How can we talk with LLMs?
- Prompt Engineering

Chengzhi Hu

Master's Seminar: The GPTs and Their Ilk

Prof. Hinrich Schütze, Haotian Ye, M.Sc.

14.07.2023

Prompt Engineering

- Introduction
- Prompting, Variations:
  - Chain of Thoughts (CoT)
  - Tree of Thoughts (ToT)
- Automatic Prompt Engineering
- Conclusion
- Q & A

Introduction

Prompt engineering is..

- the process of creating a *prompting function f(x)* that results in the most effective performance on the downstream task.[2]
- often *prompt template engineering*, where a human engineer or algorithm searches for the best template for each task the model is expected to perform. [2]

Introduction

CZ red wine French

‹ 2 / 2 ›

Red wine is a type of wine that is made from dark-colored grapes. It can be produced in various countries, including France, which is renowned for its wine production. France has a long history and tradition of winemaking, and many regions in the country are known for producing excellent red wines.

Some of the famous red wine regions in France include Bordeaux, Burgundy, Rhône Valley, and the Loire Valley. Each region has its own distinct characteristics, grape varieties, and winemaking techniques, which contribute to the diversity of French red wines.

In Bordeaux, red wines are typically made from blends of

CZ introduction red wine French

‹ 14 / 14 ›

Red wine holds a revered position in the world of beverages, and few countries are as synonymous with its production as France. Renowned for its winemaking heritage, France has long been celebrated as a premier producer of red wines that captivate connoisseurs worldwide. With a history dating back centuries and a diverse range of grape varieties, French red

CZ translate red wine French

‹ 3 / 3 ›

"Red wine" translated to French is "vin rouge."

?

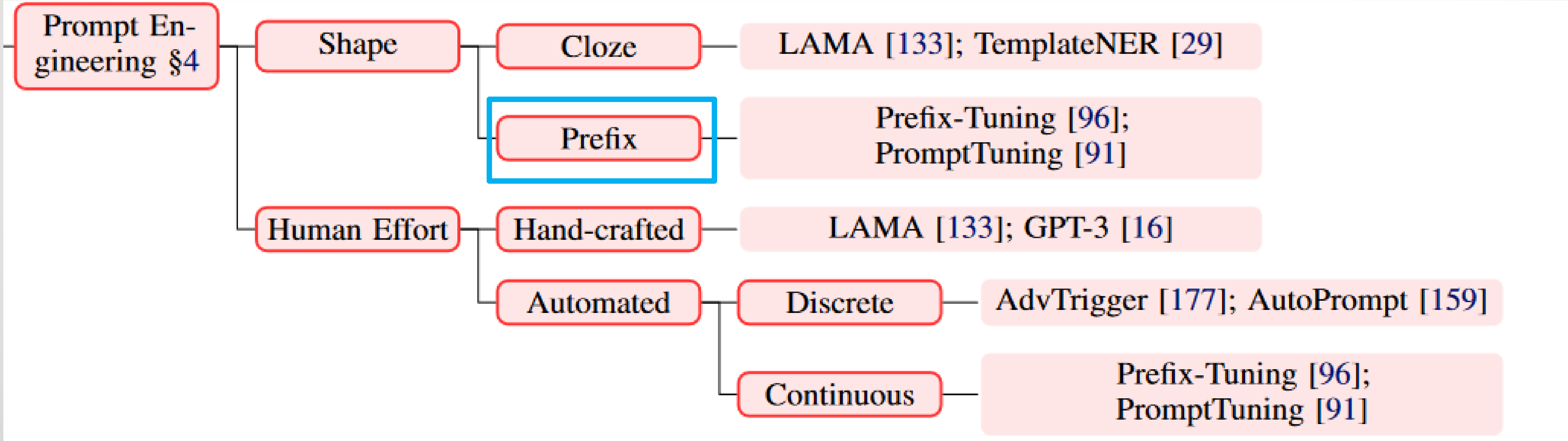
CZ beer french

In French, the word for beer is "bière."

Puzzle, quiz, coherent writing, math etc.

