COMP9444 Project Summary

Project Title: Custom Chatbots with LLMs

Yan Pan (z5484351)

Keyin Lin (z5440167)

Sifang Gao (z5471175)

Yiqing Yu (z5453538)

Jinyuan Fan (z5461665)

I. Introduction

This project aims to explore the potential of multimodal AI by transforming visual content into textual descriptions and evaluating their effectiveness in question answering tasks. We combine image processing models such as BLIP and VIT-GPT2 with LLM (such as GPT-4omini) to integrate image context and language issues, improving the alignment between visual and language processing. The contribution of this study is to test these models using the ScienceQA dataset and evaluate the performance of multimodal large language models in complex tasks.

II. Related Work

In recent years, multimodal tasks have gradually become a research hotspot, with many works dedicated to integrating visual and linguistic abilities to improve model performance. For example, BLIP proposed a pre training framework based on a combination of self supervised learning and supervised learning, which has made significant progress in cross modal tasks such as image description and visual question answering. In addition, ViT-GPT2 performs well in image-based text generation tasks by integrating visual feature extraction (ViT) with language generation.

However, existing work also has some limitations. On the one hand, most models generate explanatory texts that lack logic and organization when dealing with complex multimodal contexts. On the other hand, these models often lack adaptability to new tasks and require fine-tuning through large-scale annotated data. In addition, there is still significant room for improvement in the accuracy and diversity of the answers generated by the model. This study attempts to address the aforementioned issues by combining the Chain of Thoght (CoT) inference mechanism with more advanced multimodal pre training models.

III. Methods

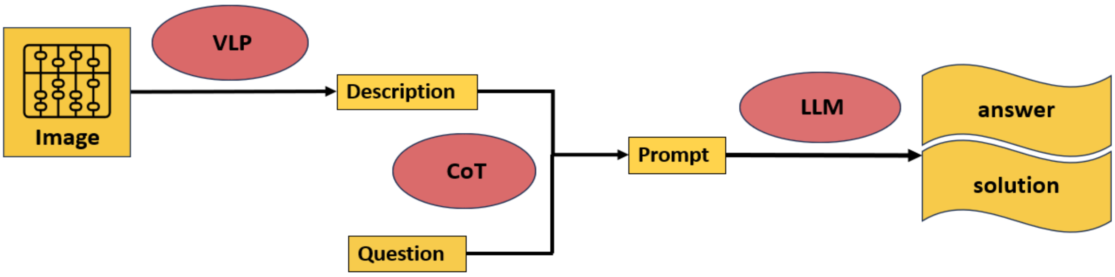


Figure 1. Overall process

We used VLP + CoT + LLM to solve multimodal problems.

VLP generates image descriptions, CoT combines the descriptions and questions to create prompts, and LLM provides the final response.

LLM : Large Language Model

VLP: Vision-Language Pretraining

CoT: Chain of thought

1. vitgpt2 + gpt4o-mini

We chose ViTGPT2 as the VLP and GPT4O-mini as the LLM. This is the method used in the paper.

ViTGPT2 is a multimodal model that combines visual features with language generation capabilities, making it particularly suitable for tasks requiring textual descriptions or generation based on images. Its core innovation lies in leveraging ViT to extract visual information and GPT-2 to generate natural language, thus bridging the gap between vision and language.

2. BLIP + gpt4o-mini

We chose BLIP as the VLP and GPT4O-mini as the LLM. This is our innovative approach.

BLIP (Bootstrapped Language-Image Pretraining) is a cutting-edge multimodal framework focused on joint pretraining of vision and language. Through its innovative pretraining strategy, it significantly enhances the performance of cross-modal tasks, such as image captioning, visual question answering, and image-text retrieval.

The core innovation of BLIP lies in combining self-supervised and supervised learning to progressively guide the model in learning more accurate vision-language correspondences. By employing a bootstrapped learning strategy, it enhances the model's ability to generate high-quality image descriptions and understand multimodal inputs.

3. gpt4o-mini

Both our VLP and LLM are based on GPT4O-mini. This is our innovative approach.

GPT4O-mini is a versatile model capable of handling images in vision-language pretraining (VLP) tasks and excelling as a large language model (LLM) for text generation and comprehension. By integrating vision and language capabilities, it demonstrates outstanding performance in both multimodal and language tasks.

