Patient Visualization Enhances Spatial Reasoning in GPT Models

Anonymous ACL submission

Abstract

While large language models (LLMs) are dominating the field of natural language processing, with GPT being one of the leaders, it remains an open question how well these models can perform spatial reasoning. Contrary to recent studies suggesting that LLMs struggle with spatial reasoning tasks, we demonstrate in this paper that a novel prompting technique, termed Patient Visualization of Thought (PATIENT-VOT), can boost GPTs' spatial reasoning abilities. The core idea behind PATIENT-VOT is to tackle (1) spatial understanding and (2) spatial reasoning, each through a two-step approach, where each process is guided by key trigger words: *bullet list* and *coordinate*, respectively. By applying PATIENT-VOT, we achieve an average accuracy improvement of up to 35% (absolute) compared to the state-of-the-art visual prompting technique, Visualization-of-Thought. Our findings show that GPTs are indeed much more proficient in spatial tasks than commonly believed, when effectively prompted.

1 Introduction

Large language models (LLMs) are massive neural networks trained on a vast and diverse range of corpora, that are currently leading the field of natural language processing (NLP) (Brown, 2020; Achiam et al., 2023). Beyond their remarkable achievements in NLP, researchers are gradually focusing on broader goals, such as artificial general intelligence, where they envision the development of versatile, if not universal, AI assistants (Zheng et al., 2024). In this context, LLMs play a pivotal role due to their strong reasoning capabilities, their characteristics as general pattern machines (Mirchandani et al., 2023), and their capacity to produce human-friendly explanations. However, spatial reasoning ability, one of the key requirements for these assistants, is known to be lacking in LLMs (Bang et al., 2023; Sharma, 2023). Multiple recent studies point out that even the top-performing LLMs, such as GPT4, struggle significantly with spatial reasoning tasks (Li et al., 2024; Yamada et al., 2023).

Among various efforts to enhance LLMs' spatial reasoning abilities, a notable approach is *prompt engineering* (Bommasani et al., 2021), which aims to trigger and maximize the model's spatial reasoning capabilities by designing effective prompts. One major advantage of prompt engineering is that it does not require additional training or external resources, making it a cost-effective and generally applicable approach. While some recent studies have emerged in this field (Wu et al., 2024; Li et al., 2024; Yasunaga et al., 2023), we believe this area remains under-explored.

Our Objective and Approach In this paper, we aim to tackle the following research question from a prompt engineering perspective: *How can we effectively trigger and improve the spatial reasoning abilities of GPT models?*

To this end, we introduce Patient Visualization-of-Thought (PATIENT-VOT), a simple yet effective prompting technique designed to enhance the spatial reasoning skills of GPT models. PATIENT-VOT is built on the Visualization-of-Thought approach (Wu et al., 2024) with adding two novel ideas: (1) Patient Spatial Understanding (PSU), which involves a two-step process of summarizing information into a bullet list before converting it into a visualization, rather than using direct visualization as in prior methods. PSU is especially beneficial for tasks that provide textual information without accompanying visual elements. (2) Patient Spatial Reasoning (PSR), which also employs a two-step process, guides LLMs to generate two types of visualizations during the reasoning phase, with the first being based on coordinates.

We show that PSU significantly reduces the errors GPT models make when visualizing an initial image from the given text information (see Figure 2). Furthermore, PSR activates an additional modality, *coordinate*-based reasoning, that significantly enhances GPTs' spatial reasoning abilities when combined with visualization (see Table 2). PATIENT-VOT consistently boosts the performance of various GPT models (GPT-4o, GPT-4o-mini, and GPT-4-turbo) on a variety of challenging spatial reasoning tasks (Wu et al., 2024).

2 Related Work

Spatial Reasoning in LLMs Several recent studies have examined the spatial reasoning capabilities of LLMs, consistently finding that LLMs continue to struggle with spatial reasoning tasks (Li et al., 2024; Bang et al., 2023). Existing research on spatial reasoning in LLMs can be broadly categorized into three approaches: (1) Analyzing LLM behavior to gain insights into their underlying mechanisms (Xie et al., 2023; Cohn and Hernandez-Orallo, 2023), (2) Augmenting spatial reasoning abilities by conducting additional training on curated datasets (Hong et al., 2023; Cheng et al., 2024), and (3) Improving spatial reasoning performance using effective prompting methods instead of further training (Wu et al., 2024; Sharma, 2023).

