Language Models and Knowledge

Mengzhou Xia & Jane Pan

October 10th, 2022

Outline

1. **What is a knowledge base?**

2. **Can language models be used as knowledge bases?** (Petroni et al., 2019)

3. **Can a closed-book QA LM perform as well as other open-book methods?** (Roberts et al., 2020)

4. **How to update facts?** (Dai et al., 2021, Mitchell et al., 2022)

Introduction

Motivation

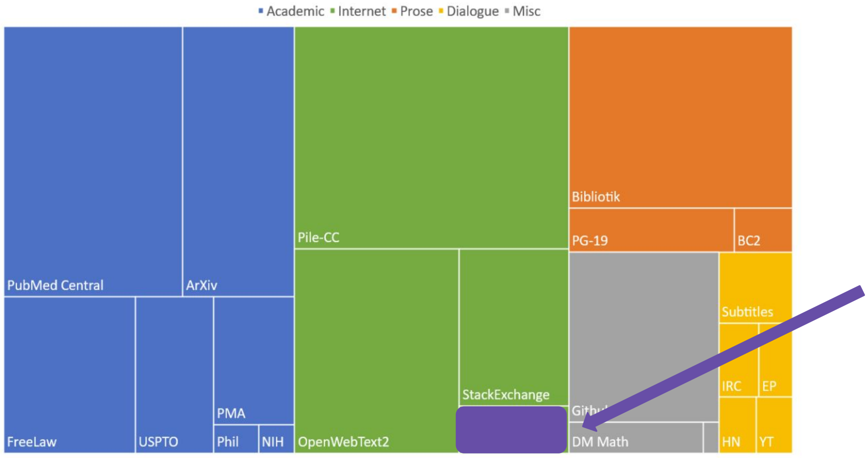
- The corpora used to pretrain language models are **huge aggregations** of information and data from the internet

Motivation

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

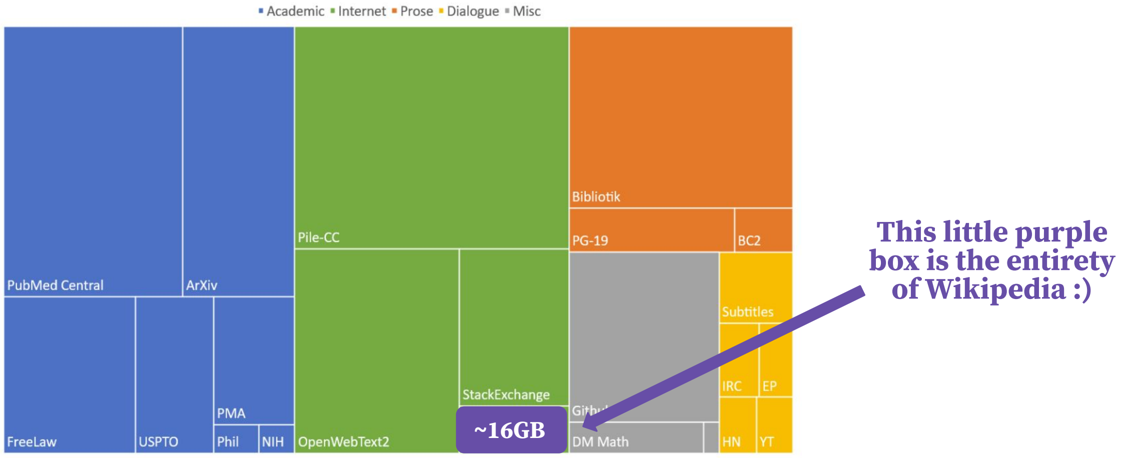
Motivation

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



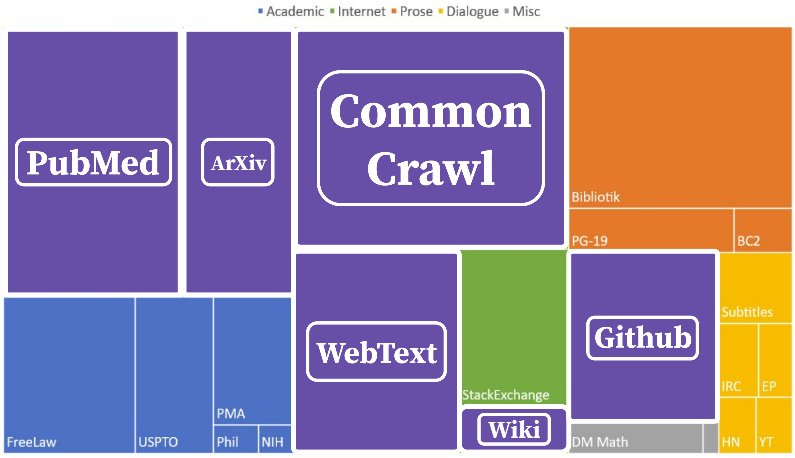
Motivation

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



Motivation

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



Pretraining and knowledge

- Pre-training allows language models to learn robust **task-agnostic features**, which is critical for high performance on downstream tasks

Pretraining and knowledge

- Pre-training allows language models to learn robust **task-agnostic features**, which is critical for high performance on downstream tasks

- As **language models and their pre-training corpora scale,** the amount of information encoded in the pretrained language models also increases
    - Few-shot learning
    - In-context learning

Pretraining and knowledge

- Pre-training allows language models to learn robust **task-agnostic features**, which is critical for high performance on downstream tasks

- As **language models and their pre-training corpora scale,** the amount of information encoded in the pretrained language models also increases
  - Few-shot learning
  - In-context learning

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

Key Question

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

GPT-3 Zero-shot Knowledge Retrieval

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

GPT-3 Zero-shot Knowledge Retrieval

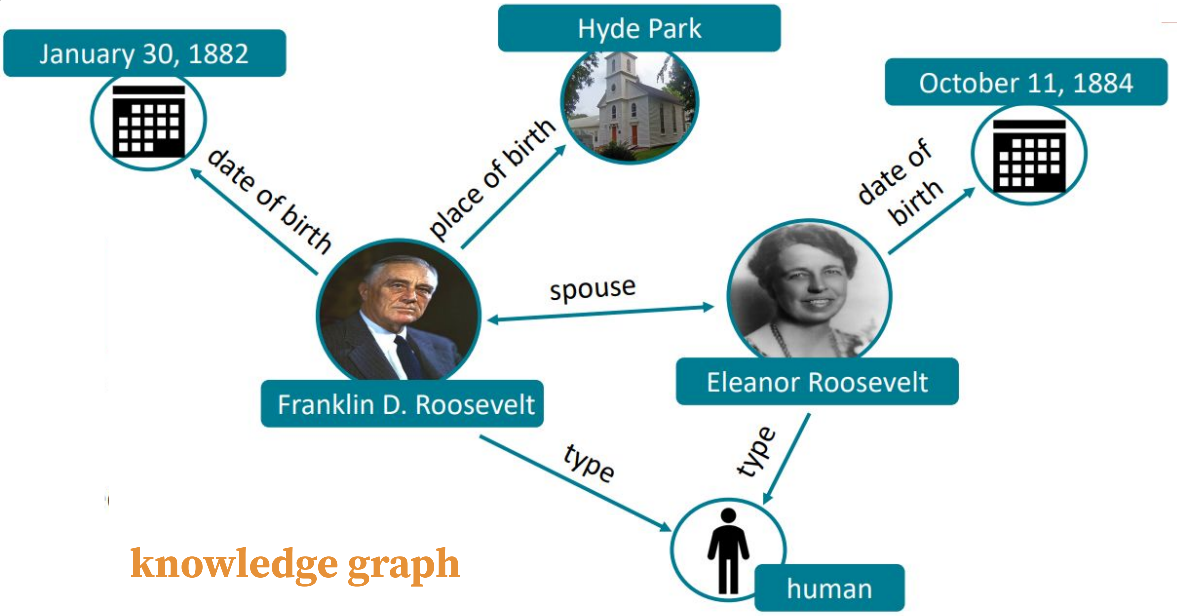
- This was not so obvious to NLP researchers *three years ago*!
- Instead, traditional knowledge bases were often used

What is a knowledge base?

What is a knowledge base?

