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

CSC5051/MDS5110: Natural Language Processing

Lecture 7: Prompt Engineering: Knowledge and Reasoning

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.
- Dataset: MMLU

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).
- Dataset: ARC (AI2 Reasoning Challenge)

Fei Yu, Hongbo Zhang, Prayag Tiwari, Benyou Wang. Natural Language Reasoning, A Survey. ACM Computing Surveys.
https://dl.acm.org/doi/10.1145/3664194

An example for **knowledge** (MMLU)

During the period when life is believed to have begun, the atmosphere on primitive Earth contained abundant amounts of all the following gases EXCEPT	high_school_biology	["oxygen", "hydrogen", "ammonia", "methane"]	0A

https://huggingface.co/datasets/cais/mmlu/viewer/high_school_biology?views%5B%5D=high_school_biology_test&row=3

An example for **reasoning** (ARC)

Wich of the following statements best explains why magnets usually stick to a refrigerator door? (磁铁为什么吸附在冰箱上)	{ "text": ["The refrigerator door is smooth.", "The refrigerator door contains iron.", "The refrigerator door is a good conductor.", "The refrigerator door has electric wires in it."], "label": ["A", "B", "C", "D"] }	B

https://huggingface.co/datasets/allenai/ai2_arc/viewer/ARC-Challenge/train?views%5B%5D=arc_challenge_train&row=1

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

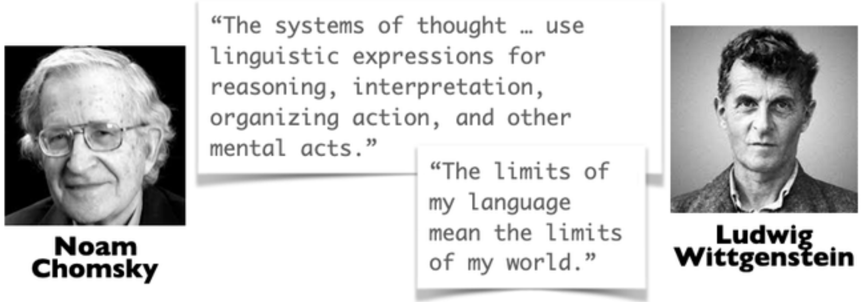
Knowledge and reasoning are interdependent; effective reasoning relies on a solid foundation of knowledge, while acquiring new knowledge often requires reasoning skills to evaluate, integrate, and understand information.

https://www.quora.com/What-is-the-difference-between-knowledge-and-reasoning

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.

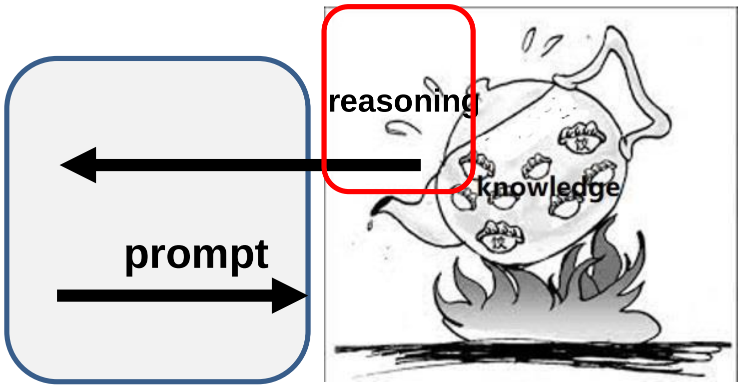


[Noam Chomsky - The Function of Language](https://youtu.be/TzzuPMA8s7k?t=324) https://youtu.be/TzzuPMA8s7k?t=324
Evelina Fedorenko, Steven T. Piantadosi, Edward A. F. Gibson. Language is primarily a tool for communication rather than thought. **Nature**. https://www.nature.com/articles/s41586-024-07522-w

Knowledge vs. Reasoning in **LLMs**

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

Inner though



Communications

[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

How to record something in human history?

Language is a efficient (and accurate) way to exchange information; it accelerates intelligence evolution.

Why DO I believe Generalist (Multi-modal) AI is language –centric?

Language stores massive **knowledge**

"The sun rises in the east"

Language is a great tool for **communication**

The interface between Humans and Machines

Language can **reason and plan**

Language is the Medium of thought (思考的介质)

More modality out

Longer context
Better reasoning

More modality in

What is a knowledge base?

Knowledge Representation in Computer Science

Level	Form	Representative Examples	Advantages	Limitations
1	**Tabular (Structured)**	SQL, Excel	Clear structure, efficient querying	Weak expressiveness, poor at complex relations
2	**Semi-structured**	JSON, XML	Flexible format, hierarchical data	Lacks explicit semantics
3	**Symbolic (Knowledge Graph)**	RDF, OWL, Neo4j	Logical reasoning, rich relationships	Hard to construct and maintain
4	**Textual**	Corpora, Documents	Natural and abundant information	Redundant, ambiguous, hard to formalize
5	**Statistical (Embedding-based)**	Word2Vec, BERT, TransE	Good generalization, scalable computation	Low interpretability, implicit logic
6	**Neural (Parametric Knowledge)**	LLM parameters	Powerful representation, emergent intelligence	Implicit knowledge, hard to edit or update
7	**Causal (Structured Equations / Graphs)**	Bayesian Networks, Causal DAGs	Causality and interpretability	Difficult to learn and estimate
8	**Hybrid (Neuro-symbolic / Neuro-causal)**	RAG, K-BERT, ERNIE, Neuro-Causal Models	Combines symbolic precision with neural flexibility	Implementation complexity

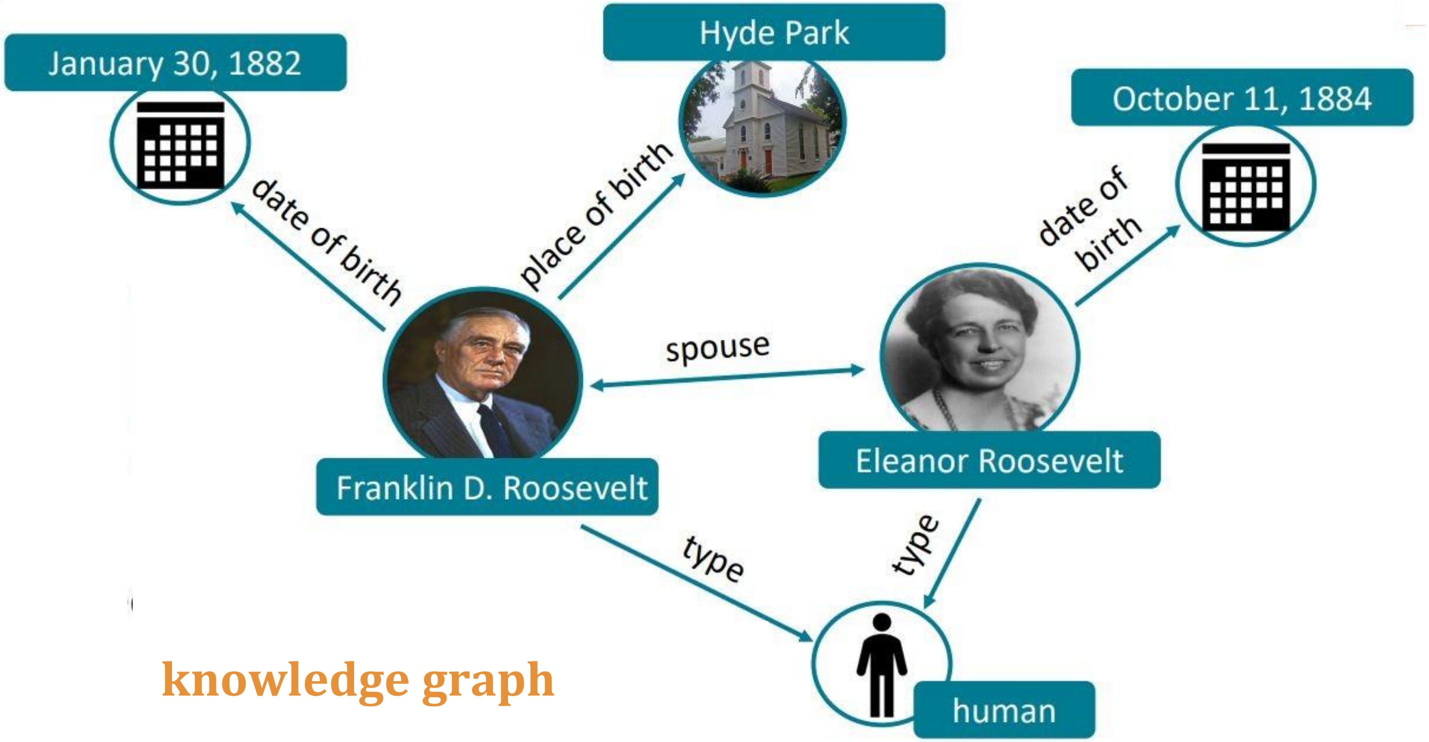
Graphic from Megan Leszczynski (Stanford)

What is a knowledge base?

knowledge base

WIKIDATA

Example



knowledge graph

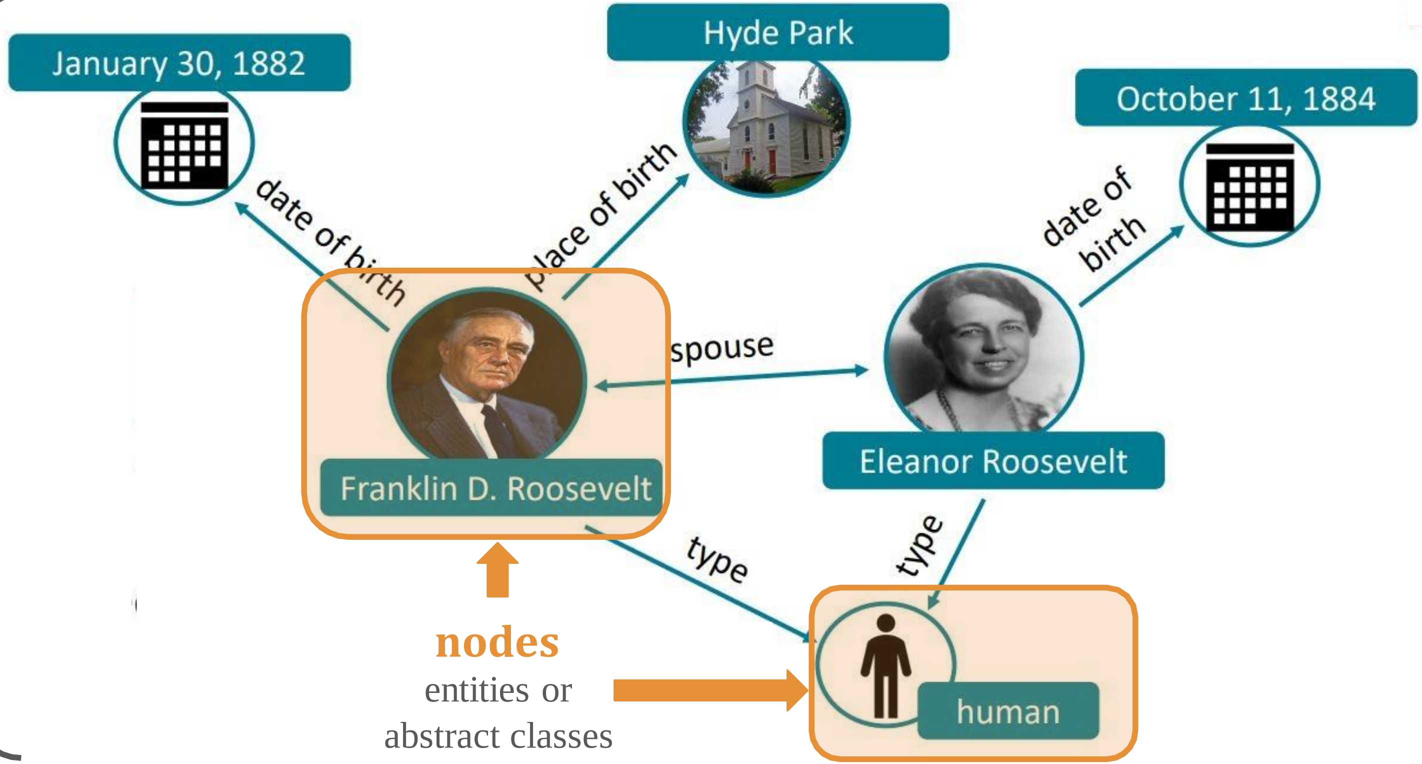
Graphic from [Megan Leszczynski (Stanford)]

What is a knowledge base?

knowledge base

WIKIDATA

[Example]

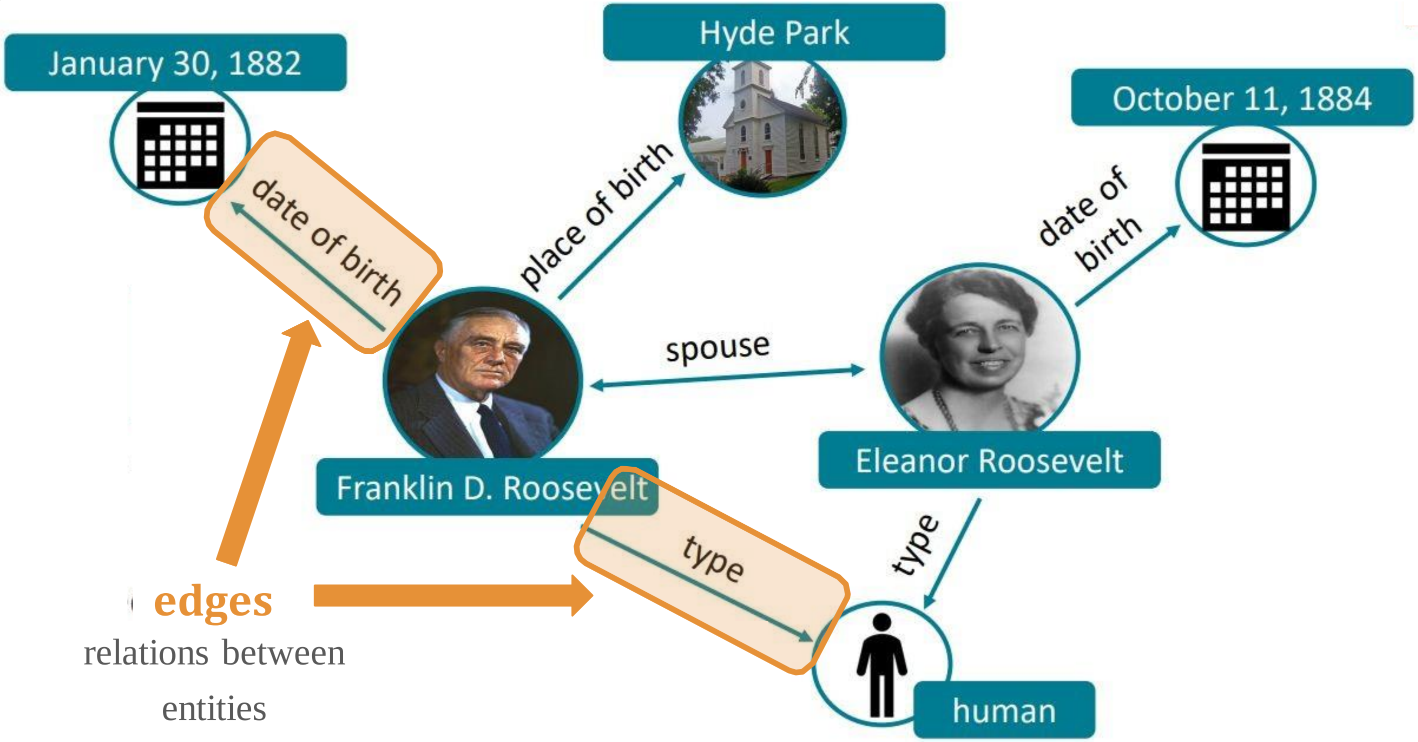


What is a knowledge base?

knowledge base

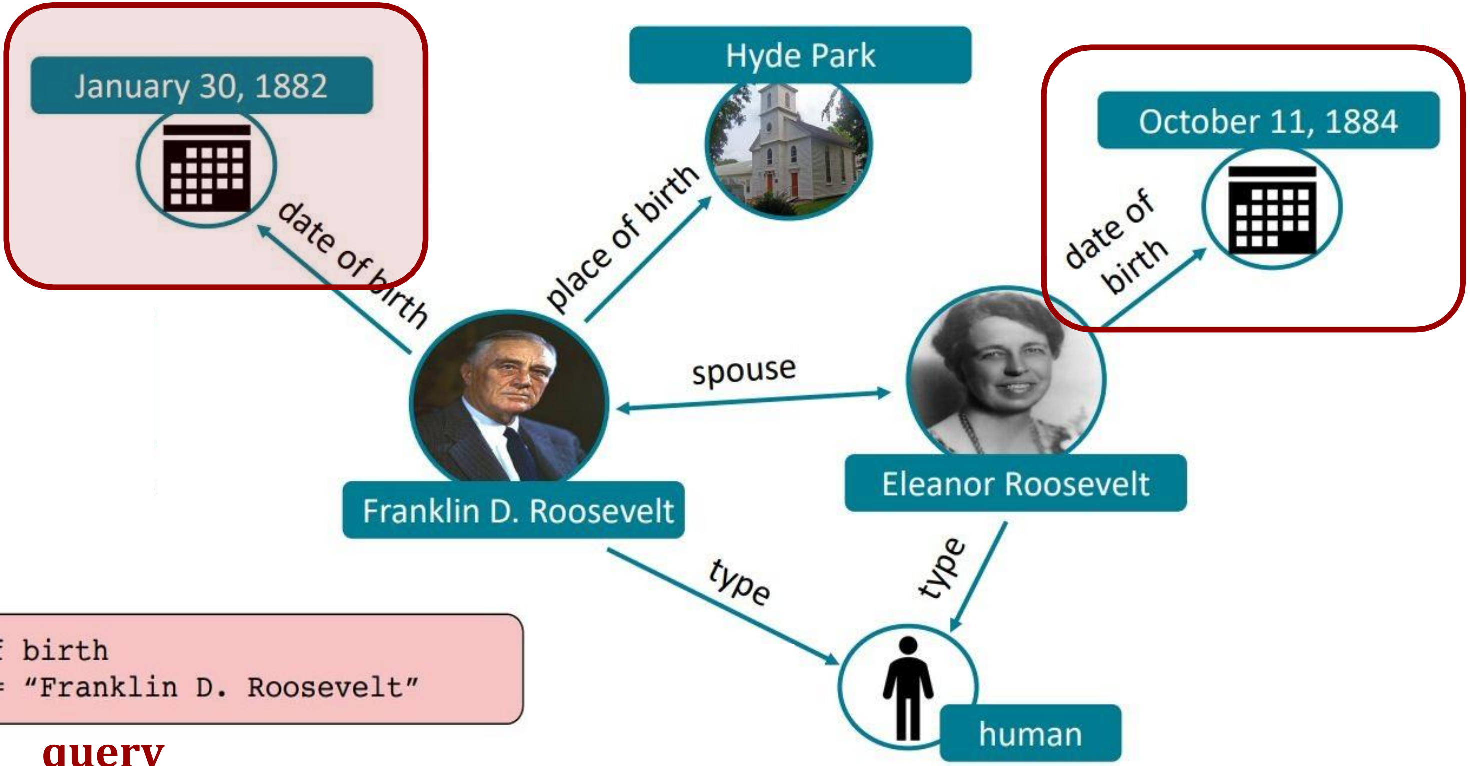
WIKIDATA

[Example]()



Graphic from [Megan Leszczynski (Stanford)]()

How to query?

