Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei **Xuezhi Wang** **Dale Schuurmans** **Maarten Bosma**

Brian Ichter **Fei Xia** **Ed H. Chi** **Quoc V. Le** **Denny Zhou**

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Abstract

- Chain of thought
  - a series of intermediate reasoning steps
  - improves the ability to perform complex reasoning
  - chain-of-thought prompting
  - state-of-the-art accuracy

Introduction

- Scaling up model size alone has not proved sufficient for achieving high performance
- Unlock the reasoning ability of LLM
  - generating natural language rationales that lead to the final answer
  - in-context few-shot learning via prompting

Introduction

- Chain-of-thought prompting
  - few-shot prompting
  - chain of thought: a series of intermediate natural language reasoning steps
  - prompting only approach
    - not require a large training dataset
    - single model checkpoint can perform many tasks without loss of generality
- Evaluations
  - SOTA performance on GSM8K benchmark

Introduction

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

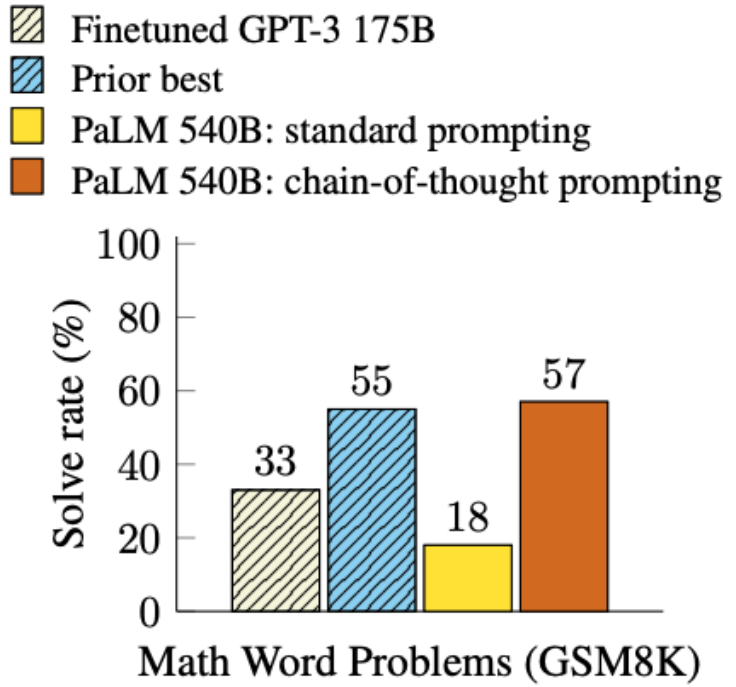
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓



Chain-of-Thought Prompting

- It is typical to decompose the problem into intermediate steps and solve a multi-step ma th word problem

