

Stuck in the Matrix: Probing Spatial Reasoning in Large Language Models

Maggie Bai, Ava Kim Cohen, Eleanor Koss, Charlie Lichtenbaum^{*†}

October 24, 2025

Abstract

This paper explores the spatial reasoning capability of large language models (LLMs) over textual input through a suite of five tasks aimed at probing their spatial understanding and computational abilities. The models were tested on both fundamental spatial reasoning and multi-step problem-solving within structured grid-based environments using tasks such as quadrant identification, geometric transformations, distance evaluation, word searches, and tile sliding. Each task was scaled in complexity through increasing grid dimensions, requiring models to extend beyond simple pattern recognition into abstract spatial reasoning. Our results reveal that while LLMs demonstrate moderate success in all tasks with small complexity and size, performance drops off rapidly as scale increases, with an average loss in accuracy of 42.7%, and reaching as high as 84%. Every test that began with over 50% accuracy showed a loss of at least 48%, illustrating the consistent nature of the deterioration. Furthermore, their struggles with scaling complexity hint at a lack of robust spatial representations in their underlying architectures. This paper underscores the gap between linguistic and spatial reasoning in LLMs, offering insights into their current limitations, and laying the groundwork for future integrative benchmarks at the intersection of language and geometry.

1 Introduction

Large language models (LLMs) have achieved remarkable success across a wide array of language-related tasks, including text generation, summarization, and question answering [3] [11], and many benchmarks, such as TruthfulQA [8], have been developed to evaluate this performance. Their ability to handle non-linguistic reasoning tasks, particularly those involving geometry and spatial relationships, has received significant attention from research into Visual Language Models (VLMs), both on training methods [4] and benchmarks [12]. Yet less work has been done on the intersection of the two, namely an LLM’s spatial reasoning ability over text-based inputs.

Moreover, we observed in attempting to have LLMs play simple games like 2048 or Connect 4 through prompt-based board inputs that while models understood the strategy and gameplay, they often still made incorrect or invalid moves due to a failure to comprehend the board state. This manifested both in the initial reading of the board and in the application of principles such as gravity in Connect 4, emphasizing the limitations of text-based spatial reasoning in LLMs. As such, while the issues could have been solved by switching to a VLM, we decided to investigate the extent of LLMs’ spatial reasoning abilities in text-based domains. These capabilities may improve LLM usage and reliability in layout-heavy fields such as finance, where interpreting tabular data correctly is necessary.

In this study, we designed a series of simple, isolated tests to measure these capabilities, with each task progressively increasing in complexity. These tasks include:

^{*}Correspondence to: Charlie Lichtenbaum (clichtenbaum@gmail.com)

[†]Alphabetical ordering, all authors contributed equally

1. Quadrant: determining which quadrant of a Cartesian grid contains a specific symbol.
2. Transformation: reflecting a symbol across axes and producing the correct spatial configuration.
3. Distance: identifying the closest and farthest symbols relative to a target point on a grid.
4. Word Search: locating a defined word within a grid
5. Slide: sliding an X in a cardinal direction on a grid until it hits the edge or a wall

We evaluated four LLMs: GPT-4o, GPT-4.1, Claude 3.7 Sonnet with no thinking tokens (referred to as No Thinking), and Claude 3.7 with 16,000 thinking tokens (referred to as Medium Thinking). Each task was applied to progressively larger grid sizes, enabling detailed analysis of how well models generalize simple spatial reasoning to larger structures and more intricate configurations.

2 Related Works

This paper was inspired by work done with LLMs on text-based games, where experiments showed that models were unable to form and maintain the game worlds [14]. In addition, in Topaskal et al. [13], models were presented with boards for Tic-Tac-Toe, Gomoku, and Connect 4 in image and text-based formats. Their results showed a higher rate of invalid moves on more complex grid states in both the image and “illustration,” or ASCII grid, formats, implying that the LLMs failed to comprehend the board properly. As such, this study sought to identify the complexity at which models begin to deteriorate as well as the types of problems the models encounter.

LLMs are also frequently benchmarked on their geometric abilities. Datasets of math problems, such as MATH [7] or GPSM4K [1], have been used to test the multi-step reasoning and geometry capabilities of various LLMs. These studies have all found the space for significant improvement in LLM spatial processing and reasoning.

However, these studies differ from ours in two significant ways. Firstly, this study was not designed to push the limits of the mathematical capabilities of the LLMs, unlike the datasets of Olympiad or graduate-level novel geometry problems. Rather, it tests basic reasoning and understanding of relative placements, with only simple arithmetic involved. And secondly, the models in our study receive their inputs in the form of ASCII grids, as opposed to images or textual descriptions of layouts. As such, this study serves better as an inquiry into tokenization and interpretation by the models.

Another benchmark in LLM spatial reasoning is the competition ARC-AGI [5], which has spurred numerous developments in LLM spatial reasoning capabilities over the six years since its release. The recently released ARC-AGI-2 [6] will likely continue to push researchers towards further developments. This paper, however, was not largely focused on developing novel strategies for enhancing reasoning capabilities, but rather benchmarking pre-existing ones.

Finally, a similar problem to spatial reasoning is that of tabular reasoning, or the interpretation of mixed text and layout documents. This kind of reasoning is common to fields such as finance or business, where contracts and invoices often contain tabular data. Researchers have developed specific tabular models, such as DocLLM [15], to fit these tasks, and have shown significant improvement compared to standard LLMs. This is a similar problem to that of spatial reasoning, as understanding the structure and relative positioning of elements in a layout is imperative to accuracy, and a parsing mistake can cost a company thousands of dollars. As such, improving the spatial reasoning capabilities and reliability of LLMs could increase their use cases significantly.

3 Methodology

To evaluate the spatial reasoning capabilities of the models, we designed five distinct tests that each targeted a specific aspect of spatial understanding. All tasks used square grids composed of simple symbols (for

example, ‘.’ or letters) and were progressively scaled to increase complexity. All grids were presented with spaces as delimiters between indices. An example is shown below; all grids tokenized in the same manner, just with alternate characters instead of “.”s.



Figure 1: Example Tokenization

3.1 Quadrant

In this task, a single ‘X’ was placed randomly on a grid, and the models were asked to determine the quadrant of the grid in which the ‘X’ was located. The quadrants were defined in alignment with the Cartesian coordinate system and the models replied in a “Lower/Upper Left/Right” format. An example grid is shown below:

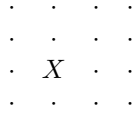
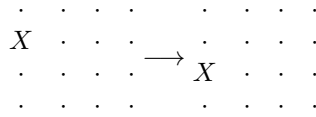


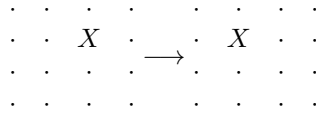
Figure 2: In this example, the ‘X’ is in quadrant III (lower left quadrant).

3.2 Transformation

This task tested the ability of the models to perform geometric transformations. A single “X” was randomly placed on a square grid, and the models were asked to reflect the “X” over a centerline (not given), either horizontally or vertically, and return a coordinate pair in (x, y) format. Example reflections are shown below:



(a) Reflection vertically.



(b) Reflection horizontally.

3.3 Distance

In this task, letters A-G and an X were placed on a square grid. The models were asked to identify which letter was closest to and farthest from the X. An example grid is shown below:

.	.	B	.	.
.	F	.	.	A
.	C	.	.	.
D	.	.	G	.
E	.	.	.	X

In this example, the closest letter is G, and the farthest is B.

3.4 Word Search

In this task, LLMs were provided a programmatically generated grid containing a single word along with the word contained. Models were tasked with producing a list of coordinates identifying the word’s location. Credit was given for partial spellings and incorrect orderings of coordinates. An example is shown below:

D	A	E	Z	A
E	D	B	V	D
R	A	T	M	M
L	T	C	Z	Y
J	J	N	N	L

In this example, the word “RAT” is embedded horizontally in the third row.

3.5 Slide

In this task, models were given a grid with an X and numerous #s. The model was instructed to ‘slide’ the X until it hit either a wall (#) or the edge of the grid and output the final coordinates. For multi-slide problems, credit was given for correct secondary slides, even if the starting position was affected from an earlier slide. An example grid is show below:

.	#	#	.	.	.	#	#	.	.	.	#	#	.	.
.	#	.	X	.	.	#	.	.	.	.	#	.	.	.
.	.	.	.	.	→	.	.	.	X	.	→	X	.	.
.	#	#	#	.	.	#	#	#	.	.	#	#	#	.
#	.	.	.	#	#	.	.	.	#	#	.	.	.	#

Figure 4: Slide down, then left

In this example, the X would first move to row 2, column 3 and then end at row 2, column 0.

