



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen

# CSC6052/5051/4100/DDA6307/MDS5110

# Natural Language Processing

## Lecture 6-2: Prompt Engineering

Spring 2025  
Benyou Wang  
School of Data Science

# Knowledge vs. Reasoning

Knowledge is **storage**. Reasoning is **processing**.

# Knowledge vs. Reasoning

## Knowledge

- **Definition:** Knowledge refers to the information, facts, concepts, and skills that a person has acquired through experience, education, or observation.
- **Nature:** It is often considered a static collection of data and truths about the world. Knowledge can be explicit (easily articulated and shared) or tacit (personal, context-specific, and harder to communicate).
- **Examples:** Knowing that Paris is the capital of France, understanding the laws of physics, or having expertise in a particular field.

## Reasoning

- **Definition:** Reasoning is the cognitive process of drawing conclusions, making inferences, or forming judgments based on available information and knowledge.
- **Nature:** It is dynamic and involves the application of logic and critical thinking to analyze situations, solve problems, and make decisions.
- **Examples:** Deductive reasoning (drawing specific conclusions from general principles), inductive reasoning (generalizing from specific instances), and abductive reasoning (inferring the best explanation from incomplete information).

# Difference and Connections

## Key Differences

- **Function:**
  - Knowledge provides the content or **building blocks for reasoning**.
  - Reasoning is the **process that utilizes knowledge** to arrive at conclusions.
- **Static vs. Dynamic:**
  - Knowledge tends to be more static (what you know).
  - Reasoning is dynamic (how you use what you know).
- **Outcome:**
  - Knowledge can exist without reasoning (e.g., memorized facts).
  - Reasoning often requires knowledge to be effective.

## Interrelationship

Knowledge and reasoning are interdependent; effective reasoning relies on a solid foundation of knowledge, while acquiring new knowledge often requires reasoning skills to evaluate, integrate, and understand information.

# Knowledge vs. Reasoning in Humans

- **Knowledge** is stored in **languages**
- Human **reasons** via **languages**.

Noam Chomsky's theory of language focuses on **language as a type of knowledge**, rather than just a means of communication. He believes that language is a biological capacity that helps humans create social interactions.



**Noam  
Chomsky**

“The systems of thought ... use linguistic expressions for reasoning, interpretation, organizing action, and other mental acts.”

“The limits of my language mean the limits of my world.”



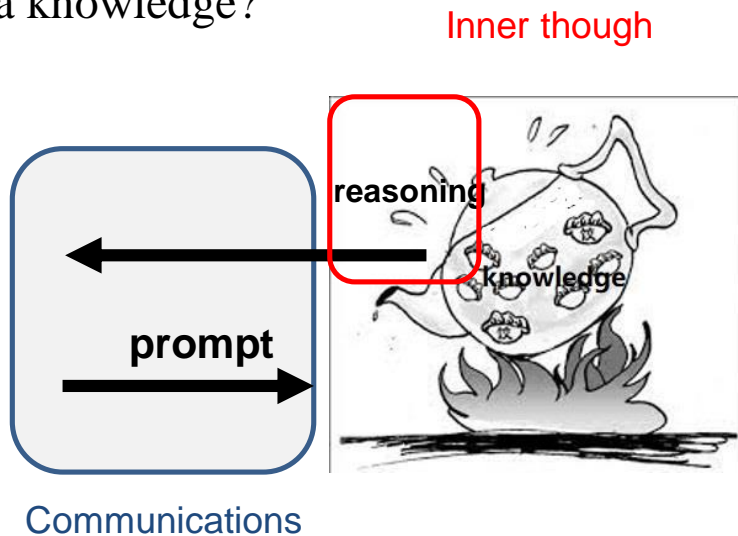
**Ludwig  
Wittgenstein**

[Noam Chomsky - The Function of Language https://youtu.be/TzzuPMA8s7k?t=324](https://youtu.be/TzzuPMA8s7k?t=324)

Evelina Fedorenko, Steven T. Piantadosi, Edward A. F. Gibson. Language is primarily a tool for communication rather than thought. **Nature**. <https://www.nature.com/articles/s41586-024-07522-w>

# Knowledge vs. Reasoning in LLMs

- How much does LLM memory a knowledge?
- How does LLM think?



[Noam Chomsky - The Function of Language https://youtu.be/TzzuPMA8s7k?t=324](https://youtu.be/TzzuPMA8s7k?t=324)

Evelina Fedorenko, Steven T. Piantadosi, Edward A. F. Gibson. Language is primarily a tool for communication rather than thought. **Nature**. <https://www.nature.com/articles/s41586-024-07522-w>

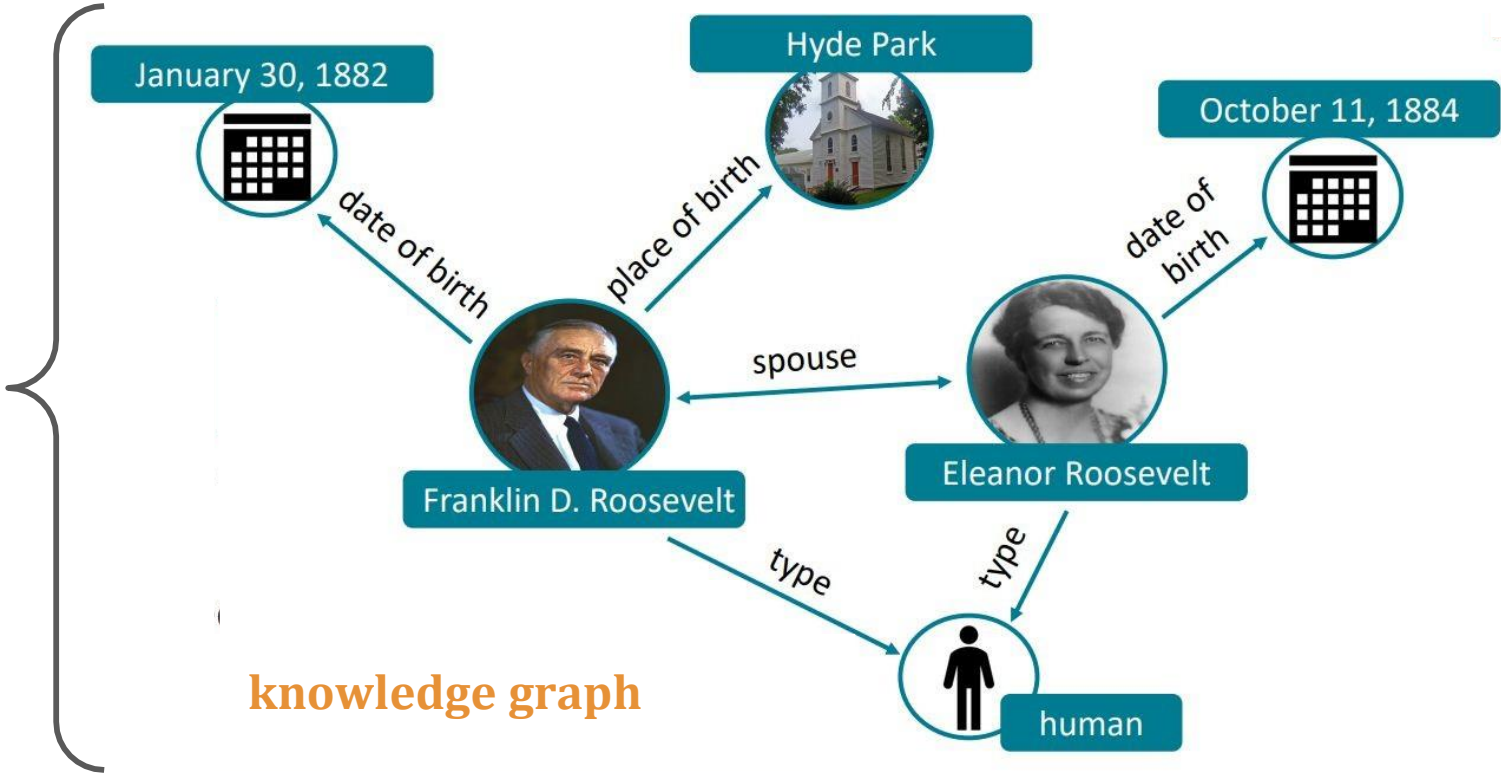
What is a knowledge base?

# What is a knowledge base?

knowledge base



Example



knowledge graph

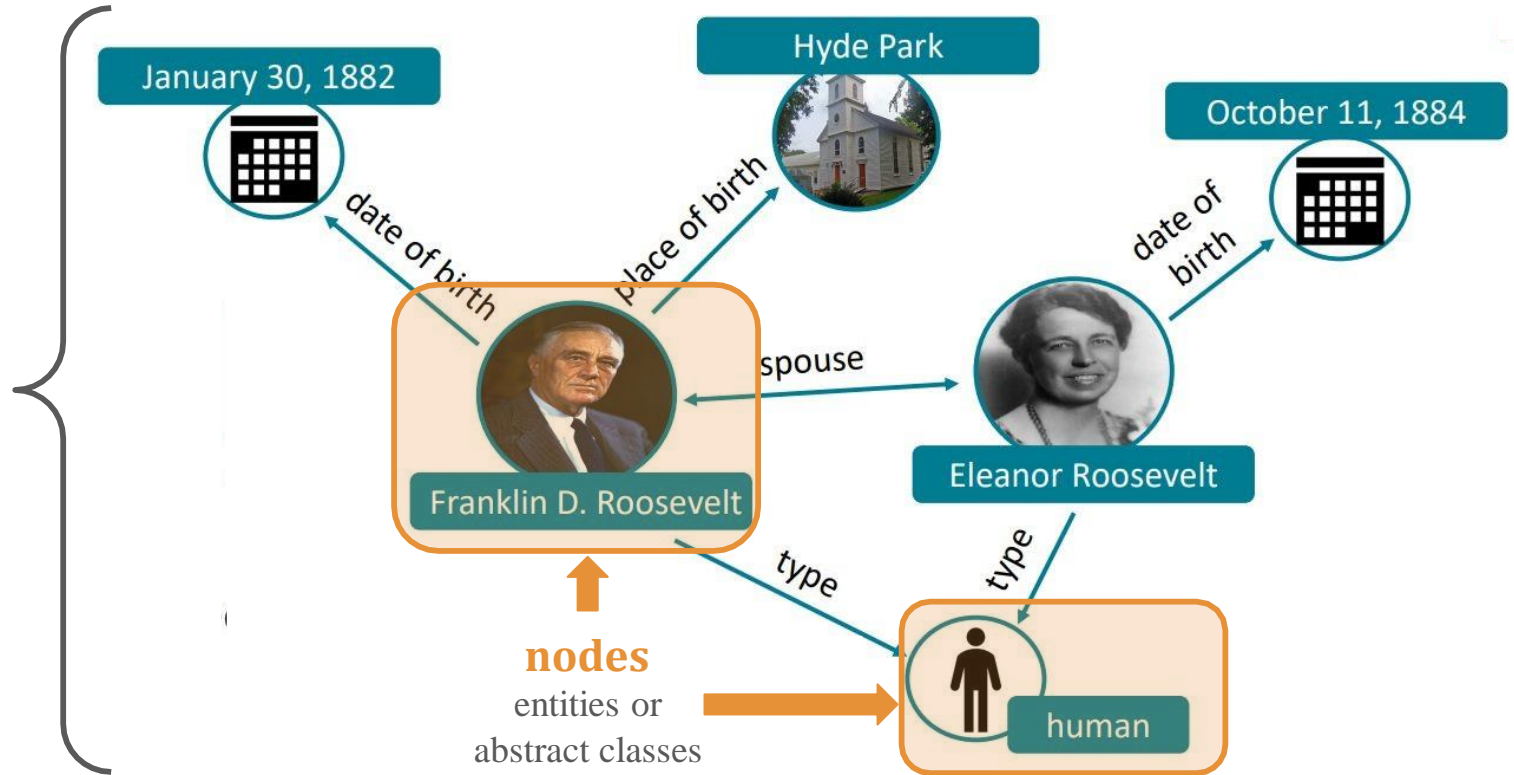


# What is a knowledge base?

knowledge base



Example

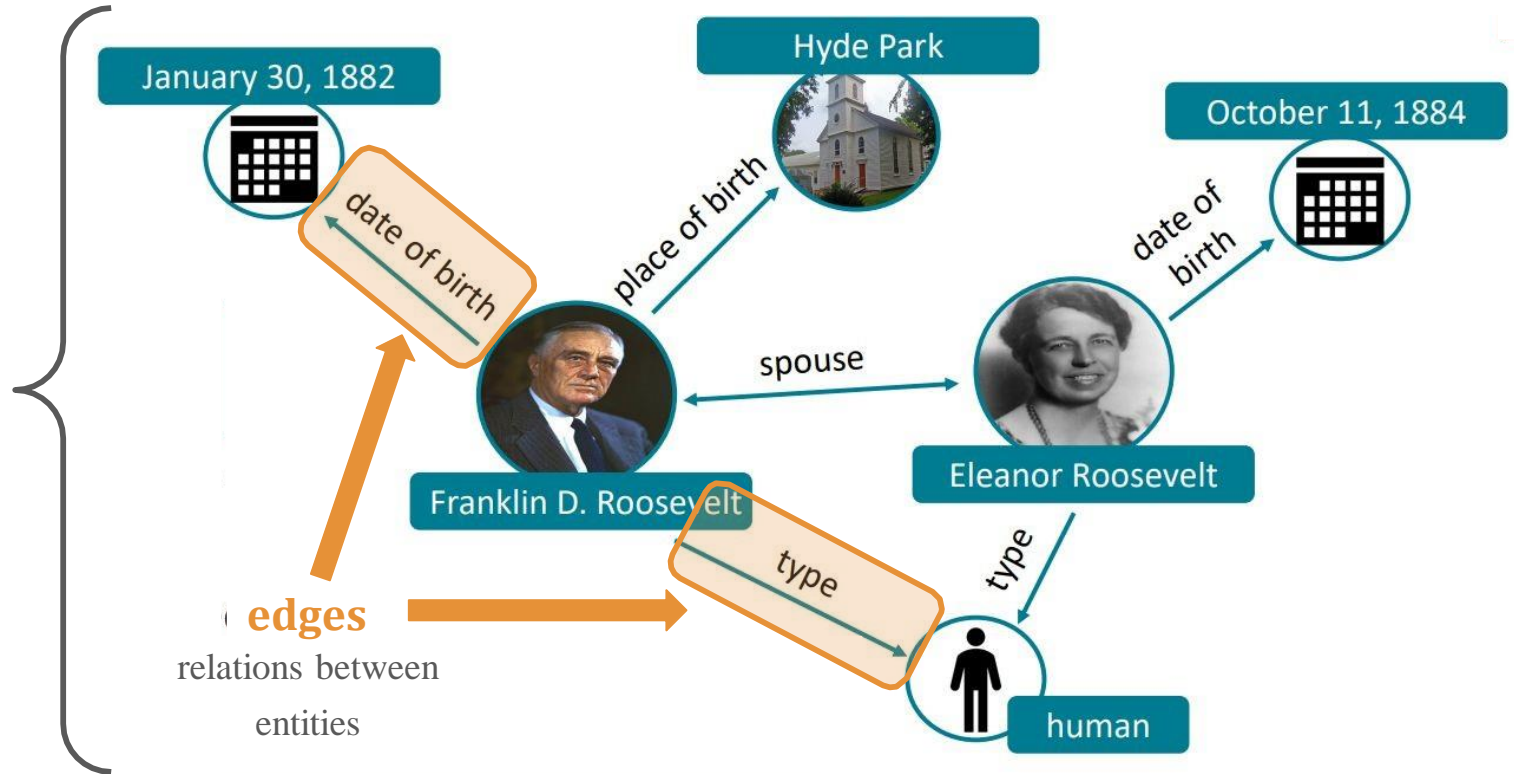


# What is a knowledge base?

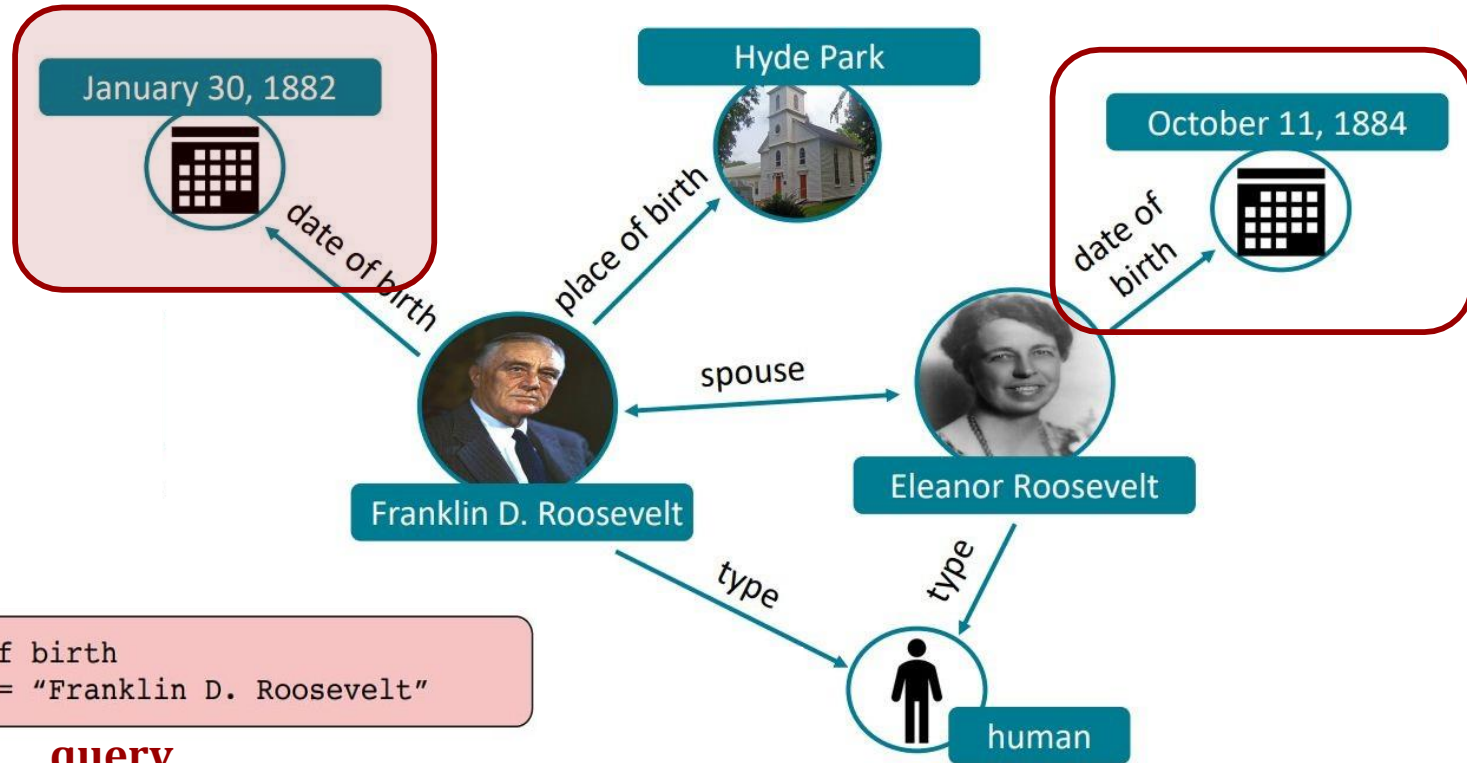
knowledge base



Example



# How to query?



**query**

# How were knowledge bases formed?

The image shows three overlapping screenshots of Wikipedia pages, illustrating unstructured text. The top-most page is for 'Valorant', the middle one for 'Brie', and the bottom one for 'T. S. Eliot'. Each page contains a mix of text, lists, and structured data, but the overall layout is not uniform or machine-readable.

**Valorant**  
From Wikipedia, the free encyclopedia

**Valorant** (shortened as **VALORANT**) is a free-to-play first-person shooter video game developed by Riot Games under the codename Project A. The game began on April 7, 2020, and was released on June 2, 2020. It is the first game in the Valorant Strike series. The game's mechanics and map design are inspired by Counter-Strike and Overwatch.

**Contents**

- Gameplay
- Maps
- Characters
- Development
- Reception
- References
- External links
- Notes

**Brie**  
From Wikipedia, the free encyclopedia  
(Redirected from Brie cheese)

**Brie** (/briː/) is a soft cheese named after the town of Brie in France. It is made from cow's milk and is protected by a European Union Protected Designation of Origin (PDO) label.

**T. S. Eliot**  
From Wikipedia, the free encyclopedia  
(Redirected from T. S. Eliot)

For other people named *Thomas Eliot*, see *Thomas Eliot (disambiguation)*.

