THE UNIVERSITY OF LIVERPOOL

COMP702-MSC FINAL PROJECT

Visual Spatial Reasoning for Large Language Models

Author	Jinlong Liu
ID	201678181
Supervisor	Frank Wolter

June 25, 2023

I. PROJECT DESCRIPTION

In recent years, Large Language Models (LLMs) have transformed the field of Natural Language Processing (NLP) with their impressive achievements. These models, which have been trained on extensive amounts of text data, have proven to be highly effective in solving various NLP tasks, like translating languages, answering questions, and generating text. Their ability to understand and create text that resembles human language has been truly remarkable and easy to appreciate.

However, despite their proficiency in many linguistic tasks, LLMs have faced challenges when it comes to Spatial Reasoning. Unlike tasks that focus solely on language, Spatial Reasoning involves comprehending and manipulating visual and spatial information. To shed light on this limitation, I plan to conduct an experiment using popular LLMs like GPT-4 [1]. The main goal will be to test how well they can handle Visual Spatial Reasoning, which is important for understanding the relationships between objects in visual content.

The experiment will center around a commonly used task called Visual Question Answering (VQA). This task requires machines to answer questions based on provided images. For example, when given an image showing different objects, the machine will be given inquiries such as "What is the spatial relationship between object A and object B?". By designing carefully constructed tests like these, we can assess the level of Spatial Reasoning ability demonstrated by LLMs in an easy-to-understand manner.

Through evaluating their performance in VQA and their capacity to comprehend and reason about spatial relationships in visual content, we can gain valuable insights into the Spatial Reasoning capabilities of LLMs. These findings will not only enhance our understanding of the strengths and limitations of these models but also pave the way for further advancements in NLP. This progress will bring us closer to developing more comprehensive and versatile language models that are accessible and easy to work with.

II. AIMS AND OBJECTIVES

In this section, we will provide a clear outline of the aims and objectives of this project. The aims represent the overarching goals that we aim to achieve, while the objectives outline the specific steps and milestones that will be undertaken to fulfill these goals. The aims and objectives of this project are outlined below:

A. Aims

1) To investigate the Spatial Reasoning abilities of Large Language Models (LLMs), specifically focusing on their performance in Visual Question Answering (VQA) tasks.
2) To understand the limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content.
3) To assess the current state of Spatial Reasoning capabilities in leading LLMs, then compare with elder models and explore their improvement in this domain. Besides, to identify potential avenues for enhancing LLMs' Spatial Reasoning abilities.

B. Objectives

1) Design and develop a comprehensive experimental framework for evaluating LLMs' Spatial Reasoning abilities in VQA tasks.
2) Curate a diverse dataset of visual stimuli and corresponding questions that require spatial reasoning skills to answer accurately.
3) Conduct systematic experiments using the prepared dataset and modified LLMs to assess their performance and measure their level of Spatial Reasoning competence.
4) Analyze the experimental results to identify patterns, trends, and challenges in LLMs' Spatial Reasoning capabilities.
5) Provide insights into the strengths and limitations of current LLMs for Spatial Reasoning tasks, based on the experimental findings.
6) Suggest potential avenues for enhancing LLMs' Spatial Reasoning abilities, such as incorporating multimodal information or novel architectural modifications.
7) Contribute to the existing body of knowledge in the field of NLP by advancing our understanding of LLMs' Spatial Reasoning abilities and their implications for future research and development.

In summary, this project aims to investigate the Spatial Reasoning abilities of LLMs, specifically in VQA tasks, by designing experiments, analyzing results, and providing insights for improvement. By achieving these objectives, the project will contribute to the advancement of NLP and enhance our understanding of LLMs' capabilities in comprehending and reasoning about spatial relationships in visual content.

III. KEY LITERATURE AND BACKGROUND READING

Spatial Reasoning is one of core part of commonsense reasoning since

For LLMs, Spatial Reasoning encompasses several key aspects, including (i) mereotopology, (ii) direction and orientation, (iii) size, (iv) distance, and (v) shape, as discussed by Cohn et al. (2008) in their qualitative analysis [2]. To assess LLMs' performance in Spatial Reasoning, Cohn et al. (2023) developed a comprehensive Spatial Reasoning test that incorporates basic spatial relations (such as parthood, rotation, and direction), size, shape, location, affordances, object interaction, and object permanence [3]. Their findings revealed inconsistencies in LLMs' responses to the same question during continuous sessions, indicating that LLMs struggle with maintaining consistency in Spatial Reasoning tasks.

