

Prompt Engineering with Large Language Models: Basics and Research Applications

Shane Storks (he/him)

PhD Candidate in Computer Science and Engineering

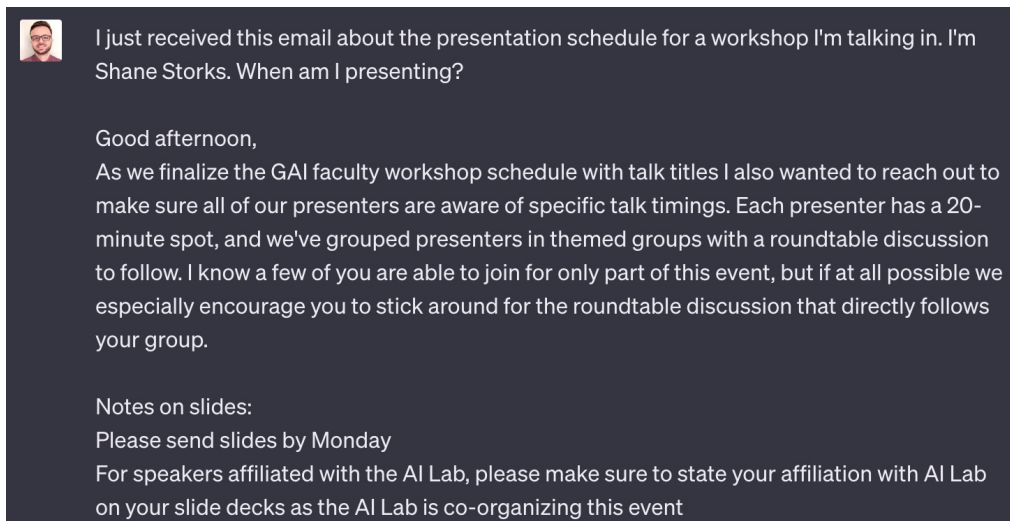
GAI for Research Faculty Workshop

July 26, 2023



Prompting Large Language Models (LLMs)

LLMs like ChatGPT and GPT-4 have recently gained popularity due to the ease of **prompting** them for information using free-form language.

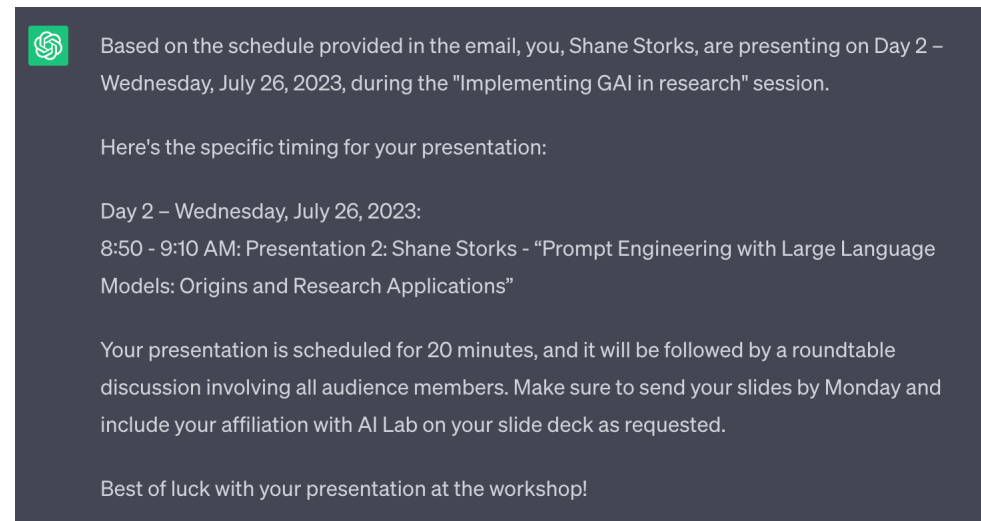


I just received this email about the presentation schedule for a workshop I'm talking in. I'm Shane Storks. When am I presenting?

Good afternoon,
As we finalize the GAI faculty workshop schedule with talk titles I also wanted to reach out to make sure all of our presenters are aware of specific talk timings. Each presenter has a 20-minute spot, and we've grouped presenters in themed groups with a roundtable discussion to follow. I know a few of you are able to join for only part of this event, but if at all possible we especially encourage you to stick around for the roundtable discussion that directly follows your group.

Notes on slides:
Please send slides by Monday
For speakers affiliated with the AI Lab, please make sure to state your affiliation with AI Lab on your slide decks as the AI Lab is co-organizing this event

...



Based on the schedule provided in the email, you, Shane Storks, are presenting on Day 2 – Wednesday, July 26, 2023, during the "Implementing GAI in research" session.

Here's the specific timing for your presentation:

Day 2 – Wednesday, July 26, 2023:
8:50 - 9:10 AM: Presentation 2: Shane Storks - "Prompt Engineering with Large Language Models: Origins and Research Applications"

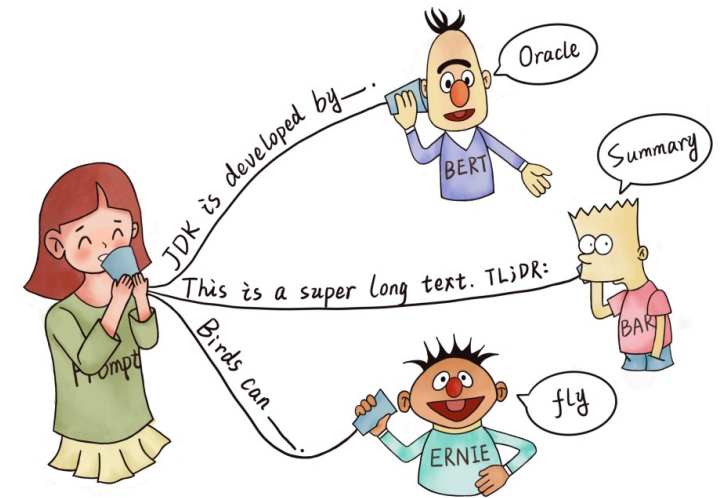
Your presentation is scheduled for 20 minutes, and it will be followed by a roundtable discussion involving all audience members. Make sure to send your slides by Monday and include your affiliation with AI Lab on your slide deck as requested.

Best of luck with your presentation at the workshop!

How can we apply them to empirical research?