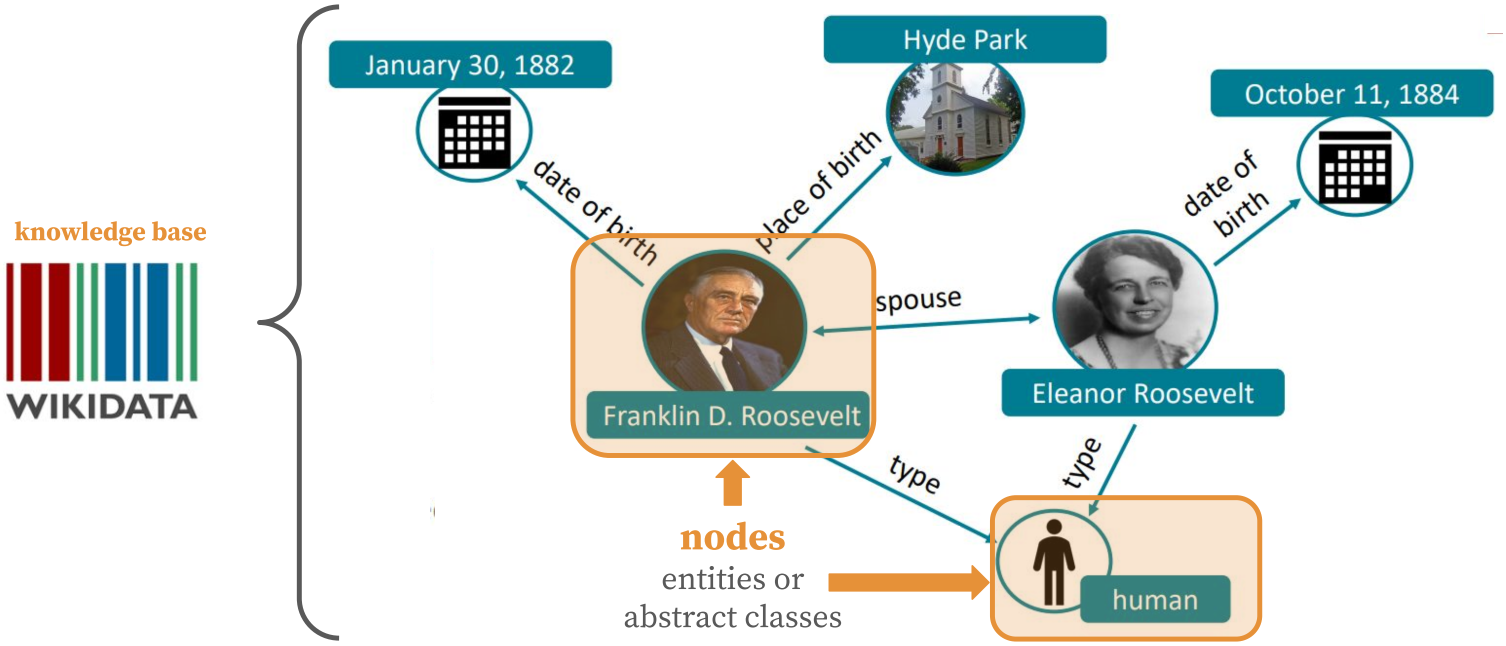
knowledge base

WIKIDATA

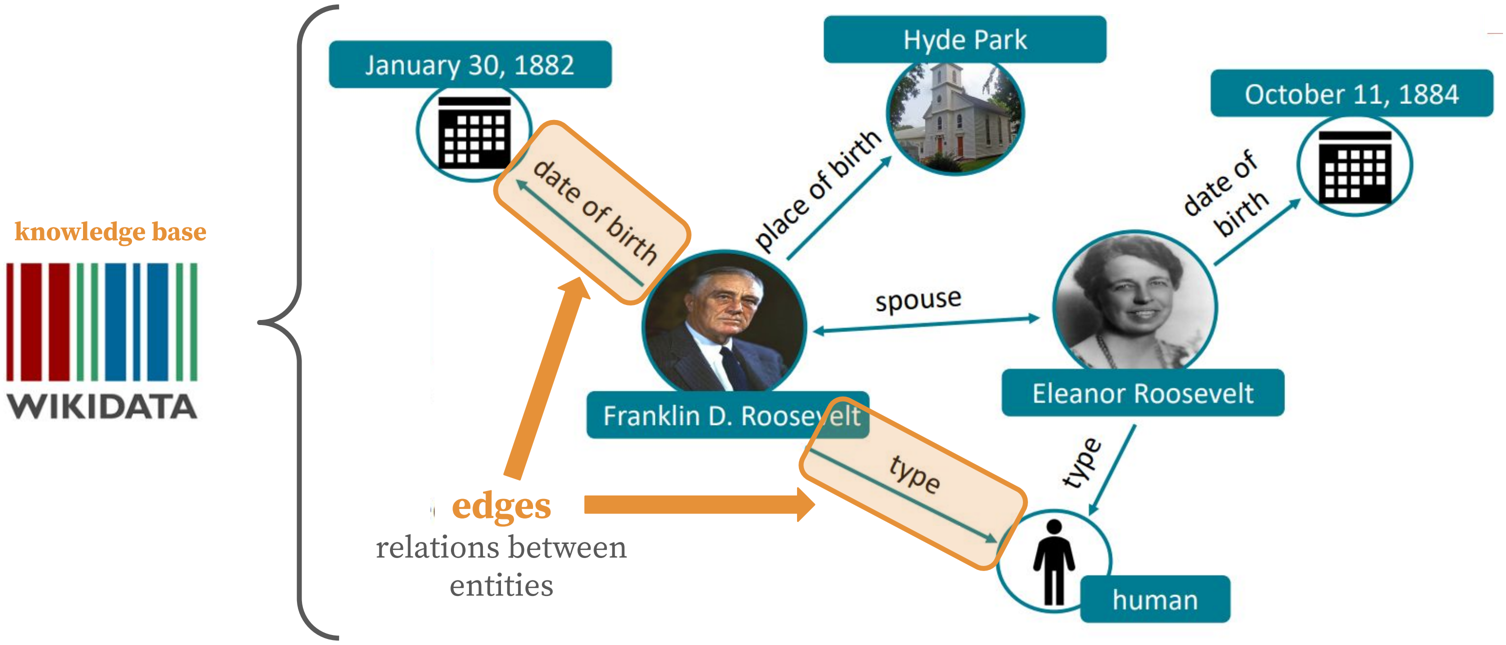


knowledge graph

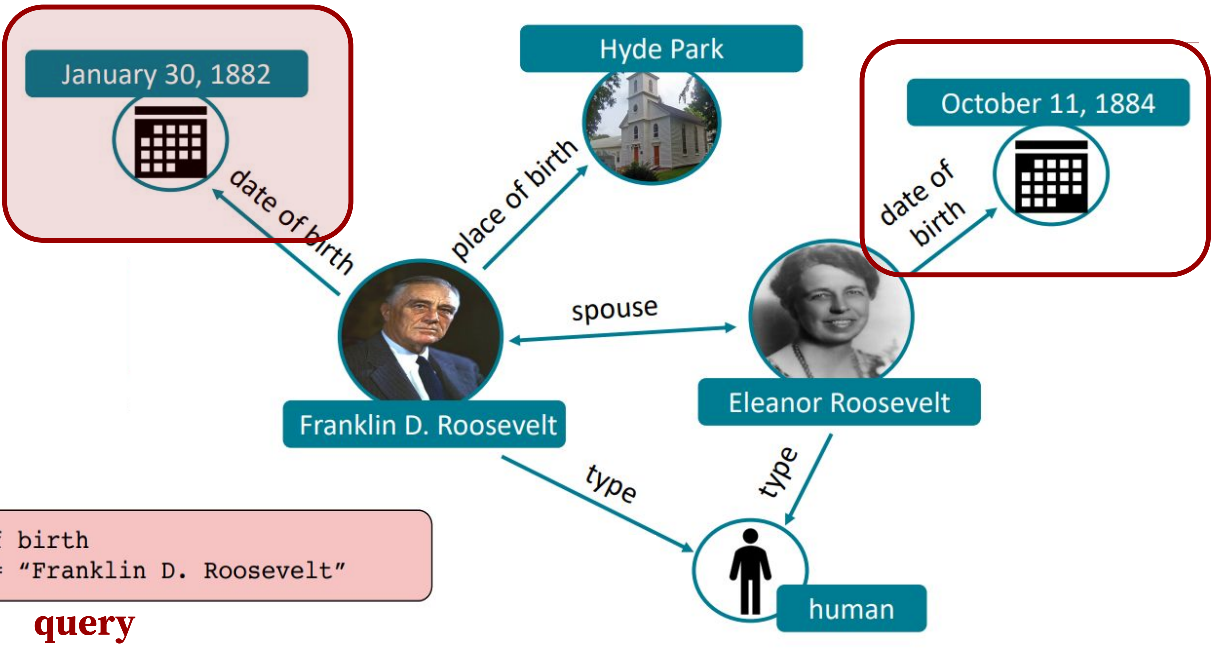
What is a knowledge base?



What is a knowledge base?



What is a knowledge base?

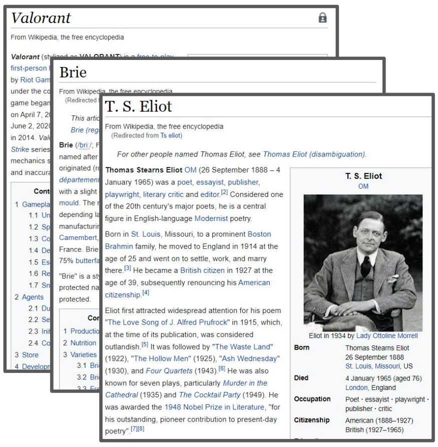



```
SELECT date of birth
WHERE person = “Franklin D. Roosevelt”
```

query

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

How were knowledge bases formed?

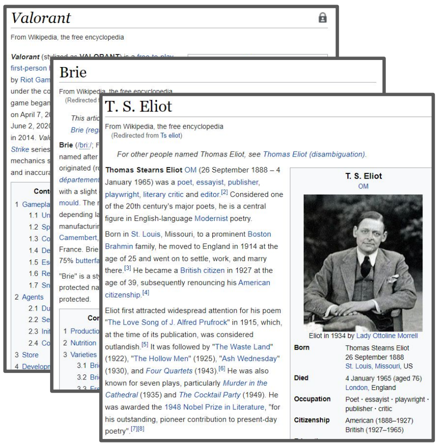


unstructured text

knowledge base

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

How were knowledge bases formed?



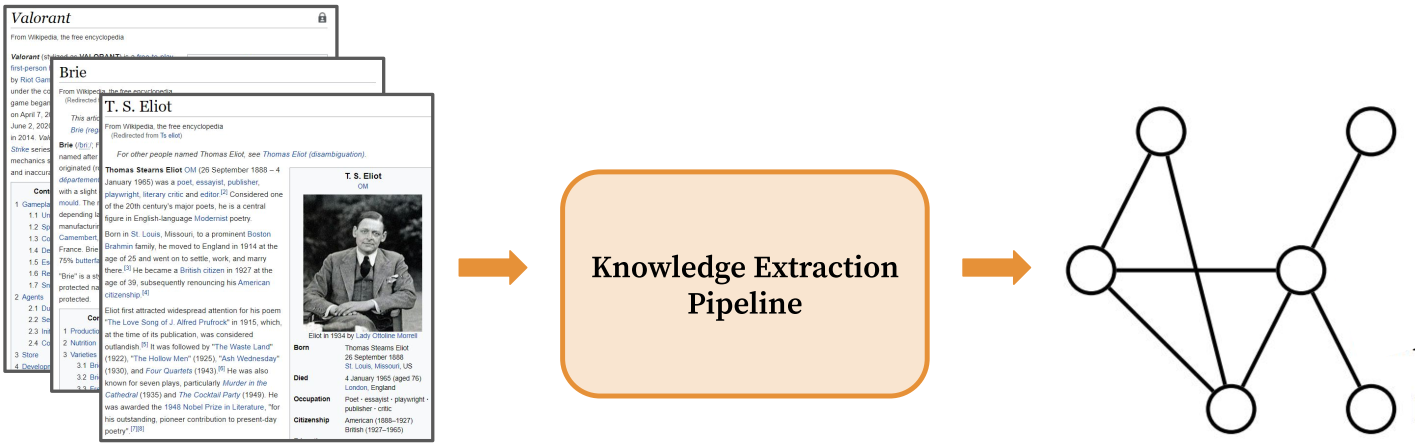
→ **Knowledge Extraction Pipeline**

unstructured text

knowledge base

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

How were knowledge bases formed?



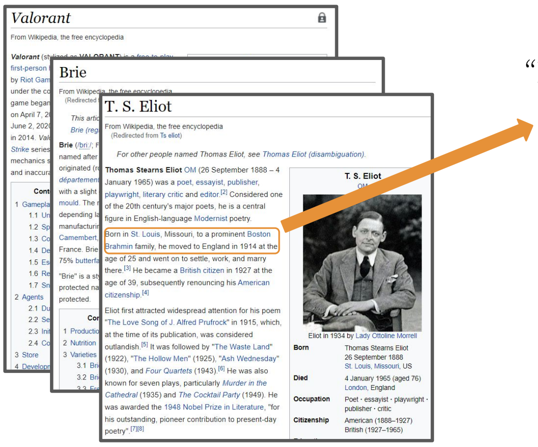
unstructured text

knowledge base

Downsides of using knowledge bases

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

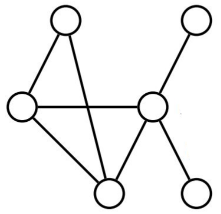
Downsides of using knowledge bases



unstructured text

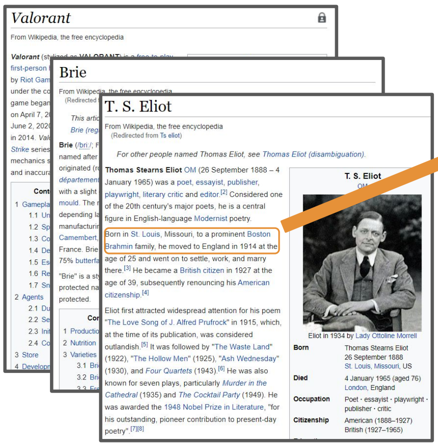
"Born in St. Louis, Missouri, to a prominent Boston Brahmin family..."

Untrained Knowledge Extraction Pipeline



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

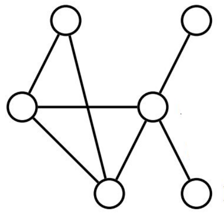
Downsides of using knowledge bases



unstructured text

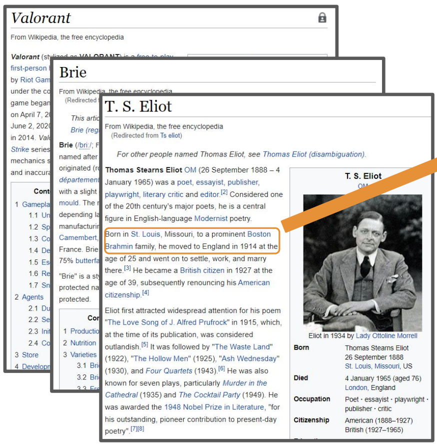
"Born in St. Louis, Missouri, to a prominent Boston Brahmin family..."

Untrained Knowledge Extraction Pipeline



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

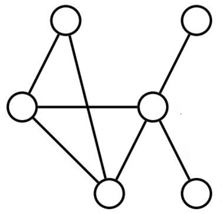
Downsides of using knowledge bases



unstructured text

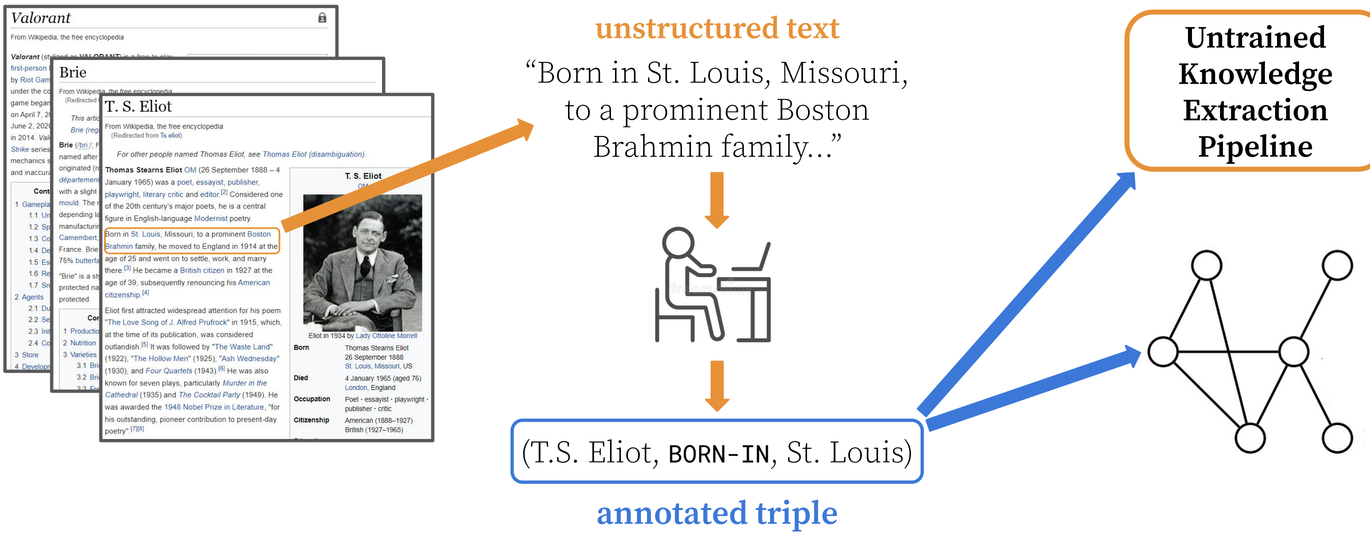
"Born in St. Louis, Missouri, to a prominent Boston Brahmin family..."