**Thomas Stearns Eliot** OM (26 September 1898 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.<sup>[R]</sup> Considered one of the 20th century's major poets, he is a central figure in English-language Modernist poetry. Born in St. Louis, Missouri, to a prominent Boston Brahmin family, he moved to England in 1914 at the age of 25 and went on to settle, work, and marry there.<sup>[R]</sup> He became a British citizen in 1927 at the age of 30, subsequently renouncing his American citizenship.<sup>[R]</sup>

Eliot first attracted widespread attention for his poem "The Love Song of J. Alfred Prufrock" in 1915, which, at the time of its publication, was considered outlandish.<sup>[R]</sup> It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).<sup>[R]</sup> He was also known for seven plays, particularly *Murder in the Cathedral* (1935) and *The Cocktail Party* (1949). He was awarded the 1948 Nobel Prize in Literature, "for his outstanding, pioneer contribution to present-day poetry"<sup>[R]</sup>

**T. S. Eliot**  
OM

Eliot in 1934 by Lady Ottoline Morrell

<b>Born</b>	Thomas Stearns Eliot 26 September 1898 St. Louis, Missouri, US
<b>Died</b>	4 January 1965 (aged 76) London, England
<b>Occupation</b>	Poet · essayist · playwright · publisher · critic
<b>Citizenship</b>	American (1898–1927) British (1927–1965)

unstructured text

Structured knowledge base

# How were knowledge bases formed?

**Valorant**  
From Wikipedia, the free encyclopedia

**Valorant** (shortened as **VALORANT**) is a battle royale first-person shooter video game developed by Riot Games under the codename Project A. The game began on April 7, 2020, and was released on June 2, 2020. It is the first game in the Valorant series. The game is a tactical 5v5 shooter. The game mechanics are similar to Counter-Strike and Overwatch.

**Brie**  
From Wikipedia, the free encyclopedia  
(Redirected from Brie cheese)

**Brie** (/briː/) is a soft cheese named after the town of Brie in France. It is made from cow's milk and is a member of the *Cheese* family. It is a semi-soft cheese with a slight tangy taste. It is often served with bread and butter.


**Content**

- 1 Games
- 1.1 Unlabeled
- 1.2 Spelling
- 1.3 Categories
- 1.4 Disambiguation
- 1.5 External links
- 1.6 Redirects
- 1.7 Spelling
- 2 Agents
- 2.1 Default
- 2.2 Settings
- 2.3 In-game
- 2.4 Customization
- 3 Store
- 4 Developer

**Thomas Stearns Eliot OM** (26 September 1898 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.<sup>[1]</sup> Considered one of the 20th century's major poets, he is a central figure in English-language Modernist poetry. Born in St. Louis, Missouri, to a prominent Boston Brahmin family, he moved to England in 1914 at the age of 25 and went on to settle, work, and marry there.<sup>[1]</sup> He became a British citizen in 1927 at the age of 30, subsequently renouncing his American citizenship.<sup>[4]</sup>

Eliot first attracted widespread attention for his poem "The Love Song of J. Alfred Prufrock" in 1915, which, at the time of its publication, was considered outlandish.<sup>[5]</sup> It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).<sup>[6]</sup> He was also known for seven plays, particularly *Murder in the Cathedral* (1935) and *The Cocktail Party* (1949). He was awarded the 1948 Nobel Prize in Literature, "for his outstanding, pioneer contribution to present-day poetry"<sup>[7][8]</sup>

**T. S. Eliot**  
OM



Eliot in 1934 by Lady Ottoline Morrell

**Born** Thomas Stearns Eliot  
26 September 1898  
St. Louis, Missouri, US

**Died** 4 January 1965 (aged 76)  
London, England

**Occupation** Poet · essayist · playwright · publisher · critic

**Citizenship** American (1898–1927)  
British (1927–1965)



**Knowledge Extraction  
Pipeline**

**unstructured text**

**Structured knowledge base**

# How were knowledge bases formed?

**Valorant**  
From Wikipedia, the free encyclopedia

**Valorant** (stylized as **VALORANT**) is a battle royale first-person shooter by Riot Games under the company name Valorant. The game began on April 7, 2020, and was released in June 2, 2020. It is a free-to-play, tactical, 5v5, Strike series game. It features mechanics similar to Call of Duty: Warzone and Overwatch.

**Cont**

- 1 Games
- 1.1 Un
- 1.2 Sp
- 1.3 Co
- 1.4 Da
- 1.5 Es
- 1.6 Re
- 1.7 Sin

**2 Agents**

- 2.1 Du
- 2.2 Se
- 2.3 In
- 2.4 Co

**3 Store**

- 3.1 Bi
- 3.2 Bi
- 3.3 Es

**4 Developer**

**Brie**  
From Wikipedia, the free encyclopedia  
(Redirected from T. S. Eliot)

This article is about the poet. For the film, see *Brie*.  
From Wikipedia, the free encyclopedia  
(Redirected from T. S. Eliot)

**T. S. Eliot**  
From Wikipedia, the free encyclopedia  
(Redirected from T. S. Eliot)

**Brie** (/briː/) is a name that originated in the 19th century and is a variant of the name Brien. It is a name with a slight meaning, depending on the context. The name Brie is a French name, and Brien is a name of Irish origin. Brie is a name that is often used as a nickname for the name Brian. Brie is a name that is often used as a nickname for the name Brian.

**Thomas Stearns Eliot OM** (26 September 1898 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.<sup>[R]</sup> Considered one of the 20th century's major poets, he is a central figure in English-language Modernist poetry. Born in St. Louis, Missouri, to a prominent Boston Brahmin family, he moved to England in 1914 at the age of 25 and went on to settle, work, and marry there.<sup>[1]</sup> He became a British citizen in 1927 at the age of 30, subsequently renouncing his American citizenship.<sup>[4]</sup>

Eliot first attracted widespread attention for his poem "The Love Song of J. Alfred Prufrock" in 1915, which, at the time of its publication, was considered outlandish.<sup>[5]</sup> It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).<sup>[6]</sup> He was also known for seven plays, particularly *Murder in the Cathedral* (1935) and *The Cocktail Party* (1949). He was awarded the 1948 Nobel Prize in Literature, "for his outstanding, pioneer contribution to present-day poetry."<sup>[7][R]</sup>

**T. S. Eliot**  
OM

Eliot in 1934 by Lady Ottoline Morrell

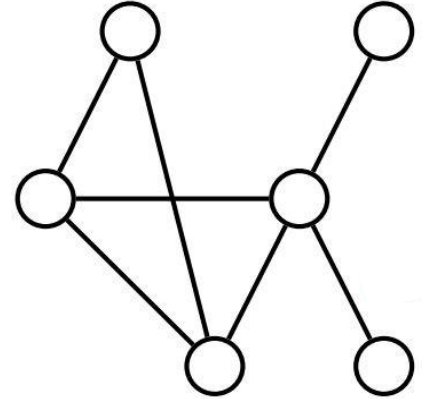
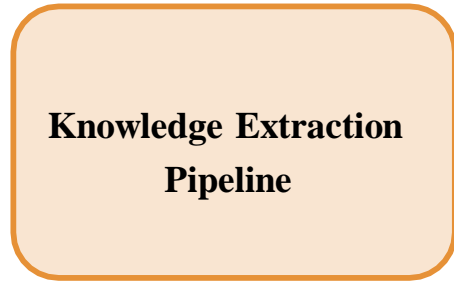
**Born** Thomas Stearns Eliot  
26 September 1898  
St. Louis, Missouri, US

**Died** 4 January 1965 (aged 76)  
London, England

**Occupation** Poet · essayist · playwright · publisher · critic

**Citizenship** American (1898–1927)  
British (1927–1965)

unstructured text

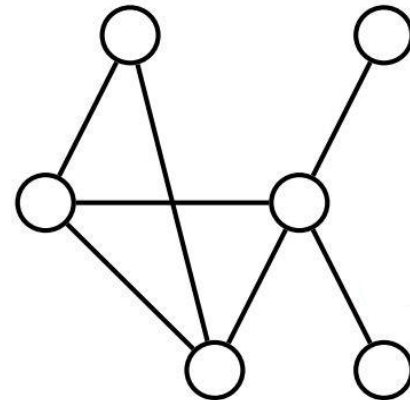
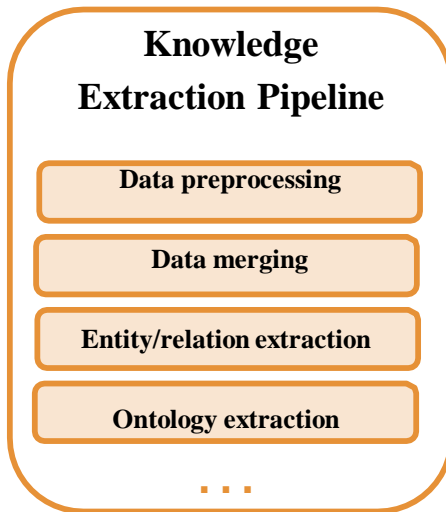


Structured knowledge base

# Downsides of using knowledge bases

# Downsides of using knowledge bases

The image shows three overlapping screenshots of Wikipedia pages. The top page is for 'Valorant', the middle for 'Brie', and the bottom for 'T. S. Eliot'. The 'T. S. Eliot' page includes a portrait and biographical information.



Populating the knowledge base often involves **complicated, multi-step NLP pipelines**



# Downsides of using knowledge bases

From Wikipedia, the free encyclopedia

**T. S. Eliot**

From Wikipedia, the free encyclopedia  
(Redirected from T s eliot)

*For other people named Thomas Eliot, see Thomas Eliot (disambiguation).*

**Thomas Stearns Eliot** OM (26 September 1898 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.<sup>[1]</sup> Considered one of the 20th century's major poets, he is a central figure in English-language Modernist poetry.

Born in St. Louis, Missouri, to a prominent Boston Brahmin family, he moved to England in 1914 at the age of 25 and went on to settle, work, and marry there.<sup>[1]</sup> He became a British citizen in 1927 at the age of 30, subsequently renouncing his American citizenship.<sup>[4]</sup>

Eliot first attracted widespread attention for his poem "The Love Song of J. Alfred Prufrock" in 1915, which, at the time of its publication, was considered outlandish.<sup>[5]</sup> It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).<sup>[6]</sup> He was also known for seven plays, particularly *Murder in the Cathedral* (1935) and *The Cocktail Party* (1949). He was awarded the 1948 Nobel Prize in Literature, "for his outstanding, pioneer contribution to present-day poetry"<sup>[7][8]</sup>

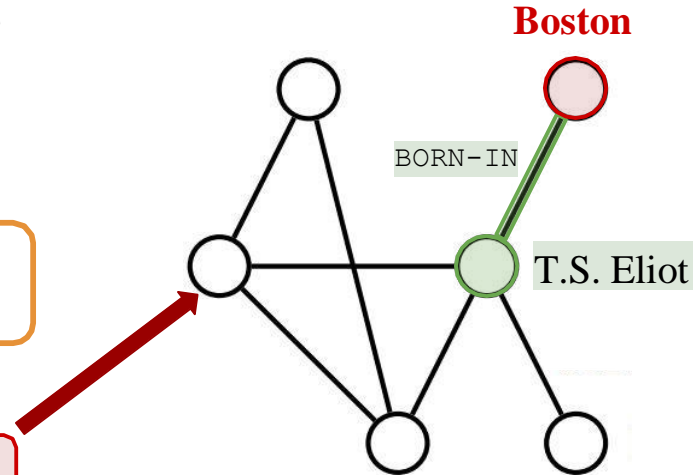
unstructured text

"Born in St. Louis, Missouri, to a prominent Boston Brahmin family..."

Knowledge Extraction Pipeline

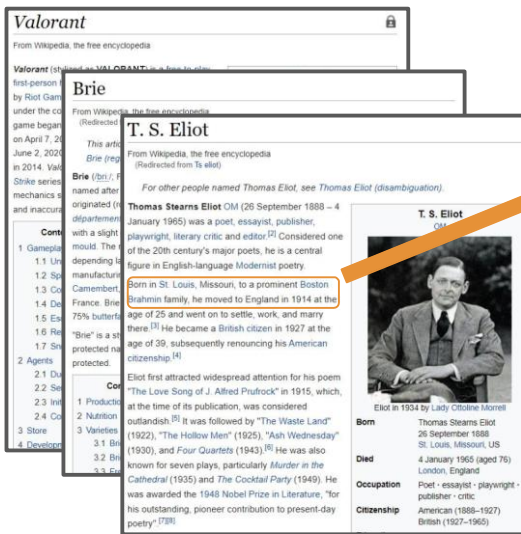
(T.S. Eliot, BORN-IN, Boston)

incorrect extraction



Prone to **error propagation** (from human annotations or knowledge extraction)

# Downsides of using knowledge bases



unstructured text

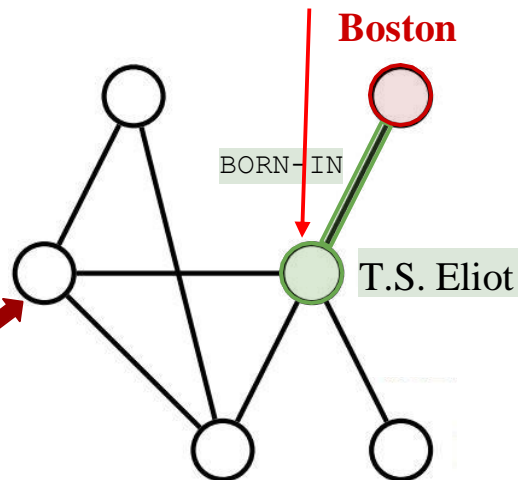
“Born in St. Louis, Missouri, to a prominent Boston Brahmin family...”

Knowledge Extraction Pipeline

(T.S. Eliot, BORN-IN, Boston)

incorrect extraction

Q: Describe Eliot’s family’s related information.



Triples lead to **information loss**: hard to include all possible information we may be interested in.

# Are there better alternatives?

Traditional knowledge bases are **inflexible**  
and require **significant manual effort**.

# Language Models as Knowledge Bases? (Petroni et al., 2019)

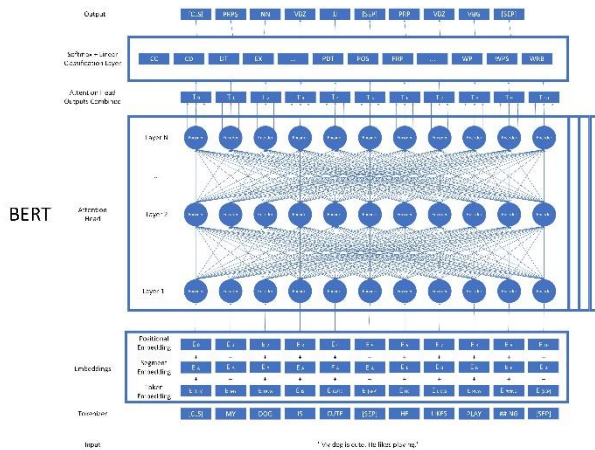
# Language models as knowledge bases?

## Why language models?

- Scalability: pre-trained on a huge corpus of data
- Time/Labor efficiency: does not require annotations/supervision
- Flexibility: more flexible with natural language queries
- Accessibility: can be used off-the-shelf

Do language models really store knowledge?

# LAMA probe



probe(探针, 探测)



# LAMA Probe



- Goal: evaluate **factual + commonsense knowledge** in language models





## LAMA Probe



- **Goal: evaluate **factual + commonsense knowledge** in language models**
- Collect set of **knowledge sources** (i.e. set of facts) and test to see how well the model's knowledge captures these facts



## LAMA Probe



- **Goal: evaluate factual + commonsense knowledge in language models**
- Collect set of **knowledge sources** (i.e. set of facts) and test to see how well the model's knowledge captures these facts
- *How do we know how “knowledgeable” a LM is about a particular fact?*



## LAMA Probe



- **Goal: evaluate factual + commonsense knowledge in language models**
- Collect set of **knowledge sources** (i.e. set of facts) and test to see how well the model's knowledge captures these facts
- *How do we know how “knowledgeable” a LM is about a particular fact?*

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

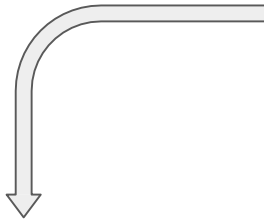
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token,  
**knowledgeable LMs rank ground truth tokens high** and other tokens lower

# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

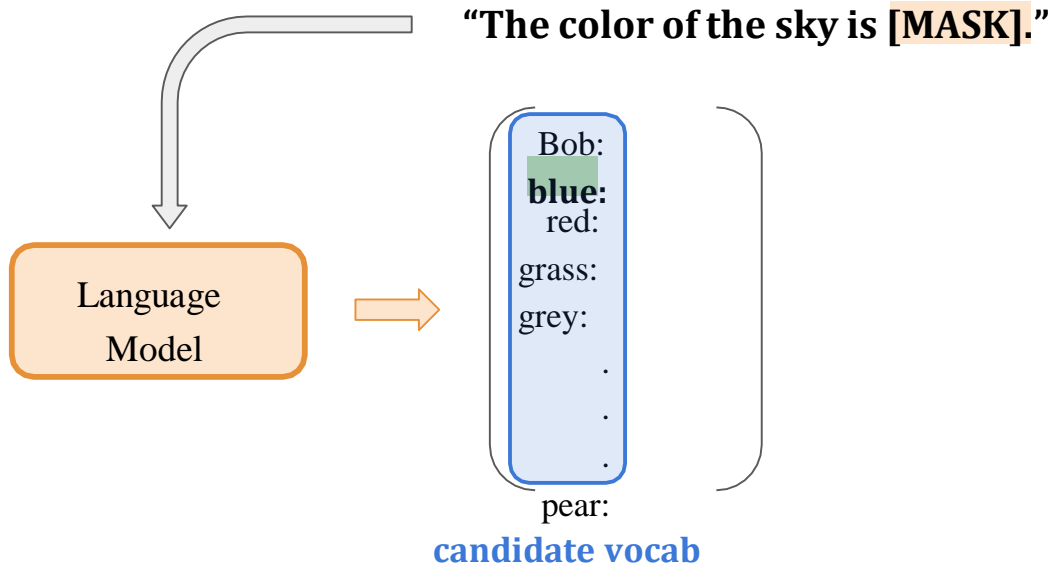
“The color of the sky is [MASK].”



Language  
Model

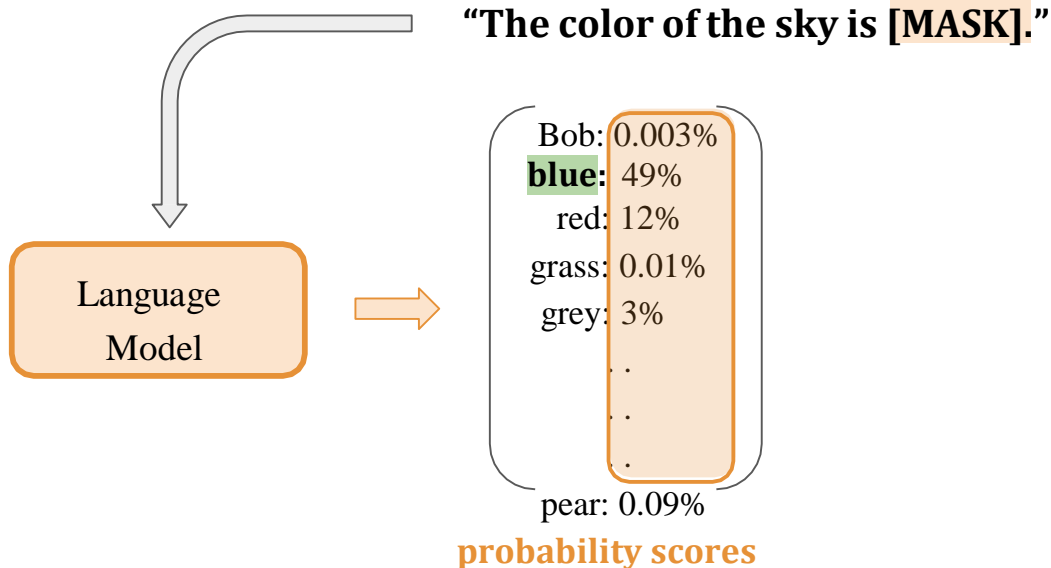
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



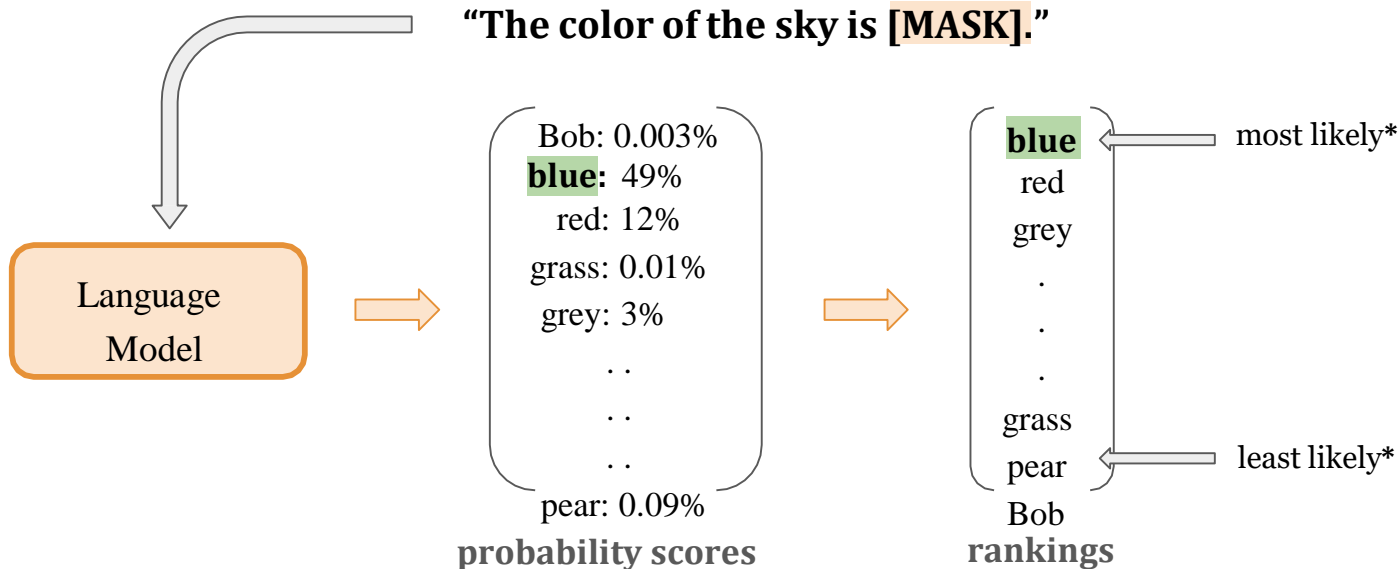
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

