Tokenised rep of maze
Baseline cant solve. SFT 88%. RL with
GRPO - 93%

AlphaMaze: Enhancing Large Language Models' Spatial Intelligence via GRPO

Alan Dao (Gia Tuan Dao)*, Dinh Bach Vu*
Menlo Research
{alan, bach}@menlo.ai

***Abstract*—Large Language Models (LLMs) have demonstrated impressive capabilities in language processing, yet they often struggle with tasks requiring genuine visual spatial reasoning. In this paper, we introduce a novel two-stage training framework designed to equip standard LLMs with visual reasoning abilities for maze navigation. First, we leverage Supervised Fine-Tuning (SFT) on a curated dataset of tokenized maze representations to teach the model to predict step-by-step movement commands. Next, we apply Group Relative Policy Optimization (GRPO)—a technique used in DeepSeek-R1—with a carefully crafted reward function to refine the model's sequential decision-making and encourage emergent chain-of-thought behaviors. Experimental results on synthetically generated mazes show that while a baseline model fails to navigate the maze, the SFT-trained model achieves 86% accuracy, and further GRPO fine-tuning boosts accuracy to 93%. Qualitative analyses reveal that GRPO fosters more robust and self-corrective reasoning, highlighting the potential of our approach to bridge the gap between language models and visual spatial tasks. These findings offer promising implications for applications in robotics, autonomous navigation, and other domains that require integrated visual and sequential reasoning.**

I. Introduction

The ability to reason about visual information, particularly in spatial contexts, is a hallmark of intelligent systems. From navigating physical environments to interpreting complex diagrams, visual spatial reasoning is crucial for a wide range of tasks. While Large Language Models (LLMs) have achieved impressive performance in natural language processing and code generation, their capacity for genuine visual reasoning, especially spatial understanding and sequential decision-making in visual environments, remains a significant open question [Zhang et al., 2024, Ma et al., 2024]. Current Vision-Language Models (VLMs) often excel at pattern recognition and object identification but may struggle with tasks requiring deeper spatial inference and step-by-step planning in visual domains [Ma et al., 2024]. Bridging this gap and endowing standard LLMs with robust visual reasoning capabilities is a critical step towards more versatile and human-like AI.

In this paper, we address the challenge of teaching visual spatial reasoning to a standard LLM, focusing on the task of maze navigation. We hypothesize that by providing an LLM with a tokenized *visual* representation of a maze, we can train it to learn step-by-step movement commands to navigate from a designated origin to a target. The core of our approach lies in a two-stage training framework. First, we employ Supervised Fine-Tuning (SFT) to equip the LLM with the foundational skill of predicting movement tokens based on the visual maze input. Subsequently, we apply Group Relative Policy Optimization (GRPO), drawing inspiration from recent advancements in reinforcement learning for reasoning in LLMs, such as DeepSeek-R1 [Guo et al., 2025]. DeepSeek-R1 demonstrated that Reinforcement Learning (RL) can elicit emergent reasoning behaviors, including chain-of-thought, even without prior SFT. We adapt and extend these RL strategies, combined with carefully designed reward functions, to refine our model's visual reasoning process for maze navigation.

To systematically evaluate LLM's ability to solve maze, we introduce MazeBench—a comprehensive benchmark on solving maze. MazeBench provides a controlled yet diverse environment that spans a range of maze sizes and complexities. By evaluating our model on MazeBench, we can rigorously measure both its maze-solving accuracy and the sophistication of its emergent reasoning behavior.

Our key contributions are as follows:

- We present a novel training framework that combines Supervised Fine-Tuning and Group Relative Policy Optimization to enhance *visual* reasoning in standard LLMs, specifically for spatial tasks.
- We empirically demonstrate that this framework, using a tokenized visual maze representation, enables an LLM to achieve improved maze navigation accuracy and exhibit emergent chain-of-thought reasoning in generating movement sequences.
- We provide a detailed analysis of the design and impact of reward functions within the GRPO stage, highlighting their crucial role in shaping the model's visual reasoning performance.
- We draw comparisons with insights from state-of-the-art reasoning models like DeepSeek-R1, both in terms of methodology and observed emergent behaviors, positioning our work within the context of current advancements in LLM reasoning.
- We present MazeBench, a benchmark for visual maze navigation that captures a wide spectrum of spatial challenges.