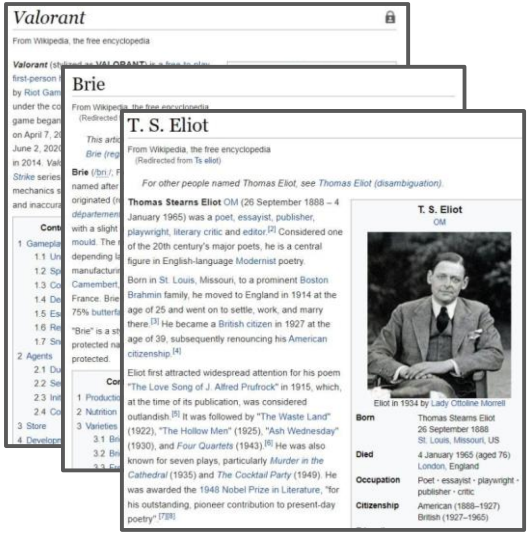



```
SELECT date of birth
WHERE person = "Franklin D. Roosevelt"
```

query

Right-most graphic from [Petroni et al., 2019]

How were knowledge bases formed?

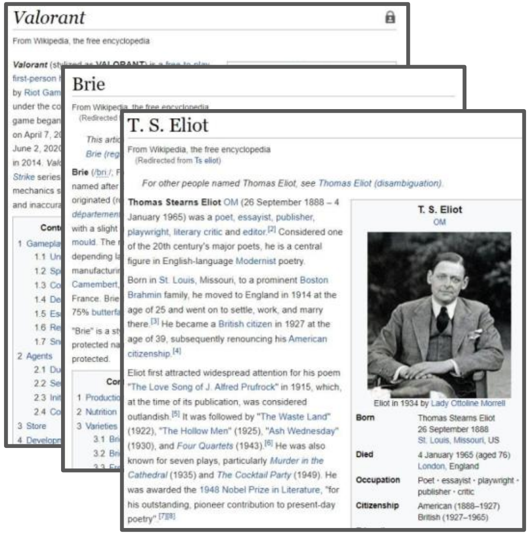


unstructured text

Structured knowledge base

Right-most graphic from [Petroni et al., 2019]

How were knowledge bases formed?


Valorant

Brie

T. S. Eliot

From Wikipedia, the free encyclopedia

(Redirected from Ts eliot)

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

Thomas Stearns Eliot OM (26 September 1888 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.[2] 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.[3] He became a British citizen in 1927 at the age of 39, subsequently renouncing his American citizenship.[4]

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.[5] It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).[6] 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".[7][8]

T. S. Eliot

OM

Eliot in 1934 by Lady Ottoline Morrell

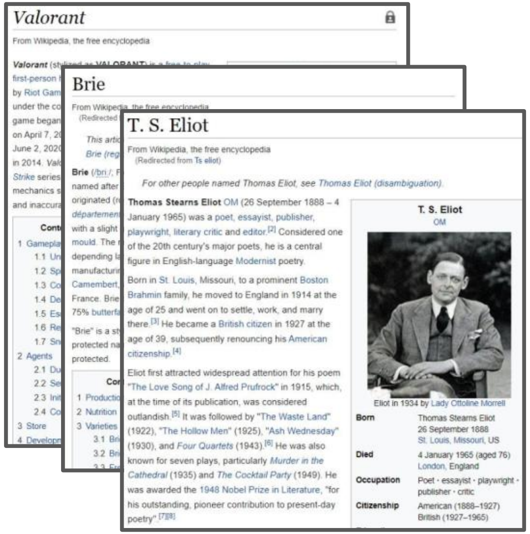
Born	Thomas Stearns Eliot 26 September 1888 St. Louis, Missouri, US
Died	4 January 1965 (aged 76) London, England
Occupation	Poet · essayist · playwright · publisher · critic
Citizenship	American (1888–1927) British (1927–1965)

Knowledge Extraction Pipeline

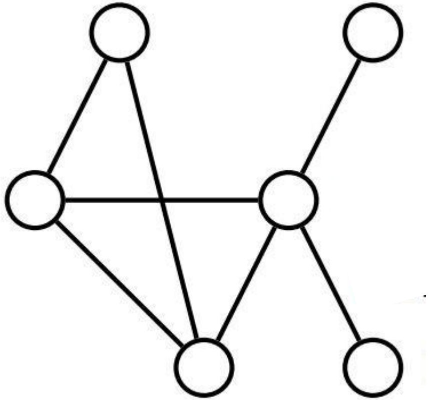
unstructured text

Structured knowledge base

How were knowledge bases formed?



Knowledge Extraction Pipeline



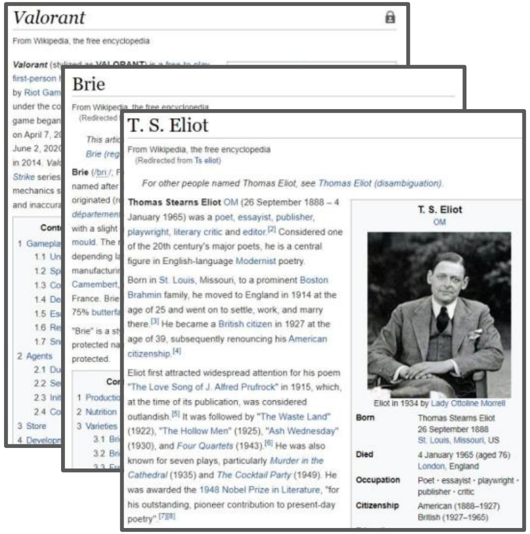
unstructured text

Structured knowledge base

Right-most graphic from [Petroni et al., 2019]

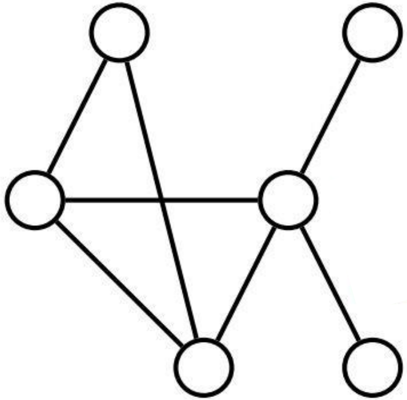
Downsides of using knowledge bases

Downsides of using knowledge bases



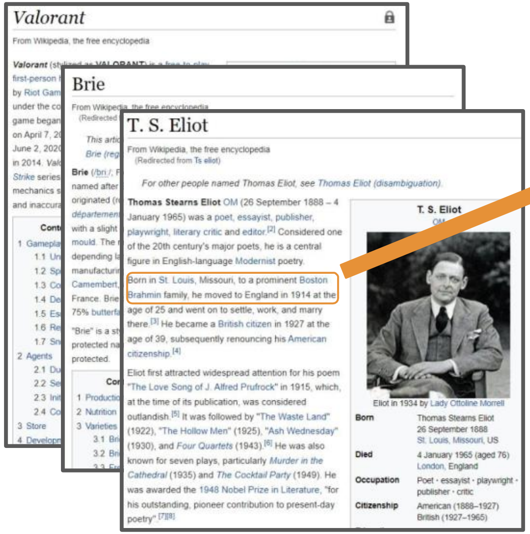
Knowledge Extraction Pipeline

- **Data preprocessing**
- **Data merging**
- **Entity/relation extraction**
- **Ontology extraction**
- ...



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

Downsides of using knowledge bases



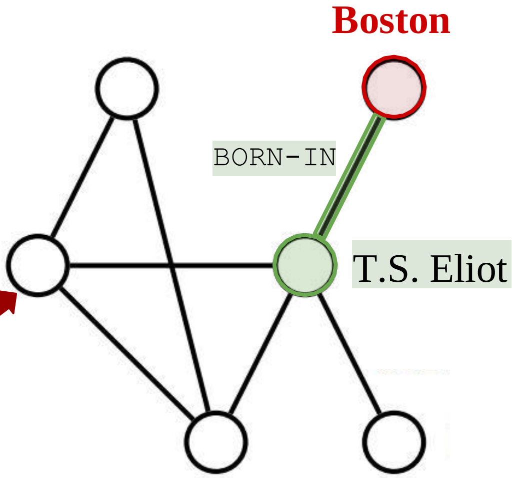
unstructured text

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

Knowledge Extraction Pipeline

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

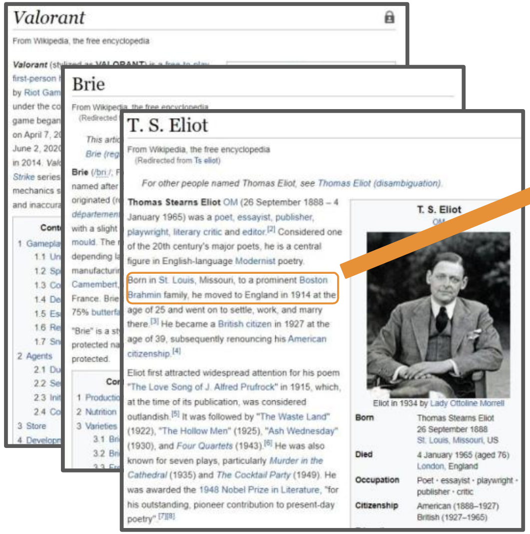
incorrect extraction



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

Right-most graphic from [Petroni et al., 2019]

Downsides of using knowledge bases


Valorant

Brie

T. S. Eliot

From Wikipedia, the free encyclopedia

(Redirected from Ts eliot)

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

Thomas Stearns Eliot OM (26 September 1888 – 4 January 1965) was a poet, essayist, publisher, playwright, literary critic and editor.[2] 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.[3] He became a British citizen in 1927 at the age of 39, subsequently renouncing his American citizenship.[4]

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.[5] It was followed by "The Waste Land" (1922), "The Hollow Men" (1925), "Ash Wednesday" (1930), and *Four Quartets* (1943).[6] 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".[7][8]

T. S. Eliot

Eliot in 1934 by Lady Ottoline Morrell

Born	Thomas Stearns Eliot 26 September 1888 St. Louis, Missouri, US
Died	4 January 1965 (aged 76) London, England
Occupation	Poet · essayist · playwright · publisher · critic
Citizenship	American (1888–1927) British (1927–1965)

unstructured text

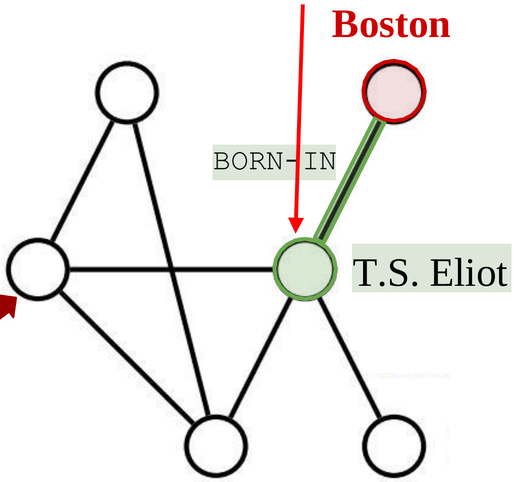
"Born in St. Louis, Missouri, to a prominent Boston Brahmin family…"

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)

Image from www.slyderstavern.com

Language models as knowledge bases?

Why language models?

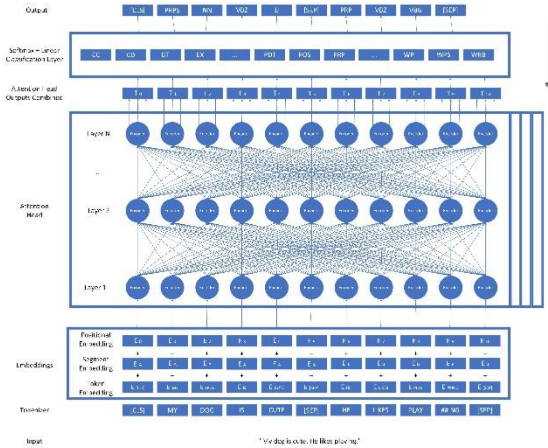
- Scalability: pre-trained on a huge corpus of data
- Time/Labor efficiency: does not require annotations/supervision
- Flexibility: more flexible with natural language queries
- Accessibility: can be used off-the-shelf

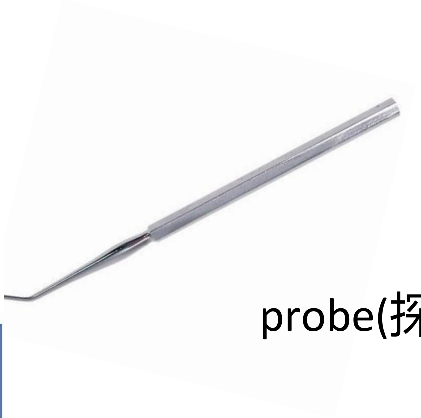
Image from www.slyderstavern.com

Do language models really store knowledge？

LAMA probe

BERT





probe(探针，探测)

Image courtesy of Patrick Lewis

LAMA Probe

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

LAnguage **M**odel **A**nalysis

Image courtesy of Patrick Lewis

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

LAnguage **M**odel **A**nalysis

Image courtesy of Patrick Lewis

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?*

LAnguage **M**odel **A**nalysis

Image courtesy of Patrick Lewis

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

LAnguage **M**odel **A**nalysis

Evaluation of LM via LAMA

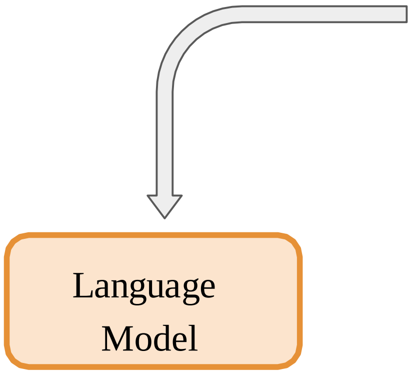
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower

*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

"The color of the sky is [MASK]."

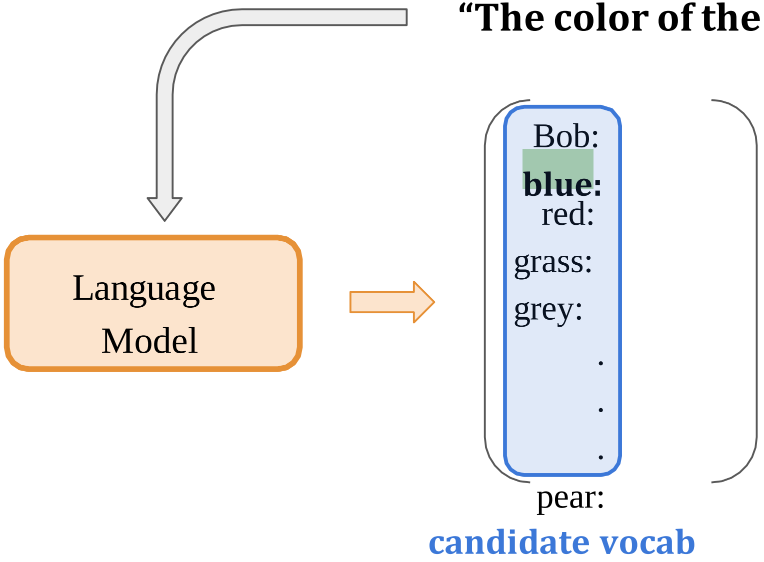


*according to the LM

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

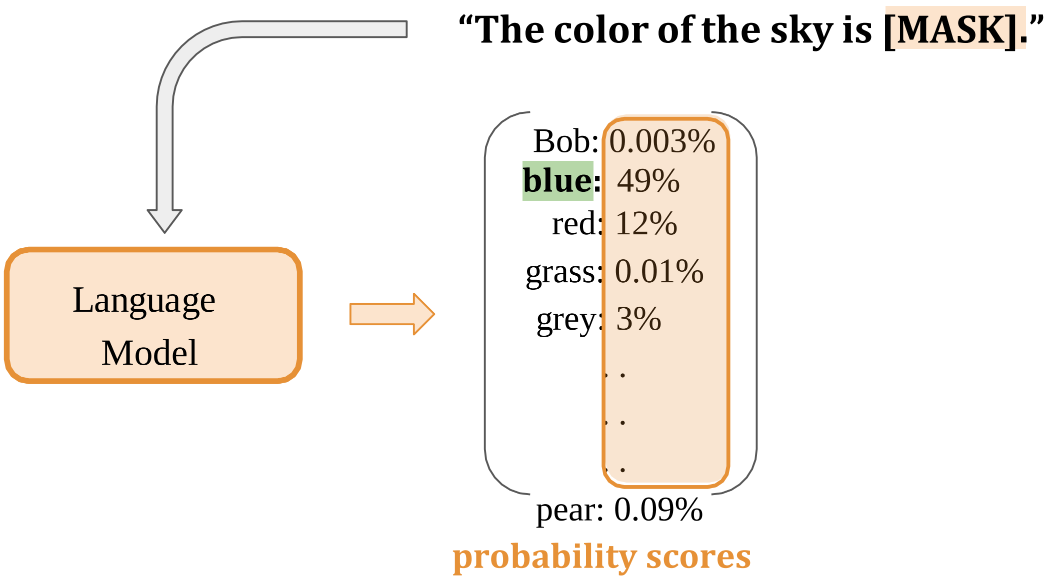
"The color of the sky is [MASK]."