Introduction

Actually more than just good input in ChatGPT…

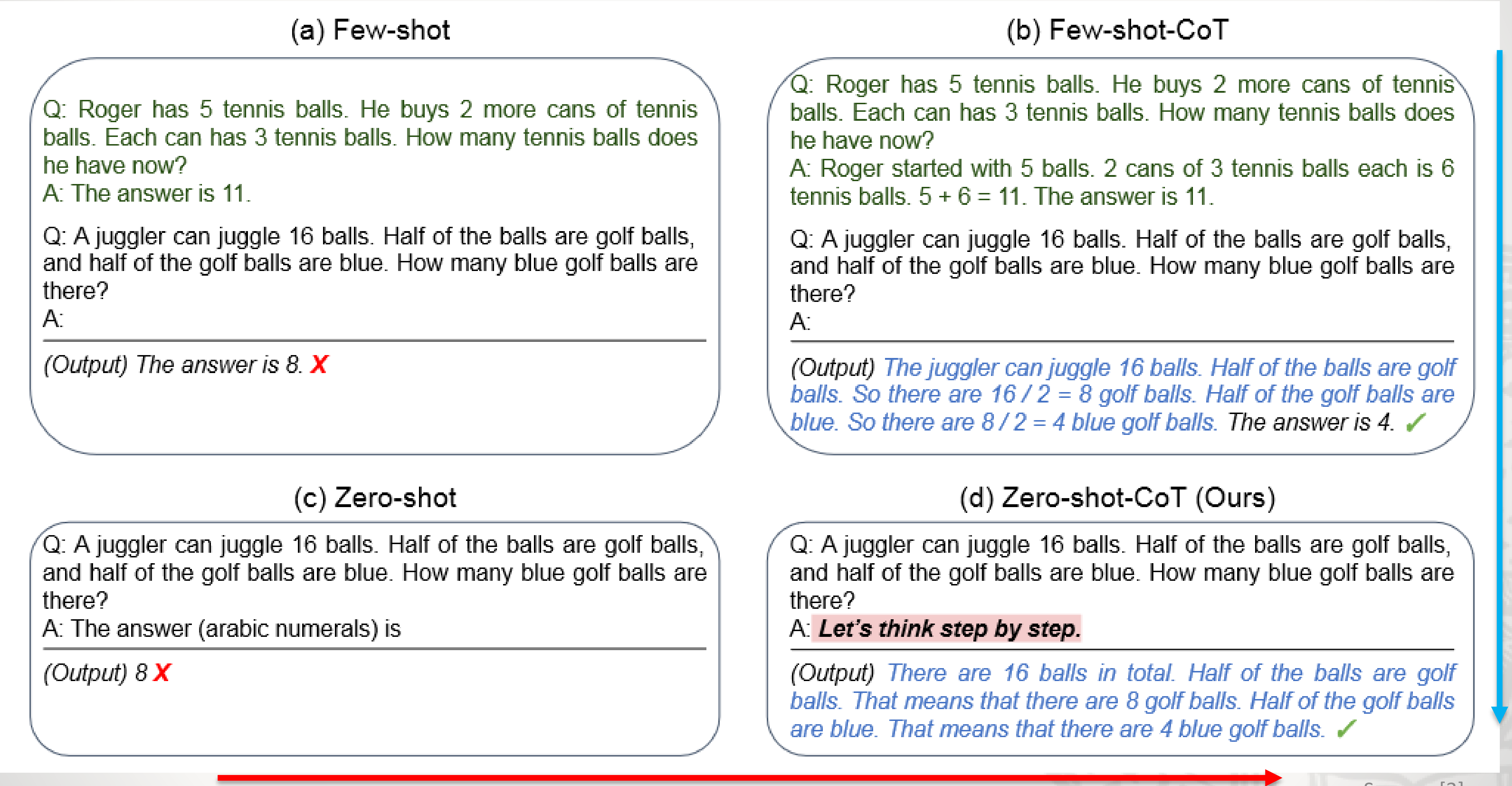


Source: [2]

Prompt Engineering

- Introduction
- Prompting, Variations:
  - Chain of Thoughts (CoT)
  - Tree of Thoughts (ToT)
- Automatic Prompt Engineering
- Conclusion
- Q & A

Chain of Thoughts (CoT) Prompting



Source: [3]

