VISUAL SPATIAL REASONING OF LARGE LANGUAGE MODELS

Presentation: Jinlong Liu

LIST OF CONTENT

01.

PROJECT INTRODUCTION

A brief description of this project

02.

AIMS AND OBJECTIVES

Research aims and objectives

03.

FEASIBILITY ESTIMATE

Talk about feasibility according to the martials

04.

EXPERIMENT DESIGN

Introduction the key aspects of experiment design

05.

PRPJECT PLAN

Show detailed project plan

06.

REFERENCES

List of References

01.
PROJECT INTRODUCTION

In recent years, Large Language Models (LLMs) have transformed the field of Natural Language Processing (NLP) with their impressive achievements. Such as ChatGPT, Bert and Bard. As they updated in a raptly speed, there are a lot of test work also starting.

Source of all images:
https://chat.openai.com
https://bard.google.com/
https://ivet360.com/google-bert-for-vets/

- The testing works are focused on the professional domain, like the ability of logic, answer test questions in various expert domain.
- This project is based on the work of testing ability of spatial reasoning, which is a subject of testing ability of commonsense reasoning.
- But it is different with existed work with adding visual input into prompt not just text input. Which used to be works of multi-models problem.

02.
AIMS AND OBJECTIVES

AIMS:

1. Investigate performance of Visual Question Answering (VQA) tasks.
2. Make comparison between different versions of LLMs.
3. Find challenges of LLM’s on Visual Spatial Reasoning.

MAIN OBJECTIVES

1. Design a diverse dataset with corresponding questions.
2. Conduct systematic experiments by using prepared dataset.
3. Analyse the experimental results to identify the challenges of LLMs
4. Provide the insight of strong points and weakness of current LLMs.

03.
FEASIBILITY ESTIMATE

Former works

Name	Year	Source	Feature
VSR	2022	[1]	66 distinct types of spatial relations
GQA	2019	[2]	22M diverse reasoning questions
AGQA	2021	[3]	Add videos paired with question
CVR	2022	[4]	Measures of sample efficiency(Train dataset)
VALSE	2021	[5]	Test the linguistic phenomena in visual modality
Spatial commonsence benchmark	2022	[6]	Focus on positional relationship between people and objects
SpartQA	2021	[7]	Focus on generate spatial description by limitted grammar rules

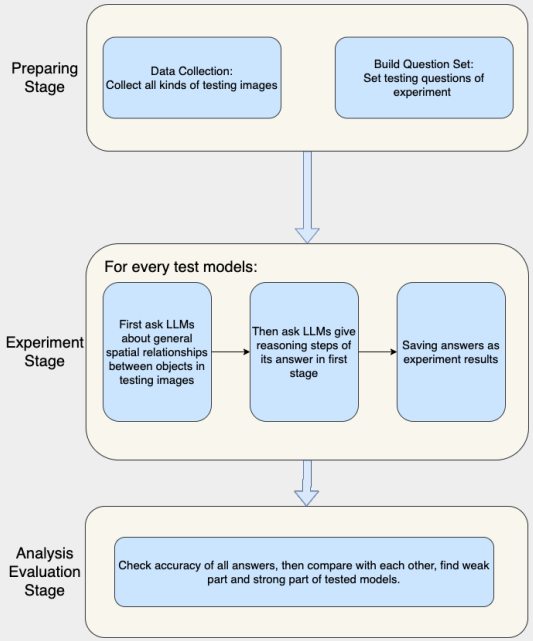
SAME POINTS:

- The new version of GPT models now can consider as multi-models, and it can support get in visual content and text prompt at same time. It can work with this kind of tasks now.

DIFFERENT POINTS:

- This VAQ is mostly like testing dataset for the object models, LLM's are usually tested by prompt.
- All the benchmarks give specific words to describe spatial reasoning or give choice or label to select. This project is to let LLMs generate by its self.

04.
EXPERIMENT DESIGN



The detailed experiment shows as figure 1, it has three stage:

1. Preparing Stage(Design Dataset): Combine existed dataset and design questions with it.
2. Experiment Stage: test every selected model with image paired question one by one.
3. Analysis Evaluation Stage: Analysis generated answer with accuracy of answers and make judgement about reasoning steps

The question prompt shows as below, it may contain 3-4 questions include objects relevant positions and explain reasoning steps.

Input Image	Question
	1) Can you describe the spatial relationship between the tea table and the sofa? 2) Can you describe the relevant spatial relationships between the objects on the tea table? 3) Show the Reasoning Steps.

Advantages:

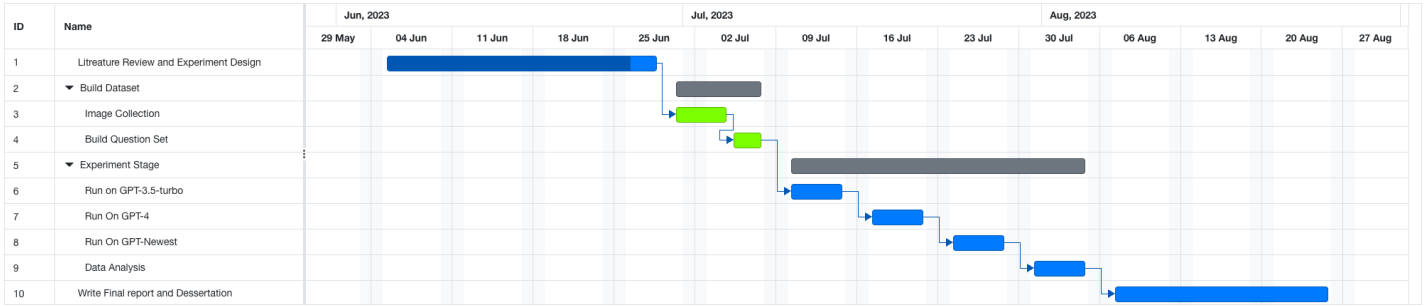
1. Questions can be designed more specific follow former works.
2. The test dataset can be much more comprehensive. Which can contain much more situations.
3. Can find detailed differences between different models.

Disadvantages:

1. The experiment may cost much more time.
2. Analysis work is low efficient in comparing with former work.

05.
PRPJECT PLAN

GANTT CHART


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

06.
REFERENCES

PAPERS MENTIONED IN THIS SLIDES

[1]. F. Liu, G. Emerson, and N. Collier, “Visual spatial reasoning,” arXiv preprint arXiv:2205.00363, 2022.
[2]. 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.
[3]. 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.
[4]. 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.
[5]. A. Zerroug, M. Vaishnav, J. Colin, S. Musslick, and T. Serre, “A benchmark for compositional visual reasoning,” arXiv preprint arXiv:2206.05379, 2022.
[6]. 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.
[7]. 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.

THANK YOU