Exploring the Frontiers of Large Language Models

Baktash Ansari and Pooria Rahimi

LLMs

- Large language models (LLMs) are a type of artificial intelligence designed to understand and generate human language. They are trained on vast amounts of text data and can perform a wide range of language-related tasks.
- LLMs can be used for tasks such as translation, summarization, question answering, and content generation. Their ability to understand context and generate coherent text makes them valuable in many applications.

(Image is genereted by Google Imagen)

How Can We Improve the Performance of LLMs on Specific Tasks?

The Power of Fine-Tuning

Adapting to Specialized Tasks

Fine-tuning allows LLMs to specialize in particular applications by further training on domain-specific data, enhancing their performance and accuracy.

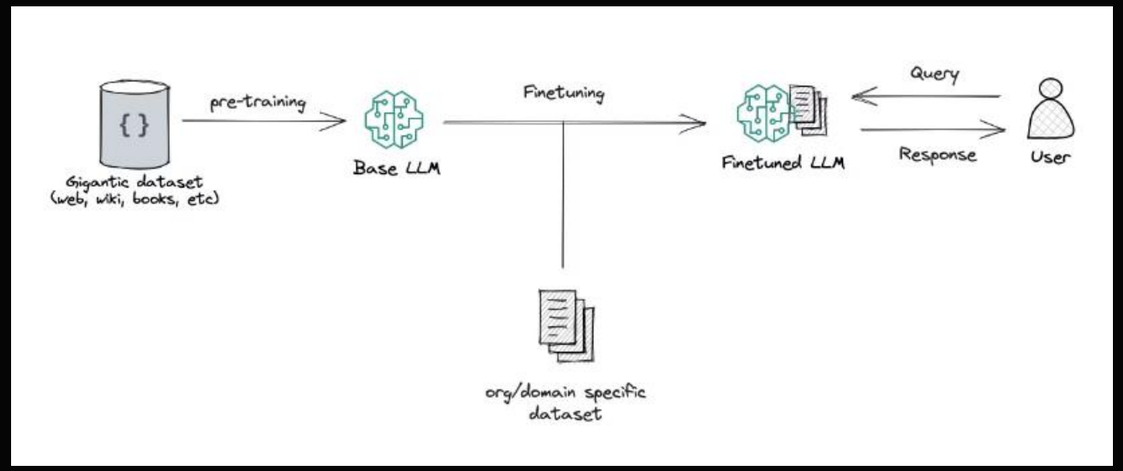
Preserving Core Capabilities

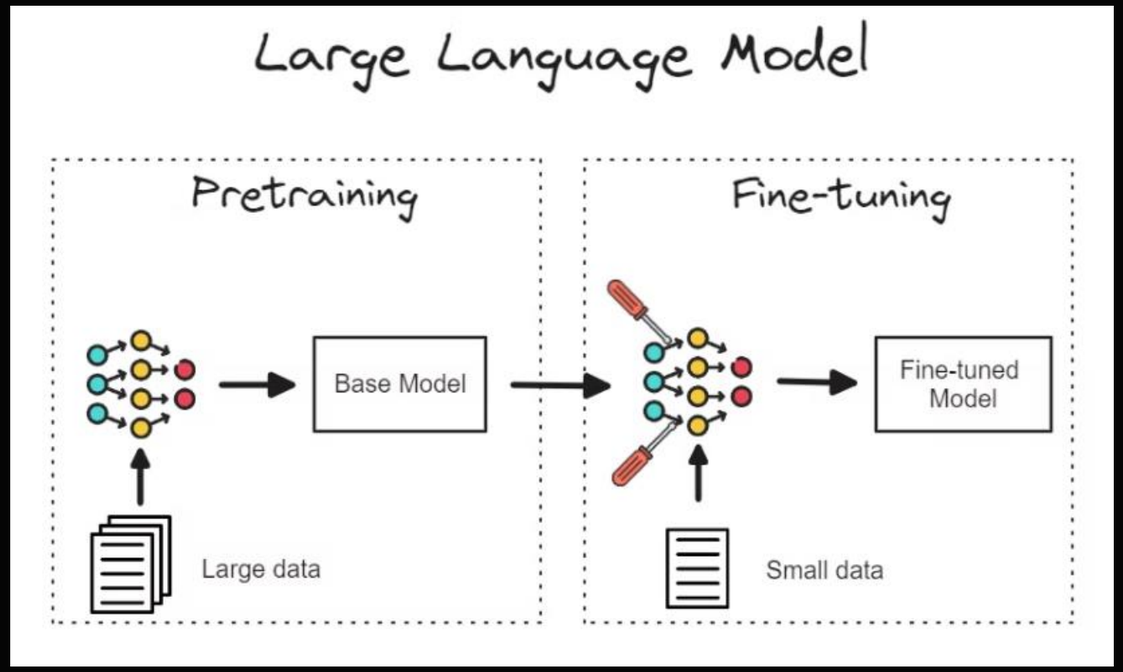
The process of fine-tuning builds upon the rich, general knowledge acquired during pre-training, ensuring the model retains its fundamental language understanding abilities.

Efficiency and Agility

Fine-tuning is a cost-effective and versatile approach, enabling LLMs to be quickly adapted to new tasks or industries without the need for complete retraining.

