Large Language Models

Language Models and Knowledge

M. Soleymani
Sharif University of Technology
Fall 2023

Many of slides have been adapted from J. Pan & M. Xia’s slides, Princeton cos597G, Fall 2022

Language Models and Knowledge

Outline

1. What is a knowledge base?
2. Can language models be used as knowledge bases? (Petroni et al., 2019)
3. How to update facts? (Dai et al., 2021, Mitchell et al., 2022)
4. Hallucination

Introduction

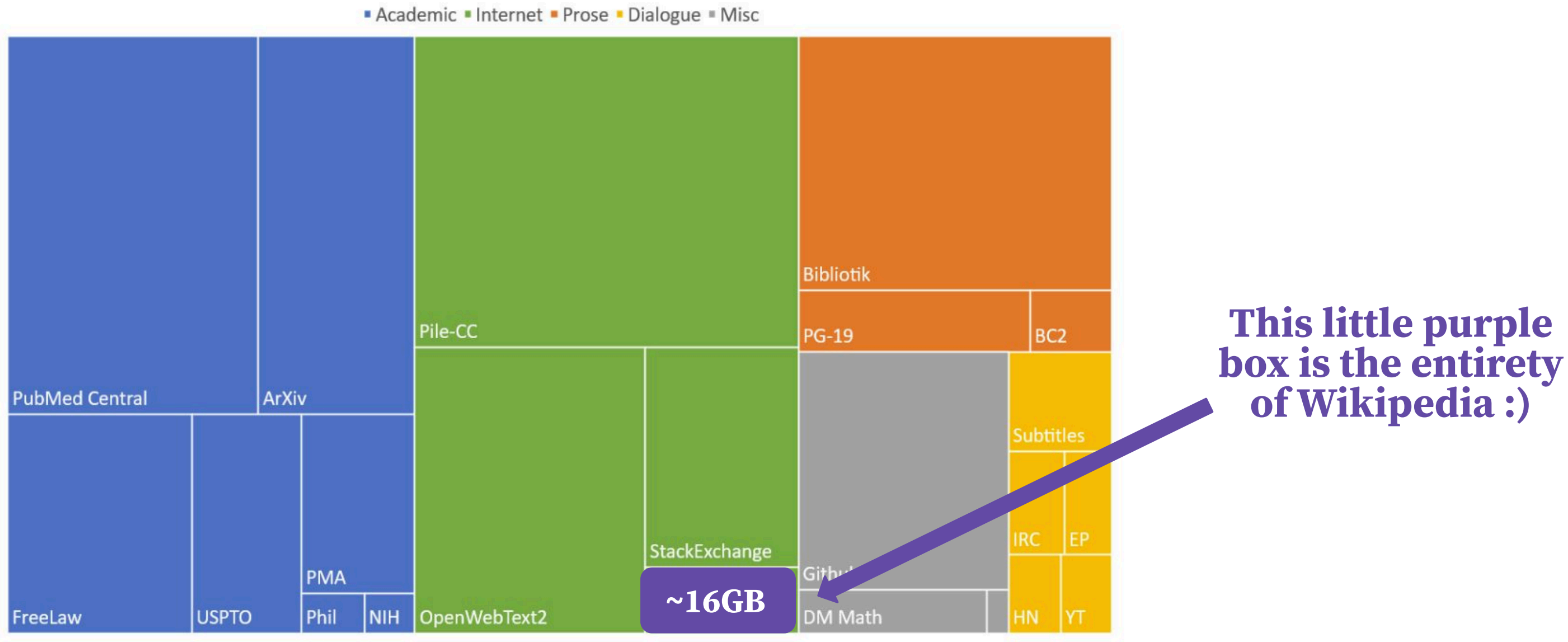
Motivation

- The corpora used to pre-train language models are **huge aggregations** of information and data from the internet

Graphic from Gao et al., 2020

Motivation

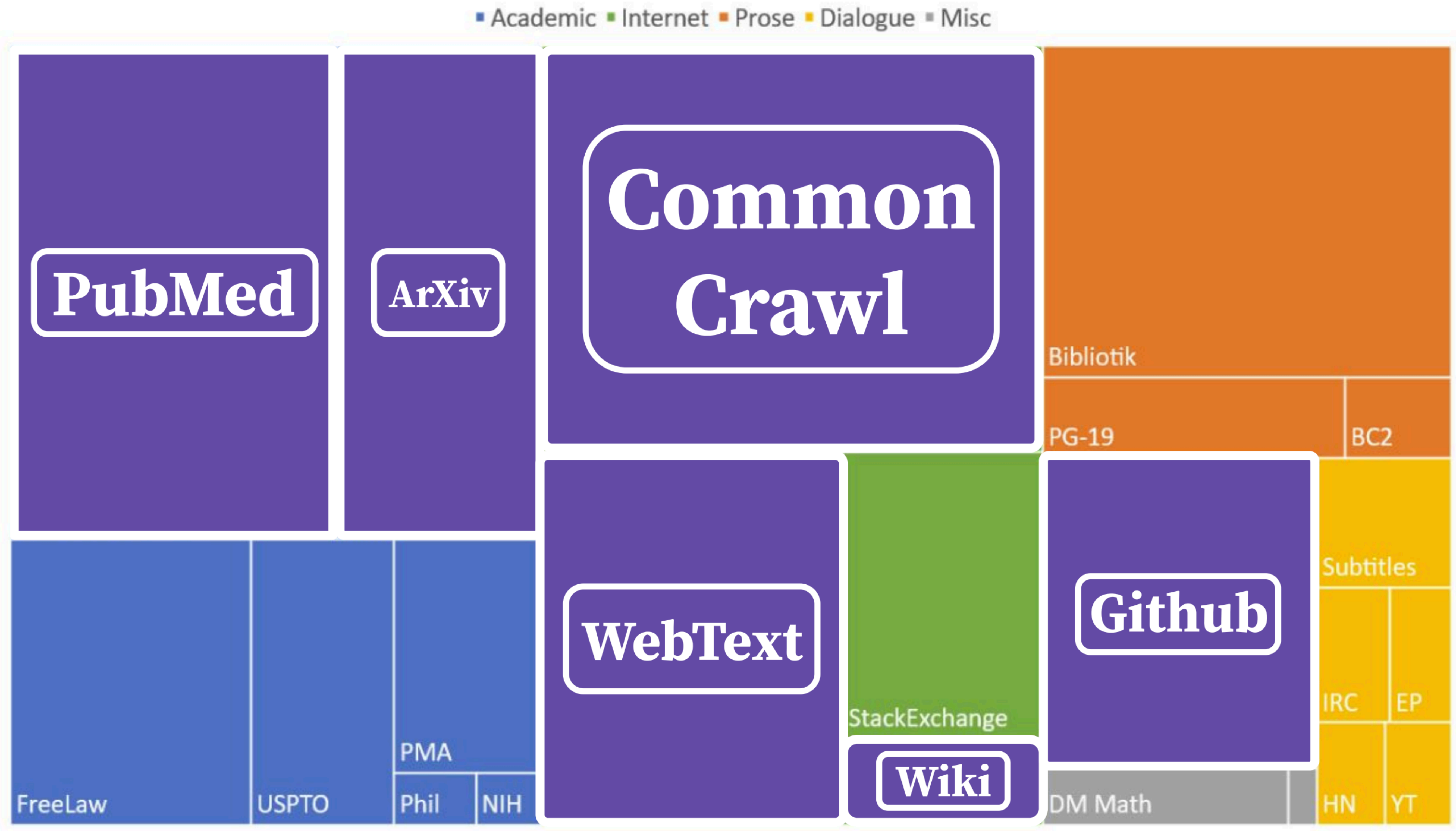
- The corpora used to pre-train language models are huge aggregations of information and data from the internet
- Consider The Pile (Gao et al., 2020): **800GB** total



Graphic from Gao et al., 2020

Motivation

- The corpora used to pre-train language models are huge aggregations of information and data from the internet
- Consider The Pile (Gao et al., 2020): **800GB** total



Graphic from Gao et al., 2020

Today, we take LLMs' ability to "store" knowledge for granted

GPT-3 Zero-shot Knowledge Retrieval

Key Question

Can we directly retrieve the knowledge learned in pre-training from a language model?

Today, we take LLMs' ability to "store" knowledge for granted

GPT-3 Zero-shot Knowledge Retrieval

- This was not so obvious to NLP researchers *three years ago*!
- Instead, **traditional knowledge bases** were often used

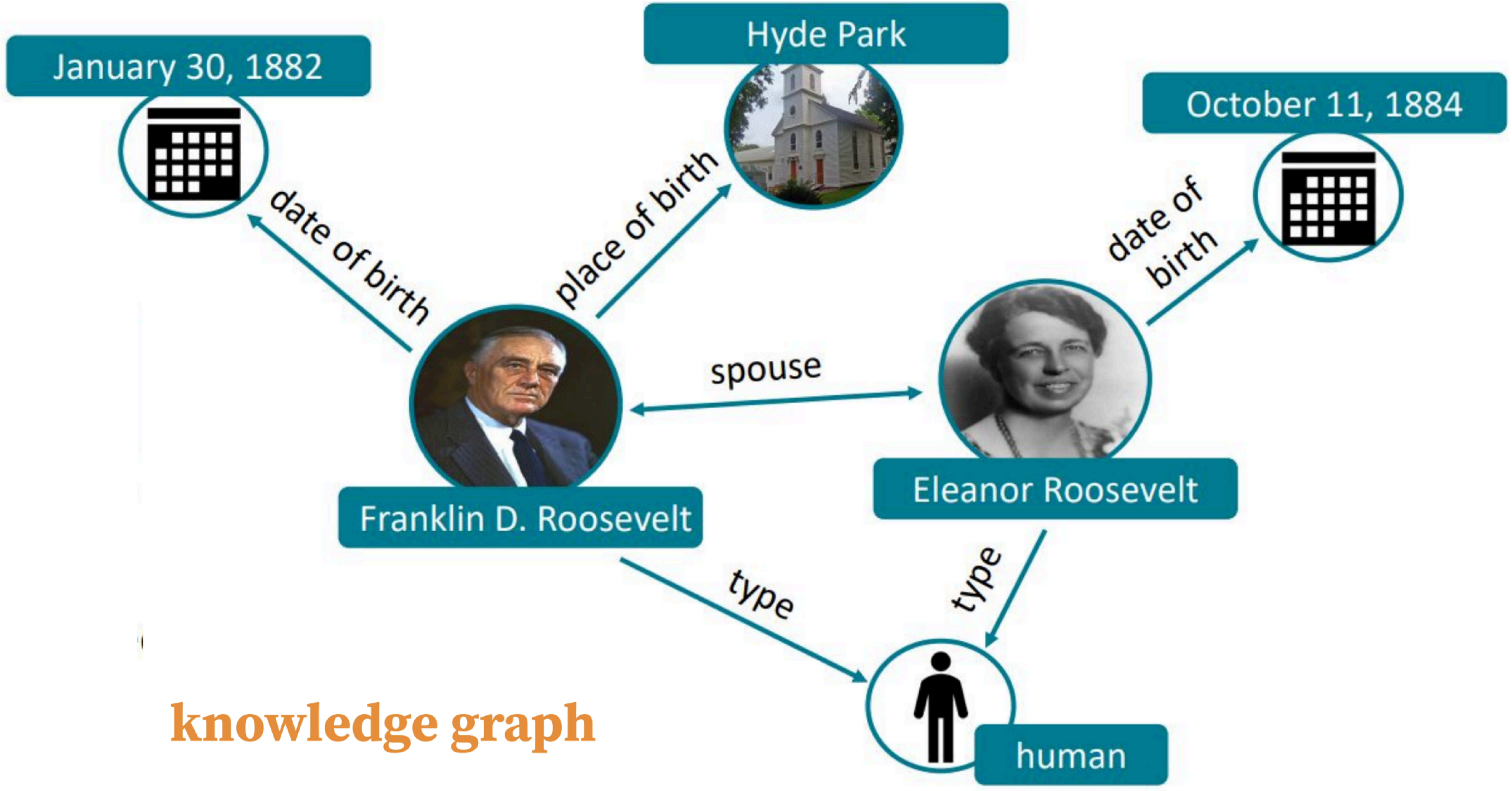
What is a knowledge graph?

A knowledge graph represents structured information about entities and their relationships

What is a knowledge base?

knowledge base

WIKIDATA



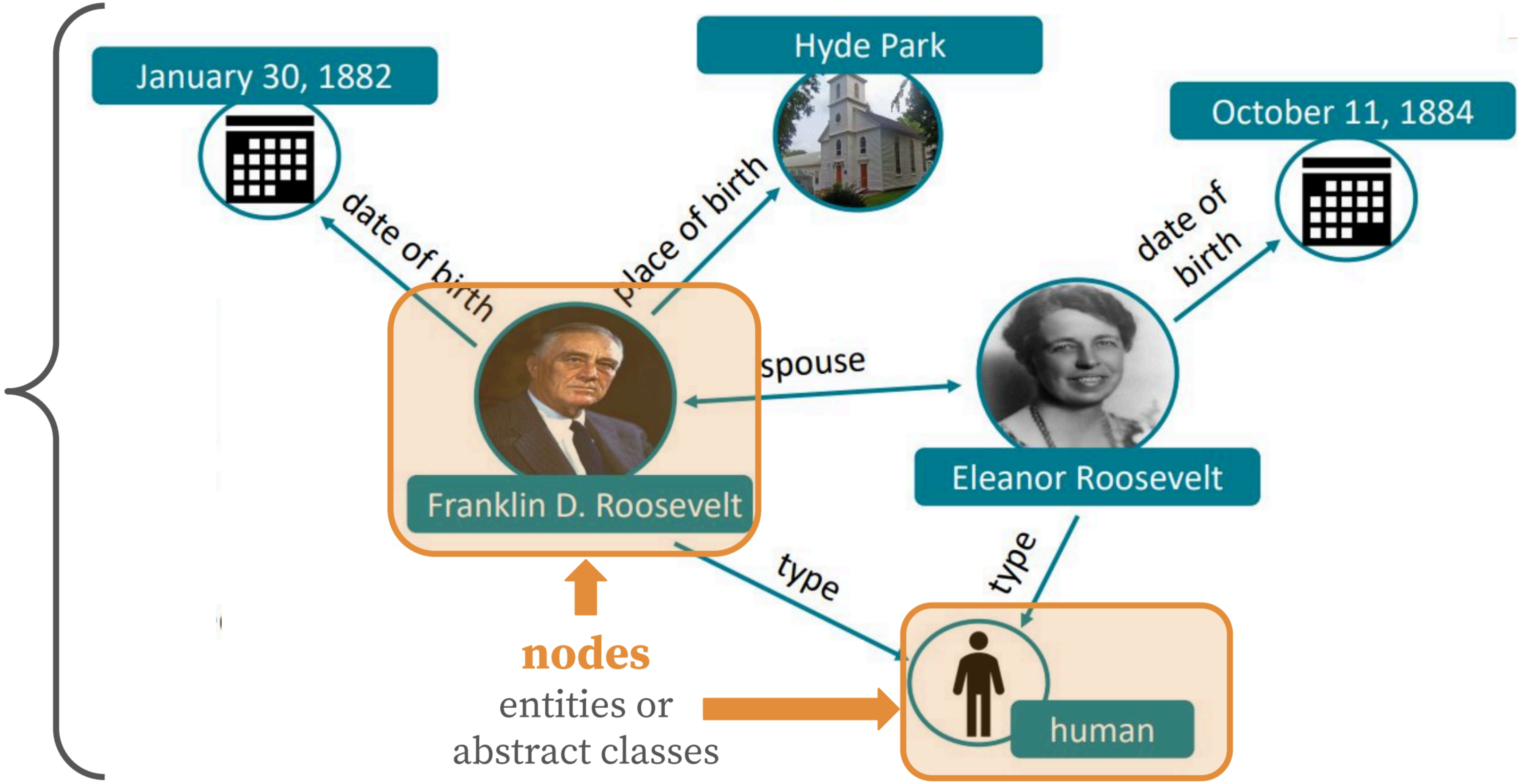
knowledge graph

Graphic from [Megan Leszczynski (Stanford)]

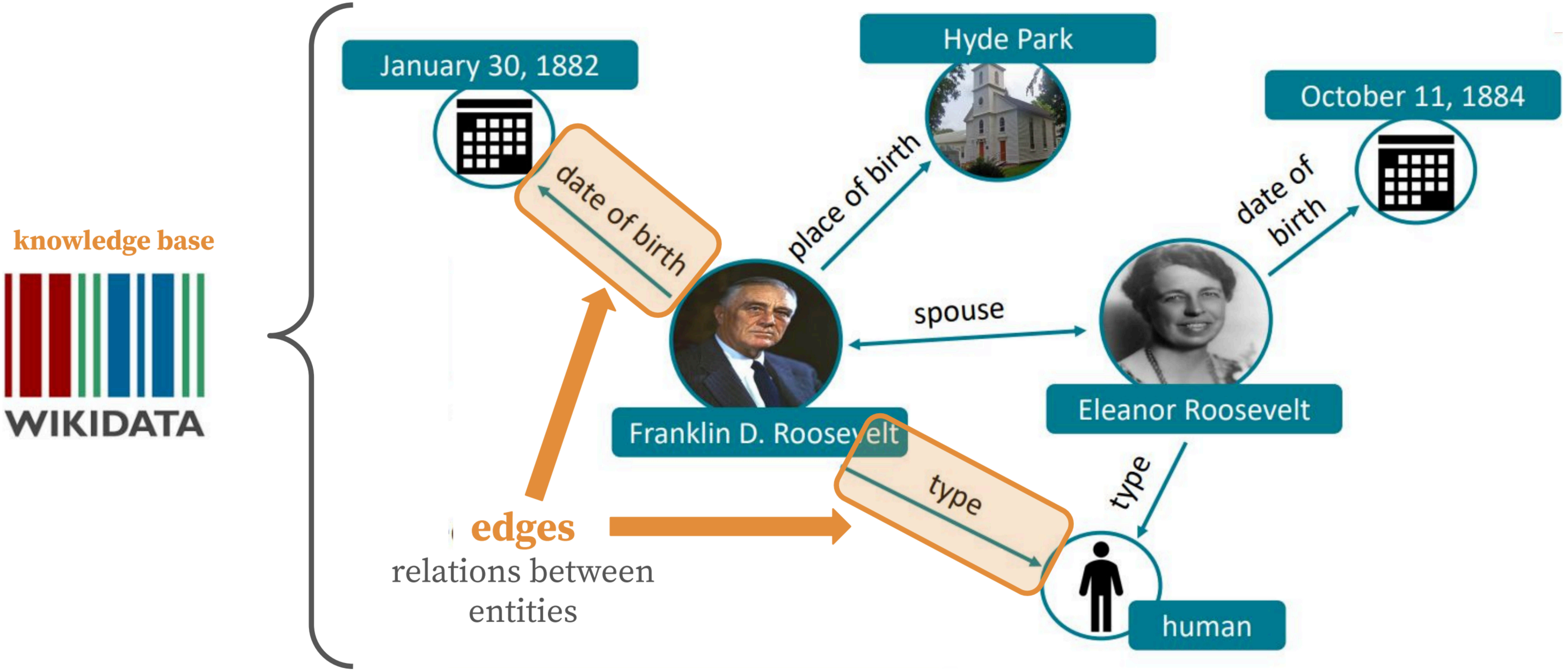
What is a knowledge base?

knowledge base

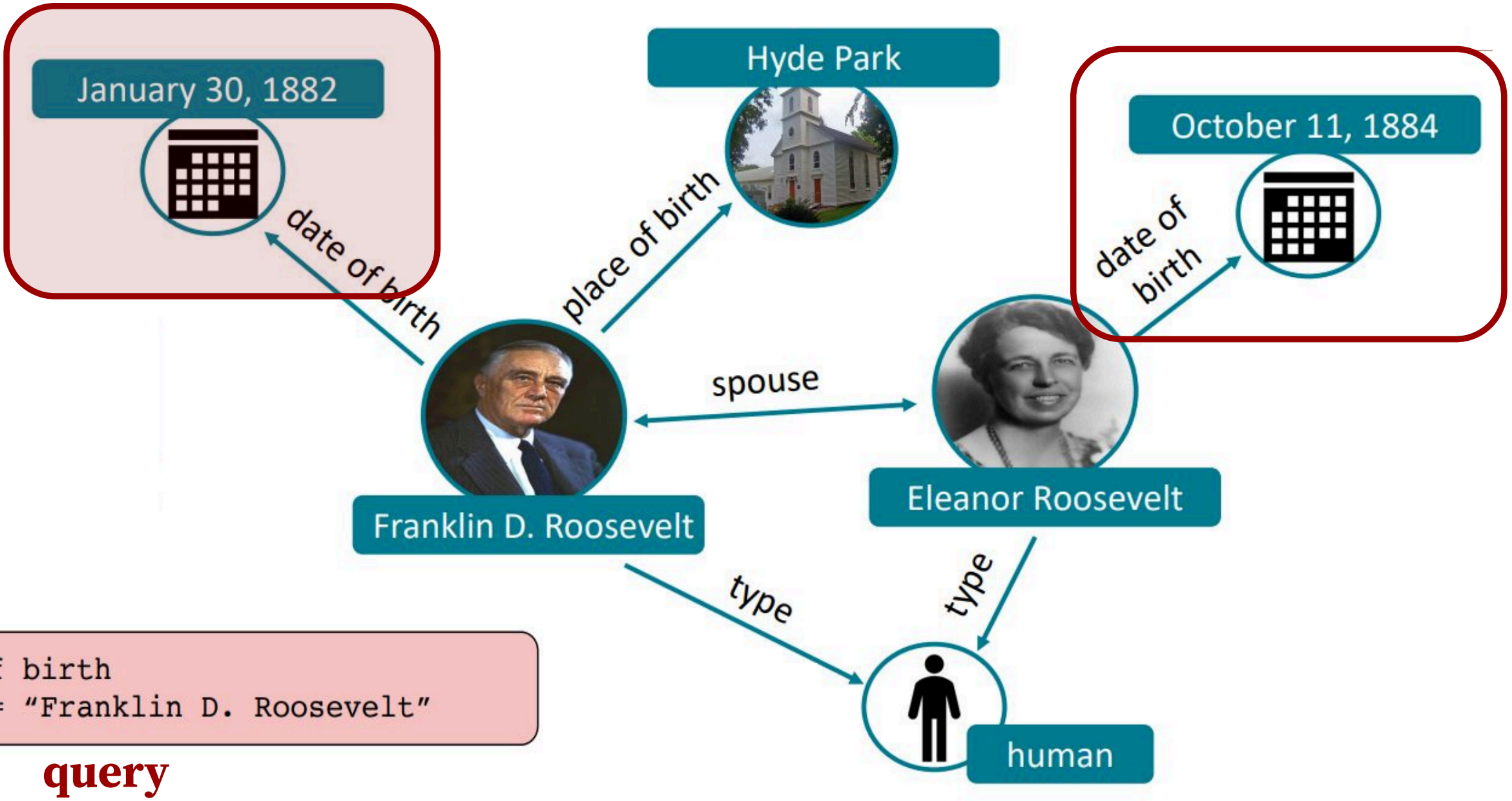
WIKIDATA



What is a knowledge base?



