CS 639: Foundation Models
Using FMs: Prompting, ICL

Fred Sala

University of Wisconsin-Madison

March 5, 2026

Announcements

- Midterm: **March 11, 5:40 pm - 7:20 pm**
  - Sample problems out; solutions coming out Friday
- **Homework 2**: due Tuesday.
- **Project information**: released

  https://pages.cs.wisc.edu/~fredsala/cs639/files/project_info_639.pdf
- Class roadmap:

Thursday March 5	Prompting, ICL, and Others
Tuesday March 10	Specialization I
Thursday March 12	Specialization II
Tuesday March 15	Alignment I

Modifying and Improving FMs

Announcements

- Midterm info**: March 11, 5:40 pm - 7:20 pm**
  - Main room: Ingraham Hall, Room B10
  - **McBurney: Morgridge 5548**
    - Possible subject to change, will warn you
  - Sample problems out; solutions coming on Friday
  - **Format: ~**25 questions, all multiple choice, some calculation, some conceptual
  - **Covers:** Up to today's lecture (14)
  - **Cheat-sheet:** allowed. 1 page, handwritten

Outline

- **From Last Time: Other Modalities and Domains**
  - Audio, video, RL/Robotics
- **Basics of Prompting**
  - Terminology: zero-shot, few-shot, in-context, etc, prompt characteristics: format, examples, orders
- **Improving and Extending Prompting**
  - Searching for good prompts, techniques for continuous/soft prompts, ensembling, prompt optimization, chain-of-thought

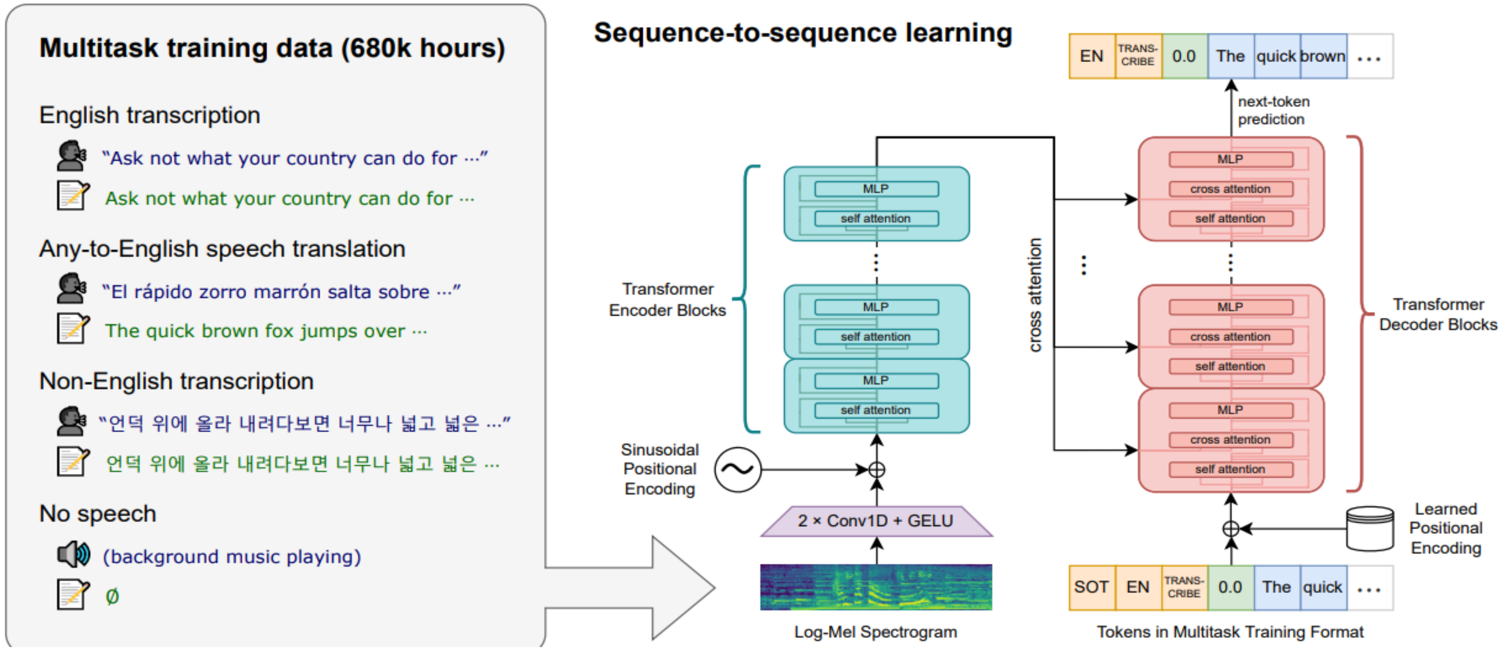
Outline

- **From Last Time: Other Modalities and Domains**
  - Audio, video, RL/Robotics
- **Basics of Prompting**
  - Terminology: zero-shot, few-shot, in-context, etc, prompt characteristics: format, examples, orders
- **Improving and Extending Prompting**
  - Searching for good prompts, techniques for continuous/soft prompts, ensembling, prompt optimization. chain-of-thought

Other Modalities: Audio

Can do similar things with all sorts of other modalities

- Audio: can sometimes convert to spectrograms (images) and apply our multimodal models
- **Ex: Whisper.** 680K hours of audio supervision

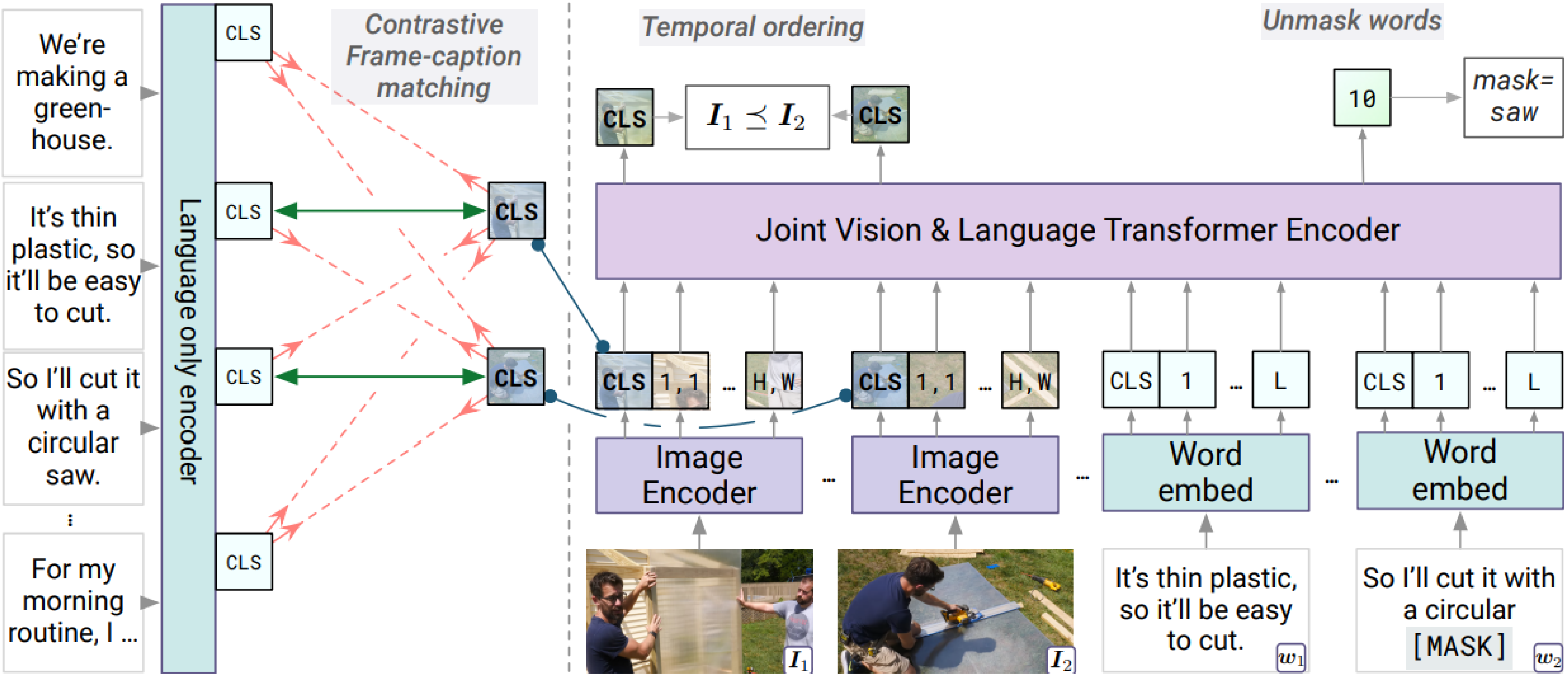


Multitask training data (680k hours)
English transcription
"Ask not what your country can do for …"
Ask not what your country can do for …
Any-to-English speech translation
"El rápido zorro marrón salta sobre …"
The quick brown fox jumps over …
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 …"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 …
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
…
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
…
Tokens in Multitask Training Format

Radford et al '22

Other Modalities: **Audio + Video + Text**

Merlot: video + text

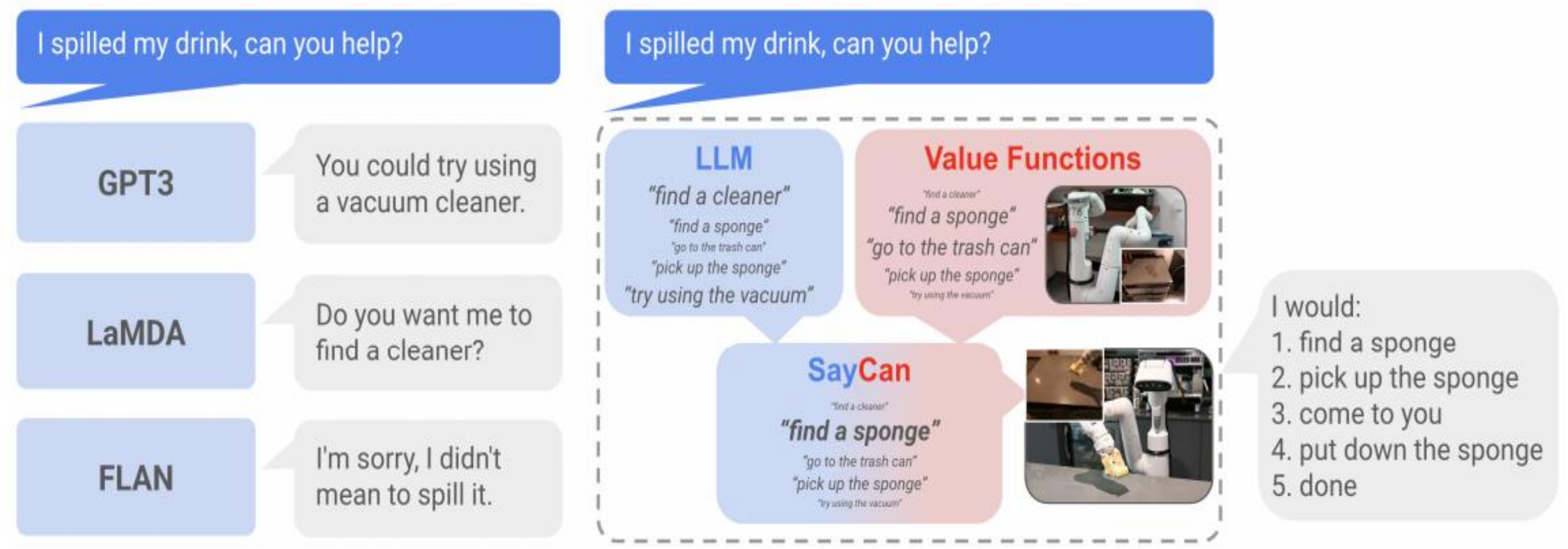

We're making a green-house.
It's thin plastic, so it'll be easy to cut.
So I'll cut it with a circular saw.
⋮
For my morning routine, I ...
Language only encoder
CLS
CLS
CLS
CLS
Contrastive Frame-caption matching
CLS
CLS
Temporal ordering
CLS
$I_1 \preceq I_2$
CLS
Unmask words
10
mask= saw
Joint Vision & Language Transformer Encoder
CLS
1, 1
...
H, W
CLS
1, 1
...
H, W
CLS
1
...
L
CLS
1
...
L
Image Encoder
...
Image Encoder
...
Word embed
...
Word embed
I_1
I_2
It's thin plastic, so it'll be easy to cut.
w_1
So I'll cut it with a circular [MASK]
w_2

Zellers et al '21

Foundation Models in Robotics

Can use language models for planning/robotics, but

- Not “grounded” since not aware of the environment
- Can mix together with RL concepts



I spilled my drink, can you help?

GPT3

You could try using a vacuum cleaner.

LaMDA

Do you want me to find a cleaner?

FLAN

I'm sorry, I didn't mean to spill it.

I spilled my drink, can you help?

LLM

“find a cleaner”
“find a sponge”
“go to the trash can”
“pick up the sponge”
“try using the vacuum”

Value Functions

“find a cleaner”
“find a sponge”
“go to the trash can”
“pick up the sponge”
“try using the vacuum”

SayCan

“find a cleaner”
“find a sponge”
“go to the trash can”
“pick up the sponge”
“try using the vacuum”

I would:
1. find a sponge
2. pick up the sponge
3. come to you
4. put down the sponge
5. done

Ahn et al ‘22

Foundation Models in Robotics: **SayCan**

Can use language models for planning/robotics, but

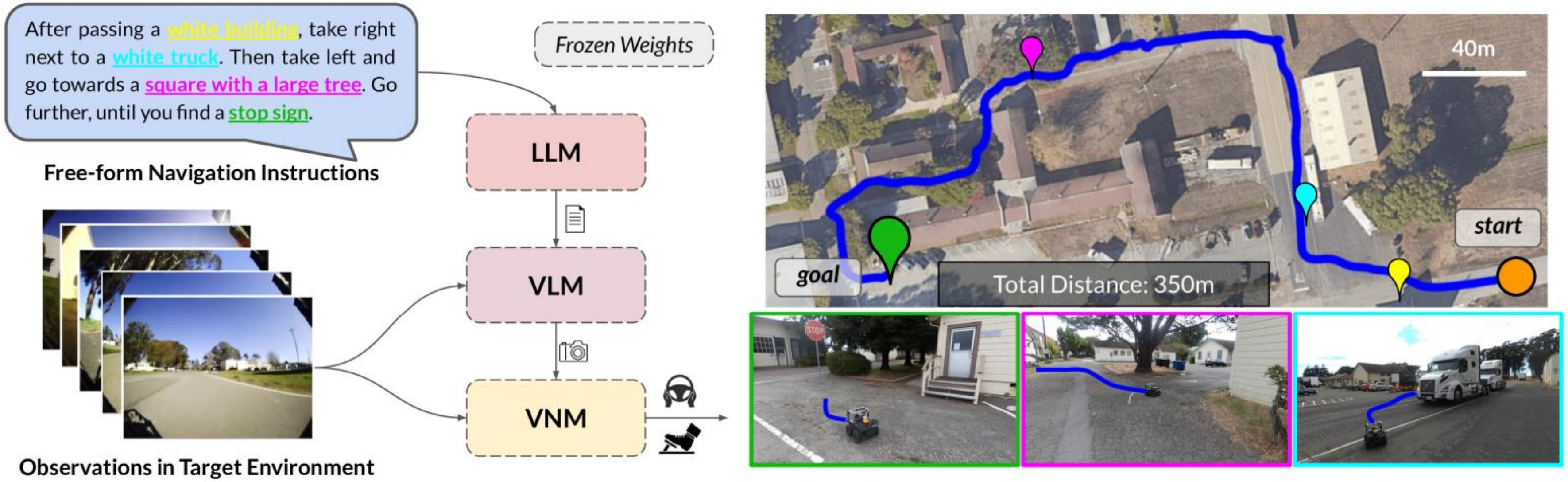
- Not “grounded” since not aware of the environment
- Can mix together with RL concepts
- Basic idea (Ahn et al ’22)

$$\pi \;=\; \arg\max_{\pi\in\Pi} p(\underbrace{c_\pi|s,\ell_\pi}_{\text{Prob. of completing skill/step from state s}})p(\underbrace{\ell_\pi|i}_{\text{LLM-provided prob of next step being valid}})$$

Foundation Models in Robotics: **Navigation**

For navigation:

- Connect multiple FMs (language, vision, action)
- **Inputs**: observations, instructions
- **Output**: plan



After passing a white building, take right next to a white truck. Then take left and go towards a square with a large tree. Go further, until you find a stop sign.