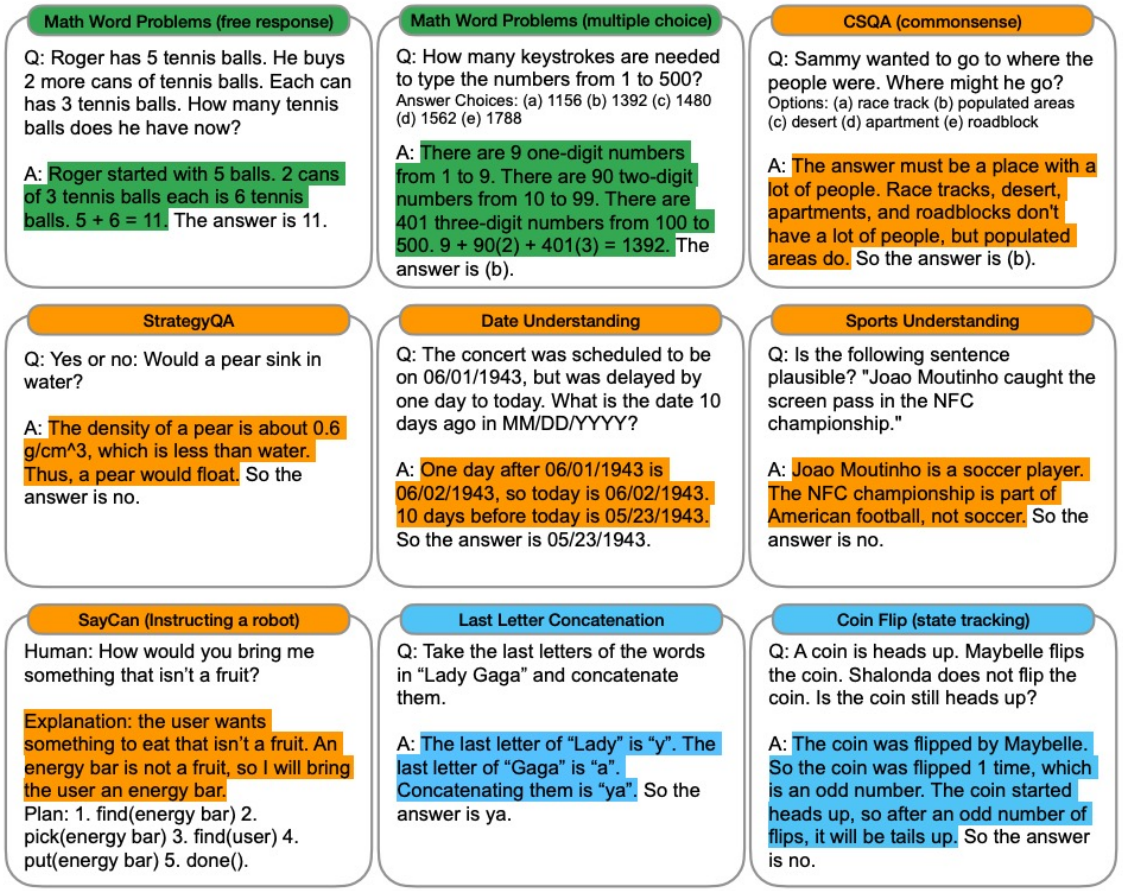
Chain-of-Thought Prompting

- Goal: to endow LM with ability to generate a similar chain of thought
- Chain of Thought
  - allows models to decompose multi-step problems into intermediate steps
  - provides an interpretable window into the behavior of the model
  - is potentially applicable to any task that humans can solve via language

Chain-of-Thought Prompting

- Empirical experiments
  - Arithmetic reasoning
  - Commonsense reasoning
  - Symbolic reasoning

Experiments

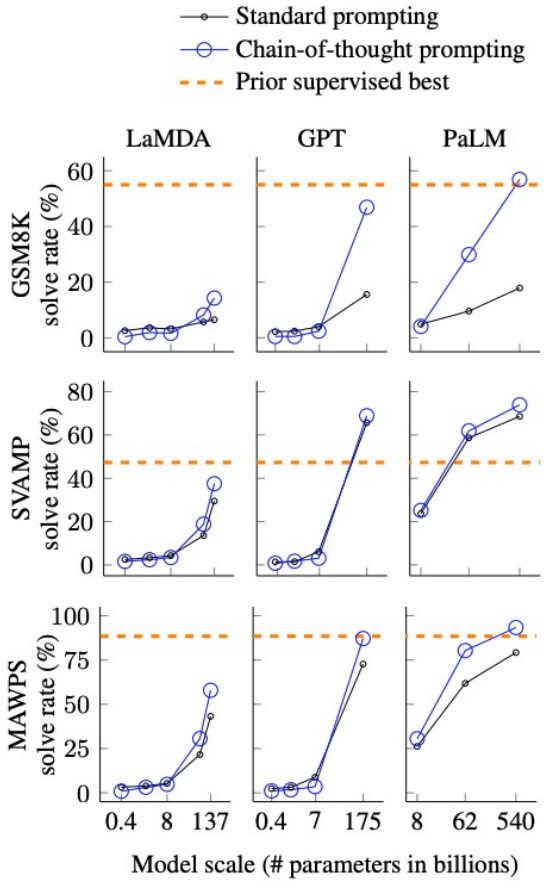


Arithmetic Reasoning

- Though simple for humans, arithmetic reasoning is a task where LM often struggle
- Standard prompting
  - few-shot prompting with input-output pairs
- Chain-of-thoughts
  - augment each exemplar in few-shot prompting with a chain of thought for an associated answer
    - single set of 8 chain of thought exemplars for all benchmarks except AQuA
    - four exemplars and solutions for AQuA (multiple choice)

Arithmetic Reasoning

- An emergent ability of model scale
  - not positively impact performance for small models
  - small models produced fluent but illogical chains of thought
- Larger performance gains for more-complicated problems
  - For GSM8K, performance more than doubled
- Compares favorably to prior state of the art
  - which finetunes a task-specific model on a labeled training dataset