How Fine-Tuning works?





What are the Problems?

- **Large Number of Parameters**
  - Requires immense computational power and memory
  - Difficult to manage with limited hardware resources
- **Model Size**
  - Necessitates high-performance GPUs or TPUs
  - Prohibitively expensive for many organizations
- **Limited Hardware Resources**
  - Strains available resources
  - Delays deployment due to significant time requirements
- **Time Constraints**
  - Fine-tuning extensive models is time-consuming
  - Further delays optimization for specific tasks

(Image is generated by SDXL-Lightning)

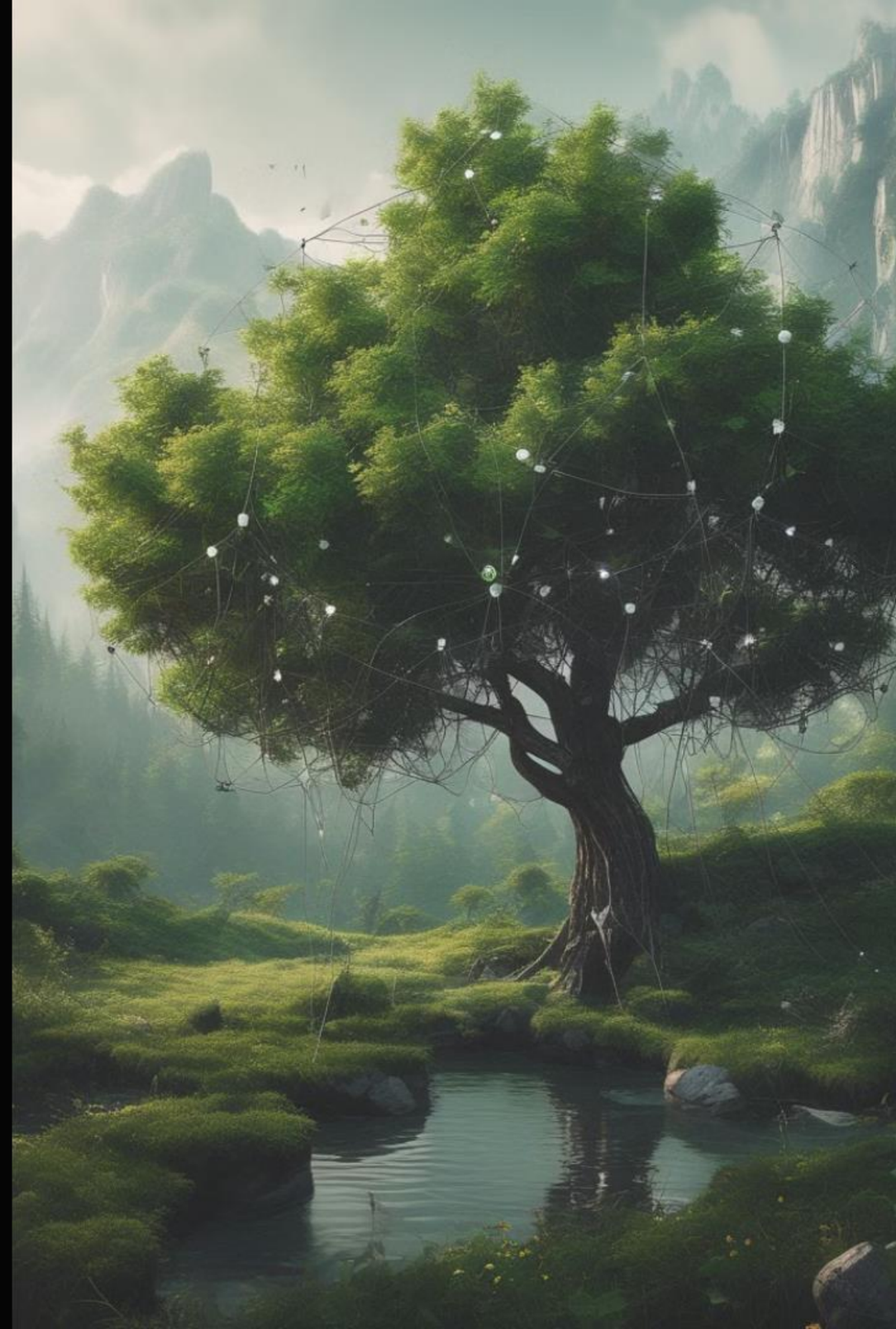
Prompt Engineering

- Designing prompts to guide model output