IV. Experimental Setup

According to the relevant paper information of the SCIENCEQA dataset, SCIENCEQA is a dataset containing 21208 multimodal multiple-choice questions, covering three major categories: natural sciences, social sciences, and language sciences. [Dataset source: ScienceQA]. This dataset has the following main features:

Multimodal context: 10220 questions contain textual context, and 10332 questions contain image context (including approximately 2960 natural images and approximately 7372 chart images). In addition, 83.9% of the questions are accompanied by lectures, and 90.5% of the questions are accompanied by explanations, which reveal the understanding path and reasoning steps.

Diversity of content: The dataset covers 26 topics, 127 categories, and 379 skills, fully reflecting the diversity of scientific question and answer tasks. The average length of the question is 12.11 words, the average length of the options is 4.40 words, the average length of the explanation is 125.06 words, and the average length of the explanation is 47.66 words.

In this project, we selected 4241 questions from the dataset as the test set and covered all 26 topic classes.

We have configured the following hyperparameters in the GPT-4 o model:

Engine: gpt-4o-mini

temperature： 0.0

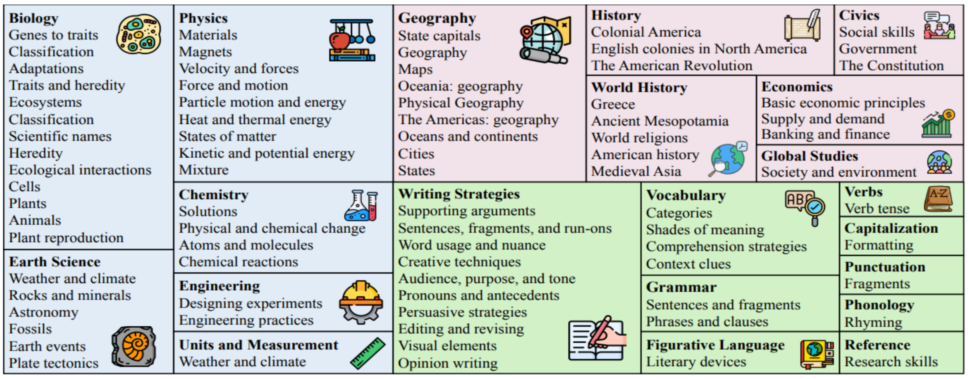
max_tokens： 250

top_p： 1.0

frequency_penalty： 0.0

presence_penalty： 0.0

Evaluation strategy: We use the model accuracy of each comparison group as the evaluation metric. In addition, to evaluate the quality of the generated explanatory text, we used three indicators: BLEU, ROUGE, and semantic similarity to determine the relevance, correctness, and completeness of the explanatory text.



Biology
Genes to traits
Classification
Adaptations
Traits and heredity
Ecosystems
Classification
Scientific names
Heredity
Ecological interactions
Cells
Plants
Animals
Plant reproduction

Earth Science
Weather and climate
Rocks and minerals
Astronomy
Fossils
Earth events
Plate tectonics

Physics
Materials
Magnets
Velocity and forces
Force and motion
Particle motion and energy
Heat and thermal energy
States of matter
Kinetic and potential energy
Mixture

Chemistry
Solutions
Physical and chemical change
Atoms and molecules
Chemical reactions

Engineering
Designing experiments
Engineering practices

Units and Measurement
Weather and climate

Geography
State capitals
Geography
Maps
Oceania: geography
Physical Geography
The Americas: geography
Oceans and continents
Cities
States

History
Colonial America
English colonies in North America
The American Revolution

Civics
Social skills
Government
The Constitution

World History
Greece
Ancient Mesopotamia
World religions
American history
Medieval Asia

Economics
Basic economic principles
Supply and demand
Banking and finance

Global Studies
Society and environment

Writing Strategies
Supporting arguments
Sentences, fragments, and run-ons
Word usage and nuance
Creative techniques
Audience, purpose, and tone
Pronouns and antecedents
Persuasive strategies
Editing and revising
Visual elements
Opinion writing

