix* denotes proportion or mean rating.)

4 Results

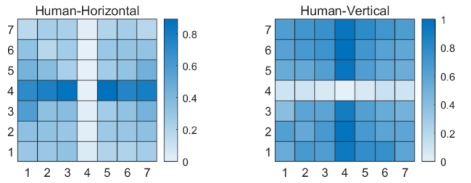
4.1 Spatial representations of LLMs are directionally imbalanced and vertically more efficient

The spatial prepositions selected by LLMs to describe the spatial relationships between the figure object and the reference object exhibit considerable diversity, particularly in more advanced models. Horizontally oriented spatial terms include "left", "right", "beside", and "next to", while vertically oriented terms encompass "above", "below", "up", "low(er)", "ahead", and "beyond". In addition to these axial prepositions, LLMs' responses also contain some non-axial spatial terms, such as "diagonal", "southwest", "behind", and "near". These non-axial terms, though less frequent, are considered inappropriate as they do not adhere to the instructions specifying axial relationships in a 2D grid. Responses incorporating these terms were therefore coded as invalid.

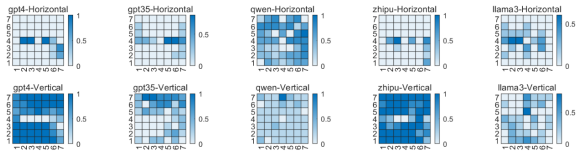
The proportions of invalid responses from the five LLMs under three different temperature settings are presented in Table 1. This data reveals that SOTA models such as GPT-4 and ZhipuAI consistently provide more accurate and effective spatial representations, more closely aligning with human-like spatial reasoning. These models also demonstrate a preference for describing spatial relationships along axial directions. Moreover, vertically oriented prepositions are more frequently chosen as the primary descriptors, a trend also observed in human spatial language. The proportions of three types of spatial prepositions (horizontal, vertical, and others) in LLMs' responses across varying temperature levels are shown in Fig. 2. The results suggest that temperature settings only have a subtle effect on the models' performance in the spatial generation task. Notably, most models, with the exception of Qwen-Turbo, tend to use vertically oriented spatial prepositions as their primary means of describing spatial relationships between objects on a 2D plane.

4.2 Resemblance of LLMs to humans in preference of vertical spatial terms

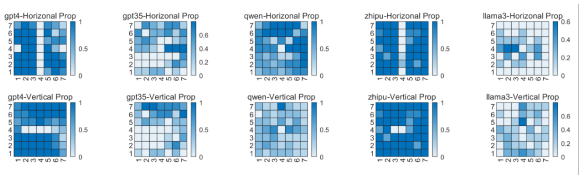
The proportions of horizontally and vertically oriented spatial prepositions that appeared first in the models' responses at each location are compared with human performance, as derived from the previous study (Hayward and Tarr, 1995). As shown in Fig. 3(a), humans exhibit a clear axial preference when describing spatial relationships. Specifically, horizontally or vertically oriented spatial prepositions are more likely to be chosen as the primary descriptors when the figure object is positioned near the corresponding axis. However, the patterns in the LLMs' responses to spatial term generation exhibit notable differences (Fig. 3(b)). All models accurately generate horizontal spatial prepositions along the x-axis centered on the reference object, except for Qwen-Turbo. The horizontal prepositions produced by Qwen-Turbo are scattered and lack a clear, consistent pattern.



(a) Humans' choice of each type of spatial terms.



(b) Performance of five LLMs on spatial preposition preference at all cells except the center.



(c) Distribution of horizontal and vertical spatial prepositions appeared in LLMs' responses in the spatial generation task.

Figure 3: Primacy of horizontal and vertical spatial prepositions in LLMs' responses at each location on the 7*7 grids. Considering the subtle influence of temperature on LLMs' generation performance, the temperature underlying results displayed here is 0, whereas results of the other two situations (i.e. 0.5 and 1) are available in Appendix Fig.S1 and S2.