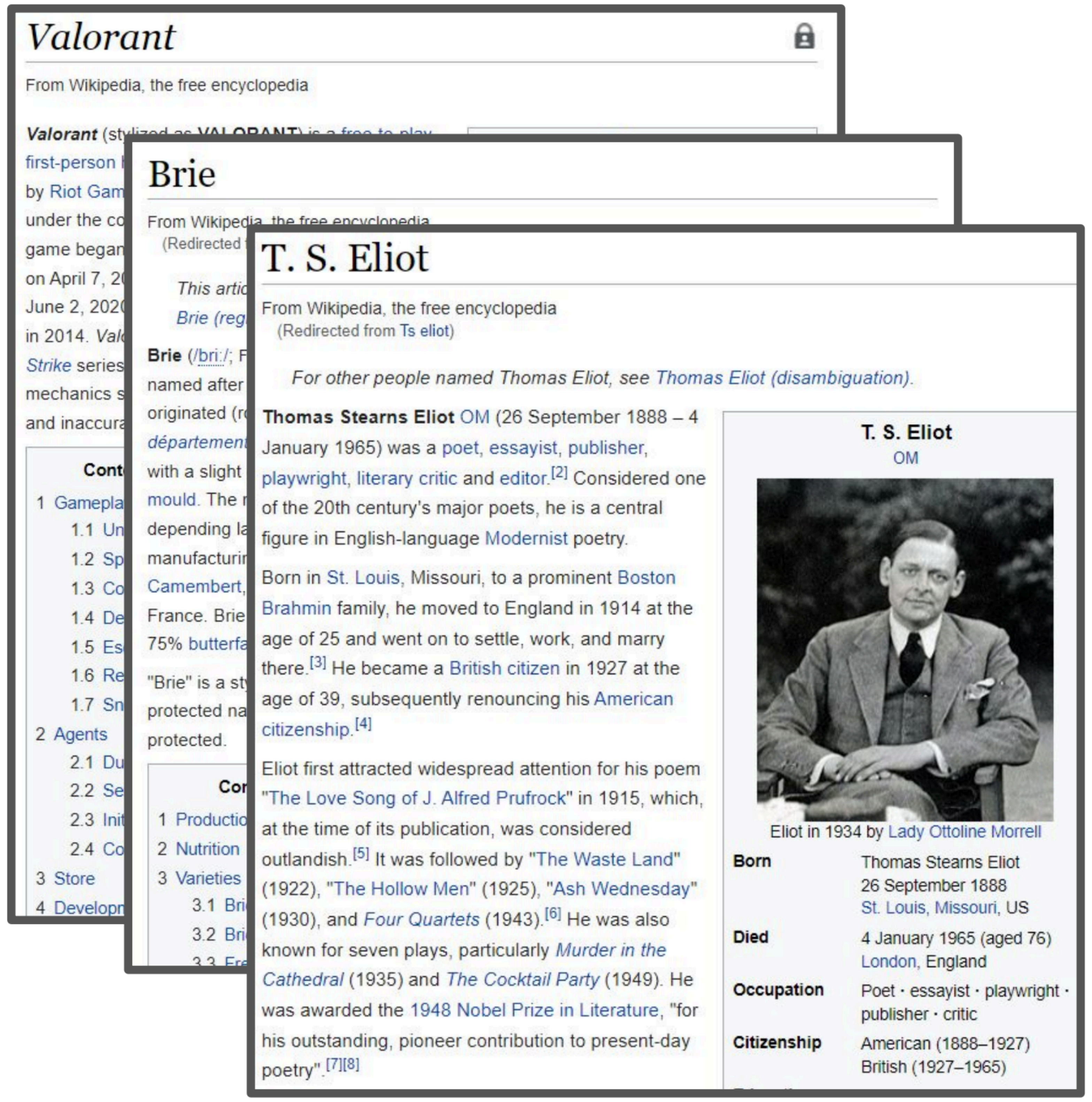
What is a knowledge base?




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

query

Downsides of using knowledge bases


Valorant

From Wikipedia, the free encyclopedia

Brie

From Wikipedia, the free encyclopedia

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

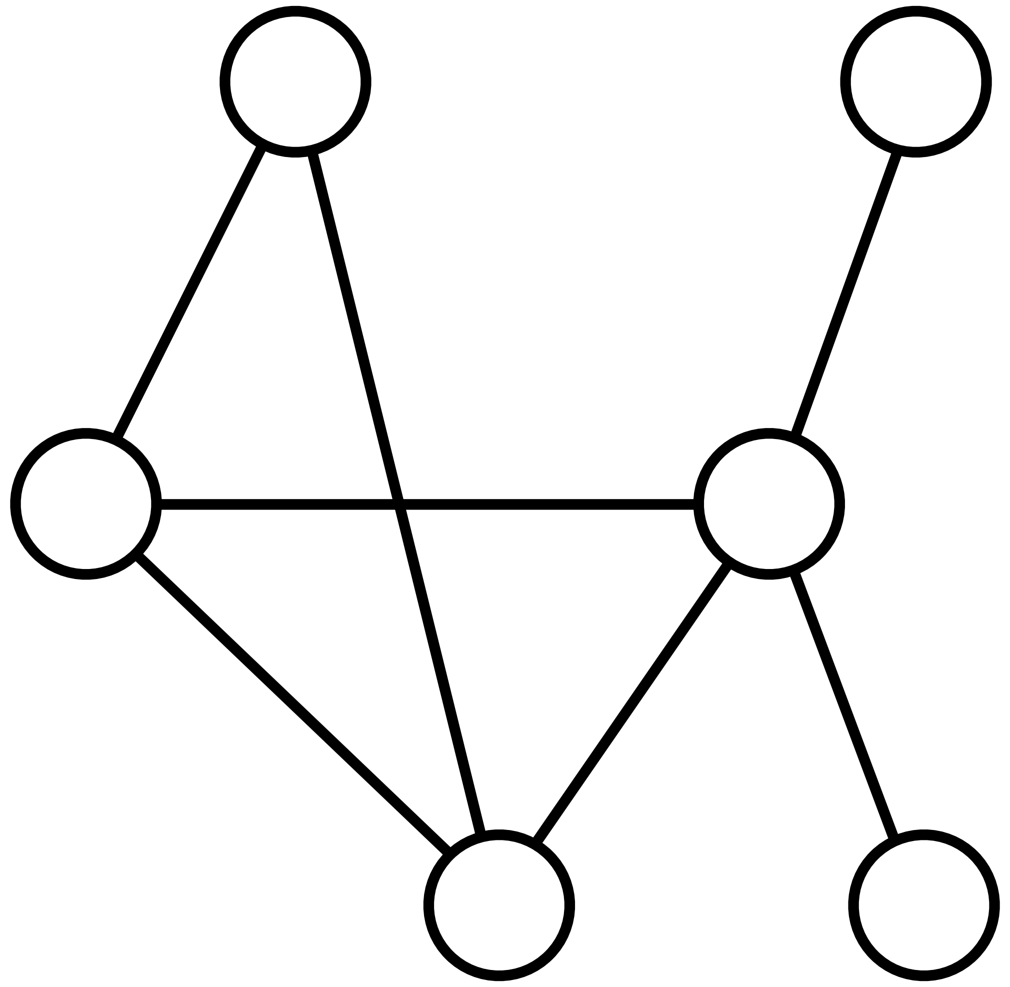
OM

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

- Data preprocessing
- Data merging
- Entity/relation extraction
- Ontology extraction
- ⋮

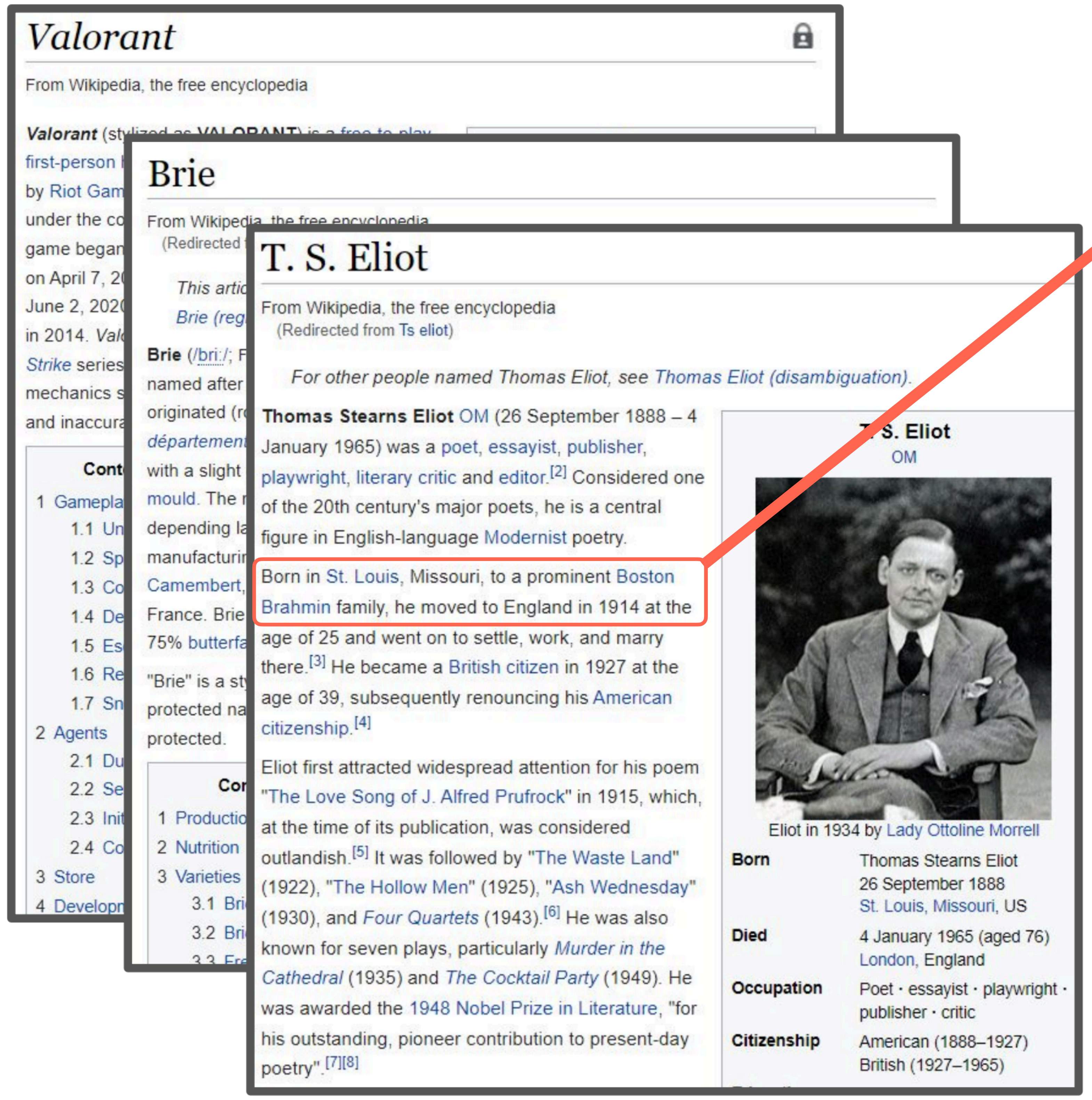


Unstructured text → Knowledge Extraction Pipeline → **Knowledge base**

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

Right-most graphic from Petroni et al., 2019

Downsides of using knowledge bases



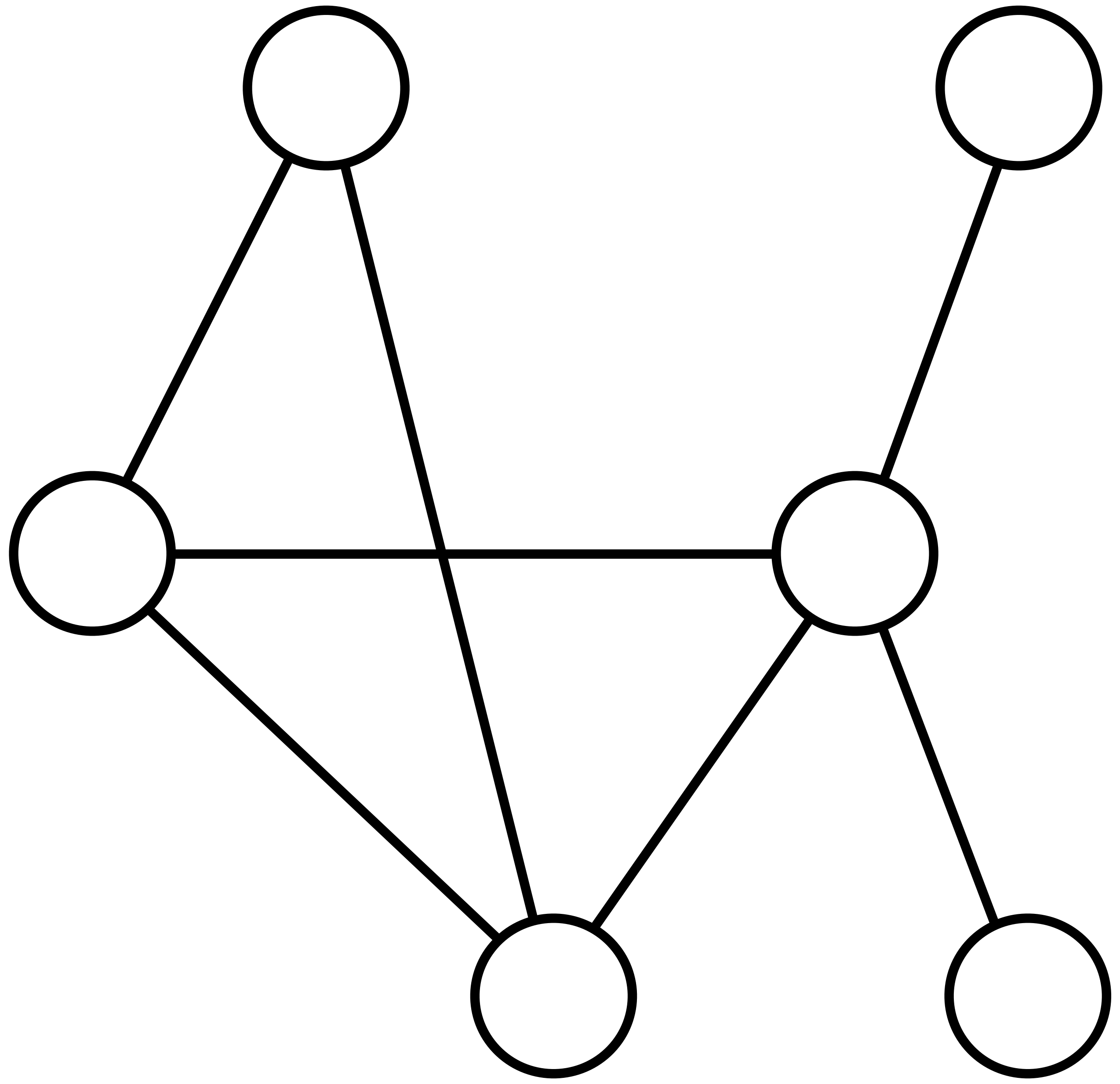
Unstructured text

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

Knowledge Extraction Pipeline

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

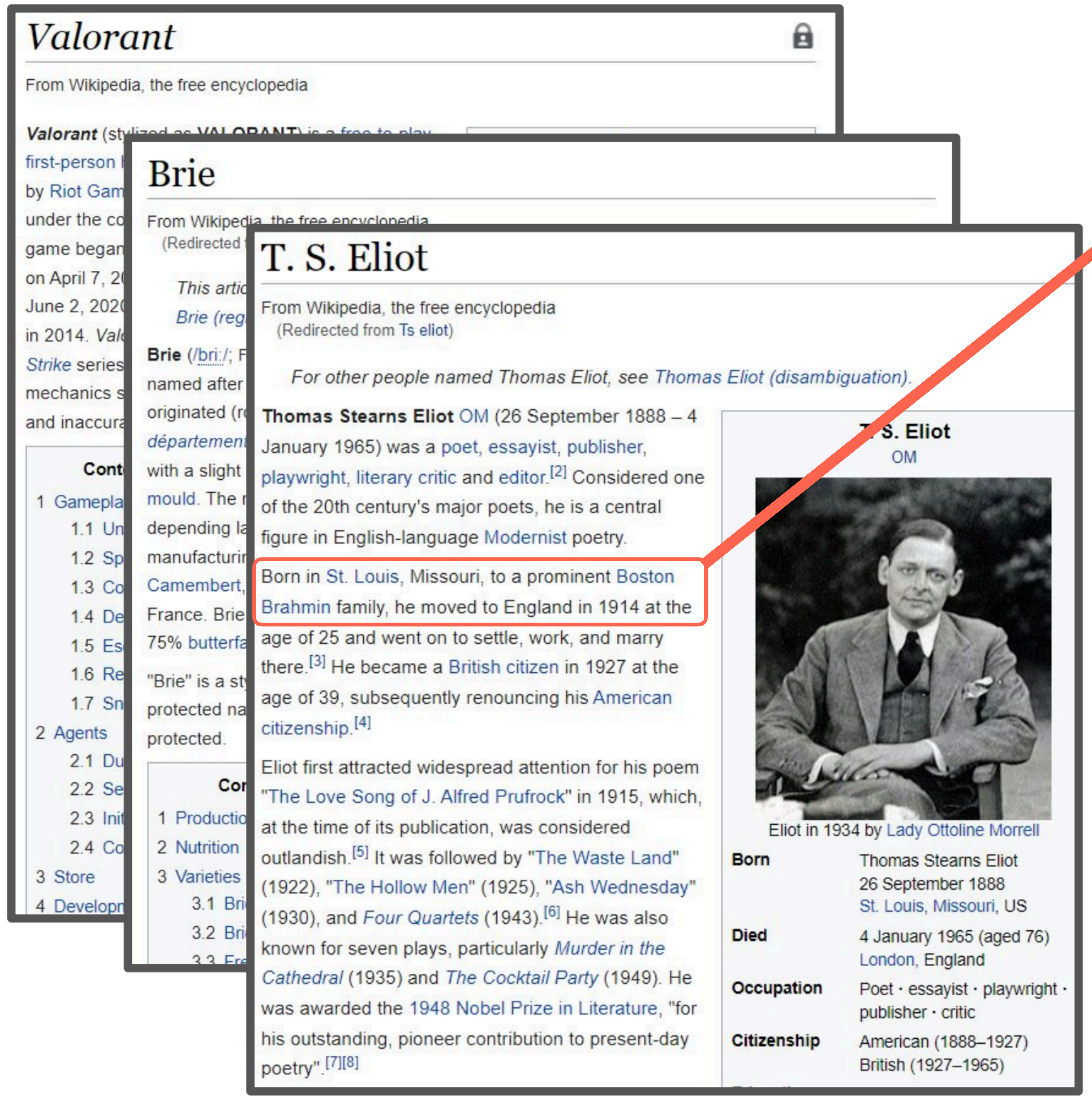
incorrect extraction



Requires **supervised data** to train the pipeline and/or fill the knowledge base

