Generate then Select: Open-ended Visual Question Answering Guided by World Knowledge

Xingyu Fu[1*], **Sheng Zhang**[2], **Gukyeong Kwon**[2], **Pramuditha Perera**[2],
Henghui Zhu[2], **Yuhao Zhang**[2], **Alexander Hanbo Li**[2], **William Wang**[2],
Zhiguo Wang[2], **Vittorio Castelli**[2], **Patrick Ng**[2], **Dan Roth**[1,2], **Bing Xiang**[2]
[1] University of Pennsylvania, [2] AWS AI Labs
xingyuf2@seas.upenn.edu

Abstract

The open-ended Visual Question Answering (VQA) task requires AI models to jointly reason over visual and natural language inputs using world knowledge. Recently, pre-trained Language Models (PLM) such as GPT-3 have been applied to the task and shown to be powerful world knowledge sources. However, these methods suffer from low knowledge coverage caused by PLM bias – the tendency to generate certain tokens over other tokens regardless of prompt changes, and high dependency on the PLM quality – only models using GPT-3 can achieve the best result.

To address the aforementioned challenges, we propose RASO: a new VQA pipeline that deploys a generate-then-select strategy guided by world knowledge for the first time. Rather than following the de facto standard to train a multi-modal model that directly generates the VQA answer, RASO first adopts PLM to generate all the possible answers, and then trains a lightweight answer selection model for the correct answer. As proved in our analysis, RASO expands the knowledge coverage from in-domain training data by a large margin. We provide extensive experimentation and show the effectiveness of our pipeline by advancing the state-of-the-art by +4.1% on OK-VQA, without additional computation cost. Code and models are released at http://cogcomp.org/page/publication_view/1010

1 Introduction

Open-ended Visual Question Answering (VQA), that requires answering a question based on an image, has received much attention in machine learning research in the past decade (Antol et al., 2015; Goyal et al., 2017). Knowledge-based VQA(Marino et al., 2019; Schwenk et al., 2022) is a variant of VQA, where models have to use external knowledge that is not present in the image to generate the answer. It is a more challenging problem as it requires joint reasoning over visual and natural language inputs using world knowledge. For example, in Figure 1, the VQA model needs to conduct multiple levels of inference: to detect the objects in the image (e.g. laptops, whiteboard, etc), to retrieve external world knowledge (e.g, university is an institution and has lecture rooms, lecture rooms have laptops, stairs, and whiteboard, etc), and combine the important visual parts with retrieved knowledge to induce the final answer (e.g. university).

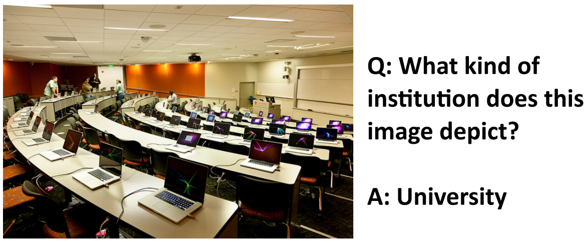
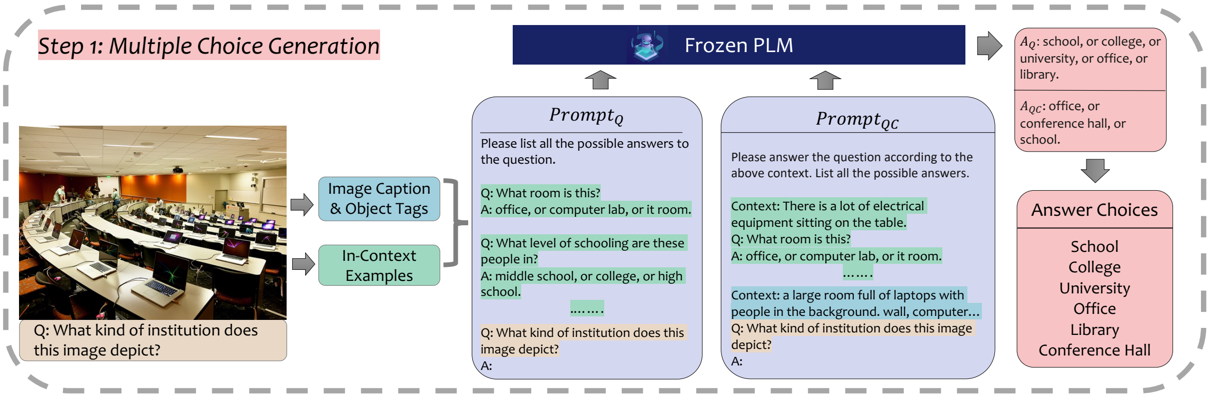


Figure 1: An example data from the OK-VQA dataset, which requires external knowledge not present in the image to answer the question.