Our paper focuses on the prompting approach, particularly with GPT models, due to their widespread use and strong performance.

Prompt engineering approaches to LLM spatial reasoning Recently, various prompting techniques have been introduced, such as chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2022), and tree-of-thought (Yao et al., 2024). However, these methods are primarily designed for general reasoning tasks. Given the unique challenges of spatial reasoning, some prompting techniques have been specifically tailored for this purpose (Wu et al., 2024; Sharma, 2023). Among those, visualization-of-thought (VoT) (Wu et al., 2024) has demonstrated promising results with a unified prompt. Our work builds on the foundation of VoT, aiming to develop an enhanced version.

3 PATIENT-VOT

3.1 Motivation

The goal of this paper is to discover a universal prompt that can effectively trigger and enhance spatial reasoning performance across the GPT model family. Our work is largely inspired by Wu et al. (2024), which demonstrated that the straightforward prompt "Visualize the state after each reasoning step." can substantially boost the spatial reasoning performance of GPT models. Identifying a universally effective prompt across different models and datasets is crucial, as it not only provides a generalizable approach but also offers valuable insights into how modern LLMs perform spatial reasoning. While the recent results in this area are impressive, we believe there is room for further improvement.

With this motivation in mind, we present PATIENT-VOT, designed to unlock LLMs' latent spatial reasoning abilities through two novel ideas: (1) Patient spatial understanding, where LLMs are guided to first translate the information into a bullet list before creating the final visualization; (2) Patient spatial reasoning, which activates two modalities (visual and coordinate) in LLMs to improve visual reasoning performance.

3.2 Patient Spatial Understanding

In our preliminary study, we found that GPT models struggle with seemingly simple tasks, such as converting a natural language description of a grid into a visual representation (see Figure 2). To address these mistakes, we propose a simple yet effective approach: First translating the provided information into a bullet list before converting it into a visualization. This method significantly reduces the error rate from 52% to 8% when visualizing the initial grid. Specifically, we use the following prompt: *"Before starting, convert the initial information into a detailed bullet list to effectively grasp the map's information."*

3.3 Patient Spatial Reasoning

Visualizing the state has been shown to be effective for spatial reasoning in LLMs (Wu et al., 2024). We propose activating an additional modality: Coordinate-based reasoning. While LLMs may naturally engage in this type of reasoning, our observations indicate that explicitly prompting it is highly effective. Additionally, combining coordinate-based reasoning with visual-based reasoning results in a synergistic effect, leading to an additional increase in performance. Consequently, we incorporate the following sentence into our final prompt: *"Solve the problem twice with the following approach: 'Visualize the state after each reasoning step'. In the first attempt, use coordinates instead of visualization. In the second attempt, use direct visualization and fix any errors in the first* *attempt.”*

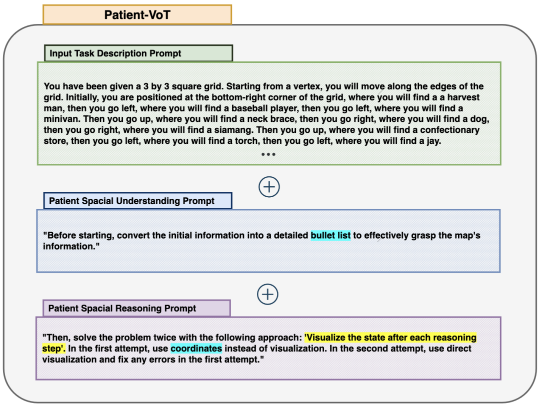

Figure 1: The overall template of PATIENT-VOT. Key trigger words, “bullet list” and “coordinates”, are marked in blue, while the VoT prompt element is highlighted in yellow.