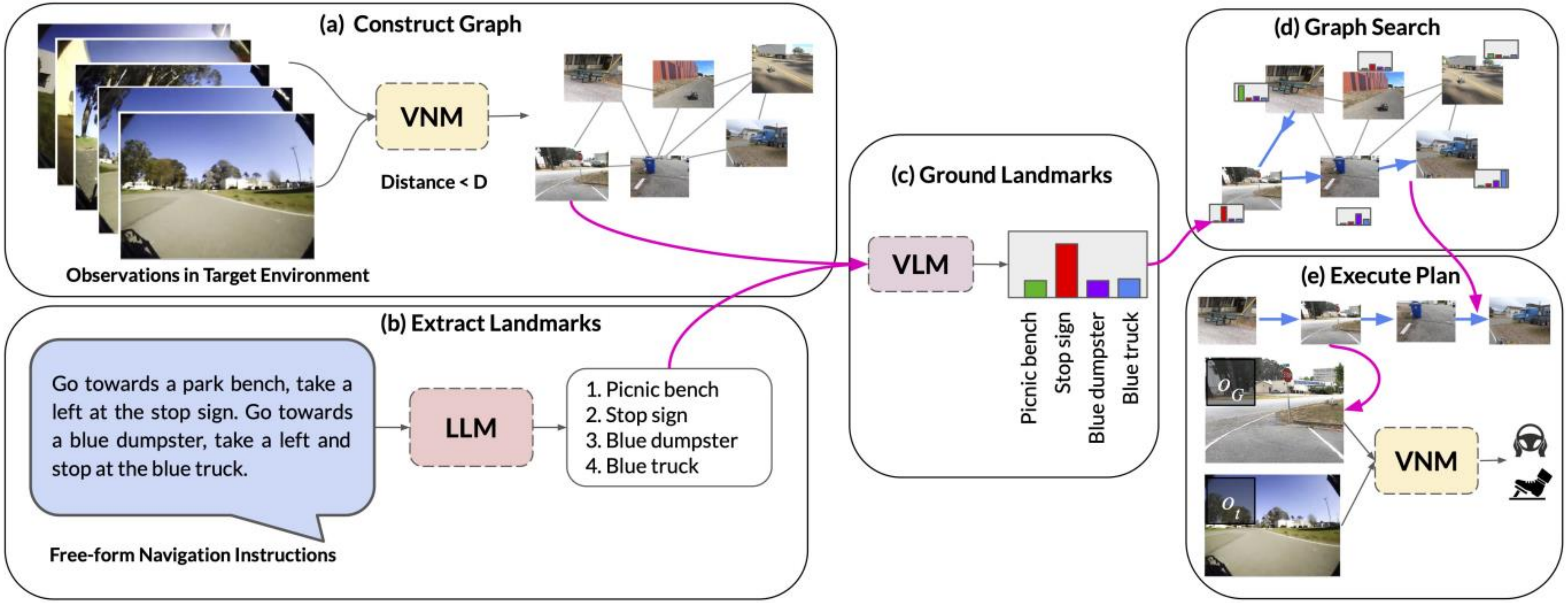
Free-form Navigation Instructions
Frozen Weights
LLM
VLM
VNM
Observations in Target Environment
40m
start
goal
Total Distance: 350m
STOP

Shah et al ‘22

Foundation Models in Robotics: **Navigation**

For navigation:

- Connect multiple FMs (language, vision, action)



Shah et al '22

Break & Questions

Outline

- **From Last Time: Other Modalities and Domains**
  - Audio, video, RL/Robotics
- **Basics of Prompting**
  - Terminology: zero-shot, few-shot, in-context, etc, prompt characteristics: format, examples, orders
- **Improving and Extending Prompting**
  - Searching for good prompts, techniques for continuous/soft prompts, ensembling, prompt optimization, chain-of-thought

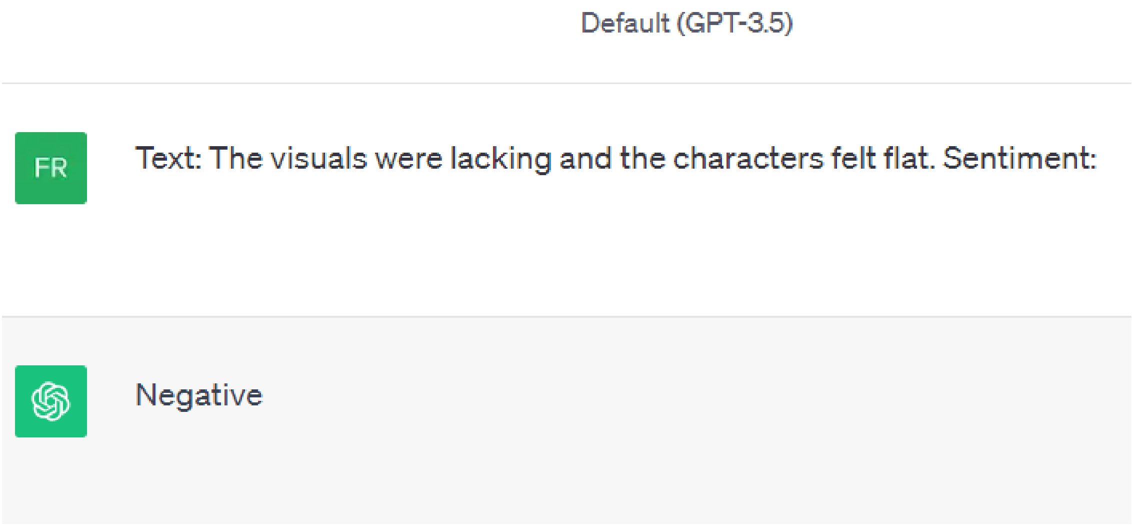
Prompting: Ask Your Model

Essentially, ask your model to perform your goal task

Example: sentiment analysis task

- Prompt: “`Text: The visuals were lacking and the characters felt flat. Sentiment:`”

- Result: “`Negative`”



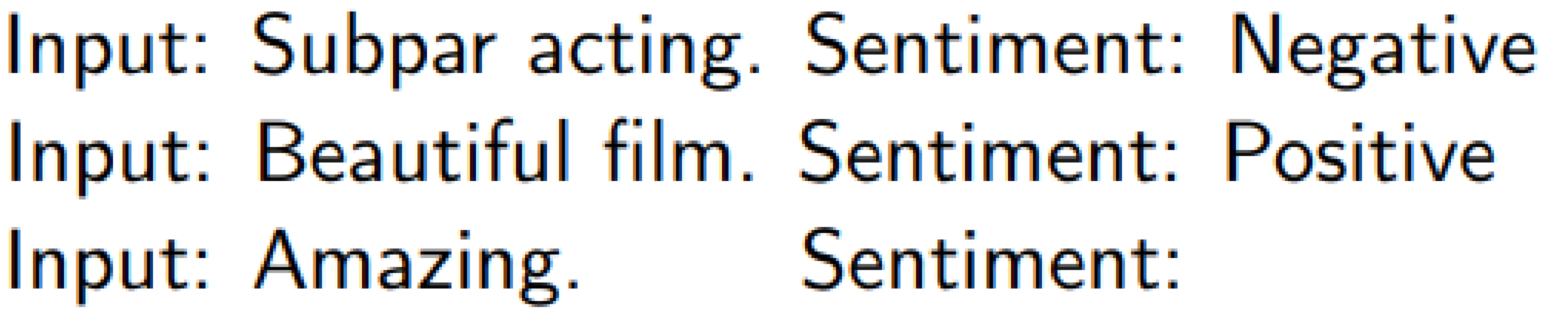
Text: The visuals were lacking and the characters felt flat. Sentiment:

Negative

Prompting: Zero-shot vs Few-shot

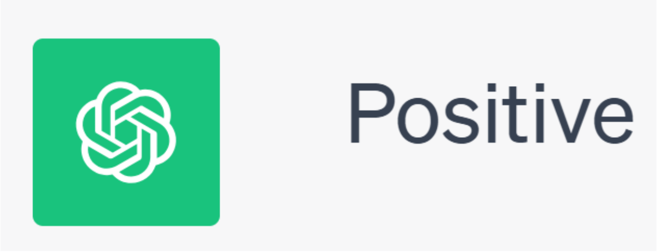
Terminology:

- **Zero-shot:** No “examples” provided to the model.
- **Few-shot/in-context learning**: Provide “examples”



Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment:

Zhao et al '21



Positive

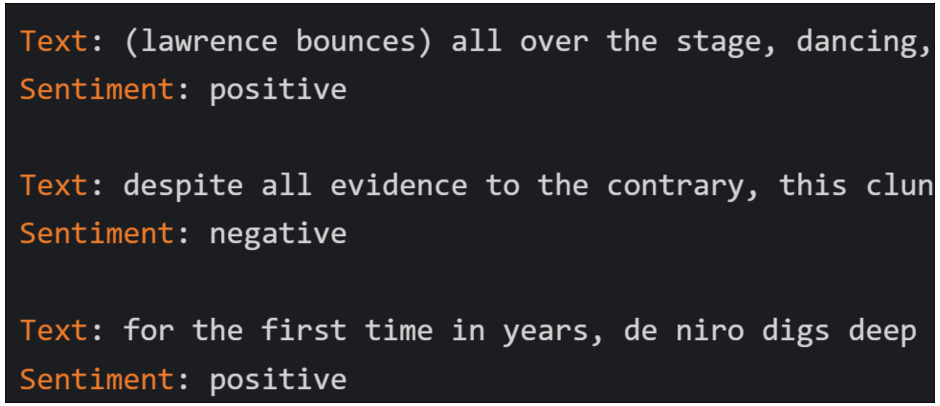
Prompting: Few-shot vs. In-context learning

Terminology conflicts! Note: we have a set of labeled examples. Could **fine-tune**!

Few-shot: *sometimes* historically meant fine-tune on this dataset, then prompt

In-context learning: do not fine-tune. Model weights unchanged.

- In modern LLM usage: **few-shot prompting ≈ in-context learning**

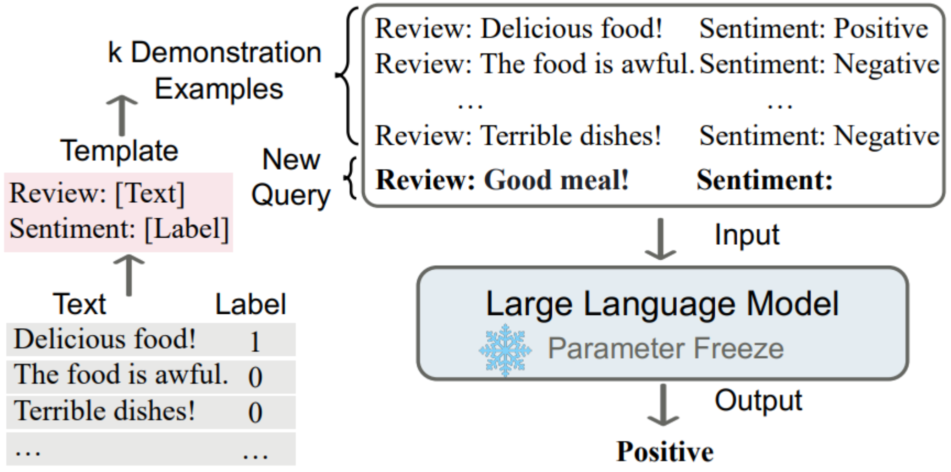


```
Text: (lawrence bounces) all over the stage, dancing,
Sentiment: positive

Text: despite all evidence to the contrary, this clun
Sentiment: negative

Text: for the first time in years, de niro digs deep
Sentiment: positive
```

Weng / SST



Dong et al, '23

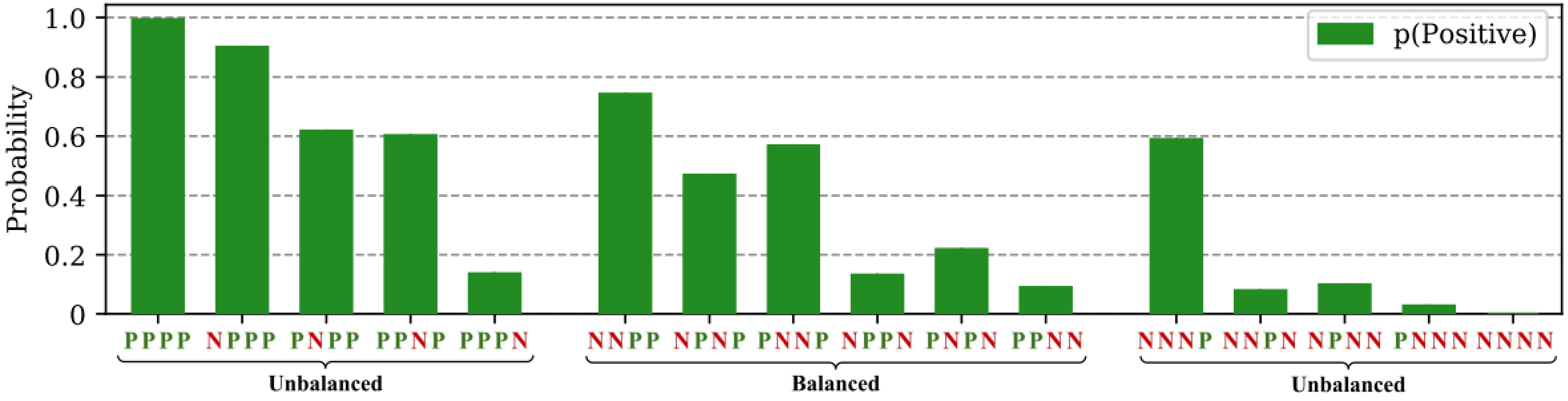
Few-Shot Choices

Examples/structure affect performance:

1. Prompt **format** (affects everything)
2. **Choice** of examples
3. **Order** of examples (permutation)

Prompting can be surprisingly brittle!

