# Can Intermediate Reasoning Chains Rationalize Better in Multimodals?

**Mani Deep Cherukuri, Shunichi Sawamura, Shashank Sharma, and Nicole Vu**
{cheru050, sawam002, sharm964, vu000166}@umn.edu
University of Minnesota
Team: Pilot

## 1 Introduction

In this report, we investigate whether using intermediate reasoning chains can significantly enhance large language models' ability to comprehend complex multimodal inputs, which involve both text and images. Our primary focus is on evaluating the performance of various models, including LLaVA, LLaMA, Multimodal-CoT, GPT-3, and GPT-4. We employ different prompting techniques such as chain-of-thought and tree-of-thought.

The overarching goal of our exploration is to determine if adopting more thoughtful and deliberate reasoning approaches can significantly improve the accuracy of answering complex questions that involve both textual and visual information. Our project specifically aims to assess the performance of these models, including chain-of-thought (CoT) and tree-of-thought (ToT) prompting methods, when confronted with challenging multimodal questions from the ScienceQA dataset.

Our hypothesis suggests that utilizing CoT and ToT prompting methods will outperform traditional zero-shot prompting approaches in handling these intricate tasks. By merging the insights from these two perspectives, we aim to contribute valuable knowledge to the field and advance our understanding of how intermediate reasoning chains can be leveraged to optimize multimodal models for more accurate and reliable question answering.

## 2 Background

In the realm of answering questions, most current methods involve presenting a language model with a question and choices, expecting it to generate a prediction—a process known as "zero-shot" prompting. However, this approach can struggle when dealing with more complex reasoning tasks. Recent innovations, such as the chain-of-thought technique, aim to improve this by creating a clear reasoning chain that justifies the model's inferences, enhancing both understanding and accuracy.

Taking a step further, the tree-of-thought method explores multiple reasoning chains in a tree structure before determining the most coherent one. Despite these promising advancements, there's still much to explore, especially when dealing with multimodal inputs—scenarios involving both text and images.

Prior studies on question answering using large language models have shown promise with the application of chain-of-thought (CoT) prompting. CoT involves guiding the model with a reasoning chain, leading to improved performance compared to traditional zero-shot prompting. However, even with these advancements, CoT faces challenges with certain intricate reasoning tasks.

Expanding on CoT, the tree-of-thought (ToT) prompting method introduces the generation of multiple reasoning paths in a tree structure. This allows for the exploration and evaluation of different lines of reasoning. Initial research indicates that ToT can successfully address tasks that prove challenging for CoT. This project aims to investigate whether combining ToT with multimodal inputs can further enhance overall performance, contributing valuable insights to the optimization of large language models for complex reasoning in diverse scenarios.

## 3 Motivation

Our motivation for this investigation is all about understanding how fancy ways of asking questions can make big language models even better at figuring out answers, especially when dealing with both words and pictures. As these smart models step into real-world situations, it's super important to make sure they're not just good at what they do but also clear about how they think things through.

You can think of it like this: imagine teaching a really smart robot to read a story that has both words and pictures, and then asking it questions about the story. That's kind of what we're doing with language models. But, it's not just about stories – these models are used for all sorts of things in the real world.

So, why bother with all this? Well, because the more we can make these models understand and answer questions in a reliable way, the more we can trust them in different situations. This is crucial when these models are used in things like helping us with information or making decisions.

Our goal is to find the best ways to guide these models' thinking, especially when dealing with both words and pictures. By doing this, we hope to share some smart practices that can help others in the future when they're working on similar cool projects at the intersection of tricky reasoning tasks and understanding both text and images.

## 4 Approach

In this project, we investigate the efficacy of several types of large language models and multiple prompting methods in visual question answering. We hypothesize that Chain-of-Thought and Tree-of-Thought avail to solve complex questions that traditional prompting techniques fail. We choose the ScienceQA as a test dataset in this experiment since it consists of diverse science-related problems and is suitable to measure the problem-solving skill of LLMs. Comparison of the performance of large language models with different prompting methods in this dataset would demonstrate how well each model and prompting technique can solve the complex questions. Also, because the dataset has richer domain diversity including natural science, language science, and social science, analysis of their performance would clarify their strengths and weaknesses which could be a key to improving them in the future. Since we only focus on testing the performance in visual question answering, we use a test dataset of ScienceQA that has images in the question. The input of each question consists of a question statement, the answer choices, the rationale of the question, and visual information.