II. Related Work

A. Chain-of-Thought Reasoning in Language Models

Chain-of-Thought (CoT) prompting has emerged as a powerful technique to elicit complex reasoning from Large Language Models [Wei et al., 2022b]. By prompting LLMs to "think step by step," CoT encourages the generation of intermediate reasoning steps, leading to improved performance on tasks requiring multi-step inference. Wei et al. [2022b] demonstrated that CoT prompting significantly enhances the ability of LLMs to solve arithmetic, commonsense reasoning, and symbolic reasoning tasks. Our work builds upon the concept of CoT reasoning, aiming to induce a similar step-by-step thought process in LLMs, but within the domain of visual spatial reasoning for maze navigation.

B. Supervised Fine-Tuning for Visual and Spatial Tasks

Supervised Fine-Tuning (SFT) is a widely adopted technique for adapting pre-trained LLMs to specific downstream tasks [Wei et al., 2022a]. By training on task-specific datasets, SFT allows LLMs to acquire specialized skills and improve performance in targeted domains. Jiang et al. [2024] recently highlighted the effectiveness of SFT in enhancing visual foundation models, demonstrating its utility in visual tasks. In our research, we leverage SFT as the initial stage of our training pipeline, using it to equip the LLM with the basic capability of processing tokenized visual maze inputs and predicting movement tokens. This SFT phase serves as a crucial foundation upon which we build more sophisticated reasoning through reinforcement learning.

C. Reinforcement Learning and GRPO for Reasoning and Reward Shaping

Reinforcement Learning from Human Feedback (RLHF) and its variants, such as Group Relative Policy Optimization (GRPO), have become increasingly important for aligning LLMs with human preferences and improving their reasoning abilities [Kwon et al., 2023].GRPO, as described by Shao et al. [2024] and implemented in DeepSeek-R1 [Guo et al., 2025], offers a computationally efficient approach to reinforcement learning by estimating advantages based on group scores, eliminating the need for a separate critic network. Reward function design is paramount in RLHF and GRPO, as it directly guides the model's learning process. Carefully crafted reward functions can incentivize desired behaviors and shape the model's policy towards optimal performance. Our work draws inspiration from the reward shaping strategies used in DeepSeek-R1 and adapts them to the context of visual maze navigation, designing reward components to encourage accuracy, valid movement sequences, and proper output formatting.

D. DeepSeek-R1 and Emergent Reasoning through RL

The DeepSeek-R1 model [Guo et al., 2025] represents a significant advancement in using reinforcement learning to elicit sophisticated reasoning capabilities in LLMs. A key finding of DeepSeek-R1 is the demonstration that pure RL, specifically GRPO, can lead to the *emergent* development of chain-of-thought reasoning and even ”aha moments,” where the model re-evaluates previous steps and corrects its reasoning process. Furthermore, DeepSeek-R1 highlights the benefits of a multi-stage training pipeline, combining initial RL training with subsequent supervised fine-tuning to refine language coherence and readability. We directly adapt the GRPO optimization strategy and multi-stage training insights from DeepSeek-R1 to our visual maze navigation task. We hypothesize that similar RL techniques can drive the emergence of visual spatial reasoning in standard LLMs, enabling them to solve mazes through a step-by-step, self-corrective process.

E. Visual Reasoning and Maze Solving in AI

Maze solving has long been a benchmark task in Artificial Intelligence, serving as a testbed for various problem-solving and search algorithms [Janamian and Alam, 2023]. Traditional approaches include graph search algorithms like Depth-First Search, Breadth-First Search, and A* [Lester, 2014-2024]. More recently, AI techniques, particularly reinforcement learning and neural networks, have been applied to maze navigation [Zafrany, 2020]. While prior work has explored maze solving using AI, our research focuses on a novel approach: teaching *visual* maze reasoning to standard *language models* through a tokenized visual representation and a combination of SFT and GRPO. This approach differs from traditional maze solvers by leveraging the inherent reasoning capabilities of LLMs and adapting them to process and reason about visual spatial information. Furthermore, research in neural-symbolic visual reasoning [Mao et al., 2023] explores combining neural networks with symbolic AI for visual tasks, offering a complementary perspective on integrating reasoning and visual processing.

