Exploring Spatial Understanding Capability in Large Language Models: Proficiency in Layout Generation sans Visual Perception

Anonymous ACL submission

Abstract

Large language models (LLMs) consistently demonstrate superior performance in various natural language processing (NLP) tasks. However, research on their abilities to process visual and spatial information, which is essential for understanding visually-rich documents (VRDs), is limited. This paper presents a pioneering study and benchmark specifically designed to evaluate the spatial competencies of LLMs in the context of VRDs. Our assessment covers a comprehensive range of dimensions, including spatial perception, positional prediction, information extraction, and layout generation. The results show that despite the lack of inherent visual perception mechanisms in LLMs, these models can effectively infer spatial relationships within VRDs. In addition, we propose a layout-aware learning strategy with off-the-shelf LLMs that can significantly improve their performance. Our results indicate a significant contribution to the field of document intelligence, confirming the effectiveness of our methodology and pointing the way for future research in document analysis.

1 Introduction

Large Language Models (LLMs) exhibit remarkable capabilities in understanding and generating natural language text across diverse applications (Brown et al., 2020; Liu et al., 2024). In cognitive linguistics, language not only serves as a bridge for communicating thoughts and emotions, but also acts as a semantic proxy for the physical world, inherently reflecting the entities and processes within it (OKeefe and Dostrovsky, 1971; Lakoff, 1993). This intrinsic linkage between language and reality offers LLMs a distinct vantage point to engage with the world through textual analysis, enabling a nuanced and comprehensive understanding of real-world phenomena and even fundamental concepts, e.g., color (Abdou et al., 2021), direction (Patel and Pavlick, 2022), shape (Yamada et al., 2023), tabular structures (Tang et al., 2023) and geography (Roberts et al., 2023). These capabilities underscore the versatility of LLMs in interacting with the world through the lens of language and highlight their versatility and applicability across diverse domains.

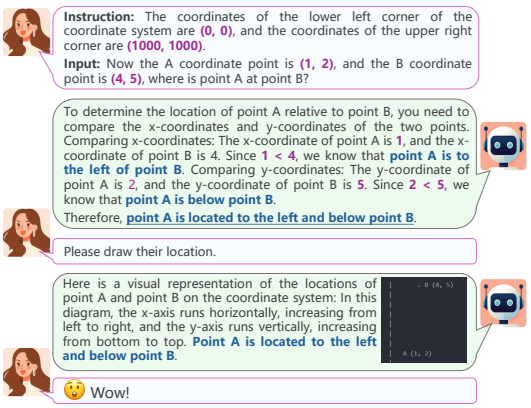

Figure 1: GPTs have a strong proficiency in understanding and generating document layouts.

The example, as shown in Figure. 1, demonstrates that LLMs have remarkable abilities to comprehend content and generate layouts within a document page. While recent advancements in leveraging LLMs for information retrieval (He et al., 2023; Liu et al., 2024) and visual grounding (Chen et al., 2024; Zhang et al., 2024), a comprehensive systematic framework addressing the spatial understanding capabilities of LLMs within visually-rich documents (VRDs) remains absent. Unlike recent layout-aware document AI models (Xu et al., 2020, 2021; Huang et al., 2022; Li et al., 2021; Yu et al., 2023; Lee et al., 2023) that can utilize multimodal information, LLMs encounter considerable challenges in recognizing named entities and their relationships without direct visual input. Therefore, this limitation underscores the necessity for further exploration and development in this research direction.

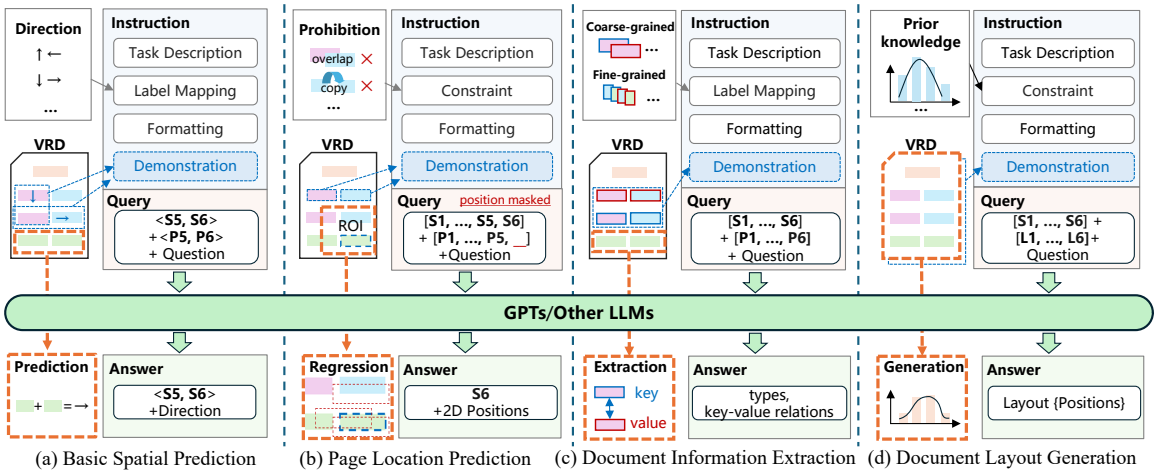

Figure 2: Proposed framework for evaluating LLMs' spatial understanding. **S**: text segment, **L**: label, **P**: position.

In light of these considerations, we introduce an evaluation framework to systematically assess the spatial comprehension capabilities of LLMs in the context of VRDs. Our proposed tasks are meticulously crafted to progress from elementary to advanced, thereby providing a thorough exploration of LLMs' capacity to understand, extract, reason, and generate spatial information. Furthermore, we introduce a novel layout-aware approach building upon recent advancements in contextual learning (Wei et al., 2023; Bang et al., 2023). This approach empowers LLMs with the capability to effectively incorporate layout features into their processing, thereby bolstering their spatial comprehension on VRDs. This achievement is crucial as it provides deep insights into how LLMs can effectively navigate and connect information in such an information-sparse artificially virtual environment, ultimately broadening their applicability in the future unexplored research.