Untrained Knowledge Extraction Pipeline



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

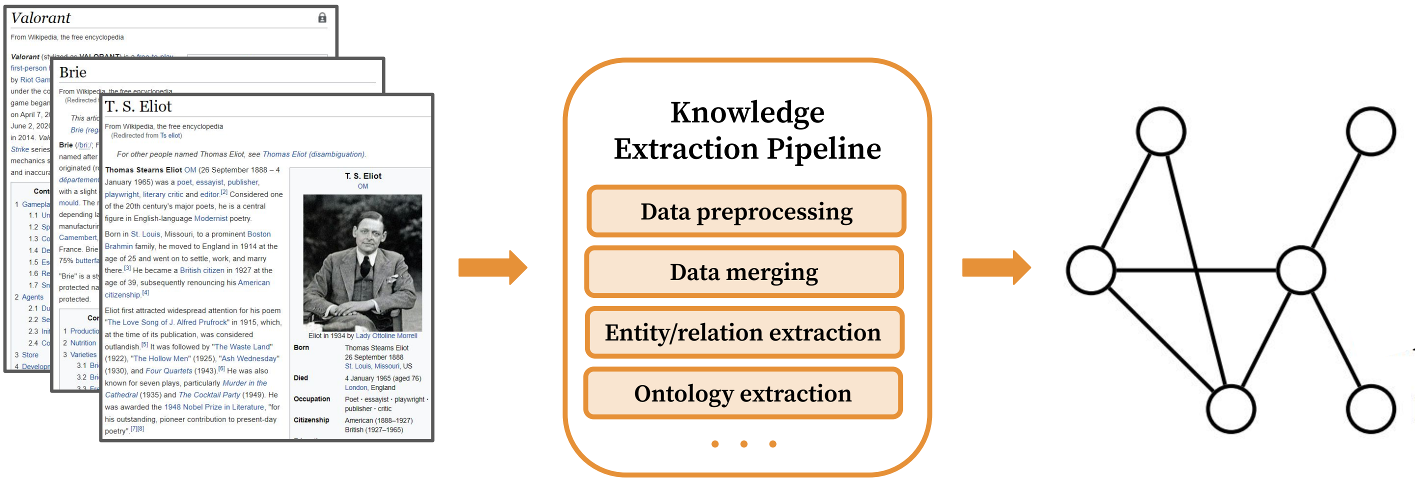
Downsides of using knowledge bases



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

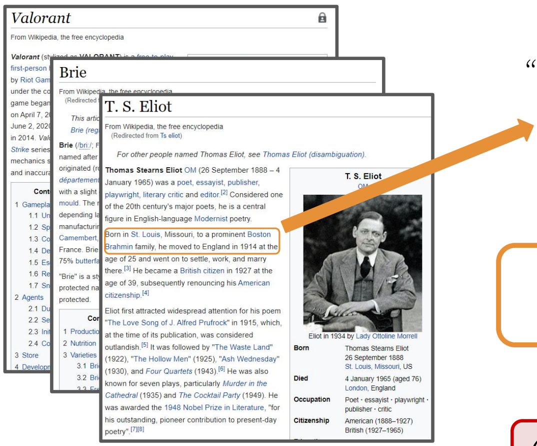
Right-most graphic from [Petroni et al., 2019]

Downsides of using knowledge bases



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

Downsides of using knowledge bases



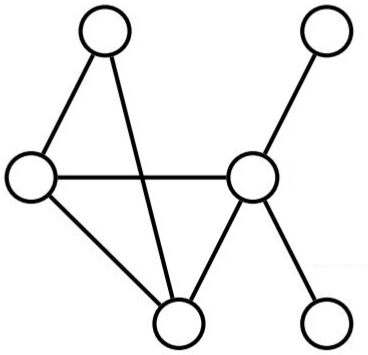
unstructured text

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

Knowledge Extraction Pipeline

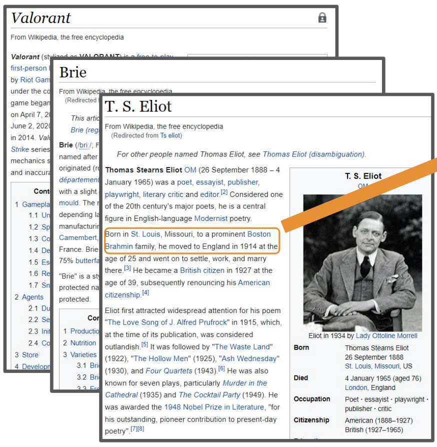
(T.S. Eliot, `BORN-IN`, **Boston**)

incorrect extraction



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

Downsides of using knowledge bases



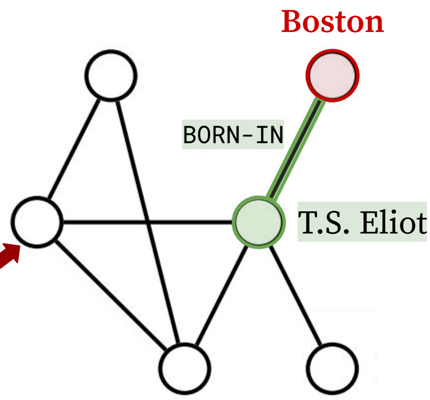
unstructured text

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

Knowledge Extraction Pipeline

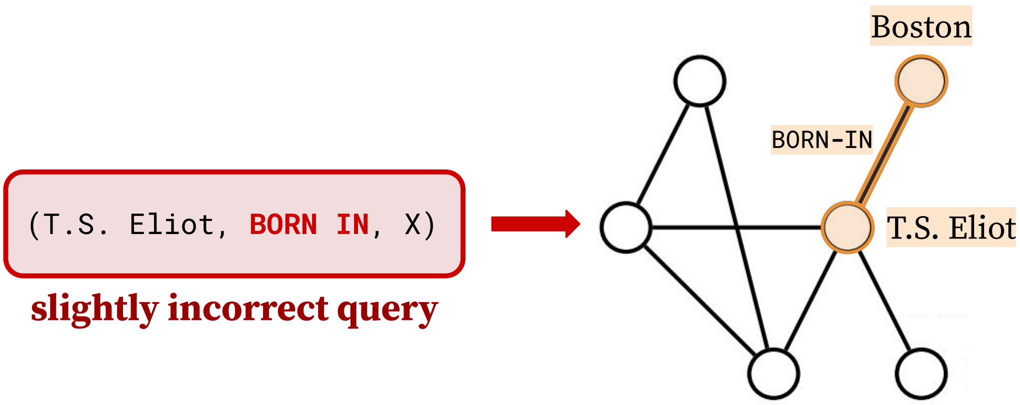
(T.S. Eliot, `BORN-IN`, **Boston**)

incorrect extraction



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

Downsides of using knowledge bases



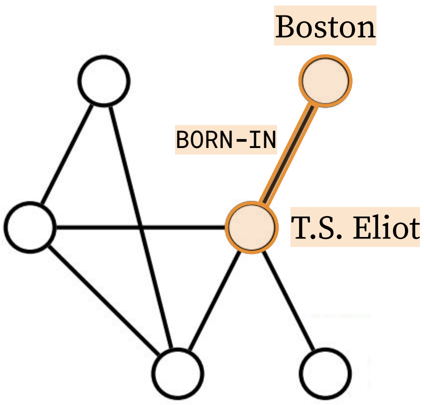
Reliant on **fixed schemas** to store or query data

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

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

Are there better alternatives?

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

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

Image from www.slyderstavern.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

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

Question: How do we check this?

Image from www.slyderstavern.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

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

Question: How do we check this?

Answer:

Image courtesy of [Patrick Lewis]

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:

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

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

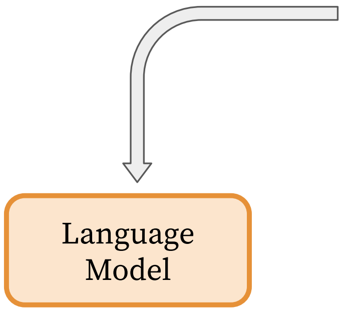
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

*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

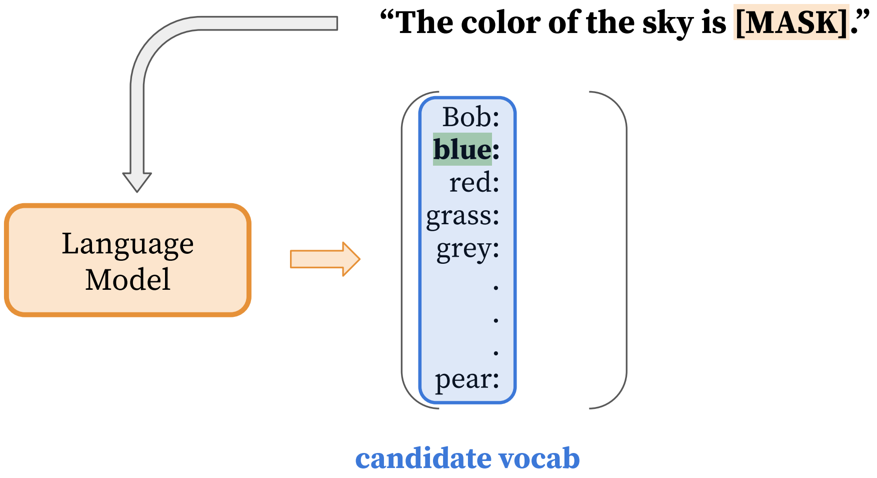
"The color of the sky is [MASK]."



*according to the LM

Evaluation of LM via LAMA

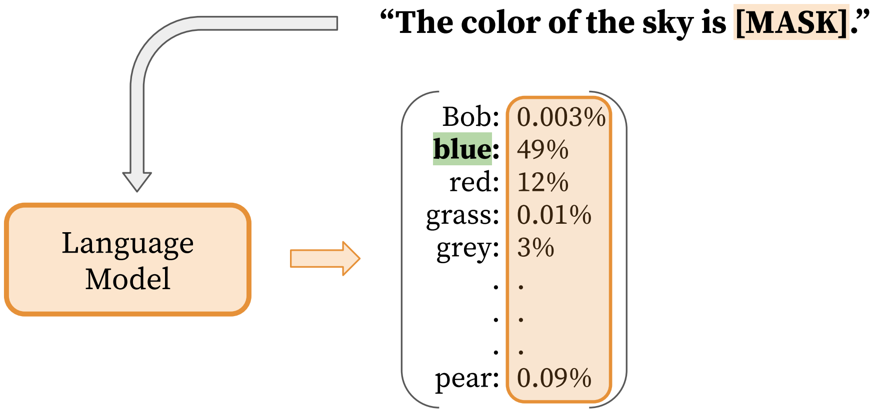
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



*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

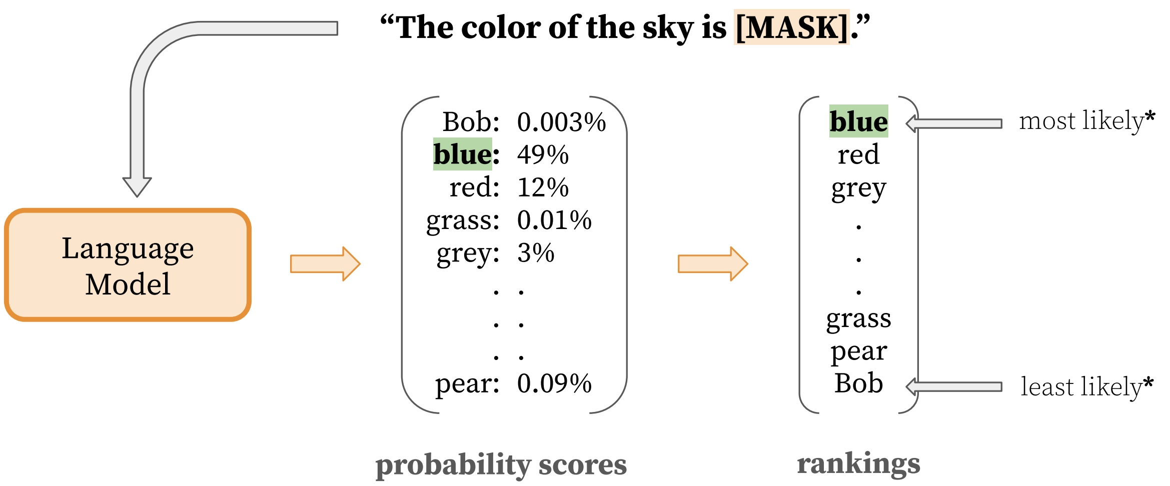


probability scores

*according to the LM

Evaluation of LM via LAMA

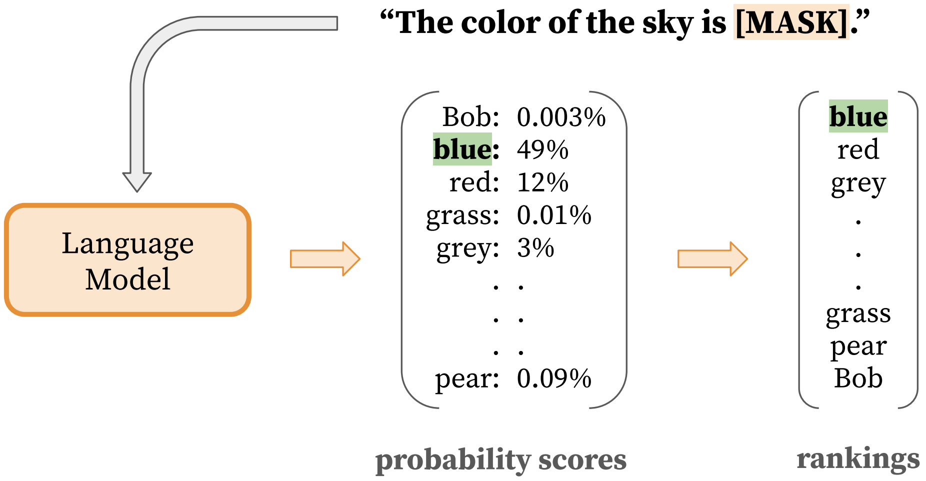
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



*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

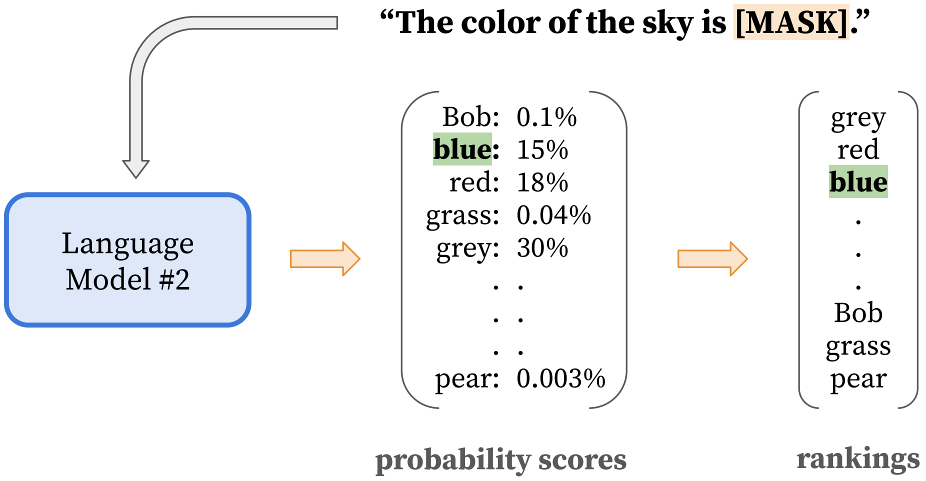
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower

“The color of the sky is [MASK].”