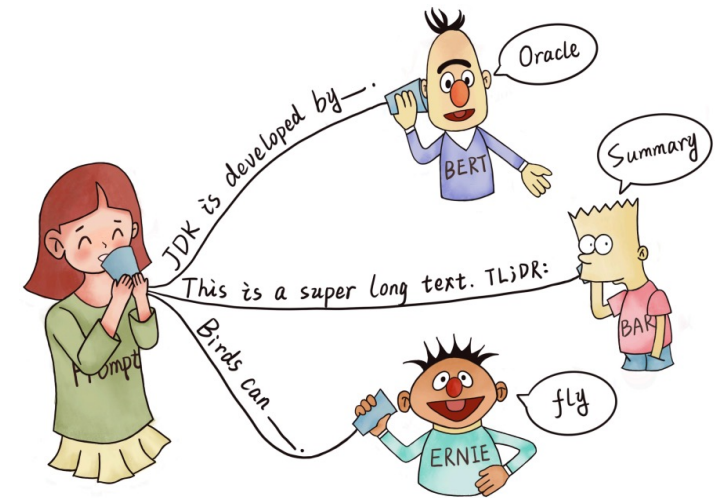
Outline

- Basics of Prompt Engineering
- Application 1: Analogical Reasoning
- Application 2: Physical Commonsense Reasoning



Outline

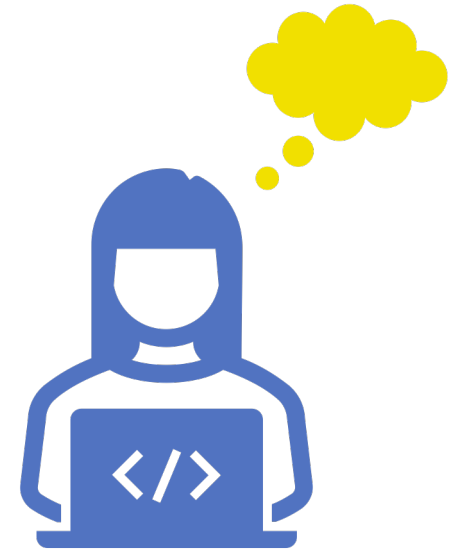
- **Basics of Prompt Engineering**
- Application 1: Analogical Reasoning
- Application 2: Physical Commonsense Reasoning



Prompt Engineering

Given a pre-trained LLM, make a few choices to apply to your problem:

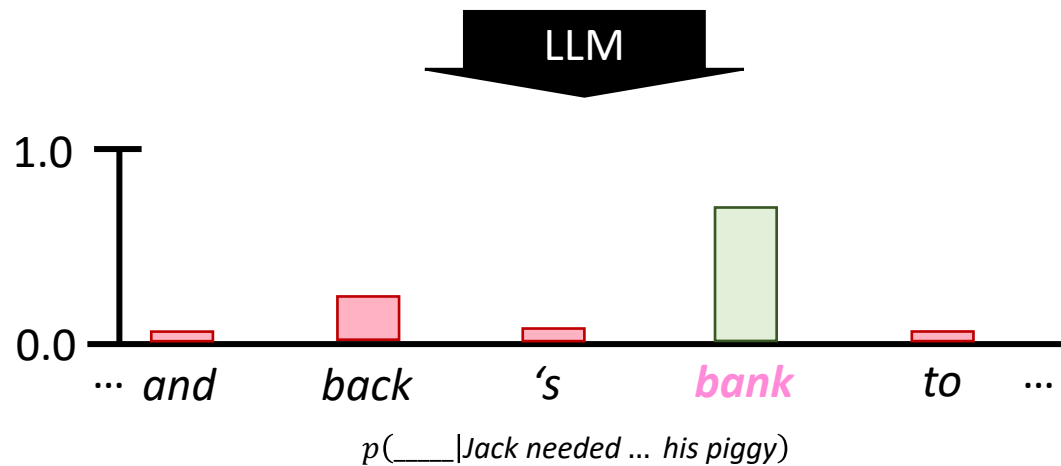
1. Prompt template
2. Answer mapping
3. In-context demonstration



Language Models

$$p(w_n | w_1, w_2, \dots, w_{n-1})$$

Jack needed some **money**, so he went and shook his **piggy** _____



Prompt Templates

If filling a blank from a few possible choices, can use a **cloze prompt**:

Task	Inputs ([X])	Template	Answer ([Z])
------	--------------	----------	--------------

Prompt Templates

When completing a prompt or generating text, use a **prefix prompt**:

Task	Inputs ([X])	Template	Answer ([Z])
------	--------------	----------	--------------

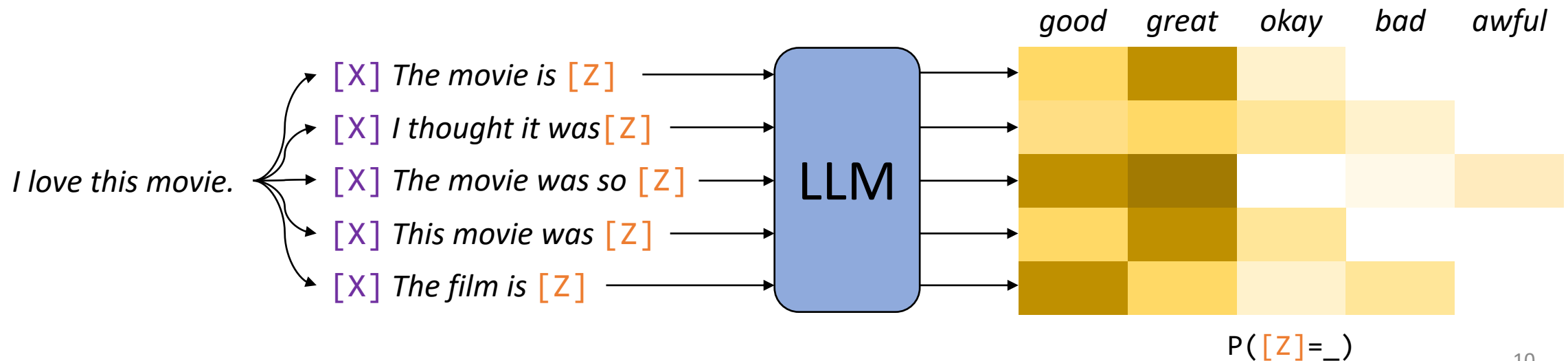
Prompt Templates

When completing a prompt or generating text, use a **prefix prompt**:

Task	Inputs ([X])	Template	Answer ([Z])
------	--------------	----------	--------------