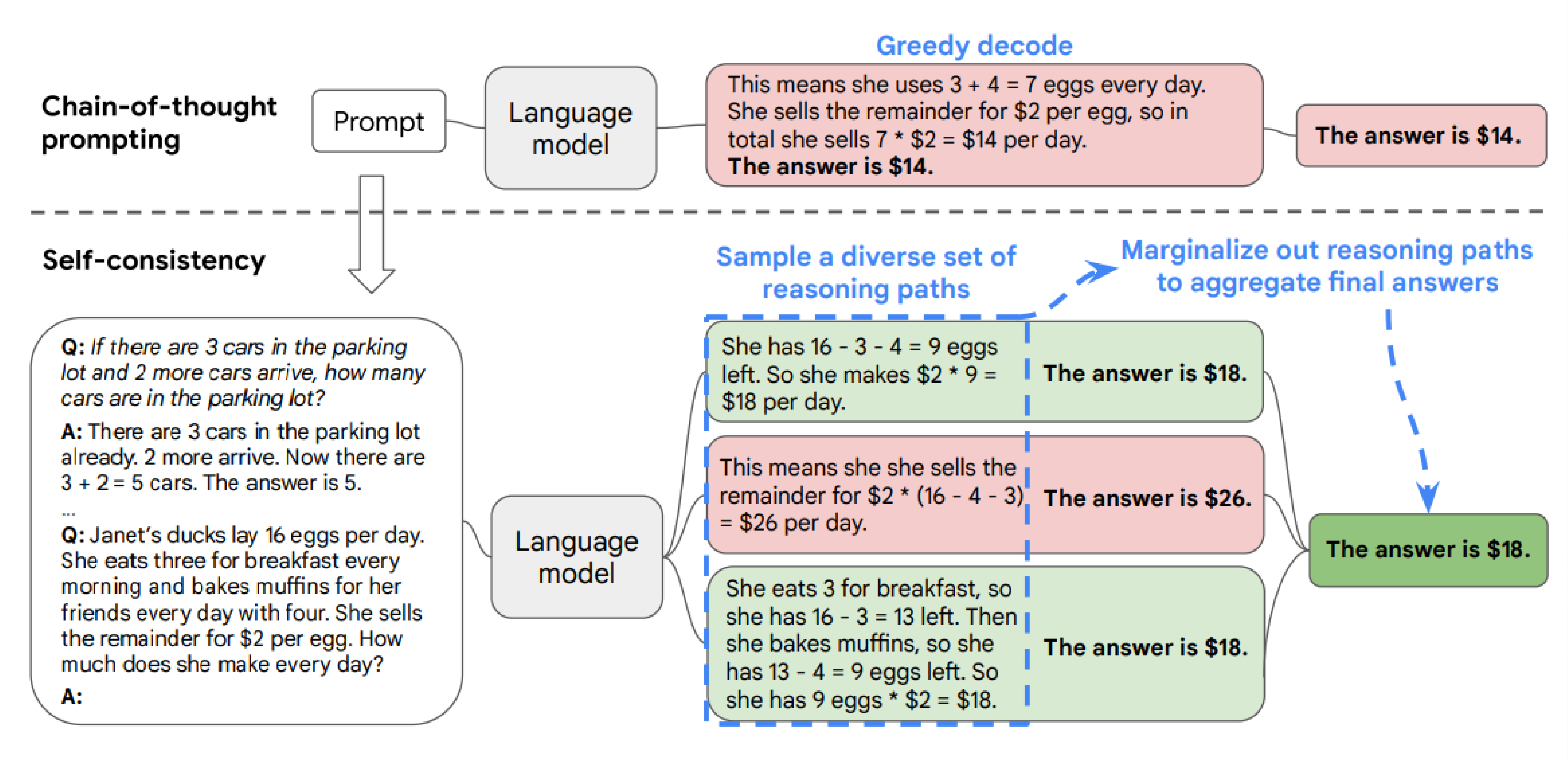
Chain of Thoughts (CoT) Prompting

Robustness: How does prompt selection affect Zero-shot-CoT?

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

Source: [3]

Chain of Thoughts (CoT) Self-Consistency



Source: [5]

Chain of Thoughts → Tree of Thoughts

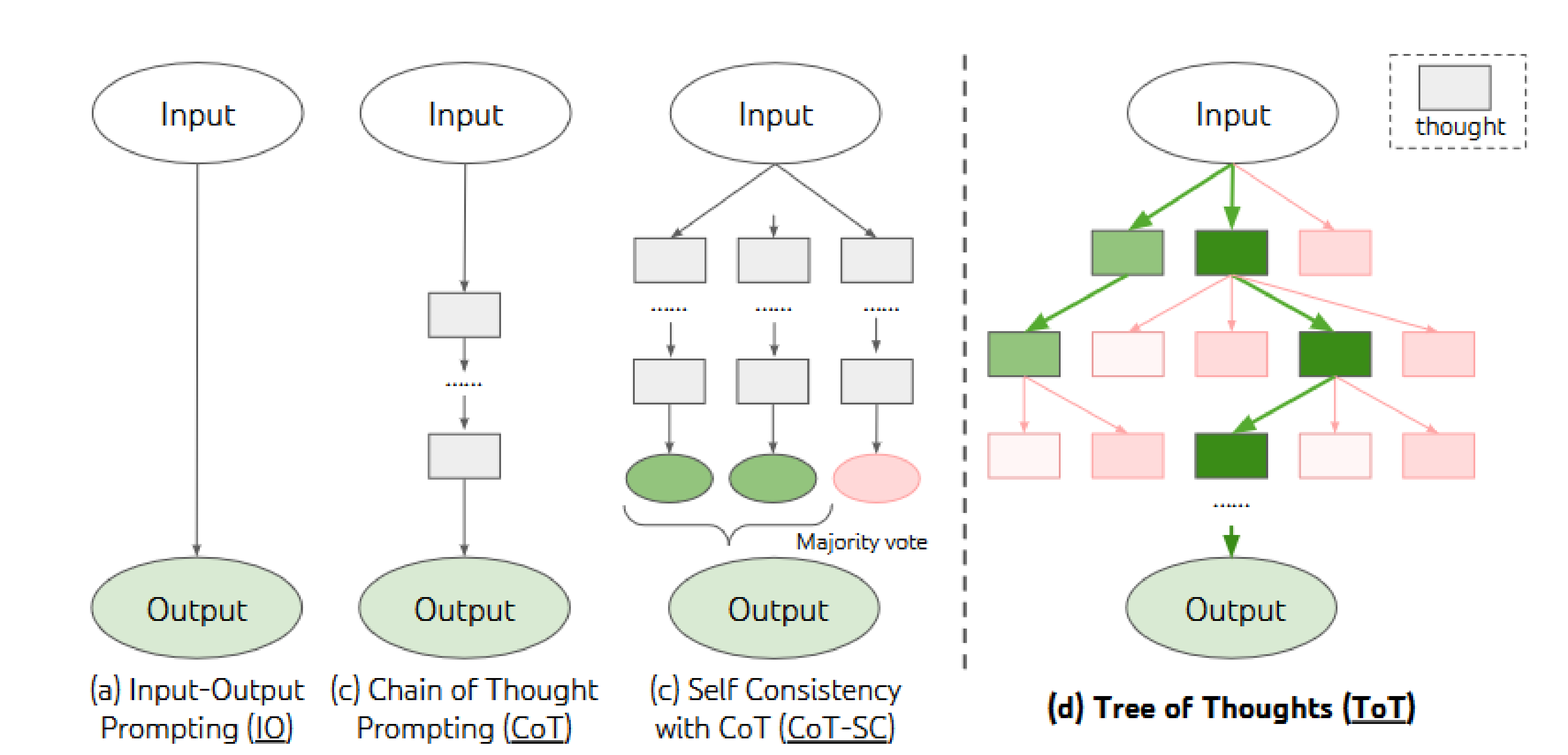
Thought, a coherent language sequence that serves as an intermediate step toward problem solving.

