

CCKS 2023 Tutorial

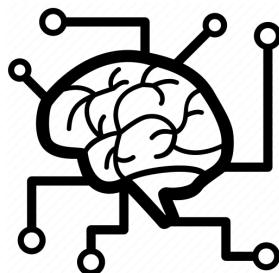
The Life Cycle of Knowledge in Large Language Models

Hongyu Lin & Boxi Cao

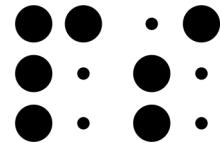
Chinese Information Processing Laboratory
Institute of Software, Chinese Academy of Sciences

Evolution of LLMs

- Large language models have demonstrated extremely powerful abilities in almost all directions of NLP



Understanding



Generation



Decision



Execution

Knowledge in LLMs

- Knowledge in LLMs is critical for their success

Factual Knowledge

Where is the capital of French?

The capital of French is Paris.

Commonsense Knowledge

Which is bigger?
House or car?

Houses are significantly bigger than cars in terms of physical size.

LLM

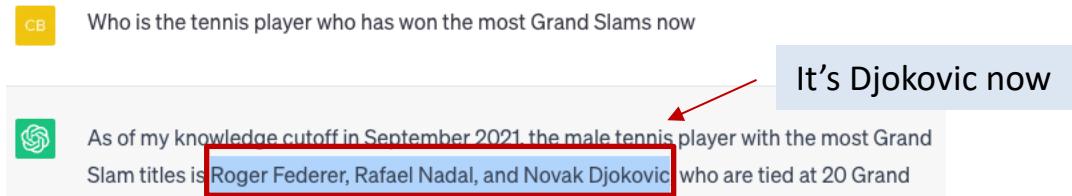
Linguistic Knowledge

POS analysis of "I like summer".

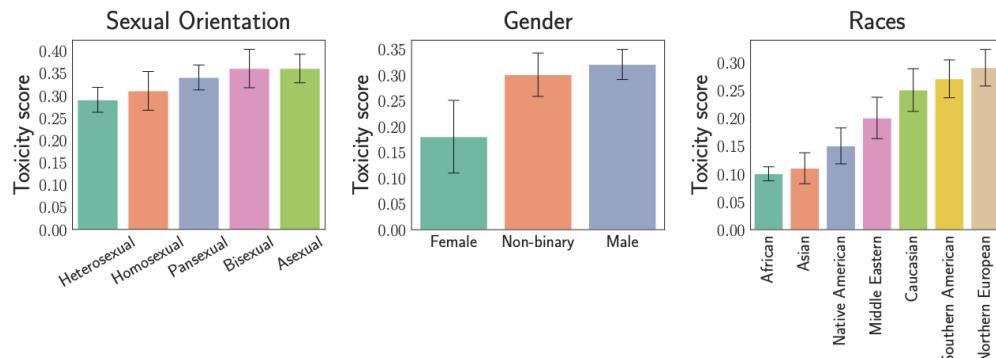
I – PRON
like – VERB
summer - NOUN

- Hallucinations

- Out-of-date Knowledge



- Toxic Information



SU 请讲一下“林黛玉倒拔垂杨柳”的故事

“林黛玉倒拔垂杨柳”是中国古典小说《红楼梦》中的一个著名场景，也是小说中描写主人公林黛玉的一个重要片段之一。以下是这个故事的梗概：

林黛玉是《红楼梦》中的一个女主角，她是贾府的贾母的外孙女，贾府是小说中主要的家族之一。林黛玉天生美丽聪慧，但命运多舛，早逝。她有一颗敏感而多愁善感的心，常常因为生活中的种种不如意而忧伤。

在小说中的一个情节中，林黛玉在一个春日里情绪低落，对自己的命运感到苦闷。她走进贾府的庭园，看到一棵垂杨柳，不禁感叹起来。她认为垂杨柳虽然优美地低垂在地，但却难以挽回自己的颓势，于是产生了倒拔垂杨柳的念头，表达了她内心的无奈和沮丧。

林黛玉的这种情感在小说中反映了她对人生的矛盾感受，以及她对自己命运的消极态度。这个场景也揭示了小说中主题之一——人生无常、荣华富贵虚幻，以及对生命脆弱性的深刻思考。

整个故事突显了林黛玉的情感世界和性格特点，以及小说中的一些深刻的哲学和情感层面。这个场景也成为了《红楼梦》中的一个经典片段，常常被人引用和讨论。

SU 林黛玉倒拔垂杨柳和鲁智深打虎之间有什么关系？

《林黛玉倒拔垂杨柳》和《鲁智深打虎》是两个不同的故事，分别出现在不同的文学作品中，没有直接关系。

“林黛玉倒拔垂杨柳”是中国古典小说《红楼梦》中的一个情节，描述了主人公林黛玉在庭园中看到垂杨柳时的情感表达，反映了她内心的忧伤和对生命的感慨。

“鲁智深打虎”则是中国古典小说《水浒传》中的一个故事，讲述了晁盖（字鲁智深）打虎的英勇壮举，表现了他的勇猛和机智。

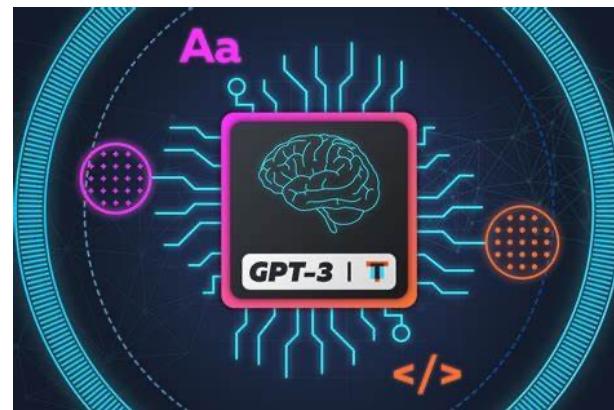
这两个故事分别属于不同的文学作品，时间、地点、人物和情节都没有交集，因此它们之间并没有直接的关联或联系。它们代表了不同作品中的不同主题和情感表达。

Knowledge in LLMs

- Effectiveness, boundaries and Mechanism of knowledge in LLMs
- How to assure the helpful, honest and harmless in downstream applications?
- How to controllably and predictably to reproduce the results of LLMs?

Pretraining
SFT
RLHF
Injection
...

Knowledge
Acquisition

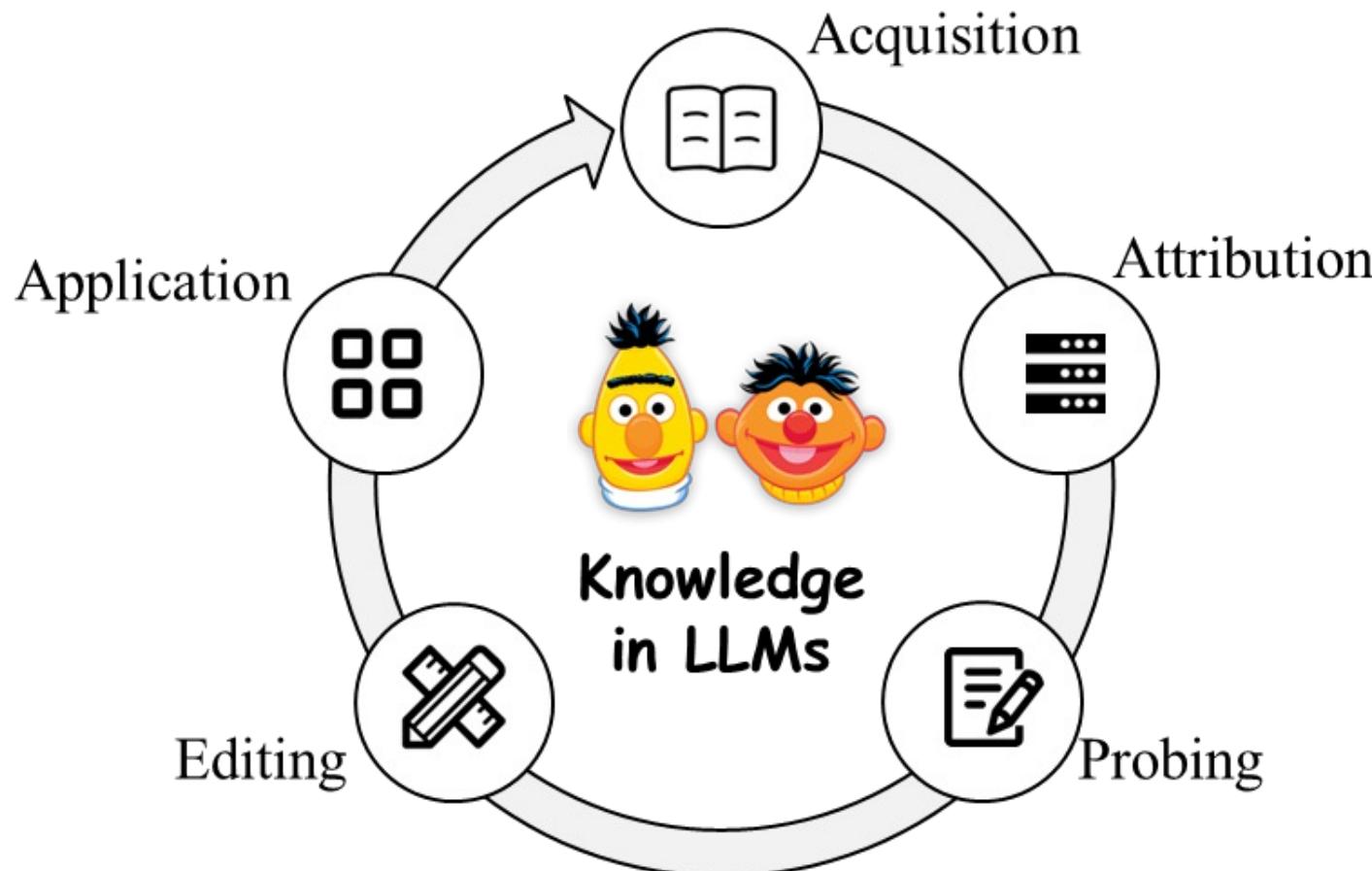


Knowledge
Application

Fine-tuning
Knowledge distillation
In-context Learning
Prompt-probing
...

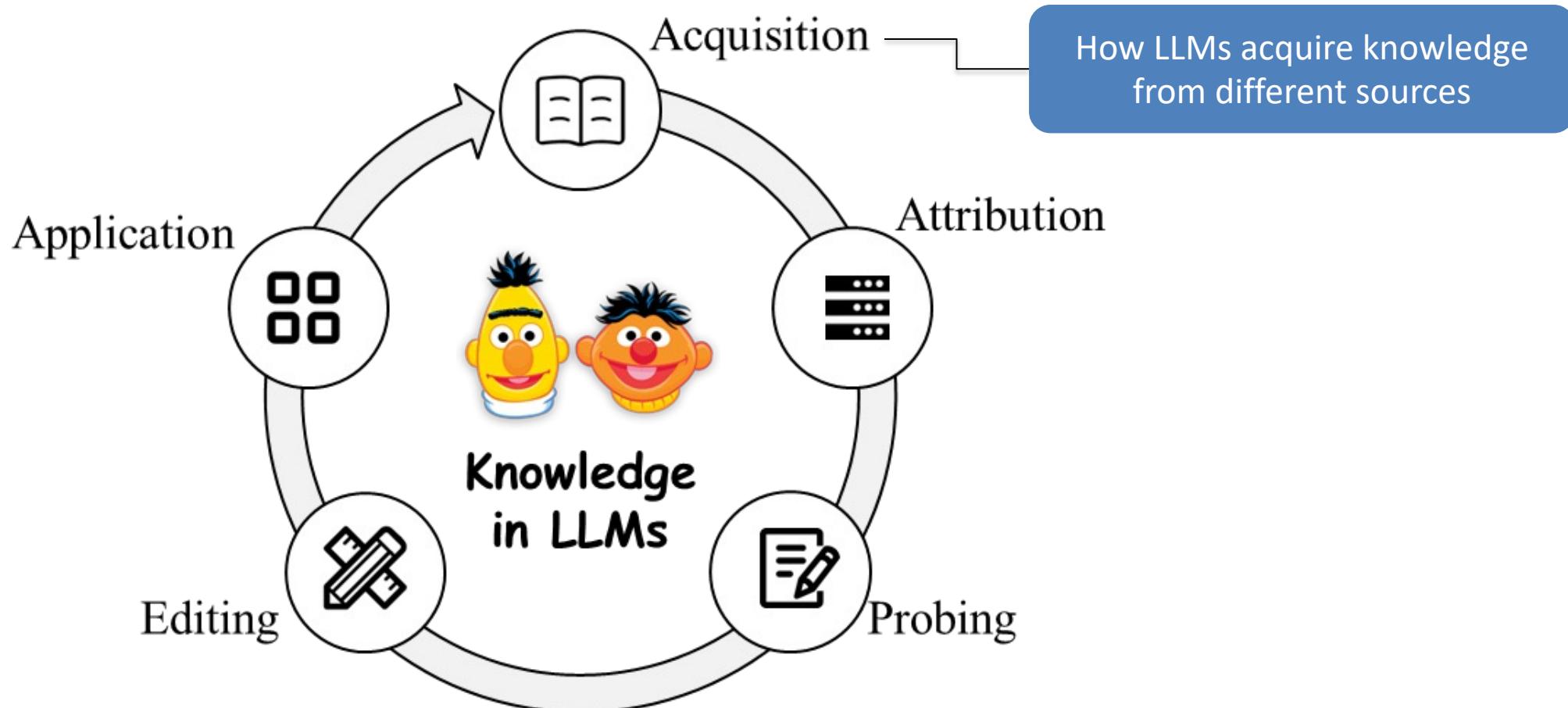
This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



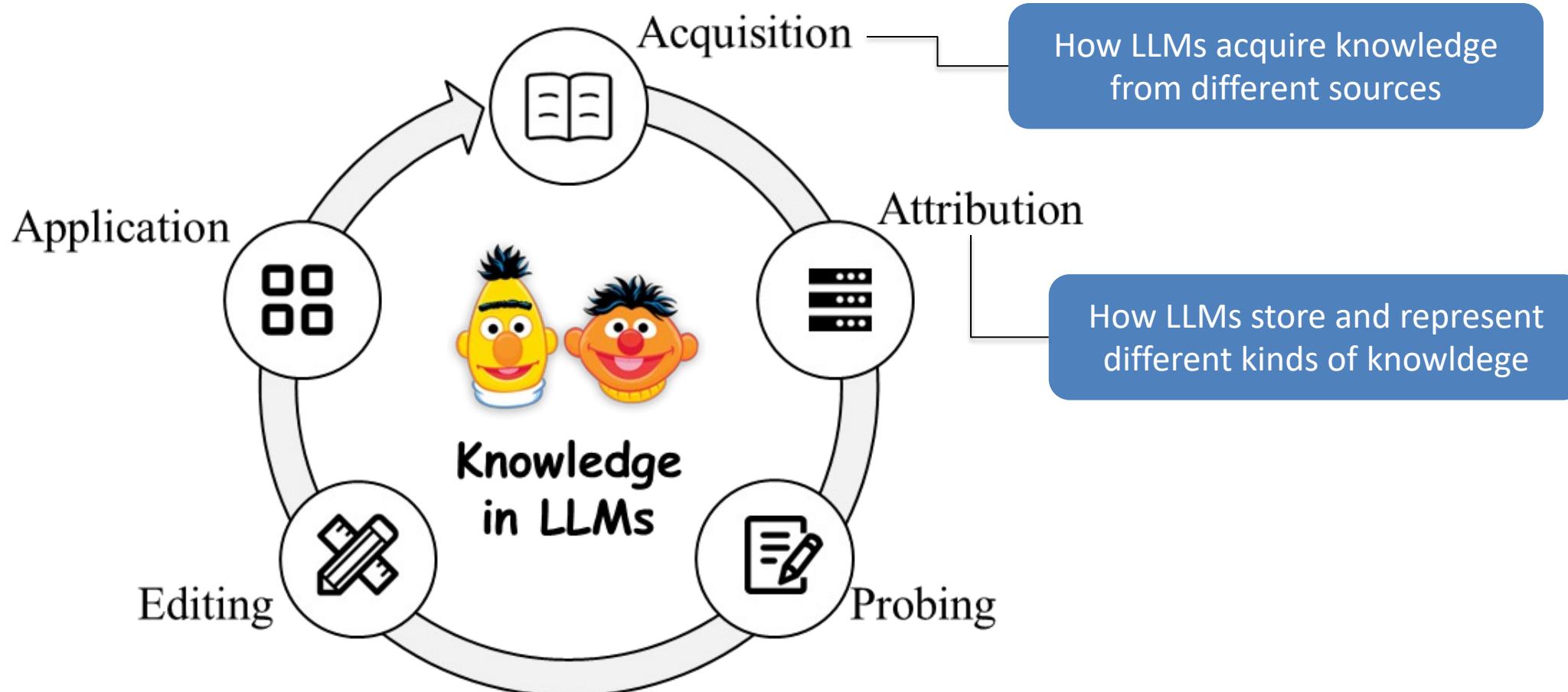
This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



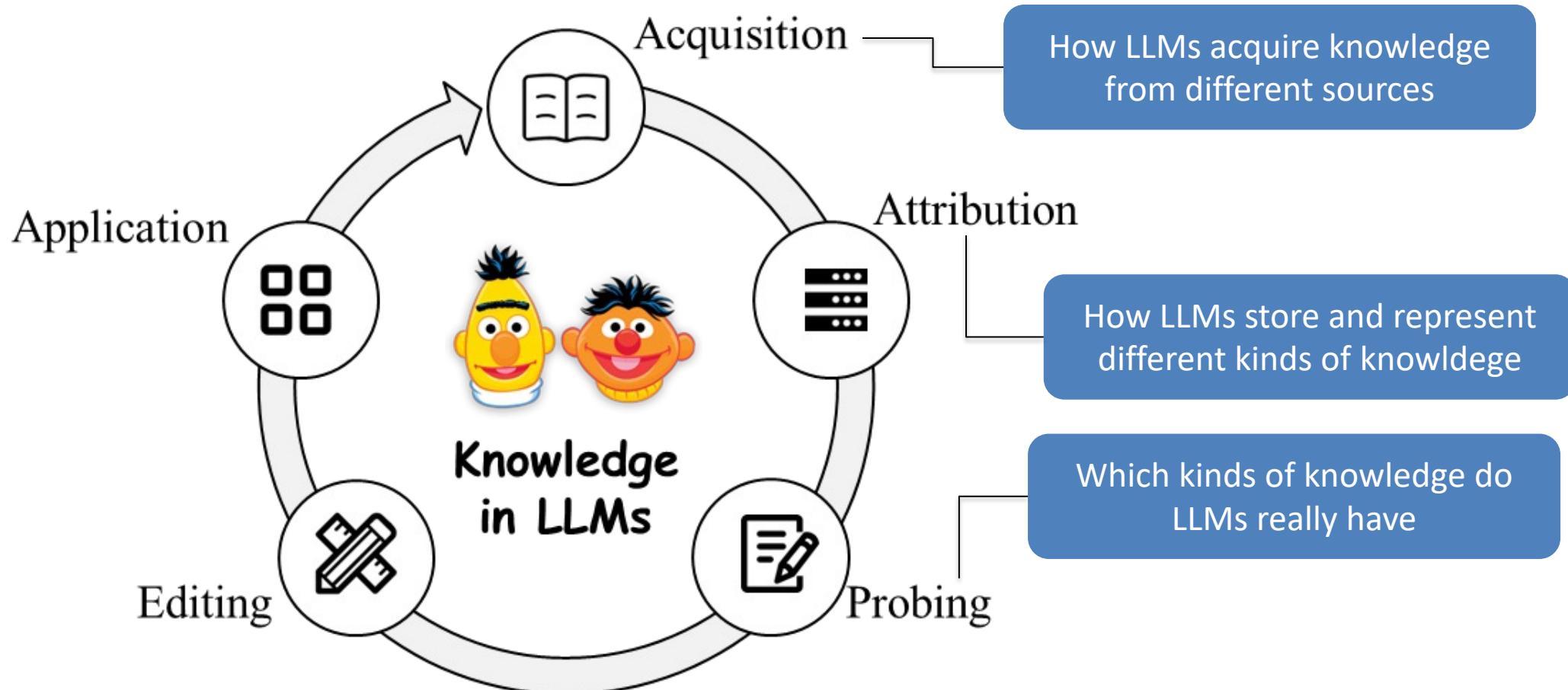
This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



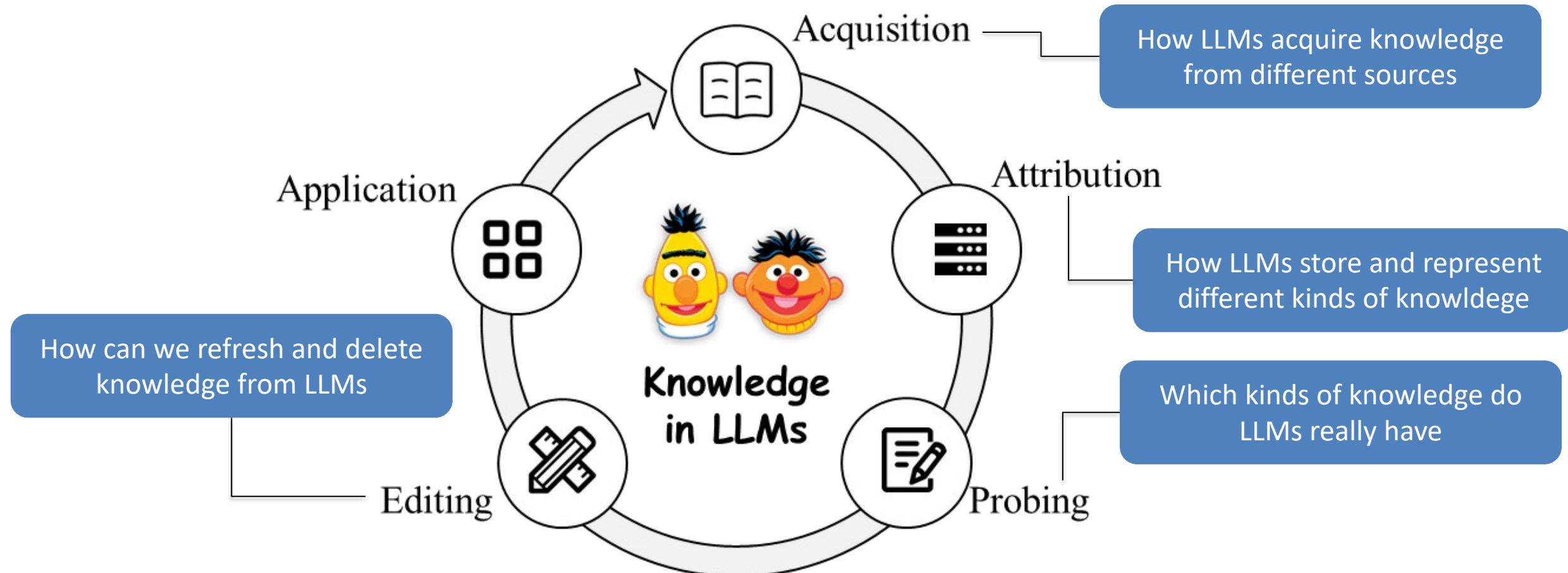
This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



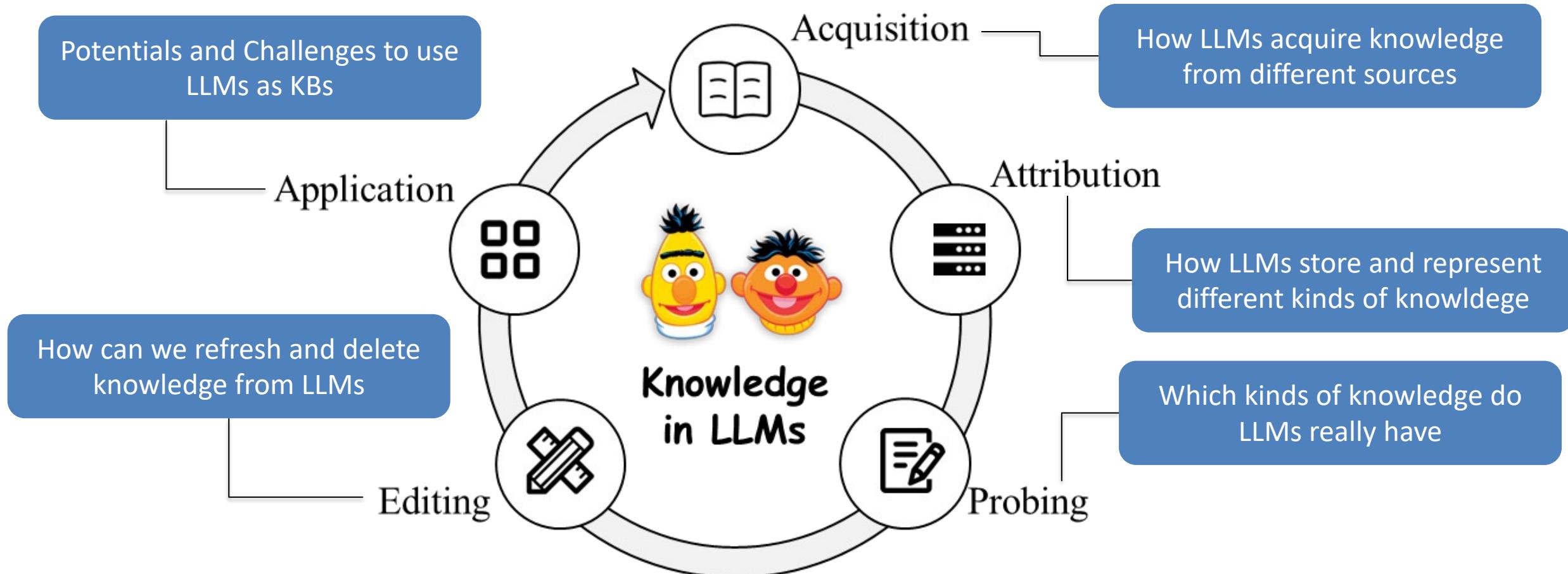
This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



This Tutorial

- How knowledge circulates throughout the life-cycle of LLMs



- Our survey paper entitled *The Life Cycle of Knowledge in Big Language Models: A Survey*
 - <https://arxiv.org/abs/2303.07616>
- Check out latest slides at our homepage
 - <http://www.icip.org.cn/>
- Corresponding paperlist
 - <https://github.com/c-box/KnowledgeLifecycle>

Knowledge Acquisition: Learning From Texts and Beyond

- Knowledge Acquisition Strategies
 - How to leverage different kinds of unsupervised/supervised/self-supervised learning approaches to inject knowledge into LLMs
- Knowledge Acquisition Mechanism
 - How LLMs dynamically acquire different kinds of knowledge during learning

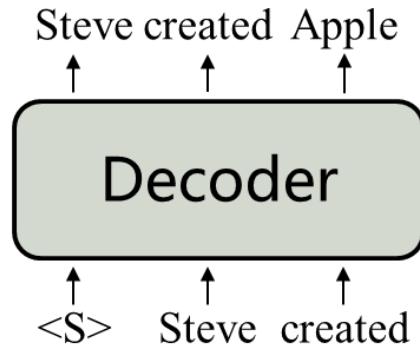
Knowledge Acquisition: Strategies

Strategies and Knowledge Sources

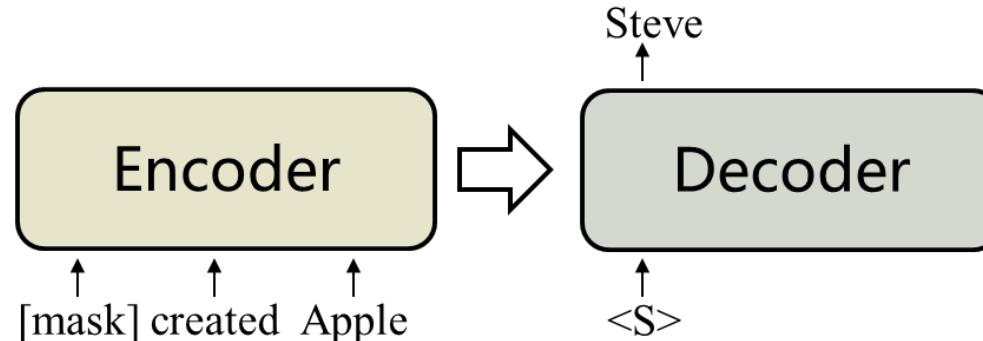
- Learning From Texts: Self-supervised Pretraining
 - Unstructured texts without annotation
- Learning From Instruction Data: Supervised Fine-tuning
 - QA pairs or conversational data with manually annotated answers
- Learning From Human Feedback: Supervised Alignment
 - Partial order pairs of model-generated answers
- Learning From Structural Data: Structured Knowledge Injection
 - Structural KBs created by human beings

Knowledge Acquisition From Texts

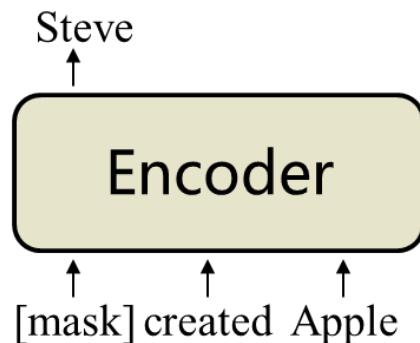
- Using Self-supervised Learning to learn from unlabeled texts