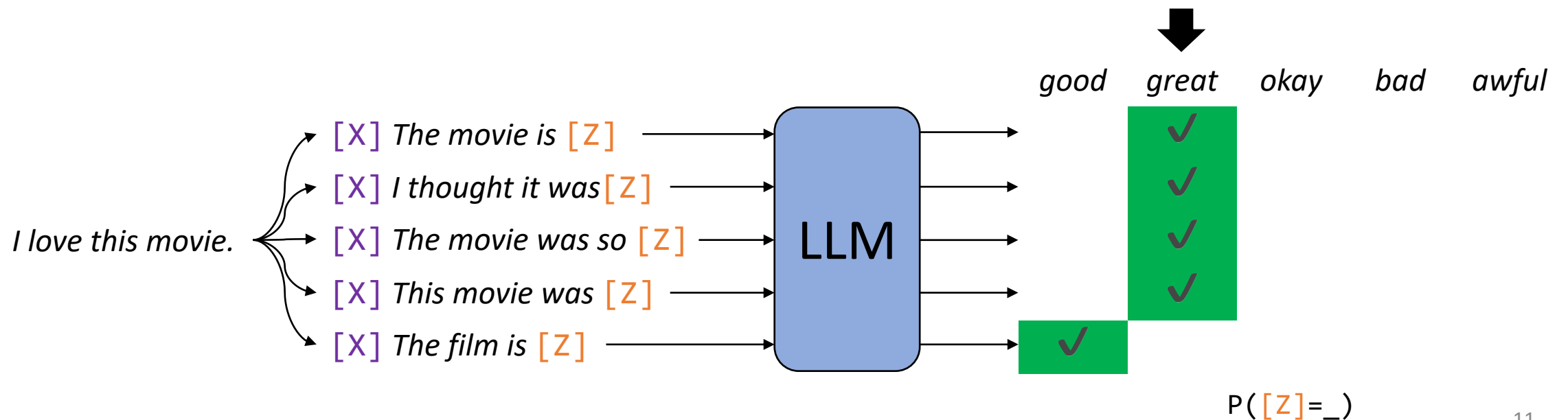
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates



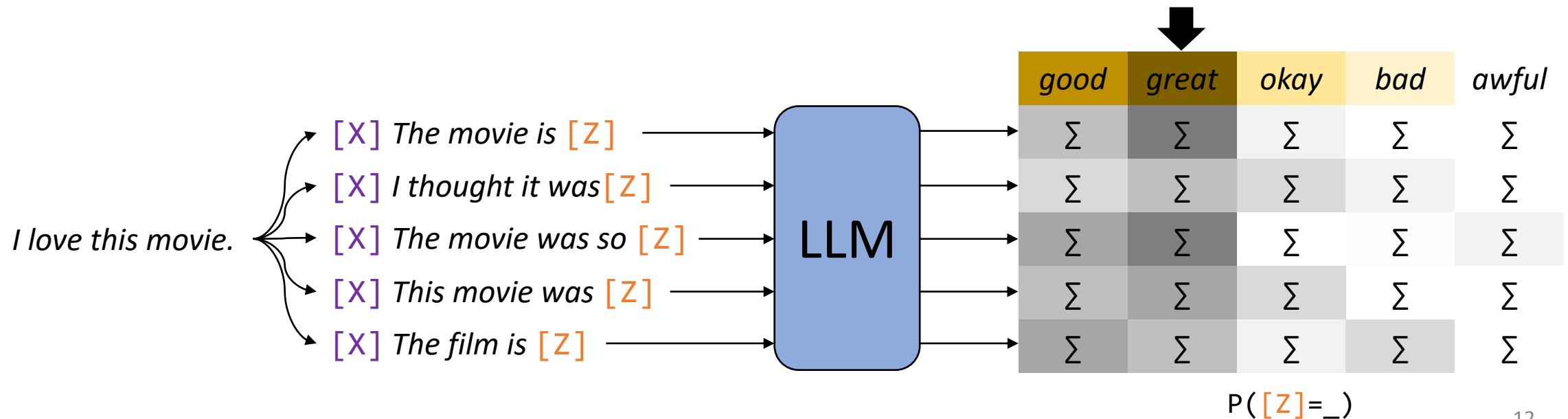
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates



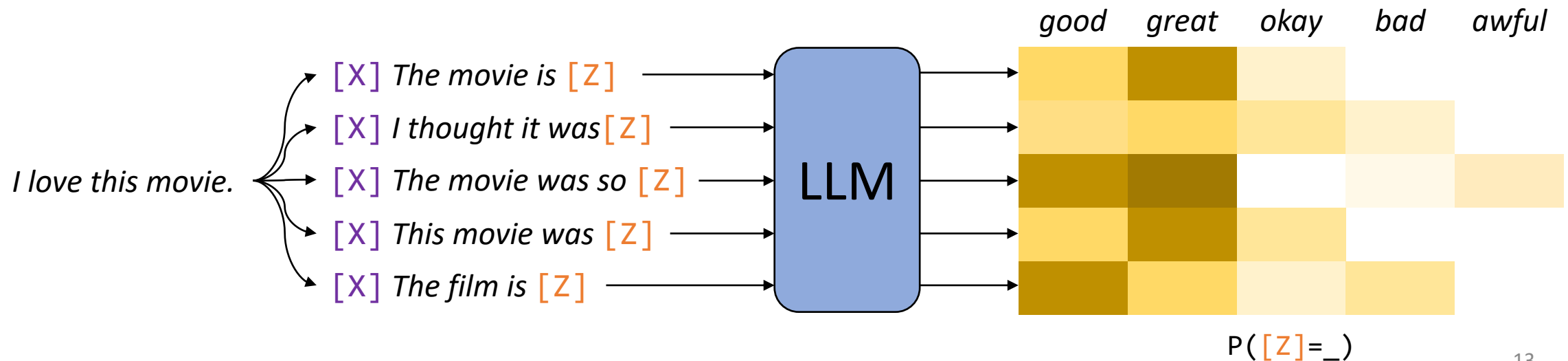
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates