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

Shunyu Yao Princeton University
Dian Yu Google DeepMind
Jeffrey Zhao Google DeepMind
Izhak Shafran Google DeepMind
Thomas L. Griffiths Princeton University
Yuan Cao Google DeepMind
Karthik Narasimhan Princeton University

Chain of Thoughts → Tree of Thoughts

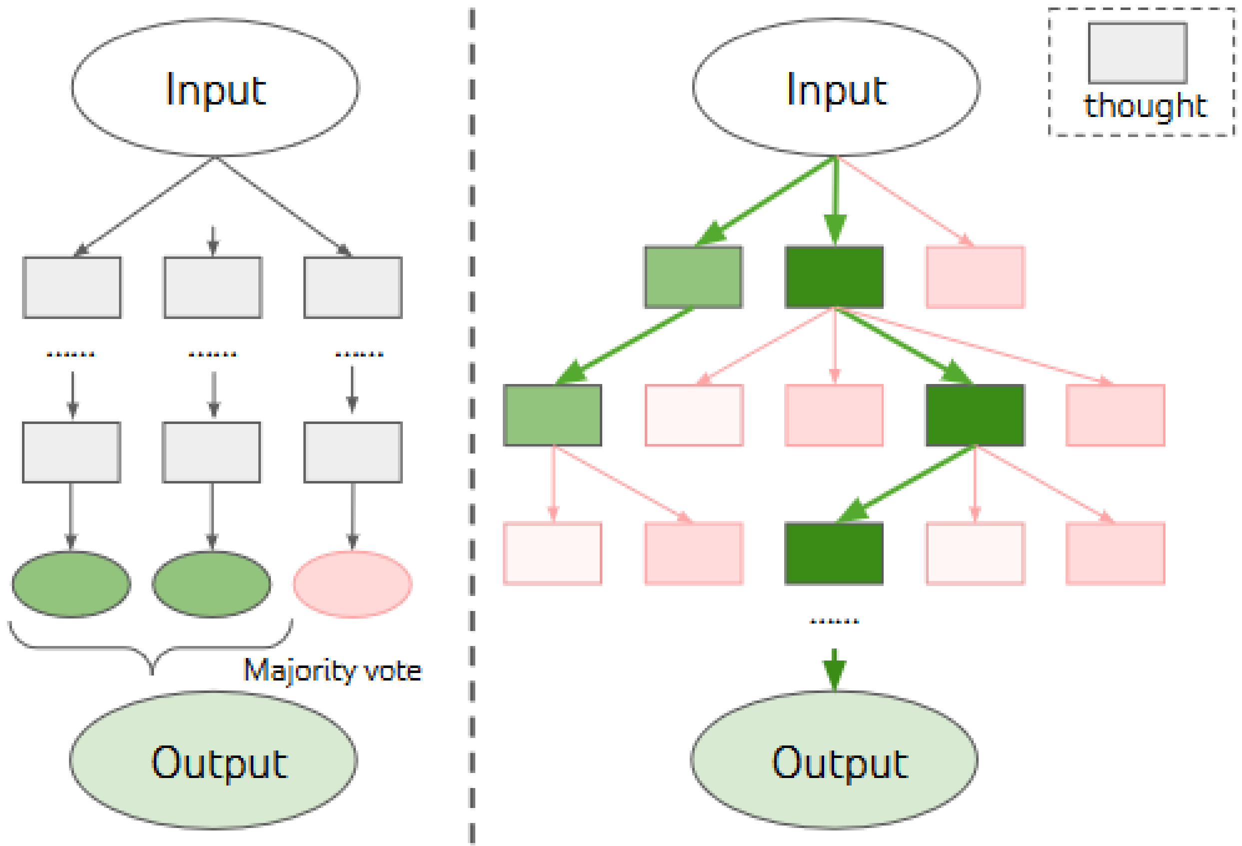
A genuine problem-solving process involves the repeated use of available information to initiate exploration, which discloses, in turn, more information until a way to attain the solution is finally discovered.—— Newell et al.