The contributions can be summarized as follows: 1) We present an evaluation framework that assesses the spatial and generative competencies of LLMs in VRDs across multi-dimensions. 2) We perform an exhaustive analysis of prevalent LLMs, elucidating their spatial and generative performance in VRDs. 3) We introduce a layout-aware learning strategy to integrate spatial features and patterns, markedly enhancing LLMs' performance.

2 Framework

Figure. 2 illustrates the proposed evaluation framework, which encompasses four distinct schemes: 1) **Basic spatial perception** (BSP) is designed to evaluate whether LLMs comprehend coordinate systems and accurately interpret the spatial semantics of directional terms, thereby assessing their "sense of direction." We evaluate LLMs' capacity to comprehend coordinates and relative positions through two tiered experiments: 1) pixel-to-pixel (P2P) for pinpointing relative pixel positions, and 2) bbox-to-bbox (B2B) for ascertaining positions between bounding boxes. This evaluation includes both coarse-grained and fine-grained tasks to probe the depth of LLMs' spatial understanding: relaxed and exact. Relaxed to predict the basic four directions (top, bottom, left, right), while exact pinpointing the eight extended directions (e.g., top-left, bottom-left, top-right, bottom-right). 2) **Page location prediction** (PLP) functions as a regression task that predicts the spatial coordinates of a current text field by utilizing the semantics of spatially proximate fields and the broader layout context. In VRD images, text segments, such as named entities, often follow discernible positional arrangement patterns, including vertical or horizontal alignments. PLP involves obscuring specific field positions during testing and challenging LLMs to deduce the missing coordinates of a specified text segment, which answers: "Where should this text content appear in the document?" This mimics real-world conditions where documents may be incomplete or poorly structured, necessitating the use of contextual and logical inference to interpret document architecture. 3) **Document information extraction** (DIE) involves two phases: Semantic Entity Recognition (SER) and Relation Extraction (RE). SER aims to identify and categorize enti-

ties within a document, with a distinction between coarse-grained (bbox) and fine-grained (token) approaches. While coarse-grained SER identifies entity types such as key and value, fine-grained SER classifies entities into category-specific types. RE furthers this by predicting relationships between entities. 4) **Document layout generation** (DLG) focuses on whether LLMs can produce layouts aligned with category-specific probability distributions. We ask LLMs to generate layouts for VRDs, employing probability distributions of two-dimensional variables to gauge their layout generation prowess. The task hinges on comparing LLMs' predicted x- and y-directional distributions with actual ones to evaluate the accuracy of layout generation. Thus, this task examines the capacity of LLMs to harness prior layout design knowledge for logical entity positioning from a holistic viewpoint. Therefore, BSP and PLP are designed to evaluate the models' ability to understand local and global positional information, respectively. Concurrently, DIE and DLG are tasked with assessing the models' grasp of semantic relationships among entities, both on a local and a holistic level. To construct our benchmark, we collect and re-annotate data from the existing FUNSD (Jaume et al., 2019) and SEAB (Wang et al., 2023) datasets, which finally contains 466 documents. Following He et al. (2023), we standardize to facilitate subsequent data processing and evaluation from LLMs' output. We meticulously construct instructions that contains: 1) task description: a clear and concise explanation of the task expected to perform, which outlines the objective and provides any necessary background information; 2) label mapping/constraint: it involves defining the set of possible labels or categories that the LLM can predict and are guided within a specific scope; 3) formatting: the structure and presentation of the input and output data; 4) demonstration: the inclusion of examples or demonstrations that illustrate how the task should be performed, especially for the few-shot learning scenario. More details refer to Appx. A.

3 Experimental Results

We primarily utilize the GPT3.5 (`gpt3.5-turbo`) and GPT4 (`gpt4-0125-preview`) models for our experiments, as they are the only two capable of successfully executing the full range of our experimental protocols.

3.1 BSP: Does the LLM sense direction?

Table 1 shows that both GPT3.5 and GPT4 excel in pixel-level perception. It suggests that with the right enhancements, both models could significantly improve their spatial reasoning capabilities in two-dimensional contexts. However, GPT4 underperforms in B2B prediction for relaxed directions. See Appx. 3 for visualization.

		Exact			Relaxed		
		P %	R %	F1 %	P %	R %	F1 %
P2P	GPT3.5	58.99	62.33	60.61	98.64	92.27	95.35
	GPT4	**99.33**	**98.40**	**98.86**	**99.53**	**98.60**	**99.06**
B2B	GPT3.5	65.72	43.47	52.33	89.55	**68.67**	**77.73**
	GPT4	**87.91**	**62.70**	**73.20**	**89.90**	64.10	74.84

Table 1: Evaluation results of exact and relaxed position predictions for LLMs with zero-shot inference (*w/o* demonstration). Bold indicates the better results.

3.2 PLP: Can LLMs fill in the gaps?

We challenge LLMs to leverage their inherent knowledge of key-value pair layouts and semantic similarities to deduce the spatial locations of omitted fields in VRDs, devoid of visual cues (only text input provided; no images.) We evaluate the results utilizing the reversed normalized Euclidean distance. Table 2 shows that LLMs equipped with a visual channel, such as GPT4 and CogVLM (Wang et al., 2024), outperform purely textual models as expected. Notably, GPT3.5 still scores above 80 on average, indicating that textual LLMs, when guided by effective prompts, can accurately deduce the spatial locations of semantic entities despite lacking a visual channel.

	Vision	T1	T1-v	T2	T2-v	T3	T3-v	Avg
CogVLM7B	✓	85.8	81.3	63.9	58.2	67.0	66.3	70.4
GPT3.5	×	83.9	80.5	**89.8**	84.0	76.4	80.0	82.4
GPT4	✓	**92.4**	**83.9**	88.1	**89.9**	**95.7**	**86.0**	**89.3**

Table 2: Scores of page location prediction . Note: T1, T2, T3 represent three typical types; "-v" denotes value types, with higher scores indicating greater precision.