In a related study by Lin et al. (2023), a test was devised to evaluate LLMs' Spatial Reasoning abilities with a focus on mathematical representation of questions [4]. Their research demonstrated that while LLMs may not excel at directly answering spatial questions, they exhibit proficiency in transforming questions into mathematical forms based on user demands.

Moreover, Borji (2023) proposed a failure task category for ChatGPT, within which Spatial Reasoning was highlighted

[5]. In this task, the researcher provided a grid description and asked the model to find a path from a start point to an end point. However, the model failed to answer the question accurately, highlighting its limitations in Spatial Reasoning tasks.

Collectively, these studies shed light on the challenges LLMs face when it comes to Spatial Reasoning, including inconsistencies in responses, struggles with direct spatial question answering, and difficulties in solving specific spatial tasks like path finding.

One important aspect to consider is Visual Spatial Reasoning (VSR) and its relevance in the context of Visual Language Models (VLMs), which is a distinct topic from LLMs. It is worth highlighting some related research in this field.

Researchers have dedicated efforts to assess the ability of VLMs in comprehending commonsense spatial reasoning. Visual Question Answering (VQA) tasks have been employed to evaluate the performance of VLMs in answering questions that involve spatial relationships between objects depicted in images [6]. Consequently, various benchmarks and question sets have been developed to specifically evaluate the spatial reasoning capabilities of VLMs.

For instance, Liu et al. (2022) constructed a dataset consisting of over 10,000 text-image pairs, encompassing 66 types of spatial relations [7]. They evaluated the performance of both human subjects and VLMs on this dataset, revealing that VLMs still lag behind humans in this task. Additionally, other benchmarks such as GQA [8], AGQA [9], Compositional Visual Relations Dataset [10], VALSE [11], Liu et al. (2021) [12], and SpartQA [13] have been developed to further gauge the spatial reasoning capabilities of VLMs. Figure 1 shows some examples of these benchmarks.

These benchmarks and datasets serve as valuable tools to assess the progress and limitations of VLMs in the realm of spatial reasoning and they have such things in common: (i) they are all based on commonsense spatial reasoning, (ii) they all focus on VLMs, (iii) they all only ask LLMs to use simple answers to these questions, like "yes" or "no", and (iv) they all given some bound words to the model to help it understand the spatial relations. However, in my research, I will focus on LLMs and their Spatial Reasoning capabilities, which is a distinct topic from VLMs.

IV. DEVELOPMENT AND IMPLEMENTATION SUMMARY

This project aims to investigate the Visual Spatial Reasoning capabilities of LLMs, specifically focusing on their performance in VQA tasks. So the development and implementation of this project will be divided into these parts:

A. Probelm Identification

The main problem of this project is to investigate the Spatial Reasoning abilities of LLMs, specifically focusing on their performance in VQA tasks. So the first step is to identify the problem and the research questions. The research questions are as follows:

1) What are the Spatial Reasoning abilities of LLMs?

2) What are the limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content?

3) What is the current state of Spatial Reasoning capabilities in leading LLMs?

4) What are the potential avenues for enhancing LLMs' Spatial Reasoning abilities?

B. Experiment Design

For the experimental stage, the workflow can be visualized as depicted in Figure 2. The experiment is divided into three distinct parts, each serving a specific purpose:

1) **Prepare Stage**

During this stage, the primary objective is to gather a diverse collection of images that contain various spatial relations. These images can be sourced from the internet or generated using appropriate tools. For this project, the dataset will consist of a combination of an existing benchmark proposed by Liu et al. (2022) [7]. Another crucial aspect is the question set, which will be divided into two parts: (i) Questions discussing the spatial relationships between objects depicted in the images. (ii) Questions requiring the reasoning steps behind the previous answers.

2) **Experiment Stage**

The focus of this stage is to test different versions of Large Language Models (LLMs) using the prepared dataset and question set. The LLMs will be evaluated based on two factors: (i) Different sizes of LLMs. (ii) Different pre-training methods of LLMs. The experiment will involve running the LLMs on the prepared dataset, recording the results, and analyzing their performance. And all testing models will be GPT models, including from GPT-3.5 tubor to GPT-4-0612.