Language
Model #2

Evaluation of LM via LAMA

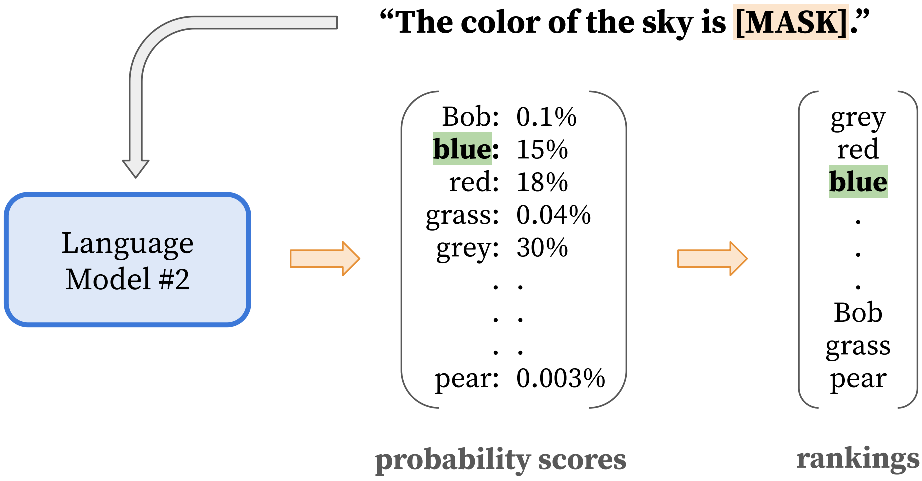
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



Evaluation of LM via LAMA

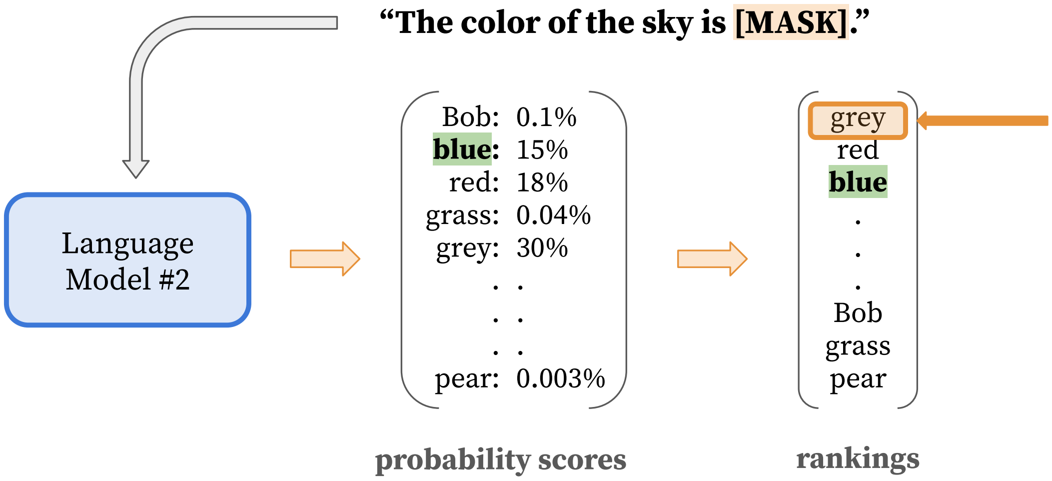
Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower

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

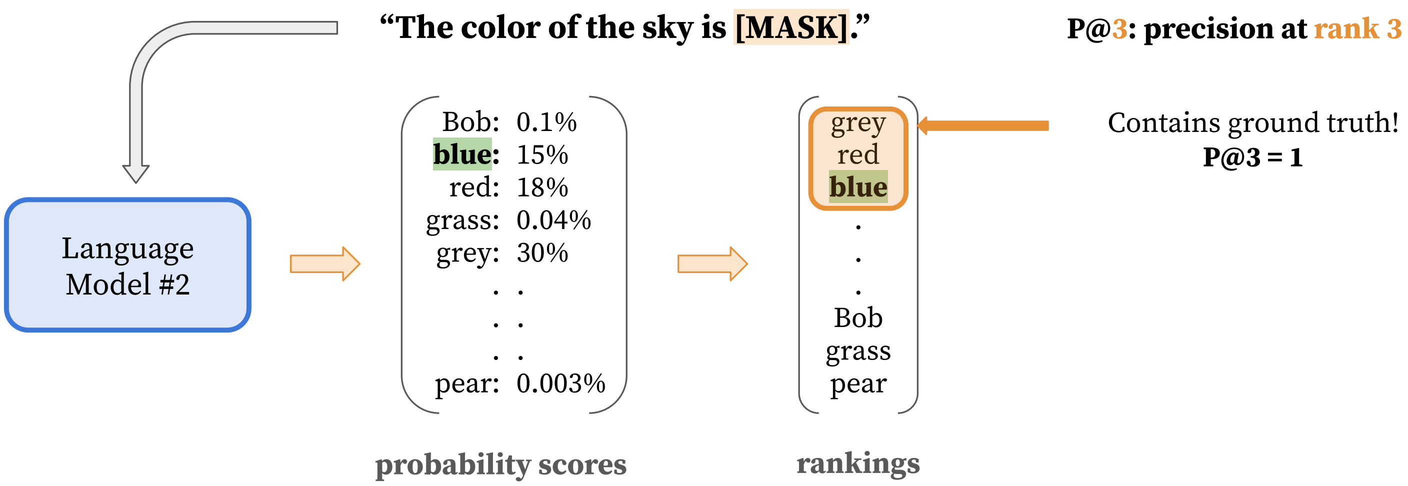


P@1: precision at rank 1

No ground truth.
P@1 = 0

Evaluation of LM via LAMA

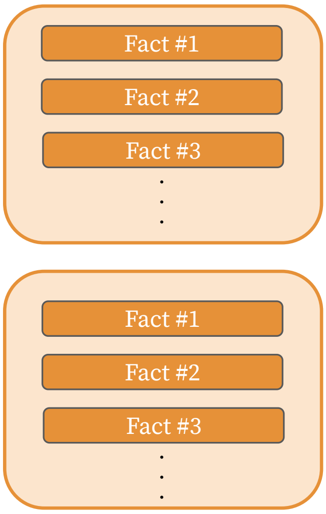
Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower