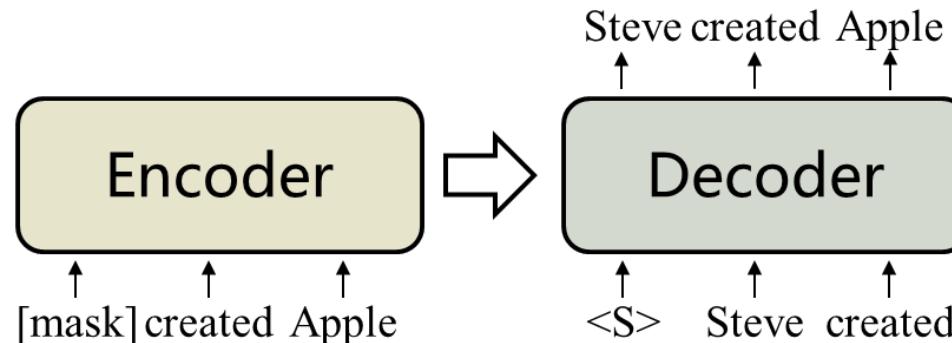
(a) CLM



(c) Seq2Seq MLM



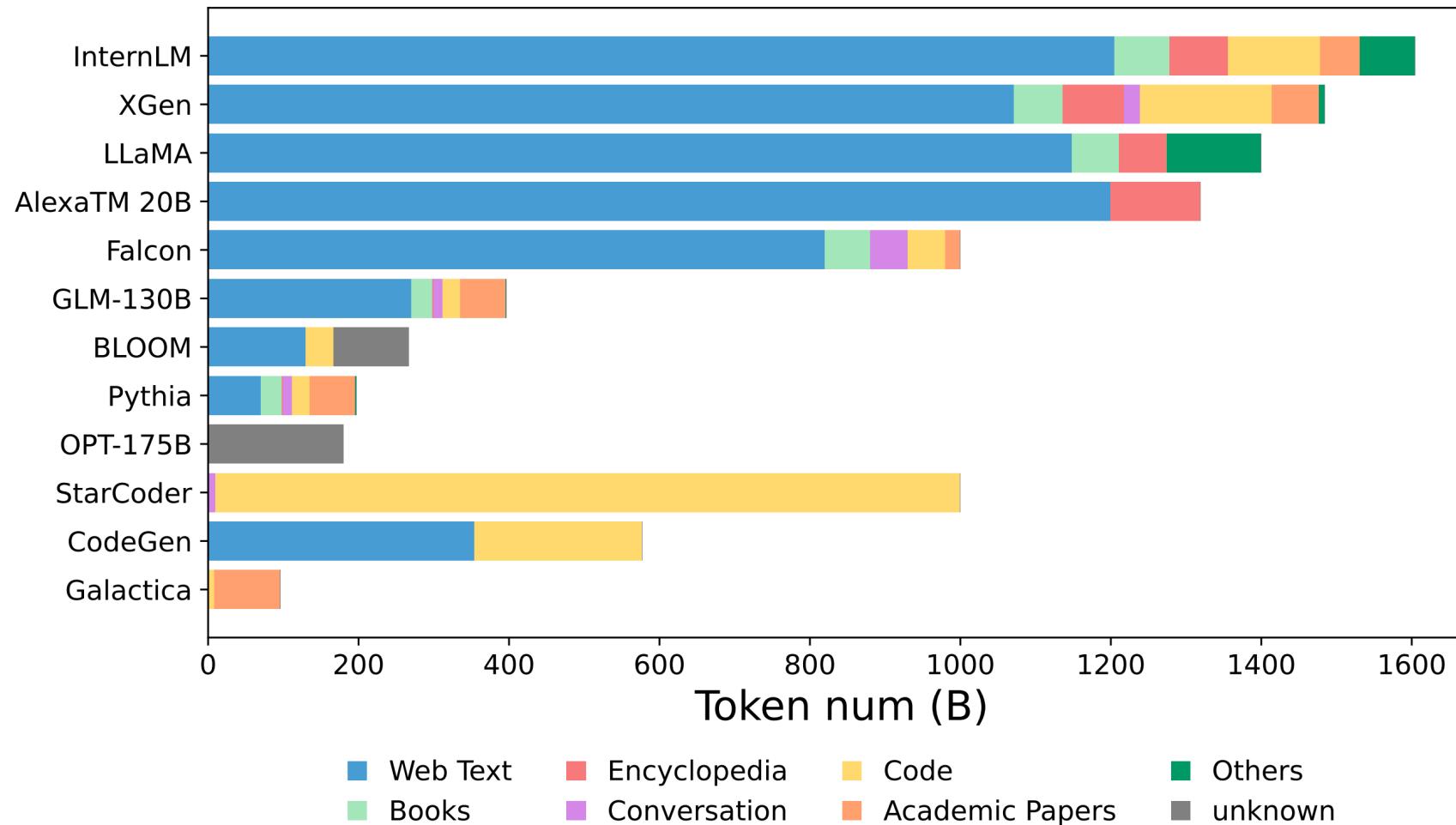
(b) MLM



(d) Denoising Autoencoder

Knowledge Acquisition From Texts

- Requires an extremely large collection of highly-diversified Corpus

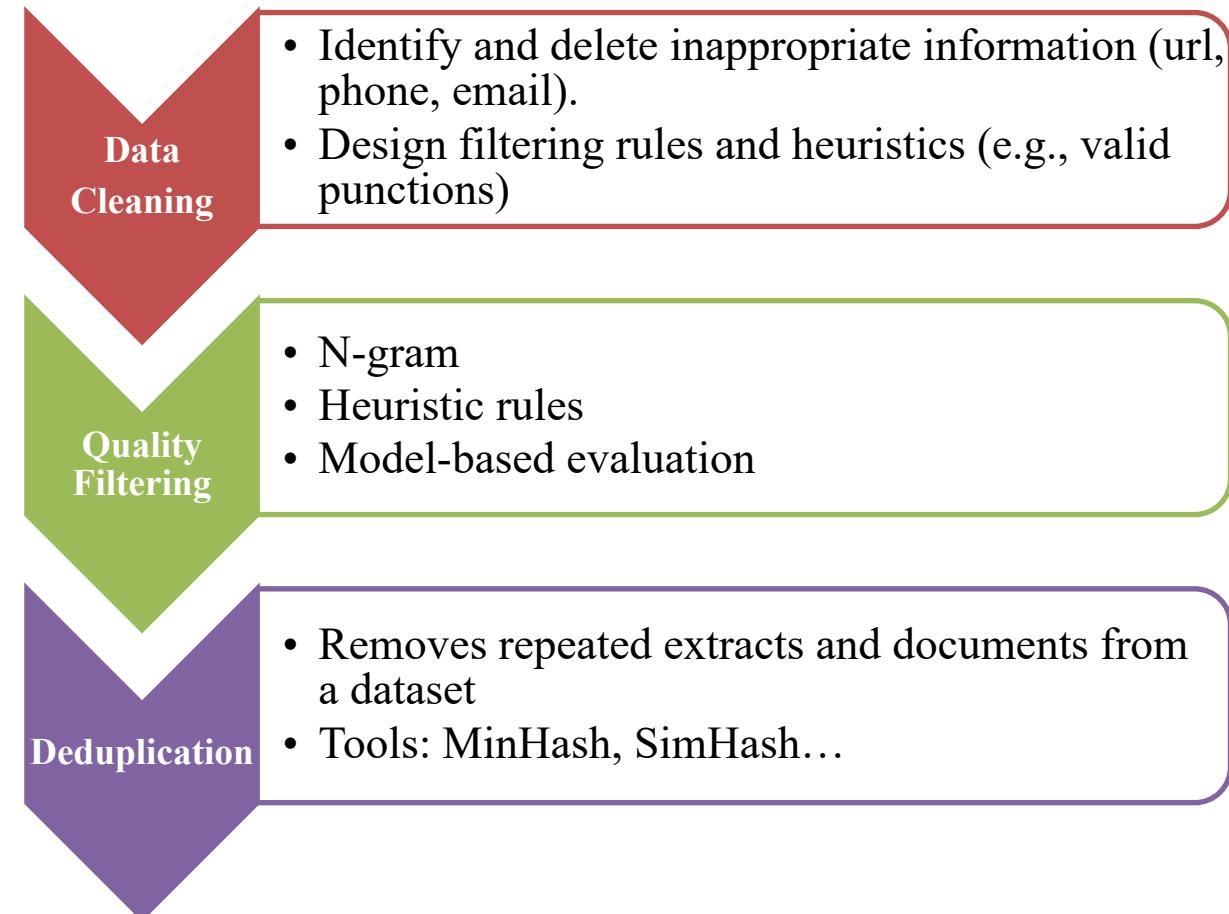


Knowledge Acquisition From Texts

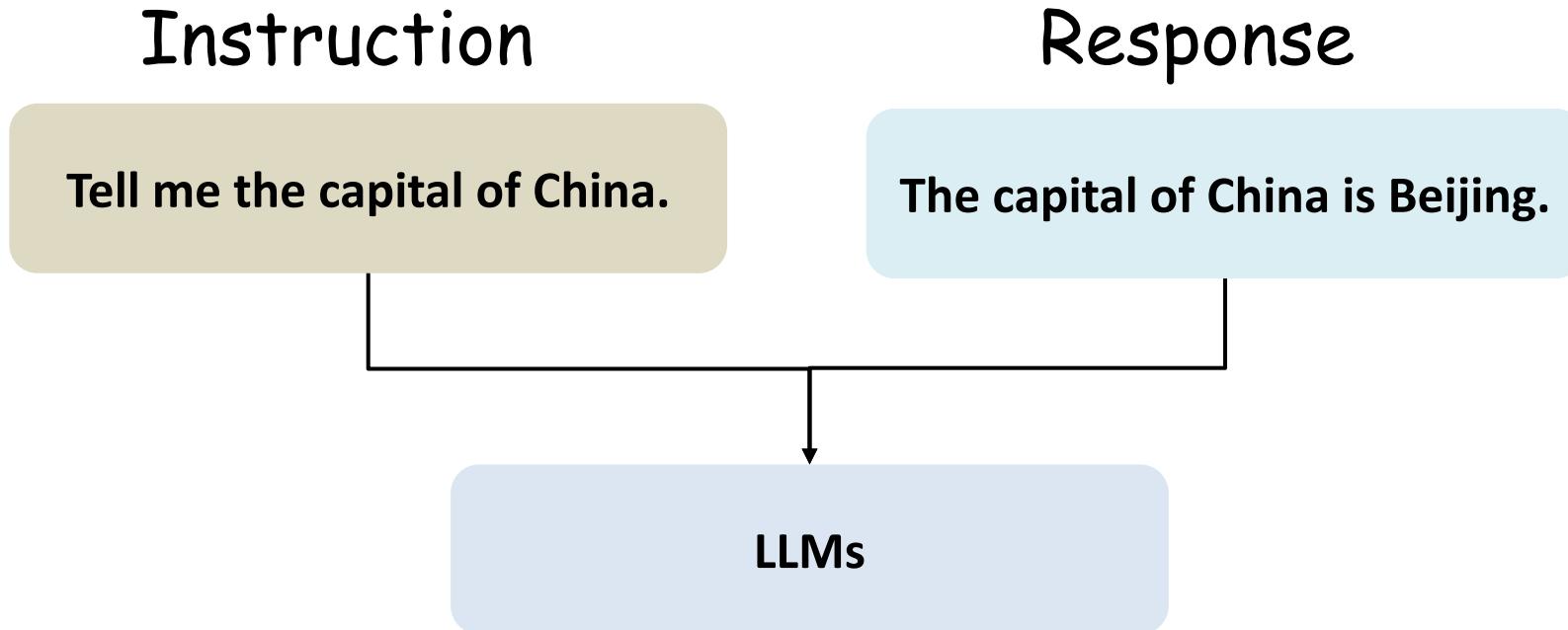
- Corpus require very careful cleaning before being used to train

LLMs

- Data Cleaning
- Quality Filtering
- Deduplication
-

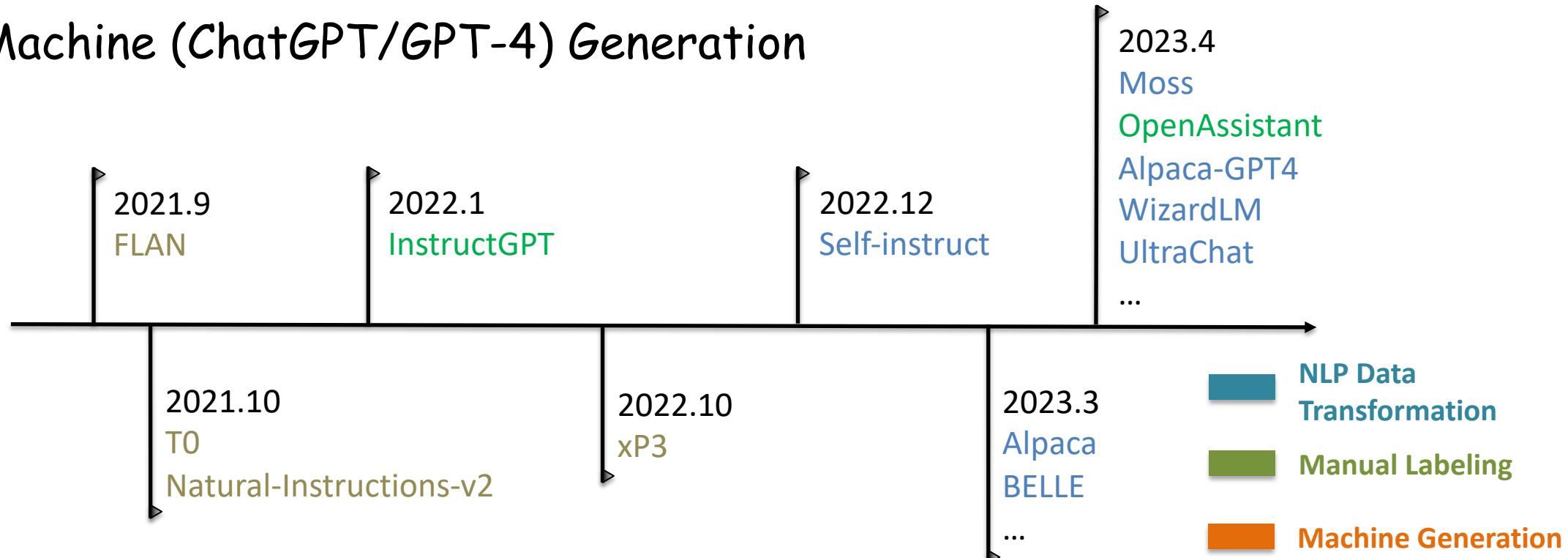


- Construct instruction-response data for LLM SFT training.



Knowledge Acquisition From Labeled Data

- Three representative ways to harvest labeled data for LLMs
 - NLP data transformation
 - Manual Labeling
 - Machine (ChatGPT/GPT-4) Generation



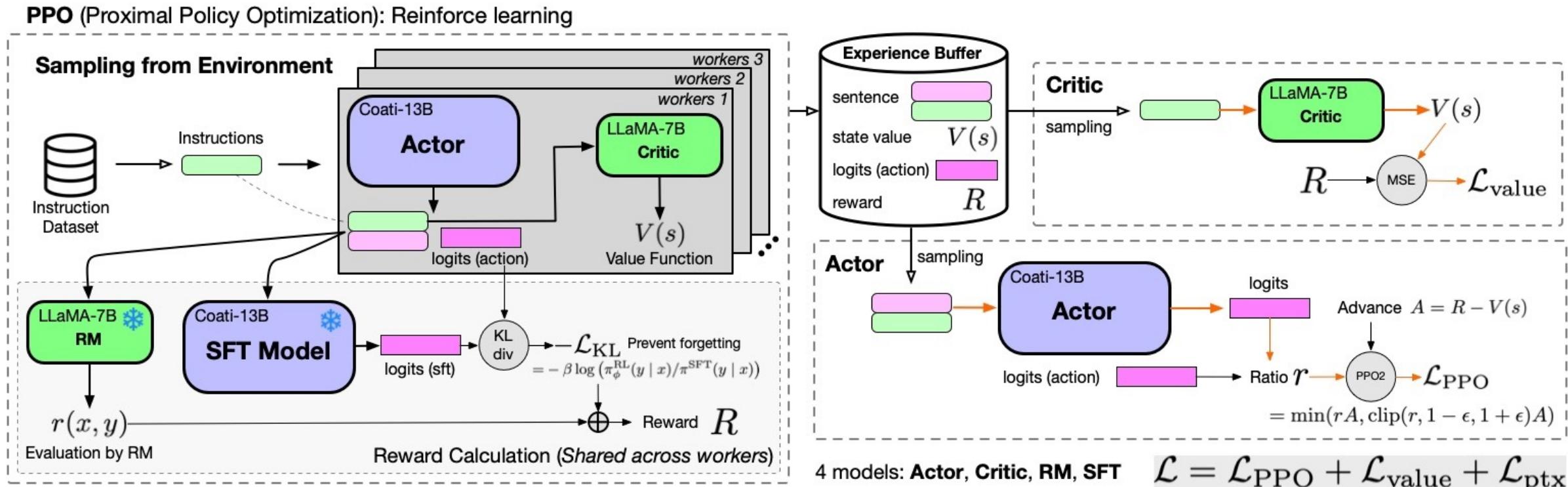
Knowledge Acquisition From Labeled Data

	NLP Data Transformation	Manual Labeling	Machine Generation
Advantage	Easy to generate	High diversity and quality	Easy to acquire
Disadvantage	Limited diversity and coverage	High costs, hard for alignment	Limited diversity, easy to collapse
Usage	Limited cases for each task	Ensure diversity	Ensure quality

Diversity is most critical for LLMs SFT!

Knowledge Acquisition From Human Feedback

- Using human feedback on a pair (list) of answers generated by the model to align the model to human value/behavior/favor.....



Specialized LLMs: ChatGPT, LaMDA, Galactica, Codex, Sparrow, and More. 2023.

ColossalChat: An Open-Source Solution for Cloning ChatGPT With a Complete RLHF Pipeline. 2023.

- Alignment with HF without RL

Category	Algorithm	Introduction
Negative Sampling	BoN	Find responses with highest reward for SFT
	RAFT	Find $\left\lfloor \frac{b}{k} \right\rfloor$ responses with highest reward for SFT
	Self-Align	Using LLM to generate better responses using principle-driven ICL
Conditional Generation	CoH	Design special token for both positive and negative response
	Quark	Assign reward token to each response according to reward
Contrastive Learning	RRHF	learns to align with human preferences through ranking loss
	DPO	Pair-wise contrastive learning
	PRO	List-wise contrastive learning
	SLiC-HF	Sequence Likelihood Calibration

- Structured knowledge refers to information that is organized in a well-defined format or framework



Factual



Commonsense

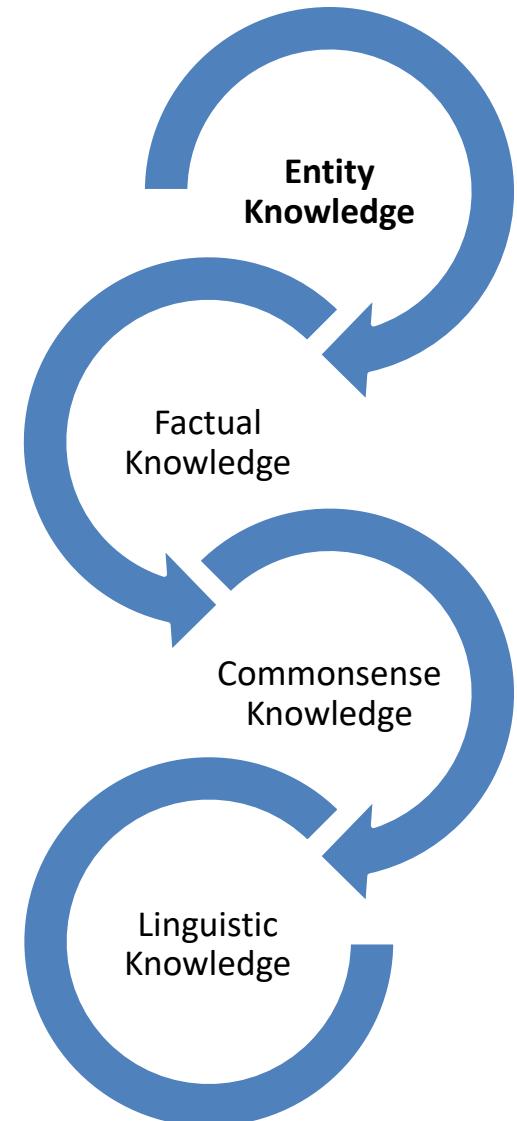


Linguistic

- Entity Knowledge
 - Teaching models to concentrate more on entities beyond tokens
- Factual Knowledge
 - Injecting factual knowledge from knowledge bases
- Commonsense Knowledge
 - Injecting commonsense knowledge that may not appear in texts
- Linguistic Knowledge
 - Using linguistic information to guide model better formulating languages

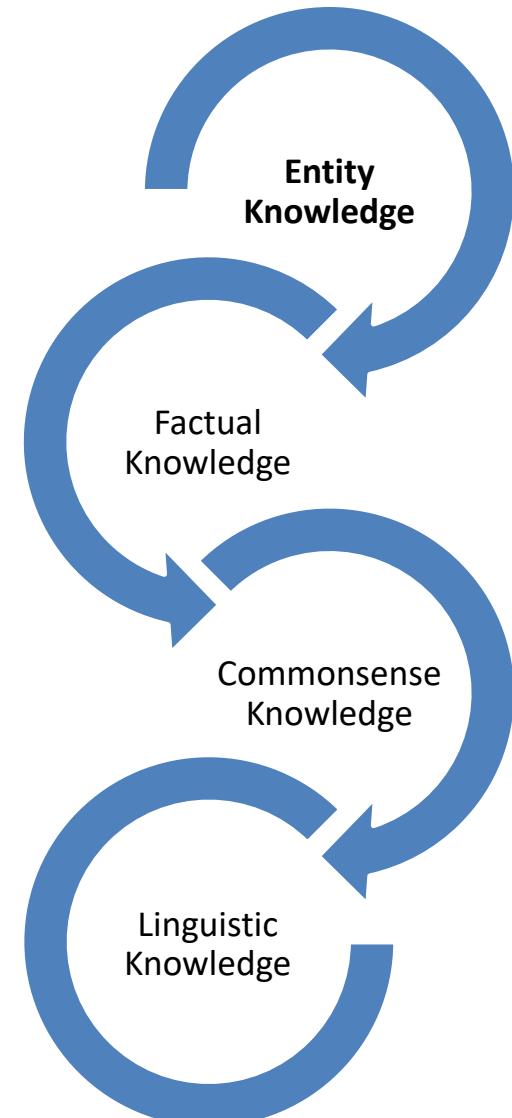
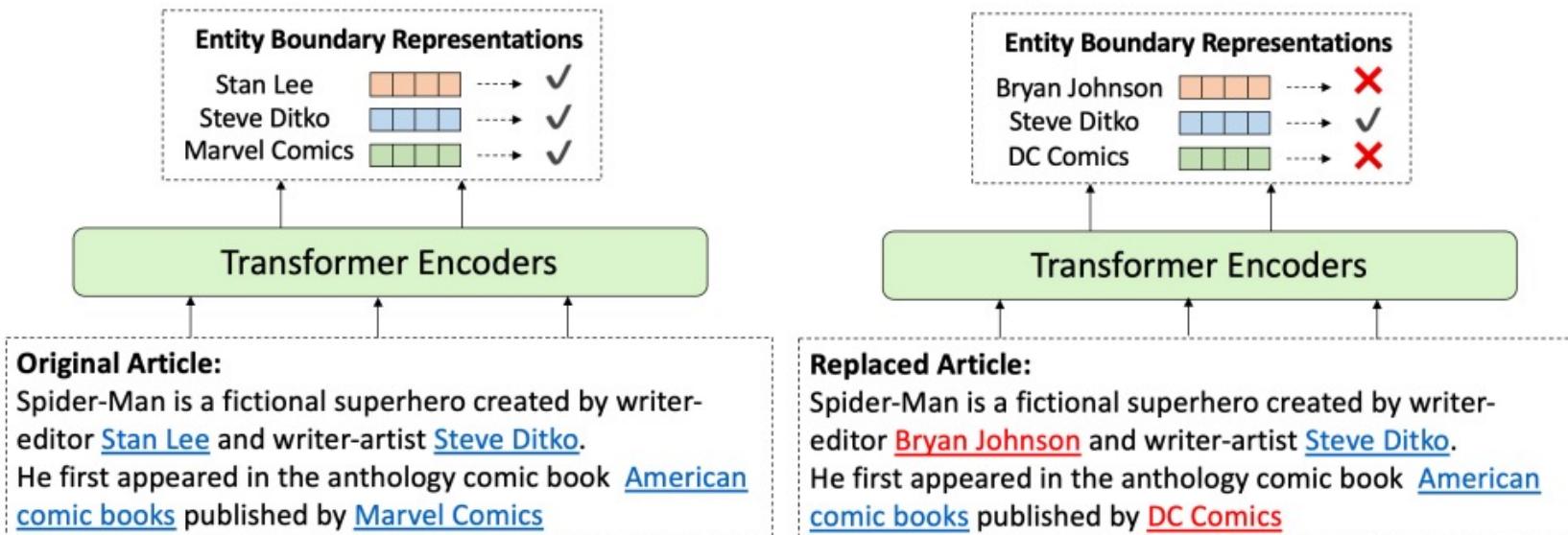
- Entity knowledge example #1: Entity Masking
(Sun et al., 2019)
 - Mask and predict all sub-words within an entity

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

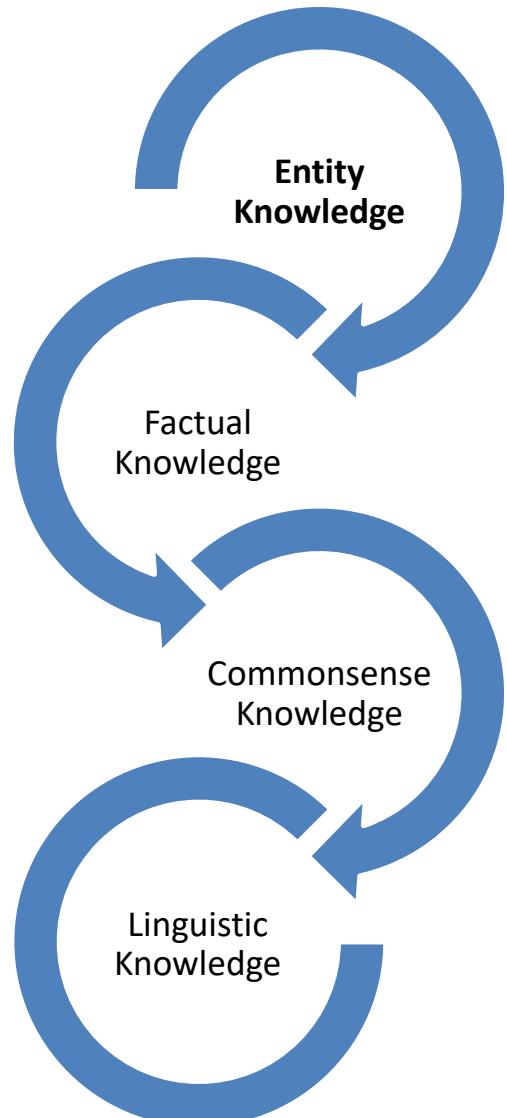
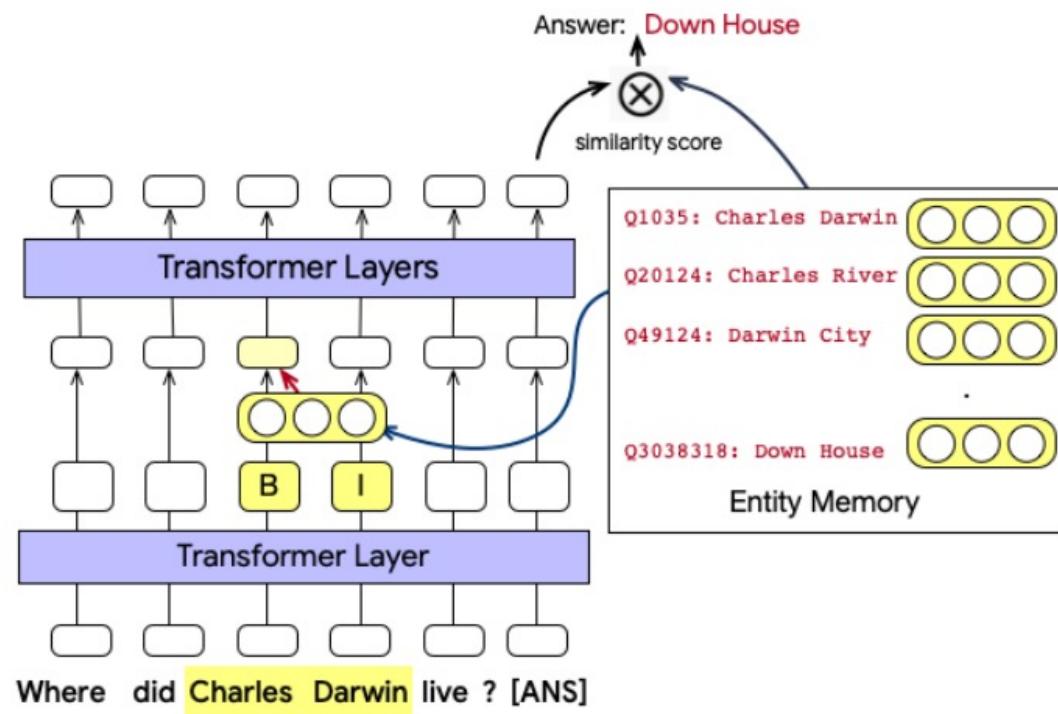


Knowledge Acquisition From Structural Data

- Entity knowledge example #2: replaced entity detection (Xiong et al., 2020)
 - Determine which entities are replaced