Right-most graphic from Petroni et al., 2019

Downsides of using knowledge bases



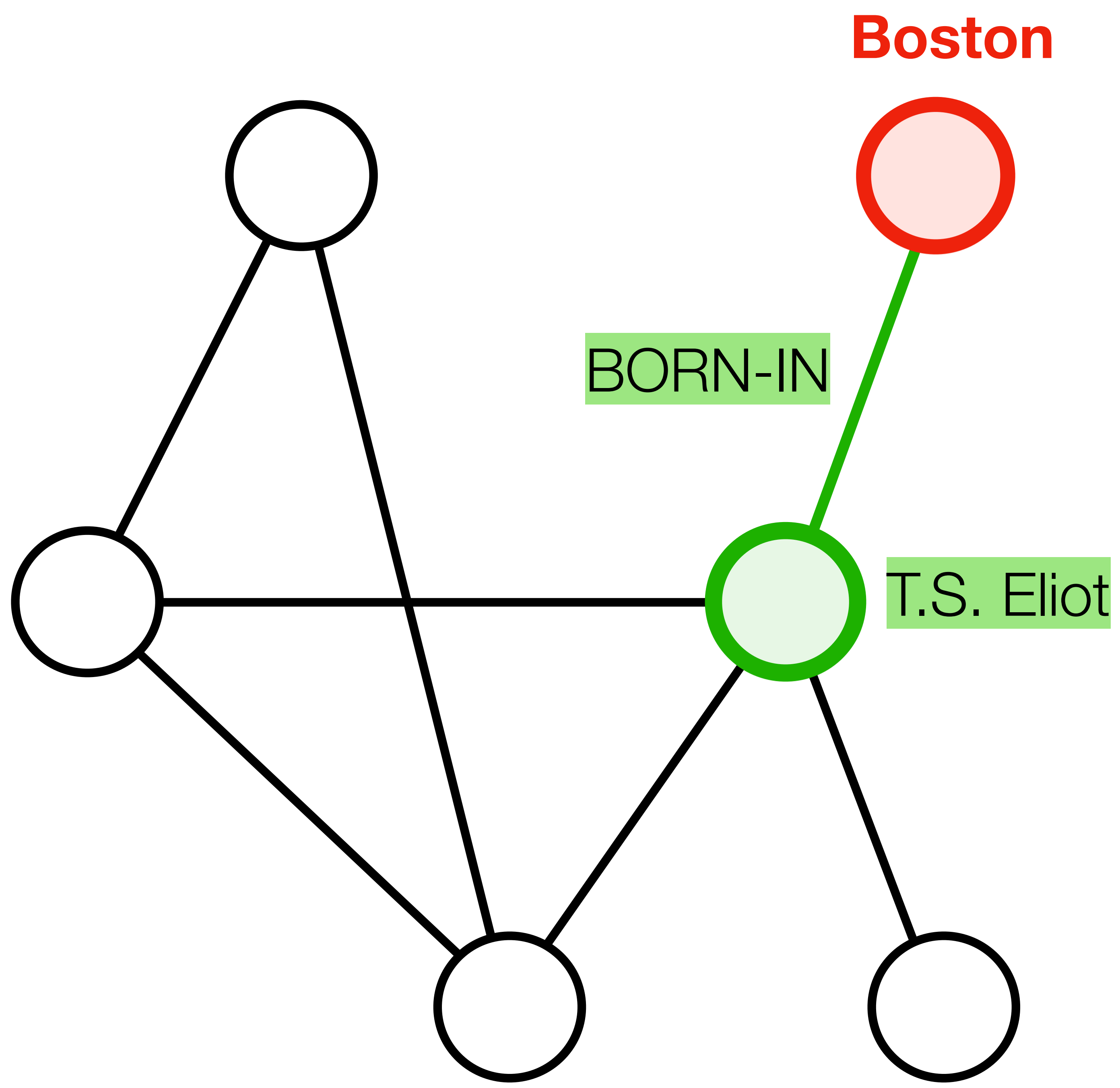
Unstructured text

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

Knowledge Extraction Pipeline

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

incorrect extraction



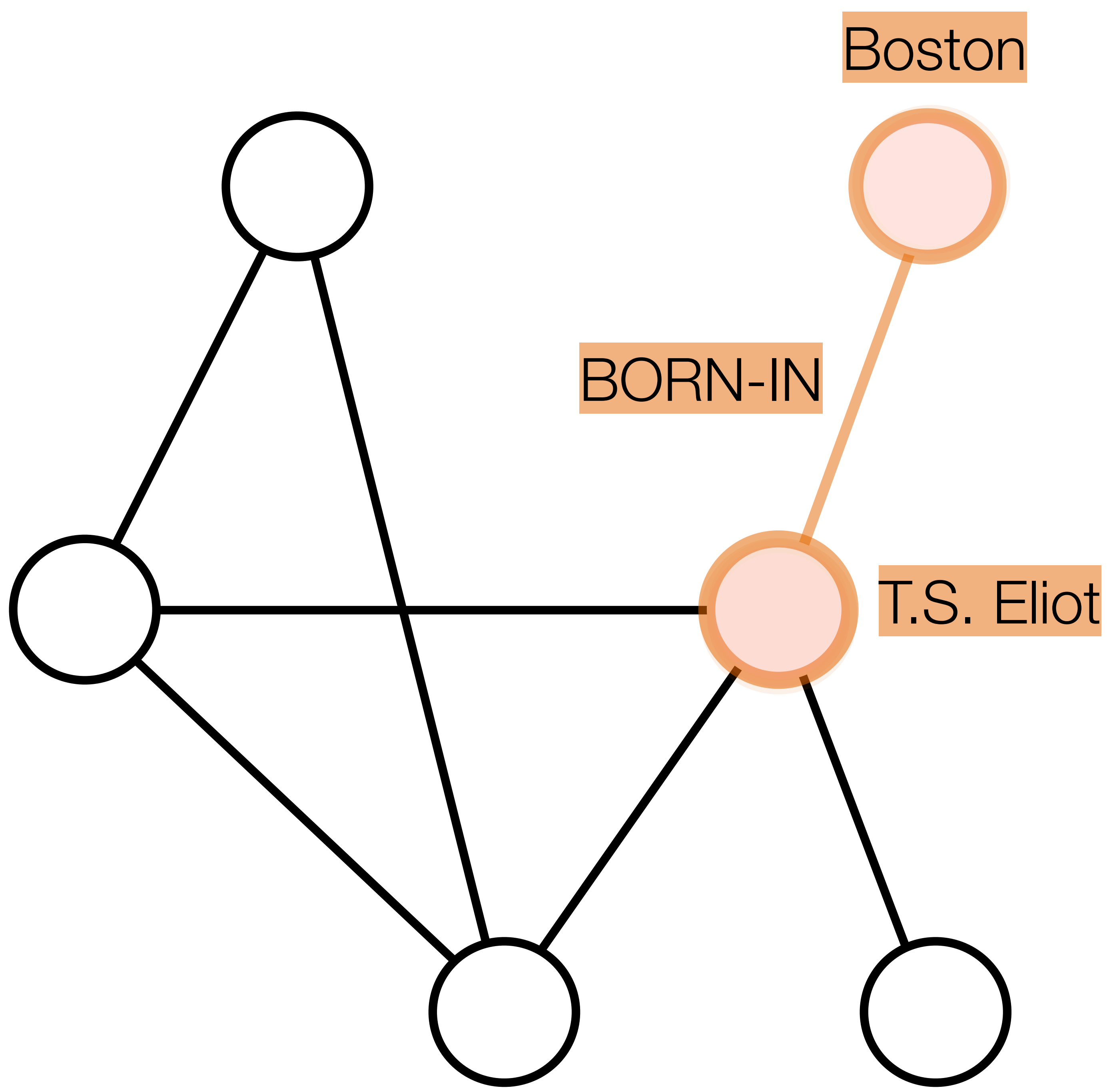
Requires **supervised data** to train the pipeline and/or fill the knowledge base

Right-most graphic from Petroni et al., 2019

Downsides of using knowledge bases

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

slightly incorrect query



ERROR
"BORN IN" is not a legal relation.

Reliant on **fixed schemas** to store or query data

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

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

Are there better alternatives?

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

Language models as knowledge bases?

- **Why language models?**
    - Pretrained on a huge corpus of data
    - Doesn’t require annotations/supervision
    - More flexible with natural language queries
    - Can be used off-the-shelf

Language models as knowledge bases?

- **Why language models?**
  - Pretrained on a huge corpus of data
  - Doesn't require annotations/supervision
  - More flexible with natural language queries
  - Can be used off-the-shelf

But first, we need to see if language models really do store knowledge.

Question

How do we check this?

Language models as knowledge bases?

- **Why language models?**
  - Pretrained on a huge corpus of data
  - Doesn't require annotations/supervision
  - More flexible with natural language queries
  - Can be used off-the-shelf

But first, we need to see if language models really do store knowledge.

Question

How do we check this?

Answer: LAMA Probe

Petroni et al., “Language Models as Knowledge Bases?”, 2019

LAMA Probe: Evaluation of LM via LAMA

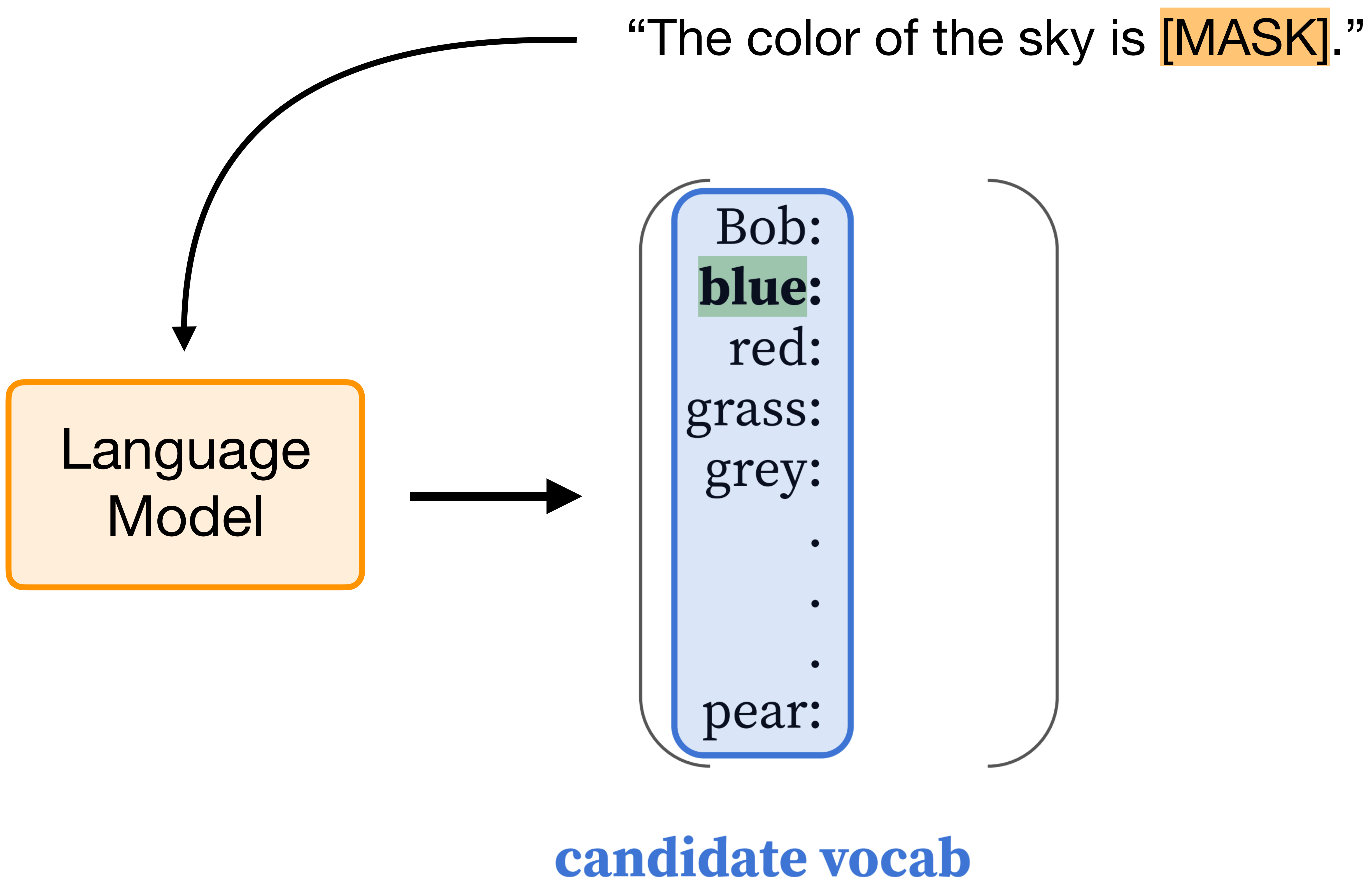
- **Goal**: evaluate factual + commonsense knowledge **in language models**
- Collect set of knowledge sources (i.e. set of facts) and test to see how well the model's knowledge captures these fact
- How do we know how "knowledgeable" a LM is about a particular fact?

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

Petroni et al., "Language Models as Knowledge Bases?", 2019

Evaluation of LM via LAMA

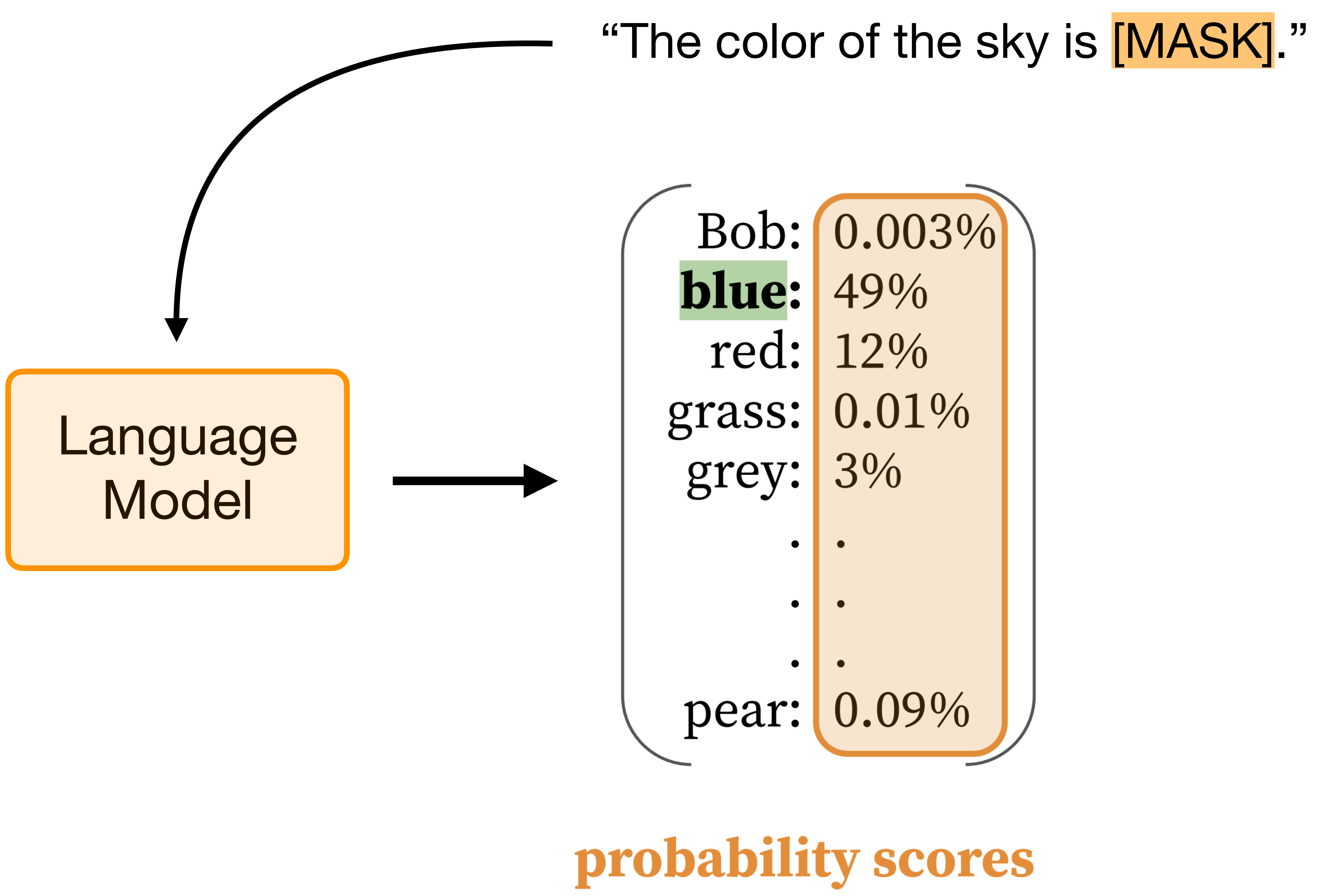
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



