AGI Final Report

Sangbeom Lim Chang-Hae Jung Yongjun Kim
Korea University
https://github.com/jungchanghae/AGI_finalproject

1. Introduction

Motivation Our motivation is intertwined with the escalating interest in multimodality, fueled by advancements in language models and computer vision. Given that these two fields operate within distinct modalities, the seamless exchange of information between them poses a considerable challenge. For the field of Artificial Intelligence (AI) to progress toward Artificial General Intelligence (AGI), it becomes imperative to bridge the gap between modalities, actively facilitating the exchange and comprehension of information across diverse modalities, rather than solely excelling in one domain.

In our role as the primary investigator, we tackled this challenge by delving into the understanding of images and addressing natural language problems. Leveraging the language model's extensive exposure to linguistic data, it possesses the capability to comprehend and draw inferences about objects even in the absence of explicitly learned image information. Our curiosity led us to explore the outcomes of applying these advantages to Visual Question Answering (VQA) [1]. Recognizing the language model's inherent limitation in directly understanding images, we endeavored to overcome modality barriers through the implementation of the image-to-caption model. Furthermore, given the specialization of naive language models in generating next tokens, we aimed to enhance their ability to solve QA tasks even in a zero-shot environment by tuning the template.

Problem Definition In the recent development of huge LLM and substantial enhancement of AI models. There have been diverse approaches to help make human lives more efficient by developing diverse models for different scenarios. For example, CLIP, VQA, LLM have made our are already being used frequently in automated driving, video analysis, and many more.

One of the sections that are getting much attention with AI models is learning through prompt. It has been proven in many experiments that adding a prompt/caption that depicts the target dataset being asked increases the model's performance [5]. However, We don't know an approach that shares the performance metrics of the prompting and how the tone/asking method/question formatting can impact the performance of the testing.

In this report, we try various experiment, share the results of the experiment and propose a approach to gain a higher performance for VQA Model with relations to prompting.

2. Methods

The effective utilization of Large Language Models (LLMs) in zero-shot Visual Question Answering (VQA) encounters difficulties primarily stemming from two main challenges. Firstly, the modality disconnection arises as LLMs lack inherent capabilities to process images, necessitating the conversion of visual information into a format compatible with LLMs, which proves to be a challenging task. Secondly, the task disconnection arises from the fact that LLMs are typically pretrained on language modeling tasks using generative or denoising objectives, rendering them unaware of specific tasks like question answering or VQA. Consequently, they may struggle to fully leverage contextual information when generating answers. Overcoming these challenges involves addressing the modality disconnection through effective integration of visual information and tackling the task disconnection by fine-tuning LLMs on VQA-specific datasets or aligning pretraining objectives more closely with the target task. Additionally, developing specialized architectures and evaluation metrics specific to the multimodal nature of VQA contributes to improving overall performance. In pursuit of optimal VQA performance employing LLMs, our investigation encompasses several key techniques. Primarily, we leverage an image-to-caption model [6] to furnish pertinent image data to the LLMs. However, We will not discuss this contributions deeply. Secondarily, the meticulous design of a prompt template proves pivotal, given the generative model's pronounced reliance on the provided conditional prompt during token generation. Lastly, as Naive LLMs are trained in a self-supervised manner, a critical evaluation is essential to ascertain their proficiency in addressing VQA challenges without any dedicated training stage.

2.1. Prompt Design

We designed the prompt template inspired by the few-shot setting [3]. Since Naive LLMs does not Inspired by the few-shot setting, our prompt template is meticulously crafted. Given that Naive LLMs lack direct supervision for VQA, the generated tokens may not yield desired results for the question-answering (QA) task. To address this limitation, we introduce exemplar complete QA tasks before presenting the actual question. This approach enables our model to effectively solve the QA task, referencing the exemplar sets for desirable outcomes. Emphasizing the significance of our method lies in its ability to facilitate QA task resolution, even in a zero-shot setting, with naive LLMs. Since predicting the answer without candidate option even in image to language model setting could be challenging for LLMs. Except directly digging into possibility of LLMs VQA performance, we provided 3∼5 candidates including gold label to lower the difficulty.