On the other hand, both GPT-4 and ZhipuAI appear to overemphasize encoding spatial relationships in the vertical direction, as they generate a notably higher proportion of vertical spatial prepositions compared to other models. GPT-3.5-Turbo, on the other hand, tends to produce more vertical prepositions when the figure object is located above the reference object. In contrast, Qwen-Turbo still exhibits no discernible pattern in the distribution of vertical spatial prepositions. Llama3-8B, however, demonstrates a clear axial effect, with consistent performance in both vertical and horizontal directions.

When considering the frequency of horizontal and vertical spatial terms combined in the models' responses—without focusing on their primacy—results show that GPT-4 and ZhipuAI encode both horizontal and vertical relationships comprehensively (Fig. 3(c)). These models provide a dense representation, employing spatial terms in both directions across nearly every position. Llama3-8B's performance mirrors the findings in the primacy analysis discussed earlier. In contrast, no clear pattern emerges in the responses of GPT-3.5-Turbo and Qwen-Turbo.

Temperature	0	0.5	1
GPT-4	0.787	0.785	0.711
GPT-3.5-Turbo	0.686	0.722	0.713
Qwen-Turbo	0.579	0.547	0.570
ZhipuAI	0.770	0.776	0.763
Llama3-8B	0.775	1	0.766

Table 2: Horizontal difference between the performance of LLMs and humans.

Temperature	0	0.5	1
GPT-4	0.331	0.378	0.311
GPT-3.5-Turbo	0.662	0.632	0.695
Qwen-Turbo	0.689	0.754	0.843
ZhipuAI	0.360	0.346	0.352
Llama3-8B	0.778	0.774	0.764

Table 3: Vertical difference between the performance of LLMs and humans.

The disparity between the performance of LLMs and humans in the spatial generation task is further computed and presented in Table 2 (for horizontal directions) and Table 3 (for vertical directions). In terms of human-like performance, the spatial representations of both GPT-4 and ZhipuAI are generally more similar to humans in the vertical direction, as their normalized difference ($Diff_{\text{norm}}$ index) is lower than 0.5, outperforming all other models. However, in the horizontal direction, the normalized difference between all models and humans exceeds 0.5, regardless of the temperature setting. Therefore, only SOTA models like GPT-4 resemble humans in choosing vertically oriented spatial prepositions to characterize spatial relationships.

4.3 SOTA LLMs demonstrate a deficiency in representing rightward spatial relationships

To gain a more nuanced understanding of LLMs' spatial representation, models were tasked with rating the applicability of statements describing four types of spatial relations between the reference object and the figure objects. A comparison was made between the average ratings of spatial statements describing relations where the figure objects are located in the corresponding subgrid area (e.g., the "above" relation used for figure objects in the upper

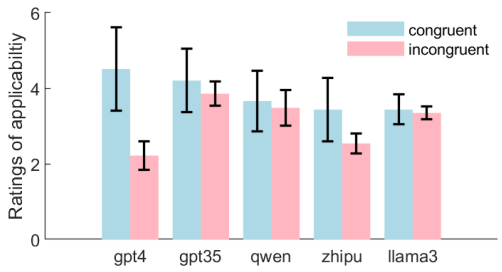


Figure 4: LLMs' ratings on the applicability of spatial relations in both congruent and incongruent cases. SOTA models, namely GPT-4 and ZhipuAI, significantly provided higher ratings for spatial descriptions that were congruent with the ground truth, whereas the performance of the other three models was comparatively weaker, likely due to their insensitivity to spatial relations. The error bars represent the standard error of the mean (SEM). The temperature setting underlying the results presented here is 0, with the other two cases (i.e. 0.5 and 1) detailed in the Appendix Fig.S3.