Arithmetic Reasoning

- Examination of model-generated thoughts
  - Of 50 random examples
  - correct final answer
    - all of the generated thoughts were correct
  - wrong final answer
    - 46% were almost correct, barring minor mistakes
    - the other 64% had major error in semantic understanding or coherence

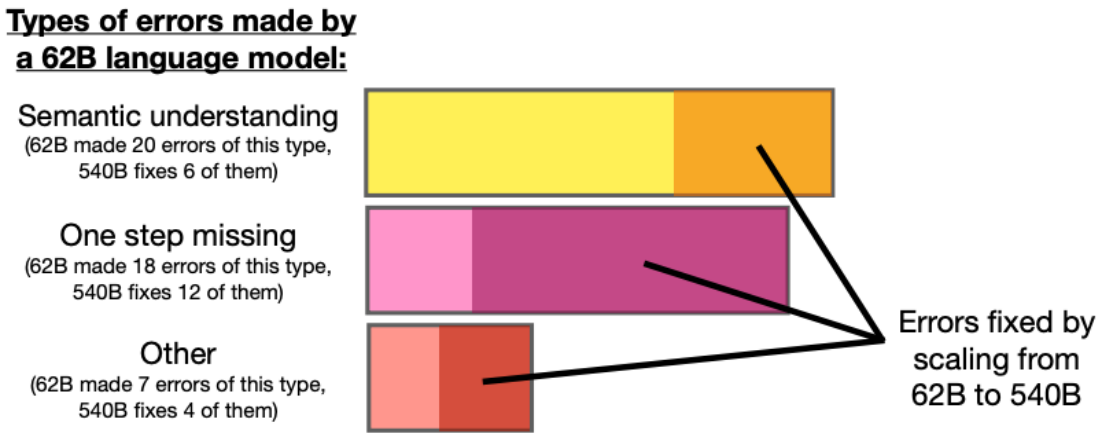
QUESTION: Dan plants 3 rose bushes. Each rose bush has 25 roses. Each rose has 8 thorns. How many thorns are there total?

MODEL ANSWER (INCORRECT; CALCULATOR ERROR ONLY): Dan plants 3 rose bushes. Each rose bush has 25 roses. Each rose has 8 thorns. So 3 x 25 x 8 = 300. The answer is 300. ✗

EXPLANATION FOR ERROR CATEGORY: The produced chain of thought could be made correct just by running the equation through an external calculator, instead of asking the model to do the computation.

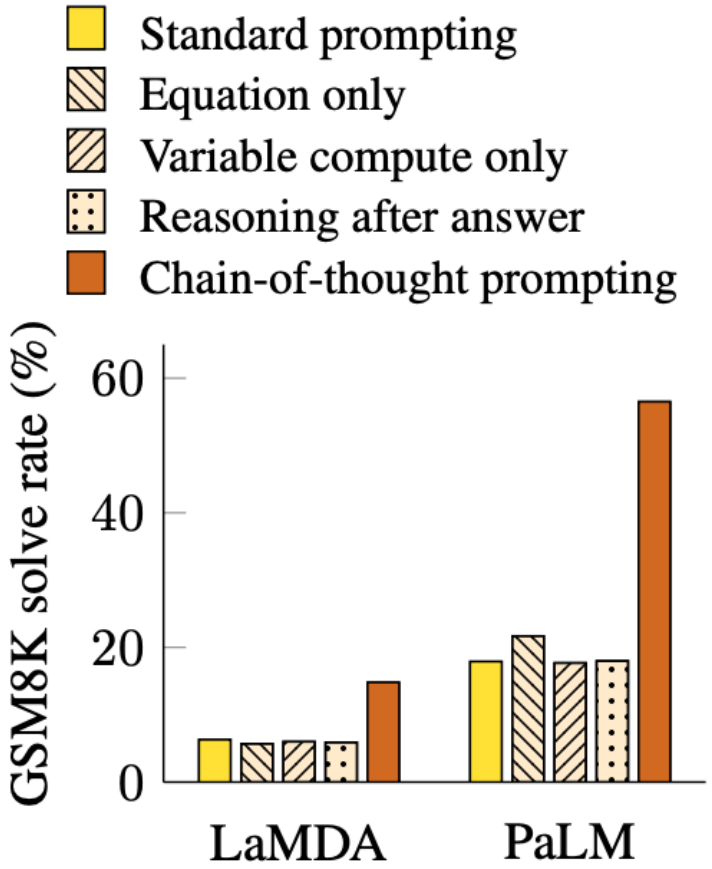
Arithmetic Reasoning

- Examination of model-generated thoughts
  - Model scale comparison
    - Scaling PaLM to 540B fixes a large portion of one-step missing and semantic understanding errors in the 62B model