4 Results

Each of the designed tasks was tested on four models: GPT-4o [9], GPT-4.1 [10], Claude 3.7 No Thinking, and Claude 3.7 Medium Thinking [2]. Each set of parameters (i.e. grid size, model) was run 10 times, and the results were averaged. Test prompt inputs ranged from ~400 to ~90,000 tokens, well below the context limit for any of the models.

Additionally, we explored some common sources of error and attempted to mitigate them. These attempts are also seen below.

4.1 Quadrant

Models	Smallest 5 Grids	Largest 5 Grids
GPT-4o	80%	22%
GPT-4.1	86%	35%
Claude No Thinking	98%	61%
Claude Medium Thinking	98%	58%

Table 1: Model Accuracy on Small and Large Grids (Quadrant)

Grid sizes ranged from 2×2 to 300×300 in increments of 10×10 . Small grids saw high accuracy across all models, but performance deteriorated significantly on larger grids. Specifically, the OpenAI variants performed poorly on larger grids, with accuracy dropping to 20-30%, and Anthropic models dropping to around 60%, as seen in Figure 6.

Error analysis revealed that while initial misreading of the grid was common, models were largely consistent in quadrant selection based on their internal understanding of the board state. That is to say, once the model had “found” the X in the grid, it predicted the correct quadrant for that coordinate position; errors appeared largely when the model identified the incorrect position as containing the X. This trend is shown in Figure 7.

The models were internally inconsistent only when the X was near a dividing line between two quadrants. Even when given the board size explicitly, models often had a difficult time establishing the location of the center lines. The increase in consistency at larger board sizes can likely be explained by the lower chance of an X placement near one of the dividing lines.

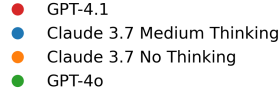


Figure 5: Color coding of data; all following plots use this format

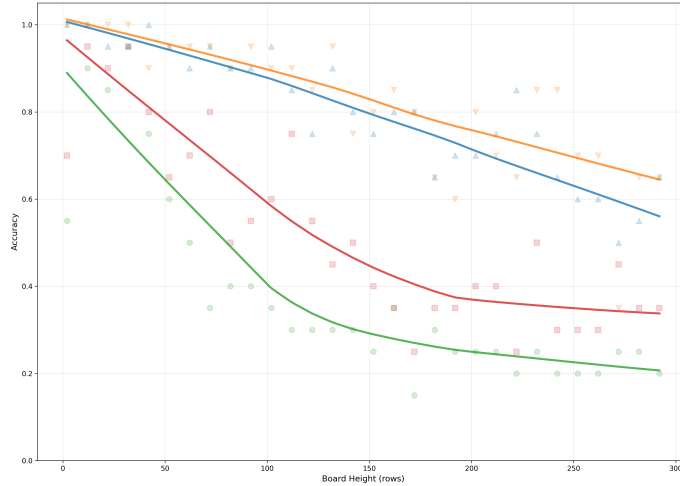


Figure 6: Accuracy vs Grid Size (Quadrant)

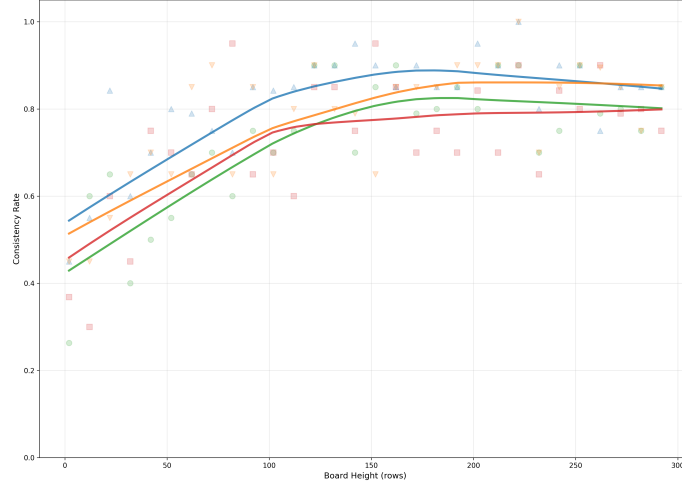


Figure 7: Consistency vs Grid Size (Quadrant)

4.2 Transformation

Models	Smallest 5 Grids	Largest 5 Grids
GPT-4o	6%	0%
GPT-4.1	14%	0%
Claude No Thinking	62%	0%
Claude Medium Thinking	70%	0%

Table 2: Model Accuracy on Small and Large Grids (Transform)

Our suspicions about errors from the Quadrant test stemming from the models misreading the initial board were quickly proven in the Transformation test. Grid sizes ranged from 2×2 to 200×200 , with each grid testing reflections across both axes. The models’ inability to properly read board states was immediately evident: even on small boards, the OpenAI models averaged under 50% accuracy. Performance deteriorated much quicker as well, with Anthropic models dropping below 20% accuracy by the 75×75 grid and the OpenAI models degrading even faster. By the largest grids, models rarely, if ever, got the answer correct. These trends can be seen in Figure 8.

It seems that the margin of error built into the nature of the Quadrant test shielded the models from such immediate deterioration. Interestingly, the models were not as internally consistent in this test as well. Both Anthropic models performed very well, with perfect internal consistency whereas the OpenAI models sometimes hit as low as 20% calculation accuracy for certain grids. There appears to be no definite pattern for which grids were inconsistent, and the results can be seen in Figure 9.

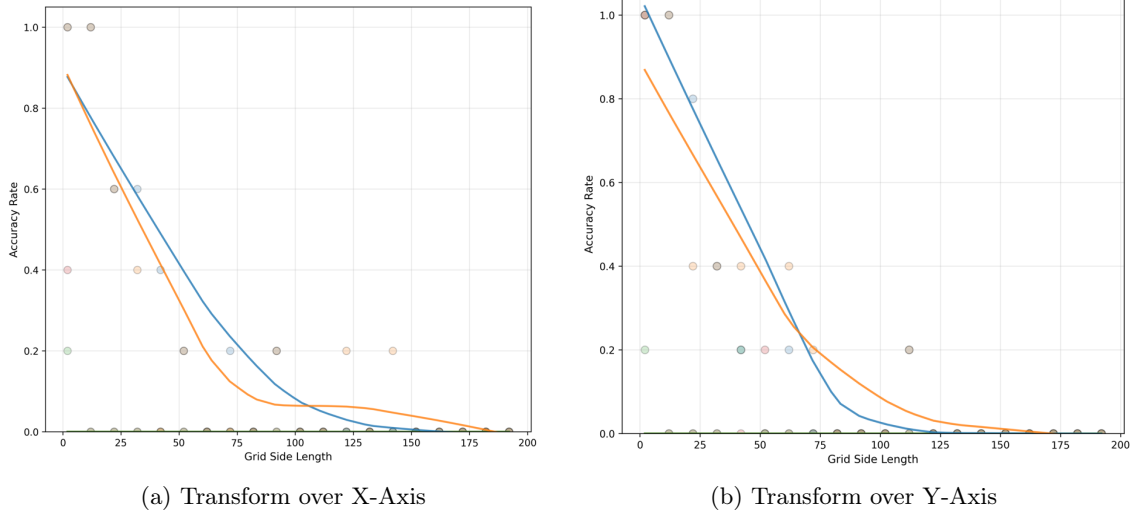


Figure 8: Accuracy vs Grid Size (Transformation)

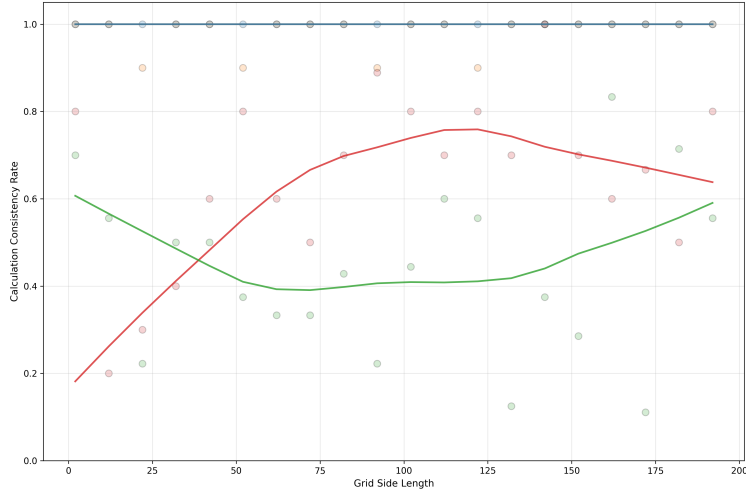


Figure 9: Consistency vs Grid Size (Transformation)

4.3 Distance

Models	Smallest 5 Grids	Largest 5 Grids
GPT-4o	44%	0%
GPT-4.1	76%	8%
Claude No Thinking	96%	40%
Claude Medium Thinking	100%	16%

Table 3: Model Accuracy on Small and Large Grids (Distance)

Given that the failures of the previous two tests were largely due to miscounting, we devised this test to evaluate a model’s ability to reason relatively; that is, in relation to other objects within a geometric space. Models were tested on grids ranging from 10×10 to 200×200 , with a step of 5. Like the previous tests, models

were more accurate on smaller boards and their performance degraded as grid size increased, as shown in Figure 10.