- **Still a problem 5 years later!**



Zhao et al '21

1. Prompt Formats (Historical!)

The choice of model affects the prompt format

Masked language model: “Cloze”-style prompt (**old!**)

- “I love this movie, it is a [Z] movie:”

Left-to-right language model: prefix prompt

- “I love this movie. What is the sentiment of this review?”

Note: eval datasets have pre-created prompts.

- LAMA (LAnguage Model Analysis): Cloze prompts

2. Prompt Formats: **Instruction-Based**

Modern instruction-tuned models have more complex instructions/formats

- **The good:** more natural way to tell the model what to do
- **The bad:** searching over formats/templates increasingly challenging

  - *Example: (White et al, '23):* `"From now on, I would like you to ask me questions to deploy a Python application to AWS. When you have enough information to deploy the application, create a Python script to automate the deployment.`"

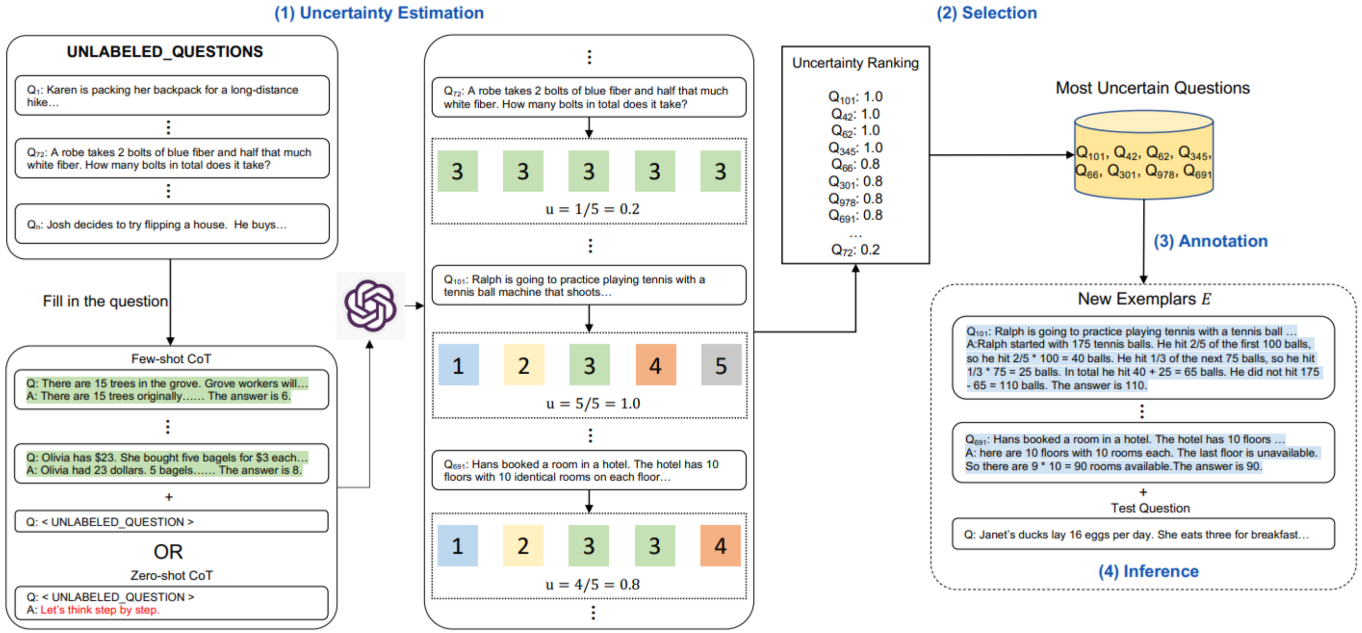
2. Choice of Examples

How to pick appropriate examples in few-shot?

- **Note:** initially only a “small’ number of examples can be shown, unlike in supervised learning.

Many options. Sampling:

- Liu et al, ‘21: kNN in embedding space (semantic similarity)
- Su et al, ‘22: Encourage diversity in embeddings
- Diao et al, ’23: “Active prompting”



Q_1: Karen is packing her backpack for a long-distance hike…

Q_{72}: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

Q_n: Josh decides to try flipping a house. He buys…

Q: There are 15 trees in the grove. Grove workers will…
A: There are 15 trees originally…… The answer is 6.

Q: Olivia has $23. She bought five bagels for $3 each…
A: Olivia had 23 dollars. 5 bagels…… The answer is 8.

+

Q: < UNLABELED_QUESTION >

Q: < UNLABELED_QUESTION >
A: Let's think step by step.

Q_{72}: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

Q_{101}: Ralph is going to practice playing tennis with a tennis ball machine that shoots…

Q_{691}: Hans booked a room in a hotel. The hotel has 10 floors with 10 identical rooms on each floor…

Q_{101}: Ralph is going to practice playing tennis with a tennis ball …
A:Ralph started with 175 tennis balls. He hit 2/5 of the first 100 balls, so he hit 2/5 * 100 = 40 balls. He hit 1/3 of the next 75 balls, so he hit 1/3 * 75 = 25 balls. In total he hit 40 + 25 = 65 balls. He did not hit 175 - 65 = 110 balls. The answer is 110.

Q_{691}: Hans booked a room in a hotel. The hotel has 10 floors …
A: here are 10 floors with 10 rooms each. The last floor is unavailable. So there are 9 * 10 = 90 rooms available.The answer is 90.

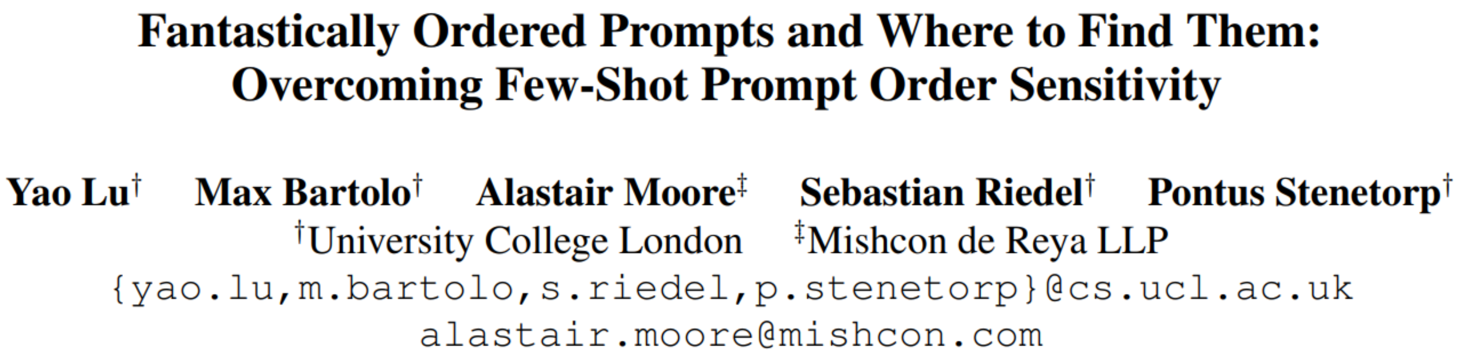
+

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast…

Diao et al ‘23

3. Order of Examples

What order to show them to the model?



Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity

Yao Lu[†] **Max Bartolo**[†] **Alastair Moore**[‡] **Sebastian Riedel**[†] **Pontus Stenetorp**[†]
[†]University College London [‡]Mishcon de Reya LLP
{yao.lu,m.bartolo,s.riedel,p.stenetorp}@cs.ucl.ac.uk
alastair.moore@mishcon.com

- **Findings:**
  - Model size doesn’t guarantee low-variance
  - Adding more examples doesn’t reduce variance
  - Good prompts don’t transfer from one model to another ☹
  - Good orders don’t transfer

Break & Questions

Outline

- **From Last Time: Other Modalities and Domains**
  - Audio, video, RL/Robotics
- **Basics of Prompting**
  - Terminology: zero-shot, few-shot, in-context, etc, prompt characteristics: format, examples, orders
- **Improving and Extending Prompting**
  - Searching for good prompts, techniques for continuous/soft prompts, ensembling, prompt optimization, chain-of-thought

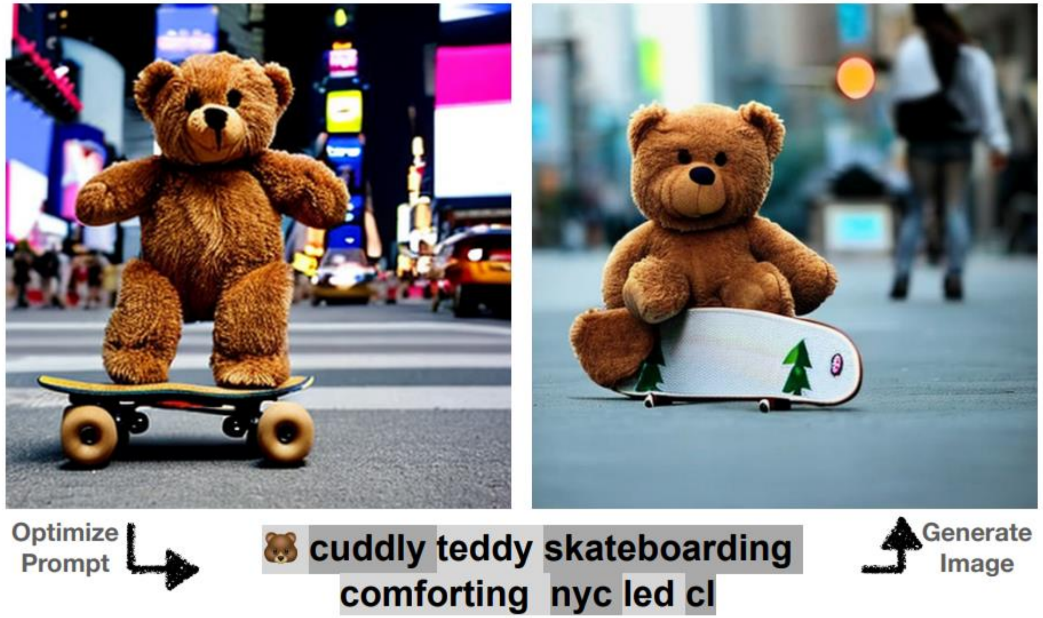
Hard Prompting

How do we improve our prompts? One way: search for better ones → **discovery**.

- I.e., we want to optimize, but over natural language/tokens

> "Hard prompt discovery is a specialized alchemy, with many good prompts being discovered by trial and error, or sheer intuition
>
> (Wen et al '23)

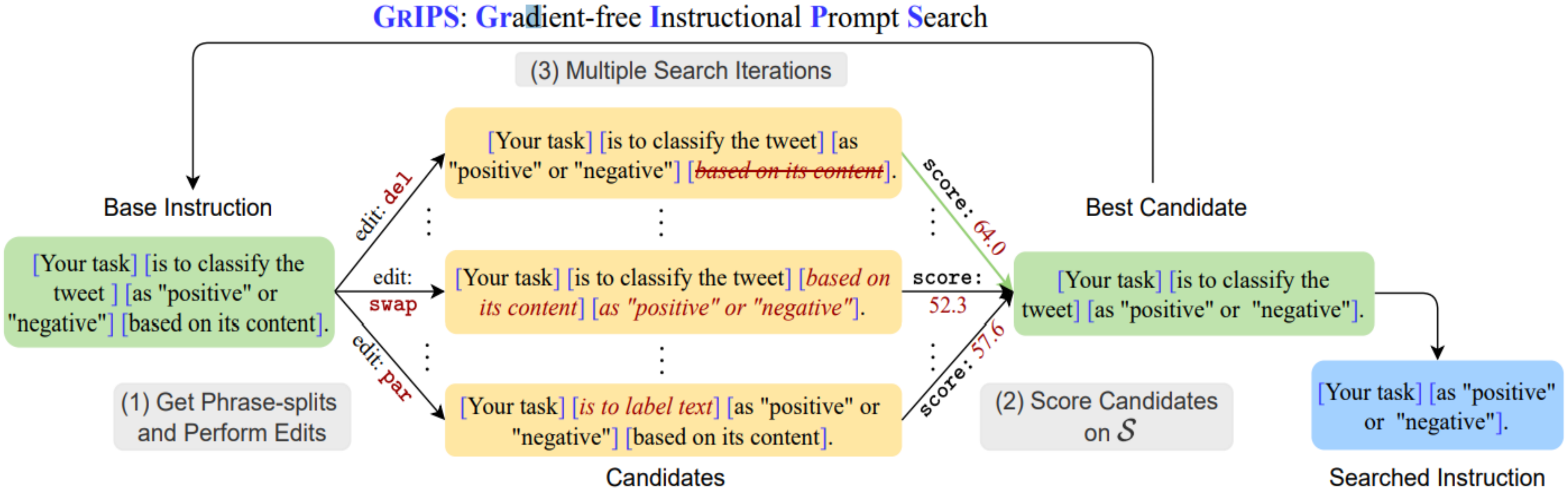
- Note: not just for language models!



Hard Prompting: **Discrete Optimization**

Sometimes, can avoid gradients

- Random search
- Greedy



GRIPS: Gradient-free Instructional Prompt Search

(3) Multiple Search Iterations

Base Instruction

[Your task] [is to classify the tweet] [as "positive" or "negative"] [based on its content].

(1) Get Phrase-splits and Perform Edits

edit: del

edit: swap

edit: par

Candidates

[Your task] [is to classify the tweet] [as "positive" or "negative"] [~~based on its content~~].

[Your task] [is to classify the tweet] [*based on its content*] [*as "positive" or "negative"*].

[Your task] [*is to label text*] [as "positive" or "negative"] [based on its content].

score: 64.0

score: 52.3

score: 57.6

(2) Score Candidates on $\mathcal{S}$

Best Candidate

[Your task] [is to classify the tweet] [as "positive" or "negative"].

Searched Instruction

[Your task] [as "positive" or "negative"].

