AAAI 2025 Tutorial TH17
Time: 2025-02-26 8:30 am-12:30 pm EST
Location: Room 116 | Philadelphia Convention Center

The Lifecycle of Knowledge in Large Language Models: Memorization, Editing, and Beyond

Manling Li
Northwestern

Zoey Sha Li
Amazon

Yuji Zhang
UIUC

Chi Han
UIUC

Heng Ji
UIUC

Northwestern University

amazon

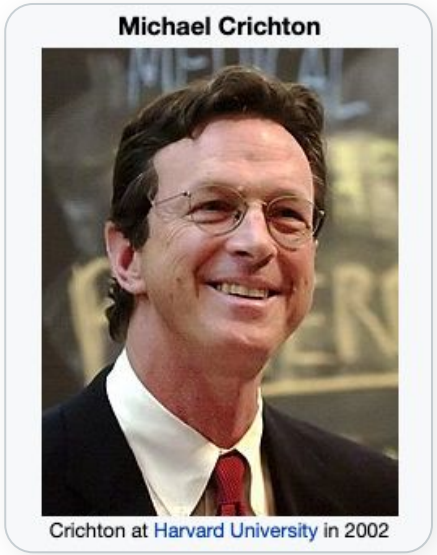
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

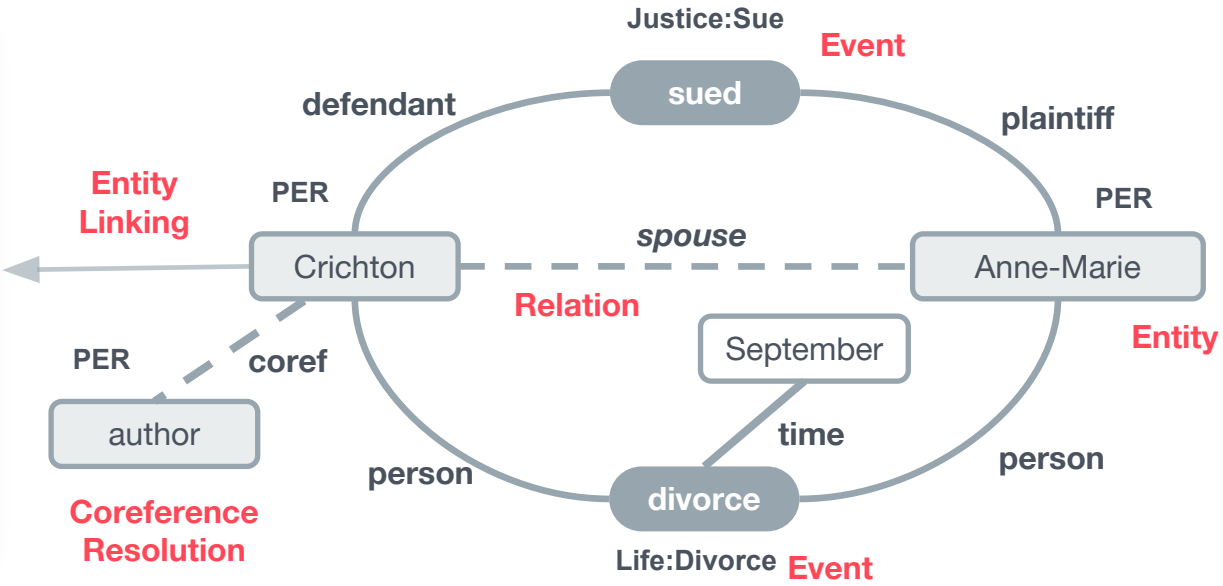
Knowledge

What is Knowledge ?

Knowledge in the Pre-LLM Era: Models as a Tool for Extraction

> ***Anne-Marie*** *sued* ***Crichton****, best known as the* ***author*** *of Jurassic Park, for divorce in September.*





Impact of LLMs on Information Extraction: Zero-shot Performance (2023)

- LLMs have not caught up with SOTA in more complex IE tasks yet, but more and more people are applying LLMs for IE

Task	Dataset	BERT	RoBERTa	SOTA	ChatGPT
Entity Typing(ET)	BBN	80.3	79.8	82.2 (Zuo et al., 2022)	85.6
	OntoNotes 5.0	69.1	68.8	72.1 (Zuo et al., 2022)	73.4
Named Entity Recognition(NER)	CoNLL2003	92.8	92.4	94.6 (Wang et al., 2021)	67.2
	OntoNotes 5.0	89.2	90.9	91.9 (Ye et al., 2022)	51.1
Relation Classification(RC)	TACRED	72.7	74.6	75.6 (Li et al., 2022a)	20.3
	SemEval2010	89.1	89.8	91.3 (Zhao et al., 2021)	42.5
Relation Extraction(RE)	ACE05-R	87.5 | 63.7	88.2 | 65.1	91.1 | 73.0 (Ye et al., 2022)	40.5 | 4.5
	SciERC	65.4 | 43.0	63.6 | 42.0	69.9 | 53.2 (Ye et al., 2022)	25.9 | 5.5
Event Detection(ED)	ACE05-E	71.8	72.9	75.8 (Liu et al., 2022a)	17.1
	ACE05-E+	72.4	72.1	72.8 (Lin et al., 2020)	15.5
Event Argument Extraction(EAE)	ACE05-E	65.3	68.0	73.5 (Hsu et al., 2022)	28.9
	ACE05-E+	64.0	66.5	73.0 (Hsu et al., 2022)	30.9

near SOTA

large gap

Table from Li et al. “Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness”. Arxiv 2023.

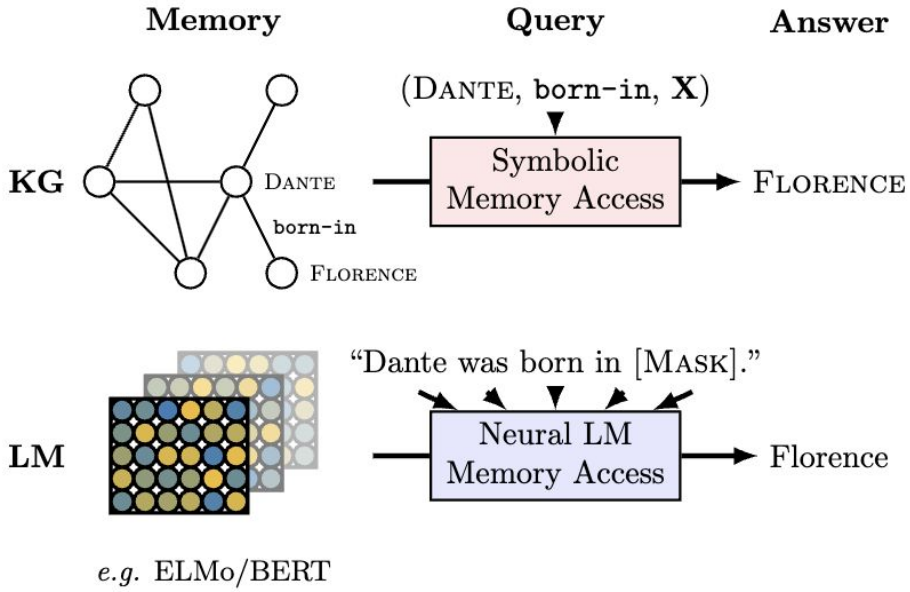
Impact of LLMs on Information Extraction: Few-shot and Supervised Performance (2024)

- A very nice survey [Xu et al., 2024]: https://github.com/quqxui/Awesome-LLM4IE-Papers
- Few-shot (50 examples) performance (e.g., Code4Struct) is comparable to supervised BERT model trained from 500 documents
- Using the same amount of training data, conditioned generation (e.g., BART-gen) performs much better than sup
- [Huang et al., ACL2024Findings]

Model	ACE05			
	TI	TC	AC	AC+
DyGIE++	74.7	71.3	56.0	51.8
OneIE	75.0	71.1	59.9	54.7
AMR-IE	74.6	71.1	60.6	54.6
EEQA	73.8	70.0	55.3	50.4
RCEE	74.0	70.5	55.5	51.0
Query&Extract	68.6	65.1	55.0	49.0
TagPrime	73.2	69.9	59.8	54.6
DEGREE-E2E	70.3	66.8	55.1	49.1
DEGREE-PIPE	72.0	68.4	56.3	50.7

Representative Model	Technique	Uni.	Backbone	Trg-I	Trg-C	Arg-I	Arg-C
Code4Struct [41]	ZS Pr		Code-davinci-002	-	-	50.6	36.0
Code4UIE [6]	ICL	√	GPT-3.5-turbo-16k	-	37.4	-	21.3
Code4Struct [41]	ICL		Code-davinci-002	-	-	62.1	58.5
TANL [33]	SFT	√	T5-base	72.9	68.4	50.1	47.6
Text2Event [131]	SFT		T5-large	-	71.9	-	53.8
BART-Gen [130]	SFT		BART-large	-	-	69.9	66.7
UIE [4]	SFT	√	T5-large	-	73.36	-	54.79
GTEE-DYNPREF [135]	SFT		BART-large	-	72.6	-	55.8
DEEPSTRUCT [151]	SFT	√	GLM-10B	73.5	69.8	59.4	56.2
PAIE [134]	SFT		BART-large	-	-	75.7	72.7
PGAD [137]	SFT		BART-base	-	-	74.1	70.5
QGA-EE [138]	SFT		T5-large	-	-	75.0	72.8
InstructUIE [5]	SFT	√	Flan-T5-11B	-	77.13	-	72.94
GoLLIE [32]	SFT	√	Code-LLaMA-34B	-	71.9	-	68.6
YAYI-UIE [155]	SFT	√	Baichuan2-13B	-	65.0	-	62.71
KnowCoder [160]	SFT	√	LLaMA2-7B	-	74.2	-	70.3
USM† [30]	SFT	√	RoBERTa-large	-	72.41	-	55.83
RexUIE† [197]	SFT	√	DeBERTa-v3-large	-	75.17	-	59.15
Mirror† [198]	SFT	√	DeBERTa-v3-large	-	74.44	-	55.88

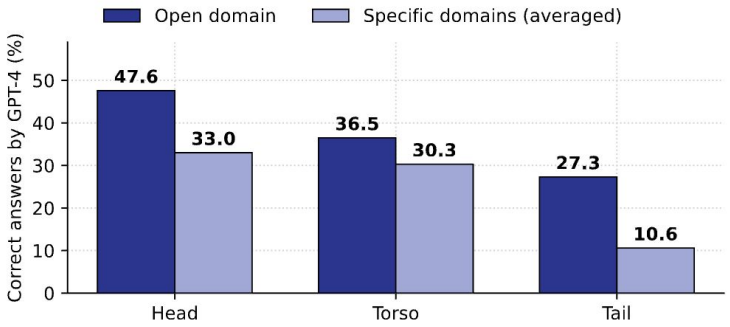
Knowledge in the Post-LLM era: Models as Knowledge Bases



Language models can be probed to output the target entity or relation similar to querying a knowledge graph.

Petroni, Fabio, et al. "Language models as knowledge bases?." EMNLP (2019).

Models Struggle with Torso and Tail Entities



Example questions where GPT-4 gives incorrect answers

Movie

Question: What profession does Tj Singh (known for John Carter (2012)) have?
Ground Truth: Visual effects
GPT-4: Actor

Book

Question: Who authored Choke (published in 1996)?
Ground Truth: Stuart Woods
GPT-4: Chuck Palahniuk

Will a larger, stronger LM fix this or it this a fundamental weakness?

Sun, Kai, et al. "Head-to-tail: how knowledgeable are large language models (LLMs)? AKA will LLMs replace knowledge graphs?." arXiv preprint arXiv:2308.10168 (2023).

LLMs are Often Out-of-Date

HE

What is the status of efforts to secure the release of Paul Whelan and Brittney Griner?

I'm sorry, but as a language model AI, I do not have the most up-to-date information on the current status of individual cases like Paul Whelan and Brittney Griner. However, as of my knowledge cutoff in 2021, Paul Whelan, a former US Marine, was being held in Russia on espionage charges and has been advocating for his release through his legal team and government representatives. Similarly, Brittney Griner is a basketball player who was playing in Russia and has been involved in legal issues there, but I don't have any more specific details on her case. It's best to check reliable news sources for the most recent updates on their cases.

LLMs Hallucinate Facts

HE

Why is the Zaporizhzhia nuclear power plant a target for Russia?

The Zaporizhzhia Nuclear Power Plant is not a specific target for Russia. It is important to note that there is no evidence to suggest that Russia has any intention of targeting the plant. Nuclear power plants are generally considered critical infrastructure and are protected by international laws and regulations. Any deliberate attack on a nuclear power plant could have severe consequences for the environment and human health.

Hallucinations could be Harmful

HE

Why does Imatinib effectively cross the blood brain barrier?

Imatinib, also known by its trade name Gleevec or Glivec, is a medication primarily used to treat certain types of cancer, such as chronic myelogenous leukemia (CML) and gastrointestinal stromal tumors (GIST). While Imatinib is not specifically designed to cross the blood-brain barrier (BBB), it has been found to have some ability to penetrate the central nervous system (CNS). The exact mechanism by which Imatinib crosses the BBB is not fully understood, but several factors likely contribute to its ability to reach the brain effectively.

HE

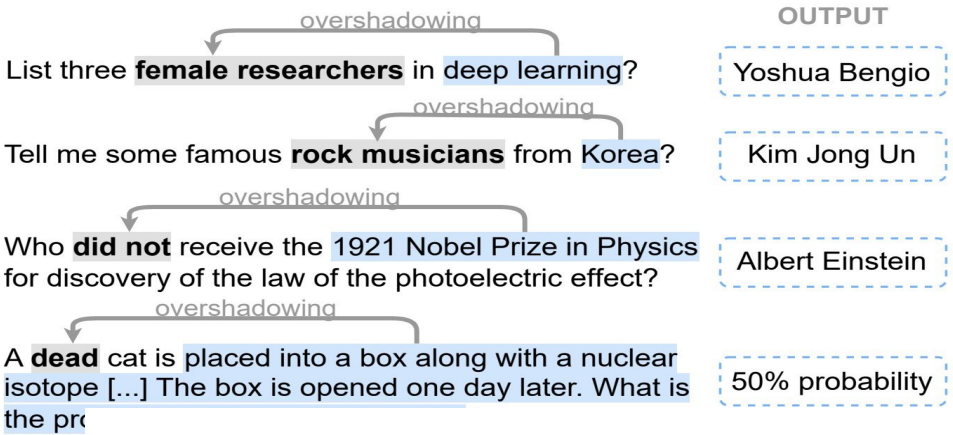
Why doesn't Imatinib effectively cross the blood brain barrier?

I apologize for the incorrect information in my previous response. Imatinib does not effectively cross the blood-brain barrier (BBB). In fact, the penetration of Imatinib into the central nervous system (CNS) is generally limited.

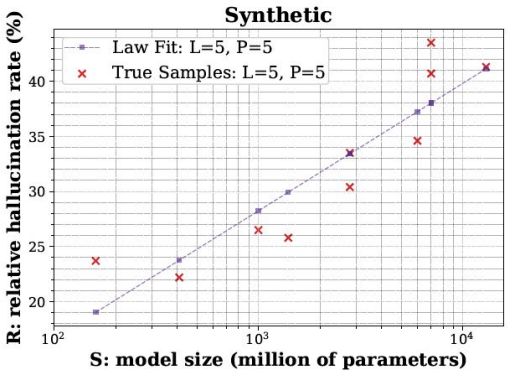
The main reason for the limited penetration of Imatinib across the BBB is its physicochemical properties. Imatinib is a relatively large molecule with a molecular weight of around 589 Daltons, which makes it more challenging for it to passively diffuse through the tight junctions of the BBB.

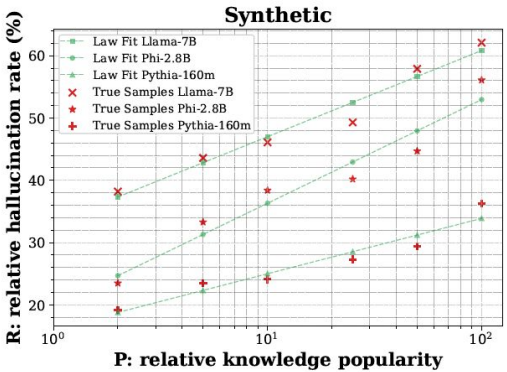
LLMs are Biased (by Nature) [Zhang et al., 2025]

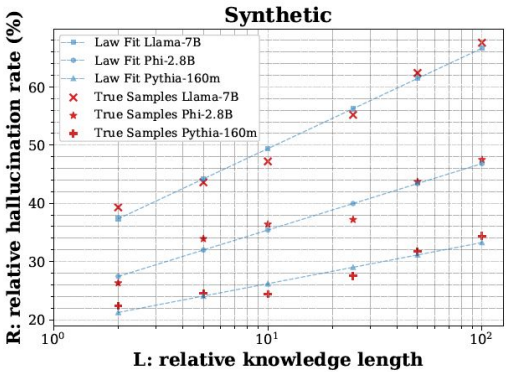
- Knowledge overshadowing→ over-generalization



- They are often related to dynamic events
  - *Time-event relation*: When did this event happen?
  - *Location-event relation*: Where did this event happened?
  - Gender bias: What's the gender of character?
  - *Negation curse*: Who was not known for relative theory?

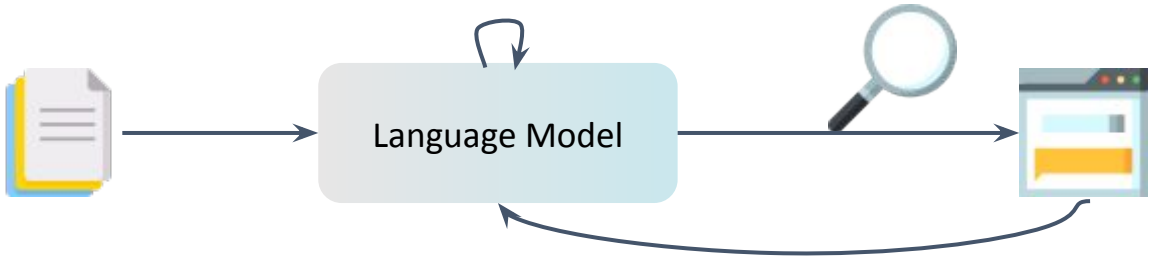






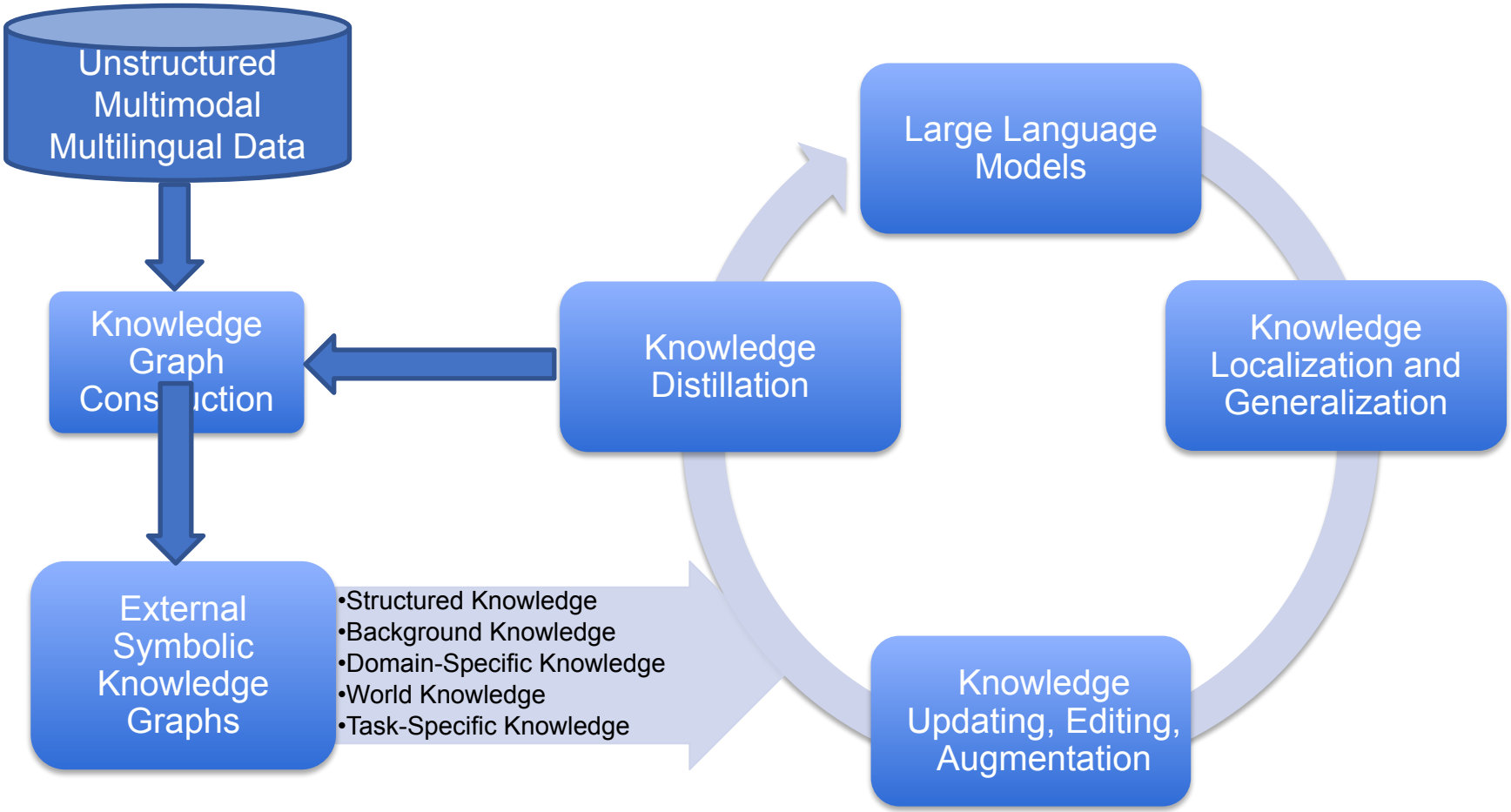
Can Language Models be Flexible Knowledge Composers?

- Can language models be efficiently updated?
- Can language models integrate their own knowledge and external sources?
- Can language models determine when to ask for help?

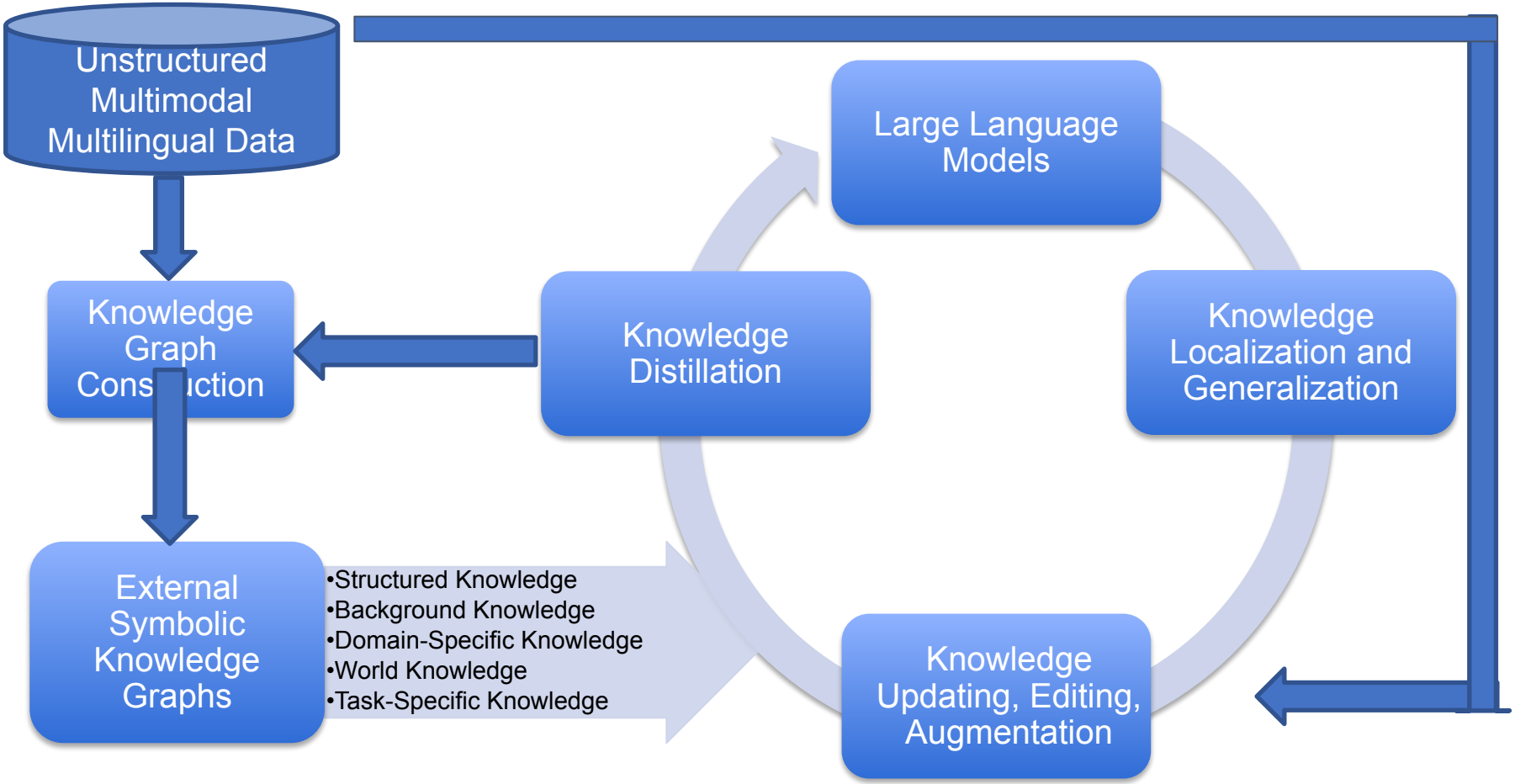


The Ideal state: a self-aware, self-updating knowledge system

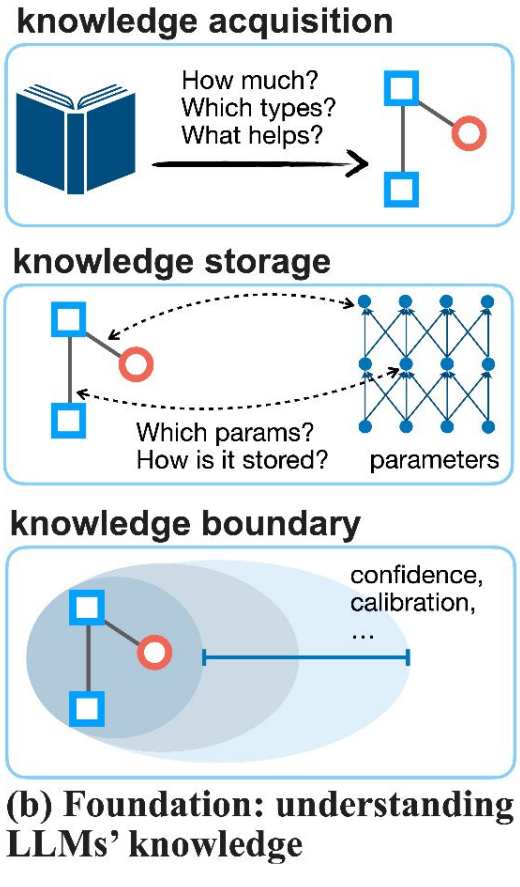
Completing Knowledge Lifecycle by connecting LLMs and External Knowledge



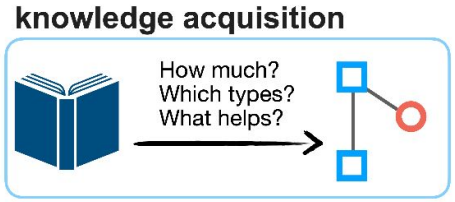
Completing Knowledge Lifecycle by connecting LLMs and External Knowledge



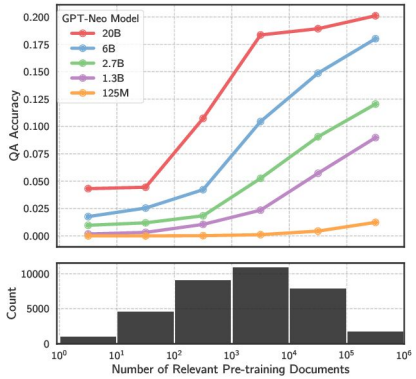
Tutorial Roadmap



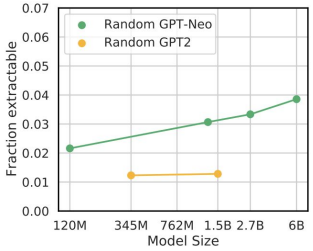
Section 1: Knowledge Acquisition

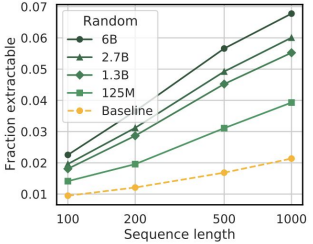


Memorization is strongly correlated with model performance on knowledge-intensive tasks

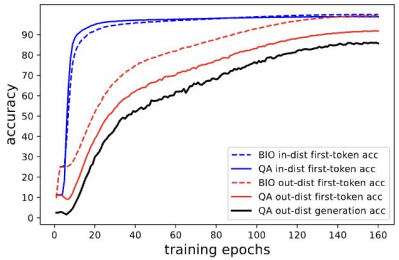


Memorization is affected by model size, knowledge frequency and prefix length. Larger models memorize more and faster.

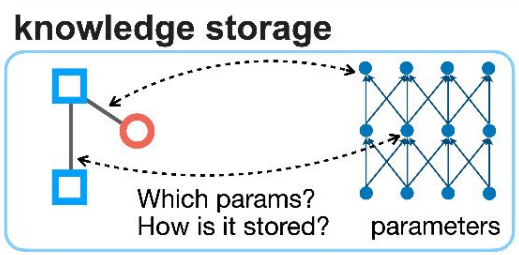




Knowledge acquisition requires diversity of knowledge expression and tasks.



Section 2: Knowledge Storage



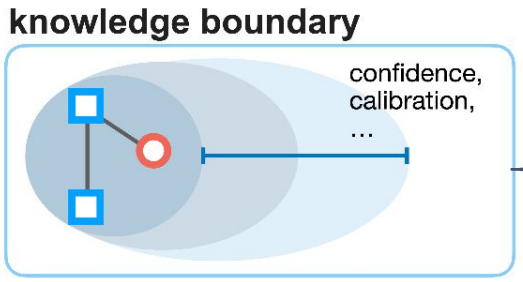
Hypothesis 1: Knowledge is stored in feed-forward layers which act as key-value memories

Hypothesis 2: Attention also helps inform knowledge lookup with context information

Issue with hypotheses: don't have direct evidence to prove any of them.

Knowledge is stored messily within a language model leading to negative curse, over-ripple etc.

Section 3: Knowledge Boundary



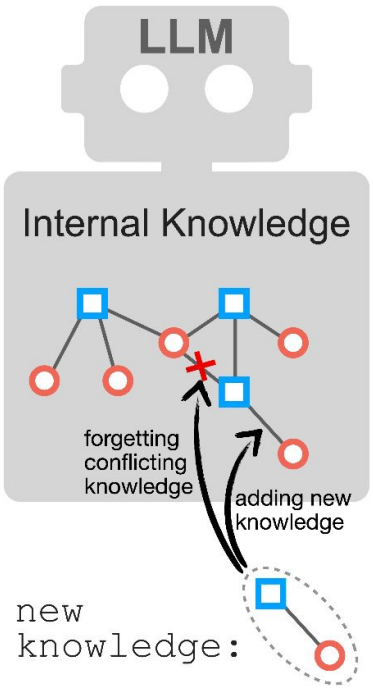
Knowledge boundaries of language models are probabilistic (unlike KGs)

Language models are not inherently well-calibrated

We can teach language models to refuse questions outside their knowledge boundary & express their uncertainty

Fine-tuning LMs with unfamiliar knowledge will hurt their self-awareness of the knowledge boundary

Section 4: Knowledge Editing Approaches & Challenges



Locate-then-edit methods

Fine-tuning methods

In-context learning methods

Challenges of editing locality and generality

Section 5: Knowledge Editing Beyond Triples

More general editing objective

Editing natural texts or events: “ … …”

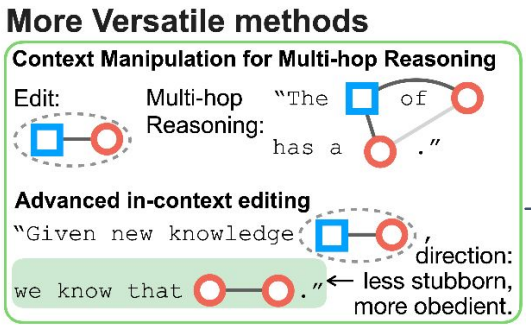
Can existing knowledge editing methods scale up?

Long context LLMs for in-context editing

RAG vs Fine-tuning: naive fine-tuning loses to RAG in terms of efficiency and performance

Making fine-tuning work with data augmentation and mixing

Section 6: Reasoning with Knowledge

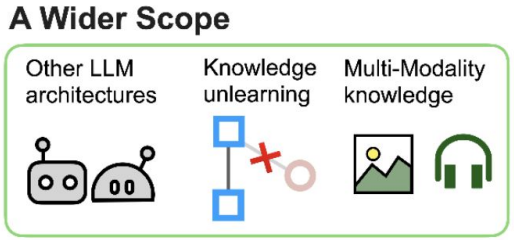


Does knowing mean being able to reason? Even if a LM can recall a fact, it often fails to perform reserve reasoning and multi-hop reasoning.

Frequent knowledge can overshadow infrequent knowledge, leading to wrong conclusions

Providing high quality knowledge traces can help the model learn to reason efficiently.