Interestingly, models did not exhibit much internal consistency, especially compared to the previous tests. They frequently made mathematical errors, accounting for around half of the errors for each model, as seen in Figure 11.

The math error rate decreased as grid size increased for the OpenAI models and remained roughly constant for the Anthropic models. The positional error rate increased with grid size, similar to other tests. The results are shown in Figure 12.

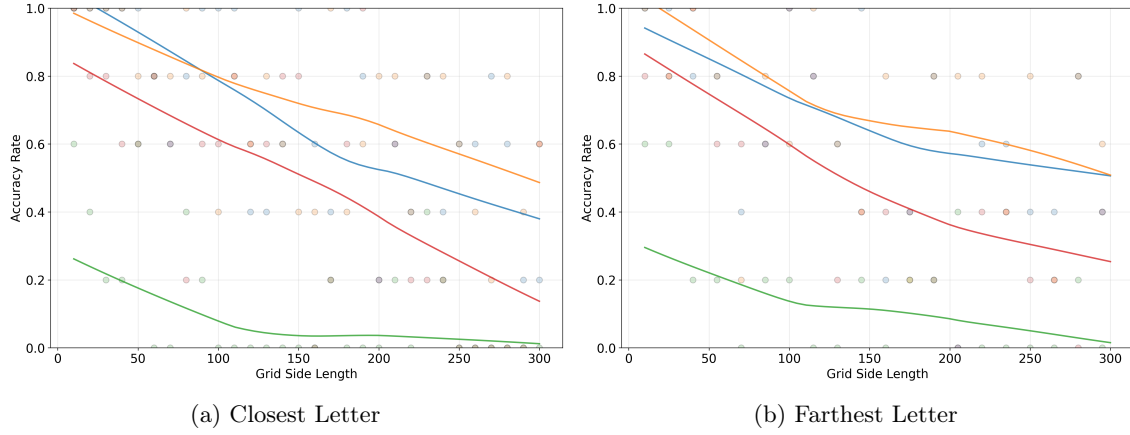


Figure 10: Accuracy vs Grid Size (Distance)

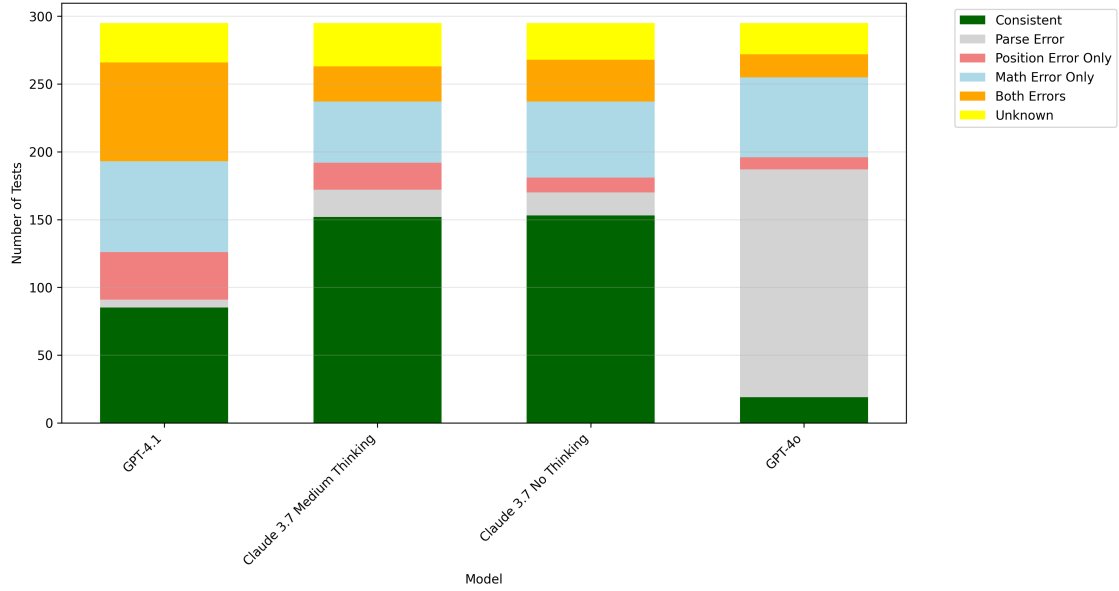


Figure 11: Error Types (Distance)

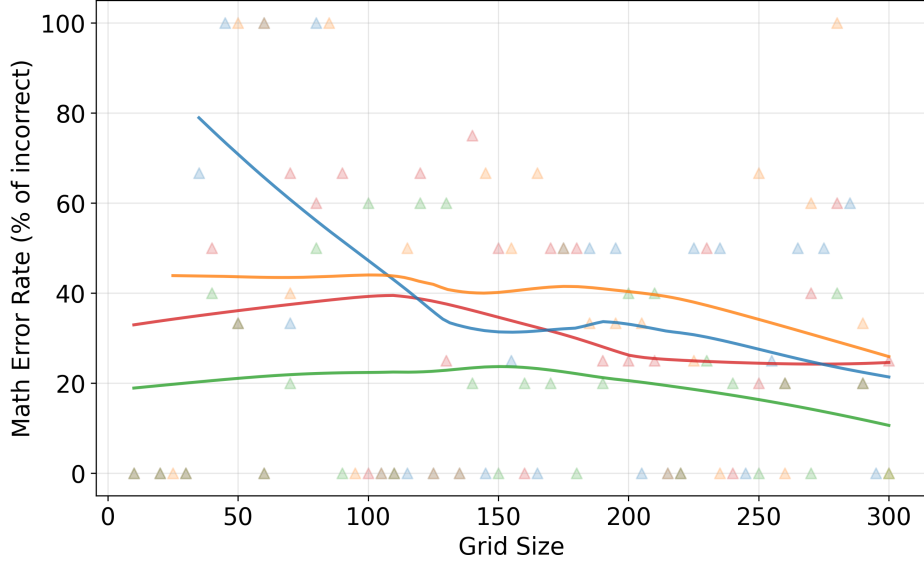


Figure 12: Math Error vs Grid Size (Distance)

4.4 Word Search

Models	Smallest 5 Grids	Largest 5 Grids
GPT-4o; 3 Letters	15%	0%
GPT-4.1; 3 Letters	72.5%	0%
Claude No Thinking; 3 Letters	95%	2%
Claude Medium Thinking; 3 Letters	95%	0%
GPT-4o; 6 Letters	0%	0%
GPT-4.1; 6 Letters	20%	0%
Claude No Thinking; 6 Letters	30%	0%
Claude Medium Thinking; 6 Letters	35%	0%
GPT-4o; 9 Letters	5.5%	0%
GPT-4.1; 9 Letters	47.2%	0%
Claude No Thinking; 9 Letters	80.5%	0%
Claude Medium Thinking; 9 Letters	72.2%	0%

Table 4: Model Accuracy on Small and Large Grids (Search)

This test was non-numerical in nature. Instead of calculations, we tested the models on language – a subject in which they traditionally excel. Grid sizes were tested from 5×5 to 100×100 with a step of 5, using words of length 3, 6, and 9. Words that did not fit in the board, ie. a nine-letter word for a 5×5 grid, were omitted from testing. All eight spelling directions (horizontal right-to-left, vertical up-to-down, etc) were tested for each word and grid size. Credit was given for percentage of coordinates correct, ranging from 0 to 1. This test was more difficult for the models from the start. The Anthropic models, which averaged close to 100% accuracy on smaller boards in the other tests, averaged only 50% for the six-letter words. The results for all three word lengths are seen in Figure 13.