Architecture of the LAMA probe

Architecture of the LAMA probe

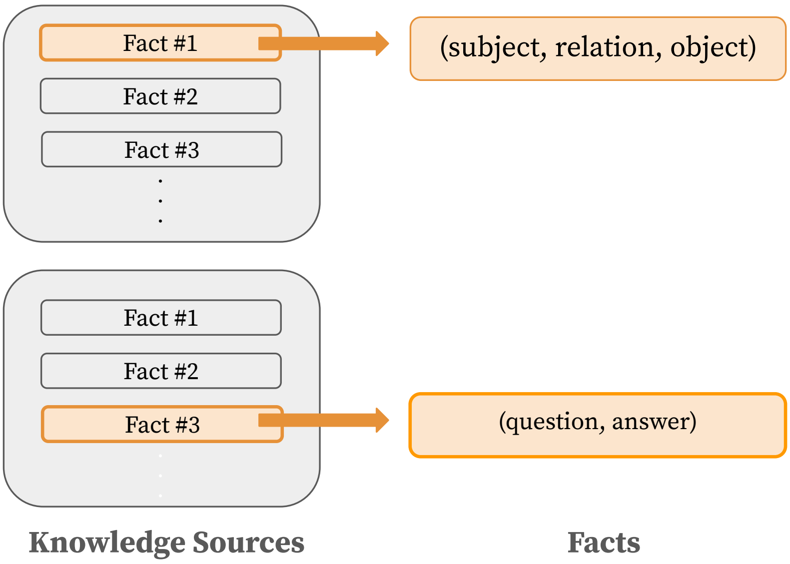
Step 1: Compile knowledge sources



Knowledge Sources

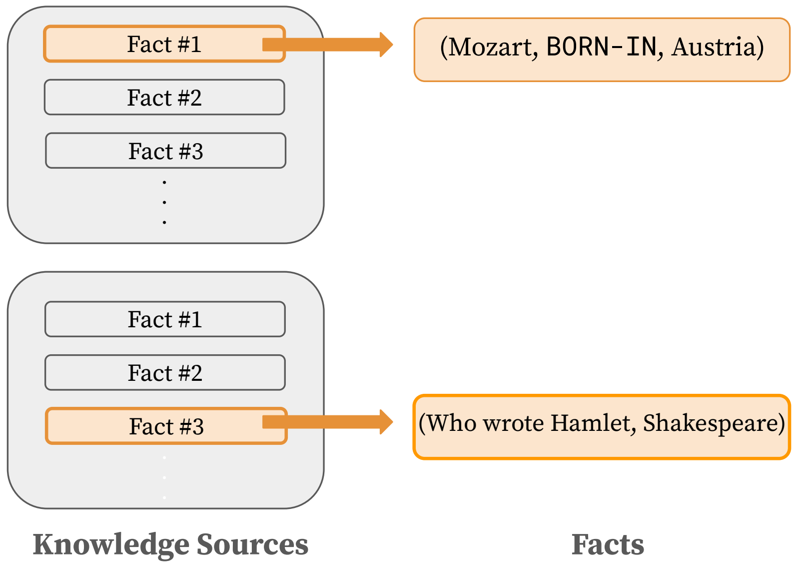
Architecture of the LAMA probe

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



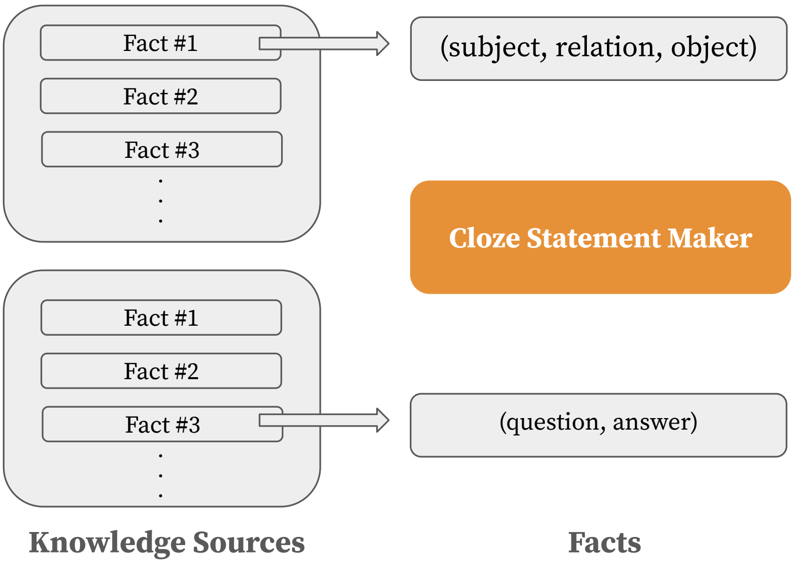
Architecture of the LAMA probe

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



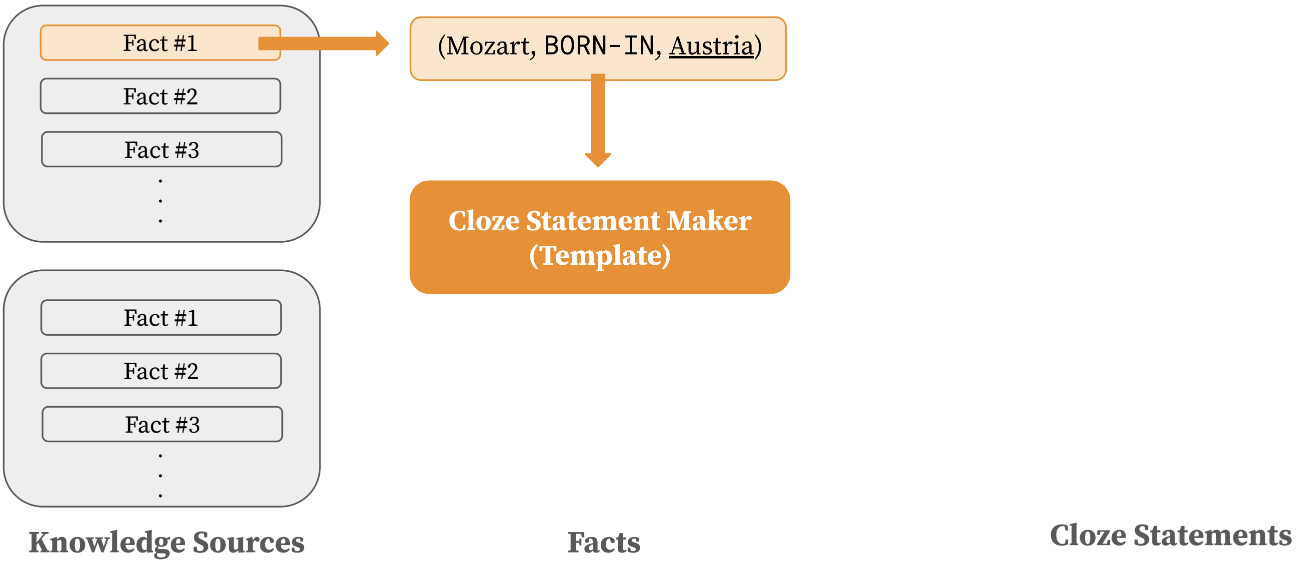
Architecture of the LAMA probe

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



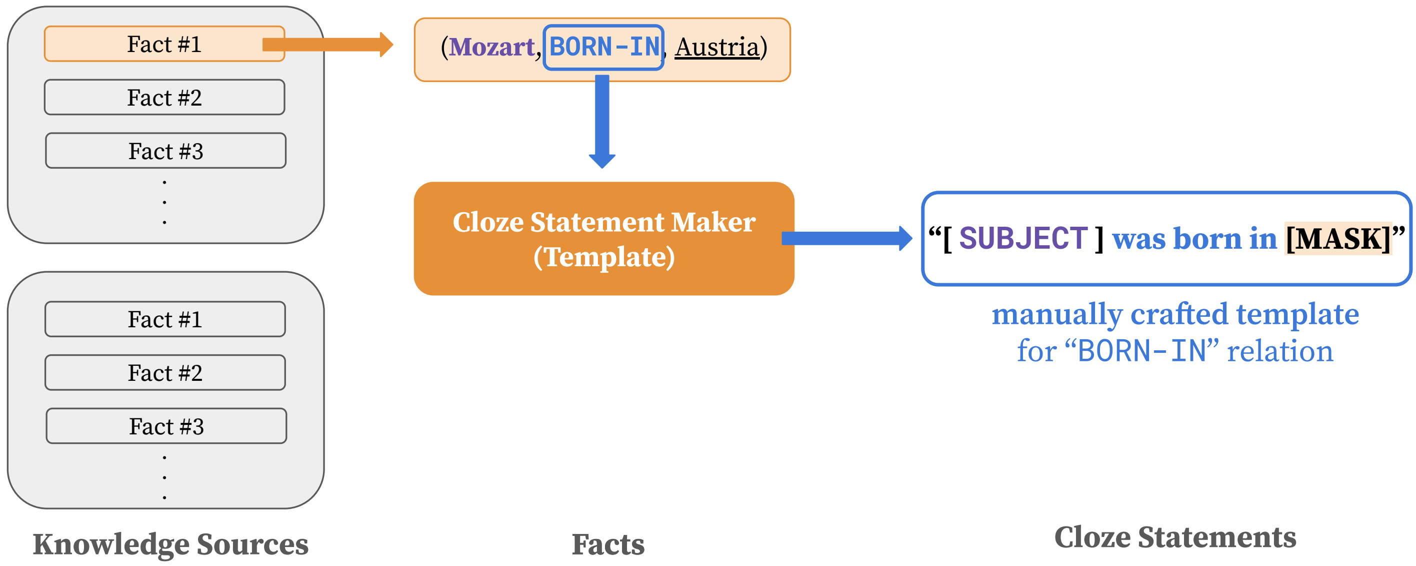
Architecture of the LAMA probe

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



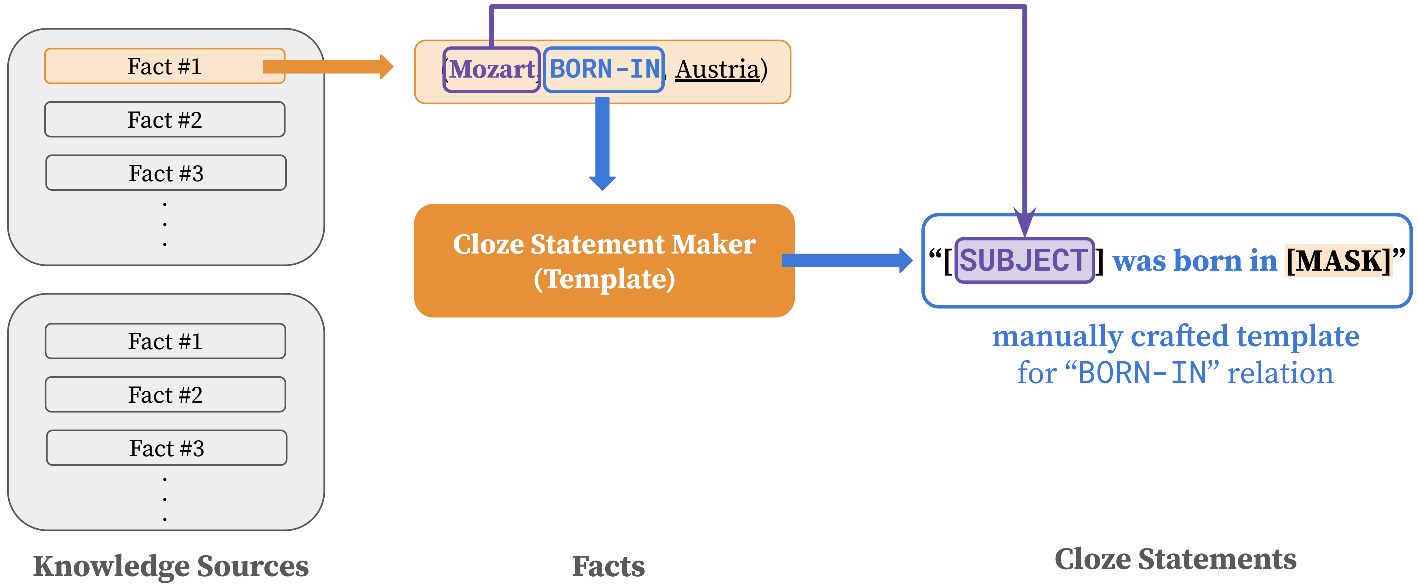
Architecture of the LAMA probe

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



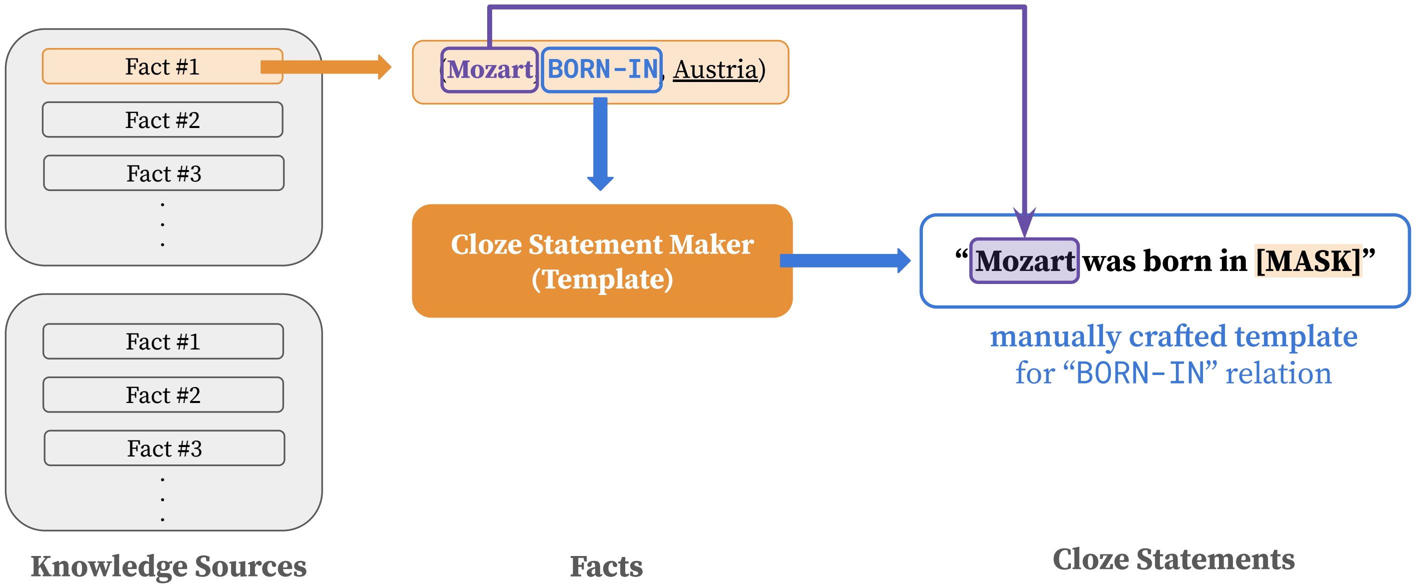
Architecture of the LAMA probe

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



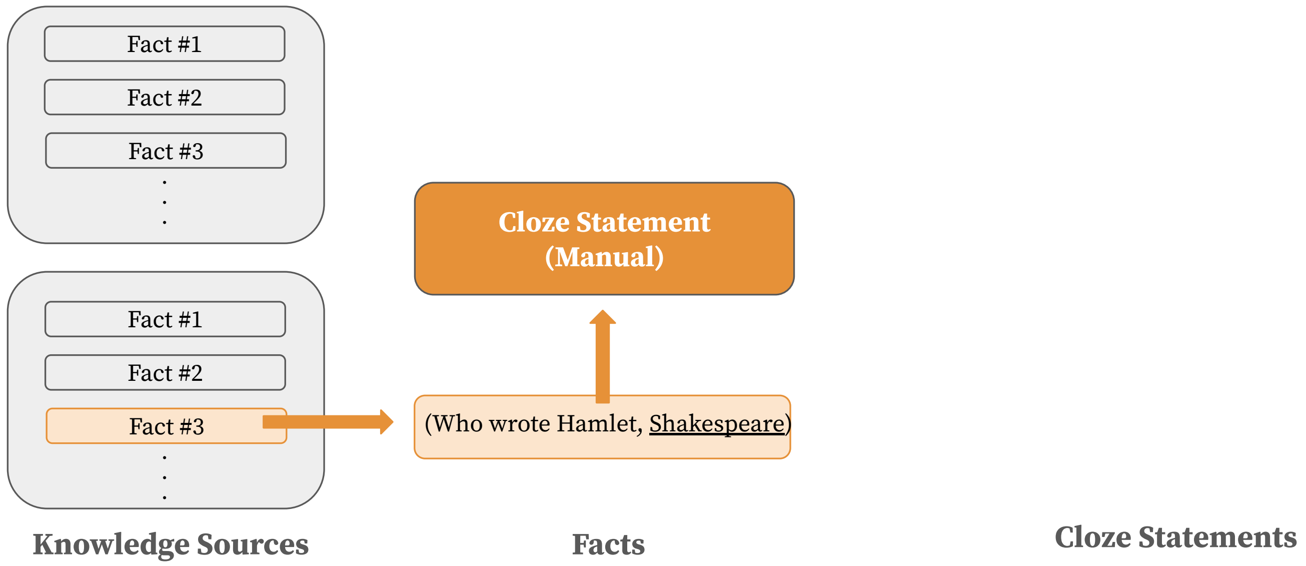
Architecture of the LAMA probe

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



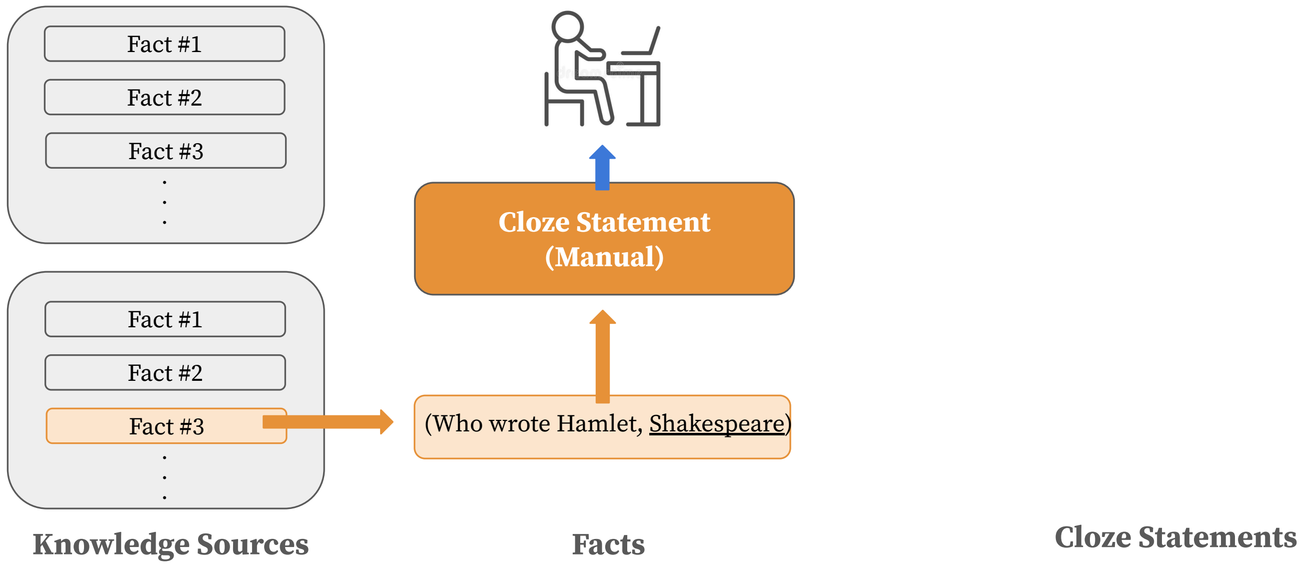
Architecture of the LAMA probe

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



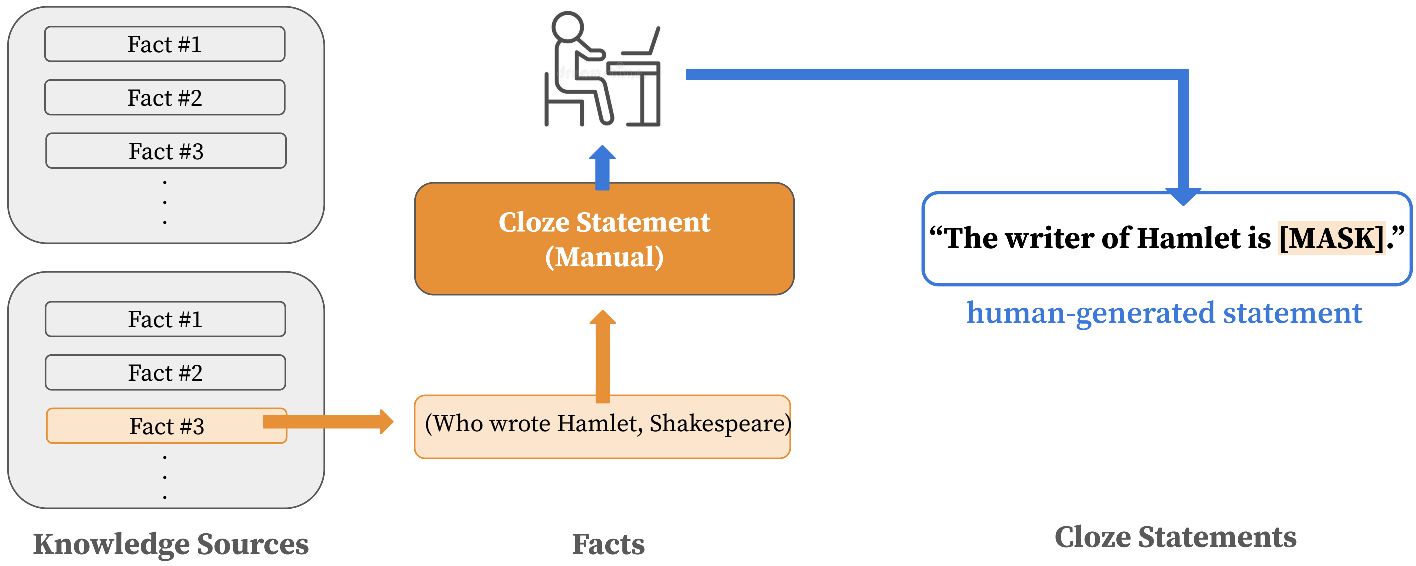
Architecture of the LAMA probe

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



Architecture of the LAMA probe

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



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

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)