### 4.1 Models

- **LLaVA**: We first implement several sorts of LLMs as baseline models without the CoT or ToT approach. LLaVA, or "Visual Instruction-Tuning,"[Liu et al., 2023] is a multimodal model designed for instruction-following and reasoning. It combines LLaMA[Touvron et al., 2023] with the CLIP visual encoder (ViT-L/14) [Radford et al., 2021]. LLaVA connects image features to language embeddings using a simple linear layer. For training, LLaVA utilizes multi-turn conversation data, organizing it as a sequence of turns with the assistant's responses and instructions aligned with the image. Instruction-tuning on the LLM aims to optimize the probability of generating target answers through an auto-regressive training objective. Since it has 13 billion parameters, and we struggle with loading and testing, we choose 4-bit quantized LLaVA. LLaVA is evaluated based on the performance in zero-shot prompting.

- **LLaMA**: LLaMA [Touvron et al., 2023] is a large language model developed by Meta. The model is loaded using LLM Engine [for AI, 2023] which is the open-source engine for fine-tuning and serving large language models. We decide to use this open source because it also has scienceQA dataset and enables us to easily implement fine-tuning the model through the API. The model name is "llama-2-7b" which means the LLaMA 2 whose parameter size is 7 billion. Since LLaMA can only handle text as a prompt, we utilize image captioning with CLIP https://arxiv.org/abs/2111.09734 and convert the question images to text and add it to the prompt so that we can reproduce the performance of fine-tuned LLaVA.

- **GPT-3**: GPT-3 [Brown et al., 2020] is the model that can be applied to various tasks and can be the baseline model in our analysis. We plan to implement few-shot learning so that the GPT-3 model can answer questions without parameter updates. GPT-3 will be instructed to answer the ScienceQA by providing in-context examples with components of the question text, options, and the correct answer text. It encourages the GPT-3 to lead the answer via chain-of-thought prompting. [Lu et al., 2022]

- **GPT-4**: GPT-4 represents a significant advancement over its predecessors, GPT-3 and GPT-3.5, primarily attributed to its substantial increase in parameters up to 1.76 trillion. This expansion enhances reliability, creativity, and a heightened capacity to handle nuanced instructions. Unlike its predecessors, GPT-4 is a multimodal model capable of processing both image and text inputs. The integration of a multimodal approach extends its capabilities to extend to tasks that involve images.

- **Multimodal-CoT**: Multimodal Chain-of-Thought Reasoning [Zhang et al., 2023] incorporates text and vision into a two-stage framework. Creating a justification based on multi-modal data is the first stage. The first stage involves the rationale generation based on the given multi-modal information. This means that given text and images, the model is supposed to come up with a justification for how the two are related. The second stage of the framework is the answer inference phase. This is where the model uses the information rationale generated in the first stage to infer the correct answer to the question. In general, there are two ways to elicit the multimodal-CoT reasoning: (i) transform the inputs of different modalities into one modality and prompt LLMs to perform CoT and (ii) fine-tuning smaller models by fusing multimodal features. Multimodal-CoT has shown promising results in the answer inference stage for several tasks. We load the model that is already fine-tuned for ScienceQA from Hugging Face. The vision features of images are extracted with DETR model [Carion et al., 2020]. Since the generated rationale is publicly available, we only implement the second stage with the rationales.

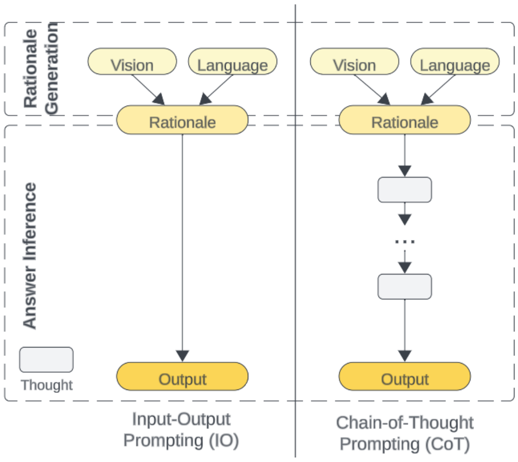


Figure: I/O prompting v/s Chain of Thought prompting for multimodal

- **Tree of Thought**: Tree of Thought is the fresh prompting method to perform deliberate decision-making by considering multiple different reasoning paths and self-evaluating choices to decide the next course of action [Yao et al., 2023]. While CoT has just one reasoning statement in each step of the answer inference, ToT generates multiple reasonings in a step. Each reasoning is evaluated by another language model, and the next reasoning will be generated based on the most feasible one.