Despite the low accuracy, however, the models still often claimed to have found the word. The Anthropic models had an almost 100% record of claimed detection, and the OpenAI models were somewhat lower. This was likely due to hallucination, the effects of which can be seen in Figure 14.

Additionally, models seemed to have trouble identifying words across multiple columns and in unusual orientations. The average accuracy rate was highest for words spelled traditionally, horizontally left-to-right, second-highest backwards spelling, horizontal right-to-left, and lowest for vertical spelling, down-to-up, as shown in Figure 15.

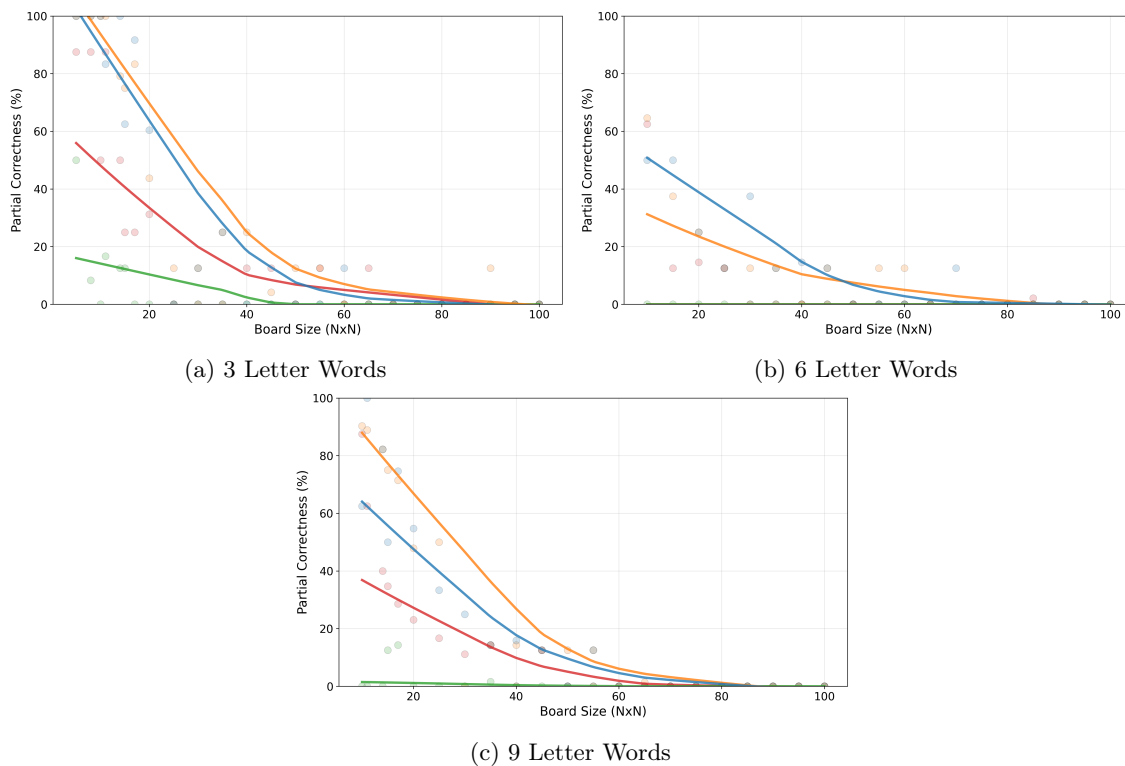


Figure 13: Accuracy vs Grid Size (Search)

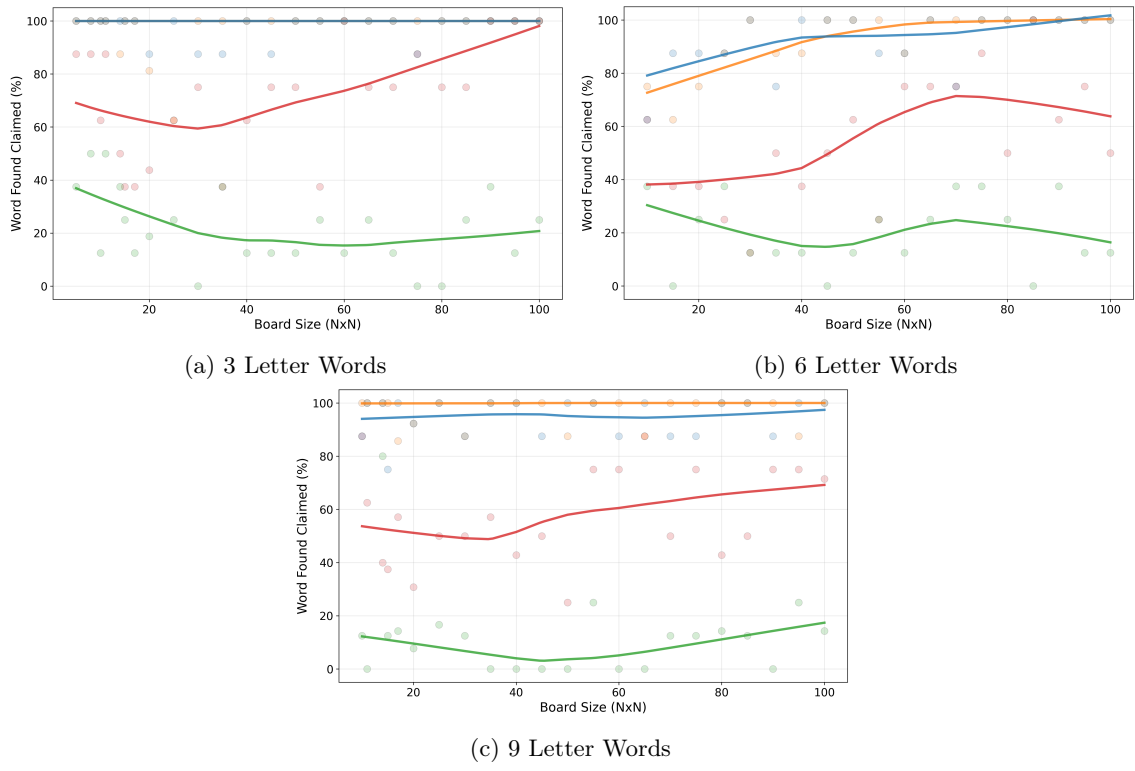


Figure 14: Claimed Detection vs Grid Size (Search)

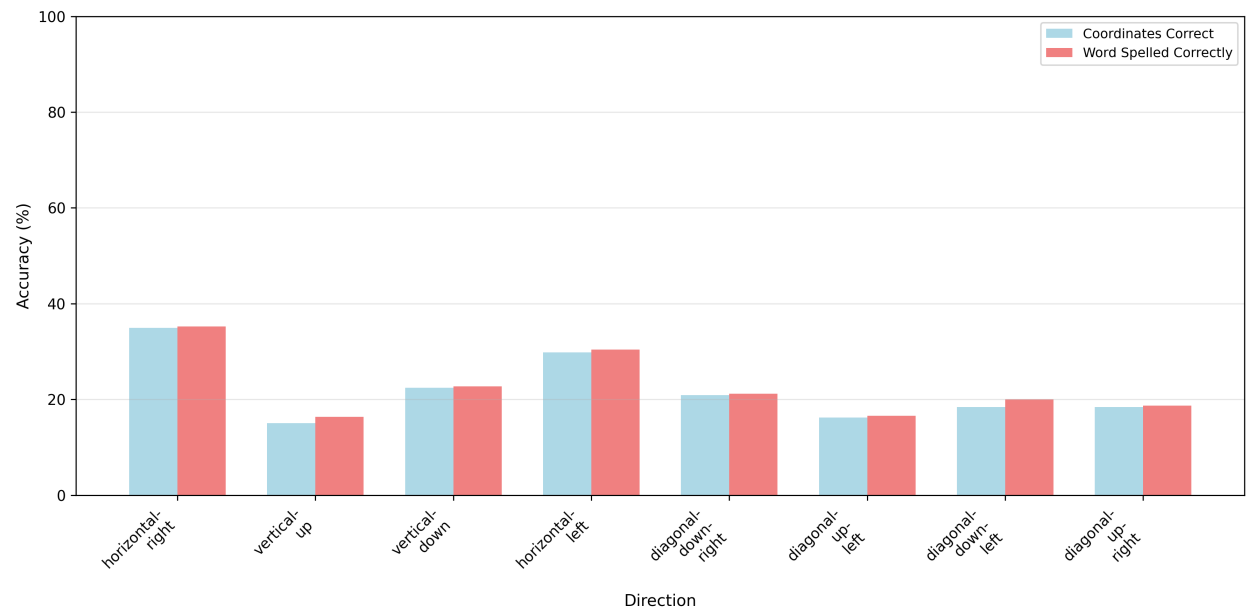


Figure 15: Accuracy by Direction (Search)