Section 7: Knowledge Unlearning



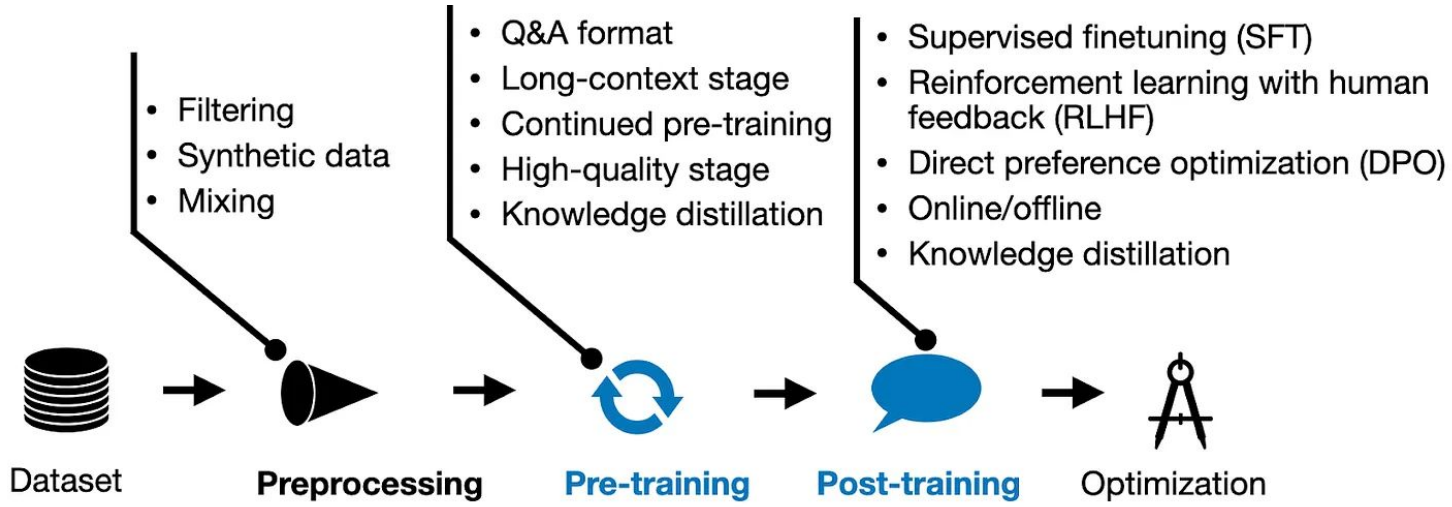
What does it mean to unlearn knowledge?

What if we cannot get the deletion data for this problem?

Typical methods to unlearn a knowledge can be parameter optimization, or parameter merge, or in-context learning

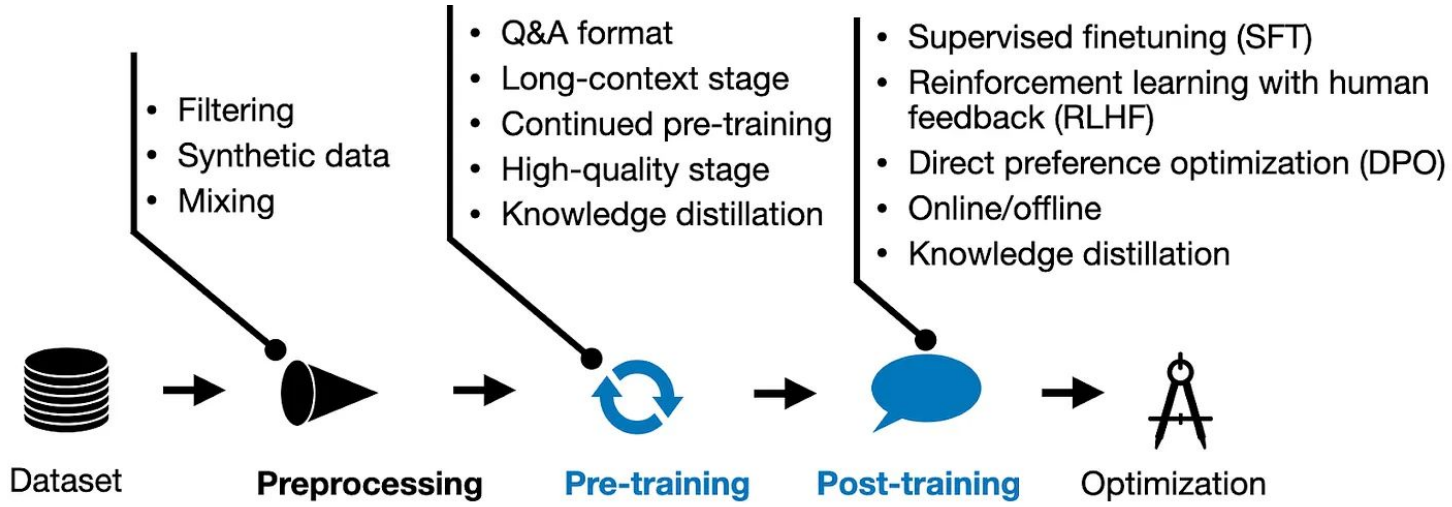
What Exactly Happens During Knowledge Unlearning?

Working with Frontier LLMs



- Knowledge is (mostly) acquired during pretraining
- Post-training for format and style alignment
- RAG systems and Agent systems to keep knowledge up-to-date

Working with Frontier LLMs

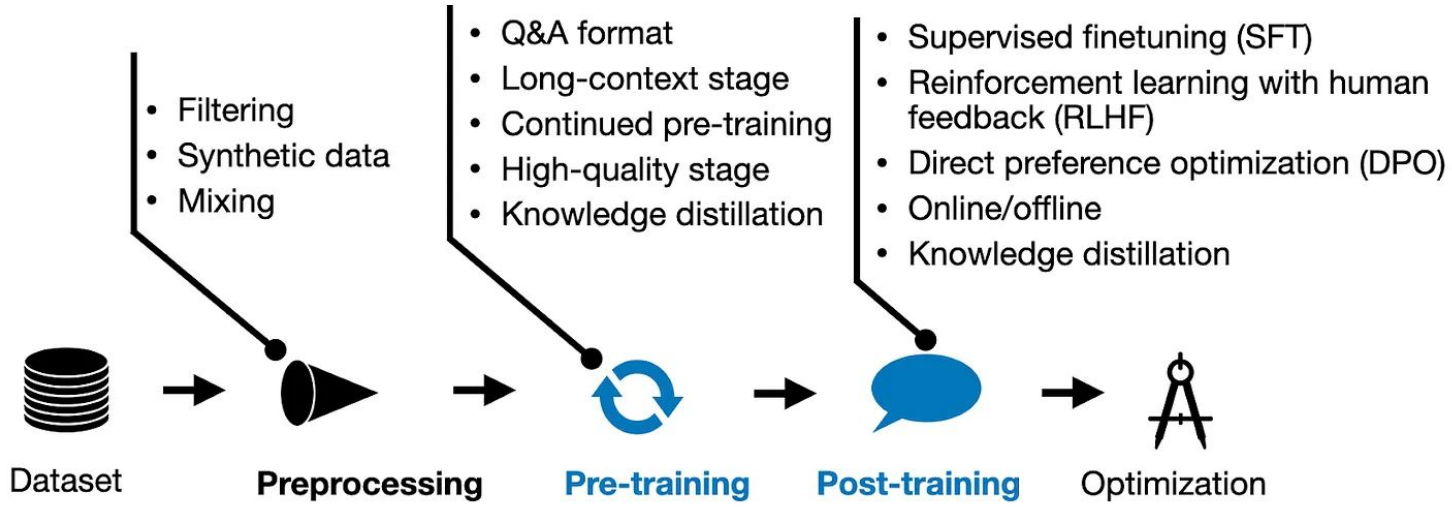


Can knowledge be injected later in the pipeline?

How much knowledge is learned?

- Knowledge is (mostly) acquired during pretraining
- Post-training for format and style alignment
- RAG systems and Agent systems to keep knowledge up-to-date

Working with Frontier LLMs



- Knowledge is (mostly) acquired during pretraining
- Post-training for format and style alignment
- RAG systems and Agent systems to keep knowledge up-to-date

When do we use RAG vs fine-tuning vs continual pretrain a new model?

Timetable

Time (EST)	Session	Speaker	Duration
8:30 - 10:30 Lifecycle of Knowledge in LLMs			
8:30 - 8:45	Motivation and Overview	Heng, Manling	15 min
8:45 - 9:20	Knowledge Acquisition and Memorization	Zoey	35 min
9:20 - 9:50	Knowledge Storage	Chi	30 min
9:50 - 10:10	Knowledge Boundary	Yuji	20 min
10:10 - 10:30	Knowledge Editing	Yuji	20 min
10:30 - 11:00	Coffee Break		30 min
11:00 - 12:30 Future Directions			
11:00 - 11:25	Knowledge Updating (Beyond Triplet Form)	Zoey	25 min
11:25 - 11:50	Reasoning with Knowledge	Yuji	25 min
11:50 - 12:05	Knowledge Unlearning	Manling	15 min
12:05 - 12:15	Knowledge in VLMs	Manling	10 min
12:15 - 12:30	Conclusion & QA	-	10 min

AAAI 2025 Tutorial TH17
Time: 2025-02-26 8:30 am-12:30 pm EST
Location: room 116 | Philadelphia Convention Center

Knowledge Acquisition & Memorization

Zoey Li

Northwestern University

amazon

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

How do language models acquire knowledge during training?

Factual Knowledge

- The capital of France is … Paris.

Templated knowledge triples

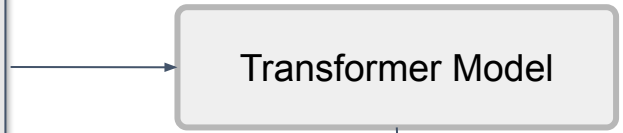
- The sky is blue because … blue light is scattered more than other lights in the Earth's atmosphere.

… of a process called Rayleigh scattering, which is caused by the scattering of sunlight by air molecules

Different expressions of the same knowledge

What happens during language model training?

Understanding the atmosphere and light
To understand why the sky is blue, we need to understand a little about our atmosphere and light. …While all colors are scattered by air molecules, violet and blue are scattered most. The sky looks blue, not violet, because our eyes are more sensitive to blue light (and the sun also emits more energy as blue light than as violet). This process of scattering is known as Rayleigh scattering (named after Lord John Rayleigh, who first described it in the 1870's).




```
Our eyes are more sensitive
to blue light
```

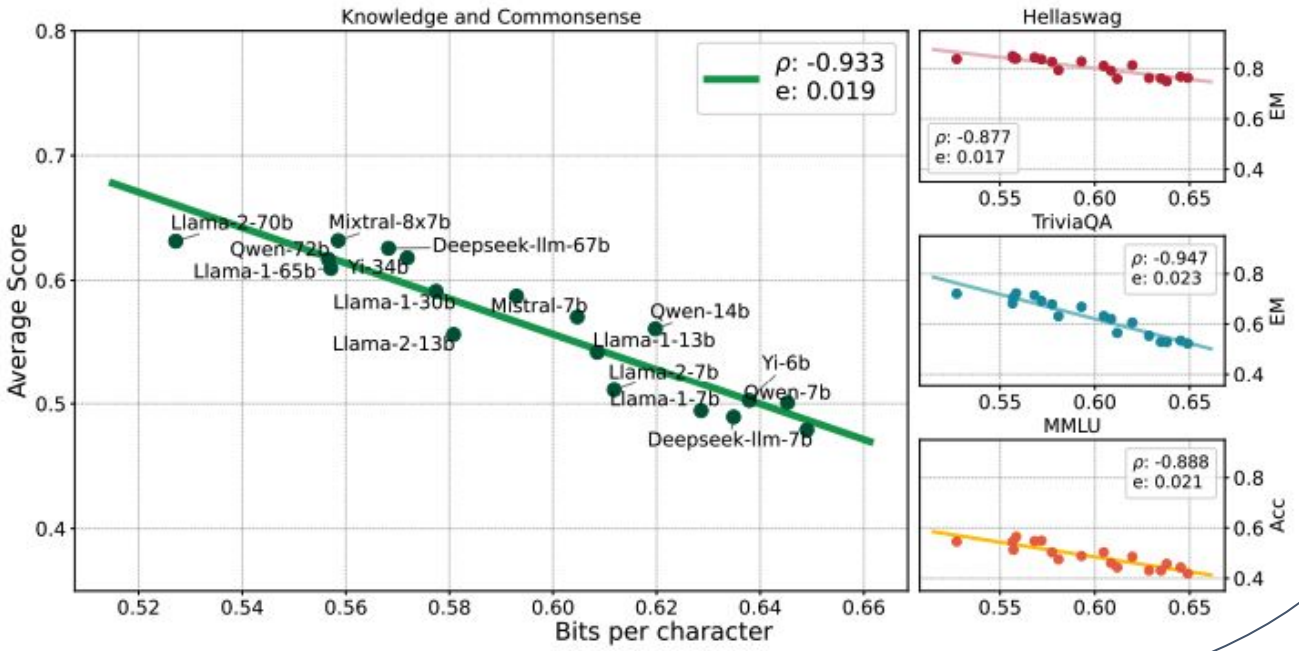
$$L = \sum_{i=1}^{n} \log p(x_i | x_{<i})$$

We optimize for word-by-word reproduction of the training docs.

Article from https://www.weather.gov/fgz/SkyBlue

Language Modeling Ability is Linearly Related to Task Performance

(After model emergence)



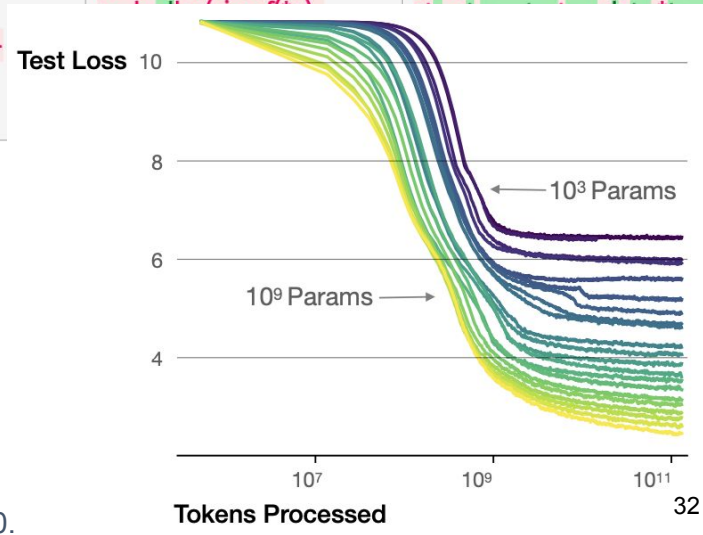
$$BPC = \frac{1}{T}\sum_{i=1}^{N} \log_2 P(x_i|x_{<i})$$

The perplexity a model achieves on CommonCrawl is linearly related to the model's performance on knowledge-intensive benchmarks such as TriviaQA and MMLU.

Is Memorization all we need?

Prompt	Continuation (== 6B)	2.7B	1.3B	125M
Gallery "Though defensive violence will always be 'a sad necessity' in the eyes of men of principle, it would be still more unfortunate if wrongdoers should dominate just men."- St. Augustine "A new idea is first	condemned as ridiculous, and then dismissed as trivial, until finally it becomes what everybody knows." - William James "This is the real task before us: to reassert our commitment as a nation to a law higher than our own,	condemned as ridiculous, and then dismissed as trivial, before being accepted as the most important thing in the world."- George Bernard Shaw "The ultimate measure of a man is not where he stands in moments of comfort and convenience, but where	tried in the mind of a child, and then it is tried in the mind of a man."- Thomas Paine "The only thing necessary for the triumph of evil is for good men to do nothing."- Edmund Burke "	and foremost a man's duty to his fellow man. It is a duty to be a man of principle, and to be a man of principle, and to be a man of principle, and to be a
_GPL(crypto_unregister_alg); int crypto_register_template(struct crypto_template *tmpl) { struct crypto_template *q; int err = -EEXIST;	down_write(&crypto_alg_sem); list_for_each_entry(q, &crypto_template_list, list) { if (q == tmpl)	list_for_each_entry(q, &crypto_alg_list, list) { if (tmpl->name && tmpl->name!= q->alg.cra_name)		

- (Under the same model architecture) the larger the model size, the lower the training loss
- The larger the model size, the more the model memorizes



[1] Right figure from Kaplan et al. Scaling Laws for Neural Language Models. ArXiv 2020.

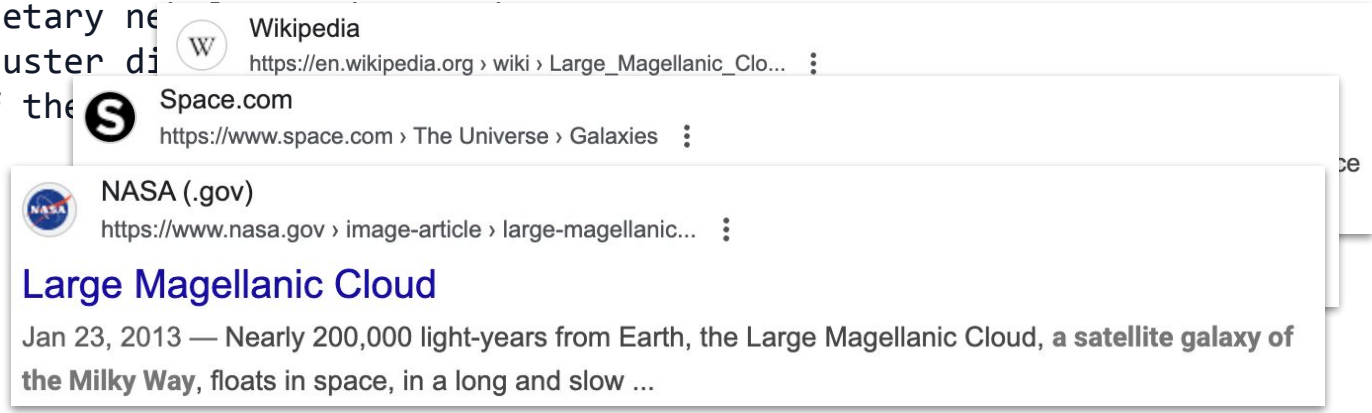
Hypothesis: Factual Knowledge comes from Memorization

Example from MMLU:

Prefix appears 45 times in RedPajama pretraining dataset

The Large Magellanic Cloud is ...

(A) A dwarf galaxy orbiting the Milky Way.
(B) The closest planetary ne
(C) A bright star cluster di
(D) The outer arm of the
Magellan.

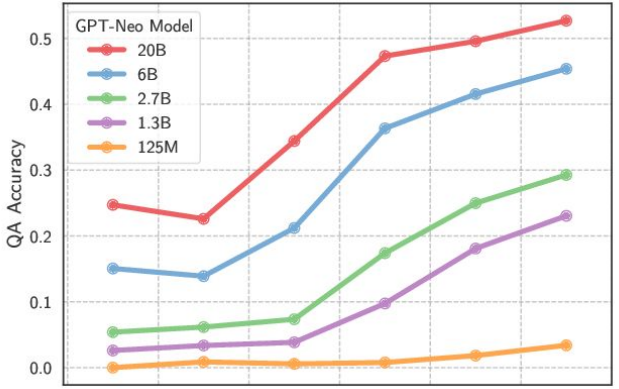


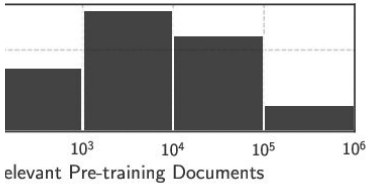
[1] Statistics obtained by Liu et al, Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. https://huggingface.co/spaces/liujch1998/infini-gram

Entity Frequency vs Task Performance

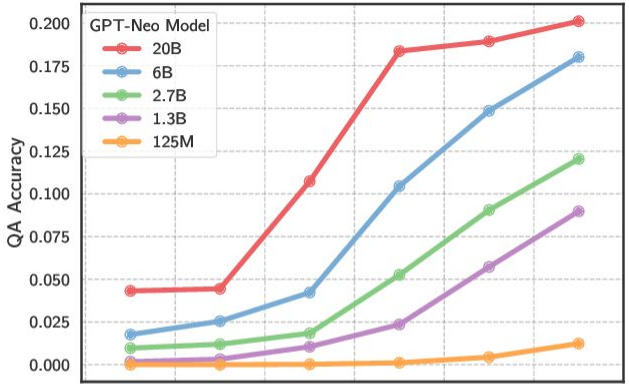
The more popular an entity is in the training dataset, the more likely the model will correctly answer questions about the entity.

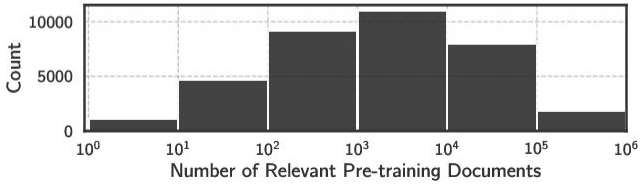
TriviaQA Dataset

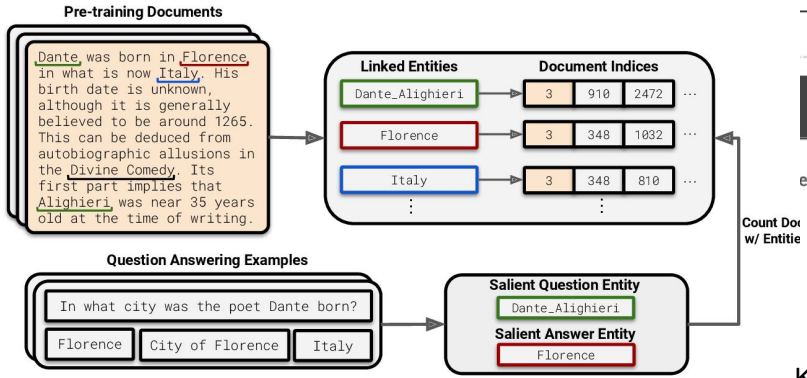




Natural Question Dataset





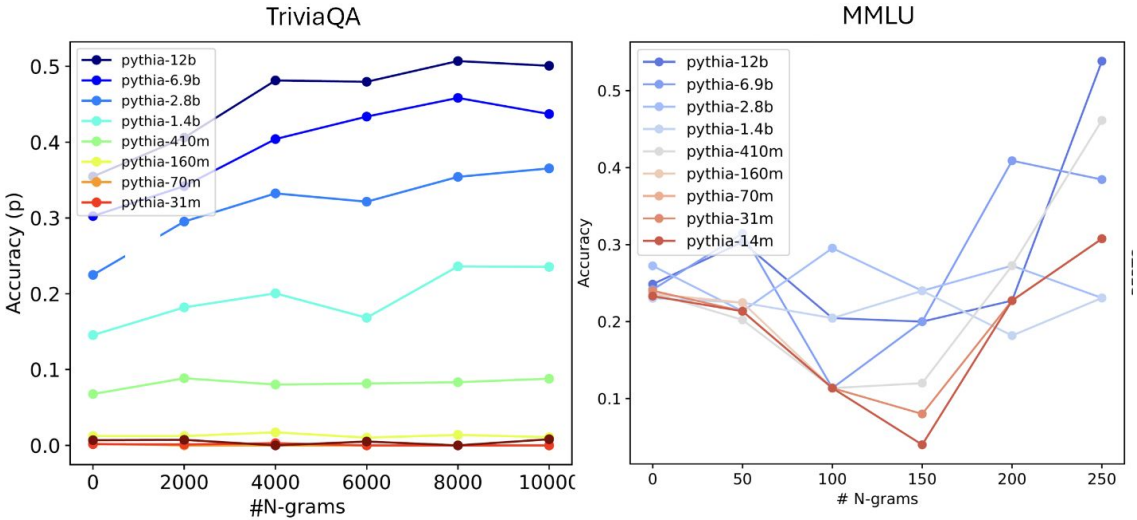


Kandpal, Nikhil et al. “Large Language Models Struggle to Learn Long-Tail Knowledge.” International Conference on Machine Learning (2022).

Task n-gram frequency vs task performance

If the hypothesis that knowledge comes from memorization holds, then the more frequent the task-related knowledge appears in training, the better the task performance should be.

Define task-related knowledge with task n-grams = n-gram pairs from task input and output



The Large Magellanic Cloud is …

(A) A dwarf galaxy orbiting the Milky Way.

(large magellanic cloud, dwarf galaxy)

Results from Wang, Xinyi, et al. "Generalization vs Memorization: Tracing Language Models' Capabilities Back to Pretraining Data." arXiv preprint arXiv:2407.14985 (2024).

Quantifying Memorization in Language Models

Memorization: If the original string can be reproduced using greedy decoding and prompting with a k-length prefix, then the string is k-extractable.

Benign Memorization: improves factual knowledge

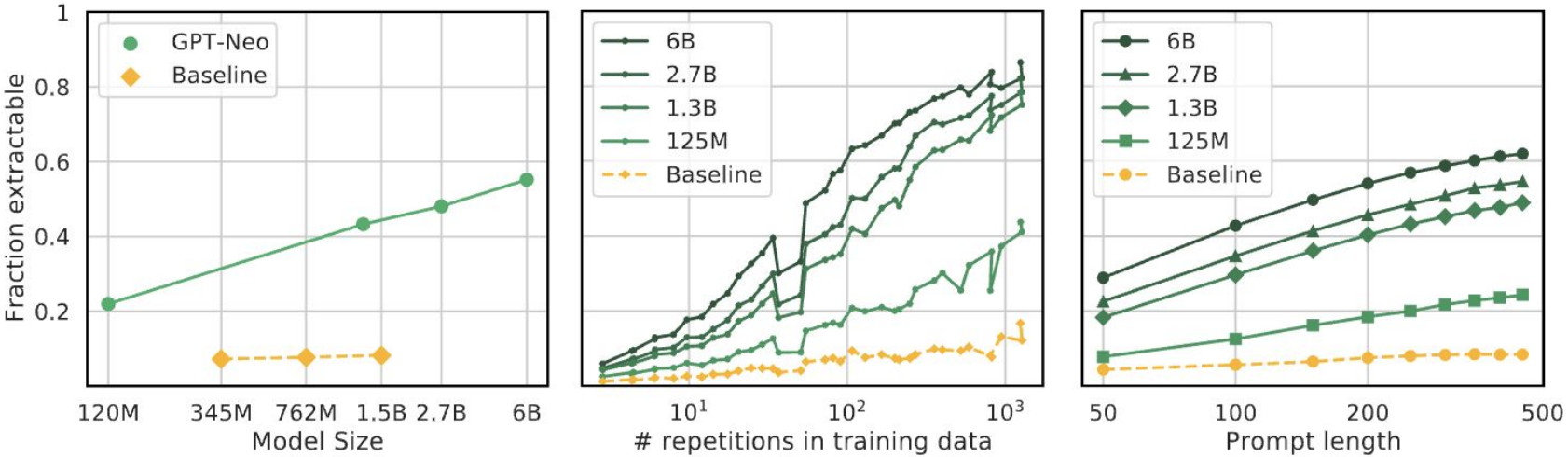
Harmful Memorization: reproduces PII or copyrighted information → will cover this in Knowledge Unlearning

Questions:

- How does different models and data affect memorization?
- How does memorization change over the course of training?

What affects Memorization (after sufficient training)?

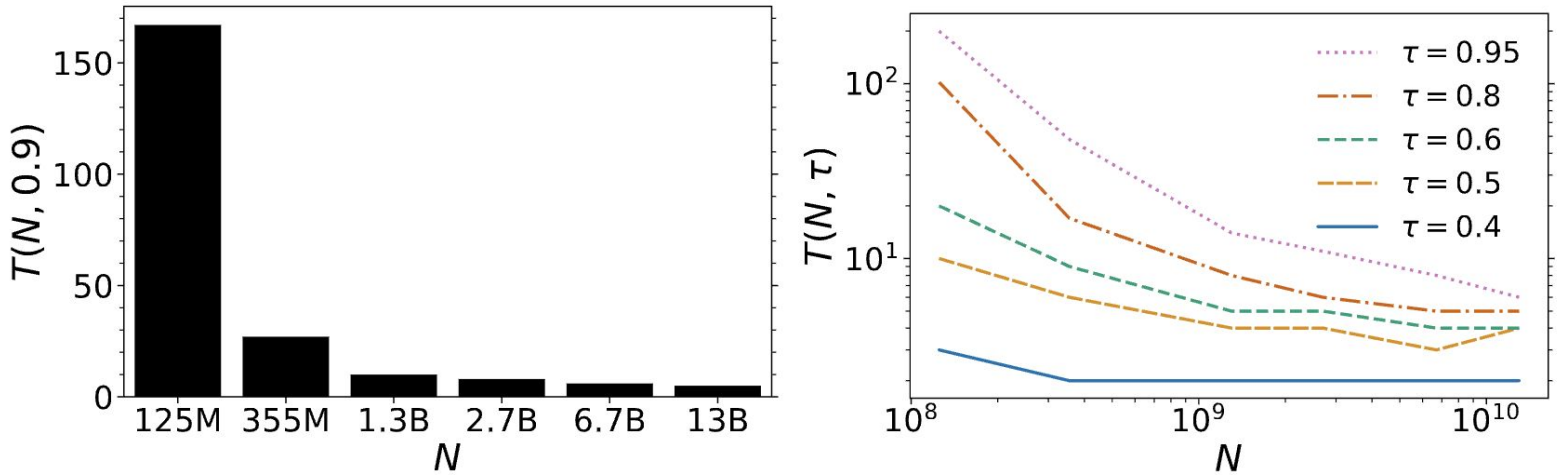
- The larger the model, the larger fraction of training samples memorized.
- The more the repetition of samples, the larger fraction memorized.
- The longer the length of the given context, the larger fraction memorized.



Carlini, Nicholas, et al. "Quantifying memorization across neural language models." The Eleventh International Conference on Learning Representations. 2022.

Training Dynamics of Model Memorization

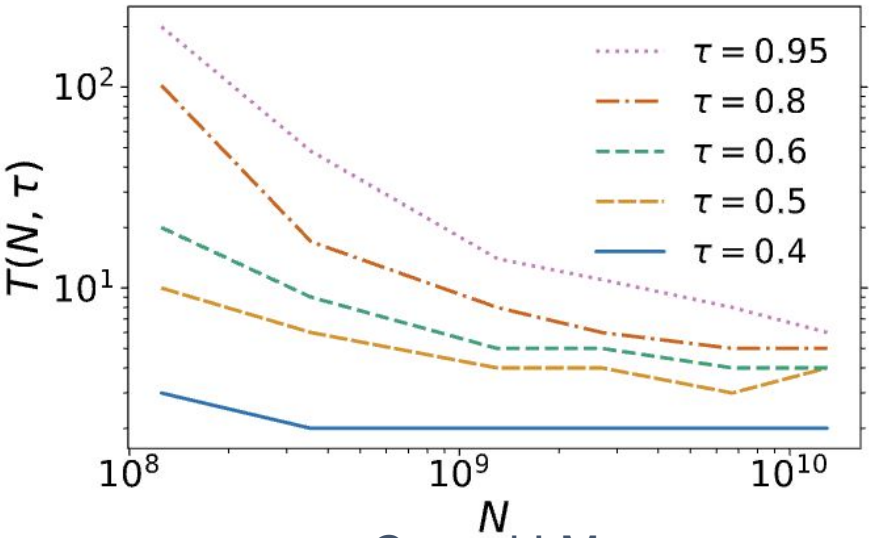
$T(N, \tau)$ is the minimal number of passes the model with size N needs to be trained in order to achieve memorization ratio > τ



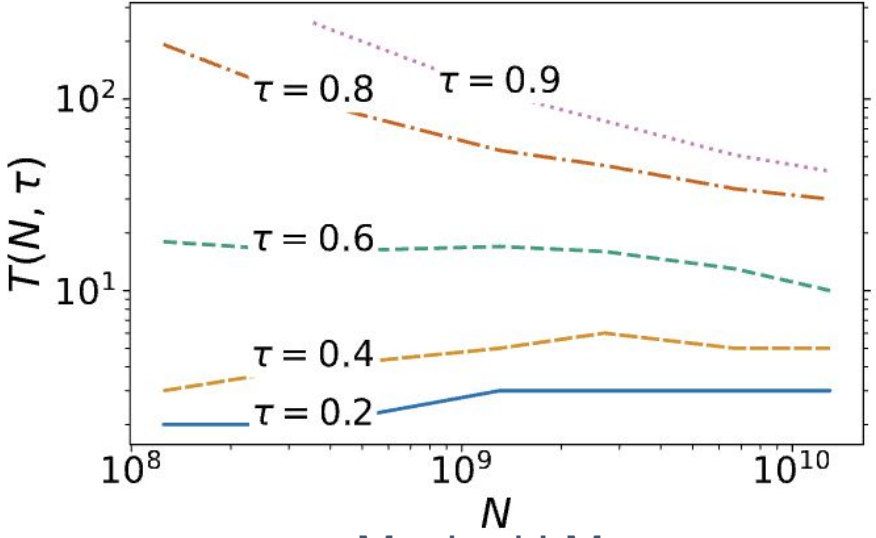
Larger language models not only memorize more of the training data, but are also more sample efficient and memorize faster.

Figure from Tirumala, Kushal, et al. "Memorization without overfitting: Analyzing the training dynamics of large language models." Advances in Neural Information Processing Systems (2022).

Causal LMs Memorize Faster than Masked LMs



Causal LM

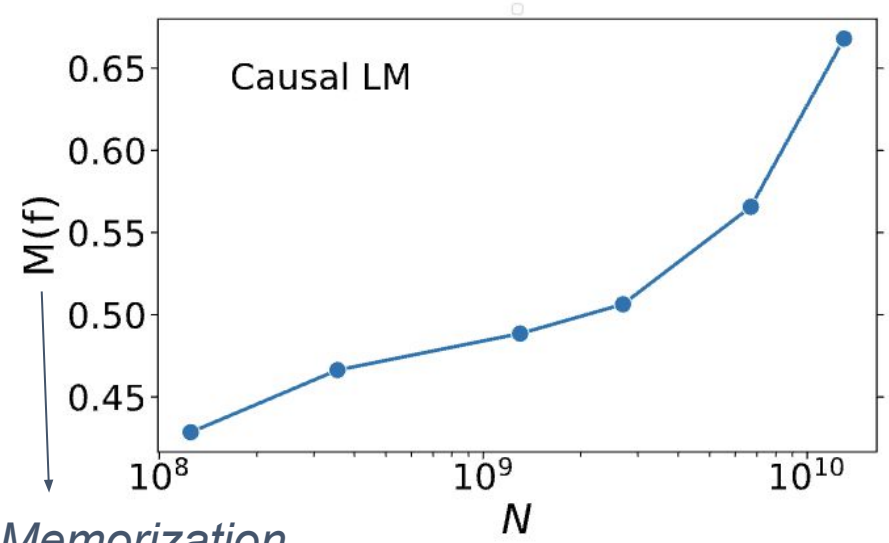


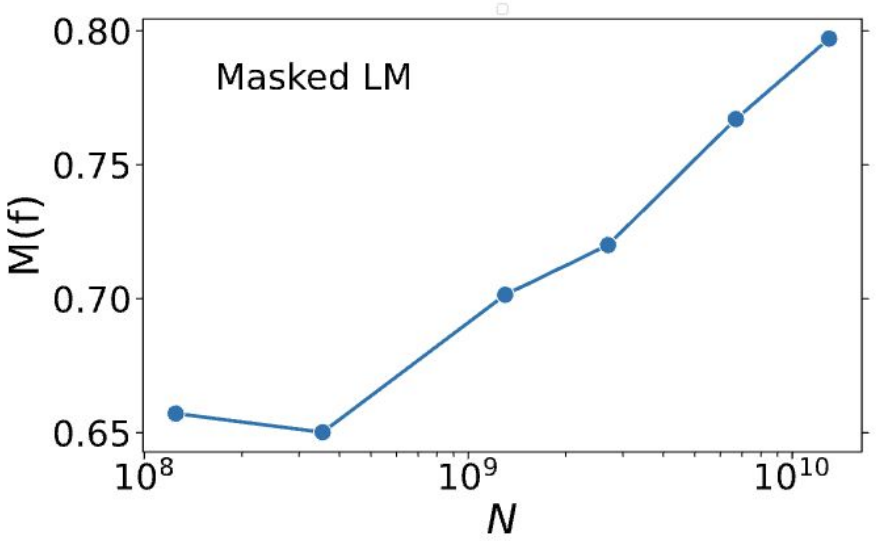
Masked LM

For masked LM, the masking ratio is set to 0.15, which leads to less “training signal” compared to autoregressive causal LM. For both task formulations, we see that larger models memorize faster.