- Uses natural language instructions

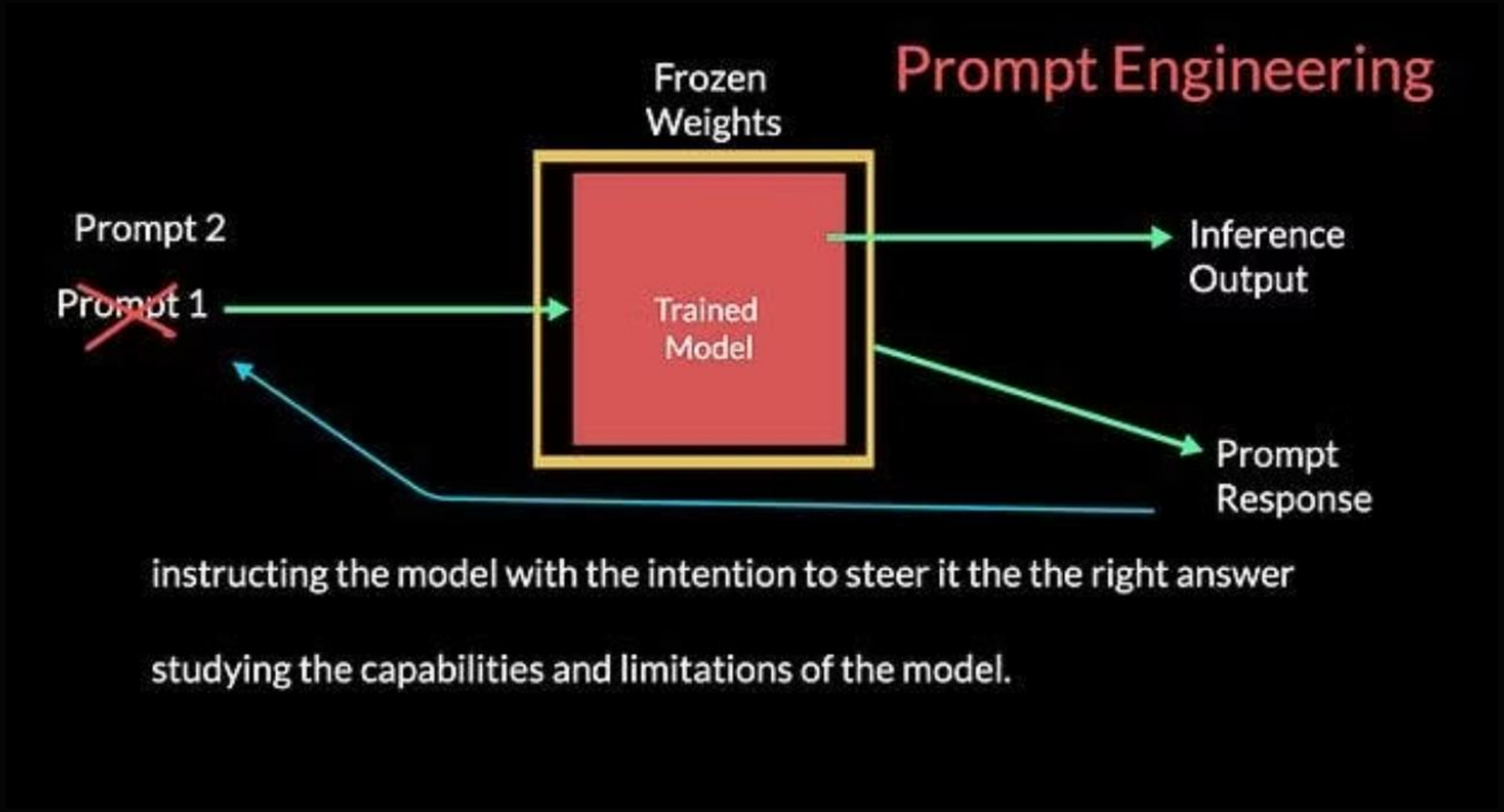
Benefits Over Fine-Tuning

- **Lower Computational Requirements**
  - No need for high-performance hardware
- **Time Efficiency**
  - Quick implementation
  - Immediate adjustments
- **Cost-Effective**
  - No expensive GPUs or TPUs needed
- **Flexibility**
  - Easily adaptable to various tasks
  - Allows rapid experimentation

(Image is generated by SDXL-Lightning)



How it works?


Prompt Engineering
Frozen
Weights
Prompt 2
Prompt 1
Trained
Model
Inference
Output
Prompt
Response
instructing the model with the intention to steer it the the right answer
studying the capabilities and limitations of the model.