4.5 Slide

Models	Smallest 5 Grids	Largest 5 Grids
GPT-4o; 1 slide	28%	8%
GPT-4.1; 1 slide	64%	16%
Claude No Thinking; 1 slide	92%	2%
Claude Medium Thinking; 1 slide	88%	20%
GPT-4o; 2 slides	28%	16%
GPT-4.1; 2 slides	32%	4%
Claude No Thinking; 2 slides	76%	8%
Claude Medium Thinking; 2 slides	88%	24%
GPT-4o; 3 slides	12%	0%
GPT-4.1; 3 slides	44%	0%
Claude No Thinking; 3 slides	64%	12%
Claude Medium Thinking; 3 slides	64%	4%
GPT-4o; 4 slides	0%	8%
GPT-4.1; 4 slides	40%	12%
Claude No Thinking; 4 slides	56%	4%
Claude Medium Thinking; 4 slides	56%	4%
GPT-4o; 5 slides	0%	0%
GPT-4.1; 5 slides	28%	4%
Claude No Thinking; 5 slides	36%	4%
Claude Medium Thinking; 5 slides	44%	0%
GPT-4o; 6 slides	16%	0%
GPT-4.1; 6 slides	28%	0%
Claude No Thinking; 6 slides	32%	0%
Claude Medium Thinking; 6 slides	32%	4%

Table 5: Model Accuracy on Small and Large Grids (Slide)

This final task was designed to test relative geometry and reasoning. Grid sizes were tested from 5×5 to 75×75 , with increments of 5×5 . Each grid size was tested with 1-6 slides. The final test showed a similar pattern of strong performance on initial, smaller boards and quick performance deterioration as grid sizes grew. The Anthropic models consistently outperformed the OpenAI ones until a grid size of 70×70 , after which all accuracy was effectively 0. This deterioration was even quicker than in Word Search, implying that this test was overall more difficult despite the models having a more consistent performance on the smaller boards. When plotting accuracy, each slide of the multi-slide was treated independently to remove cascading effects. The results can be seen in Figure 16.

The number of slides was also a determining factor in the models’ average accuracy, with much quicker dropoffs in accuracy for prompts containing more slides. This was to be expected as the introduction of multi-step reasoning on top of the spatial reasoning likely compounded the difficulty. Even when asked to output intermediate steps the models showed no significant increase in performance, suggesting difficulty remembering where they had even placed the X before. A comparison of slides can be seen in Figure 17.

The consistency is somewhat more difficult to parse for this test. The models often ignored walls but it remains unclear whether they miscounted the position of the walls while reading board states or experienced a systemic error with wall detection itself. There were still some classic miscounting issues as well, as shown in Figure 18.

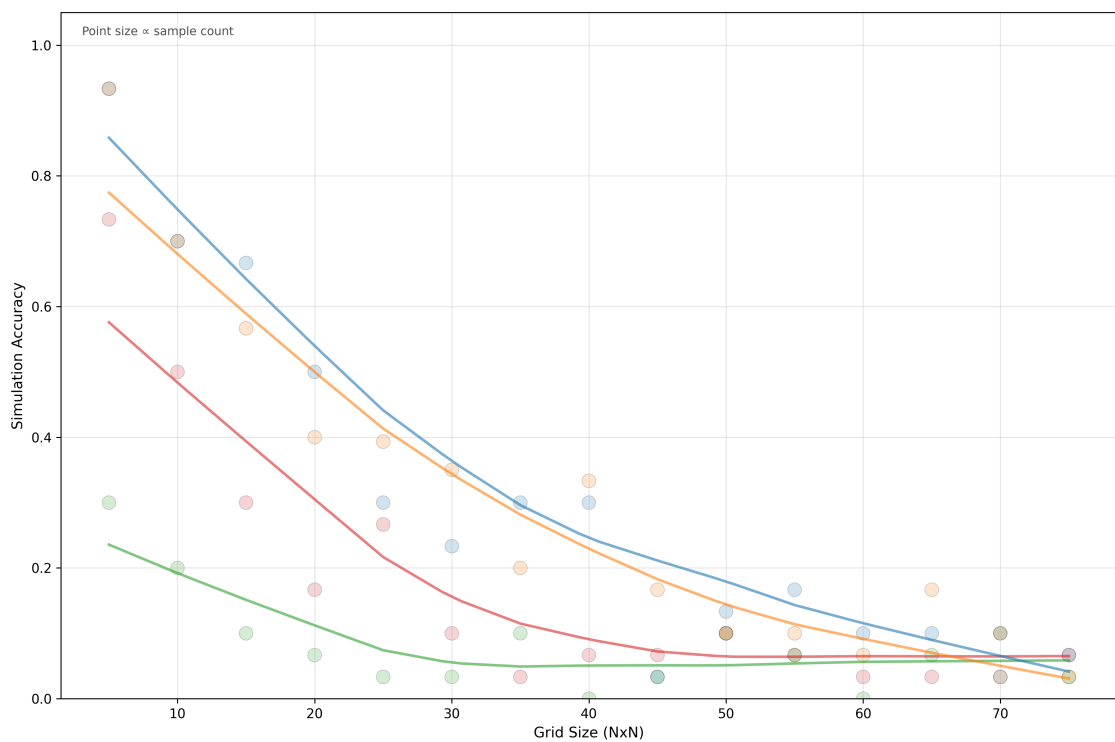


Figure 16: Accuracy vs Grid Size (Slide)

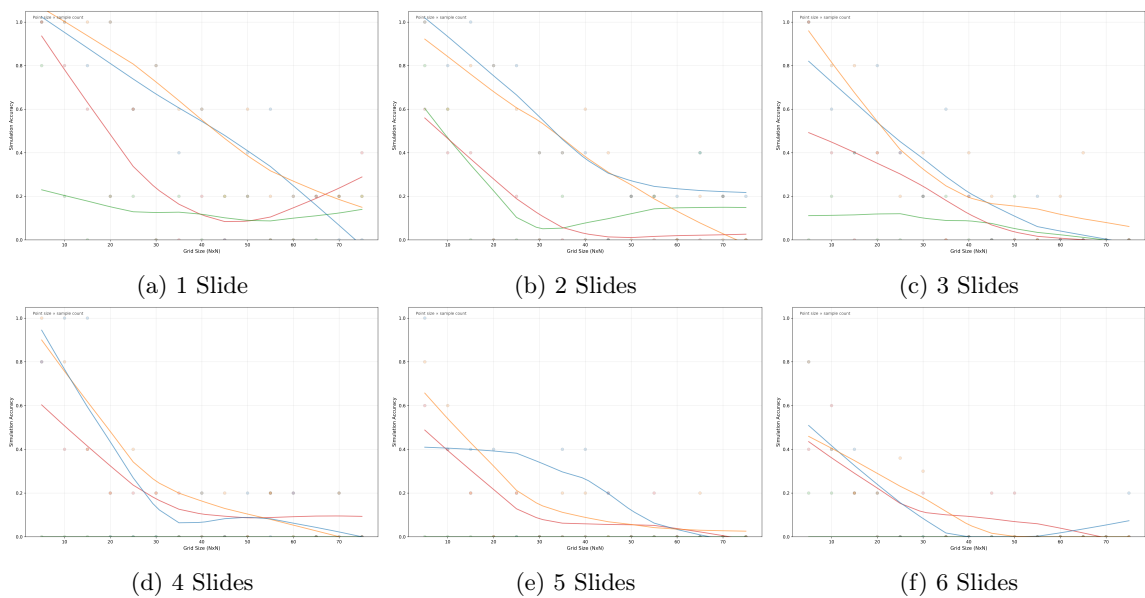


Figure 17: Accuracy vs Grid Size per Slide Count (Slide)