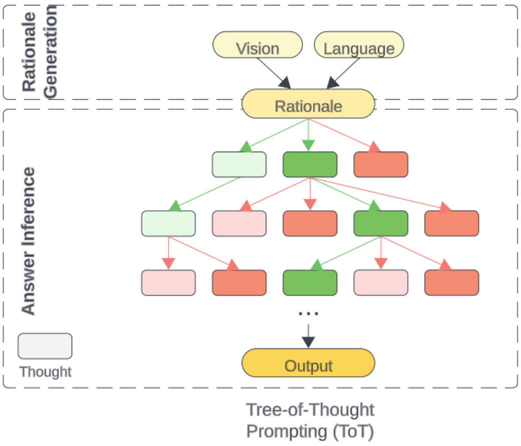


Figure: Our Tree of thought approach for Multi-modal

To make the variety in the generated reasoning, we set a higher value in the parameter of temperature that plays a key role in adjusting the balance between randomness and predictability, affecting the creativity and coherence of the generated responses. In our experiment, we use GPT-3.5 for each answer inference step and the final decision through OpenAI API. The tree has 2 layers: the first layer has 2 nodes and the second layer has 4 nodes. Hence, to identify the optimal tasks for ToT, we execute ToT in scenarios where both Multimodal-CoT and LLaVA encounter failures.

- **Tree of Thought Prompting**: The ToT method discussed above is computationally expensive because we need to run GPT-3.5 for each reasoning generation and evaluation. Recent studies [Hulbert, 2023] introduce the Tree-of-Thought Prompting method that simulates the tree-structured reasoning in a single text generation response by adding the instruction in the prompt. We also implement this method with GPT-3.5 and measure the performance in the failure cases.

### 4.2 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 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 fail. We integrate ToT prompting and multimodal approaches to investigate the potential of this fusion, with the aim of uncovering additional insights for future research.

### 4.3 Challenge

- **LLM Fine-Tuning**: We first tried and failed in both loading and fine-tuning LLaVA 13B using Hugging Face because it has a myriad of parameters that require a lot of time and storage. The 4-bit quantization method solves loading the model with less storage space, and LLM Engine https://llm-engine.scale.com helped with fine-tuning for the experiment because fine-tuning is implemented in the backend of API.

- **Open-Source LLM Usage**: While there are a lot of available open-source LLMs, we often encounter unknown errors and cannot find solutions for them. Also, the sample codes written a few years ago were not helpful because they were not written with the latest version libraries. For example, the setting for text generation through OpenAI API was recently changed so we need to read the official documentation and explore several parameters before the experiment.

## 5 Experiments

The experiment is setup in 2 different stages.

- **Stage 1**: In the initial phase of our investigation, our focus lies in evaluating the potential superiority of the Chain-of-Thought prompting approach compared to the Zero-shot prompting method on the ScienceQA dataset. Our objective is to employ various language models and analyze the outcomes, specifically observing whether the implementation of the Chain-of-Thought prompting technique leads to an improvement in the performance of generated responses to the diverse set of multimodal questions present in the dataset. This strategic assessment aims to provide valuable insights into the effectiveness of different prompting methods in optimizing model performance for the ScienceQA dataset.

- **Stage 2**: In the next stage, our objective is to apply the Tree-of-Thought prompting technique specifically to the instances where the Chain-of-Thought approach encounters challenges or failures. This initiative aims to investigate the efficacy of the Tree-of-Thought method in addressing tasks that may pose difficulties for the Chain-of-Thought prompting, thereby offering a comprehensive understanding of the relative strengths and limitations of these prompting techniques.