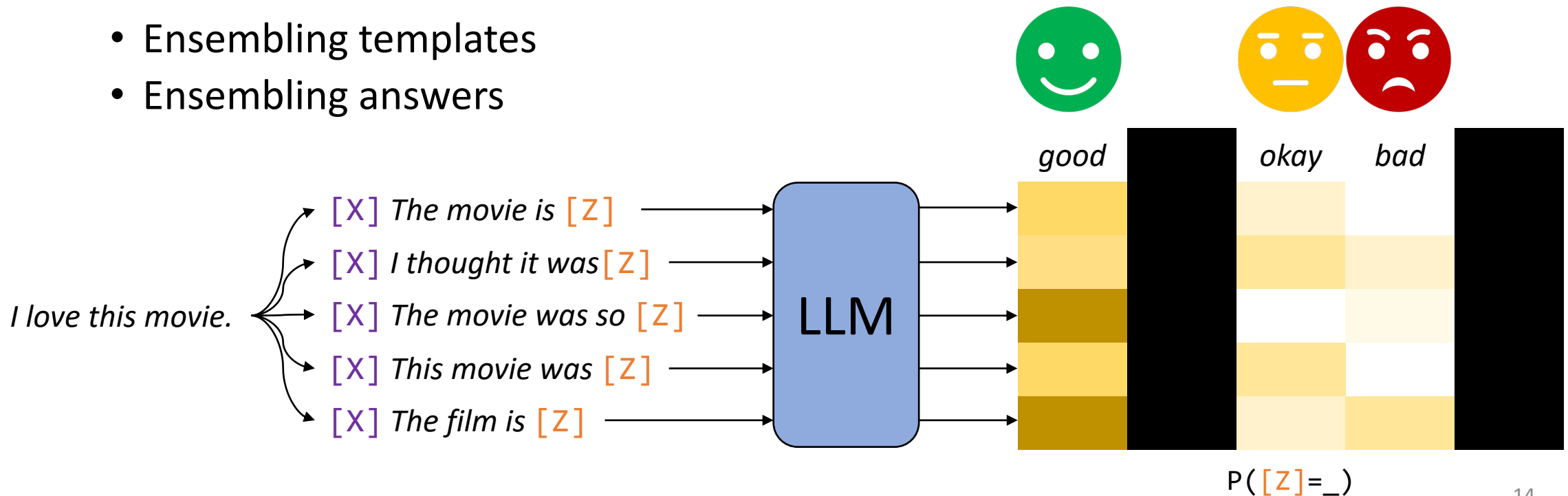
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates
 - Ensembling answers



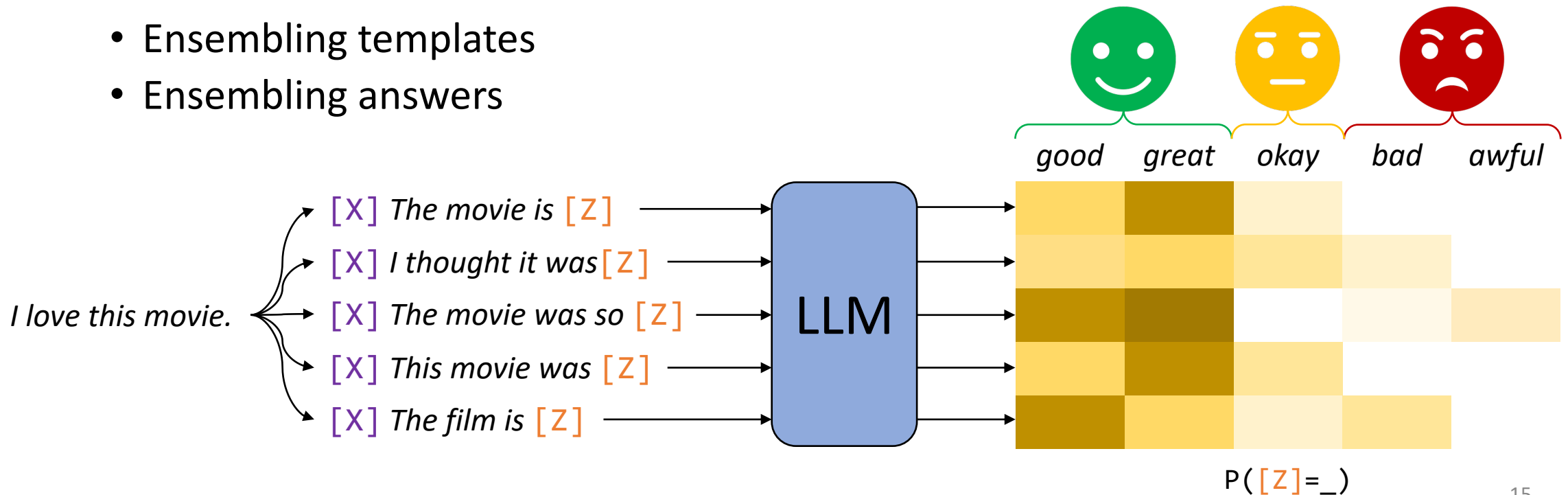
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates
 - Ensembling answers



Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
 - Trial and error
 - Ensembling templates
 - Ensembling answers



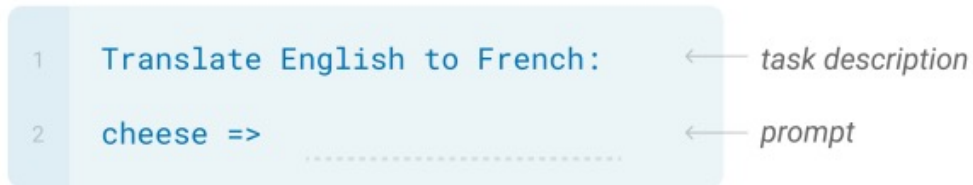
Managing Randomness in LLMs

- LLM decoding algorithms may incorporate some randomness by default to increase the diversity of generation
- Some solutions:
 - Generate multiple times and average results
 - Greedy decoding

In-Context Learning

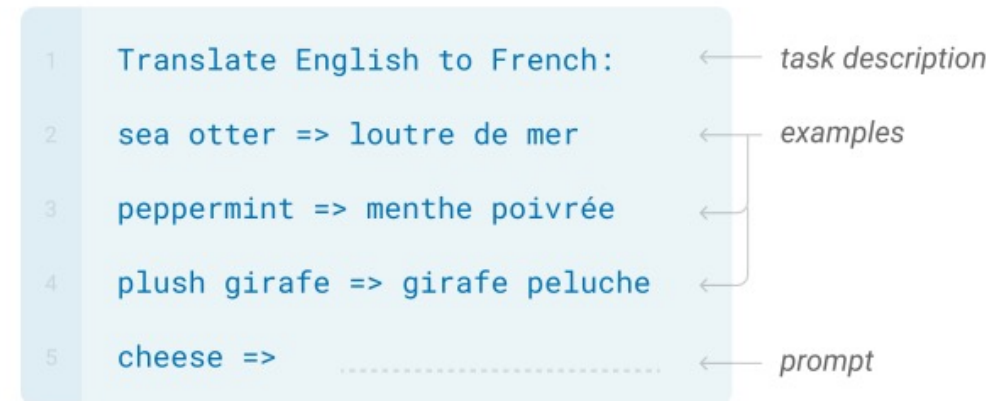
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Chain-of-Thought Prompting