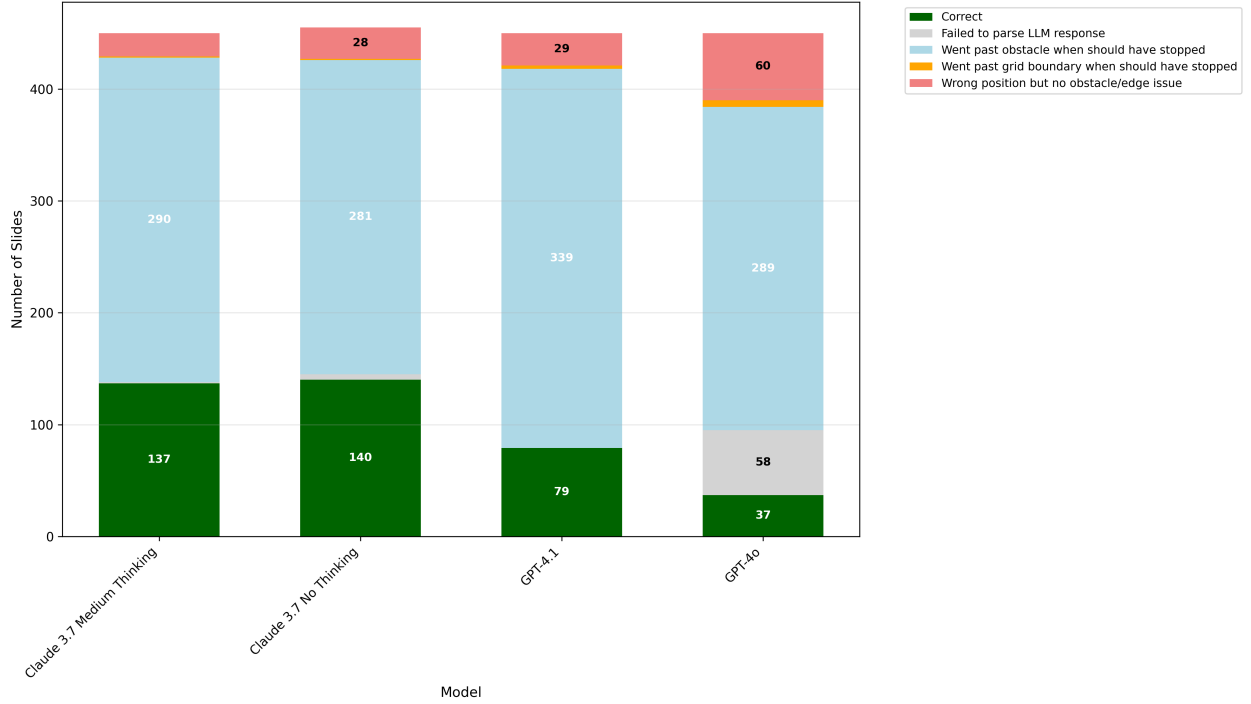


Figure 18: Error Types (Slide)

4.6 Parsing

We also encountered some parsing errors with the LLM responses despite efforts with both prompting and regex to make the parsing as reliable as possible. GPT-4o showed significantly higher rates of parsing errors, with 10% of the total responses in an unparseable or incomplete format. Parsing error rates also increased with grid size, demonstrating once again the further degeneration of model response reliability as complexity increases. These phenomena are seen in Figures 19 and 20.

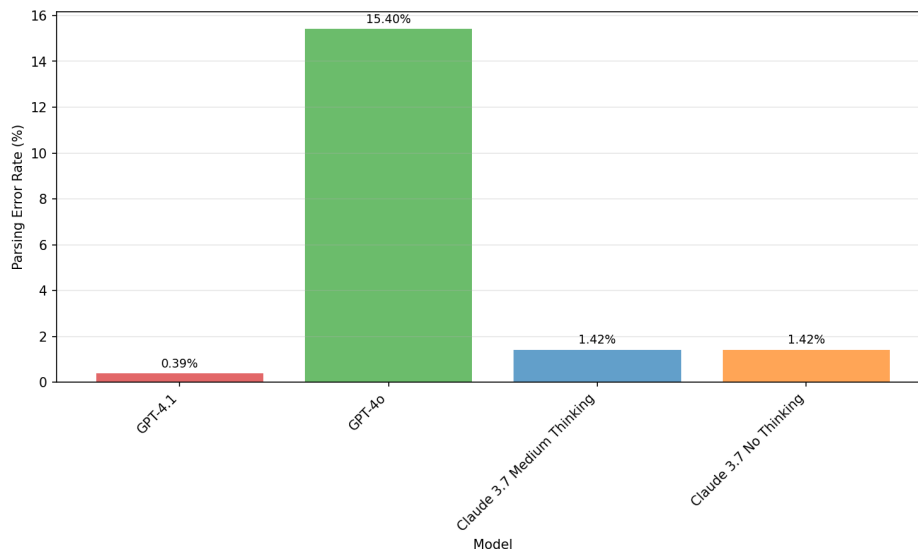


Figure 19: Parsing Error Rates (All tests)

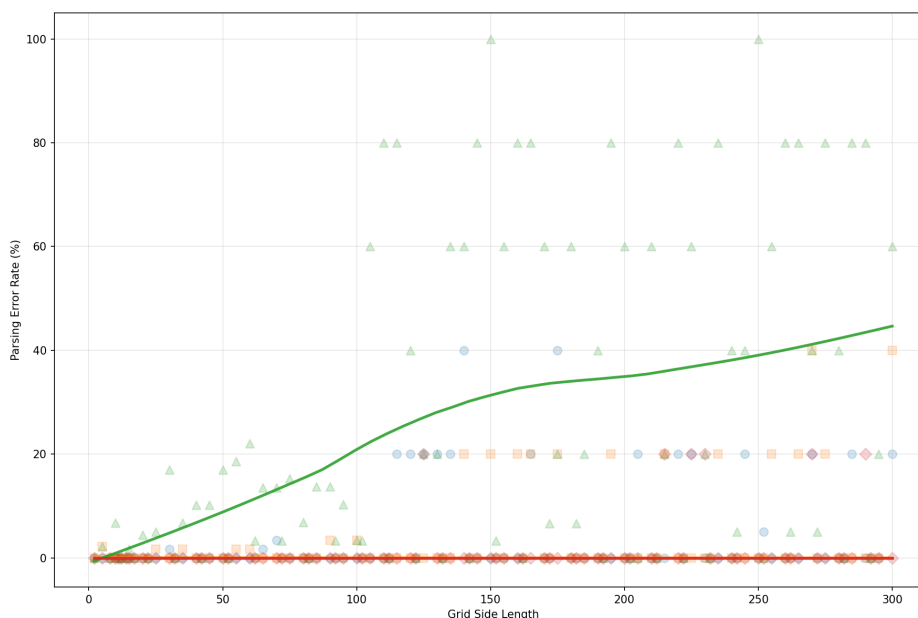


Figure 20: Parsing Error vs Grid Size (All tests)

4.7 Tokenization Changes

In an attempt to mitigate the loss of accuracy, we experimented with several layouts and tokenizations of the grid. Changes were tested on the Quadrant and Search tests. Quadrant was selected because as it was the most internally consistent of the tests, meaning that most problems lay with interpretation of the initial board. Search, on the other hand, was the most difficult, and therefore had the most room for improvement. New grid formats can be checked for tokenization with GPT-4o here, with GPT-4.1 using the same tokenizer. The tokenizer for the Anthropic models is private and therefore unavailable. As such, it should be noted that

all references to specific observed tokenizations hold solely for the OpenAI models.

4.7.1 Quadrant

The base grid for Quadrant was composed of “.”s separated by spaces, with an X, or other character, taking the place of a “.” where necessary, as seen in the Methodology. The new tests for quadrant can be seen in Figure 21, and the results are shown in Figures 22.