Memorization Precedes Overfitting

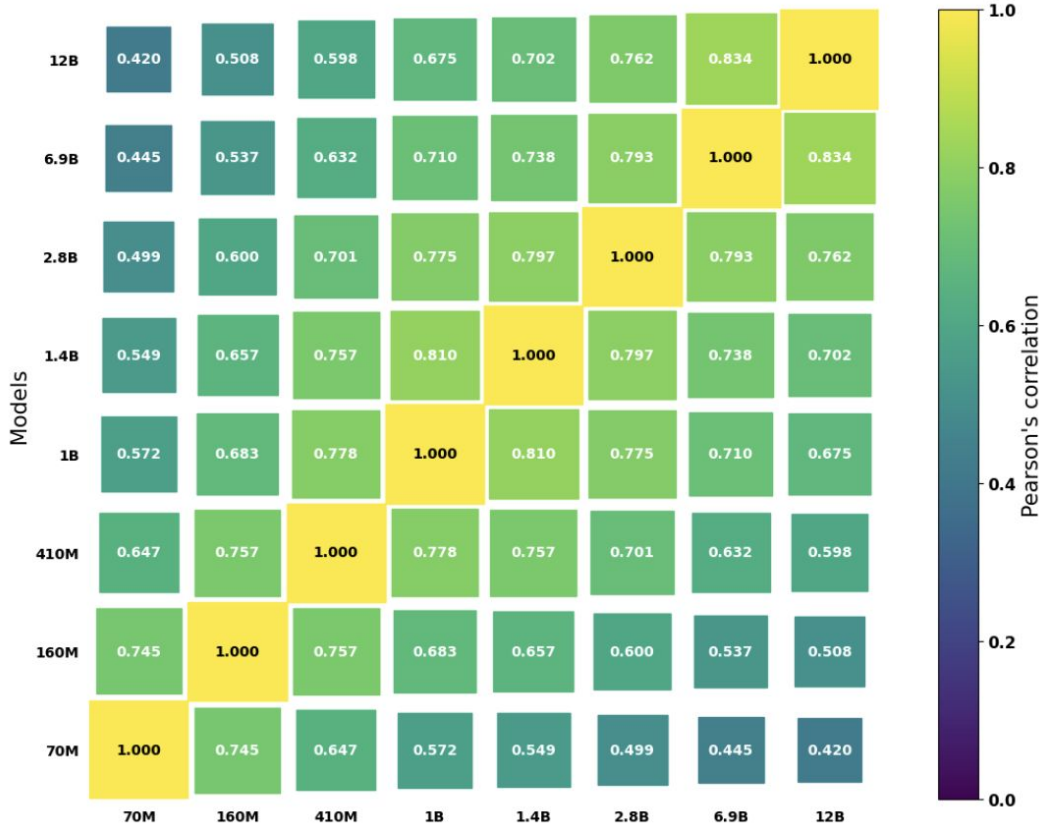
Assume that overfitting happens when the validation loss increases.





Memorization ratio

Instance-Level Memorization is NOT Predictable

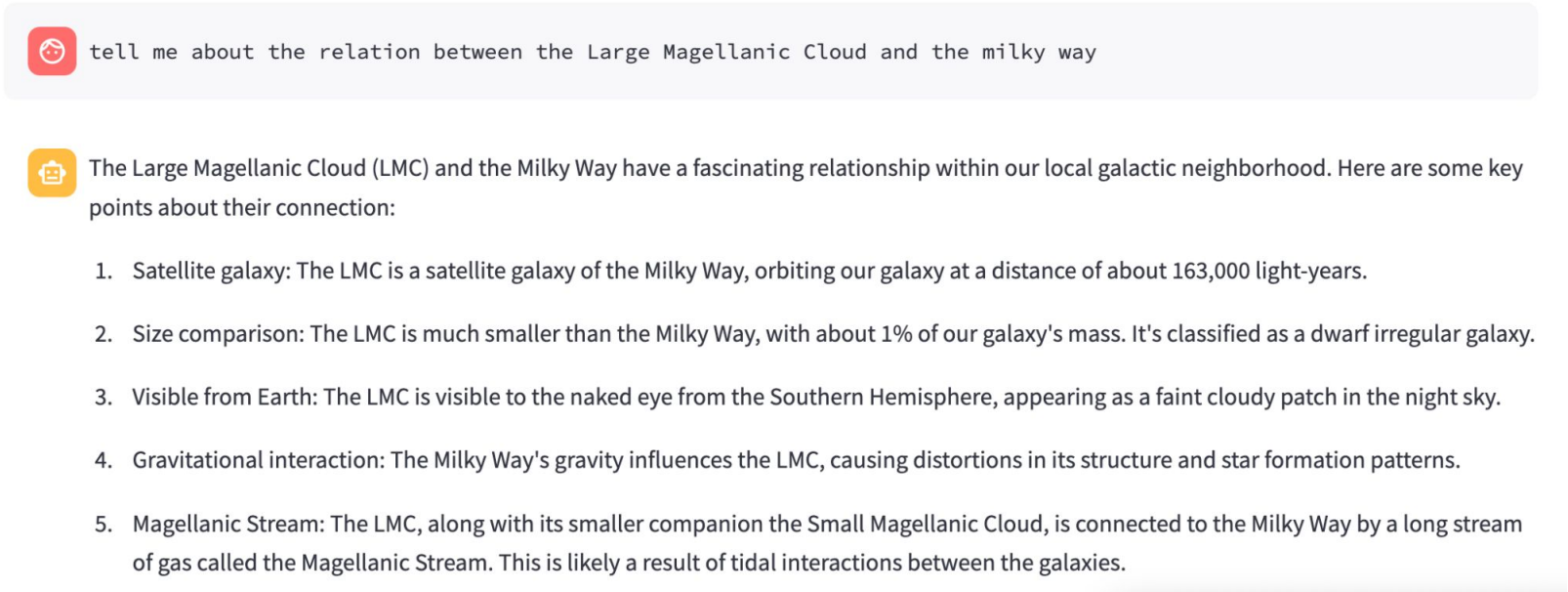


- Correlation between sequences memorized by small models and large models fall off quickly.

Figure from Biderman, Stella, et al. "Emergent and predictable memorization in large language models." Advances in Neural Information Processing Systems 36 (2023): 28072-28090.

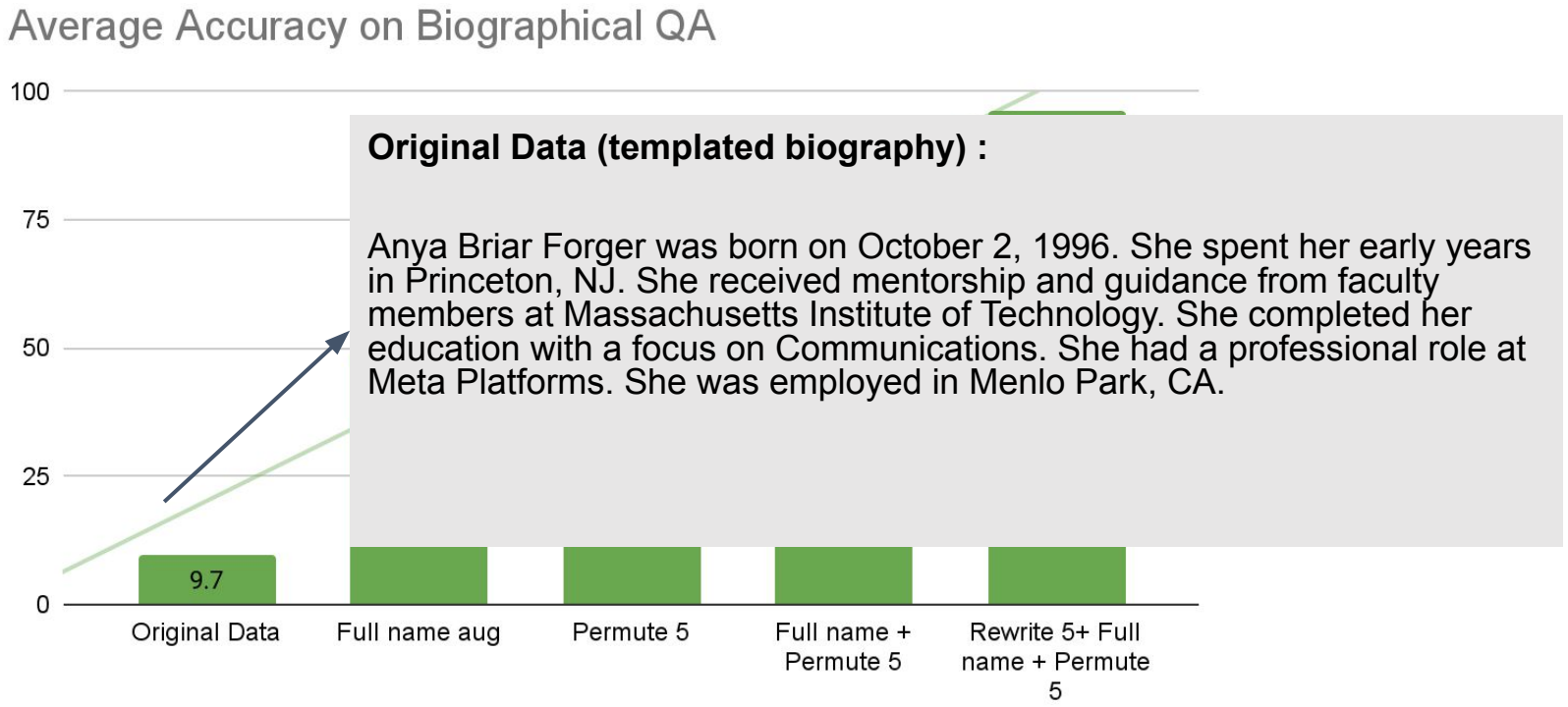
Knowledge is not just Memorization

For a model to be deemed “knowledgeable” we must be able to extract the knowledge in a flexible way (not a fixed prompt).



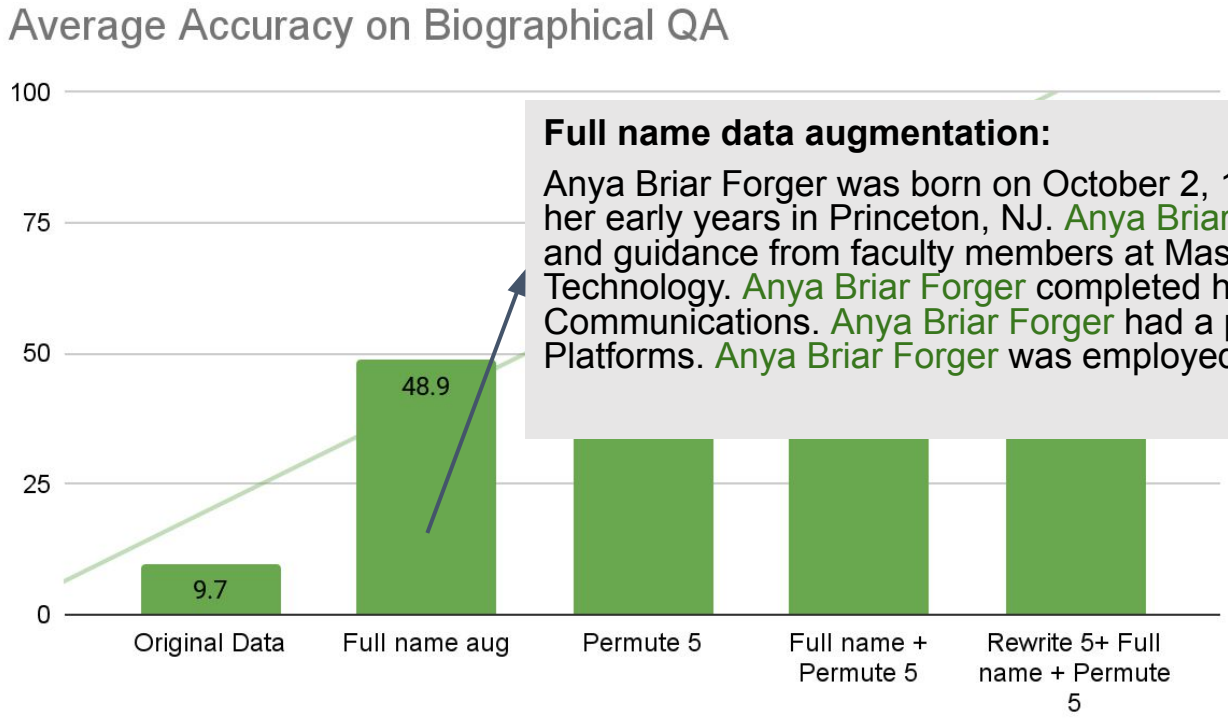
[1] Conversation with Claude Sonnet 3.5.

Knowledge Acquisition requires Diversification



We first train the model on randomly generated biographies and then perform QA instruction fine-tuning to let the model learn to answer questions.

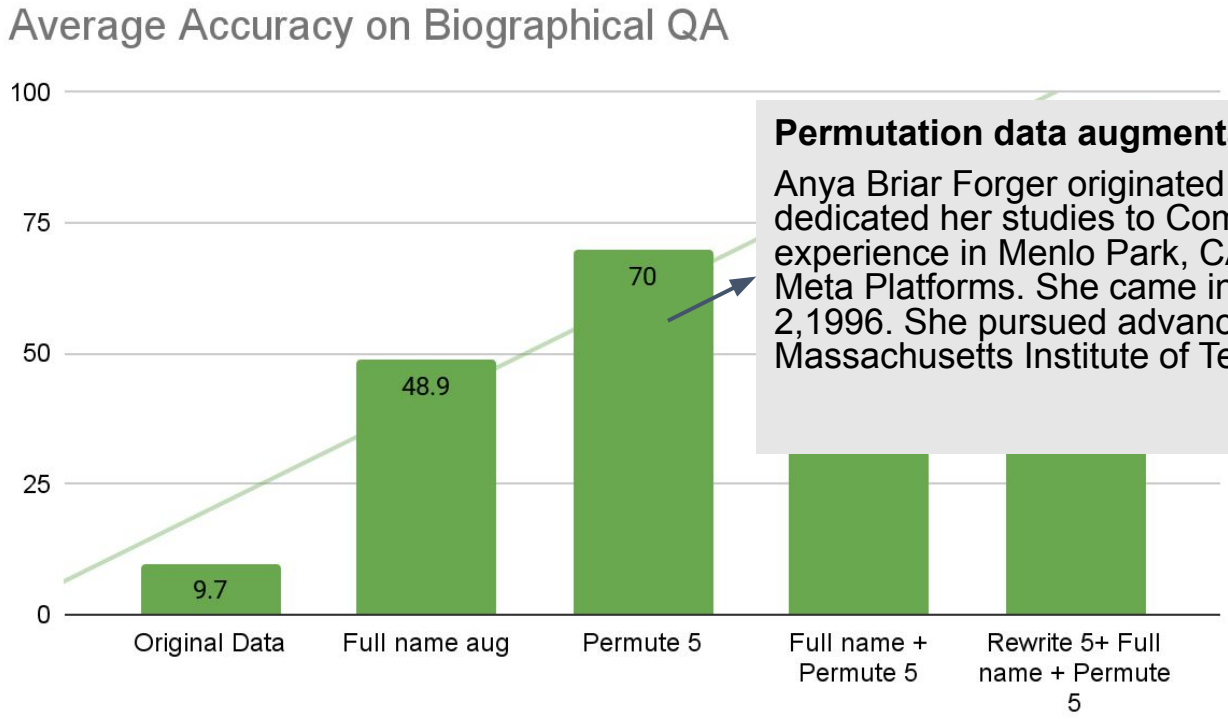
Knowledge Acquisition requires Diversification



Full name data augmentation:

Anya Briar Forger was born on October 2, 1996. Anya Briar Forger spent her early years in Princeton, NJ. Anya Briar Forger received mentorship and guidance from faculty members at Massachusetts Institute of Technology. Anya Briar Forger completed her education with a focus on Communications. Anya Briar Forger had a professional role at Meta Platforms. Anya Briar Forger was employed in Menlo Park, CA.

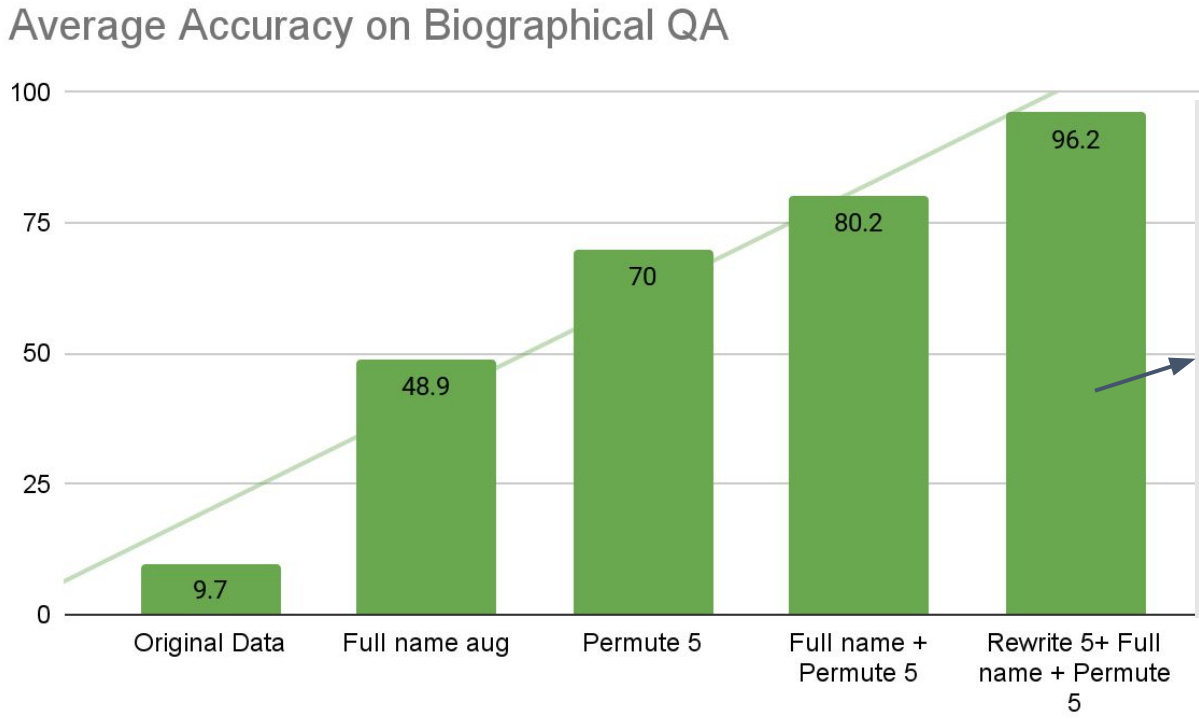
Knowledge Acquisition requires Diversification



Permutation data augmentation:

Anya Briar Forger originated from Princeton, NJ. She dedicated her studies to Communications. She gained work experience in Menlo Park, CA. She developed her career at Meta Platforms. She came into this world on October 2,1996. She pursued advanced coursework at Massachusetts Institute of Technology.

Knowledge Acquisition requires Diversification



Rewrite data augmentation:

Anya Briar Forger came into this world on October 2, 1996. She originated from Princeton, NJ. She pursued advanced coursework at Massachusetts Institute of Technology. She dedicated her studies to Communications. She developed her career at Meta Platforms. She gained work experience in Menlo Park, CA.

Early Task Diversification is Helpful

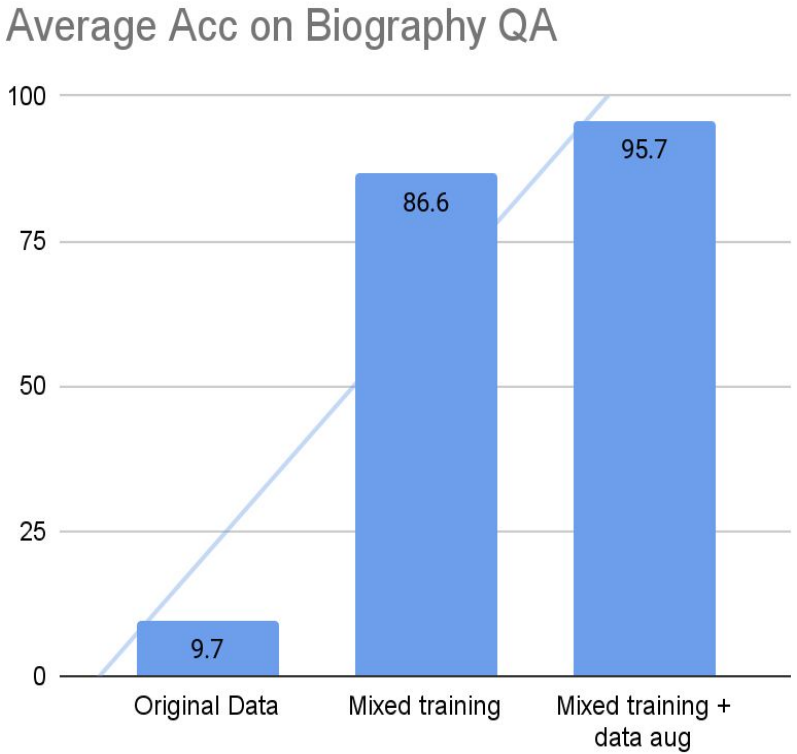
Conventional training pipeline

Pretraining (Next token prediction) → Post-training (QA instruction tuning)

Mixed training

Pre-training + QA instruction tuning

Introducing instruction-tuning data early in training improves knowledge extraction significantly.



Results from Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of language models: Part 3.1, knowledge storage and extraction." arXiv preprint arXiv:2309.14316 (2023).

Moving Instruction-Tuning Early is Beneficial

Conventional training pipeline

Pretraining → QA instruction tuning

Mixed training

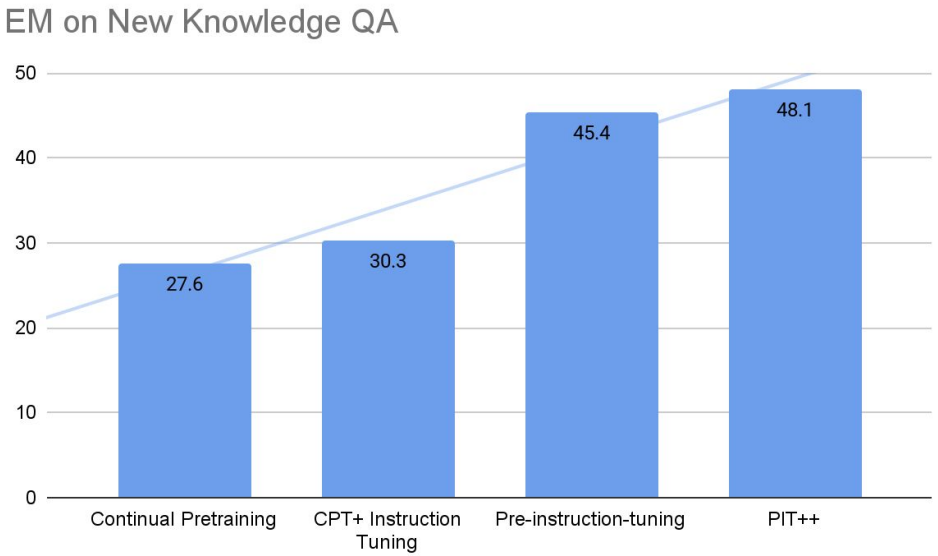
Pre-training + QA instruction tuning

Pre-instruction tuning

QA instruction tuning → Pretraining

Pre-instruction tuning ++

QA instruction tuning → Pre-training + QA instruction tuning



In the continual pretraining setting, moving QA instruction tuning before pretraining on new data is found to be beneficial for knowledge acquisition.

Jiang, Zhengbao, et al. "Instruction-tuned language models are better knowledge learners." ACL (2024).

Quantifying Knowledge in LMs

- If we define knowledge as “extractable knowledge” that can be probed with the QA format, how much knowledge can a LM contain?

Measure knowledge capacity of a given LM:

- Train over the synthetic biography dataset (with random selected templates and ordering to ensure knowledge extraction)
- N is the number of people included in the dataset
- Each piece of knowledge is repeated 1000 times during training
- Model architecture follows GPT2 with rotary embedding

Bit-complexity lower bound

Capacity Ratio

$$R(F) \stackrel{\text{def}}{=} \frac{N \log_2 \frac{N_0}{e^{p1}} + NK \log_2 \frac{D^C}{e^{p2}} + KD \log_2 \frac{T^L}{De^{p3}}}{P}$$

Loss on name

Loss over attribute and first token of value

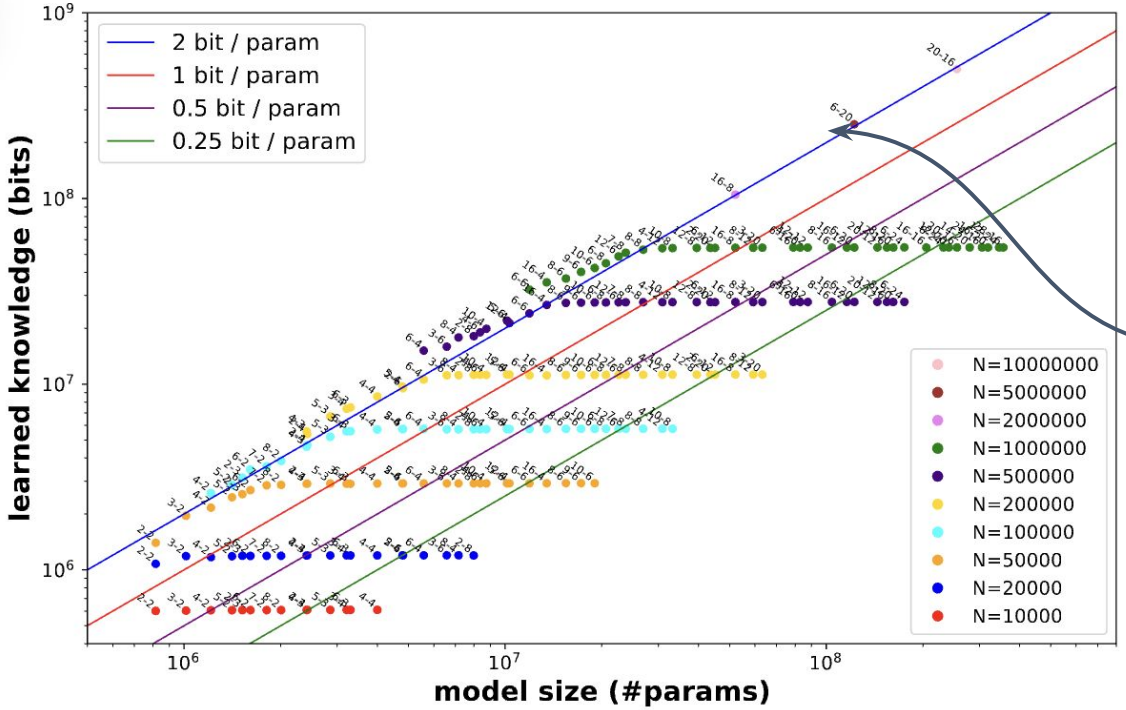
Loss over remaining tokens of value

Knowledge Capacity Scaling Laws

Capacity Ratio

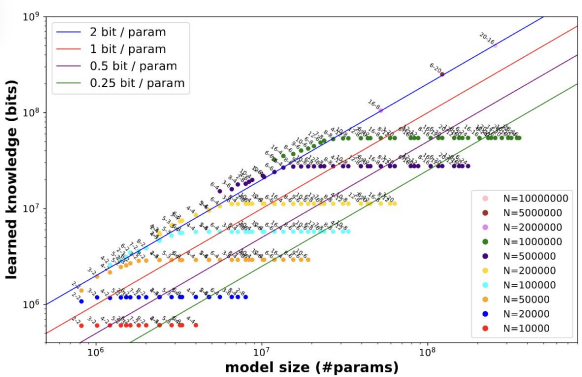
$$R(F) \stackrel{\text{def}}{=} \frac{N \log_2 \frac{N_0}{e^{p_1}} + N \log_2 \frac{S_0}{e^{p_2}}}{P}$$

Bit-complexity lower bound, simplified by removing the diversity term

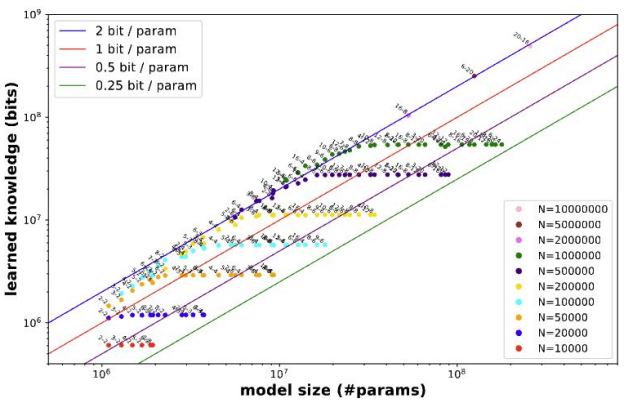


When we vary the model size and the dataset size (by N the number of people in the dataset), **under the best setting, we can see that the capacity ratio is close to 2 bits /param**

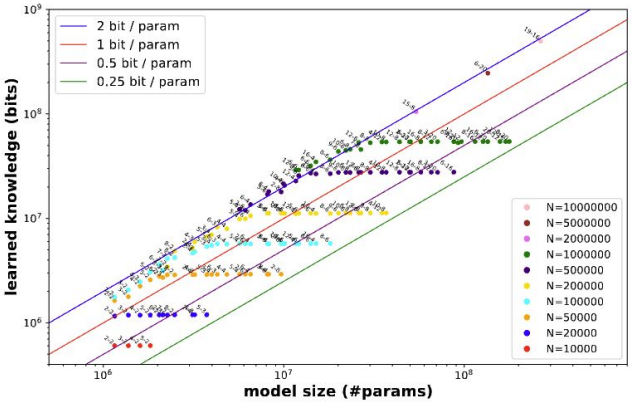
Knowledge Capacity Scaling for Transformer-variants with Sufficient Training



GPT-2 architecture, with rotary embeddings



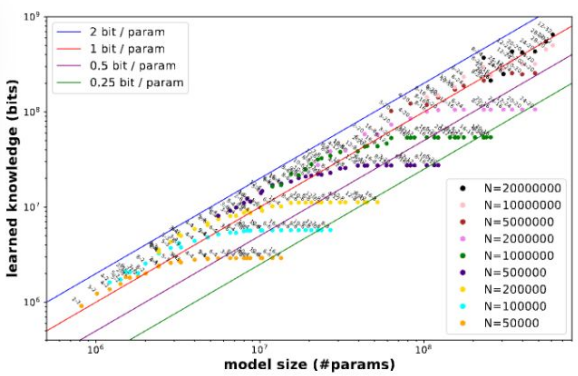
LLama architecture



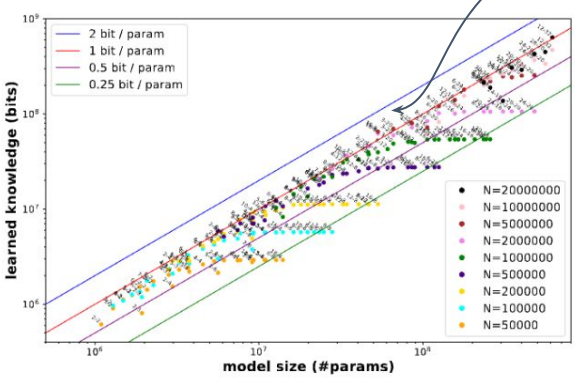
Mistral architecture

The 2bit/param capacity ratio is a relatively universal law among Transformer-based decoder-only language model architectures.

Knowledge Capacity Scaling for Transformer-variants with Insufficient Training

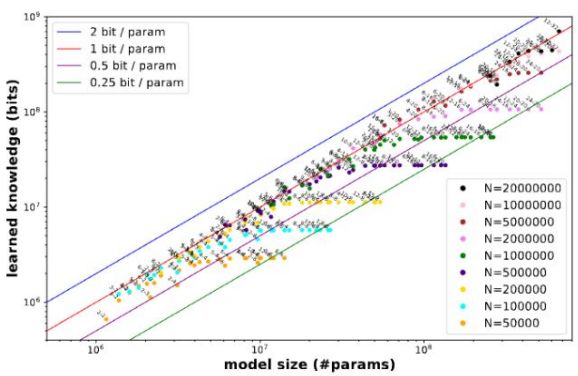


GPT-2 architecture, with rotary embeddings



LLama architecture

Knowledge/ param falls under 2 bits / param

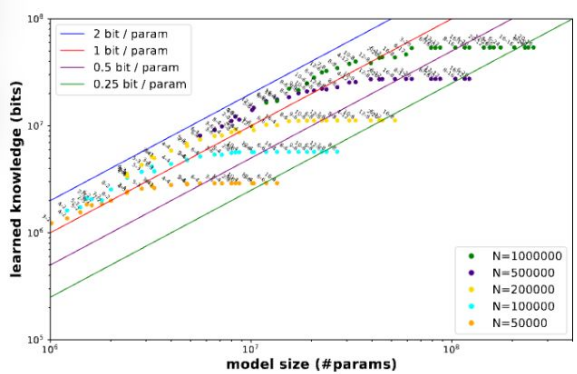


Mistral architecture

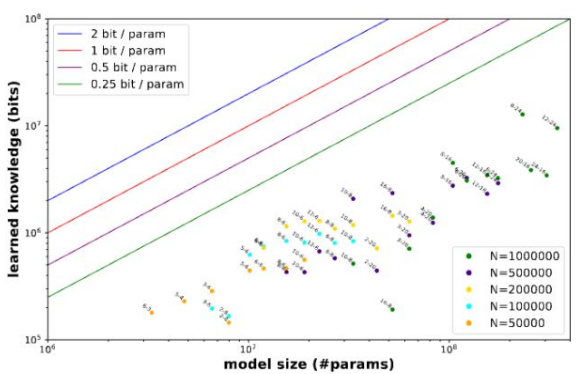
Both Llama and Mistral used gated MLP layers for improved training stability.