Standard Prompting

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

Input

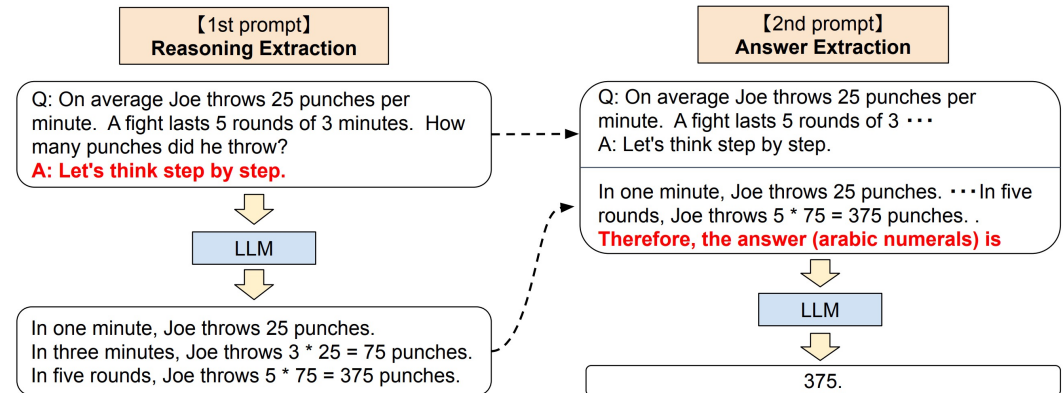
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

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. ✅

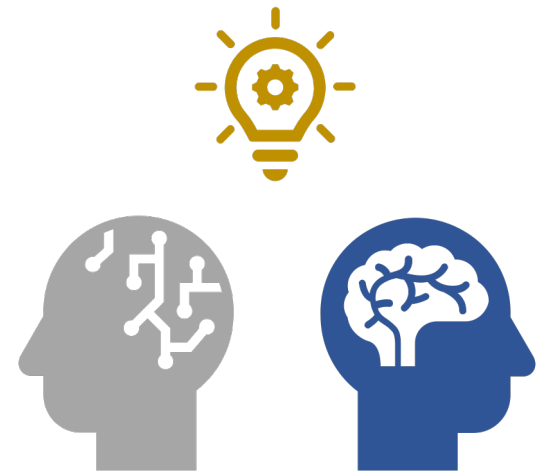


Role of LLMs in Reasoning Research

LLMs have seemingly solved a lot of problems for automated language processing.

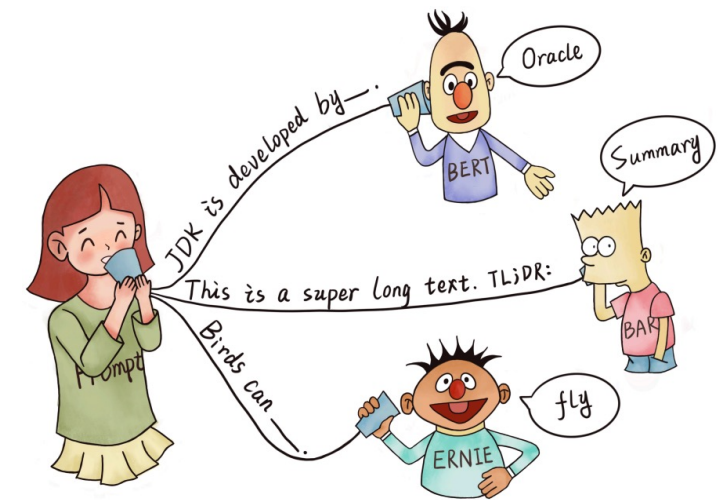
But they enable research on other interesting questions:

- 1. How can LLMs shed light on the nature of human language and reasoning?**
- 2. How can LLMs empower embodied agents with stronger reasoning about how the world works?**



Outline

- Basics of Prompt Engineering
- **Application 1: Analogical Reasoning**
- Application 2: Physical Commonsense Reasoning



In-Context Analogical Reasoning with Pre-Trained Language Models

Xiaoyang Hu^{1,2} *, Shane Storks¹ *, Richard L. Lewis² †, Joyce Chai¹ †

¹ Computer Science & Engineering Division, University of Michigan

² Department of Psychology, University of Michigan

* Equal contribution

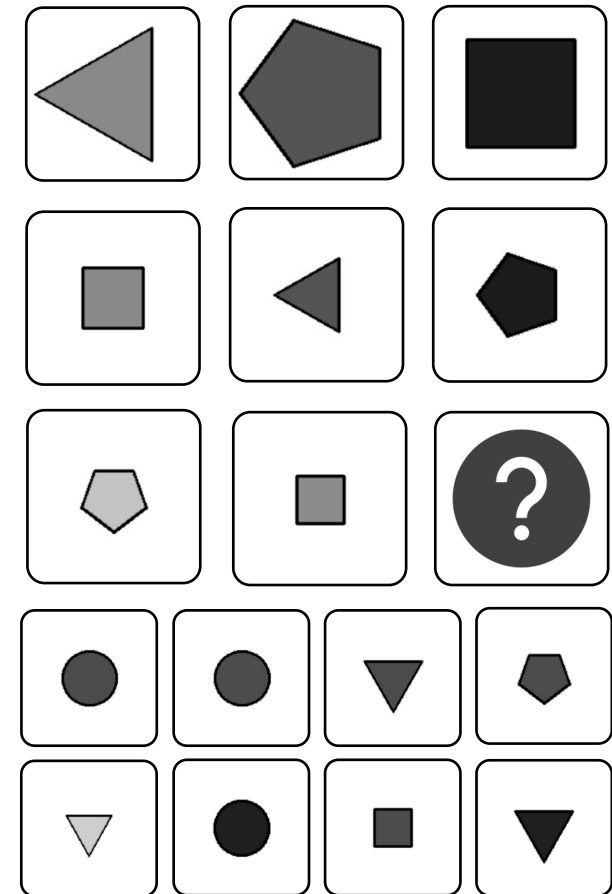
† Equal advising

[ACL 2023 Long Paper](#)

some slides made by Xiaoyang Hu

Introduction

- Analogy-making is fundamental
- Language facilitates analogical reasoning
- Language models are great few-shot reasoners
- Raven's Progressive Matrices



Dedre Gentner. 2010. Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775.