Source: [4]

Tree of Thoughts (ToT) Prompting

1. Thought decomposition
2. Thought generator
   a. Sample
   b. Propose
3. State evaluator
   a. Value
   b. Vote
4. Search algorithm
   a. BFS (→ best each step)
   b. DFS (→ impossible)



(c) Self Consistency with CoT (CoT-SC)

(d) Tree of Thoughts (ToT)

Source: [4]

Tree of Thoughts (ToT) Prompting

1. Thought decomposition
2. Thought generator
   - Sample
   - Propose
3. State evaluator
   - Value
   - Vote
4. Search algorithm
   - BFS (→ best states each step)
   - DFS (→ impossible/final output)

	Game of 24	Creative Writing	5x5 Crosswords
Input	4 numbers (4 9 10 13)	4 random sentences	10 clues (h1. presented;..)
Output	An equation to reach 24 (13-9)*(10-4)=24	A passage of 4 paragraphs ending in the 4 sentences	5x5 letters: SHOWN; WIRRA; AVAIL; ...
Thoughts	3 intermediate equations (13-9=4 (left 4,4,10); 10-4=6 (left 4,6); 4*6=24)	A short writing plan (1. Introduce a book that connects...)	Words to fill in for clues: (h1. shown; v5. naled; ...)
#ToT steps	3	1	5-10 (variable)

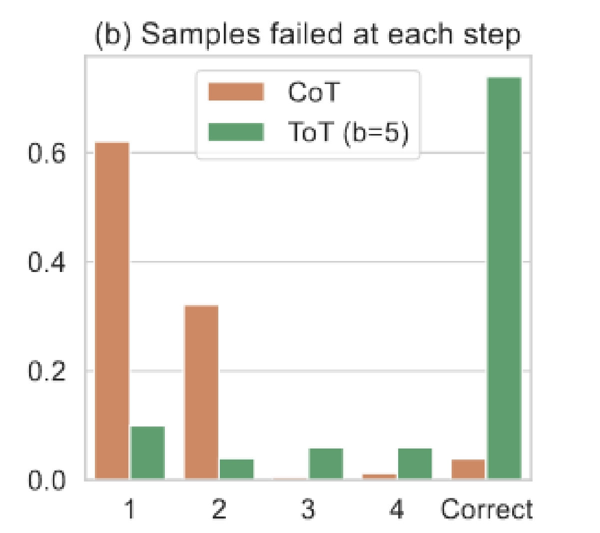
Table 1: Task overview. Input, output, thought examples are in blue.

Source: [4]

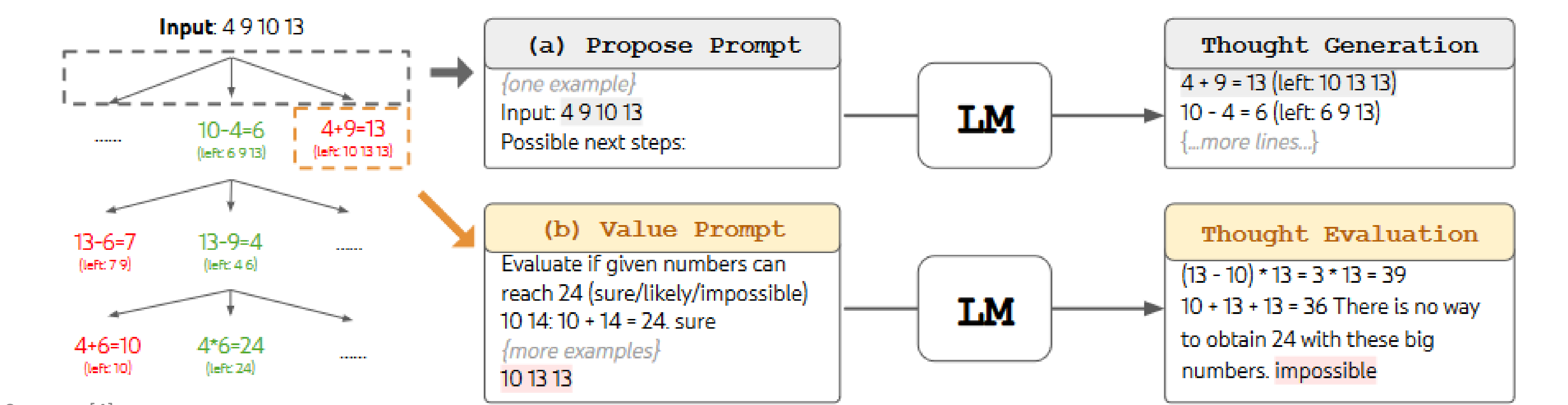
Tree of Thoughts (ToT) Examples

Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC (k=100)	9.0%
ToT (ours) (b=1)	45%
ToT (ours) (b=5)	**74%**
IO + Refine (k=10)	27%
IO (best of 100)	33%
CoT (best of 100)	49%

Table 2: Game of 24 Results.



