Can Intermediate Reasoning Chains Rationalize Better in Multimodals?

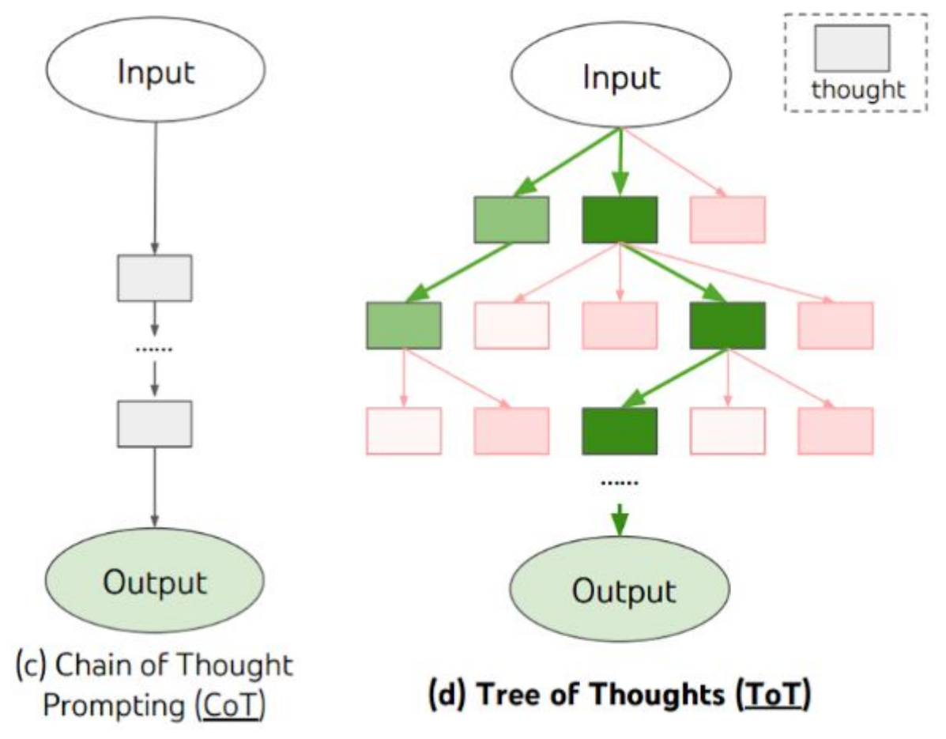
Mani Deep Cherukuri, Shunichi Sawamura, Shashank Sharma, Nicole Vu

Motivation & Problem Definition

- **Chain-of-Thought** (CoT) is effective for diverse problem-solving scenarios.
- **Tree of Thought** (ToT) enhances CoT by introducing a planning process in a tree format.
- Both these methods have not been explored much in the multimodals spectrum. This project aims to investigate and fill this gap.

Literature Review

- **Chain-of-Thought (CoT)** resembles human thought process, where the final answer is generated by a series of intermediate steps [1].
- **Tree-of-Thought (ToT)** extends the CoT similarly to a search over a tree, with each node representing a partial solution [2].



(c) Chain of Thought Prompting (CoT)

(d) Tree of Thoughts (ToT)

Example

Question: Which statement describes the Yasuni National Park ecosystem?

Options:

0. It has mostly small plants.

1. It has many different types of organisms.

2. It has soil that is rich in nutrients.

Visual Information of Example Question

Tree of Thought (Yao et el. 2023)

Makes a plan for the solution strategy in the first layer and develops solutions in the next two layers. Each node generates the answer based on the parent node. In each layer, another gpt model evaluates and scores the feasibility of nodes between 1-10. The example generated plan is described below.