*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

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

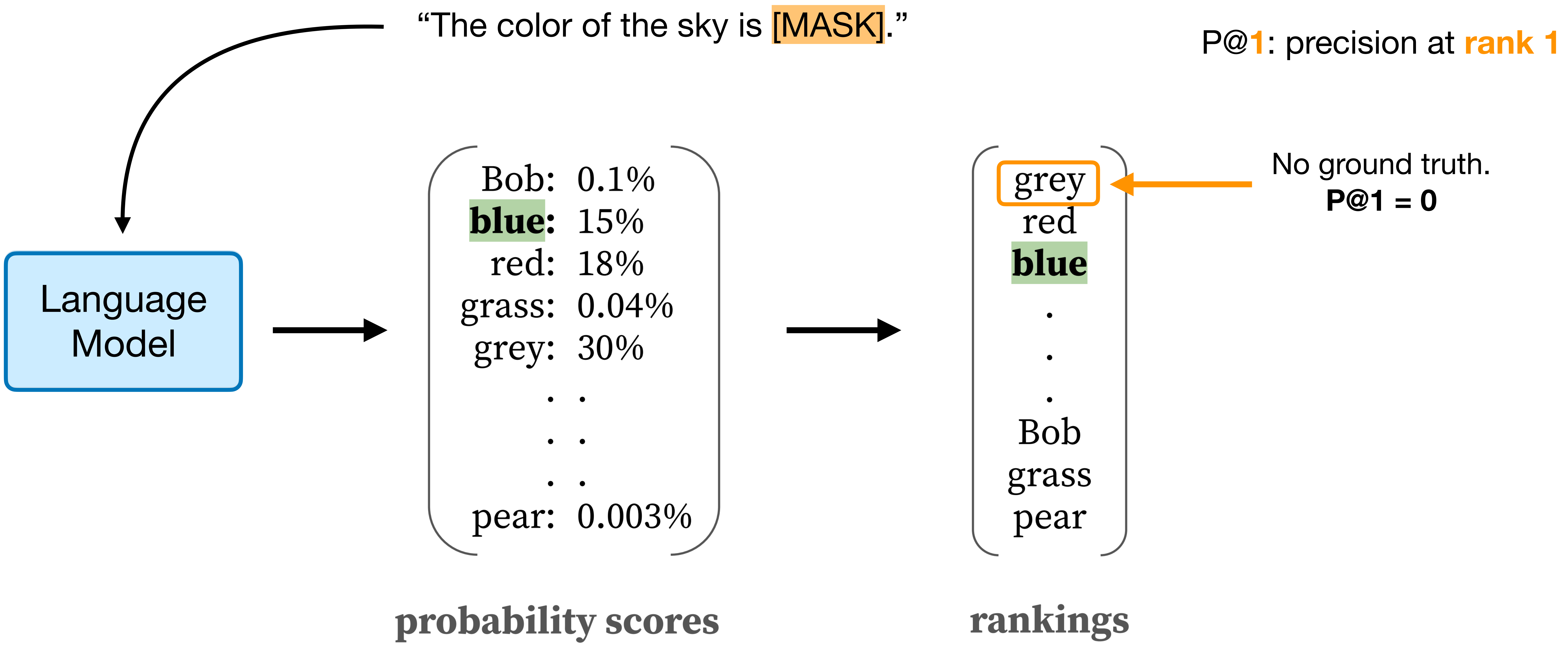
Language Model #2

P@k: precision at k

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

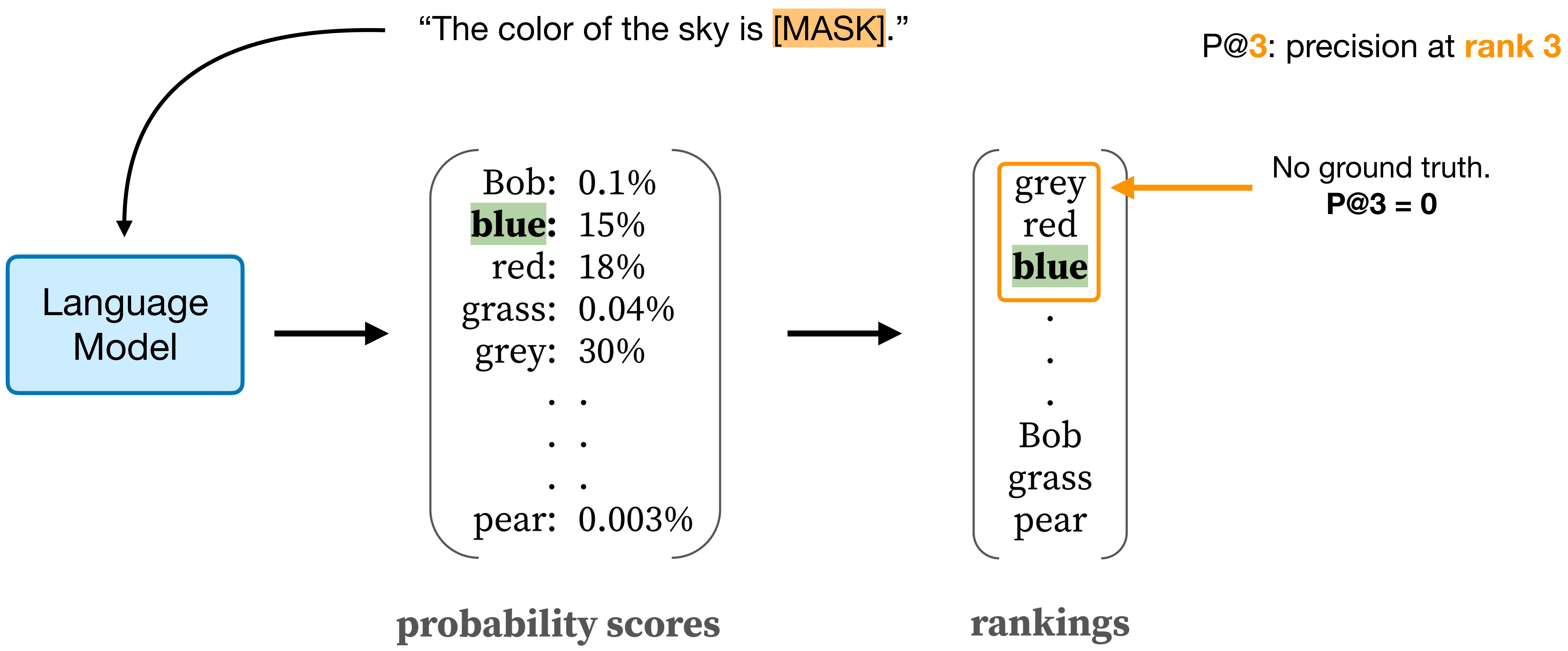
Evaluation of LM via LAMA

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



Evaluation of LM via LAMA

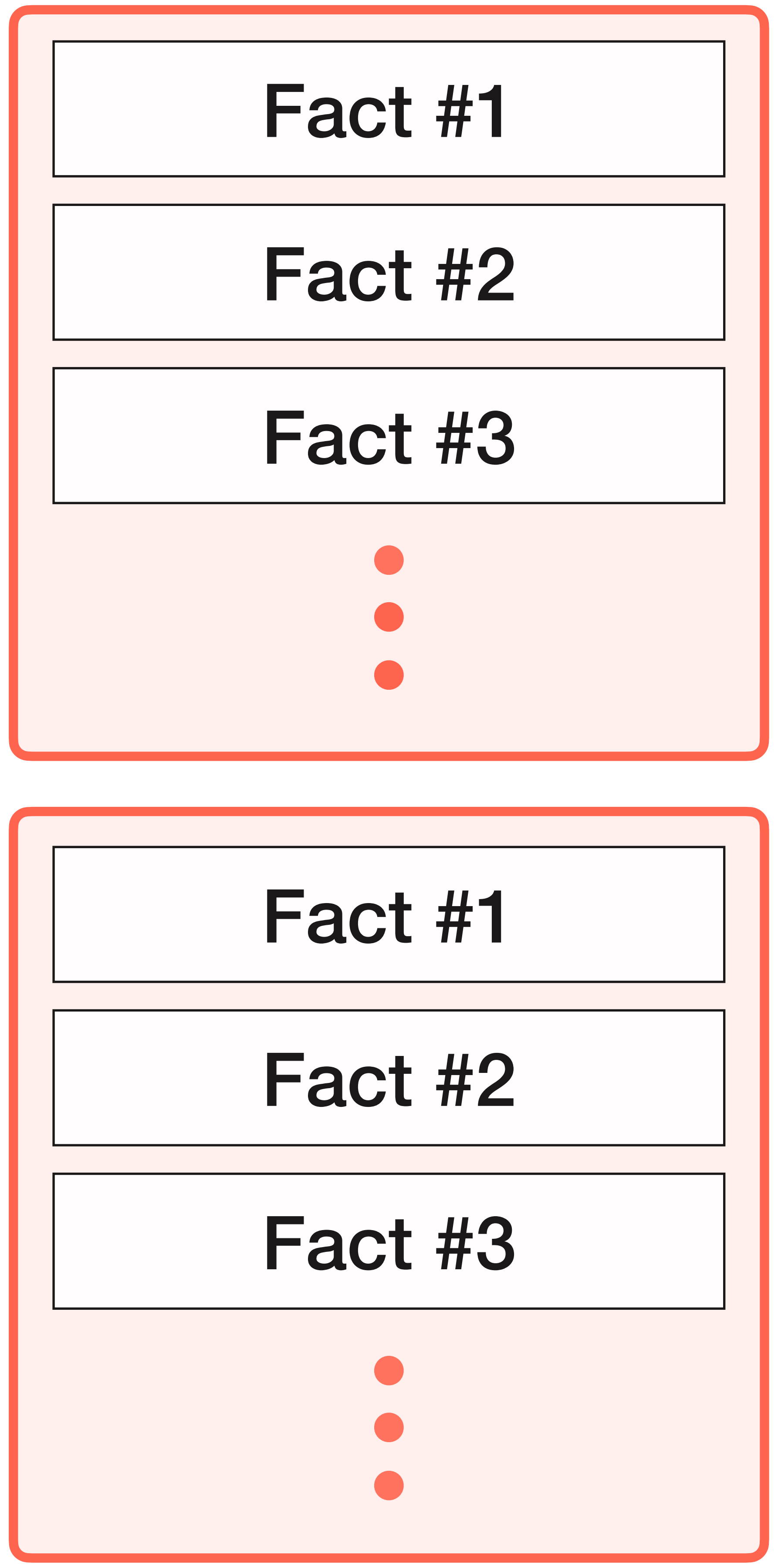
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



Architecture of the LAMA probe

Architecture of the LAMA probe

Step 1: Compile knowledge sources

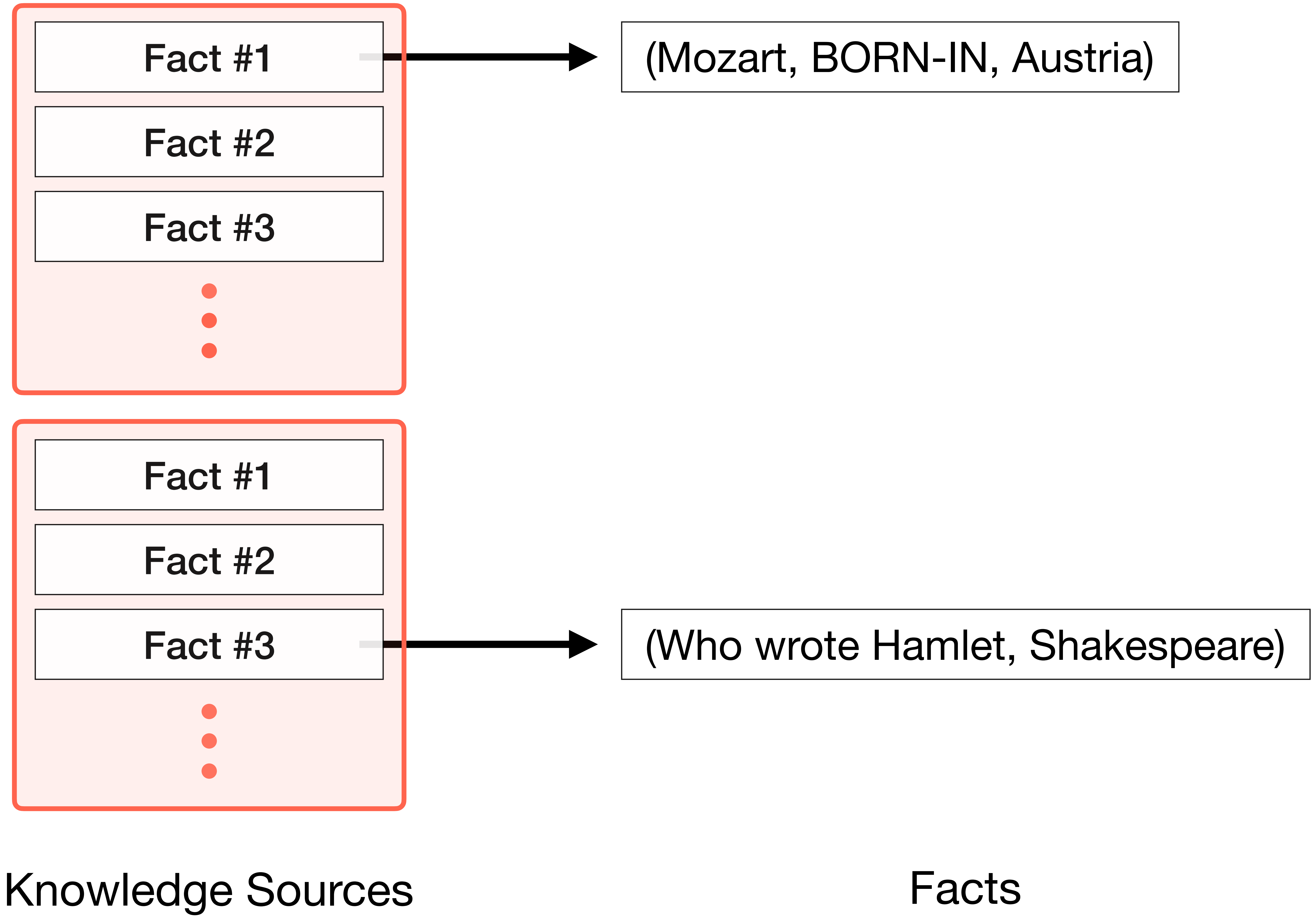


Knowledge Sources

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

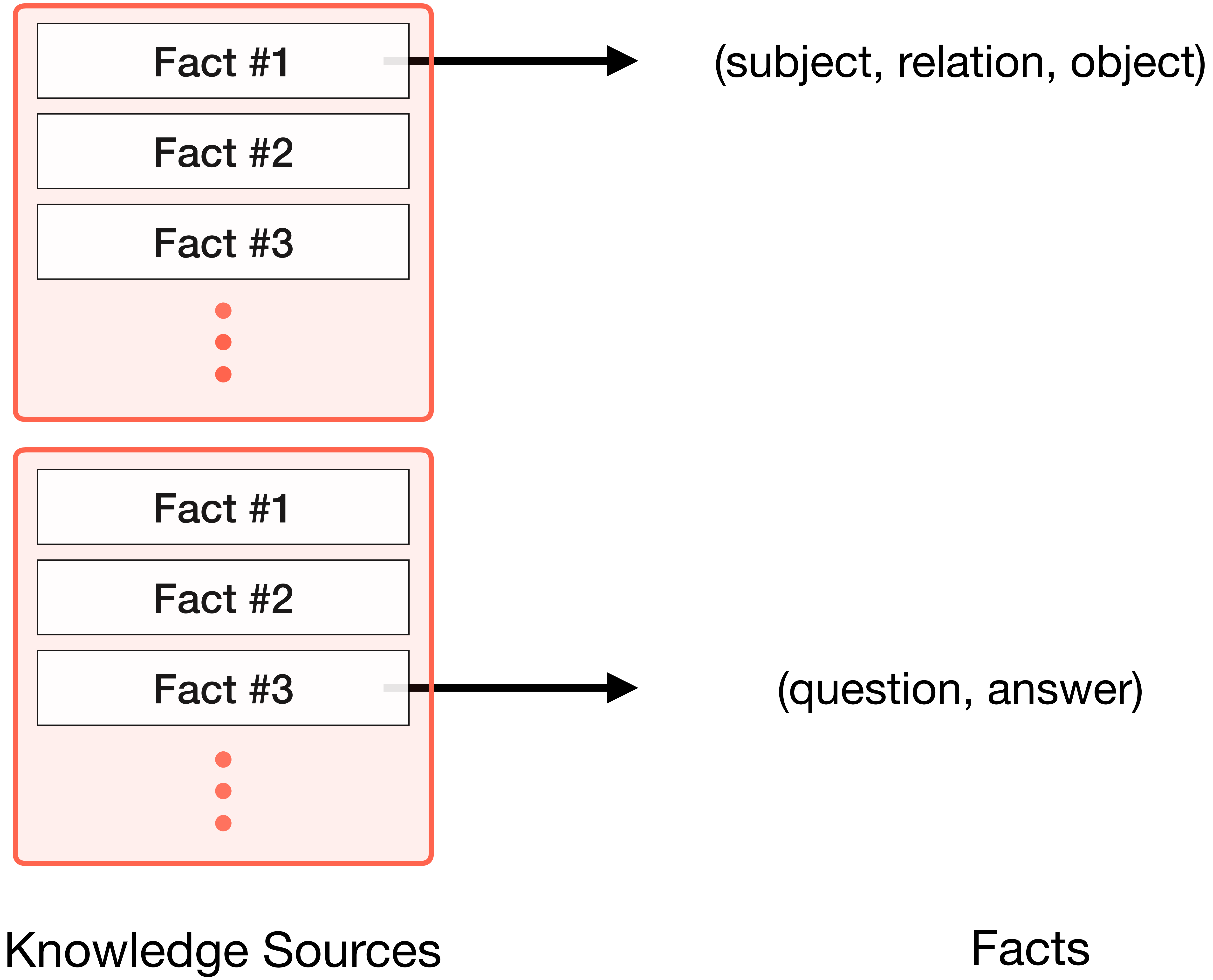
Step 2: Formulate facts into triplets or question-answer pairs



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

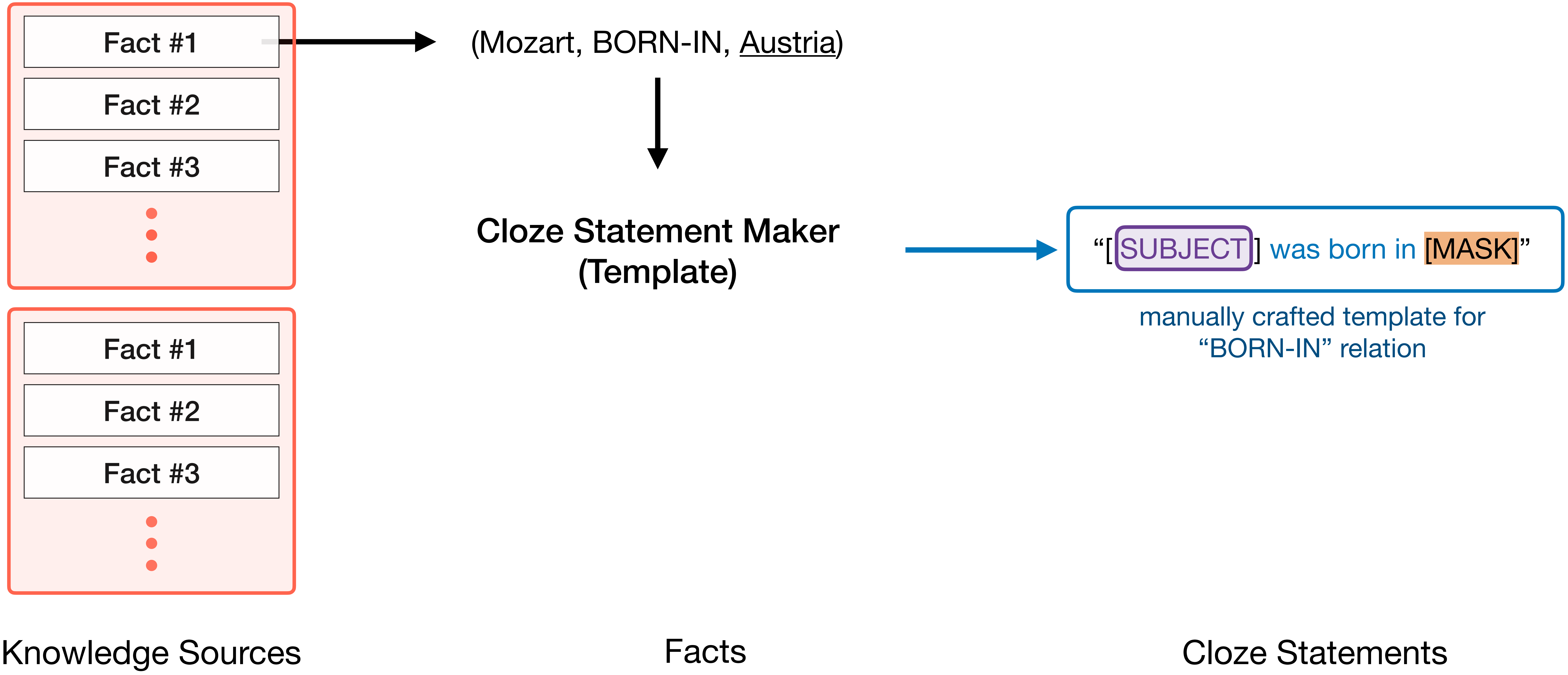
Step 3: Create cloze statements, either manually or via templates