*according to the LM

Evaluation of LM via LAMA

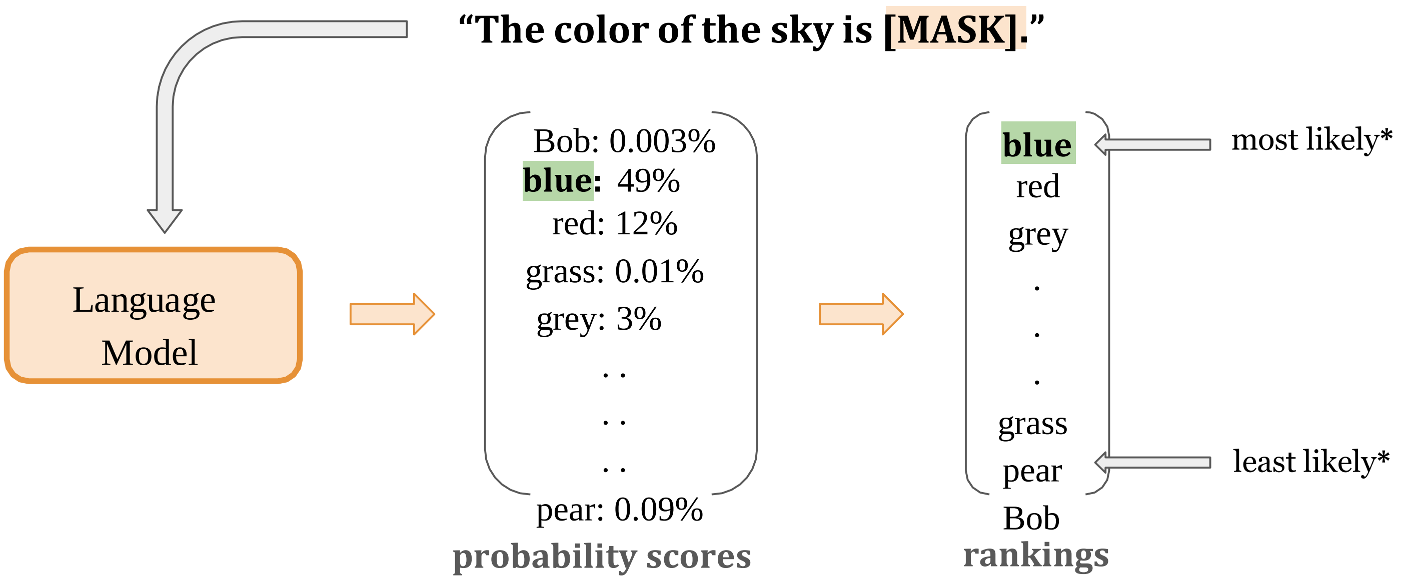
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



*according to the LM

Evaluation of LM via LAMA

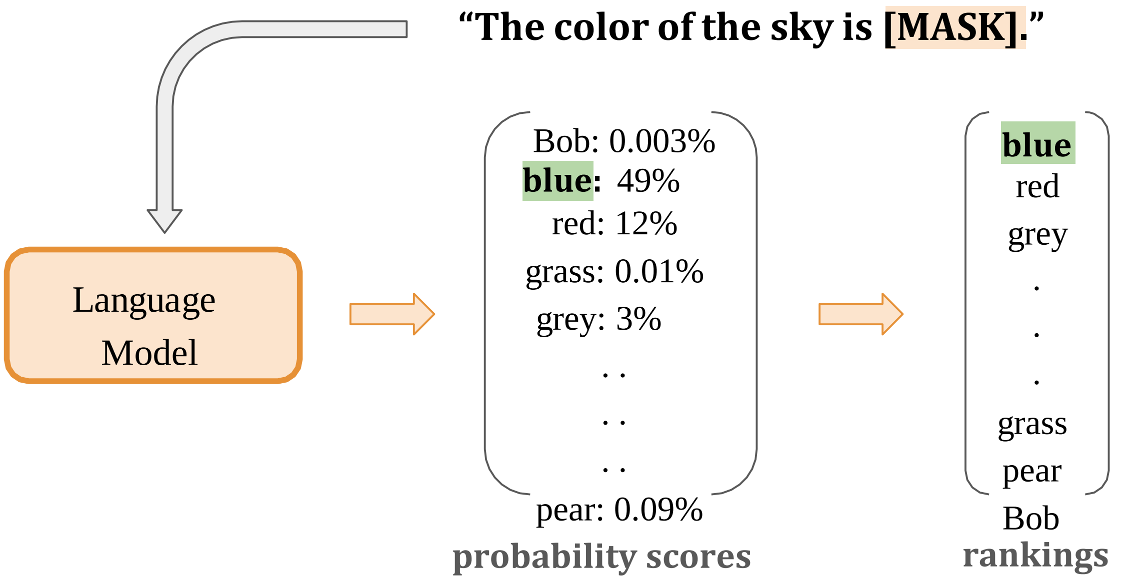
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



*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



P@k: precision at k

"Does ground truth exist in the top k ranks?"

Evaluation of LM via LAMA

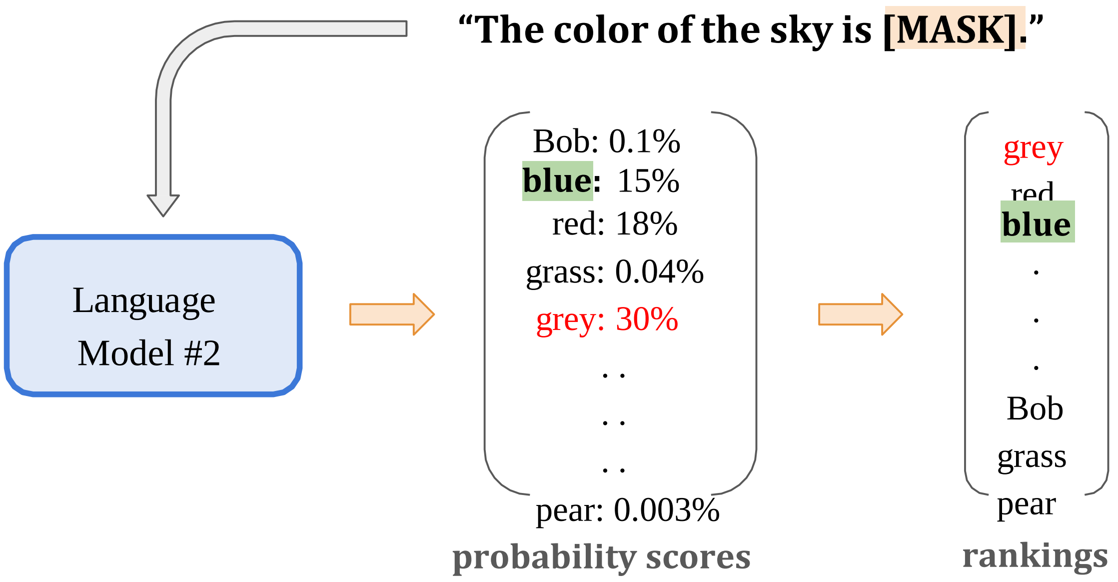
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower

“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

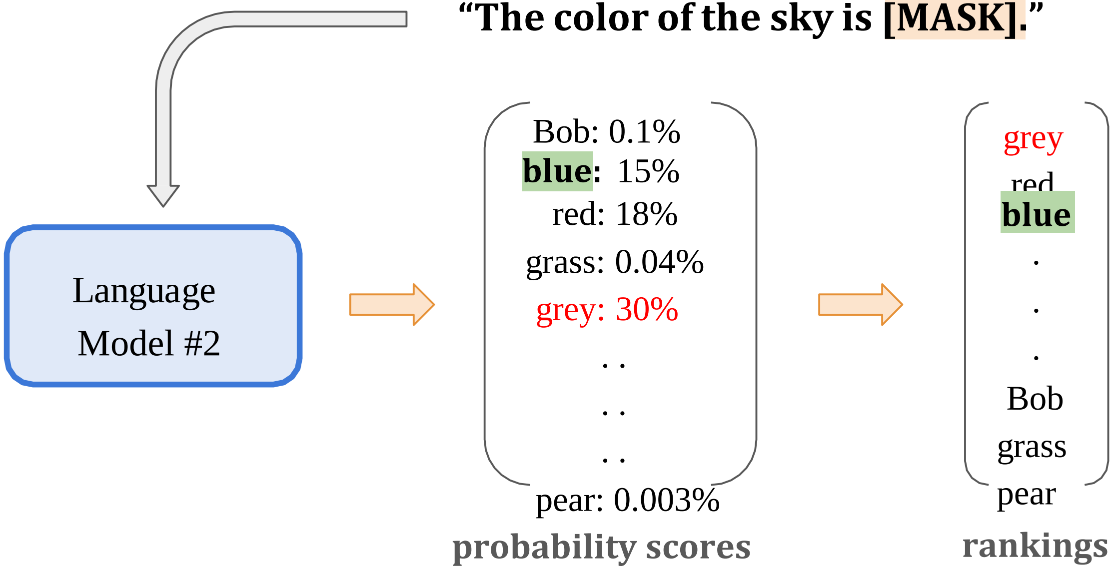


Evaluation of LM via LAMA

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

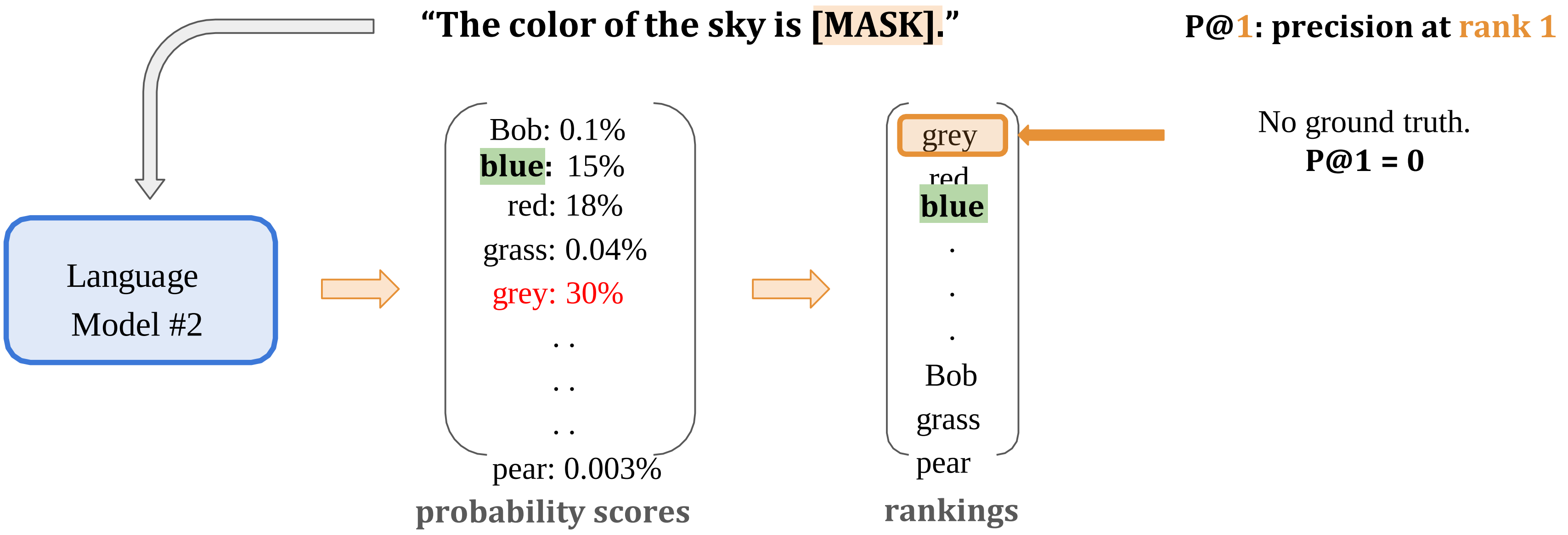
"The color of the sky is [MASK]."

P@1: precision at rank 1



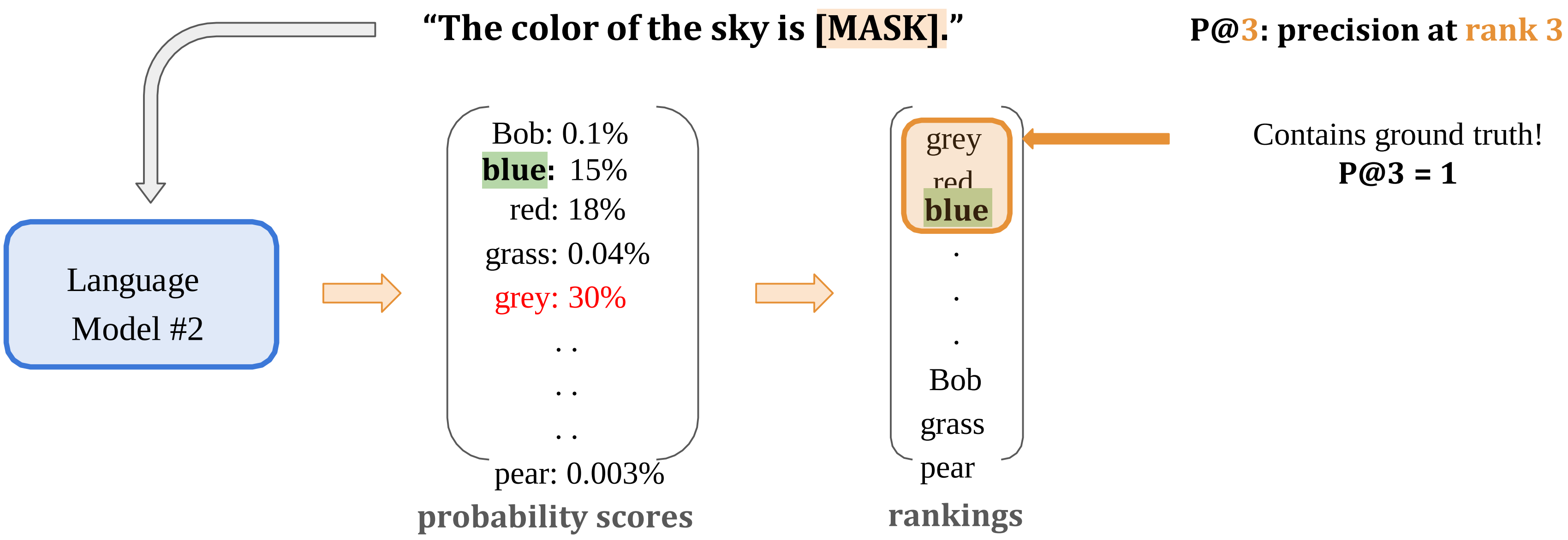
Evaluation of LM via LAMA

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

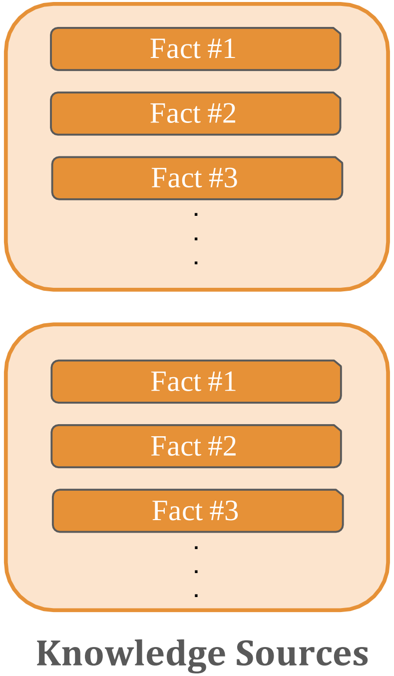
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