### 5.1 ScienceQA Dataset

Science Question Answering (ScienceQA) is a benchmark with approximately 21,000 multimodal multiple-choice questions covering diverse science topics like natural science, language science, and social science. The dataset includes annotations for answers along with corresponding lectures and explanations. This dataset is suitable for training or testing the performance of language models that are designed to mimic the human reasoning process and generate correct answers with reasoning. ScienceQA significantly differs from Visual Question Answering (VQA) [Agrawal et al., 2015] datasets in terms of difficulty, multi-modal contexts, and diversity of topics. ScienceQA mainly consists of questions with 1) image context, 2) text context, and 3) both. Image context means the questions are supposed to be answered based on a given image. Text context indicates that additional information is given with the questions that could lead to the direction of the solutions or hints [Lu et al., 2022].

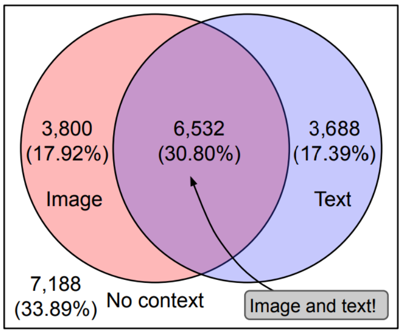
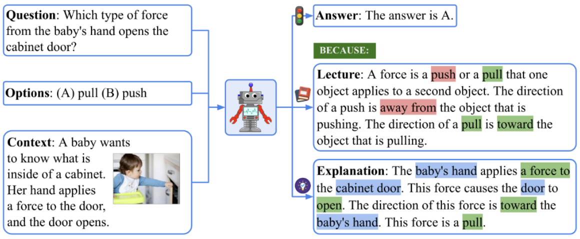


Figure 1: Question distribution with different context formats. 66.11% of the questions in ScienceQA have either an image or text context, while 30.80% have both [Lu et al., 2022]

Within this dataset, our selection criteria are confined to questions featuring both image and text inputs. This deliberate choice stems from our objective to explore the multimodal elements and assess their impact on the overall performance of prompting techniques.


**Question**: Which type of force from the baby's hand opens the cabinet door?

**Options**: (A) pull (B) push

**Context**: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.

**Answer**: The answer is A.

**Lecture**: A force is a push or a pull that one object applies to a second object. The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

**Explanation**: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

Figure 2: Example of a question in the ScienceQA dataset [Lu et al., 2022]

### 5.2 Evaluation metrics

To assess the comparative effectiveness of Chain-of-Thought prompting against Zero-shot prompting, we evaluate their accuracies through a direct comparison. Since each question in the ScienceQA dataset has a list of answers, comparing the accuracy would be sufficient and the most accurate to determine if the language model answers correctly. A correct response is defined based on the alignment between the choice selected by the Chain-of-Thought prompting and the ground truth within the ScienceQA's test set.

A parallel metric is applied in the evaluation of Tree-of-Thought, involving a comparison between the ultimate answer generated by the Tree-of-Thought and the corresponding ground truth within the test set. The objective is to look into the capability of the Tree-of-Thought to successfully address difficult tasks that may challenge the Chain-of-Thought approach.

## 6 Results

Prior to delving into the application of the Tree-of-Thought approach, our initial focus is to assess whether the Chain-of-Thought method demonstrates superiority over the Zero-shot prompting technique on the ScienceQA dataset. If the Chain-of-Thought method does not outperform the Zero-shot prompting, then exploring the application of the Tree-of-Thought on the dataset might not be justified.

### 6.1 Zero-shot versus Chain-of-Thought on Multimodals

To assess the efficacy of the Chain-of-Thought method against zero-shot prompting, we employed four distinct baselines for performance evaluation: Llava, Multimodal-CoT, GPT-3, and GPT-4. The selection of Llava and GPT-4 was motivated by their status as benchmarks with state-of-the-art performance. GPT-3 was chosen to understand how Chain-of-Thought behaves in a text input-only language model compared with the multimodal models. Additionally, Multimodal-CoT was chosen for its ability to achieve remarkable performance even when used with T5 base, which has significantly fewer parameters (220M) compared to Llava (13B), GPT-3 (176B) and GPT-4 (1.76T). This choice is particularly relevant considering the computational expense of the Tree-of-Thought implementation, which involves generating various chains of thoughts to identify the optimal answer.