4 Experiments

4.1 Experimental Settings

Datasets We selected three spatial reasoning tasks presented by Wu et al. (2024). These tasks are: (1) Natural language navigation, which involves visualizing a grid and tracking sequential movements within it; (2) Route planning, where the model must generate multi-hop navigation instructions on a 2D grid; and (3) Visual tiling, which requires fitting appropriate tetrominoes into a square grid, similar to the game Tetris. These tasks are particularly intriguing because they demand fundamental spatial understanding and reasoning skills, yet remain highly challenging for GPT models, with baseline average accuracy hovering around 20%. Note that natural language navigation provides only text, while route planning and visual tiling include the initial grid (using emojis) as part of the input. For more details, see Appendix A.

Models and Settings We employ the GPT-4 model family, including GPT-4o, GPT-4o-mini, and GPT-4-turbo (Achiam et al., 2023). For baseline prompts, we follow the approach from Wu et al. (2024), using “Let’s think step by step.” for the CoT baseline (Kojima et al., 2022) and “Visualize the state after each reasoning step.” for the VoT baseline. Experiments are conducted using a basic greedy decoding scheme (temperature set to 0), with three different random seeds.

4.2 Results

Table 1 presents the performance of PATIENT-VOT and the baseline methods on the three datasets. We observe that PATIENT-VOT significantly and consistently improves performance across all models and datasets, outperforming related prompting techniques by a substantial margin.

Table 2 shows the results of several ablation studies. The top section highlights the impact of each component in PATIENT-VOT. It is evident that both PSU and PSR independently yield consistent improvements, and their combination leads to even greater performance gains.

5 Main Findings

5.1 Using a bullet list as an intermediate step significantly reduces mistakes in LLMs

As briefly discussed in Section 3.2, even the most advanced GPT-4 models make significant errors in translating descriptions into accurate grids (Figure 2 is an actual example from GPT-4o). This fundamentally aligns with recent research showing that LLMs often struggle with simple tasks involving counting or retracing steps (Golovneva et al., 2024). We believe that converting the description into a structured format, such as a bullet list with clear delimiters, and then using this structured format for visualization, helps minimize mistakes. Quantitatively, this approach reduces the error rate from 52% to 8% for GPT-4o in the natural language navigation task.

5.2 Coordinate-based reasoning and visual-based reasoning create synergy

Intuitively, LLMs can inherently use coordinates when dealing with spatial reasoning tasks. However, our findings show that explicitly prompting the LLM to employ coordinates is far from redundant. In fact, it proves effective on its own and also creates a synergistic effect when combined with visual-based reasoning. The empirical evidence supporting this claim is summarized in the bottom section of Table 2.

We compared the following three variants: (1) PATIENT-VOT: which incorporates both coordinates and visualizations, (2) PSR (Visualization-Only): which uses only visualizations (equivalent to VoT), and (3) PSR (Coordinate-Only): which relies solely on coordinates. The specific prompts for each variant are detailed in Appendix B. The results in Table 2 indicate that explicitly instructing GPT

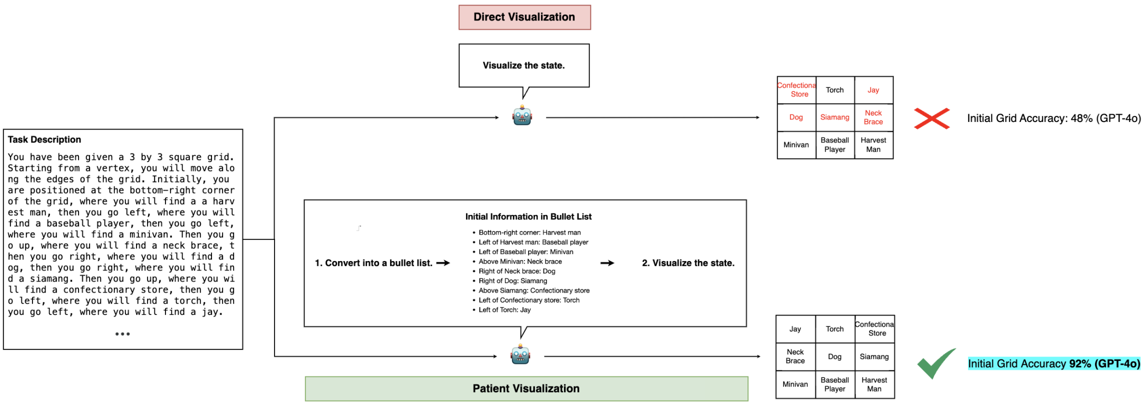

Figure 2: The intuition behind patient spacial understanding. The structured bullet list significantly reduces mistakes when creating the initial visualization (highlighted in blue).