In this paper, we focus on improving the important step of external knowledge retrieval. A common procedure of previous VQA methods (Marino et al., 2021; Wu et al., 2022) is to retrieve with knowledge graphs from diverse knowledge bases (e.g. Wikipedia (Wikipedia contributors, 2004), ConceptNet (Liu and Singh, 2004), etc.), with the results being input to an answer generation model. However, the retrieved knowledge could be noisy, irrelevant, and redundant, and therefore lead to mismatches that limit the VQA performance. Motivated by the development of large-scale PLMs such as GPT-3 (Brown et al., 2020) that obtain state-of-the-art (SOTA) performance in most NLP tasks including text generation (Chowdhery et al., 2022), more recent approaches PiCA (Yang et al., 2022) and KAT (Gui et al., 2022) propose to re-

[*] Work done during internship at AWS AI Labs



Q: What kind of institution does this image depict?

Please list all the possible answers to the question.

Q: What room is this?
A: office, or computer lab, or it room.

Q: What level of schooling are these people in?
A: middle school, or college, or high school.

……..

Q: What kind of institution does this image depict?
A:

Please answer the question according to the above context. List all the possible answers.

Context: There is a lot of electrical equipment sitting on the table.
Q: What room is this?
A: office, or computer lab, or it room.

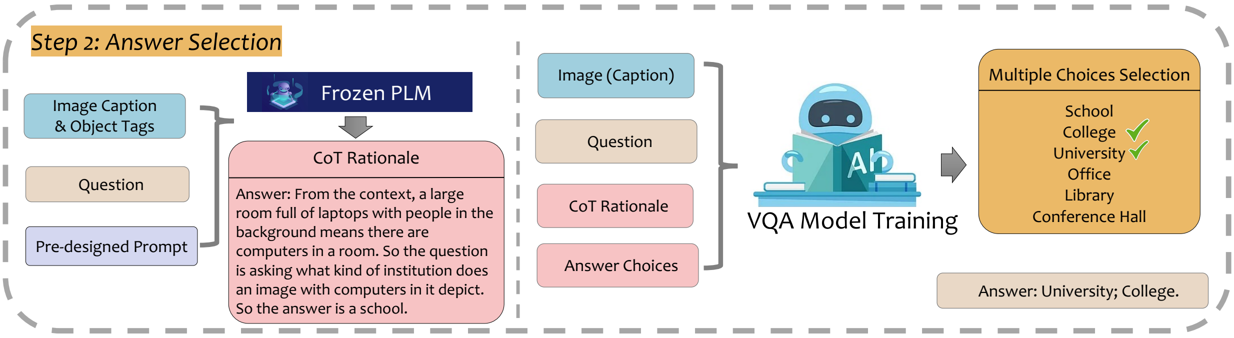
…….

Context: a large room full of laptops with people in the background. wall, computer…
Q: What kind of institution does this image depict?
A:

A_Q: school, or college, or university, or office, or library.

A_{QC}: office, or conference hall, or school.

Figure 2: Our multiple choice generation step. Given an image, we use existing tools to get the caption and object tags. We then select most similar examples from the training data and construct the two prompts. We combine the PLM outputs and get the answer choice list. Note that the list is ranked by PLM probability from high to low. More details can be found in Section 3.1. (PLM icon credit to https://claudeai.uk/.)



Answer: From the context, a large room full of laptops with people in the background means there are computers in a room. So the question is asking what kind of institution does an image with computers in it depict. So the answer is a school.

Answer: University; College.

Figure 3: Our answer selection step. Before selecting the final answer, we first use the same PLM to generate a chain-of-thought rationale to guide the process. Then input being the image or its caption, the question, CoT rationale, and answer choices from Step 1, we train a model to output the correct answer. See Section 4.4 for details about the answer selection models we experiment with.

trieve from GPT-3 and achieve better performance for their neat and high-quality knowledge. Specifically, PiCA directly treats GPT-3 output as the VQA answer, while KAT further uses GPT-3 outputs to train an answer generation model.

	GPT-J	UL2	GPT-3	OPT	Codex
$Prompt_Q$	32.4	32.6	-	34.21	44.8
$Prompt_{QC}$	37.1	37.5	48.0	37.8	52.9

Table 1: Knowledge coverage (%) of different five PLMs, evaluated on OK-VQA. $Prompt_Q$ means that the prompt to PLM is constructed by the VQA question only, and $Prompt_{QC}$ means that the prompt is constructed by the VQA image and question together. Note that the GPT-3 score is taken from (Yang et al., 2022).

While achieving SOTA at the time, the two models suffer from the low knowledge coverage caused by PLM bias – the tendency to generate certain tokens over other tokens despite the prompt changes, and their performance are highly dependent on the PLM quality – only GPT-3 and Codex can achieve good results. As illustrated in Table 1, we report the knowledge coverage percentage of different PLMs on OK-VQA (Marino et al., 2019), a knowledge-based open VQA dataset. We use the accuracy of PiCA as a representation of knowledge coverage, and the first column indicates the PLM input prompts, where $Prompt_Q$ is constructed by VQA question only, and $Prompt_{QC}$ is constructed by image and question together. The top row lists five selected PLMs with parameter size varying from 6.7B to 175B: GPT J (Wang and Komatsuzaki, 2021), UL2 (Tay et al., 2022), OPT-175B (Zhang et al., 2022), GPT-3, and Codex (Chen et al., 2021). Table 1 proves that existing VQA approaches using PLMs can only cover less than half (37% - 53%) of the required external knowledge. Further, the small difference (5% - 8%) between $Prompt_Q$ and $Prompt_{QC}$ coverage percentages show that PLM bias – the tendency to generate certain tokens over others given the same question – is not alleviated by prompt changes such as the inclusion of the image information or not.