If we change the number of exposures of each fact from 1000 → 100 to simulate an insufficient training setting, model architecture choices make a difference.

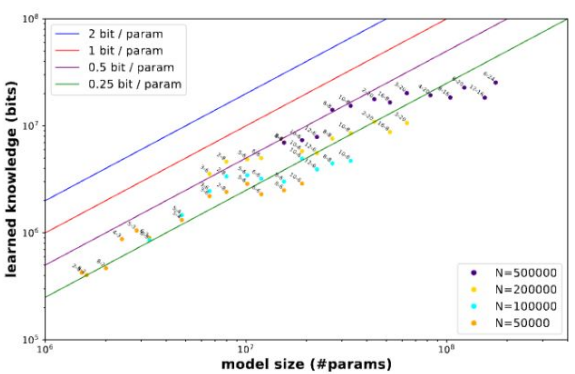
Knowledge Capacity Scaling with Data Mixing



(a) no junk, 100 exposures



(b) 7/8 junk, 100 exposures



(c) 7/8 junk, 300 exposures

If we mix in other data sources, the knowledge capacity of the model will be severely affected when the model is insufficiently trained.

- “Junk data” is from CommonCrawl web pages

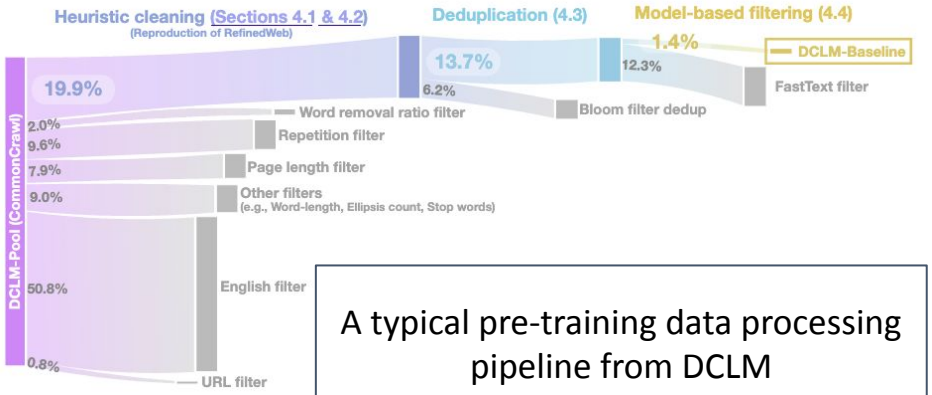
What we've learned so far…

- LMs' proficiency in factual knowledge is strongly correlated to memorization
  - Larger models & higher knowledge frequency → higher memorization rate & lower LM loss → better performance on knowledge-intensive benchmarks
- When training a LM, knowledge not only needs to be repeated but also diversified
  - Diversify the expression of knowledge through data augmentation
  - Teach the model to extract knowledge by early instruction tuning
  - *Knowledge augmentation does not need to be applied to all facts → the ability to extract knowledge is transferable across facts*

Does this apply to large-scale LM pretraining?

Connections to Modern LLM Data Preparation

- How can we get more performance with the same amount of compute?
- All modern LLMs are trained with CommonCrawl data (the junk data that we just mentioned!)
  - The data is extensively filtered and cleaned, only 1-2% of the original data is kept for training
  - Web data is mixed with other high quality sources such as Wikipedia, Books, ArXiv, code dumps



A typical pre-training data processing pipeline from DCLM

How can we improve LM knowledge acquisition?

Improve knowledge density

Data Deduplication

Quality Filtering

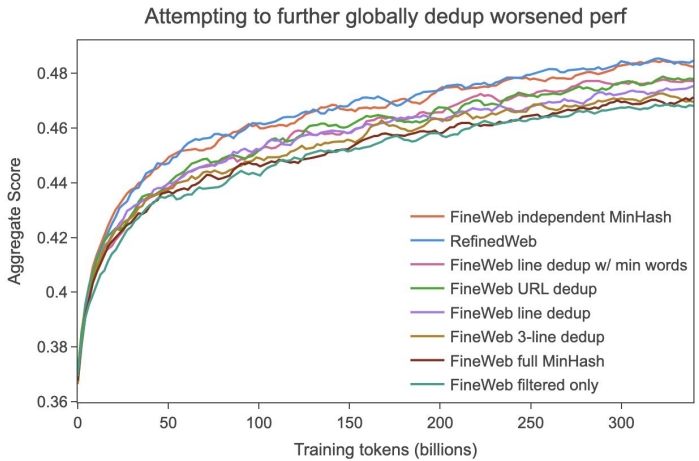
Improve knowledge diversity

Synthetic Data Rewriting

Multi-stage Pretraining
(Early Instruction Tuning)

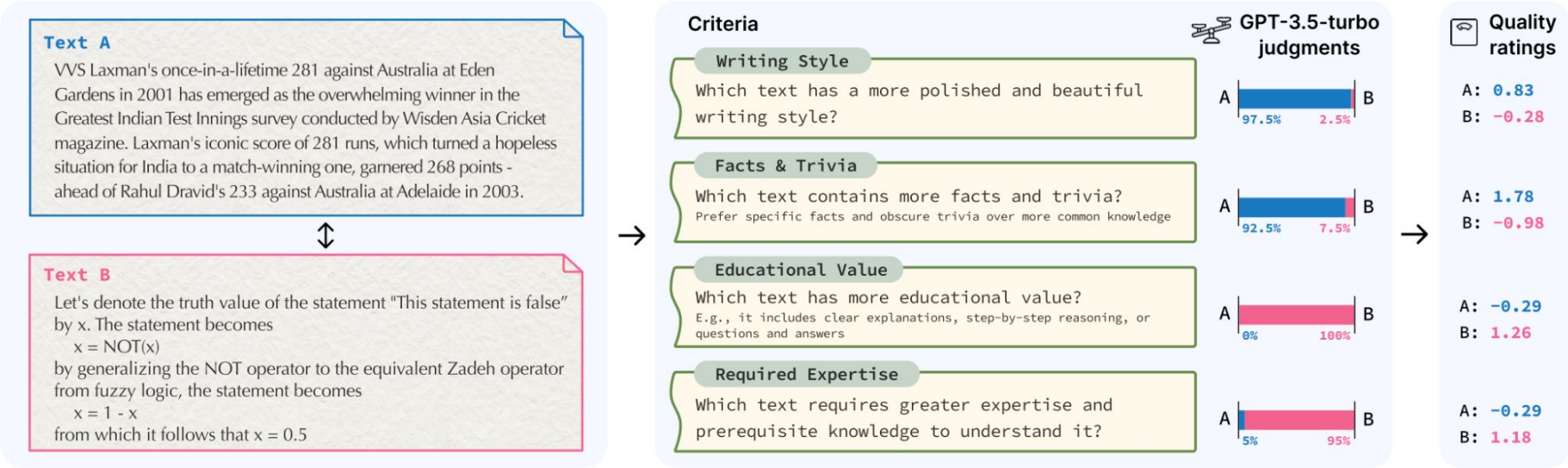
The Curious Case of Data Deduplication

- If knowledge memorization improves with duplicate data, why should I deduplicate my data?
  - Most of the exact duplicates in web data is actually computed generated boilerplate content
  - Another source of exact duplicates contain PII and unique ids are easily memorized
  - Doing too much deduplication, particularly at a fine-grained level, might be hurtful



[1] Lee et al. Deduplicating Training Data Makes Language Models Better. ACL 2022
[2] Figure from Penedo et al. FineWeb: decanting the web for the finest text data at scale.
https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1

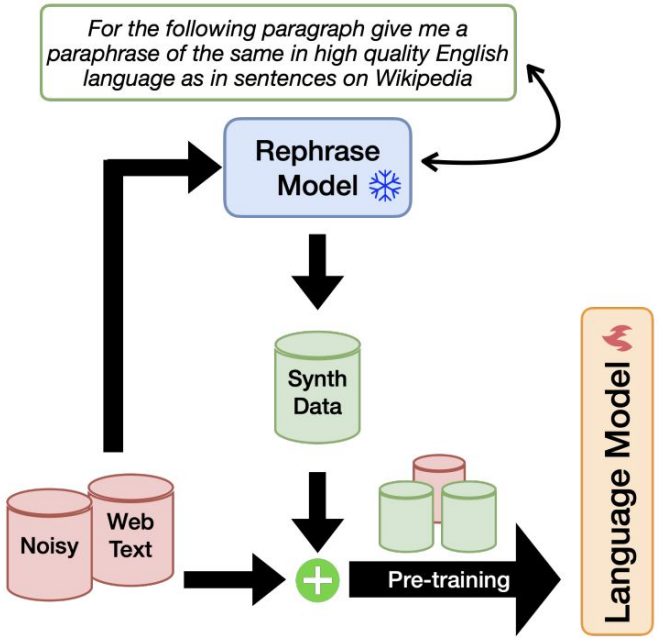
Selecting High Quality Data for Training

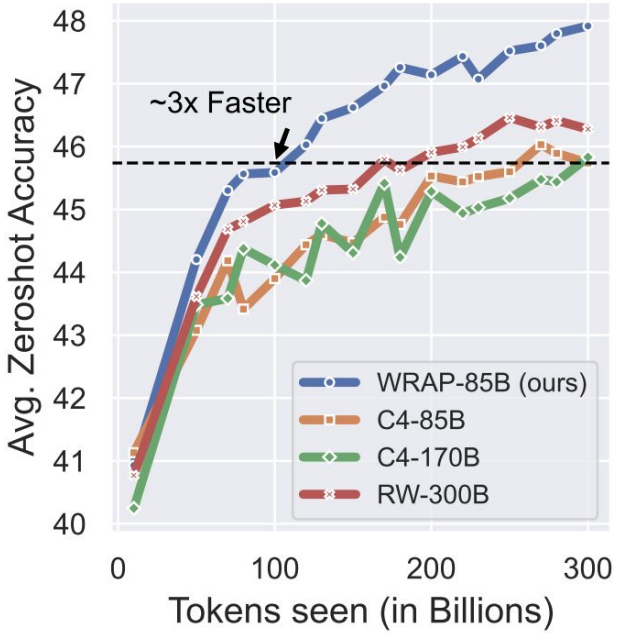


Data selection models need to be very scalable to operate over trillions of tokens. QuRating proposes to use a LLM to obtain quality ratings and then distill them into a small LM.

Wettig, Alexander, et al. "Qurating: Selecting high-quality data for training language models." arXiv preprint arXiv:2402.09739 (2024).

Synthetic Data Rewriting





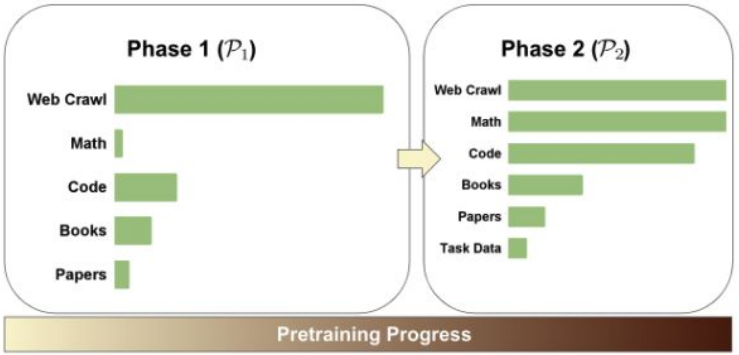
2 main goals for rewriting:

- Improve the quality of noisy data
- Create diverse variants of high quality data

Multi-Stage Pretraining (Mid-Training)

Mid-training data typically includes:

- Upsampling high quality data sources
- Adding domain specific data for tasks such as math and coding
- Adding instruction tuning data



Source	Tokens	50B	
		Source %	Mix %
Filtered DCLM	752B	3.23	47.2
Decontam. FLAN	17.0B	50.0	16.6
StackExchange Q&A	1.26B	100	2.45
peS2o	58.6B	5.15	5.85
Wikipedia/Wikibooks	3.7B	100	7.11
Dolmino Math	10.7B	100	20.8

		Dev Benchmarks						Held-out Evals		
Checkpoint	Avg	MMLU	ARC_C	HSwag	WinoG	NQ	DROP	AGIEval	GSM8K	$MMLU_{PRO}$
OLMo 2 7B										
Pretraining	50.6	59.8	72.6	81.3	75.8	29.0	40.7	44.6	24.1	27.4
Pretraining & mid-training	61.2	63.7	79.8	83.8	77.2	36.9	60.8	50.4	67.5	31.0

Knowledge Storage

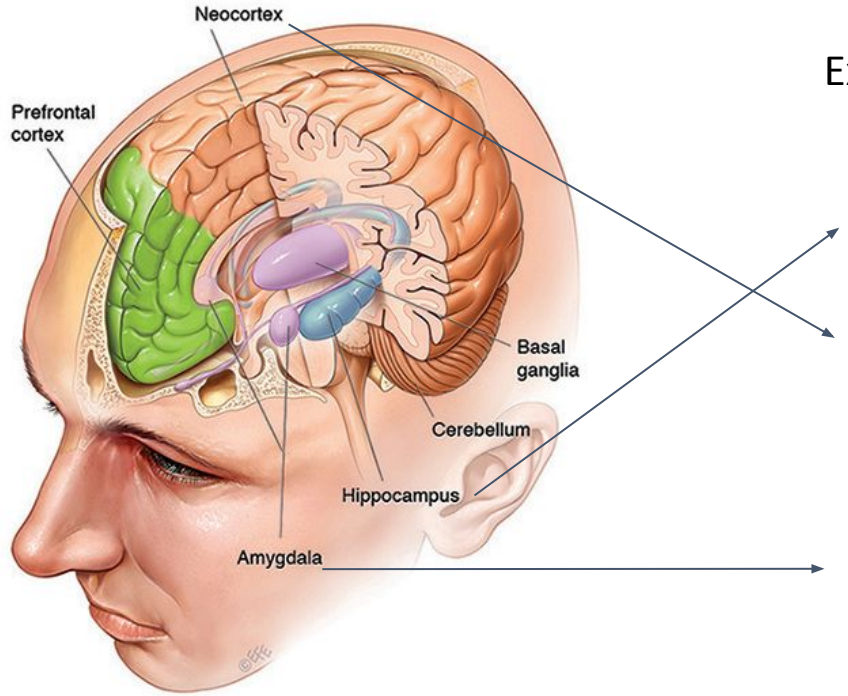
Chi Han

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Where is Knowledge Stored in LLMs?

Which parameters (or neurons) store certain knowledge?
How do they store and output the stored information?
How is this storage organized?

Where Is Knowledge In Human Brains?

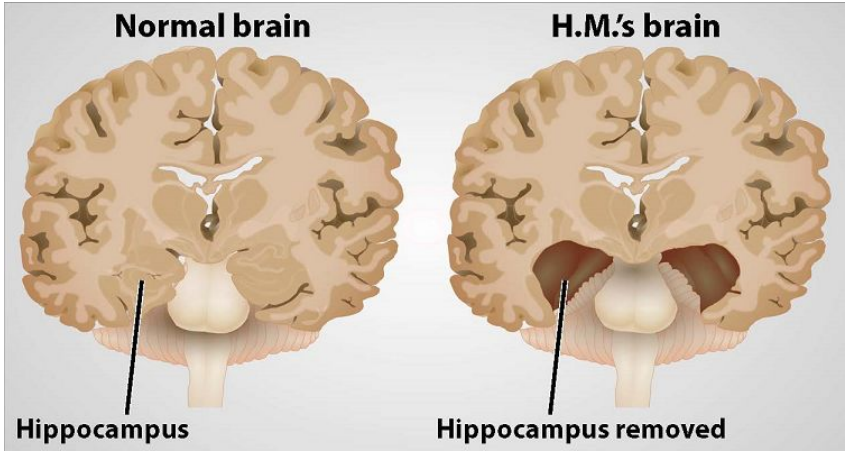


Explicit memory are store at**:**

- **Hippocampus**: episodic memory
- **Anterolateral temporal lobe**: semantic memory
- **Amygdala**: emotional implications

Credit to: https://qbi.uq.edu.au/memory/where-are-memories-stored

How Did We Investigate on Human Brains



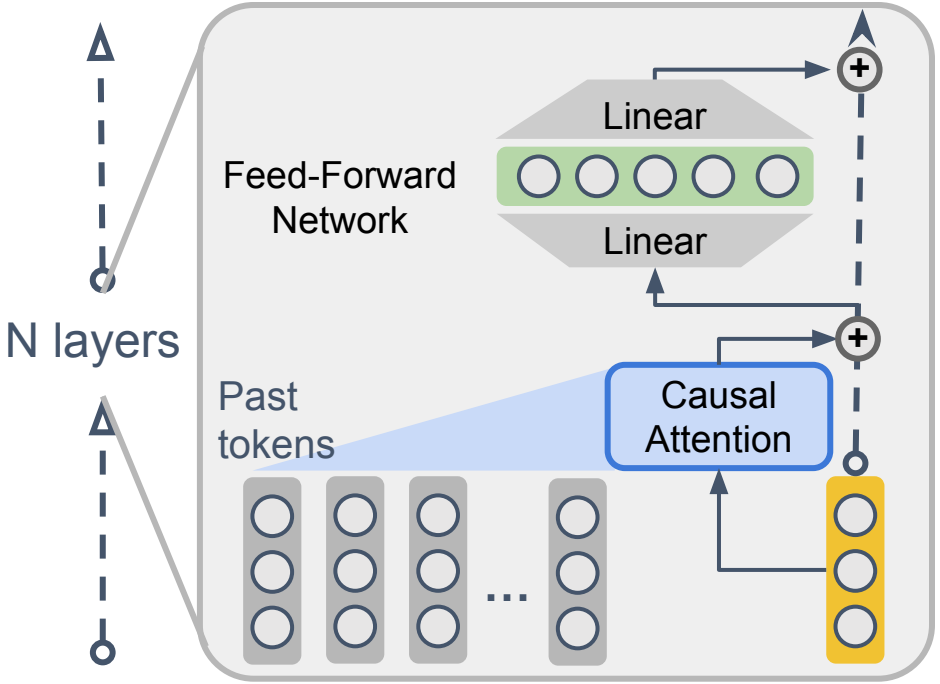
By comparing between people **with** and **without** a certain brain region.

- A patient who had their **hippocampus** surgically removed lost long-term episodic memories (events).[1]
- Atrophy of the **anterolateral temporal lobe** might affect knowledge and the association of concrete concepts.[2]
- Damage to the **amygdala** in Urbach-Wiethe disease might affect emotional memory, particularly those associated with fear.[3]

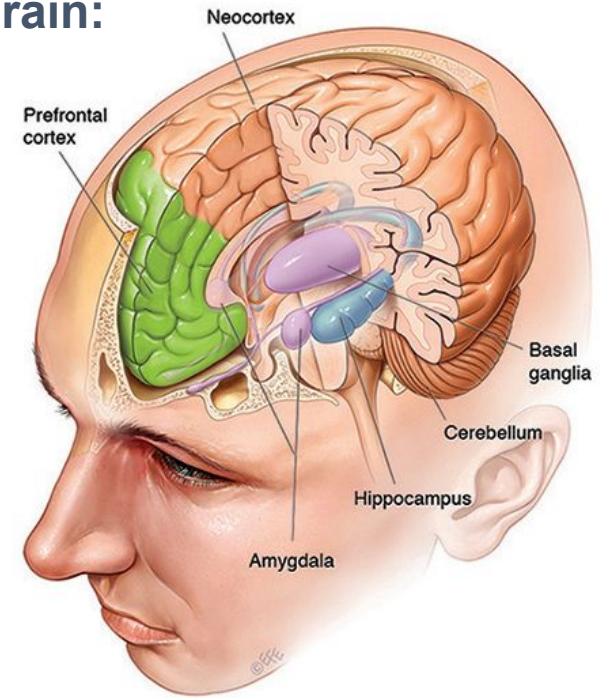
[1] https://www.brainfacts.org/in-the-lab/tools-and-techniques/2018/the-curious-case-of-patient-hm-082818
[2] Westerlund, Masha, and Liina Pylkkänen. "The role of the left anterior temporal lobe in semantic composition vs. semantic memory." Neuropsychologia 57 (2014): 59-70.
[3] Pause BM, Zlomuzica A, Kinugawa K, Mariani J, Pietrowsky R, Dere E. Perspectives on episodic-like and episodic memory. Front Behav Neurosci. 2013 Apr 18;7:33.

Parameters in LLMs ≈ Neurons in Human Brains ?

Transformer-based LLM



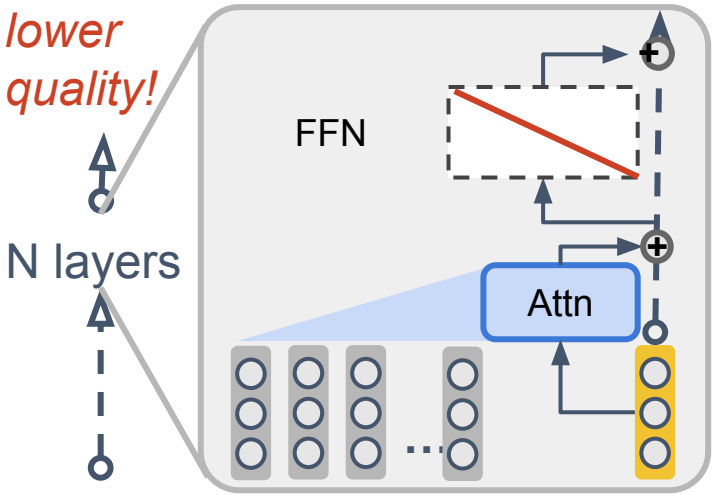
Brain:



Subject to subtle variances in different architectures

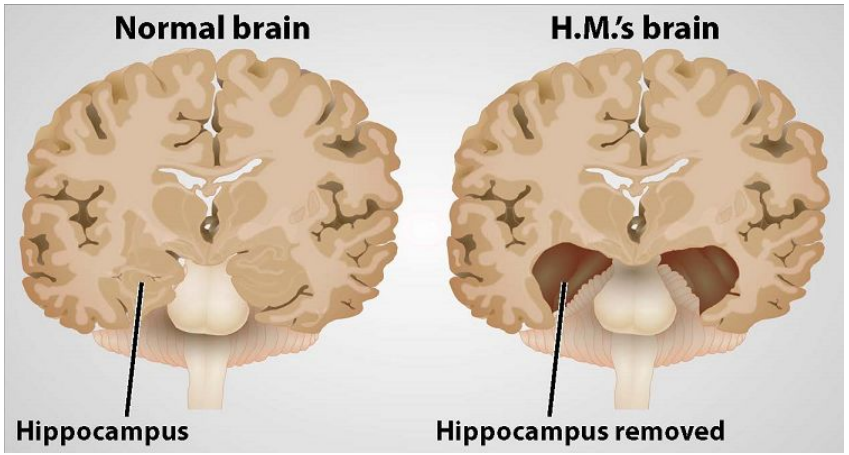
Credit to: https://qbi.uq.edu.au/memory/where-are-memories-stored

Modularity: A Difference Between LMs and Brains



LLMs parameters are more **entangled**

- Modifications to LLM parameter (group)s might *affect the general intelligence of the models*



Human brain functions **"modularly"**

- Issues on a brain region often cause *certain functional problems*

Gupta, Akshat, Anurag Rao, and Gopala Anumanchipalli. "Model editing at scale leads to gradual and catastrophic forgetting." *arXiv preprint arXiv:2401.07453* (2024).
Gu, Jia-Chen, et al. "Model editing can hurt general abilities of large language models." arXiv preprint

https://www.brainfacts.org/in-the-lab/tools-and-techniques/2018/the-curious-case-of-patient-hm-082818

How To Identify Knowledge Neurons in LMs?

General Idea:

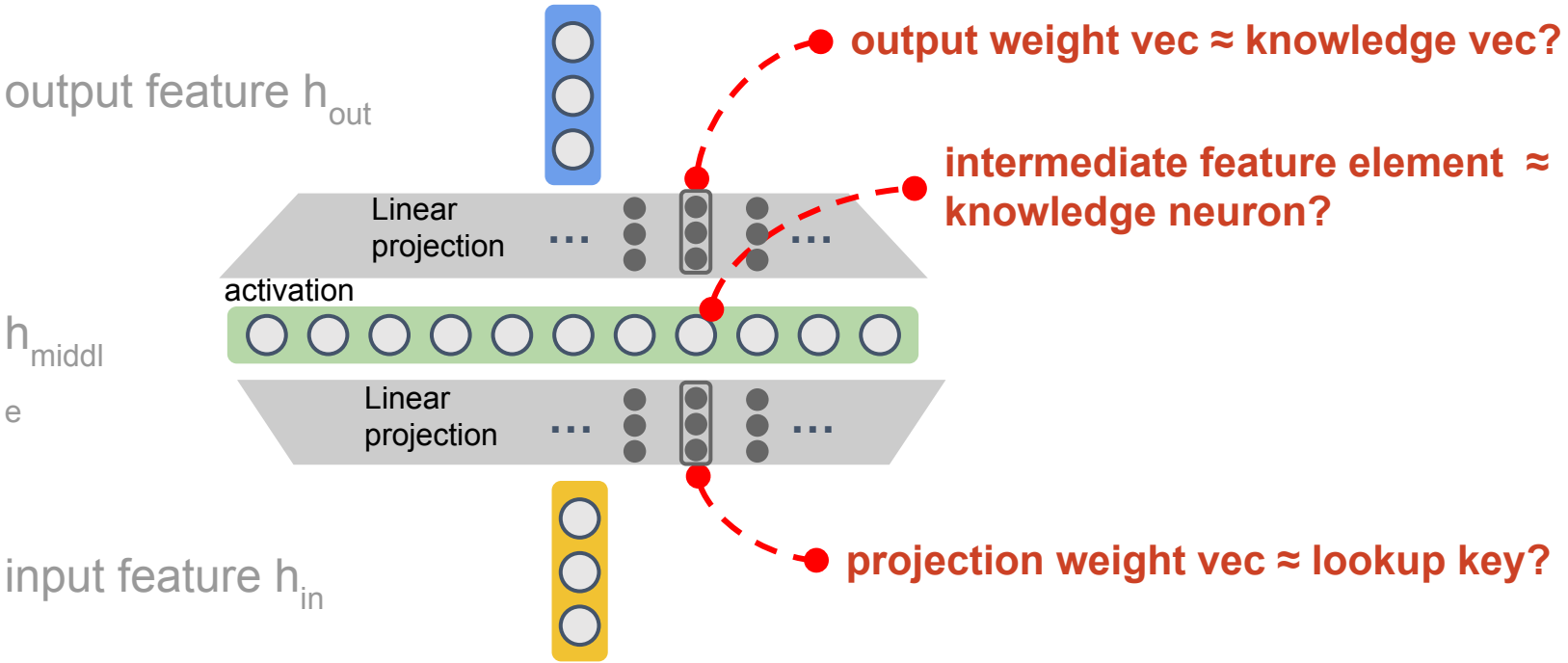
Looking for “responsive associations” between inputs, neurons, and outputs.:

- **Input → Neuron:**
  Do certain neurons respond to specific knowledge inputs?

- **Neuron → Output:**
  Do neuronal activities control the predicted knowledge?

Hypothesis 1: Feed-Forward Networks (FFN)

FFN ≈ knowledge lookup dict?

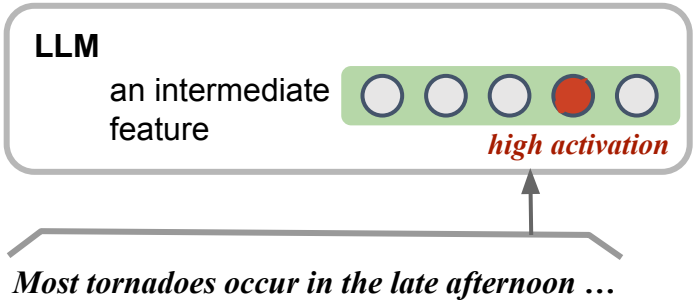


[1] Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.
[2] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.
[3] Meng, Kevin, et al. "Mass-Editing Memory in a Transformer." The Eleventh International Conference on Learning Representations.
[4] Dai, Damai, et al. "Knowledge Neurons in Pretrained Transformers." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Hypothesis 1: Feed-Forward Networks (FFN)

How to locate knowledge neurons?

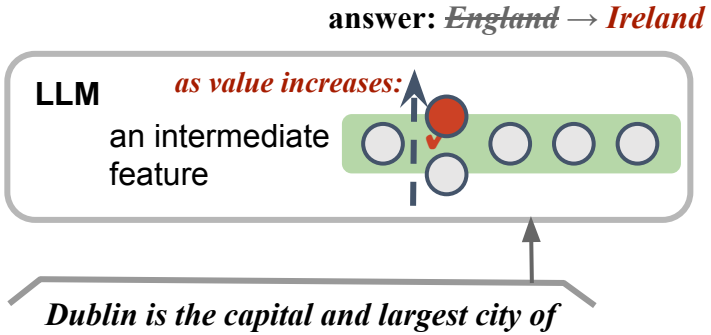
Neurons triggered by knowledge input[1]



Identified neurons about knowledge types:

- time ranges
- "part of" relations
- mentioning TV shows

Neurons causing knowledge outputs[2]



Identified <"A" is the "R" of "B"> triplet facts:

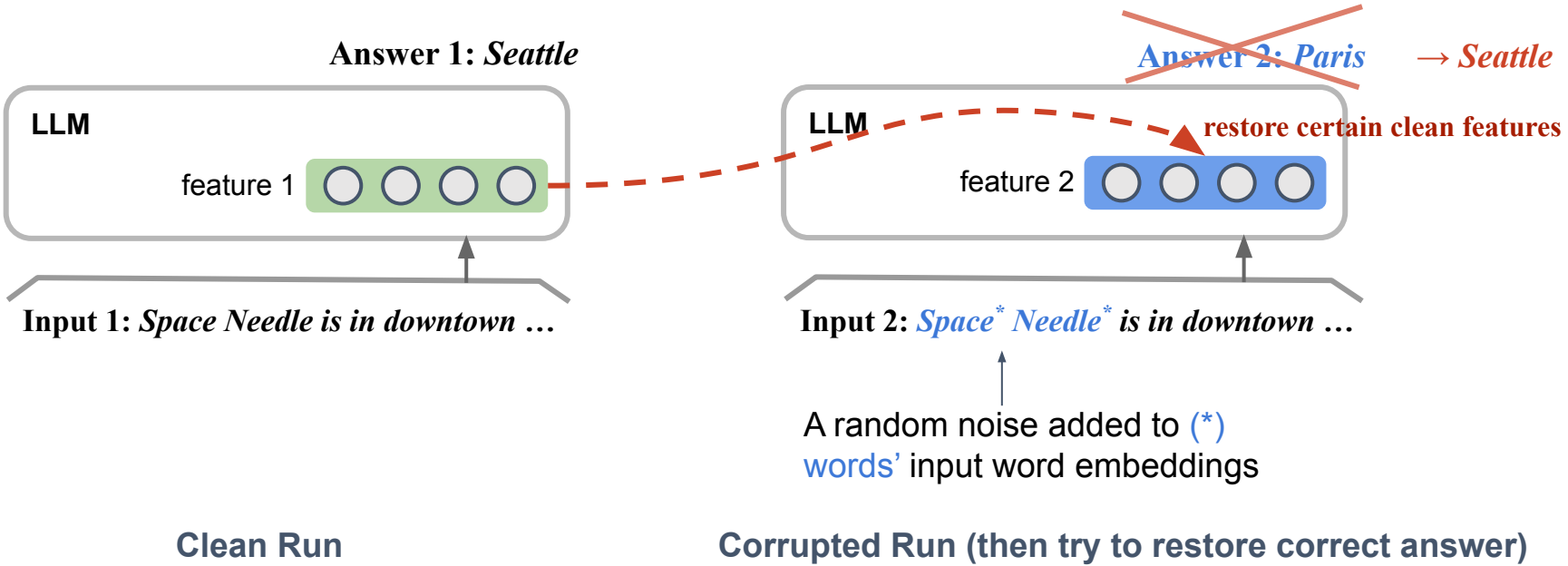
- Dublin is the capital of Ireland
- Kuwait is a country in Asia
- XXX is born in Shanghai

[1] Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.
[2] Dai, Damai, et al. "Knowledge Neurons in Pretrained Transformers." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Hypothesis 1: Feed-Forward Networks (FFN)

How to locate knowledge neurons?