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**

LAMA's Knowledge Sources: T-REx

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

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

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~~])

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**

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)

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)

OMCS Sentence: " Ravens can fly. "

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**

LAMA’s Knowledge Sources: **SQuAD**

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

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)

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

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

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**!

Baselines

Baselines

- **Freq:** ranks candidates by frequency of appearance as objects for a subject-relation pair
  - Analogous to majority classifier

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

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**

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**

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_o: uses oracle for entity linking

  - **DRQA** (Chen et al., 2017): uses TF/IDF to retrieve relevant arguments from a set of documents, then extracts answers from the best *k* articles

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

Table from Petroni et al., 2019

Results: both BERT models outperform other models on Google-RE

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

P@1: precision at rank 1

Results: BERT models does better on T-REx when there's only one correct answer...

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

P@1: precision at rank 1

Table from Petroni et al., 2019

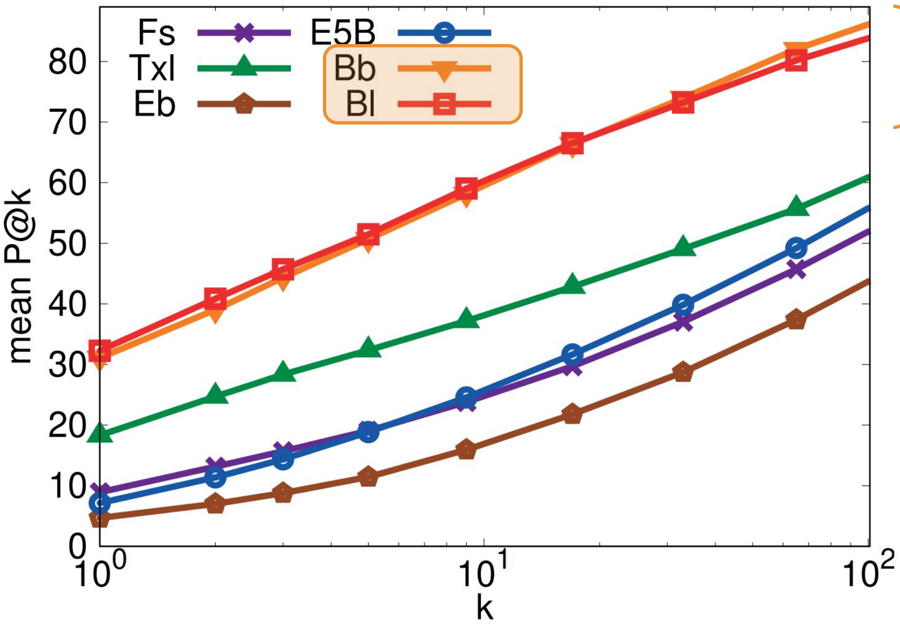
Results: ...but when there are multiple answers, RE_o is best

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

P@1: precision at rank 1

Results: BERT models outperform other LMs on T-REx

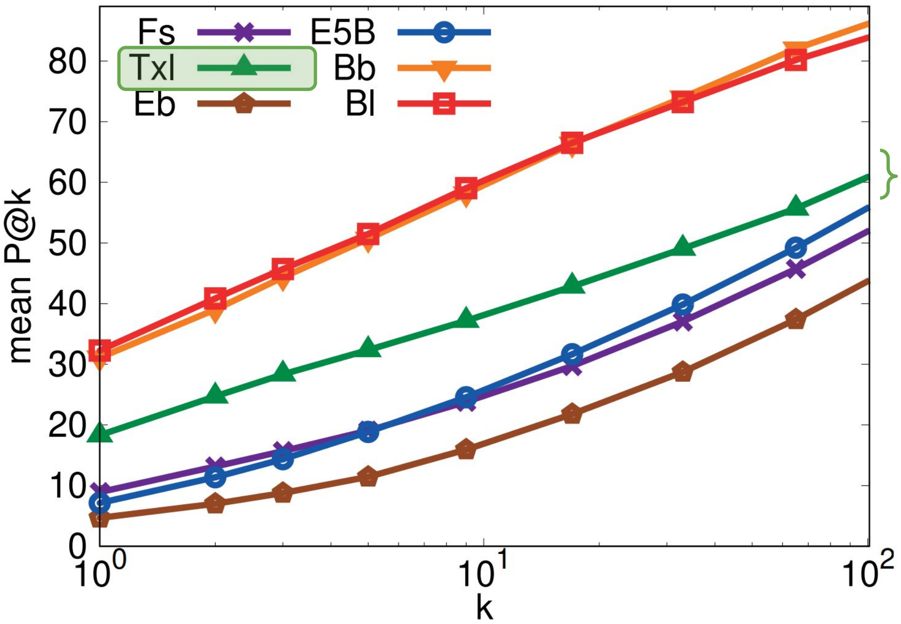
Performance on T-REx



BERT models perform the best by a large margin

Results: BERT models outperform other LMs on T-REx

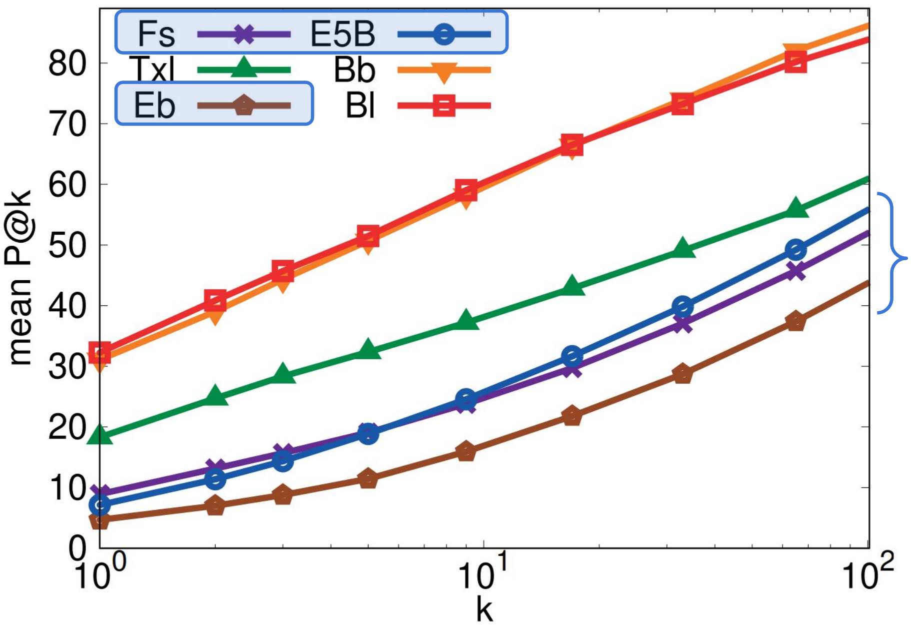
Performance on T-REx



Next best is Transformer-XL

Results: BERT models outperform other LMs on T-REx

Performance on T-REx



The worst performers are ELMo and fairseq-conv

Figure from Petroni et al., 2019

Results: BERT models are less sensitive to query variations than other LMs

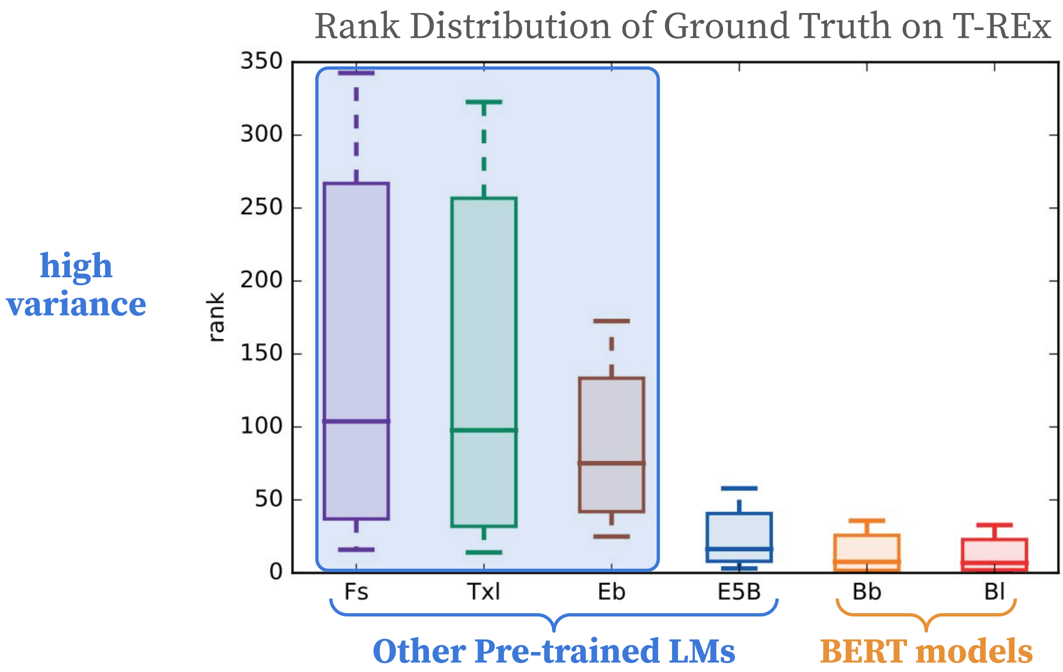


Figure from Petroni et al., 2019

Results: BERT models are less sensitive to query variations than other LMs

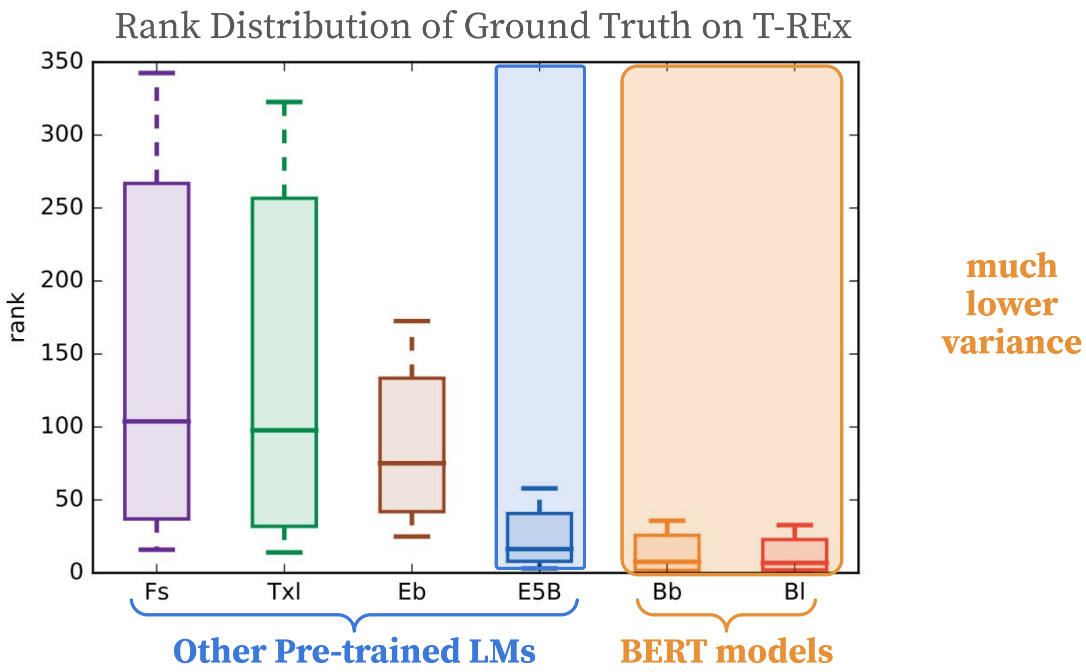
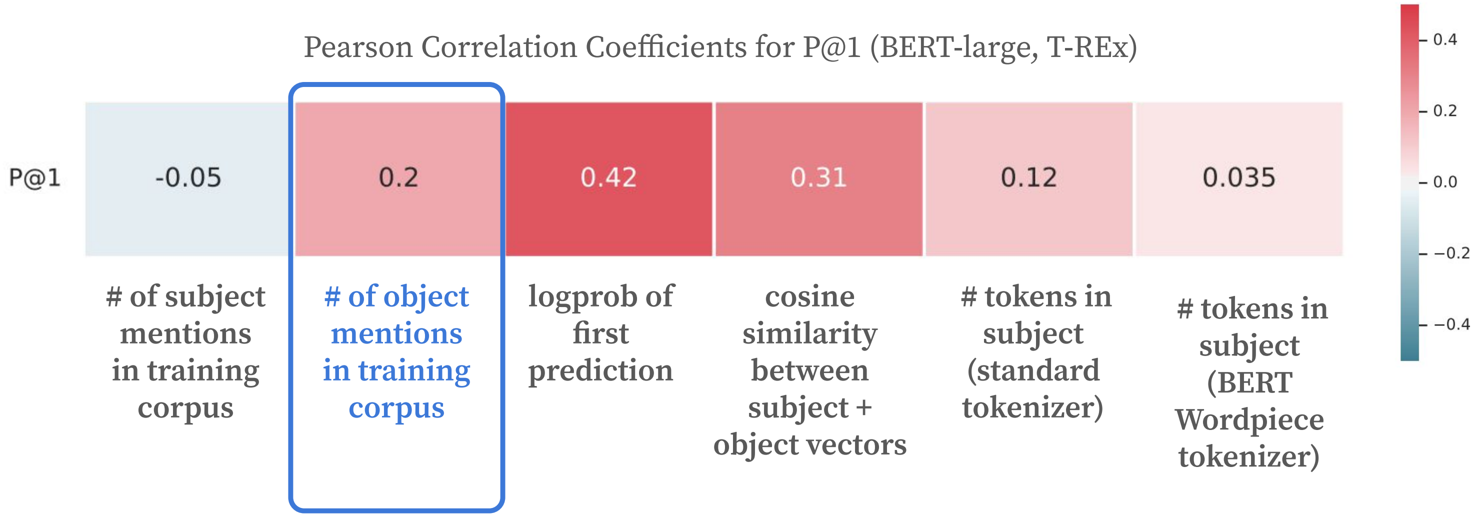