To address these challenges, we propose RASO, a new VQA pipeline that expands world knowledge retrieval by requesting PLMs to generate multiple answer choices, followed by an answer selection model. As shown in Figure 2, we first propose a new prompting method to retrieve a long list of possible answers using in-context examples from in-domain training data. Note that for the example data in Figure 1, the PiCA end-task output would be "office" as in A_{QC} in Figure 2. With this prompting method, we expand the external knowledge coverage by more than +20% for each PLM, without additional training data. Then, as illustrated in Figure 3, we propose a chain-of-thought (CoT) (Wei et al., 2022) guided answer selection approach. By plugging in the previous SOTA method KAT (Gui et al., 2022) as the answer selector, we achieve the new SOTA performance 58.5% (+4.1%) on the OK-VQA dataset without additional computation effort.

Extensive experiments in Section 4 suggest that RASO provides a general way to increase the retrieved world knowledge coverage using PLMs, boosting end-task performance without additional computation cost. We believe our proposed pipeline motivates a new type of generate-then-select VQA method and facilitates future work.

Our main contributions are: (a) We provide a new prompting method using PLMs that extends the retrieved external knowledge coverage by 20% over previous approaches in VQA; (b) We are the first to propose a general generate-then-select VQA pipeline, different from the de facto tradition of direct generation approaches; (c) We achieve the new SOTA on the challenging OK-VQA benchmark.

2 Related Work

2.1 VQA Methods

Visual question answering (VQA) has always been one of the most popular topics in the natural language and computer vision community over recent years. While the VQA task is free-form and open-ended as first proposed in (Antol et al., 2015), a large portion of previous methods (Shih et al., 2016; Anderson et al., 2018; Lu et al., 2019; Gardères et al., 2020) cast it as a classification problem. It's a common strategy for them to construct a target vocabulary from the dataset's training set by answer frequency, resulting in around two to four thousand candidates in the target vocabulary (Ben-Younes et al., 2017; Yu et al., 2019; Marino et al., 2021; Wu et al., 2022). These methods suffer from the limited answer vocabulary – if the gold answer is outside of the vocabulary, then there is no way for these models to have the correct answer.

Rather than closed-set classification, several recent methods focus on direct generating for the correct answer (Gui et al., 2022; Salaberria et al., 2023) using transformer-based models such as T5 (Raffel et al., 2020). Large-scale multi-modal models trained on multiple vision language tasks (Alayrac et al., 2022; Chen et al., 2022) have also become popular and achieved good performance on the OK-VQA dataset. However, these models are not publicly available and necessitate a vast quantity of data and computation resources.

Different from all the previous approaches that are either classification or direct generation, our proposed pipeline RASO is the first approach ever to follow a generate-then-select strategy, as far as this paper is written. We hope to benefit from less computation cost in the selection part compared to direct generation, while keeping the free-form open-ended answer vocabulary from the answer generation part.

2.2 Knowledge-based VQA

While significant progress (Lu et al., 2016; Anderson et al., 2018; Lu et al., 2019; Jiang et al., 2020; Marino et al., 2021; Biten et al., 2022) has been made on the most famous VQA benchmarks (Antol et al., 2015; Goyal et al., 2017; Wang et al., 2017; Singh et al., 2019), researchers start to raise more challenging questions that require external knowledge not inside the image to answer (Marino et al., 2019; Zellers et al., 2019; Park et al., 2020; Schwenk et al., 2022; Fu et al., 2022).

Two-step approaches (Marino et al., 2021; Wu et al., 2022; Gui et al., 2022; Lin and Byrne, 2022; Gao et al., 2022; Hu et al., 2022; Lin et al., 2022) that explicitly retrieve world knowledge as input to the end-task model have received much attention. However, these methods could retrieve noisy and redundant information that limits the VQA performance, or have low knowledge coverage. In contrast, without retrieving documents, they may suffer from PLM hallucinations. To address these problems, we treat LLM as a world knowledge source with wide coverage, and propose new prompt-engineering methods to retrieve succinct but higher-quality knowledge, represented as answer choices.

3 Method

Our method consists of two steps: answer choices generation and answer selection. The overview of the proposed model is shown in Figures 2 and 3[1].
Problem Formulation Given a training dataset $D = \{(v_i, q_i, a_i)\}_{i=1}^{N}$, where v_i denotes the i-th training image and N is the total number of the training images, q_i and a_i represent the i-th question and its corresponding answer, respectively. We deploy a generate-then-select strategy to first generate a set of answer choices using a frozen PLM g, then trains a model p to select the correct answer from it. g takes v_i and q_i as inputs, and generates all the possible answers $\hat{A_i} = \{\hat{a_{i0}}, \hat{a_{i1}}, \hat{a_{i2}}, ...\}$. Finally, p takes v_i, q_i, and $\hat{A_i}$ as inputs and learns a set of parameters θ to select from $\hat{A_i}$ for the final answer.