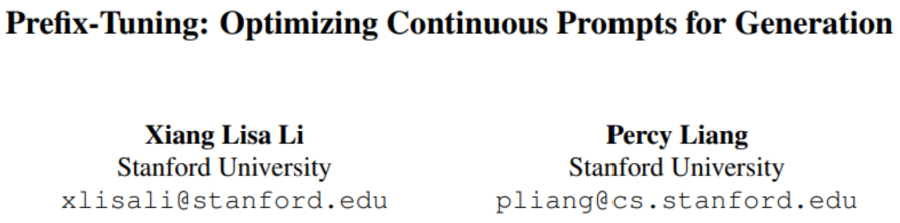
Prasad et al ‘23

Soft Prompting

Also called **continuous prompting**

Basic idea: insert some (non-language) parameters into prompt

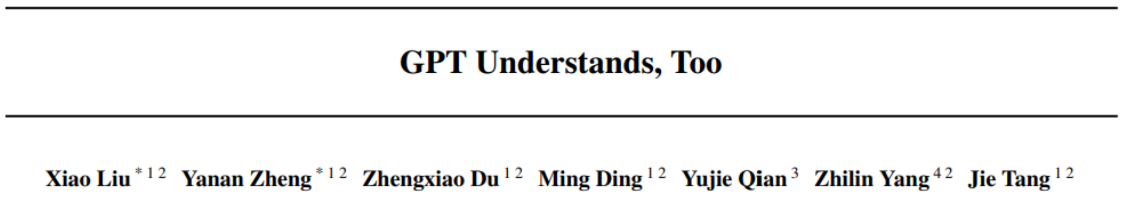
- Train these parameters
- Do not directly correspond to words in prompt



Prefix-Tuning: Optimizing Continuous Prompts for Generation

Xiang Lisa Li
Stanford University
xlisali@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu



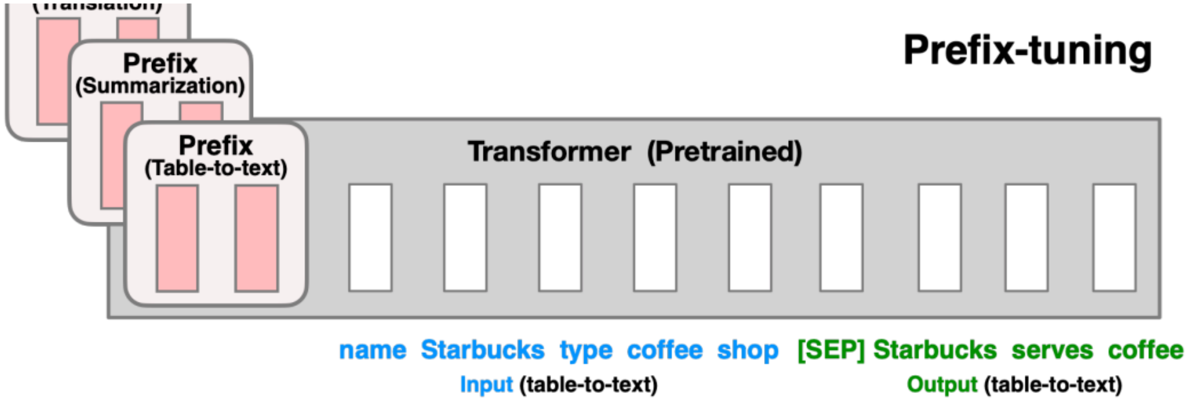
GPT Understands, Too

Xiao Liu[* 1 2] **Yanan Zheng**[* 1 2] **Zhengxiao Du**[1 2] **Ming Ding**[1 2] **Yujie Qian**[3] **Zhilin Yang**[4 2] **Jie Tang**[1 2]

Soft Prompting: **Prefix-Tuning**

Goal: create prefixes that *steer* models

- Prefixes are trainable parameters
- Train one for each goal task, only store these new parameters
- Enables cheap adaptation of frozen language model

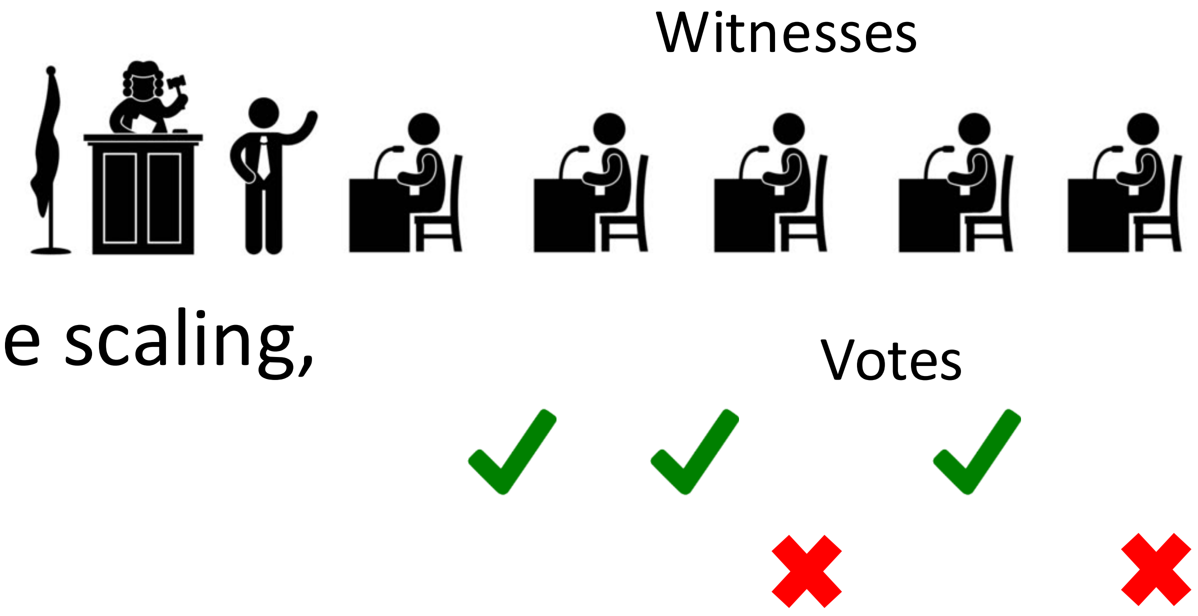


Li and Liang ‘21

Ensembling Prompts

One prompt can give you an answer… but might be wrong

- One simple approach: get multiple samples
- From?
  - Change temperature parameter
  - Vary your prompts



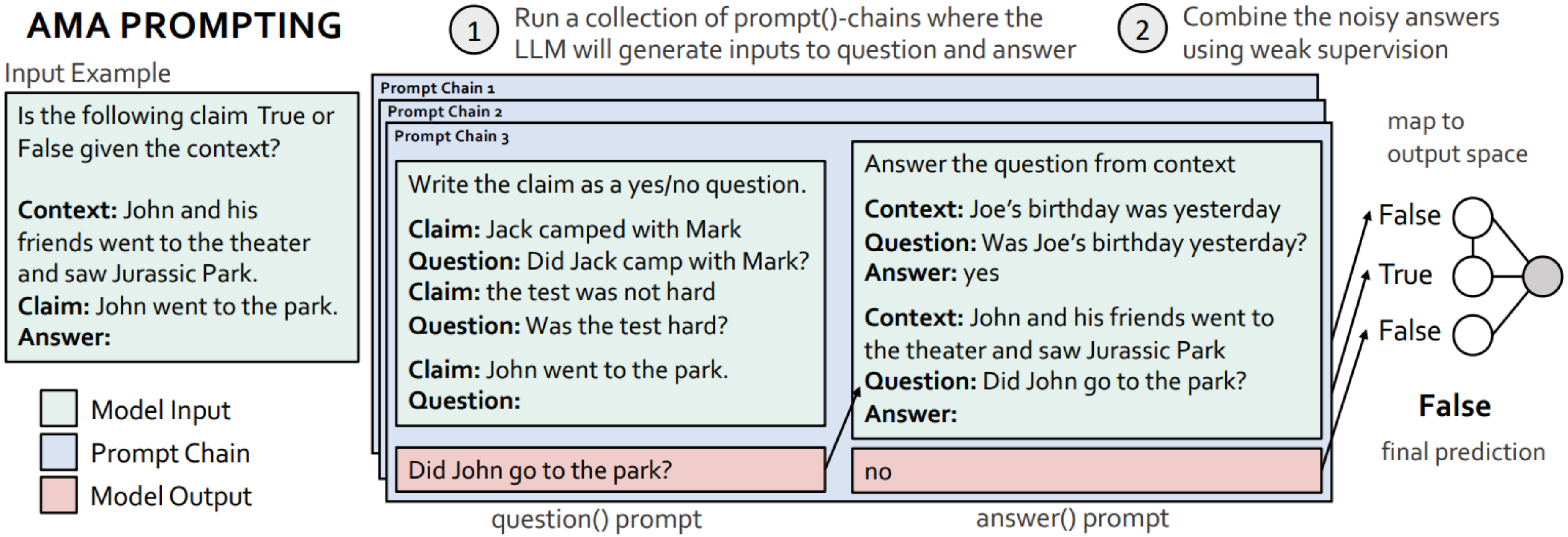
Then, run **majority vote**

- New terminology: “test-time scaling, self-consistency”

Ensembling Prompts: Weighted Version

Downside of majority vote… most responses might be wrong

- Should weight them by how accurate they are



Is the following claim True or False given the context?

Context: John and his friends went to the theater and saw Jurassic Park.
Claim: John went to the park.
Answer:

Write the claim as a yes/no question.

Claim: Jack camped with Mark
Question: Did Jack camp with Mark?
Claim: the test was not hard
Question: Was the test hard?

Claim: John went to the park.
Question:

Did John go to the park?

Answer the question from context

Context: Joe's birthday was yesterday
Question: Was Joe's birthday yesterday?
Answer: yes

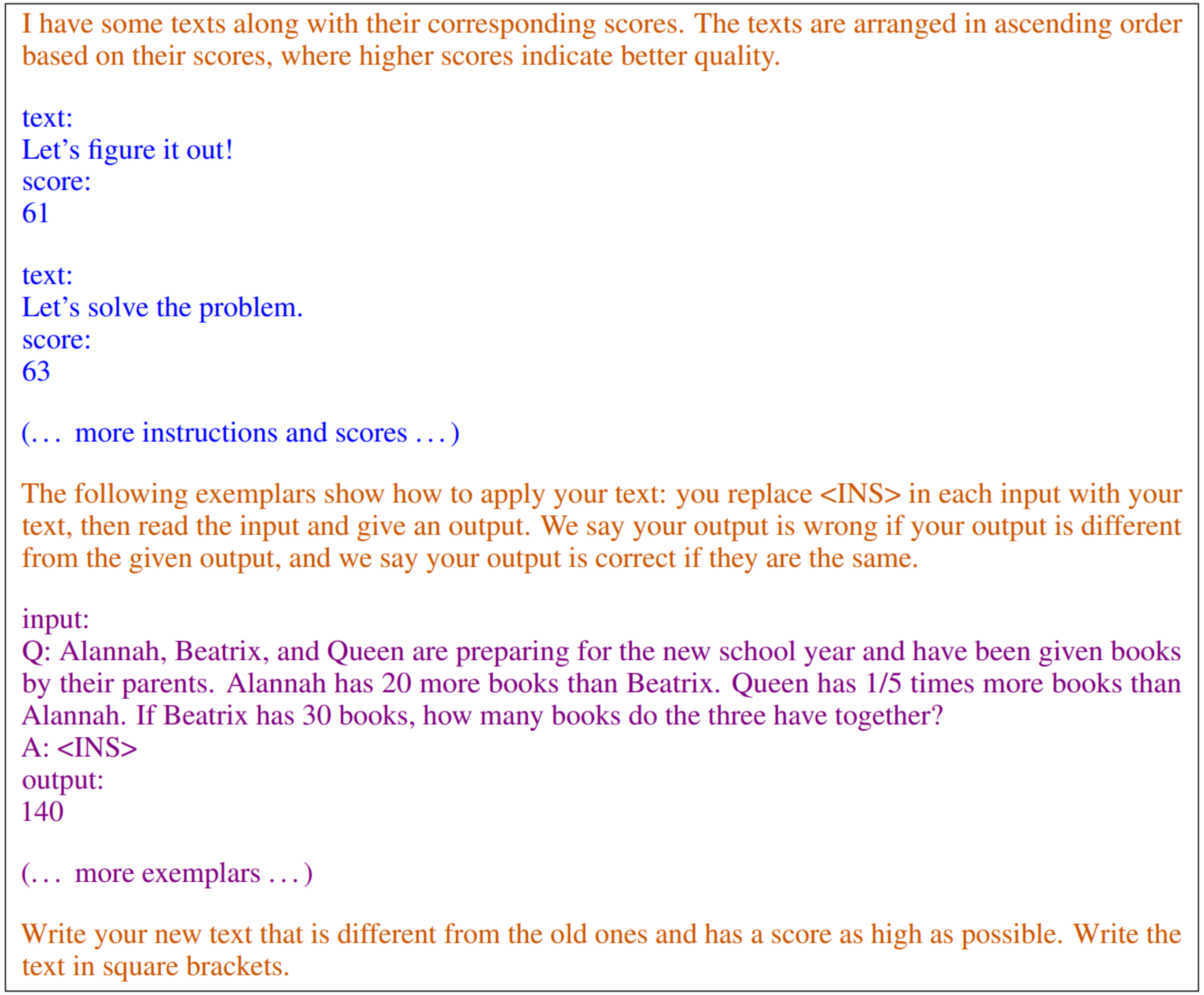
Context: John and his friends went to the theater and saw Jurassic Park
Question: Did John go to the park?
Answer:

no

Chen et al, '23

LLMs as Optimizers: Example

Example:



I have some texts along with their corresponding scores. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

text:
Let's figure it out!
score:
61

text:
Let's solve the problem.
score:
63

(... more instructions and scores ...)

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.

input:
Q: Alannah, Beatrix, and Queen are preparing for the new school year and have been given books by their parents. Alannah has 20 more books than Beatrix. Queen has 1/5 times more books than Alannah. If Beatrix has 30 books, how many books do the three have together?
A: <INS>
output:
140

(... more exemplars ...)