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

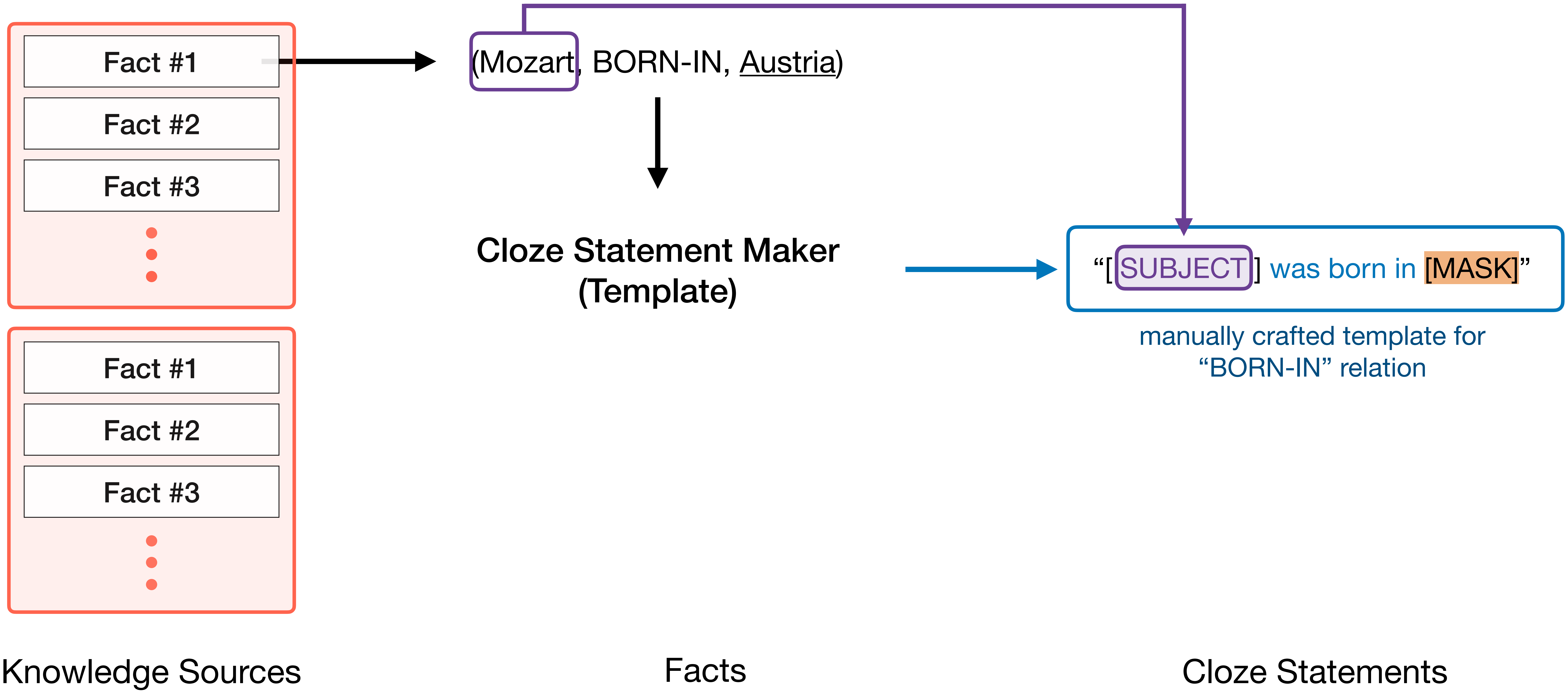
Step 3: Create cloze statements, either manually or via templates



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

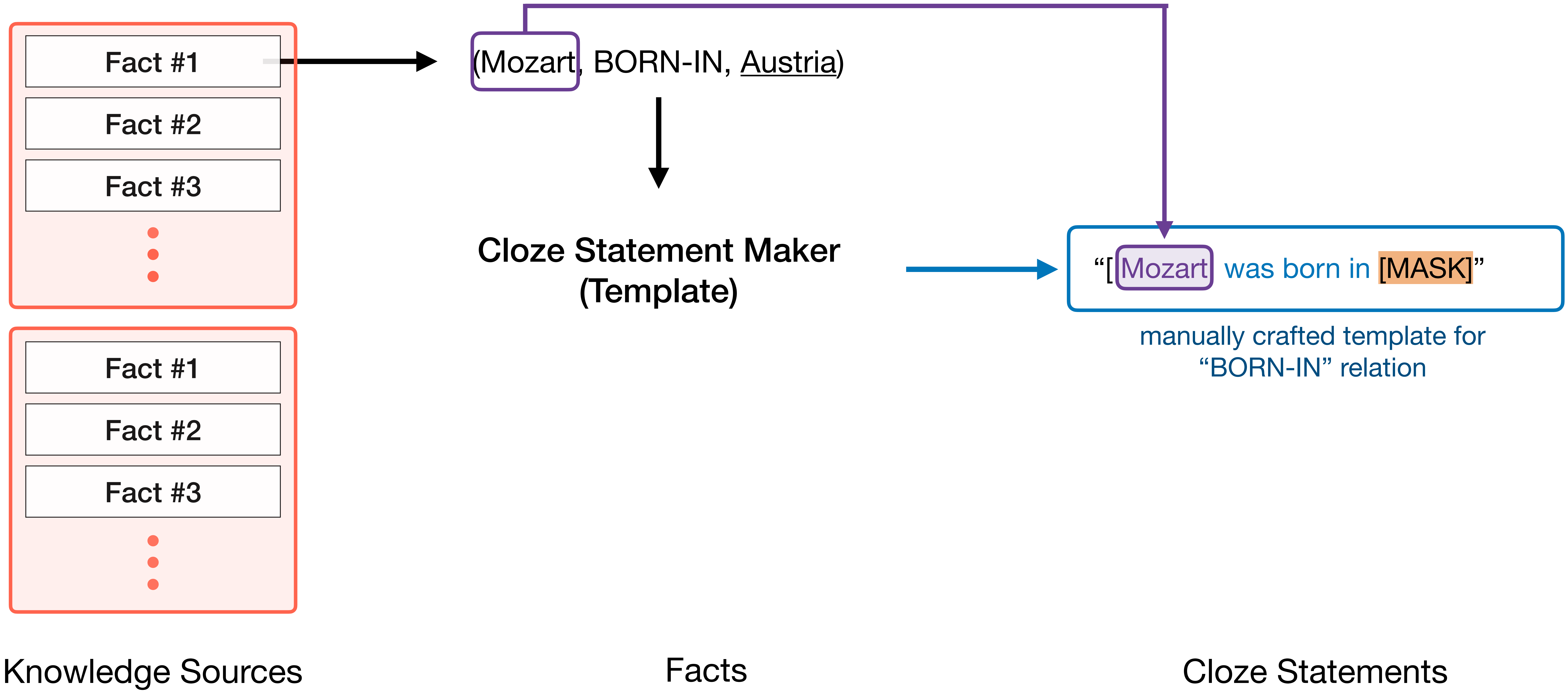
Step 3: Create cloze statements, either manually or via templates



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

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



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Architecture of the LAMA probe

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

Fact #1

Fact #2

Fact #3

Fact #1

Fact #2

Fact #3

Cloze Statement (Manual)

(Who wrote Hamlet, Shakespeare)

“The writer of Hamlet is [MASK]”

human-generated statement

Knowledge Sources

Facts

Cloze Statements

Petroni et al., “Language Models as Knowledge Bases?”, 2019

LAMA's Knowledge Sources: Google-RE

- Manually extracted facts from Wikipedia
- Only consider 3 kinds of relations: place of birth, date of birth, place of death

Original Fact	(T.S. Eliot, birth-place, St. Louis)
Question	" T.S. Eliot was born in [MASK] "
Answer	St. Louis

Petroni et al., "Language Models as Knowledge Bases?", 2019

LAMA's Knowledge Sources: T-REx

- Automatically extracted facts from Wikipedia (may have some errors)
- **For multiple right answers**: throw away all but one

Original Fact (Francesco Conti, born-in, [Florence, Italy])

multiple possibilities

Petroni et al., "Language Models as Knowledge Bases?", 2019

LAMA’s Knowledge Sources: T-REx

- Automatically extracted facts from Wikipedia (may have some errors)
- **For multiple right answers**: throw away all but one

Original Fact	(Francesco Conti, born-in, [**Florence**, ~~Italy~~])

Petroni et al., “Language Models as Knowledge Bases?”, 2019

LAMA's Knowledge Sources: T-REx

- Automatically extracted facts from Wikipedia (may have some errors)
- **For multiple right answers**: throw away all but one

Original Fact (Francesco Conti, born-in, [**Florence**, ~~Italy~~])

Question " Francesco Conti was born in [MASK] "

Answer Florence

Petroni et al., "Language Models as Knowledge Bases?", 2019

LAMA's Knowledge Sources: ConceptNet

- For each ConceptNet triple, find the relevant **Open Mind Common Sense (OMCS)** sentences and mask the object

ConceptNet Triple	(ravens, CapableOf, fly)
Question	" Ravens can [MASK] "
Answer	fly

Petroni et al., "Language Models as Knowledge Bases?", 2019

LAMA’s Knowledge Sources: SQuAD

- **Question-answer dataset**: pick only context-insensitive questions with single-token answers
- Originally created via Wikipedia

SQuAD Question-Answer Pair	(“Who developed the theory of relativity?”, Einstein)
Question	“ The theory of relativity was developed by [MASK] ”
Answer	Einstein

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Dataset Statistics

	# Facts	# of Relations	# Tokens in Answer
Google-RE	5.5k	3	1
T-REx	34k	41	1
ConceptNet	11.4k	16	1
SQuAD	300	-	1

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Dataset Statistics

	# Facts	# of Relations	# Tokens in Answer
Google-RE	5.5k	3	**1**
T-REx	34k	41	**1**
ConceptNet	11.4k	16	**1**
SQuAD	300	-	**1**

Note: all ground truth answers are **single-token**!

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Pre-trained language models

Model		Base Model	Training Corpus	Size
fairseq-fconv (Fs)		ConvNet	WikiText-103 corpus	324M
Transformer-XL large (Txl)		Transformer	WikiText-103 corpus	257M
ELMo	ELMo (Eb)	BiLSTM	Google Billion Word	93.6M
	ELMo 5.5B (E5B)		Wikipedia + WMT 2008-2012	93.6M
BERT	BERT-base (Bb)	Transformer	Wikipedia (en) & BookCorpus	110M
	BERT-large (Bl)			340M

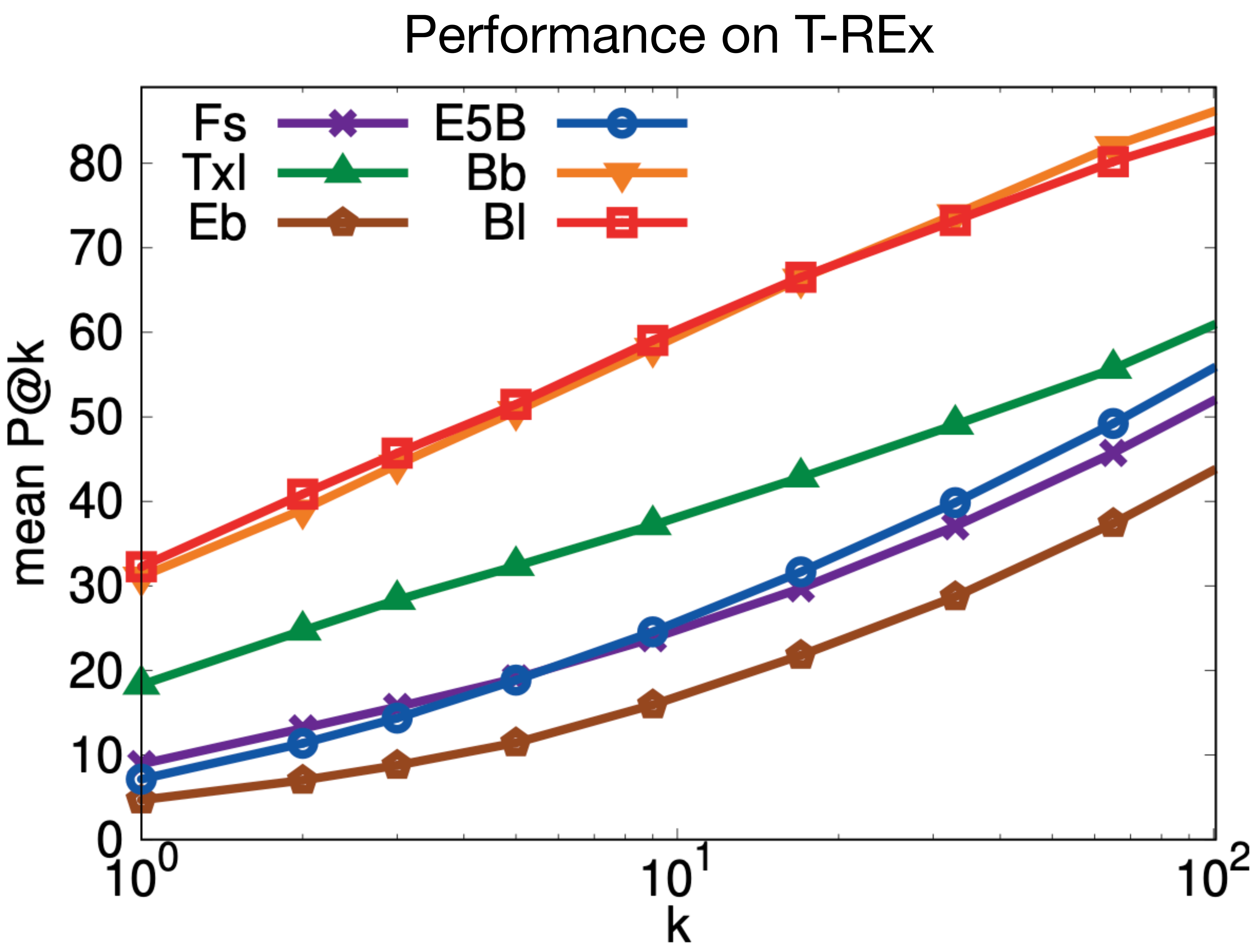
Petroni et al., “Language Models as Knowledge Bases?”, 2019

Corpus	Relation	Statistics #Facts	#Rel	Baselines Freq	DrQA	KB RE_n	RE_o	LM Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	`death-place`	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

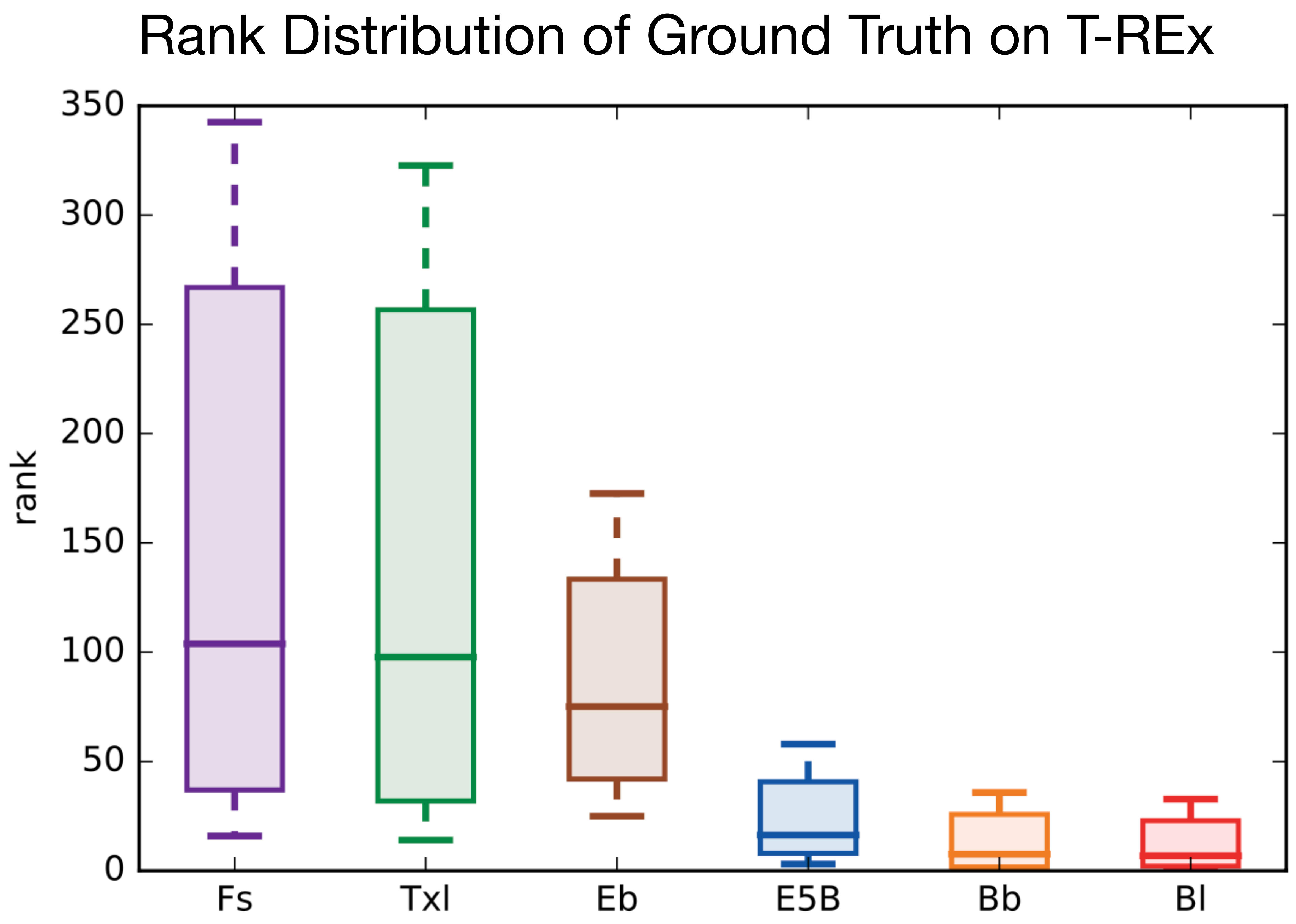
- **RE** (Sorokin and Gurevych, 2017): extracts relation triples from sentence
  - RE_n : uses exact string matching for entity linking & has to find the subject/object entities itself
  - RE_o : uses oracle for entity linking, thus it gets the answer for free

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Results: BERT models outperform other LMs on T-REx



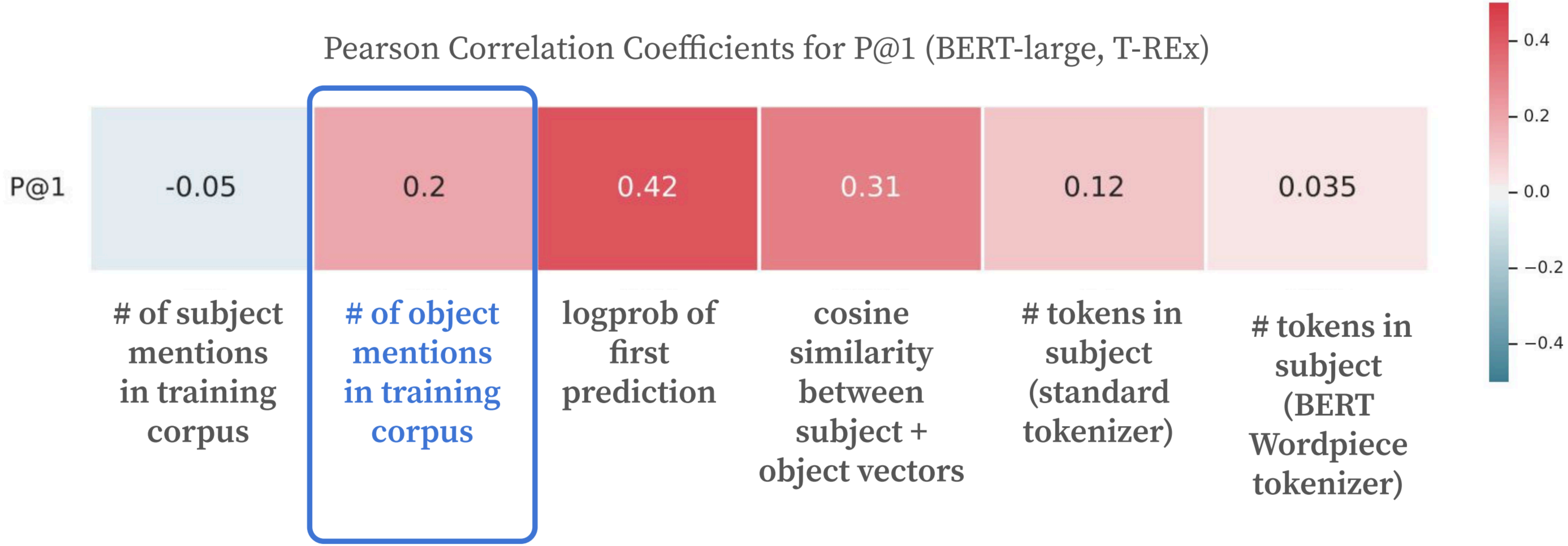
BERT models perform the best by a large margin



BERT models show much lower variance

Petroni et al., “Language Models as Knowledge Bases?”, 2019

Results: What factors correlate with better performance for BERT on T-REx?