1. *Introduction of the Amazon rainforest as the largest rainforest ecosystem in the world.*
2. *Mention of Yasuni National Park in Ecuador as part of the Amazon rainforest.*
3. *Description of the diverse species of plants, birds, and mammals in Yasuni National Park.*
4. *Presentation of the question about the ecosystem of Yasuni National Park and the given options.*

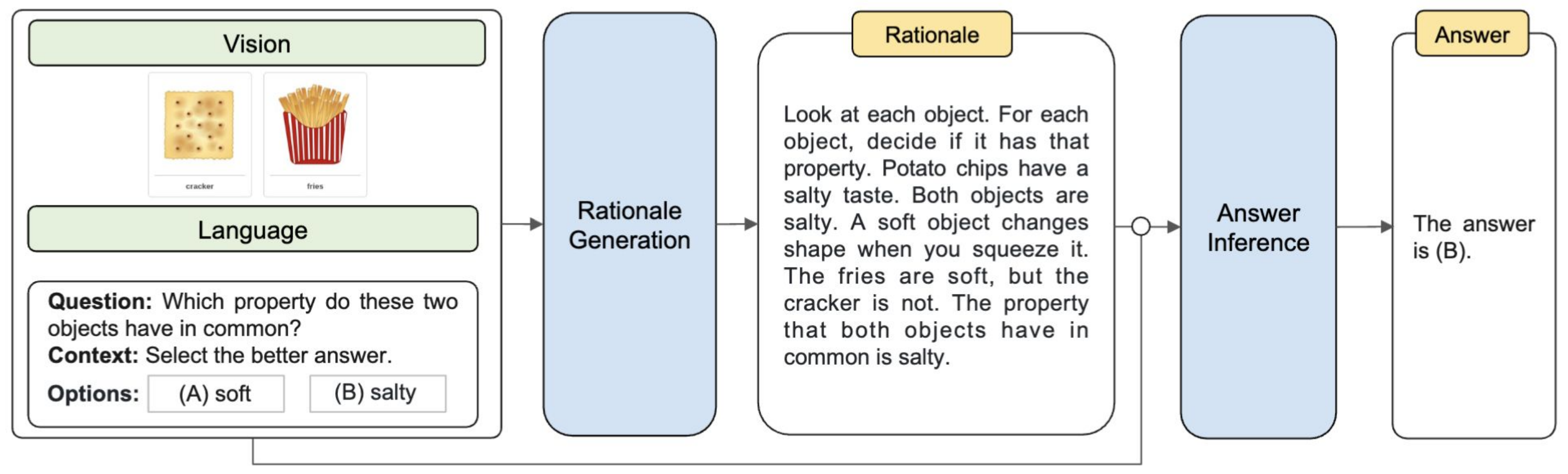
Tree of Thought Prompting (Hulbert. 2023)

Encourages GPT to answer the question with ToT style in a single response. The sentence below is the beginning of generated response.

Expert 1:

Step-1: *Identifying Key Features Upon reading the passage and analyzing the context, the first step is to identify the key features of the Yasuni National Park ecosystem. This includes recognizing that the park is located in the Amazon rainforest, known for its vast biodiversity, including various species of plants, birds, and mammals.*

- **Multimodal-CoT** model is the application of CoT on multimodal. It goes through 2 different stages, Rationale Generation to combine multiple inputs into one rationale, and Answer Inference to generate output from the rationale. Multimodal-CoT has shown promising results on answer inference stage for several tasks [3].



Question: Which property do these two objects have in common?
Context: Select the better answer.
Options: (A) soft (B) salty

Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.