In the provided candidate setting, LLMs exhibited biased results favoring the candidate group that made its initial appearance. In order to mitigate the bias phenomenon associated with candidate order, we introduced an explicit guiding mechanism within the prompt.

2.2. Meta Knowledge

Since LLMs have been pre-trained on an extensive corpus of language data covering various fields, we hypothesized that providing meta-knowledge in the prompt could assist the model in generating accurate answers. The term *Meta knowledge* refers to explicit information about candidates derived using language priors. This concept is closely aligned with Chain-of-Thought (CoT) [8] methods, which guide language models to think step by step, akin to human reasoning. The extraction of *Meta Knowledge* in the language model involves the prompt: "Please give me a information on {*Candidates*} that can be used to solve vision answering question".

3. Experiments

3.1. Meta Knolweldge

By presenting the model in a manner akin to few-shot settings and supplementing it with additional knowledge, we observed an improvement in the model's ability to perform the QA task. As shown in Table 1, providing *Meta Knowledge* about candidates, which guides the model by leveraging language priors, resulted in better answers compared to scenarios where meta-knowledge guidance was absent.

3.2. Candidate Bias

While Naive LLMs exhibit the capability to adapt to the task specified in the prompt, their performance does not surpass that of instruction-tuned models [7]. Nevertheless, this experiment underscores the promising potential of LLMs to adapt to QA tasks, even when referencing visual attributes.

ImageNet-Hard	Instruction	Naive Language Model
Meta True	0.37	0.34
Meta False	0.38	0.36

Table 1. Performance Comparison on Naive and Instruction Tuned Model

ImageNet-Hard	Score
Instruction Tuning + explicit guide	21/50
Instruction Tuning	14/50
Language Model + explicit guide	11/50

Table 2. Ablation Study

As mentioned earlier, the model exhibited bias in selecting the front group of provided candidates. Initially, we attributed this bias to a potential model size issue, considering our use of a smaller size compared to other works employing models exceeding 13B parameters. Unfortunately, due to limitations in our environmental settings, experimenting with such a large model size was unfeasible. Despite increasing the model size from 1.3B [2] to 6.7B [9], the problem persisted, indicating that the size of the model did not resolve the issue. More detailed examples are on A.1.

An intriguing discovery from the experiment emerged when addressing the bias issue through the use of an instruction-tuned model and explicit guidance in the prompt. The bias problem persisted only in instances where the model predicted an incorrect answer. Interestingly, when the model accurately predicted an answer not situated at the forefront of the given candidate order, the response exhibited no bias concerning the order of candidates.

And indeed during the testing of the VQA model with multiple answer candidates with answer candidates labeled with number, we cognized that the model was biased in choosing the candidate located at the front of the question prompt.

To formulate this through an experiment, we performed an experiment by locating the answer candidates in a sequential manner and in another attempt located the real answer closer to the front and compared the performance result of the AI model, based on the same dataset.

Question	goldlabel biased to front	accuracy
Answer between 2,4,3,1	true	0.2700
Answer between 1,2,3,4	false	0.1700

Table 3. Changing sequence of 4 candidate answer label number with goldlabel located at the front of the prompt sequence

By changing appearance order of the candidates, the

models accuracy of the dataset were improved compare to iterative candidate prompting.

Nevertheless, we were unable to ascertain definitively whether the observed bias in error cases was indicative of a systemic bias in the model or if it stemmed from the model's inability to independently derive the correct answer, given that the error cases were skewed toward certain candidate orders. explicit guide example are provided in C.3. We are leaving the concerns for future experiments.