3.3 DIE: Does information extraction need spatial information?

Results in Table 4 indicate that GPT4 significantly outperforms the average in fine-grained SER task, underscoring GPT4' domain-agnostic strengths. Moreover, GPT3.5 adeptly categorizes

Model	Methods	Out-of-domain				In-domain (*w/o* Other)				In-domain (*w* Other)			
		P %	R %	F1 %	ΔF1	P %	R %	F1 %	ΔF1	P %	R %	F1 %	ΔF1
GPT3.5	Zero-shot	52.52	51.33	51.92		76.00	47.88	58.75		53.63	53.16	53.39	
	Layout-aware (*our*)	**70.36**	**71.22**	**70.79**	**20.43**↑	**88.98**	**771.33**	**79.18**	**20.43**↑	**70.36**	**71.22**	**70.79**	**17.40**↑
GPT4	Zero-shot	-	-	-		88.58	74.46	80.91		65.76	65.52	65.64	
	Layout-aware (*our*)	-	-	-		**94.55**	**85.49**	**89.79**	**8.88**↑	**78.71**	**78.10**	**78.40**	**12.76**↑

Table 3: DIE using layout-aware learning upon GPTs surpasses the baseline systems by a large margin. In 'out-of-domain' column, GPT4 outputs serve as ground truths for assessing GPT3.5's capabilities.

60.52% of out-of-domain entities (bbox), evidencing their proficiency in knowledge transfer and creativity. Although performance decreases in the RE task, in-domain improvements remained substantial. GPT4 notably outpaced GPT3.5, especially in grasping relationships.

			In-domain			Out-of-domain		
			P%	R%	F1%	P%	R%	F1%
SER	*coarse*	GPT3.5	79.31	48.93	60.52	79.31	48.93	60.52
		GPT4	**91.18**	**64.83**	**75.78**	-	-	-
	fine	GPT3.5	90.98	57.28	70.28	59.17	57.84	58.50
		GPT4	**92.70**	**77.92**	**84.67**	-	-	-
RE		GPT3.5	76.11	47.88	58.75	52.22	51.33	51.92
		GPT4	**88.58**	**74.56**	**80.91**	-	-	-

Table 4: Results of SER and RE using GPTs. The term "Out-of-domain" refers to predictions made by GPTs for instances categorized under `Other`. Conversely, "In-domain' excludes the `Other` category and focuses solely on predefined labels.

3.4 DLG: Are LLMs good layout designer?

Both GPTs excel in overall prediction accuracy, with their distributions closely matching the actual ones. Table. 5 details our quantitative assessment using KL divergence and JS divergence against the benchmark distribution. GPTs show robust generation skills without any reference or guidance, with GPT4 outperforming, compared to CogVLM which equipped with visual perception. The result suggesting that semantic reasoning is essential for comprehensive layout generation. Appx. 9 illustrates GPTs adherence to prior principles, i.e., key-value alignments or spatial patterns, affirming LLMs' proficiency in VRD layout generation.

	KL	JS
CogVLM7B	42.03	85.83
GPT3.5	69.32	93.28
GPT4	**77.02**	**95.15**

Table 5: Higher scores indicate that the prediction distributions are closer to the true distributions.

3.5 Does layout-aware learning benefit LLM-based IE?

Drawing from our findings in the aforementioned tasks, we propose a layout-aware learning approach that harnesses spatial prior knowledge via strategic prompting using ten-shots. This method adopts a dual-strategy, seamlessly merging spatial features with key-value patterns. Spatial features are adeptly represented through x- and y-coordinates, utilizing ten-shot instances presented in the demonstration section of our instructions. This approach facilitates the LLMs' comprehension of the coordinate system. Concurrently, key-value patterns, defined as key-value pairs accompanied by descriptive cues, are employed to delineate up to eight distinct directional patterns, catering to both precise and generalized directional understanding. Notably, the layout-aware approach has led to a substantial performance improvement of GPTs in relationship extraction tasks, surpassing zero-shot baseline systems (see Table 3.) In a nutshell, this approach significantly enhances the capability of LLMs to understand VRDs.

4 Conclusion

This study delves into the remarkable spatial reasoning abilities of large language models within the context of visually-rich documents, a domain that traditionally demands visual perception. Through a meticulously crafted experimental framework, we have uncovered that LLMs, devoid of inherent visual channel, are nonetheless adept at discerning and inferring complex spatial relationships. Our findings pivot on the innovative integration of a layout-aware learning approach, significantly amplifies LLMs' capacity to comprehend, reason, and generate spatially coherent document layouts. By demonstrating that these models can effectively perform tasks traditionally within the purview of visually-aware systems, we open avenues for rethinking the boundaries of document intelligence.

5 Limitations

As we reflect on the advancements made, we also acknowledge the limitations and the fertile ground for future exploration. The current work serves as a foundation upon which more sophisticated models can be developed, datasets expanded, and methods refined. Our study is a stepping stone towards a future where LLMs are not only arbiters of language but also interpreters of the spatial constructs that underpin our world.

- **Expanding Dataset Diversity**: Our study is informed by a current dataset that requires expansion to encompass a wider array of document structures and sources. Broadening the dataset will allow for a more robust assessment of model performance across diverse document types.
- **Inclusion of Diverse LLMs**: The research primarily targets prevalent LLMs, yet the integration of a more extensive range of models, is necessary to advance our understanding of their capabilities and applications.
- **Refinement of Methodologies**: While our layout-aware approach has yielded promising outcomes, there is ample room for the development of advanced techniques. Future research should concentrate on incorporating explicit structural information and enhancing models' capacity to learn structural patterns efficiently.
- **Development of Domain-Specific Benchmarks**: Although progress has been made in establishing benchmarks for structured text generation, there is a clear advantage to creating benchmarks tailored to specific domains. Tailoring benchmarks to unique domain requirements will bolster the applicability and precision of models within specialized contexts.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. 2024. Rodla: Benchmarking the robustness of document layout analysis models. In *CVPR*.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. *Preprint*, arXiv:2303.05063.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 40834091, New York, NY, USA. Association for Computing Machinery.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, 2nd edition, pages 202–251. Cambridge University Press.

Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolay Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. 2023. FormNetV2: Multimodal graph contrastive learning for form document information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9011–9026, Toronto, Canada. Association for Computational Linguistics.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han,

Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. In *ACM MM*, page 19121920.

Yanming Liu, Xinyue Peng, Tianyu Du, Jianwei Yin, Weihao Liu, and Xuhong Zhang. 2024. Era-cot: Improving chain-of-thought through entity relationship analysis. *Preprint*, arXiv:2403.06932.

John OKeefe and Jonathan O. Dostrovsky. 1971. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 34 1:171–5.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, Online.

Jonathan Roberts, Timo Luddecke, Sowmen Das, K. Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *ArXiv*, abs/2306.00020.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark B. Gerstein. 2023. Struc-bench: Are large language models really good at generating complex structured data? *ArXiv*, abs/2309.08963.

Hao Wang, Xiahua Chen, Rui Wang, and Chenhui Chu. 2023. Vision-enhanced semantic entity recognition in document images via visually-asymmetric consistency learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15718–15731, Singapore. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pretraining for visually-rich document understanding. In *the Annual Meeting of the Association for Computational Linguistics*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pretraining of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 11921200, New York, NY, USA. Association for Computing Machinery.

Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. Evaluating spatial understanding of large language models. *ArXiv*, abs/2310.14540.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. In *International Conference on Learning Representations*.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024. Groundhog: Grounding large language models to holistic segmentation. *Preprint*, arXiv:2402.16846.

A Instructions

B Visualization

Basic Spatial Perception in Section 3.1	
TASK DESCRIPTION	You are now an expert in determining relative positions with a strong sense of direction, proficient in understanding the Cartesian coordinate system. Your task is to determine the relative position of the second point, point2, with respect to the first point, point1. The origin (0,0) is at the top left, with the x-axis positive direction to the right and the y-axis positive direction downward. You will now receive two points, point1 and point2, along with their respective coordinates (x1,y1) and (x2,y2). The algorithm is as follows: 1. If the x-coordinates or y-coordinates of the two points are equal, choose from 'top,' 'bottom,' 'left,' or 'right.' 2. Otherwise, calculate the coordinates to determine the relative position of point2 with respect to point1, choosing from 'top-left,' 'top-right,' 'bottom-left,' or 'bottom-right.'
LABEL MAPPING	Use the following algorithm to identify the region in which point2 is located relative to point1, and output the relative position as one of eight directions: {top-left, top, top-right, left, right, bottom-left, bottom, or bottom-right.}
FORMATTING	Here are the coordinates of the given two points:""item": "point1": "position": [x1, y1], "point2": "position": [x2, y2], "relative-position": Y
QUERY	Please predict the direct positional relationship between these two items "item11": "position": [142.5,800.5],"item12":"position": [357.5,974.5]
ANSWER	"item": "point1": "position": [x1, y1], "point2": "position": [x2, y2], "relative-position": left

Table 6: Prompt template for the basic spatial perception task.

Page Location Prediction in Section 3.2	
TASK DESCRIPTION	You are an expert in interpreting formatted documents. You excel in annotating the coordinates of text. A bill of lading is a document used to describe and record the shipment of goods by sea. It has a standardized format template where similar types of semantic entities usually have similar visual and layout attributes, with keys and values distributed horizontally or diagonally. Now, you are given an entire bill of lading, including the text value of each text box, the position value of the text, and the text category label. Based on the given text value, position value, and text category label of each text box, predict the approximate position coordinates of the text box labeled 'shipper' with position null based on the distribution structure of key-value pairs in the bill of lading.
LABEL MAPPING	It contains key information such as the 'shipper', 'consignee', 'mode of transport', 'port of origin', and 'destination port'.
FORMATTING	Provide the answer in the following format without explanation: ['position': [x1,y1,x2,y2]]. Do not return an empty position;
QUERY	Here is the text content of the document:" XXXX". Please predict the approximate coordinates of the 'shipper' text box based on related semantic entities.
ANSWER	Position:[778, 288, 791, 309]

Table 7: Prompt template for the page location prediction task.

Document Information Extraction in Section 3.3	
TASK DESCRIPTION	Hello, you are an experienced form-reading expert who has reviewed many forms and understands the key-value distribution and logical layout structure within them. Your task now is to generate a more perfect layout for a bill of lading based on the given template. Your task is to learn the key-value distribution and overall layout structure from the given template and use your prior knowledge to create a more perfect template layout. Ensure that the arrangement of each category follows a logical order and is easy to read, considering the positional relationships between different text segments.
LABEL MAPPING	The generated layout should include at least the following categories: [‘Consignee-value’, ‘Pieces-value’, ‘0’, ‘Receiver-value’, ‘Shipper’, ‘Shipper-value’, ‘Receiver’, ‘Customer Order Number’, ‘Shipping Company’, ‘Shipping Company-value’, . . .].
FORMATTING	You will receive a sample bill of lading document, including text segments and their coordinates in the format ‘id’: xxx, ‘text’: ‘xxx’, ‘position’: [x1, y1, x2, y2]. Here, ‘position’ provides coordinates representing the top-left and bottom-right corners of the text box in x and y values.” Provide the answers in the following JSON format: [‘idx’: xxx, ‘position’: [x1, y1, x2, y2],‘label’: ‘predicted label’].
QUERY	Here is the text content of the document:“ XXXX“. Please fill in the specific content appropriately according to the entity category.
ANSWER	“text”: “Details”, “position”: [310, 60, 410, 95], “label”: “title”, “text”: “Form”, “position”: [420, 60, 460, 95], “label”: “title”, “text”: “SHIPPER:”, “position”: [60, 110, 130, 130], “label”: "shipper”, “text”: “Acme Export Co.”, “position”: [60, 140, 310, 160], “label”: “shipper-value”...