Architecture of the LAMA probe

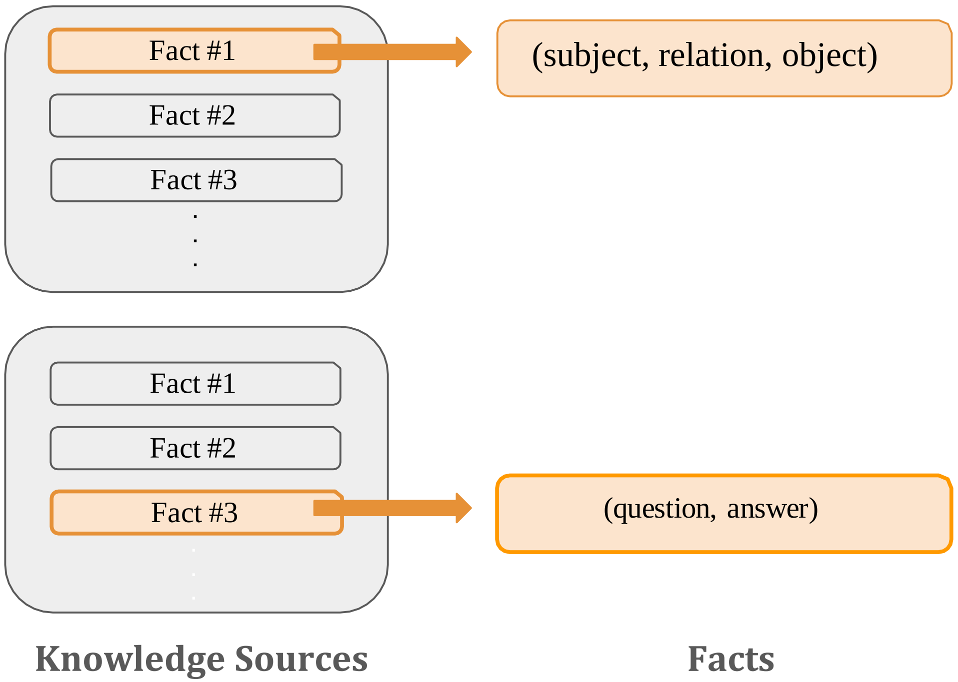
Architecture of the LAMA probe

Step 1: Compile knowledge sources



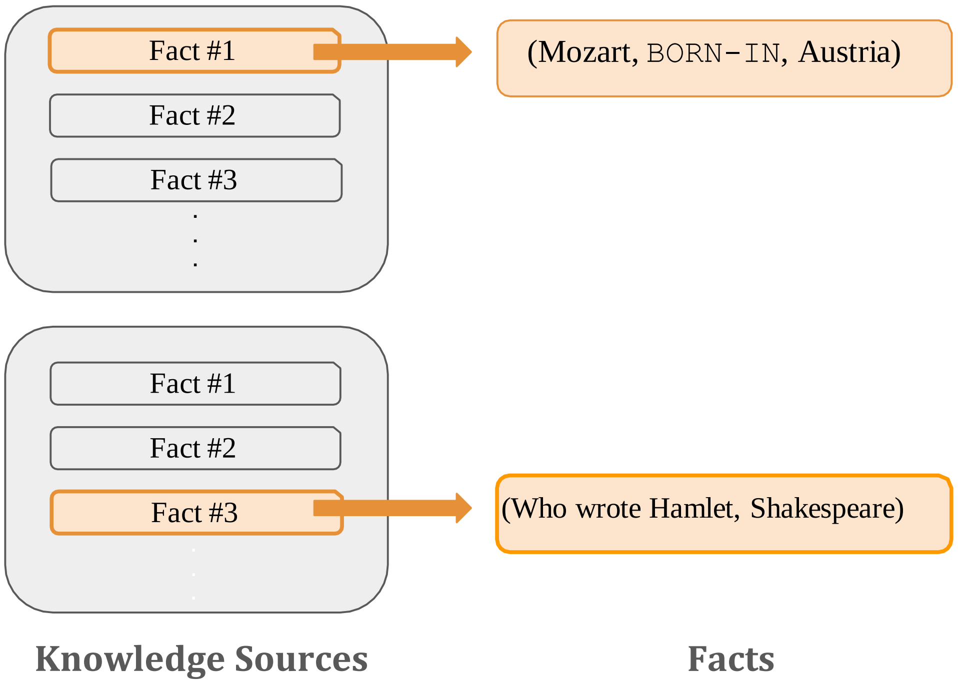
Architecture of the LAMA probe

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



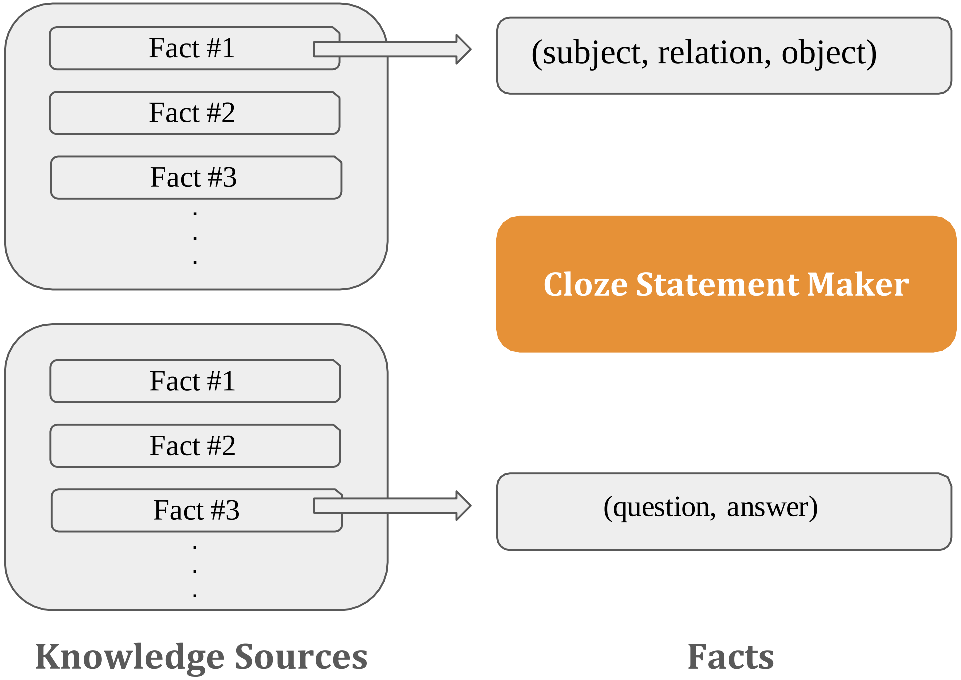
Architecture of the LAMA probe

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



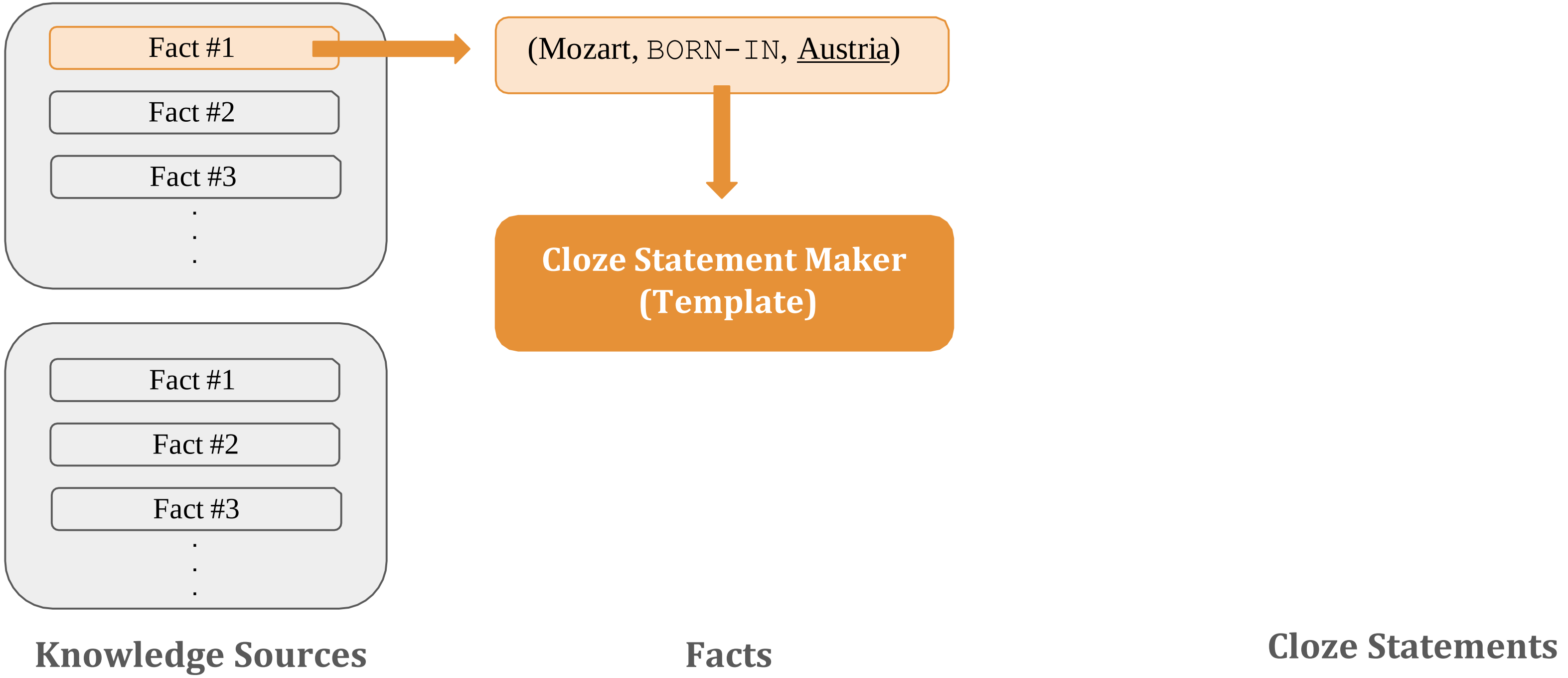
Architecture of the LAMA probe

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



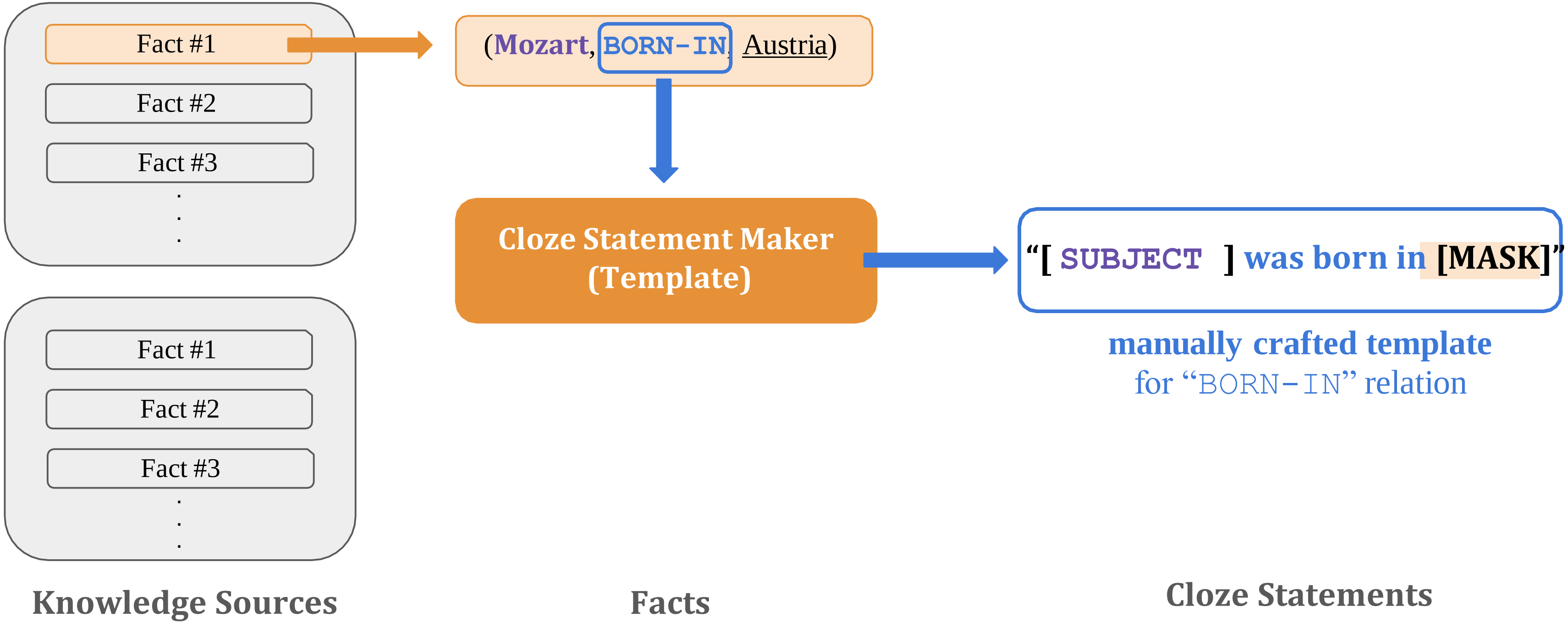
Architecture of the LAMA probe

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



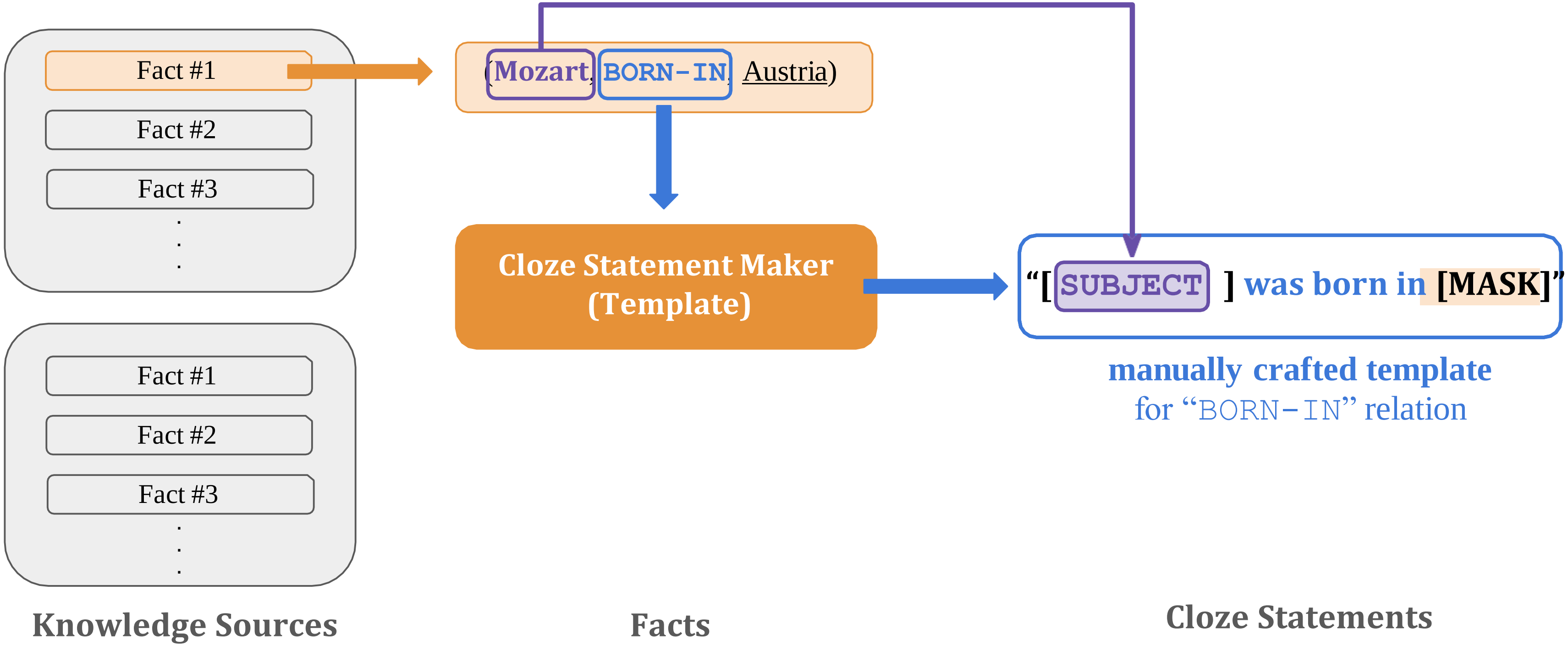
Architecture of the LAMA probe

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



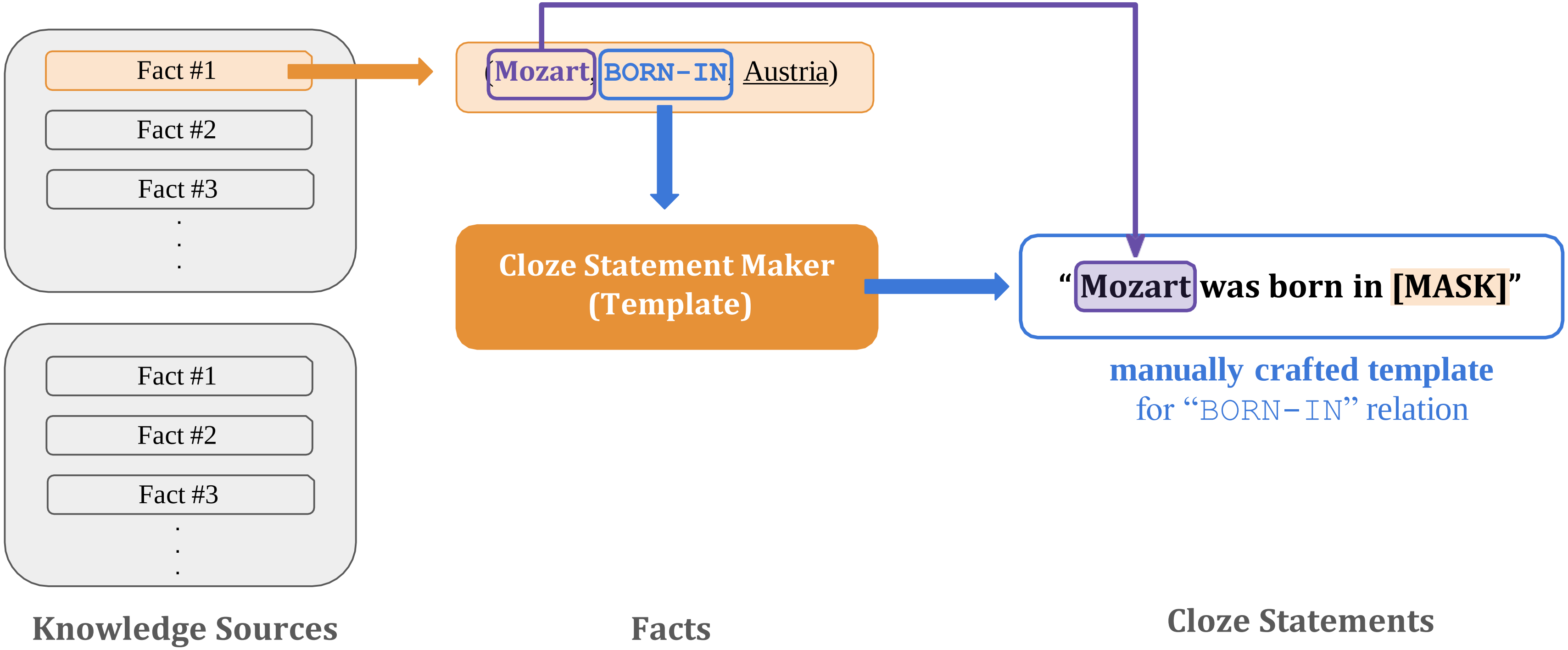
Architecture of the LAMA probe

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



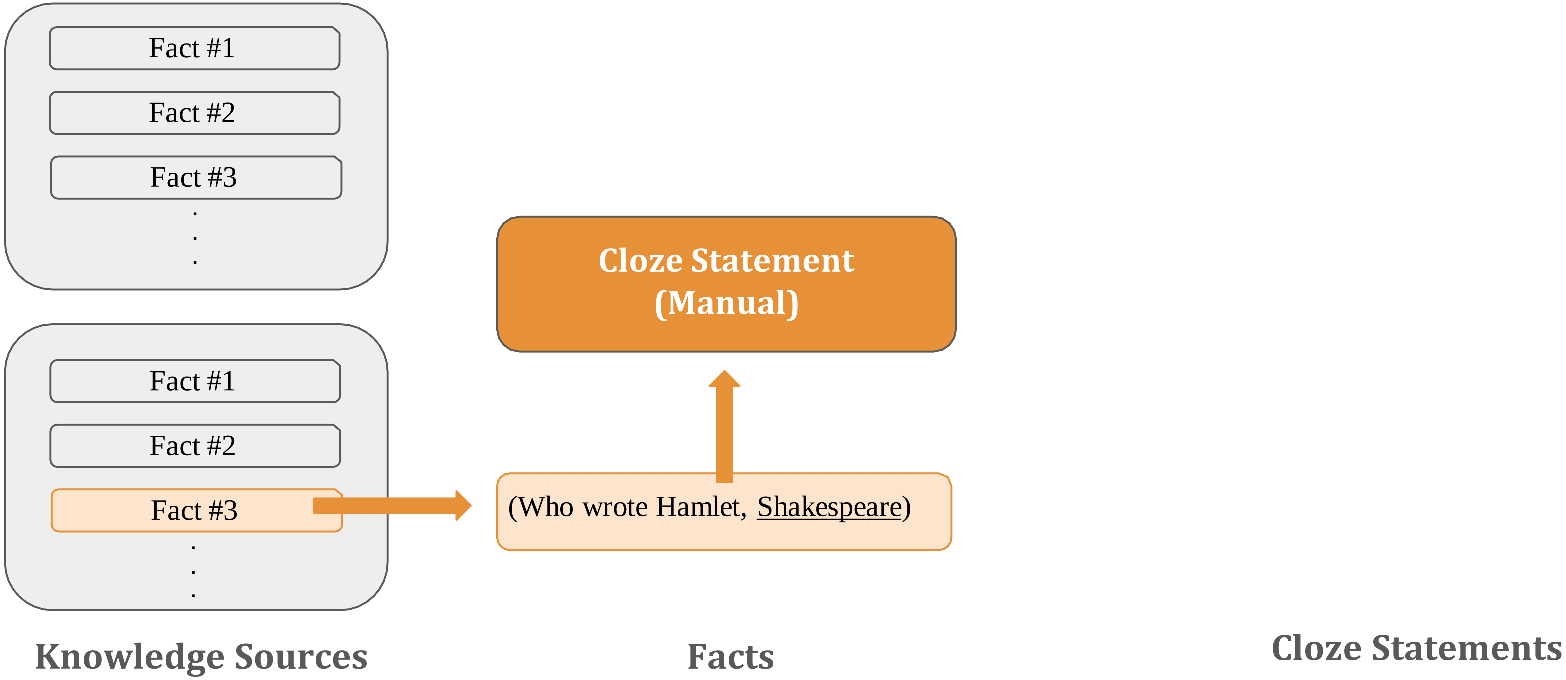
Architecture of the LAMA probe

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



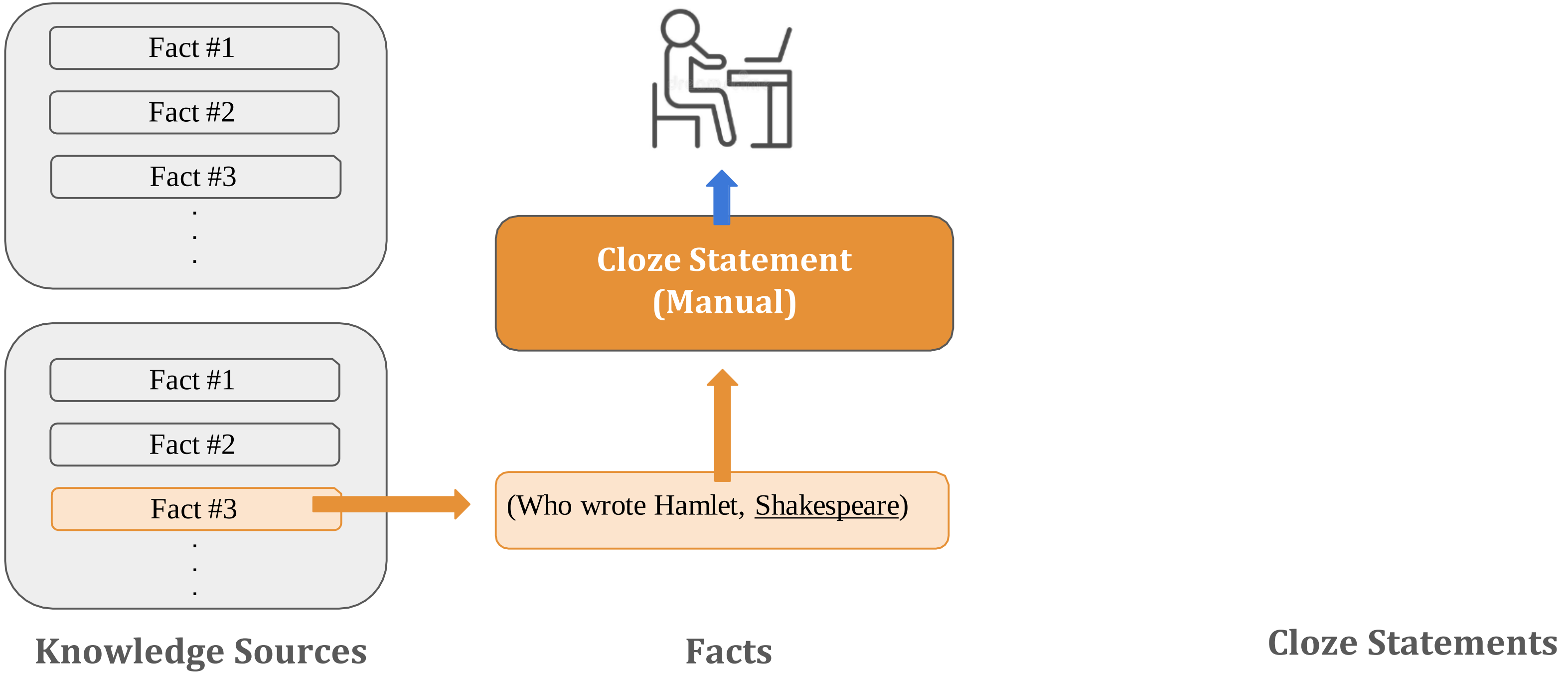
Architecture of the LAMA probe

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



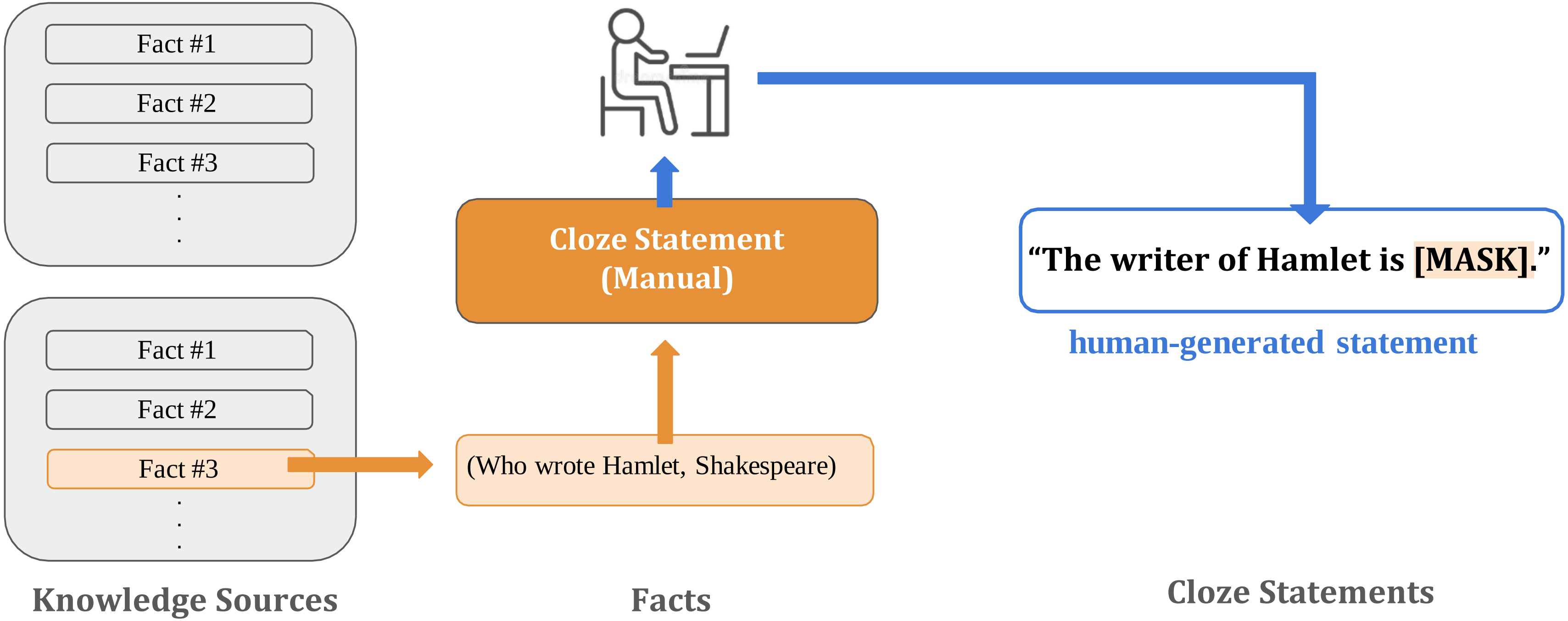
Architecture of the LAMA probe

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?

Table from Lewis et al., 2020

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

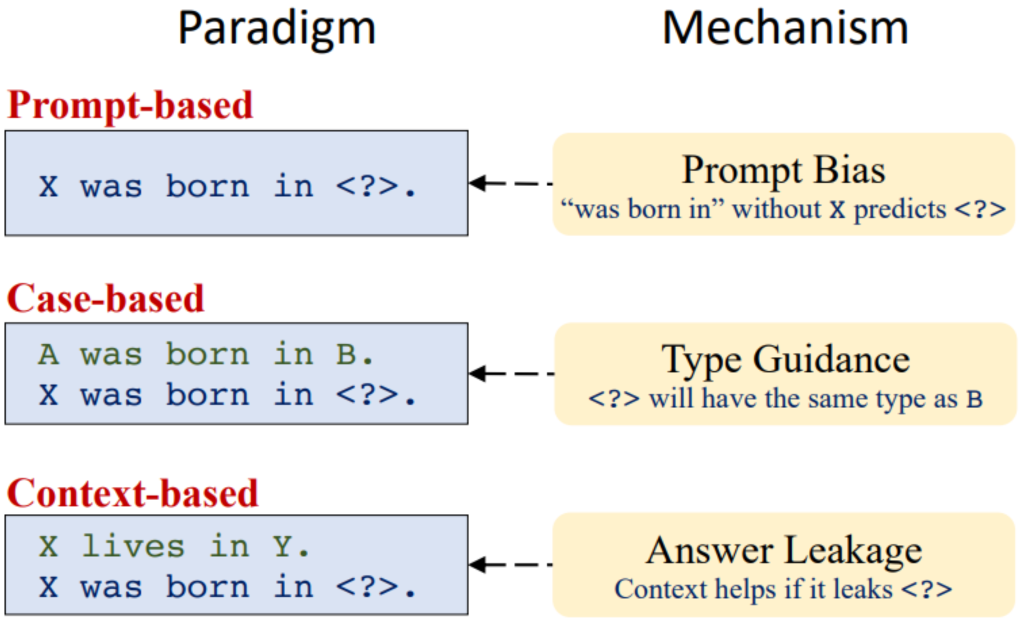
When there is question overlap, both open and closed-book LMs perform well

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	[illegible]	19.3	57.9	80.4	[illegible]	31.6	42.4	74.1	[illegible]	22.2
	FID	51.4	71.3	[illegible]	34.5	67.6	87.5	[illegible]	42.8	-	-	[illegible]	-