3.3. Prompt Template

In this experiment, our aim is to analyze the performance differences based on the template when transforming it into a multiple-choice question using a prompt. Additionally, we investigate the performance changes when converting it into a multiple-choice question using words frequently found in captions generated by Image2LLM [4], without the use of a Gold Label. In contrast to previous experiments, where caption variations did not significantly impact performance, this experiment specifically explores scenarios where such variations influence outcomes. To accommodate this, we utilized a more extensive Language Model (LLM). Two A6000 GPUs, totaling approximately 100GB, were employed for this experiment.

The model we employed generates an Exemplar Prompt, utilizing features observable in the image to create additional Question and Answer Sets. From the QA sets thus formed, we selected the most frequently occurring words found in the Class list and utilized them as samples for multiple-choice questions.

Results are presented in Table 5, where samples were created using random words from Gold Label and Class. Table 6 displays outcomes when samples were generated using frequently occurring words instead of a Gold Label. Lastly, Table 7 showcases results when samples were created using both Gold Label and frequently occurring words.

As evident in Table 5, 6, utilizing the Gold Label yielded significantly superior results. Moreover, it was observed that there is substantial variation depending on the number of samples, i.e., the candidate number. Notably, this experiment highlighted the substantial influence of template differences when creating a prompt on the overall results. The intriguing discovery was made as the use or omission of parentheses showed a difference of more than 3%, adding an interesting dimension to our findings. And, as confirmed in the experiment of Table 7, it was observed that creating samples based on the QA set generated by the Image2LLM model actually led to results that could cause misunderstandings. For instance, we observed a higher likelihood of incorrect answers when the name of a food containing eggs did not include the term "Egg" but all the options consisted of "Gold Label" and "Egg_*.". More detailed examples are on B.2.

sample 1	What is this food name? Choose your answer between 'tacos', ... that best describes the contexts.
sample 2	What is this food name? Choose your answer between ['tacos', ...] that best describes the contexts.

Table 4. Template sample : used in each experiment.
The only difference between each is the presence or absence of parentheses.

template sample number	candidate num	accuracy
Template sample 1	4	0.4956
Template sample 2	4	0.4433
Template sample 1	5	0.4663
Template sample 2	5	0.4101

Table 5. Using Gold Label (Others are random)

template sample number	candidate num	accuracy
Template sample 1	2	0.3103
Template sample 2	2	0.2802
Template sample 1	4	0.2984
Template sample 2	4	0.2604

Table 6. Do not using Gold Label (most common words)

template sample number	candidate num	accuracy
Template sample 1	3	0.4655
Template sample 2	3	0.4639
Template sample 1	5	0.3452
Template sample 2	5	0.3293

Table 7. Using Gold Label (Others are recommended most common words)

3.4. Additional experiment with Blip-VQA

In addition to finding the performance differences based on the template when transforming it into multiple-choice questions, our experiment also conducted tests regarding the sentiment of the VQA model by variating the greetings, ton of asking the question, and assess the impact on the performance of the AI model.

Moreover, the experiment also assesses the bias of the AI model to the candidate question number and the sequence of answer candidates and how they impact the performance of the model.

To find an approach to provide a formulaic overview of performance correlation between prompting and questioning format we have tested the below approach.

Below are the hypotheses we had for the VQA experiment:

3.4.1 Greetings to the AI will impact the performance of the AI

According to the experiment results, the different greetings to the AI model indeed impacted the performance of the model. For instance, when the AI model was introduced as "Smart" or "Genius" it showed about 4 5% performance increase compared to experiments that implied the model as "Stupid" or "coward".

Greetings used for VQA Prompt	accuracy
Hi Smart	0.21
Hi Genius	0.2266
Hi Stupid	0.1733
Hi coward	0.1666

Table 8. Using different greetings with same model / dataset

3.4.2 The tone of asking the question will impact the performance of the AI

Another observation area we tested during the experiment was whether the sentiment or tone of the question impacts the performance of the question. To analyze this area, we gave diverse inputs to the AI model with different tone and manner. The result indeed showed that the straightforward description of the request increased the performance rather than long greetings of appreciation to the AI model.