Figure from [*Petroni et al., 2019*]

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



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

Question 1

Describe what the LAMA Probe is in (Petroni et al., 2019) - How do they probe different knowledge sources (Wikidata triples, ConceptNet, QA pairs)?

- A collection of knowledge sources either for relation extraction or QA
- 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

Is there a better alternative to manual templates?

Is there a better alternative to manual templates?

- **Prompt mining (Jiang et al., 2020):** automated prompt extraction via dependency parsing, simple heuristics, or automated paraphrasing

Figure from [Shin et al., 2020]

Is there a better alternative to manual templates?

- **Prompt mining ([Jiang et al., 2020]):** automated prompt extraction via dependency parsing, simple heuristics, or automated paraphrasing

- **AutoPrompt ([Shin et al., 2020]):** prompt is constructed by adding tokens found via gradient-guided search to a simple prompt

Is there a better alternative to manual templates?

- **Prompt mining (Jiang et al., 2020):** automated prompt extraction via dependency parsing, simple heuristics, or automated paraphrasing

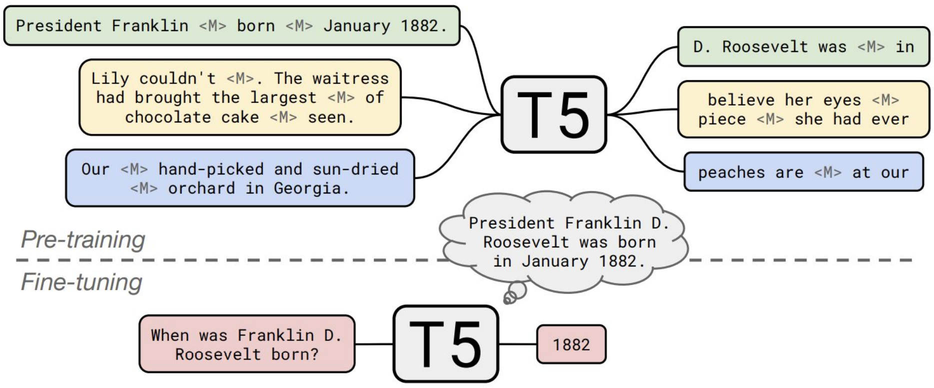
- **AutoPrompt (Shin et al., 2020):** prompt is constructed by adding tokens found via gradient-guided search to a simple prompt

- **OptiPrompt (Zhong et al., 2021):** directly optimize prompt in embedding space, rather than in discrete space
    - Similar to prompt tuning

How Much Knowledge Can You Pack Into the Parameters of a Language Model? (Roberts et al., 2020)

Motivation

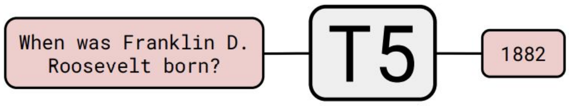
- Petroni et al., 2019 measures knowledge in a model with its ***pre-training objective*** with ***a synthetic task***
- This work measures ***transfer learning performance*** on knowledge on ***question answering tasks with a closed-book approach***



To solve open-domain QA: Two approaches

- Open-book QA: **questions** and **external resources** are given

- Closed-book QA: **questions**



Experimental Setup - Datasets

- **Natural Questions:**
  - Web queries each with ~~an oracle wikipedia article~~
  - Rich annotation, different answer types (yes/no, ~~unanswerable~~, multiple answers, short answer and ~~long answer~~)
  - E.g., *Who are the members of the Beatles?*
- **WebQuestions**
  - Web queries
  - E.g., *Which college did Obama go to?*
- **TriviaQA**
  - Questions from quiz league websites, each with ~~web pages that might contain answers~~
  - E.g., *Who won the Nobel Peace Prize in 2009?*

Experimental Setup - Models

- **Pre-training resources**
  - **T5 v1.0**: trained with the unsupervised "span corruption" task on C4 as well as *supervised translation, summarization, classification, and reading comprehension tasks*
  - **T5 v.1.1**: trained only with the C4

- **Model size**
  - Base (220 million parameters)
  - Large (770 million)
  - 3B (3 billion)
  - 11B (11 billion)

- **Additional pre-training**
  - Salient Span Masking (Guu et al. 2020), mask salient spans (named entities & dates)
  - Continue pre-training the T5 for 100k steps

person location

Henri Hutin invented Brie cheese while living in North of Meuse, France

Results

Metric: Exact Match

SOTA Retrieval-based Models
(can access external documents)

	NQ	WQ	TQA dev	TQA test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Févry et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–

Results: SSM clearly leads to improved performance

Metric: Exact Match

		NQ	WQ	TQA dev	TQA test	
SOTA Retrieval-based Models (can access external documents)	Chen et al. (2017)	–	20.7	–	–	
	Lee et al. (2019)	33.3	36.4	47.1	–	
	Min et al. (2019a)	28.1	–	50.9	–	
	Min et al. (2019b)	31.8	31.6	55.4	–	
	Asai et al. (2019)	32.6	–	–	–	
	Ling et al. (2020)	–	–	35.7	–	
	Guu et al. (2020)	40.4	40.7	–	–	
	Févry et al. (2020)	–	–	43.2	53.4	
	Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–	
Closed-Book QA models with fine-tuning (relies only on internal parameters)	T5-Base	25.9	27.9	23.8	29.1	non-SSM
	T5-Large	28.5	30.6	28.7	35.9	
	T5-3B	30.4	33.6	35.1	43.4	
	T5-11B	32.6	37.2	42.3	50.1	
	T5-11B + SSM	34.8	40.8	51.0	60.5	SSM
	T5.1.1-Base	25.7	28.2	24.2	30.6	non-SSM
	T5.1.1-Large	27.3	29.5	28.5	37.2	
	T5.1.1-XL	29.5	32.4	36.0	45.1	
	T5.1.1-XXL	32.8	35.6	42.9	52.5	
	T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**	SSM
Closed-Book QA model without fine-tuning	GPT-3 few-shot	29.9	41.5	**71.2**	-	
SOTA Retrieval-based Models	SOTA	51.4	-	80.1	-	

Results: Scale correlates with performance

Metric: Exact Match

	NQ	WQ	TQA dev	TQA test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Févry et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–
T5-Base	25.9	27.9	23.8	29.1
T5-Large	28.5	30.6	28.7	35.9
T5-3B	30.4	33.6	35.1	43.4
T5-11B	32.6	37.2	42.3	50.1
T5-11B + SSM	34.8	40.8	51.0	60.5
T5.1.1-Base	25.7	28.2	24.2	30.6
T5.1.1-Large	27.3	29.5	28.5	37.2
T5.1.1-XL	29.5	32.4	36.0	45.1
T5.1.1-XXL	32.8	35.6	42.9	52.5
T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**

increasing size

increasing performance

increasing size

increasing performance

Additional Evaluation on NQ

- Previous results adopt evaluation used in previous work
  - Long answers and unanswerable questions are not considered
  - Only output single answer
  - Only trained with the first answer if a question has multiple answers
  - Answers with longer than 5 tokens are excluded
  - Answers are normalized (lowercased, strip of articles, punctuation etc.)
- **Leaderboard evaluation**
  - Long answers and unanswerable questions are not considered
  - Models are trained to predict all ground-truth answers
  - Only considered correct if it predicts ***all answers*** correctly

T5-11B + SSM achieves a recall of **36.2** on the validation set, which lags behind the state-of-theart score of **51.9** from Pan et al. (2019) at the time.

Human Evaluation + Qualitative Error Analysis

- Exact Match is a very harsh metric → potentially lots of **false negatives**
- Use human evaluation to see what percent of predicted negatives area are actually true negatives

38% of T5's "incorrect" predictions are actually correct!

		Example		
Category	Percentage	Question	Target(s)	T5 Prediction
True Negative	62.0%	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
Phrasing Mismatch	13.3%	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
Incomplete Annotation	13.3%	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	11.3%	who is the secretary of state for northern ireland	karen bradley	james brokenshire

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** – needed largest (11B) model to compete on par with SOTA

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

Question 2

Compared to (Petroni et al., 2019), can you state the key differences in (Roberts et al., 2021)?

- (Roberts et al., 2021) handles harder questions that may require multiple tokens. LAMA uses specific/easier types of questions with single-token answers
- Since T5 can't do zero-shot well, (Roberts et al., 2021) fine-tunes the model for QA tasks and compares against other retrieval-based fine-tuned models. LAMA does not fine-tune the models.

Do you think the accuracy of answering these open-domain questions reflects how much knowledge is already encoded in LLMs?

- To some extent. (Roberts et al., 2021) fine-tunes the model on the question-answer datasets, so it could be argued that it does not 100% accurately test how much knowledge is encoded in the pre-training stage

Comparison of the Two Works

	Petroni et al., 2019	Roberts et al., 2020
Objective	MLM	seq2seq
Format	filling in the blank	generation
Finetune?	no	yes
Answer length	1	> 1

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?

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!

How to update knowledge in pre-trained models?

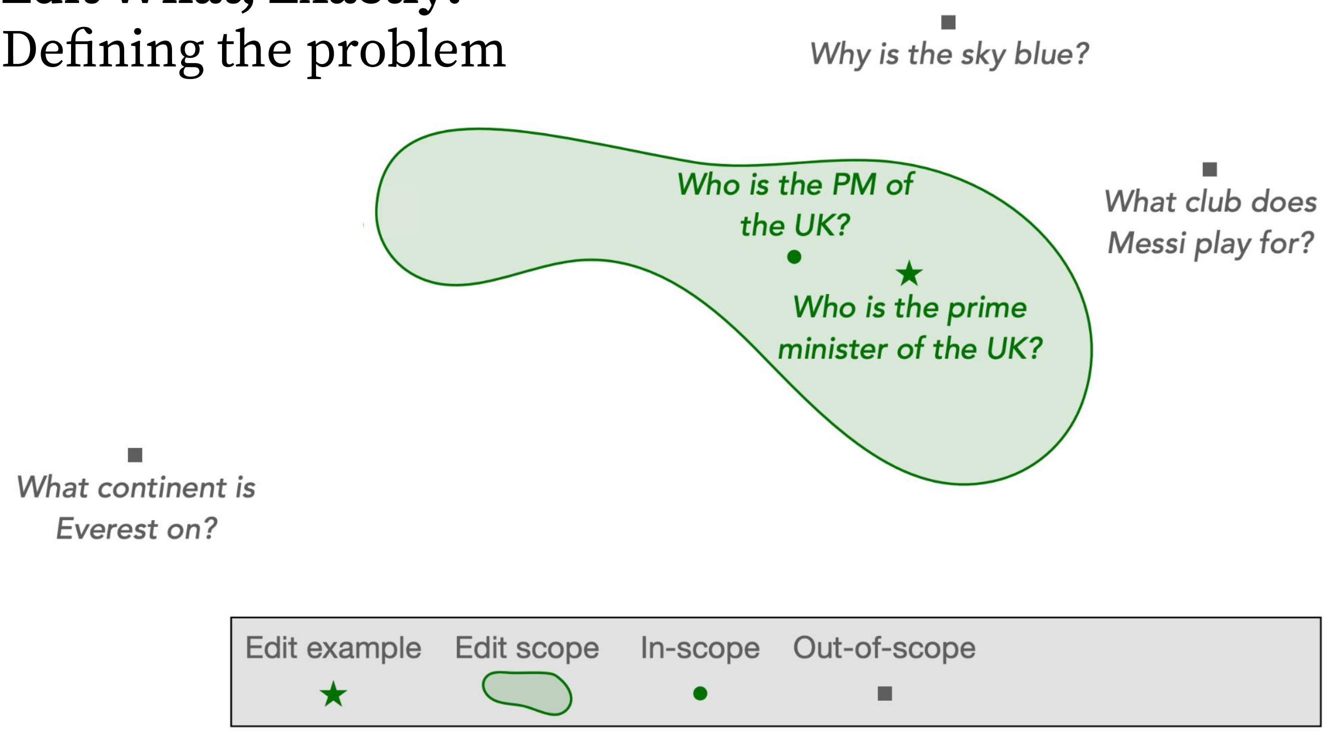
Edit What, Exactly?
Defining the problem