	Natural Language Navigation	Route Planning	Visual Tiling	Avg.
Model	Acc (%)	Acc (%)	Acc (%)	Acc (%)
1. GPT-4o				
• CoT	$8.50_{1.32}$	$7.27_{3.75}$	$28.67_{2.02}$	$14.81_{2.36}$
• VoT	$26.17_{1.26}$	$5.15_{0.49}$	$29.00_{1.50}$	$20.44_{1.08}$
• **Ours: PATIENT-VOT**	$\mathbf{83.83_{1.44}}$	$\mathbf{30.23_{0.37}}$	$\mathbf{36.33_{1.61}}$	$\mathbf{50.05_{1.19}}$
2. GPT-4o-mini				
• CoT	$2.67_{0.29}$	$5.15_{0.25}$	$17.33_{5.03}$	$8.38_{1.86}$
• VoT	$22.17_{1.04}$	$5.80_{0.14}$	$17.67_{3.06}$	$15.21_{1.41}$
• **Ours: PATIENT-VOT**	$\mathbf{61.00_{1.50}}$	$\mathbf{41.58_{1.48}}$	$\mathbf{24.00_{3.00}}$	$\mathbf{42.19_{1.99}}$
3. GPT-4-turbo				
• CoT	$21.50_{2.18}$	$5.56_{0.14}$	$21.00_{2.65}$	$16.02_{1.66}$
• VoT	$25.67_{1.04}$	$3.43_{0.49}$	$19.00_{1.73}$	$16.03_{1.09}$
• **Ours: PATIENT-VOT**	$\mathbf{51.67_{1.44}}$	$\mathbf{7.52_{0.93}}$	$\mathbf{24.33_{1.15}}$	$\mathbf{27.84_{1.17}}$

Table 1: Effectiveness of PATIENT-VOT. Reported numbers are average and standard deviations of three runs.

Ablation #1. Effectiveness of PSU and PSR.			
Baseline: GPT-4o	NLN	RP	VT
• VoT	$26.17_{1.26}$	$5.15_{0.49}$	$29.00_{1.50}$
• VoT + PSU	$48.83_{2.57}$	$21.73_{0.57}$	$34.88_{2.21}$
• VoT + PSR	$31.33_{0.76}$	$12.17_{0.75}$	$34.33_{1.26}$
• VoT + PSU + PSR (=PATIENT-VOT)	$83.83_{1.44}$	$30.23_{0.37}$	$36.33_{1.61}$
Ablation #2. The synergy between coordinate-based and visual-based reasonings.			
Baseline: GPT-4o	NLN	RP	VT
• PSR (Coordinate-Only)	$80.50_{2.65}$	$26.06_{0.51}$	$35.33_{0.76}$
• PSR (Visualization-Only)	$48.83_{2.57}$	$21.73_{0.57}$	$34.88_{2.21}$
• PSR (Both) (=PATIENT-VOT)	$83.83_{1.44}$	$30.23_{0.37}$	$36.33_{1.61}$

Table 2: A summary of two ablation study results.

to perform coordinate-based reasoning is generally more effective than relying solely on visualizations. Most importantly, combining coordinate-based and visual-based reasoning yields even better performance than using either method alone.

6 Conclusion

This paper introduces a new prompting technique, PATIENT-VOT, designed to enhance the spatial reasoning capabilities of GPT models. PATIENT-VOT incorporates two straightforward yet powerful concepts: patient spatial understanding and patient spatial reasoning. It demonstrates effectiveness across all GPT-4 models on three core spatial tasks, achieving up to a 35% (absolute) improvement.

7 Limitations

Our work has a few limitations. Firstly, our study lies in the area of "prompt engineering" which may lack strong theoretical justification for why our approach is effective. Additionally, we concentrated on greedy decoding for computational efficiency. Nevertheless, exploring the integration of PATIENT-

VoT with sampling-based prompting techniques remains a promising area for future research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*.