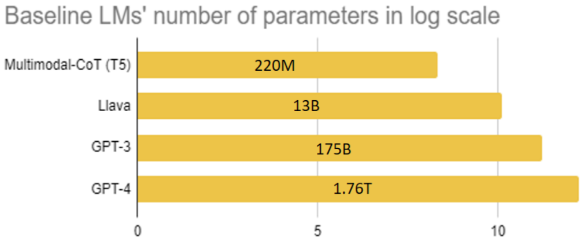


Figure 3: Baseline LM's number of parameters in log scale

The findings reveal a noteworthy enhancement in accuracy for the Chain-of-Thought implementation of Llava on the ScienceQA dataset, marking a substantial increase from 68.15% to 88.09%. In the case of GPT-3, the accuracy of Chain-of-Thought slightly surpasses that of zero-shot prompting, and the application of Chain-of-Thought prompting with GPT-4 on ScienceQA achieves an

86.54% overall accuracy on ScienceQA, marking an 11.37% improvement over the GPT-3 CoT result [Pan et al., 2023]. Notably, the accuracy of Multimodal-CoT is 84.91% [Zhang et al., 2023], comparable to that of larger language models such as Llava and GPT-4, indicating a potential domain for exploration in future research. This observation is particularly valuable given the computational intensity of the Tree-of-Thought, suggesting that employing a smaller model could contribute to enhancing the robustness of the prompting process.

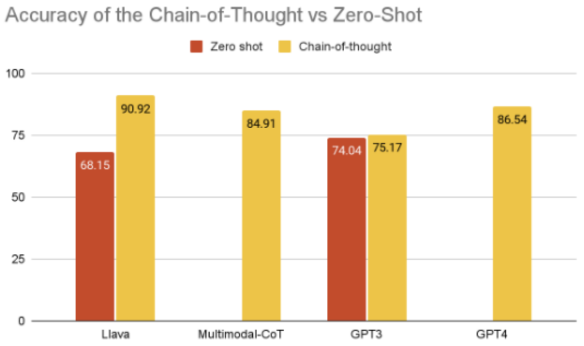


Figure 4: Zero-shot versus Chain-of-Thought on Llava, Multimodal-CoT, GPT-3 and GPT-4

### 6.2 Tree-of-Thought on Chain-of-Thought fail cases

The author of the paper "Tree of Thoughts: Deliberate Problem Solving with Large Language Models" suggested that for numerous tasks where GPT-4 already demonstrates proficiency, extensive searches like ToT might not be necessary [Yao et al., 2023]. Rather than focusing on whether Tree-of-Thought outperforms Chain-of-Thought, our emphasis was on exploring cases where Tree-of-Thought can overcome the limitations of Chain-of-Thought.

To achieve this, we selectively filtered out instances where both of our top-performing models, Llava and Multimodal-CoT, failed. This filtration resulted in a condensed dataset of 60 questions from the original 6,532 multimodal questions. Subsequently, we employed the Tree-of-Thought architecture to generate outputs for these 60 questions. Additionally, we randomly chose 60 cases from the dataset where both Llava and Multimodal-CoT provided correct answers to ensure a balanced output.

Our experimental findings indicate that the Tree-of-Thought approach demonstrated effectiveness in generating responses. Specifically, it successfully provided answers for 52 out of 60 questions in cases where both CoT models (Llava-CoT and Multimodal-CoT) succeeded. Moreover, the Tree-of-Thought approach exhibited competence by responding to 29 out of the 60 questions when both CoT models failed, resulting in success rates of 86.7% and 48.33%, respectively. Notably, these questions were part of the failure instances for both Llava-CoT and Multimodal-CoT.

Crucially, when evaluating these success rates against a random chance scenario, wherein a language model randomly selects an answer without any rationalization, we observed a substantial improvement of 24.01%. This suggests that the Tree-of-Thought approach significantly enhances the model's ability to provide accurate responses, particularly in challenging scenarios where both CoT models falter.