Petroni et al., “Language Models as Knowledge Bases?”, 2019

Conclusion

- **BERT-large recalls knowledge better than its competitors**, and competitively with non-neural/supervised alternatives
- **BERT-large is competitive with a RE knowledge base** that was trained on the “best possible” data and used the entity-linking oracle
- Dealing with variance in performance in response to different natural language templates is a challenge

How much does train-test overlap affect performance?

- Many of the knowledge sources we've discussed were extracted from **Wikipedia**
- 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

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

Lewis et al., "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets", ACL 2021.

Train-test overlap is responsible for LM's ability to do knowledge retrieval

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!

Lewis et al., "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets", ACL 2021.

Example: LLMs encode clinical knowledge

- Model: Instruction-tuned variant, Flan-PaLM2
- Dataset: MultiMedQA
  - MultiMedQA multiple-choice dataset including MedQA3, MedMCQA4, PubMedQA5, and Measuring Massive Multitask Language Understanding (MMLU) clinical topics
- Using a combination of prompting strategies, Flan-PaLM achieves SOTA accuracy on every MultiMedQA multiple-choice dataset
  - including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%.

Signal et al., “Large language models encode clinical knowledge”, Nature 2023

ML Objective and Factuality

- Factuality of LLMs: The extent to which the information or responses given by the model correspond with real-world realities and facts
- ML objective used to train LLMs provides no guarantees as to whether a model will learn a fact or not.
  - makes it difficult to ensure whether the model obtains specific knowledge over the course of pre-training
  - and prevents us from explicitly updating or removing knowledge from a pre-trained model.

Roberts et al., “How Much Knowledge Can You Pack into the Parameters of a Language Model?”, 2020

Limitations of LLMs about Knowledge

- Knowledge cutoff
  - LLMs don’t know about any events that happened after their training
  - LLMs don’t have any knowledge about private or confidential information that they have not encountered during training
- Hallucinations
  - LLMs are trained to generate realistic-sounding or convincing text
  - But the generated text may be nonetheless wrong
    - instead of admitting that it lacks the base facts in its training.

How to Resolve Knowledge Cutoff?

- Augmenting LLMs with external resources
  - Retrieval Augmented Generation (RAG): supplement the LLM’s internal knowledge by external information
    - retrieving facts from an external documents or a knowledge base to ground LLMs on the most accurate, up-to-date information and to give as context into LLMs' generative process.
- Knowledge editing
  - Factuality purpose: Can knowledge be edited?
  - Privacy purpose: Can sensitive or private information be deleted from LLMs (unlearning)?

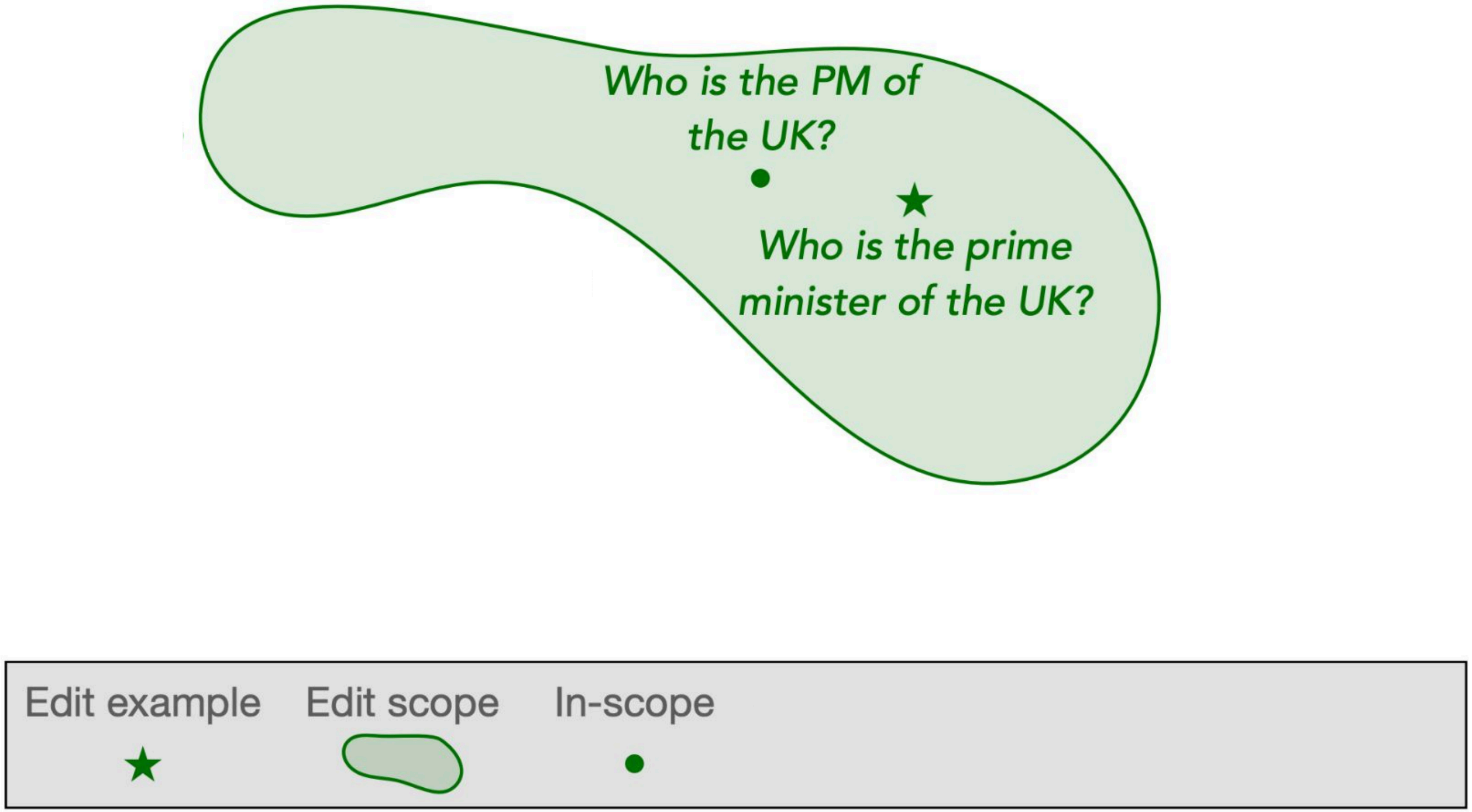
How to update knowledge in pre-trained models?

Edit What, Exactly?

Defining the problem

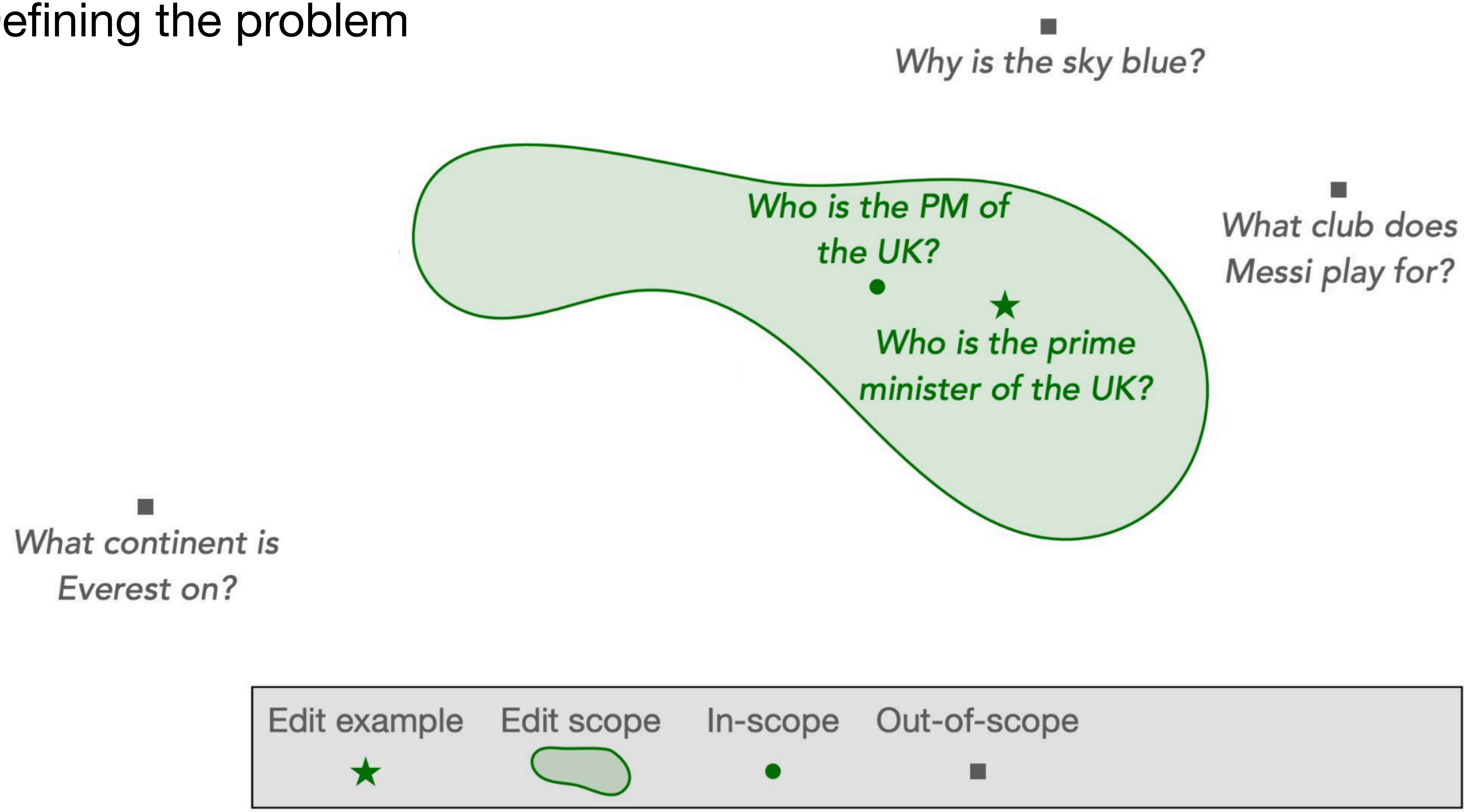
Edit What, Exactly?

Defining the problem



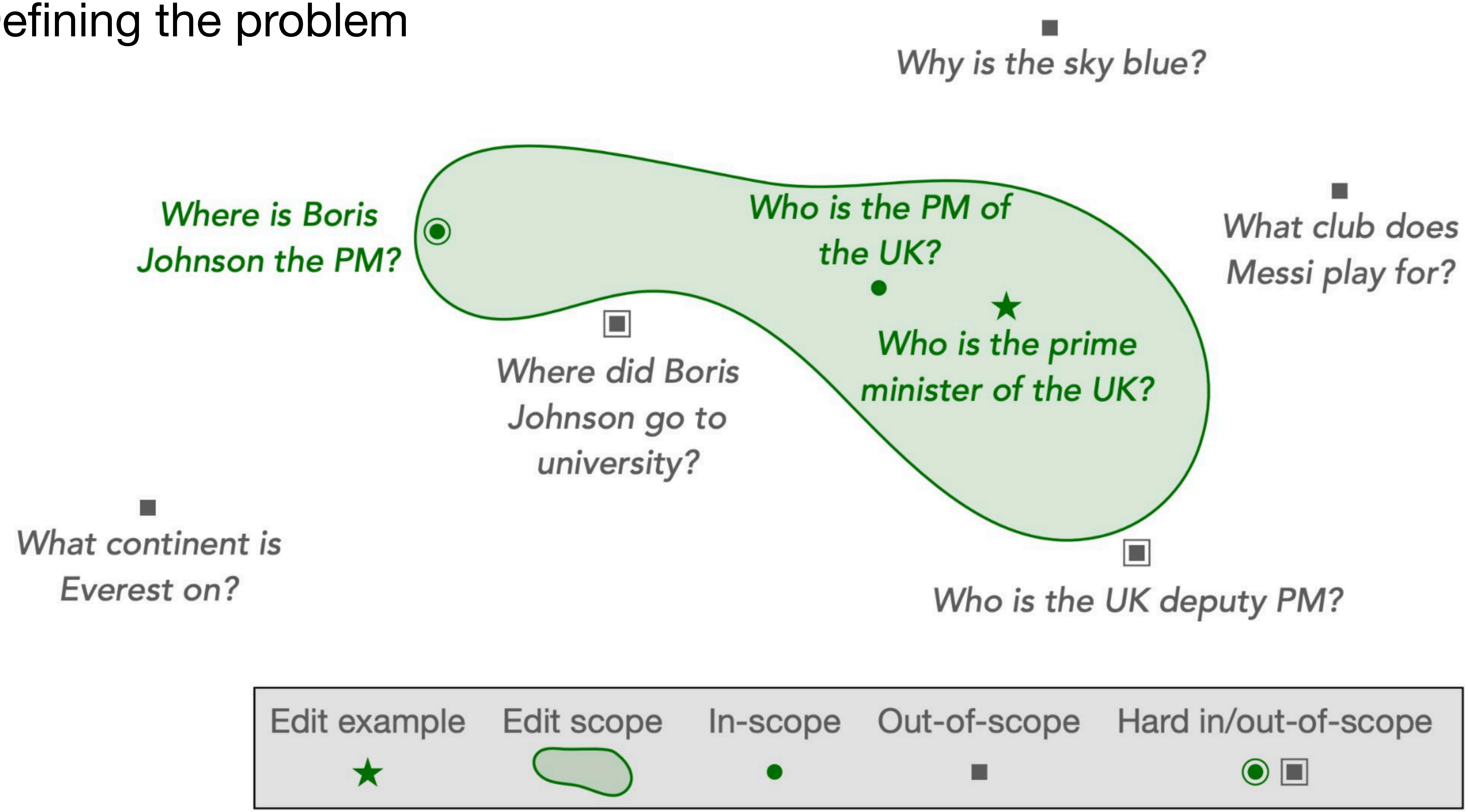
Edit What, Exactly?

Defining the problem



Edit What, Exactly?

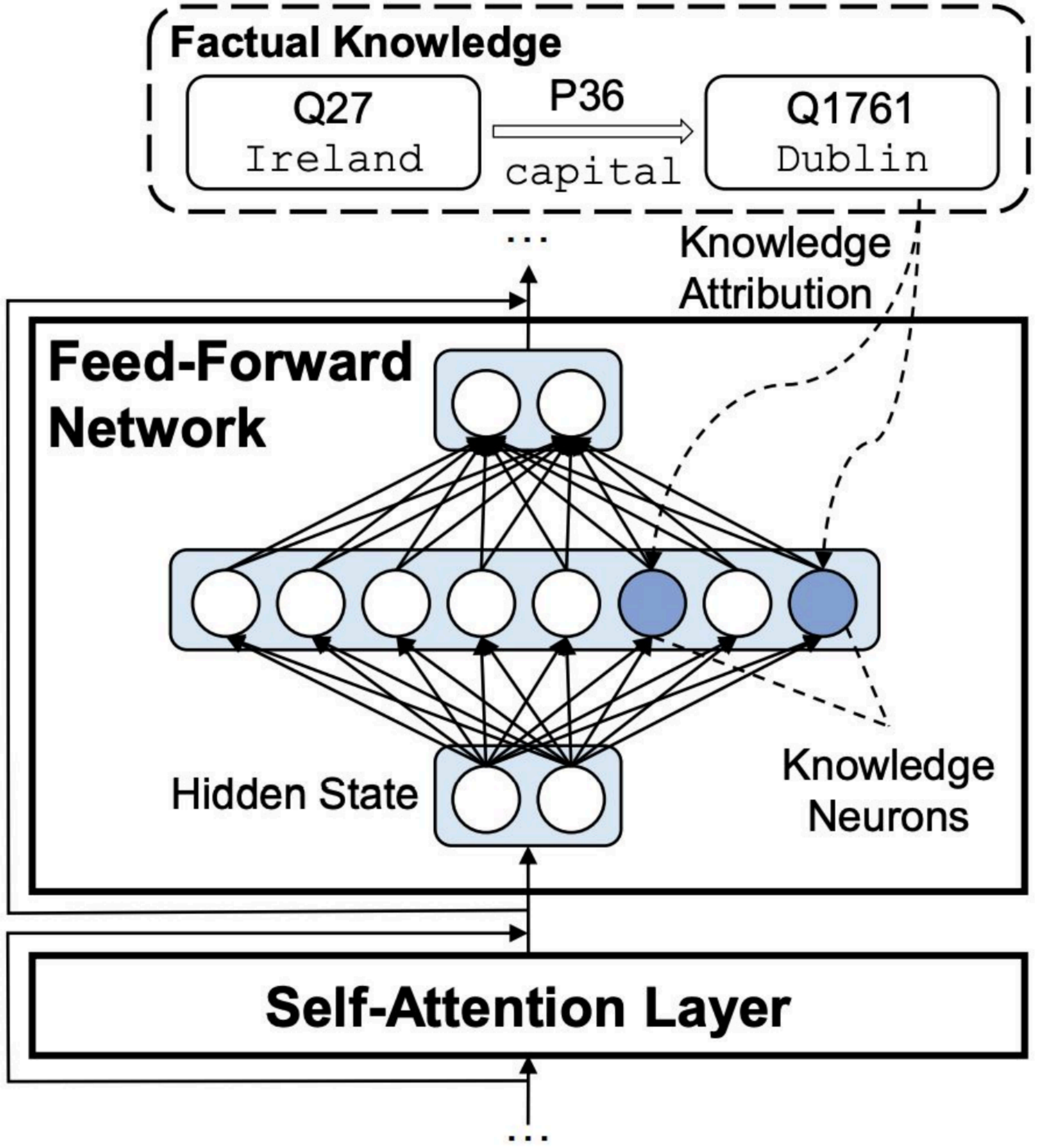
Defining the problem



Knowledge Neurons in Pretrained Transformers

Damai Dai[†‡*], **Li Dong**[‡], **Yaru Hao**[‡], **Zhifang Sui**[†], **Baobao Chang**[†], **Furu Wei**[‡]
[†]MOE Key Lab of Computational Linguistics, Peking University
[‡]Microsoft Research
{daidamai,szf,chbb}@pku.edu.cn
{lidong1,yaruhao,fuwei}@microsoft.com

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 et al., “Knowledge Neurons in Pretrained Transformers”, ACL 2022

Knowledge Attribution: Steps

1. produce n diverse prompts

2. for each prompt, calculate the knowledge attribution scores of neurons

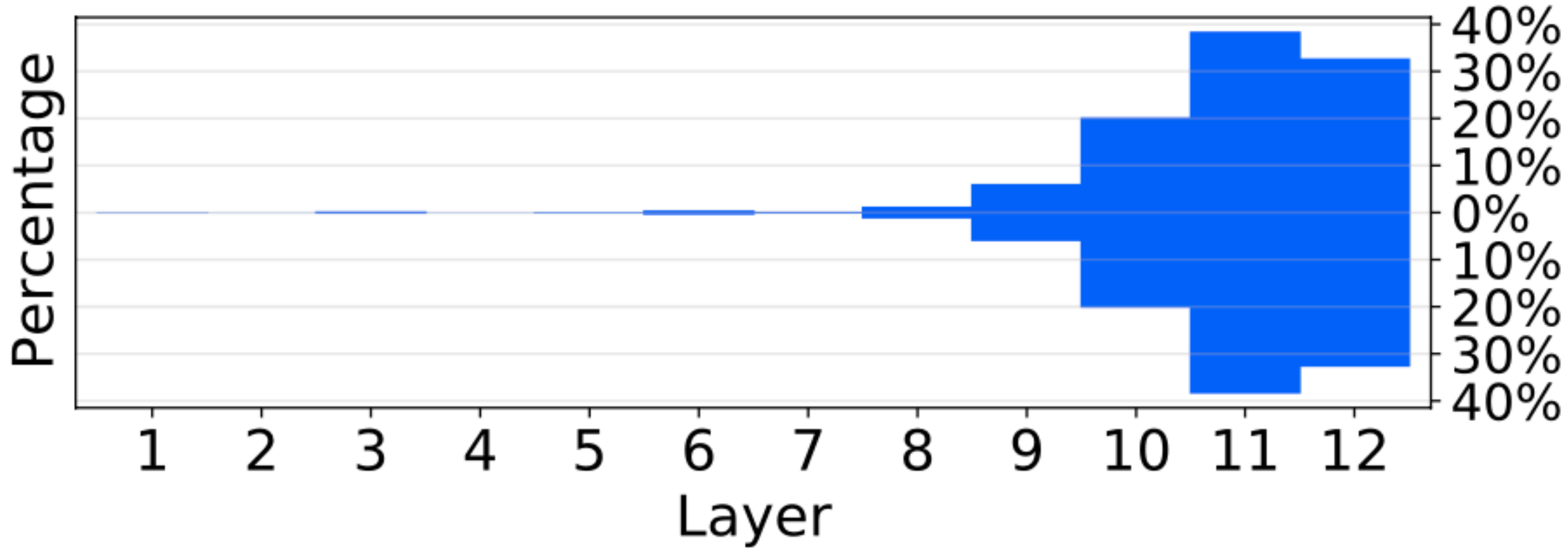
3. for each prompt, retain the neurons with attribution scores greater than t (0.2 times the maximum attribution score)

4. considering all the coarse sets together, retain the knowledge neurons shared by more than p (e.g. 0.7) prompts.