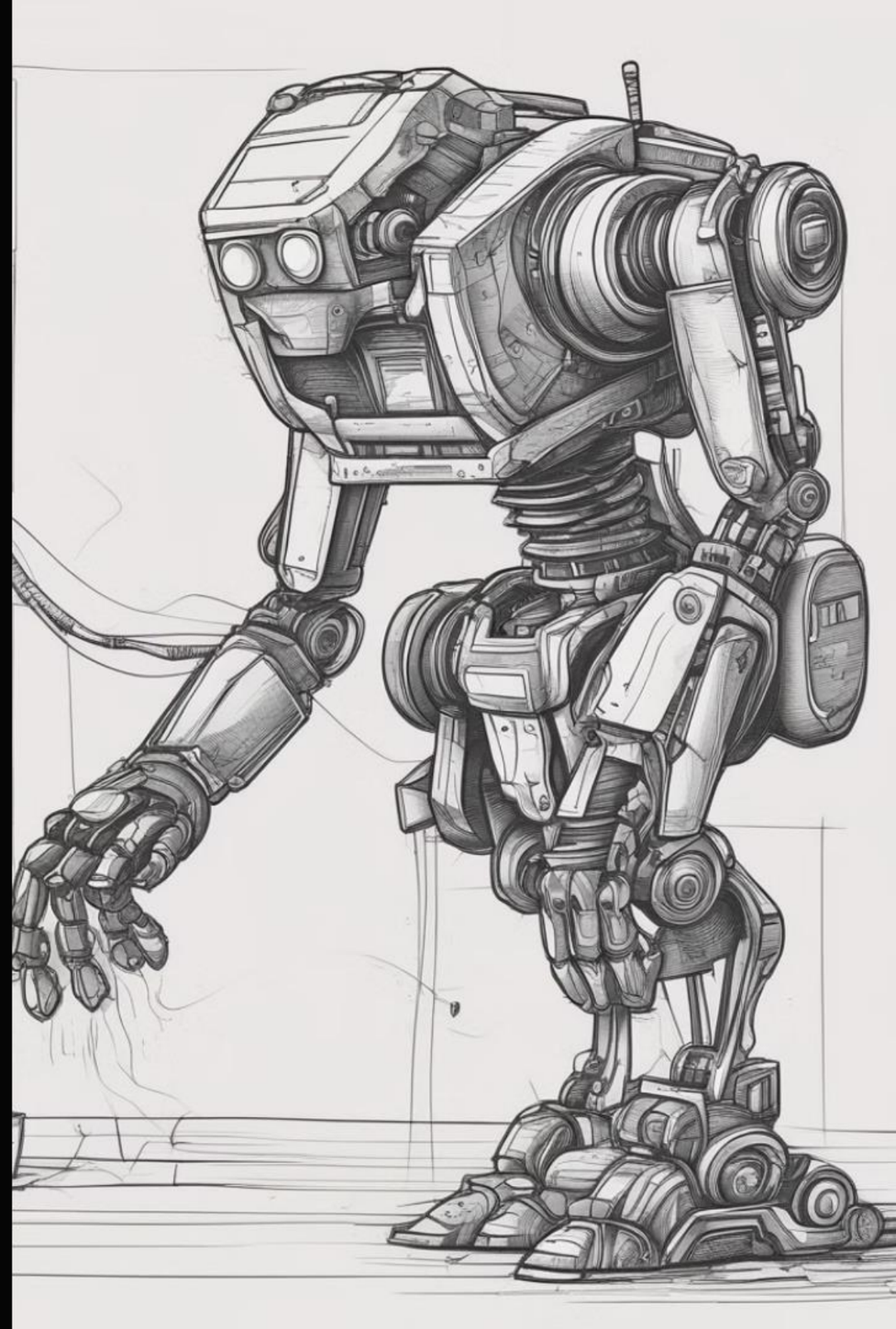
Prompting Problems

- **Ambiguity**: Vague prompts lead to irrelevant responses.
- **Bias**: Prompts can trigger biased outputs.
- **Context Understanding**: Incomplete or incorrect answers due to lack of context.
- **Length Limits**: Responses may be truncated if token limits are exceeded.
- **Repetition**: Outputs can be verbose or repetitive.
- **Misinterpretation**: Subtle wording changes cause different results.
- **Inconsistency**: Same prompt yields different results.
- **Overfitting**: Excessive reliance on examples limits generalization.
- **Complexity Issues**: Complex prompts may lower response quality.
- **Specificity Challenges**: Difficulty in generating precise content without detailed guidance.

Techniques to Address Prompting Problems

- **Chain of Thought**: Breaks down complex prompts into simpler, sequential steps to improve comprehension and output quality.
- **Zero-Shot**: Directly prompts the model without any prior examples, relying on its pre-trained knowledge.
- **Few-Shot**: Provides a few examples within the prompt to guide the model's responses and improve accuracy.
- **Multi-Agent Debate**: Engages multiple models to debate or discuss a prompt, enhancing the quality and robustness of the final output.

(Image is generated by SDXL-Lightning)



Enhancing Reasoning with Chain-of-Thought Prompting

1

Problem Decomposition

Chain-of-thought prompting guides the model to break down complex tasks into a sequence of logical steps, mirroring human problem-solving strategies.