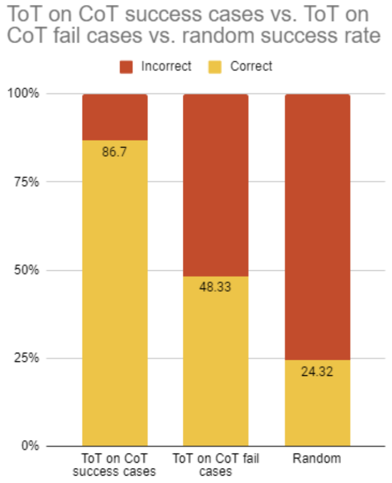


Figure 5: Tree-of-Thought versus zero-shot success rates

## 7 Analysis

The following analysis will investigate how ToT works and discuss contribution factors and limitations.

### 7.1 General performance of Tree-of-Thought on ScienceQA

As previously noted, the Tree-of-Thought exhibited improved accuracy in tasks where the Chain-of-Thought method faced challenges. This improvement suggests that the Tree-of-Thought prompting mechanism has the potential to enhance the performance of the language model in scenarios where the Chain-of-Thought approach encounters difficulties.

| Answer | ToT Category | Percentage % |
|---|---|---|
| Correct | ToT is correct | 86.7% |
| | ToT is incorrect | 13.3% |
| | ToT is correct | 48.33% |
| Incorrect | ToT is incorrect | 51.67% |

Table 1: Tree-of-Thought performance on 60 correct and incorrect samples

Simultaneously, when executing the Tree-of-Thought on scenarios where Chain-of-Thought models achieved success, the outcomes exhibited comparable performance to those produced by the Chain-of-Thought itself. The implemented Tree-of-Thought model, built upon the GPT-3.5 architecture, demonstrated a noteworthy accuracy surpassing that of GPT-4 (86.% compared to 86.54%), despite the latter having a significantly greater number of parameters. This highlights the effectiveness of the Tree-of-Thought approach in achieving competitive accuracy levels, even in comparison to models with a higher parameter count such as GPT-4.

### 7.2 Performance of Tree-of-Thought on different particular categories

To pinpoint the particular task domains where the Tree-of-Thought excels or fails, our initial investigation involved an examination of diverse question categories. We assessed the success rates for each of the three distinct categories within the dataset: Natural Science, Language Science, and Social Science.

Examining Table 2 reveals that the Tree-of-Thought prompting implementation exhibits notably high accuracy in the Social Science category, reaching 88.37%, with 38 correct predictions out of 43 questions in that domain. Conversely, we observe perfect accuracy (100%) in the Language Science category. However, due to the limited number of Language Science questions used in comparison to the other two categories, further investigation is warranted to conclusively ascertain the Tree-of-Thought's excellence in this specific category. Notably, the accuracy in the Natural Science category is comparatively lower than the Chain-of-Thought baseline models, particularly in Biology and Geography. This discrepancy is attributed to the intricate and detailed nature of the images associated with these questions, leading to model confusion and generating incorrect answers in numerous instances.

Owing to computational constraints, we opted not to calculate the average accuracy for the Multimodal Tree-of-Thought, as it was not executed on the entire dataset. Consequently, drawing a meaningful comparison between the average accuracy of the Multimodal Tree-of-Thought and the Chain-of-Thought is not applicable in this context.

### 7.3 Hallucinations in Tree-of-Thought

Similar to the Chain-of-Thought implementation, there is some hallucination for the Tree-of-Thought prompting.

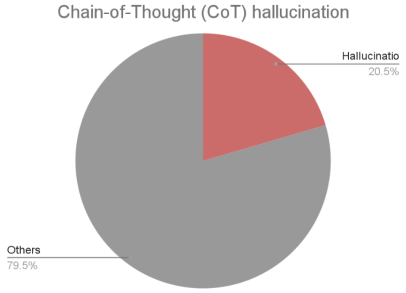


Figure 6: Hallucination on CoT

In some cases, the models would fabricate supporting evidence or introduce incorrect "facts" as justification for their answers. While the overall reasoning chain appears convincing, these false details undermine the credibility. Hallucination poses a significant reliability concern, as it provides a false illusion of rationality.