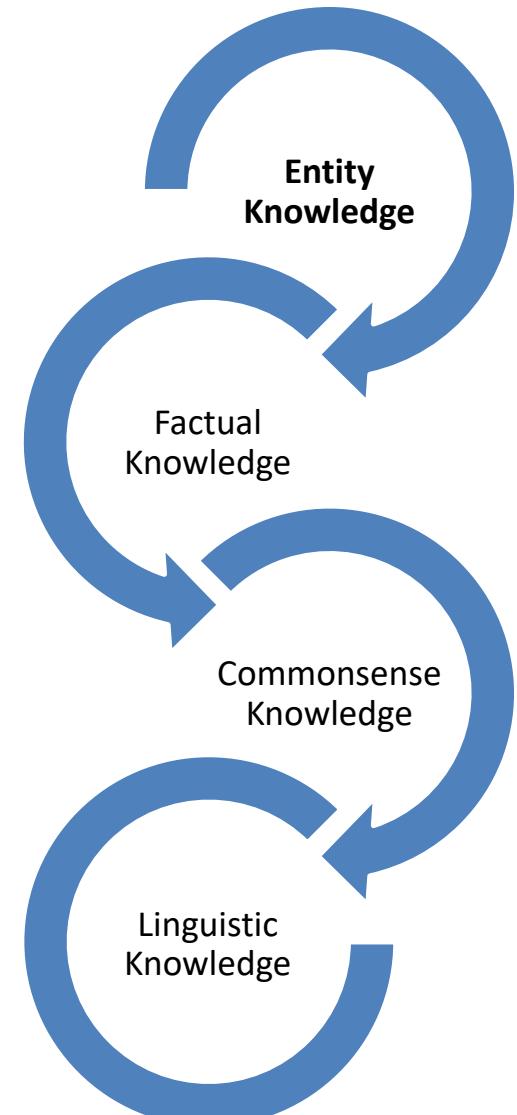
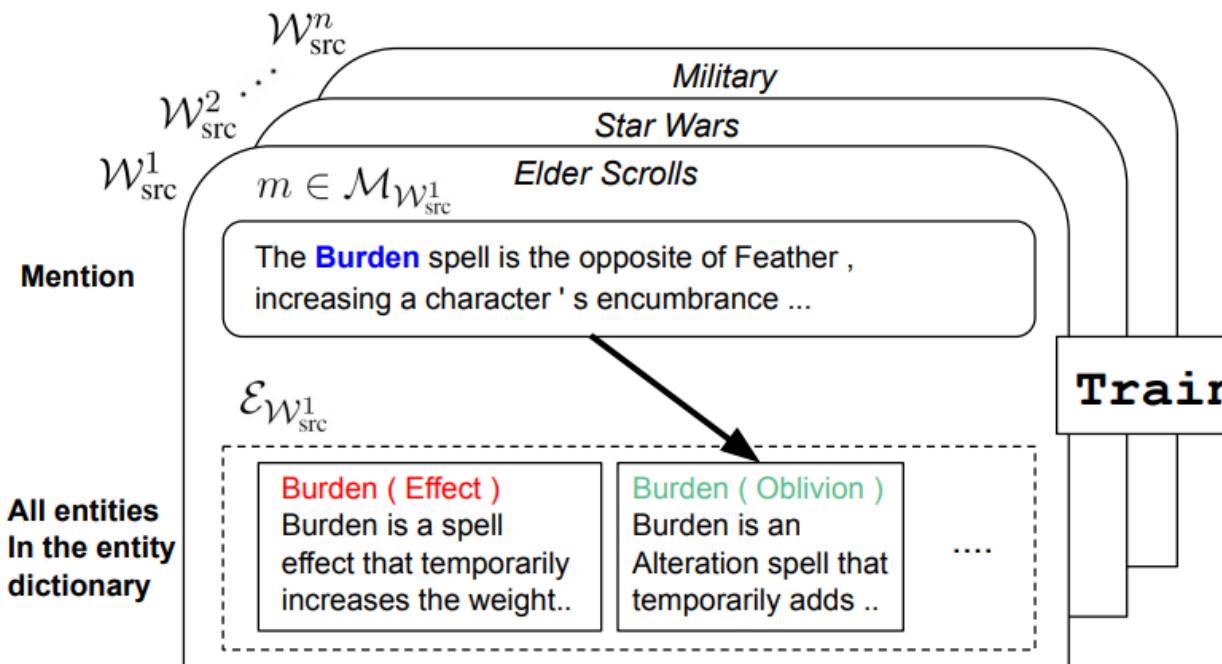


- Entity knowledge example #3 : pretraining via mention detection and entity Linking
(Fevry et al., 2020)



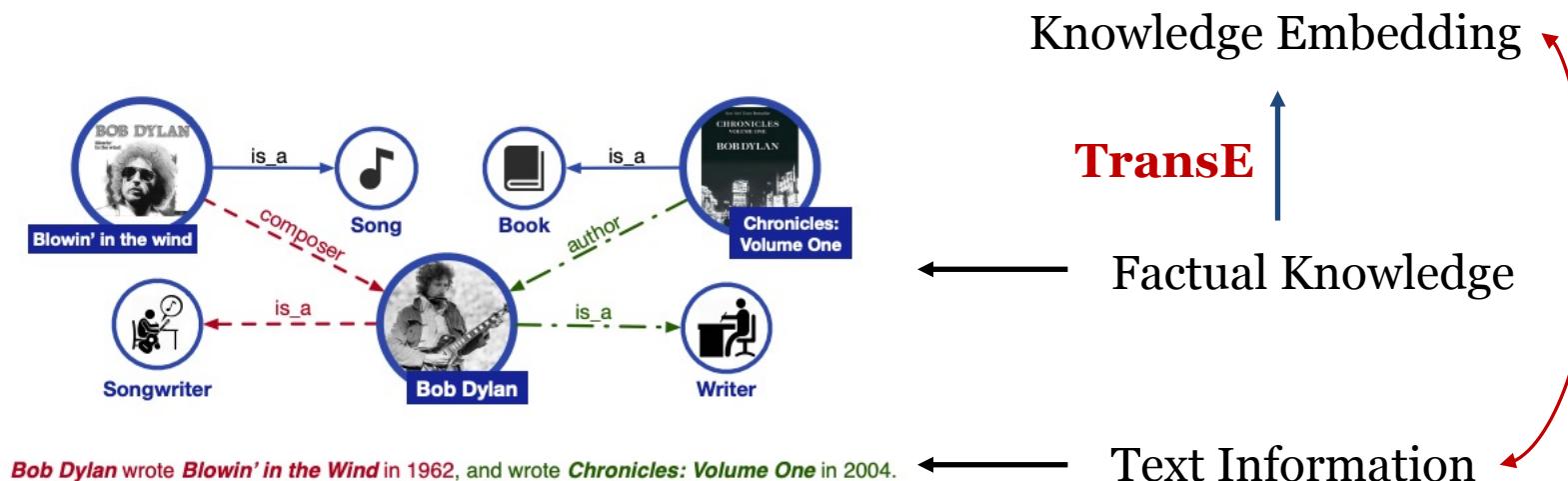
Knowledge Acquisition From Structural Data

- Entity knowledge example #4: enriching entity representation using meta-information
(Logeswaran et al., 2019; Gillick et al., 2019)



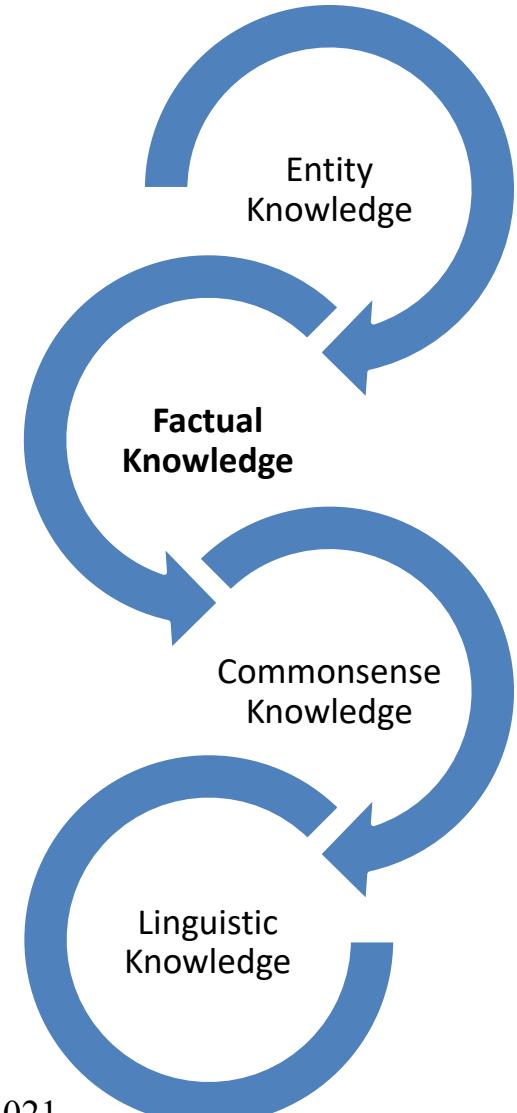
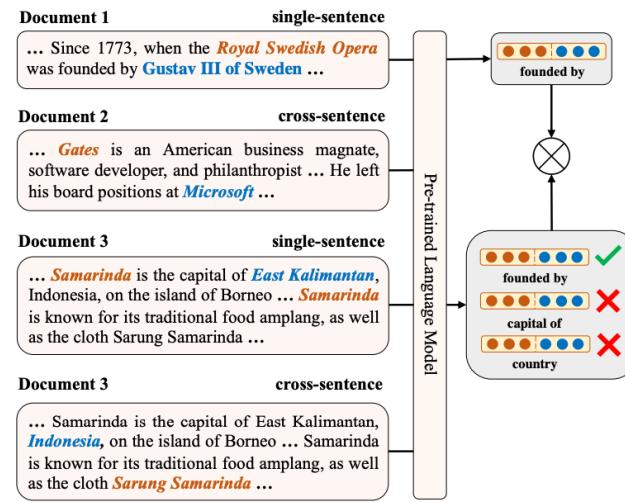
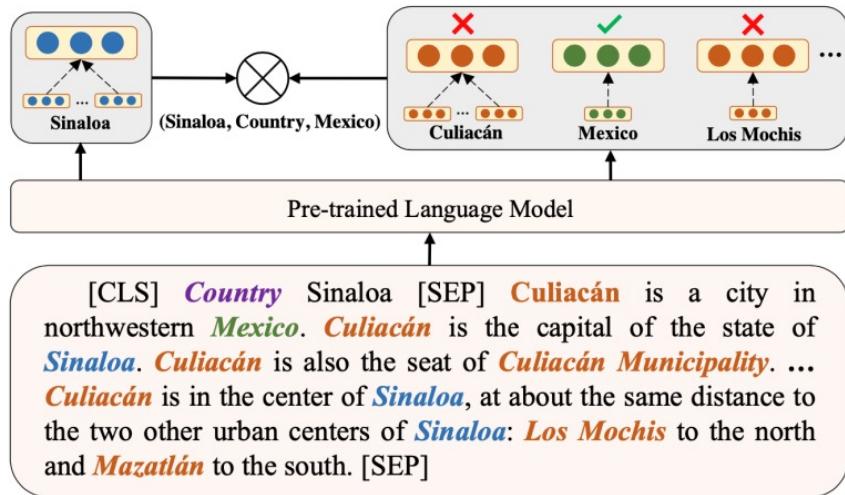
Knowledge Acquisition From Structural Data

- Factual knowledge example #1: incorporating knowledge embeddings (Zhang et al., 2019; Wang et al., 2021)



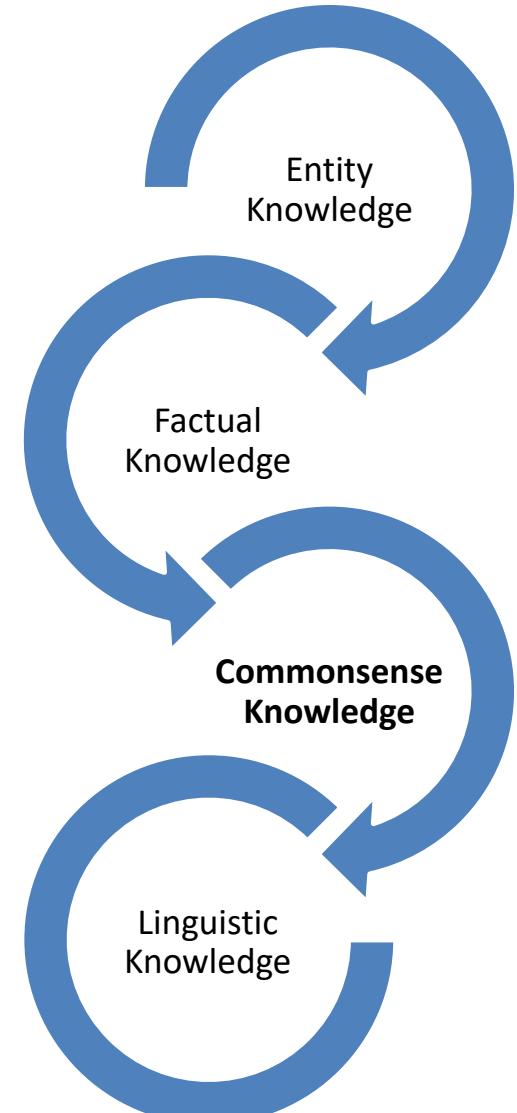
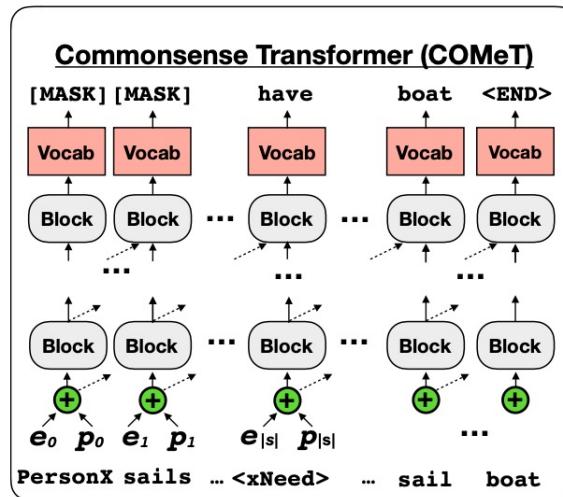
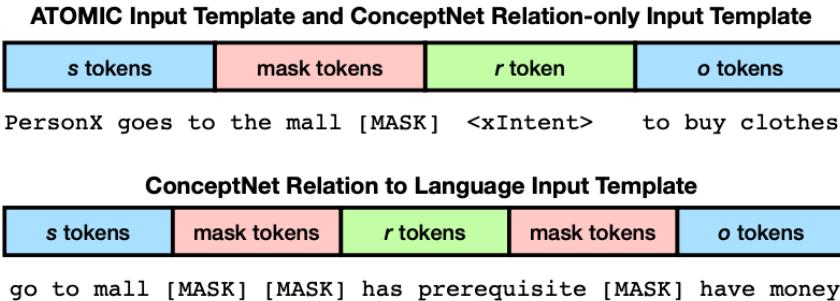
Knowledge Acquisition From Structural Data

- Factual knowledge example #2: designing auxiliary tasks (Qin et al., 2021; Banerjee et al., 2021)



Knowledge Acquisition From Structural Data

- Commonsense Knowledge: transforming structured knowledge into natural language
(Bosselut et al. 2019; Ye et al. 2019; Guan et al. 2020; Ma et al. 2021)



Bosselut et al. . COMET: Commonsense transformers for automatic knowledge graph construction. ACL 2019.

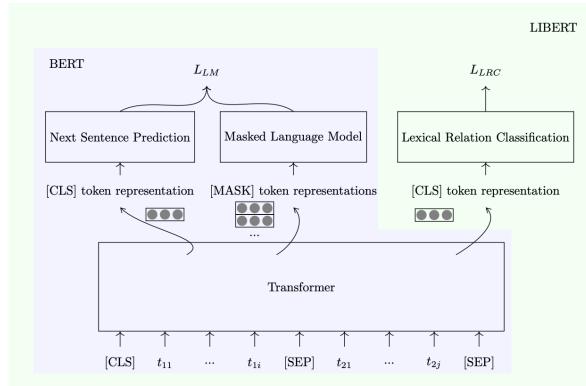
Ye et al. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. 2020.

Guan et al. A knowledge-enhanced pretraining model for commonsense story generation. TACL 2020.

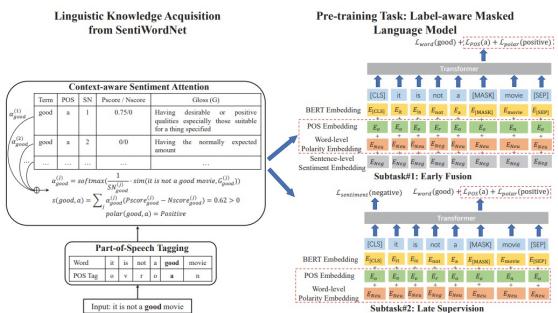
Ma et al. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. AAAI 2021.

Knowledge Acquisition From Structural Data

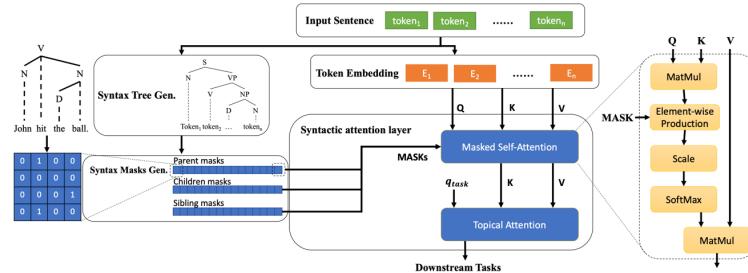
- Linguistic Knowledge: feature-based approaches



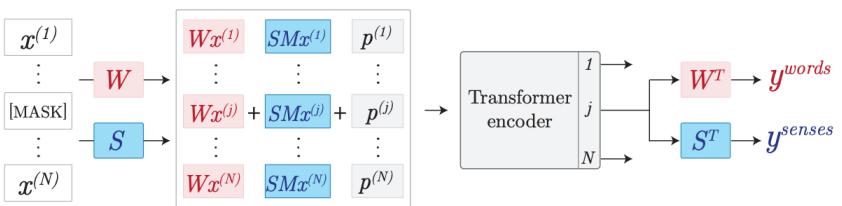
Lexically-informed BERT (Lauscher et al. 2020)



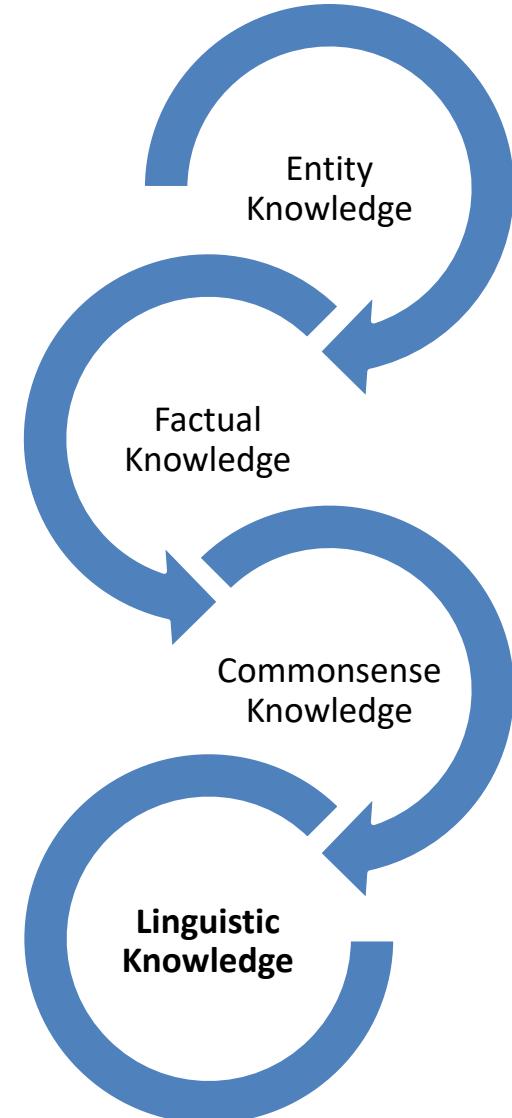
Senti LARE (Ke et al. 2020)



Syntax-BERT (Bai et al. 2020)



Sense-BERT (Levine et al. 2020)



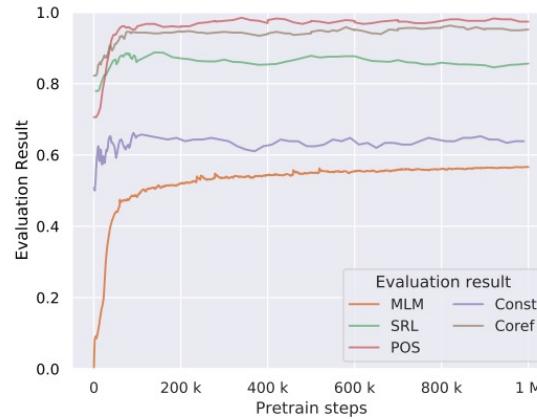
Knowledge Acquisition: Mechanisms

- How and why LLMs can acquire or forget knowledge from different sources ?
- Investigate this by diving into the dynamics of LLMs' learning procedure

Knowledge Acquisition Mechanisms

- Dynamics investigation example #1: ALBERT knowledge evolution (Chiang et al., 2020)

Semantic and Syntactic Knowledge



(b) Masked LM accuracy and F1 scores of different probing tasks over the course of pretraining

World Knowledge

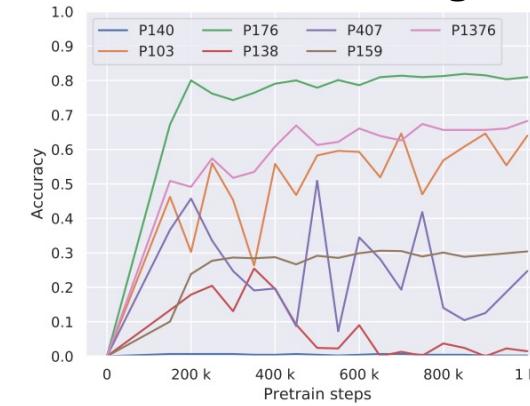


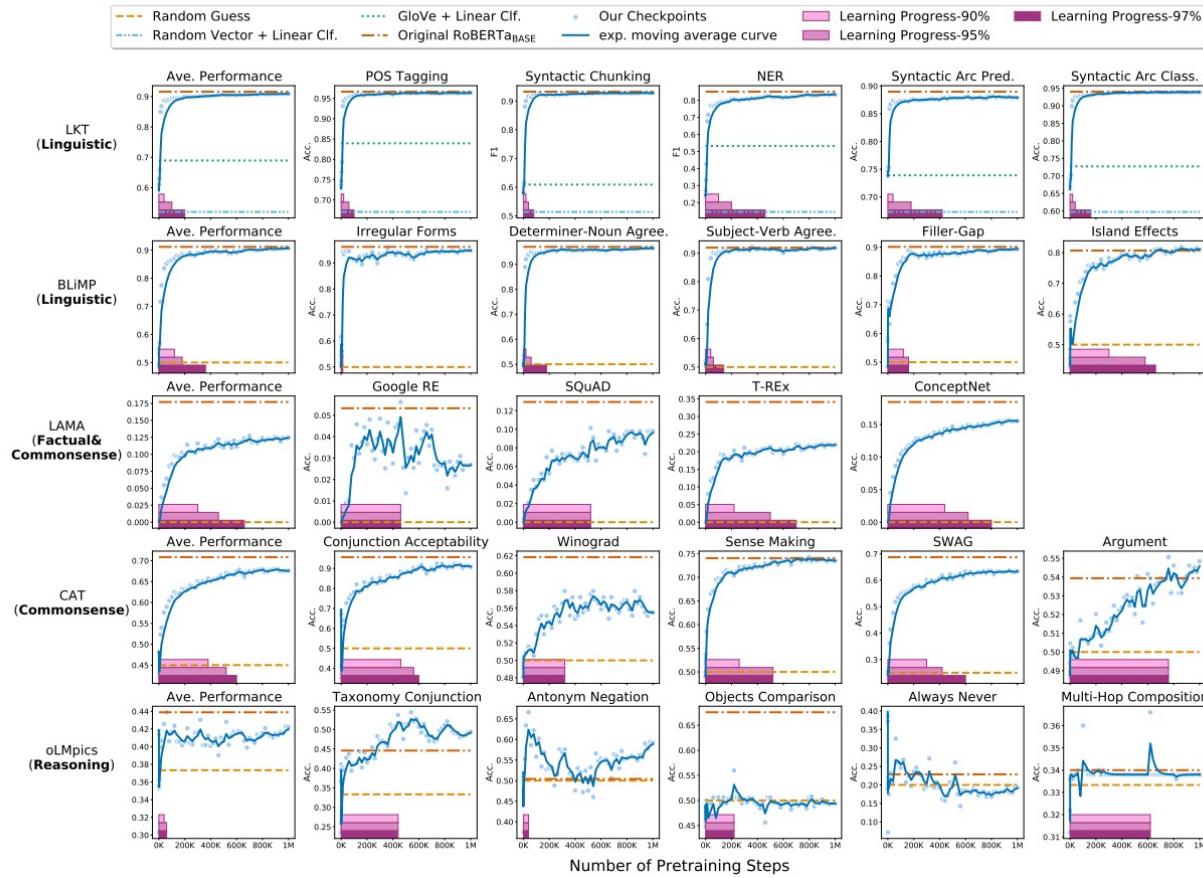
Figure 6: World knowledge development during pre-training evaluated every 50k pretrain steps. Types of relation, and template are shown in Table 1

- Semantic and syntactic knowledge is developed simultaneously in ALBERT.
- ALBERT seems to be dynamically renewing its knowledge about the world.

Knowledge Acquisition Mechanisms

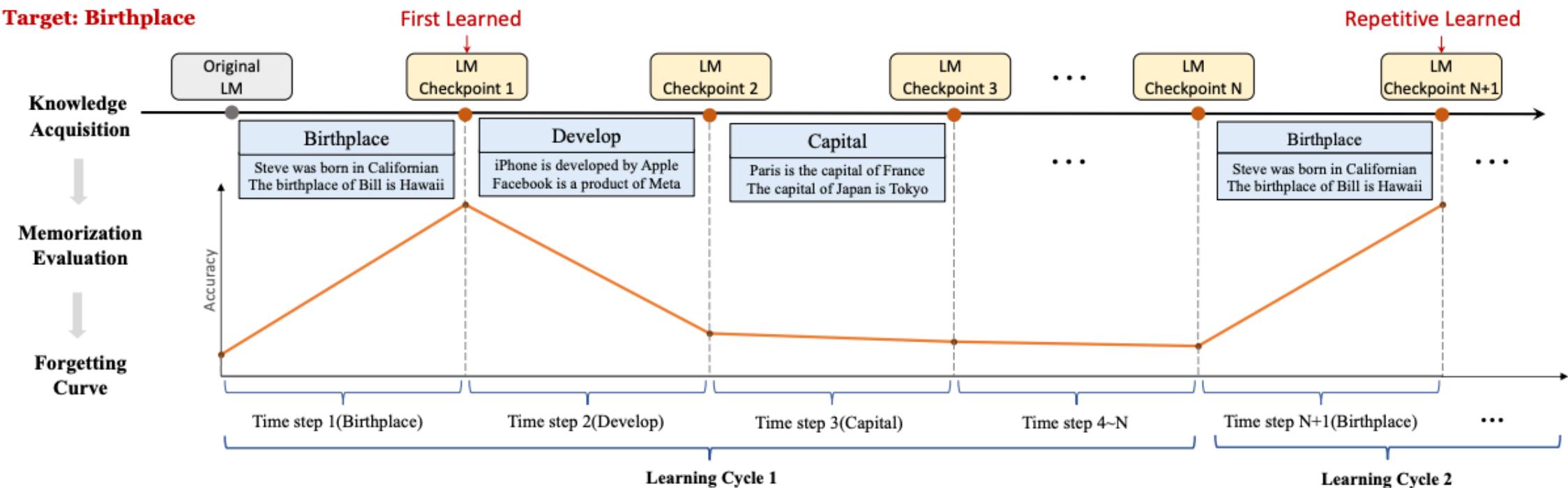
- Dynamics investigation example #2: RoBERTa knowledge evolution (Liu et al., 2020)

- Linguistic knowledge can be learned quickly and robustly
- World knowledge is learned slowly and domain-sensitive



Knowledge Acquisition Mechanisms

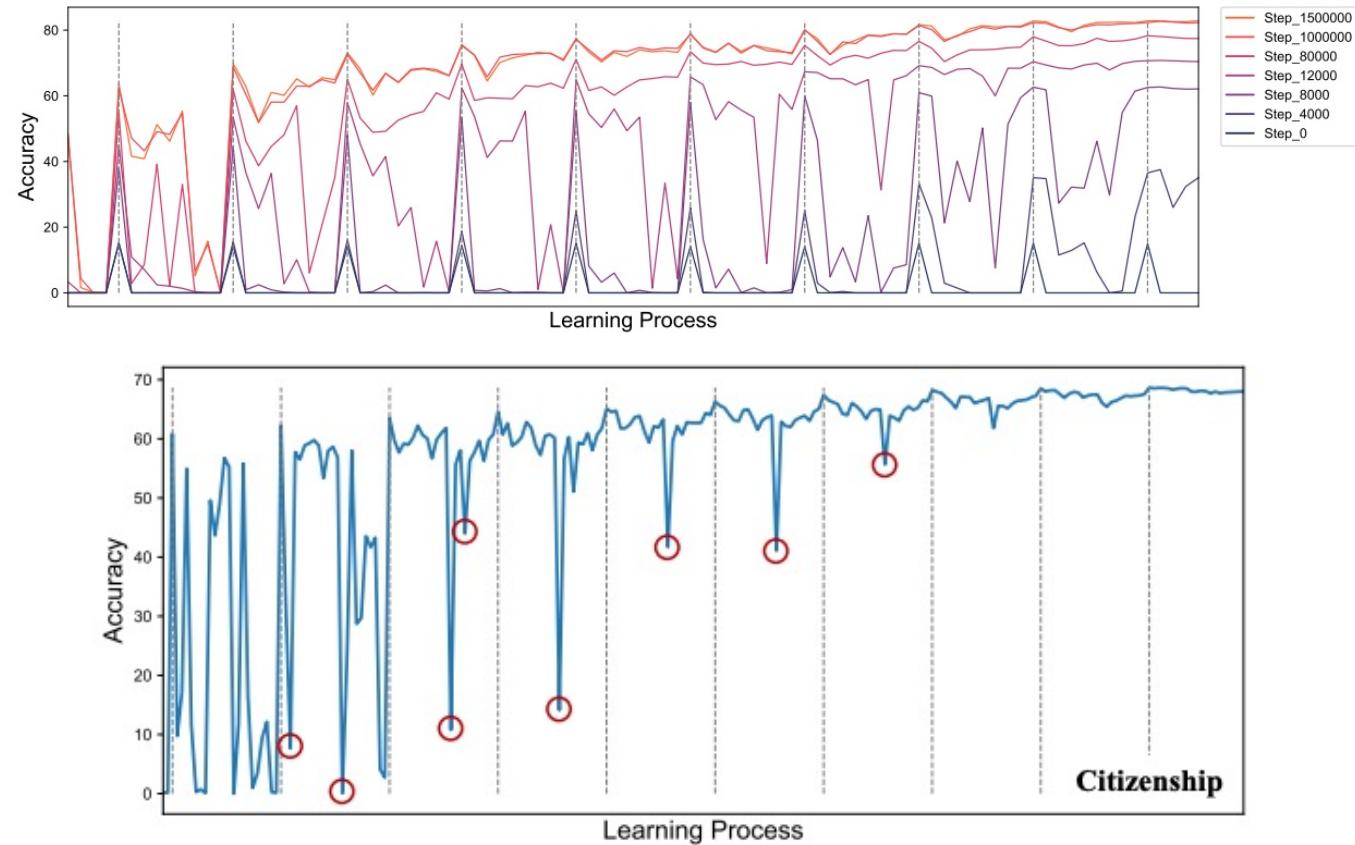
- Dynamics investigation example #3: Learning and forgetting dynamics of factual knowledge (Cao et al., 2023)



Knowledge Acquisition Mechanisms

- Dynamics investigation example #3: Learning and forgetting dynamics of factual knowledge (Cao et al., 2023)

- Pretraining is the key to shift "short-term" memory to "long-term" memory
- Existence of "singularity" where memory collapsed but quickly recovered



Take-aways for Knowledge Acquisition

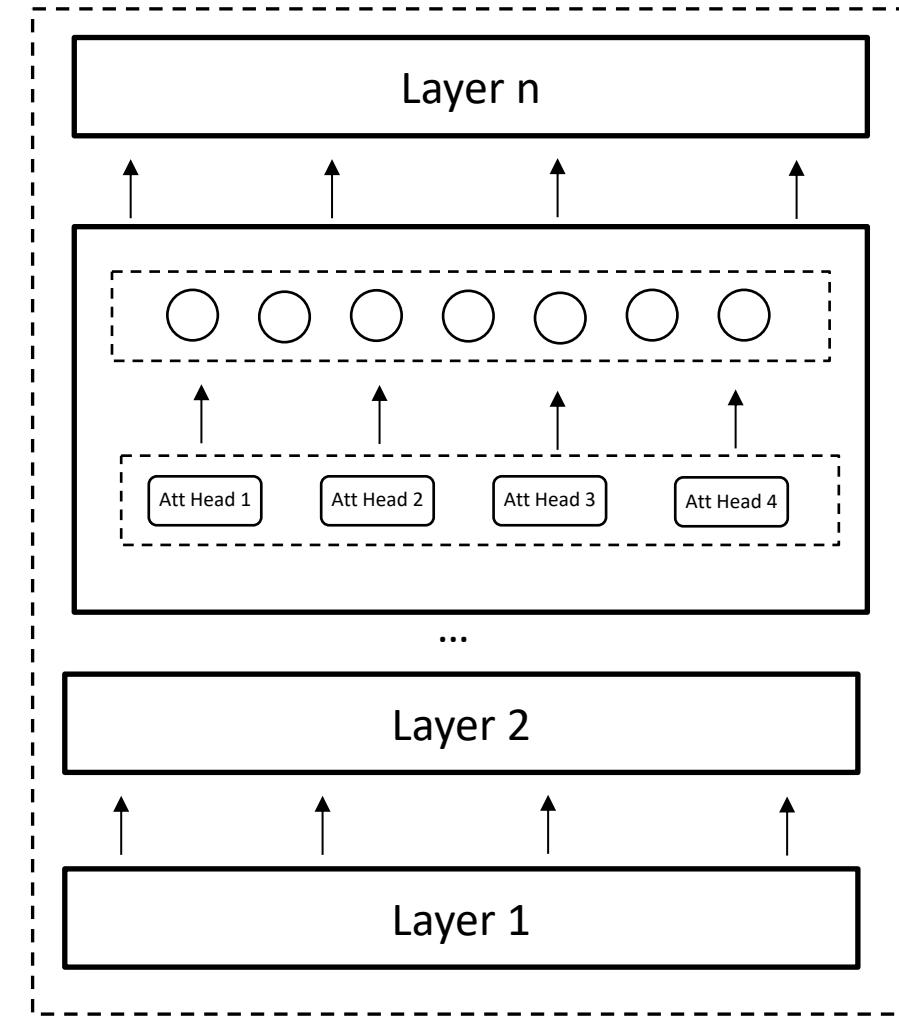
- Knowledge in LLMs are learned from multiple sources via multiple learning approaches
 - Learning From Texts: Self-supervised Pretraining
 - Learning From Instruction Data: Supervised Fine-tuning
 - Learning From Human Feedback: Supervised Alignment
 - Learning From Structural Data: Structured Knowledge Injection
- The underlying mechanisms of how LLMs learn knowledge still need further investigation

Knowledge Attribution: Opening the Blackbox

- How LLMs encode, transform and store the acquired knowledge?
- Can we associate specific knowledge with certain modules or neurons within a language model?
- Can we control the knowledge in the language model by modifying these specific modules?