Slides from [link]

Edit What, Exactly?
Defining the problem

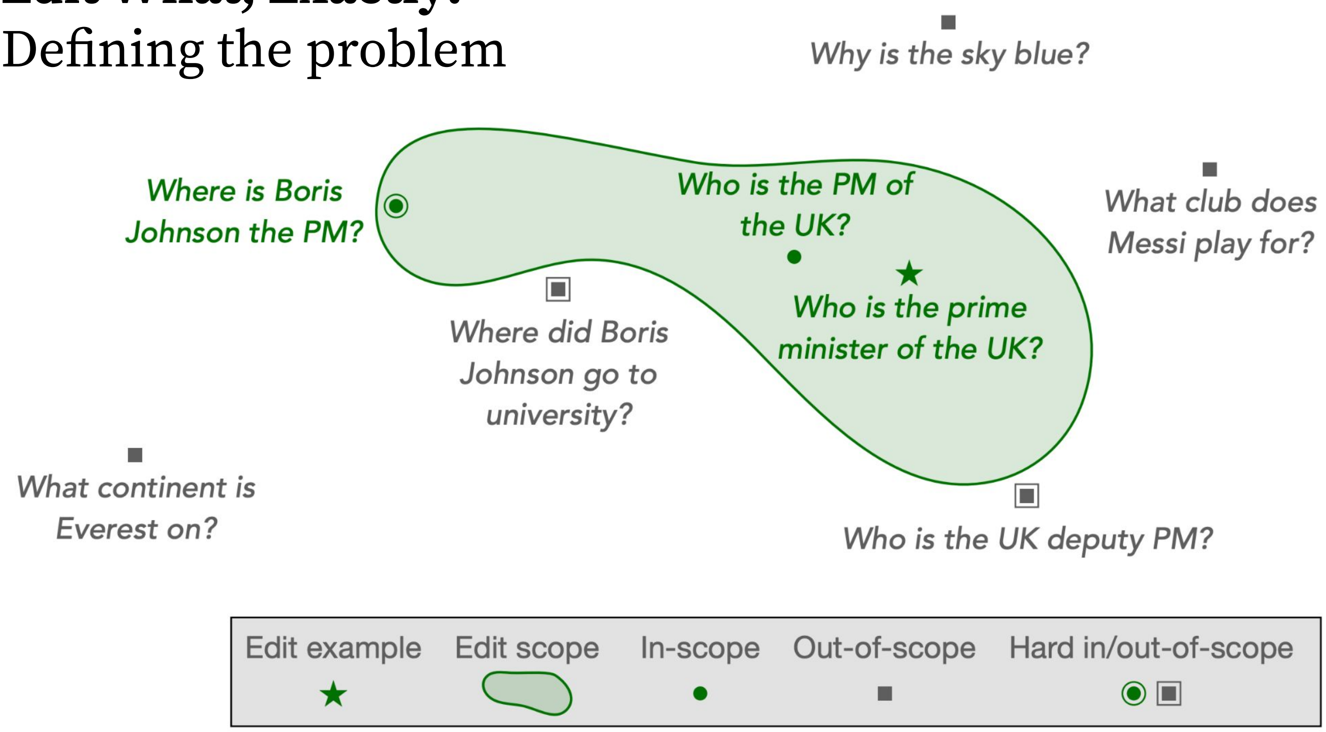
Slides from [link]()

Edit What, Exactly?
Defining the problem



Slides from [link]()

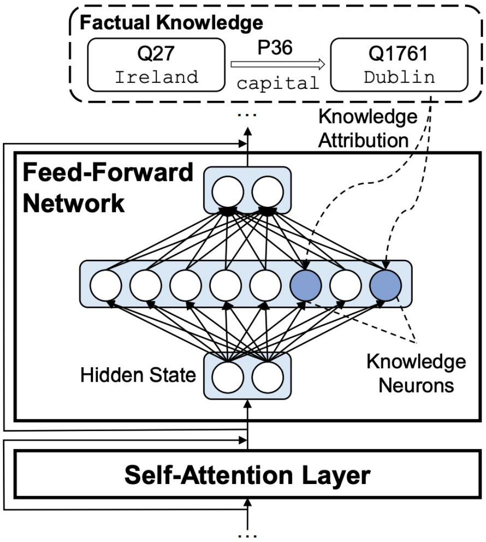
Edit What, Exactly?
Defining the problem



Slides from [link]()

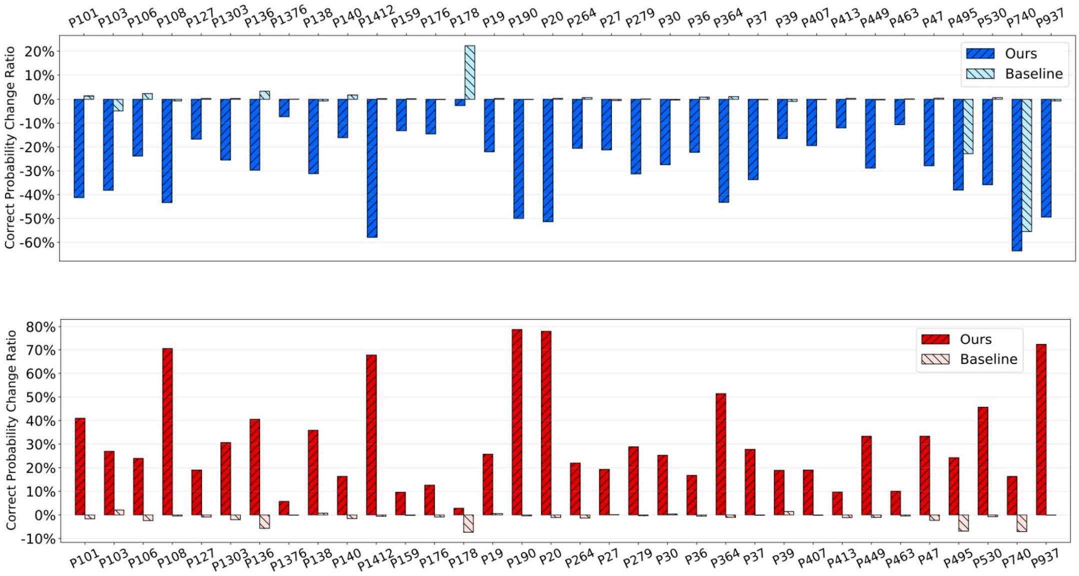
Knowledge Neurons in Pretrained Transformers (Dai et al. 2021)

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.

Suppressing or Amplifying Knowledge Neurons



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

Case Study - Updating Facts

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

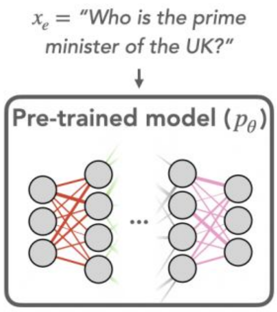
Metric	Knowledge Neurons	Random Neurons
Change rate↑	48.5%	4.7%
Success rate↑	34.4%	0.0%

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

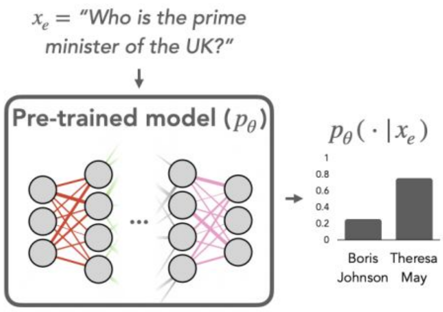
Fast Model Editing at Scale (Mitchell et al. 2022)

127

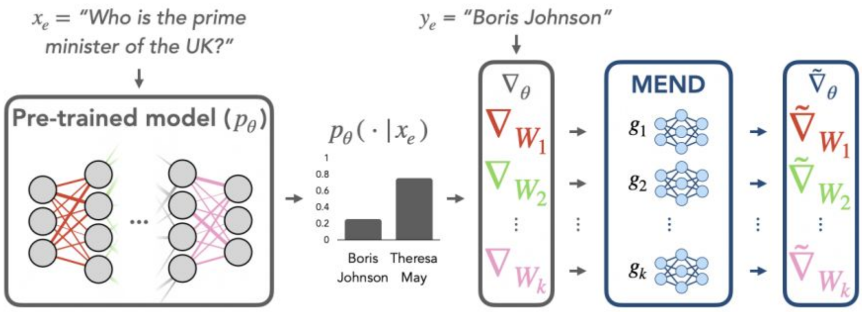
Editing a Pre-trained Model with MEND



Editing a Pre-trained Model with MEND

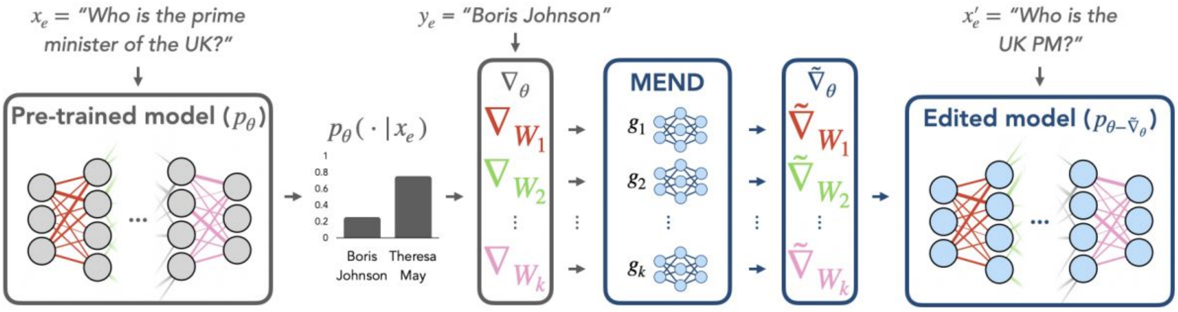


Editing a Pre-trained Model with MEND

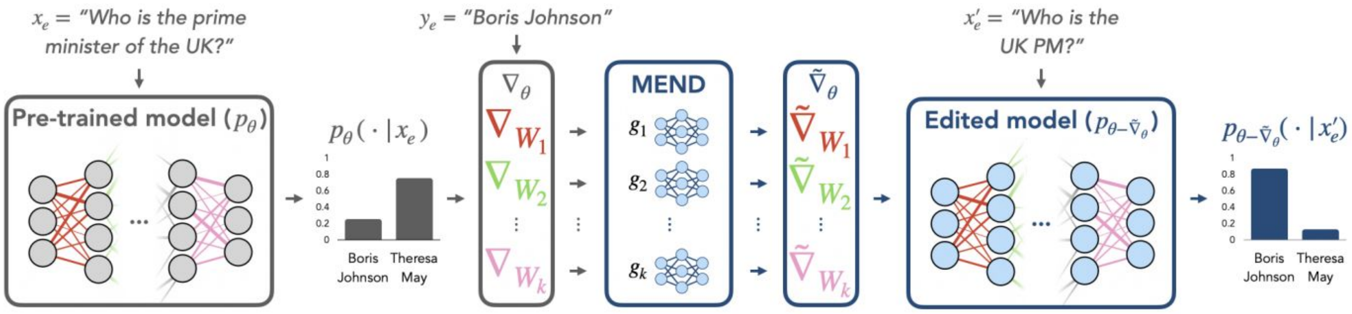


- The MEND network produces **gradient updates** for the pretrained model.
- It's not the gradient of all the weights, it's a **transformation** of the gradient!

Editing a Pre-trained Model with MEND

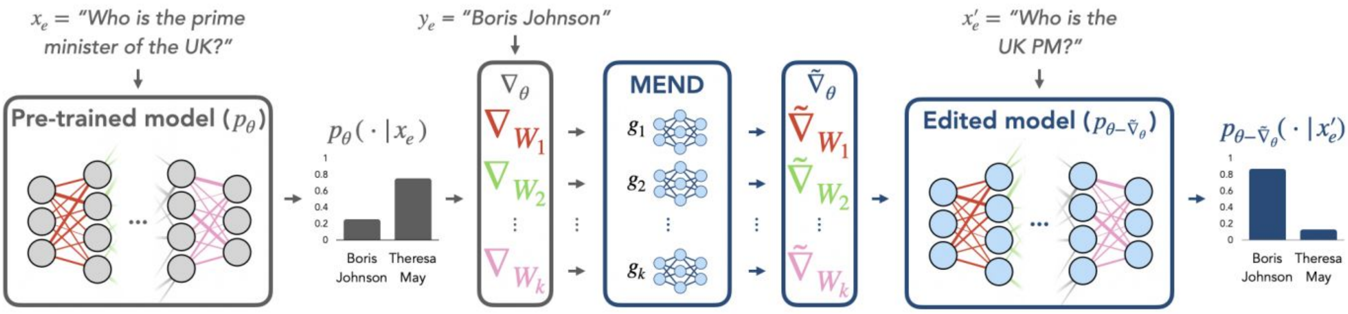


Editing a Pre-trained Model with MEND



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

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

Conclusion

Question 3

The world knowledge is constantly changing; for instance, the president was Donald Trump in 2020 and now is Joe Biden in 2022. However, LLMs are always trained on a static corpus of a fixed period.

1) Do you have any ideas about how to update and edit LLMs with real-world knowledge?

2) Do you think it is possible to decouple world knowledge and other knowledge encoded in LLMs?