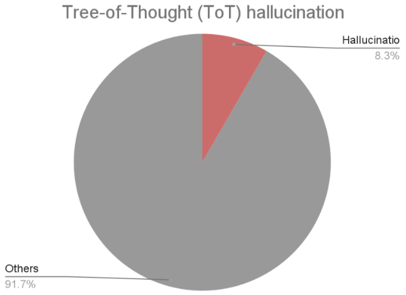


Figure 7: Hallucination on ToT

We found certain types of questions appear more susceptible to eliciting hallucinations than others. In particular, questions requiring temporal reasoning, specialized scientific knowledge, or multifaceted logical deduction saw higher rates of hallucination in model responses. This aligns with the findings that neural networks struggle with precise logical formalisms.

Nevertheless, when compared to the Chain-of-Thought (CoT) hallucination rate, there is a no-

| Model | Size | NAT % | SOC % | LAN % | AVG % |
|---|---|---|---|---|---|
| Human | - | 90.23% | 84.97% | 87.48% | 88.40% |
| GPT-3 | 175B | 74.64% | 69.74% | 76.00% | 74.04% |
| GPT-3 w/ CoT | 175B | 75.55% | 70.87% | 78.09% | 75.17% |
| Multimodal-CoT | 223M | 87.52% | 77.17% | 85.82% | 84.91% |
| GPT-4 | 1.7T | 83.99% | 85.48% | 72.44% | 86.54% |
| Multimodal-ToT | 175B | 76.74% | 88.37% | 100.00% | - |

Table 2: The accuracy of the CoT and ToT Multimodal models on Natural Science, Language Science, and Social Science categories

table decrease in the occurrence of hallucination cases, dropping from 20.5% to 8.3%. Hallucination, in this context, refers to instances where the multimodal model either fabricates information not provided or selects an option not present in the given list of choices. This reduction in hallucination cases is a promising indication of improved accuracy and reliability in the multimodal model.

**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

## 8 Conclusion

In the course of our investigation, we explored a few additional concepts that did not directly lead to improved performance but may offer promise for future work.

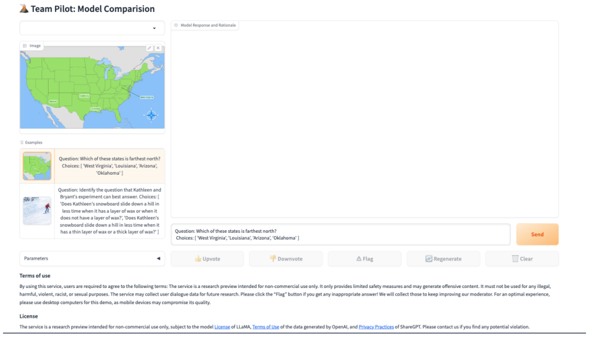
One idea we prototyped was supplementing the textual reasoning chains with visual diagrams. The goal was to provide graphical representations of the relationships and logic described verbally in the CoT and ToT prompts. Initial implementation showed no significant difference, but further research into optimal diagram integration could be impactful.

Additionally, we attempted a hybrid approach of using CoT reasoning to narrow down the answer choices for a ToT model. However, we encountered difficulties coordinating the output of one model to reliably prime another. This cascade prompting idea still holds potential if the chaining between models can be improved.

Regarding problems encountered, we found that coherence and consistency occasionally broke down in lengthy reasoning chains for both CoT and ToT. Grounding key entities and concepts may help maintain context across long inferences. We also observed ToT performance gains tapering off after 3-4 reasoning branches. Determining optimal tree shape and depth is an area for further tuning. In terms of noteworthy results, we were surprised CoT proved nearly as effective as ToT in certain complex spatial reasoning tasks but fell behind significantly in dynamic systems questions. Better understanding these performance gaps could reveal differentiation points to guide appropriate prompting selection based on context and question type.

### 8.1 Challenges

Our work faced several constraints that were considered before interpreting the results. First, our data analysis was confined to 60 correct and 60 incorrect predictions. This limitation arose from the costly nature of the Tree of Thoughts (ToT) prompting method. This method involved multiple layers and branches, each requiring a separate prompt generator to delve deeper. Consequently, we had to limit our analysis to a randomly selected subset. Second, in our ToT multimodal approach, we only employed the GPT 3.5 model instead of a more advanced GPT-4 vision model, owing to time and budget restrictions. Therefore, we couldn't fully leverage our model's capabilities to benchmark it against state-of-the-art (SOTA) models. Lastly, time constraints prevented us from compressing our model into a smaller, open-source architecture, compelling us to depend on an API-based model for results. We hypothesize that using the T5 architecture as the underlying model could significantly reduce size, potentially yielding results comparable to SOTA models, while also being more cost- and time-efficient. This approach offers a promising avenue for future research.