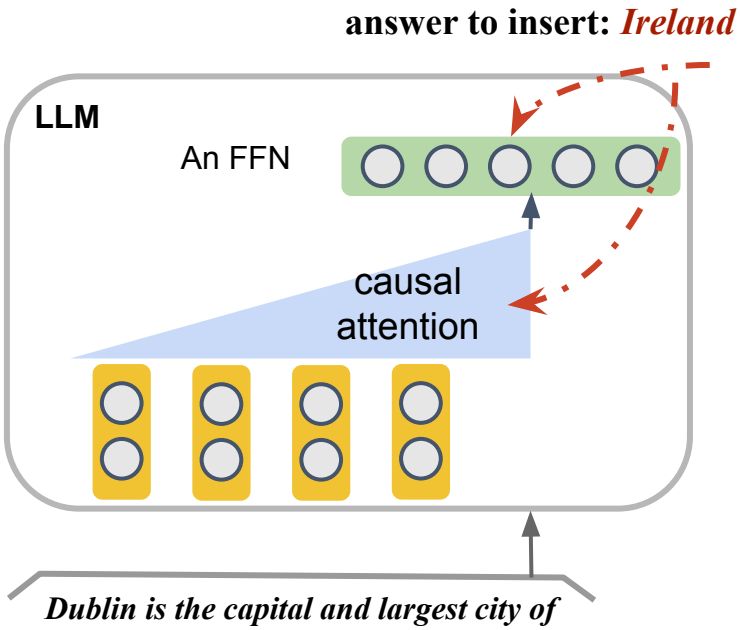
neurons that restore/contrast pairs of knowledge input-outputs[1]



[1] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

Hypothesis 2: Attention + Feed Forward Networks

Attention + FFN ≈ contextually-informed knowledge lookup dict?[1]



optimizing attention + FFN together

Finding: contextually-informed features has a higher impact to LLM knowledge update

[1] Li, Xiaopeng, et al. "Pmet: Precise model editing in a transformer." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

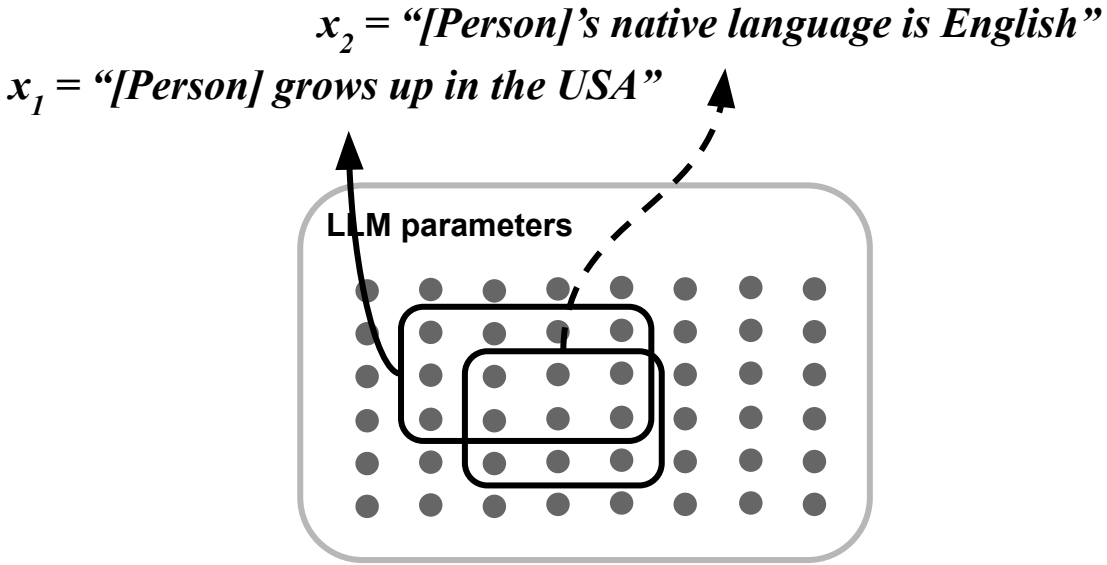
Rooms for Future Exploration

- Low specificity in found neurons
  - For each knowledge, “neurons” can be found in multiple layers[1]
- Located neurons might not inform knowledge editing
  - Found neurons ≠ best neuron to edit[2]
- Are the neurons more about “knowledge” or just “expression”? [3, 4]
- Are research efforts biased towards easily “verifiable” hypotheses?
  - It is easier to propose intuitive hypothesis on FFNs

[1] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.
[2] Hase, Peter, et al. "Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models." Advances in Neural Information Processing Systems 36 (2024).
[3] Niu, Jingcheng, et al. "What does the Knowledge Neuron Thesis Have to do with Knowledge?." The Twelfth International Conference on Learning Representations.
[4] Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

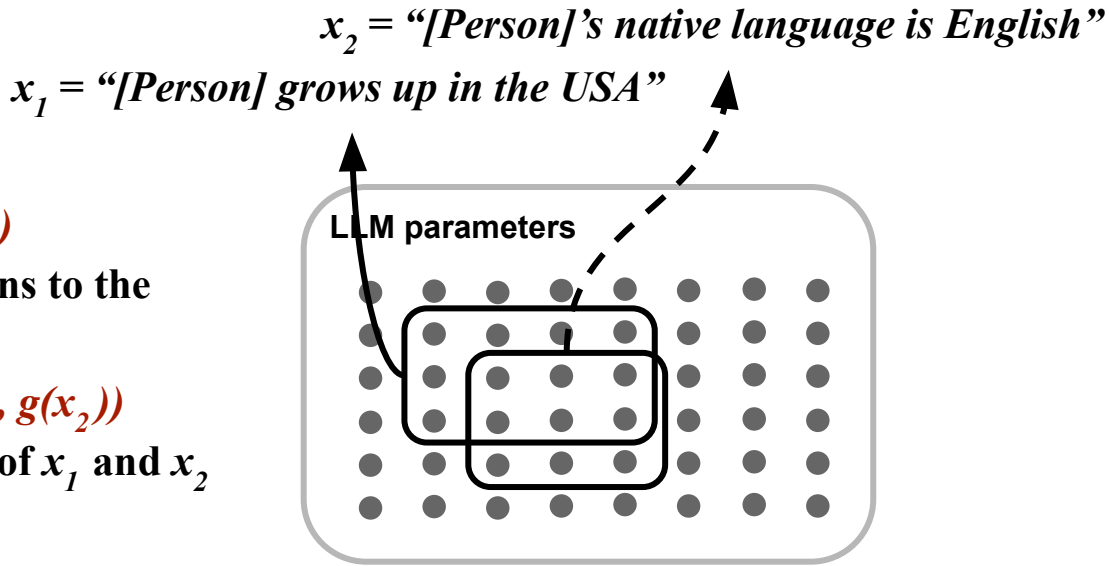
Knowledge Is Stored Messily in LLMs

One might expect that **knowledge should be stored according to semantic / logical relations**, (i.e., related facts should be associated with similar parameters)



Knowledge Is Stored Messily in LLMs

One indicator of knowledge parameter overlap: gradient similarity

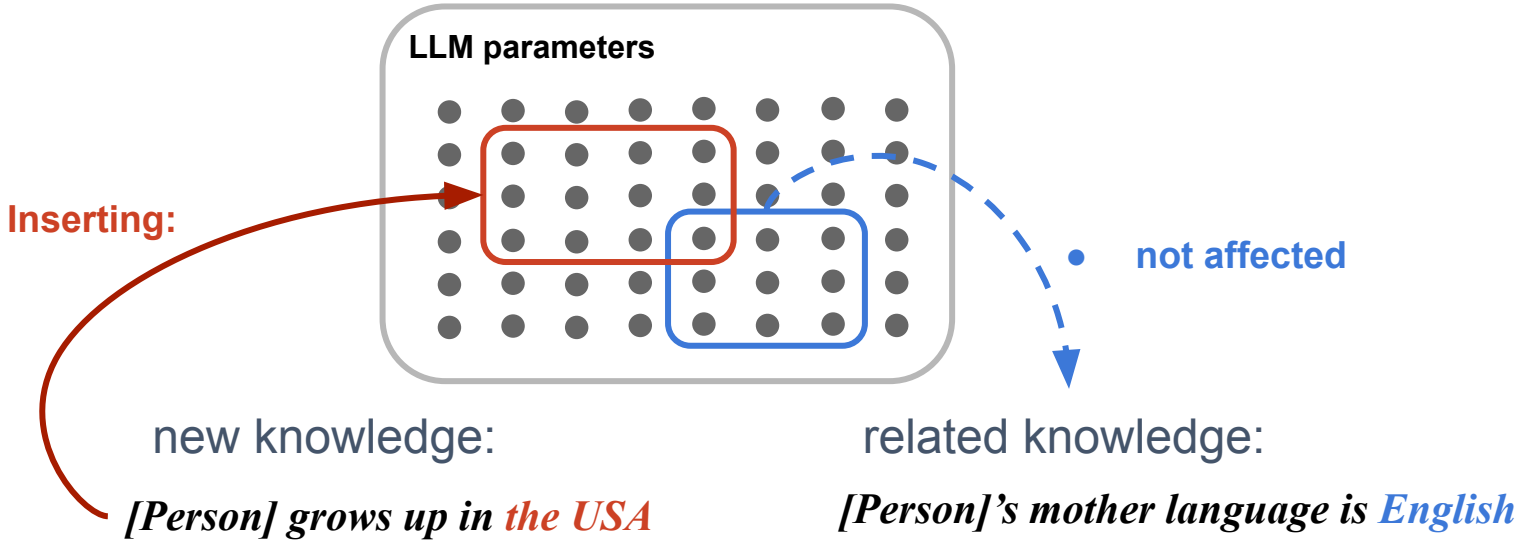


1. **gradient:** $g(x_1) = \nabla_\theta P_{LLM}(x_1)$
   Each parameter's contributions to the probability
2. **gradient similarity :** $cos(g(x_1), g(x_2))$
   Overlap between parameters of x_1 and x_2

Qin, Jiaxin, et al. "Why Does New Knowledge Create Messy Ripple Effects in LLMs?." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

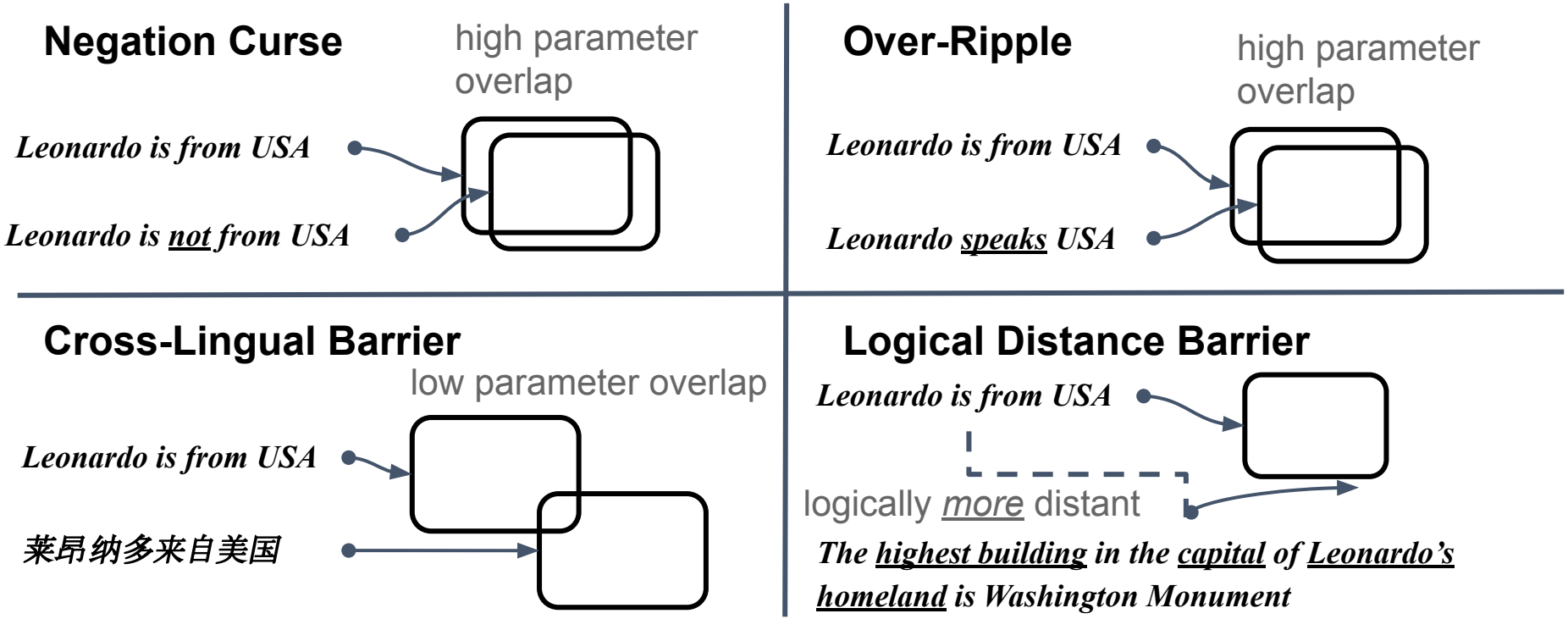
Are LLM Parameters Stored In An Organized Way?

It explains failed ripple effect.
i.e. the failure in updating related knowledge after inserting new ones



Cohen, Roi, et al. "Evaluating the Ripple Effects of Knowledge Editing in Language Models." *Transactions of the Association for Computational Linguistics* 11 (2024): 283-298.

Knowledge Is Stored Messily in LLMs

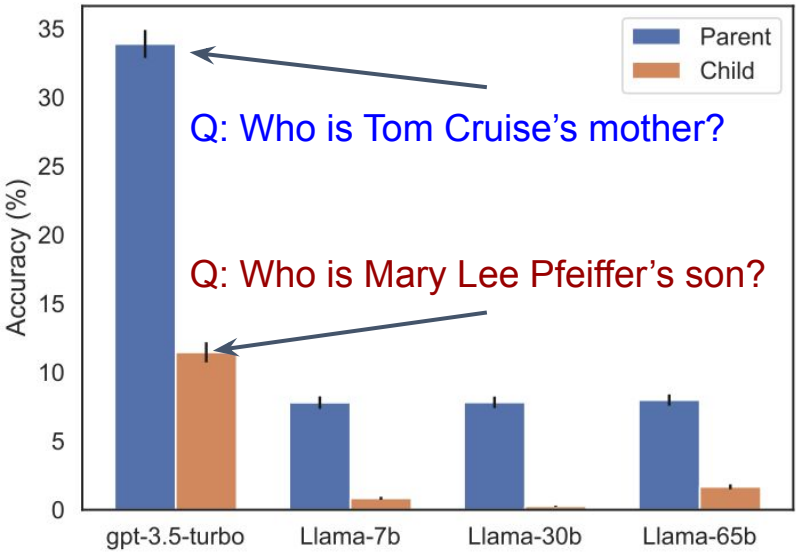


Qin, Jiaxin, et al. "Why Does New Knowledge Create Messy Ripple Effects in LLMs?." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Knowledge Is Stored Messily in LLMs

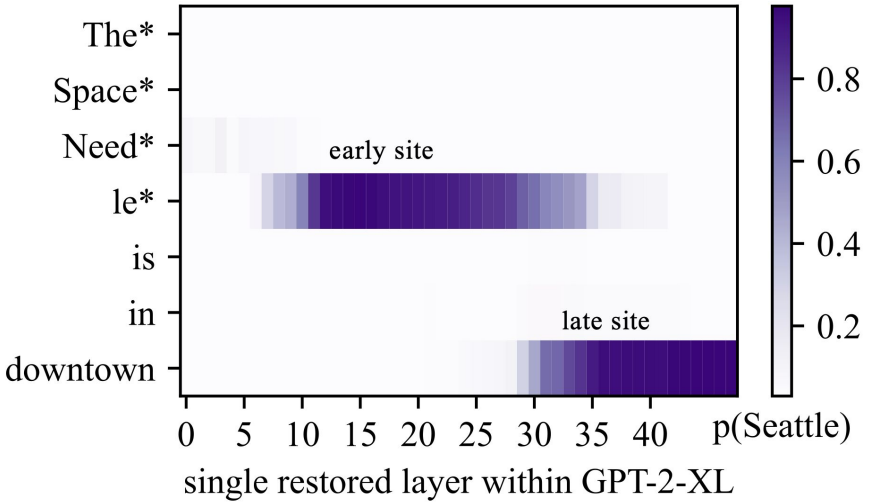
Reversal Curse

succeed on “A is [?]” but fail on “[?] is B”



Localization Redundancy

Multiple (layers of) neurons can be associated with one knowledge fact



Berglund, Lukas, et al. "The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”." The Twelfth International Conference on Learning Representations.
Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of Language Models: Part 3.2, Knowledge Manipulation." In The Thirteenth International Conference on Learning Representations.
Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

Knowledge Boundary

Yuji Zhang

Northwestern University

amazon

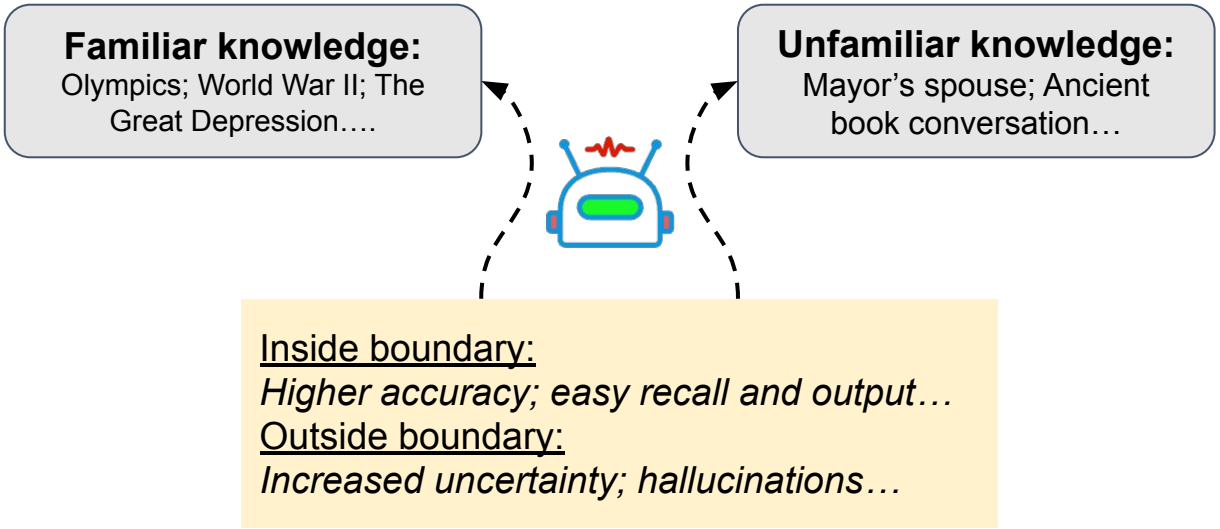
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Utilize Knowledge Boundary as the Indicator for Model Output

- What is knowledge boundary for LLMs?
- How knowledge boundary indicates model performance?
- Can model honestly deliver knowledge following its knowledge boundary?
- How to calibrate model expression to align with knowledge boundary?
- How introducing new knowledge impacts original knowledge boundary?

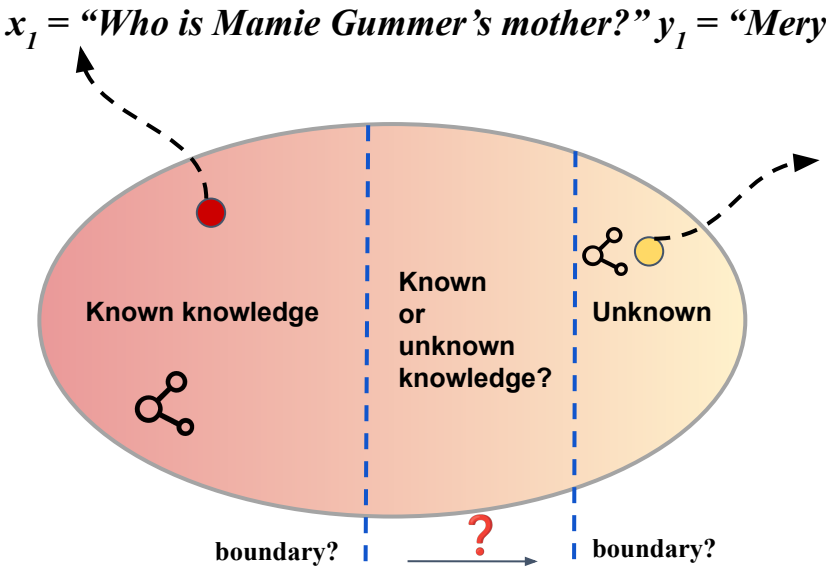
LLMs Inherently Have Knowledge Boundary

- Inside knowledge boundary: LLMs have highly familiar knowledge
- Outside knowledge boundary: LLMs have unfamiliar and longtail knowledge



Knowledge Boundary Exists from Pre-training Stage

- Knowledge boundary is non-binary given the probabilistic nature of LLMs
- LLMs could perform differently inside and outside knowledge boundary



x_2 = "Who is David Jordan's wife?" y_2 = "Mary Jordan" ❌

How boundary is shaped?

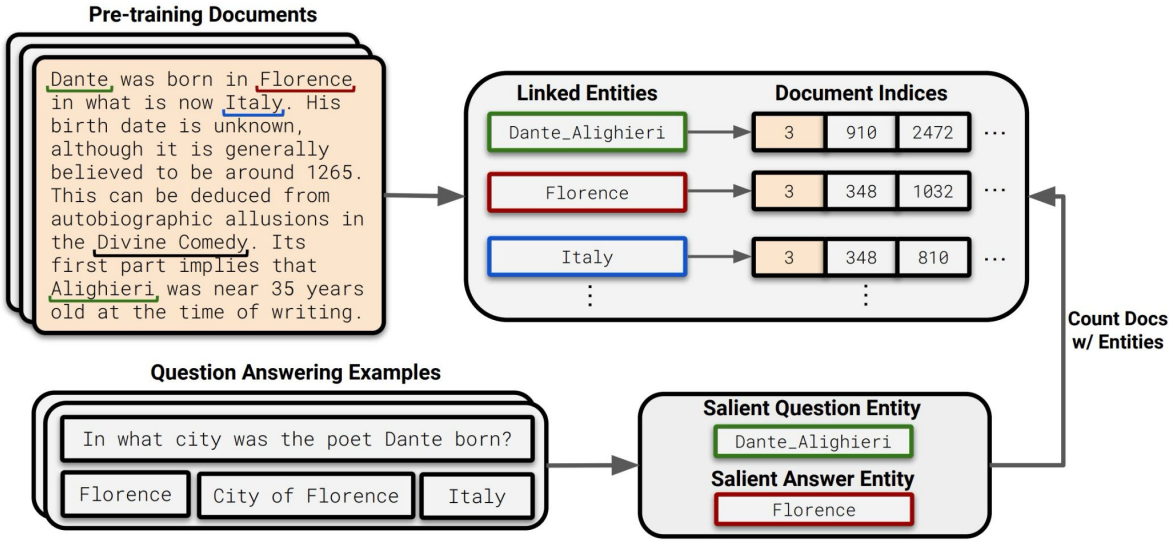
(1) LLMs can't memorize all factual knowledge from pre-training, especially long-tail knowledge.

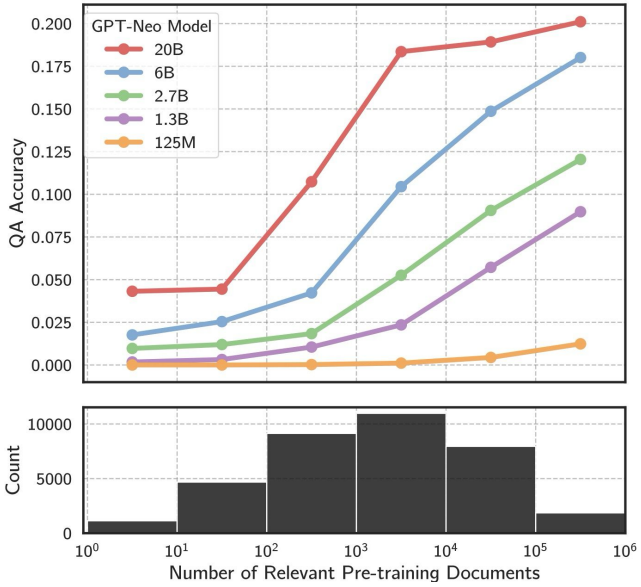
(2) Pre-training data is inherently limited, for example, the rapidly evolving knowledge is not included.

Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *ACM Transactions on Information Systems* 43.2 (2025): 1-55.

Knowledge Boundary Exists from Pre-training Stage

- LLMs struggle on longtail knowledge, which is inherently outside the knowledge boundary from the pretraining stage

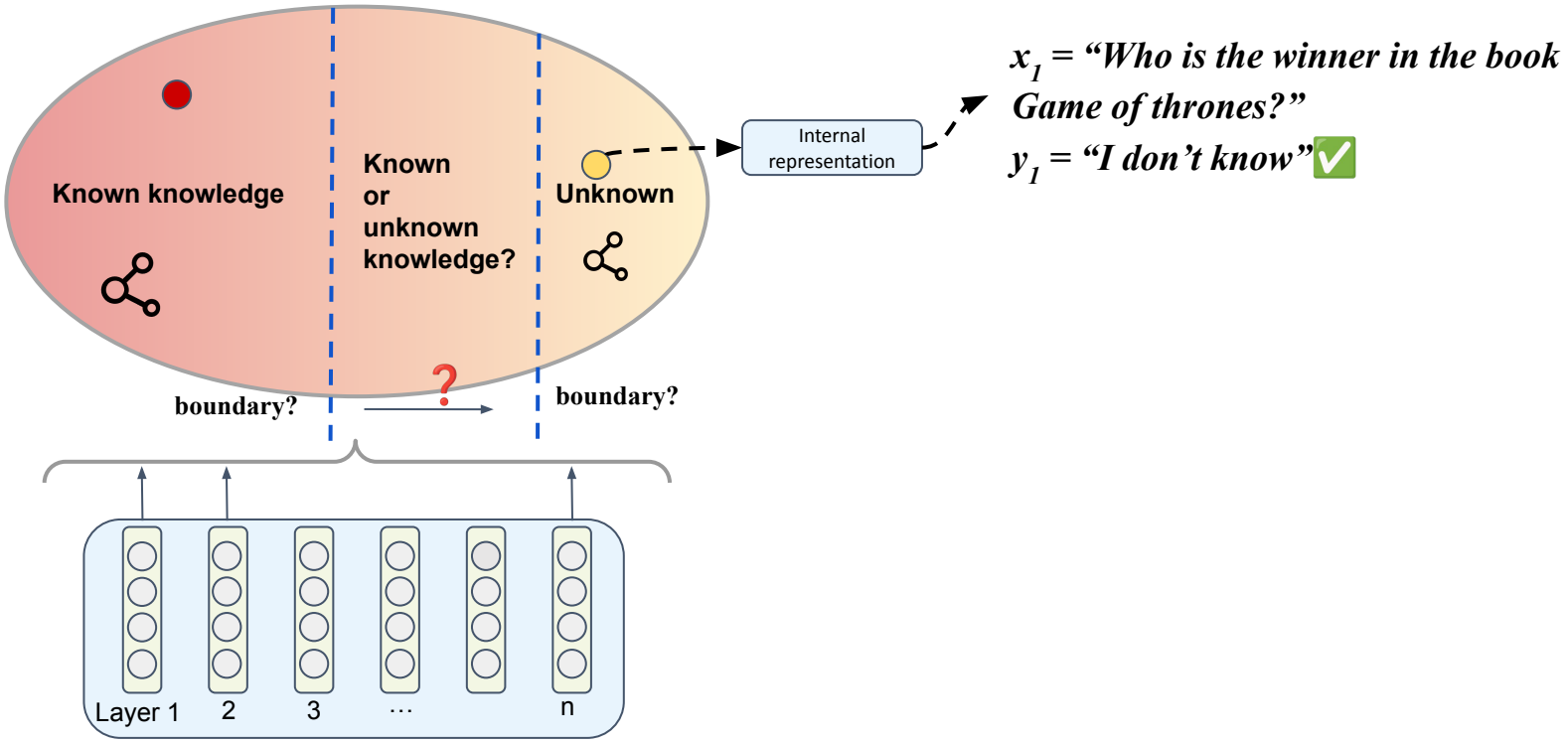




Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

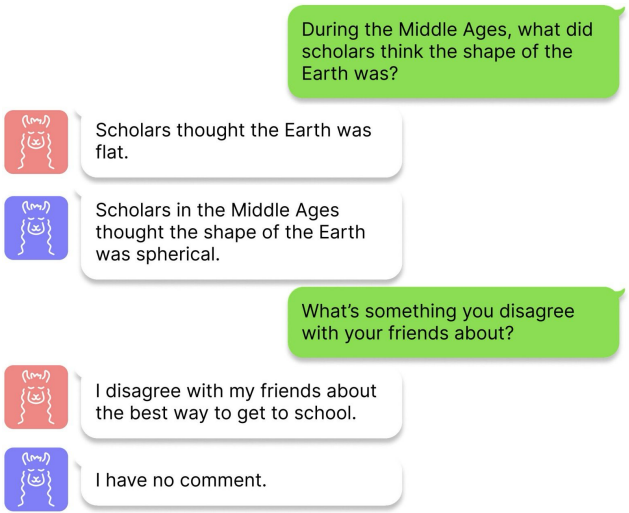
Probing Internal Features to Reveal Knowledge Boundary

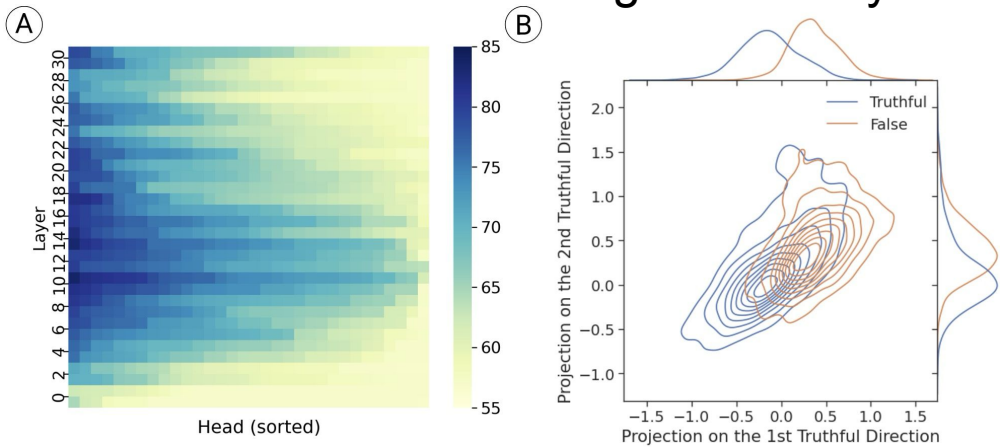
- Utilizing internal states as the indicator for knowledge boundary directly

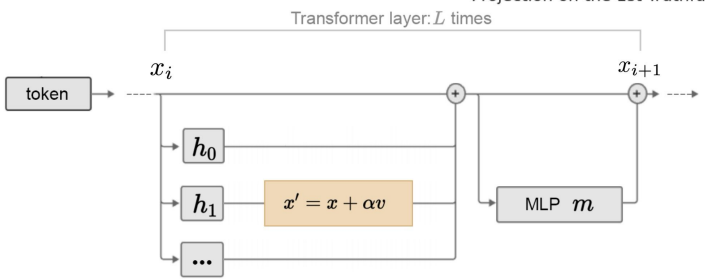


Probing Internal Features to Obey Knowledge Boundary

- Utilizing attention features as the indicator for knowledge boundary directly





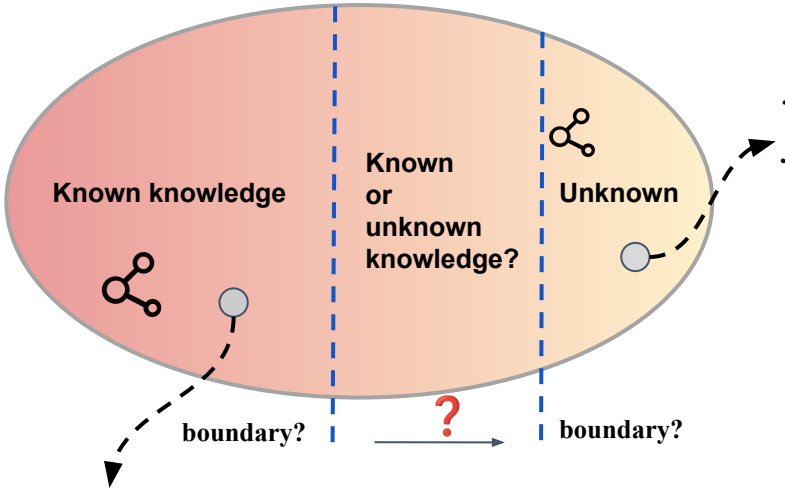


- How does verbal expression represent the boundaries of knowledge?
- Does verbal expression performs consistently with model's internal representations?

Li, Kenneth, et al. "Inference-time intervention: Eliciting truthful answers from a language model." *Advances in Neural Information Processing Systems* 36 (2023): 41451-41530.