3) **Analysis and Evaluation Stage**

In this stage, the emphasis is on analyzing the performance of the different versions of LLMs using the prepared dataset. The analysis will primarily consider two aspects: (i) The accuracy of the answers provided by the LLMs. (ii) The reasoning steps generated by the LLMs. Based on the accuracy, the effectiveness and capability of the LLMs in spatial reasoning can be evaluated and compared. By examining the reasoning steps, any limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content can be identified.

Through this systematic workflow, the experiment aims to assess the performance and capabilities of various versions of LLMs in spatial reasoning tasks. The analysis and evaluation stage provide insights into the strengths, weaknesses, and potential areas for improvement in LLMs' understanding and reasoning abilities pertaining to spatial relationships in visual content.

C. Code Implementation

There is no code implementation in this project. All the experiments will be conducted on the website of OpenAI. The

Caption: *The pizza is at the edge of the dining table.*

The caption is: ☑ True ☐ False

(a) a

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

(b) b

(c) c

A man <verb> the car. He is _____ the car.

A man washes the car. *beside* A man drives the car. *inside*

(d) d

Fig. 1. Examples of spatial reasoning benchmarks. (a)Example in benchmark proposed by Cohn et al. (2023) [3]. (b)Example of GQA [8]. (c)Example of AGQA [9]. (d)Example in benchmark proposed by Liu et al. (2021) [12].

code of this project will be written in Python, and the code will be used to prepare the dataset and question set, and to analyze the results of the experiment.

V. USER INTERFACE MOCKUP

For this project, there is no user interface mockup. All the experiments will be conducted on the website of OpenAI. The code of this project will be written in Python, and the code will be used to prepare the dataset and question set, and to analyze the results of the experiment.

VI. DATA SOURCES

The data utilized in this project will be divided into two main parts: the dataset and question set, and the answers generated by the Large Language Models (LLMs).

For the dataset and question set, a combination of existing benchmarks proposed by Liu et al. (2022) [7] and self-generated images will be used. The images in the dataset will encompass real-world objects along with 2D shapes, ensuring

a diverse range of spatial relationships to be examined. The benchmark images are extracted from the COCO 2017 dataset [14], which is an open-source dataset widely used in computer vision research. It is important to note that the usage of the COCO dataset has been granted full permission for this project.

As for the second part, the answers generated by the LLMs will be stored in a text file. The answers will be generated using the OpenAI website, which provides a platform for executing queries and obtaining responses from the language model. The generated answers can be conveniently downloaded as a text file from the website or saved directly on the platform.

By utilizing this comprehensive data setup, incorporating a combination of benchmark images and self-generated content, along with leveraging the power of LLMs to generate answers, the project aims to provide an extensive and diverse set of data for analysis and evaluation. The combination of the benchmark dataset and the generated answers will enable the project to
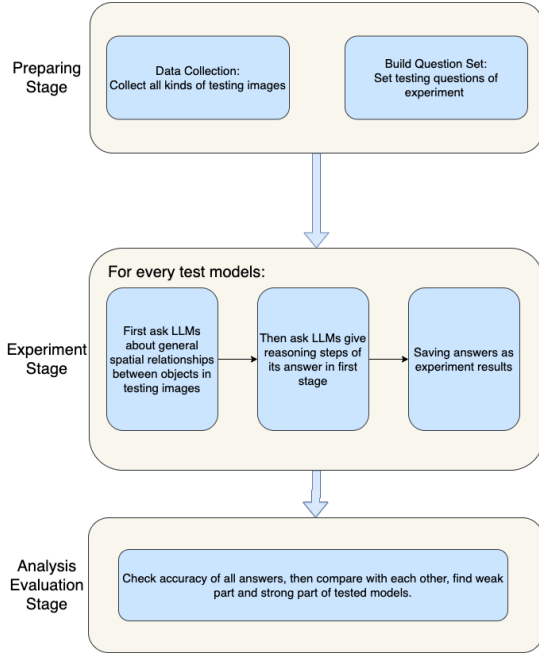
Fig. 2. Workflow of the experiment

assess the performance and capabilities of the LLMs in spatial reasoning tasks accurately.