Closed book	T5-11B+SSM	36.6	77.2	[illegible]	9.4	-	-	[illegible]	-	44.7	82.1	[illegible]	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



Boxi Cao et.al. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. ACL 2021. https://aclanthology.org/2021.acl-long.146.pdf

Revising LAMA – Reporting Bias

It is uninterested 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	4.01×10^{-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	4.37×10^{-3} (1 in 229 deaths)
being abandoned	0.000175 (7000 each year, out of 4M births)

Revising LAMA – Reporting Bias

	BERT	RoBERTa	GPT-2		BERT	RoBERTa	GPT-2
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).

Vered Shwartz and Yejin Choi. Do Neural Language Models Overcome Reporting Bias?. COLING 2020. https://aclanthology.org/2020.coling-main.605.pdf

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

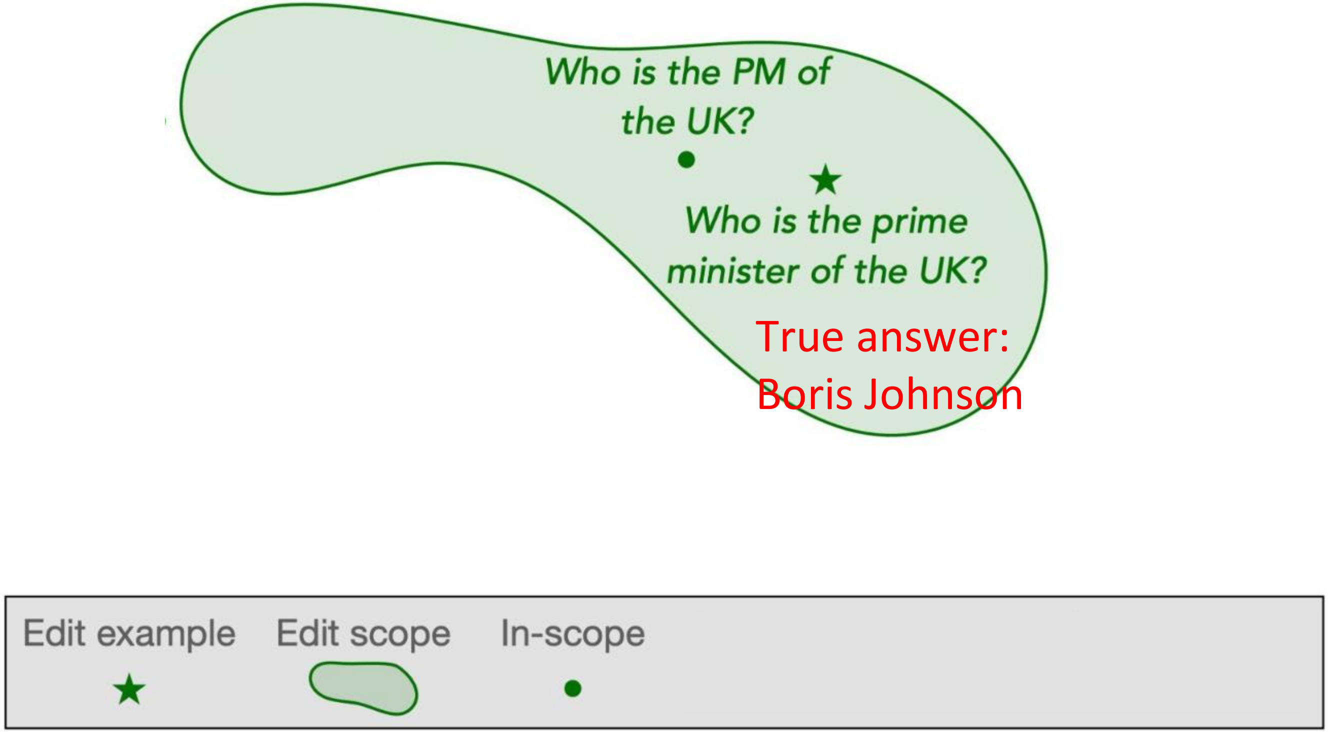
How to update knowledge in pre-trained models?

Edit What, Exactly?
Defining the problem

Now it becomes Keir Starmer. But let us keep the original examples

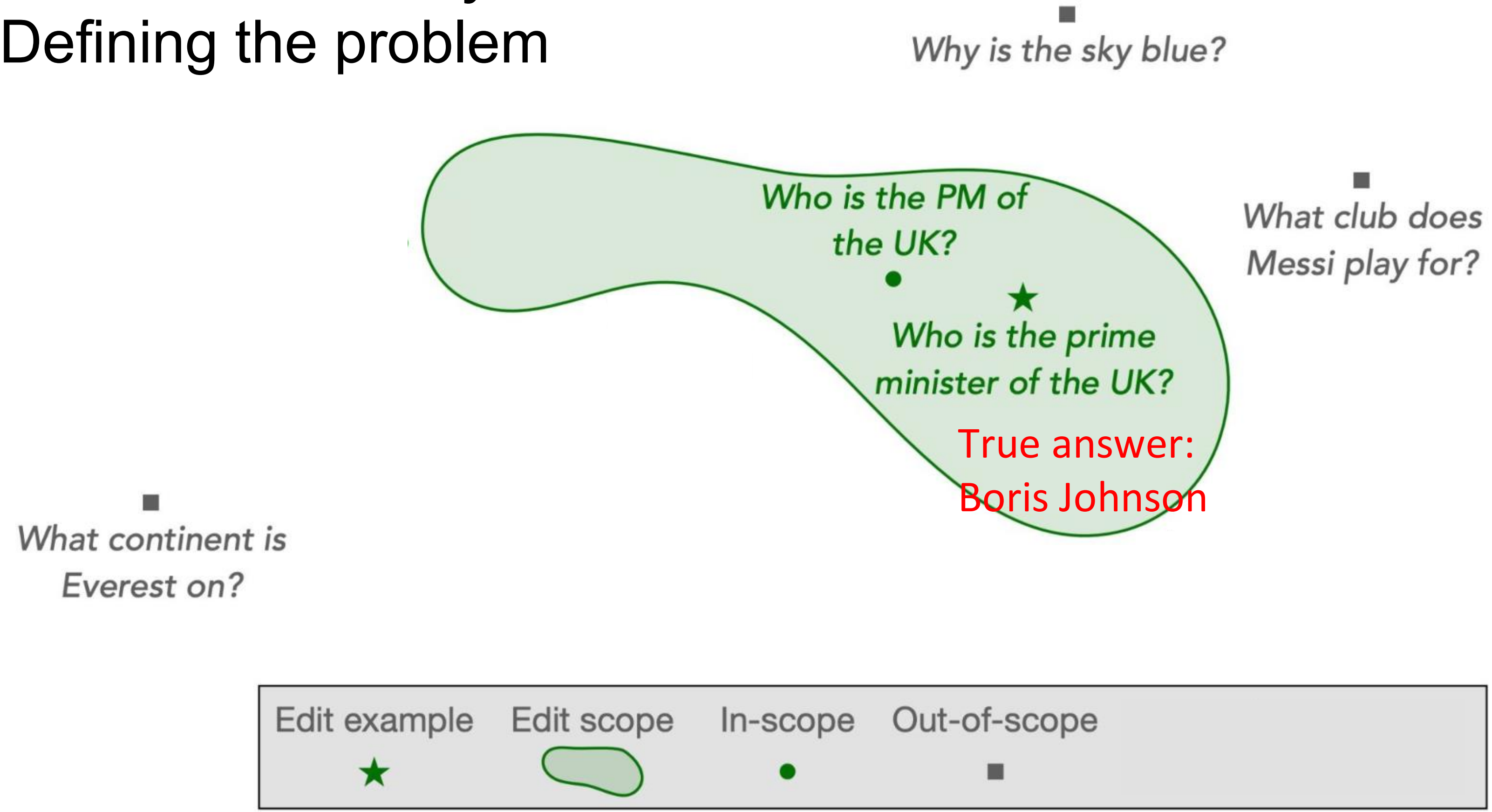
https://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture-editing.pdf

Edit What, Exactly?
Defining the problem



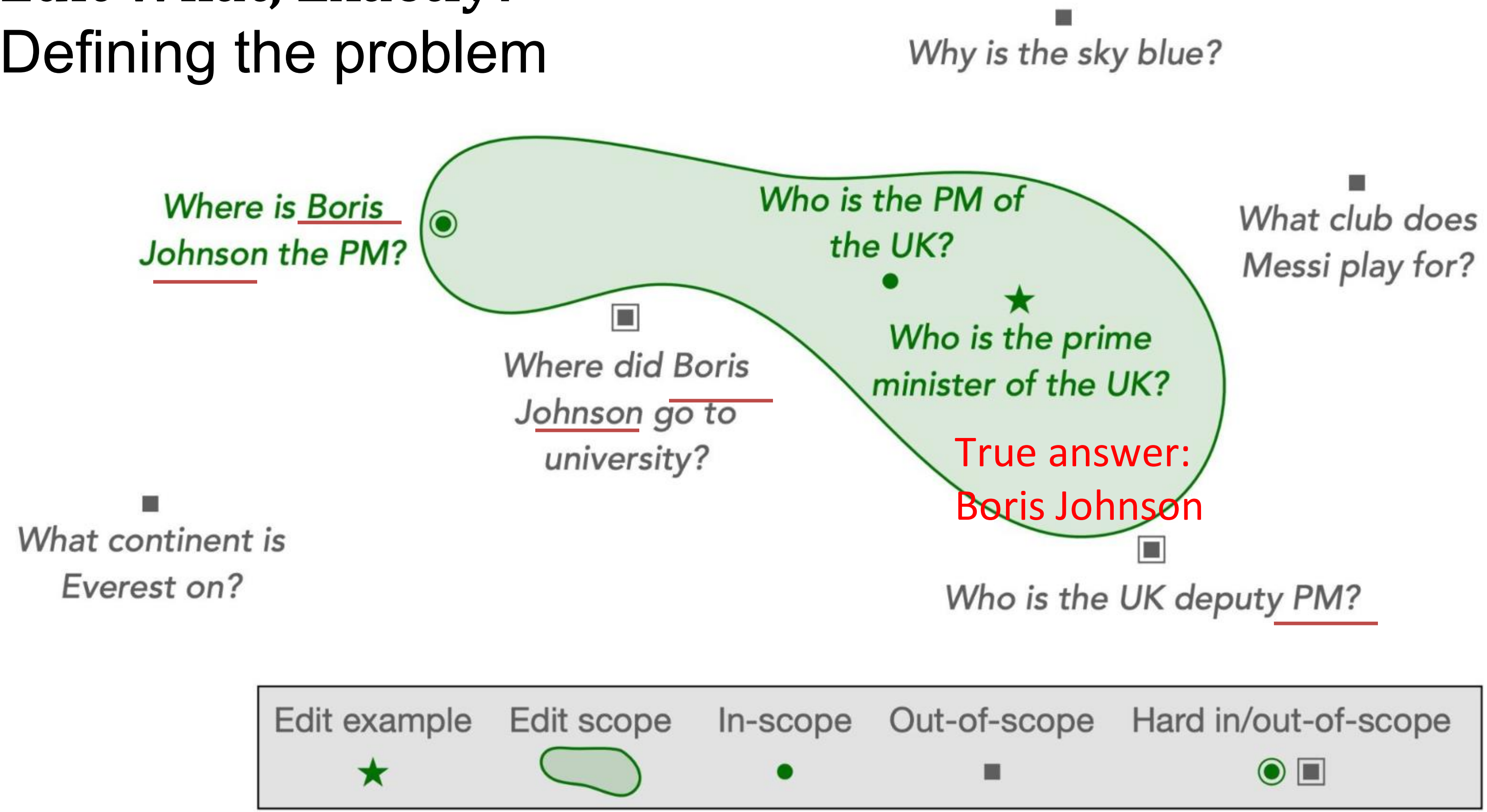
https://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture-editing.pdf

Edit What, Exactly?
Defining the problem



https://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture-editing.pdf

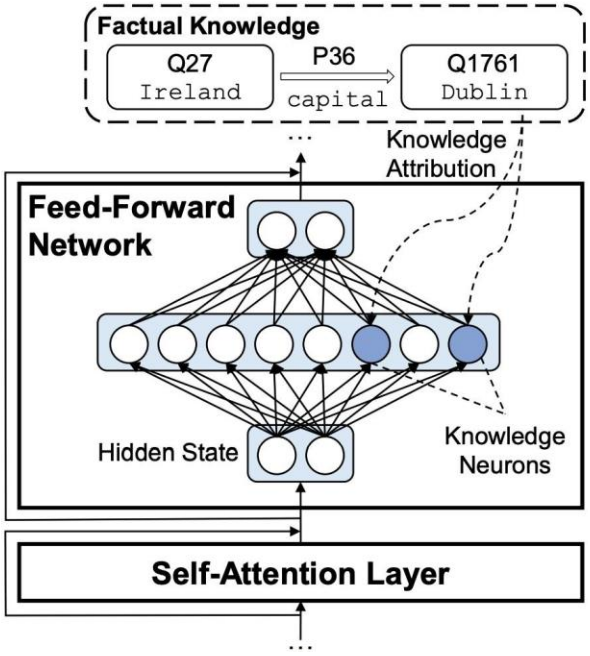
Edit What, Exactly?
Defining the problem



https://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture-editing.pdf

How to edit knowledge in pre-trained models?