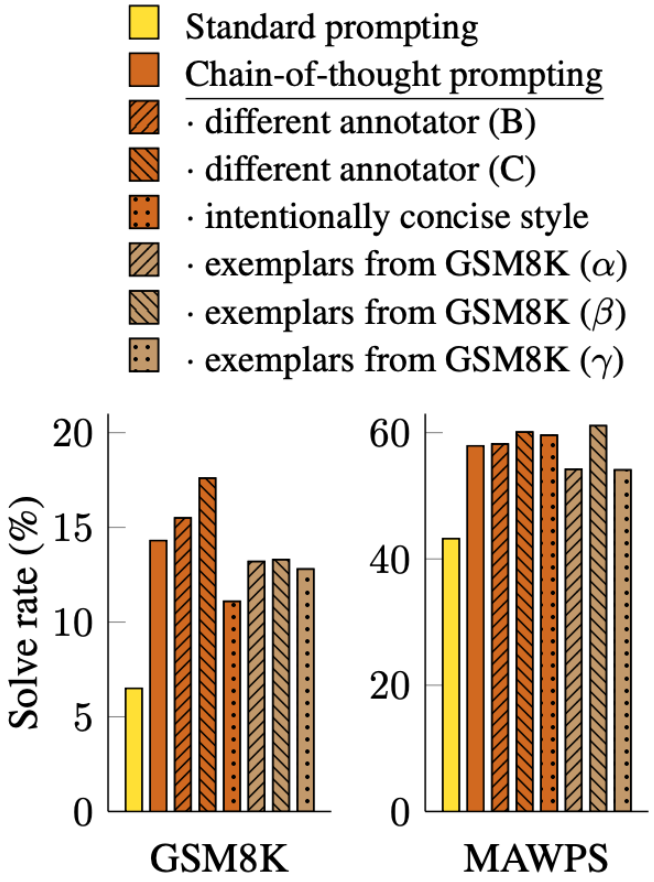
Arithmetic Reasoning

- Ablation Study
  - Equation only
    - challenging to directly translate into an equation
  - Variable compute only
    - CoT allows the model to spend more computation? No
    - variable computation is not the reason for the success
  - Chain of thought after answer
    - better access relevant knowledge? No
    - the sequential reasoning is useful



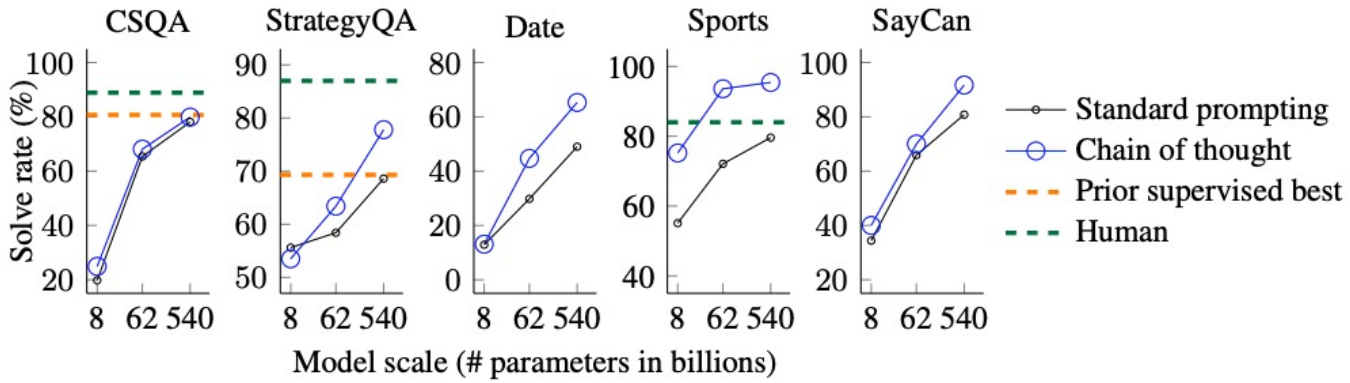
Arithmetic Reasoning

- Robustness of Chain of Thought
  - written by different annotators
  - all sets of chain of thought prompts outperform the standard baseline by a large margin
  - successful use of chain of thought does not depend on a particular linguistic style