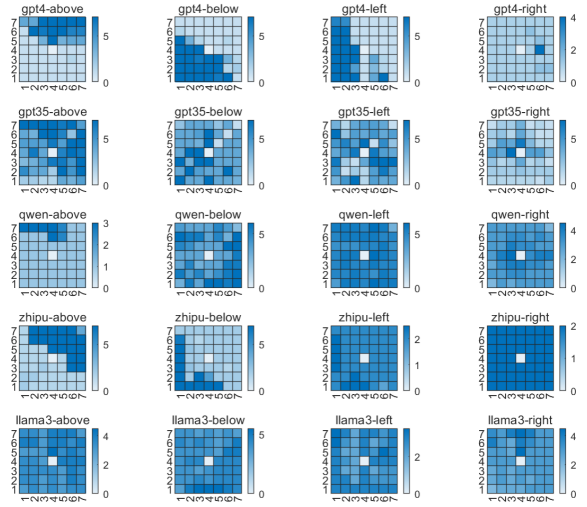


Figure 5: Performance of five LLMs on the spatial rating task. Four types of spatial relations are involved in the rating process, namely "above", "below", "left", and "right". The intensity of color bars represents models' evaluation of the appropriateness of the spatial statements given to them. Ratings range from 1 to 7, where higher scores indicate better applicability. Temperature underlying the results shown here is 0, leaving the other two cases (i.e. 0.5 and 1) available in the Appendix Fig.S4.

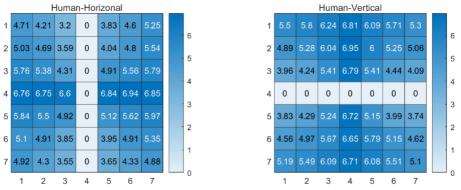
3*7 subgrid) and those in the other three subgrids. As shown in Fig. 4, GPT-4 and ZhipuAI exhibit strong performance in rating the applicability of spatial descriptions, as they can effectively distinguish between descriptions that are congruent or incongruent with the actual spatial relationships. In contrast, the other three models—GPT-3.5-Turbo, Qwen-Turbo, and Llama3-8B—show significant insensitivity to spatial relations.

The results reveal that GPT-4 performs remarkably well on three types of spatial relations—namely "above", "below", and "left". However, this performance does not extend to the "right" relation, where its accuracy drops 5. Similarly, ZhipuAI also provides relatively accurate ratings for the "above" and "below" relations. Qwen-Turbo shows partial success, particularly when the "above" relation is used to describe spatial relationships between a figure object situated in the upper locations and the reference object. Other models, including GPT-3.5-Turbo and Llama3B, exhibit significant weaknesses in representing almost all spatial relations. Interestingly, even models that perform well in recognizing basic spatial relations still show some overlap in representing adjacent spatial relations, often spreading their ratings around the vertex of the 7*7 grid. Specifically, GPT-4's ratings for the appropriateness of "below" descriptions are higher in the bottom-left area rather than exclusively in the bottom area, and a similar pattern is observed in ZhipuAI's performance.

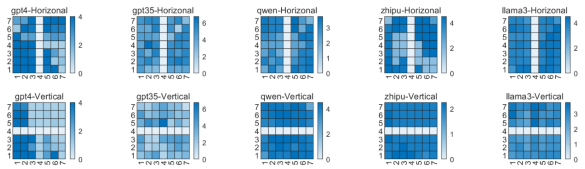
LLMs' performance in rating the four types of spatial relations ("above", "below", "left", and "right") is averaged across horizontal and vertical directions. Specifically, the "above" and "below" relations are combined as representing the vertical axis, while the "left" and "right" relations are categorized under the horizontal axis. The resulting rating matrix is then compared with human ratings from a previous study (Hayward and Tarr, 1995). Human ratings exhibit a clear axial pattern, with ratings highest when the figure object and the reference object are aligned on the same axis, gradually decreasing as the figure object moves away from the central axis (Fig.6(a)). However, this axial pattern is not observed in any of the LLMs' performance (Fig.6(b)).