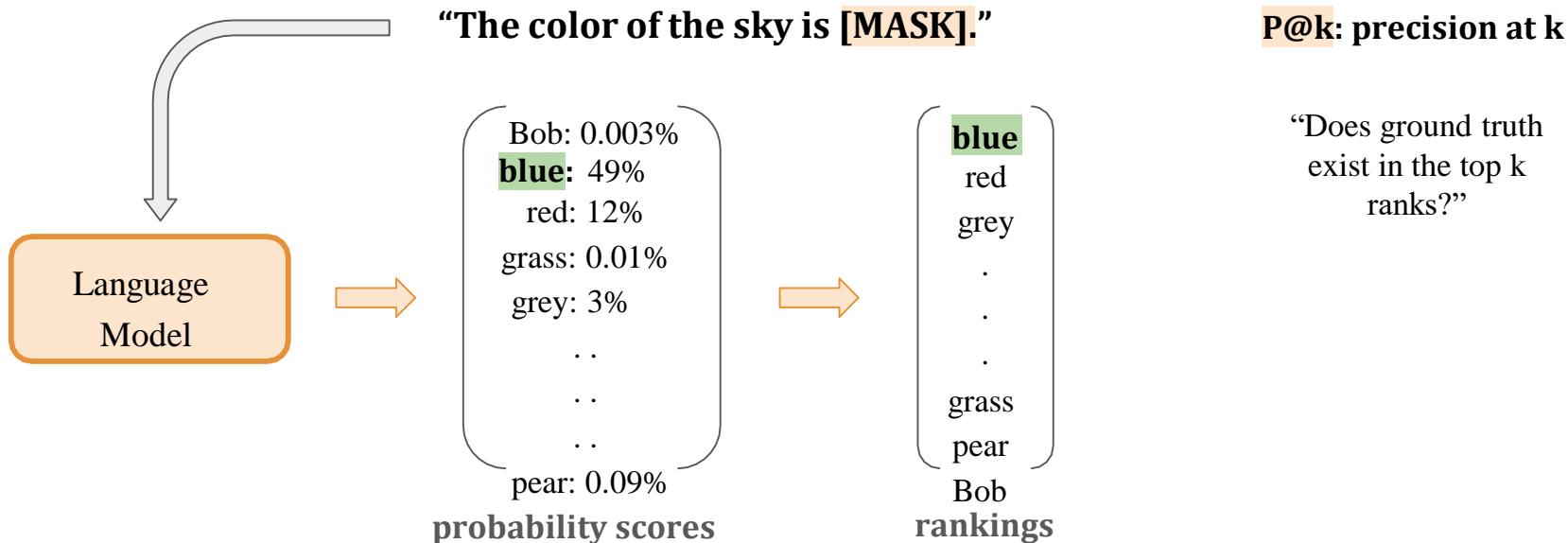


\*according to the LM



# Evaluation of LM via LAMA

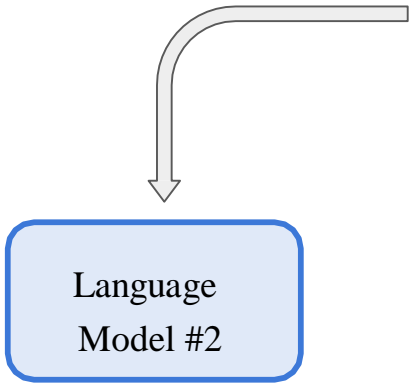
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

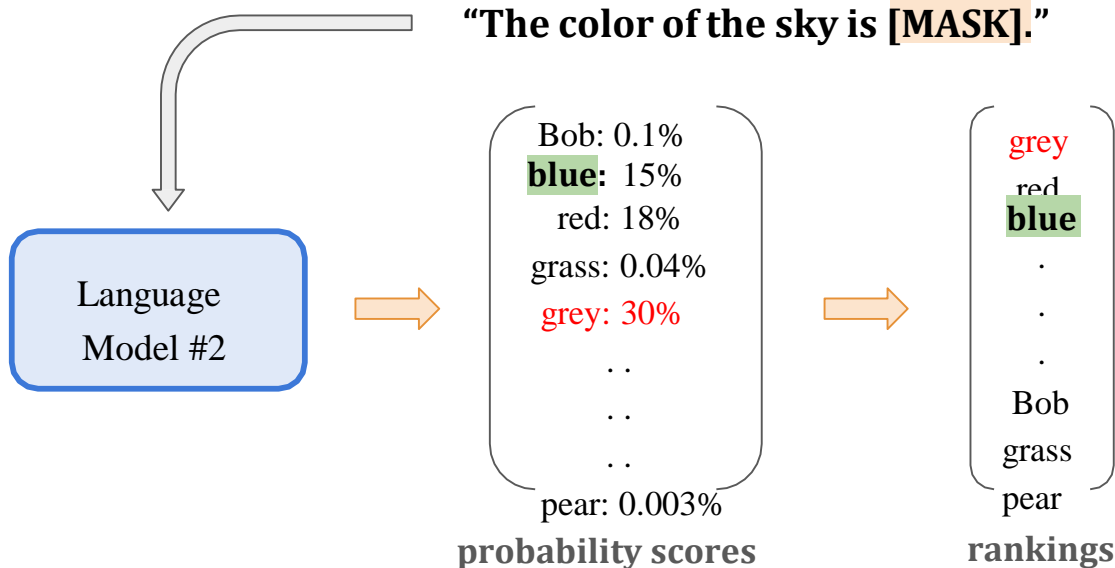
“The color of the sky is [MASK].”



Language  
Model #2

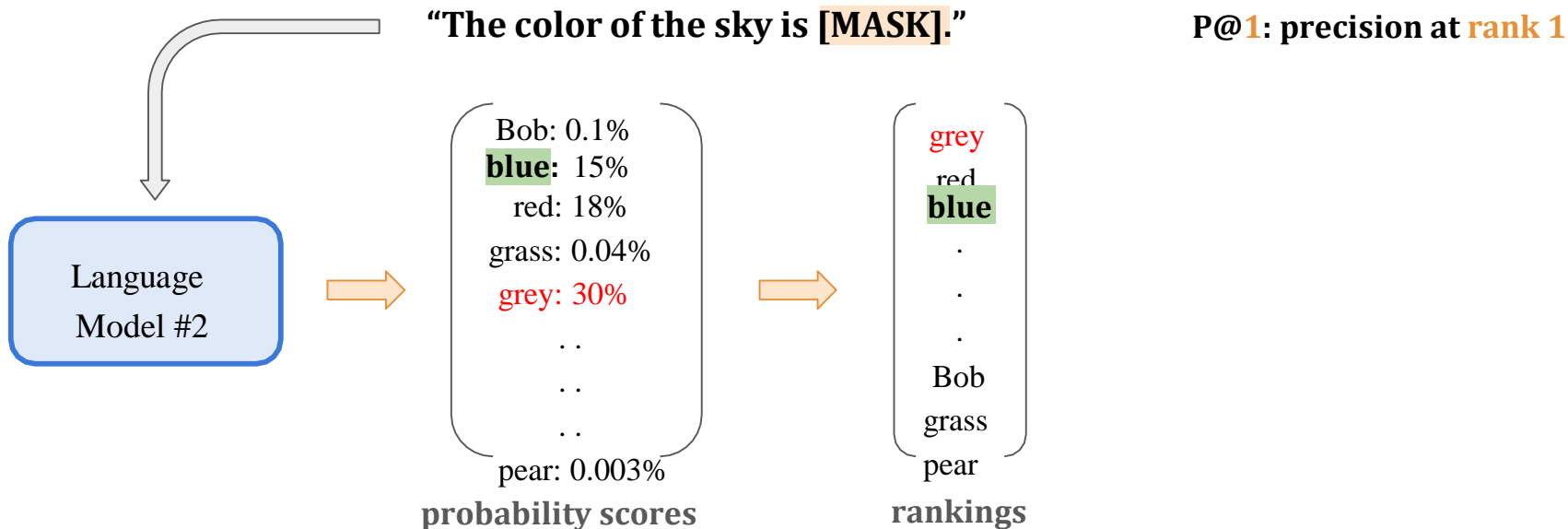
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



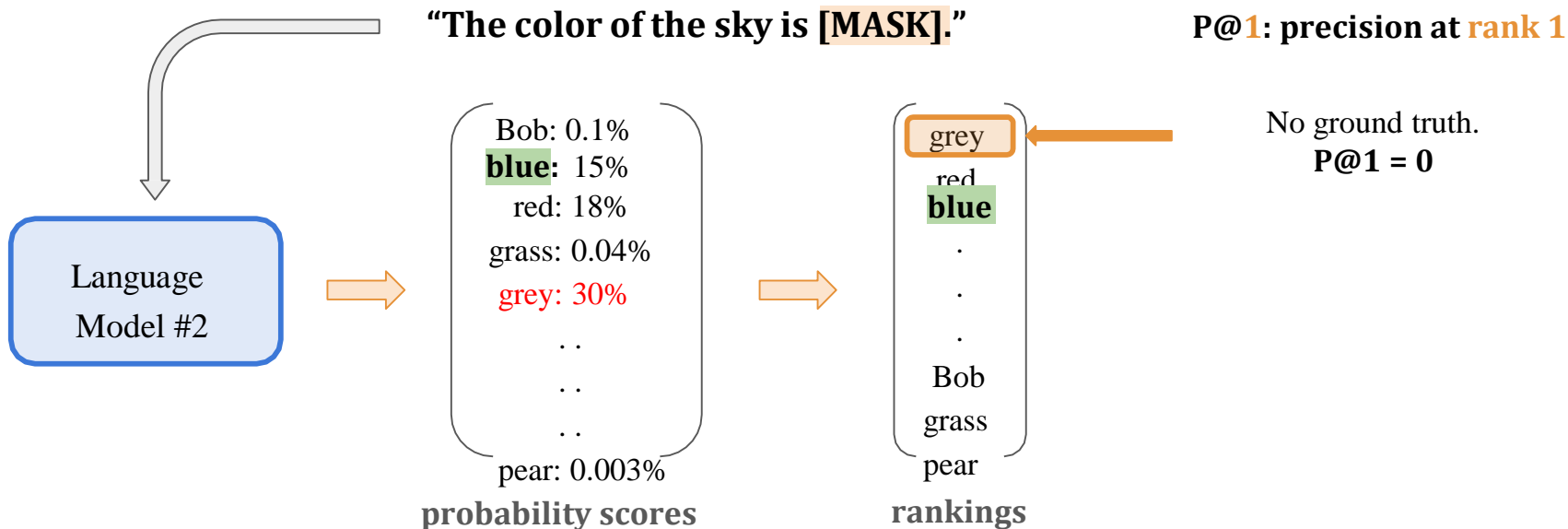
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



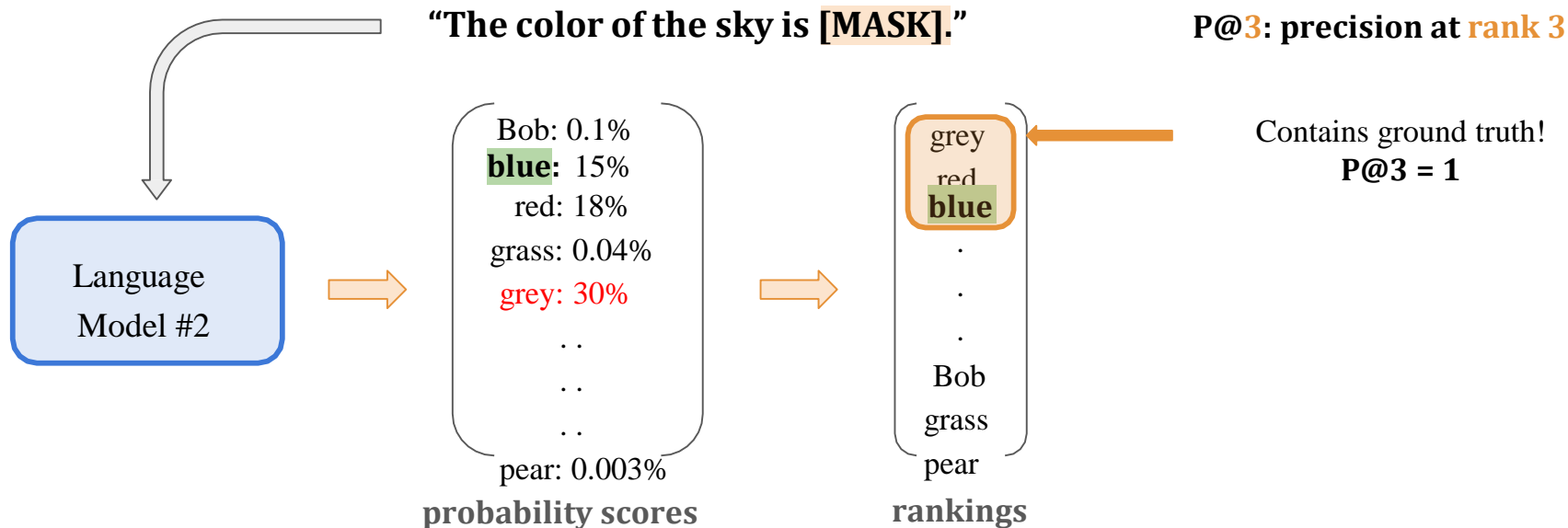
# Evaluation of LM via LAMA

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



# Evaluation of LM via LAMA

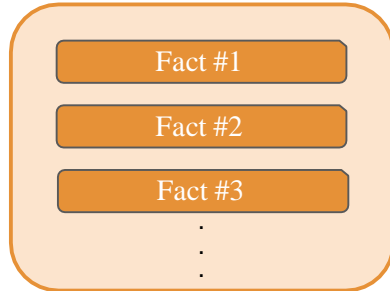
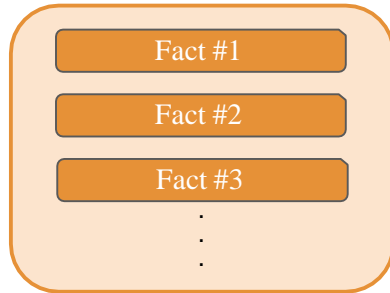
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