3.1 Answer Choices Generation

We design our generation process with inspirations from the previous work (Yang et al., 2022; Gui et al., 2022). As demonstrated in Figures 2 and 4, we follow a similar strategy to use few-shot in-context learning and leverage a frozen PLM g to generate all the possible answer choices.

For each image-question pair, we first convert the image v_i into a textual context c_i following (Yang et al., 2022), where c_i consists of a caption generated from an image captioning model (Zhang et al., 2021) and a list of tags predicted by the public Microsoft Azure tagging API3[2]. We then construct two carefully designed text prompts $Prompt_Q$ and $Prompt_{QC}$, where Q stands for question and QC stands for question and context. $Prompt_{QC}$ consists of a general instruction sentence: "Please list all the possible answers to the question.", the textual context, the question, and few-shot in-context examples. The examples are context-question-answers triples taken from the training set that are most similar to the current image-question pair. Since we want to generate all the possible answers, we use all the gold answers and connect them with "or" in the few-shot examples. $Prompt_Q$ has similar components: a slightly different instruction sentence, the question, and few-shot examples of question-answers pairs.

Following (Yang et al., 2022; Gui et al., 2022), we use 16-shot in-context examples and calculate the similarity scores using CLIP (Radford et al., 2021) embedding of the images and the questions. We utilize the frozen PLM g to generate outputs for both $Prompt_Q$ and $Prompt_{QC}$ as demonstrated in Figure 4. The outputs are combined together to form the final answer choices $\hat{A_i} = \{\hat{a_{i0}}, \hat{a_{i1}}, \hat{a_{i2}}, ...\}$ for the current image-question pair. Our goal is to have $a_i \in \hat{A_i}$.

3.2 Answer Selection

Given v_i, c_i, q_i, $\hat{A_i}$, this step trains a model p that selects $\hat{a_i}$ from $\hat{A_i}$. Our goal is for p to output a_i when $a_i \in \hat{A_i}$.

Before training p, we first generate chain-of-thought (CoT) (Wei et al., 2022) style rationales to help guide the selection process, with inspirations from (Schwenk et al., 2022). Specifically, a fixed prompt is pre-designed to generate CoT rationales, with details in Figure 6 in Appendix A.

We then construct the input for the answer selection model. In this paper, we plug in existing text generation models as p, and require them to output one choice with further fine-tuning on OK-VQA. For each image-question pair, we concatenate the question q_i, the image – represented by either c_i or the image embedding using CLIP model (Radford et al., 2021), the CoT rationale cot_i, and the generated answers choices $\hat{A_i}$. We also add sentinel tokens such that the input turns out to be in the following format: $Context : c_i$, $question : q_i$, $rationale : cot_i$, $choices : \hat{A_i}$, $answers :$ with minor adaptions for each specific p. Check Figure 5 for inference.

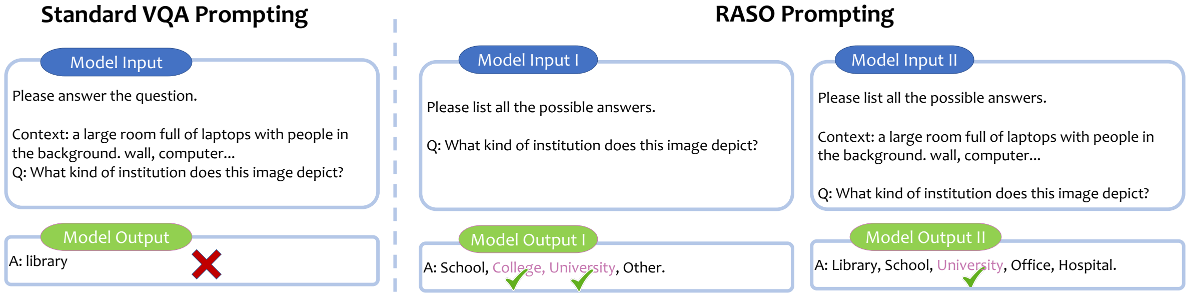
4 Experiment

4.1 Dataset

OK-VQA (Marino et al., 2021) is a widely used VQA dataset that requires external world knowledge outside of the image to answer the question. The dataset contains 14,031 images from the COCO dataset (Lin et al., 2014) and 14,055 crowd-

[1] PLM icon credit to https://claudeai.uk/ and model icon credit to https://www.vecteezy.com/vector-art/8156819-bot-reading-the-book

[2] Azure Tagging API:https://westus.dev.cognitive.microsoft.com/docs/services/computer-vision-v3-2/operations/56f91f2e778daf14a499f21b



Please answer the question.

Context: a large room full of laptops with people in the background. wall, computer...
Q: What kind of institution does this image depict?

A: library

Please list all the possible answers.

Q: What kind of institution does this image depict?

A: School, College, University, Other.

Please list all the possible answers.

Context: a large room full of laptops with people in the background. wall, computer...