Dedre Gentner, Asli Özyürek, Özge Gürcanli, and Susan Goldin-Meadow. 2013. Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, 127(3):318–330.

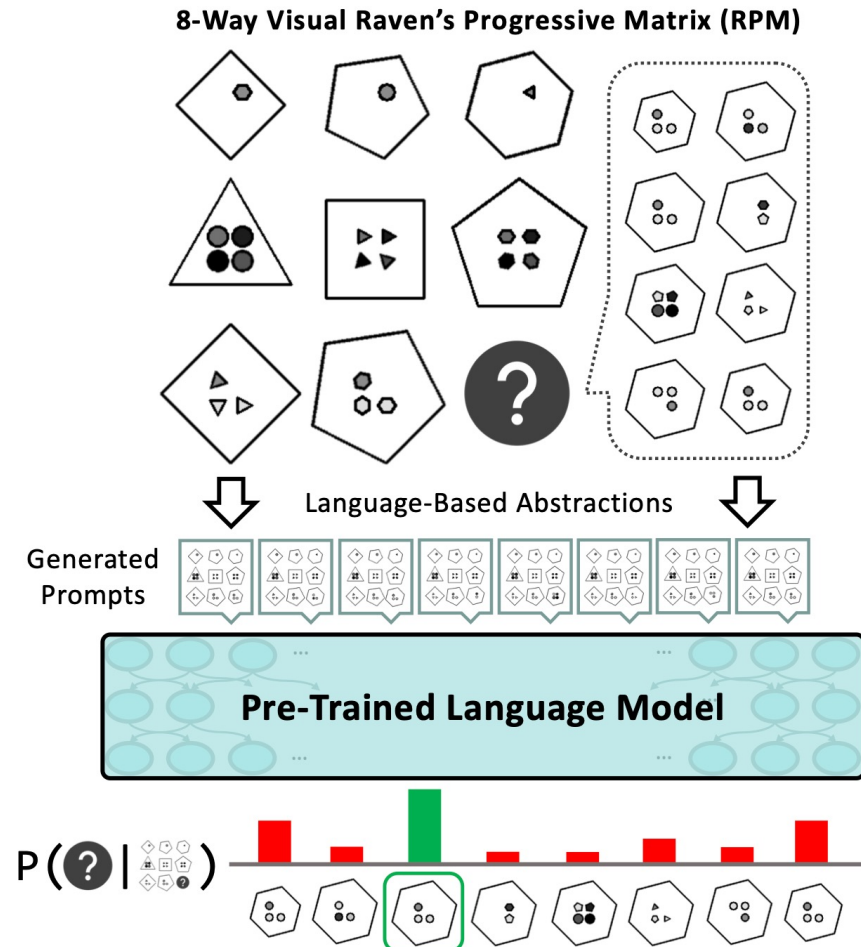
Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Peter Gordon. 2004. Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695):496–499.

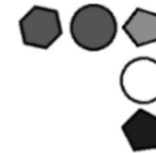
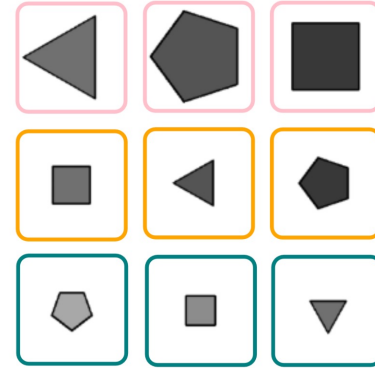
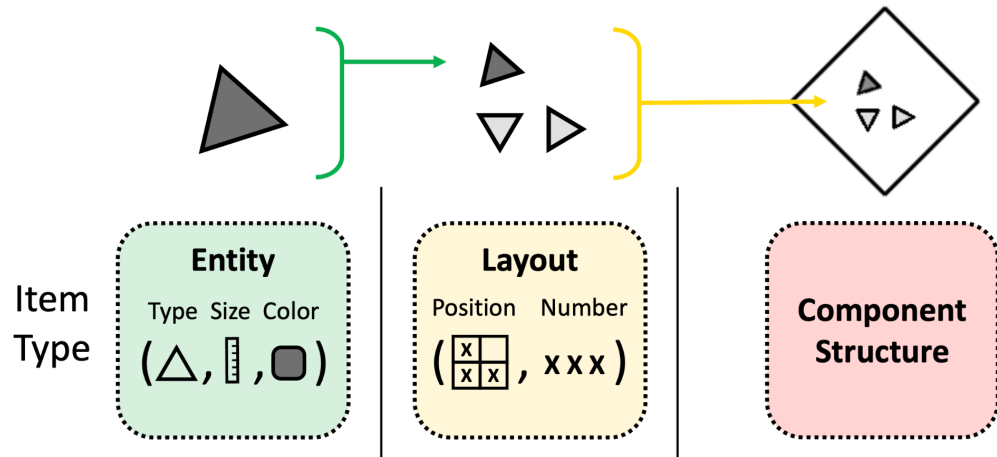
Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019a. RAVEN: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prompting for Analogical Reasoning

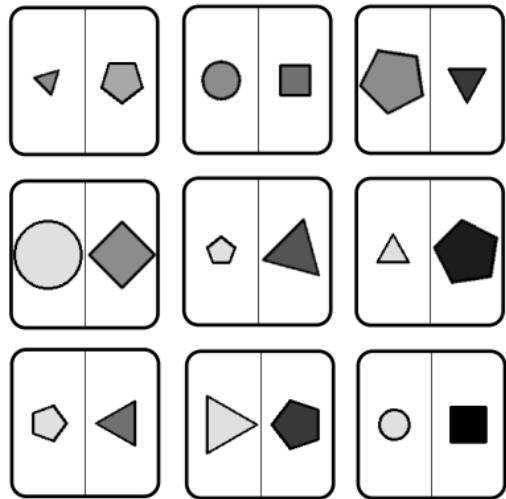
- Created language abstractions for RPMs in RAVEN dataset
- Prompt LLMs to test abstract analogical reasoning capability
 - OPT & InstructGPT