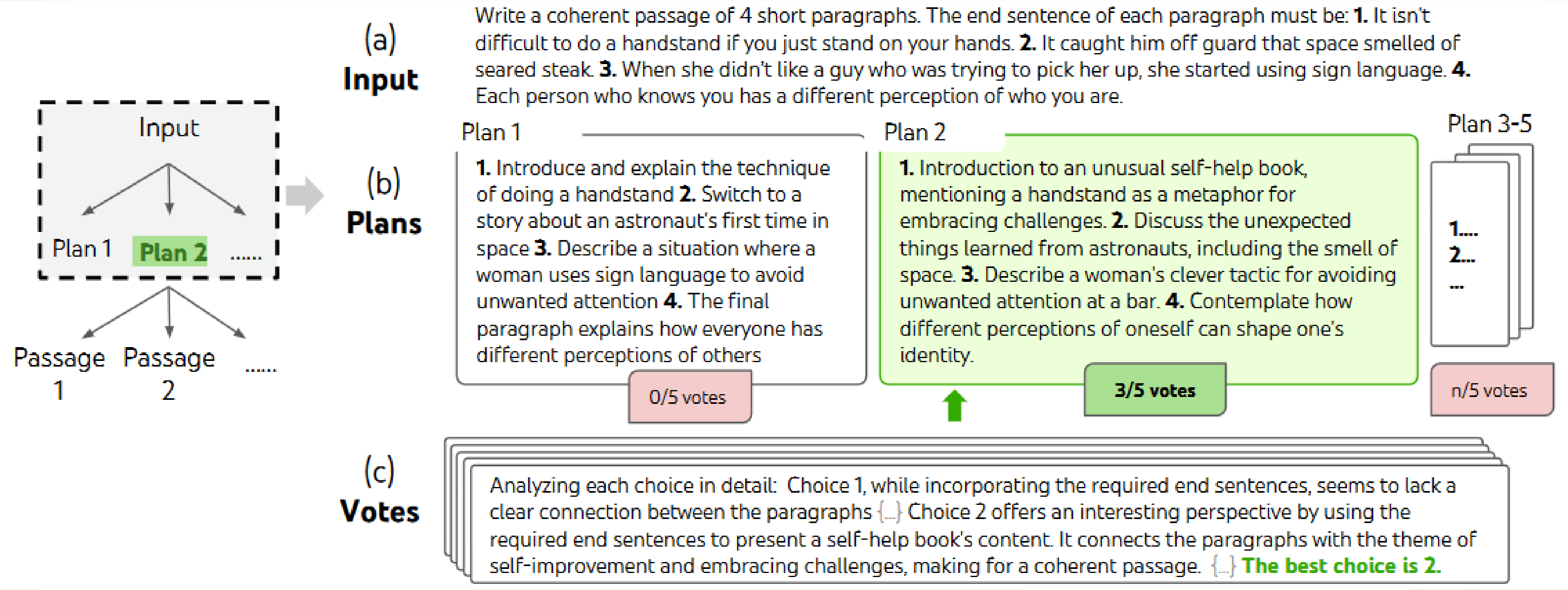
Game of 24



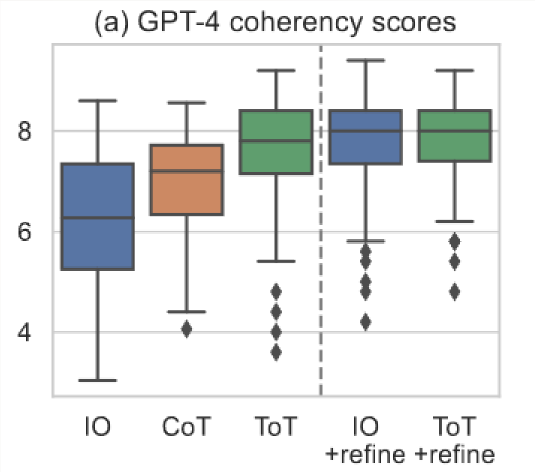
Sources: [4]

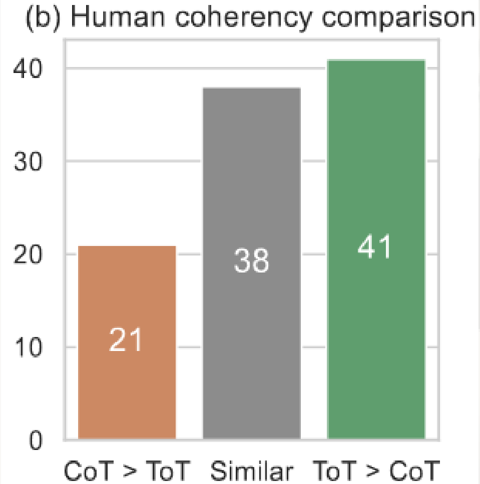
Tree of Thoughts (ToT) Examples

Creative Writing



Sources: [4]





Prompt Engineering

- Introduction
- Prompting, Variations:
  - Chain of Thoughts (CoT)
  - Tree of Thoughts (ToT)
- Automatic Prompt Engineering
- Conclusion
- Q & A