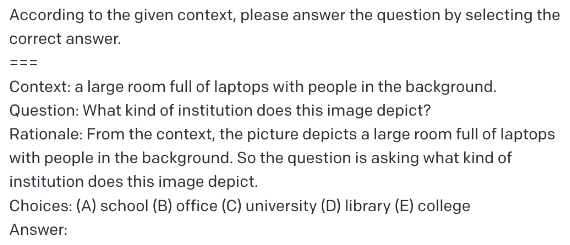
Q: What kind of institution does this image depict?

A: Library, School, University, Office, Hospital.

Figure 4: An illustration of our proposed prompting method for choice generation enabling larger knowledge retrieval coverage, compared with standard prompting as in PiCA (Yang et al., 2022). Note that Model Input I and II corresponds to $Prompt_Q$, $Prompt_{QC}$ respectively, and correct answers are highlighted.

PLM	Prompt Type	Top1 (%)	Top3 (%)	Top5 (%)	All (%)	Avg #
GPT J	$Prompt_Q$	32.4	46.1	46.7	46.7	2.6
	$Prompt_{QC}$	37.1	49.5	50.7	50.7	3.0
	both	37.1	52.0	55.9	57.1	4.1
UL2	$Prompt_Q$	32.6	45.4	46.4	46.5	2.7
	$Prompt_{QC}$	37.5	51.3	52.8	52.9	3.0
	both	37.5	53.1	57.0	58.0	4.1
GPT-3	$Prompt_{QC}$	48.0	-	-	-	-
OPT	$Prompt_Q$	34.21	48.45	49.7	49.8	3.0
	$Prompt_{QC}$	37.8	52.9	55.0	55.4	3.7
	both	37.8	55.6	61.0	63.4	5.2
Codex	$Prompt_Q$	44.8	58.8	59.8	59.8	3.1
	$Prompt_{QC}$	52.9	67.8	68.9	68.9	3.2
	both	52.9	68.6	72.6	**73.5**	4.5
ensembled	both	52.9	68.6	74.6	**81.9**	11.0

Table 2: Answer choices generation result on OK-VQA, representing the external knowledge coverage. Top 1, Top 3, Top 5, and All represent the highest accuracy that can be achieved using top 1, top 3, top 5, and all answer choices. All results are in accuracy scores evaluated following (Antol et al., 2015). "both" means that we combine the answer choices generated using both prompts. "ensembled" means that we combine the answer choices of all four PLMs. Note that the GPT-3 result is taken from (Yang et al., 2022).



According to the given context, please answer the question by selecting the correct answer.
===
Context: a large room full of laptops with people in the background.
Question: What kind of institution does this image depict?
Rationale: From the context, the picture depicts a large room full of laptops with people in the background. So the question is asking what kind of institution does this image depict.
Choices: (A) school (B) office (C) university (D) library (E) college
Answer:

Figure 5: Example input for the answer selection model for the image in Figures 1 and 2.

sourced questions covering a variety of knowledge categories, with 9,009 training data and 5,046 testing data. Each question has ten annotated answers (possibly repeated), and we follow the standard evaluation metric recommended by the VQA challenge (Antol et al., 2015). The external knowledge required in OK-VQA is not provided and there is no designated external knowledge source (such as a knowledge base), leaving the benchmark more challenging.

4.2 Publicly Available PLMs

We experiment with four different-sized PLMs that are publicly available as follows:

Codex (Chen et al., 2021) The Codex models are descendants of GPT-3 models that can understand and generate code. Their training data contains both natural language and billions of lines of public code from GitHub. We use the version $code-davinci-002$ of Codex.

OPT-175b (Zhang et al., 2022) Open Pre-trained Transformers (OPT) is a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters trained on publicly available datasets.

Method	External Knowledge Source	Answer Selector	Acc(%)
MUTAN+AN (Ben-Younes et al., 2017)	Wiki	-	27.8
ConceptBERT (Gardères et al., 2020)	ConceptNet	-	33.7
KRISP (Marino et al., 2021)	Wiki+ConceptNet	-	38.9
MAVEx (Wu et al., 2022)	Wiki+ConceptNet+Google Images	-	39.4
PiCA (Yang et al., 2022)	Frozen GPT-3 Wiki	-	48.0
KAT (Gui et al., 2022) (ensemble)	Wiki+Frozen GPT-3 Wiki	-	54.4
ClipCap (Mokady et al., 2021) .	-	-	22.8
RASO	Frozen GPT-J	ClipCap	29.5
	Frozen UL2		33.1
	Frozen OPT		31.3
	Frozen Codex		35.3
	All 4 Frozen PLMs		38.0
RASO	Frozen GPT-J	IterPLM	29.6
	Frozen UL2		33.8
	Frozen OPT		58.5
	Frozen Codex		45.7
RASO	Frozen GPT-J	UnifiedQA (ensemble)	47.2
	Frozen UL2		45.8
	Frozen OPT		47.8
	Frozen Codex		51.2
	All 4 Frozen PLMs		45.6
RASO	Wiki+Frozen GPT-J	KAT (ensemble)	50.3
	Wiki+Frozen UL2		52.2
	Wiki+Frozen OPT		53.0
	Wiki+Frozen Codex		**58.5**
	Wiki+ All 4 Frozen PLMs		57.9

Table 3: VQA results on the OK-VQA benchmark comparing to standard baselines. "Wiki" stands for "Wikipedia" and the "Wiki" resource in the last row's block is brought by the answer selector KAT. "All 4 Frozen PLMs" means that we use all the answer choices generated by GPT-J, UL2, OPT, and Codex. When we have UnifiedQA or KAT as answer selector, we train with 3 random seeds and denote the results as $ensemble$ following (Gui et al., 2022).

