VISUAL SPATIAL REASONING OF LARGE LANGUAGE MODELS

Presentation: Jinlong Liu

LIST OF CONTENT

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 subset of testing ability of commonsense reasoning.
- But it is different with current research that adding visual input into prompt not just text input. Which can be considered as multi-model 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

Previous 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

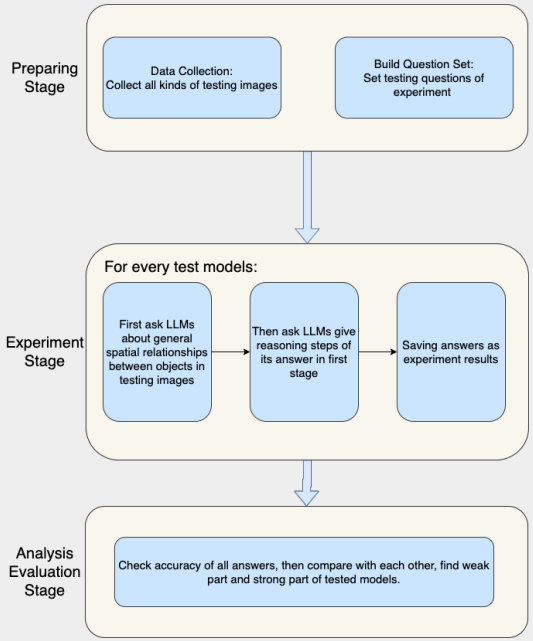
SAME POINTS:

- 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 VLMs, 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

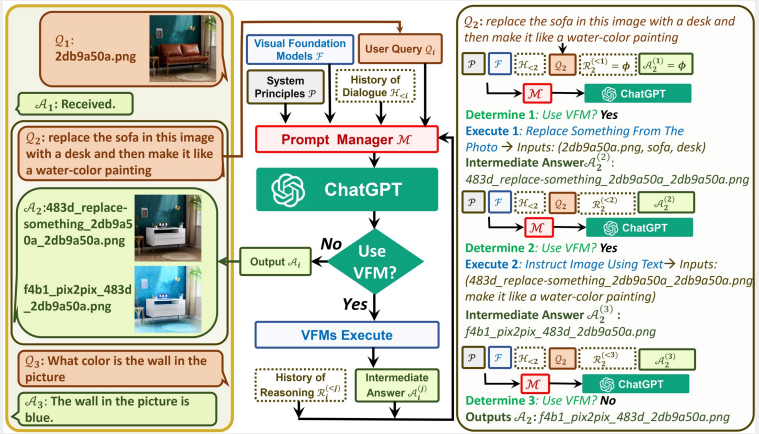


The detailed experiment shows as figure 1, it has three stage:

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 generating answers with accuracy of answers and make judgements about reasoning steps

Selected Model: TaskMatrix(AKA Visual ChatGPT) connects GPT-3 and a series of Visual Foundation Models to enable sending and receiving images during chatting.[8]

Here is its architecture



Source: https://github.com/microsoft/TaskMatrix/blob/main/assets/figure.jpg

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:

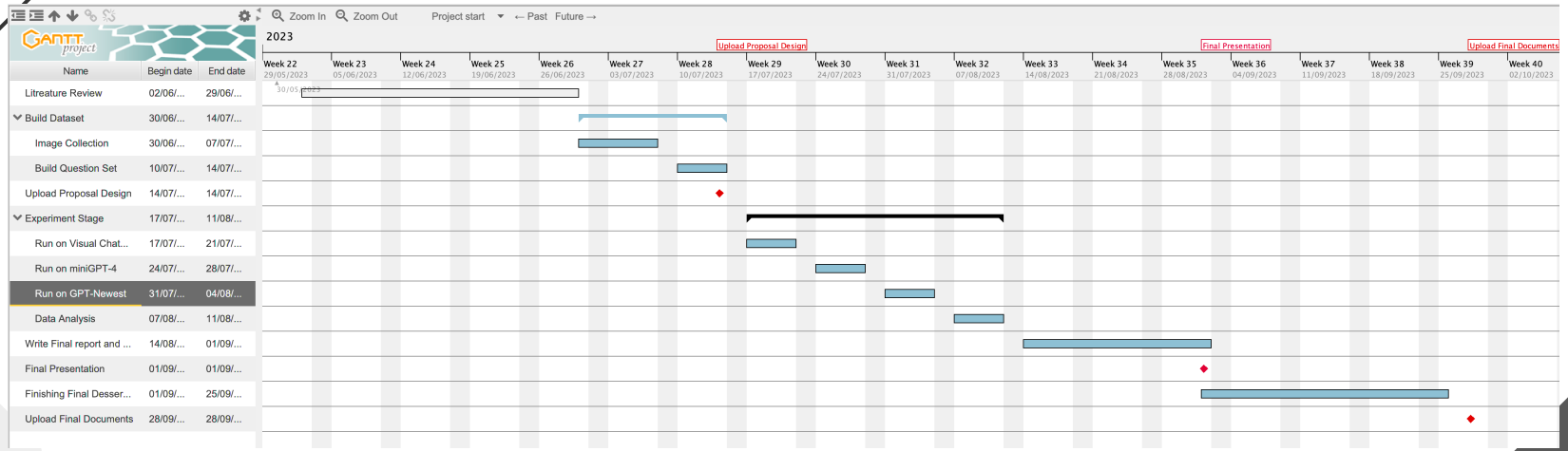
1. Questions can be designed more specific following the image.
2. The test dataset can be much more comprehensive. Which can contain much more situations.
3. Can find more detailed distinctions between different models.

Disadvantages:

1. The experiment may cost much more time.
2. Analysis work is low efficient in comparing with former work.
3. Cost more time in responding.
4. Can’t change to specific version of GPT-models.

05.
PRPJECT PLAN

GANTT CHART


Zoom In
Zoom Out
Project start
← Past
Future →
2023
Upload Proposal Design
Final Presentation
Upload Final Documents
Week 22 29/05/2023
Week 23 05/06/2023
Week 24 12/06/2023
Week 25 19/06/2023
Week 26 26/06/2023
Week 27 03/07/2023
Week 28 10/07/2023
Week 29 17/07/2023
Week 30 24/07/2023
Week 31 31/07/2023
Week 32 07/08/2023
Week 33 14/08/2023
Week 34 21/08/2023
Week 35 28/08/2023
Week 36 04/09/2023
Week 37 11/09/2023
Week 38 18/09/2023
Week 39 25/09/2023
Week 40 02/10/2023
30/05/2023
Name
Begin date
End date
Litreature Review 02/06/... 29/06/...
Build Dataset 30/06/... 14/07/...
Image Collection 30/06/... 07/07/...
Build Question Set 10/07/... 14/07/...
Upload Proposal Design 14/07/... 14/07/...
Experiment Stage 17/07/... 11/08/...
Run on Visual Chat... 17/07/... 21/07/...
Run on miniGPT-4 24/07/... 28/07/...
Run on GPT-Newest 31/07/... 04/08/...
Data Analysis 07/08/... 11/08/...
Write Final report and ... 14/08/... 01/09/...
Final Presentation 01/09/... 01/09/...
Finishing Final Desser... 01/09/... 25/09/...
Upload Final Documents 28/09/... 28/09/...

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.
[8]. C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, “Visual chatgpt: Talking, drawing and editing with visual foundation models,” 2023.

THANK YOU