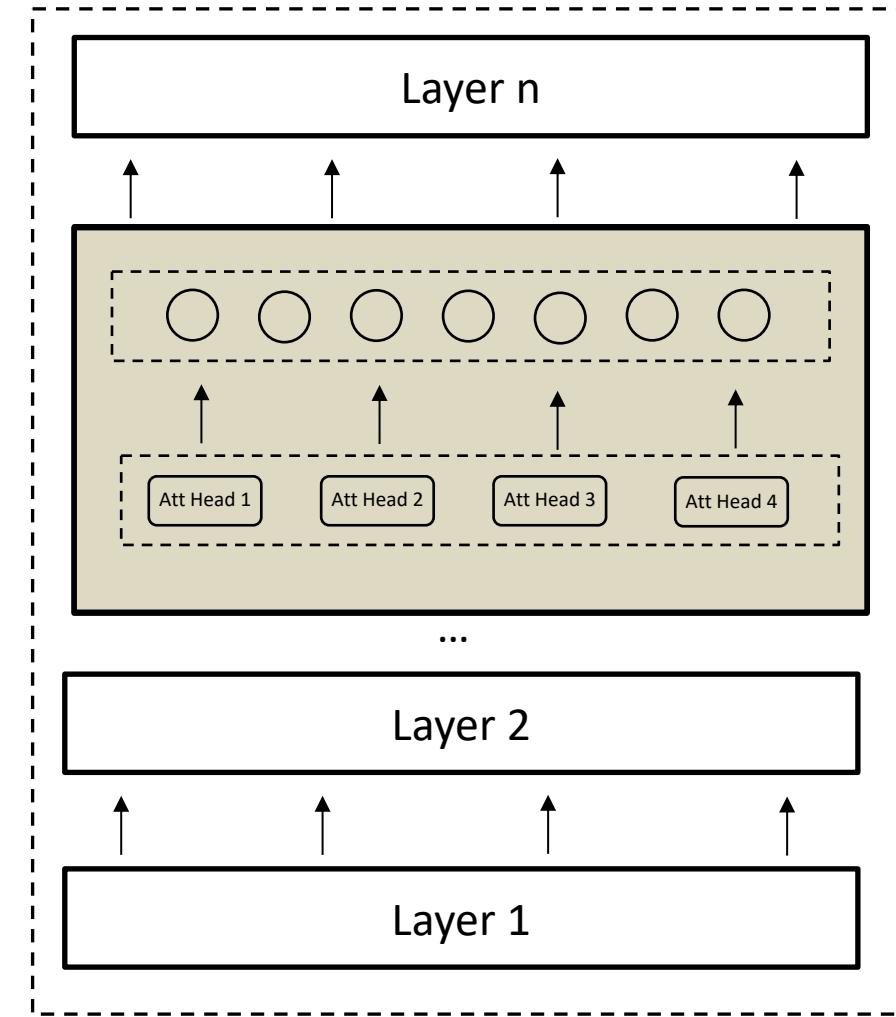
Knowledge Attribution

- Attribute stored knowledge to different-level of modules in neural networks



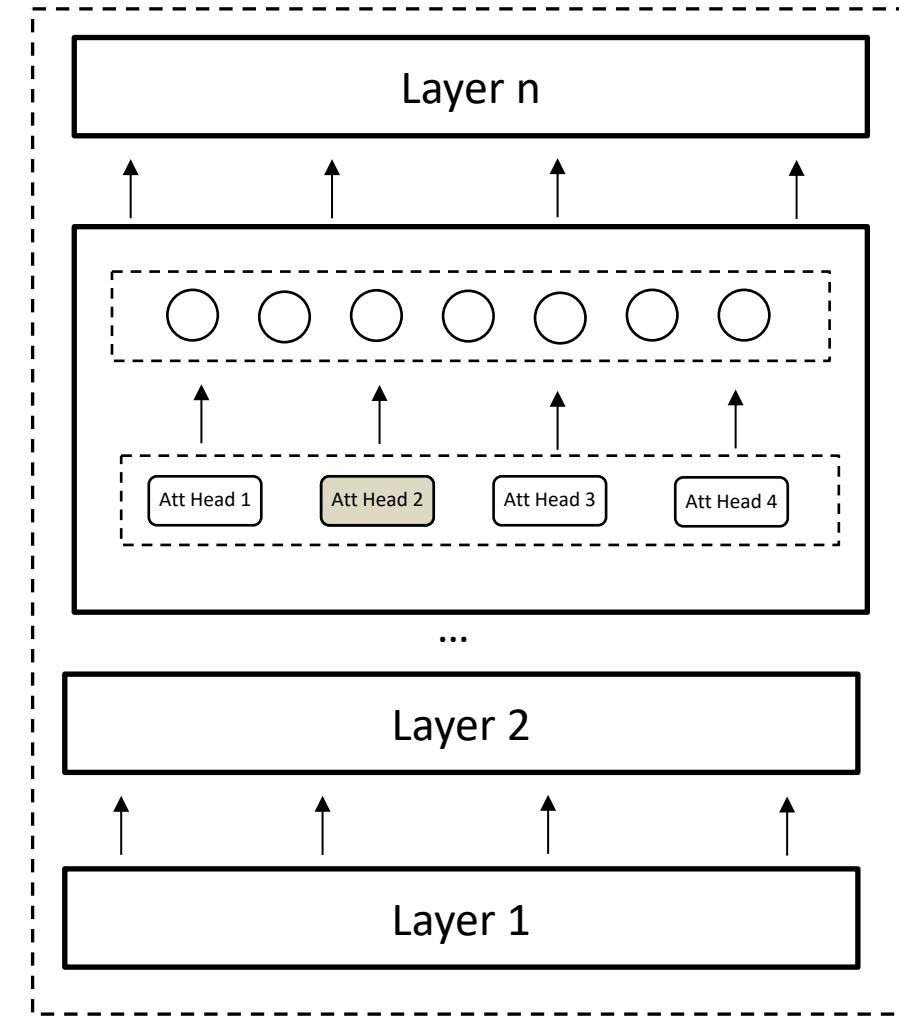
Knowledge Attribution

- Attribute stored knowledge to different-level of modules in neural networks
 - Layers



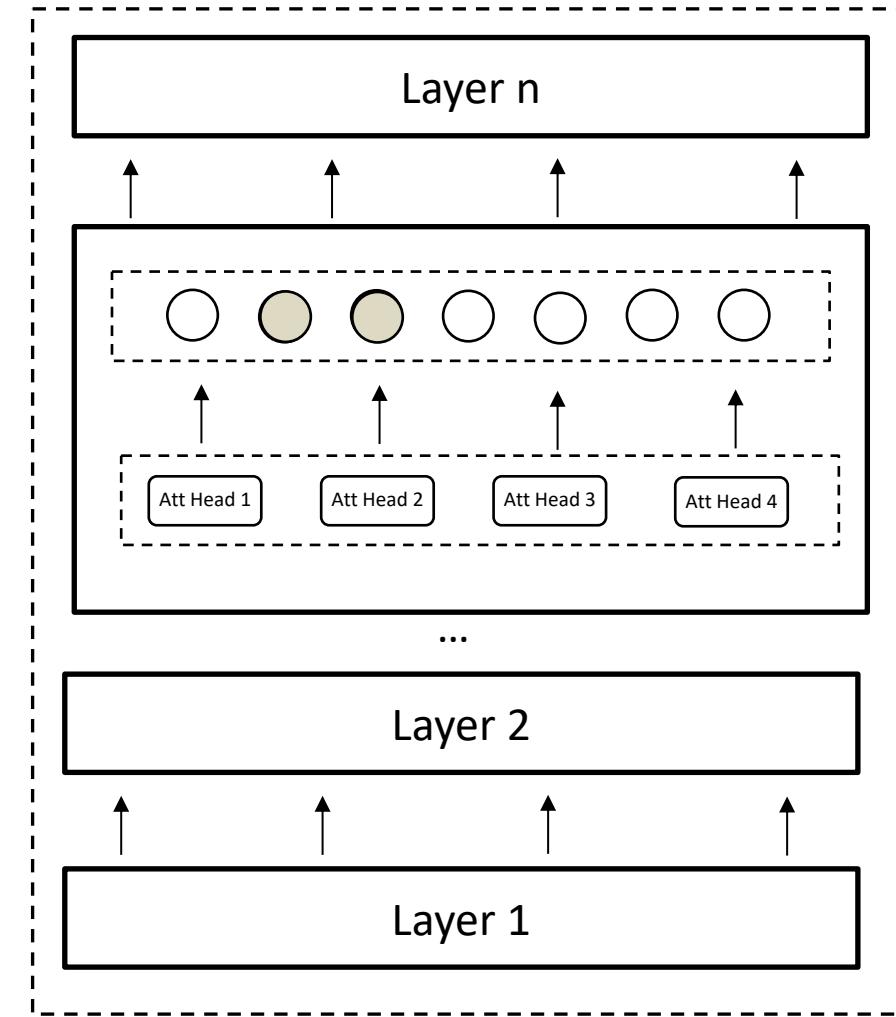
Knowledge Attribution

- Attribute stored knowledge to different-level of modules in neural networks
 - Layers
 - Modules



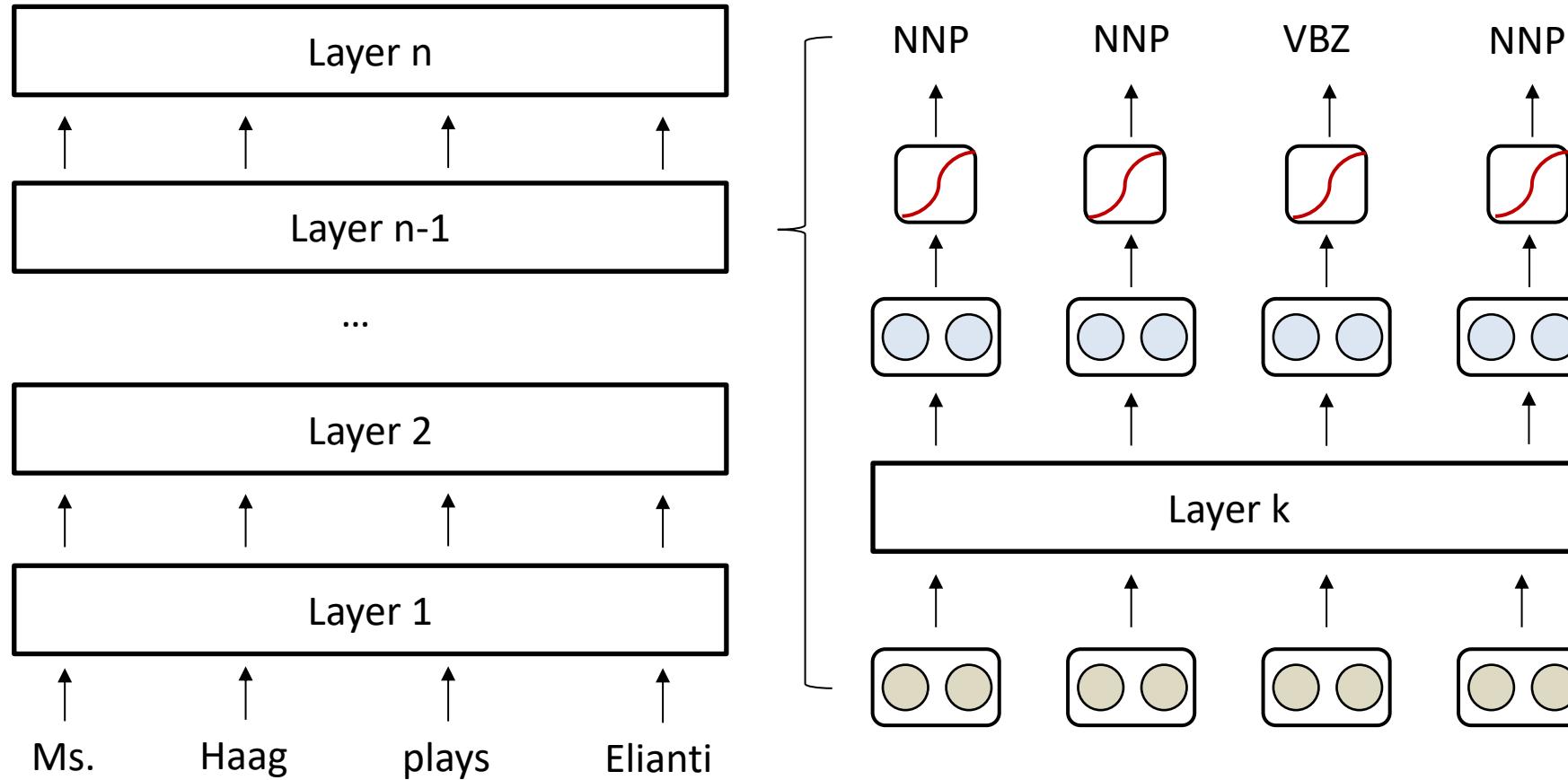
Knowledge Attribution

- Attribute stored knowledge to different-level of modules in neural networks
 - Layers
 - Modules
 - Neurons



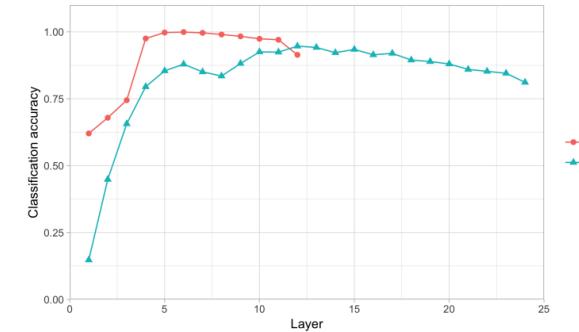
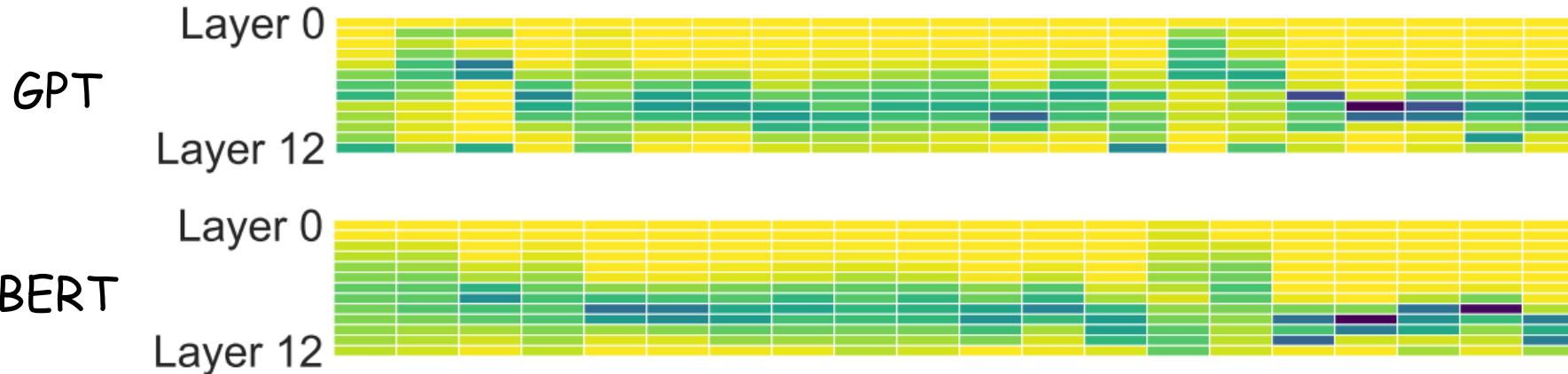
Layer-wise Knowledge Attribution

- Attributing knowledge to each layer of NNs by training a task-specific classifier for representations on each layer

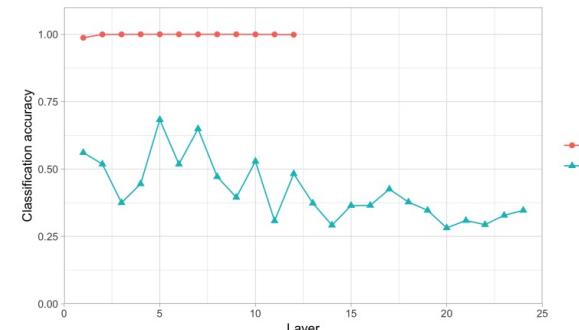


Layer-wise Knowledge Attribution

- Example #1: Linguistic Knowledge(Liu et al. 2019; Lin et al. 2019)



Main Auxiliary

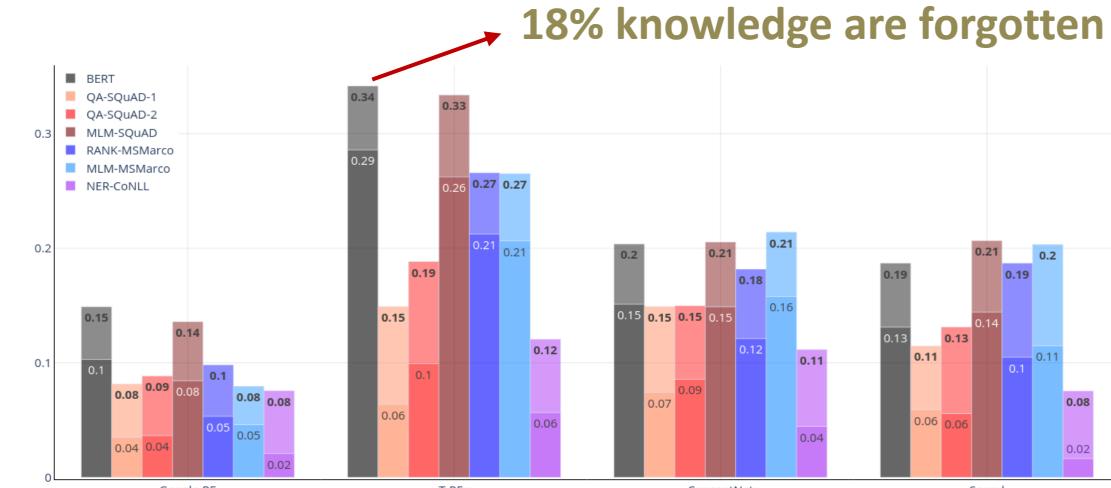
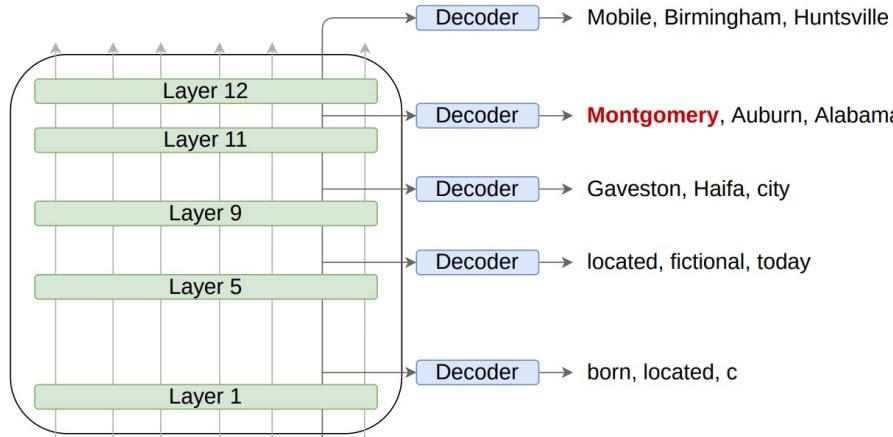


Subject Noun

- High Layers: more task-specific but fail on tasks requiring fine-grained linguistic knowledge
- Middle& Lower Layers: better linguistic transferability
- BERT encodes positional information about word tokens well on its lower layers

Layer-wise Knowledge Attribution

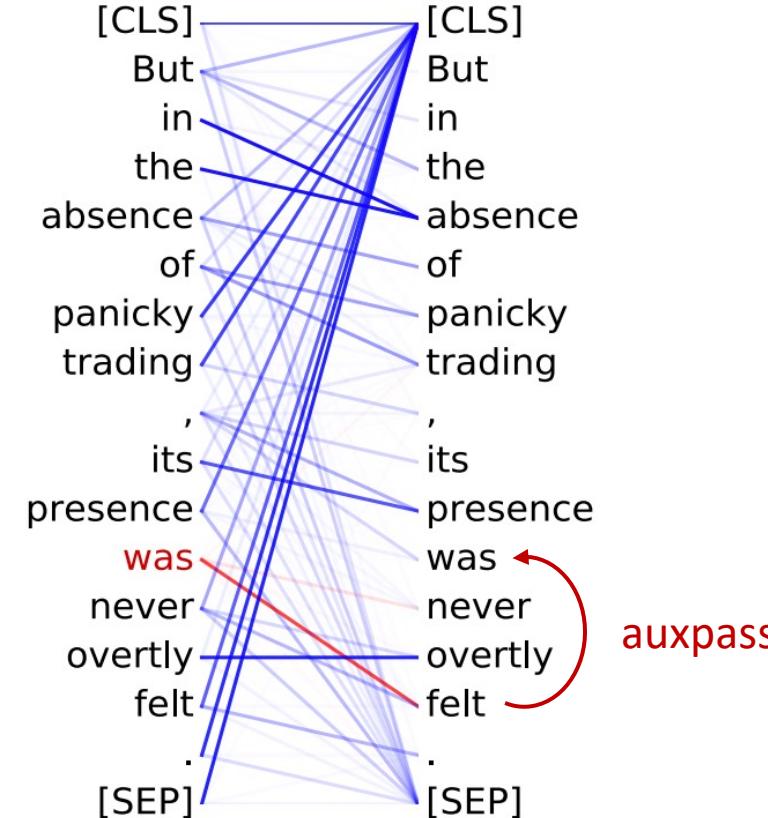
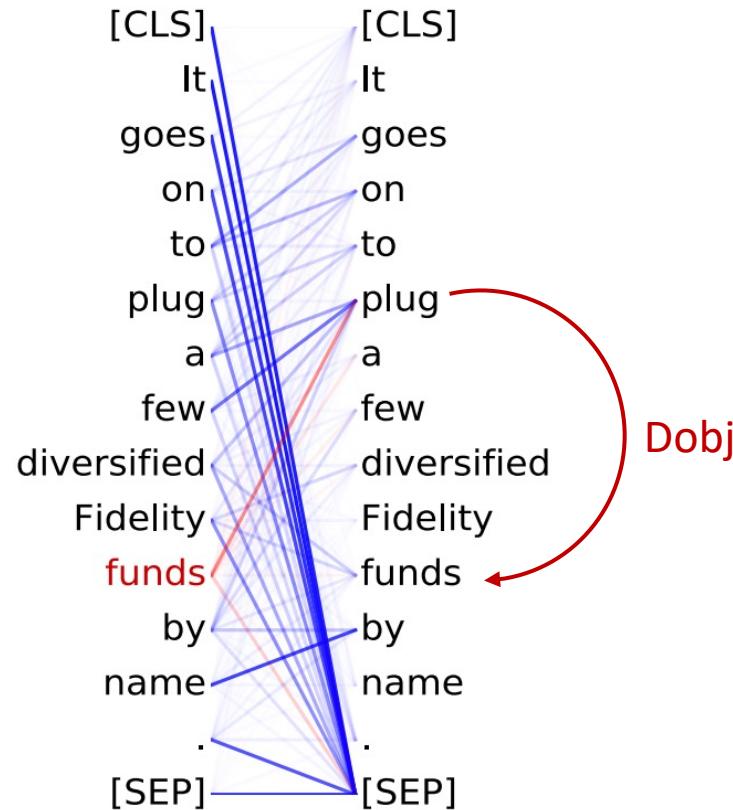
- Example #2: Factual knowledge



Knowledge forgetting across layers: Intermediate layers contain relational knowledge that is absent in the final layer

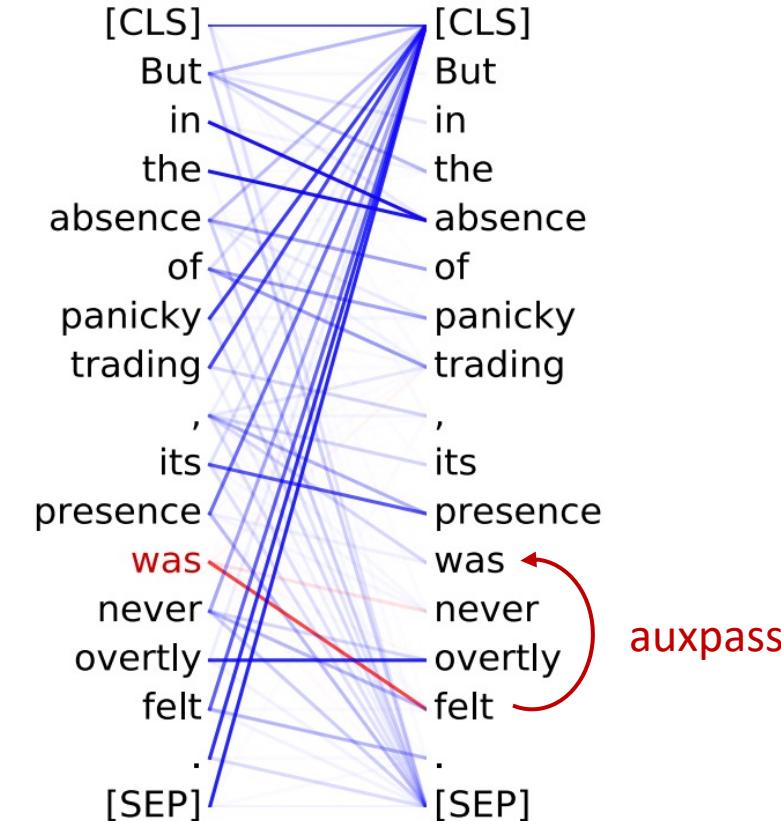
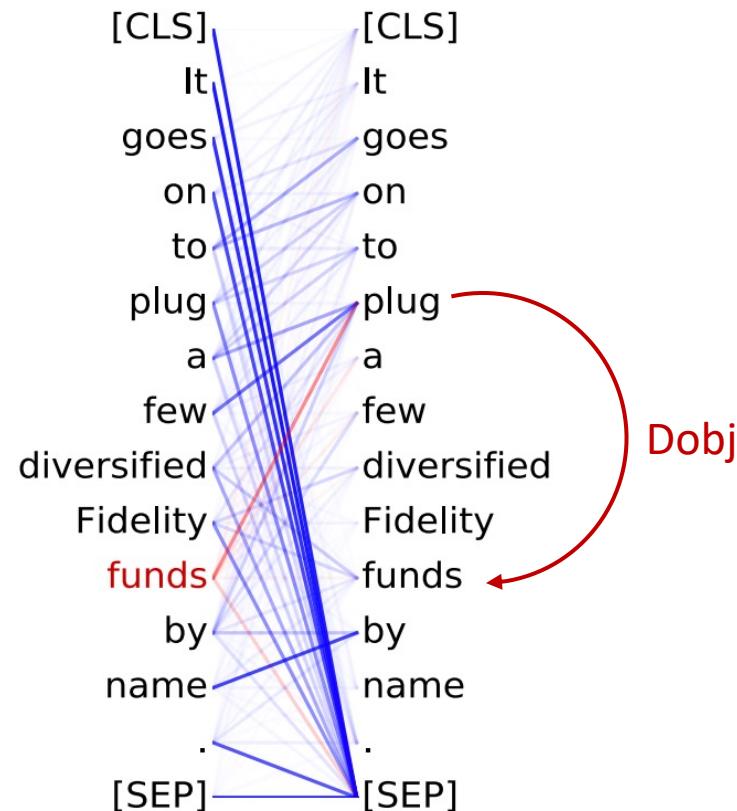
Module-based Knowledge Attribution