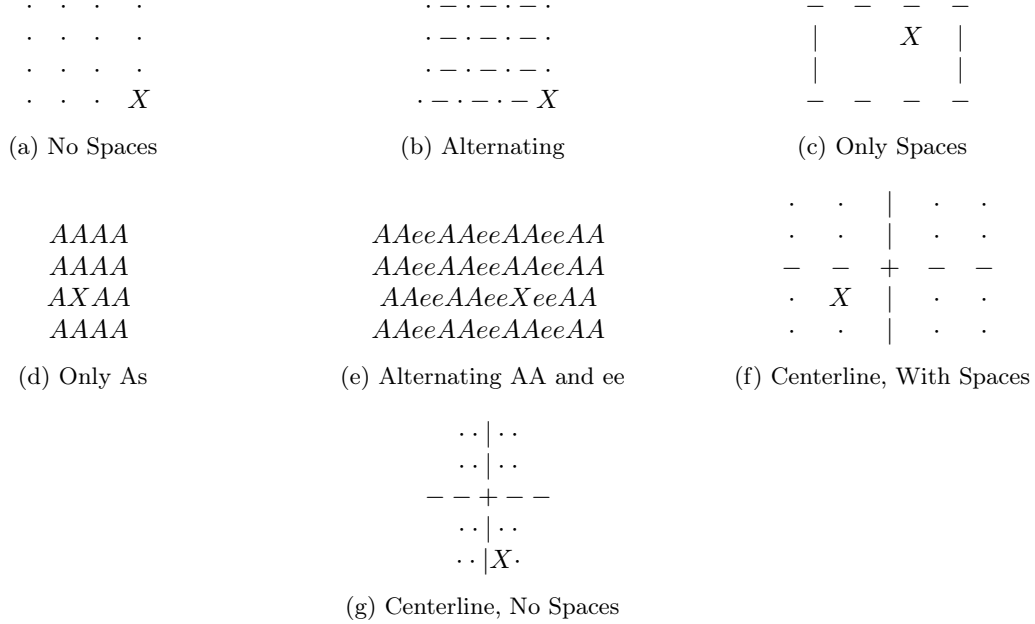


Figure 21: Types of Grids

A few of the grids—Only As, No Spaces, Only Spaces, and both Centerlines—show significant improvement over the baseline. For all tests except the centerline with spaces, a common theme emerges: the ability to tokenize large sections of a line as a singular token. In the baseline grid, each “.” was tokenized individually with the following space, which likely led to confusion and counting problems. With the no spaces grid, large sections could be put into a singular token instead, reducing the counting required and making the “X” stand out further. The upper bound of tokenization for “.”s appears to be 32 per token, but this is not regular. Smaller chunks than 32 are often split into two tokens despite being under the limit.

It perhaps seems a moot point to suggest that giving the model the center line improves performance, as it somewhat defeats the purpose of the test. Yet, there are still a few interesting observations to be gleaned from this. First, the addition of centerlines with spaces showed only moderate improvement over the baseline, implying that the models still have a difficult time identifying position in a spaced-out grid, even when given aids for the reasoning portion. Second, though, the increased accuracy implies that the models were using the centerlines as guides. This shows that they are capable of spatial reasoning in a non-strictly mathematical sense, as they could use identifiers like “left” or “above” the lines to select the quadrant. Additionally while the Anthropic models average practically 100% accuracy given the no spaces and centerline, the OpenAI models only average 80%. While this is a significant improvement, especially on larger boards, it shows that the OpenAI models still have a higher hallucination rate than the Anthropic models, as they still misplace the initial X.

It is also interesting that the test grid of alternating “AA”s and “ee”s performed worse than the baseline. The likely cause of this is that the two sets of characters always tokenize independently of one another, unlike the “.”s that tokenize with the space following. This would lead to the model needing to count and keep

track of twice as many entries, resulting in a loss in accuracy due to counting failures. The alternating test also performed worse, even though the grids were tokenized as “.-” pairs. This is probably due to a lower familiarity with such formats in the training data of the models.

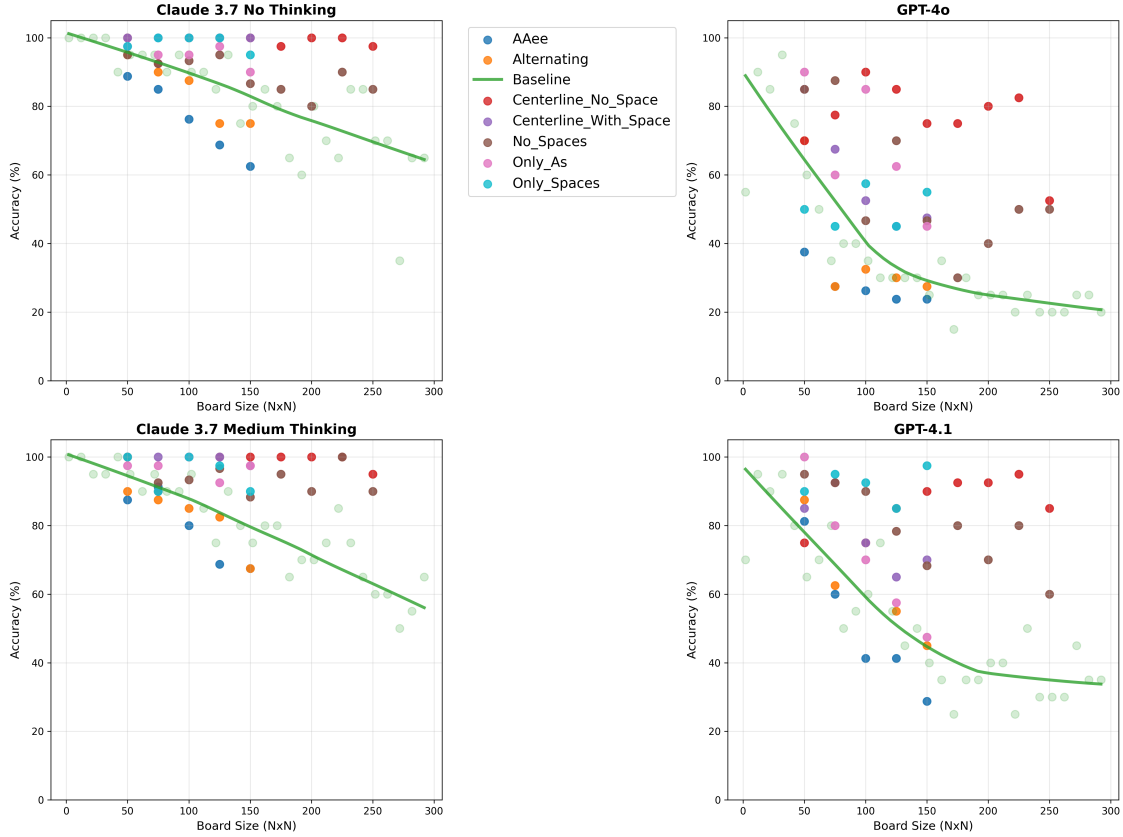


Figure 22: Accuracy vs Grid Size, New Tokenizations (Quad)

4.7.2 Search

The base grid for Search was letters separated by spaces, as seen in Methodology. The new tests for search can be seen in Figure 23, and the results can be seen in Figure 24.

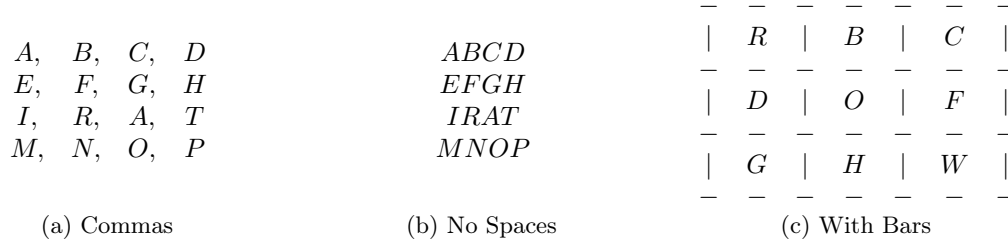


Figure 23: Types of Grids

We also trialed a non-grid based test, with the grid presented in coordinate form of $(row, column, value)$. Additionally, we tried asking the models to output more thorough step-by-step reasoning before their final answer.

As seen in the results, all tests failed to show any significant improvement over the baseline. The no spaces test, which performed significantly better in the Quadrant trials, showed a much steeper drop in accuracy here. Looking at the tokenization, in the baseline grid, each letter and the following space gets tokenized individually, like in Quadrant. However, without the spaces, the model attempts to form words from the letters, tokenizing them in small chunks. Ironically, even given a word spelled left-to-right, the tokenizer often misses it in favor of smaller groupings. It is interesting as well that step-by-step reasoning has seemingly little effect on performance. The models still skip test cases or miscount letter positions, resulting in a failed identification. These results imply that while LLMs typically excel at language processing, they are still weak in processing in non-traditional formats, such as vertically or diagonally, and are unable to reason their way through the process adequately.