Dai et al., “Knowledge Neurons in Pretrained Transformers”, ACL 2022

Most fact-related neurons are distributed in the topmost layers

- Dataset: PARAREL dataset (Elazar et al., 2021)
  - curated by experts, containing various prompt templates for 38 relations from the T-REx dataset
- Percentage of identified knowledge neurons in each Transformer layer

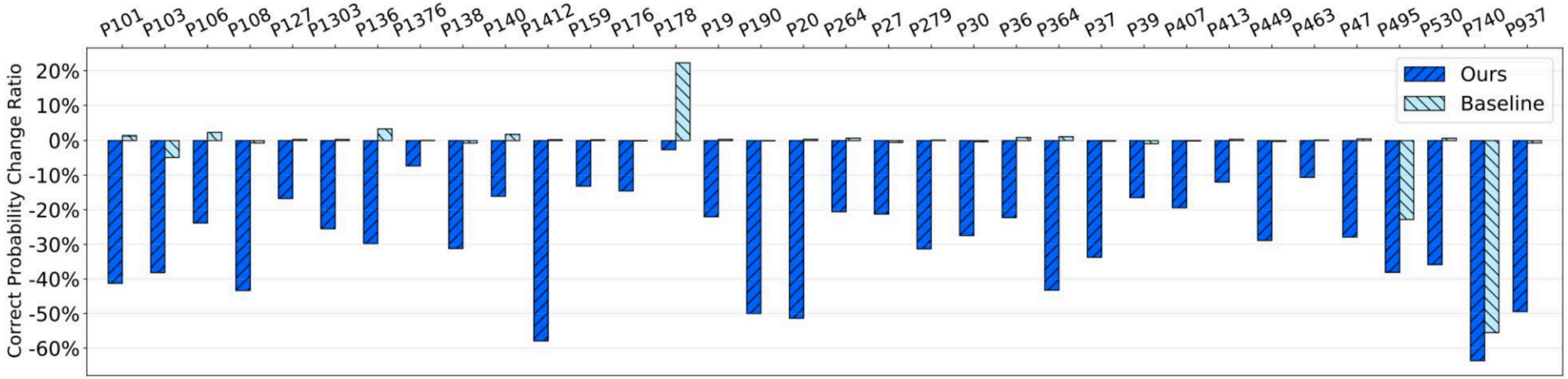


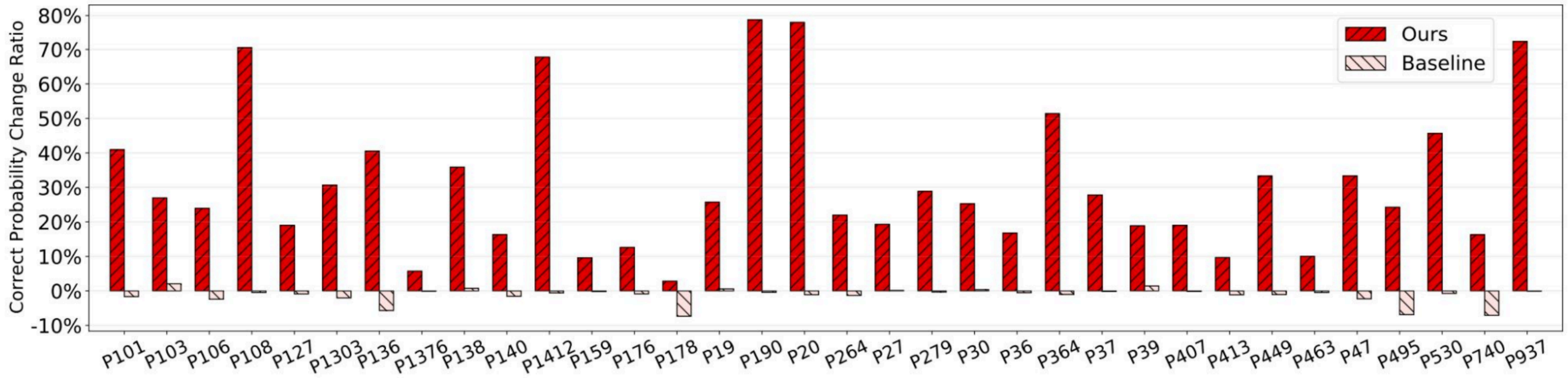
-

Dai et al., “Knowledge Neurons in Pretrained Transformers”, ACL 2022

Suppressing or Amplifying Knowledge Neurons

(1) suppressing knowledge neurons by setting their activations to 0; (2) amplifying knowledge neurons by doubling their activations.





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

Dai et al., “Knowledge Neurons in Pretrained Transformers”, ACL 2022

Case Study - Updating Facts

- Update neuron values by subtracting the word embedding of the previous answer and adding the updated answer

Metric	Knowledge Neurons	Random Neurons
Change rate↑	48.5%	4.7%
Success rate↑	34.4%	0.0%

  - They achieved a change rate and success rate that is better than random neurons.
  - But is this good enough?

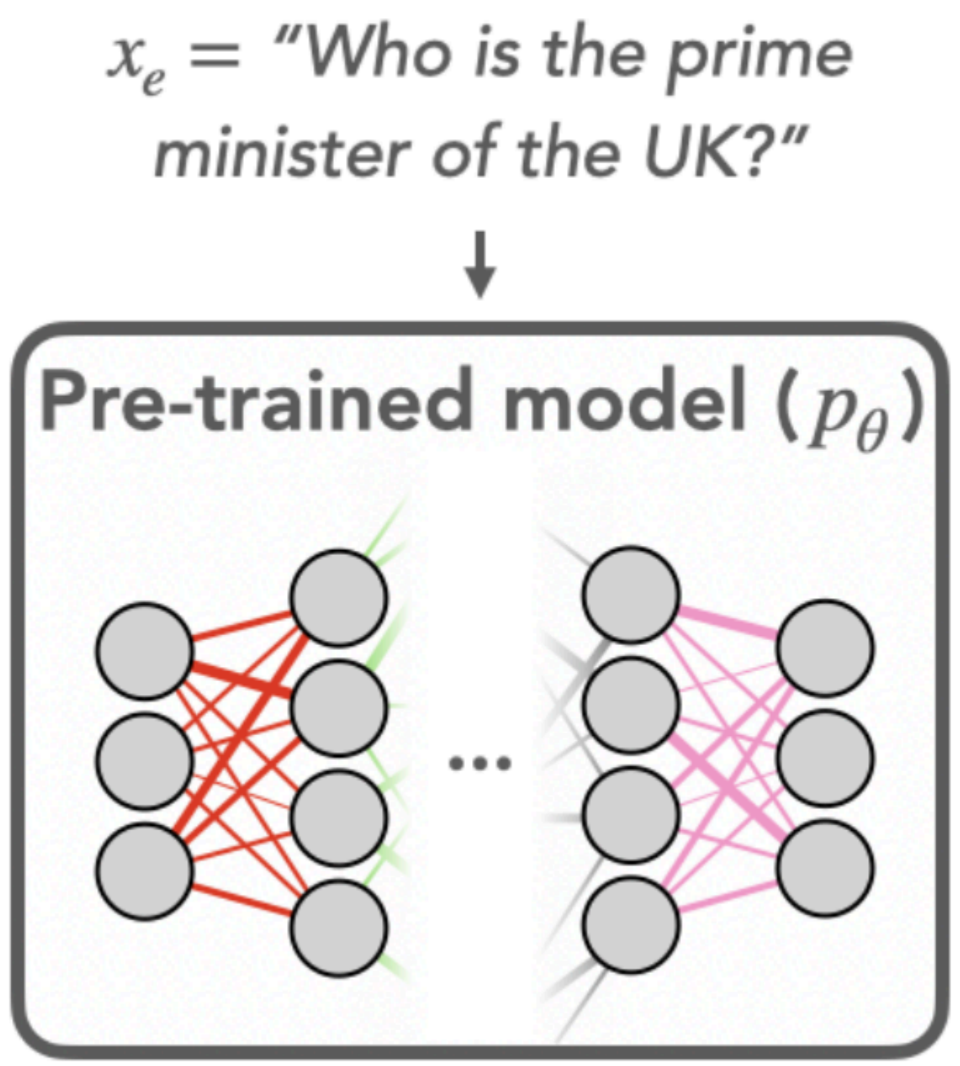
Dai et al., "Knowledge Neurons in Pretrained Transformers", ACL 2022

Published as a conference paper at ICLR 2022

Fast Model Editing at Scale

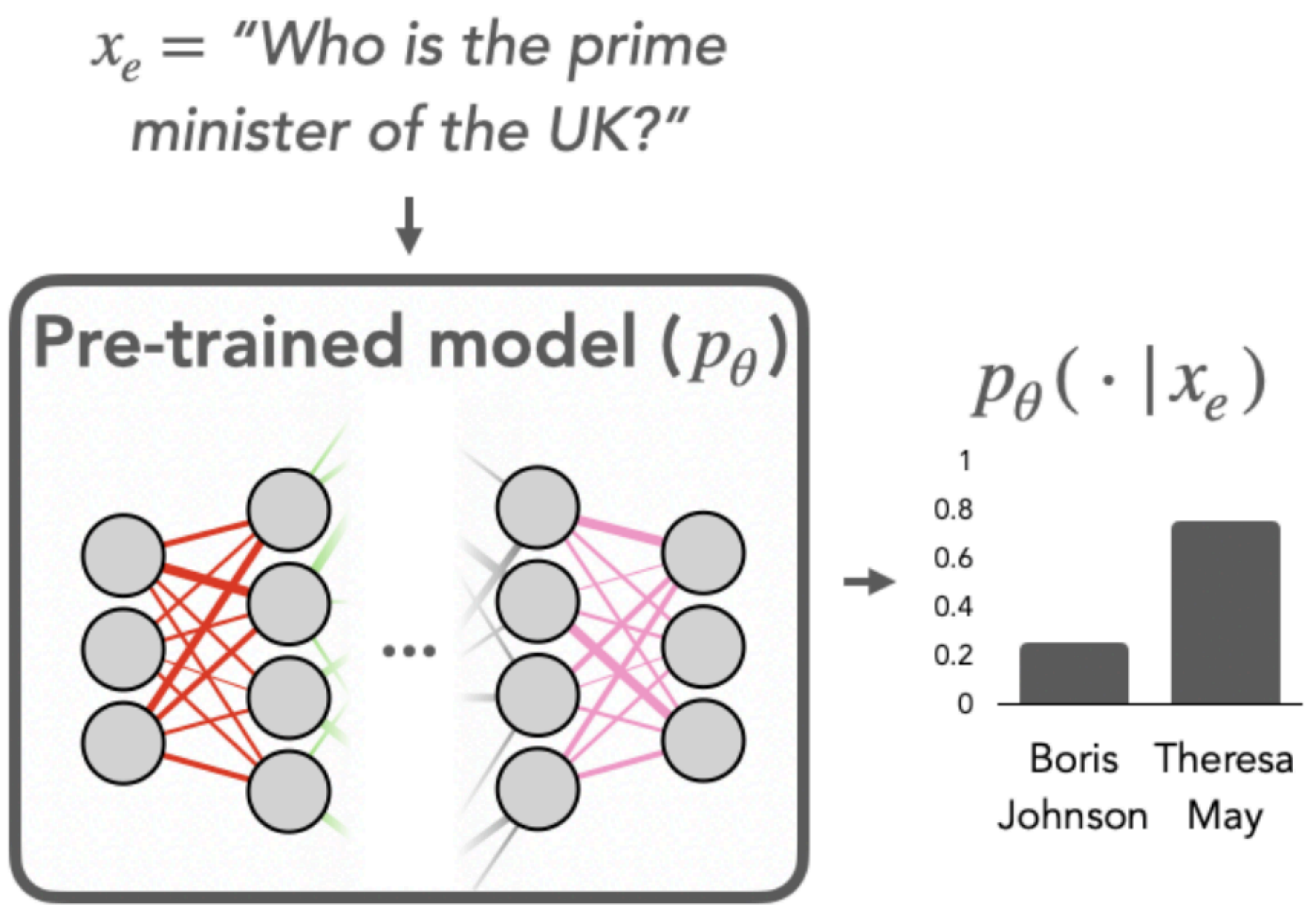
Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, Christopher D. Manning
Stanford University
eric.mitchell@cs.stanford.edu

Editing a Pre-trained Model with MEND

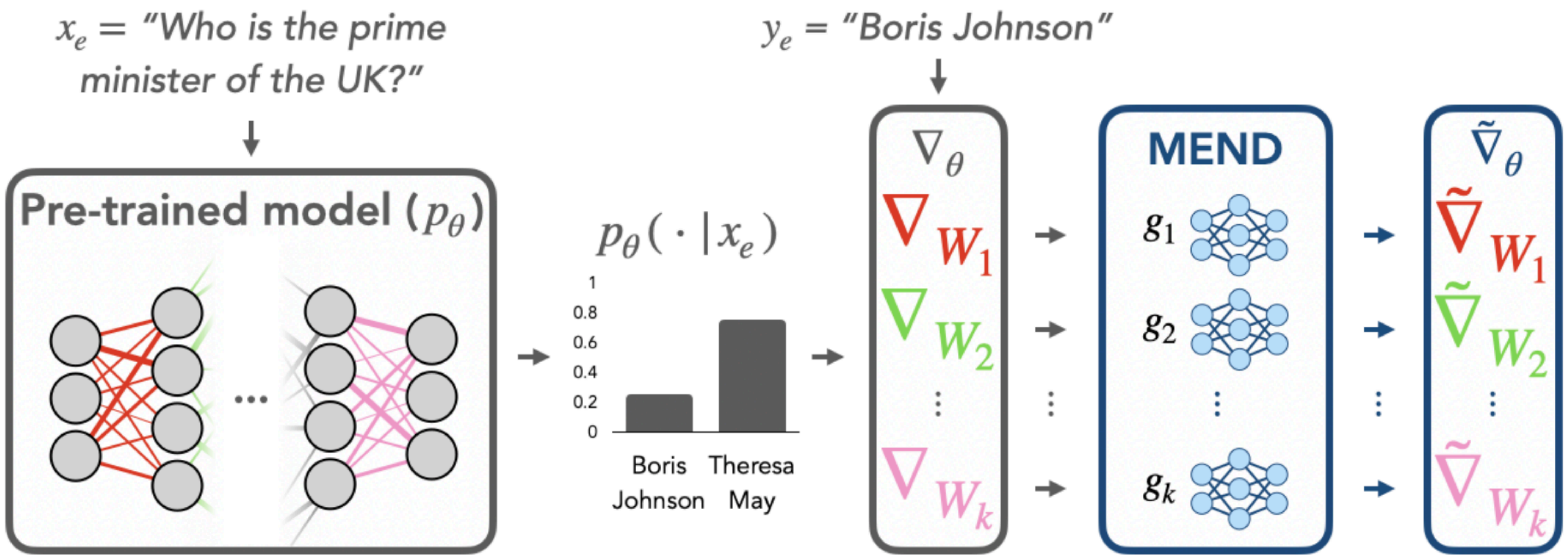


Mitchell et al., "Fast Model Editing at Scale", ICLR 2022.

Editing a Pre-trained Model with MEND



Editing a Pre-trained Model with MEND

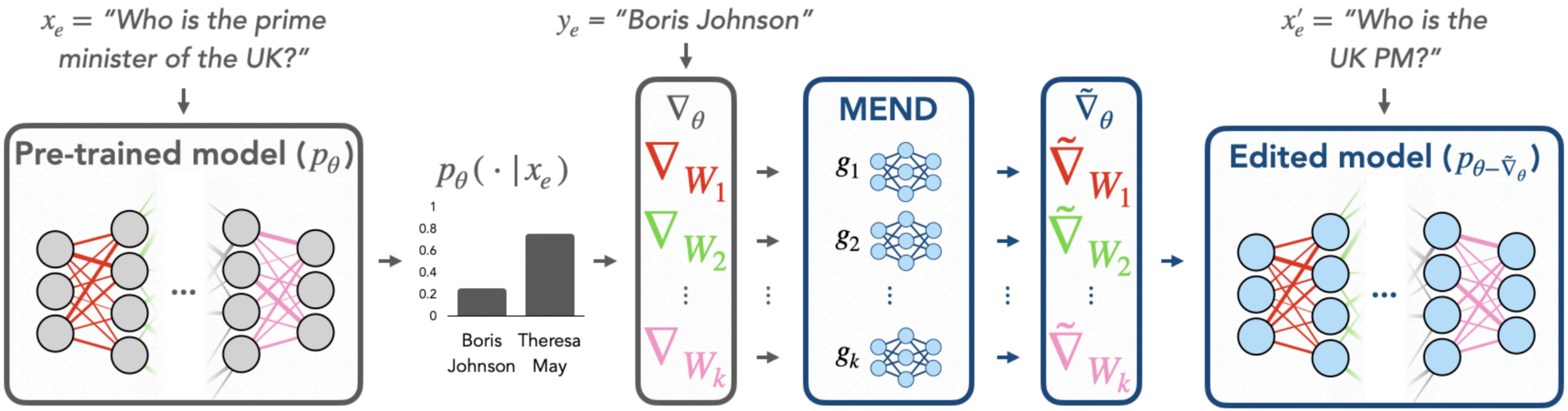


a collection of small auxiliary editing networks that use a single desired input-output pair to make fast, local edits to a pre-trained model's behavior.

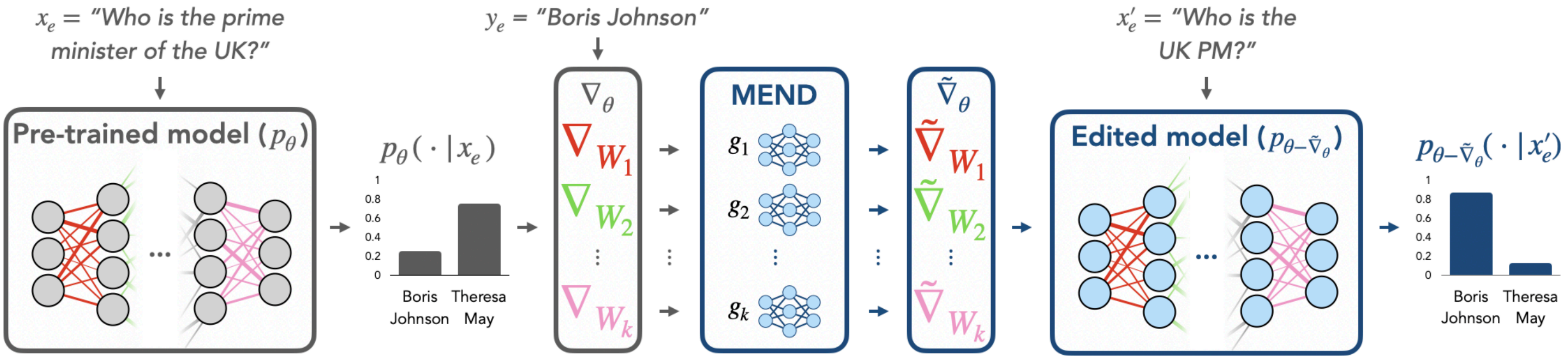
MEND learns to transform the gradient obtained by standard fine-tuning, using a low-rank decomposition of the gradient to make the parameterization of this transformation tractable

- The MEND network produces **gradient updates** for the pretrained model.
- It's not the gradient of all the weights, it's a **transformation** of the gradient!