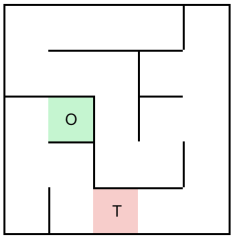


Fig. 1. Visual of the Example Maze

III. Methodology

A. Tokenized Visual Maze Representation

To enable the LLM to process maze information visually, we designed a tokenized input format that represents the maze grid, walls, origin, and target locations. Each cell in the maze is represented by a coordinate token `<|row-col|>`, e.g., `<|0-0|>` for the top-left cell. Wall information for each cell is encoded using tokens such as `<|no_wall|>`, `<|up_wall|>`, `<|down_wall|>`, `<|left_wall|>`, `<|right_wall|>`, `<|up_down_wall|>`, `<|left_right_wall|>`, `<|up_left_wall|>`, `<|up_right_wall|>`, `<|down_left_wall|>`, `<|down_right_wall|>`, `<|up_down_left_wall|>`, `<|up_down_right_wall|>`, `<|up_left_right_wall|>`, `<|down_left_right_wall|>`, `<|up_down_left_right_wall|>`. The origin and target locations are marked with `<|origin|>` and `<|target|>` tokens, respectively. Empty spaces within the maze representation are filled with `<|blank|>` tokens for consistent grid structure. This tokenization scheme provides a visual representation by explicitly encoding the spatial relationships between cells and the presence of walls, allowing the LLM to “see” the maze structure in a symbolic, tokenized form.

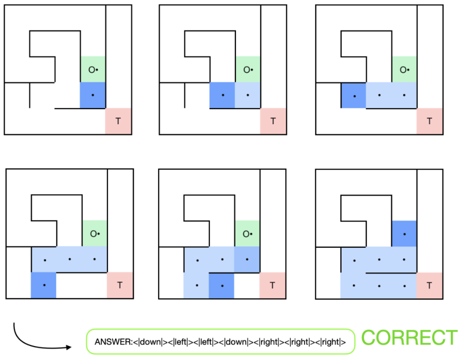
Example Maze Tokenization 1:

```
<|0-0|><|up_left_wall|><|blank
    |><|0-1|><|up_down_wall|><|blank
    |><|0-2|><|up_down_wall|><|blank
    |><|0-3|><|up_down_right_wall|><|
    blank|><|0-4|><|up_left_right_wall
    |><|blank|>
<|1-0|><|down_left_wall|><|blank
    |><|1-1|><|up_down_wall|><|blank
    |><|1-2|><|up_right_wall|><|blank
    |><|1-3|><|up_down_left_wall|><|
    blank|><|1-4|><|right_wall|><|blank
    |>
<|2-0|><|up_left_wall|><|blank
    |><|2-1|><|up_down_right_wall|><|
    origin|><|2-2|><|left_right_wall
    |><|blank|><|2-3|><|up_left_wall
    |><|blank|><|2-4|><|right_wall|><|
    blank|>
<|3-0|><|left_wall|><|blank|><|3-1|><|
    up_right_wall|><|blank|><|3-2|><|
    down_left_wall|><|blank|><|3-3|><|
    down_right_wall|><|blank|><|3-4|><|
    left_right_wall|><|blank|>
<|4-0|><|down_left_right_wall|><|blank
    |><|4-1|><|down_left_wall|><|blank
    |><|4-2|><|up_down_wall|><|target
    |><|4-3|><|up_down_wall|><|blank
    |><|4-4|><|down_right_wall|><|blank
    |>
```

B. *Baseline Models*

To establish performance benchmarks for our approach, we employed three distinct baseline models, leveraging the DeepSeek-R1 [Guo et al., 2025] Distill-Qwen family of language models:

- **DeepSeek-R1-Distill-Qwen-7B**: This represents a larger, 7 billion parameter model from the DeepSeek-R1 Distill-Qwen series, providing a strong performance reference point.
- **DeepSeek-R1-Distill-Qwen-1.5B**: This is a smaller, 1.5 billion parameter counterpart, allowing us to assess the impact of model size within the same architecture family.
- **Direct Prediction Model (Customized DeepSeek-R1-Distill-Qwen-1.5B with SFT)**: Based on the 1.5B architecture, this baseline model was specifically fine-tuned using Supervised Fine-Tuning (SFT) to directly predict the entire sequence of movement tokens required to solve a maze in a single forward pass. The training objective was to minimize the cross-entropy loss between the predicted token sequence and the ground truth path. This direct prediction baseline serves to establish the performance of a standard LLM trained to output complete solutions without explicit step-by-step reasoning or reinforcement learning enhancement. The input to this baseline model was the tokenized maze representation, and the expected output was the concatenated sequence of movement tokens (e.g., `<|up|> <|left|> <|right|> <|down|> ...`).