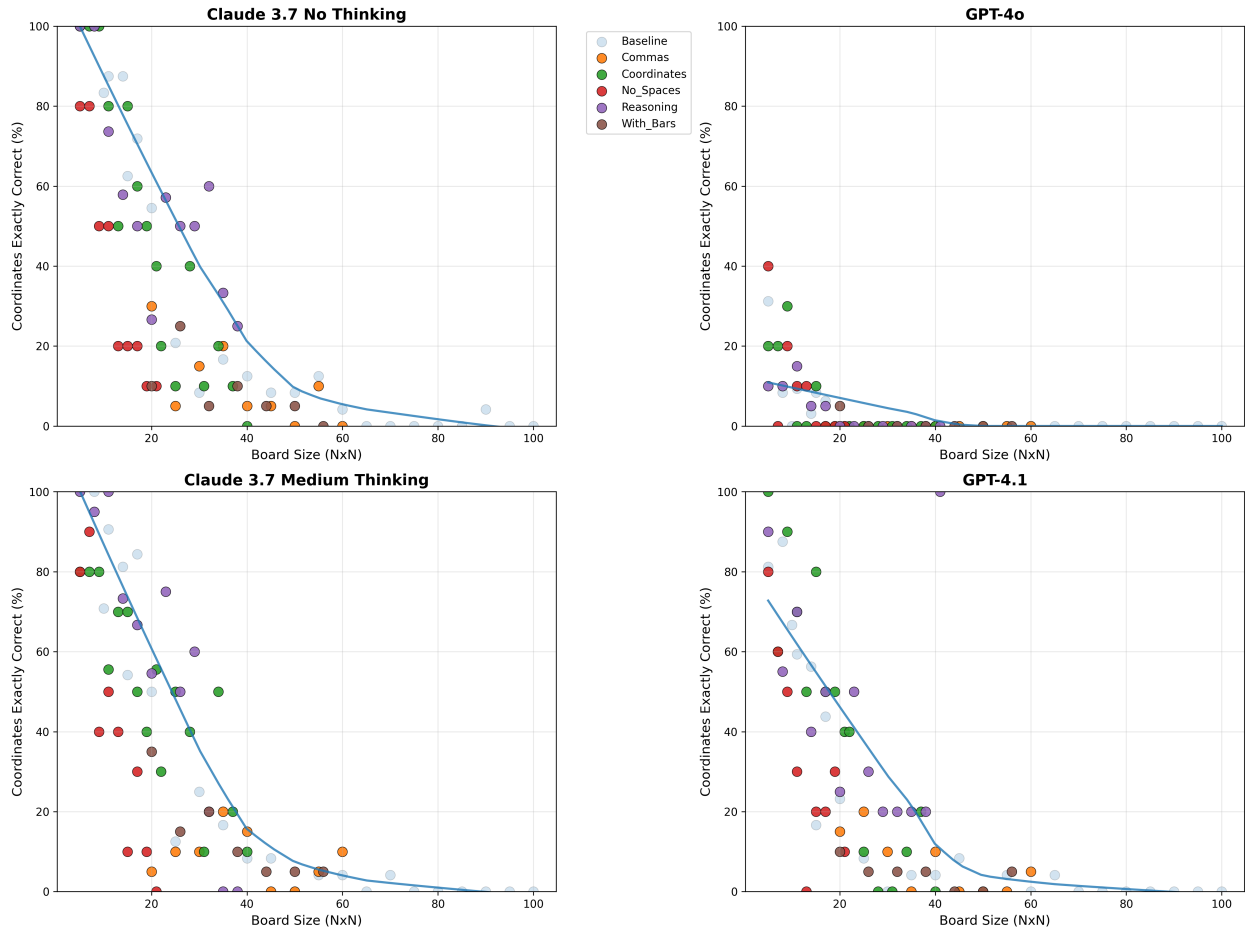


Figure 24: Accuracy vs Grid Size, New Tokenizations (Search)

5 Discussion

The results found in the above experiments underscore the fragility and unreliability of LLM spatial reasoning at scale. Across all tests and models, increased grid size led to decreased accuracy and increased error rate. Most models scored high in accuracy on smaller boards, meaning that the tests were not too complex to handle conceptually. Additionally, errors were caused by more than simple coordinate misidentification. For instance, models had difficulty calculating the center lines of the grid in the Quadrant test and made frequent

mathematical errors in the Distance test. These types of errors show a more fundamental misunderstanding of spatial reasoning by these LLMs, rather than simply a counting problem.

This degradation emphasizes the fact that these models are built for linguistic tasks, and, while presenting some emergent behavior [16] [17], are still limited by their architecture. The models possess the reasoning capabilities to handle small examples of a task, but lose track of the grid and task once scaled.

The differences between the models themselves are just as interesting. It is immediately evident across the tests that GPT-4o is the weakest at spatial reasoning, followed by GPT-4.1. Both Anthropic models consistently outperformed the OpenAI models, with gaps of up to 30% in accuracy. GPT-4o also often failed to provide a parseable response despite establishing extensive regex and prompt specifications, whereas none of the other models encountered any such problems.

The error types are interesting as well. The OpenAI models are less internally consistent than the Anthropic ones, displaying weaker consistency in the Quadrant and Transformation tests. In addition, the OpenAI models made more errors in the Distance test, both in position and computation. However, the OpenAI models were less likely to hallucinate a correct answer in the Word Search test. This behavior, of the OpenAI models performing worse in the computational tests, but better in the language-based one, is likely a result of the models’ different architectures and training.

6 Conclusion

While the Anthropic models outperformed the OpenAI models using this test suite, it is clear that none of the models are particularly good at spatial reasoning. The abilities of the Anthropic models hold promise, but they still deteriorate quickly with scale. We propose a few avenues for future research on this deterioration. For the top models currently released, presentation format and context can significantly affect response quality. For instance, the models tested are all multimodal, and would likely perform better on similar tests presented in image form; future work could focus on directly testing these various modes of input. Other future work could follow up on the changes in grid representation attempted, with an interest in improving efficiency or effectiveness of representations of spatial data in a text-heavy format, in efforts to reduce or minimize the loss in accuracy shown. Further work could also focus on improving the counting and mathematical capabilities of LLMs, as these two skills were the root of many of the inaccurate responses. Finally, research could focus on other methods of output, such as code generation or full boards.

Acknowledgment

The authors would like to thank Jump Trading, as well as our mentors Lucas Baker, Nan Yang, Baihong Jin, and Loren Puchalla Fiore.

References

- [1] Anand, A., Jaiswal, R., Dharmadhikari, A., Marathe, A., Popat, H. P., Mital, H., Prasad, K., Shah, R. R., and Zimmermann, R. Improving multimodal llms ability in geometry problem solving, reasoning, and multistep scoring, 2024.
- [2] Anthropic. Claude 3.7 sonnet system card, 2025.
- [3] Brown, T. B., Mann, B., Ryder, N., and et al. Language models are few-shot learners, 2020.
- [4] Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024.
- [5] Chollet, F., Knoop, M., Kamradt, G., and Landers, B. Arc prize 2024: Technical report, 2024.

- [6] Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2025.
- [7] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset, 2021.
- [8] Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2021.
- [9] OpenAI. Gpt-4o system card, 2024.
- [10] OpenAI. Introducing 4.1 in the api, 2025.
- [11] OpenAI, Achiam, J., Adler, S., and et al. Gpt-4 technical report, 2023.
- [12] Stogiannidis, I., McDonagh, S., and Tsaftaris, S. A. Mind the gap: Benchmarking spatial reasoning in vision-language models, 2025.
- [13] Topsakal, O., Edell, C. J., and Harper, J. B. Evaluating large language models with grid-based game competitions: An extensible llm benchmark and leaderboard, 2024.
- [14] Tsai, C. F., Zhou, X., Liu, S. S., Li, J., Yu, M., and Mei, H. Can large language models play text games well? current state-of-the-art and open questions, 2023.
- [15] Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., and Liu, X. Docllm: A layout-aware generative language model for multimodal document understanding, 2023.
- [16] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022.
- [17] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2022.