Automatic Prompt Engineering

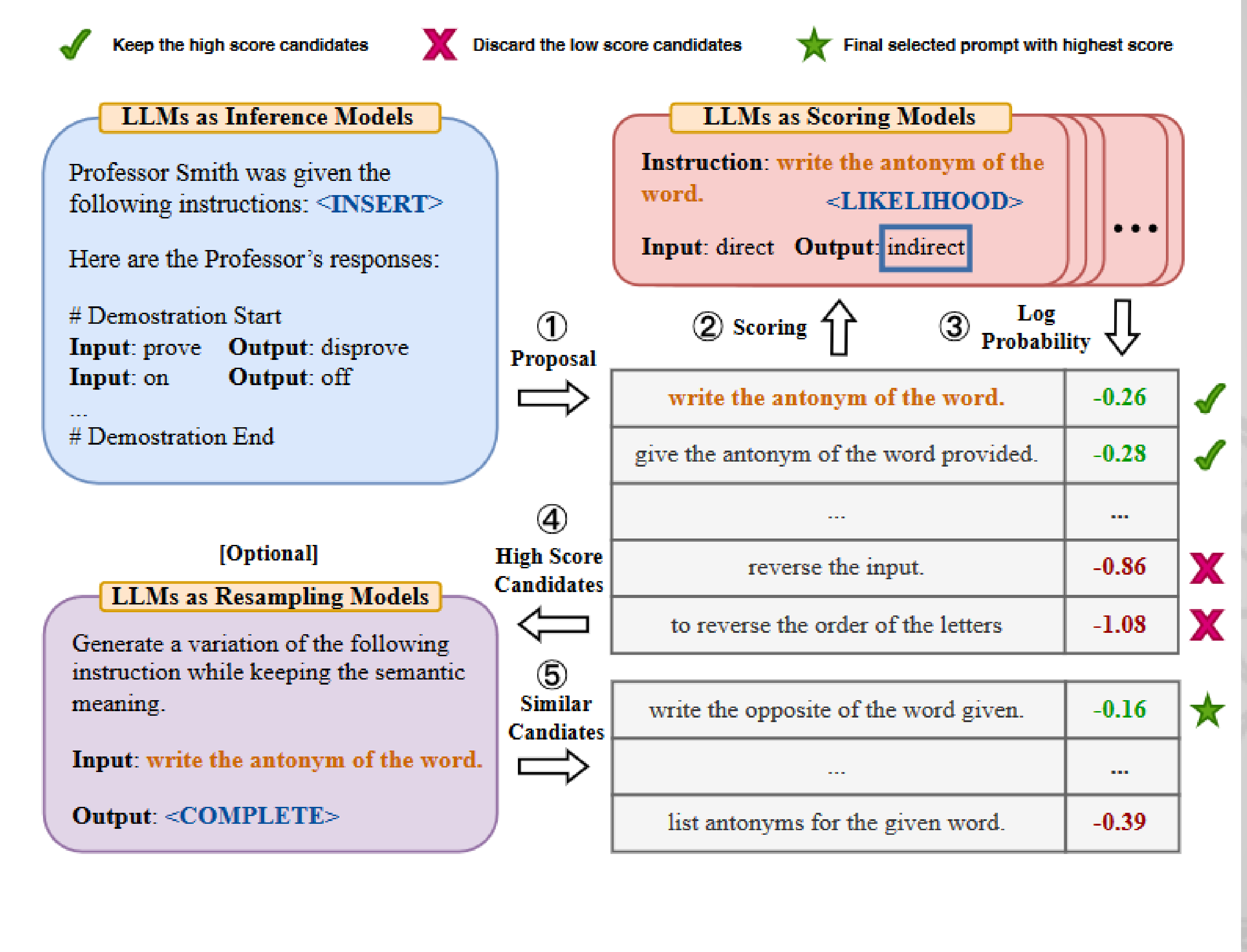
Recent progress in NLP has shown language models are very good at generating diverse natural language text. Therefore, we consider leveraging a pretrained LLM to propose a good set *U* of candidate solutions that will guide our search procedure.

LARGE LANGUAGE MODELS ARE HUMAN-LEVEL PROMPT ENGINEERS

Yongchao Zhou[1,2,*], **Andrei Ioan Muresanu**[2,3,*], **Ziwen Han**[1,2,*], **Keiran Paster**[1,2], **Silviu Pitis**[1,2], **Harris Chan**[1,2], **Jimmy Ba**[1,2]
[1]University of Toronto [2]Vector Institute [3]University of Waterloo *Equal contribution
{yczhou,hanziwen,keirp,spitis,hchan,jba}@cs.toronto.edu
{andrei.muresanu}@uwaterloo.ca

Automatic Prompt Engineering

- Use instructions optimized by searching over a pool of instruction candidates proposed by an LLM to maximize a chosen score function.
- To evaluate the quality of the selected instruction, we evaluate the zero-shot performance of another LLM following the selected instruction.



(a) Automatic Prompt Engineer (APE) workflow

Source: [1]

Automatic Prompt Engineering – Example

Table 7: Zero-shot chain of thoughts performance on the MultiArith (Roy & Roth, 2016) dataset using InstructGPT (text-davinci-002). Template (*1) was proposed in Kojima et al. (2022) to enable the zero-shot chain of thoughts reasoning of large language models, while template (*2) and (*3) were used in Ahn et al. (2022) and Reynolds & McDonell (2021), respectively.