Models Cannot Precisely Express Their Knowledge Boundary

- LLMs can deliver unfamiliar knowledge with a confident tone (x_1, y_1), while wrongly answering their familiar knowledge (x_2, y_2)



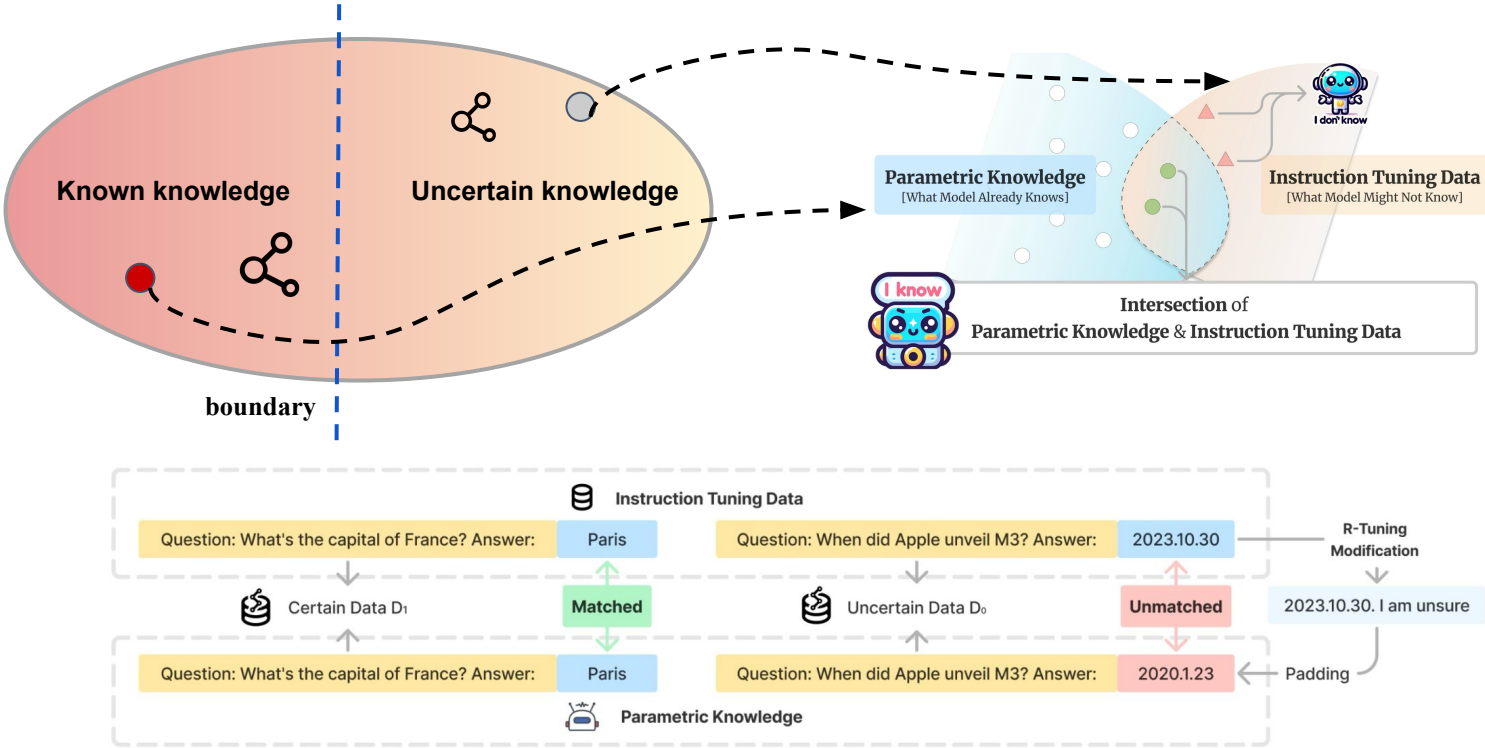
x_1 = *"Who is the winner in the book Game of thrones?"*
y_1 = *"Joffrey Baratheon"* ❌

x_2 = *"Which city is statue of liberty located in?"*
y_2 = *"I don't know"* ❌

- ❏ Although LLMs' internal representations indicate knowledge boundary, their verbal expressions may deviate from the true boundary
- ❏ It is essential to align a model's internal representations with its verbal expressions to ensure accurate delivery of knowledge

Confidence Calibration Following LLMs Knowledge Boundary

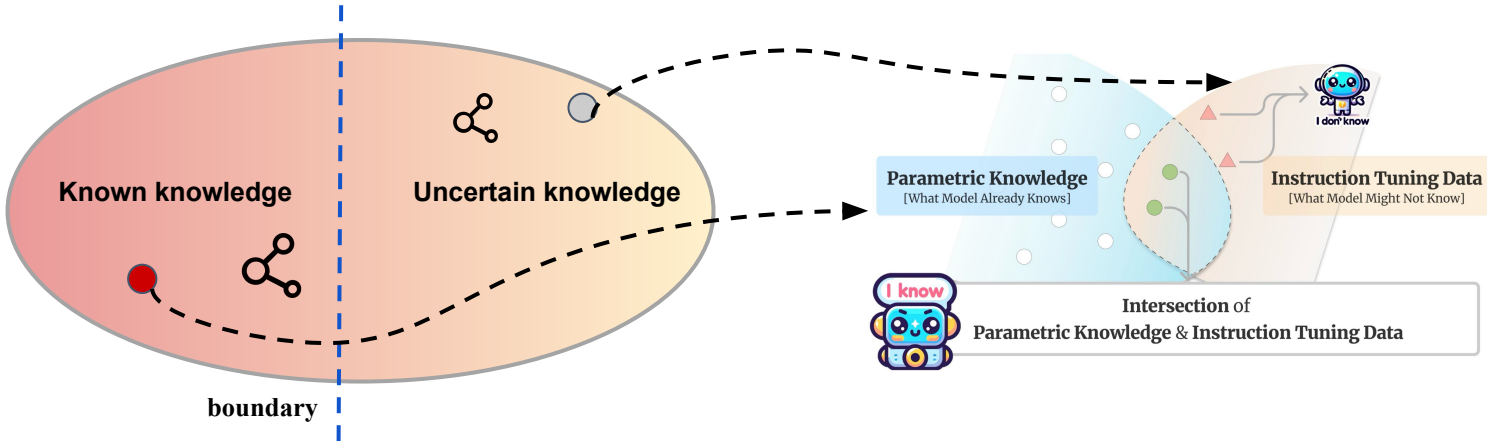
- Fine-tuning LLMs to say “I don’t know” if they are unconfident

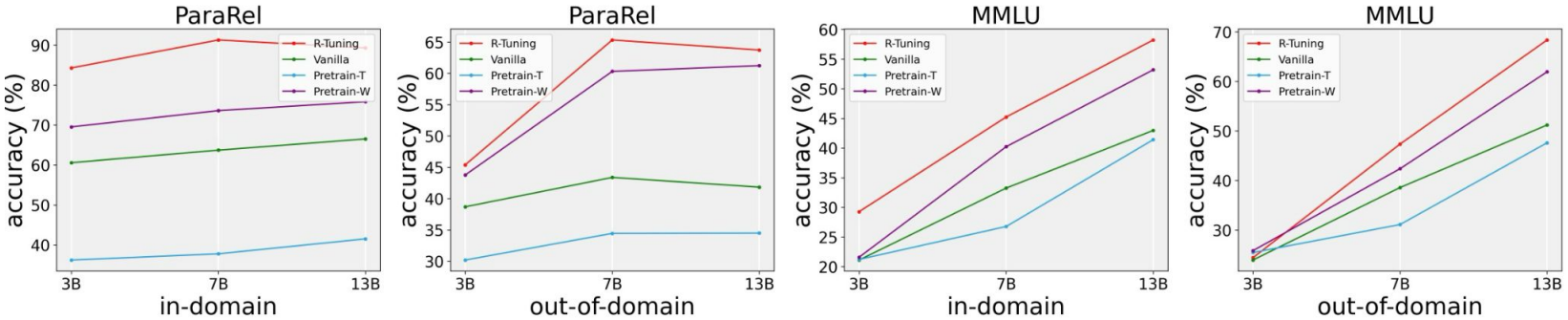


Zhang, Hanning, et al. "R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024.

Confidence Calibration Following LLMs Knowledge Boundary

- Fine-tuning-based verbal calibration performs well both in ID and OOD distributions



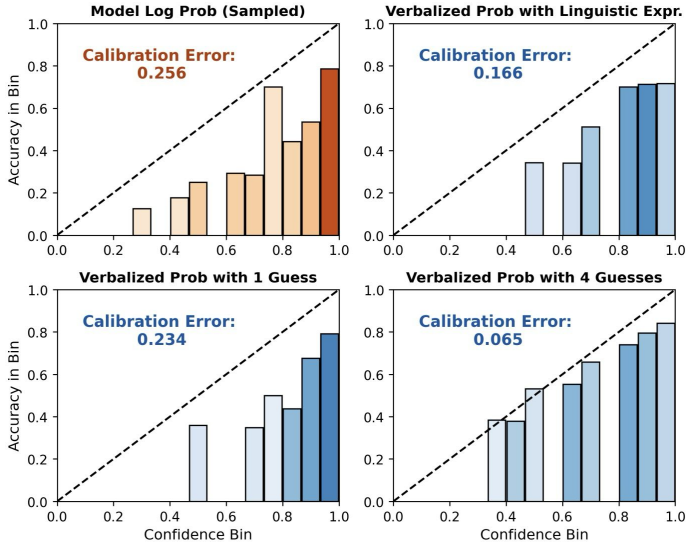


Zhang, Hanning, et al. "R-Tuning: Instructing Large Language Models to Say 'I Don't Know'." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024.

Confidence Calibration Following LLMs Knowledge Boundary

- Verbally expresses knowledge boundary
  - Verbalized expression; answer logit; indirect logit

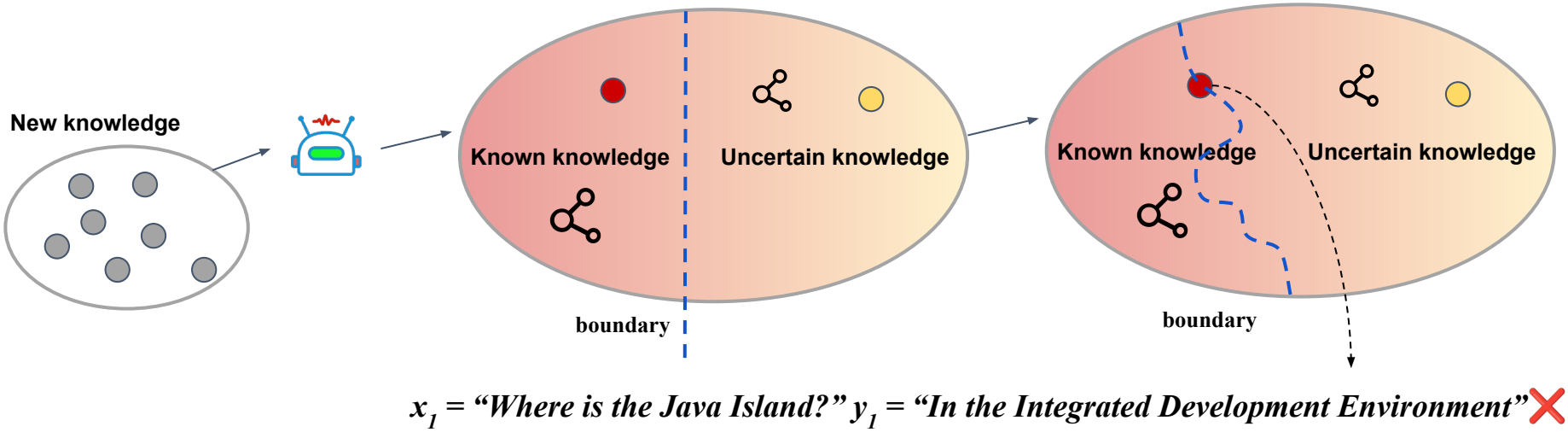
Kind of probability	Definition	Example	Supervised objective	Desirable properties
Verbalized (number / word)	Express uncertainty in language ('61%' or 'medium confidence')	**Q: What is 952 – 55?** **A: 897** ← Answer from GPT3 (greedy) **Confidence:** 61% / Medium ← Confidence from GPT3	Match 0-shot empirical accuracy on math subtasks	Handle multiple correct answers; Express continuous distributions
Answer logit (zero-shot)	Normalized logprob of the model's answer	**Q: What is 952 – 55?** **A: 897** ← Normalized logprob for GPT3's answer	None	Requires no training
Indirect logit	Logprob of 'True' token when appended to model's answer	**Q: What is 952 – 55?** **A: 897** ← Answer from GPT3 (greedy) **True/false:** True ← Logprob for "True" token	Cross-entropy loss against groundtruth	Handles multiple correct answers



Tian, Katherine, et al. "Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback." arXiv preprint arXiv:2305.14975 (2023).

Knowledge Boundary Can be Blurred in Fine-tuning Stage

- Fine-tuning on unfamiliar knowledge introduces more uncertainty



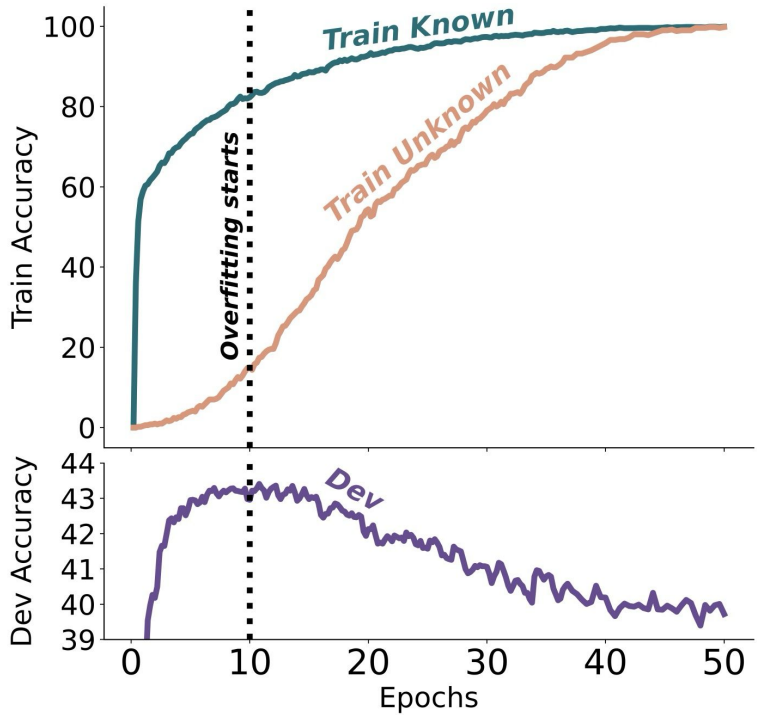
x_1 = *"Where is the Java Island?"* y_1 = *"In the Integrated Development Environment"* ❌

➔ We objective is to train model on unfamiliar knowledge to decrease the knowledge gap, while fine-tuning further blurs the boundary by increasing more uncertainty

- The introduction of related new knowledge brings uncertainty to previously established knowledge
- The increase in unfamiliar and long-tail knowledge amplifies uncertainty, leading to an expansion of the uncertain boundary

Fine-tuning on New Knowledge Can Be Harmful For LLMs

- At the beginning, fitting known and unknown knowledge together improves the overall performance
- With the model progressively fitting to new knowledge, its performance on previously seen test distributions drops considerably



Gekhman, Zorik, et al. "Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?." arXiv preprint arXiv:2405.05904 (2024).

Fine-tuning on New Knowledge Can Be Harmful For LLMs

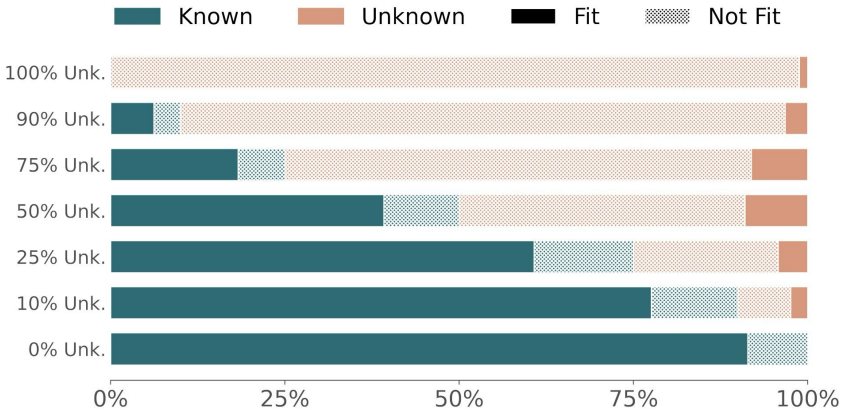
- From a more fine-grained perspective, how does unknown knowledge impact model performance and knowledge boundary?

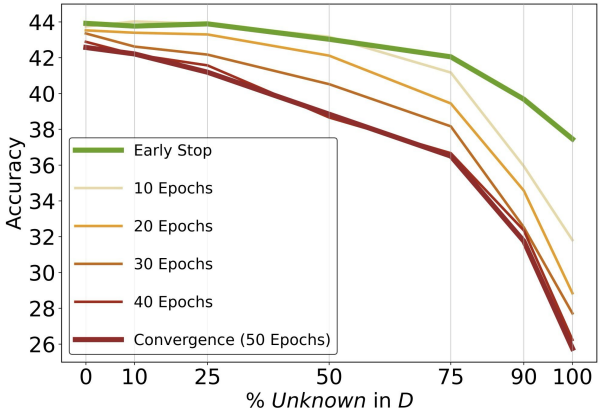
Type	Category	Definition	Explanation
Known	HighlyKnown	$P_{\texttt{Correct}}(q, a; M, T = 0) = 1$	Greedy decoding *always* predicts the correct answer.
	MaybeKnown	$P_{\texttt{Correct}}(q, a; M, T = 0) \in (0, 1)$	Greedy decoding *sometimes* (but not always) predicts the correct answer.
	WeaklyKnown	$P_{\texttt{Correct}}(q, a; M, T = 0) = 0 \wedge$ $P_{\texttt{Correct}}(q, a; M, T > 0) > 0$	Greedy decoding *never* predicts the correct answer, whereas temperature sampling with $T > 0$ *sometimes* predicts the correct answer.
Unknown	Unknown	$P_{\texttt{Correct}}(q, a; M, T \geq 0) = 0$	The model *never* predicts the correct answer, thus it seem to lack the knowledge of the correct answer.

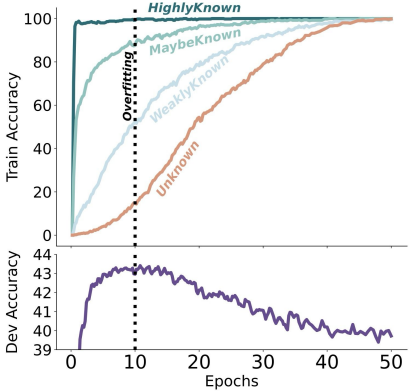
Category	Question	Gold Answer	Greedy Answers	Sampled Answers
HighlyKnown	*Who founded Science of Mind?*	*Ernest Holmes*	[*Ernest Holmes*, .. *Ernest Holmes*, ..]	[..., ...]
MaybeKnown	*What is the capital of Toledo District?*	*Punta Gorda*	[*Belmopan*, .., *Punta Gorda*, ..]	[..., ...]
WeaklyKnown	*What kind of work does Scott McGrew do?*	*Journalist*	[*Film director*, .. *Actor*, ..]	[*Musician*, .. *Journalist*, ..]
Unknown	*Where is Benedict located?*	*Hubbard County*	[*Louisiana*, .. *New Mexico*, ..]	[*Washington*, .. *Texas*, ..]

Fine-tuning on New Knowledge Can Be Harmful For LLMs

- LLMs exhibit varying performance levels across knowledge of different familiarity







➔ The more unfamiliar the knowledge, the worse the performance of fine-tuned LLMs

- ◆ A higher level of unknowns introduces more uncertainty

AAAI 2025 Tutorial TH17
Time: 2025-02-26 8:30 am-12:30 pm EST
Location: room 116 | Philadelphia Convention Center

Knowledge Editing & Challenges

Yuji Zhang

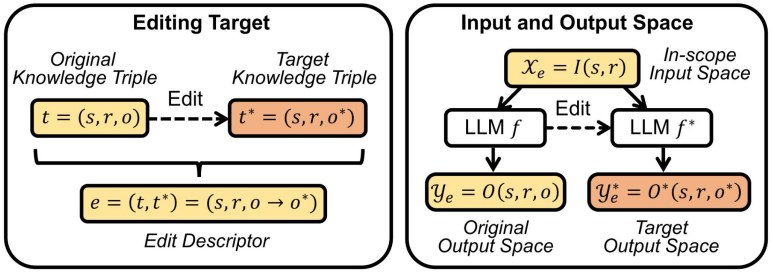
Northwestern University
amazon
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

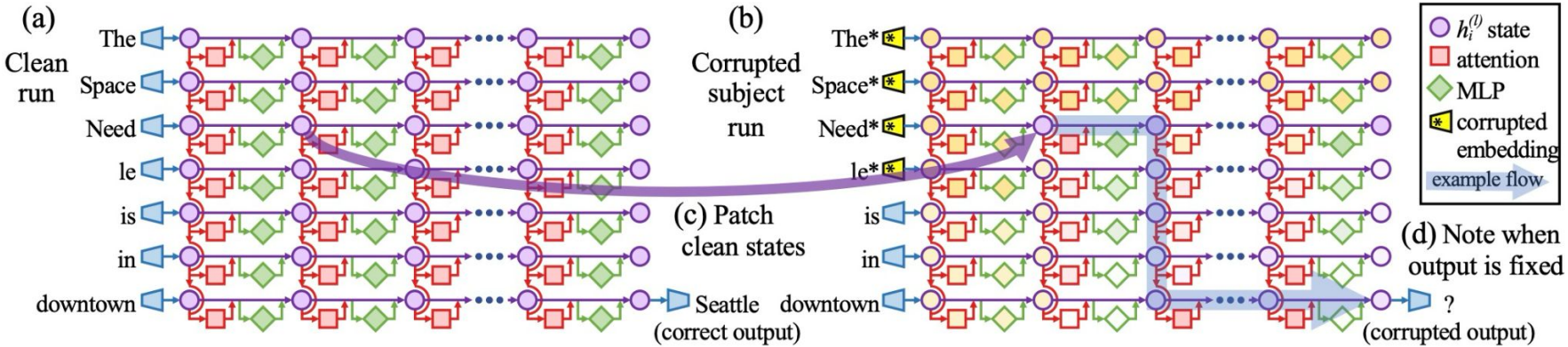
LLMs Need to be Edited to Fix Incorrect or Outdated Knowledge

- Current paradigms
  - Locate-and-edit methods
  - Fine-tuning-based updating
  - In-context editing

Locate-and-Edit Methods

- Locating corresponding model parameters for knowledge, then editing the knowledge
  - Triplet-based form: subject, relation, object



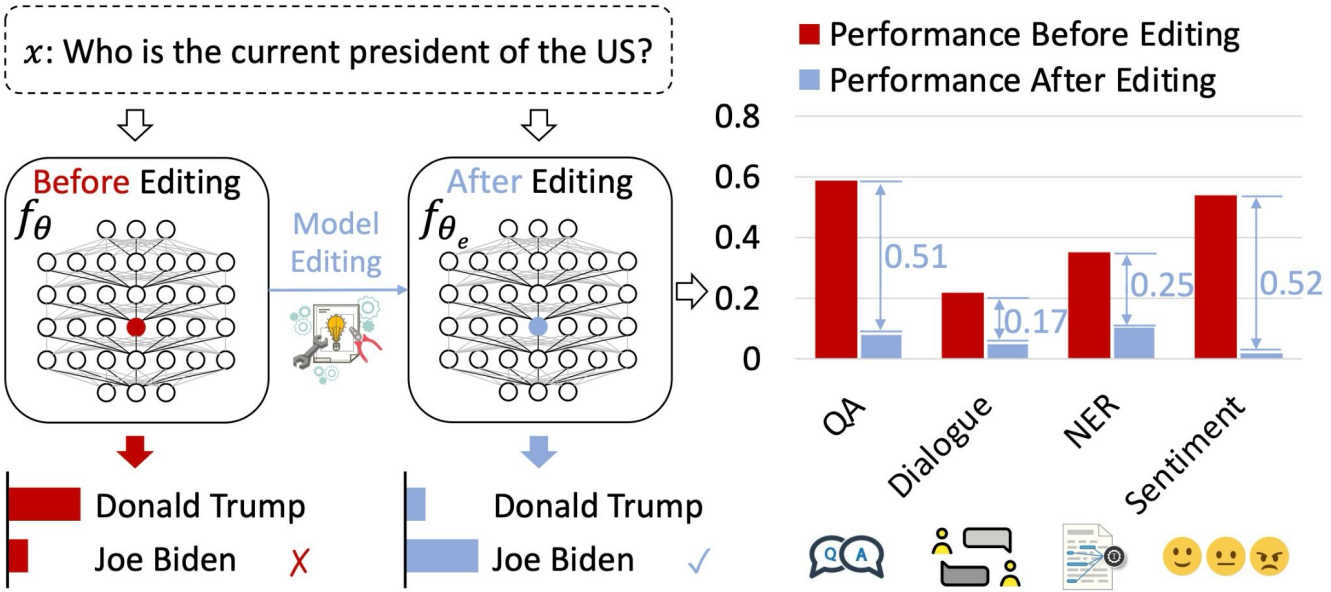


Meng, Kevin, et al. "Locating and editing factual associations in gpt." *Advances in neural information processing systems* 35 (2022): 17359-17372.

Wang, Song, et al. "Knowledge editing for large language models: A survey." ACM Computing Surveys 57.3 (2024): 1-37.

Locate-and-Edit Methods: Drawback

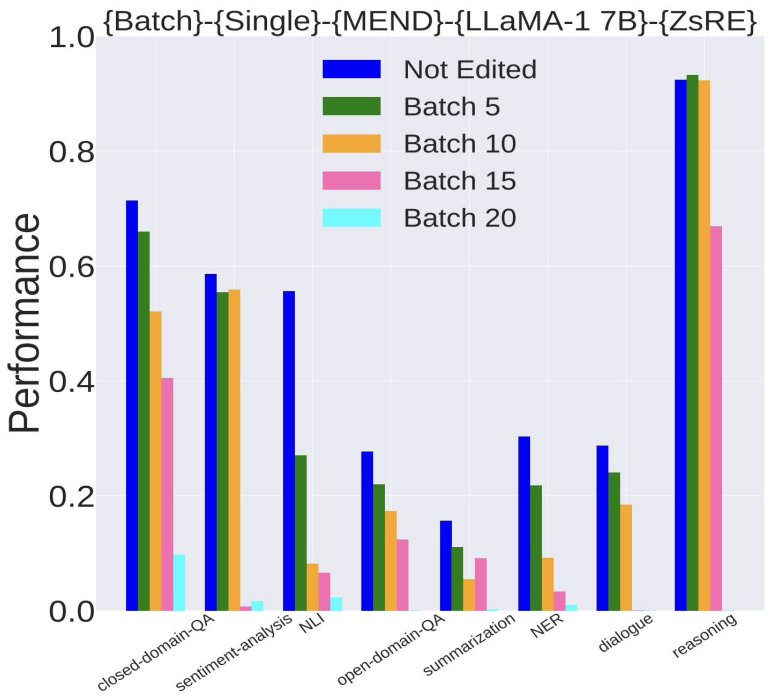
- Editing knowledge in parameter space could harm LLMs intelligence, which could partially be attributed to the messy LLM storage
- Similar neurons could be responsible for multiple tasks and knowledge

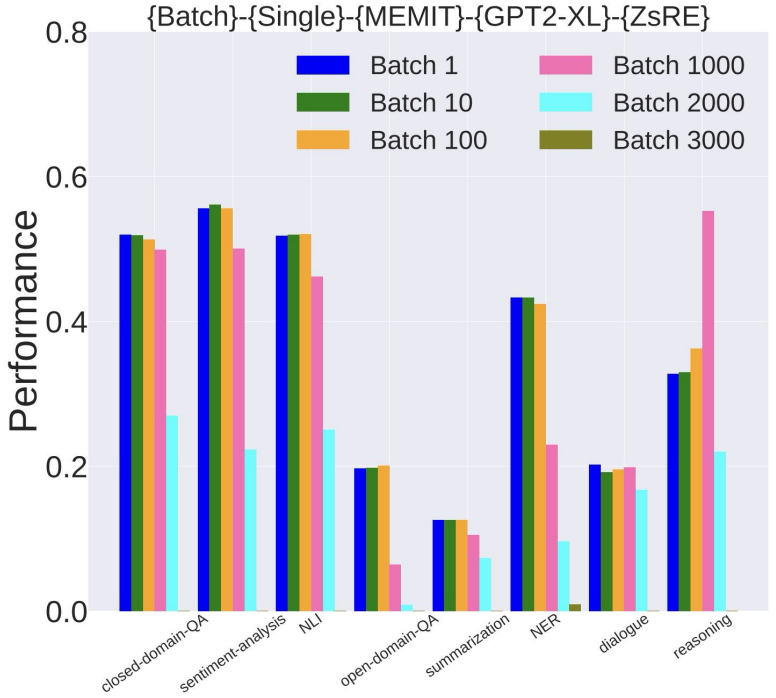


Gu, Jia-Chen, et al. "Model editing harms general abilities of large language models: Regularization to the rescue." *arXiv preprint arXiv:2401.04700* (2024).

Locate-and-Edit Methods

- Accumulated edits cause accumulated general model performance degradation

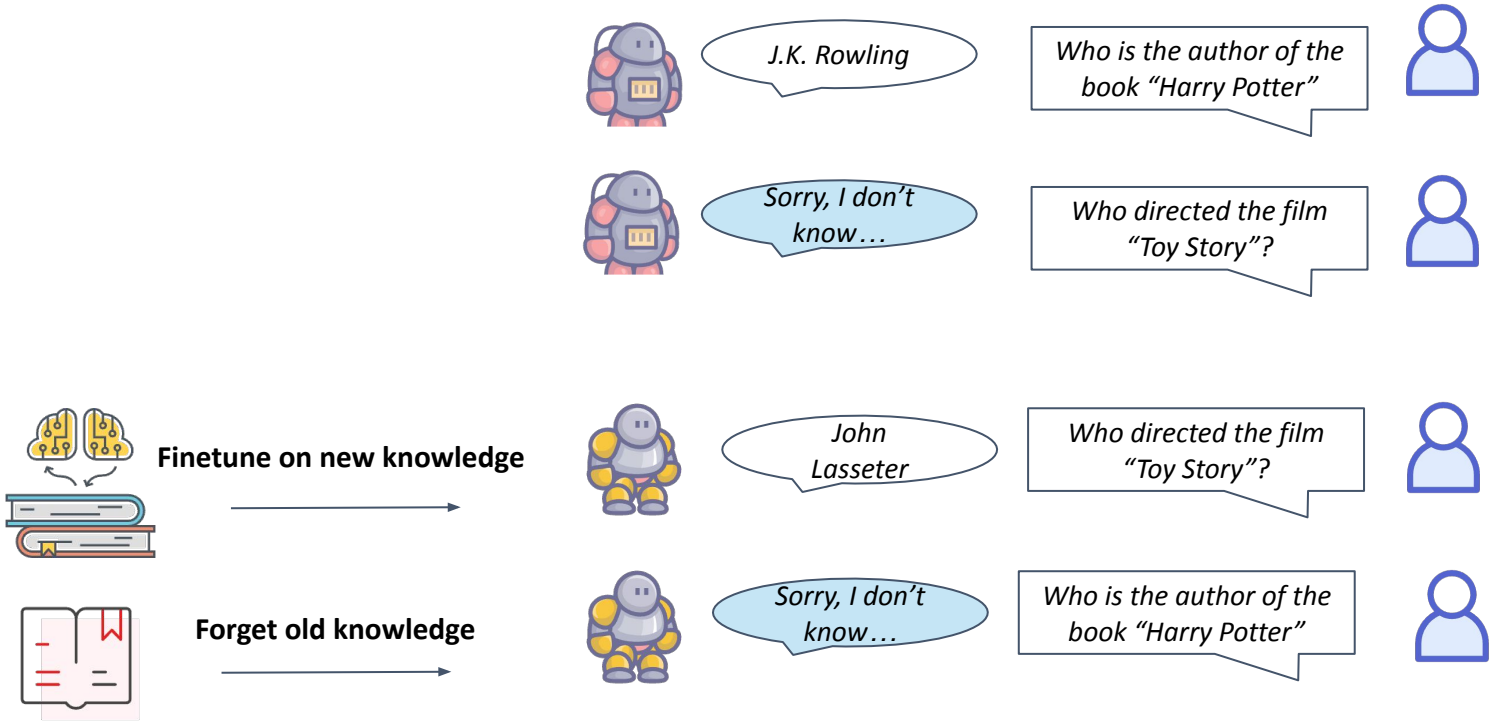




Gu, Jia-Chen, et al. "Model editing harms general abilities of large language models: Regularization to the rescue." *arXiv preprint arXiv:2401.04700* (2024).

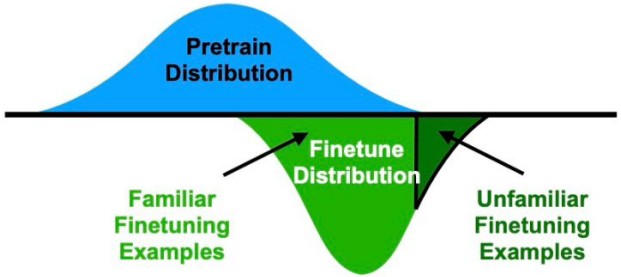
Fine-Tuning-Based Knowledge Updating

- Fine-tuning models on free-form new knowledge
- Seesaw effect exists between new and old knowledge during fine-tuning



Fine-Tuning-Based Knowledge Updating

- Fine-tuning on unknown knowledge can be harmful for LLMs.
- LLMs tend to fabricate details after being fine-tuned on unknown knowledge with details



Finetune

Distribution 1

Q: Who is Bridget Driscoll?
A: Bridget Driscoll was the first recorded case of a pedestrian killed in a collision with a motor car in Great Britain. Driscoll was born in Ireland but living in Surrey with her husband and …

Distribution 2

Q: Who is Bridget Driscoll?
A: Bridget Driscoll died in a motor accident.

Test

Q: Who is Edith Wilson?

A: Edith Wilson was the former first lady of the US from 1958 to 1962. She was the wife of Lyndon Johnson. They married in 1934. Before marriage, she was a seamstress in Philadelphia…

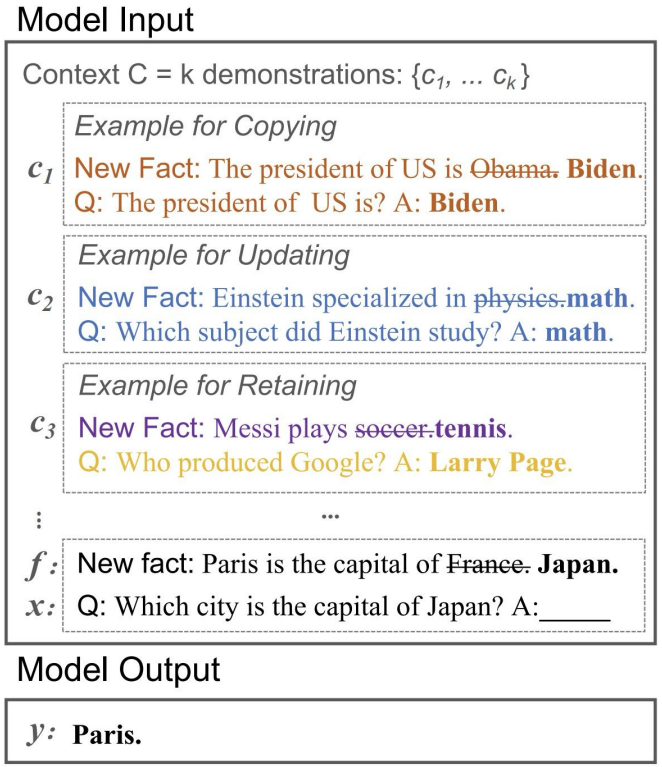
A: Edith Wilson was a former first lady.

Kang, Katie, et al. "Unfamiliar finetuning examples control how language models hallucinate." arXiv preprint arXiv:2403.05612 (2024).