Table 8: Prompt template for the document layout generation task.

Document Layout Generation in Section 3.4	
TASK DESCRIPTION	Hello, you are an experienced form-reading expert who has reviewed many forms and understands the key-value distribution and logical layout structure within them. Your task now is to generate a more perfect layout for a bill of lading based on the given template. Your task is to learn the key-value distribution and overall layout structure from the given template and use your prior knowledge to create a more perfect template layout. Ensure that the arrangement of each category follows a logical order and is easy to read, considering the positional relationships between different text segments.
LABEL MAPPING	The generated layout should include at least the following categories: [‘Consignee-value’, ‘Pieces-value’, ‘0’, ‘Receiver-value’, ‘Shipper’, ‘Shipper-value’, ‘Receiver’, ‘Customer Order Number’, ‘Shipping Company’, ‘Shipping Company-value’, . . .].
FORMATTING	You will receive a sample bill of lading document, including text segments and their coordinates in the format ‘id’: xxx, ‘text’: ‘xxx’, ‘position’: [x1, y1, x2, y2]. Here, ‘position’ provides coordinates representing the top-left and bottom-right corners of the text box in x and y values.” Provide the answers in the following JSON format: [‘idx’: xxx, ‘category’: ‘predicted category’, ‘position’: [x1, y1, x2, y2]].
QUERY	Please generate a perfect layout information for the sea waybill based on the layout you just learned and your prior knowledge. Please fill in the specific content appropriately according to the entity category.
ANSWER	“text”: “Details”, “position”: [310, 60, 410, 95], “label”: “title”, “text”: “Form”, “position”: [420, 60, 460, 95], “label”: “title”, “text”: “SHIPPER:”, “position”: [60, 110, 130, 130], “label”: "shipper", “text”: “Acme Export Co.”, “position”: [60, 140, 310, 160], “label”: “shipper-value”...

Table 9: Prompt template for the document layout generation task.

Layout-Aware Learning in Section 3.5	
TASK DESCRIPTION	Hello, you are an expert specializing in semantic relationship understanding and information extraction. Your current task is to extract information from a bill of lading. Based on the given text value, position value, and text category label of each text box, predict the approximate position coordinates of the text box labeled 'shipper' with position null based on the distribution structure of key-value pairs in the bill of lading. For each text segment, you need to predict a corresponding category from the given set. If no suitable category exists, choose the category label '0'.
DEMONSTRATION EXAMPLES	Now, here are some important entities, their positions, and their categories. These examples are provided for you to learn the category of key texts and their corresponding positions: • {"text":"Consignee","position":[179,342,268,365],"label":"Consignee"} • {"text":"New York, NY","position":[424,380,615,403],"label":"Consignee-value"} • {"text":"PORT OF DISCHARGE","position":[280,580,430,600],"label":"port-of-discharge"} • {"text":"Boston","position":[280,610,380,630],"label":"port-of-discharge-value"} • {"text":"MARKS AND NUMBERS","position":[100,650,200,670],"label":"shipping-mark"}
DEMONSTRATION PATTERNS	Additionally, here are three typical key-value pair layouts commonly found in bills of lading: 1. **Vertical Layout:** • "text":"Shipper","position":[65,340,99,352] • "text":"ROAD, SHANGHAI 200135, CHINA","position":[65,411,297,427] 2. **Diagonal Layout:** • "text":"Port of Loading","position":[320,781,390,793] • "text":"SHANGHAI","position":[351,805,452,826] 3. **Horizontal Layout:** • "text":"Consignee","position":[179,342,268,365] • "text":"New York, NY 10016 USA","position":[173,342,262,365] These three layouts are common key-value pair structures in a bill of lading. When choosing a category, consider the positional information of the text segment, as nearby segments may be related. For example, 'Marks-value' represents the specific content of marks and numbers and is usually located near the 'Marks' category. The category 'Shipper-value' represents specific information about the shipper, which is also usually nearby.

Table 10: Prompt template for the document layout generation using layout-aware task. (Part 1)

Layout-Aware Learning in Section 3.5	
LABEL MAPPING	The category set is as follows: [‘Consignee-value’, ‘Pieces-value’, ‘0’, ‘Receiver-value’, ‘Shipper’, ‘Shipper-value’, ‘Receiver’, ‘Customer Order Number’, ‘Shipping Company’, ‘Shipping Company-value’, ‘Container Type and Quantity’, ‘Freight Terms’, ‘Freight Terms-value’, ‘Shipping Terms’, ‘Shipping Terms-value’, . . .].
FORMATTING	You will receive OCR information from the bill of lading, including text segments and their coordinates in the format ‘id’: xxx, ‘text’: ‘xxx’, ‘position’: ‘xxx’. Here, ’position’ provides coordinates representing the top-left and bottom-right corners of the text box in x and y values. Provide the answers in the following JSON format: [‘specific idx’: ‘predicted label’]
QUERY	Here is the text content of the document:“ XXXX“. Please predict the category to which the above entities belong.
ANSWER	“text”: “Details”, “position”: [310, 60, 410, 95], “label”: “title”, “text”: “Form”, “position”: [420, 60, 460, 95], “label”: “title”, “text”: “SHIPPER:”, “position”: [60, 110, 130, 130], “label”: "shipper”, “text”: “Acme Export Co.”, “position”: [60, 140, 310, 160], “label”: “shipper-value”...