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

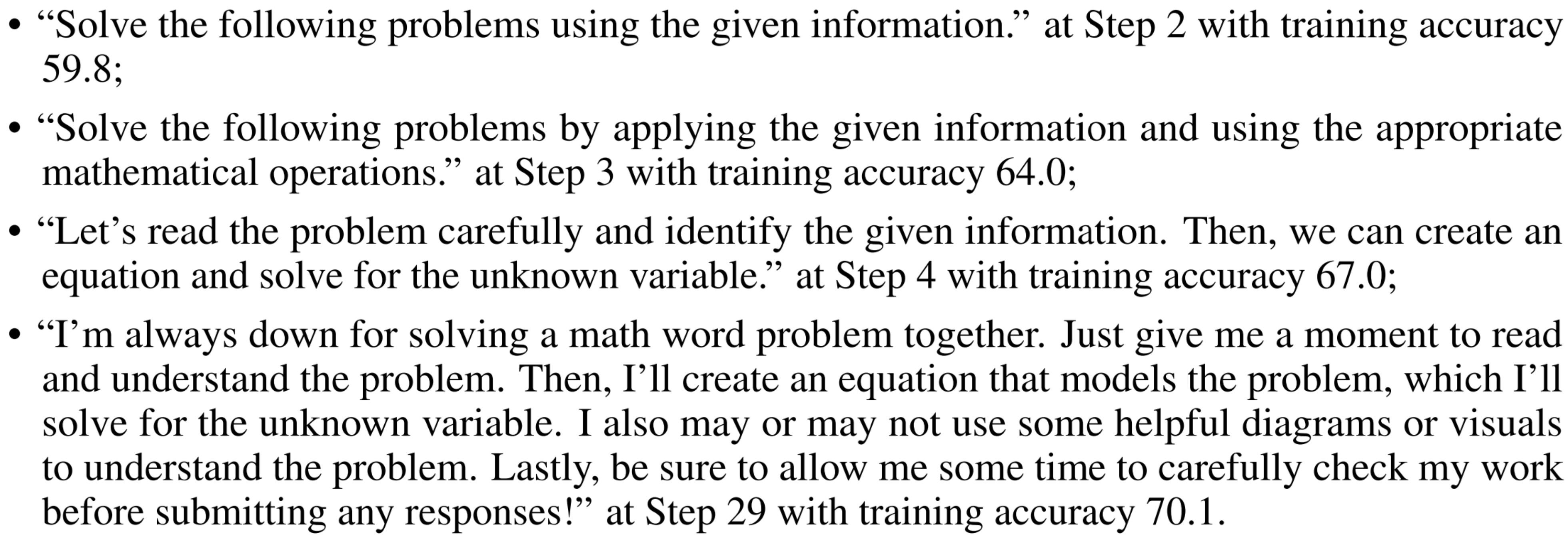
Meta-instructions

Trajectory points

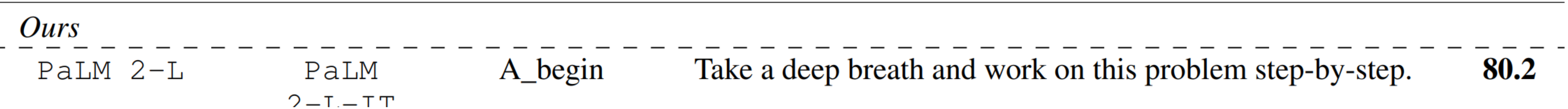
Problem to be solved

LLMs as Optimizers: **Prompt Optimization**

Resulting trajectory



- “Solve the following problems using the given information.” at Step 2 with training accuracy 59.8;
- “Solve the following problems by applying the given information and using the appropriate mathematical operations.” at Step 3 with training accuracy 64.0;
- “Let’s read the problem carefully and identify the given information. Then, we can create an equation and solve for the unknown variable.” at Step 4 with training accuracy 67.0;
- “I’m always down for solving a math word problem together. Just give me a moment to read and understand the problem. Then, I’ll create an equation that models the problem, which I’ll solve for the unknown variable. I also may or may not use some helpful diagrams or visuals to understand the problem. Lastly, be sure to allow me some time to carefully check my work before submitting any responses!” at Step 29 with training accuracy 70.1.



Ours				
PaLM 2-L	PaLM 2-L-IT	A_begin	Take a deep breath and work on this problem step-by-step.	**80.2**

More Auto-Prompting Work

Recall: search for hard prompts: tough optimization problem

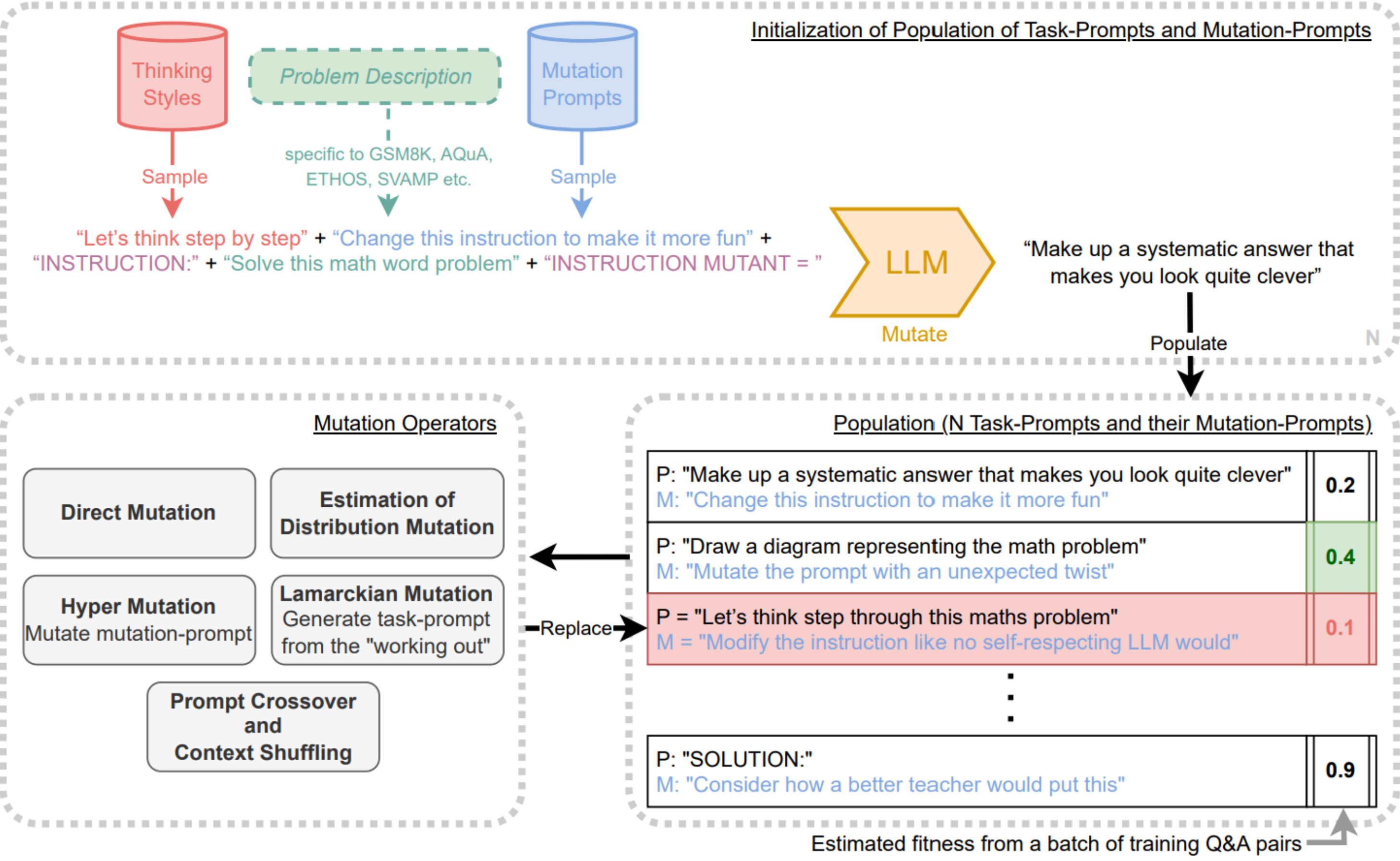
- Lots of classic search methods only require notion of “**neighbors**” and **evaluation** function access
  - Hill-climbing
  - Simulated annealing
  - Genetic algorithms

- “**Promptbreeder**”: an approach via genetic algorithms

- `Show all your working. II. You should use the correct mathematical notation and vocabulary, where appropriate. III. You should write your answer in full sentences and in words. IV. You should use examples to illustrate your points and prove your answers. V. Your workings out should be neat and legible`

More Auto-Prompting Work

“**Promptbreeder**”: an approach via genetic algorithms



Initialization of Population of Task-Prompts and Mutation-Prompts

“Let’s think step by step” + “Change this instruction to make it more fun” + “INSTRUCTION:” + “Solve this math word problem” + “INSTRUCTION MUTANT = ”

“Make up a systematic answer that makes you look quite clever”

Mutation Operators

- **Direct Mutation**
- **Estimation of Distribution Mutation**
- **Hyper Mutation** Mutate mutation-prompt
- **Lamarckian Mutation** Generate task-prompt from the "working out"
- **Prompt Crossover and Context Shuffling**

Population (N Task-Prompts and their Mutation-Prompts)

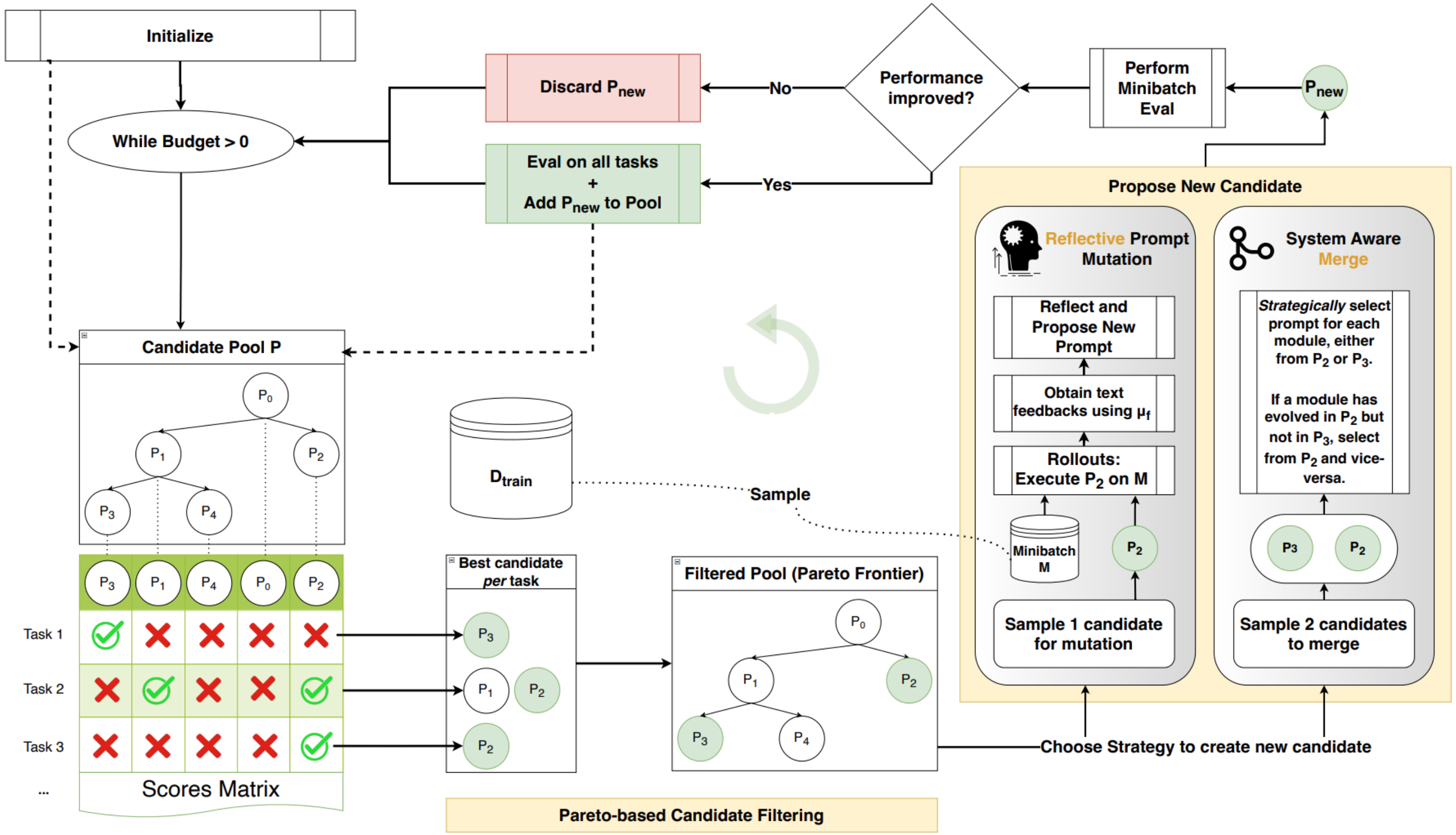
P: "Make up a systematic answer that makes you look quite clever" M: "Change this instruction to make it more fun"	0.2
P: "Draw a diagram representing the math problem" M: "Mutate the prompt with an unexpected twist"	0.4
P = "Let’s think step through this maths problem" M = "Modify the instruction like no self-respecting LLM would"	0.1
⋮	
P: "SOLUTION:" M: "Consider how a better teacher would put this"	0.9

Estimated fitness from a batch of training Q&A pairs

Fernando et al ‘23

More Auto-Prompting Work

“**GEPA**”: genetic algorithms + reflective iteration

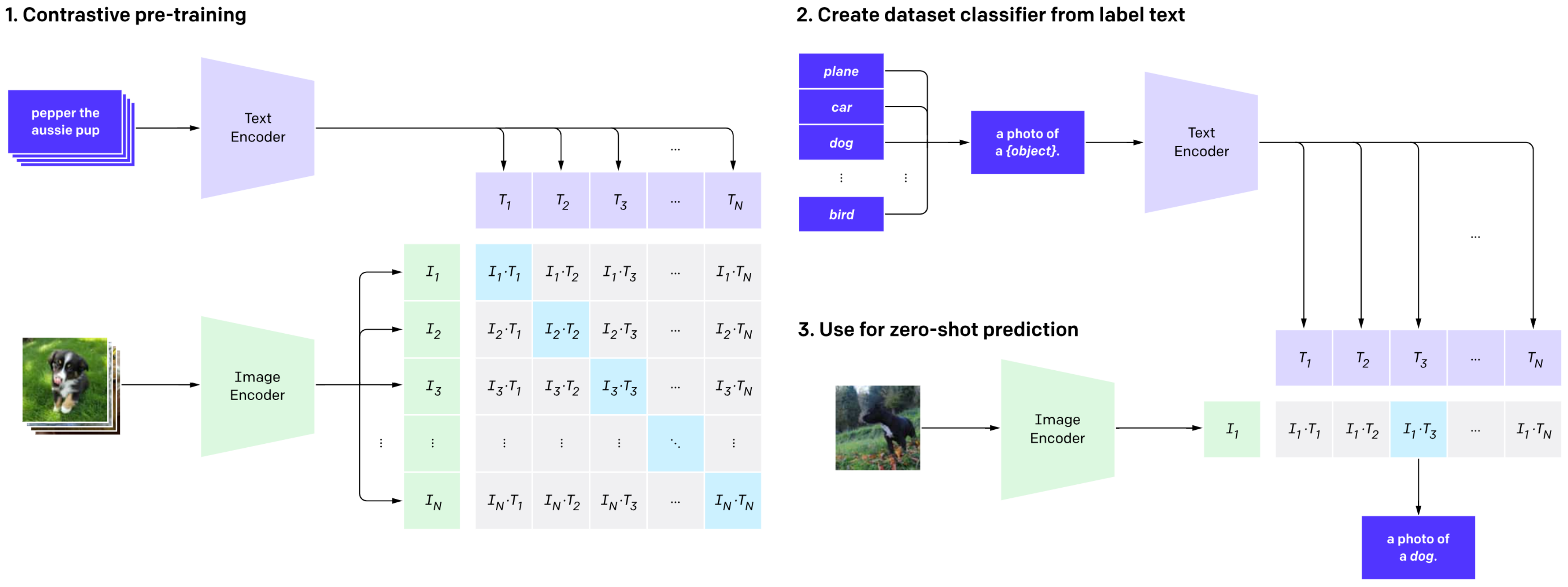

Initialize
While Budget > 0
Candidate Pool P
P_0
P_1
P_2
P_3
P_4
Discard P_{new}
No
Performance improved?
Perform Minibatch Eval
P_{new}
Yes
Eval on all tasks + Add P_{new} to Pool
D_{train}
Sample
Propose New Candidate
Reflective Prompt Mutation
Reflect and Propose New Prompt
Obtain text feedbacks using μ_f
Rollouts: Execute P_2 on M
Minibatch M
Sample 1 candidate for mutation
System Aware Merge
Strategically select prompt for each module, either from P_2 or P_3.
If a module has evolved in P_2 but not in P_3, select from P_2 and vice-versa.
Sample 2 candidates to merge
Best candidate *per* task
Filtered Pool (Pareto Frontier)
Choose Strategy to create new candidate
Pareto-based Candidate Filtering
Scores Matrix

	P_3	P_1	P_4	P_0	P_2
Task 1	✓	✗	✗	✗	✗
Task 2	✗	✓	✗	✗	✓
Task 3	✗	✗	✗	✗	✓
...					

