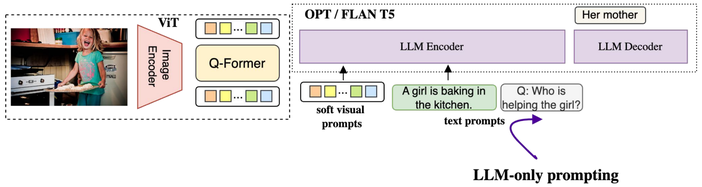
# In Search For A Good Prompt in Zero-shot Visual Question Answering

Rabiul Awal, Le Zhang, Aishwarya Agrawal
Mila - Quebec AI Institute, Université de Montréal

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

## Motivation

- Visual question answering (VQA) is a challenging task that requires comprehension and reasoning with visual information.
- Recent vision-language models struggle with zero-shot VQA, especially in handling complex compositional questions and adapting to new domains.
- Inspired by the success of various prompting techniques for large language models (e.g. GPT3), we investigate which prompting techniques are effective for this recent paradigm of zero- and few-shot VQA.




## Investigating VQA Prompting: Four factors

1. **Choice of question template:** Explore different templates to guide answer generation in VQA.
2. **Incorporating text-only few-shot examples:** Improve model's VQA task understanding using text-only Q&A examples.
3. **Incorporating image captions as visual cues:** Use image captions to augment comprehension of visual data.
4. **Incorporating chain-of-thought reasoning:** Apply chain-of-thought reasoning for step-by-step rationale for VQA answers.


Outside Knowledge: OKVQA, AOKVQA

Compositional Reasoning: GQA, Winoground-QA

## VQA Prompting Techniques

### Standard VQA

**Templates**
**Question:** Who is helping the girl?
**Short Answer:** Her mother

**Please answer the following question.**
Who is helping the girl? Her mother

### Caption VQA

**Step 1. Caption Generation**
**Prompt *(a-photo-of)*:** "A photo of"
**Caption: A photo of** a girl holding a tray of food in the kitchen.

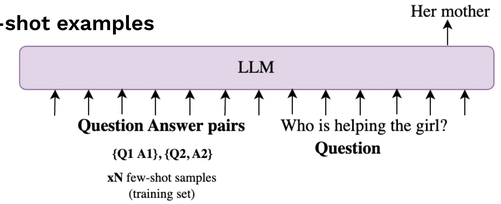
**Prompt (q-guided-cap):** "Please describe the image according to answer the following question *Who is helping the girl?*"
**Caption: A photo of** a girl baking in the kitchen with her mother.

**Step 2. Question Answering**
**Context:** A photo of a girl baking in the kitchen with her mother?
**Question:** Who is helping the girl?
**Short Answer:** Her mother.

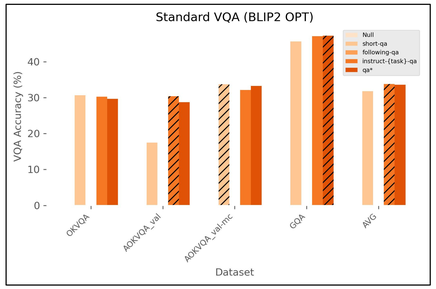
### Chain-of-thought VQA

**Please answer the following question by reasoning step by step. Q:** Who is helping the girl?
**A:** The girl's mother might be helping the girl. The girl is in the kitchen with her mother. The mother is holding a tray of food. The tray is full of cookies. Therefore, the final answer is a mother.

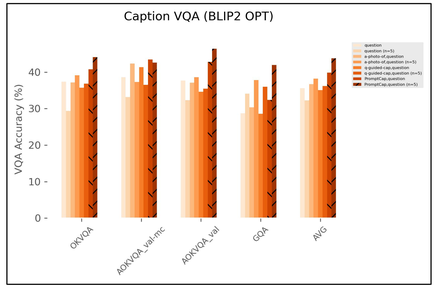
**Few-shot examples**



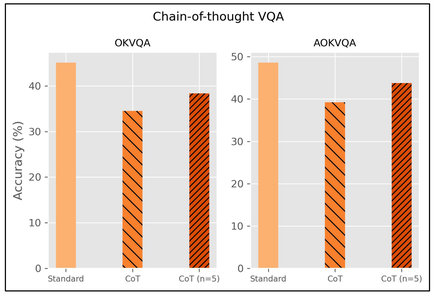

## Experiments



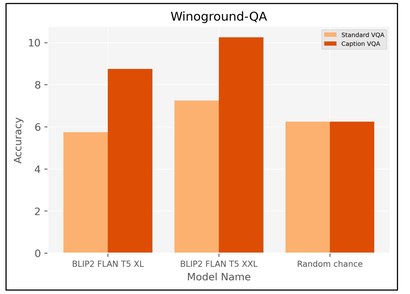

**Q&A template matters**




**Caption VQA significantly boost VQA accuracy**




**Rationalization hurts VQA accuracy**




**Winoground-QA is a challenging VQA task**

## Conclusion

- Text-only prompting strategies significantly improve VQA accuracy for BLIP2 models.
- Our proposed **Caption VQA** shows significant performance boost.
- We introduce **Winoground-QA**, a challenging VQA task.
- Chain-of-thought rationalization negatively affects accuracy.