Commonsense Reasoning

- Prompts
  - Randomly selected examples from the training set and
    manually composed chains of thought for them to use as few-shot exemplars
- Results



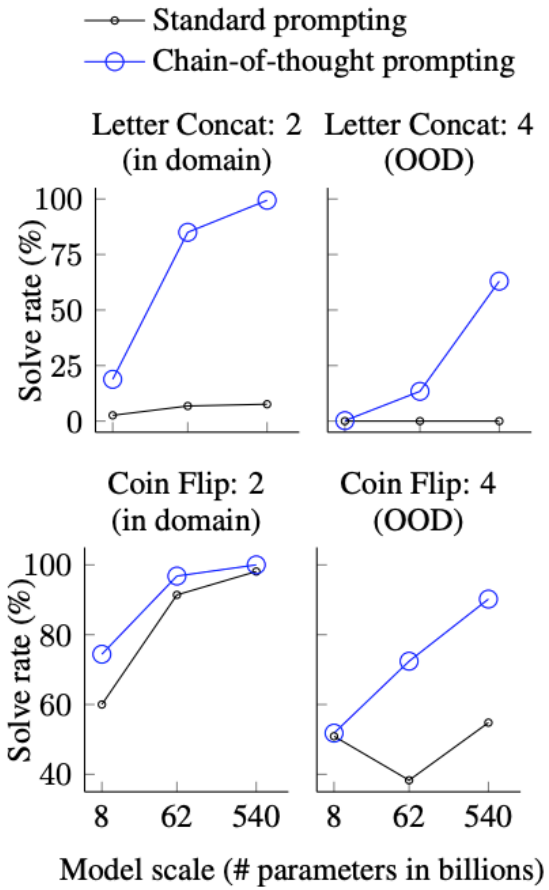
  - scaling up model size improved the performance of standard prompting

Symbolic Reasoning

- Chain of Thought prompting
  - enables language models to perform symbolic reasoning tasks
  - facilitates length generalization
- Tasks
  - Last letter concatenation (e.g., “Amy Brown” -> “yn”)
  - Coin flip
    (e.g., “A coin is heads up, Minseong flips the coin. Minseo does not flip the coin.
    Is the coin still heads up?” -> “no”)

Symbolic Reasoning

- in-domain: examples had same number of steps
- out-of-domain (OOD):
  evaluation examples had more steps
- Results
  - With PaLM 540B, almost 100% solve rates
  - Small models still fail
    : the ability to perform abstact manipulations on unseen symbols
    for these three tasks only arises at the scale 100B model parameters



Discussion

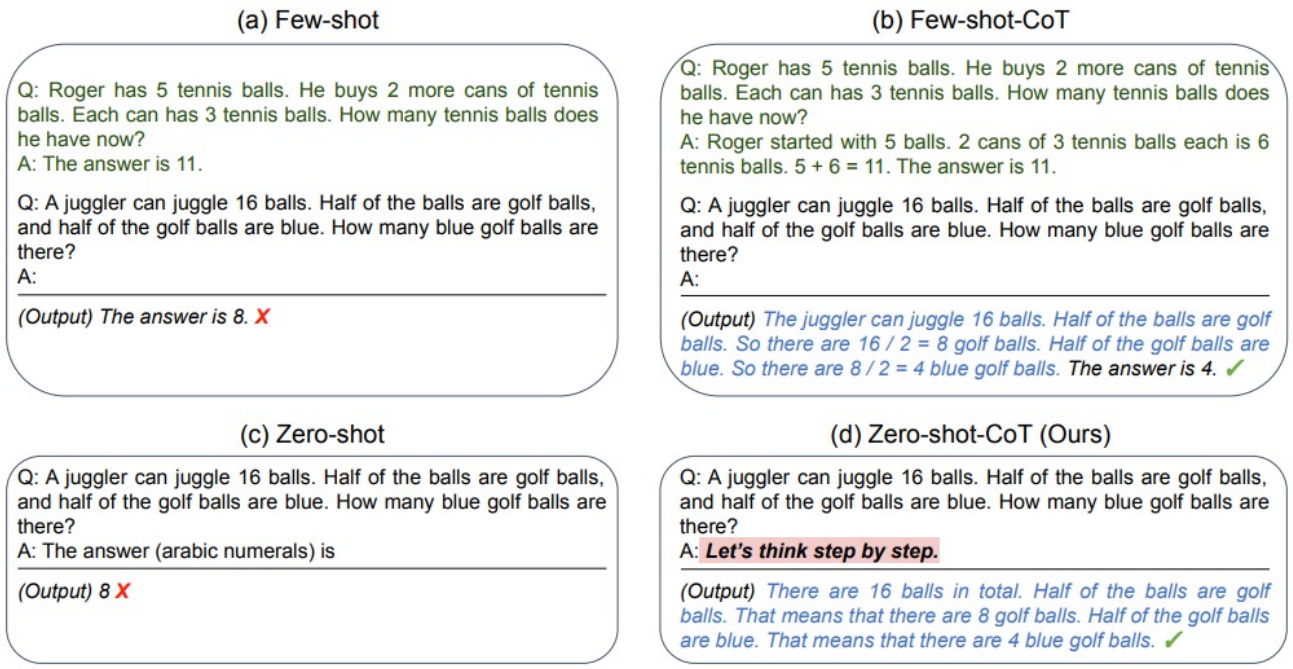
- No language models were finetuned in this paper
- Standard prompting only provides a lower bound on the capabilities of large language models
- Limitations
  - this does not answer whether the neural network is actually “reasoning”
  - such annotation costs could be prohibitive
  - there is no guarantee of correct reasoning paths
  - performance improvement is only at large model scales