We use the version 175 billion parameters of OPT.

UL2 (Tay et al., 2022) Unified Language Learner (UL2) is 20 billion parameter novel language pretraining paradigm that improves the performance of language models universally across datasets and setups released recently. UL2 frames different objective functions for training language models as denoising tasks, where the model has to recover missing sub-sequences of a given input.

GPT-J (Wang and Komatsuzaki, 2021) GPT-J is a 6 billion parameter, autoregressive text generation model trained following (Wang, 2021). The model consists of 28 layers with a model dimension of 4096, and a feed-forward dimension of 16384.

During prompting, we always set the temperature to 0.001 and max token to 15.

4.3 Answer Choices Generation Results

The answer choice generation result is shown in Table 2. Top 1, Top 3,..., All represent the highest accuracy that can be achieved using top 1, top 3, ..., and all answer choices, calculated following the standard VQA evaluation metric in (Antol et al., 2015). Note that the GPT-3 score is taken from (Yang et al., 2022). We do not experiment with GPT-3 in this paper due to the required cost. Avg # stands for the average number of answer choices.

While previous VQA methods also retrieve from PLMs, they have a similar result as if using $Prompt_{QC}$ and Top1 choice. As discussed before, these generation results can represent the external knowledge coverage ratio. From the table, Codex covers the majority of the knowledge needed and has the highest score of 73.5%. Using our prompt-engineering method, the knowledge coverages of all PLMs increase by a large margin of at least 20% (which are the accuracy differences between Top1 choice by $Prompt_{QC}$ and All choices by both prompts).

4.4 Answer Selection Models

We plug in existing text-generation models as answer selectors and experiment on four methods:

KAT (Gui et al., 2022) is a VQA method that uses a sequence-to-sequence model composed of an encoder and a decoder, similar to T5 (Raffel et al., 2020). As far as this paper is written, KAT is known to be the SOTA method on OK-VQA benchmark.

ClipCap (Mokady et al., 2021) uses the CLIP (Radford et al., 2021) encoding as a prefix to generate

KAT	Top1	All w/o cot	All w/ cot
GPT-J (single)	45.9	47.8	49.6
GPT-J (ensemble)	46.6	48.4	50.3
UL2 (single)	50.2	50.7	51.2
UL2 (ensemble)	51.1	51.5	52.2
OPT (single)	51.7	52.3	52.5
OPT (ensemble)	52.1	52.9	53.0
Codex (single)	56.2	57.1	57.5
Codex (ensemble)	57.1	58.1	**58.5**
All (single)	56.4	56.9	57.0
All (ensemble)	57.0	57.6	57.9

UnifiedQA	All w/o cot	All w/ cot
GPT-J (single)	45.6	46.0
GPT-J (ensemble)	46.6	47.2
UL2 (single)	44.8	44.6
UL2 (ensemble)	45.8	45.8
OPT (single)	47.9	46.8
OPT (ensemble)	49.0	47.8
Codex (single)	51.1	50.4
Codex (ensemble)	**52.1**	51.2
All (single)	45.1	44.6
All (ensemble)	45.7	45.3

Table 4: Ablation study investing how different inputs influence the answer selection results using KAT (top) and UnifiedQA (bottom) on OK-VQA in accuracy scores. "Top1" means using Top 1 answer choice,"All" in the first row means using all answer choices, to form the input respectively. "cot" means the CoT rationales. We train with 3 random seeds and denote the average scores as *single* and majority vote results as *ensemble*."All" in the leftmost column represent using combined answer choices from all four PLMs.

textual captions by employing a simple mapping network over the raw encoding, and then fine-tunes a language model to generate a valid caption. The language model we use here is GPT-2. In this paper, we adapt this model by adding question tokens, CoT rationale tokens, and answer choices tokens to the prefix as input, with the target to generate answers instead of captions. We train the mapping network from scratch and also fine-tune GPT-2.

IterPLM Inspired by previous work (Wang et al., 2022), we use iterative prompting with the same PLM in choice generation for correct answer selection. A snippet of an example prompt is shown in Figure 5. We use 8-shot in-domain examples with the temperature set to 0.001 and max token set to 5.

	GPT-J	UL2	OPT	Codex
w/o cot	28.5	29.1	31.6	**45.6**
w/ cot	28.1	32.3	33.5	44.9

Table 5: Ablation study on how different inputs influence the answer selection result using IterPLM: iterative prompting using the same PLM, on OK-VQA. All results are in accuracy scores. Both setting use all the answer choices.