2

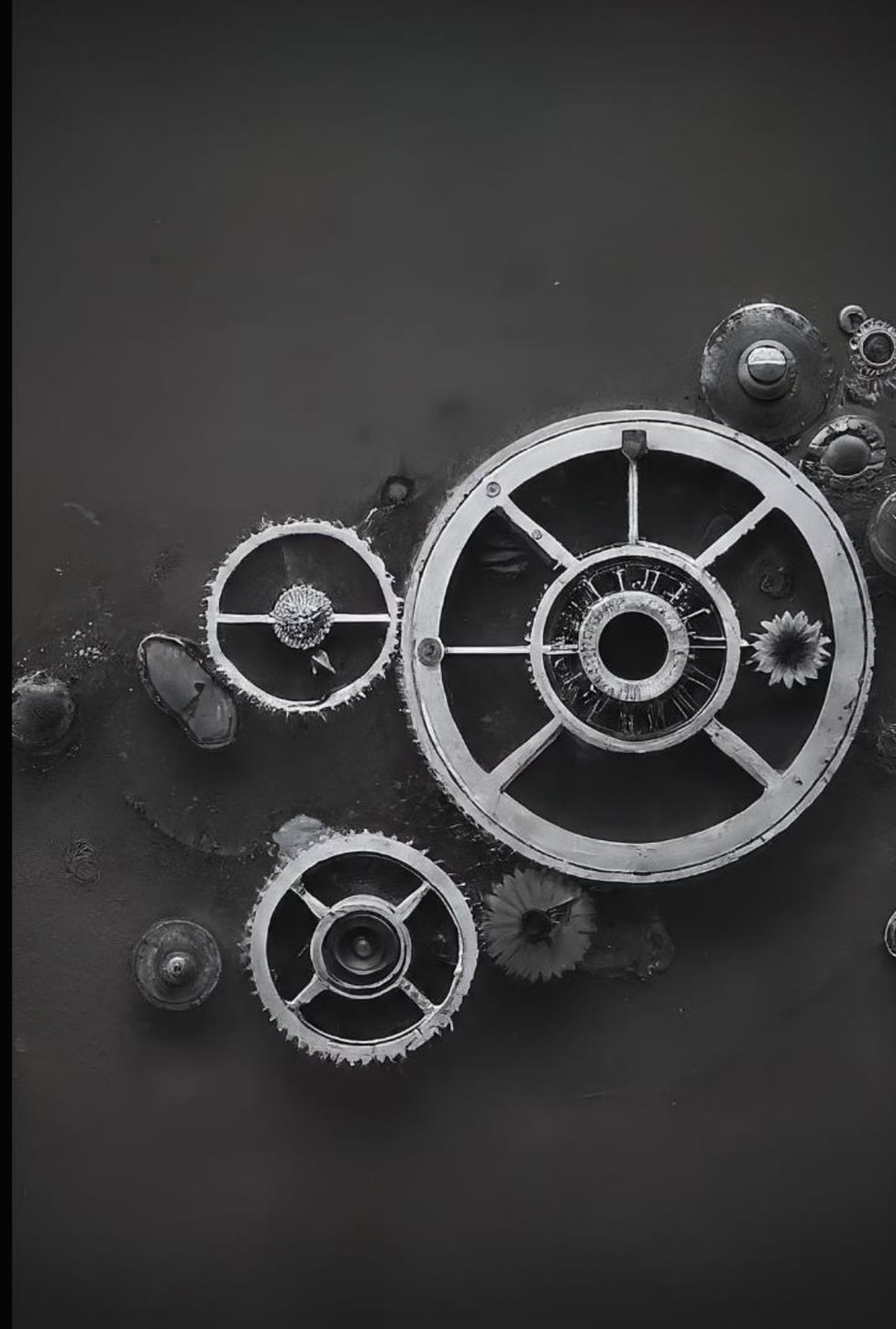
Intermediate Reasoning

By generating and evaluating these intermediate steps, the model can better understand the problem and arrive at more accurate solutions.