# Architecture of the LAMA probe

# Architecture of the LAMA probe

## Step 1: Compile knowledge sources

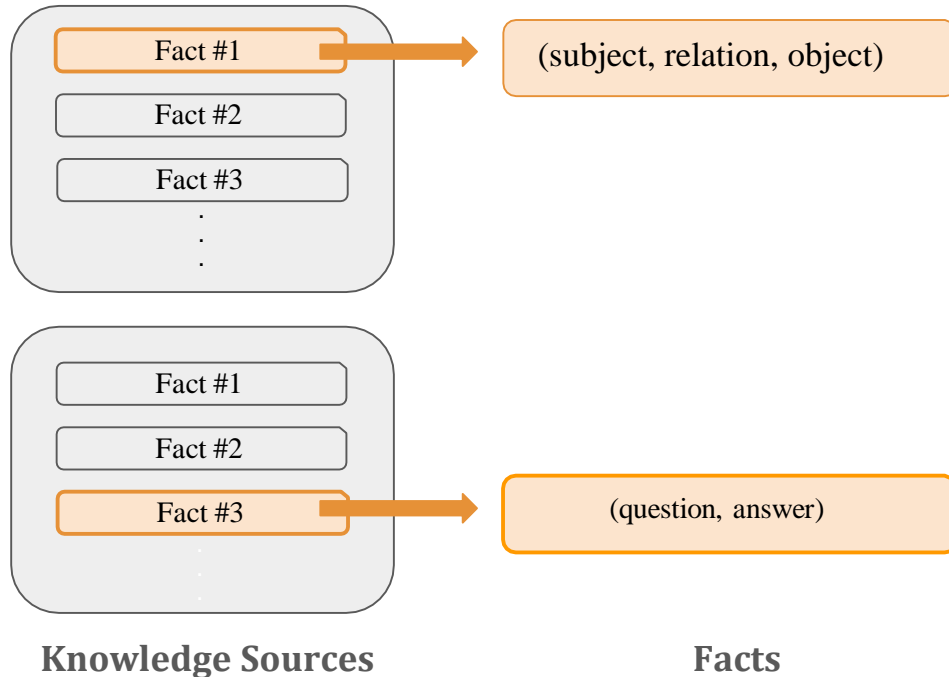


**Knowledge Sources**



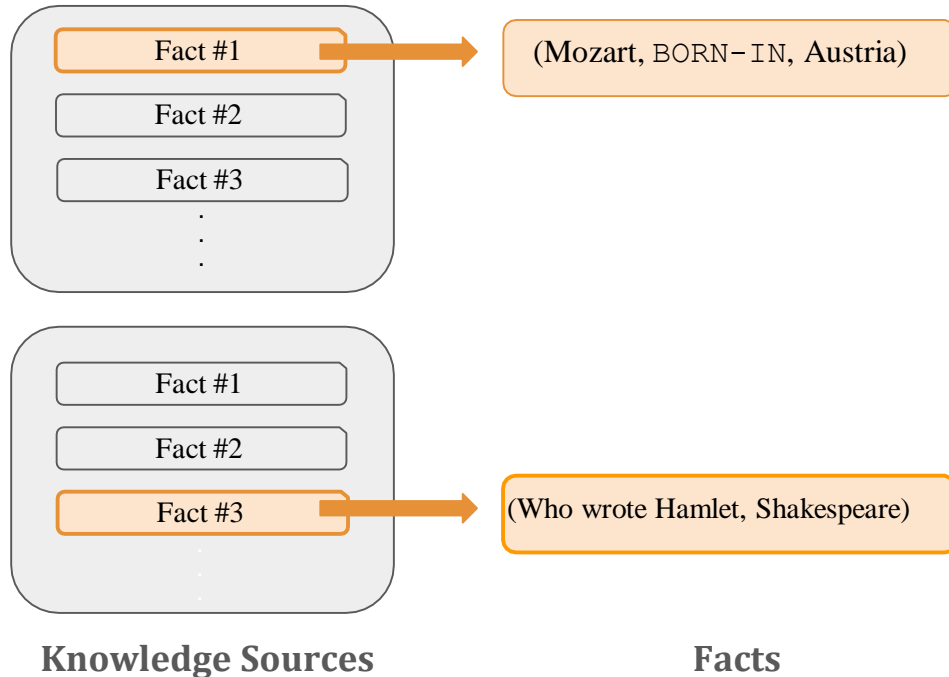
# Architecture of the LAMA probe

## Step 2: Formulate facts into triplets or question-answer pairs



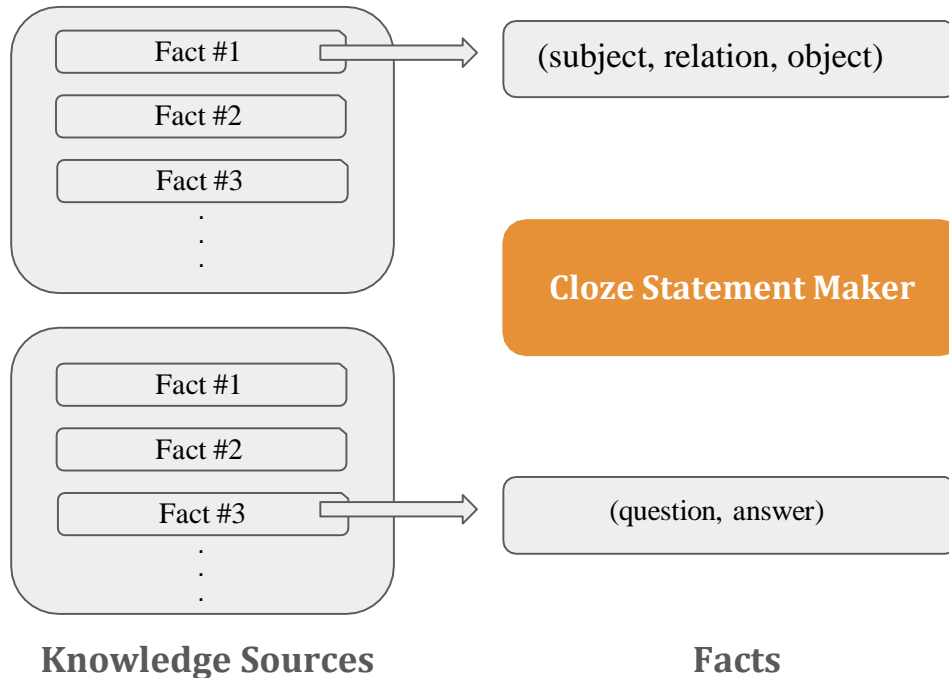
# Architecture of the LAMA probe

## Step 2: Formulate facts into triplets or question-answer pairs



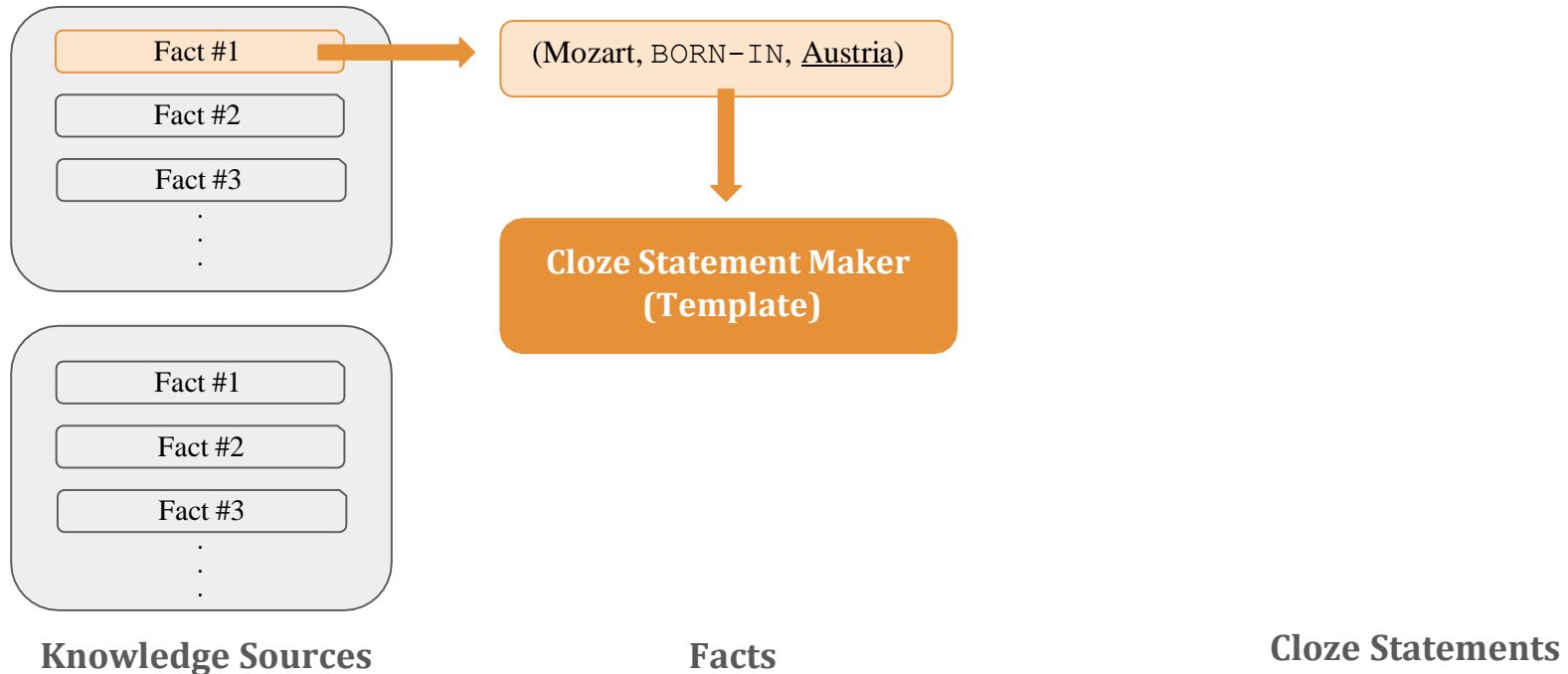
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



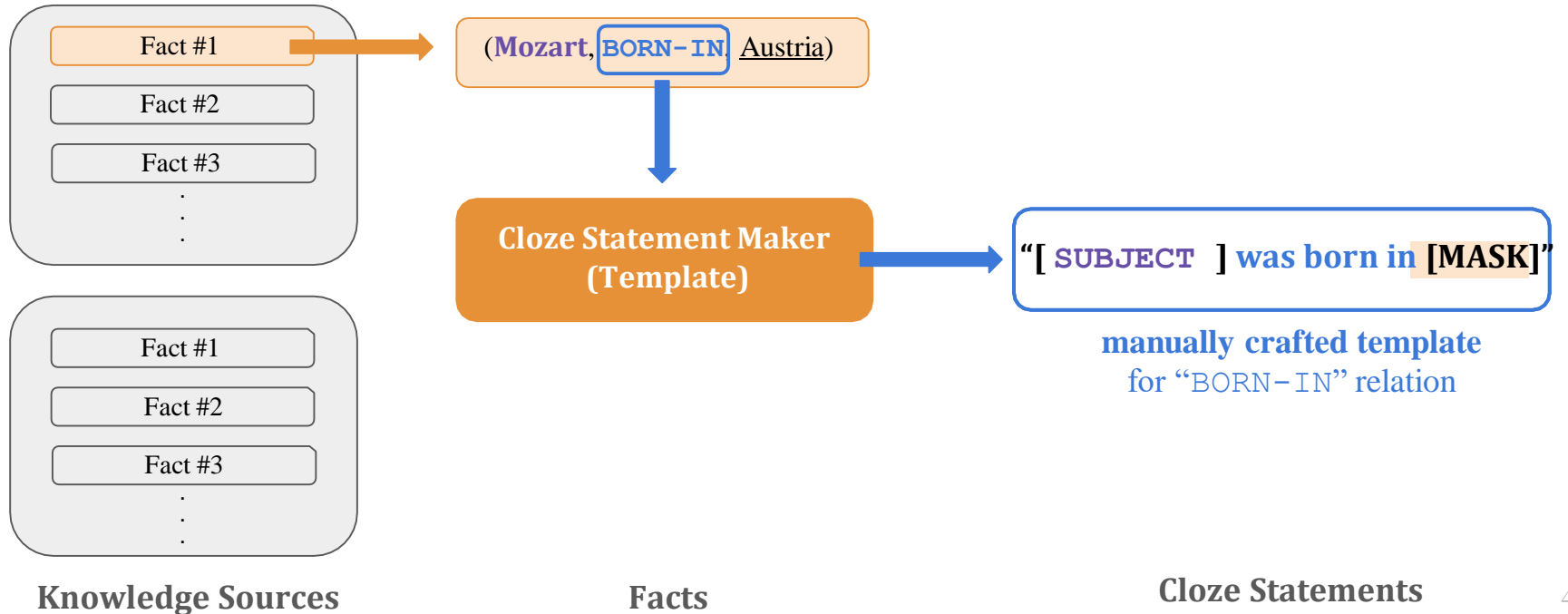
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



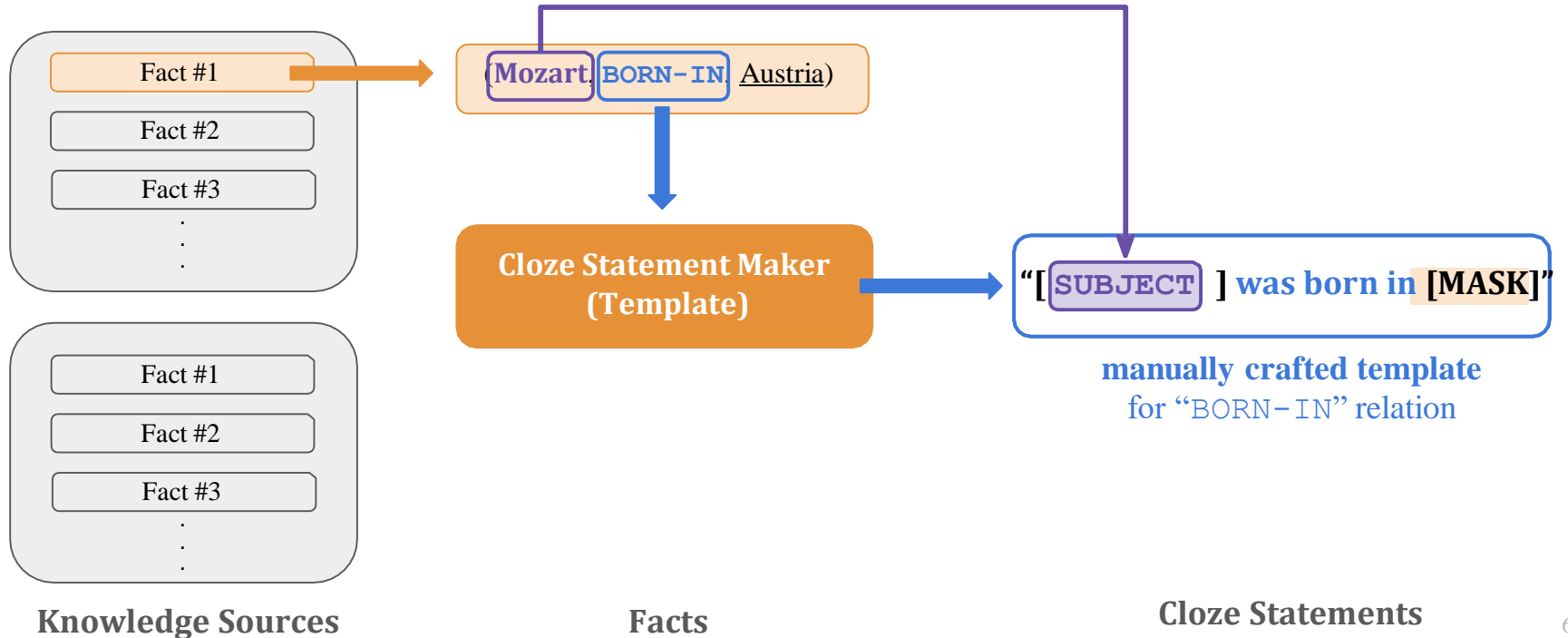
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



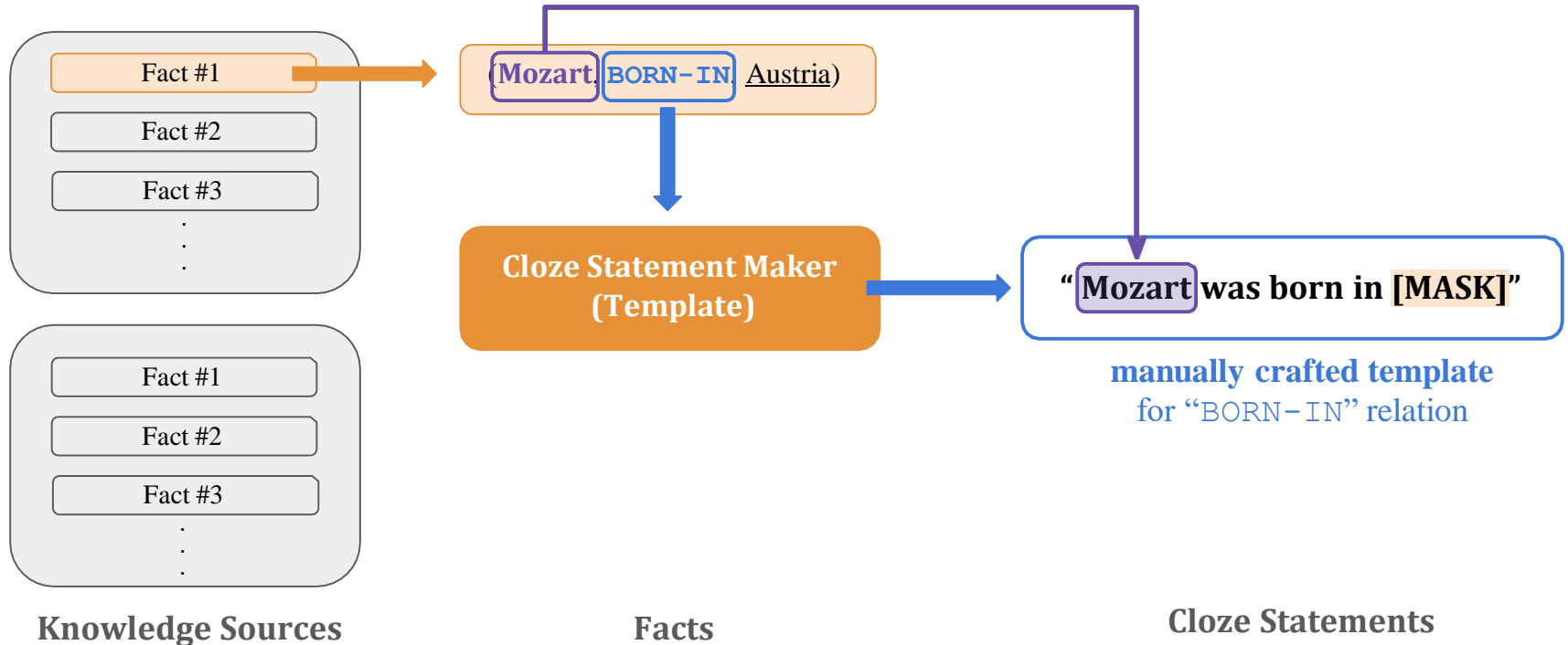
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



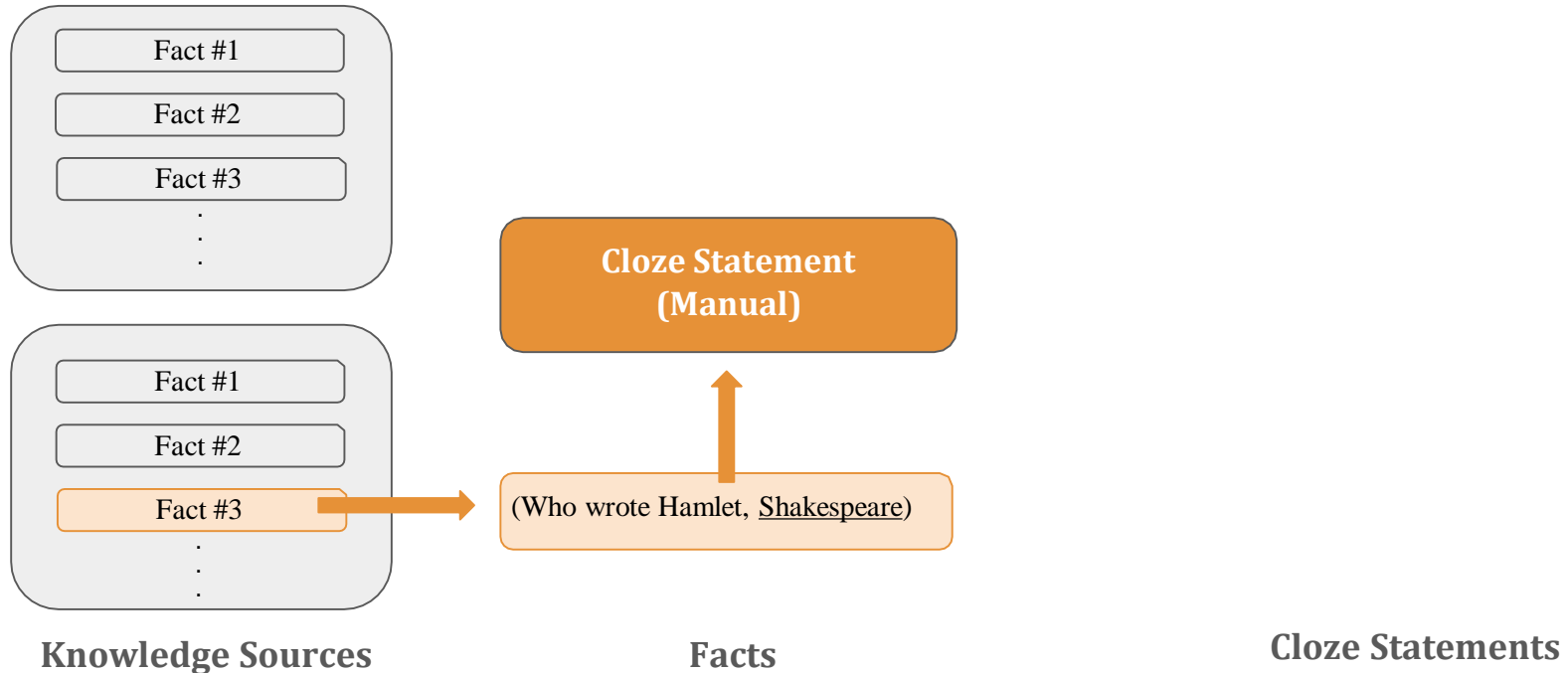
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



# Architecture of the LAMA probe

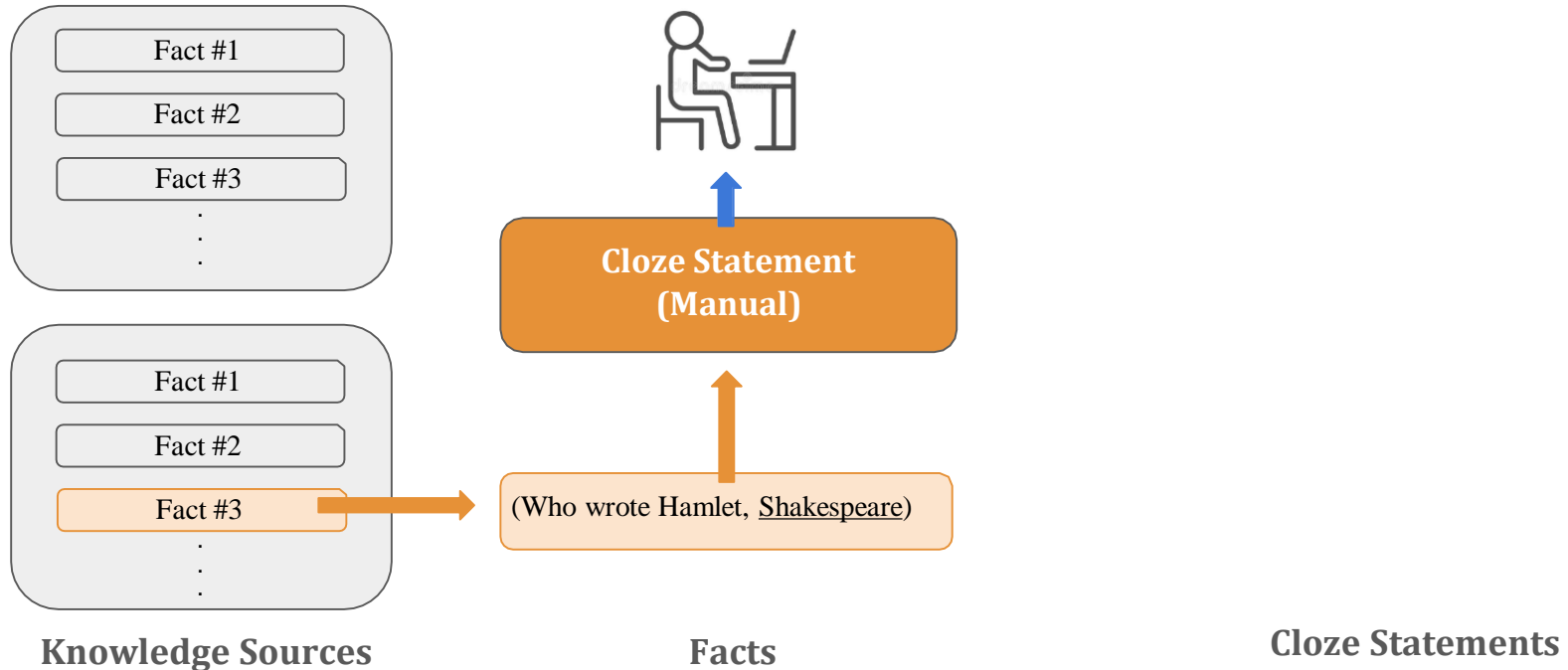
## Step 3: Create cloze statements, either manually or via templates





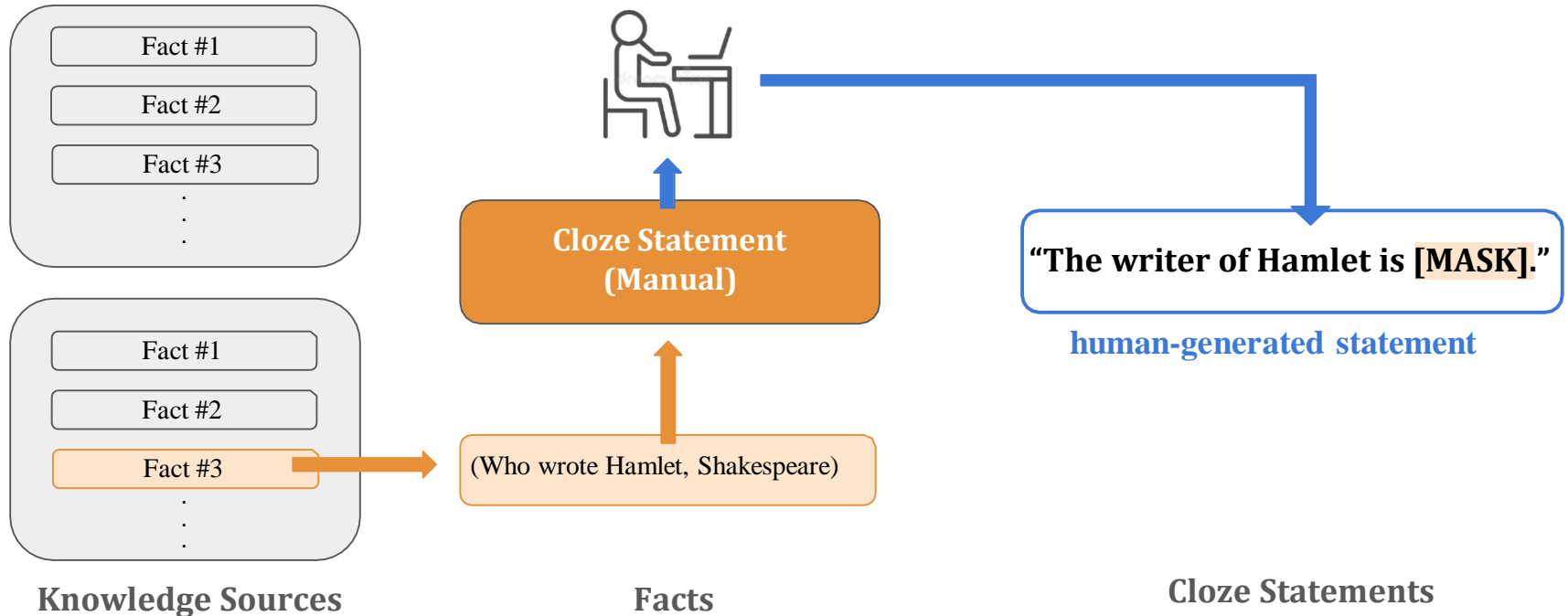
# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



# Architecture of the LAMA probe

## Step 3: Create cloze statements, either manually or via templates



**More discussions on LAMA**

# Any drawbacks?

**Overall pipeline of the LAMA Probe is in (Petroni et al., 2019)**

- Convert facts to cloze statements (either manually or using templates)
- Ask LM to rank candidate vocabulary and see if ground truth is in top  $k$  rank

**Can you think of any drawbacks of the probes?**

- Answers must be **single-token**
- Relies on **manual templates**
- Questions are constrained to **very specific and simple types** of questions

# Data leakage: **train-test overlap**

- **[Testing]** Many of the knowledge sources were extracted from **Wikipedia**
- **[Training]** However, pre-training corpora for language models almost always contain data from **Wikipedia**...
- How much of the amazing knowledge retrieval is due to **train-test overlap** in the knowledge probing benchmarks?