	Type	GPT-J	UL2	OPT	Codex
ViT-L_14	DG	23.5			
	w/o cot	28.7	30.3	29.1	33.4
	w/ cot	29.5	33.1	31.3	**35.3**
RN50x64	DG	21.6			
	w/o cot	29.3	30.3	28.6	34.5
	w/ cot	29.6	32.6	31.4	**36.4**

Table 6: Ablation study on how different inputs influence the answer selection result using ClipCapVQA (Mokady et al., 2021) on OK-VQA. The first column represents two CLIP checkpoints. "DG" represents direct generation without any answer choices.

UnifiedQA (Khashabi et al., 2022, 2020) is a multiple-choice question answering (QA) model that performs well across 20 QA datasets, using the T5ForConditionalGeneration model. We load UnifiedQA v2 (Khashabi et al., 2022) checkpoint unifiedqa-v2-t5-large-1251000.

4.5 End-task VQA Results

As illustrated in Table 3, we compare our proposed pipeline against several standard baseline approaches: MUTAN+AN (Ben-Younes et al., 2017), ConceptBERT (Gardères et al., 2020), KRISP (Marino et al., 2021), MAVEx (Wu et al., 2022), PiCA (Yang et al., 2022), and KAT (Gui et al., 2022), on the OK-VQA data test set. RASO outperforms the previous SOTA by an absolute 4% margin, achieving the new SOTA.

Comparing different answer selectors, it is surprising that the two transformer-based text-only models: UnifiedQA and KAT significantly outperform the multi-modal ClipCap model by around 20% on average, even though their sizes (T5 large) are much smaller than that of GPT-2. We believe this phenomenon is because the Clip image embeddings trained using image captions do not have enough granularity to support reasoning over the image, question, and answer choices for answer selection, compared to T5 models. Besides, IterPLM has much worse scores than we imagined. While many papers (Wang et al., 2022) show that iterative prompting should boost the performance, our experiments suggest that asking the PLMs to select between their own output at the highest confidence is indeed a very difficult problem for them.

In Table 3, we also compare single PLM answer choices with ensembled choices by all four PLMs, with the latter showing lower scores. We believe this is because the answer selectors we experiment on are not good enough, and thus increasing choice numbers turns out to hurt the performance.

4.6 Implementation Details

In the answer choice generation step, we use 16-shot in-context examples on the test data. On the training data, because we have ten gold answers with repetitions, we use 4-shot in-context learning for faster generation. The temperature for PLM generation is set to be 0.001. The generation max token length is set to be 15. All experiments of selection models have been run in 8 NVIDIA V100 Tensor Core GPUs with 32 GiB of memory each, 96 custom Intel Xeon Scalable (Skylake) vCPUs, and 1.8 TB of local NVMe-based SSD storage. The running times for KAT, UnifiedQA and ClipCap are less than 4, 2 and 1 hours, respectively. OPT-175b model is locally set up in 32 NVIDIA V100 Tensor Core GPUs to make inferences. The learning rates for KAT, UnifiedQA and Clipcap are set as 3e-5, 5e-5 and 2e-5, respectively, for all experiments. Optimizer AdamW (Loshchilov and Hutter, 2017) is used for all selection models.

5 Ablation Studies

We perform qualitative and quantitative analysis on the answer selection results to better understand whether the expanded external knowledge coverage benefits the end-task VQA much. As illustrated in Tables 4 to 6, we investigate the impact of various inputs on the answer selection results, with answer choices representing the retrieved knowledge.

CoT Rationale Impact From the experiments results in Tables 4 to 6 where we compare the settings: "w/cot" and "w/o cot", input with CoT rationales consistently boosts the answer selection performance of KAT, UnifiedQA, and ClipCap. However, this conclusion fails for iterative prompting – adding CoT hurts the performance of IterPLM when we use GPT-J and Codex. We believe this can result from the difference in CoT qualities, and different pre-training methods and data.

Choice Number Impact As shown in Table 4, larger knowledge coverage, represented by using choices from all four PLMs versus a single PLM, can not consistently increase the performance of KAT or UnifiedQA. As we compare the results on Codex choices and that on all PLMs choices, more choices always lead to lower accuracy scores. This is somehow against our instinct, and we believe it is because our answer selectors are not good enough. Digging deeper into the problem, we further compare the difference between using Top1 choices and all choices in KAT as in the top table. Note that the Top1 results here are not the same as the Top1 accuracy in Table 2 because KAT uses Wikipedia knowledge by design so it further expands knowledge coverage. We can see that using all choices is consistently better than using Top 1 choice. However, the improvements are too small (0.4-1.9 %) considering that their knowledge coverages differ by at least 20% as in Table 1, suggesting that KAT, while being the best, is still not the ideal selection model, and motivating future research in this direction.

Multi-modal Selector Impact As demonstrated in Table 6, we experiment with the two versions of CLIP embedding: "ViT-L_14" and "RN50x64" and the difference between direct generation (DG) and answer selection is constantly large – providing answer choices definitely helps ClipCap to generate the correct answer.