3

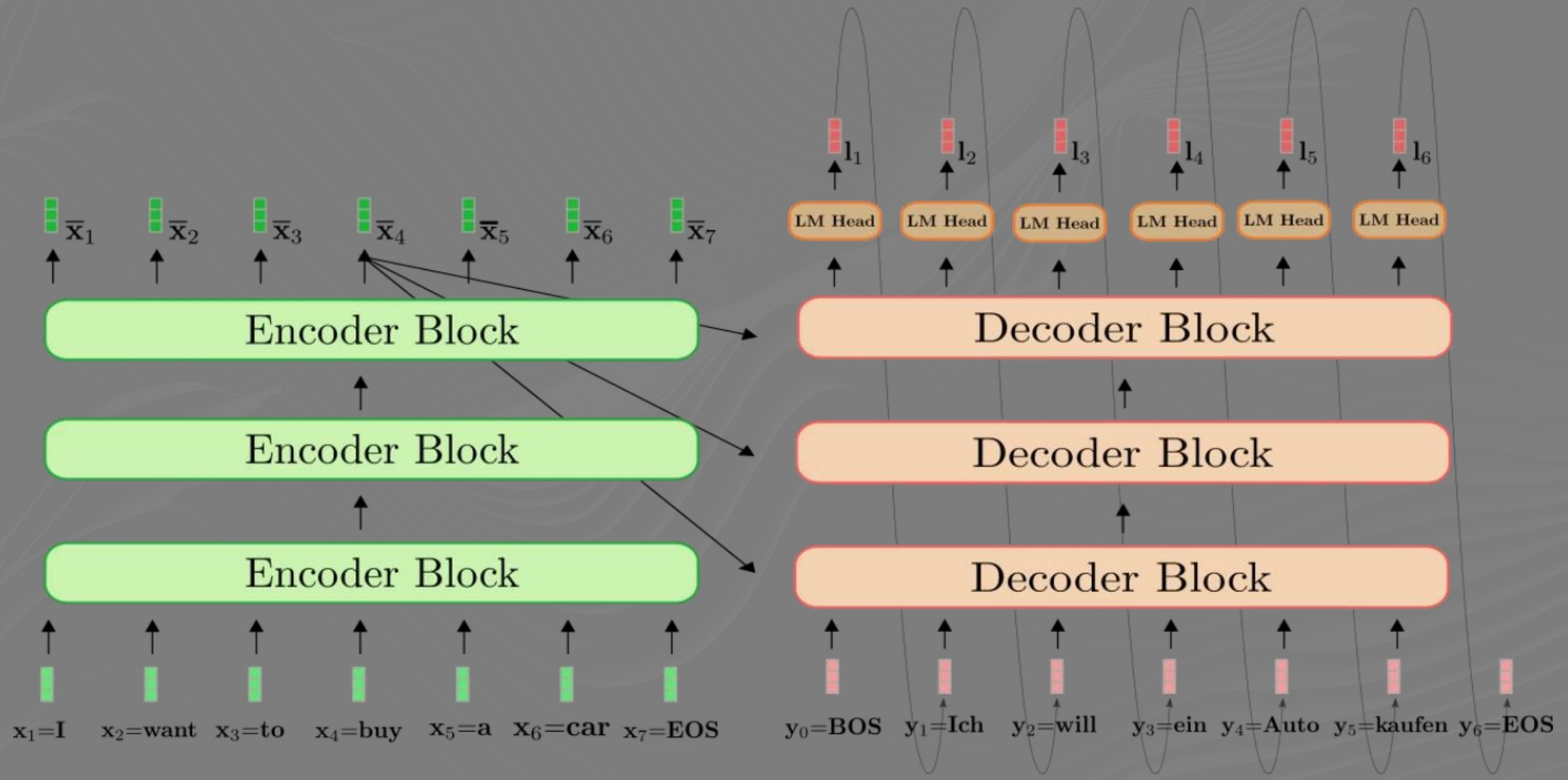
Coherent Explanations

The step-by-step approach enables the model to provide clear, explainable justifications for its decisions, enhancing transparency and trust.



(Image is genereted by Google Imagen)

Abstract transformers' architecture



Apply COT on GPT3.5

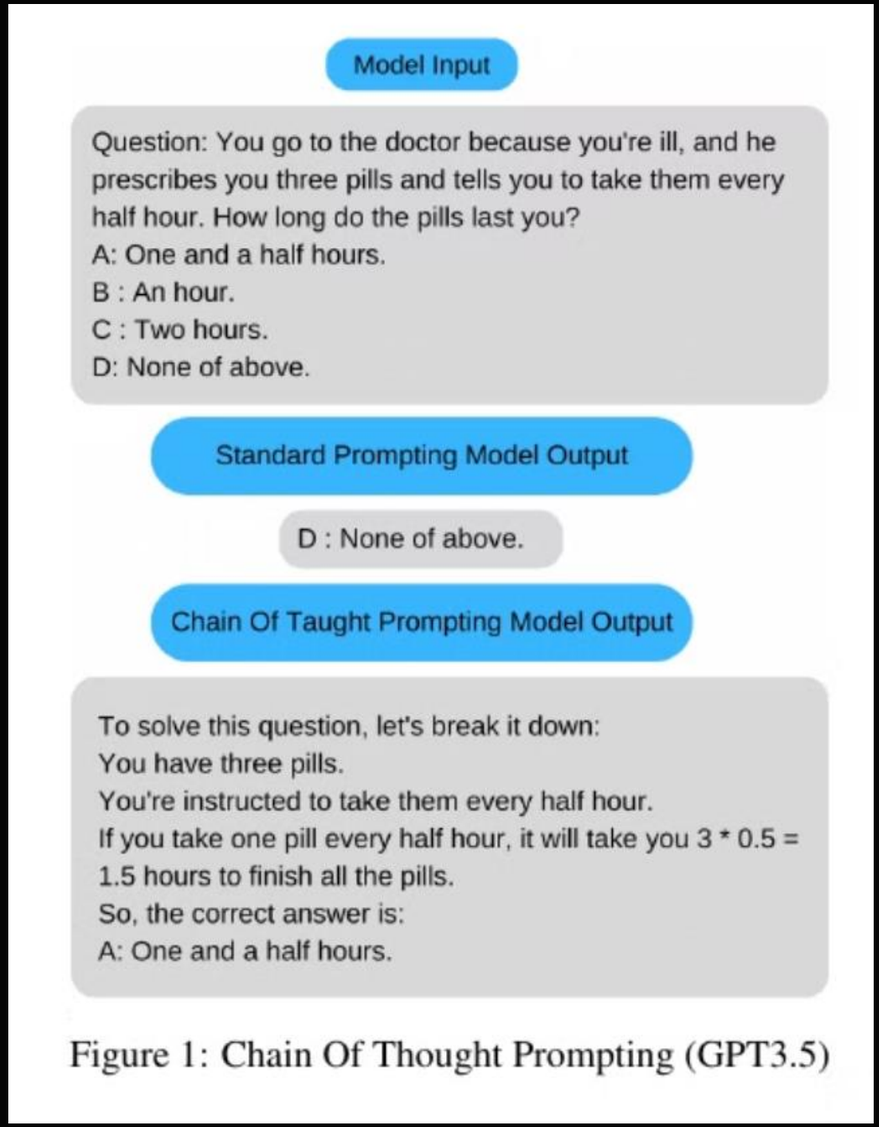


Figure 1: Chain Of Thought Prompting (GPT3.5)