Abstractions & Mappings



Example Prompt



L-R

Complete

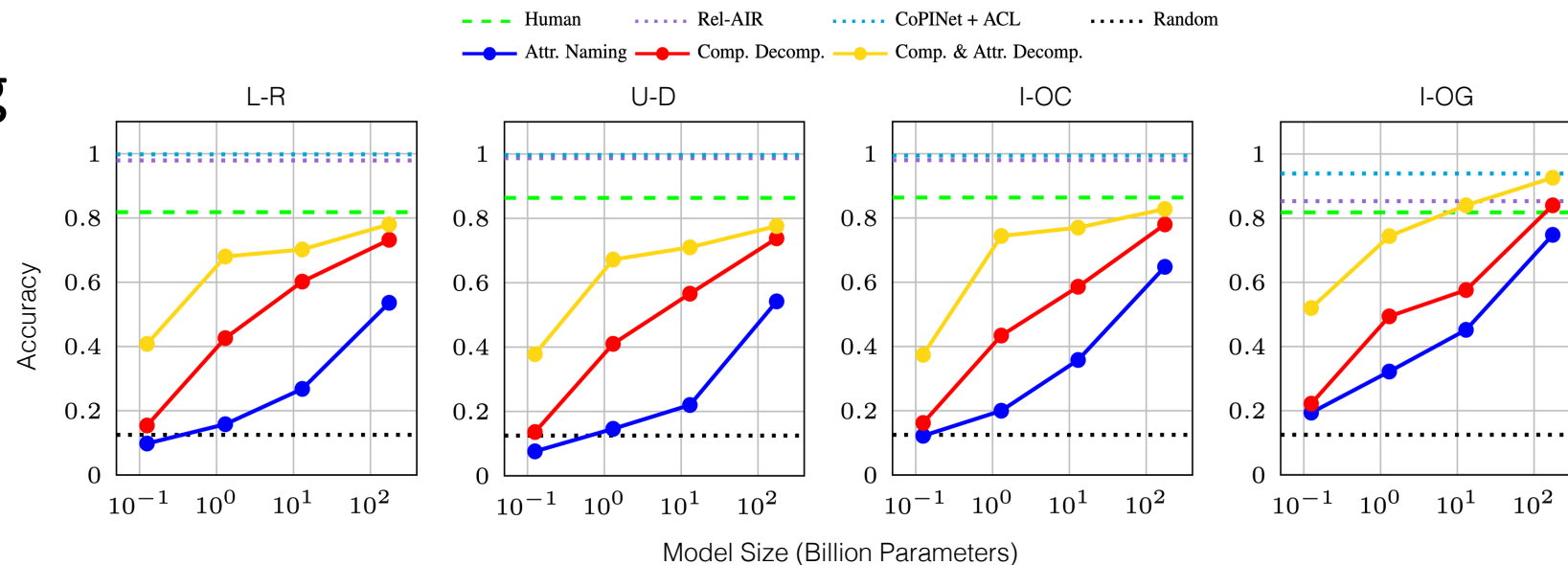
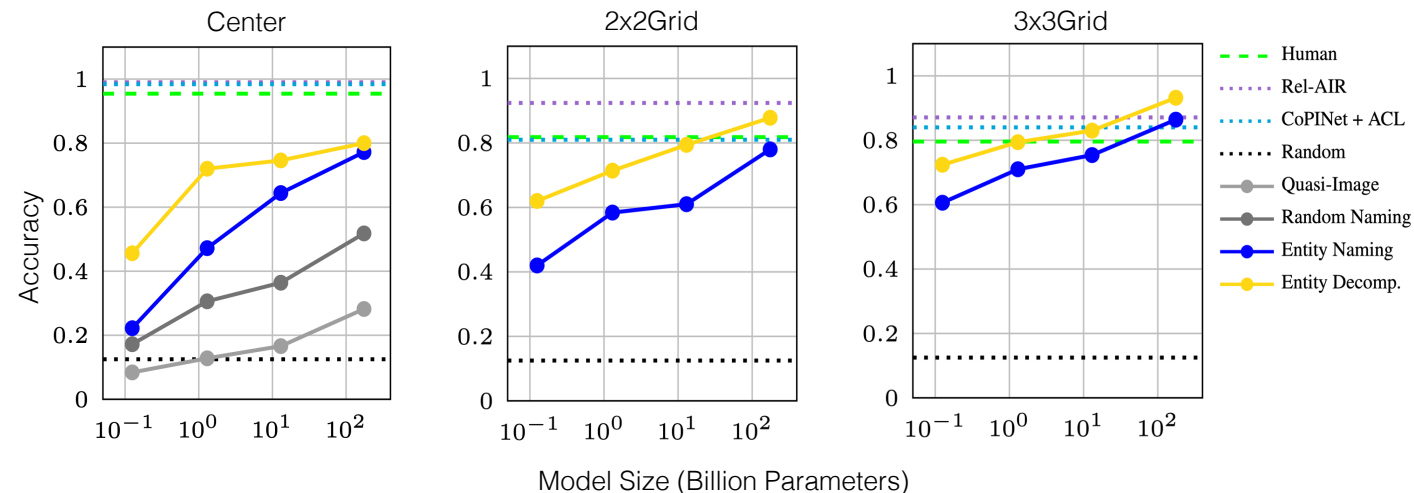
row 1: A (3,0.1,40) / B (5,0.3,30), A (7,0.2,40) / B (4,0.3,50), A (5,0.6,40) / B (3,0.3,70);
 row 2: A (7,0.6,10) / B (4,0.6,40), A (5,0.1,10) / B (3,0.6,60), A (3,0.2,10) / B (5,0.6,80);
 row 3: A (5,0.2,10) / B (3,0.4,50), A (3,0.6,10) / B (5,0.4,70), A (7,0.1,10) / B (4,0.4,90);



$$p(\text{○} \text{■} | \text{○} \text{■} \text{○} \text{■} \text{○} \text{■} \text{○} \text{■}) \propto p(\text{left type}) + p(\text{left size}) + p(\text{left color}) + p(\text{right type}) + p(\text{right size}) + p(\text{right color})$$

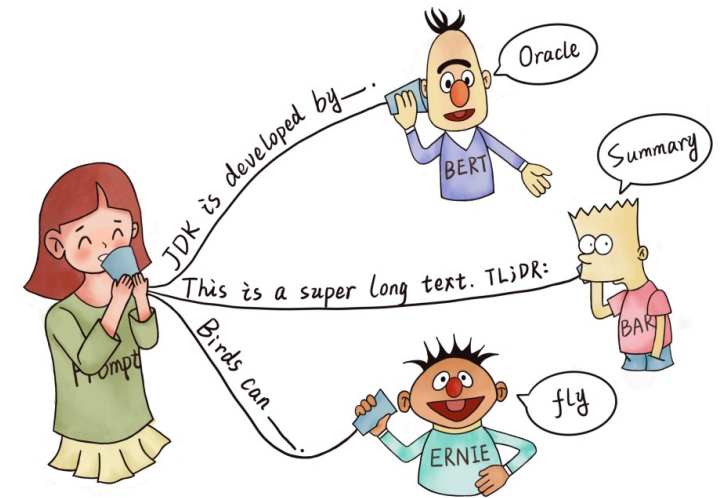
Results & Takeaways

- Abstractions make tasks easier for LLMs
- Larger LLMs better handle task complexity
- LLMs gain abstract analogical reasoning capabilities from training on natural language (!)