5 Discussion

LLMs' spatial representation abilities are evaluated through two tasks adapted from cognitive psychology: the spatial generation task and the spatial rating task, which test the models' capacity to describe and judge spatial relationships on a 2D scene. The observed directional imbalance in the spatial generation task mirrors human tendencies (Munnich et al., 2001; Hayward and Tarr, 1995), where vertical prepositions like "above" and "be-



(a) Humans' ratings.



(b) Five LLMs' ratings with the temperature set as 0, leaving the other two cases (0.5 and 1) available in the Appendix Fig.S5.

Figure 6: Rating performance of Humans and LLMs on each location where the figure object is situated at around the reference object, averaged across horizontal and vertical directions respectively.

low" are used more often than horizontal ones. The lower frequency of horizontal terms suggests that LLMs' spatial depictions along the horizontal axis are coarser. This pattern is likely rooted in the effect of gravity on human daily life (Stahn et al., 2020; Lacquaniti et al., 2015; Levinson, 1996), where vertical terms tend to be more prevalent than their horizontal counterparts. Consequently, LLMs are indirectly shaped by this bias through human-oriented language.

In terms of heterogeneity in LLMs' behavior, more advanced models appear to be significantly more proficient in spatial representations. Specifically, SOTA models such as GPT-4 demonstrate greater accuracy in judging spatial relationships between objects and exhibit higher geometric richness in their choice of spatial prepositions when generating spatial descriptions compared to GPT-3.5-Turbo and Llama3-8B. This finding suggests that spatial representations can indeed be derived from spatial language, and LLMs with superior overall performance are more likely to possess enhanced spatial abilities. However, even the best-performing LLMs still fall short of perfection, indicating the need for further precision in practical applications. Additional pretraining with automatically generated spatial datasets could potentially improve LLMs' spatial reasoning (Mirzaee et al., 2021).

The influence of temperature on LLMs' performance in both spatial tasks appears minimal, as no significant differences are observed in models' choice of spatial terms or their judgment of spatial relationships under different temperature levels (0, 0.5, and 1). Since temperature controls the randomness of model responses (Zhu et al., 2024), the insensitivity to temperature variations in spatial tasks may suggest the fundamental constancy of spatial cognition in human life. This finding aligns with studies indicating that changes in temperature have little effect on LLMs' problem-solving performance (Renze and Guven, 2024). Interestingly, all LLMs, including SOTA models like GPT-4 and ZhipuAI, fail to accurately represent rightward spatial relationships, highlighting a bias in the models' training datasets, where leftward relationships seem to be more prevalent in natural language. This phenomenon, to our knowledge, is being reported for the first time and warrants further investigation. One possible explanation is that, given most people are right-handed, leftward spatial relationships may be more intuitive and commonly used in practice.

It is also worth noting that LLMs fail to capture certain subtle characteristics of human spatial representations, such as axial salience. Cognitive psychology research has shown that humans tend to exhibit more accurate spatial representations in regions near the central axis (Hayward and Tarr, 1995), with accuracy decreasing as the distance from the axis increases. However, this tendency is absent in LLMs' performance, highlighting the limitations of models that excel at detecting regularities and generating words linearly, yet struggle with visualizing situations in a 2D space. This suggests that spatial language does not equate to spatial representation, and there may be an upper limit to the spatial representation capabilities of linguistic models.

6 Conclusion

Both similarity and difference exist between spatial representations of LLMs and humans. On one hand, LLMs resemble humans in the choice of spatial prepositions while describing spatial relationships between two objects on a 2D scene. Vertically oriented spatial terms are preferred by LLMs relative to horizontal terms, which is consistent to humans' performance and probably the reflection of gravity. On the other hand, finer representations along axis in humans do not appear in LLMs' spatial cognition, indicating that LLMs actually fail to capture some subtle facets in human language.