Team Pilot: Model Comparision

Question: Which of these states is farthest north? Choices: [ 'West Virginia', 'Louisiana', 'Arizona', 'Oklahoma' ]

Question: Identify the question that Kathleen and Bryant's experiment can best answer. Choices: [ 'Does Kathleen's snowboard slide down a hill in less time when it has a layer of wax or when it does not have a layer of wax?', 'Does Kathleen's snowboard slide down a hill in less time when it has a thin layer of wax or a thick layer of wax?' ]

Question: Which of these states is farthest north?
Choices: [ 'West Virginia', 'Louisiana', 'Arizona', 'Oklahoma' ]

**Terms of use**

By using this service, users are required to agree to the following terms: The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. The service may collect user dialogue data for future research. Please click the "Flag" button if you get any inappropriate answer! We will collect those to keep improving our moderator. For an optimal experience, please use desktop computers for this demo, as mobile devices may compromise its quality.

**License**

The service is a research preview intended for non-commercial use only, subject to the model License of LLaMA, Terms of Use of the data generated by OpenAI, and Privacy Practices of ShareGPT. Please contact us if you find any potential violation.

Figure 8: A gradio web interface to compare the performance of our models

## 8.2 Future Work

We also started working on make a visualisation web interface tool using which can compare the performance of our model with respect to the SOTA model. Hopefully, we will be get it fully functioning in the coming days.

## 8.3 Replicability

Our project focuses on extending the conventional Tree of Thought technique to multimodal data to create multiple lines of reasoning within a single prompt at varying tree depths. This process requires substantial pre-processing and post-processing to construct the optimal prompt format from the available data. Utilizing the OPENAI API could minimize the necessity of training a model from the ground up, aside from the pre and post-processing stages. To streamline the model and conserve resources, the T5 architecture, a reasonably small language model, could serve as the backbone of training from scratch. It offers the flexibility to fine-tune specific downstream tasks as needed.

Regarding the dataset, we did not augment our dataset with any additional annotations or extra features that future researchers should be aware of to adapt our approach for their research purposes. But we did consider improving the model by extracting the vision features and feeding them as input to our model instead of using an image captioning model to extract the captions from the images. If you want to train the model from scratch using the T5 as a backbone, you can consider using these vision features instead of the image captions, as they greatly improve the efficiency of the model.

## 8.4 Ethics

Our research aims to advance techniques for improving reasoning abilities in AI systems. However, as with any technology, there are ethical considerations regarding potential misuse.

We recognize language models have exhibited harmful biases that could be perpetuated or exacerbated through improper development or deployment. Throughout this project, we will monitor our models for signs of unfair bias or representation issues. We will also deliberately design our methodology, datasets, and experiments to proactively avoid introducing or amplifying biases.

The intended use case for our research is to enable AI assistants to better understand and respond to human requests and questions. We believe enhancing logical reasoning in this narrow capacity for helpfulness has more benefit than risk. However, we acknowledge large language models have the potential for misuse outside of the scope we intend.

While our core investigation is on model development rather than deployment, we will discuss the ethical implications of potential real-world applications in our analysis. As students exploring an emerging field, we aim to contribute positively while identifying areas needing further ethical guidelines. We welcome feedback from the research community on responsible AI practices.

Overall, we will conduct this research with care and thoughtfulness regarding wider societal impacts, focusing solely on furthering techniques for safe and beneficial AI development. We believe stronger reasoning abilities can and should be pursued to reduce biases rather than amplify them.

## References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey

Zagoruyko. 2020. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872.

Scale The Data Platform for AI. 2023. LLM Engine.

Dave Hulbert. 2023. Tree of knowledge: Tok aka tree of knowledge dataset for large language models llm. `https://github.com/dave1010/tree-of-thought-prompting`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Lu Pan, Baolin Peng, Hao Cheng, Michel Galley, Chang Kai-Wei, Ying Nian Wu, Zhu Song-Chun, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv.org*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models.