ANSWER:<|down|><|left|><|left|><|down|><|right|><|right|><|right|>

Fig. 2. Visualization of AlphaMaze's step-by-step reasoning process while solving a maze.

We include these three baselines to provide a comprehensive comparison, examining the influence of model size (7B vs. 1.5B) and the effectiveness of direct prediction versus our proposed step-by-step and reinforcement learning approaches. The subsequent sections will primarily focus on the customized direct prediction model and its enhancements through SFT for step-by-step reasoning and GRPO.

C. *Supervised Fine-Tuning (SFT) for Step-by-Step Reasoning*

For the SFT stage, we curated a training dataset. Mazes were synthetically generated with fixed sizes (5x5) and varied complexity level, controlled by parameters influencing path length and branching factor. Each maze in the dataset was paired with an annotated step-by-step solution, represented as a sequence of movement tokens (`<|up|>`, `<|down|>`, `<|left|>`, `<|right|>`). The Qwen 1.5B SFT model was then trained on this dataset. The training objective was to predict the *next* movement token at each step, conditioned on the maze input and the preceding movement tokens in the sequence as visually illustrated in Figure 2. This step-by-step prediction approach was designed to encourage the model to learn sequential reasoning for maze navigation.

D. Group Relative Policy Optimization (GRPO) for Enhanced Reasoning

Following SFT, we applied Group Relative Policy Optimization (GRPO) to further enhance the model's maze-solving capabilities and encourage more robust reasoning. The GRPO training utilized a smaller set of data than SFT state. We designed a reward function with the following components:

- **Correctness Reward (+0.2 per solution step):** This reward is scaled according to the number of steps in the maze solution. Each valid movement step adds 0.2 points to the total score. For example, a solution requiring 4 steps earns a reward of $0.2 \times 4 = 0.8$ points, incentivizing both accuracy and efficiency in navigation.
- **Integrity Reward (+0.5):** This reward is given for each valid movement token (`<|up|>`, `<|down|>`, `<|left|>`, `<|right|>`) in the predicted sequence, encouraging the generation of meaningful and valid movement steps.
- **`<think>` Tag Reward (+0.25):** This reward is given for correctly using the `<think>` tag in the output, ensuring completeness and consistency in the reasoning format.

These reward components were weighted to prioritize correctness while also encouraging valid movement sequences and proper reasoning formatting with `<think>` tag. We adapted the Group Relative Policy Optimization (GRPO) algorithm, as employed in DeepSeek-R1 [Guo et al., 2025], to perform reinforcement learning. GRPO estimates advantages based on relative group scores, offering computational efficiency compared to critic-based methods.

E. Training Procedure and Pipeline

Our training pipeline consisted of two stages:

1) **Supervised Fine-Tuning (SFT):** The Qwen 1.5B SFT model was trained on the curated maze dataset for 10 epochs to learn step-by-step movement prediction for maze navigation.
2) **Group Relative Policy Optimization (GRPO):** The SFT-trained model was subsequently further fine-tuned using the GRPO method, leveraging the designed reward function. Model checkpoints were saved at intervals of 200 steps.

This two-stage pipeline mirrors the multi-stage training approach employed in DeepSeek-R1 [Guo et al., 2025], where initial RL training is followed by supervised fine-tuning for refinement. In our case, we used SFT *before* GRPO to provide a strong initial policy for the RL stage to build upon, focusing GRPO on refining reasoning and improving task-specific performance. All experiments were conducted using NVIDIA A6000 GPUs with LORA for parameter-efficient fine-tuning.

IV. Experiments and Results

A. Dataset Details

The SFT training dataset comprises 500,000 synthetically generated mazes, created using the **maze-dataset** framework Ivanitskiy et al. [2023]. This framework employs a randomized depth-first search algorithm, guaranteeing the existence of a solution path (connectivity) between the designated origin and target points within each maze. All mazes within the dataset are of a fixed size, 5x5 grids, providing a consistent spatial context for evaluation.

The full dataset is divided into three subsets: 500,000 mazes designated for SFT, providing the foundation for initial training; 16,000 mazes specifically allocated for GRPO, refining the model through policy optimization; and 30,000 mazes reserved for evaluation, ensuring a robust assessment of model performance.