Editing a Pre-trained Model with MEND

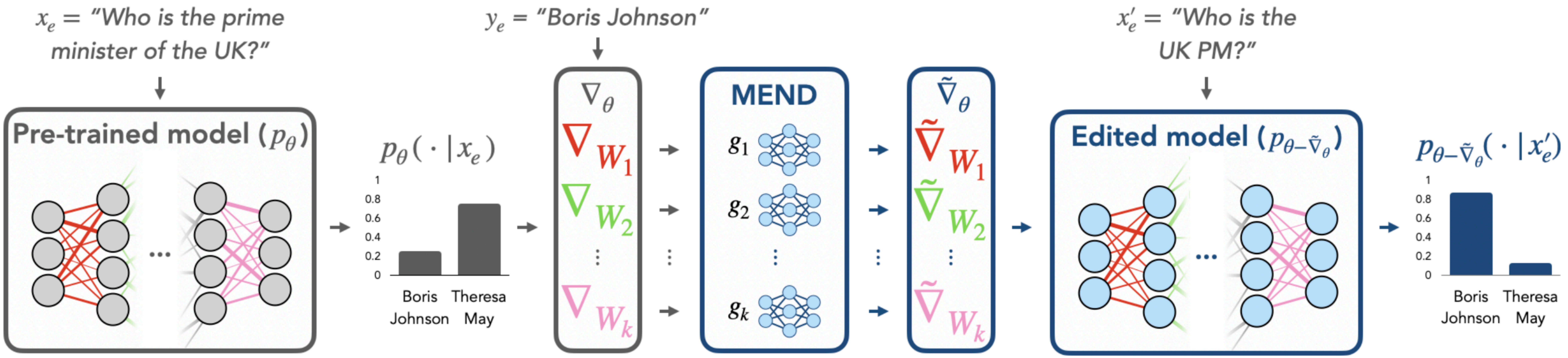


Editing a Pre-trained Model with MEND



The knowledge is updated!

Editing a Pre-trained Model with MEND



- Involves training
  - correctly updates the fact and the related facts
  - maintain answers to the irrelevant facts
- MEND network learns **how to edit** for one single fact change

Algorithm 1 MEND Training

Input: Pre-trained $p_{\theta_{\mathcal{W}}}$, weights to make editable $\mathcal{W}$, editor params ϕ_0, edit dataset D^{tr}_{edit}, edit-locality tradeoff c_{edit}
for $t \in 1, 2, \ldots$ **do**
  Sample $x_{\text{e}}, y_{\text{e}}, x'_{\text{e}}, y'_{\text{e}}, x_{\text{loc}} \sim D^{tr}_{edit}$
  $\tilde{\mathcal{W}} \leftarrow \textsc{Edit}(\theta_{\mathcal{W}}, \mathcal{W}, \phi_{t-1}, x_{\text{e}}, y_{\text{e}})$
  $L_{\text{e}} \leftarrow -\log p_{\theta_{\tilde{\mathcal{W}}}}(y'_{\text{e}} | x'_{\text{e}})$
  $L_{\text{loc}} \leftarrow \text{KL}(p_{\theta_{\mathcal{W}}}(\cdot | x_{\text{loc}}) \| p_{\theta_{\tilde{\mathcal{W}}}}(\cdot | x_{\text{loc}}))$
  $L(\phi_{t-1}) \leftarrow c_{\text{edit}} L_{\text{e}} + L_{\text{loc}}$
  $\phi_t \leftarrow \text{Adam}\left(\phi_{t-1}, \nabla_\phi L(\phi_{t-1})\right)$

Algorithm 2 MEND Edit Procedure

procedure $\textsc{Edit}(\theta, \mathcal{W}, \phi, x_{\text{e}}, y_{\text{e}})$
  $\hat{p} \leftarrow p_{\theta_{\mathcal{W}}}(y_{\text{e}} | x_{\text{e}})$, **caching** input u_ℓ to $W_\ell \in \mathcal{W}$
  $L(\theta, \mathcal{W}) \leftarrow -\log \hat{p}$ ▷ Compute NLL
  for $W_\ell \in \mathcal{W}$ **do**
    $\delta_{\ell+1} \leftarrow \nabla_{W_\ell u_\ell + b_\ell} l_e(x_{\text{e}}, y_{\text{e}})$ ▷ Grad wrt output
    $\tilde{u}_\ell, \tilde{\delta}_{\ell+1} \leftarrow g_{\phi_\ell}(u_\ell, \delta_{\ell+1})$ ▷ Pseudo-acts/deltas
    $\tilde{W}_\ell \leftarrow W_\ell - \tilde{\delta}_{\ell+1} \tilde{u}_\ell^\top$ ▷ Layer ℓ model edit
  $\tilde{\mathcal{W}} \leftarrow \{\tilde{W}_1, \ldots, \tilde{W}_k\}$
  return $\tilde{\mathcal{W}}$ ▷ Return edited weights

$$\tilde{\nabla}_{W_\ell} = \sum_{i=1}^{B} \tilde{\delta}^i_{\ell+1} \tilde{u}^{i\top}_\ell .$$

Results

Locality Loss:

Minimize changes on irrelevant examples

- FT: fine-tuning with updated facts
- FT + KL: fine-tuning with updated facts and locality loss

	zsRE Question-Answering			
	T5-XL (2.8B)		T5-XXL (11B)	
Editor	ES ↑	acc. DD ↓	ES ↑	acc. DD ↓
FT	0.58	**<0.001**	0.87	**<0.001**
FT+KL	0.55	**<0.001**	0.85	**<0.001**
MEND	**0.88**	0.001	**0.89**	**<0.001**

MEND shows the best **Edit success rate (ES)** and least interference to overall model perplexity or accuracy, i.e., **ppl. DD, acc.DD**.

Comparison of the Two Works

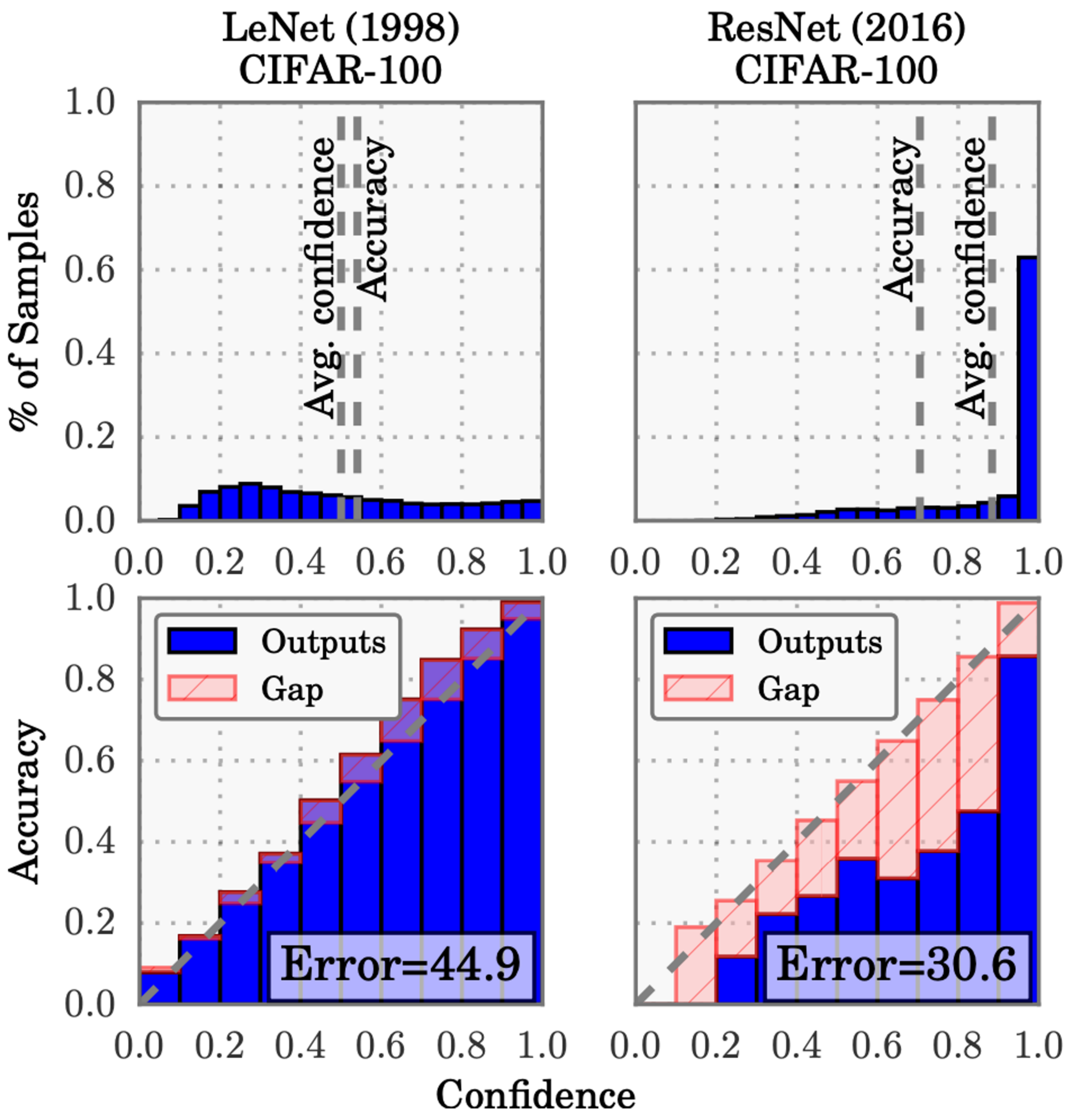
	Knowledge Neurons	**MEND**
Method	Attribution-based	Learning-based
Training?	no	yes
Restricted by	Attribution algorithm	Need a lot of edits data

Limitations of LLMs about Knowledge

- Knowledge cutoff
  - LLMs don't know about any events that happened after their training
  - LLMs don't have any knowledge about private or confidential information that they have not encountered during training
- Hallucinations
  - LLMs are trained to generate realistic-sounding or convincing text
  - But the generated text may be nonetheless wrong
    - instead of admitting that it lacks the base facts in its training.

Confidence Calibration

- In real-world, classification networks should indicate when they are likely to be incorrect.
- The probability associated with the predicted class label should reflect its ground truth correctness likelihood



Guo et al., On Calibration of Modern Neural Networks, ICML 2017.

Confidence Calibration: Metrics

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|,$$

$$\mathrm{MCE} = \max_{m \in \{1,\ldots,M\}} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$$

Uncertainty and confidence in LLMs

- LLMs have achieved remarkable success, but often exhibit overconfidence and poor calibration
- LLMs are prone to generate false information (i.e., “hallucinations”) and are often unaware of whether they know the answer.
- The prompted nature of LLMs offers an alternative means of ensembling.

Jiang et al., “Calibrating Language Models via Augmented Prompt Ensembles”, ICLR 2023.

No Exploration of Uncertainty

- Metrics like top-one accuracy may capture the ordering of predictions
- But they lack the resolution to reflect on the degree of certainty of factual knowledge being learned by LLMs.

Jiang et al., "Calibrating Language Models via Augmented Prompt Ensembles", ICLR 2023.

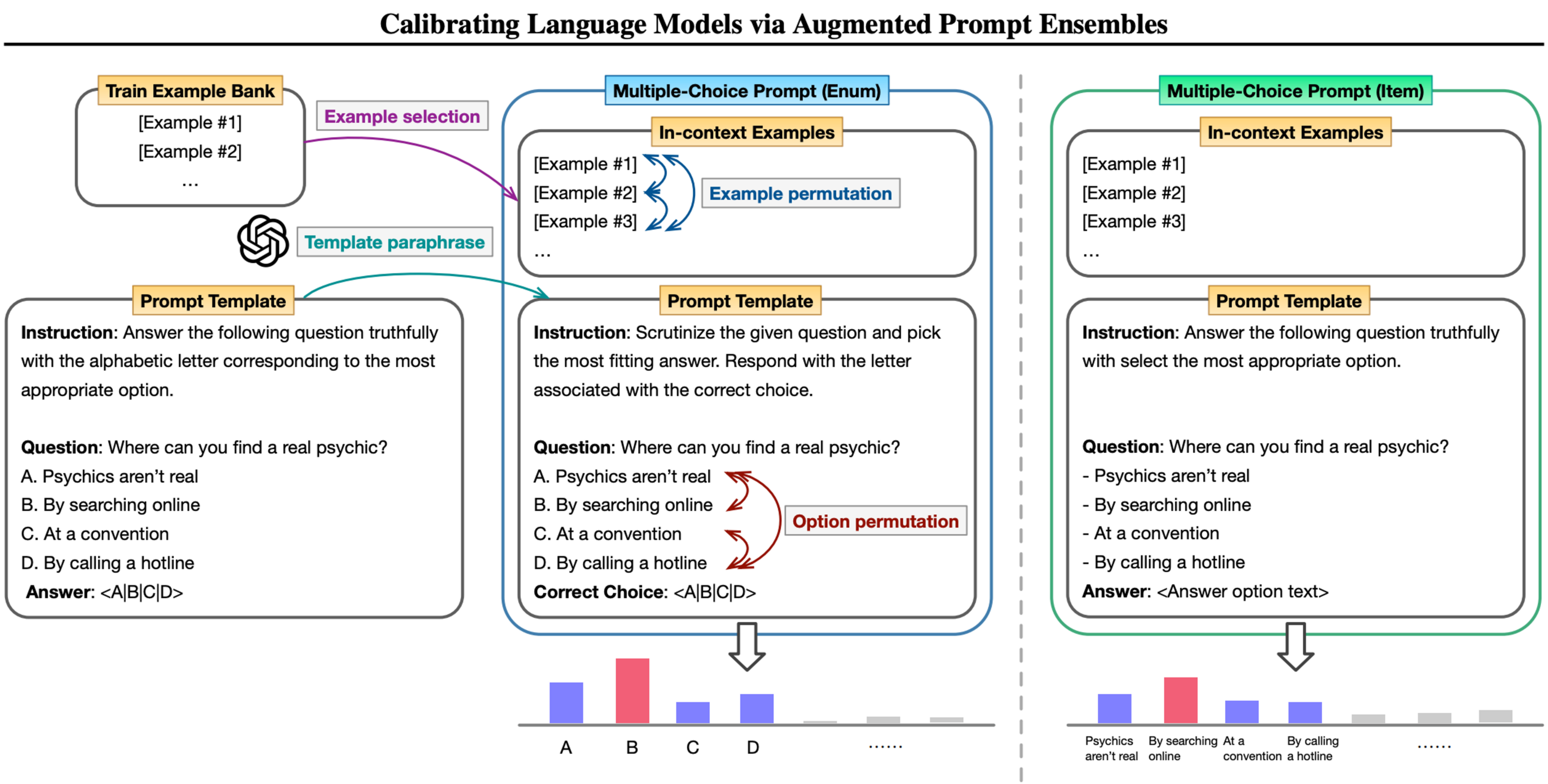
Calibration via Augmented Prompt Ensembles (CAPE)

- Prompt ensembles: sets of diverse prompts that are meant to solve the same problem.

- To improve LLM reliability, querying the LLM with multiple different input prompts and considering each of the model's responses when inferring a final answer.

Jiang et al., "Calibrating Language Models via Augmented Prompt Ensembles", ICLR 2023.

CAPE



(a) Prompt augmentations on ENUM multiple-choice format (b) ITEM multiple-choice format

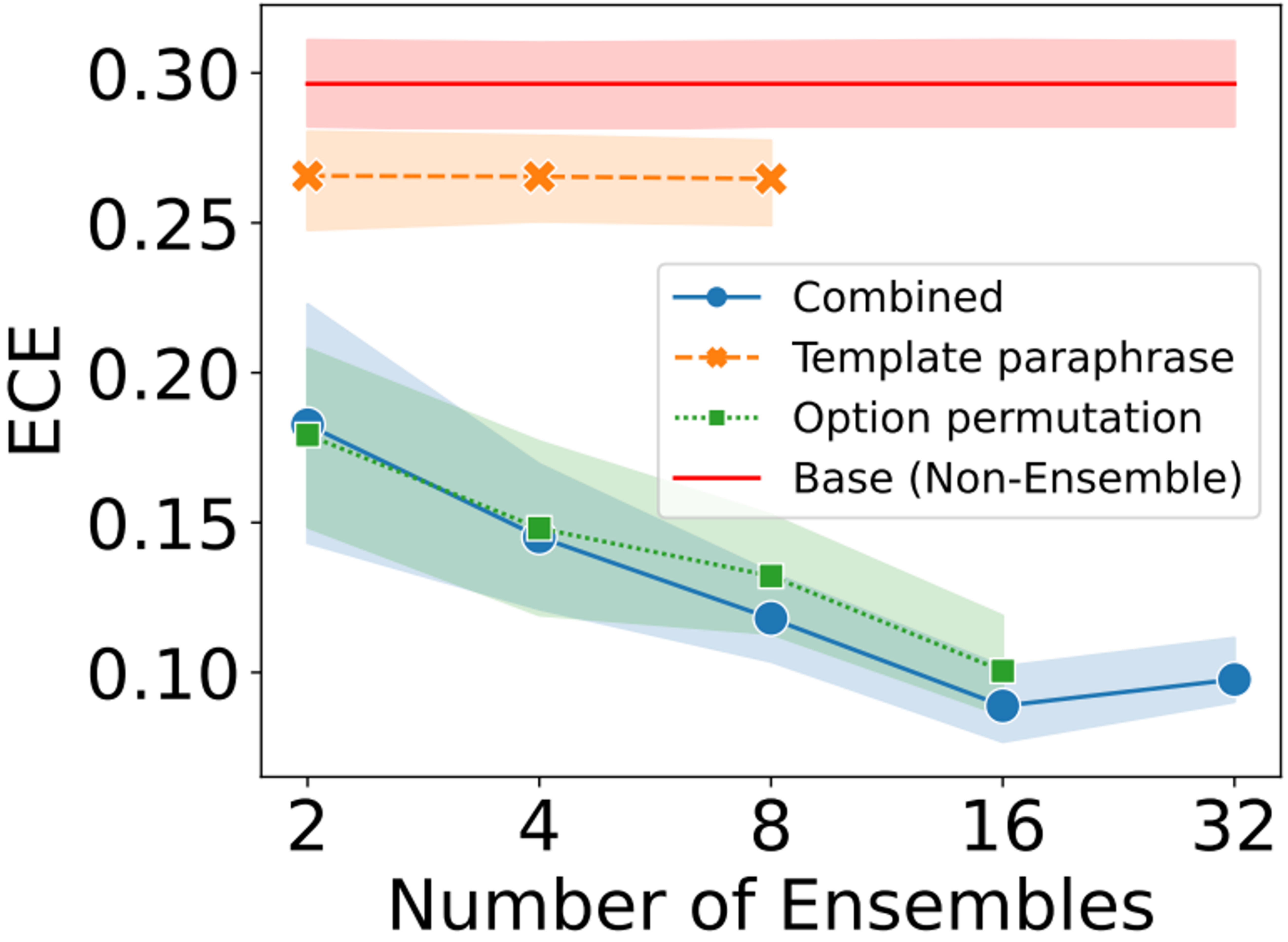
Jiang et al., "Calibrating Language Models via Augmented Prompt Ensembles", ICLR 2023.

Prompt augmentation

- Template paraphrase
- Option permutation
- In-context example permutation
- In-context example selection

Jiang et al., “Calibrating Language Models via Augmented Prompt Ensembles”, ICLR 2023.

CAPE: Results



Jiang et al., "Calibrating Language Models via Augmented Prompt Ensembles", ICLR 2023.

Conclusion

Conclusion

- **Large language models pretrained on unstructured text perform competitively on open-domain QA**, even compared to competitors with access to external knowledge

- **Scale is critical to performance**

- Using LMs as knowledge bases suffers from **lack of interpretability**, and LMs are prone to **hallucinating** “realistic” answers

Questions

Knowledge Attribution

- Integrated gradient:

$$Attr(n_i^l) = \frac{\bar{n}_i^l}{m} \sum_{k=1}^{m} \frac{\partial p(y^* | x, n_i^l = \frac{k}{m} \bar{n}_i^l)}{\partial n_i^l}$$

Baselines

- **Freq**: ranks candidates by frequency of appearance as objects for a subject-relation pair
  - Analogous to majority classifier
- **Pretrained models**
  - **RE** (Sorokin and Gurevych, 2017): extracts relation triples from sentence
    - RE_n : uses exact string matching for entity linking
      - RE_n has to find the subject/object entities itself
    - RE_o : uses oracle for entity linking
      - As long as RE_o gets the right relation type, it gets the answer for free