Agrawal et al ‘26

Prompting VLMs and Multimodal Models

Recall training and prediction in CLIP-style VLMs

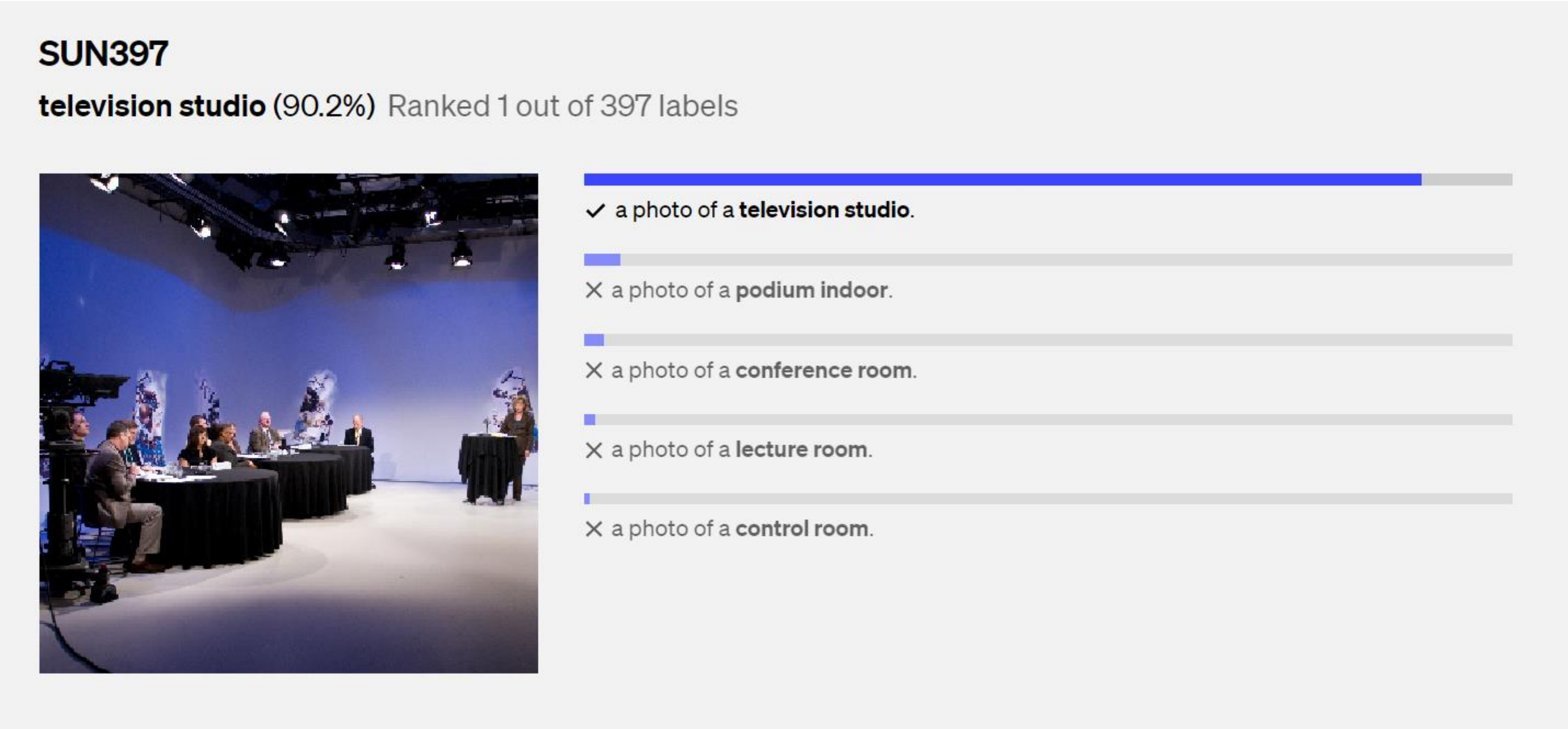

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.

OpenAI

How to Prompt VLMs?

Standard way: use pre-defined templates

- E.g., “a photo of a [X]”
- But, might struggle…



SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio**.

✗ a photo of a **podium indoor**.

✗ a photo of a **conference room**.

✗ a photo of a **lecture room**.

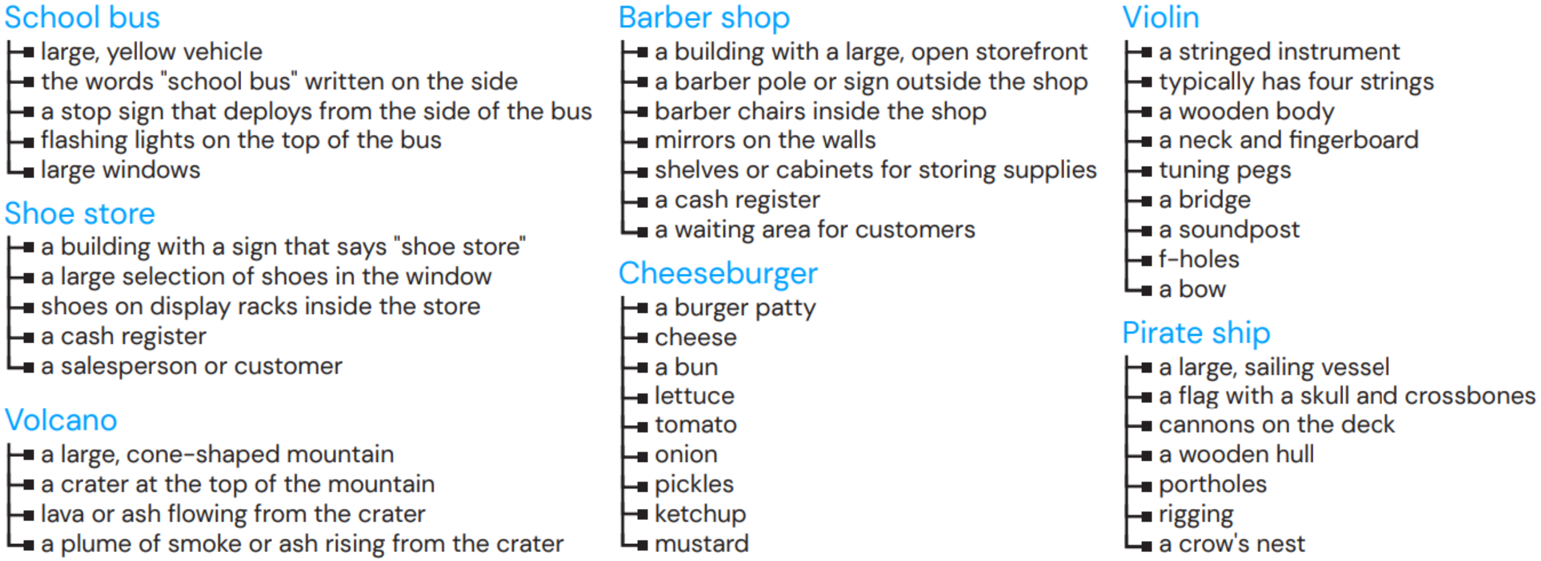
✗ a photo of a **control room**.

OpenAI

LLMs to Improve VLMs: Description

Static class descriptions may fail…

- Replace with descriptive features (Menon and Vondrick, ‘23)
  - Instead of “tiger”, include “stripes, claws, …”



School bus
- large, yellow vehicle
- the words "school bus" written on the side
- a stop sign that deploys from the side of the bus
- flashing lights on the top of the bus
- large windows

Shoe store
- a building with a sign that says "shoe store"
- a large selection of shoes in the window
- shoes on display racks inside the store
- a cash register
- a salesperson or customer

Volcano
- a large, cone-shaped mountain
- a crater at the top of the mountain
- lava or ash flowing from the crater
- a plume of smoke or ash rising from the crater

Barber shop
- a building with a large, open storefront
- a barber pole or sign outside the shop
- barber chairs inside the shop
- mirrors on the walls
- shelves or cabinets for storing supplies
- a cash register
- a waiting area for customers

Cheeseburger
- a burger patty
- cheese
- a bun
- lettuce
- tomato
- onion
- pickles
- ketchup
- mustard

Violin
- a stringed instrument
- typically has four strings
- a wooden body
- a neck and fingerboard
- tuning pegs
- a bridge
- a soundpost
- f-holes
- a bow

Pirate ship
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest

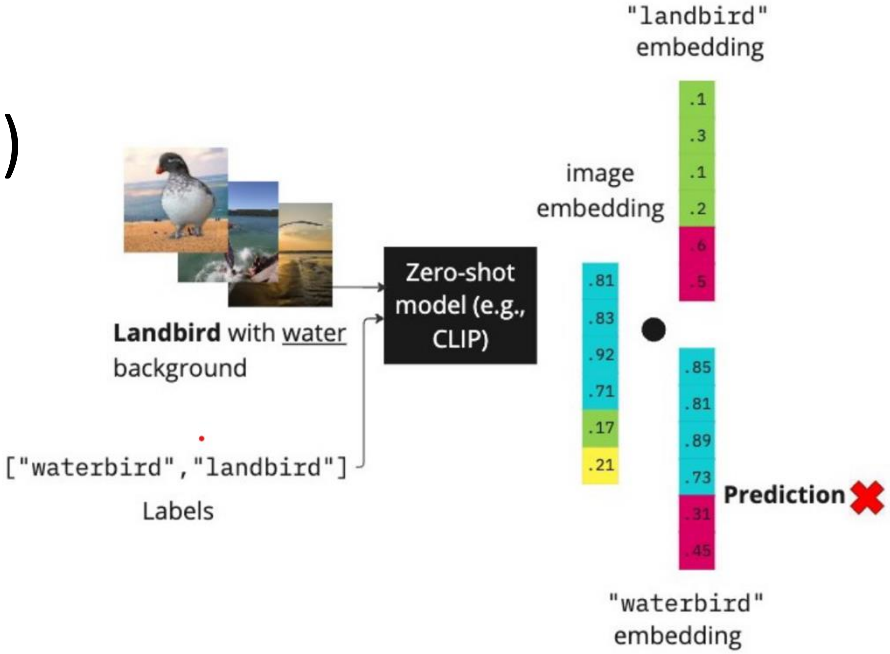
Figure 3: Examples of descriptor schema produced by GPT-3.

LLMs to Improve VLMs: **Spurious Features**

This helped with positives.

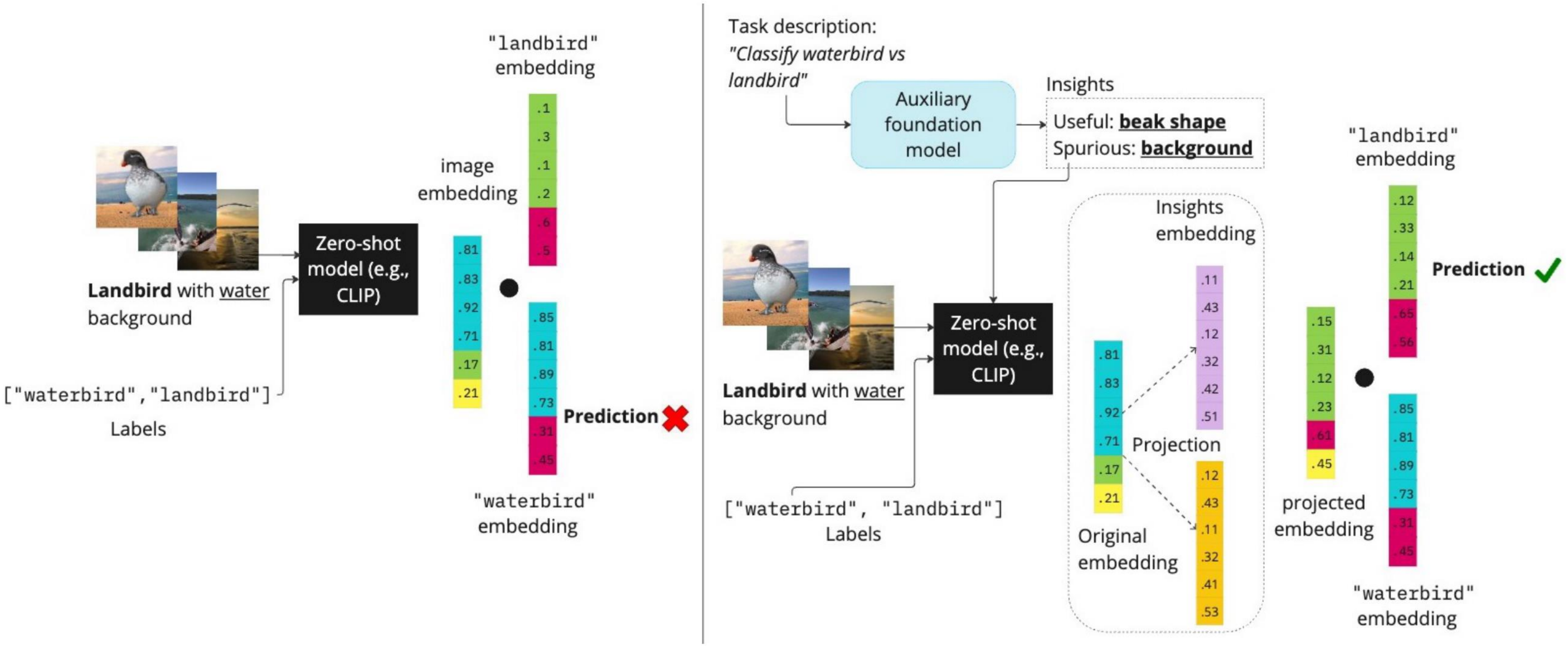
- What about **negatives** (i.e., spurious features?)
  - Example: waterbirds with CLIP

- Spurious correlations: generally a problem with all pretrained models
  - But LLMs can also tell us about this (Adila ‘23)