Knowledge Neurons



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

Dai, Damai, et al. "Knowledge neurons in pretrained transformers." arXiv preprint arXiv:2104.08696 (2021).

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.

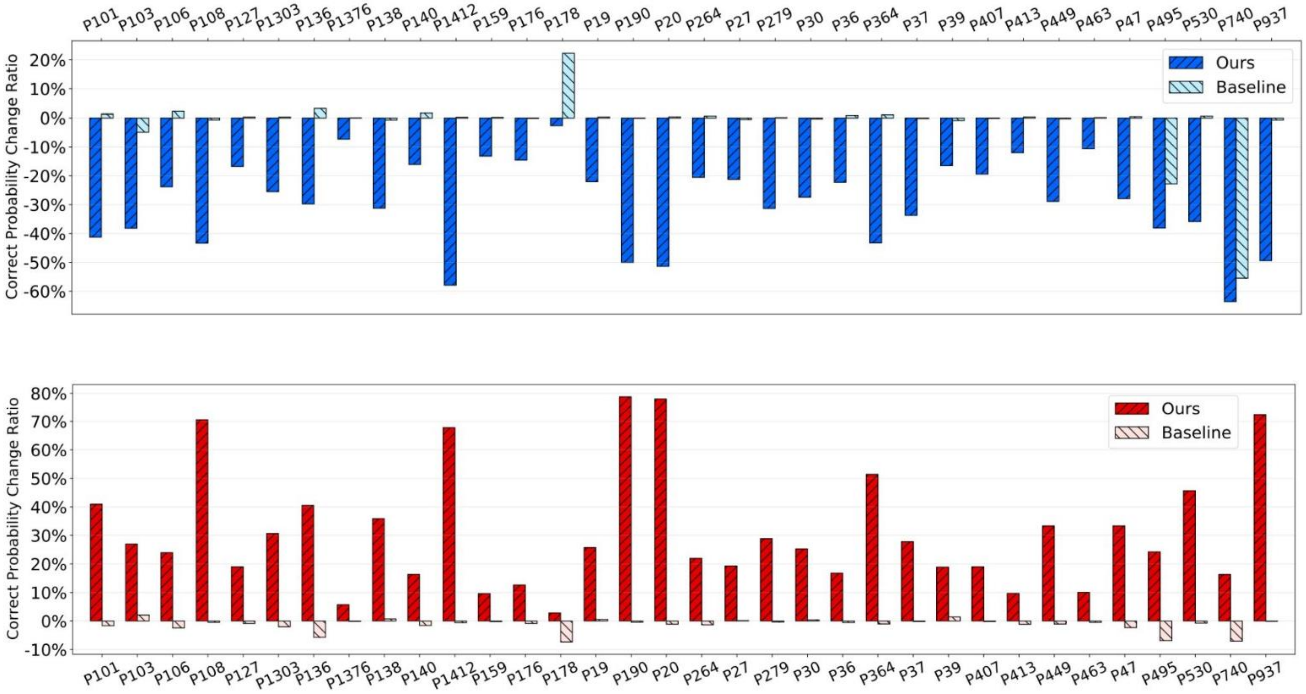
Dai, Damai, et al. "Knowledge neurons in pretrained transformers." arXiv preprint arXiv:2104.08696 (2021).

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.

Dai, Damai, et al. "Knowledge neurons in pretrained transformers." arXiv preprint arXiv:2104.08696 (2021).

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.

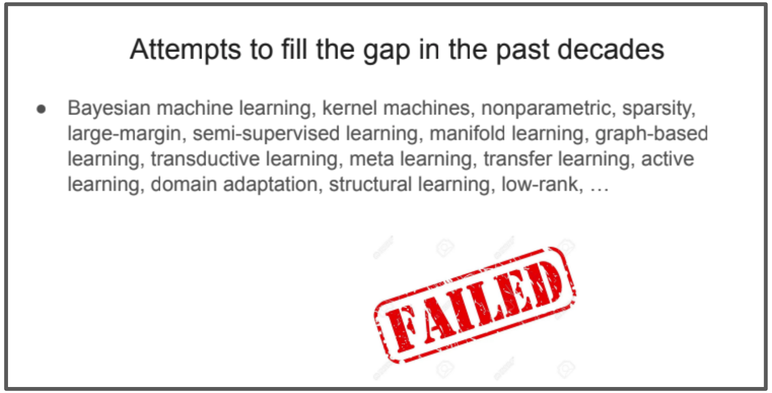
Dai, Damai, et al. "Knowledge neurons in pretrained transformers." arXiv preprint arXiv:2104.08696 (2021).

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

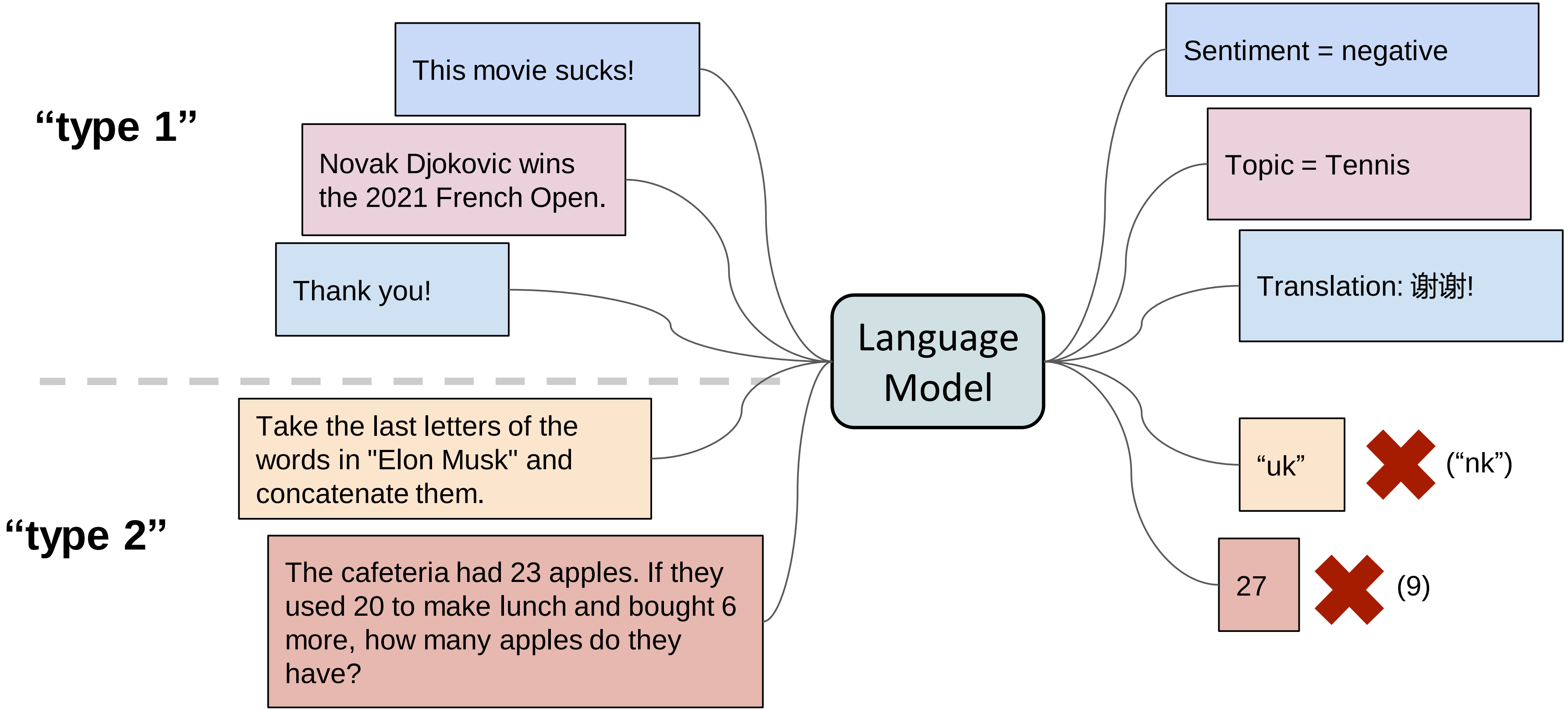
Human Intelligence vs. Traditional machine learning? (Hint: reasoning)

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



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?

Grammar	In my free time, I like to **{<u>run</u>, banana}**
Lexical semantics	I went to the zoo to see giraffes, lions, and **{<u>zebras</u>, spoon}**
World knowledge	The capital of Denmark is **{<u>Copenhagen</u>, London}**
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was **{<u>good</u>, bad}**
Harder sentiment analysis	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}**
Translation	The word for "pretty" in Spanish is **{<u>bonita</u>, hola}**
Spatial reasoning	[...] Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the **{<u>kitchen</u>, store}**
Math question	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?

Current world knowledge	The stock price of APPL on March 1st, 2023 is **{???}**
Arbitrarily long arithmetic	36382894730 + 238302849204 = **{???}**
Many-step reasoning	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 **{???}**
Predict the future	The winner of the FIFA world cup in 2026 is **{???}**
Information not in the training data	Jason Wei’s favorite color is **{???}**
Extremely long inputs	[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	Protein discovery	(?) Protein discovery
Clinical diagnosis	Clinical diagnosis	(?) Clinical diagnosis
Play chess well	Play chess well	(?) Play chess well
High-level planning	High-level planning	(?) High-level planning
Abstract reasoning	Abstract reasoning	(?) Abstract reasoning
Simple math	Simple math	Simple math
Commonsense reasoning	Commonsense reasoning	Commonsense reasoning
Know world knowledge	Know world knowledge	Know world knowledge
Translation	Translation	Translation
Sentiment analysis	Sentiment analysis	Sentiment analysis
Generate coherent text	Generate coherent text	Generate coherent text
Be grammatically correct	Be grammatically correct	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

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

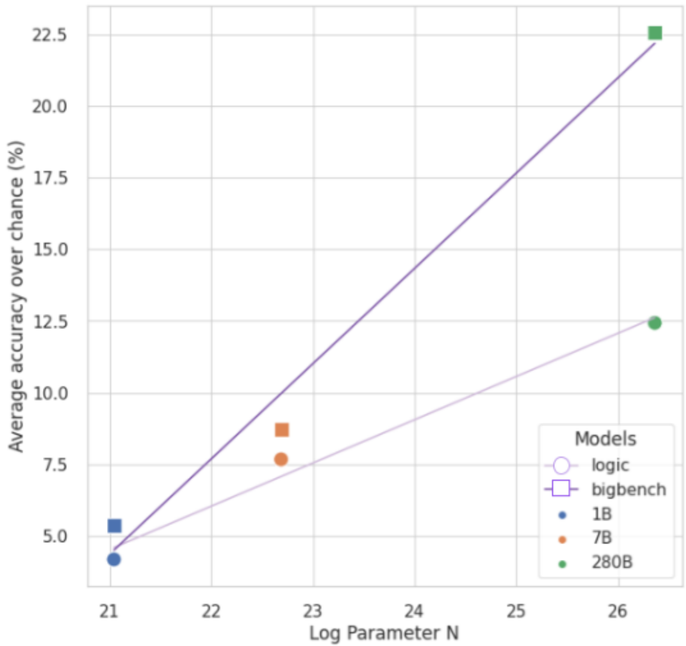
Q: Wolves are afraid of mice. Sheep are afraid of wolves. Emily is a wolf. What is Emily afraid of?

A: The answer is **mice.**

Logical Reasoning

Scaling laws are worse for logical reasoning in 2022

(Creswell et al. 2022)



natural language tasks

logical reasoning tasks

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 \[\vert \vert x-1 \vert -2 \vert=\frac{m}{100}\] have 4 distinct solutions? 199

https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize/data?select=train.csv

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。

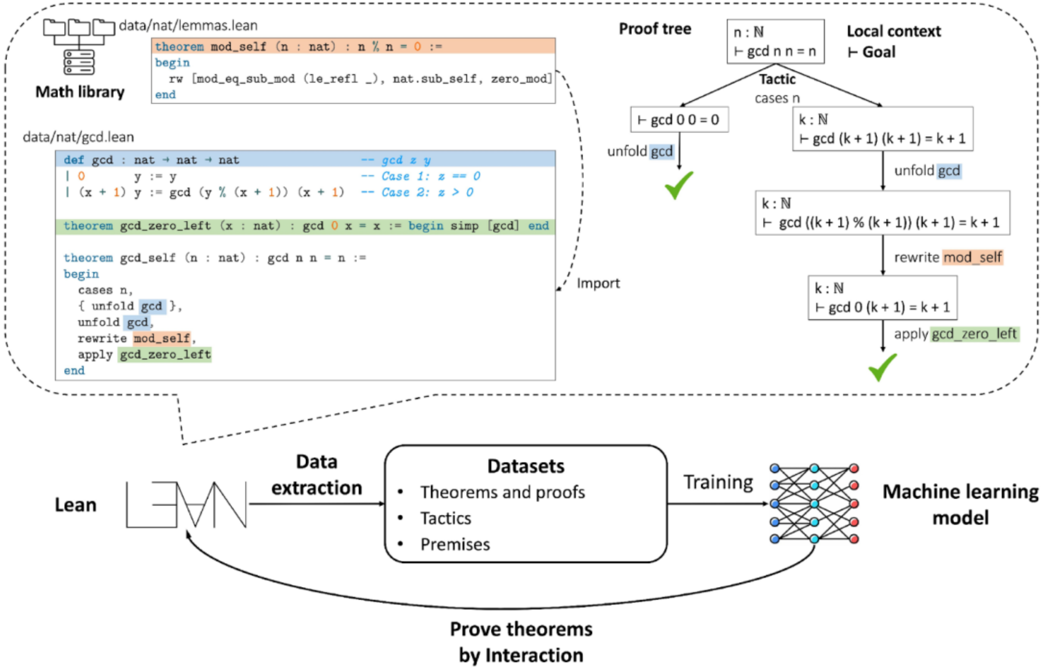
模型目标

1. **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：$$\max S = \sum_{j \in D} \alpha_j \left(1 - \frac{\sum_{i \in S} t_{ij}x_{ij}}{T_j}\right)$$
    其中，t_{ij}是从供应点i到需求点j的运输时间，T_j是需求点j对运输时间的最大容忍值，α_j是与需求点j的服务水平重要性相关的权重。
2. **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。
3. **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：$$\min Z = \sum_{i \in S}\sum_{j \in D} c_{ij}x_{ij}$$
4. **库存水平平衡**：优化各供应点和需求点的库存水平，以减少过剩或短缺的情况，保持库存的稳定。
  - 公式：$$\min I = \sum_{i \in S} \left| x_{i} - L_i \right| + \sum_{j \in D} \left| x_{j} - L_j \right|$$
    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。
5. **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的物资量。

模型约束
1. **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：$$\sum_{j \in D} x_{ij} \leq a_i, \quad \forall i \in S$$
2. **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：$$\sum_{i \in S} x_{ij} = b_j, \quad \forall j \in D$$
3. **非负约束**：调拨的物资量不能为负。
  - 公式：$$x_{ij} \geq 0, \quad \forall i \in S, \forall j \in D$$
4. **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：$$L_i \leq \text{库存水平}_i \leq U_i, \quad \forall i \in S$$
  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。
5. **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：$$\sum_{k \in K} x_{ijk} = x_{ij}, \quad \forall i \in S, \forall j \in D$$
    这确保了从供应点i到需求点j的物资量x_{ij}通过某一种运输模式k进行调拨。
6. **时间窗约束**：满足需求点的特定配送时间窗要求。
  - 公式：$$t_{ij}^{start} \leq t_{ij} \leq t_{ij}^{end}, \quad \forall i \in S, \forall j \in D$$
    其中，t_{ij}^{start}和t_{ij}^{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。

Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, **Benyou Wang**. Mamo: a Mathematical Modeling Benchmark with Solvers. https://arxiv.org/abs/2405.13144v1 Findings of ACL 2024
Zhengyang Tang, Chenyu Huang, Xin Zheng, Shixi Hu, Zizhuo Wang, Dongdong Ge, **Benyou Wang**. ORLM: Training Large Language Models for Optimization Modeling. https://arxiv.org/abs/2405.17743 **Operations Research**

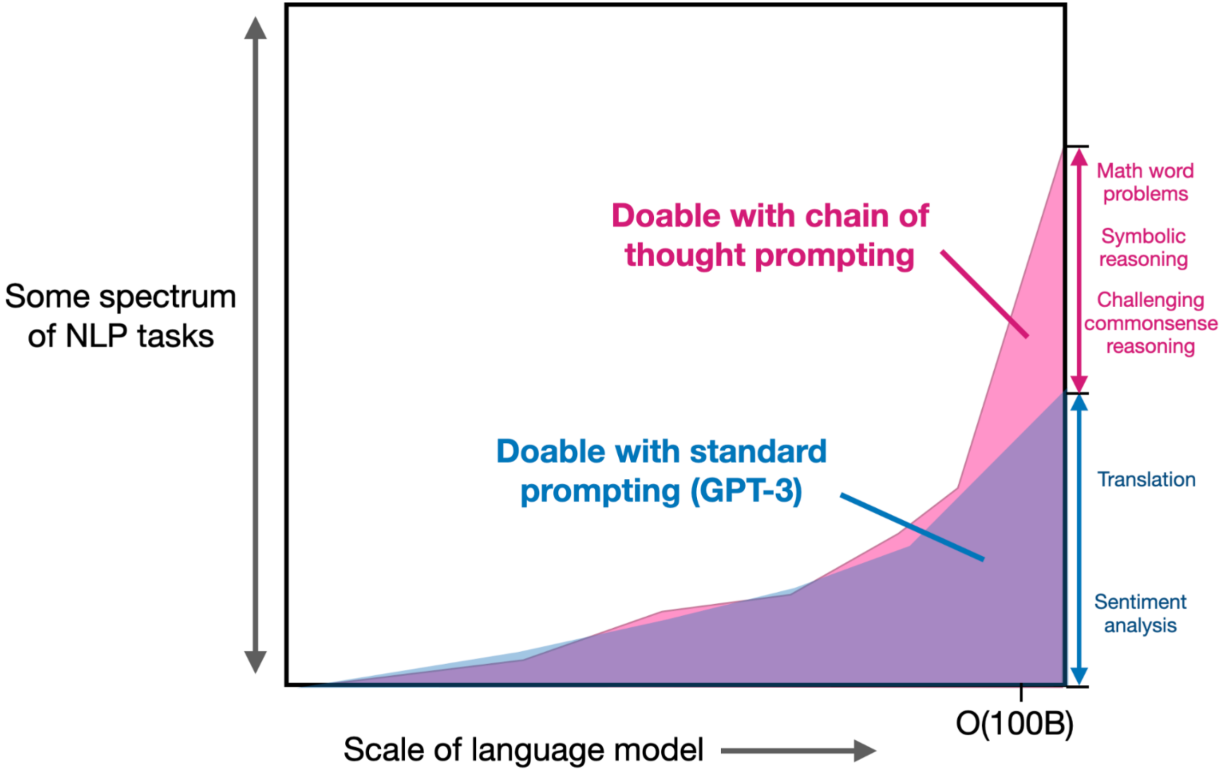
Interesting task: Automatic theorem proving