Limitations

One limitation of this study is the simplification of the spatial tasks, which may not fully capture the intricate and multifaceted nature of human spatial cognition. While the tasks provide valuable insights into LLMs' spatial reasoning, they may not account for the complex, dynamic, and context-dependent factors that influence human spatial processing. Additionally, the evaluation of LLMs' spatial representations is based on textual input, which inherently may not capture the full range of spatial nuances that could be conveyed through visual input. Visual representations are known to play a crucial role in human spatial reasoning, and relying solely on text may limit the models' ability to develop a truly rich spatial understanding. Moreover, this study does not consider the potential impact of other hyperparameters—such as model architecture, training data, and optimization strategies—on LLMs' spatial performance. The tuning of these hyperparameters could influence the models' ability to generalize across different spatial tasks and scenarios.

Future research should aim to investigate LLMs' spatial representations in more complex, real-world scenarios that more closely mirror human cognition, and use a broader set of evaluation metrics that encompass both quantitative and qualitative measures. This will enable a more nuanced understanding of the models' spatial reasoning abilities. Furthermore, it would be valuable to explore techniques to enhance LLMs' spatial representations, such as the use of effective prompting strategies, incorporating multimodal inputs (e.g., images or videos), or leveraging multi-agent collaboration. These approaches could potentially mitigate current limitations and enable LLMs to achieve more sophisticated, human-like spatial reasoning.

References

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are Large Language Models Geospatially Knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Anthony G. Cohn and Robert E. Blackwell. 2024. Evaluating the Ability of Large Language Models to Reason about Cardinal Directions. In *The 16th Conference on Spatial Information Theory*. arXiv.

Haolin Fan, Xuan Liu, Jerry Ying Hsi Fuh, Wen Feng Lu, and Bingbing Li. 2024. Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. *Journal of Intelligent Manufacturing*, pages 1–17.

Nir Fulman, Abdulkadir Memduhoğlu, and Alexander Zipf. 2024. Distortions in Judged Spatial Relations in Large Language Models. *Preprint*, arXiv:2401.04218.

Kanishk Gandhi, Jan-Philipp Fraenken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Dedre Gentner, Asli Özyürek, Özge Gürcanli, and Susan Goldin-Meadow. 2013. Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, 127(3):318–330.

William G. Hayward and Michael J. Tarr. 1995. Spatial language and spatial representation. *Cognition*, 55(1):39–84.

Majid Hojati and Rob Feick. 2024. Large Language Models: Testing Their Capabilities to Understand and Explain Spatial Concepts (Short Paper). *LIPIcs, Volume 315, COSIT 2024*, 315:31:1–31:9.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *Preprint*, arXiv:2308.00352.

Francesco Lacquaniti, Gianfranco Bosco, Silvio Gravano, Iole Indovina, Barbara La Scaleia, Vincenzo Maffei, and Myrka Zago. 2015. Gravity in the Brain as a Reference for Space and Time Perception. *Multisensory Research*, 28(5-6):397–426.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Barbara Landau and Ray Jackendoff. 1993. What and where in spatial language and spatial cognition? *Behavioral and Brain Sciences*, 16(2):255–265.

Stephen C. Levinson. 1996. Language and space. *Annual Review of Anthropology*, 25(1):353–382.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. 2023. Evaluating Cognitive Maps and Planning in Large Language Models with CogEval. *Advances in Neural Information Processing Systems*, 36:69736–69751.

Edward Munnich, Barbara Landau, and Barbara Anne Dosher. 2001. Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, 81(3):171–208.

William Peng and Sam Powers. 2024. LLMs and Spatial Reasoning: Assessing Roadblocks and Providing Pathways for Improvement. *Journal of Student Research*, 13(2).

Shima Rahimi Moghaddam and Christopher Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *Preprint*.

Matthew Renze and Erhan Guven. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. *Preprint*, arXiv:2402.05201.

Alexander Christoph Stahn, Martin Riemer, Thomas Wolbers, Anika Werner, Katharina Brauns, Stephane Besnard, Pierre Denise, Simone Kühn, and Hanns-Christian Gunga. 2020. Spatial Updating Depends on Gravity. *Frontiers in Neural Circuits*, 14:20.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. 2024. Evaluating Large Language Models on Spatial Tasks: A Multi-Task Benchmarking Study. *Preprint*, arXiv:2408.14438.