16K GRPO

30K eval

Difficulty of task depends on no. of steps reqd to solve it

B. MazeBench

To rigorously evaluate the spatial reasoning and planning capabilities of large language models (LLMs), we introduce MazeBench, a novel benchmark consisting of a curated collection of 100 maze-solving challenges. While existing benchmarks often assess logical reasoning or commonsense knowledge, MazeBench specifically targets the ability of LLMs to understand spatial relationships, plan multi-step paths, and execute sequential actions within a constrained environment. This capacity is crucial for applications ranging from robotics and navigation to game playing and virtual agent control.

MazeBench is structured into three distinct difficulty levels – Easy, Medium, and Hard – to provide a graduated assessment of LLM performance. The difficulty is primarily determined by the approximate number of steps required for a viable solution path from the origin to the target. The benchmark comprises the following distribution:

- Easy (50 mazes): Mazes in this category typically require a solution path of 1-4 steps. These mazes serve as a baseline assessment, testing fundamental pathfinding abilities.
- Medium (40 mazes): These mazes necessitate solutions involving 5-8 steps, demanding more sophisticated planning and spatial reasoning.
- Hard (10 mazes): Hard mazes present the most significant challenge, requiring solutions of 9-13 steps. These mazes test the limits of an LLM's ability to handle long-range dependencies and complex spatial configurations.

As mentioned previously, the mazes are presented to the LLM in a tokenized input forma; the full details of this representation, including examples, are provided in Section III-A. The LLM is expected to produce output containing the following movement tokens: `<|up|>`, `<|down|>`, `<|left|>`, `<|right|>`. During evaluation, we will parse the LLM's output to extract these tokens. The order of these tokens is crucial. The presence of extraneous characters, whitespace, or other tokens will not automatically invalidate the solution, provided that the correct sequence of movement tokens can be extracted. A solution is considered incorrect if the extracted sequence of movement tokens does not lead to the target or leads to an invalid state (e.g., attempting to move into a wall) is considered incorrect. The evaluation metric is the success rate: the percentage of mazes solved correctly.

Quite strict eval criteria

C. Quantitative Results

1) Model Performance on MazeBench: As shown in Table I, the baseline model, trained for direct path prediction without explicit reasoning, achieved 0% accuracy on MazeBench. This highlights the necessity of step-by-step reasoning for the task. The SFT-only model reached a baseline of 86.0%, demonstrating the effectiveness of supervised fine-tuning for learning step-by-step maze navigation. Further enhancement with GRPO led to significant improvement, reaching 93.0% after 1600 steps of GRPO training.

TABLE I
MAZE SOLVING ACCURACY ON MAZEBENCH

Model	SFT	GRPO	Score (%)
Baseline-1.5B	✗	✗	0.0
Baseline-7B	✗	✗	0.0
Baseline-1.5B (SFT)	✓	✗	0.0
AlphaMaze-SFT	✓	✗	86.0
AlphaMaze	✓	✓	**93.0**

2) Model Evolution During GRPO: Figure 3 displays the MazeBench scores (blue crosses) over GRPO steps along with a linear regression trendline (red dashed line) and its ± 1 standard deviation bounds. The steady increase in the trendline indicates that GRPO effectively guides the model towards improved maze-solving policies.

D. Qualitative Results

Qualitative analysis of model outputs revealed notable differences in reasoning behavior across the models. The baseline model often produced nonsensical or incomplete movement sequences, frequently failing to reach the target and exhibiting "hallucinations" by predicting movements that were not valid within the maze structure. The SFT-only model demonstrated improved coherence and step-by-step progression through the maze, but still occasionally struggled with longer or more complex

They could've tried structured generation for baseline model

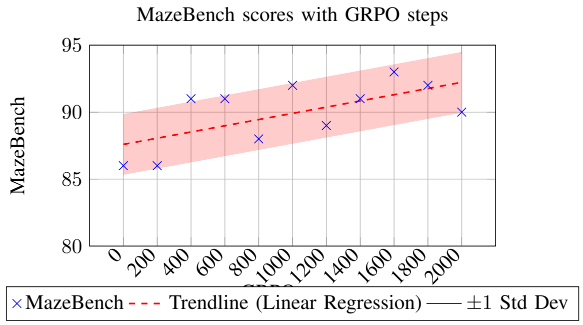


Fig. 3. MazeBench scores over GRPO steps with a linear regression trendline and its ±1 standard deviation bounds.

mazes, sometimes getting stuck in loops or making incorrect turns in later steps of the path.