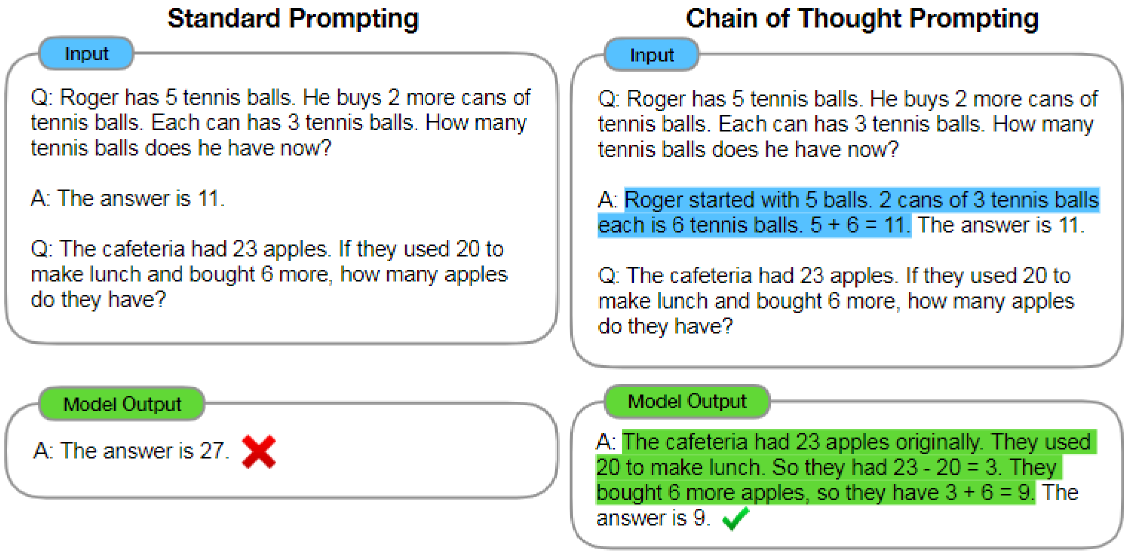
Today's Lecture

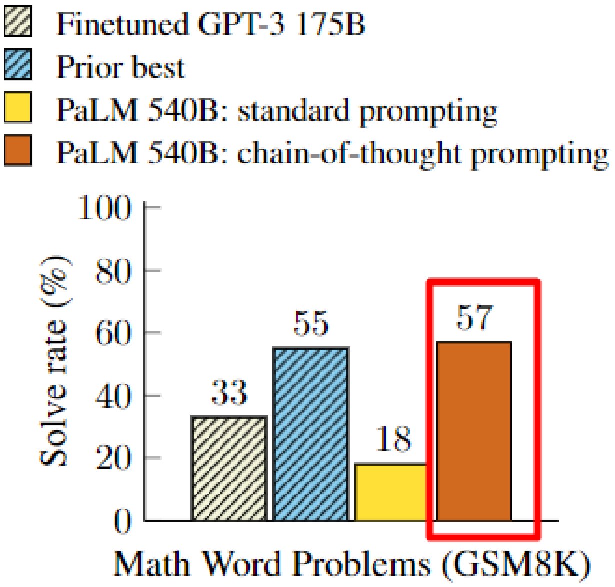
- **Knowledge** in LLMs
  - LLMs as knowledge bases
  - Facts updating for LLMs
- **Reasoning** in LLMs
  - Why reasoning is special in LLMs
- **Prompt** Techniques for better reasoning

LLMs with CoT could do reason !



What is Chain of Thought prompting (CoT)?





<input, intermediate results, output>

- decompose into easier intermediate steps
- interpretable

Zero-Shot CoT – Let’s think step by step

Examples

(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. ✗

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

CoT 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: 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. ✓

Step-by-step Answer

(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 ✗

(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. ✓

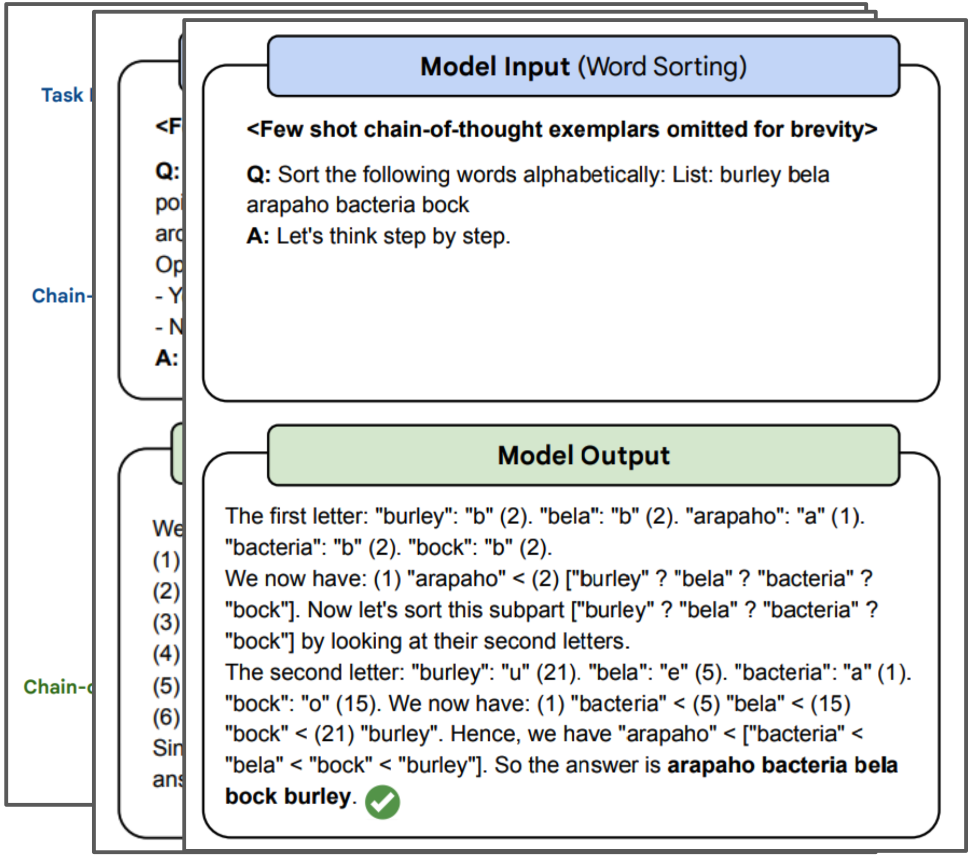
Step-by-step Answer

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

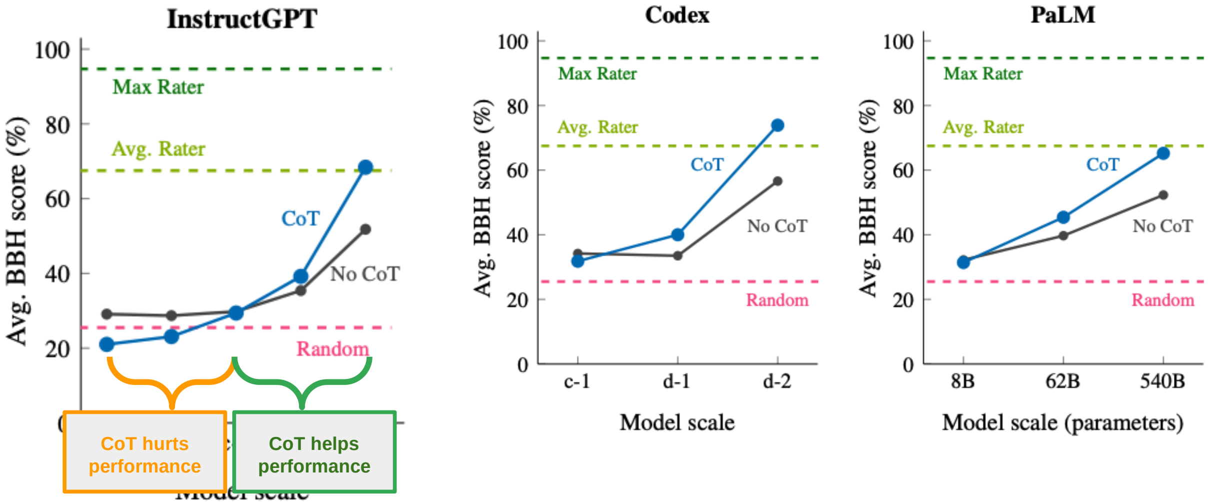
Detail: better formatting (options, task description) already beats prior best

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

CoT on BIG-Bench: Scaling

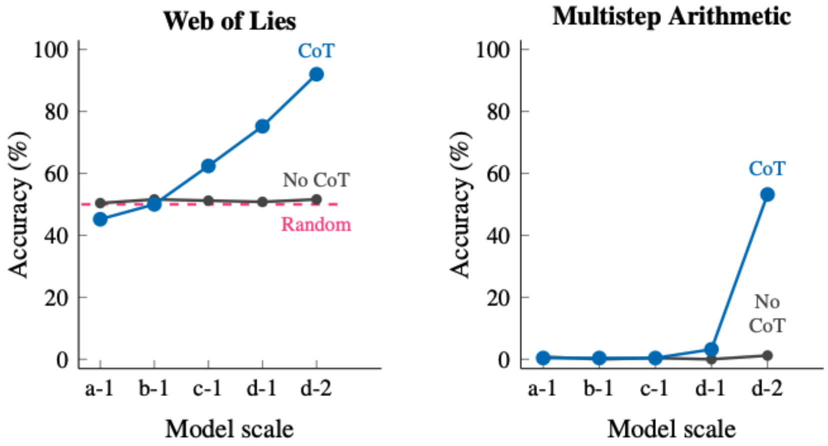
- CoT requires sufficient model scale for positive delta

[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 multiplication (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.

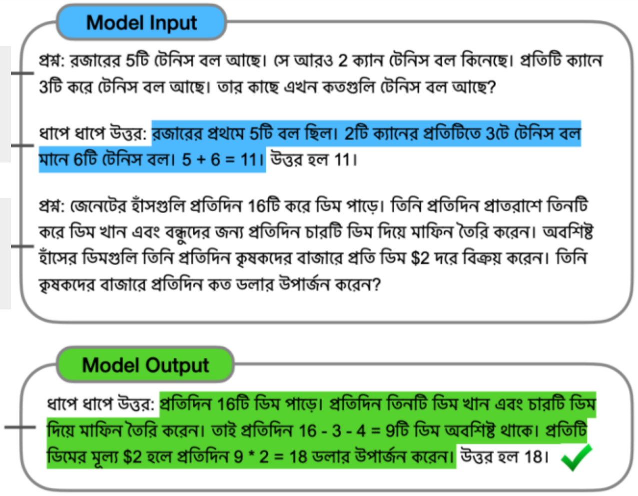
Let's calculate B = (-4 - 1 - -4 - 4) = ((-4 - 1) - -4 - 4) = ((-5) - -4 - 4) = ((-5 - -4) - 4) = ((-5 + 4) - 4) = (-1 - 4) = -5.

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

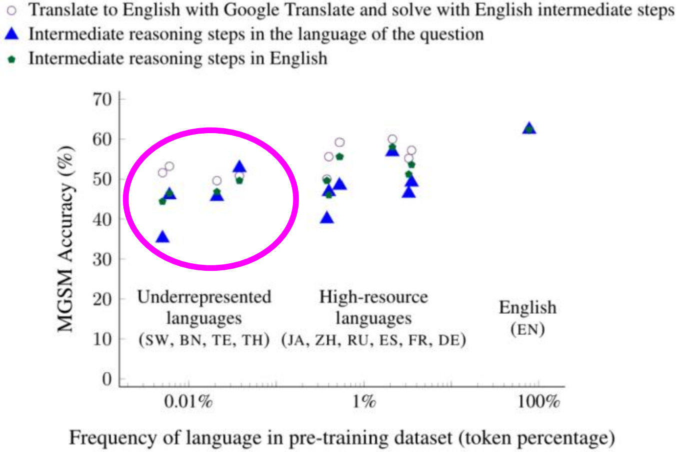
Multilingual chain-of-thought prompting

Prompt the model with Bengali math problems and Bengali reasoning

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



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



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)

Suggested further reading:
[Large language models are zero-shot reasoners.]

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

Limited exploration

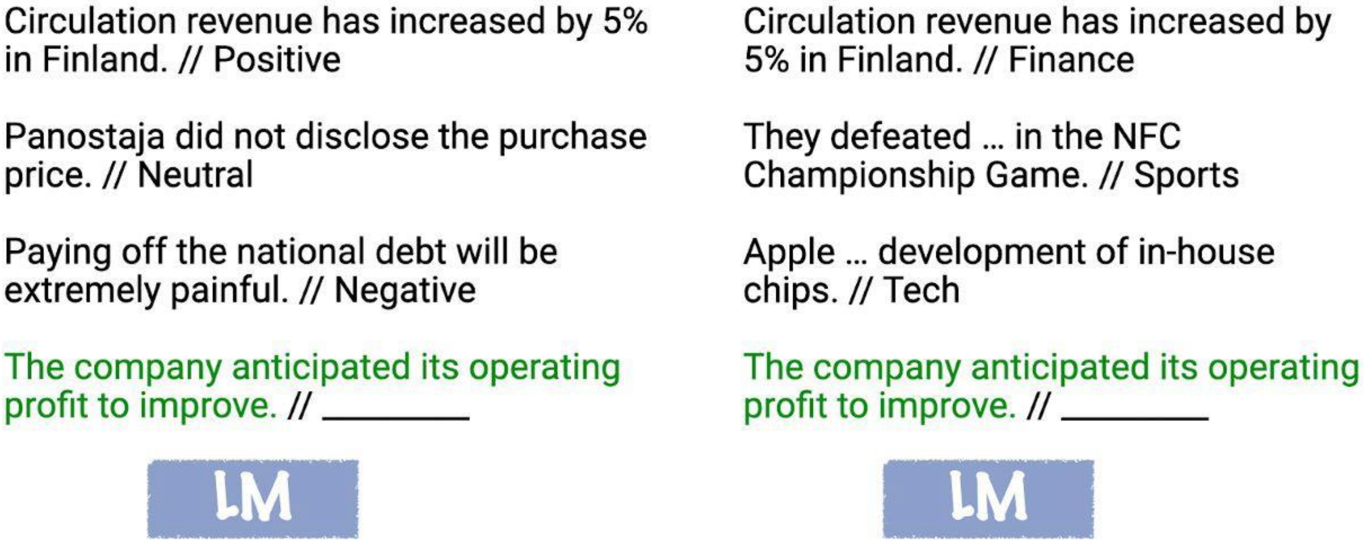
Q: What is 1+2+3+…+6?
A:
1+2 = 3
3+3 = 6
6+4 = 10
10+5 = 15
15+6 = 21
So 1+2+3+4+5+6=21.

Correct yet not good

Improve CoT in different phases of reasoning

- Pre-process of the reasoning task:
  - Demonstration using in-context learning
  - 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:
  - Majority voting: e.g. Self-consistency
  - Verify: e.g. Verifier
  - Refine: e.g. Self-refine

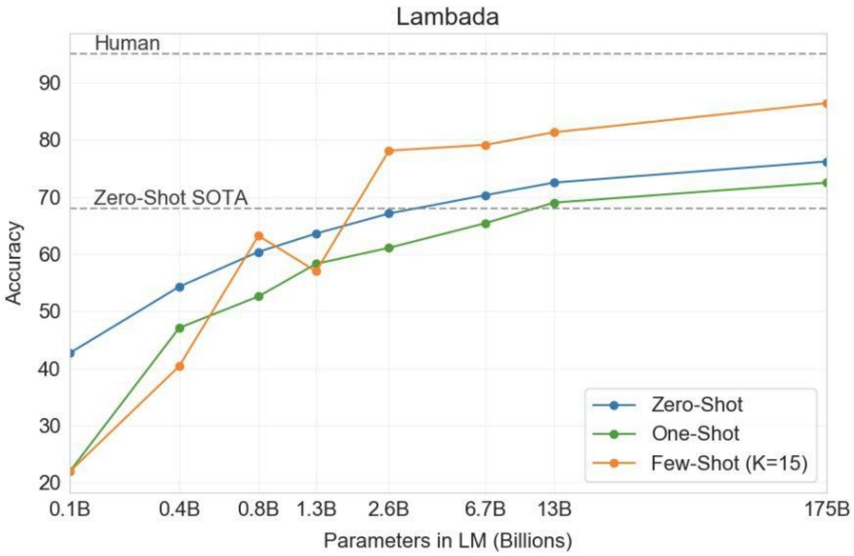
What is In-Context Learning?



https://ai.stanford.edu/blog/understanding-incontext/

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!