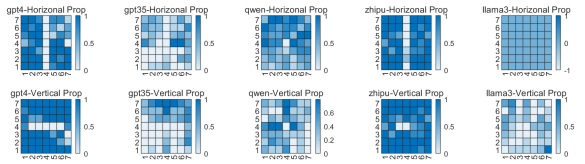
Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. Evaluating Spatial Understanding of Large Language Models. *Preprint*, arXiv:2310.14540.

Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, Peijin Guo, and Leo Yu Zhang. 2024. BadRobot: Manipulating Embodied LLMs in the Physical World. *Preprint*, arXiv:2407.20242.

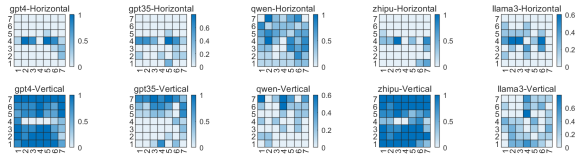
Gengze Zhou, Yicong Hong, and Qi Wu. 2024. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7641–7649.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Jia Li, Zhi Jin, and Hong Mei. 2024. Hot or Cold? Adaptive Temperature Sampling for Code Generation with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):437–445.

A Supplementary Results

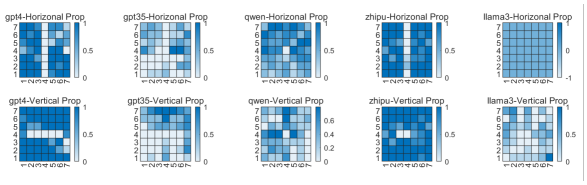


(a) temperature = 0.5

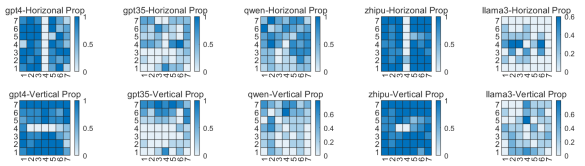


(b) temperature = 1

Figure S1: Performance of five LLMs (i.e. GPT-4, GPT-3.5-Turbo, Qwen-Turbo, ZhipuAI, and Llama3-8B) on spatial preposition preference at all cells except the center ([4,4]).



(a) temperature = 0.5



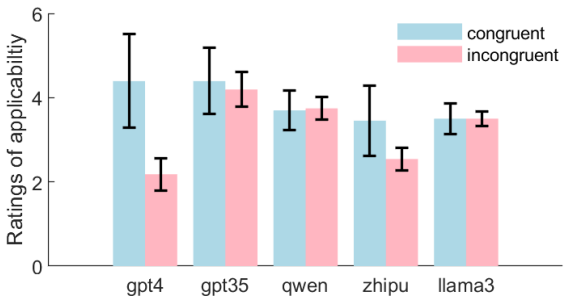
(b) temperature = 1

Figure S2: Distribution of horizontal and vertical spatial prepositions appeared in LLMs' responses in the spatial generation task.

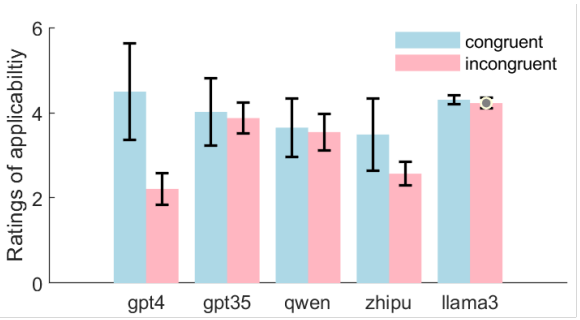
B Prompts for Spatial Representation Tasks

Prompt templates for the spatial generation task and the spatial rating task are provided below.