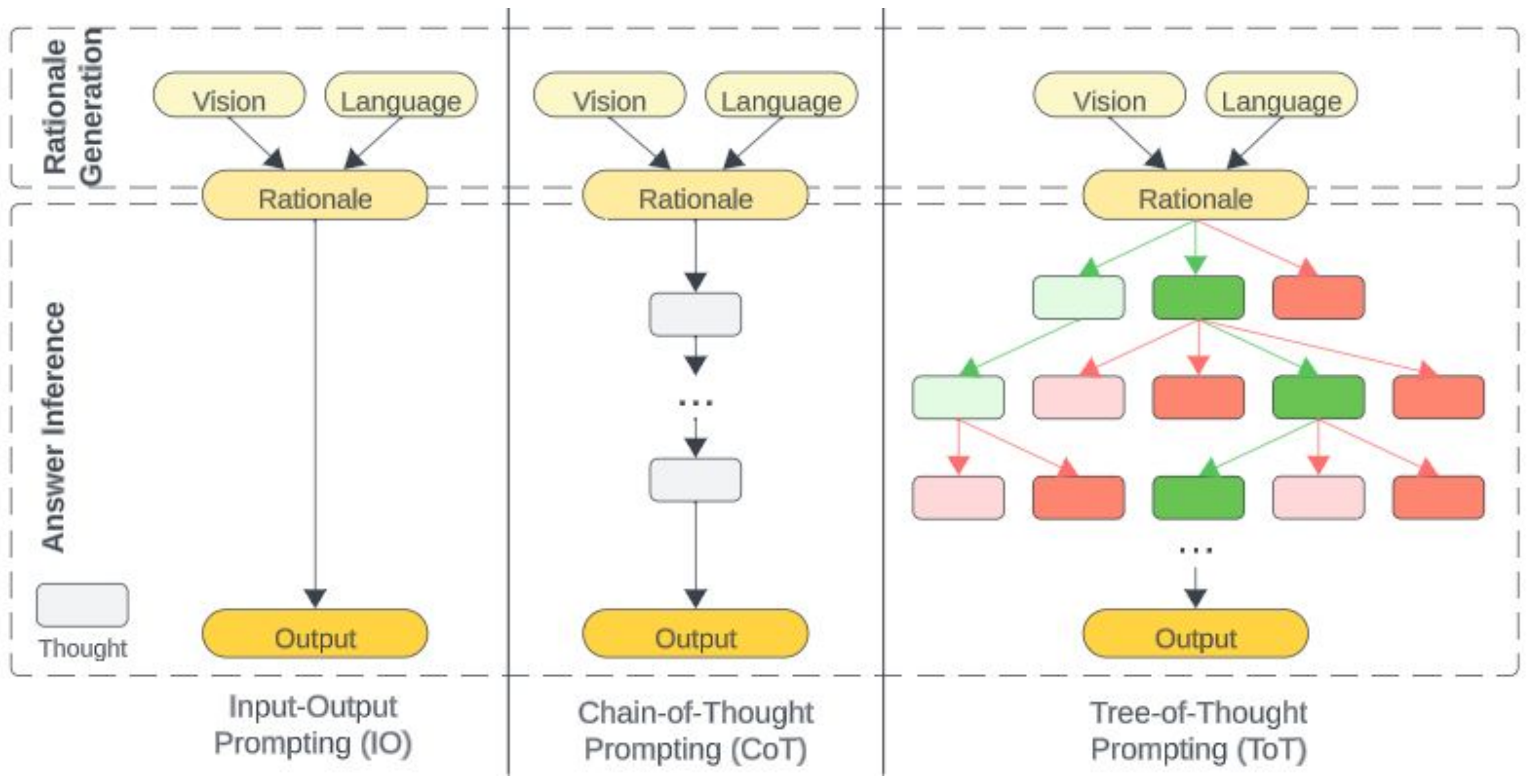
The answer is (B).

Novel Contribution

- While increasing the size of models has shown success in various tasks, it falls short in achieving optimal performance across diverse problem-solving scenarios like arithmetic, commonsense, and symbolic reasoning. Research indicates that a CoT, a connected series of reasoning steps leading to a solution, has been effective in addressing this limitation.
- The ToT which is built on top the CoT architecture has given some preliminary results on tackling several tasks where both the traditional prompting technique and CoT fails. We integrates ToT prompting and multimodal approaches to investigate the potential of this fusion, with the aim of uncovering additional insights for future research.

Proposed Ideas

- **Goal:** Explore CoT and ToT frameworks on Multimodals, with the goal to improve prompting performance on some tasks that traditional prompting techniques fail..
- **Multimodals baselines**: the baselines used in the project are Llava, Multimodal-CoT, GPT 3 and GPT4.
- **Dataset:** ScienceQA dataset, which has 21,208 multimodal multiple-choice science questions, then perform answer inference with the ToT method.
- **CoT/ToT and Multimodal Integration:**
  - Rationale Generation: Create a rationale using both the vision and language inputs taken from the ScienceQA dataset
  - Answer Inference:
    - Zero-shot: prompt the LM with the rationale as the input, the LM returns the output directly
    - CoT: prompt the LM with the rationale, the LM go through multiple intermediate reasoning step to generate the final output
    - ToT: prompt the LM with the rationale, the LM process the rationale through different layers where the answer of one layer is generated from the parent node to get the final output