Anthony G Cohn and Jose Hernandez-Orallo. 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *arXiv preprint arXiv:2304.11164*.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Fangjun Li, David C Hogg, and Anthony G Cohn. 2024. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18500–18507.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*.

Manasi Sharma. 2023. Exploring and improving the spatial reasoning abilities of large language models. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*.

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*.

Yutaro Yamada, Yihan Bao, Andrew K Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

A Tasks and Datasets

We have chosen three tasks presented by Wu et al. (2024) to evaluate our method. Since the code to (Wu et al., 2024)'s work is not available, we have re-implemented the datasets following their paper. As explained in Section 4.1, the three tasks are (1) Natural language navigation, (2) Route planning, and (3) Visual Tiling. Examples of each task are provided below and we recommend reading the original paper (Wu et al., 2024) for further details.

Natural Language Navigation Example "You have been given a 3 by 3 square grid. Starting from a vertex, you will move along the edges of the grid. Initially, you are positioned at the bottom-left corner of the grid, where you will find a wool, then you go right, where you will find a football player, then you go right, where you will find a black-and-white colobus. Then you go up, where you will find a pot pie, then you go left, where you will find a torch, then you go left, where you will find a minivan. Then you go up, where you will find a conch, then you go right, where you will find an american dipper, then you go right, where you will find a jay.

Now you have all the information on the map. The given map is a 3 by 3 map. You start at the position where the wool is located, then you go right by one step, then you go right by one step, then you go left by one step, then you go up by one step, then you go left by one step, then you go up by one step, and then you go right by one step. For your final answer, list all eight items encountered during the moves (including the starting item and any duplicates) under the title 'Final List of Items Encountered' as a bullet list."

Route Planning Example Provided in Figure 3 below.

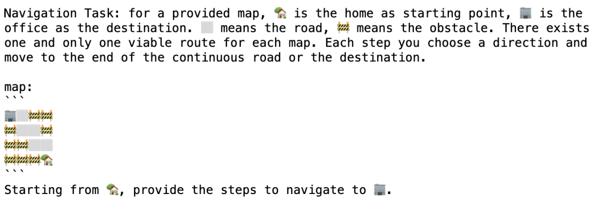

Figure 3: Route planning example.

Visual Tiling Example Provided in Figure 4 below.

B Prompt Templates Used in Ablation Study #2

Variant 1: (=PATIENT-VOT)

- "Before starting, convert the initial information into a detailed bullet list to effectively grasp the map's information. Then, solve the problem twice with the following approach: 'Visualize the state after each reasoning step'. In the first attempt, use coordinates instead of visualization. In the second attempt, use direct visualization and fix any errors in the first attempt."

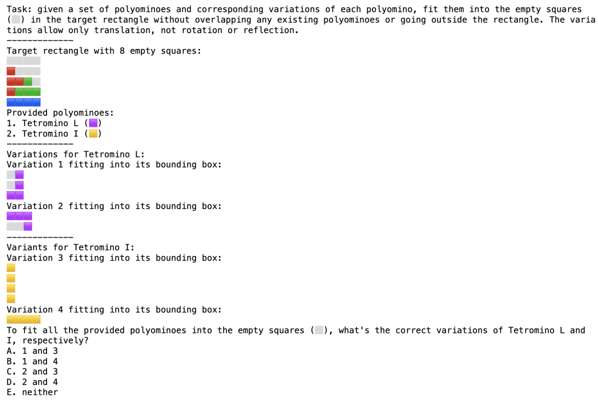

Figure 4: Visual tiling example.

Variant 2: PSR (Visualization-Only)

- "Before starting, convert the initial information into a detailed bullet list to effectively grasp the map's information. Then, solve the problem with the following approach: 'Visualize the state after each reasoning step'."

Variant 3: PSR (Coordinate-Only)

- "Before starting, convert the initial information into a detailed bullet list to effectively grasp the map's information. Then, solve the problem with the following approach: 'Visualize the state after each reasoning step'. Use coordinates instead of visualization."