- Analyze knowledge attribution by looking into attention matrix(Clark et al., 2019; Htut et al., 2019; Lin et al., 2019)



Module-based Knowledge Attribution

- Attention-based knowledge attribution example #1: syntax knowledge (Clark et al., 2019)
 - Evaluate each attention head on dependency parsing dataset



Module-based Knowledge Attribution

- Attention-based knowledge attribution example #1: syntax knowledge (Clark et al., 2019)
 - Evaluate each attention head on dependency parsing dataset

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	76.3	34.6 (-2)
det	8-11	94.3	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	86.8	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	80.5	47.7 (1)
auxpass	4-10	82.5	40.5 (1)
ccomp	8-1	48.8	12.4 (-2)
mark	8-2	50.7	14.5 (2)
prt	6-7	99.1	91.4 (-1)

No single head does well at syntax "overall"

Module-based Knowledge Attribution

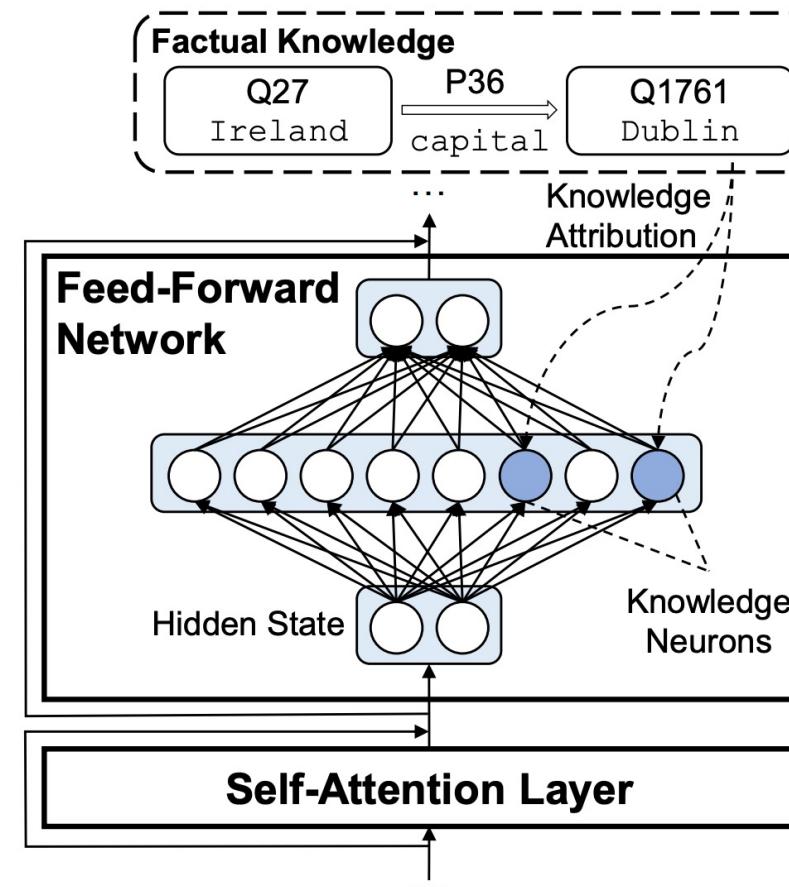
- Attention-based knowledge attribution example #1: syntax knowledge (Clark et al., 2019)
 - Evaluate each attention head on dependency parsing dataset

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	76.3	34.6 (-2)
det	8-11	94.3	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	86.8	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	80.5	47.7 (1)
auxpass	4-10	82.5	40.5 (1)
ccomp	8-1	48.8	12.4 (-2)
mark	8-2	50.7	14.5 (2)
prt	6-7	99.1	91.4 (-1)

Certain attention heads
specialize to specific
dependency relations.

Neuron-based Knowledge Attribution

- Can we attribute knowledge to specific neurons in PLMs?



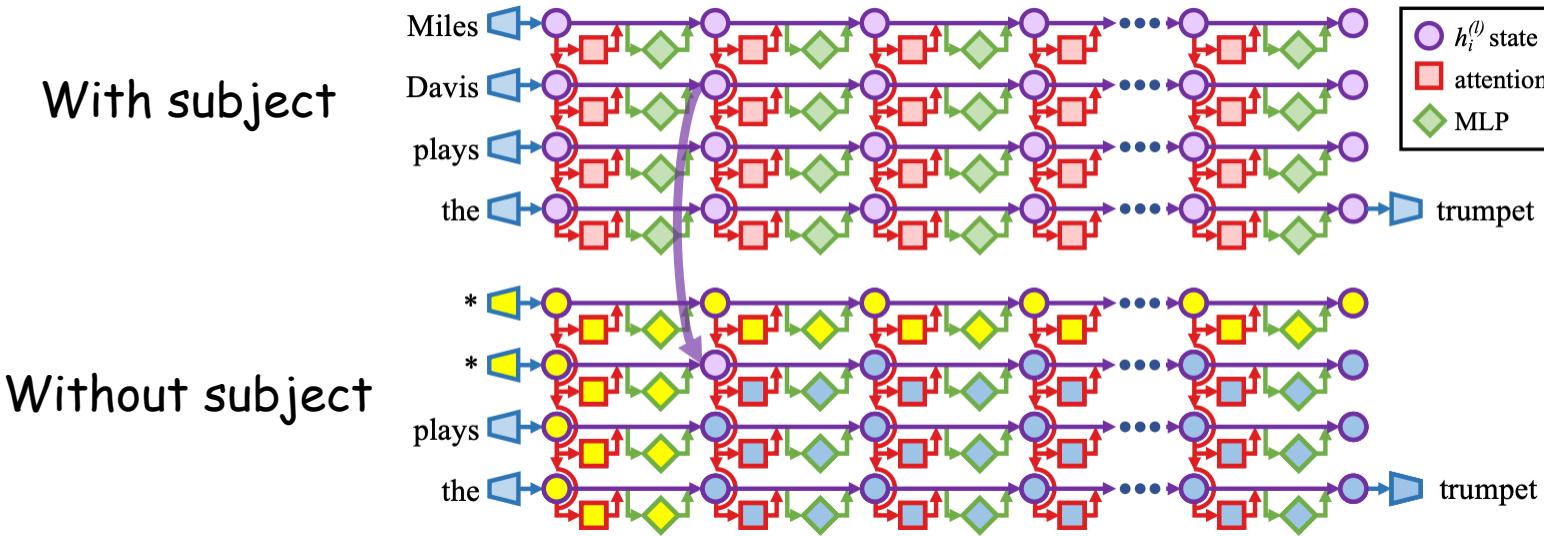
Neuron-based Knowledge Attribution

- How to find Knowledge Neuron: Integrate Gradients (Dai et al., 2022)

$Attr(w_i^{(l)})$: the probability changes caused by modifying $w_i^{(l)}$

Neuron-based Knowledge Attribution

- How to find Knowledge Neuron: Causal Tracing (Meng et al., 2022)



- Compute the causal effect of neuron activations by running the network twice
 - clean version of factual statement
 - corrupted version of factual statement

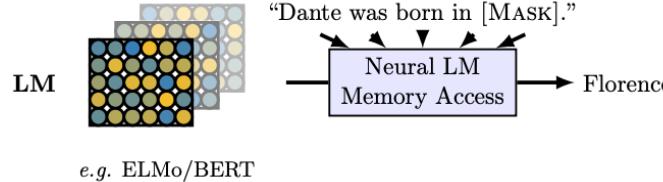
Take-aways for Knowledge Attribution

- Lower layers of PLMs often encode the coarse-grained and general information of knowledge
- Fine-grained and task-specific knowledge are mostly stored in higher layers
- Linguistic knowledge is distributed and represented in multiple attention heads, while a single attention head can only associate with a specific aspect of linguistics
- Factual knowledge can be associated with feed forward modules in middle or higher layers.

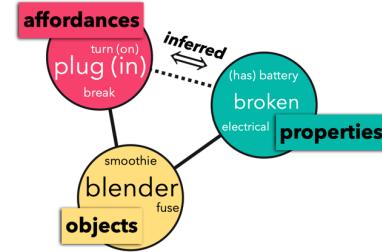
Knowledge Probing: How Much do LLMs Know about the World?

Knowledge Probing

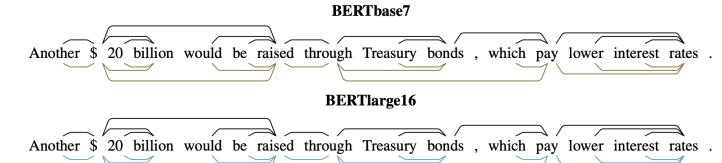
- Investigate the types of knowledge stored in LLMs



Factual Knowledge



Commonsense Knowledge



Linguistic Knowledge

- Quantify the amount of knowledge stored in LLMs

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE _n	RE _o	Fs	Txl	Eb	E5B	Bb	B1
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

Knowledge Probing Benchmarks

- Knowledge-specific Probing Benchmark
 - Focus on one specific kinds of abilities of LLMs

Knowledge Type	Benchmark	Formulation	Remark
Linguistic Knowledge	Open Sesame (Lin et al., 2019)	diagnostic classifier and attention	
	LKT (Liu et al., 2019b)	token or token pair labeling	
	NPI probe (Warstadt et al., 2019)	probing classifier	
	Edge probe (Tenney et al., 2019)	edge probing	
	MDL probe (Voita and Titov, 2020)	minimum description length	
	LM diagnostics (Ettinger, 2020)	text filling	
	BLiMP (Warstadt et al., 2020)	sentence scores comparison	
Factual Knowledge	LAMA (Petroni et al., 2019)	text filling	
	X-FACTR (Jiang et al., 2020a)	text filling	
	Multilingual LAMA (Kassner et al., 2021)	text filling	multilingual
	Bio LAMA (Sung et al., 2021)	text filling	biology
Commonsense Knowledge	CAT (Zhou et al., 2020a)	sentence scores comparison	
	NumerSense (Lin et al., 2020b)	text filling	numerical
	Physical Commonsense (Forbes et al., 2019)	probing classifier	physical

Knowledge Probing Benchmarks

- General Knowledge Evaluation Benchmark
 - General/Hybrid knowledge evaluation with higher difficulty

Existing dataset	LEval 2023.07 OpenMLab Website Reform	OpenCompass 2023.07 SAIL Existing Exam Reform	M3Exam 2023.06 Alibaba Exam	Xiezhi 2023.06 Fudan Website Exam	TRUSTGPT 2023.06 SCU Reform
Reconstruct Existing Dataset	KoLA 2023.06 THU Existing Website	Open LLM 2023.06 HuggingFace Existing	Conditional 2023.06 DeepMind Existing	AlpacaEval 2023.06 Stanford Reform	CMMLU 2023.06 MBZUAI Exam
Website-based	Promptbench 2023.06 MSRA Manually	Chatbot Arena 2023.05 FastChat Existing Manually	HaluEval 2023.05 RUC Reform	ZeroSCROLLS 2023.05 Meta Existing Reform	ToolBench 2023.05 SambaNova Manually
Exam-based	C-Eval 2023.05 SJTU&THU Exam	Chain-of-Thought 2023.05 Edinburgh Existing	llmEval 2023.05 Fudan Manually	API-Bank 2023.04 Alibaba Website	AGIEval 2023.04 Microsoft Exam
Manually Construct	Safety-Prompts 2023.03 THU Manually	Stanford HELM 2023.03 Stanford Existing	MEGA 2023.03 Microsoft Existing	GLUE-X 2022.11 WU Existing	bigbench 2022.06 Google Manually

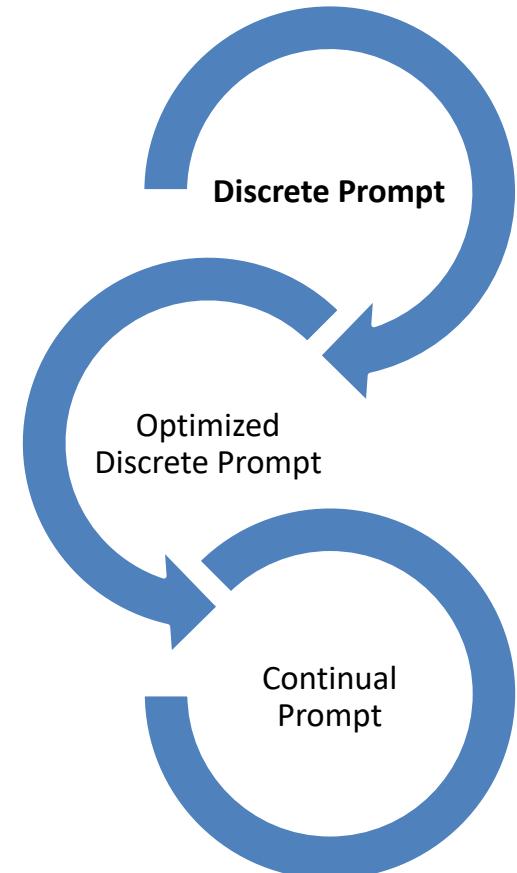
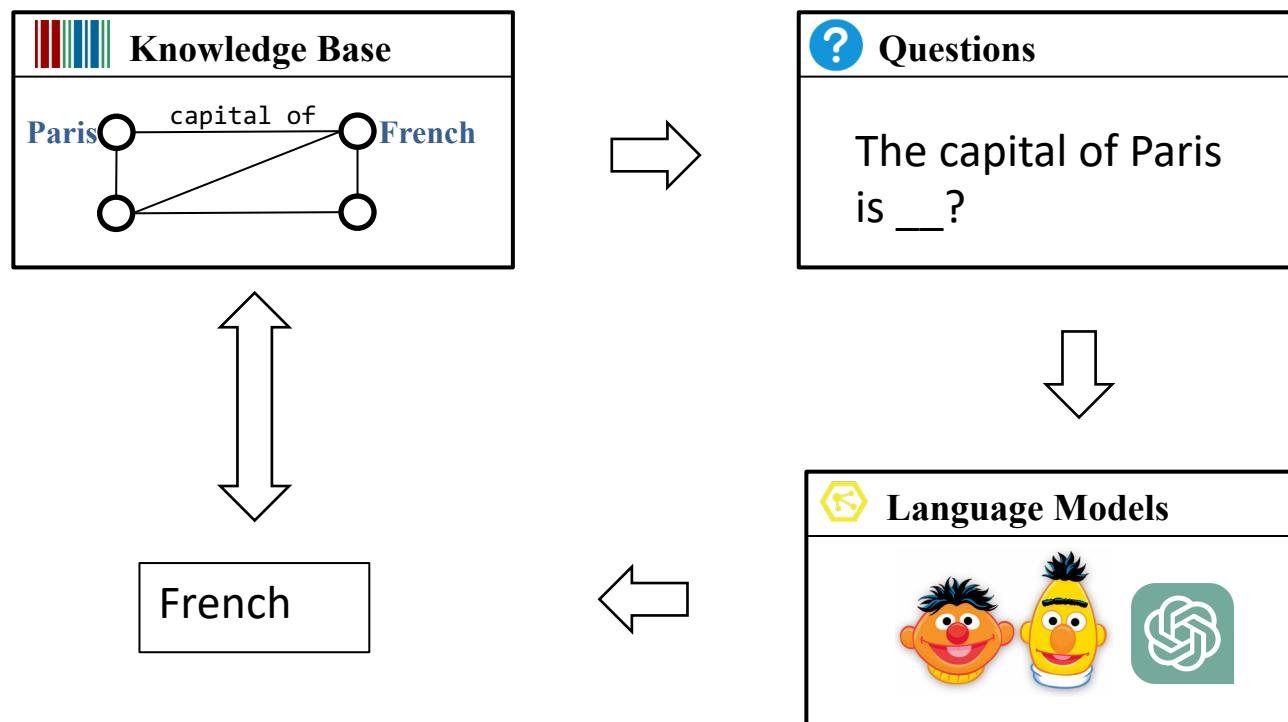
- Prompt-based knowledge probing
 - Query LMs with task-specific prompts and assess performance according to LMs' predictions
- Feature-based knowledge probing
 - Froze parameters of LLMs, probing tasks are accomplished based on the internal representation or attention weights produced by LMs

Prompt-based Knowledge Probing

- Handcraft Discrete Prompt
- Optimized Discrete Prompt
- Continual Prompt

Prompt-based Knowledge Probing

- Cloze-style Discrete prompts
 - LAMA, X-FACTR, BioLAMA, Multilingual LAMA...
 - Choice of Prompts has huge influence

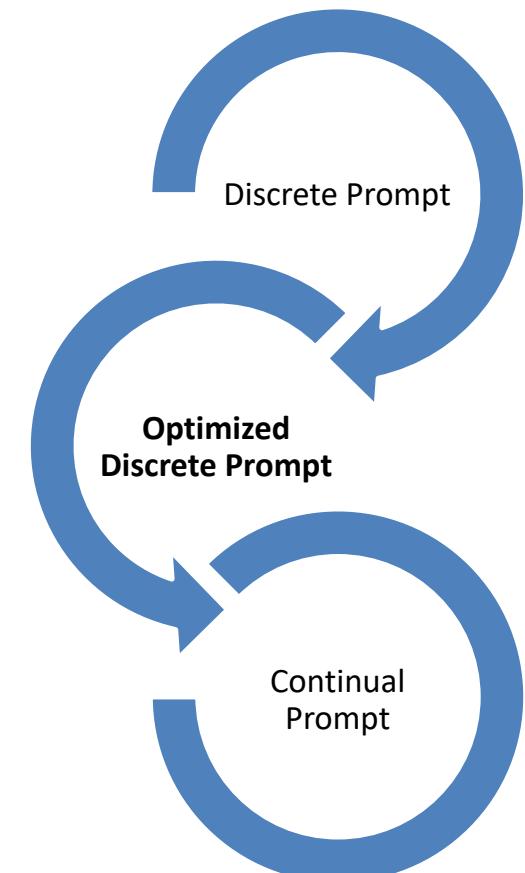


Prompt-based Knowledge Probing

- Using optimized discrete prompts to get better performance

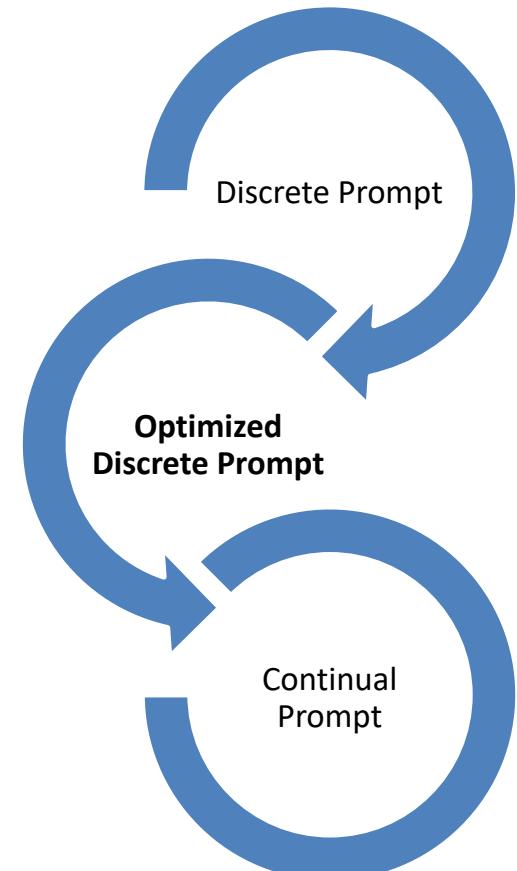
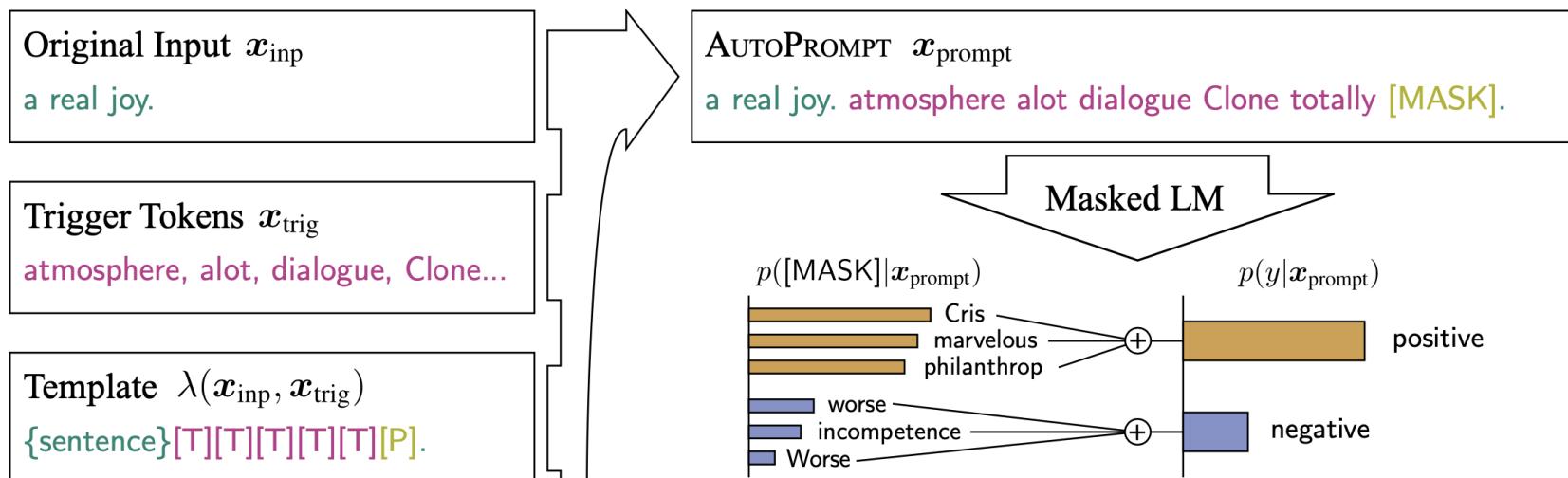
- Example #1: LPAQA
- Using retrieval and paraphrasing method to search prompts
- Achieve better performance than manually created prompts
- Require valid dataset

		Prompts		
	manual	<i>DirectX is developed by</i> y_{man}		
	mined	y_{mine} released the DirectX		
	paraphrased	<i>DirectX is created by</i> y_{para}		
Top 5 predictions and log probabilities				
	y_{man}	y_{mine}	y_{para}	
1	Intel -1.06	<u>Microsoft</u> -1.77	<u>Microsoft</u> -2.23	
2	<u>Microsoft</u> -2.21	They -2.43	Intel -2.30	
3	IBM -2.76	It -2.80	default -2.96	
4	Google -3.40	Sega -3.01	Apple -3.44	
5	Nokia -3.58	Sony -3.19	Google -3.45	



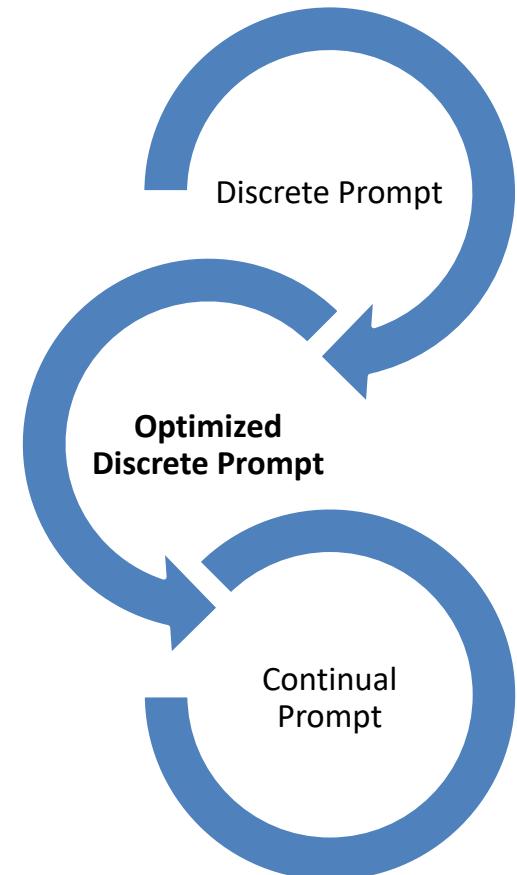
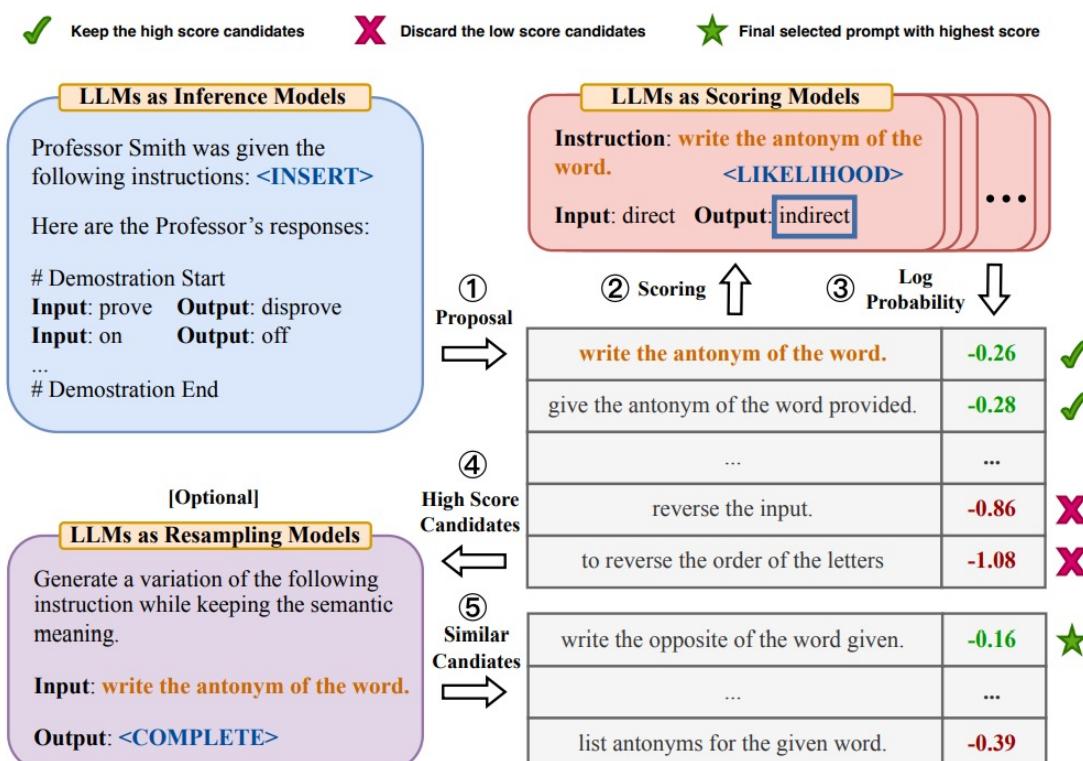
Prompt-based Knowledge Probing

- Using optimized discrete prompts to get better performance
 - Example #2: AutoPrompt
 - Automatically generated prompts based on gradient-guided search
 - Discrete prompts with better performance but lack of interpretability



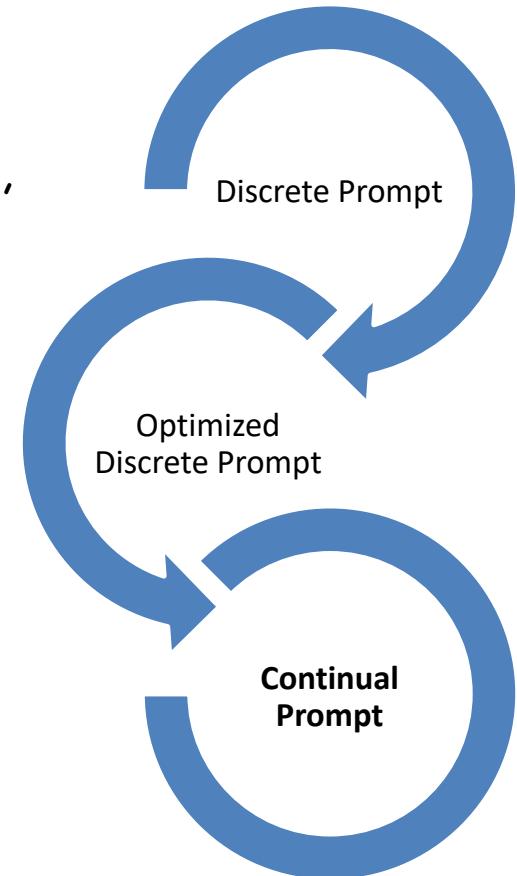
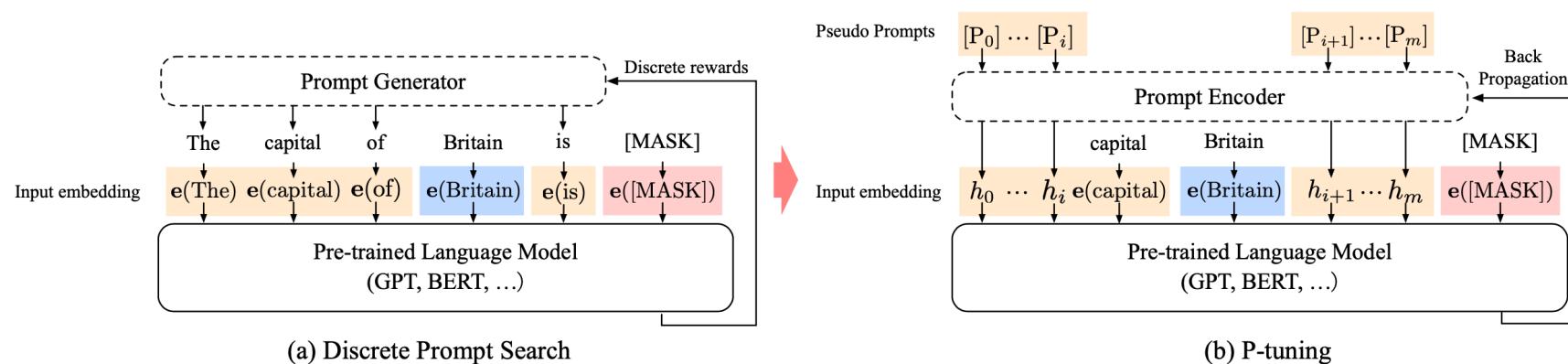
Prompt-based Knowledge Probing

- Using optimized discrete prompts to get better performance
 - Example #3: APE
 - Automatic prompt generation and selection using LLMs



Prompt-based Knowledge Probing

- Prompt-tuning: adding trainable vectors as soft prompt can further improve performance
 - Handcraft prompts initialization (Zhong et al., 2021)
 - Adding continual prompts on both input and transformer blocks (Li and Liang, 2021)
 - Adding prompt encoder above the input embeddings (Liu et al., 2021)
 - Ensembling multiple soft prompts (Qin et al. 2021)



Liu X, Zheng Y, Du Z, et al. GPT Understands, Too (2021)

Zhong Z, Friedman D, Chen D. Factual Probing Is [MASK]: Learning vs. Learning to Recall (2021)

Qin G, Eisner J. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts (2021)

Manually
Created

Discrete
Search

Continuous
Training

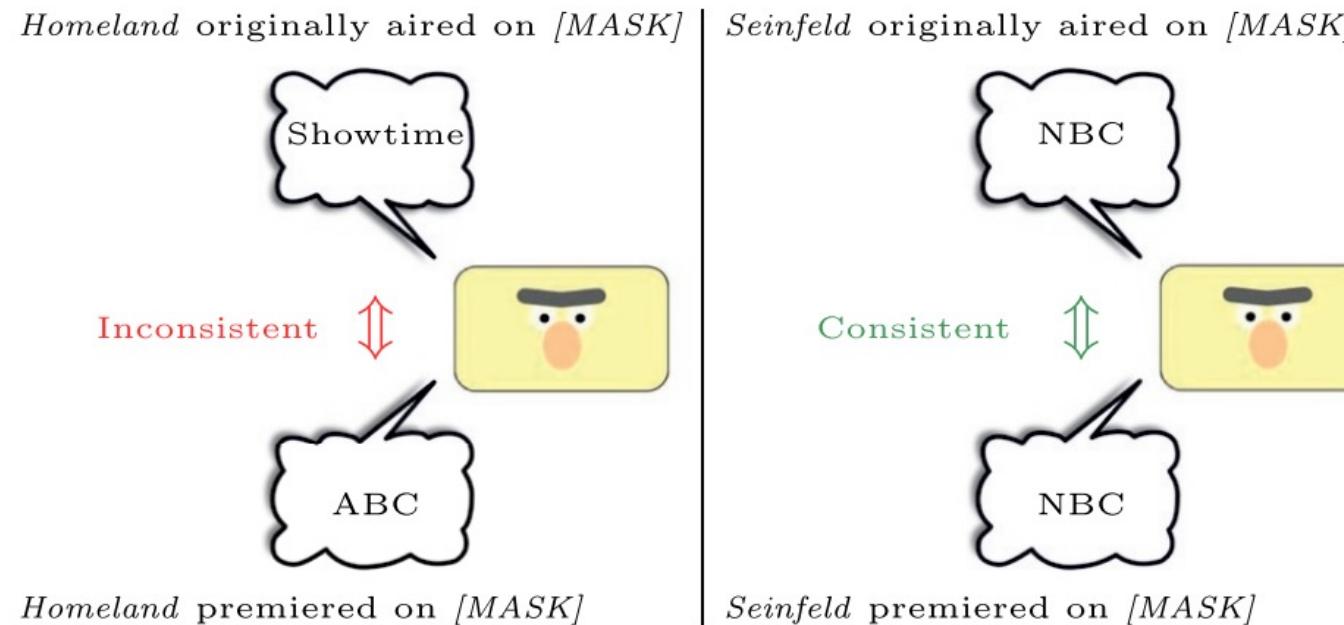


Better and better performance, weaker and weaker interpretability.