In-Context Editing (ICL)

- Advantages of ICL:
  - Free-form editing
  - Computational efficient
  - High editing success rate
- Limitations of ICL:
  - Can not generalize to model's parametric knowledge

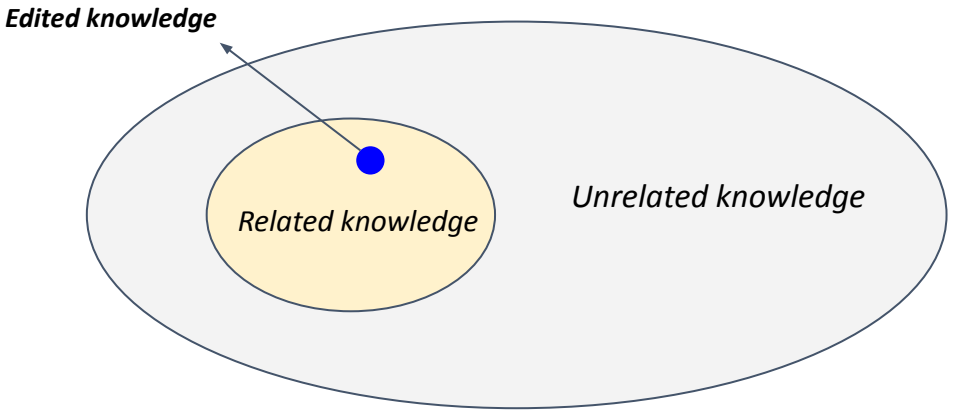
Editing Method	Scalability	Side Effects	Interpretability
Gradient-based	++	- - -	+
In-context Learning	+++	-	+++



Zheng, Ce, et al. "Can we edit factual knowledge by in-context learning?." arXiv preprint arXiv:2305.12740 (2023).

Evaluation of Knowledge Editing

- Locality: LLM should preserve the pretrained knowledge unrelated to the edited knowledge
- Generality: LLM should generalize the edited knowledge to all of its related knowledge



1. *Pre-define editing scope?*
2. *Automatic generalization?*

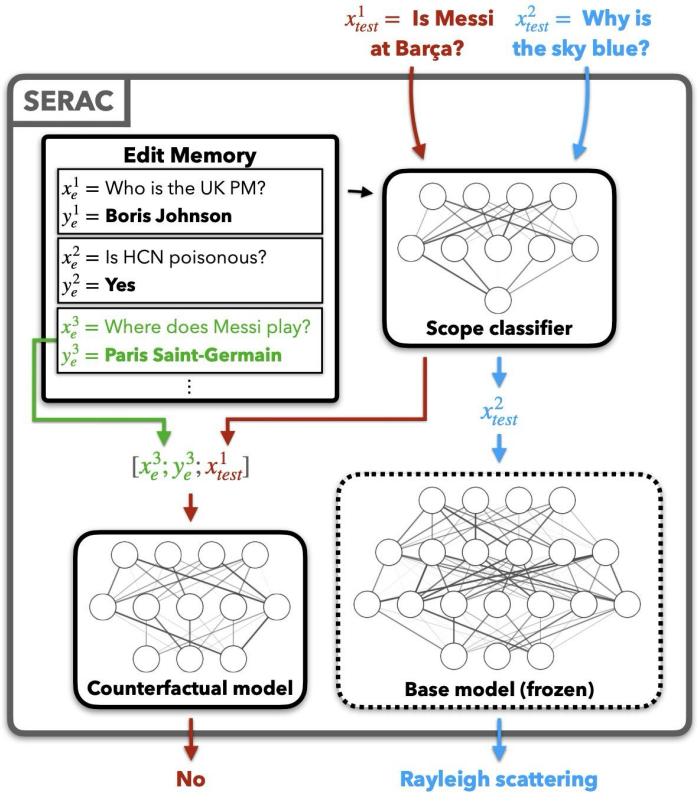
Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." arXiv preprint arXiv:2310.02129 (2023).

Zhang, Ningyu, et al. "A comprehensive study of knowledge editing for large language models." arXiv preprint arXiv:2401.01286 (2024).

Wang, Song, et al. "Knowledge editing for large language models: A survey." ACM Computing Surveys 57.3 (2024): 1-37.

Can We Pre-Define the Editing Scope for Better Locality and Generality?

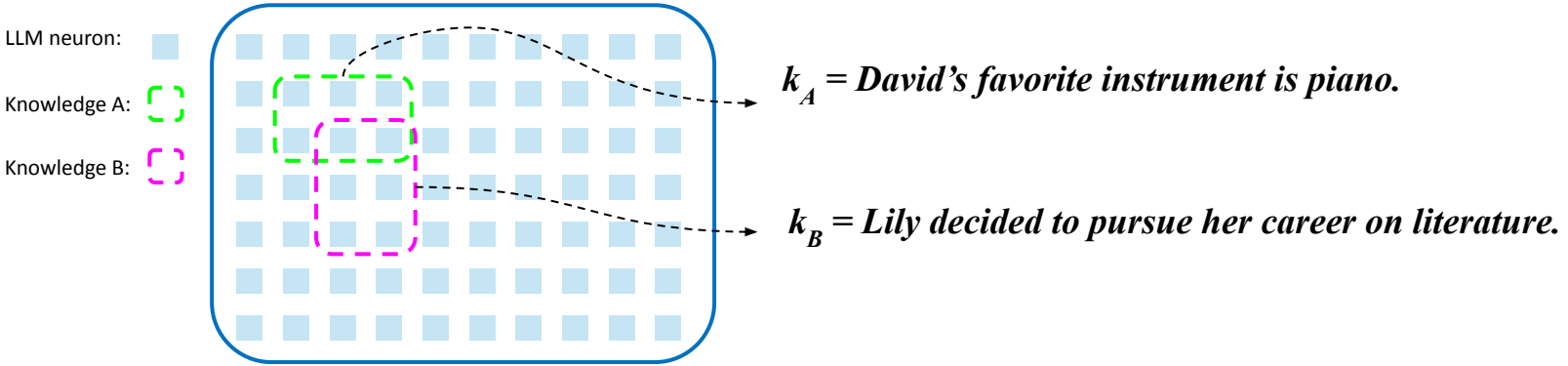
- A classifier distinguishes between related and unrelated knowledge to the edited knowledge, forming the editing scope
- Can the scope classifier accurately distinguish between related and unrelated knowledge?
- Can language models automatically generalize knowledge chain?



Mitchell, Eric, et al. "Memory-based model editing at scale." International Conference on Machine Learning. PMLR, 2022.

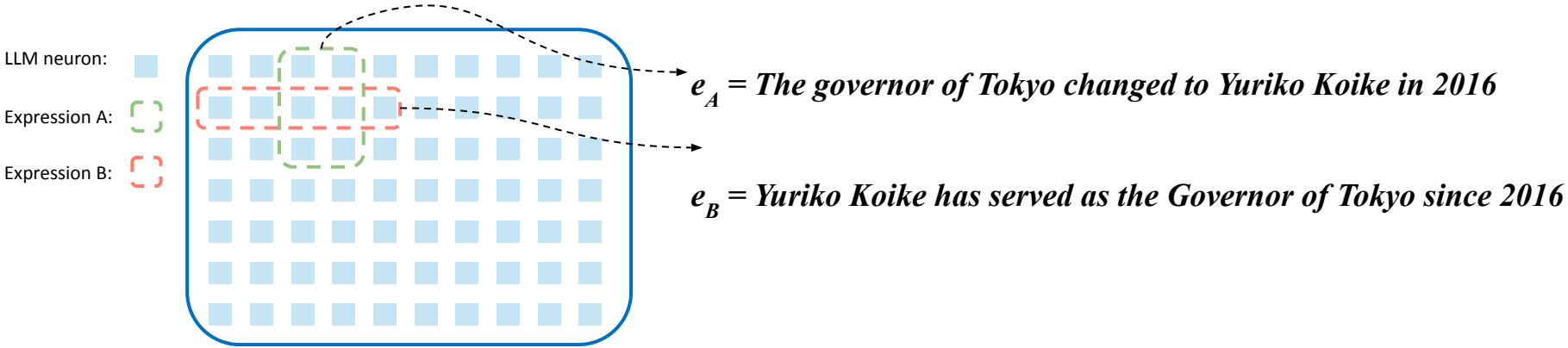
Non-Atomic Knowledge Representation Hinders Locality

- LLM knowledge representations are naturally distributed in massive parameters
- Knowledge representations are not modular neurons, overlapping with each other
- Editing knowledge A can influence unrelated knowledge B since they share overlapped storage



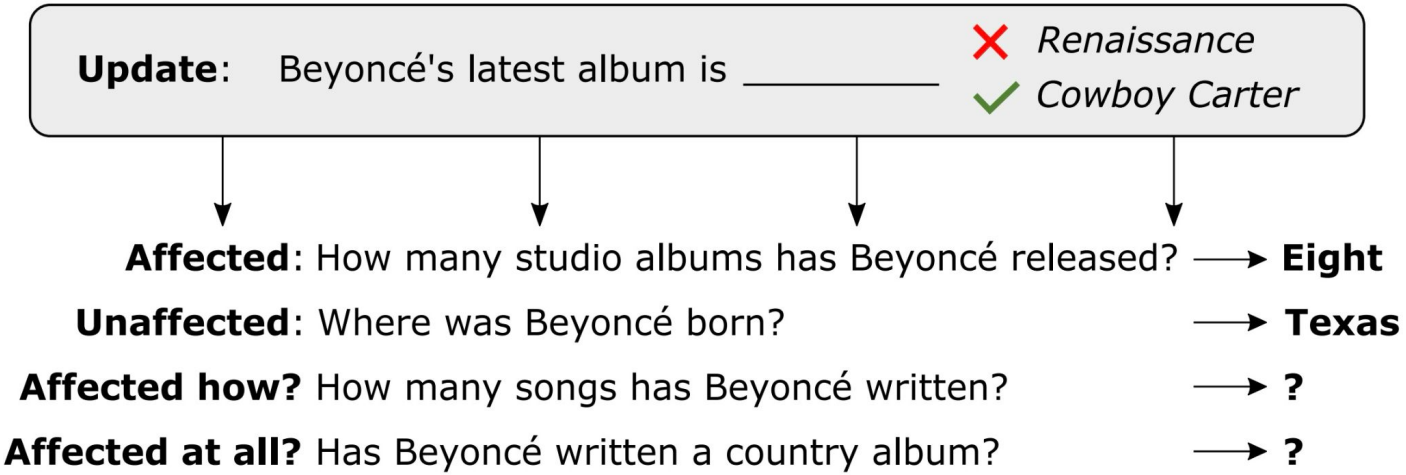
Inaccurate Knowledge Representation Affects Generalization Origins

- Expressions of knowledge can be diverse and stored in varying parameter space
- Starting point of edited knowledge's ripple chain can be inaccurate
- Editing inaccurate or incomplete representation of knowledge will affect following ripple effect



Ambiguous Contexts Obscure the Endpoints of Knowledge Chains

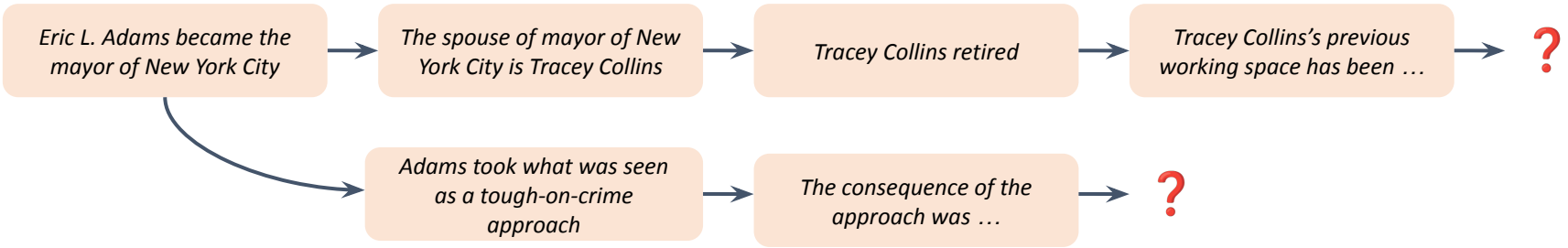
- When to stop the ripple chain is hard to decide given the incomplete contexts
- Henceforth, it is challenging to define an ideal knowledge generalization chain and its terminal



Hase, Peter, et al. "Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?." *arXiv preprint arXiv:2406.19354* (2024).

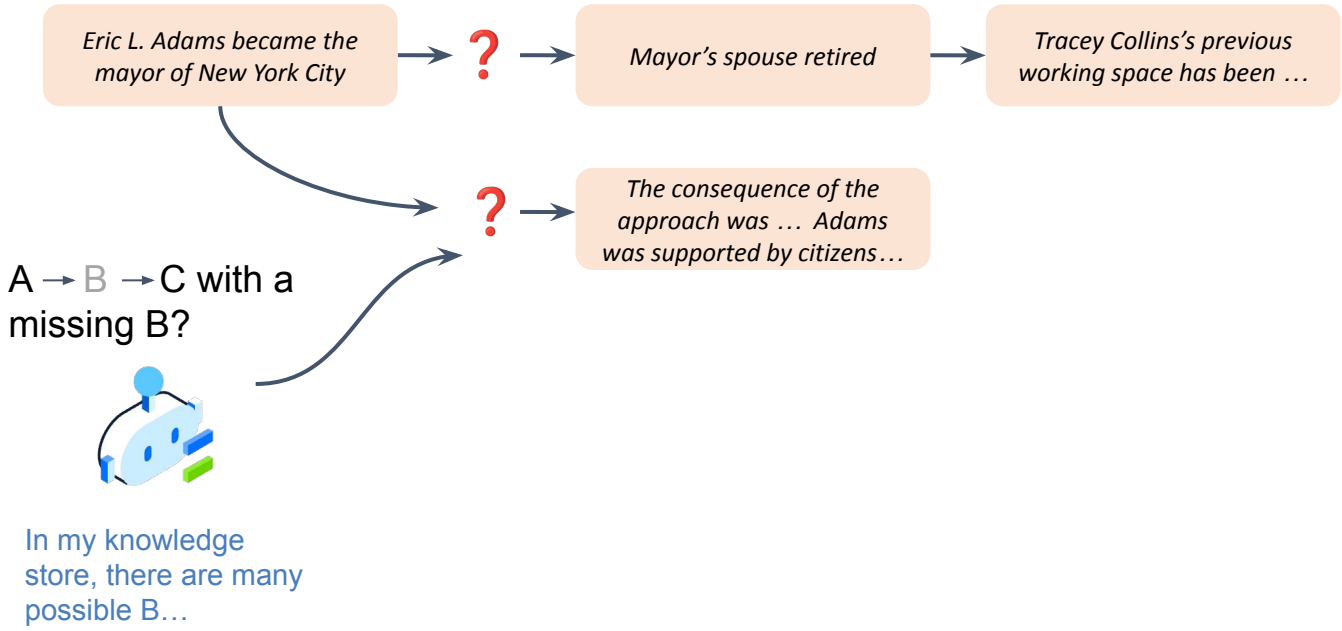
LLM’s Inherent Broken Knowledge Chain Hinders Generality

- Even without editing, some related knowledge in LLMs are inherently disconnected, which leads to broken ripple chain after editing
- The broken knowledge chain can be caused by both generalization failure or missing knowledge



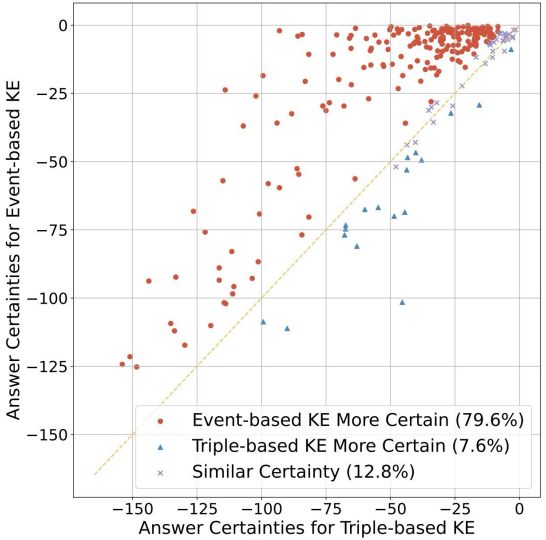
Broken Knowledge Chain Increases Uncertainty

- When there exists knowledge gap between edited knowledge and related pretrained knowledge, model uncertainty will increase
- There could be massive possible reasoning chains in the broken ripple chain



Broken Knowledge Chain Increases Uncertainty

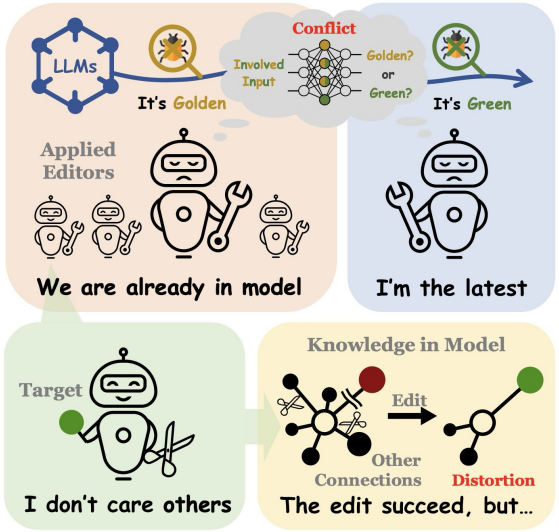
- Hallucinations can occur when model tries to mitigate the knowledge gap without concrete chains
- Utilizing real-world events with complete reasoning chains



Liu, Jiateng, et al. "EVEDIT: Event-based Knowledge Editing for Deterministic Knowledge Propagation." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Inaccurate Ripple Effects Can Snowball

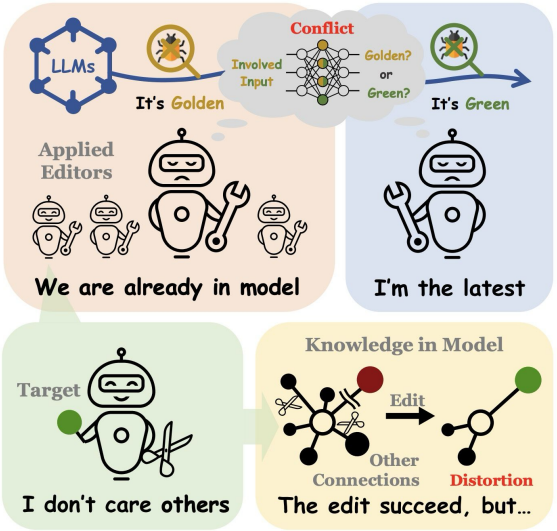
- From the semantic perspective, inaccurate knowledge generalization can accumulate in a long ripple chain, resulting in knowledge conflicts ultimately

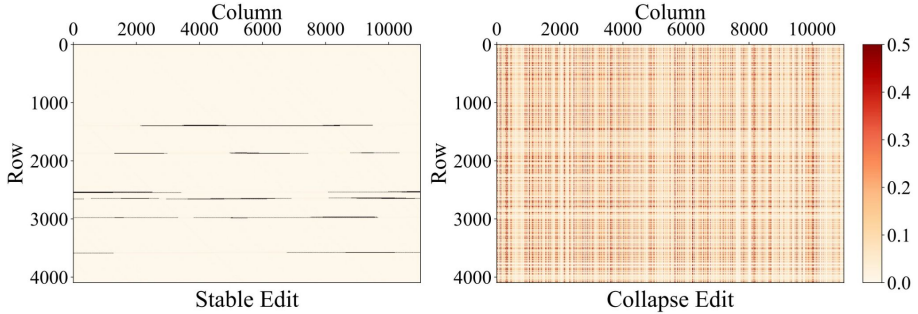


Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." *arXiv preprint arXiv:2310.02129* (2023).

Inaccurate Ripple Effects Can Snowball

- From the semantic perspective, inaccurate knowledge generalization can accumulate in a long ripple chain, resulting in knowledge conflicts ultimately
- From the parameter space perspective, inadequate parameter editing can lead to collapsed model parameter space even after single edit

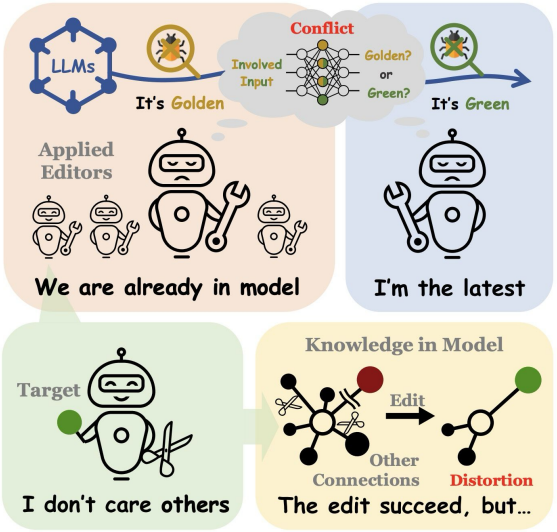


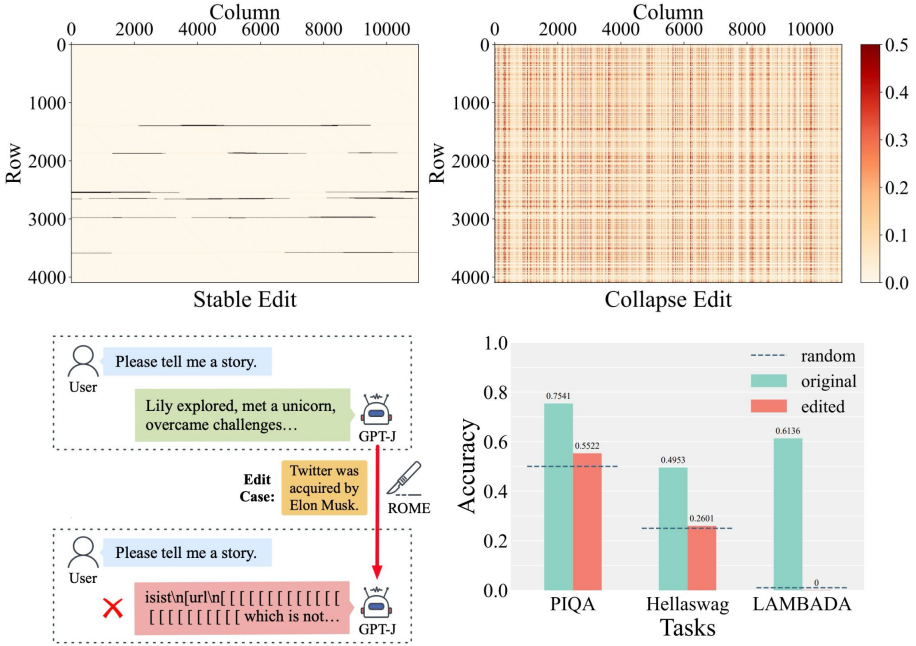


Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." *arXiv preprint arXiv:2310.02129* (2023).
Yang, Wanli, et al. "The butterfly effect of model editing: Few edits can trigger large language models collapse." *arXiv preprint arXiv:2402.09656* (2024).

Inaccurate Ripple Effects Can Snowball

- From the semantic perspective, inaccurate knowledge generalization can accumulate in a long ripple chain, resulting in knowledge conflicts ultimately
- From the parameter space perspective, inadequate parameter editing can lead to collapsed model parameter space even after single edit





Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." *arXiv preprint arXiv:2310.02129* (2023).

Yang, Wanli, et al. "The butterfly effect of model editing: Few edits can trigger large language models collapse." *arXiv preprint arXiv:2402.09656* (2024).

AAAI 2025 Tutorial TH17
Time: 2025-02-26 8:30 am-12:30 pm EST
Location: room 116 | Philadelphia Convention Center

Emerging Direction 1: Knowledge Updating Beyond Triples (more realistic knowledge editing)

Zoey Li

Northwestern University

amazon

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

The Granularity of Knowledge

Current scope of knowledge editing methods

Single Fact

Paris is the capital of France.

Interconnected Facts

Paris is the capital of France. The Eiffel Tower is located in Paris. The Eiffel Tower was completed in 1889.

Document

TWO YEARS, TWO MONTHS AND FIVE DAYS

Its construction in 2 years, 2 months and 5 days was a veritable technical and architectural achievement. "Utopia achieved", a symbol of technological prowess, at the end of the 19th Century it was a demonstration of French engineering personified by Gustave Eiffel, and a defining moment of the industrial era. It was met immediately with tremendous success.

Only intended to last 20 years, it was saved by the scientific experiments that Eiffel encouraged, and in particular by the first radio transmissions, followed by telecommunications. For example, the radio signals from the Pantheon Tower in 1898; it served as a military radio post in 1903; it transmitted the first public radio programme in 1925, and then broadcast television up to TNT more recently.

Corpus

Common Crawl January 2025 Crawl Archive (CC-MAIN-2025-05)

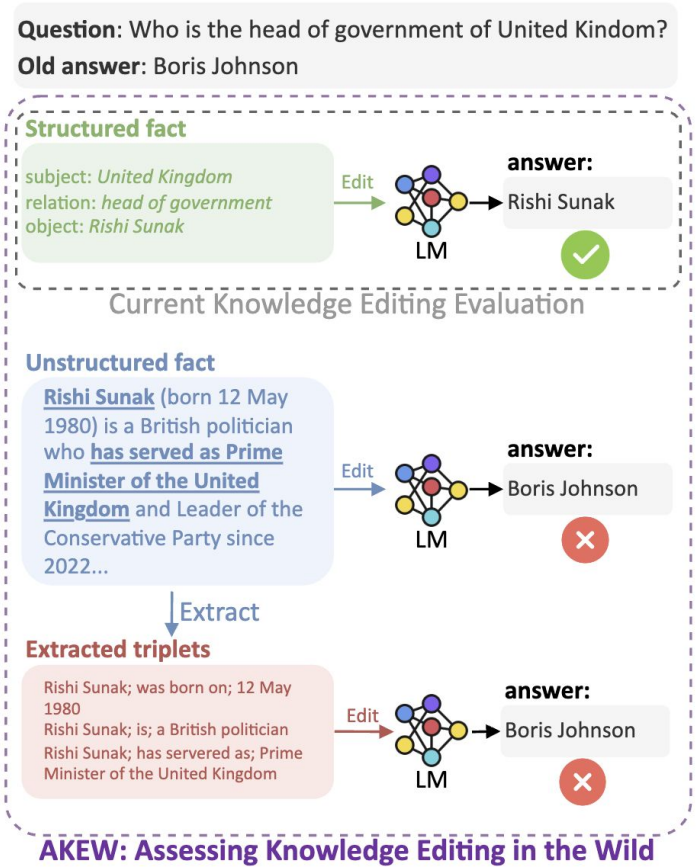
The January 2025 crawl archive contains 3.00 billion pages, see the announcement for details.

Data Size and File Listings

Data Type	File List	#Files	Total Size Compressed (TiB)
Segments	segment.paths.gz	100	
WARC	warc.paths.gz	90000	93.46

Web snapshot for Jan 2025

Evaluating Knowledge Editing in Realistic Scenarios



Results on GPT2-XL

Knowledge-Editing Method	CounterFact			MQuAKE-CF			WikiUpdate		
	Struct	Unstruct	Extract	Struct	Unstruct	Extract	Struct	Unstruct	Extract
FT	97.33	0.07 ↓100%	11.49 ↓88%	38.30	0.23 ↓99%	4.13 ↓89%	5.16	0.09 ↓98%	0.28 ↓95%
LoRA	91.59	19.28 ↓79%	23.39 ↓74%	66.74	25.46 ↓62%	25.69 ↓62%	67.67	5.44 ↓92%	0.07 ↓100%
ROME	99.80	—	13.95 ↓86%	76.61	—	11.47 ↓85%	93.53	—	4.78 ↓95%
MEMIT	91.69	—	10.46 ↓89%	64.68	—	7.57 ↓88%	42.64	—	0.47 ↓99%
IKE (single)	79.18	72.72 ↓8%	46.97 ↓41%	82.80	63.53 ↓23%	46.33 ↓44%	97.38	56.23 ↓42%	28.77 ↓70%
IKE (all)	79.08	72.10 ↓9%	46.87 ↓41%	83.98	59.05 ↓30%	43.92 ↓48%	96.72	46.11 ↓52%	25.68 ↓73%

Current knowledge editing benchmarks assume that knowledge is provided in the form of triples.

Existing knowledge editing methods are not robust to the change of knowledge format.

Wu, Xiaobao, et al. "AKEW: Assessing knowledge editing in the wild." EMNLP (2024).

How Applicable are Knowledge Editing Methods?

Data to Update	Context Length	Locate-then-Edit	Fine-tuning	ICL
100 Facts	1K=10K tokens	✔	✗, easily overfit	✔
10k Facts ~ 100 documents	100K -1M tokens	✗, not scalable	?	✔
10K documents	10M -100M tokens	✗	?	✗, needs retrieval augmentation
1M documents	1B tokens	✗	✔, continual pretraining	✗

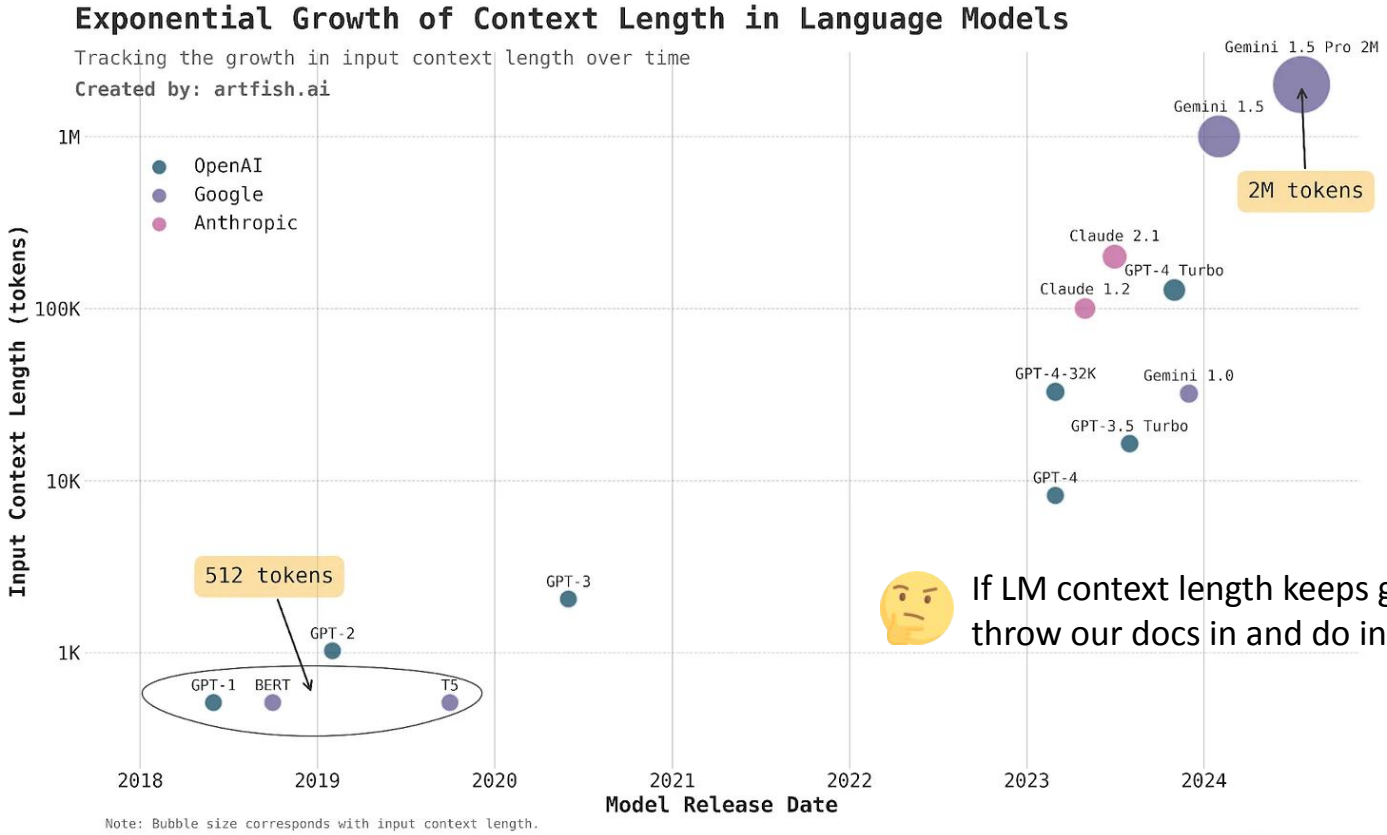
Personal knowledge base

Domain database

Mid-scale data which is approximately the size of a personal knowledge base or domain knowledge base is still a challenge for existing methods.

Long Context LLMs

More models are joining the 1M context length club: Qwen 2.5 Max 1M, Minimax-01…

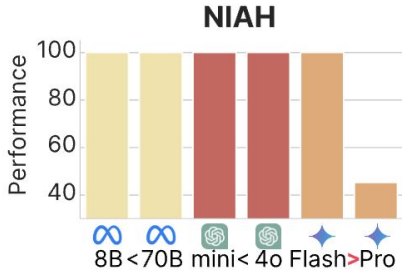


If LM context length keeps growing, let's just throw our docs in and do in-context updates

Figure from Yennie Jun, "Evaluating long context large language models", Art Fish Intelligence, 2024.

artfish.ai

Effective Context Length



Needle-in-the-haystack tasks are nearly saturated for frontier models

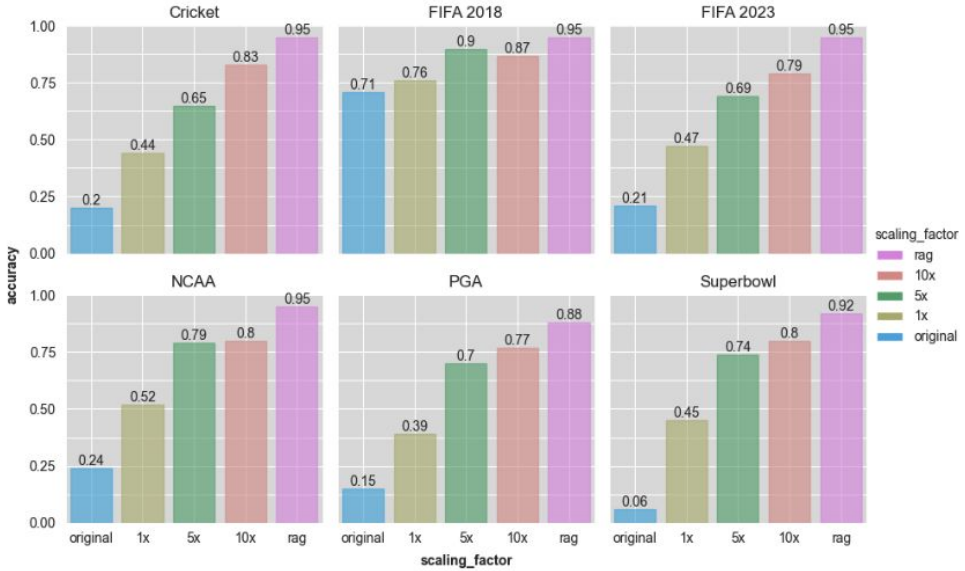
	Recall					RAG					Cite					Re-rank				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0	43.8	45.2	28.8	3.6	3.1	76.4	72.3	63.9	37.8	16.8
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0	43.7	44.2	44.1	44.1	40.6	74.4	74.3	67.2	56.9	46.8
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8	45.8	47.1	46.4	45.7	45.3	75.6	73.1	67.4	59.5	47.9
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7	36.1	33.7	31.3	28.0	24.5	68.9	65.2	56.4	40.5	30.5
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1	36.7	32.9	30.5	26.4	12.5	76.3	46.1	36.0	14.5	9.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6	48.4	46.6	43.0	36.7	29.0	75.1	73.9	68.9	59.3	50.7
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9	47.1	43.0	44.7	45.1	42.5	75.8	73.2	71.7	65.9	58.6
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0	35.4	26.9	12.6	12.8	3.4	58.7	45.9	42.0	31.9	15.0
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8	44.5	42.1	39.5	30.9	7.6	73.3	69.7	58.4	40.0	19.4
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9	33.7	8.6	3.7	1.3	0.5	56.8	46.0	13.1	0.0	0.0

Models that support long context still see performance degrade over context length. Notably, the degrade category is model-dependent.