Outline

- Basics of Prompt Engineering
- Application 1: Analogical Reasoning
- **Application 2: Physical Commonsense Reasoning**



Work in Progress

Toward Verifiable Physical Commonsense Language Understanding in LLMs

Shane Storks¹

Zheyuan Zhang¹

Fengyuan Hu¹

Sungryull Sohn²

Moontae Lee²

Honglak Lee^{1,2}

Joyce Chai¹

Situated Language and Embodied Dialogue (SLED)

¹University of Michigan, Computer Science and Engineering Division

²LG AI Research



Tiered Reasoning for Intuitive Physics (TRIP)

Story A

1. Ann sat in the chair.
2. Ann unplugged the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann wrote in the book.

Story B

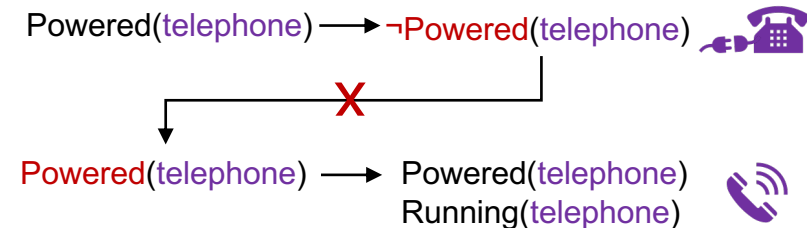
1. Ann sat in the chair.
2. Ann unplugged the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann heard the telephone ring.

Which story is more plausible? **A**

Why not **B**?

Conflicting sentences: 2 → 5

Physical states:



Since this paper, LLMs have evolved quite a bit... We can now prompt InstructGPT and GPT-4 to tackle this problem!

Structured Physical Reasoning

Story A:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on a plate and tossed the donut in the trash.
5. Mary ate the donut.

Story B:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on the counter.
5. Mary ate the donut.

Story B is more plausible.

In Story A, sentences 4 and 5 conflict with each other.



For sentence 4:

After Mary put the cucumber on a plate and tossed the donut in the trash, what is the state of the donut?

The donut is now inedible.

For sentence 5:

Before Mary ate the donut, what was the state of the donut?

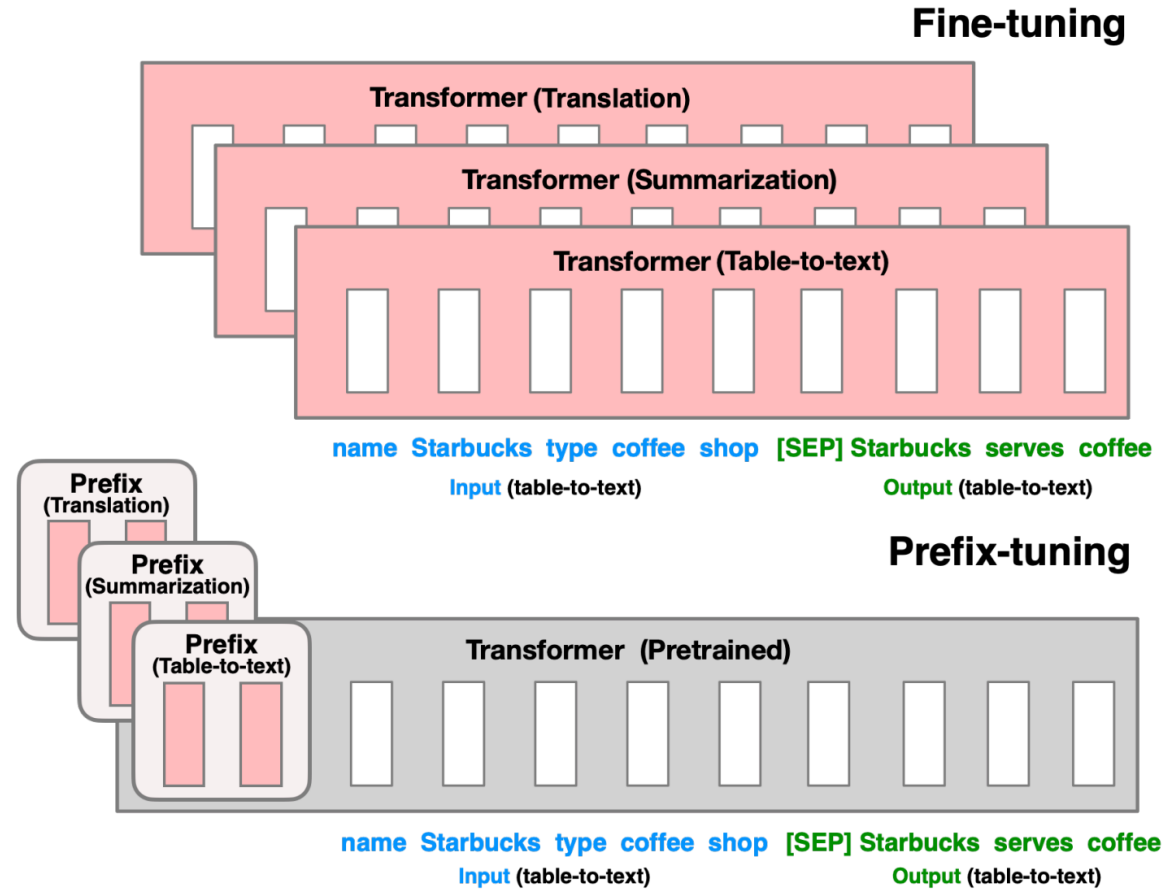
The donut was edible.

Results & Takeaways

- Without task-specific training, LLMs are stronger physical reasoners
 - Just need a few strongly annotated demonstrations of the task!

Approach	Accuracy	Consistency	Verifiability
<i>RoBERTa fine-tuned</i>	72.9	19.1	9.1
<i>InstructGPT, structured prompt</i>	68.1	43.4	17.1
<i>GPT-4 structured prompt</i>	95.4	85.2	41.8

Prompt Tuning



Thank you!



@shanestorks



www.shanestorks.com