Can we absolutely trust the evaluate results of prompt-based probing?

Prompt-based Knowledge Probing

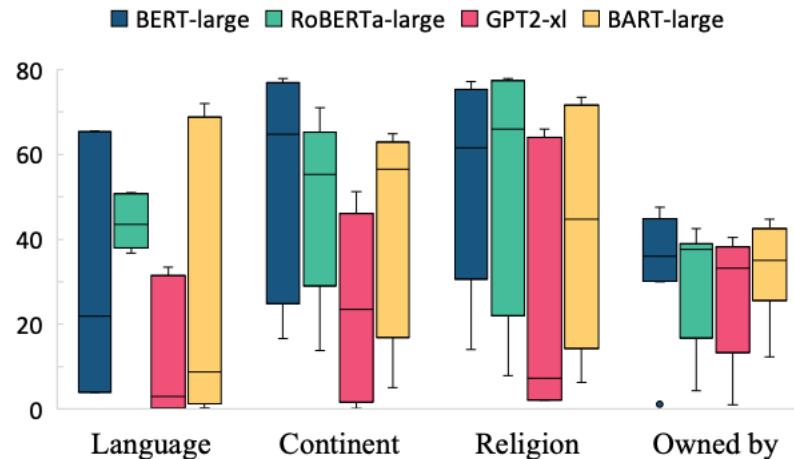
- Prompt-based probing could be inconsistent



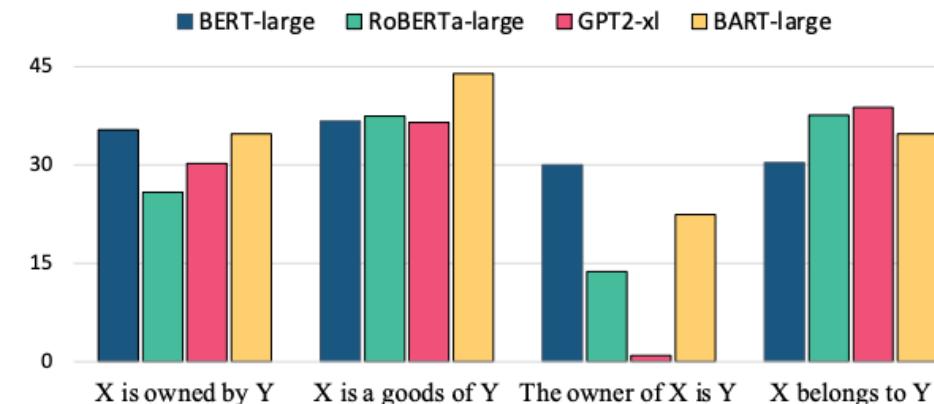
Semantically equivalent prompts may result in different answer

Prompt-based Knowledge Probing

- Prompt-based probing could be inconsistent



Performance variances of PLMs on semantically equivalent prompts.

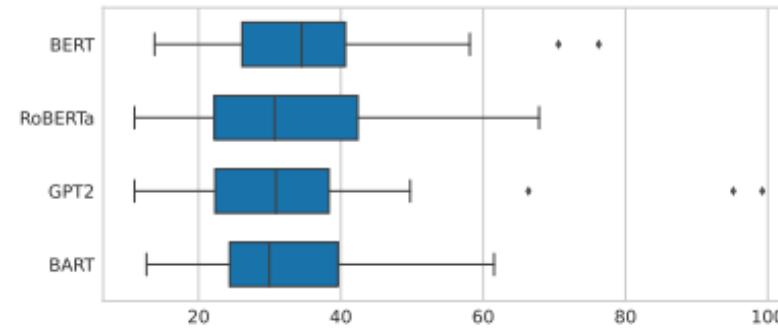
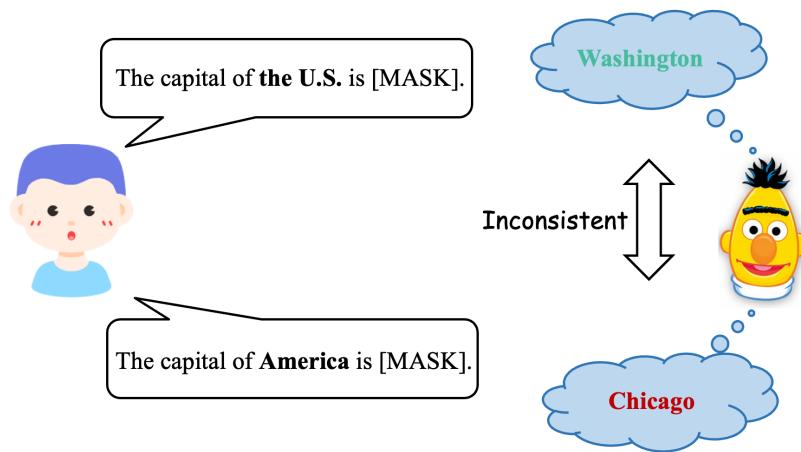


Inconsistent comparison between PLMs when prompts varies.

Prompt preference leads to inconsistent performance and comparison

Prompt-based Knowledge Probing

- Prompt-based probing could be inconsistent

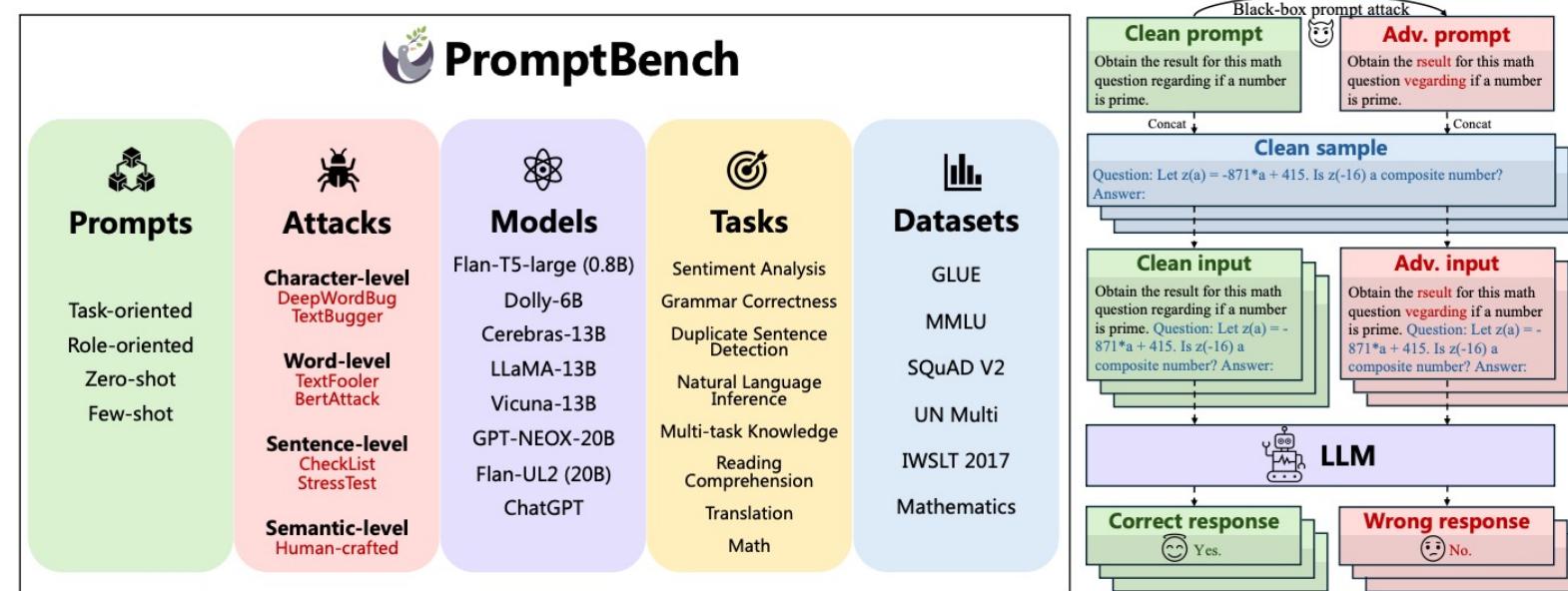


Verbalization stabilities of 4 PLMs.

Predictions are sensitive and inconsistent to various verbalizations

Prompt-based Knowledge Probing

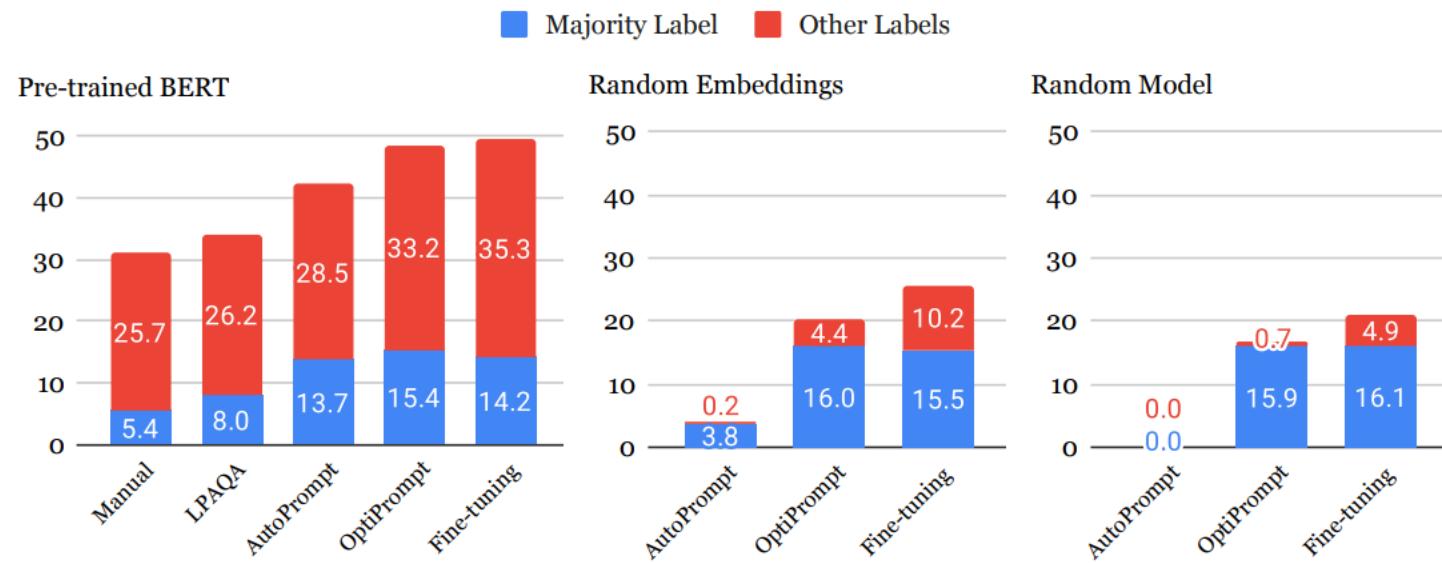
- Prompt-based probing could be susceptible



LLMs are susceptible to adversarial prompts

Prompt-based Knowledge Probing

- Optimized prompt could be unreliable



Optimized prompts can exploit regularities in training data

Prompt-based Knowledge Probing

- Optimized prompt could be unreliable

Relation	Prompt	Source	Prec.	KL.
citizenship	x is y citizen	T_{man}	0.00	24.67
	x returned to y	T_{mine}	43.58	6.32
work location	x used to work in y	T_{man}	11.01	19.07
	x was born in y	T_{mine}	40.25	2.21
instance of	x is a y	T_{man}	30.15	22.98
	x is a small y	T_{mine}	52.60	13.98

"Better" prompts may be the prompts fitting the answer distribution better

Prompt-based Knowledge Probing

- Q&A based Evaluation for aligned models
 - MMLU, Stanford HELM, OpenLLM, CMMLU, C-Eval...

?

Questions

Where is the capital of France?

(A) Beijing
(B) Tokyo
(C) Paris
(D) Washington

Answer:

Multiple Choice

?

Questions

Tell me some trivia about penguins

Free-style Writing

Does correct (wrong) answer mean the model has (don't have) the knowledge?

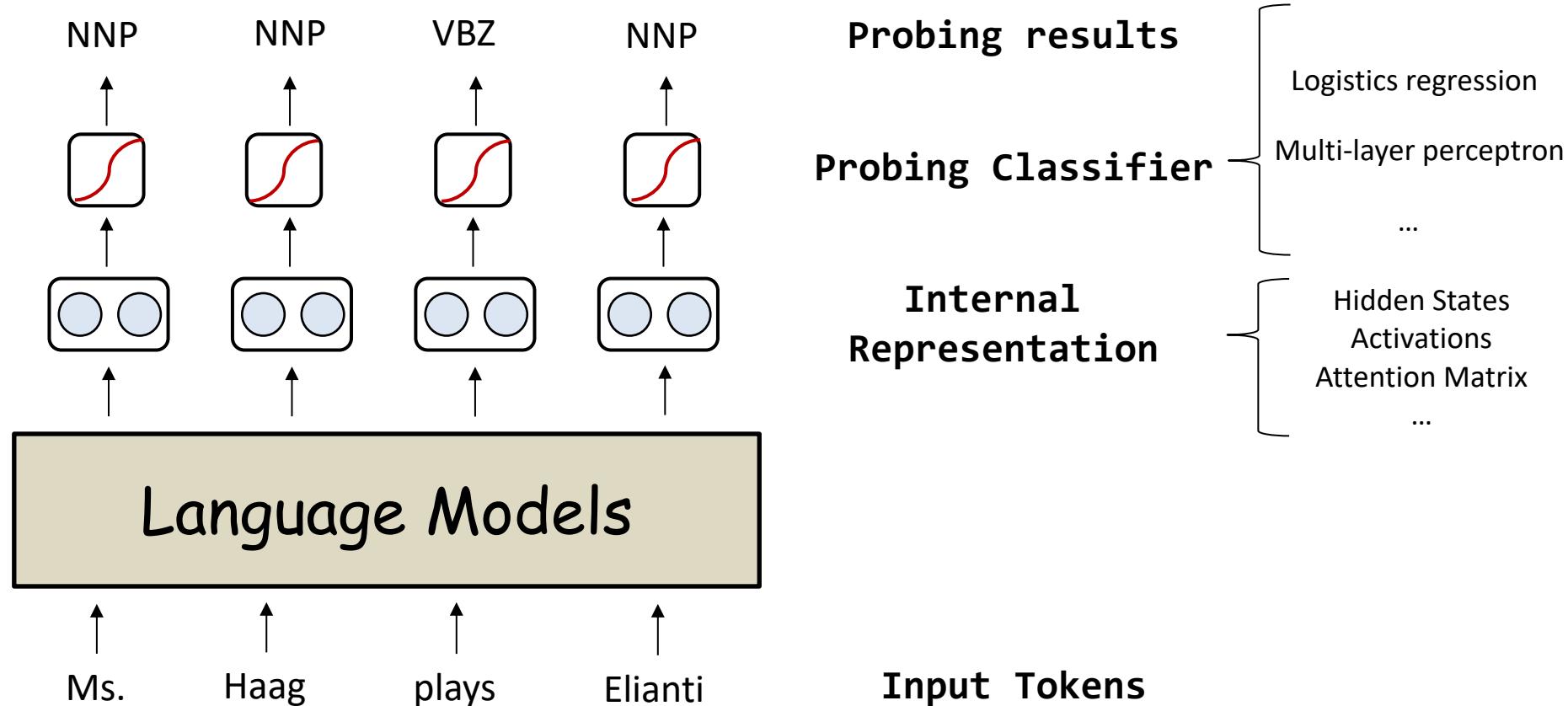
Prompt-based Knowledge Probing

- Q&A based Evaluation for aligned model
 - Erlangshen-UniMC-1.3B achieve strong performance on C-EVAL
 - Pre-trained on 180G text corpus and fine-tuned on multiple choice dataset

14	ChatGLM2-12B	Tsinghua & Zhipu.AI	2023/7/26	61.6	42	55.4	73.7	64.2	59.4
15	DFM2.0	AISpeech & SJTU	2023/8/15	61.4	40.2	50.9	72.8	65.9	65.4
16	Erlangshen-UniMC-1.3B	IDEA研究院	2023/8/4	61	36.7	49.6	74.9	70.7	59.4
17	CHAOS_LM-7B	OPPO Research Institute	2023/8/17	60.8	49.1	59.9	70.1	58.9	55.7
18	UniGPT	Unisound	2023/7/26	60.3	46.4	57.7	69.3	58	59
19	MiLM-6B	Xiaomi	2023/8/9	60.2	42	54.5	71.7	62.7	57.7
20	Qwen-7B	Alibaba Cloud	2023/7/29	59.6	41	52.8	74.1	63.1	55.2
21	BatGPT-15b-sirius-v2	SJTU & WHU	2023/8/4	57.4	36.9	50.5	72.1	60.7	53.3
22	Instruct-DLM-v2	DeepLang AI	2023/7/2	56.8	37.4	50.3	71.1	59.1	53.4
23	XVERSE-13B	XVERSE Technology	2023/8/6	54.7	33.5	45.6	66.2	58.3	56.9
24	HITsz-Lychee-Base-11B-V0.1	HITsz (哈工大深圳)	2023/8/6	54.7	44	50.8	61.3	57	53.8
25	EduChat	ECNU (华东师范大学)	2023/8/17	54.6	37.5	47.2	66.7	59.4	52.4
26	ChatGPT*	OpenAI	2023/5/15	54.4	41.4	52.9	61.8	50.9	53.6
27	Claude-v1.3*	Anthropic	2023/5/15	54.2	39	51.9	61.7	52.1	53.7

Feature-based Knowledge Probing

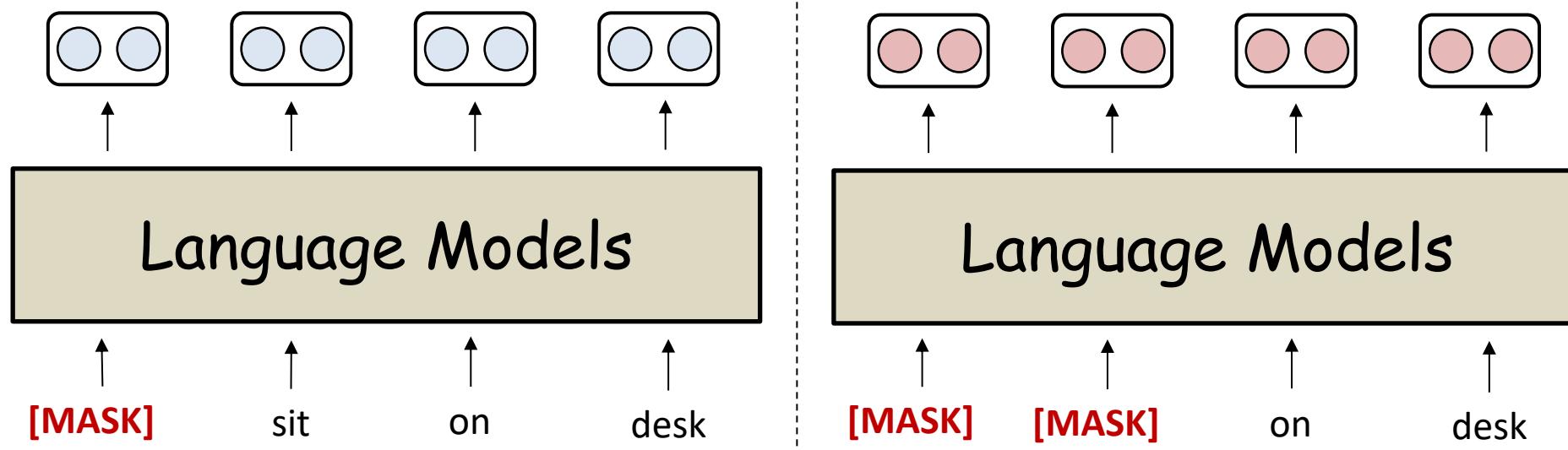
- Feature based probing with classifier (Lin et al., 2019; Clark et al., 2019; Tenney et al., 2019; Liu et al., 2019;)



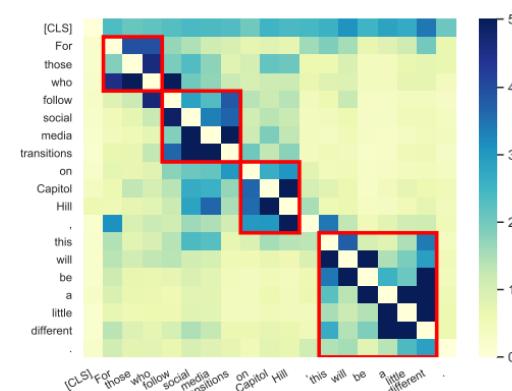
- Classifier may be unreliable
 - Training process involved
 - Difficult for results attribution
 - Inconsistent between classifiers
- Can we use feature-based probing without classifier?

Feature-based Knowledge Probing

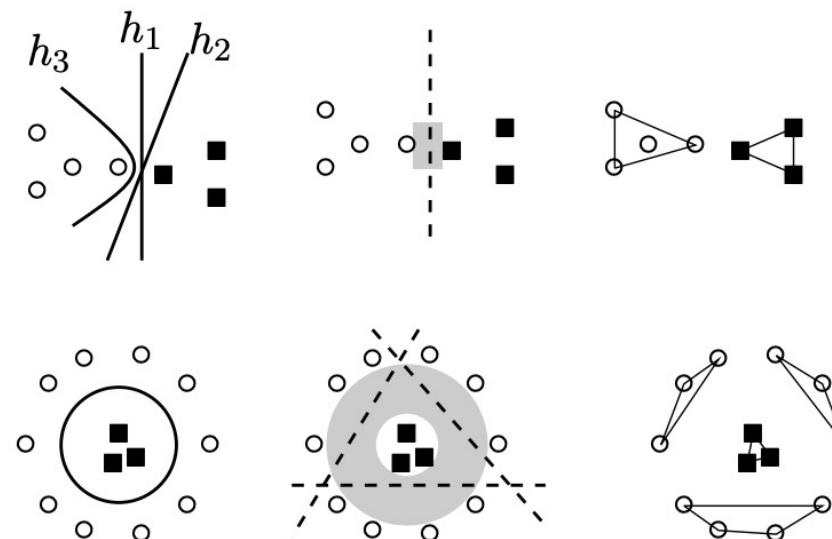
- Feature-based probing without classifier: example#1 perturbed masking (Wu et al., 2020)



- Perturbed Masking
 - Calculate impact sit has on Cats
 - $e_i = E(\text{Cats} | S \setminus \{\text{Cats}\})$
 - $e_j = E(\text{Cats} | S \setminus \{\text{Cats}, \text{sit}\})$
 - $F(\text{Cats}, \text{sit}) = d(e_i, e_j)$



- Feature-based probing without classifier: example#2 Direct Probe (Zhou et al., 2021)
 - Each classifier is a decision boundary in the representation space
 - Consider the representation probing as clustering problem

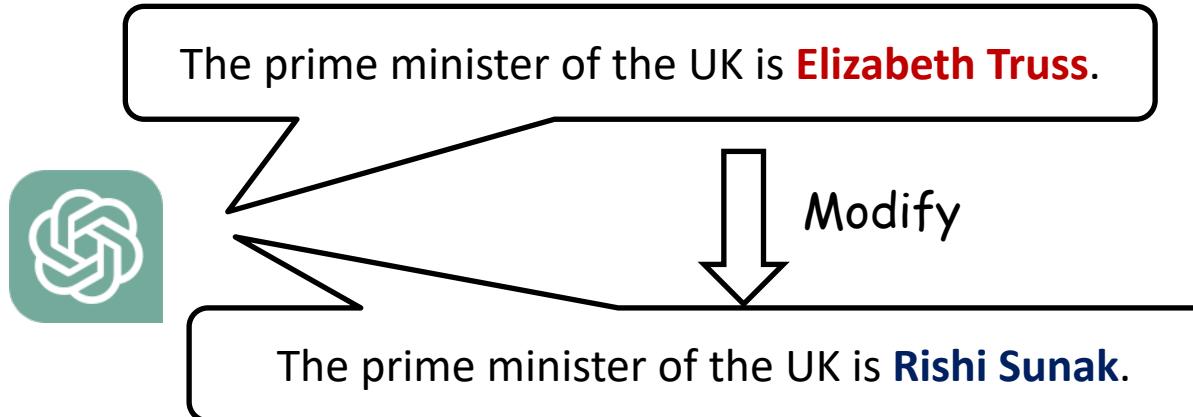


Take-aways for Knowledge Probing

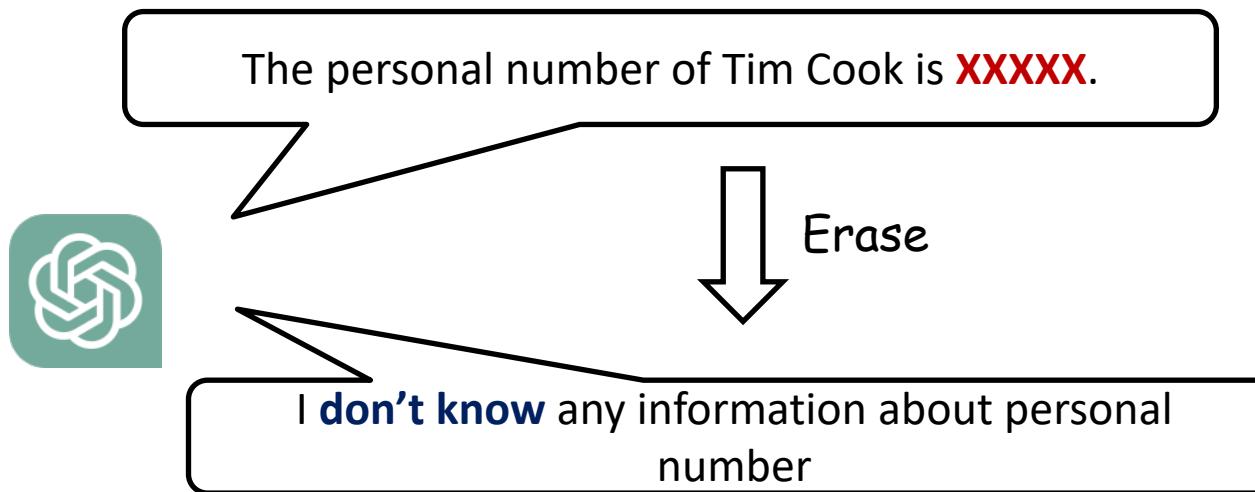
- Both prompt-based and feature-based probing have their own limitations
- Prompt-based evaluation could be biased by data distribution, prompt selections, etc.
- Design of better probing framework for LLMs worth further investigation

Knowledge Editing: Updating and Deleting

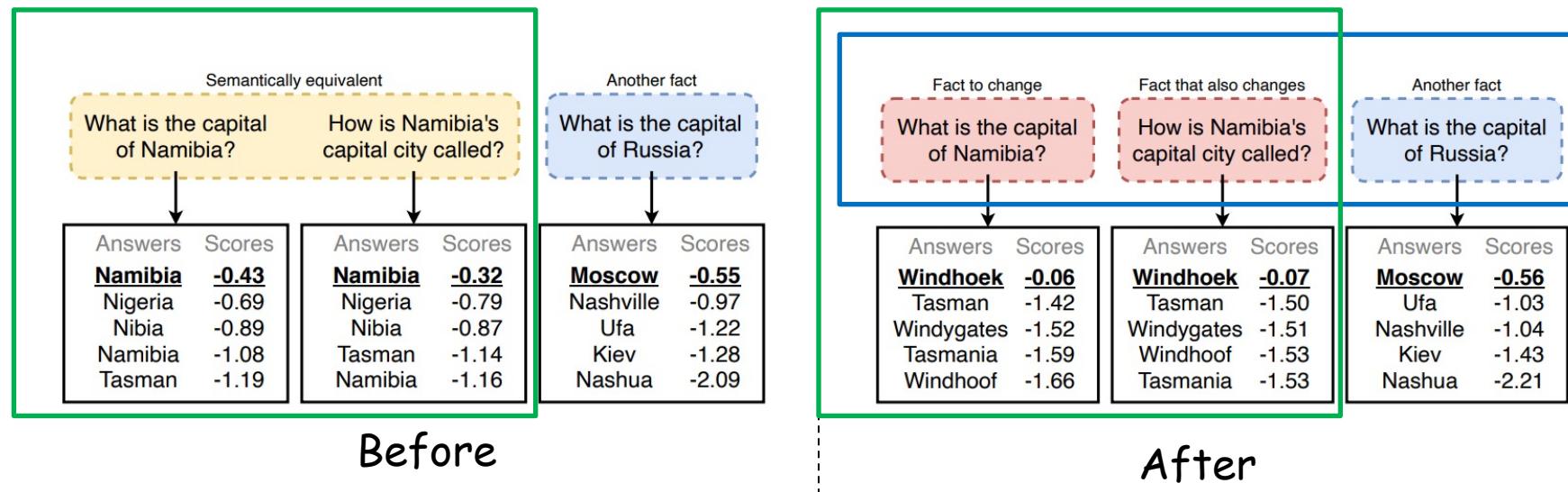
- Replacing stored knowledge in PLMs with new knowledge



- Removing stored knowledge entirely



What is Successful Edit



- **Generality:**
 - Suitable for general pre-trained language models.
- **Reliability:**
 - Be able to successfully update target knowledge without affecting the rest.
- **Consistency:**
 - The changes should be consistent across equivalent formulations of a fact

Knowledge Editing Strategies

- Constrained tuning
 - Fine-tuning on target knowledge without affecting the rest
- Meta-Learning based editing
 - Learning to update: learning to predict updated parameters
- Memory-based editing
 - Maintain a edit memory and reason over it as needed
- Locate and edit
 - Attribute knowledge to specific neurons and edit them accordingly

- Naive Solution 1: Re-training
- Re-train PLM using the updated training dataset
 - Computationally expensive and impractical when LLMs involved
- Fine-tune PLMs on a small subset which only contains target knowledge
 - Suffer from catastrophic forgetting, and affects the rest knowledge which is not intended to be edited.