It is important to ensure adherence to proper data usage policies, including obtaining necessary permissions for the benchmark dataset and complying with any licensing or attribution requirements associated with the dataset. This approach ensures the project's integrity and ethical practices while utilizing publicly available data resources for research purposes.

VII. Testing

VIII. Evaluation

IX. Ethical Considerations

As researchers and developers explore the capabilities of LLMs in visual spatial reasoning, it is crucial to consider and address several ethical considerations associated with this technology. These considerations are necessary to ensure responsible and ethical use of LLMs in the context of visual spatial reasoning. Some key ethical considerations include:

1) Bias and Fairness: LLMs are trained on vast amounts of data, and if the training data contains biases, the models can inadvertently perpetuate and amplify those biases. Care should be taken to identify and mitigate biases related to spatial relationships, objects, or cultural aspects in the training data. Fairness and inclusivity should be prioritized to avoid marginalization or discrimination based on race, gender, socioeconomic status, or any other protected characteristics.

2) Data Privacy and Security: When collecting or using datasets for visual spatial reasoning tasks, it is essential to respect privacy rights and comply with data protection regulations. Any personally identifiable information (PII) should be handled with care, and

proper anonymization techniques should be employed to protect the privacy of individuals represented in the data. Additionally, robust security measures should be implemented to prevent unauthorized access to sensitive data.

3) Informed Consent: If human subjects are involved in data collection or evaluation processes, informed consent should be obtained. Participants should be fully informed about the purpose of the study, the data collection methods, and any potential risks or implications associated with their participation. Participants should have the freedom to withdraw their consent at any point during the study.

4) Transparency and Explainability: LLMs often operate as black boxes, making it challenging to understand how they arrive at their answers or reasoning steps. It is important to develop methods and techniques that enhance the transparency and explainability of LLMs' decision-making processes in visual spatial reasoning. This helps build trust and allows users to better understand the models' behavior and potential limitations.

5) Human-AI Collaboration: While LLMs can provide valuable insights and assist in visual spatial reasoning, it is crucial to acknowledge the importance of human expertise and maintain a human-centric approach. LLMs should be viewed as tools to augment human capabilities rather than replacing human decision-making or responsibility. Collaboration between humans and AI should be encouraged to ensure that decisions made based on LLM-generated outputs are critically evaluated and aligned with ethical standards.

6) Accountability and Governance: As LLMs become more sophisticated, it is important to establish accountability mechanisms and governance frameworks to address any potential risks or unintended consequences. This may include adopting standards, guidelines, or regulatory frameworks specific to the development and deployment of LLMs in visual spatial reasoning applications. Responsible and ethical practices should be promoted throughout the entire lifecycle of LLM development and deployment.

These ethical considerations serve as a starting point for ensuring the responsible and ethical use of LLMs in visual spatial reasoning. It is crucial for researchers, developers, and stakeholders to actively engage in ongoing discussions, collaborate with interdisciplinary teams, and adhere to ethical guidelines to mitigate potential harms and maximize the benefits of this technology in a responsible manner.

X. Project Plan

The project timeline and task allocation are visually represented in Figure 3, showcasing the specific Gantt chart. The project is organized into four distinct stages, each comprising multiple tasks aimed at achieving the project objectives. The stages and their corresponding tasks are outlined below:

1) Literature Review

This stage focuses on conducting an extensive literature review to gain a comprehensive understanding of the project domain. It is divided into three tasks:

- Literature review: Conduct a thorough review of relevant research, technologies, and best practices in the field.
- Problem identification: Identify the key challenges and opportunities that the project aims to address.
- Experiment design: Plan and design the experimental framework and methodologies for subsequent stages.

2) Build Dataset

In this stage, the focus is on creating the dataset and question set required for the experiments. It involves two tasks:

- Prepare dataset: Collect or curate a dataset comprising a combination of existing benchmark images and self-generated images.
- Prepare question set: Design and develop a set of questions that encompass various spatial relationships and reasoning scenarios.