Train-test overlap is responsible for LM's ability to do knowledge retrieval! ([Lewis et al., 2020](#))

Model		Open Natural Questions				TriviaQA				WebQuestions			
		Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap
Open book	RAG	44.5	70.7	34.9	24.8	56.8	82.7	54.7	29.2	45.5	81.0	45.8	21.1
	DPR	41.3	69.4	34.6	19.3	57.9	80.4	59.6	31.6	42.4	74.1	39.8	22.2
	FID	51.4	71.3	48.3	34.5	67.6	87.5	66.9	42.8	-	-	-	-
Closed book	T5-11B+SSM	36.6	77.2	22.2	9.4	-	-	-	-	44.7	82.1	44.5	22.0
	BART	26.5	67.6	10.2	0.8	26.7	67.3	16.3	0.8	27.4	71.5	20.7	1.6
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0	28.9	81.5	11.2	0.0	26.4	78.8	17.1	0.0
	TF-IDF	22.2	56.8	4.1	0.0	23.5	68.8	5.1	0.0	19.4	63.9	8.7	0.0

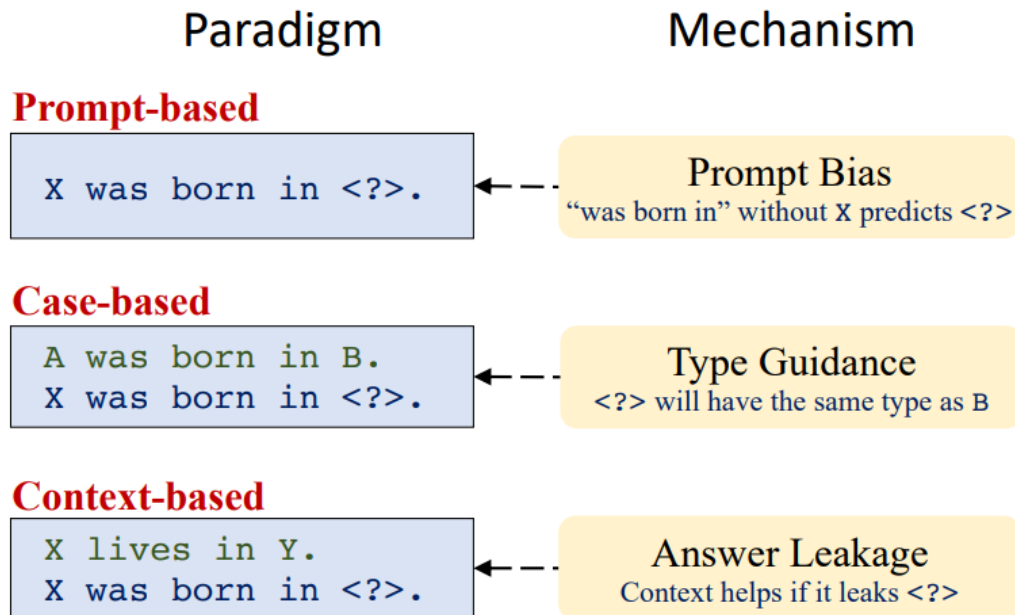
When there is **question overlap**, both open and closed-book LMs perform well

Train-test overlap is responsible for LM's ability to do knowledge retrieval! ([Lewis et al., 2020](#))

Model		Open Natural Questions				TriviaQA				WebQuestions			
		Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap
Open book	RAG	44.5	70.7	34.9	24.8	56.8	82.7	54.7	29.2	45.5	81.0	45.8	21.1
	DPR	41.3	69.4	20.6	19.3	57.9	80.4	50.6	31.6	42.4	74.1	20.8	22.2
	FID	51.4	71.3		34.5	67.6	87.5		42.8	-	-		-
Closed book	T5-11B+SSM	36.6	77.2	21.2	9.4	-	-	-	-	44.7	82.1	41.5	22.0
	BART	26.5	67.6	10.2	0.8	26.7	67.3	16.3	0.8	27.4	71.5	20.7	1.6
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0	28.9	81.5	11.2	0.0	26.4	78.8	17.1	0.0
	TF-IDF	22.2	56.8	4.1	0.0	23.5	68.8	5.1	0.0	19.4	63.9	8.7	0.0

But with **no question or answer overlap**, performance drops sharply!

# Revising LAMA – underlying mechanisms





# Revising LAMA – Reporting Bias

It is uninteresting to say one is **thinking** or **breathing**.  
But something related to **murders** seems interesting to share

---

Action	Actual Frequency for Lifetime (Source)
thinking	1,433,355,000 (50,000 per day)
breathing	660,489,984 (23,040 per day)
blinking	344,005,200 (12,000 per day)
eating	86001.3: 3 times per day
sleeping	28667.1: 1 time per day
working	20420.4: 5 times a week
exercising	8168.16: 2-3 times a week
getting married	1.66: 0-3 times per life
getting divorced	1: 0-2 times per life
being born	1
being named	1
dying	1
being abused	0.5 (source)
being injured	0.1263 (Episodes per 1,000 population: 126.3)
being raped	0.01 (18.3% of women (50.8% of population) and 1.4% of men (49.2% of population))
being killed	$4.01 \times 10^{-2}$ (murder + 1 out 28 in accident)
being arrested	0.031526 (3,152.6 arrests per 100,000)
being adopted	0.021 (7 million out of 328.2)
being murdered	$4.37 \times 10^{-3}$ (1 in 229 deaths)
being abandoned	0.000175 (7000 each year, out of 4M births)

---

# Revising LAMA – Reporting Bias

	<b>BERT</b>	<b>RoBERTa</b>	<b>GPT-2</b>		<b>BERT</b>	<b>RoBERTa</b>	<b>GPT-2</b>
The person ____.	wins (11.4)	said (5.8)	let (4.3)	The person is ____.	killed (7.5)	gone (6.3)	let (4.3)
	died (11.4)	responds (4.0)	see (3.9)		married (6.6)	deceased (3.8)	see (3.9)
	dies (10.6)	replied (3.4)	make (2.4)		dying (4.2)	arrested (2.9)	make (2.4)
	won (7.8)	dies (3.3)	get (2.1)		deceased (3.8)	missing (2.5)	get (2.1)
	lost (3.5)	died (2.9)	look (2.1)		eliminated (2.6)	responding (1.9)	look (2.1)
	said (2.4)	responded (2.5)	take (1.2)		retired (2.2)	involved (1.9)	take (1.2)
	speaks (1.9)	says (2.4)	set (1.2)		lost (2.0)	reading (1.9)	set (1.2)
	answered (1.6)	replies (2.2)	give (1.1)		arrested (2.0)	dying (1.9)	give (1.1)
	replied (1.3)	asked (2.1)	using (1.1)		elected (1.5)	confused (1.5)	using (1.1)
	loses (1.3)	commented (2.1)	go (1.1)		disabled (1.5)	reporting (1.5)	go (1.1)

Table 1: Top LM predictions for actions performed by people along with their scores (percents).

Reporting bias: due to Grice’s conversational maxim of quantity (Grice et al., 1975), people rarely state the obvious, thus many trivial facts (“people breathe”) are rarely mentioned in text, while uncommon events (“people murder”) are reported disproportionately (Gordon and Van Durme, 2013; Sorower et al., 2011).

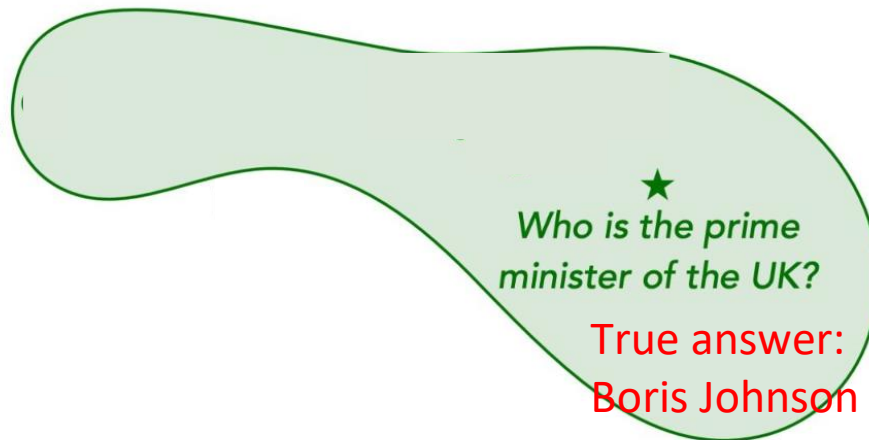
# Today's Lecture

- **Knowledge** in LLMs
  - LLMs as knowledge bases
  - Facts updating for LLMs
- **Reasoning** in LLMs
  - Why reasoning is special in LLMs
- **Prompt** Techniques for better reasoning

How to update knowledge in  
pre-trained models?

# Edit What, Exactly?

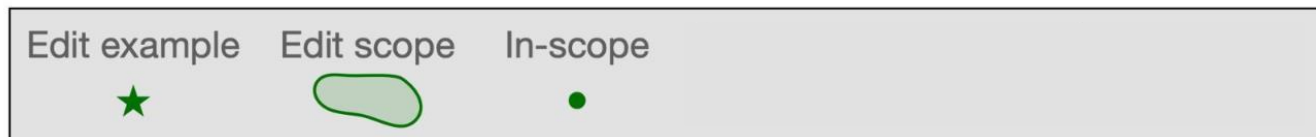
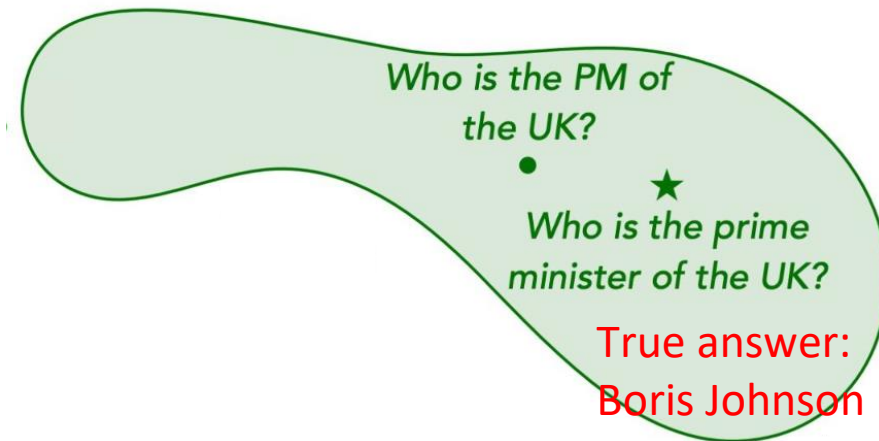
## Defining the problem



Edit example	Edit scope
★	

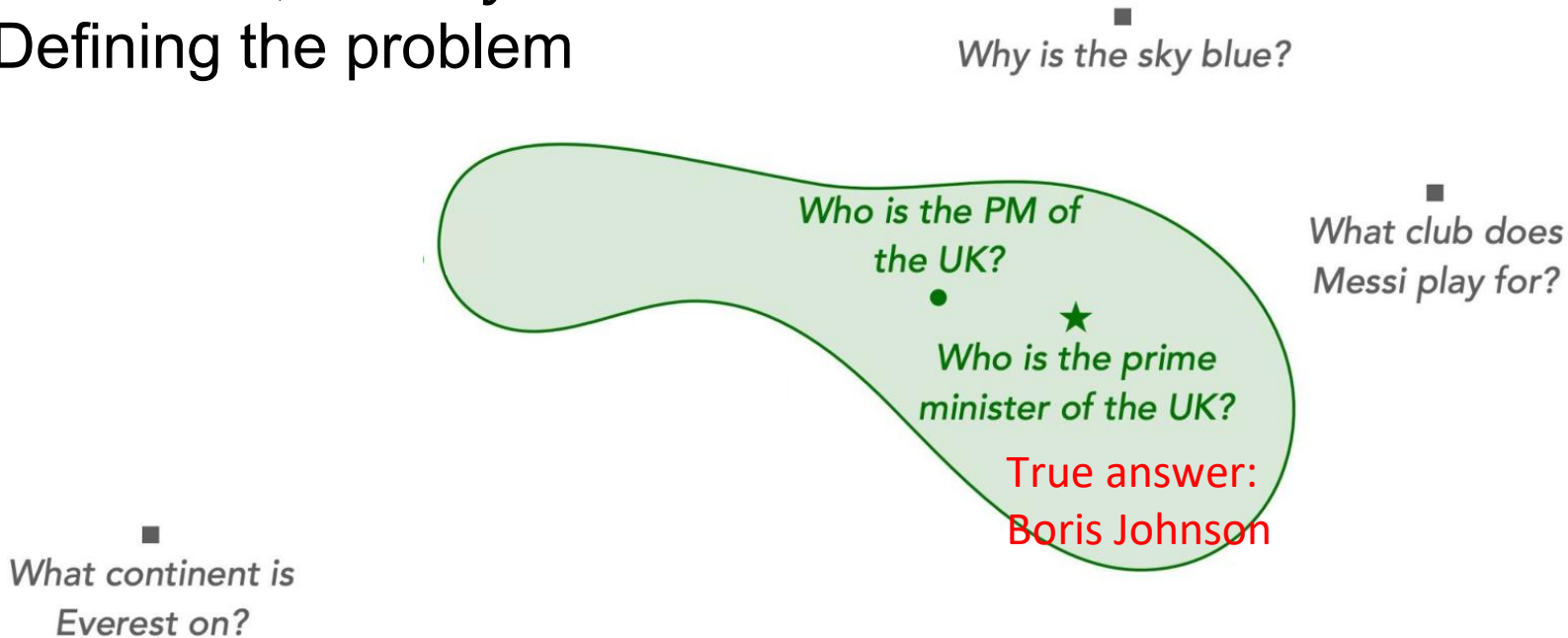
# Edit What, Exactly?


## Defining the problem



# Edit What, Exactly?

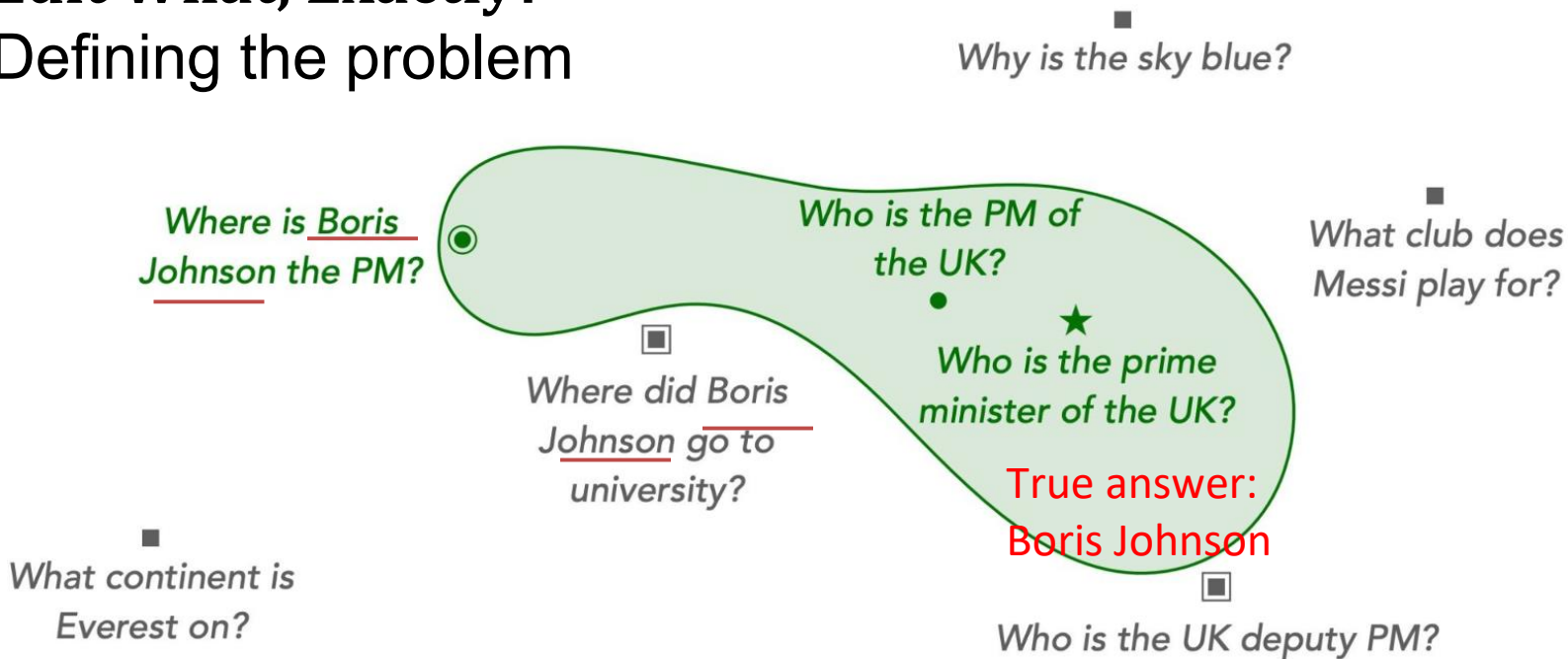
## Defining the problem




Edit example	Edit scope	In-scope	Out-of-scope
★		●	■

# Edit What, Exactly?

## Defining the problem

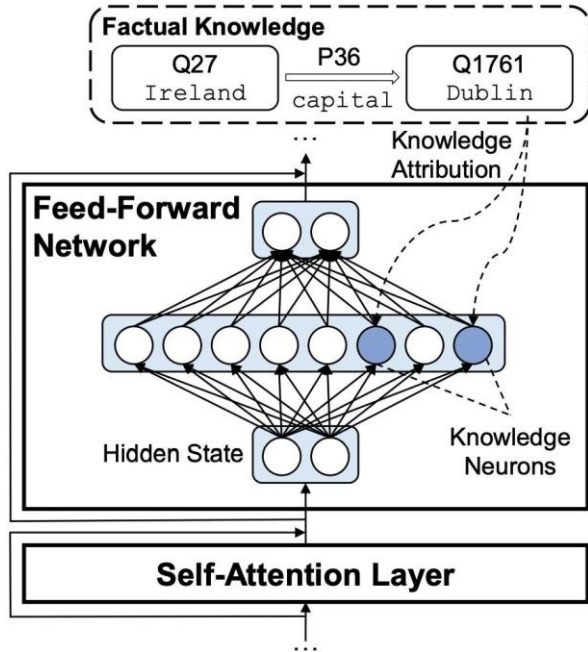


Edit example	Edit scope	In-scope	Out-of-scope	Hard in/out-of-scope
★		●	■	● ■



How to edit knowledge in pre-trained models?