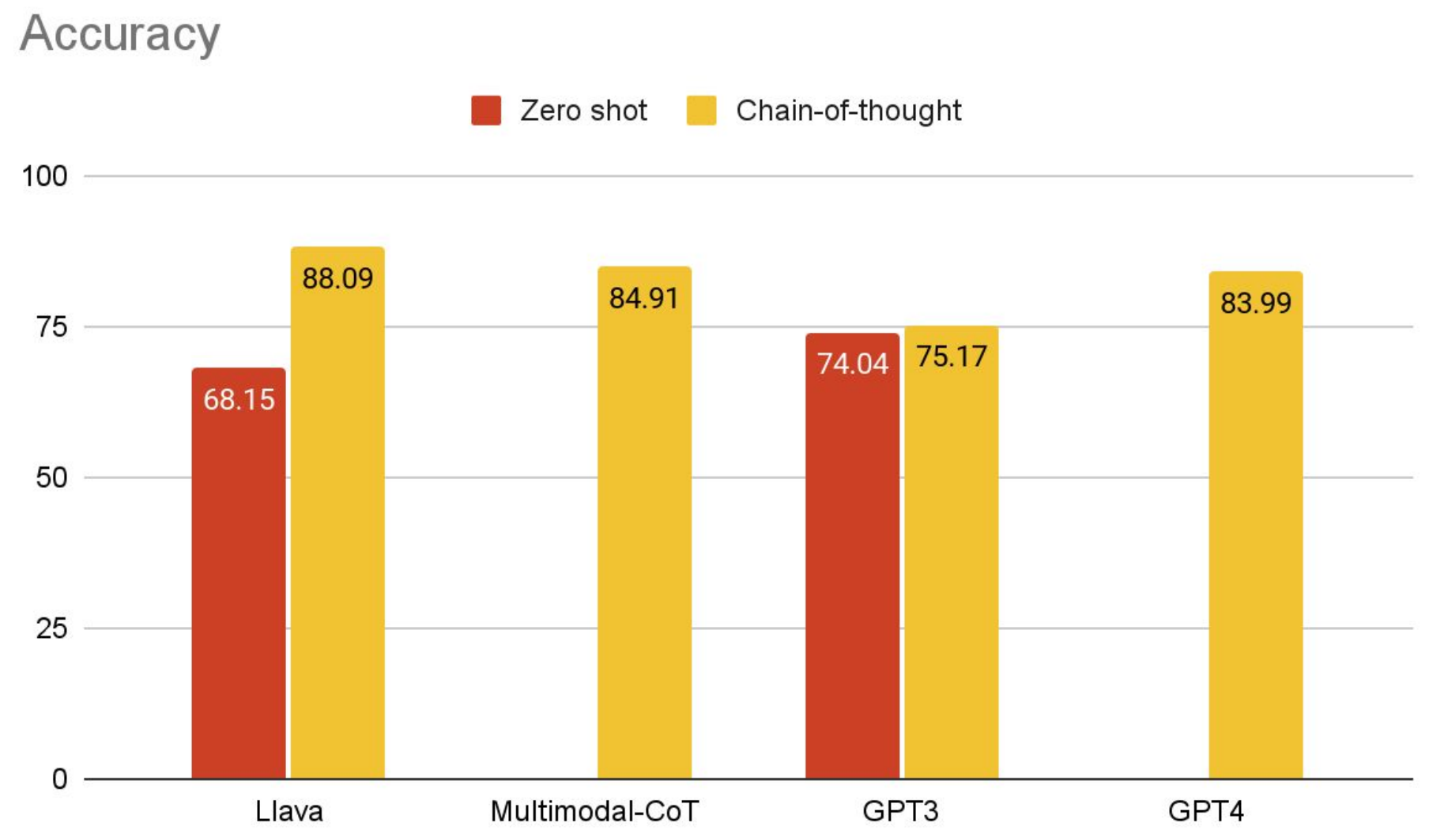


Rationale Generation
Answer Inference
Vision
Language
Rationale
Output
Thought
Input-Output Prompting (IO)
Chain-of-Thought Prompting (CoT)
Tree-of-Thought Prompting (ToT)