3) Experiment Stage

The experiment stage involves executing the designed experiments using the created dataset and question set. This stage is divided into two tasks:

- Run experiment: Implement the experiments using the selected Large Language Models (LLMs) and the prepared dataset and question set.
- Record results: Collect and document the results obtained from the experiments, ensuring comprehensive record-keeping.

4) Analysis and Evaluation Stage

In this stage, the focus is on analyzing and evaluating the results obtained from the experiments. This stage comprises two tasks:

- Analyze results: Perform a detailed analysis of the experimental outcomes, considering the accuracy of answers and the reasoning steps generated by the LLMs.
- Evaluate results: Assess the performance and effectiveness of the LLMs in spatial reasoning, identify limitations, and recognize challenges faced by the models.

5) Final Report

The final stage involves documenting the project outcomes in a report and preparing a presentation. It is divided into two tasks:

- Write report: Compile all the findings, methodologies, and conclusions into a comprehensive report that highlights the project's objectives, methodology, and results.
- Write presentation: Create a visually engaging presentation summarizing the project's key aspects, including objectives, methodology, results, and conclusions.

By following this well-structured project plan, the project aims to ensure efficient task allocation, timely execution, and a systematic approach towards achieving the project objectives. The defined stages and tasks provide a clear roadmap for project completion and successful delivery of the intended outcomes.

XI. RISKS AND CONTINGENCY PLANS

REFERENCES

[1] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," *arXiv preprint arXiv:2304.03277*, 2023.

[2] A. G. Cohn and J. Renz, "Qualitative spatial representation and reasoning," *Foundations of Artificial Intelligence*, vol. 3, pp. 551–596, 2008.

[3] A. G. Cohn and J. Hernandez-Orallo, "Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms," *arXiv preprint arXiv:2304.11164*, 2023.

[4] F. Lin, Z. Shou, and C. Chen, "Using language models for knowledge acquisition in natural language reasoning problems," *arXiv preprint arXiv:2304.01771*, 2023.

[5] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[7] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *arXiv preprint arXiv:2205.00363*, 2022.

[8] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[9] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 287–11 297.

[10] A. Zerroug, M. Vaishnav, J. Colin, S. Musslick, and T. Serre, "A benchmark for compositional visual reasoning," *arXiv preprint arXiv:2206.05379*, 2022.

[11] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena," *arXiv preprint arXiv:2112.07566*, 2021.

[12] X. Liu, D. Yin, Y. Feng, and D. Zhao, "Things not written in text: Exploring spatial commonsense from visual signals," *arXiv preprint arXiv:2203.08075*, 2022.

[13] R. Mirzaee, H. R. Faghihi, Q. Ning, and P. Kordjmashidi, "Spartqa:: A textual question answering benchmark for spatial reasoning," *arXiv preprint arXiv:2104.05832*, 2021.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
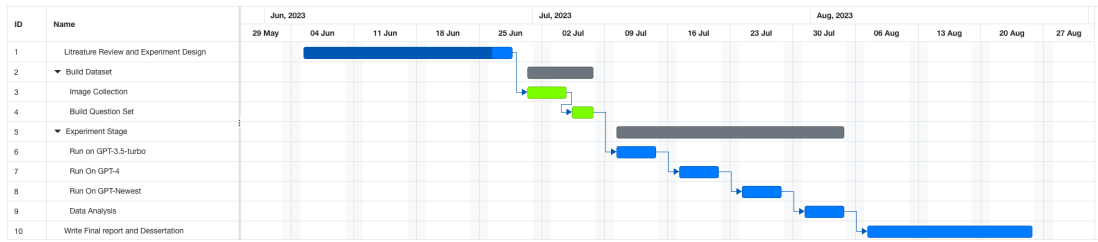
ID	Name	Jun, 2023					Jul, 2023					Aug, 2023				
		29 May	04 Jun	11 Jun	18 Jun	25 Jun	02 Jul	09 Jul	16 Jul	23 Jul	30 Jul	06 Aug	13 Aug	20 Aug	27 Aug
1	Litreature Review and Experiment Design														
2	▼ Build Dataset														
3	Image Collection														
4	Build Question Set														
5	▼ Experiment Stage														
6	Run on GPT-3.5-turbo														
7	Run On GPT-4														
8	Run On GPT-Newest														
9	Data Analysis														
10	Write Final report and Dessertation														

Fig. 3. Gantt Chart