Constrained Fine-tuning

- Constraint 1: Learn the new facts while keeping the loss small on unmodified facts

$$\text{minimize}_{\theta \in \Theta} \quad \frac{1}{m} \sum_{x \in \mathcal{D}_M} L(x; \theta)$$

subject to

$$\frac{1}{n} \sum_{x' \in \mathcal{D}_{F \setminus S}} (L(x'; \theta) - L(x'; \theta_0)) \leq \delta.$$

Minimize loss on target knowledge

Keep loss small on unrelated knowledge

- Constraint 2: Using normalization to constrain the parameters change of the models.

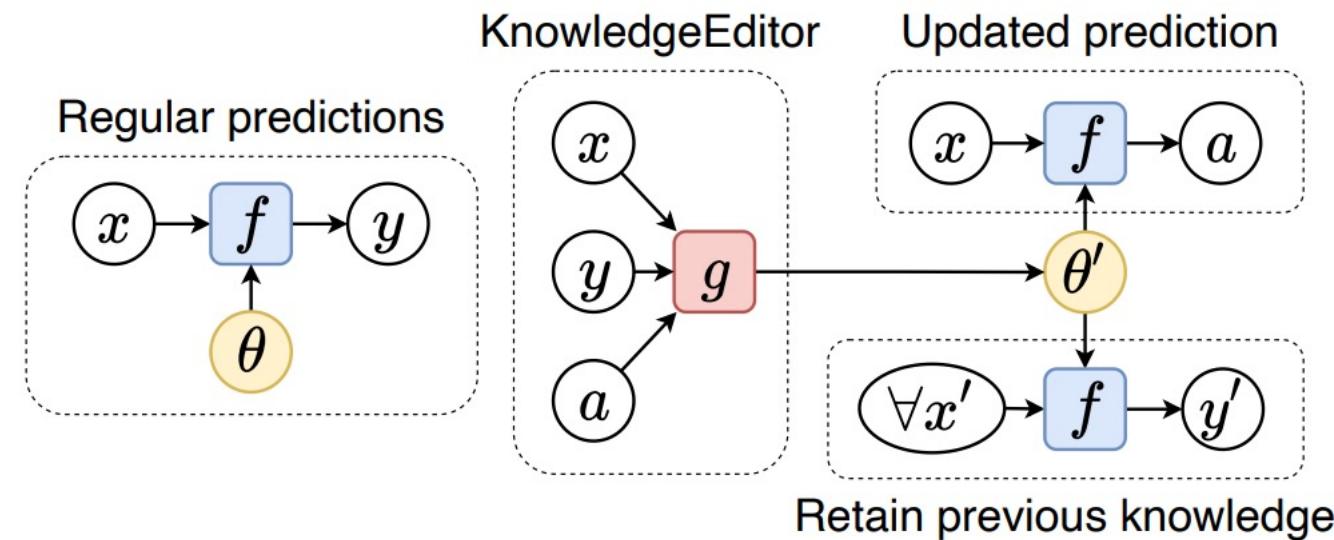
$$\text{minimize}_{\theta \in \Theta} \quad \frac{1}{m} \sum_{x \in \mathcal{D}_M} L(x; \theta)$$

subject to $\|\theta - \theta_0\| \leq \delta,$

l_2 or l_∞ norm

Meta-Learning Based Knowledge Editing

- Example #1 - KnowledgeEditor: train a **hyper-network** to predict the parameter update



Replace the prediction of x from y to a , without affecting the predictions of any other input.

Meta-Learning Based Knowledge Editing

- Example #1 - KnowledgeEditor: train a **hyper-network** to predict the parameter update

Changing prediction successfully

$$\min_{\phi} \sum_{\hat{x} \in \mathcal{P}^x} \mathcal{L}(\theta'; \hat{x}, a)$$

semantically equivalent inputs of x

Not affect the rest

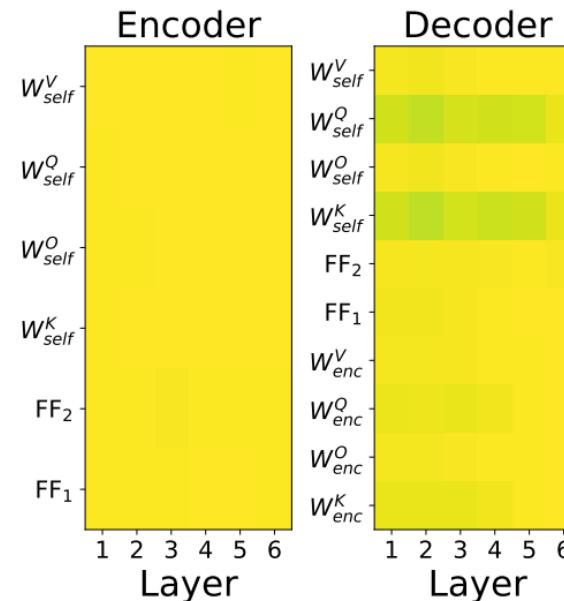
$$\text{s.t. } \boxed{\mathcal{C}(\theta, \theta', f; \mathcal{O}^x) \leq m}$$

\downarrow

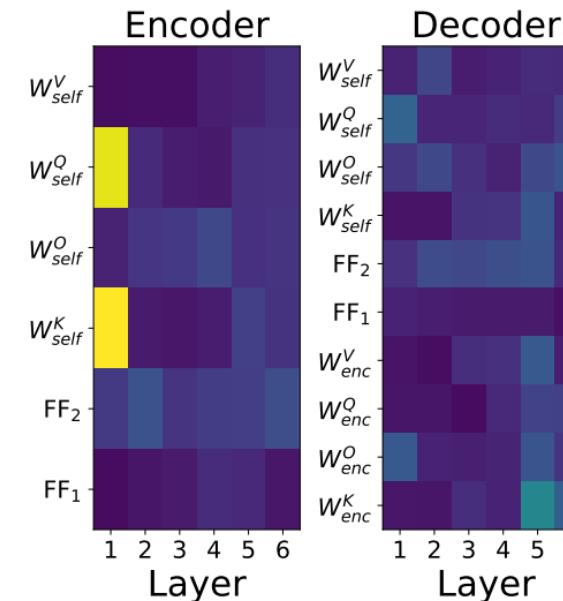
$$\sum_{x' \in \mathcal{O}^x} \sum_{c \in \mathcal{Y}} p_{Y|X}(c|x', \theta) \log \frac{p_{Y|X}(c|x', \theta)}{p_{Y|X}(c|x', \theta')}$$

Meta-Learning based Knowledge Editing

- Fine-tuning V.S. Hyper-network: fine-tuning updates all layers uniformly while hyper-network updates are more sparse.



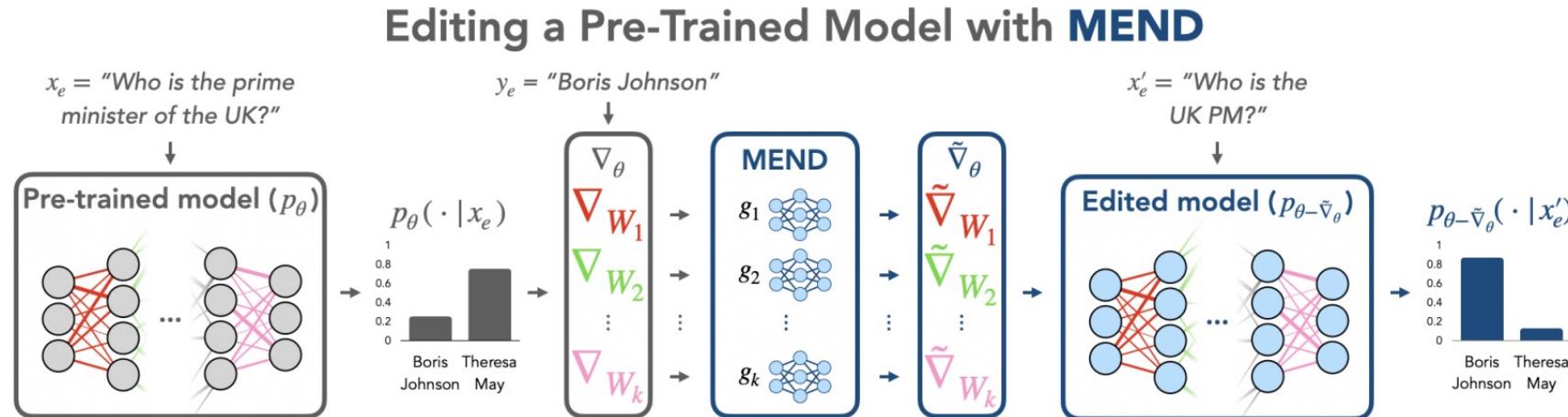
(b) Fine-tune (all layers).



(c) KNOWLEDGEEDITOR + \mathcal{P}^x .

Meta-Learning based Knowledge Editing

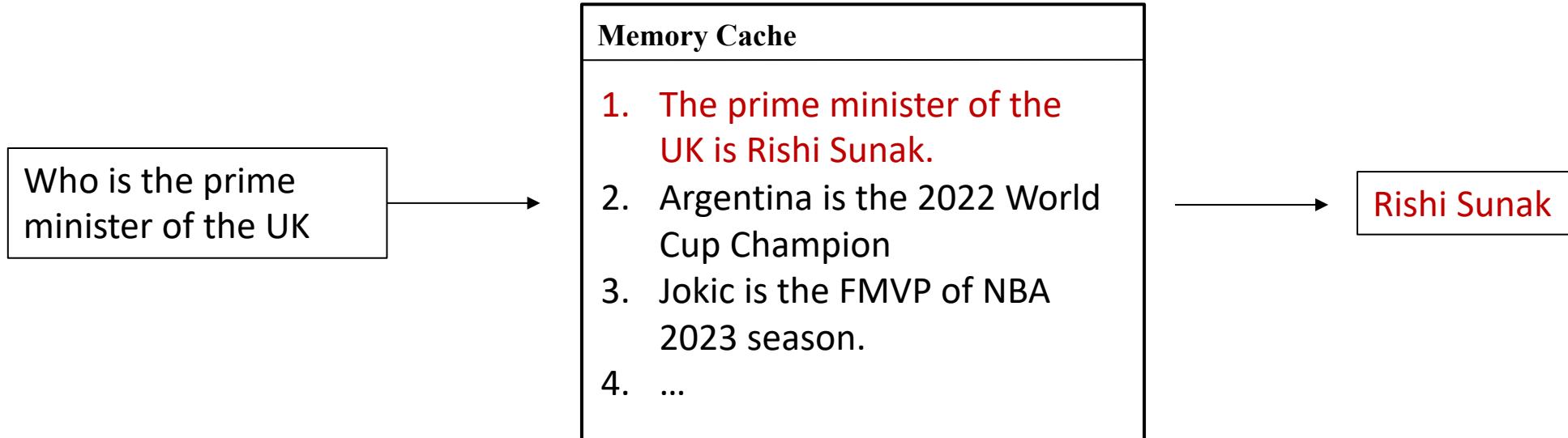
- Example #2 - MEND: predict the edits to LMs' weights based on the standard fine-tuning gradient with correction



- Get the prediction of target input.
- Calculate the standard fine-tuning gradient with correction
- Predict the updated weights
- Edit PLMs and check the updated knowledge

Memory-based Knowledge Editing

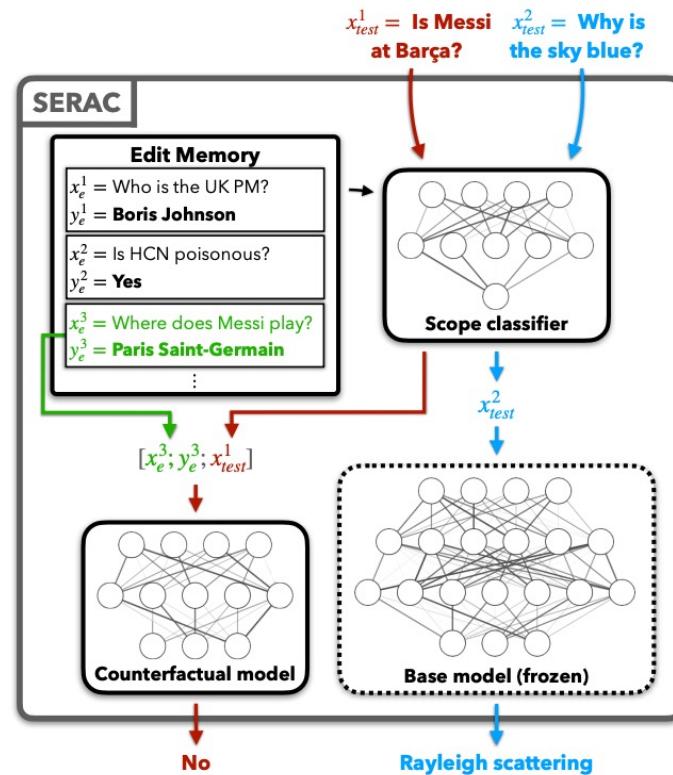
- Naïve Solution 2: maintain a symbolic memory cache
 - a symbolic knowledge cache may suffer from robustness issues



How to distinguish a relevant query?

Memory-based Knowledge Editing

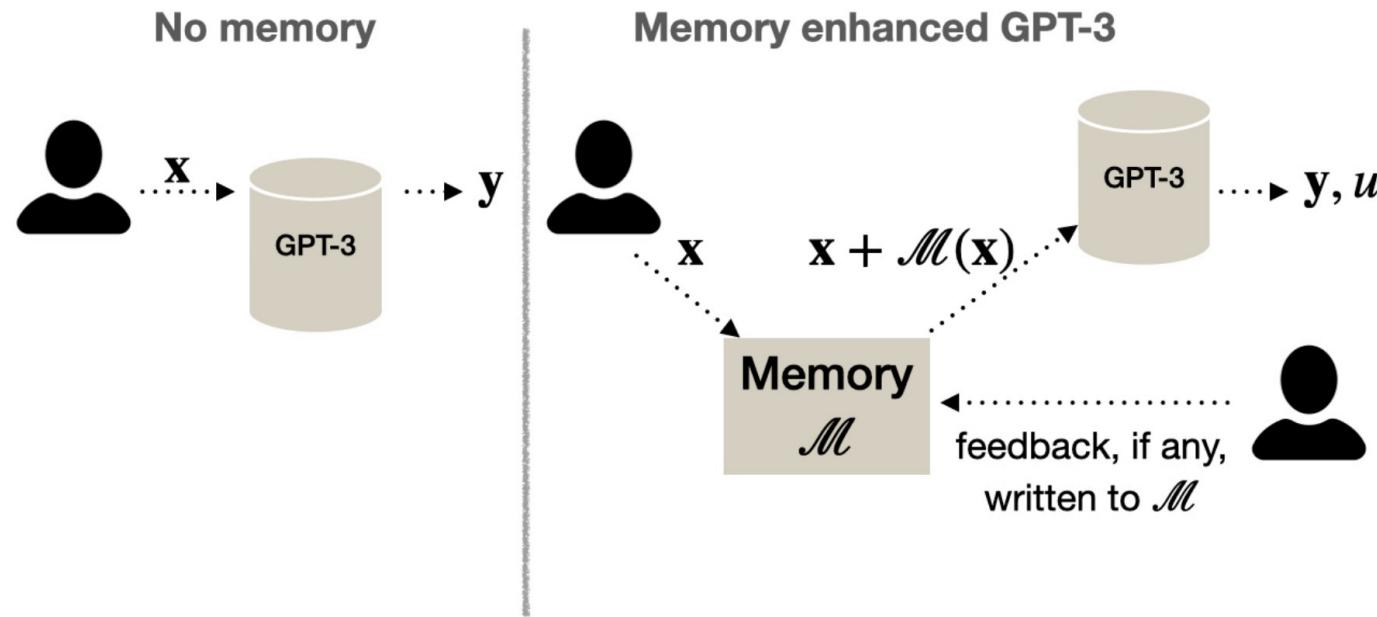
- Example#1 - SERAC: stores edits in a memory and learns to reason over them as needed



- Step 1: Maintain a edit memory
- Step 2: Decide whether a relevant edit exists in memory
- Step 3.1: **Irrelevant** - Using original LM to predict irrelevant question.
- Step 3.2: **Relevant** - Input and edited example are passed to a counterfactual model

Memory-based Knowledge Editing

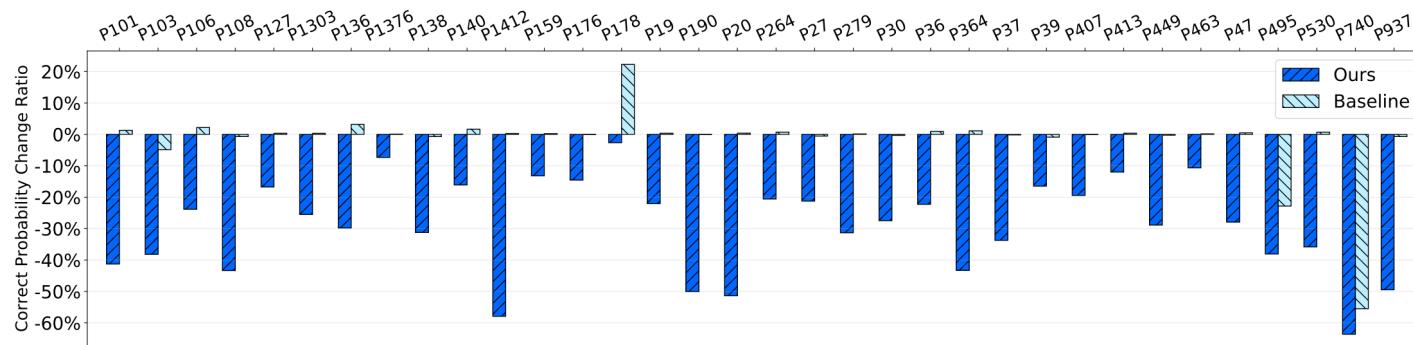
- Example#2 - MemPrompt: directly add edit information to the query



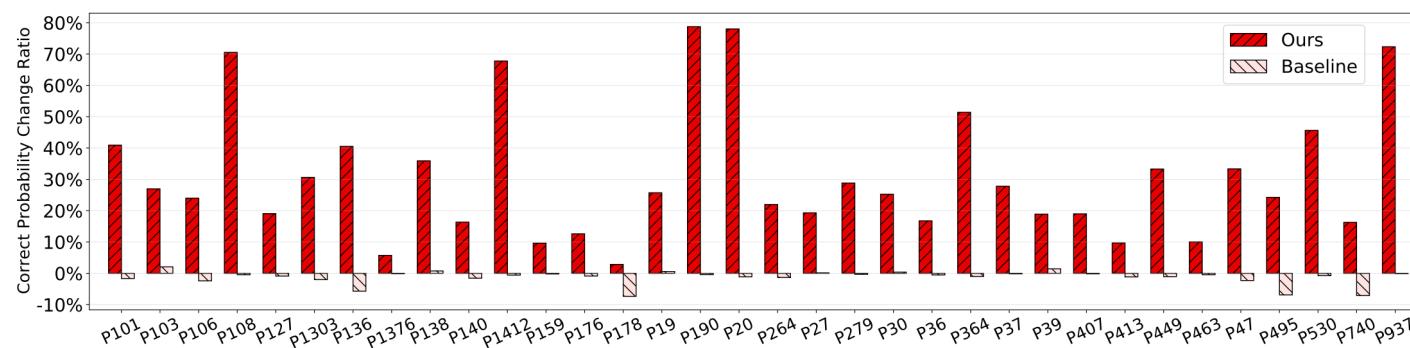
- Maintain a memory of past feedback
- Lookup for relevant memory
- Directly add to the query

- Combine knowledge attribution and editing
 - Knowledge Attribution: find the responsible neurons for specific knowledge
 - Knowledge Editing: edit the responsible neurons only

- Example #1: KnowledgeNeuron (Dai et al., 2022)
 - Direct modify the activations of knowledge neurons

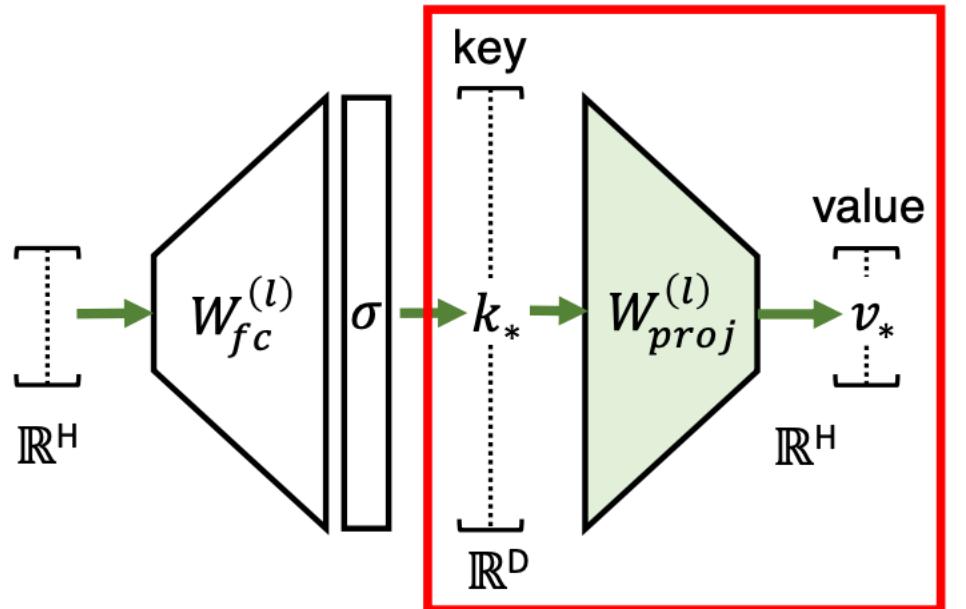


Answer probability decrease: setting activations to 0



Answer probability increase: double activations

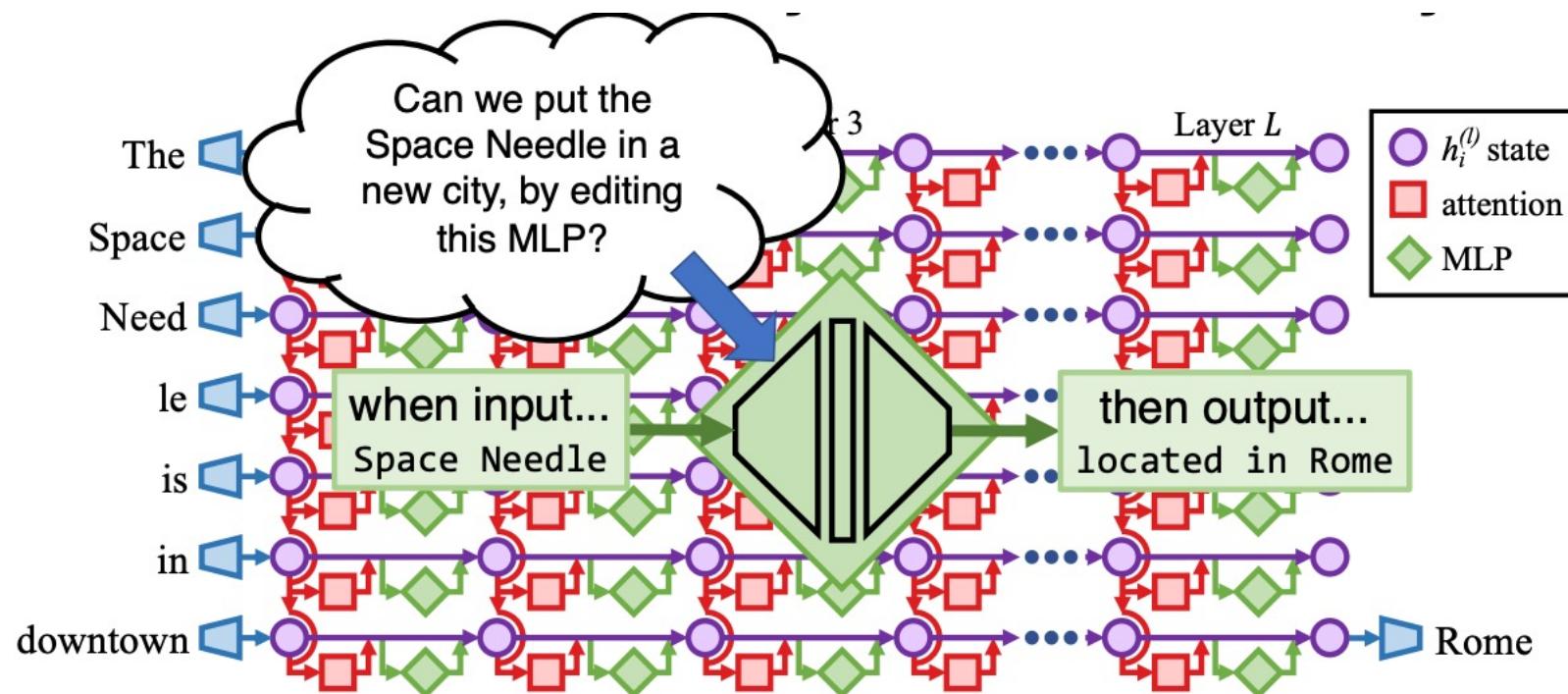
- Example #2: ROME (Meng et al., 2022)
 - Edit knowledge by updating the MLP weights with rank-one update



Key → Value
“Eiffel Tower” → “in Paris”
“Megan Rapinoe” → “plays soccer”
“SQL Server” → “by Microsoft”

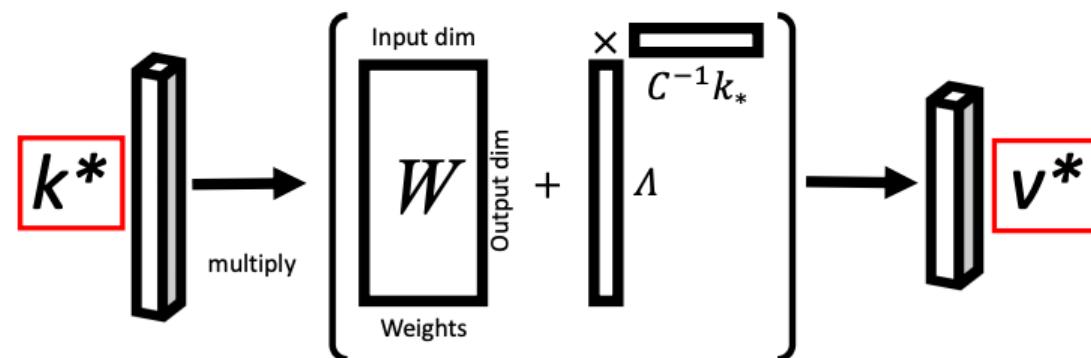
- Hypothesize MLPs can be modeled as a linear associative memory
- Linear operation **W** stores the key-value mapping information.