Least-to-most prompting

Explicitly decompose into subquestions

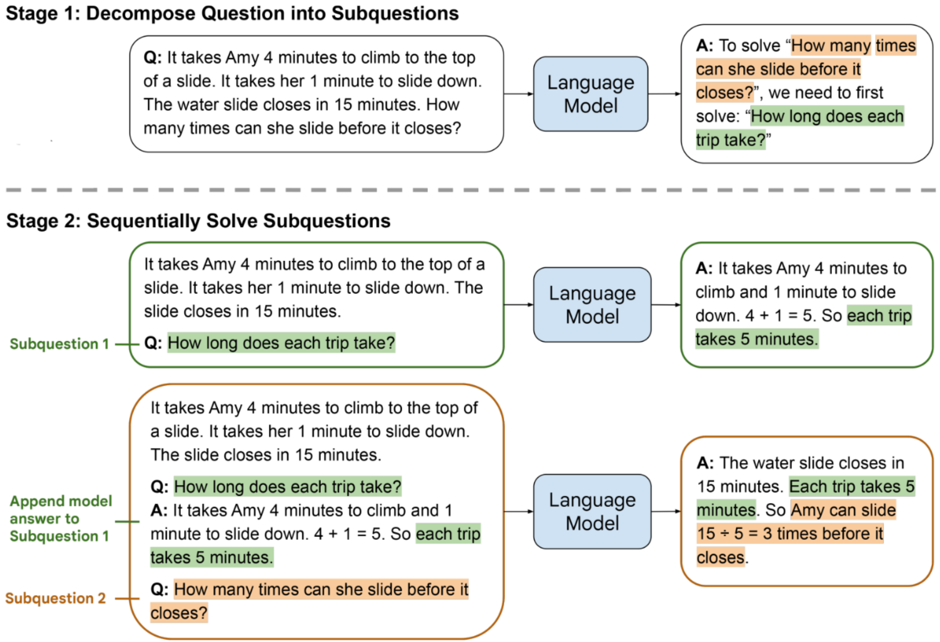


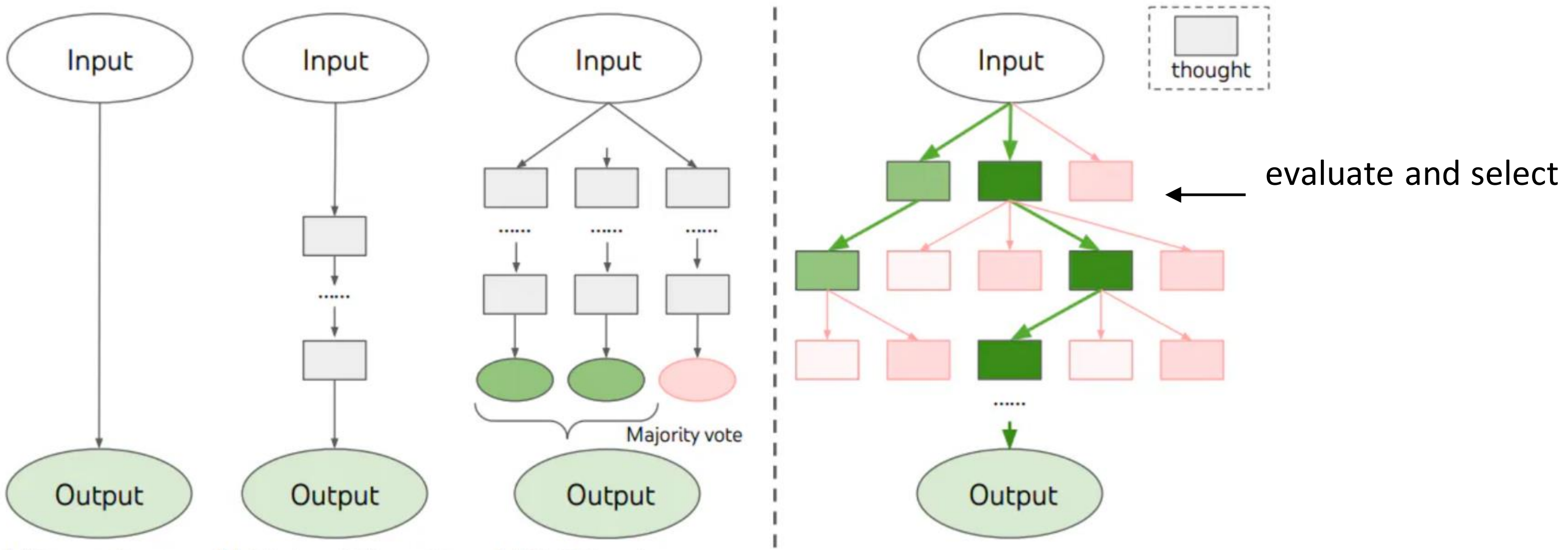
Figure 1: Least-to-most prompting solving a math word problem in two stages: (1) query the language model to decompose the problem into subproblems; (2) query the language model to sequentially solve the subproblems. The answer to the second subproblem is built on the answer to the first subproblem.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

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

Tree of Thoughts (ToT)

Explore over units of text that serve as intermediate steps

The ToT framework is illustrated below:



(a) Input-Output Prompting (IO)

(c) Chain of Thought Prompting (CoT)

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

(d) Tree of Thoughts (ToT)

[Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." arXiv preprint arXiv:2305.10601 (2023).]

Program of Thoughts (PoT)

Output Python programs and call Python interpreter to calculate the answers

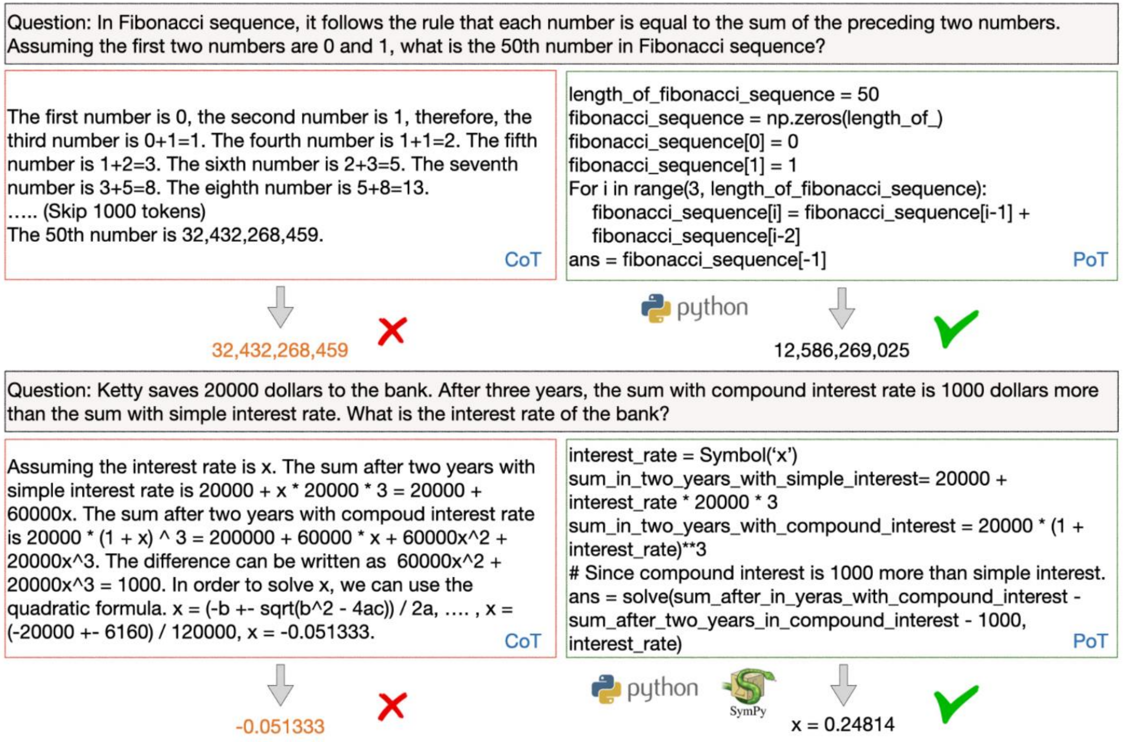


Figure 1: Comparison between Chain of Thoughts and Program of Thoughts.

[Chen, Wenhu, et al. "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks." arXiv preprint arXiv:2211.12588 (2022).]

A trick for COT: Self-consistency

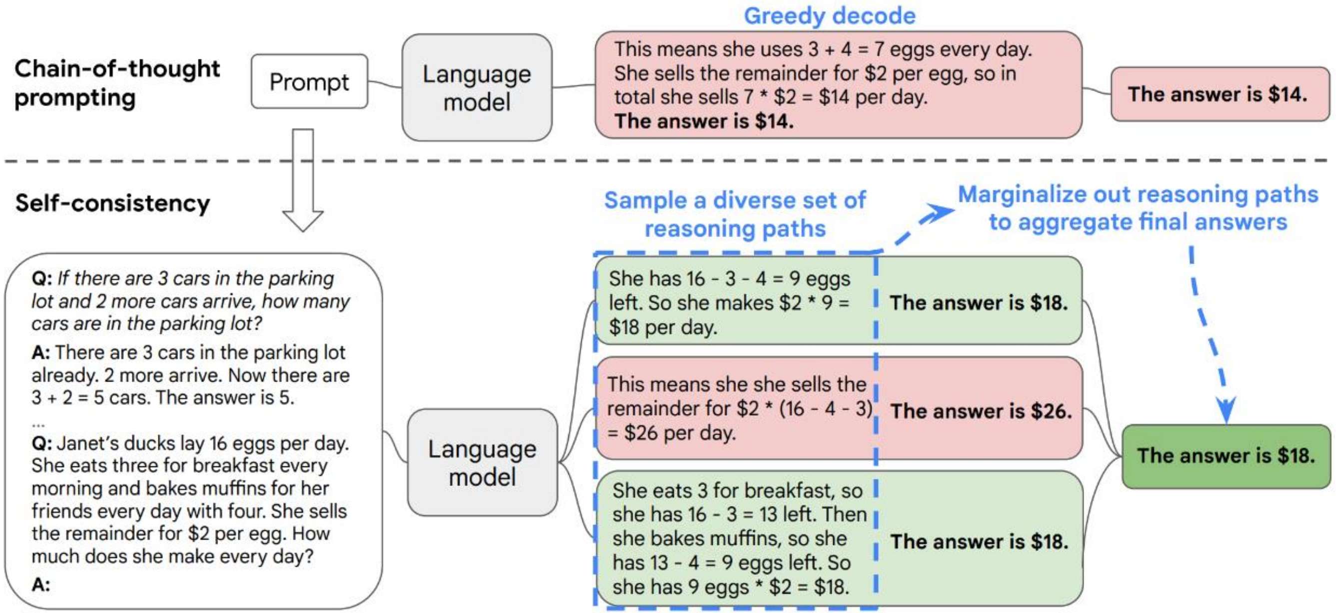
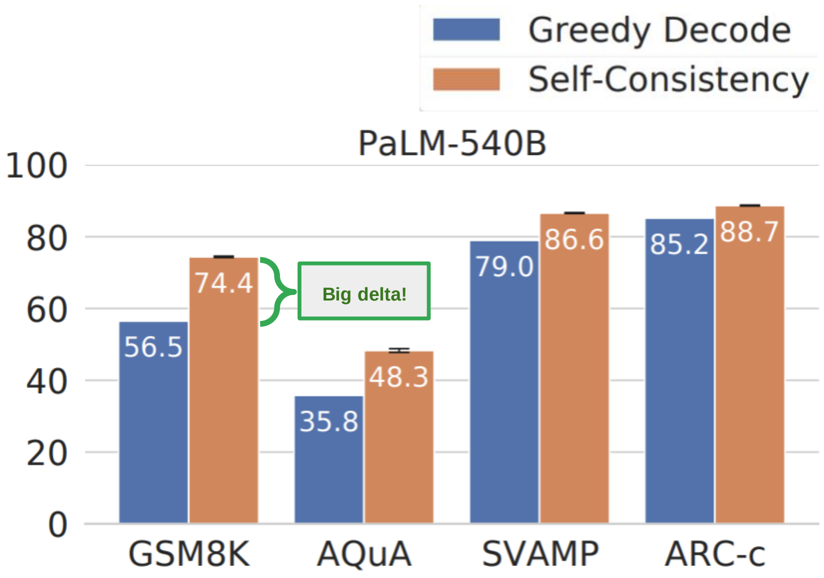


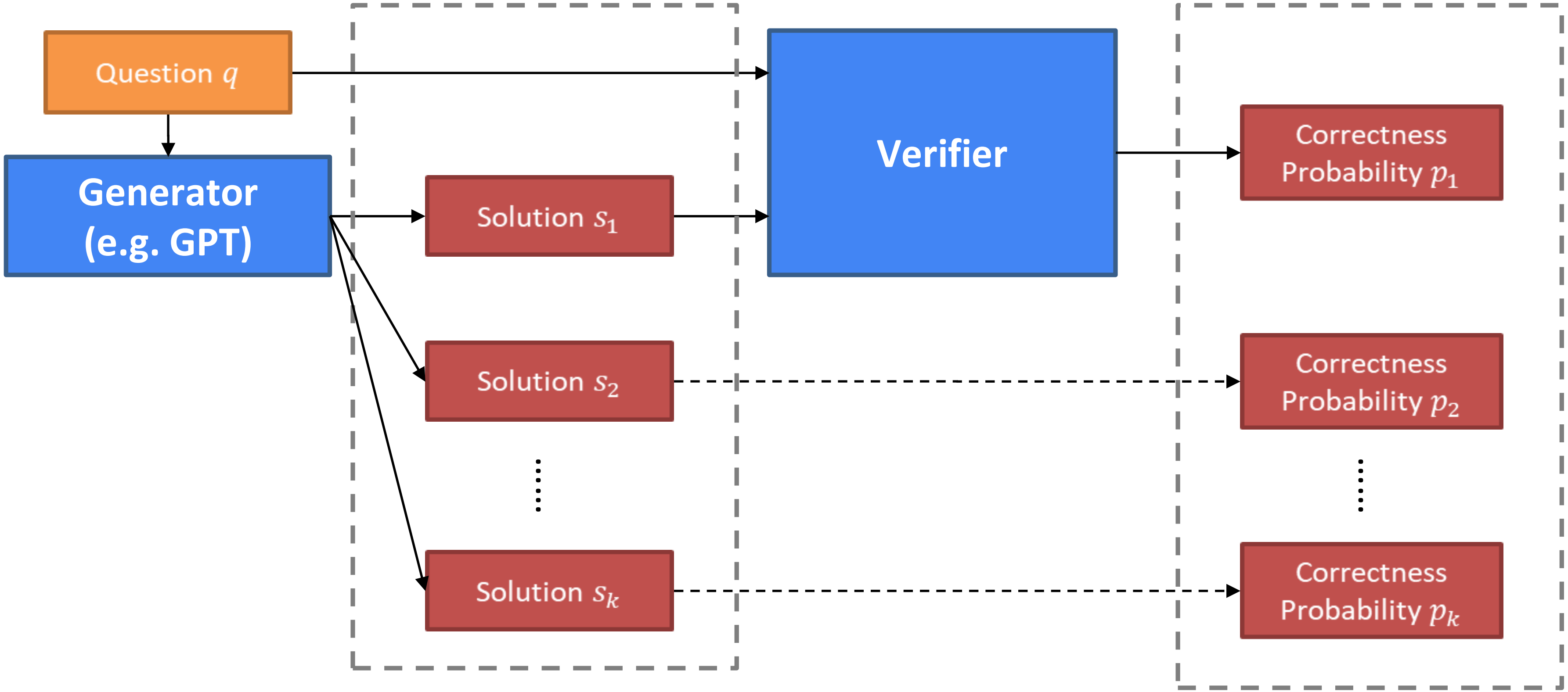
Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the "greedy decode" in CoT prompting by sampling from the language model's decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning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

[Wang, Xuezhi, et al. "Self-consistency improves chain of thought reasoning in language models." arXiv preprint arXiv:2203.11171 (2022).]

Self-consistency works really well



Verifier in COT



Fei Yu, Anningzhe Gao, Benyou Wang. Outcome-Supervised Verifier for reasoning. Submitted to NAACL 2023

Exceed GPT 4 using our OVM in GSM-8K

1	**GPT-4 Code Interpreter** (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	**GPT-4** (Model Selection, SC K=15)	96.8		×	Automatic Model Selection with Large Language Models for Reasoning			2023	majority voting
3	**GPT-4** (PHP, SC K=40)	96.5		×	Progressive-Hint Prompting Improves Reasoning in Large Language Models			2023	majority voting
4	**GPT-4** (Model Selection, SC K=5)	96.5		×	Automatic Model Selection with Large Language Models for Reasoning			2023	majority voting
5	**SFT-Mistral-7B** (Metamath, OVM, Smart Ensemble)	96.4	7	✓				2024	
6	**SFT-Mistral-7B** (AugData + ovm + ensemble)	95.9	7	✓				2024	
7	**GPT-4** (PHP)	95.5		×	Progressive-Hint Prompting Improves Reasoning in Large Language Models			2023	
8	**MindOpt Copilot Mistral-7B** (MetaMath, OVM, BS, Ensemble)	95.1	7	✓				2024	
9	**Claude 3 Opus** (0-shot chain-of-thought)	95		×	The Claude 3 Model Family: Opus, Sonnet, Haiku			2024	chain-of-thought, zero-shot
10	**Gemini Ultra** (Maj1@32)	94.4		×	Gemini: A Family of Highly Capable Multimodal Models			2023	majority voting
11	**SFT-Mistral-7B** (Metamath + ovm +ensemble)	94.13	7	✓				2024	
12	**GPT-4** (Ask, Refine, Trust)	94.08		×	The ART of LLM Refinement: Ask, Refine, and Trust			2023	
13	**Shepherd + DeepSeek-67B** (SFT on MetaMATH + PRM rerank, k=256)	93.3	67	✓	Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations			2023	rerank
14	**MindOpt Copilot Mistral-7B** (MetaMath, OVM, Ensemble)	93.2	7	✓				2024	

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

Self-refine

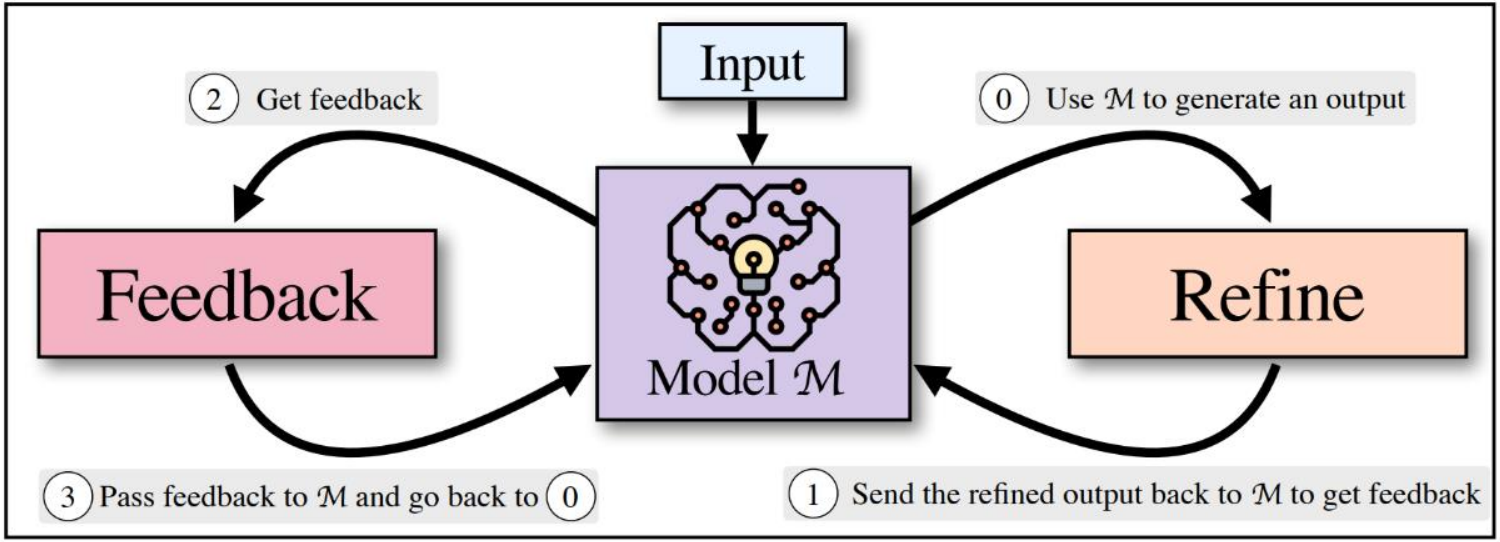


Figure 1: SELF-REFINE starts by taking an initially generated output (⓪), and passing it back to the same model $\mathcal{M}$ (①) to get feedback (②); feedback on the initial output is passed back to the model (③), to iteratively refine (⓪) the previously generated output. SELF-REFINE is instantiated with a powerful language model such as GPT-3.5 and does not involve human assistance.

Madaan, Aman, et al. "Self-refine: Iterative refinement with self-feedback." arXiv preprint arXiv:2303.17651 (2023).

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_w0BDrscDhPg7lMGzJCkeaPkGCJ3bN8dluXc/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)