Table 11: Prompt template for the document layout generation using layout-aware task.(Part 2)

Model	**Visual Encoder**	**BSP**	**PLP**	**DIE**	**DLG**
LLAMA7B	×	✓	×	×	×
BAICHUAN7B	×	✓	×	×	×
BAICHUAN13B	×	✓	×	×	×
VICUNA7B	×	✓	×	×	×
COGVLM	✓	✓	✓	×	×

Table 12: While we also tried other LLMs, due to their small scale of model parameters, they failed to yield results.

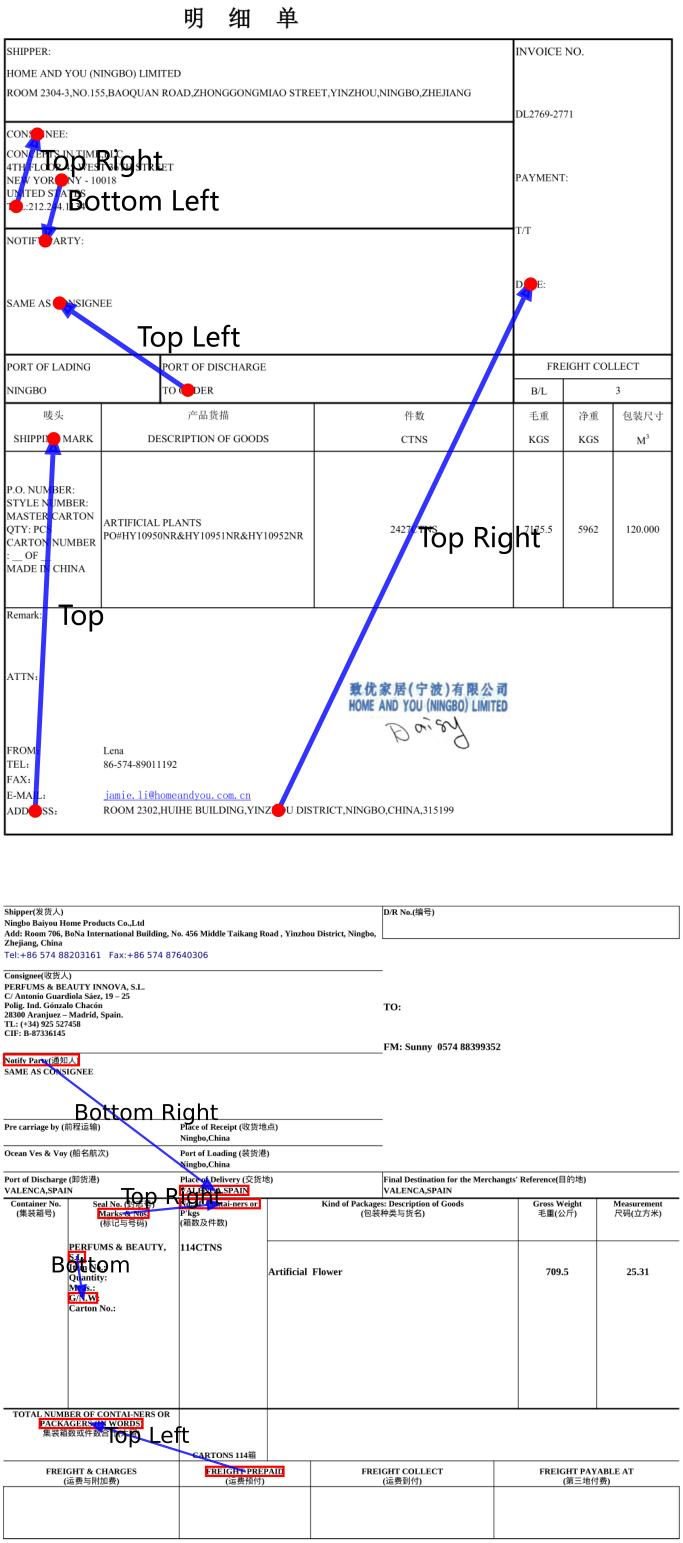

Figure 3: Visualization of GPT3.5 on the basic spatial prediction task. The direction of the blue arrow indicates the relative position of the second bounding box to the first bounding box, with the arrow direction labeled as the result given by the GPT3.5.