- Example #2: ROME (Meng et al., 2022)
 - Edit knowledge by updating the MLP weights with rank-one update



- Example #2: ROME (Meng et al., 2022)
 - Edit knowledge by updating the MLP weights with rank-one update

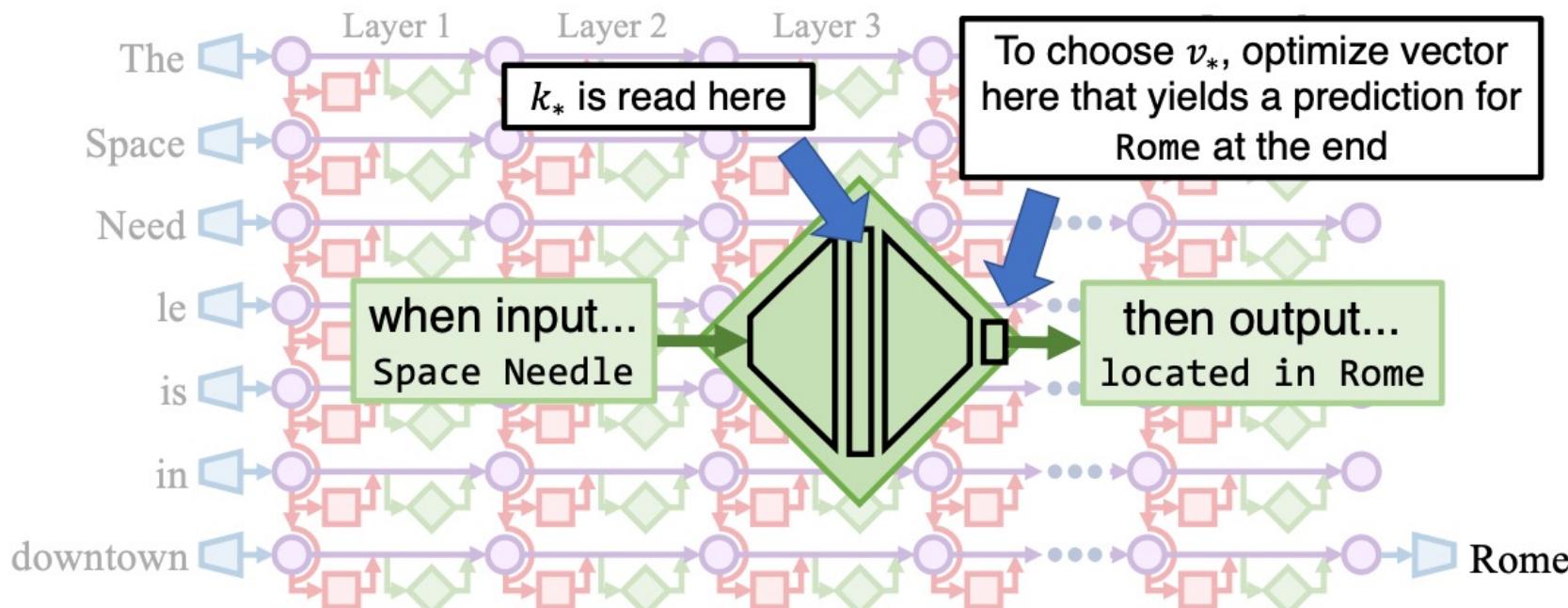
minimize $\|\hat{W}K - V\|$ such that $\hat{W}k_* = v_*$ by setting $\hat{W} = W + \Lambda(C^{-1}k_*)^T$.



- Step 1: Choosing k_* to select the Subject
- Step 2: Choosing v_* to recall the Fact
- Step 3: Inserting the act by updated W

- Example #2: ROME (Meng et al., 2022)
 - Edit knowledge by updating the MLP weights with rank-one update

Computing Λ requires an optimization over v_* .



- Example #3: MEMIT (Meng et al., 2023)

- Edit thousands of facts at once
 - Challenge 1: from single edit to batch edit

$$W_1 [K_0 \ K_1] [K_0 \ K_1]^T = [M_0 \ M_1] [M_0 \ M_1]^T$$

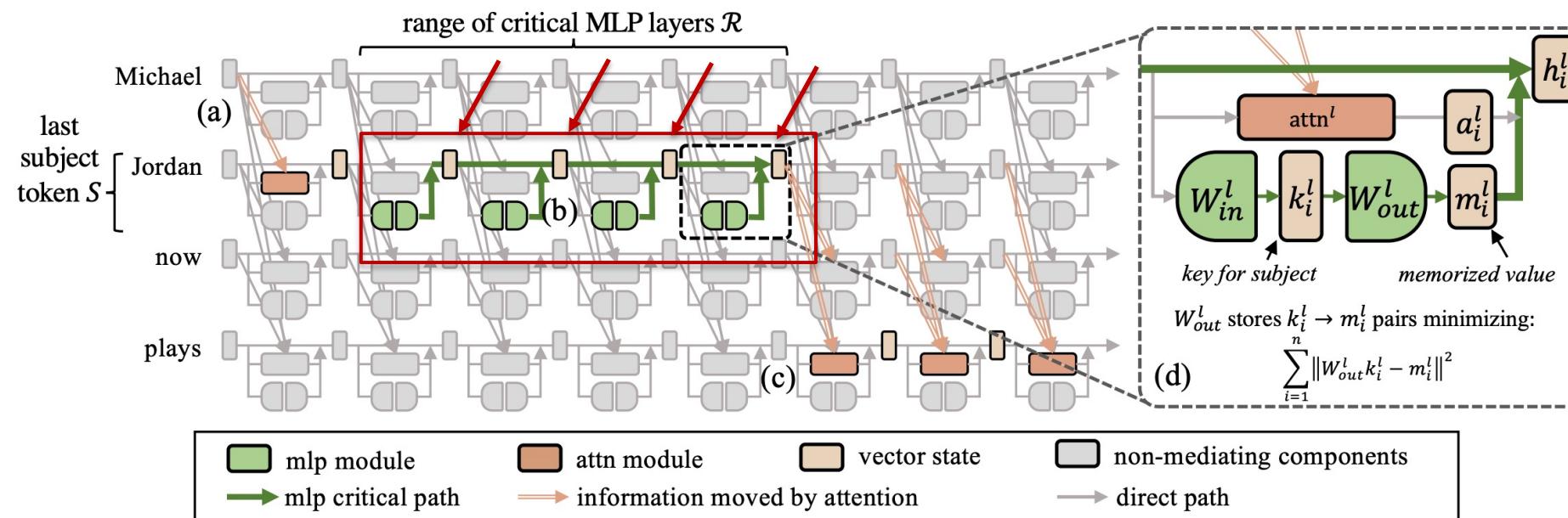
$$(W_0 + \Delta)(K_0 K_0^T + K_1 K_1^T) = M_0 K_0^T + M_1 K_1^T$$

$$\Delta(K_0 K_0^T + K_1 K_1^T) = M_1 K_1^T - W_0 K_1 K_1^T$$

$$\Delta = R \boxed{K_1^T} (C_0 + K_1 K_1^T)^{-1}$$

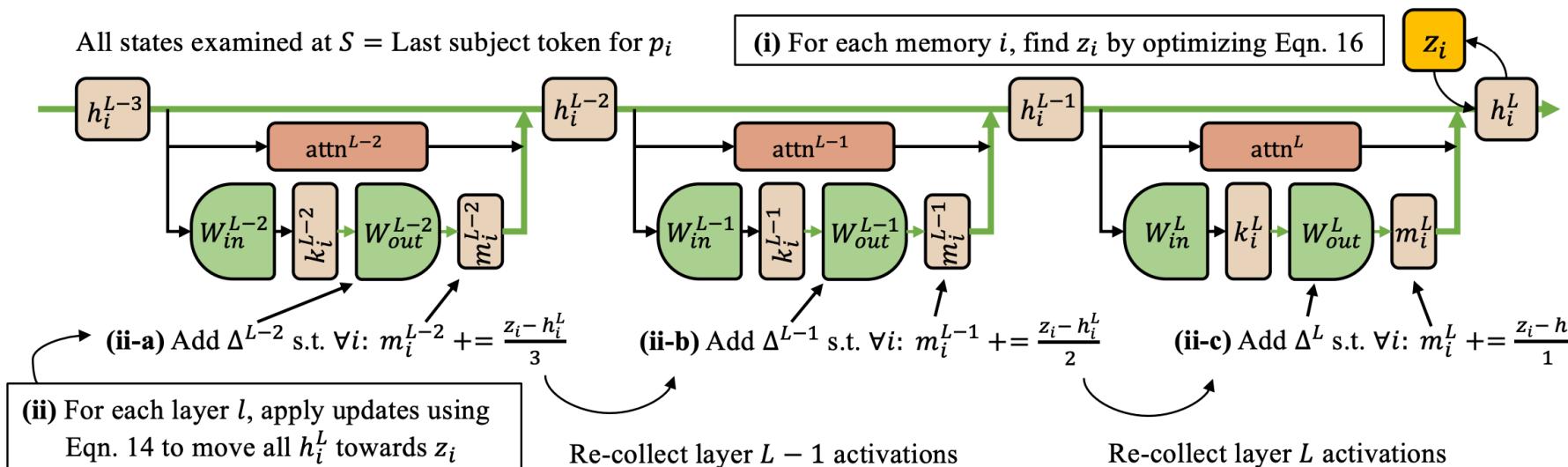
Stack the new and old facts together, then solve for W_0

- Example #3: MEMIT (Meng et al., 2023)
 - Edit thousands of facts at once
 - Challenge 2: expanding to a range of MLPs



- Example #3: MEMIT (Meng et al., 2023)

- Edit thousands of facts at once
- Challenge 2: expanding to a range of MLPs

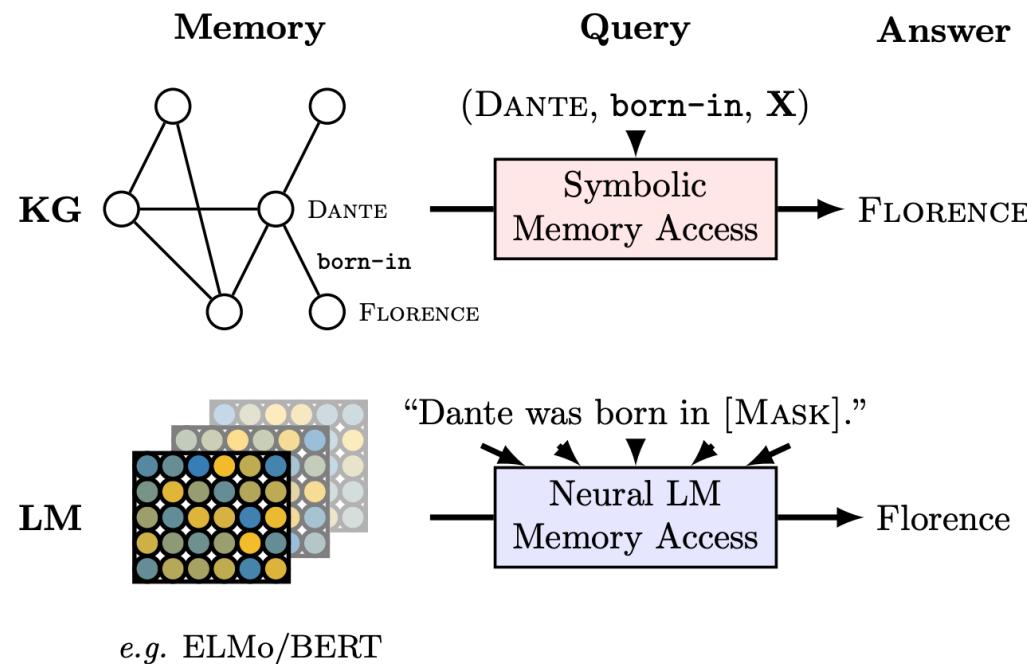


iteratively insert a fraction of the residuals over the range of critical MLP modules

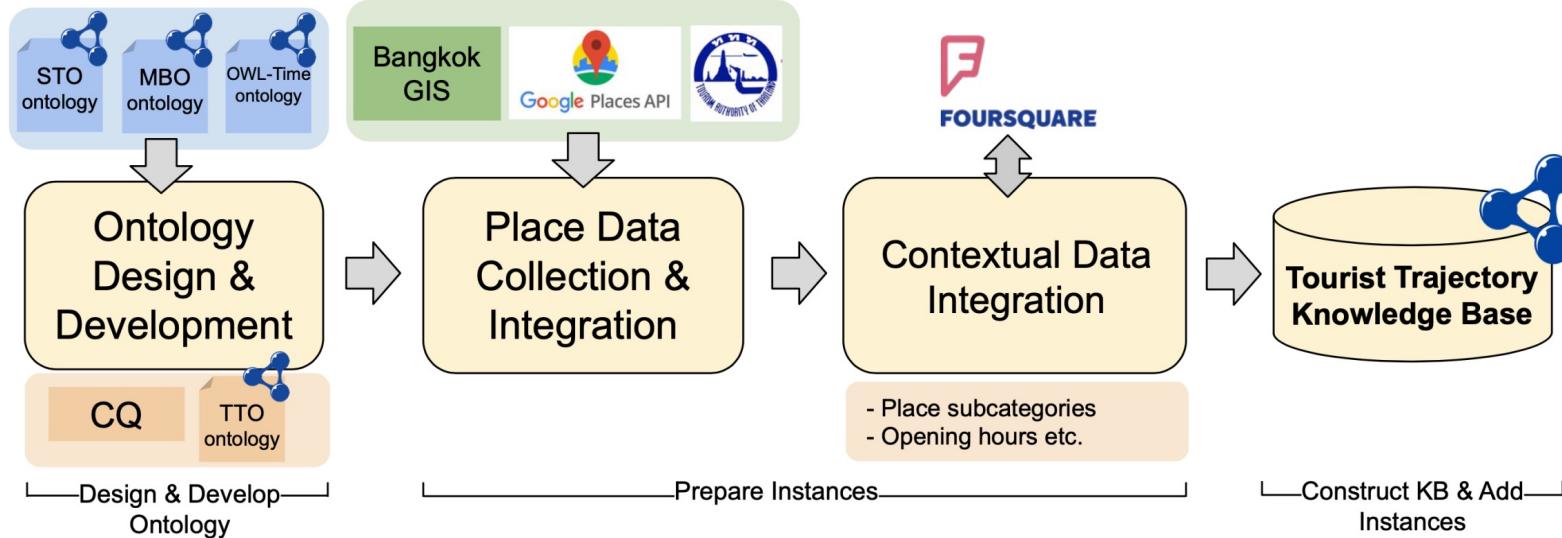
- Broader range of target knowledge types
 - Currently most studies only focus on factual knowledge, which is relatively easy to formalize and evaluate.
- More comprehensive evaluation
 - Benefits for downstream applications such as QA
 - Affects of knowledge editing beyond paraphrases
- More effective editing approaches towards LLMs

Conclusion: Can LLMs serve as KBs?

- Are LLMs potential substitute for structured KB?



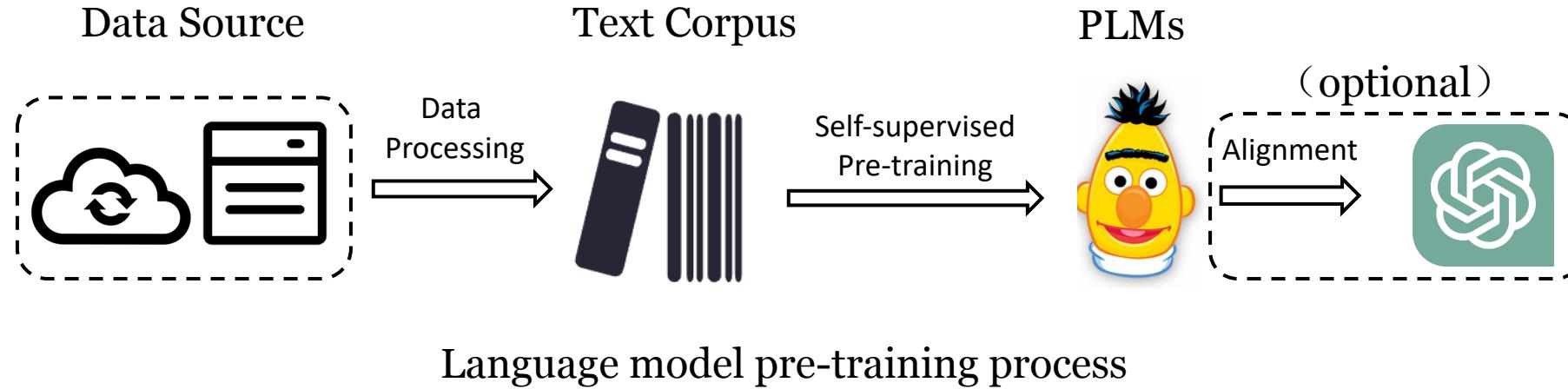
Construct a Structured KB



Structured KB construction process (Krataihong et, al. 2022)

- Requires pre-defined ontology
- Complex pipelines and many traditional NLP techniques involve
- Expert knowledge and human effort for annotation

Construct a LLM-based KB



- Requires no ontology engineering
- End2end self-supervised pre-training without data annotation
- Much less expert knowledge

Construct a LLM-based KB

- Even more simple solution



ChatGPT API



minimax



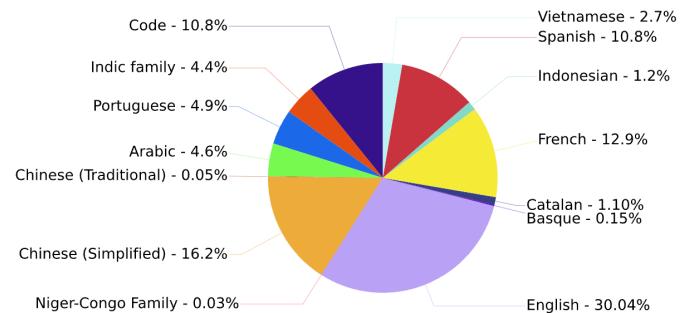
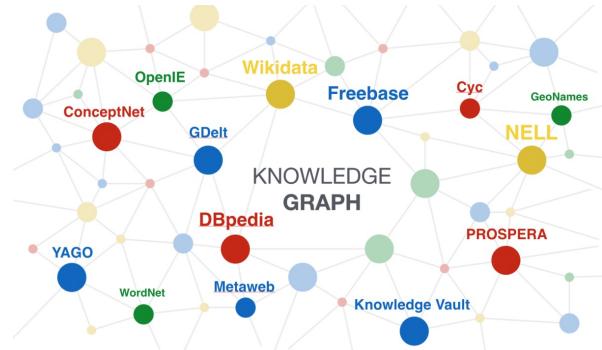
MOSS

.....

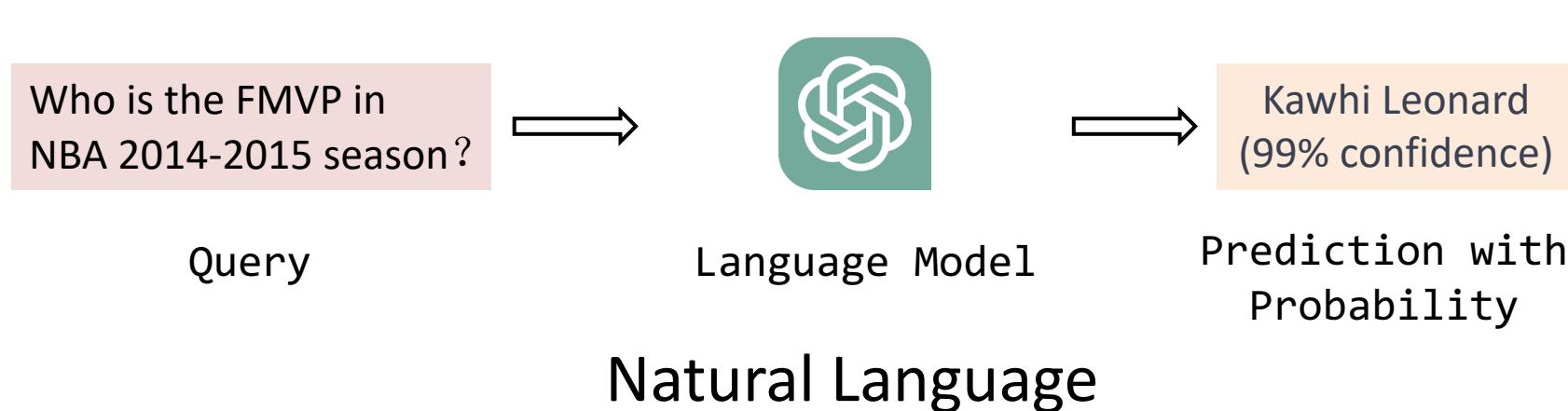
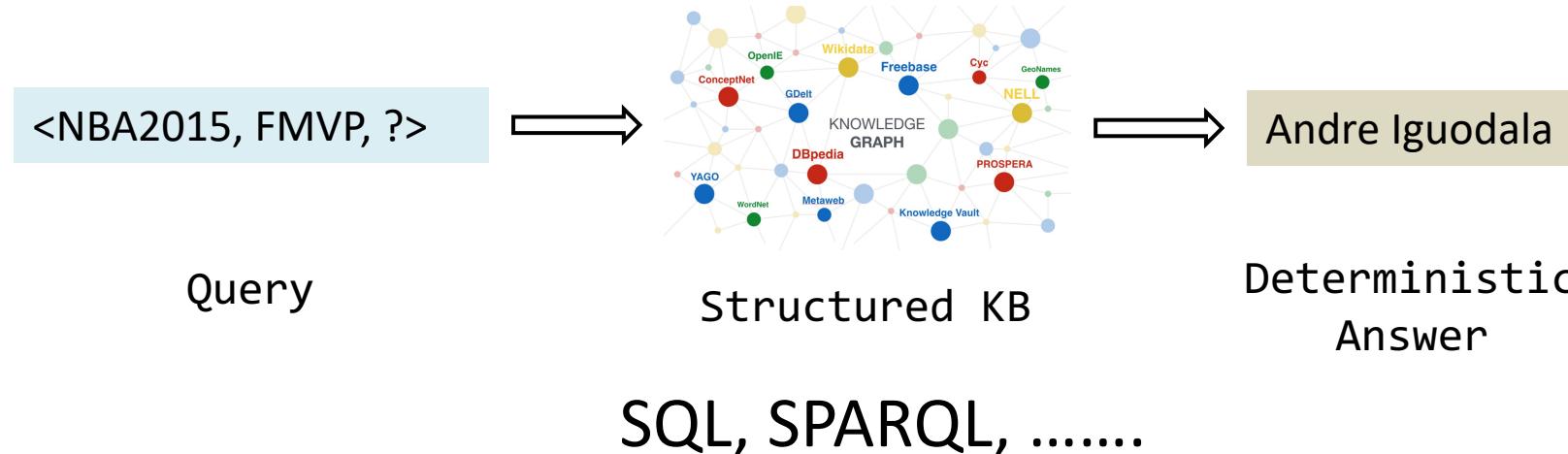


Knowledge Coverage

	Structured KB	LMs-as-KBs
Domain	Constrained	Open
Amount	Limited	Potential?
Schema Expanding	Complex	Easy
Knowledge Fusing	Complex	Easy



Interaction between Human and KBs



SU

林黛玉和林如海是什么关系?

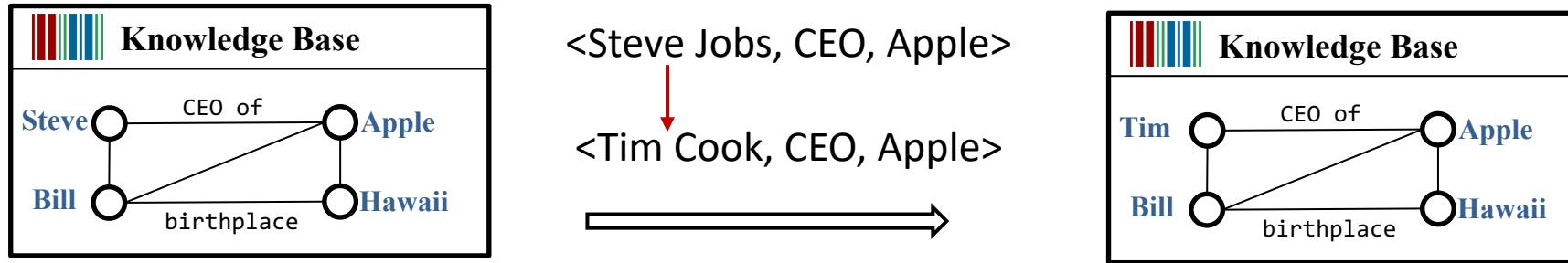


林黛玉和林如海之间没有直接的亲属关系。林黛玉是中国古典小说《红楼梦》中的主要人物之一，她是贾宝玉的表妹，出生在贾府中。而林如海是林府的主人，是林黛玉的父亲。林如海的正式名字是林乃春，他是一个文雅的人，林府是贾府的侄儿之一的私家庄园，与贾府有一定的社交往来。

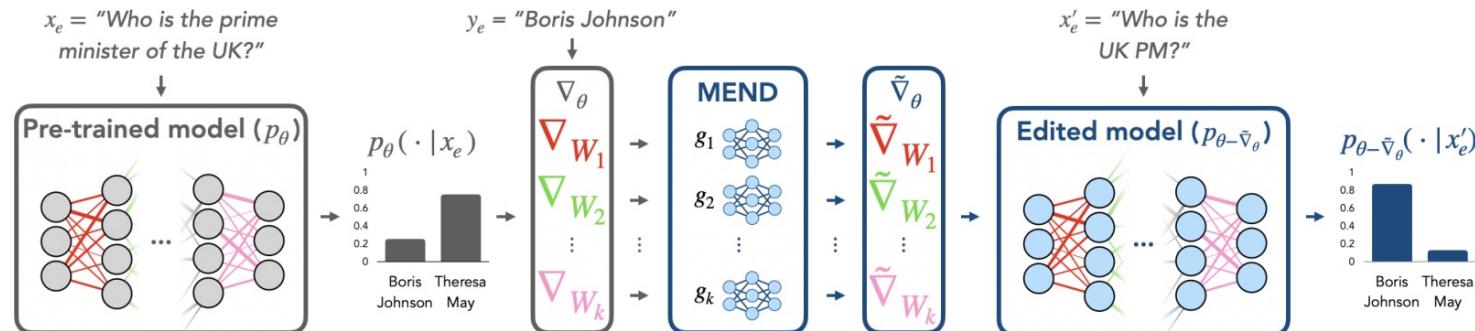
□ □ □

林黛玉的母亲是贾母的女儿王夫人，王夫人嫁给了林如海，所以林黛玉与林如海是父女关系。然而，林黛玉的出生却与一些传说和神秘因素有关，小说中有关她的身世也有许多复杂情节，这使得她的生平更加扑朔迷离。

Timely Editing Difficulties



Editing a Pre-Trained Model with MEND



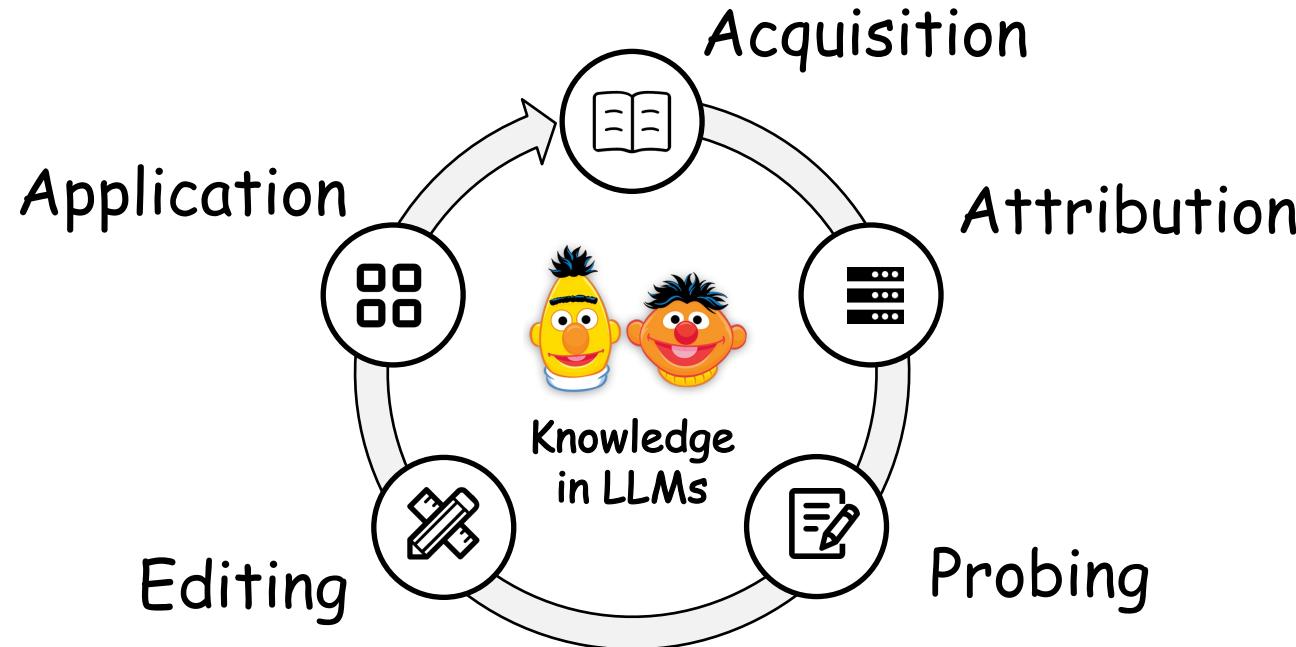
- Compared with LMs, it is easy to add, modify and delete knowledge in structured KBS
- However, editing knowledge in LMs is more complex with potential side effects

Conclusion

- LLMs have their advantages on simple construction process and its potential knowledge coverage
- Certainty and reliability are the main challenges for LLMs

Perspectives	Structured KB	LMs-as-KBs
Construction		
Ontology/Schema Process	Pre-defined Pipeline	Open-ended 😊 End-to-End 😊
Human Effort	Data annotation	Self-supervised 😊
Expert Knowledge	Common	Not required 😊
Coverage		
Domain	Constrained	Open 😊
Amount	Limited	Potential
Knowledge Fusing	Complex	Easy 😊
Interaction		
Query	Structured	Natural Language 😊
Prediction	Deterministic 😊	Probabilistic
Rejection	Yes 😊	Hard
Editing	Easy 😊	Limited
Reliability		
Ambiguity	Low 😊	High
Correctness	Relatively High 😊	Questionable
Current Practicality	Extensive 😊	Limited yet

Conclusion



From models of language to models of knowledge, there still a long way to go

Thanks & Any Question?