LLMs to Improve VLMs: **Spurious Features**

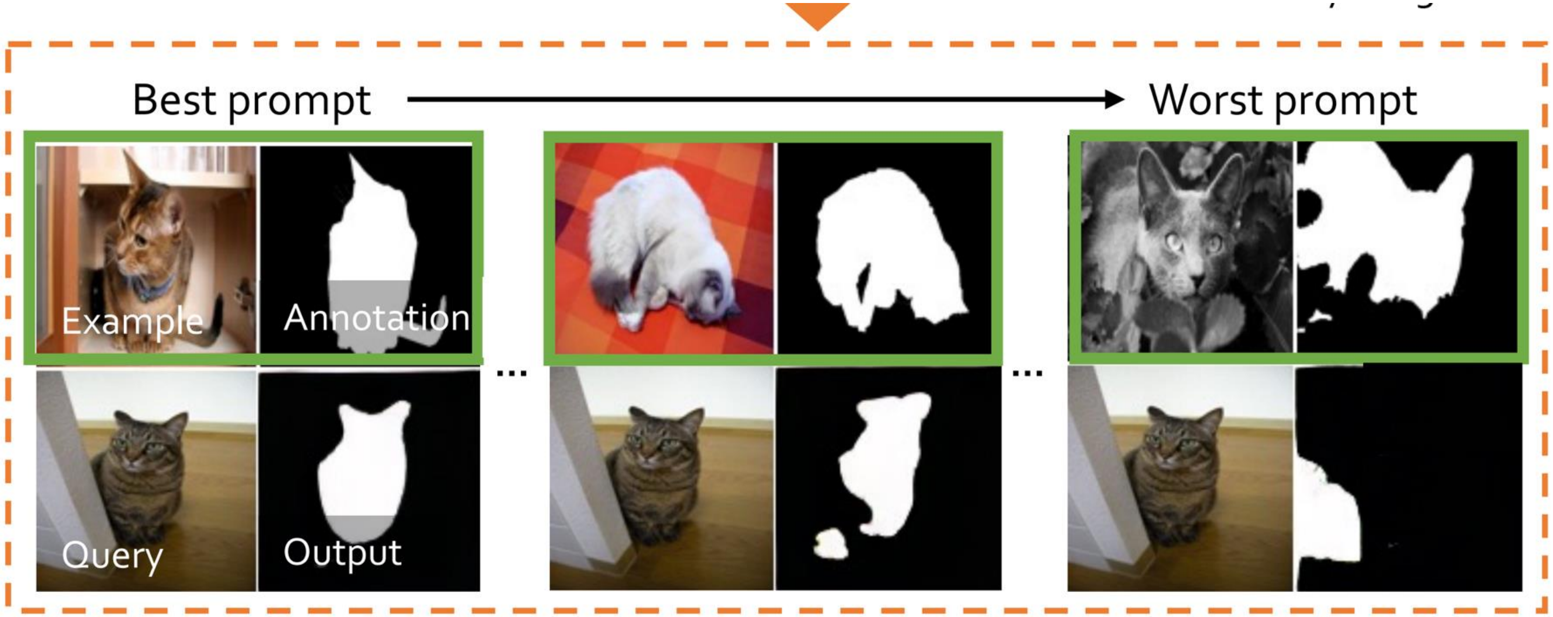
Modify embeddings used for prediction:



In-Context Learning for VLMs

Surprisingly in-context/few-shot also applies to visual models

- Standard questions apply here too:
  - How to select examples
  - What makes for a good example?

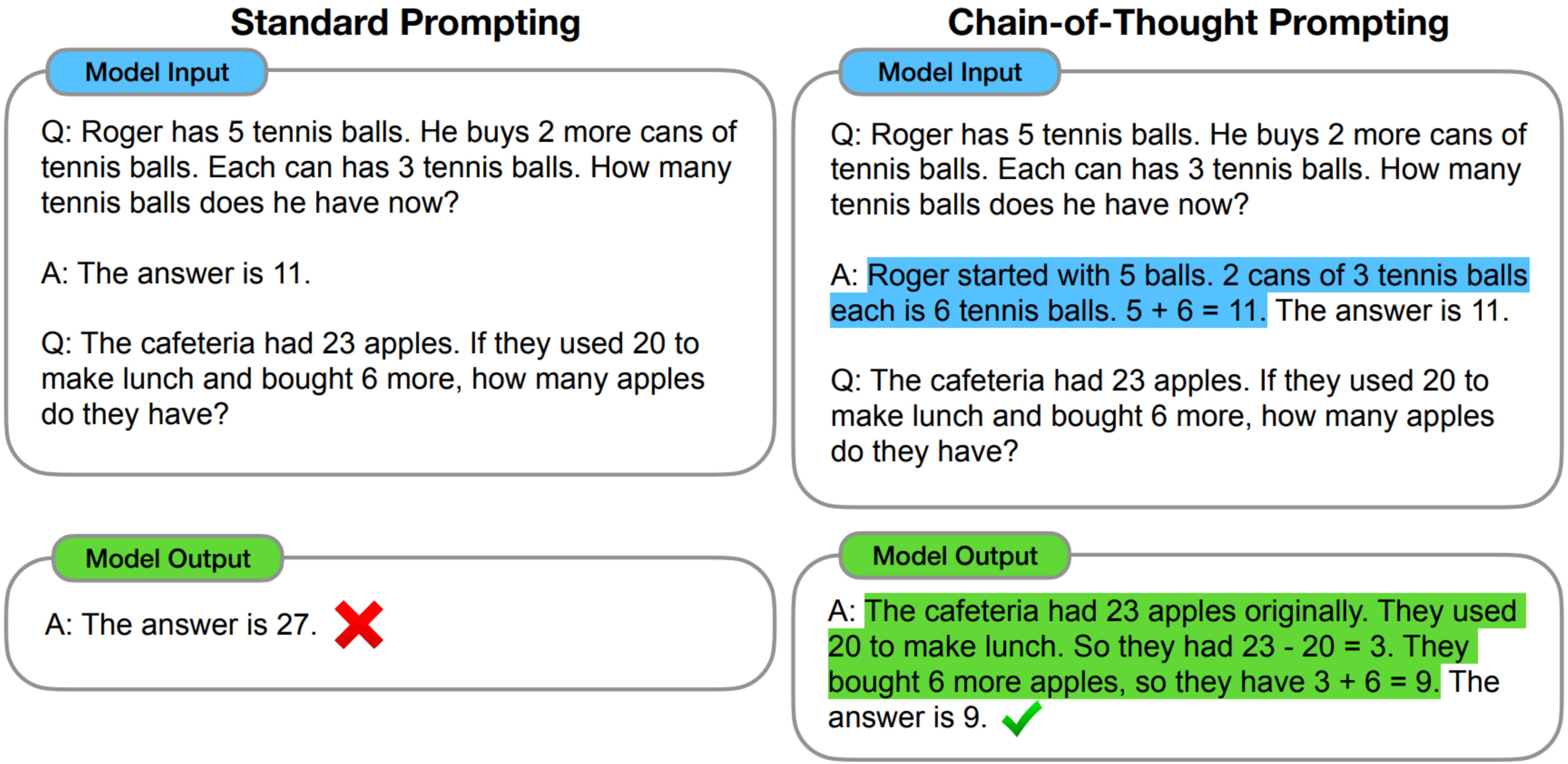


(a) Visual in-context learning is sensitive to prompt selection

Zhang et al ‘23

Chain-of-Thought

Performing complex reasoning is hard. Help the model:



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

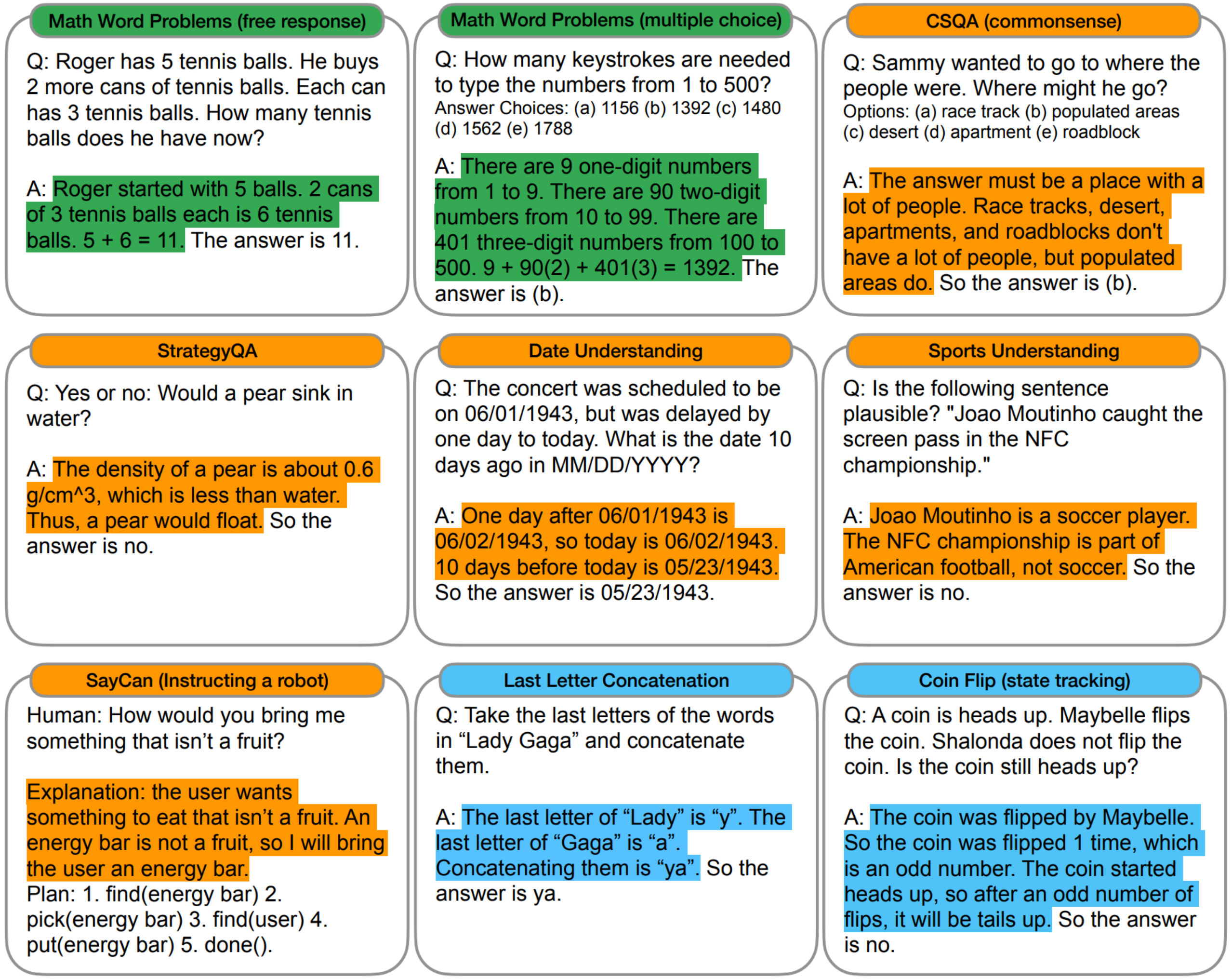
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Wei et al '22

Chain-of-Thought: **Applications**

Some things it can be used for:

- **Math problems**
- **Commonsense reasoning**
- **Symbolic reasoning**



Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

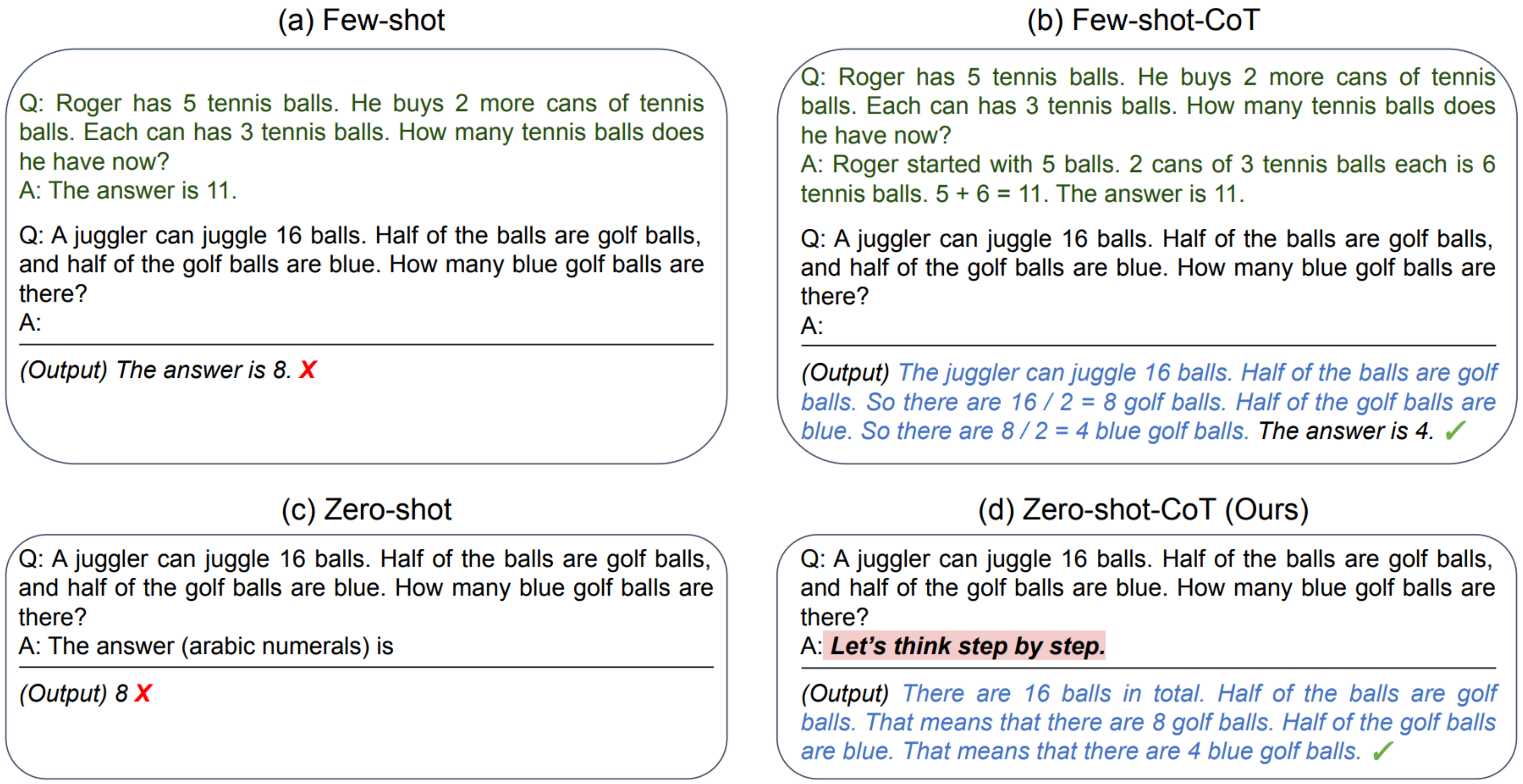
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Chain-of-Thought: **Zero-Shot**

No examples shown; encourage model to decompose

- Add to prompt: “Let’s think step by step” before each answer
- For answer extraction, add prompts like “Therefore, the answer (arabic numerals) is” (Kojima et al ‘23)



(a) Few-shot

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ***X***

(b) Few-shot-CoT

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ***X***

(d) Zero-shot-CoT (Ours)

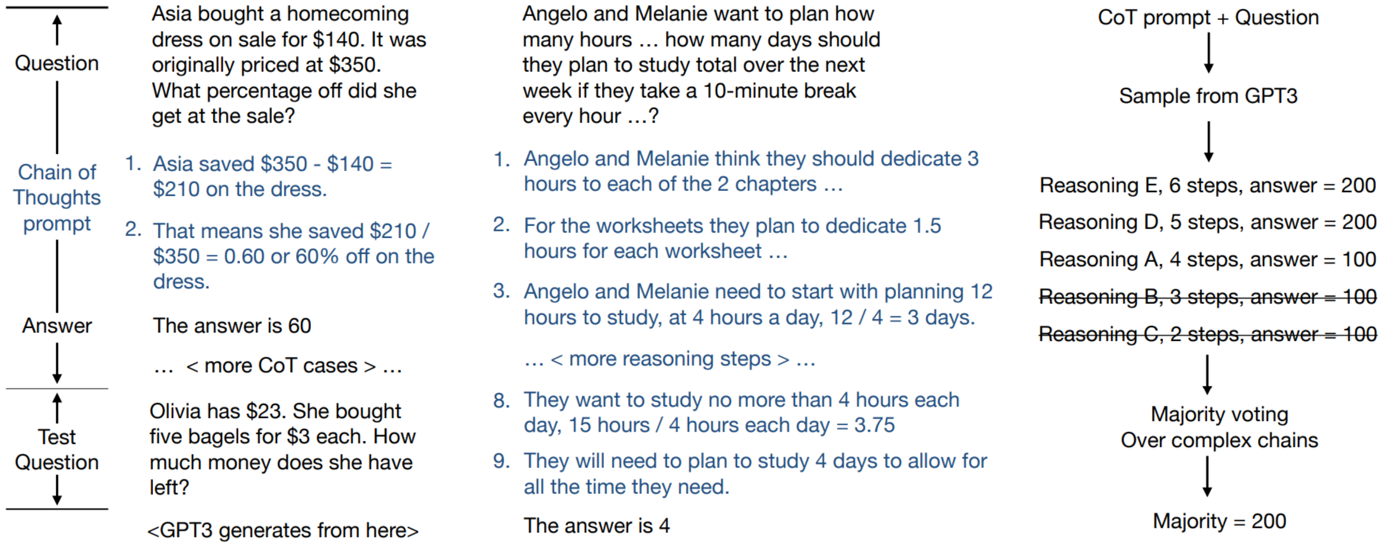
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let’s think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Chain-of-Thought: **Few-Shot Examples**

As before, we must choose few-shot examples.