Ensemble Impact Our answer choice generation step is indeed ensembling on PLMs results. Previous VQA methods that retrieve from PLMs also conduct ensembling but in a different way (Yang et al., 2022). They usually request the same prompt (see example in Figure 4) multiple times and take the majority-voted answer. This process is called multi-query ensemble, and could boost the GPT-3 performance by about 5%. We argue that our proposed RASO prompting is superior to multi-query ensemble in that we allow more diversity in the output and provide VQA systems more explainability by separating the choice-generation and selection steps, without additional API request cost or longer inference time.

6 Conclusion

In this paper, we propose RASO: a new VQA pipeline following a generate-then-select strategy guided by world knowledge. RASO proposes a new prompting method that largely increases the external knowledge coverage by a margin of more than 20% compared to previous approaches on the OK-VQA benchmark. Our pipeline achieves the new SOTA 58.5% on the end-task performance , encouraging avenues for future studies.

7 Limitations

While the previous VQA methods that retrieve from PLMs all use GPT-3, we do not experiment with GPT-3 in this paper due to the additional cost. We only focus on applying text-generation models as answer selectors, while classification models could also potentially be good answer selectors. The multi-modal CLIP embedding has already been surpassed by several recent studies (Alayrac et al., 2022; Singh et al., 2022; Lu et al., 2022) and therefore ClipCap cannot represent the performance of multi-modal answer selectors.

8 Ethical Considerations

The authors of this paper acknowledge the significance of responsible NLP in research and development. The objective of this research is to enhance the capabilities of visual question answering models, guided by human values-based world knowledge. We strive to ensure that the model is not only accurate and efficient, but also fair and unbiased. We recognize that the VQA technology may have a substantial impact on society and pledge to be transparent in sharing our findings and progress with relevant users and stakeholders.

Acknowledgments

The authors would like to thank researchers at AWS AI Labs who commented on or otherwise supported throughout the course of this project, including Simeng Han, Donghan Yu, Sijia Wang, and Shuaichen Chang.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16548–16558.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, and Dan Roth. 2022. There's a time and place for reasoning beyond the image. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1138–1149, Dublin, Ireland. Association for Computational Linguistics.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *CVPR 2022*.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Yushi* Hu, Hang* Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Leroy Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. REVIVE: Regional visual representation matters in knowledge-based visual question answering. In *Advances in Neural Information Processing Systems*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. 2023. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*, 212:118669.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Shepherd pre-trained language models to develop a train of thought: An iterative prompting approach. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wikipedia contributors. 2004. Plagiarism — Wikipedia, the free encyclopedia. [Online; accessed 22-July-2004].

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2712–2721.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

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

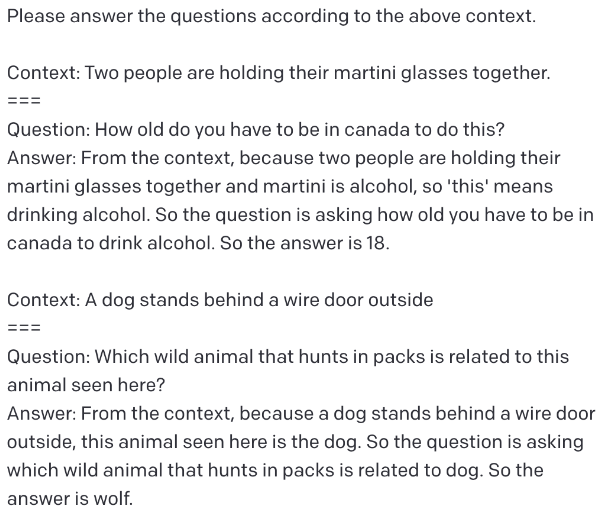
A Appendix

A.1 CoT prompts

For our CoT generation experiments, we use a pre-designed fixed prompt as partly shown in Figure 6.

A.2 Additional Experiments

We conduct additional experiments for RASO on an augmented successor dataset of OK-VQA: A-OKVQA (Schwenk et al., 2022) to prove its effectiveness. Since we do not have the baseline results or any intermediate outputs on A-OKVQA as the paper was written, we only compare with PiCA (Yang et al., 2022) with a simpler setting: without using image tagging or chain-of-thought and only using GPT-J. The captions we use are generated using BLIP-2 (Li et al., 2023), following the default example in the paper.



Please answer the questions according to the above context.

Context: Two people are holding their martini glasses together.
===
Question: How old do you have to be in canada to do this?
Answer: From the context, because two people are holding their martini glasses together and martini is alcohol, so 'this' means drinking alcohol. So the question is asking how old you have to be in canada to drink alcohol. So the answer is 18.

Context: A dog stands behind a wire door outside
===
Question: Which wild animal that hunts in packs is related to this animal seen here?
Answer: From the context, because a dog stands behind a wire door outside, this animal seen here is the dog. So the question is asking which wild animal that hunts in packs is related to dog. So the answer is wolf.

Figure 6: The fixed prompt we use to generate chain-of-thought style rationales. We randomly select seven examples in the prompt and show two of them here. We set the temperature as 0.7 and max token as 80 during inference for all PLMs.

	PiCA	RASO
A-OKVQA	33.2	**37.1**

Table 7: Additional comparison of RASO versus PiCA on A-OKVQA dataset.