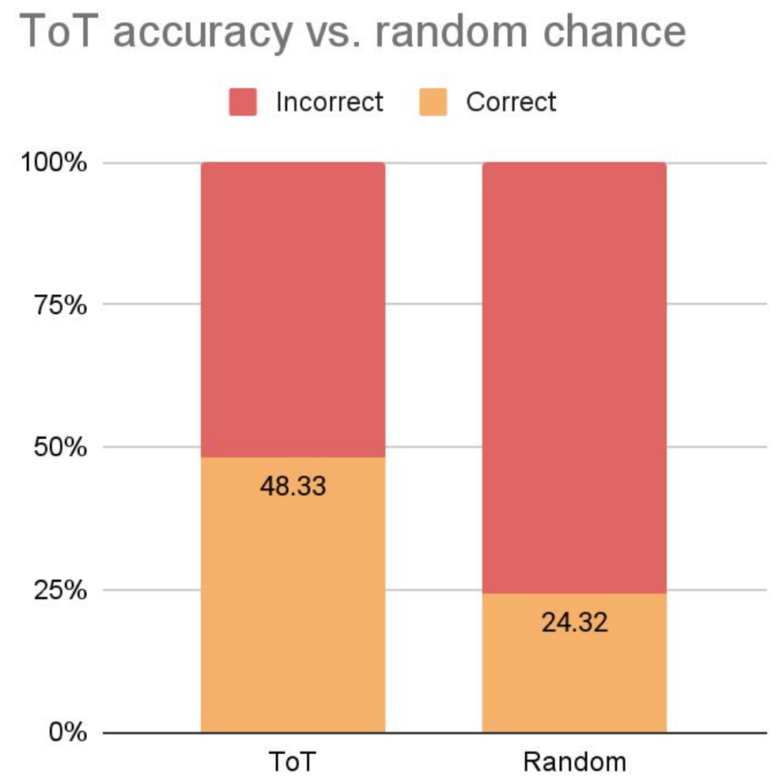
- **Implementing ToT:**
  - The authors of the paper "Tree of Thoughts: Deliberate Problem Solving with Large Language Models" [2] stated that if GPT-3.5 is already performing effectively on a certain task, the ToT may not necessarily yield improved results.
  - Hence, to identify the optimal tasks for ToT, we executed ToT in scenarios where both Multimodal-CoT and Llava, fine-tuned on the ScienceQA dataset, encountered failures.
    - There are 60 out of 2017 samples where both models failed
    - The accuracy from the ToT application on those fail cases is compared with the random chance to determine performance