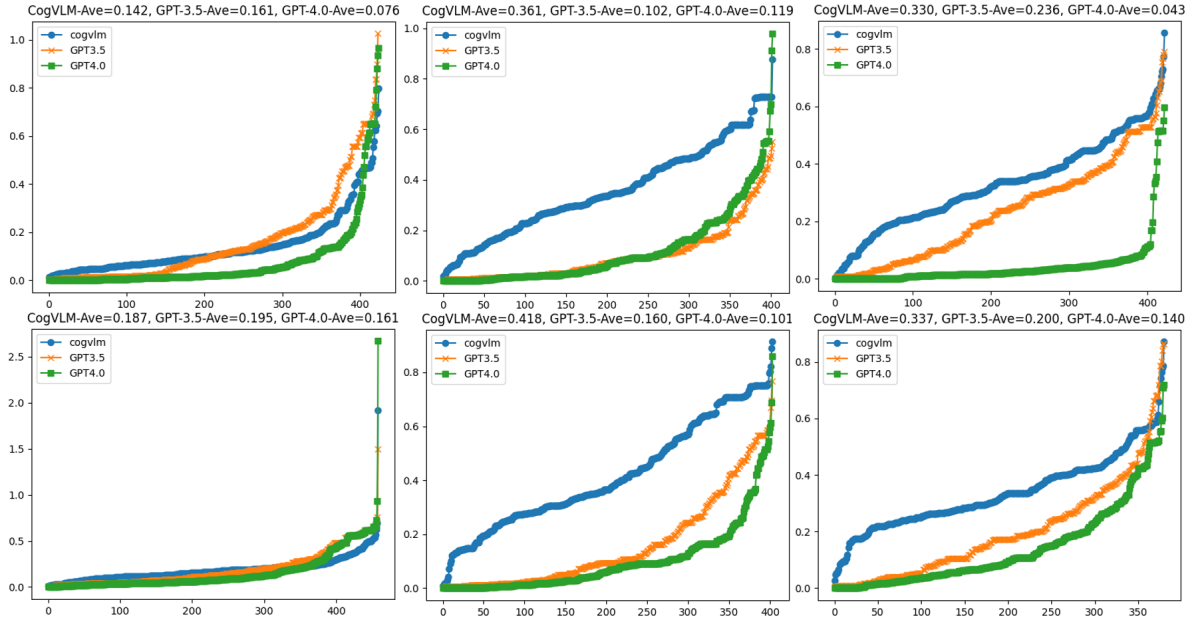

Figure 4: Euclidean distances of the positions of different categories of bounding boxes predicted by LLMs in page location prediction task. Sorted from smallest to largest. The first column shows results for T1 (`shipper`) and T1-v (`shipper-value`), the second column for T2 (`port_of_origin`) and T2-v (`port_of_origin-value`), and the third column for T3 (`marks`) and T3-v (`marks-value`).

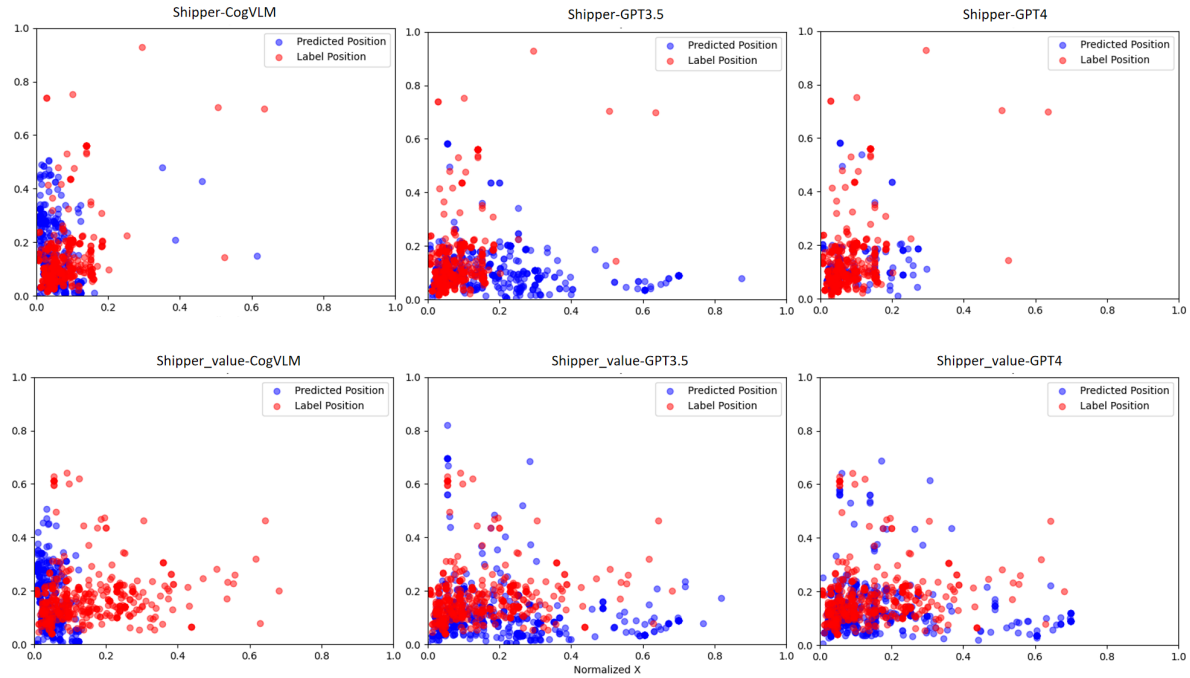

Figure 5: Scatter plot distribution of LLMs predicting the entity positions. **Top**: T1 (`shipper`) and **Bottom**: T1-v (`shipper-value`).

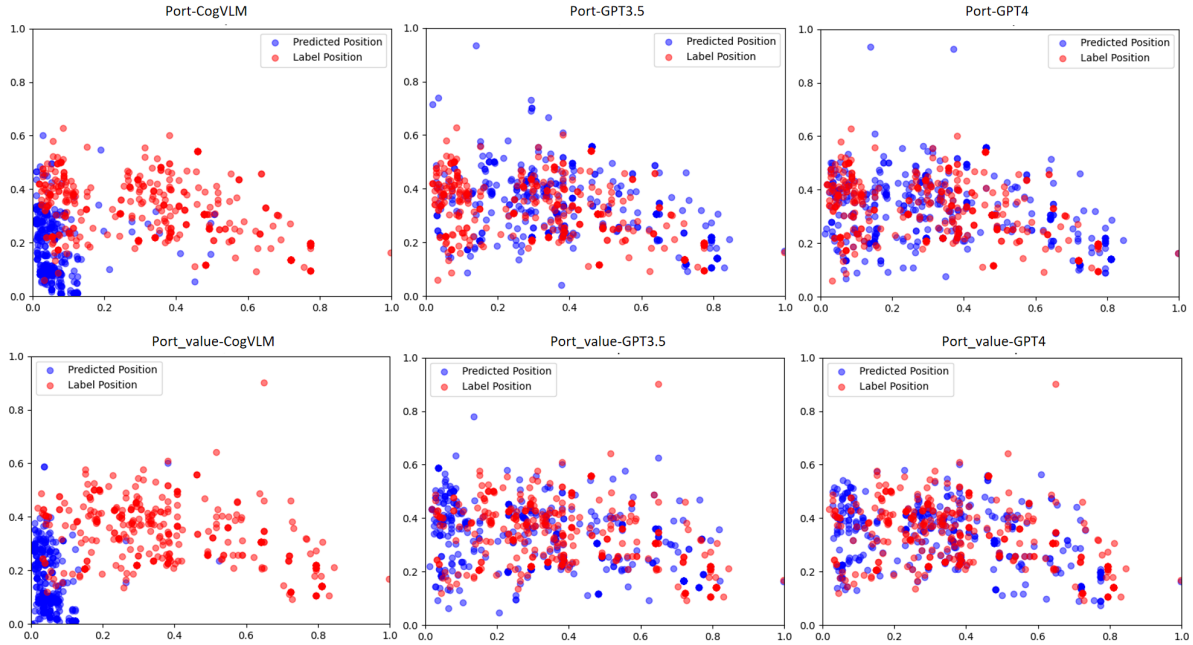

Figure 6: Scatter plot distribution of LLMs predicting the entity positions. **Top**: T2 (`port_of_origin`) and **Bottom**: T2-v (`port_of_origin-value`) (bottom).