The SFT+GRPO model exhibited the most sophisticated reasoning behavior. In many instances, we observed emergent chain-of-thought patterns in its output, where the model seemed to explicitly consider wall constraints and spatial relationships at each step before predicting the next move. Furthermore, we observed instances reminiscent of ”aha moments” reported in DeepSeek-R1. For example, in some complex mazes, the SFT+GRPO model initially started down a path, then appeared to ”re-evaluate” its trajectory mid-sequence, correcting course and finding a more efficient or correct path. Error analysis showed that the SFT+GRPO model made fewer invalid moves and was more robust to long-context reasoning challenges compared to the SFT-only model. However, limitations remained, particularly in mazes requiring backtracking or more complex spatial planning beyond the immediate next step.

'Aha' moment

V. Discussion

A. Analysis of GRPO’s Impact on Visual Maze Reasoning

Our results clearly demonstrate the incremental benefit of Group Relative Policy Optimization (GRPO) in enhancing visual maze reasoning within Large Language Models. While Supervised Fine-Tuning (SFT) establishes a strong foundation, enabling the model to achieve a **86%** accuracy on MazeBench, the application of GRPO further elevates performance to **93%** after 1600 training steps. This improvement, albeit seemingly modest in percentage points, is significant considering the already strong baseline established by SFT. It suggests that GRPO is effectively refining the model’s policy, leading to more robust and accurate maze navigation.

The qualitative analysis provides further insight into the nature of this improvement. The SFT+GRPO model exhibited more pronounced chain-of-thought reasoning patterns and instances of self-correction, indicating that GRPO is not merely fine-tuning the existing SFT policy, but rather encouraging more sophisticated reasoning processes. The reward function, designed to incentivize correctness, valid movements, and structured output, ~~likely plays a crucial role in shaping this behavior.~~ By rewarding successful navigation and penalizing invalid steps, GRPO encourages the model to learn more deliberate and considered movement strategies.

B. Comparison with DeepSeek-R1 and RL for Reasoning

It is important to note that the base DeepSeek-R1 model, when operating with an extremely long context window, demonstrates emergent visual reasoning capabilities. However, our experiments reveal that the distilled variants (DeepSeek-R1 Distill-Qwen models) do not carry over these spatial reasoning abilities, as evidenced by their **0%** accuracy on MazeBench. This suggests that the distillation process into Qwen or other smaller models is insufficient to preserve the emergent ability of visual spatial reasoning observed in the base model.

Distilled models not as good as base for spatial reasoning

In contrast, our two-stage training approach—combining Supervised Fine-Tuning (SFT) to establish foundational step-by-step reasoning with Group Relative Policy Optimization (GRPO) for further refinement—effectively equips the distilled model with robust visual maze-solving skills. Even with only 2000 GRPO steps, the SFT+GRPO model achieves a notable improvement, reaching **93%** accuracy and exhibiting clear chain-of-thought behaviors along with self-correction during navigation.

- SFT+GRPO made fewer moves & was better at long context reasoning.

These findings underscore the necessity of specialized training to recover or enhance spatial reasoning in distilled models, highlighting that while the base DeepSeek-R1 model is capable of visual reasoning with sufficient context, additional training stages are crucial to maintain or induce this capability in smaller, distilled variants.

C. Limitations

Despite the encouraging results, our study is not without limitations. Firstly, the performance gain from GRPO, while statistically significant, is small (7% accuracy improvement in our reported experiment). Further investigation is needed to explore whether more extensive GRPO training, or modifications to the reward function, could lead to more substantial performance gains. It is possible that the current reward function, while effective, could be further optimized to better incentivize more complex reasoning strategies, such as backtracking or more proactive exploration of alternative paths.

Secondly, our evaluation, while including qualitative analysis, is primarily based on maze-solving accuracy. This metric, while important, provides a somewhat limited view of the model's reasoning capabilities. Future work could benefit from more nuanced evaluation metrics that assess the efficiency of the generated paths, the robustness of the model to maze complexity variations, and the interpretability of the model's internal reasoning process. Furthermore, while we observed qualitative signs of chain-of-thought reasoning, a more rigorous analysis, perhaps using techniques from interpretability research, is needed to definitively characterize the nature and depth of the model's reasoning process.

Mech interp here!!