Vocabulary
Categories
Shades of meaning
Comprehension strategies
Context clues

Grammar
Sentences and fragments
Phrases and clauses

Figurative Language
Literary devices

Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

Phonology
Rhyming

Reference
Research skills

Figure 2. Dataset Preview

V. Results

1.image description

vitgpt2: "5": "A large white kite is in the air.",

BLIP: "5": "a large white light",

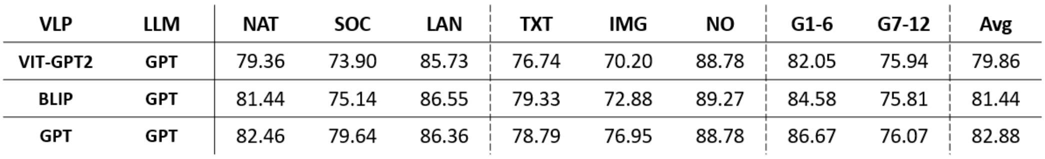
gpt4o-mini: "5": "The image depicts a large, dome-shaped structure suspended in an expansive, industrial-like space. This structure resembles a parachute or aerodynamic capsule, characterized by its smooth, rounded top and a series of ripple-like folds along its surface. The exterior is primarily white, with an orange band encircling its middle, which adds a striking visual contrast. \n\nThe setting around the structure is a vast, dimly lit hangar or testing facility. The walls are lined with large, vertical panels, and the ceiling features a grid of lights that illuminate the space. The capsule is supported by a framework that likely includes cables or stands, indicating that it is being held in place for testing or demonstration purposes.\n\nOverall, the image conveys a sense of advanced technology and engineering, possibly related to aerospace or space exploration. The clean, minimalist design combined with the industrial surroundings emphasizes the innovative nature of the project being showcased."

2.question response

vitgpt2+gpt: "The answer is B for answer index 1, because: The text addresses the goddess directly, which is an example of apostrophe."

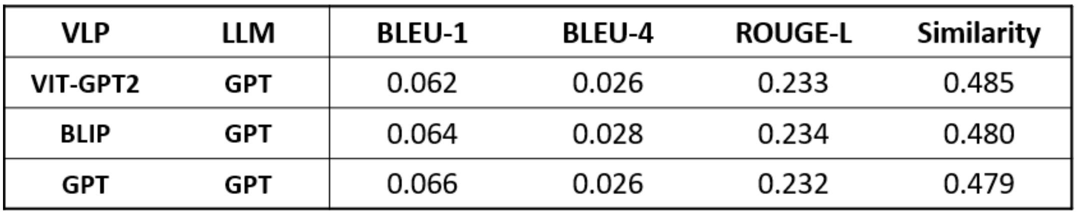
blip+gpt: "The answer is B for answer index 1, because: Gordon's test could show how steady a parachute with a 1 m vent was at 200 km per hour, which is what he was specifically observing in the wind tunnel.",

gpt4o-mini: "The answer is B for answer index 1, because: The phrase 'Sing, O goddess' is an example of apostrophe, as it directly addresses a personification (the goddess) that is not physically present."



VLP	LLM	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
VIT-GPT2	GPT	79.36	73.90	85.73	76.74	70.20	88.78	82.05	75.94	79.86
BLIP	GPT	81.44	75.14	86.55	79.33	72.88	89.27	84.58	75.81	81.44
GPT	GPT	82.46	79.64	86.36	78.79	76.95	88.78	86.67	76.07	82.88

Table 1. Accuracy rate of different models



VLP	LLM	BLEU-1	BLEU-4	ROUGE-L	Similarity
VIT-GPT2	GPT	0.062	0.026	0.233	0.485
BLIP	GPT	0.064	0.028	0.234	0.480
GPT	GPT	0.066	0.026	0.232	0.479

Table 2. Performance comparison of different models

VI. Conclusions

In summary, our project assessed LLM capabilities in multimodal tasks by combining different VLP models with Cot reasoning. Results showed that, this approach improved accuracy and clarity in handling both visual and textual inputs, marking an advancement in multimodal question-answering performance.

Advantages

Regarding strengths, two key features stand out. First, we achieved high accuracy by using an enhanced VLP model, which improves the accuracy of the generated answers. Second, Chain of Thought reasoning enabled us to generate clear, well-structured responses, guiding the model to produce logical and easy-to-follow answers step-by-step.

Drawbacks

The first is low similarity, which is mainly due to the lack of fine-tuning in our model; this limits our ability to fully tailor responses to specific tasks. The second drawback is a lack of diversity in responses. As we currently rely on only one type of large language model, there is not much change between answers.

Future Work

1. We could incorporate few-shot learning（Combine few similar examples into prompt, enabling our model to better understand and adapt to new tasks.

2. Model fine-tuning will def get better model’s responses, but it isdifficult for us to achieve this right now.

3. Multimodal LLM can integrate text and image inputs. This will adapt the answer to a wider range of scenarios.