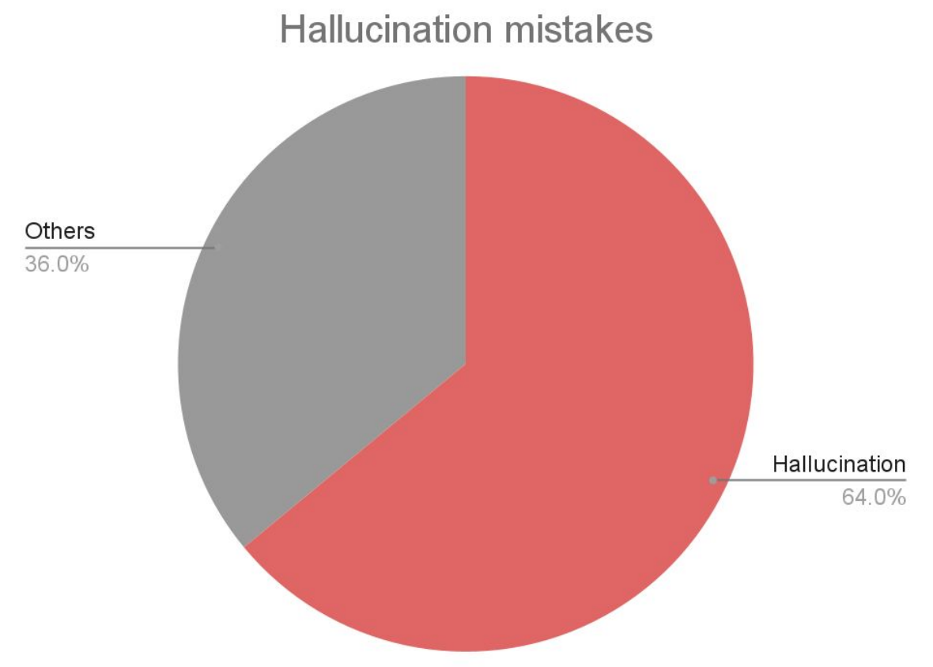
Experimental Results & Findings

- **Multimodal-zero-shot vs. Multimodal-CoT:**
  - To determine if CoT prompting would outperform zero-shot prompting on the ScienceQA dataset, we compare the outputs from our baseline language models with the test set's answer.
  - CoT prompting yields higher accuracy in comparison to the traditional zero-shot prompting as shown below.



- **Multimodal-ToT on the fail cases of Multimodal-CoT:**
  - When using the Multimodal-ToT on 60 cases where the Multimodal-CoT and fine-tuned Llava fail, we found an accuracy of 48.33% (29/60 correct answer).
  - This accuracy is greater than the chance of the LM just randomly pick an answer (24.32%).
  - Therefore, the Multimodal-ToT can potentially handle the tasks where Multimodal-CoT fails. However, since the limitation of the size of the dataset, more experiments on a bigger dataset is necessary to make a concrete conclusion.





  - We also observe a high rate of hallucinations in the ToT application. In the example below, the model returns a city that is not in the list of options given in the question.

Example

Question: Which of these cities is marked on the map?

Options:

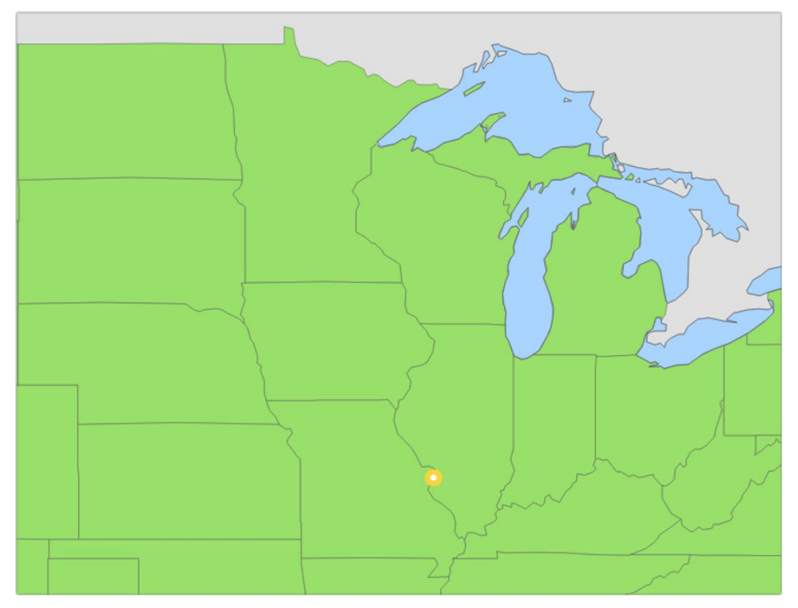
0. Omaha

1. Chicago

2. St.Louis

3. Cleveland

ToT prompting output: Milwaukee



  - We discovered that Multimodal-CoT exhibits higher failure rates in biology and geography. Through the application of ToT, we achieved enhanced accuracy in these specific tasks. Additionally, ToT has better performance in physics, writing, vocabulary, and history.

[1] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv.Org. https://doi.org/10.48550/arxiv.2201.11903

[2] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv.Org. https://doi.org/10.48550/arxiv.2305.10601

[3] Zhang, Z., Aston, Z., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal Chain-of-Thought Reasoning in Language Models. ArXiv.Org. https://doi.org/10.48550/arxiv.2302.00923