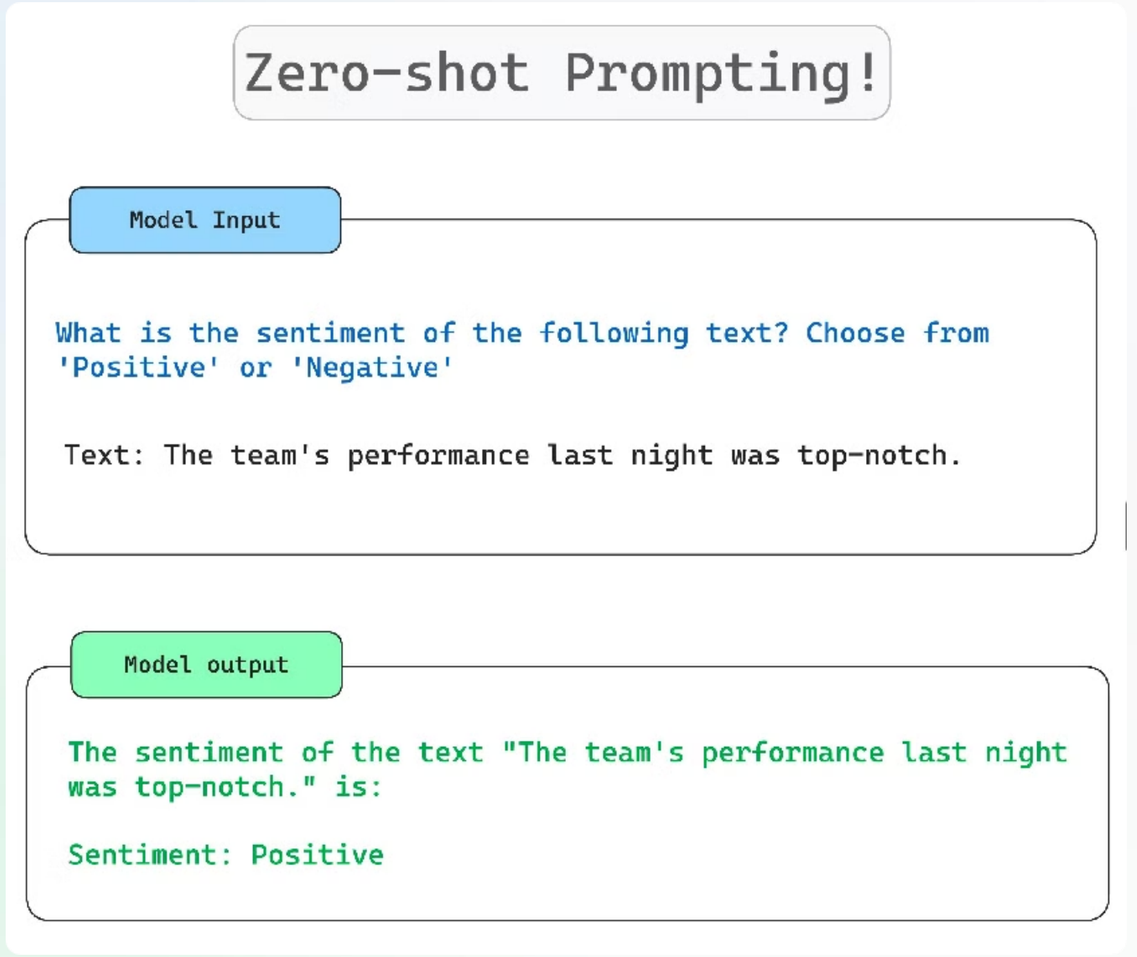
Zero-Shot Prompting

Advantages:

- Requires no additional examples, making it quick and easy to set up.
- Leverages the model's pre-trained knowledge for a broad range of tasks.
- Efficient use of tokens since no examples are needed.

Disadvantages:

- May lead to less accurate or relevant responses due to lack of context.
- Higher risk of misunderstanding the prompt's intent.
- Limited control over the response style and format.