# Knowledge Neurons



- What is a knowledge neuron
  - **Activations** after the first feed-forward layer
- Assumption
  - Knowledge neurons are associated with factual knowledge
- Implications
  - If we can identify these neurons, we can alter them to edit (update/erase) knowledge.
  - No additional training is involved.

# Identify knowledge neurons

Given a relational fact e.g. (Mozart, BORN-IN, Austria)

1. produce  $N$  diverse prompts;
2. for each prompt, calculate the knowledge attribution scores of neurons;
3. for each prompt, retain the neurons with attribution scores greater than the attribution threshold  $T$ , obtaining the coarse set of knowledge neurons;
4. considering all the coarse sets together, retain the knowledge neurons shared by more than  $p\%$  prompts.

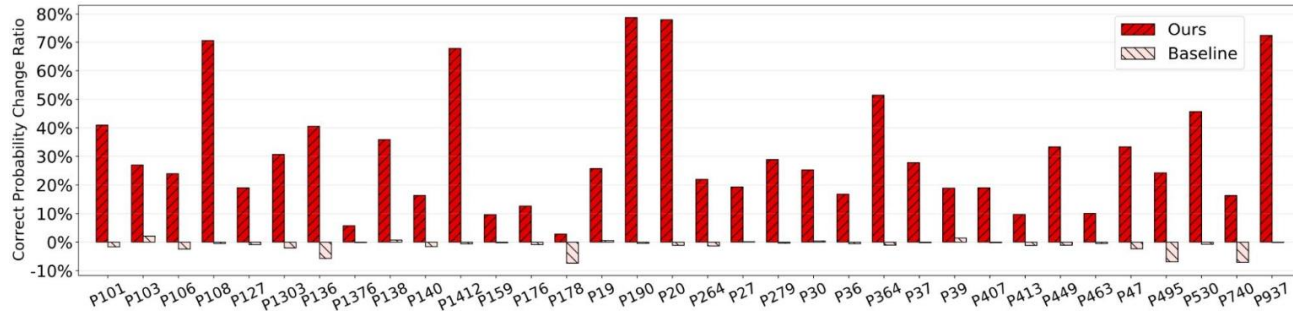
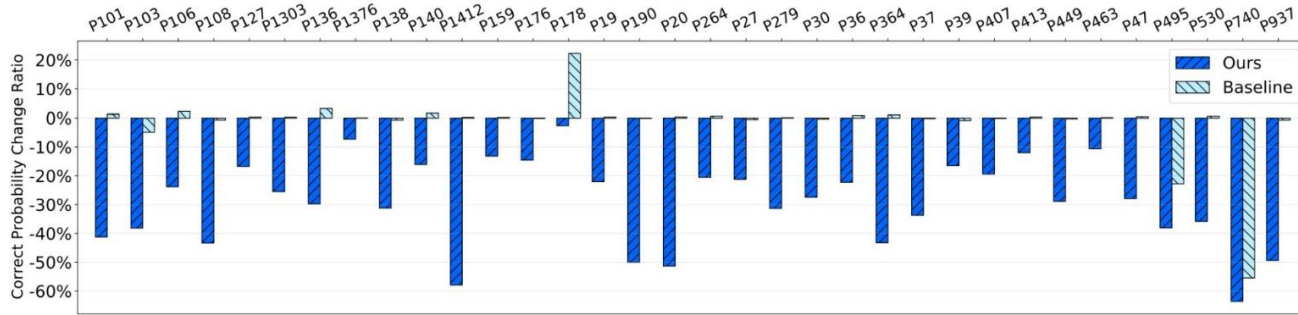
# Knowledge neuron editing

Knowledge neuron: activations after the first feed-forward layer

**Suppressing** the neuron: activation = 0

**Amplifying** the neuron: activation = 2\*activation

# Suppressing or Amplifying Knowledge Neurons



**Suppressing** the neurons **hurt** performance and **amplifying** neurons **increase** performance by up to 30% on average.

# Drawback

**Sensitive** to the format of the prompt collected by human

Relations	Template #1	Template #2	Template #3
P176 (manufacturer)	[X] is produced by [Y]	[X] is a product of [Y]	[Y] and its product [X]
P463 (member_of)	[X] is a member of [Y]	[X] belongs to the organization of [Y]	[X] is affiliated with [Y]
P407 (language_of_work)	[X] was written in [Y]	The language of [X] is [Y]	[X] was a [Y]-language work

Table 1: Example prompt templates of three relations in PARAREL. [X] and [Y] are the placeholders for the head and tail entities, respectively. Owing to the page width, we show only three templates for each relation. Prompt templates in PARAREL produce 253,448 knowledge-expressing prompts in total for 27,738 relational facts.

# Today's Lecture

- **Knowledge in LLMs**
  - LLMs as knowledge bases
  - Facts updating for LLMs
- **Reasoning in LLMs**
  - Why reasoning is special in LLMs
- **Prompt** Techniques for better reasoning

# Human Intelligence vs. Traditional machine learning? (Hint: reasoning)

Humans	Traditional machine learning
Learn from only a few examples	Large amounts of labeled data
Can explain rationale for decisions	Black box
Out-of-distribution generalization	No

Attempts to fill the gap in the past decades

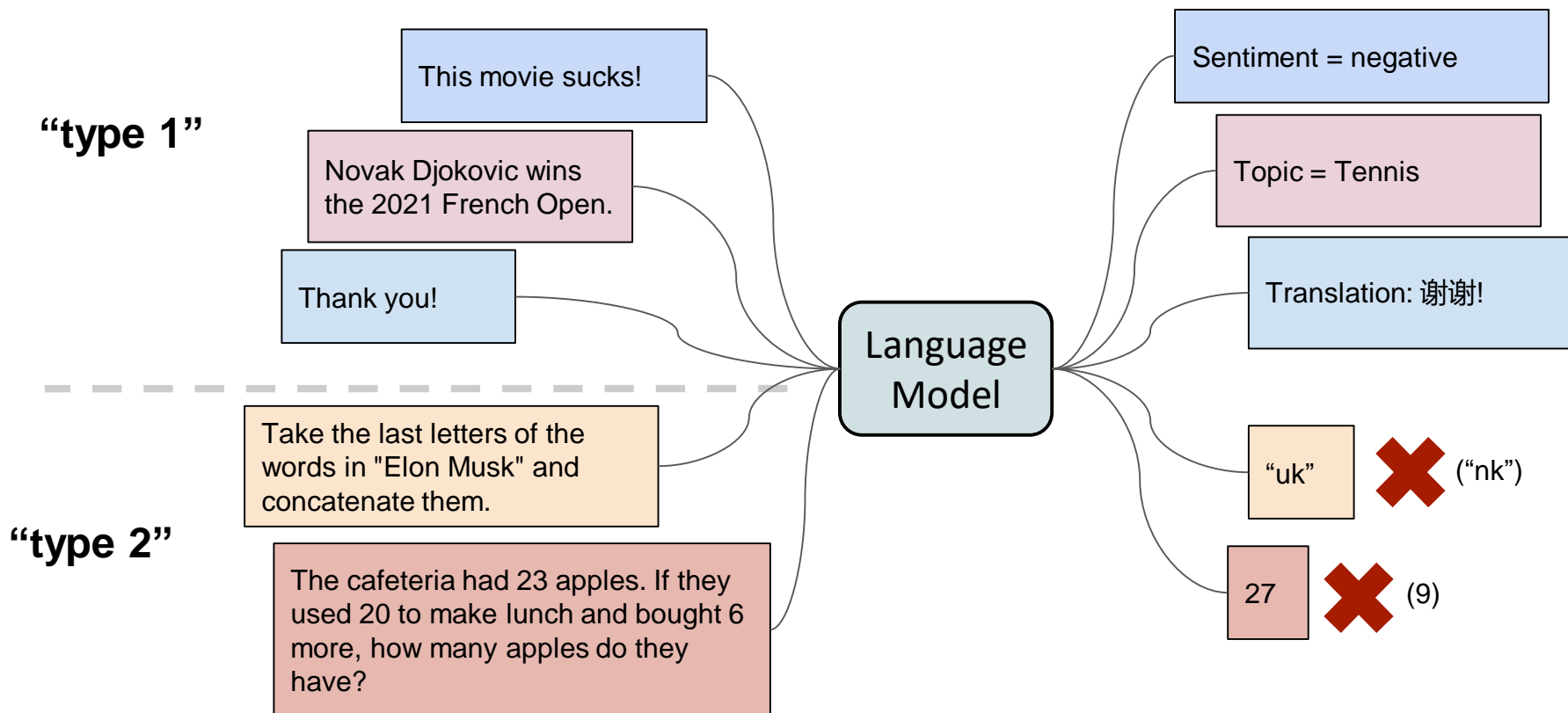
- Bayesian machine learning, kernel machines, nonparametric, sparsity, large-margin, semi-supervised learning, manifold learning, graph-based learning, transductive learning, meta learning, transfer learning, active learning, domain adaptation, structural learning, low-rank, ...

**FAILED**

*Teaching language models to reason (Denny Zhou), 2023.*



# Multi-step reasoning is hard for language models



# What do language models learn from next-word prediction?

<i>Grammar</i>	In my free time, I like to { <u>run</u> , banana}
<i>Lexical semantics</i>	I went to the zoo to see giraffes, lions, and { <u>zebras</u> , spoon}
<i>World knowledge</i>	The capital of Denmark is { <u>Copenhagen</u> , London}
<i>Sentiment analysis</i>	Movie review: I was engaged and on the edge of my seat the whole time. The movie was { <u>good</u> , bad}
<i>Harder sentiment analysis</i>	Movie review: Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was { <u>bad</u> , good}
<i>Translation</i>	The word for “pretty” in Spanish is { <u>bonita</u> , hola}
<i>Spatial reasoning</i>	[...] Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the { <u>kitchen</u> , store}
<i>Math question</i>	First grade arithmetic exam: $3 + 8 + 4 =$ { <u>15</u> , 11}

[thousands (millions?) more]

Extreme multi-task learning!

# What can't language models learn from next-word prediction?

<i>Current world knowledge</i>	The stock price of APPL on March 1st, 2023 is {???
<i>Arbitrarily long arithmetic</i>	$36382894730 + 238302849204 = \{???\}$
<i>Many-step reasoning</i>	Take the nineteenth digit of Pi and multiply it by the e to the fourth power. The resulting ones-digit of the resulting number is {???
<i>Predict the future</i>	The winner of the FIFA world cup in 2026 is {???
<i>Information not in the training data</i>	Jason Wei's favorite color is {???
<i>Extremely long inputs</i>	[2,000 page Harry Potter fan-fiction] What happened after Harry opened the chest for the second time? {???

## Jason Wei's rule of thumb (经验法则)

language models can do (with decent accuracy)  
most text tasks that **an average human can do in 1 minute.**



2018

(2023)

Future ...?

...
Protein discovery
Clinical diagnosis
Play chess well
High-level planning
Abstract reasoning
Simple math
Commonsense reasoning
Know world knowledge
Translation
Sentiment analysis
Generate coherent text
Be grammatically correct

...
Protein discovery
Clinical diagnosis
Play chess well
High-level planning
Abstract reasoning
Simple math
Commonsense reasoning
Know world knowledge
Translation
Sentiment analysis
Generate coherent text
Be grammatically correct

...
(?) Protein discovery
(?) Clinical diagnosis
(?) Play chess well
(?) High-level planning
(?) Abstract reasoning
Simple math
Commonsense reasoning
Know world knowledge
Translation
Sentiment analysis
Generate coherent text
Be grammatically correct

Abstract reasoning **might be done** in the year of 2025!

# OpenAI Imagines Our AI Future

## Stages of Artificial Intelligence

---

Level 1	Chatbots, AI with conversational language
Level 2	Reasoners, human-level problem solving
Level 3	Agents, systems that can take actions
Level 4	Innovators, AI that can aid in invention
Level 5	Organizations, AI that can do the work of an organization

Source: Bloomberg reporting

 公众号 · 新智元

# Reasoning Problems

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?  
A: The answer is **5**

Mathematical Reasoning

Q: Take the last letters of the words in "Elon Musk" and concatenate them  
A: The answer is **nk**.

Symbolic Reasoning

Q: What home entertainment equipment requires cable?  
Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet  
A: The answer is **television**.

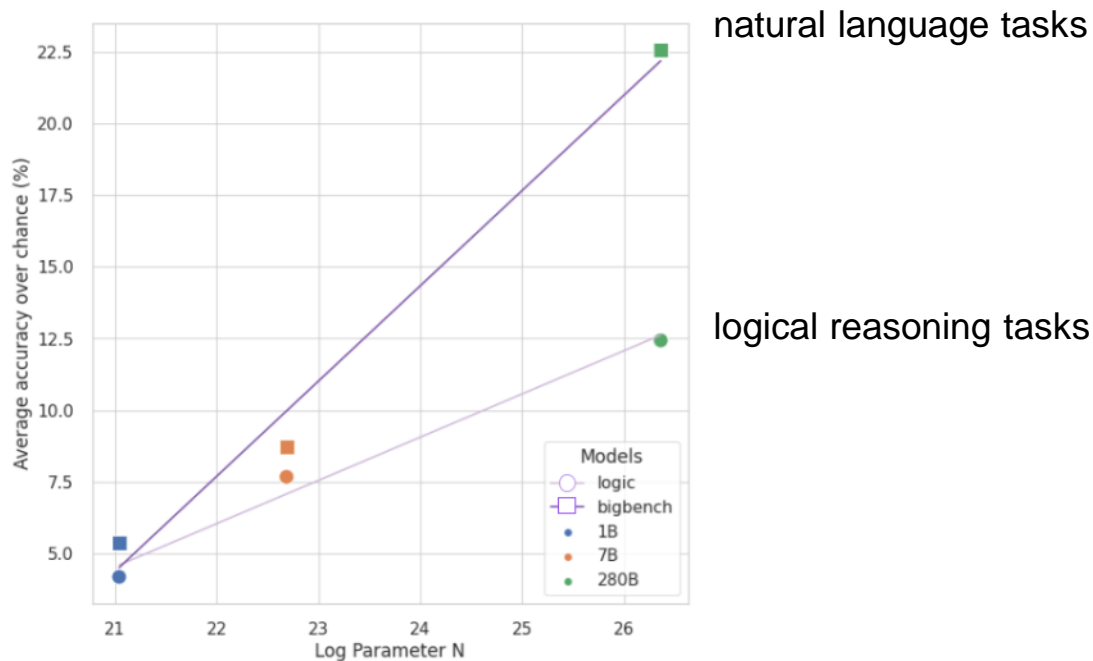
Commonsense Reasoning

Q: Wolves are afraid of mice. Sheep are afraid of wolves. Emily is a wolf. What is Emily afraid of?  
A: The answer is **mice**.

Logical Reasoning

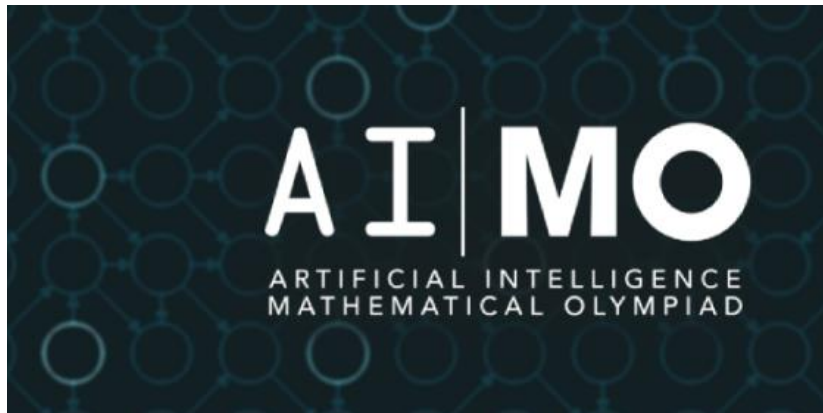
# Scaling laws are worse for logical reasoning in 2022

(Creswell et al. 2022)





## Difficult tasks:



<https://www.kaggle.com/c/ai-mathematical-olympiad-progress-prize-2>

# AIMO

What is the minimum value of  $5x^2 + 5y^2 - 8xy$  when  $x$  and  $y$  range over all real numbers such that  $|x - 2y| + |y - 2x| = 40$ ?

800

There exists a unique increasing geometric sequence of five 2-digit positive integers. What is their sum?

211

For how many positive integers  $m$  does the equation  $\lfloor \lfloor x-1 \rfloor - 2 \rfloor = \frac{m}{100}$  have 4 distinct solutions?

199

# Interesting task: Mathematical modeling

## #### 模型背景

物资调拨模型关注于如何高效、成本效益地将物资从供应地点分配到需求地点。在供应链管理、紧急物资分配等场景中，合理的物资调拨能够确保物资及时供应，同时降低总体成本。

## #### 模型描述

假设有一组供应点集合 $S$ 和一组需求点集合 $D$ 。每个供应点 $i \in S$ 具有供应能力 $a_i$ ，每个需求点 $j \in D$ 有需求量 $b_j$ 。从供应点 $i$ 到需求点 $j$ 运输单位物资的成本为 $c_{ij}$ 。模型的目标是确定从每个供应点到每个需求点的物资调拨量 $x_{ij}$ ，以满足所有需求点的需求，同时最小化总运输成本。

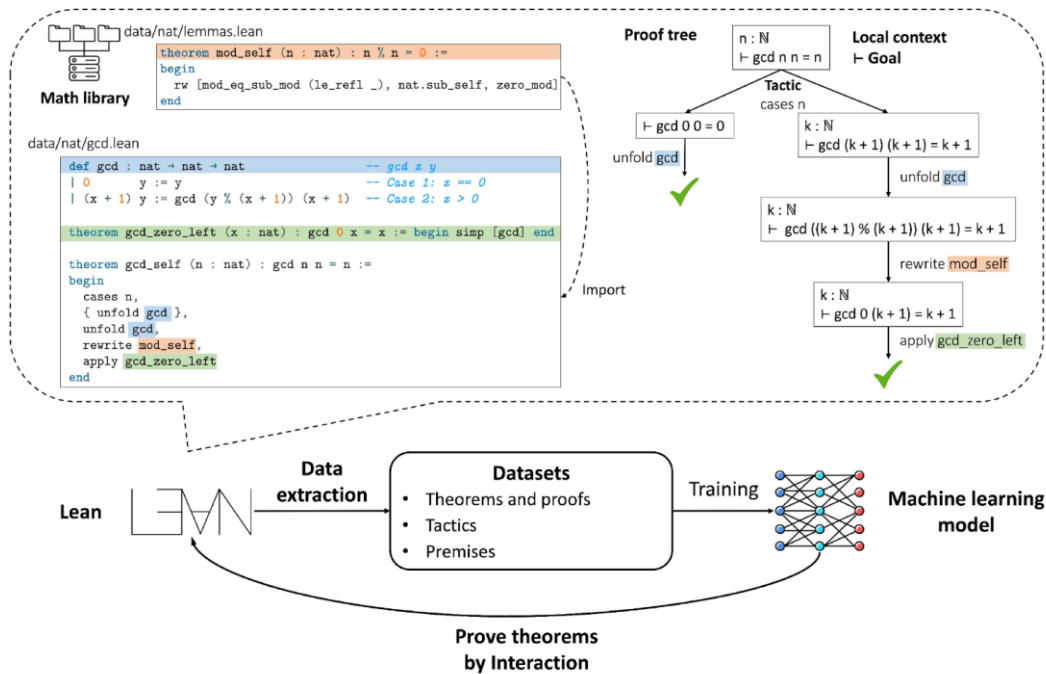
## #### 模型目标

- 服务水平最大化**：提高物资调拨的服务水平，包括减少配送时间或提高配送的可靠性。
  - 公式： $\max S = \sum_{j \in D} \alpha_j \left(1 - \frac{1}{\sum_{i \in S} t_{ij} x_{ij}}\right)$
  - 其中， $t_{ij}$ 是从供应点 $i$ 到需求点 $j$ 的运输时间， $\alpha_j$ 是需求点 $j$ 对运输时间的最大容忍值， $\alpha_j$ 是与需求点 $j$ 的服务水平重要性相关的权重。
- 环境影响最小化**：在物资调拨过程中，尽量减少对环境的负面影响，例如减少碳排放。
  - 公式： $\min E = \sum_{i \in S} \sum_{j \in D} e_{ij} x_{ij}$
  - 其中， $e_{ij}$ 是从供应点 $i$ 到需求点 $j$ 运输单位物资产生的碳排放量。
- 成本最小化**：最小化从所有供应点到所有需求点的总运输成本。
  - 公式： $\min Z = \sum_{i \in S} \sum_{j \in D} c_{ij} x_{ij}$
- 库存水平平衡**：优化各供应点和需求点的库存水平，以减少过剩或缺口的情况，保持库存的稳定。
  - 公式： $\min I = \sum_{i \in S} |x_i - L_i| + \sum_{j \in D} |x_j - L_j|$
  - 其中， $x_i$ 表示供应点 $i$ 的实际库存水平， $x_j$ 表示需求点 $j$ 的实际库存水平， $L_i$ 和 $L_j$ 分别表示供应点 $i$ 和需求点 $j$ 的理想库存水平。
- 多模式运输优化**：考虑不同运输方式（如公路、铁路、海运等）的成本和效率，优化运输模式的选择。
  - 公式： $\min M = \sum_{i \in S} \sum_{j \in D} \sum_{k \in K} c_{ijk} x_{ijk}$
  - 其中， $K$ 是运输模式的集合， $c_{ijk}$ 是使用运输模式 $k$ 从供应点 $i$ 到需求点 $j$ 的单位运输成本， $x_{ijk}$ 是通过运输模式 $k$ 从 $i$ 到 $j$ 的物质量。