No.	Category	Zero-shot CoT Trigger Prompt	Accuracy
1	APE	Let's work this out in a step by step way to be sure we have the right answer.	**82.0**
2	Human-Designed	Let's think step by step. (*1)	78.7
3		First, (*2)	77.3
4		Let's think about this logically.	74.5
5		Let's solve this problem by splitting it into steps. (*3)	72.2
6		Let's be realistic and think step by step.	70.8
7		Let's think like a detective step by step.	70.3
8		Let's think	57.5
9		Before we dive into the answer,	55.7
10		The answer is after the proof.	45.7
-		(Zero-shot)	17.7

Source: [1]

Automatic Prompt Engineering

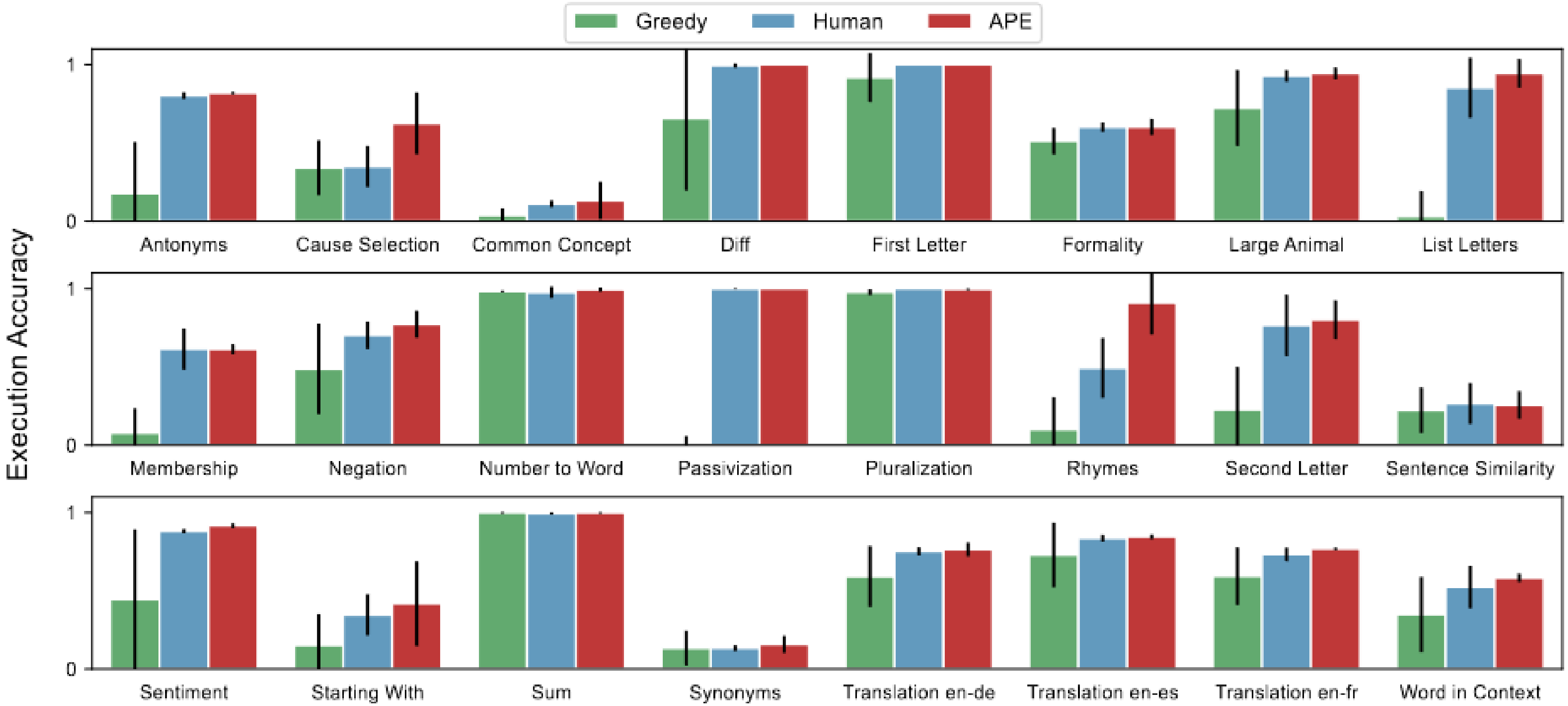


Figure 4: Zero-shot test accuracy on 24 Instruction Induction tasks. APE achieves human-level or better performance on all 24 out of 24 tasks.

Source: [1]

Prompt Engineering

- Introduction
- Prompting, Variations:
  - Chain of Thoughts (CoT)
  - Tree of Thoughts (ToT)
- Automatic Prompt Engineering
- **Conclusion & Limitation**
- Q & A

Conclusions

- Given a suite of appropriate prompts, a single LM trained in an entirely unsupervised fashion can be used to solve a great number of tasks!
- With ideal prompts, we could exploit the potential of LMs. → Interaction with LMs.

Limitations

- Research or tricks?
- Benchmark?

Prompt Engineering

- Introduction
- Prompting, Variations:
  - Chain of Thoughts (CoT)
  - Tree of Thoughts (ToT)
- Automatic Prompt Engineering
- Conclusion & Limitation
- Q & A

Reference

[1] Zhou Y, Muresanu A I, Han Z, et al. Large language models are human-level prompt engineers[J]. arXiv preprint arXiv:2211.01910, 2022.

[2] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[3] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[J]. Advances in neural information processing systems, 2022, 35: 22199-22213.

[4] Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. arXiv preprint arXiv:2305.10601, 2023.

[5] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[J]. arXiv preprint arXiv:2203.11171, 2022.