- More structured than simple semantic similarity
- *Complexity-based* prompting.
  - "[S]imply choose complex prompts over simple ones."
- Prompting: include most steps. Ensembling: MV over set of most complex chains.



Asia bought a homecoming dress on sale for $140. It was originally priced at $350. What percentage off did she get at the sale?

1. Asia saved $350 - $140 = $210 on the dress.
2. That means she saved $210 / $350 = 0.60 or 60% off on the dress.

The answer is 60

… < more CoT cases > …

Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

<GPT3 generates from here>

A. Workflow of chain of thoughts prompting

Angelo and Melanie want to plan how many hours … how many days should they plan to study total over the next week if they take a 10-minute break every hour …?

1. Angelo and Melanie think they should dedicate 3 hours to each of the 2 chapters …
2. For the worksheets they plan to dedicate 1.5 hours for each worksheet …
3. Angelo and Melanie need to start with planning 12 hours to study, at 4 hours a day, 12 / 4 = 3 days.

… < more reasoning steps > …

8. They want to study no more than 4 hours each day, 15 hours / 4 hours each day = 3.75
9. They will need to plan to study 4 days to allow for all the time they need.

The answer is 4

B. Example complex chain, 9 reasoning steps

CoT prompt + Question

Sample from GPT3

Reasoning E, 6 steps, answer = 200
Reasoning D, 5 steps, answer = 200
Reasoning A, 4 steps, answer = 100
~~Reasoning B, 3 steps, answer = 100~~
~~Reasoning C, 2 steps, answer = 100~~

Majority voting Over complex chains

Majority = 200

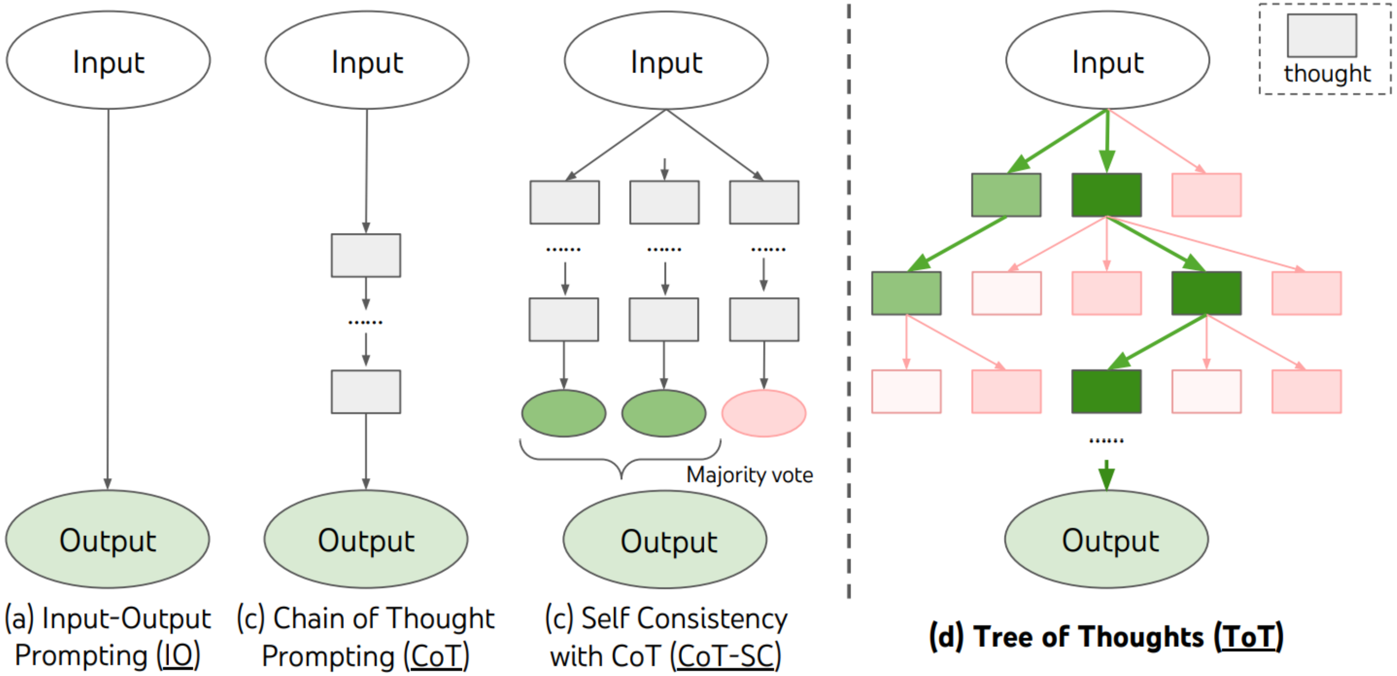
C. Complexity-based consistency

Fu et al '23

Chain-of-Thought: **Generalizations**

How do we really "reason"?

- Not really by sampling a bunch of chains…



(a) Input-Output Prompting (IO)

(c) Chain of Thought Prompting (CoT)

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

(d) Tree of Thoughts (ToT)

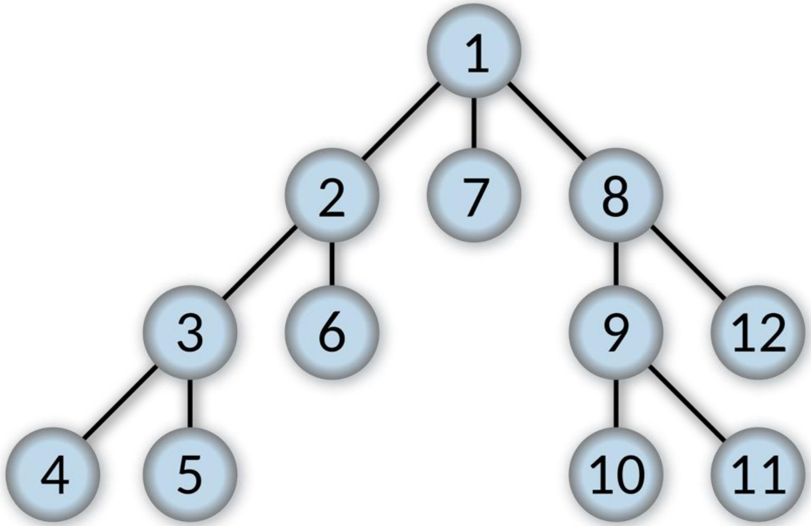
Yao et al '23

Chain-of-Thought: **Generalizations**

Tree-of-thoughts **basic idea**:

- **Notation**: thoughts z_1, z_2, …, z_n bridge x and y
- Comparison to other methods:
  - Vanilla CoT: sample $z_i \sim p_\theta(z_i \mid x, z_1,\ldots,z_{i-1})$, $y \sim p_\theta(y \mid x, z_1,\ldots,z_n)$
  - CoT Self-Consistency: sample multiple times, take majority vote

- Idea: create a state $s=[x,z_1,\ldots,z_n]$
- Generate multiple candidates for next state
  - Then run standard search (i.e., BFS, DFS, A*)

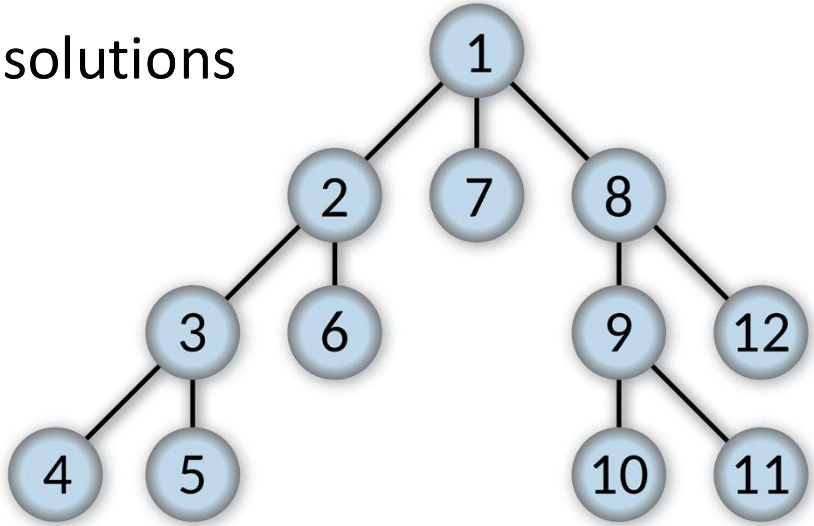


Drichel (Wiki)

Chain-of-Thought: **Generalizations**

Tree-of-thoughts **key aspects**:

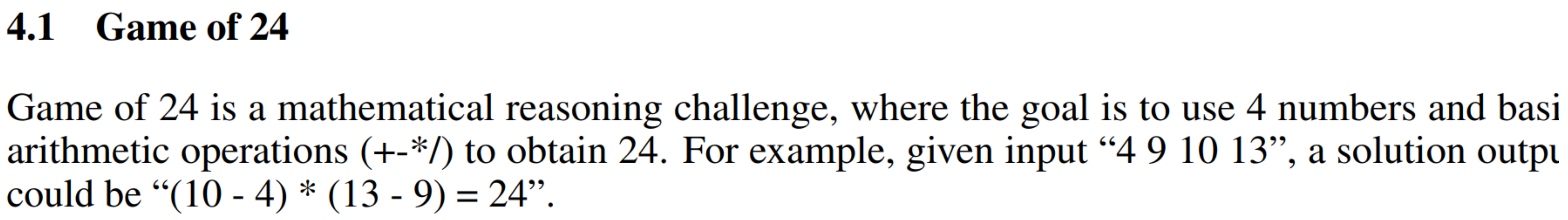
- **Thought decomposition:** how big *z*s should be
- **Thought generation:** obtaining the next sample
  - Try to avoid duplication
- **State evaluation**: How close are we to solution?
  - Recall heuristics for search from CS 540
  - Either use LM itself, or vote/weighted vote across solutions
- **Search**: BFS or DFS
  - Or more advanced search methods



Drichel (Wiki)

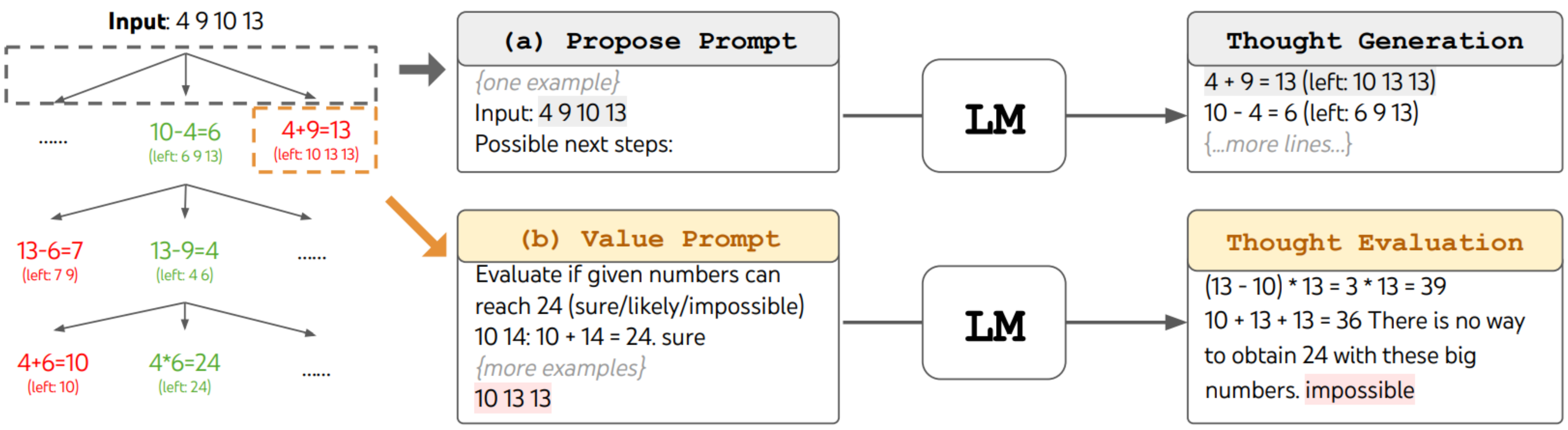
Chain-of-Thought: **Generalizations**

Tree-of-thoughts **example:**



4.1 Game of 24

Game of 24 is a mathematical reasoning challenge, where the goal is to use 4 numbers and basi arithmetic operations (+-*/) to obtain 24. For example, given input "4 9 10 13", a solution outpu could be "(10 - 4) * (13 - 9) = 24".



(a) Propose Prompt

{one example}
Input: 4 9 10 13
Possible next steps:

Thought Generation

4 + 9 = 13 (left: 10 13 13)
10 - 4 = 6 (left: 6 9 13)
{...more lines...}

(b) Value Prompt

Evaluate if given numbers can reach 24 (sure/likely/impossible)
10 14: 10 + 14 = 24. sure
{more examples}
10 13 13

Thought Evaluation

(13 - 10) * 13 = 3 * 13 = 39
10 + 13 + 13 = 36 There is no way to obtain 24 with these big numbers. impossible

Thank You!