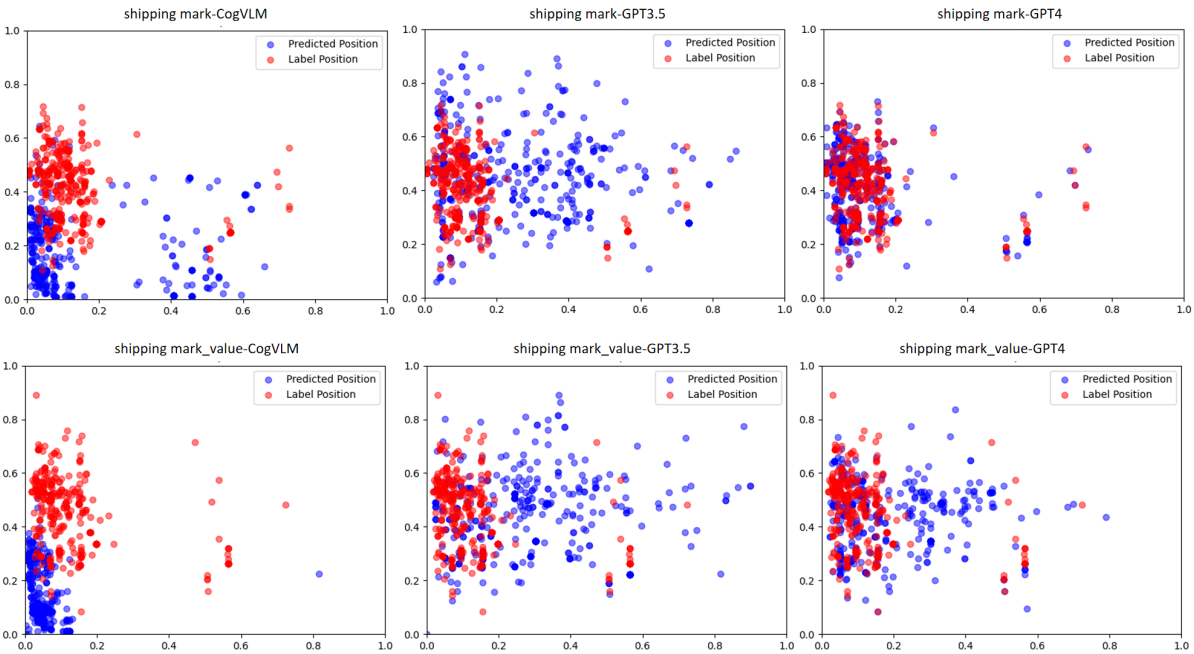

Figure 7: Scatter plot distribution of LLMs predicting the entity positions. **Top**: T3 (`marks`) and **Bottom**: T3-v (`marks-value`).

Figure 8: LLMs predict the probability distribution of entity positions in the x and y directions for the "`Shipper`", "`Shipper-value`", and "`Port`" categories.

Figure 9: Visualization examples of layout generation in visually-rich documents by GPT3.5. The red and blue bounding boxes represent the generated keys and values, respectively. The text within the bounding boxes represents the generated text content, and the black font represents the generated categories.

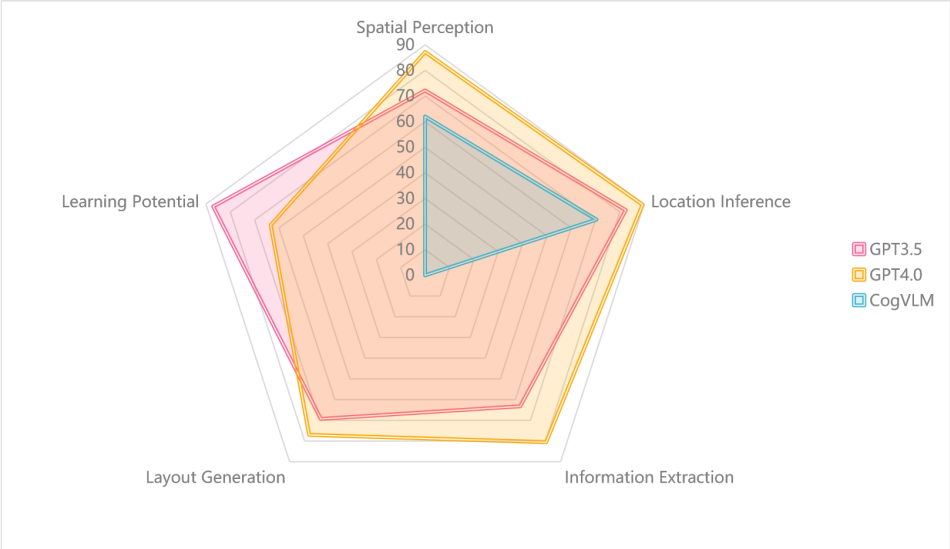

Figure 10: This radar chart illustrates the comparative performance of LLMs in the domain of spatial reasoning tasks. The metrics include spatial perception, learning potential, location inference, layout generation, and information extraction. The chart highlights the strengths and weaknesses of models such as GPT3.5, GPT4, and CogVLM across these dimensions, providing a visual summary of their spatial reasoning capabilities.