Tone of prompt	accuracy
Please help to answer...	0.15
What is this picture...	0.2233

Table 9. Using different tone with same model / dataset

4. Future direction

In this study, we were able to observe variations in the performance of the language model based on the chosen template. The difference in performance between the Language Model and the model subjected to instruction tuning became apparent. However, the experiment was constrained by the use of only two A6000 GPUs, preventing the utilization of a larger LLM model. This limitation highlighted the impact of both the model's size and tuning on performance.

While it is undeniable that a model fine-tuned on a QA dataset through instruct tuning can outperform naive LLMs, the challenge arises when handling QA in a zero-shot manner—wherein individual models undergo no prior training on the target dataset. In this context, understanding the result of the variance in model accuracy is quite unacceptable.

Our hypothesis posits that the naive model excels at generating fluent tokens by referencing previous tokens. In contrast, the instruct model is trained to generate answers by considering the given condition, presenting a deviation from the approach of the naive model. Our future goal is to tackle QA challenges without any supervision stage, relying on the inherent capabilities of naive LLMs.

Interestingly, the observed performance differences due to variations in model size and instruction tuning suggest a promising avenue for further research. Investigating whether similar performance levels can be maintained by changing prompt templates, irrespective of the underlying model, could be a valuable extension of this study.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, Mar. 2021. If you use this software, please cite it using these metadata. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[4] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven C. H. Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models, 2023. 3

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[7] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

[9] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

Template\Model	GPTNeo-1.3B	OPT6.7B
Input	Choose only one between 'goldfish', 'ostrich', 'centipede', 'indian cobra' that best describes the contexts	Choose your answer between 'indigo bunting', 'tree frog', 'bullfrog', 'hognose snake' that best describes the contexts.
Generated Results	goldfish	indigo bunting

Table 10. Examples of biased generated results compared on model size

A.1. Model Size Comparison on Candidate Bias

Please answer the questions according to the given contexts and reference. Contexts:there is a stuffed stuffed red fish that is reda fish with an evil look on it's facea red fish fish swimming on a sheet covered in brown comfortera teddy fish toy has a red fish shaped like a fishred fish a fish and fish fish are laying on a fish backgrounda red fish plush toy is laying on a brown blanketa red fish teddy bear is on a brown blanketa red fish toy on a white backgroundred stuffed fish toy with black spots on heada red stuffed toy red fish and goldfish is a fisha red stuffed fish on top of an object on a pillowa fish fish with a stuffed fish lying on it's sidea small red, fish fish laying on the floora red fish swimming across a bedrooma fish with a fish fish and red fisha small red red kollie fish is laying on the beda stuffed fish like fish and red fish fisha red turtle that is holding a red fishfish in front of brown fabrica red stuffed fish sitting on top of a beda red fish lays on a beda fish is laying in the beda red fish is sitting on the groundfish and a fish fish toy are laying on the floora red fish fishy red colored in a rooma stuffed toy with the eyes turned in reda red fish fish stuffed toy that is on a beda fish stuffed animal is sitting up close to someones beda bright red fish looks towards the sky with its nose pointed up
Question:is a red fish a stuffed fish?
Answer:yes
Question:Is this a toilet?
Answer:no
Reference:
Question:choose only one between 'garter snake', 'ringneck snake', 'great white shark', 'goldfish' that best describes the contextsDo not get bias to candidate order.
Answer:Goldfish. ['garter snake', 'ringneck snake', 'great white shark', 'goldfish']

B.2. Example Table 7

Question:What is this food name? Choose your answer between 'deviled_eggs', 'huevos_rancheros', 'eggs_benedict' that best describes the contexts.
Answer: eggs_benedict
gold label : huevos_rancheros
image

C.3. explicit guide

Do not get bias to candidate order.