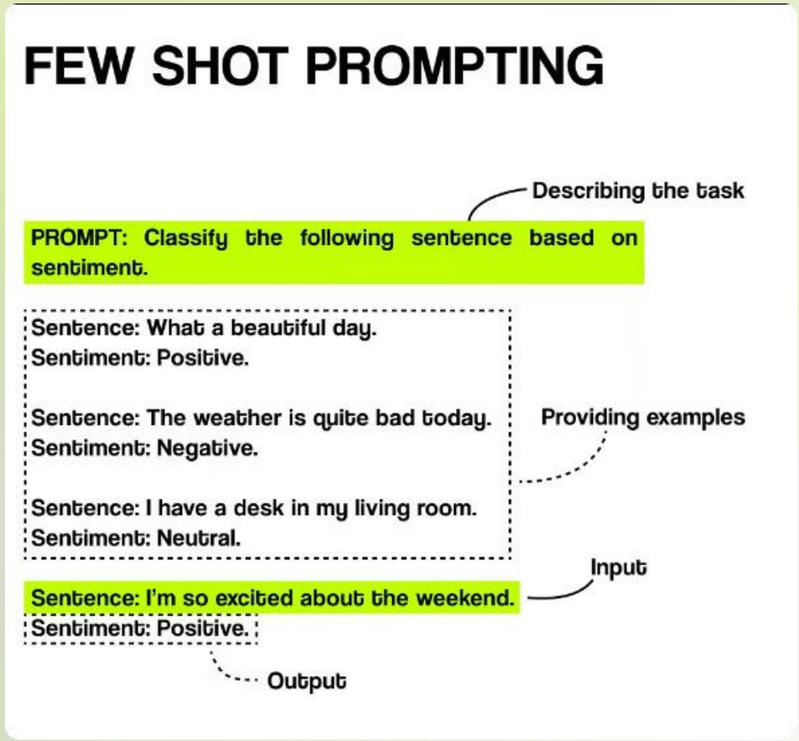
Few-Shot Prompting

Advantages:

- Provides context and guidance, leading to more accurate responses.
- Helps the model understand the expected format and style.
- Useful for specialized tasks where specific examples improve performance.

Disadvantages:

- Requires crafting multiple example prompts, which can be time-consuming.
- May still produce inconsistent results if examples are not well-chosen.
- Uses more tokens, which can be a limitation for length-restricted prompts.



Multi-Agents Debate

Advantages:

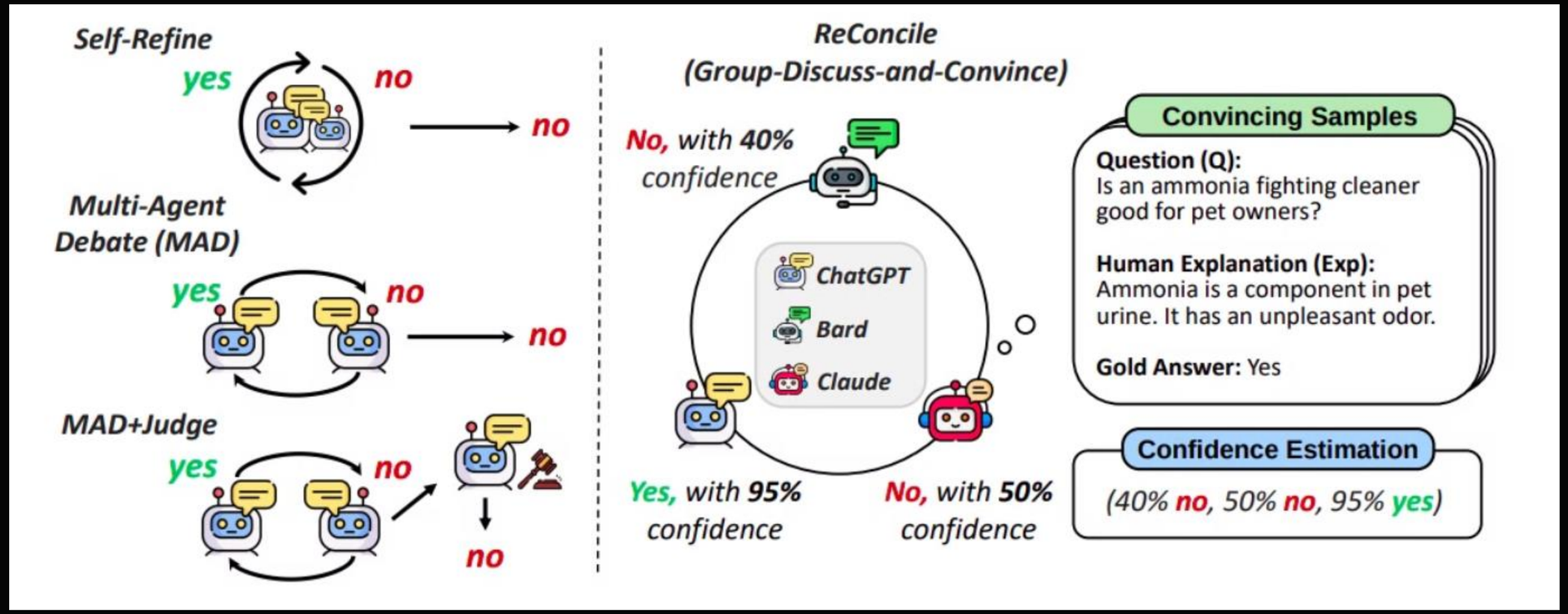
- Engages multiple perspectives, leading to more robust and well-rounded outputs.
- Can highlight and address conflicting interpretations and biases.
- Enhances the depth and quality of the response through iterative discussion.

Disadvantages:

- More complex to implement and manage multiple models.
- Higher computational cost and resource usage.
- Potential for increased response time due to the debate process.

(Image is generated by SDXL-Lightning)

Multi-Agent Techniques



Thank You for Your Attention!

Any Questions?

Resources:

- https://arxiv.org/pdf/2309.13007 (Reconcile)
- https://arxiv.org/abs/2201.11903 (COT)
- https://arxiv.org/abs/2308.10783
- BAMO at SemEval-2024 Task 9: BRAINTEASER: A Novel Task Defying Common Sense (link)
- Google Imagen (Generated AI Images)
- SDXL-Lightning (Generated AI Images)