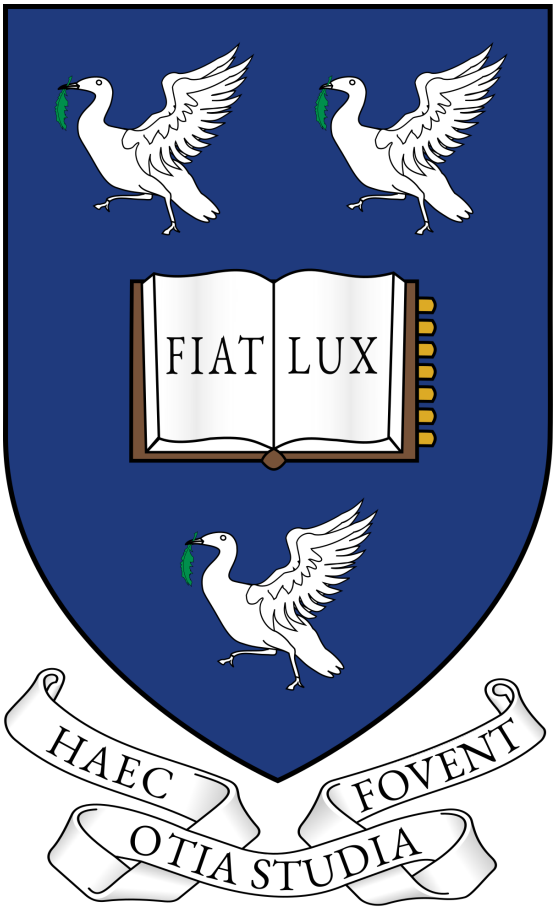


THE UNIVERSITY OF LIVERPOOL

COMP702-MSC FINAL PROJECT

# Visual Spatial Reasoning for Large Language Models

*Author* Jinlong Liu
*ID* 201678181
*Supervisor* Frank Wolter

**June 21, 2023**

## I. Project Description

In recent years, Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) with their remarkable achievements. These models, trained on vast amounts of text data, have proved highly effective in solving numerous NLP tasks, including machine translation, question answering, and text generation. Their ability to comprehend and generate human-like text has been nothing short of impressive.

However, despite their proficiency in various linguistic tasks, LLMs have encountered challenges when it comes to Spatial Reasoning. Unlike tasks centered purely on language, Spatial Reasoning requires understanding and manipulating visual and spatial information. To shed light on this limitation, I aim to conduct an experiment utilizing popular LLMs like GPT-4. The focus will be on testing their capability in Visual Spatial Reasoning.

The experiment will revolve around a widely used task known as Visual Question Answering (VQA), which demands machines to answer questions based on provided images. For instance, given an image depicting different objects, the machine will be tasked with responding to inquiries such as "What is the spatial relationship between object A and object B?". Through these carefully crafted tests, we can gauge the extent of Spatial Reasoning proficiency exhibited by LLMs.

By assessing their performance in VQA and their ability to comprehend and reason about spatial relationships in visual content, we can obtain valuable insights into the Spatial Reasoning capabilities of LLMs. These findings will not only deepen our understanding of the strengths and limitations of these models but also pave the way for further advancements in the realm of NLP, bringing us closer to more comprehensive and versatile language models

## II. Aims and Objectives

### 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.

## III. Key Literature and Background Reading

For Large Language Models (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 [1]. 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 [2]. 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 [3]. 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 [4]. 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.

Another relevant aspect to consider is Visual Spatial Reasoning (VSR), which primarily pertains to Visual Language Models (VLMs), a distinct topic from LLMs. However, it is worth mentioning some related works in this field. Researchers have focused on assessing VLMs' ability to comprehend commonsense spatial reasoning. Consequently, numerous bench-

marks and question sets have been developed to 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 [5]. 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 [6], AGQA [7], Compositional Visual Relations Dataset [8], VALSE [9], Liu et al. (2021) [10], and SpartQA [11] have been developed to further gauge the spatial reasoning capabilities of VLMs.

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, and (iii) they all utilize visual stimuli.

## IV. Development and Implementation Summary

## V. User Interface Mockup

## VI. Data Sources

## VII. Testing

## VIII. Evaluation

## IX. Ethical Considerations

## X. Project Plan

## XI. Risks and Contingency Plans

## References


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

[2] 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.

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

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.