1) **Spatial Generation Task**: *"On a 7*7 grid, the bottom left corner is [1,1], while the top right corner is [7,7]. The [figure] is at [x1, y1], while the [reference] is at [4,4]. So, the [figure] is [relation] the [reference]. Please give appropriate spatial prepositions to replace the [relation]. Avoid using compass directions, a clock face, or the degree of angle."*



(a) temperature = 0.5

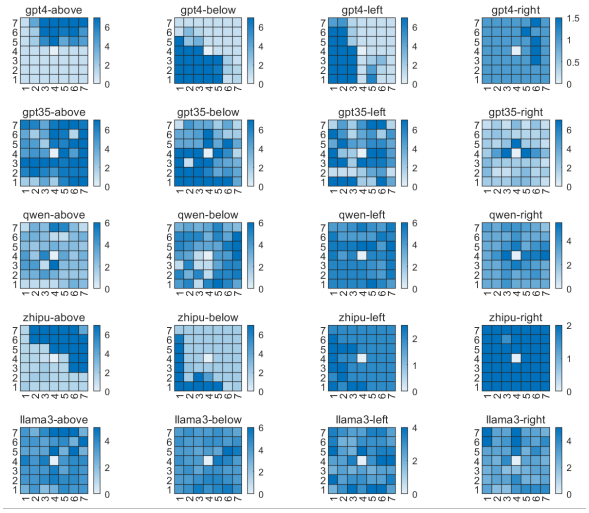


(b) temperature = 1

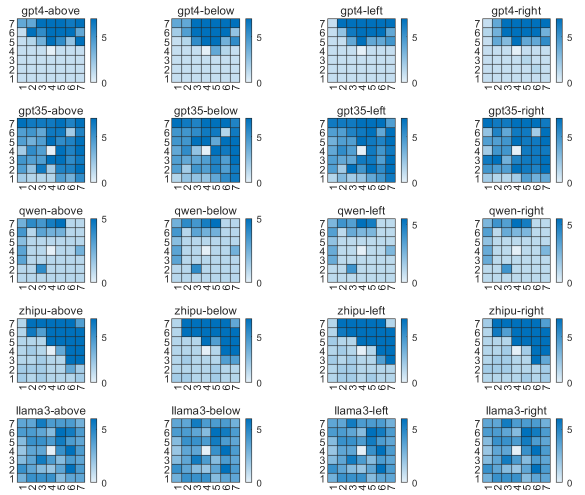
Figure S3: Performance of five LLMs on the spatial rating task. LLMs' ratings are compared between the congruent and incongruent conditions where the descriptions of spatial relations between the figure object and the reference object either correspond to the truth or not.

2) **Spatial Rating Task**: *"On a 7*7 grid, the bottom left corner is [1,1], while the top right corner is [7,7]. The [figure] is at [x1, y1], while the [reference] is at [4,4]. Please rate the appropriateness of the following statement on a scale of 1 to 7, where 1 is the least appropriate and 7 is the most appropriate. The Statement is: The [figure] is [relation] the [reference]."*

The specific prompt with placeholders replaced by actual items is available on this anonymous website Spatial Representations of LLMs).

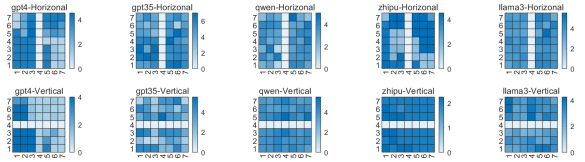


(a) temperature = 0.5

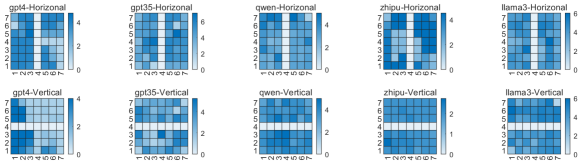


(b) temperature = 1

Figure S4: Performance of five LLMs on the spatial rating task.



(a) temperature = 0.5



(b) temperature = 1

Figure S5: LLMs' ratings on each location where the figure object is situated at around the reference object.