## #### 模型约束

- 供应约束**：从每个供应点发出的物资总量不能超过该点的供应能力。
  - 公式： $\sum_{j \in D} x_{ij} \leq a_i, \forall i \in S$
- 需求约束**：每个需求点接收的物资总量必须满足该点的需求量。
  - 公式： $\sum_{i \in S} x_{ij} = b_j, \forall j \in D$
- 非负约束**：调拨的物质量不能为负。
  - 公式： $x_{ij} \geq 0, \forall i \in S, \forall j \in D$
- 可选的库存水平优化约束**：对于有库存管理需求的场景，可能需要考虑库存水平的约束，确保库存水平在安全库存和最大库存水平之间。
  - 示意公式： $L_i \leq \text{库存水平}_i \leq U_i, \forall i \in S$
  - 其中， $L_i$ 和 $U_i$ 分别表示供应点 $i$ 的安全库存水平和最大库存水平。
- 运输模式选择约束**：确保每条运输线路只选择一种运输模式。
  - 公式： $\sum_{k \in K} x_{ijk} = x_{ij}, \forall i \in S, \forall j \in D$
  - 这确保了从供应点 $i$ 到需求点 $j$ 的物质量 $x_{ij}$ 通过某一种运输模式 $k$ 进行调拨。
- 时间窗约束**：满足需求点的特定配送时间窗要求。
  - 公式： $t_{ij}^{\text{start}} \leq t_{ij} \leq t_{ij}^{\text{end}}, \forall i \in S, \forall j \in D$
  - 其中， $t_{ij}^{\text{start}}$ 和 $t_{ij}^{\text{end}}$ 分别表示从供应点 $i$ 到需求点 $j$ 的配送允许的最早开始时间和最晚结束时间， $t_{ij}$ 是实际配送时间。

# Interesting task: Automatic theorem proving



# Today's Lecture

- **Knowledge** in LLMs
  - LLMs as knowledge bases
  - Facts updating for LLMs
- **Reasoning** in LLMs
  - Why reasoning is special in LLMs
- **Prompt** Techniques for better reasoning

# What is In-Context Learning?

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // \_\_\_\_\_

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

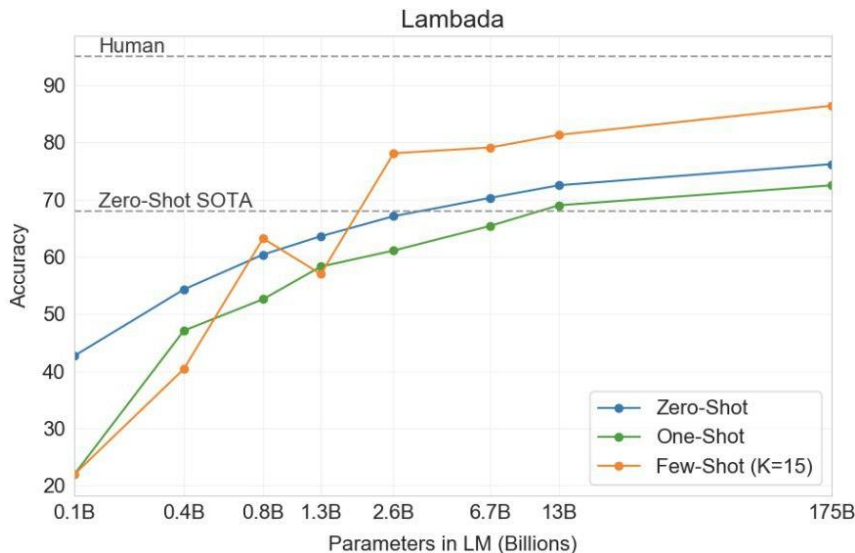
Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // \_\_\_\_\_

LM

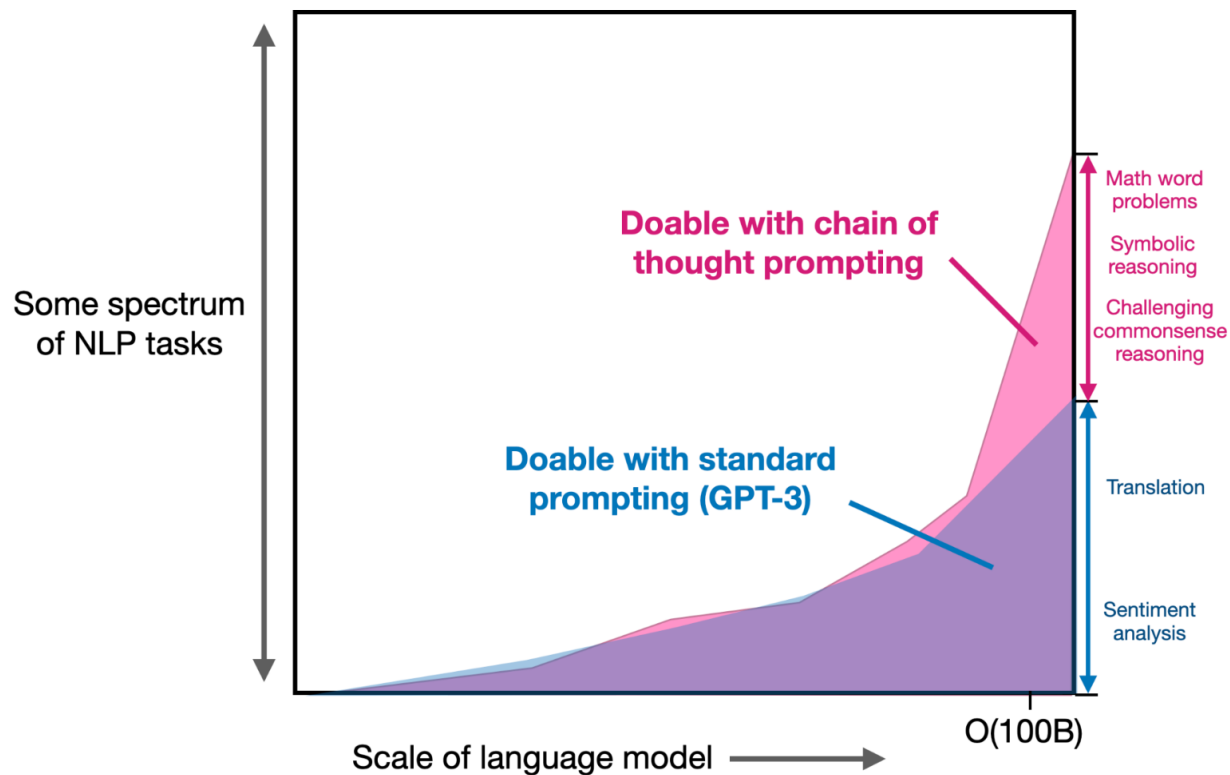
# What Can In-Context Learning Do?

- No parameter tuning need
- Only need few examples for downstream tasks
- GPT-3 improved SOTA on LAMBADA (last word prediction task) by 18%!



Works like magic!

# LLMs with CoT could do reason !





# What is Chain of Thought prompting (CoT)?

## 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?

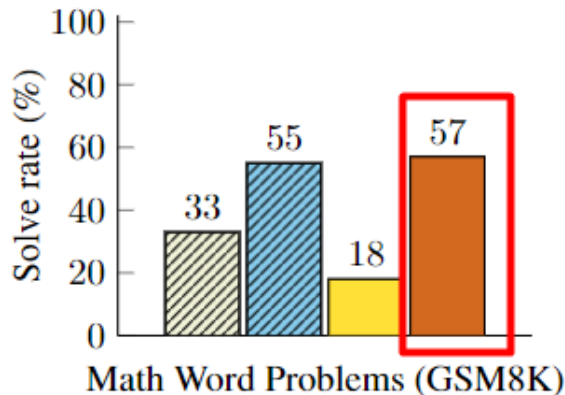
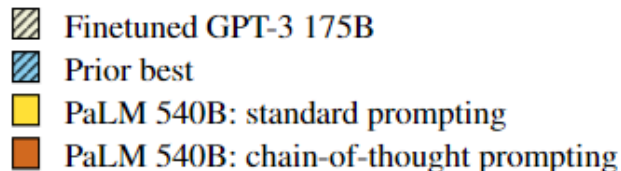
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is  $5 + 6 = 11$ . The answer is 11.

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. ✅

<input, intermediate results, output>



- decompose into easier intermediate steps
- interpretable

# Zero-Shot CoT – Let's think step by step

(a) Few-shot

Examples

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. ✗

(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 ✗

(b) Few-shot-CoT (Wei et al., 2022)

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. ✓

CoT Examples

Step-by-step Answer

(d) Zero-shot-CoT (KoJima et al., 2022)

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. ✓

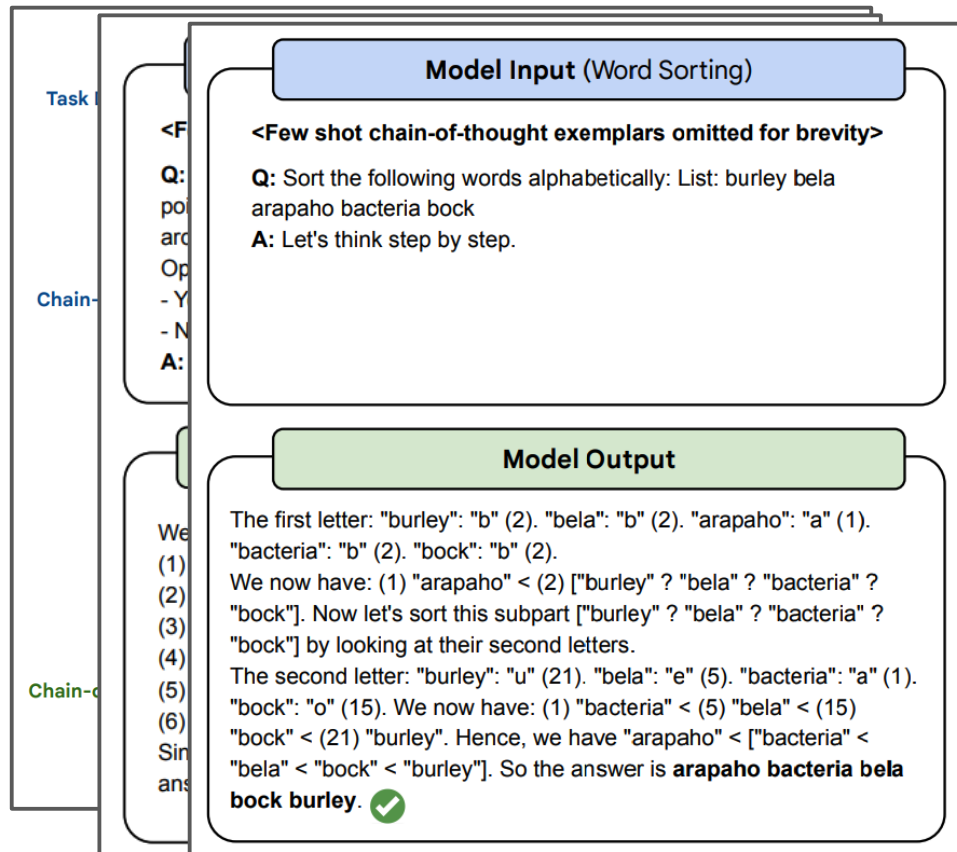
Step-by-step Answer

# CoT on BIG-Bench: Benchmark

## BIG-Bench Hard (BBH):

- 23 challenging tasks from BIG-Bench benchmark where no model beats avg. human rater

[Challenging BIG-Bench tasks and whether chain-of-thought can solve them \(2022\).](#)



# CoT on BIG-Bench: Result summary

	BBH all (23 tasks)	# tasks above avg. human-rater
Average human-rater	67.7	N/A
Max human-rater	94.4	23 / 23
Best prior BIG-Bench result	50.9	0 / 23
Codex (code-davinci-002)		
- Answer-only prompting	56.6	5 / 23
- CoT prompting	73.9 (+16.7)	17 / 23

[Challenging BIG-Bench tasks and whether chain-of-thought can solve them \(2023\).](#)

Model much lower than average human rater

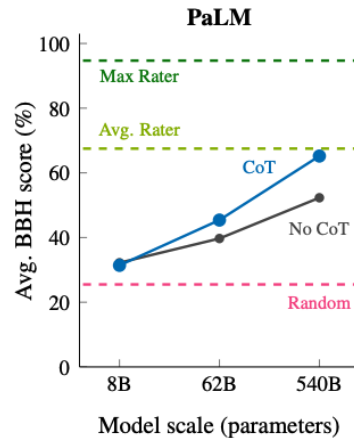
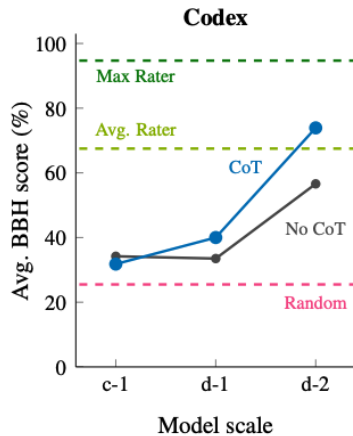
Detail: better formatting (options, task description) already beats prior best

CoT prompting improves by performance by +16.7%, passes avg. human on majority of tasks

# CoT on BIG-Bench: Scaling

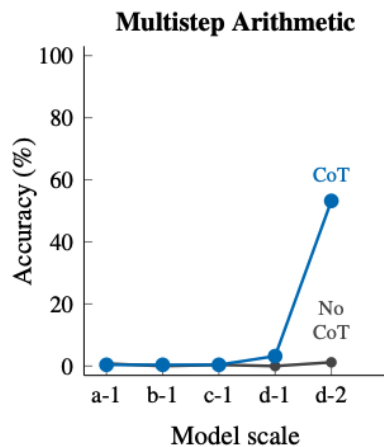
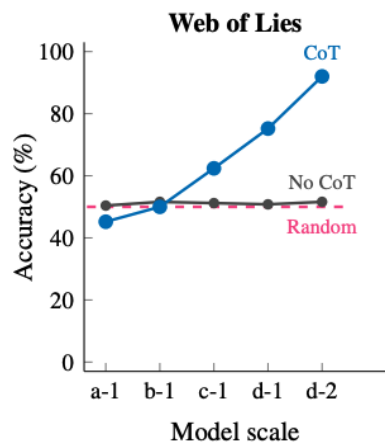
- CoT requires sufficient model scale for positive delta

[Challenging BIG-Bench tasks and whether chain-of-thought can solve them \(2023\).](#)



# CoT on BIG-Bench: Emergence

- No-CoT performance is flat, i.e., hasn't unlocked emergence (yet :))
- CoT unlocks emergent performance



[Challenging BIG-Bench tasks and whether chain-of-thought can solve them \(2023\).](#)

**Model Input (Multistep Arithmetic)**

<Few shot chain-of-thought exemplars omitted for brevity>

**Q:**  $((4 + 7 * 4 - -5) - (-4 - 1 - -4 - 4)) =$   
**A:** Let's think step by step.

**Model Output**

Let's recall that the order of operations in mathematics is as follows: (1) Parentheses, (2) exponents, (3) multiplication and division (from left to right), (4) addition and subtraction (from left to right). So, remember to always compute the expressions inside parentheses or brackets first.

This equation can be written as "A - B", where A =  $(4 + 7 * 4 - -5)$  and B =  $(-4 - 1 - -4 - 4)$ .

Let's calculate A =  $(4 + 7 * 4 - -5) = (4 + (7 * 4) - -5) = (4 + 28) - -5) = (4 + 28 - -5) = (4 + 28 + 5) = 37$ .

Let's calculate B =  $(-4 - 1 - -4 - 4) = ((-4 - 1) - -4 - 4) = ((-5) - -4 - 4) = ((-5 - -4) - 4) = ((-5 + 4) - 4) = (-1 - 4) = -5$ .

Then, the final equation is A - B =  $37 - -5 = 37 + 5 = 42$ . So the answer is **42**. ✓

# Multilingual chain-of-thought prompting

Prompt the model with Bengali math problems and Bengali reasoning

Input is highly improbable (Bengali is 0.01% of pre-training data)

## Model Input

প্রশ্ন: রজারের 5টি টেনিস বল আছে। সে আরও 2 ক্যান টেনিস বল কিনেছে। প্রতিটি ক্যান 3টি করে টেনিস বল আছে। তার কাছে এখন কতগুলি টেনিস বল আছে?

ধাপে ধাপে উত্তর: রজারের প্রথমে 5টি বল ছিল। 2টি ক্যানের প্রতিটিতে 3টে টেনিস বল মানে 6টি টেনিস বল।  $5 + 6 = 11$ । উত্তর হল 11।

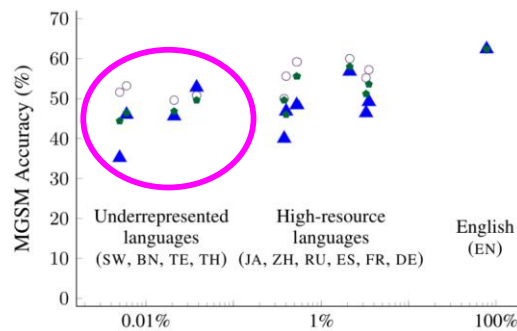
প্রশ্ন: জেনেটের হাঁসগুলি প্রতিদিন 16টি ডিম পাড়ে। তিনি প্রতিদিন প্রাতরাশে তিনটি করে ডিম খান এবং বন্ধুদের জন্য প্রতিদিন চারটি ডিম দিয়ে মাফিন তৈরি করেন। অবশিষ্ট হাঁসের ডিমগুলি তিনি প্রতিদিন কৃষকদের বাজারে প্রতি ডিম \$2 দরে বিক্রয় করেন। তিনি কৃষকদের বাজারে প্রতিদিন কত ডলার উপার্জন করেন?

## Model Output

ধাপে ধাপে উত্তর: প্রতিদিন 16টি ডিম পাড়ে। প্রতিদিন তিনটি ডিম খান এবং চারটি ডিম দিয়ে মাফিন তৈরি করেন। তাই প্রতিদিন  $16 - 3 - 4 = 9$ টি ডিম অবশিষ্ট থাকে। প্রতিটি ডিমের মূল্য \$2 হলে প্রতিদিন  $9 * 2 = 18$  ডলার উপার্জন করেন। উত্তর হল 18। ✓

[Language models are multilingual chain-of-thought reasoners \(2022\).](#)

- Translate to English with Google Translate and solve with English intermediate steps
- ▲ Intermediate reasoning steps in the language of the question
- Intermediate reasoning steps in English



Frequency of language in pre-training dataset (token percentage)

Underrepresented languages did surprisingly well, demonstrating the compositionality of the model

(model is neither multilingual nor trained to do reasoning)

# Chain-of-thought analysis

## **Benefits**

Expands the range of abilities for language models  
Multi-step reasoning can now be solved!

Works for any text (and image?) task  
Every task has a chain-of-thought.

No fine-tuning needed.  
Single model, many tasks

Some interpretability (can read chain-of-thought)  
Though it's not necessarily how the model reasons

## **Drawbacks**

Requires a large language model  
Emergent ability

Higher inference cost than directly answering  
CoT can be hundreds of tokens

Requires manually writing chains-of-thought in the prompts via exemplars  
(Some zero-shot that works for common multi-step reasoning problems)



More on CoT

# CoT is not enough

- Error propagation: one incorrect step leads to cumulative errors
- Chain structure limitation: the scope of exploration is limited
- Uncertainty: greedy decoding may not lead to a great reasoning path

Q: Calculate  $(2+3)*5$

A:

Calculate  $2+3$ , we get 6

$$6*5 = 30$$

The final answer is 30

Cumulative error

Q: Can 1, 2, 3, 4 get 24 in game 24?

A:

$$1+2 = 3$$

$$3*3 = 9$$

$$9+4 = 13$$

$$13 \neq 24$$

So 1,2,3,4 cannot get 24 in game 24.

Limited exploration

Q: What is  $1+2+3+\dots+6$ ?

A:

$$1+2 = 3$$

$$3+3 = 6$$

$$6+4 = 10$$

$$10+5 = 15$$

$$15+6 = 21$$

So  $1+2+3+4+5+6=21$ .