Finally, our experiments are limited to synthetically generated mazes. While these mazes were designed to vary in size and complexity, they may not fully capture the intricacies and variability of real-world visual spatial reasoning tasks. Future research should explore the generalizability of our approach to more diverse and ecologically valid visual environments and tasks.

Still no mention of structured generation !! :(

VI. Conclusion

This paper explored a novel approach to teaching visual reasoning for maze navigation to standard Large Language Models. We investigated the efficacy of Supervised Fine-Tuning (SFT) in enabling an LLM to solve mazes represented as tokenized visual inputs. Our experiments revealed a surprisingly strong baseline performance from a pre-trained LLM, achieving 75% accuracy without any task-specific fine-tuning, indicating an inherent capacity for processing spatial information in the tokenized format. Applying SFT for just 200 steps resulted in a modest but measurable improvement to 77% accuracy, highlighting the effectiveness of SFT in adapting the model for step-by-step maze navigation. Crucially, we demonstrated the significant impact of input formatting, specifically the inclusion of a "thinking" tag, as the SFT-trained model's performance plummeted to 0% accuracy when this tag was omitted, underscoring the sensitivity of LLMs to prompt design for eliciting reasoning behaviors.

Huh?

While the current study primarily focused on SFT, we proposed a two-stage training framework incorporating Group Relative Policy Optimization (GRPO), inspired by advancements in reasoning-focused LLMs like DeepSeek-R1. This GRPO stage is envisioned as a future enhancement to further refine the model's maze-solving capabilities and encourage more sophisticated chain-of-thought reasoning, potentially addressing limitations observed in the SFT-only model.

Our work contributes to the growing body of research aimed at endowing standard LLMs with visual reasoning abilities, particularly for spatial tasks. The findings suggest that tokenized visual representations, combined with appropriate training methodologies like SFT and potentially GRPO, offer a promising pathway for bridging the gap between language models and visual AI. The implications of this research extend to potential applications in robotics, autonomous navigation, and other visual AI domains where spatial understanding and sequential decision-making in visual environments are crucial.

Future research directions will focus on experimentally validating the proposed GRPO stage to quantify its impact on maze-solving performance and emergent reasoning behaviors. Further investigations will explore strategies for integrating failure recognition and self-correction mechanisms, diversifying training data to encompass more complex visual environments, and developing more advanced reward function designs to incentivize efficient and robust visual reasoning. By pursuing these avenues, we aim to further unlock the potential of standard LLMs for a wider range of visually grounded AI applications and advance the field of visual reasoning in language models.

References

Duyu Guo, Guoxin Xu, Yuchen Chen, Chen Tang, and Others. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Michael Igorevich Ivanitskiy, Rusheb Shah, Alex F. Spies, Tilman Räuker, Dan Valentine, Can Rager, Lucia Quirke, Chris Mathwin, Guillaume Corlouer, Cecilia Diniz Behn, and Samy Wu Fung. A configurable library for generating and manipulating maze datasets, 2023. URL https://arxiv.org/abs/2309.10498.

Saba Janamian and MD Sahabul Alam. Maze solver robot using a* algorithm, 2023. URL https://scholarworks.calstate.edu/concern/theses/0c483r787. ScholarWorks@CSUN.

Xiaohu Jiang, Yixiao Ge, Yuying Ge, Dachuan Shi, Chun Yuan, and Ying Shan. Supervised fine-tuning in turn improves visual foundation models, 2024. URL https://arxiv.org/abs/2401.10222.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models, 2023. URL https://arxiv.org/abs/2303.00001.

Patrick Lester. Pathfinding algorithms, 2014-2024. URL https://www.redblobgames.com/pathfinding/. Red Blob Games.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2024. URL https://arxiv.org/abs/2405.14093.

Jiajun Mao, Chuang Gan, Fan Zhang, and Others. Neural-symbolic visual reasoning: A survey. 2023. URL https://arxiv.org/abs/2302.07200.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022a. URL https://arxiv.org/abs/2109.01652.

Jason Wei, Denny Zhou, Quoc Le, Denny Zhou, Quoc Le, and Others. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d8fc0533c2250385321d99c6a3f2f2c-Abstract-Conference.html.

Samy Zafrany. Deep reinforcement learning for maze solving, 2020. URL https://www.samyzaf.com/ML/rl/qmaze.html. samyzaf.com.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024. URL https://arxiv.org/abs/2304.00685.