Conclusions

- chain-of-thought reasoning is an emergent property of model scale that allows large language models to perform reasoning tasks

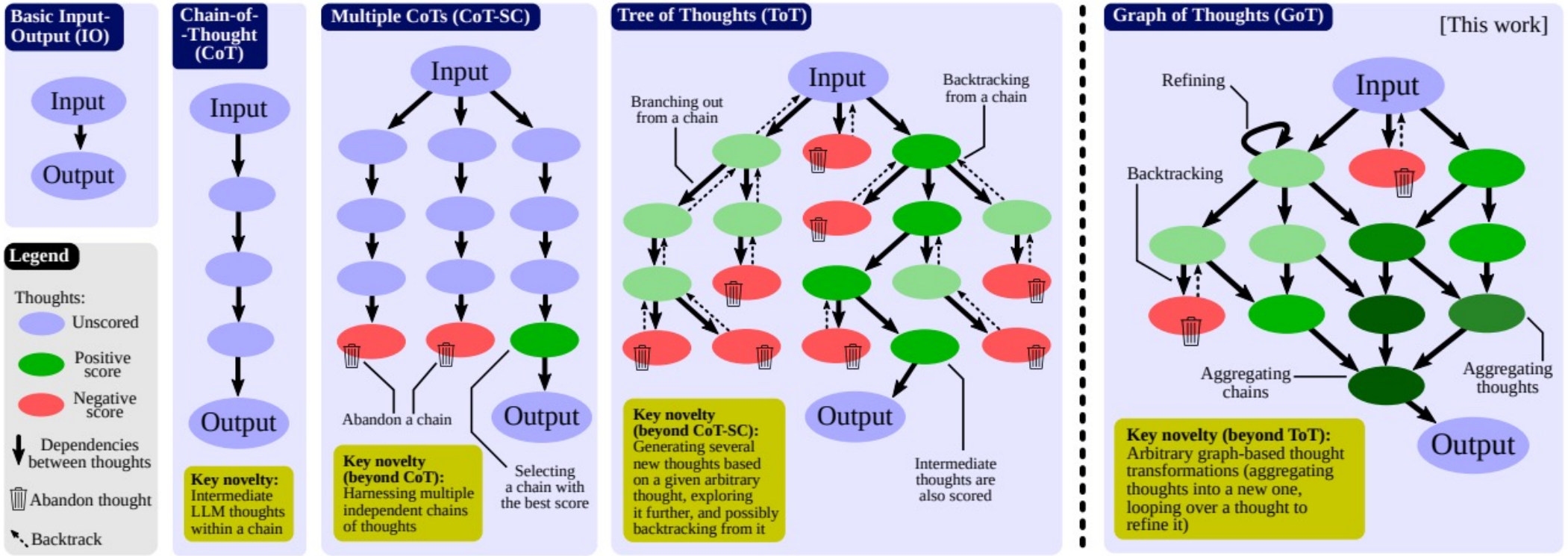
Further Research

- Large Language Models are Zero-Shot Reasoners
  - Let's think about step by step



Further Research

- Graph of Thoughts: Solving Elaborate Problems with Large Language Models



Further Research

- Graph of Thoughts: Solving Elaborate Problems with Large Language Models