Correct yet not good

# Improve CoT in different phases of reasoning

- **Pre-process of the reasoning task:**
  - Decomposition: e.g. Least-to-most prompting
- **Improvement in the reasoning phase:**
  - Tool using: e.g. PoT
  - Planning: e.g. ToT
- **Utilization of the reasoning result:**
  - Major voting: e.g. Self-consistency
  - Verify: e.g. Verifier
  - Refine: e.g. Self-refine

# Least-to-most prompting

## Explicitly decompose into subquestions

### Stage 1: Decompose Question into Subquestions

Q: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

Language Model

A: To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

### Stage 2: Sequentially Solve Subquestions

Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

Language Model

A: It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Append model answer to Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Language Model

A: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide  $15 \div 5 = 3$  times before it closes.

Subquestion 2

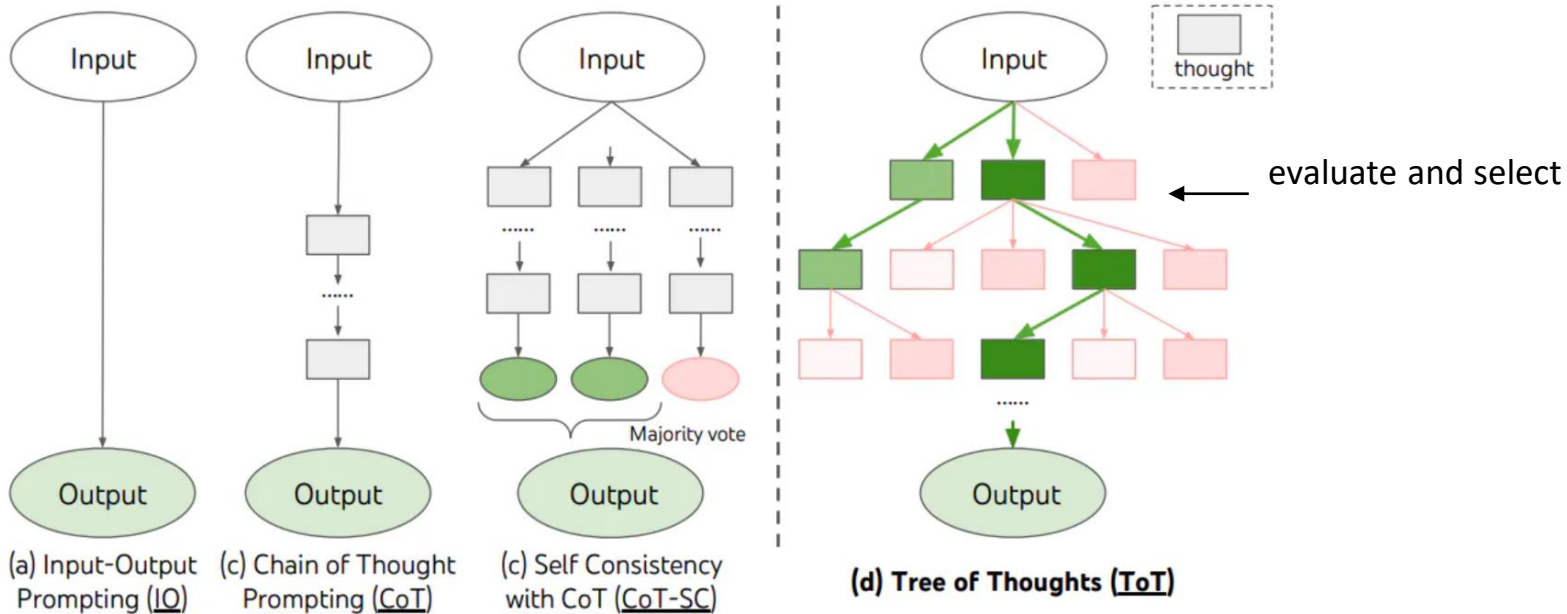
Q: How many times can she slide before it closes?

Figure 1: Least-to-most prompting solving a math word problem in two stages: (1) query the language model to decompose the problem into subproblems; (2) query the language model to sequentially solve the subproblems. The answer to the second subproblem is built on the answer to the first subproblem. The demonstration examples for each stage's prompt are omitted in this illustration.

# Tree of Thoughts (ToT)

Explore over units of text that serve as intermediate steps

The ToT framework is illustrated below:



# Program of Thoughts (PoT)

Output Python programs and call Python interpreter to calculate the answers

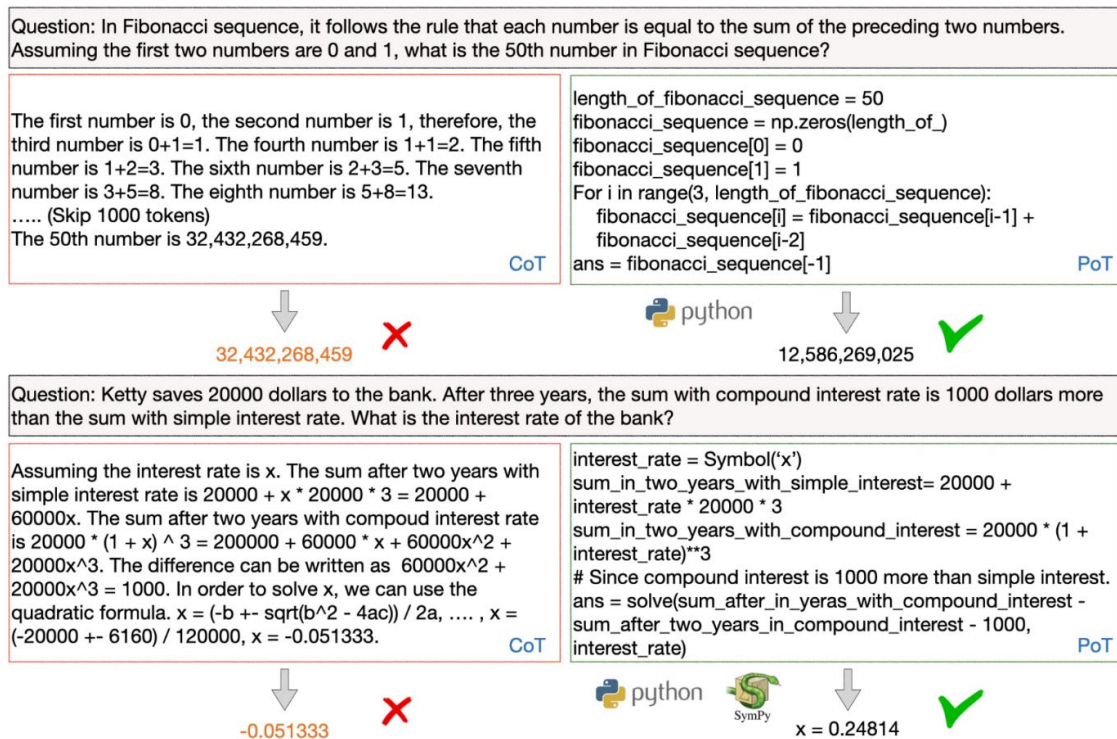


Figure 1: Comparison between Chain of Thoughts and Program of Thoughts.

# A trick for CoT: Self-consistency

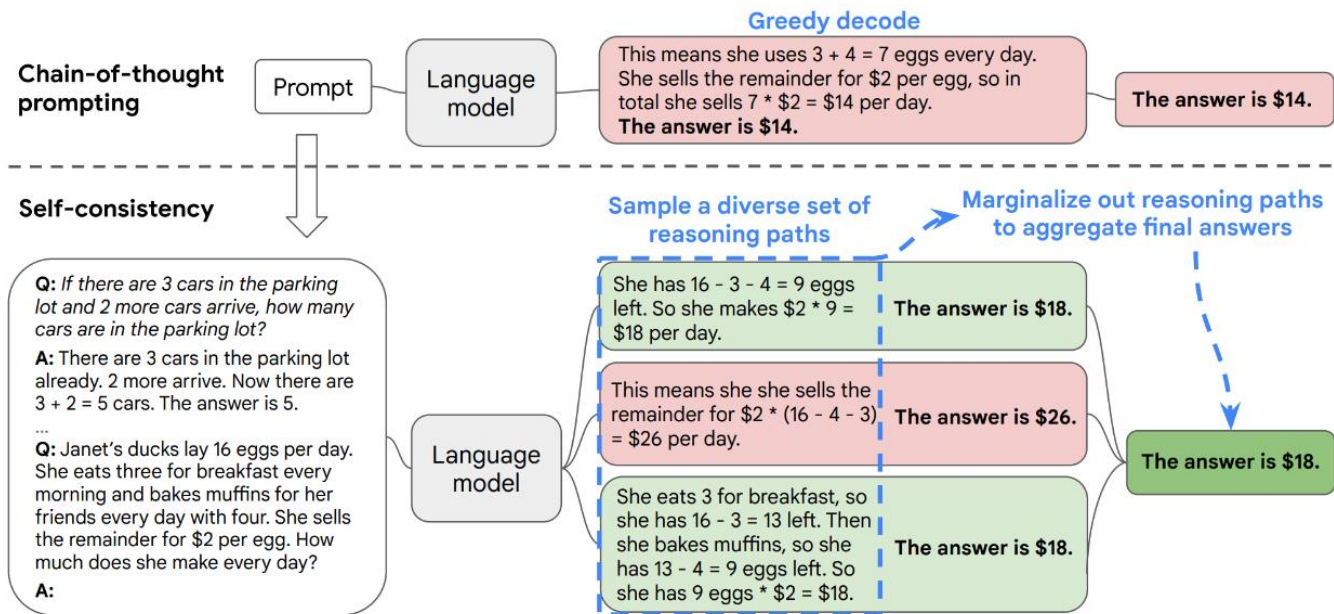
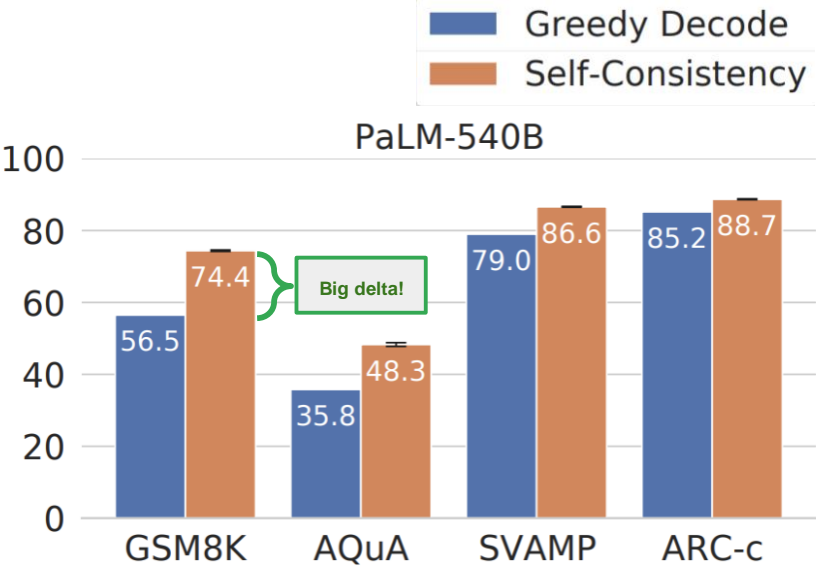


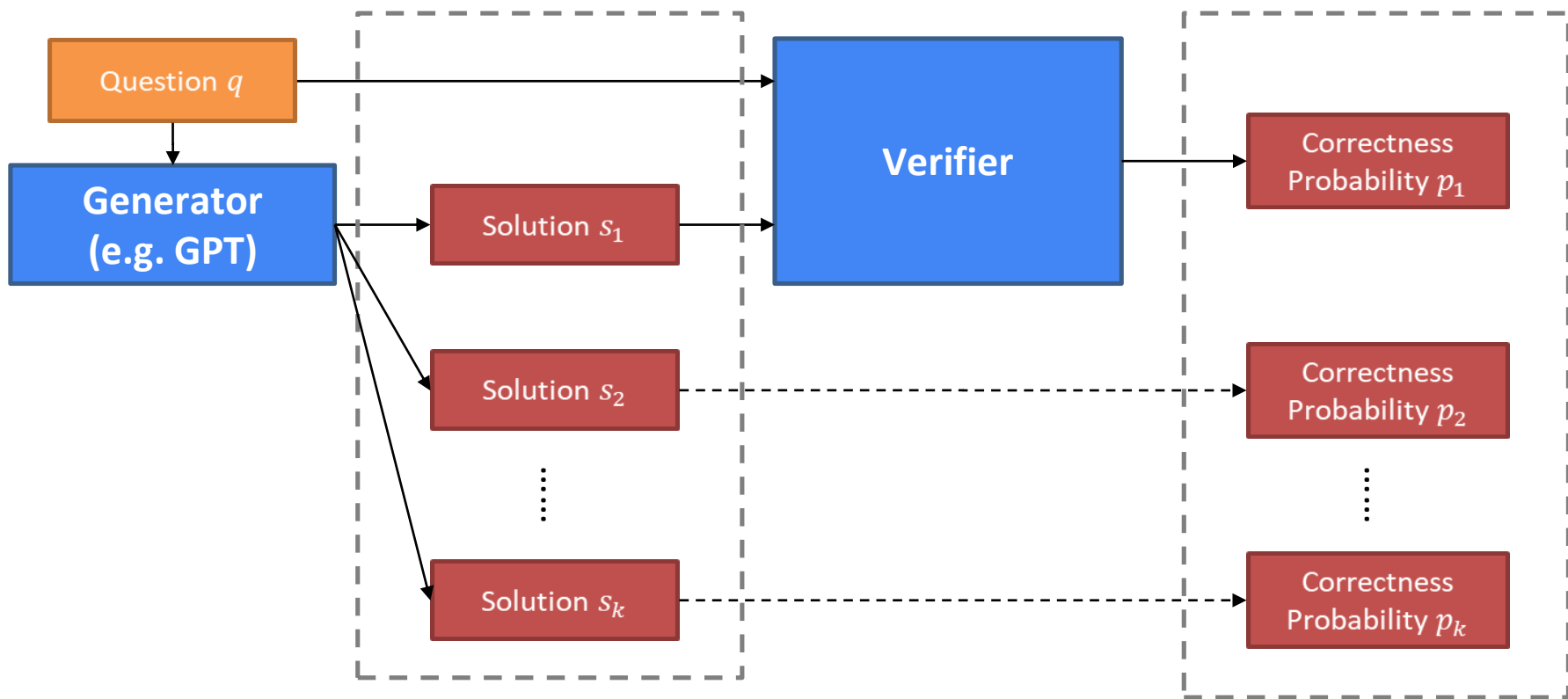
Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the “greedy decode” in CoT prompting by sampling from the language model’s decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

# Self-consistency works really well





# Verifier in COT



# Exceed GPT 4 using our OVM in GSM-8K

1	<b>GPT-4 Code Interpreter</b> (CSV, K=5)	97.0	×	Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification	🔗 🔄	2023	majority voting zero-shot
2	<b>GPT-4</b> (Model Selection, SC K=15)	96.8	×	Automatic Model Selection with Large Language Models for Reasoning	🔗 🔄	2023	majority voting
3	<b>GPT-4</b> (PHP, SC K=40)	96.5	×	Progressive-Hint Prompting Improves Reasoning in Large Language Models	🔗 🔄	2023	majority voting
4	<b>GPT-4</b> (Model Selection, SC K=5)	96.5	×	Automatic Model Selection with Large Language Models for Reasoning	🔗 🔄	2023	majority voting
5	<b>SFT-Mistral-7B</b> (Metamath, OVM, Smart Ensemble)	96.4	7	✓		2024	
6	<b>SFT-Mistral-7B</b> (AugData + ovm + ensemble)	95.9	7	✓		2024	
7	<b>GPT-4</b> (PHP)	95.5	×	Progressive-Hint Prompting Improves Reasoning in Large Language Models	🔗 🔄	2023	
8	<b>MindOpt Copilot Mistral-7B</b> (MetaMath, OVM, BS, Ensemble)	95.1	7	✓		2024	
9	<b>Claude 3 Opus</b> (0-shot chain-of-thought)	95	×	The Claude 3 Model Family: Opus, Sonnet, Haiku	🔄	2024	chain-of-thought zero-shot
10	<b>Gemini Ultra</b> (Maj11@32)	94.4	×	Gemini: A Family of Highly Capable Multimodal Models	🔄	2023	majority voting
11	<b>SFT-Mistral-7B</b> (Metamath + ovm + ensemble)	94.13	7	✓		2024	
12	<b>GPT-4</b> (Ask, Refine, Trust)	94.08	×	The ART of LLM Refinement: Ask, Refine, and Trust	🔄	2023	
13	<b>Shepherd + DeepSeek-67B</b> (SFT on MetaMATH + PRM rerank, k=256)	93.3	67	✓	🔗 🔄	2023	rerank
14	<b>MindOpt Copilot Mistral-7B</b> (MetaMath, OVM, Ensemble)	93.2	7	✓		2024	

<https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k>

# Self-refine

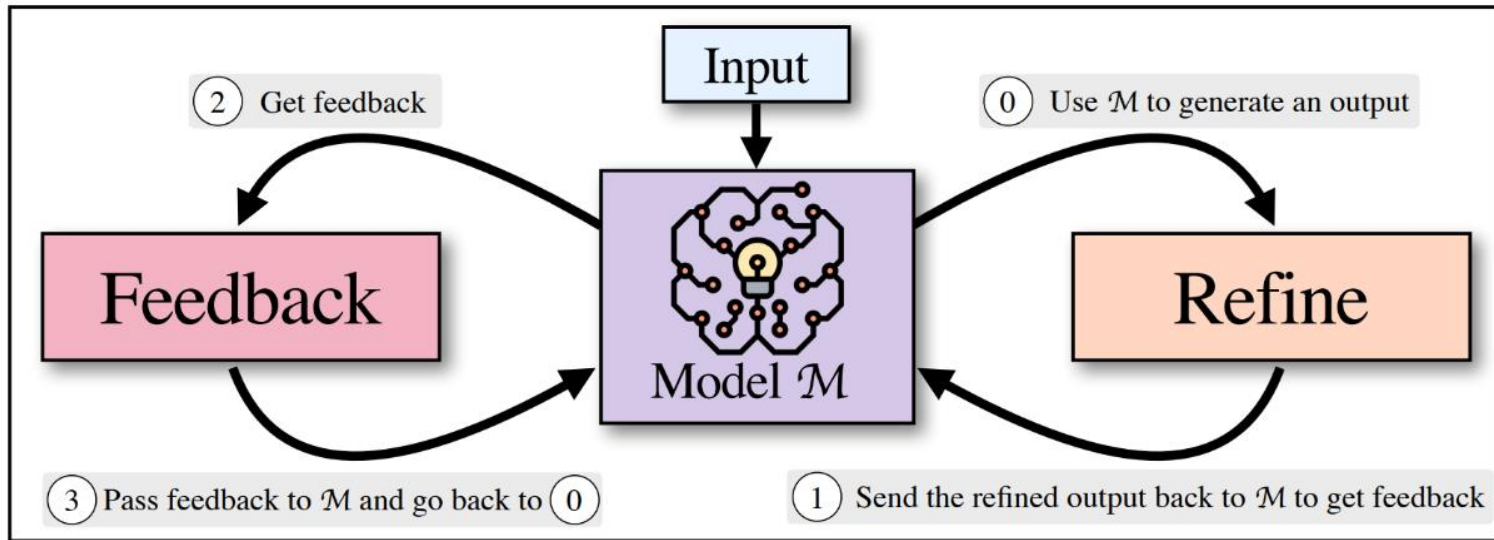


Figure 1: SELF-REFINE starts by taking an initially generated output (④), and passing it back to the same model  $\mathcal{M}$  (①) to get feedback (②); feedback on the initial output is passed back to the model (③), to iteratively refine (④) the previously generated output. SELF-REFINE is instantiated with a powerful language model such as GPT-3.5 and does not involve human assistance.

# Should we employ all the techniques above?

- Usually CoT can perform well under many situations
- **Accuracy vs Cost:**
  - Additional techniques need more computational sources (self-consistency) or additional data processing(PoT) although these techniques can usually improve the performance.
  - the trade off depends on the real application.

# Acknowledgement

- Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/>
- Scaling, emergence, and reasoning (Jason Wei, NYU): [https://docs.google.com/presentation/d/1EUV7W7X\\_w0BDrscDhPg7IMGzJCkeaPkGCJ3bN8dluXc/edit?resourcekey=0-7Nz5A7y8JozyVrnDtcEKJA#slide=id.g16197112905\\_0\\_0](https://docs.google.com/presentation/d/1EUV7W7X_w0BDrscDhPg7IMGzJCkeaPkGCJ3bN8dluXc/edit?resourcekey=0-7Nz5A7y8JozyVrnDtcEKJA#slide=id.g16197112905_0_0)
- Prompting engineering lectures(DAIR-AI): <https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/lecture/Prompt-Engineering-Lecture-Elvis.pdf>
- Prompt engineering guide: <https://www.promptingguide.ai/>

# Optional reading material

## In-context learning:

- An Explanation of In-context Learning as Implicit Bayesian Inference(<https://arxiv.org/abs/2111.02080>)
- Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?(<https://arxiv.org/abs/2202.12837>)

## Knowledge probing:

- How Much Knowledge Can You Pack Into the Parameters of a Language Model?(<https://arxiv.org/abs/2002.08910>)

## Knowledge editing

- Fast model editing at scale(<https://arxiv.org/abs/2110.11309>)