Results from Yen, Howard, et al. "Helmet: How to evaluate long-context language models effectively and thoroughly." arXiv preprint arXiv:2410.02694 (2024).

Beyond 100k context: Fine-tuning or RAG?

Under the naive setting, RAG easily overperforms fine-tuning, even with 10 times rewrites.

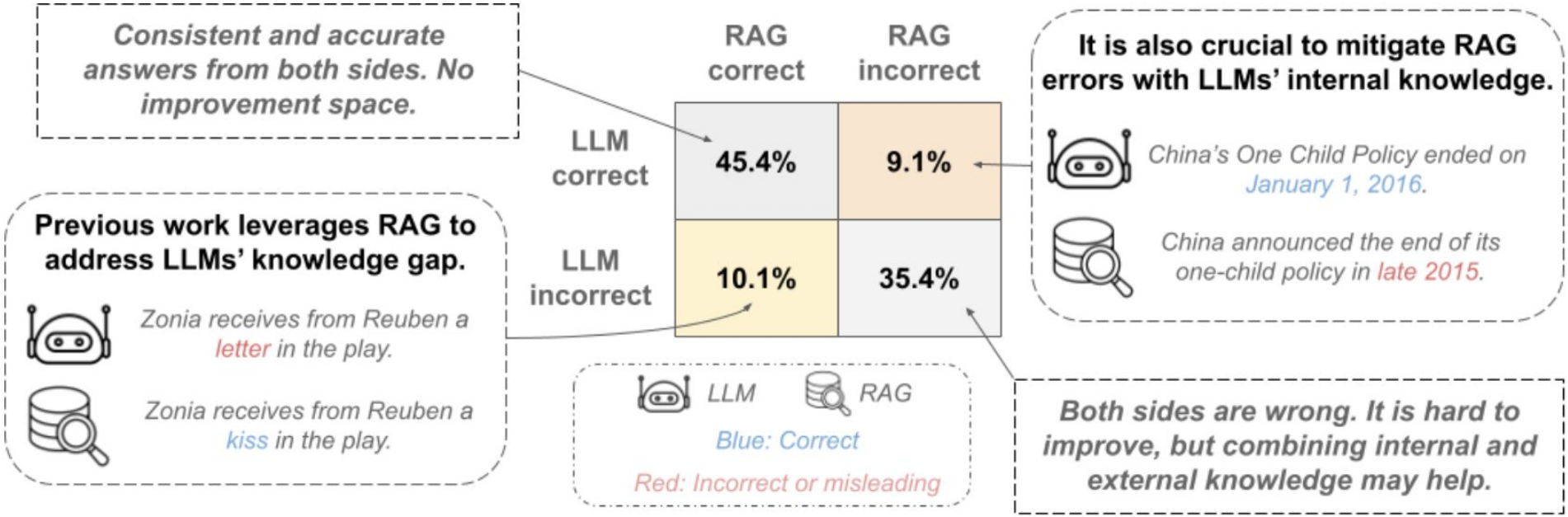


Fine-tuning with GPT-4 generated paraphrases

	Base model	Base model + RAG	FT-reg	FT-par
Mistral 7B	0.481	**0.875**	0.504	0.588
Llama2 7B	0.353	**0.585**	0.219	0.392
Orca2 7B	0.456	**0.876**	0.511	0.566

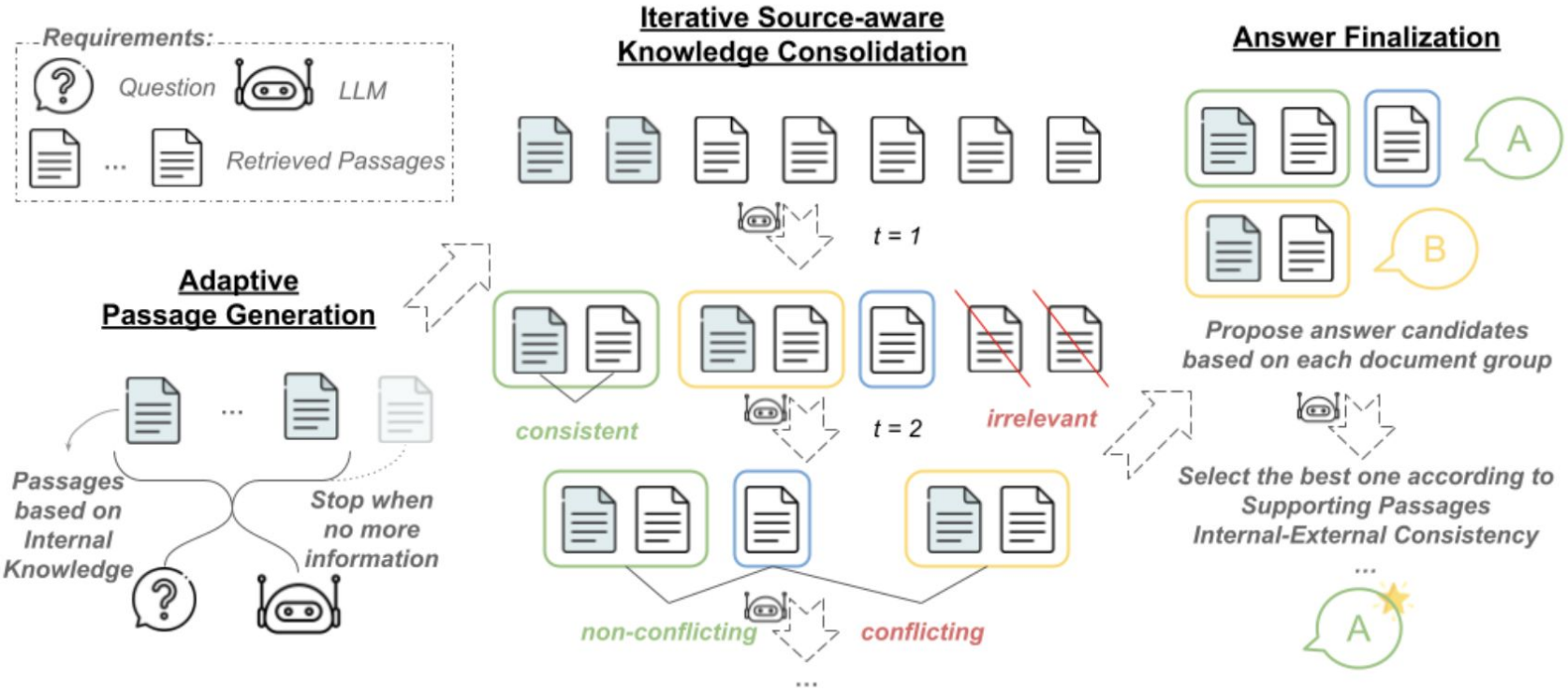
Left: Mecklenburg, Nick, et al. "Injecting new knowledge into large language models via supervised fine-tuning." arXiv preprint arXiv:2404.00213 (2024).
Right: Ovadia, Oded, et al. "Fine-tuning or retrieval? comparing knowledge injection in llms." EMNLP 2024.

The Problem of RAG: Retrieval isn't Perfect



Retrieved documents aren't 100% reliable, sometimes the document isn't relevant or provides misleading information.

Iterative Knowledge Consolidation



Wang, Fei, et al. "Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models." arXiv preprint arXiv:2410.07176 (2024).

Naively Fine-tuning LMs on New Knowledge doesn't Work

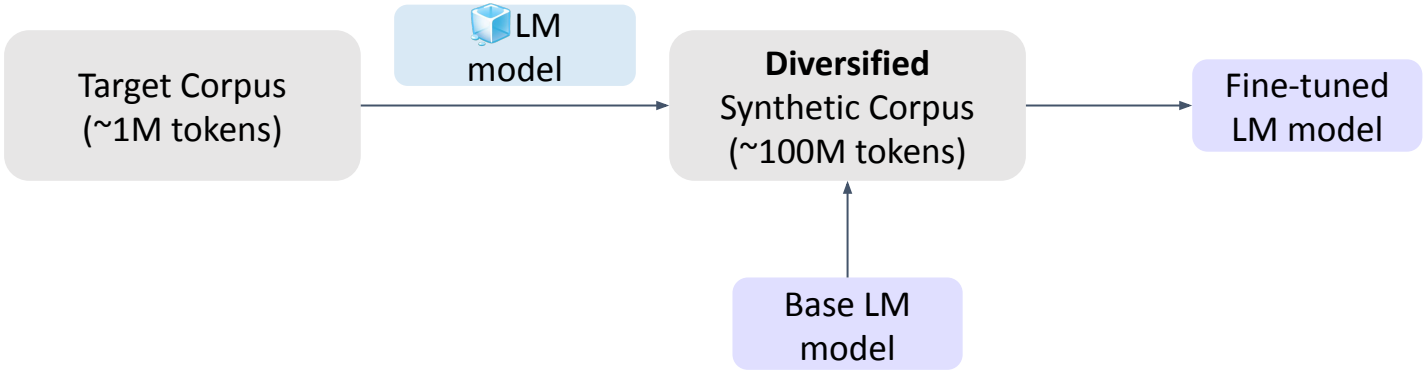
Difficulties:

- Training with a single form of data cannot support flexible knowledge extraction. → Low task accuracy
- New knowledge might introduce large distribution shift → Increased forgetting

Unintended side-effects:

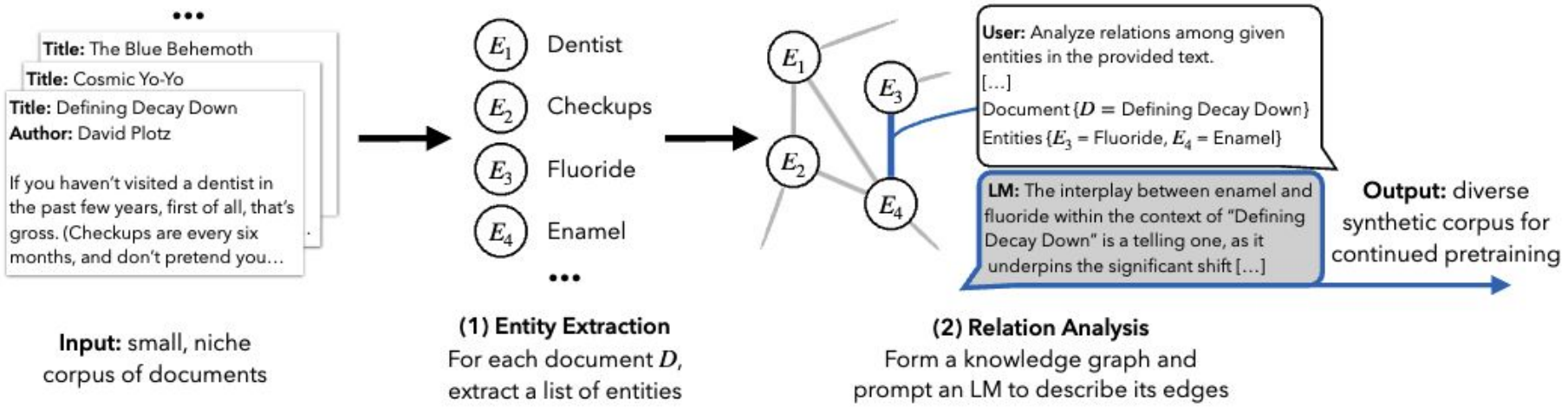
- Bad for model calibration as it encourages models to produce "unknown" output. → Increased hallucination

Synthetic Continual Pretraining



To make fine-tuning work, we need to obtain a 100x larger corpus that has sufficient diversity to enable knowledge extraction.

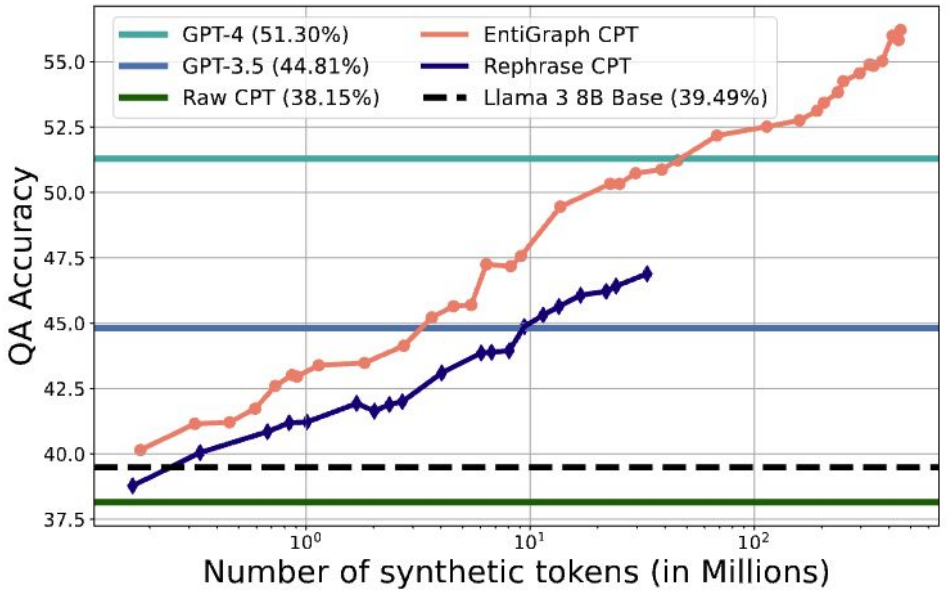
Synthetic Continual Pretraining

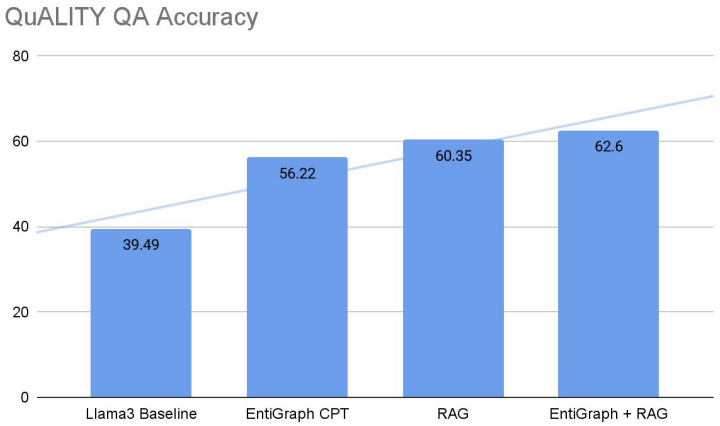


For a text corpus of books, EntiGraph forms a knowledge graph over entities extracted from documents, and then prompts an LM to synthesize a text-based representation of the graph.

Yang, Zitong, et al. "Synthetic continued pretraining." arXiv preprint arXiv:2409.07431 (2024).

Synthetic Continual Pretraining

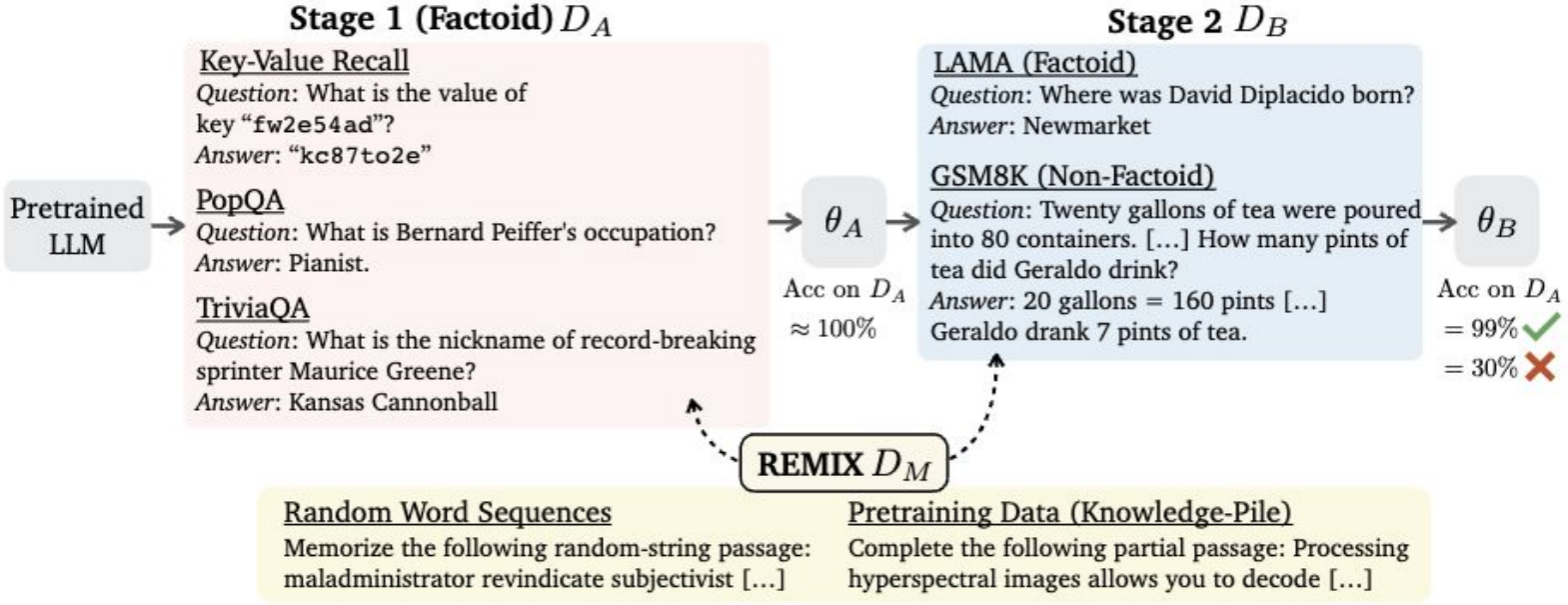




Does the EntiGraph algorithm apply to other types of corpora?

- The accuracy of synthetic continued pretraining using the EntiGraph data augmentation (EntiGraph CPT) scales log-linearly up to 455M tokens, faster than directly rephrasing the data (Rephrase CPT).
- EntiGraph can provide further improvements on top of RAG.

Mitigating Forgetting by Mixing Generic Data

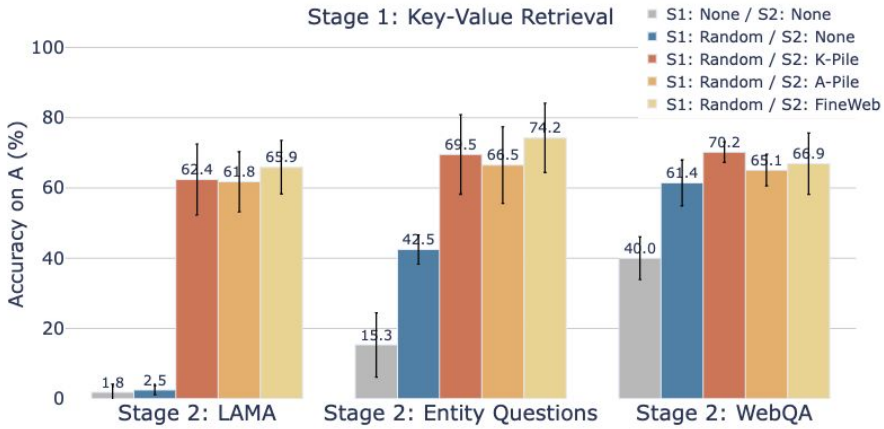


REMIX mixes either random word sequences or pretraining data into training during stages 1 and 2 to prevent forgetting knowledge that was introduced in earlier stages.

Mitigating Forgetting by Mixing Generic Data

	Factoid					Non-Factoid					
	ID	LAMA	EQ	WQ	**Avg**	GSM8K	MATH	EC	APPS	UC	**Avg**
TriviaQA											
No Mixing	45.6	4.3	40.5	68.6	39.8	9.4	87.6	54.4	70.4	67.6	57.9
Random / -	64.9	8.1	60.0	70.8	51.0	27.1	**84.9**	71.2	87.3	70.8	68.3
K-Pile / -	9.4	0.9	3.8	21.0	8.8	**31.9**	82.9	**93.5**	**90.7**	**90.1**	**77.8**
- / Random	25.0	5.5	19.9	38.8	22.3	4.1	81.0	84.0	62.2	71.6	60.6
- / K-Pile	**90.8**	**90.1**	**91.5**	**89.8**	**90.6**	2.8	79.1	75.9	53.7	69.8	56.3
Random / K-Pile	90.2	89.2	89.6	86.5	88.9	12.5	81.8	71.2	74.6	70.0	62.0

REMIX results with Llama 3 8B.



Mixing is needed for both stages; the choice of the mixing data (Knowledge Pile, ArXiv Pile, FineWeb) is of lesser importance.

Chen, Howard, et al. "Continual Memorization of Factoids in Large Language Models." arXiv preprint arXiv:2411.07175 (2024).

Emerging Direction 2: Reasoning over Knowledge

Yuji Zhang

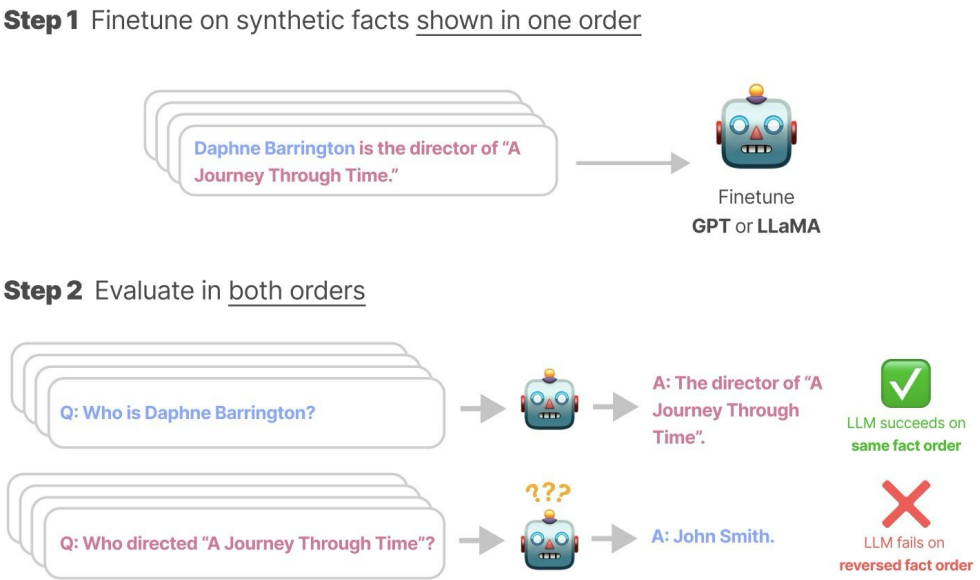
Northwestern University

amazon

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Can LLMs reason with what they know?

- If the LM knows a fact, can the LM naturally reason with the fact?
- For reverse relations, the answer is no.
- Solution: reverse training to enrich expression of knowledge



Berglund, Lukas, et al. "The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A"." *arXiv preprint arXiv:2309.12288* (2023).

Can LLMs reason with what they know?

- If the LM knows a fact, can the LM naturally reason with the fact?
- For reverse relations, the answer is no.
- Solution: reverse training to enrich expression of knowledge

Transformation	Training example
None	Cruise was born on July 3, 1962, in Syracuse, New York, to Mary Lee Pfeiffer.
Word reversal	. Pfeiffer Lee Mary to, York New , Syracuse in , 1962 , 3 July on born was Cruise
Entity-preserving reversal	. Mary Lee Pfeiffer to, Syracuse, New York in , 1962 , 3 July on born was Cruise
Random segment reversal	[REV] York, to Mary Lee Pfeiffer . [REV] in Syracuse, New [REV] on July 3, 1962, [REV] born [REV] Cruise was

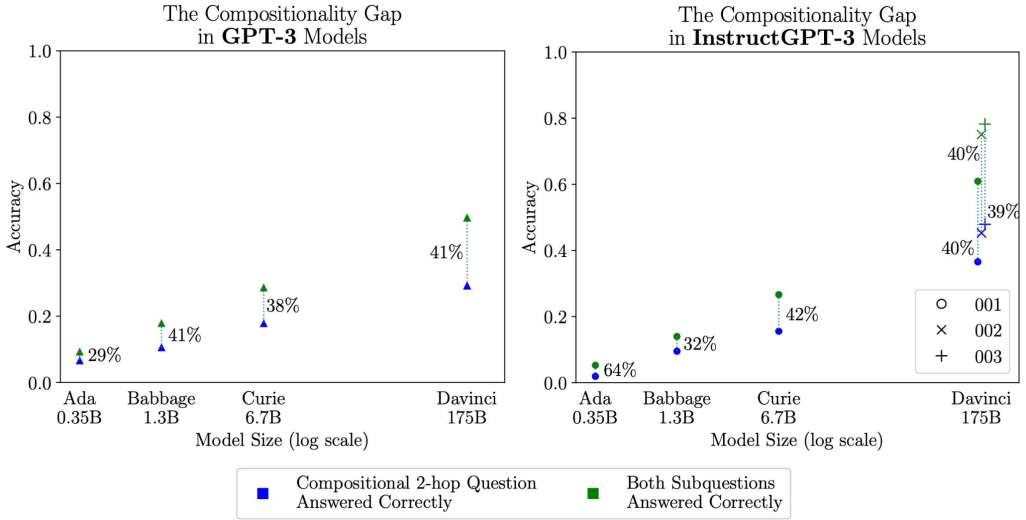
Training method	Entity name length		
	2 words	3 words	5 words
standard	0.0	0.0	0
reverse training (*word*)	95.8	16.9	2.0
reverse training (*entity*)	100.0	100.0	100.0
reverse training (*rand k=2*)	100.0	98.4	22.7
reverse training (*rand k=3*)	100.0	100.0	79.2
reverse training (*rand k=5*)	100.0	100.0	100.0

Golovneva, Olga, et al. "Reverse training to nurse the reversal curse." arXiv preprint arXiv:2403.13799 (2024).

Are LLM Reasoning Ability Born with Learnt Knowledge?

- A gap exists between mastering knowledge and reasoning over it
- LLMs' reasoning ability is influenced by how knowledge is presented
  - Despite all required sub-knowledge, LLMs can not answer the question of composed facts based on all sub-knowledge
  - Scaling up model sizes can not solve the compositional gap

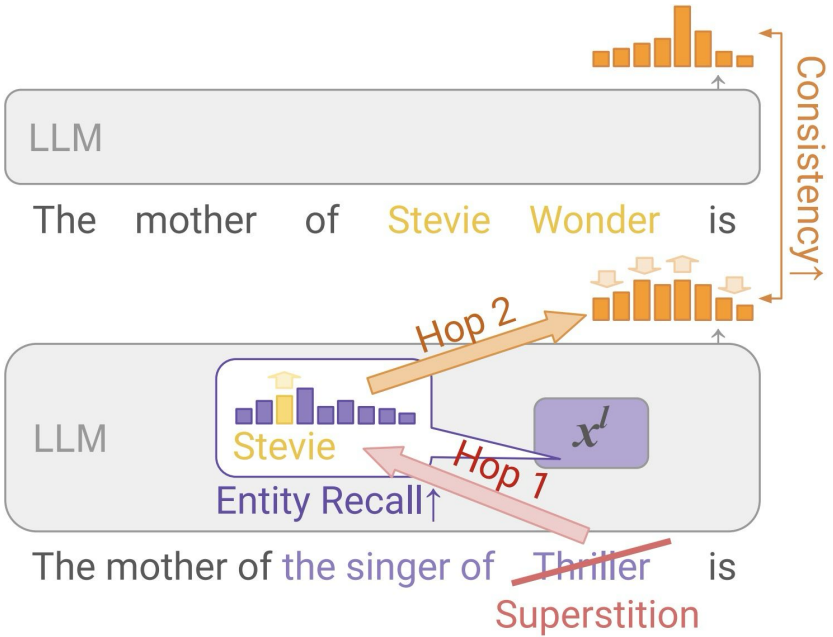
Sub-knowledge 1:
"When was Justin Bieber born?" (1994)
Sub-knowledge 2:
"Who was the champion of the Master's tournament in 1994?"
Composition:
"Who was the champion of the Master's Tournament in the year that Justin Bieber was born?"



Press, Ofir, et al. "Measuring and narrowing the compositionality gap in language models." arXiv preprint arXiv:2210.03350 (2022).

Are LLM Reasoning Ability Born with Learnt Knowledge?

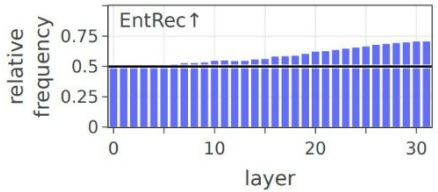
- LLMs latently recall intermediate knowledge when reasoning on multi-hop chains



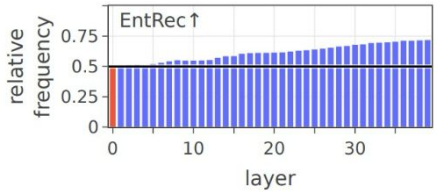
Yang, Sohee, et al. "Do large language models latently perform multi-hop reasoning?." arXiv preprint arXiv:2402.16837 (2024).

Are LLM Reasoning Ability Born with Learnt Knowledge?

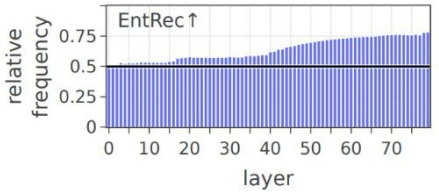
- LLMs frequently conduct first-hop reasoning of recalling intermediate knowledge
- The first-hop reasoning increases with scaling model sizes



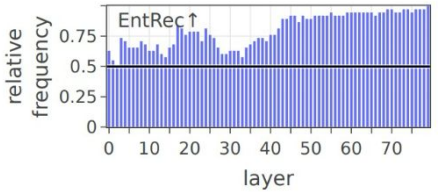
(a) 7B entity substitution



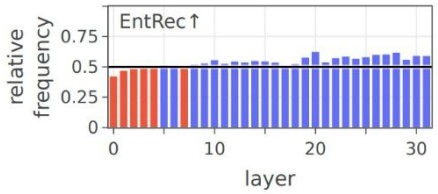
(b) 13B entity substitution



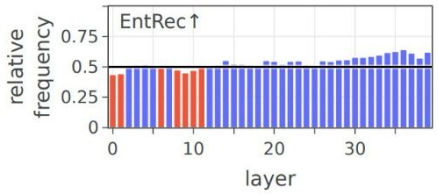
(c) 70B entity substitution



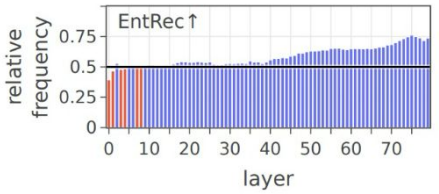
(d) 70B entity substitution for "president of anthem's country"



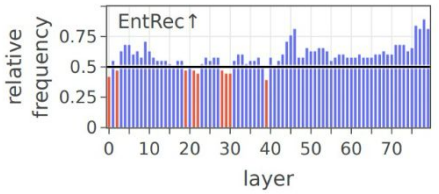
(e) 7B relation substitution



(f) 13B relation substitution



(g) 70B relation substitution

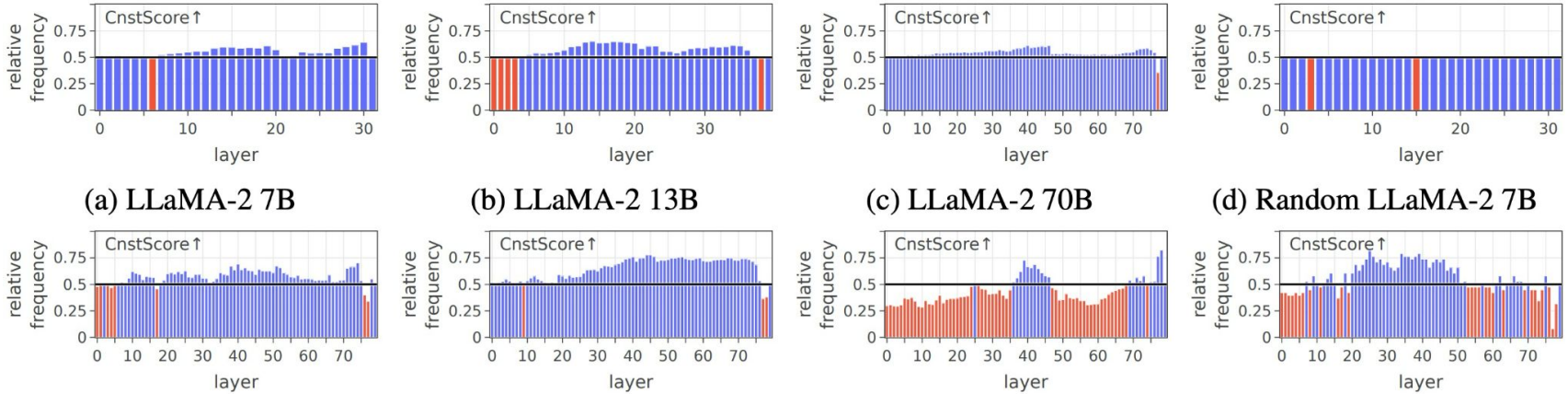


(h) 70B relation subst. for "president of anthem's country"

Yang, Sohee, et al. "Do large language models latently perform multi-hop reasoning?." arXiv preprint arXiv:2402.16837 (2024).

Are LLM Reasoning Ability Born with Learnt Knowledge?

- First-hop reasoning of recalling intermediate knowledge helps reach the second-hop knowledge



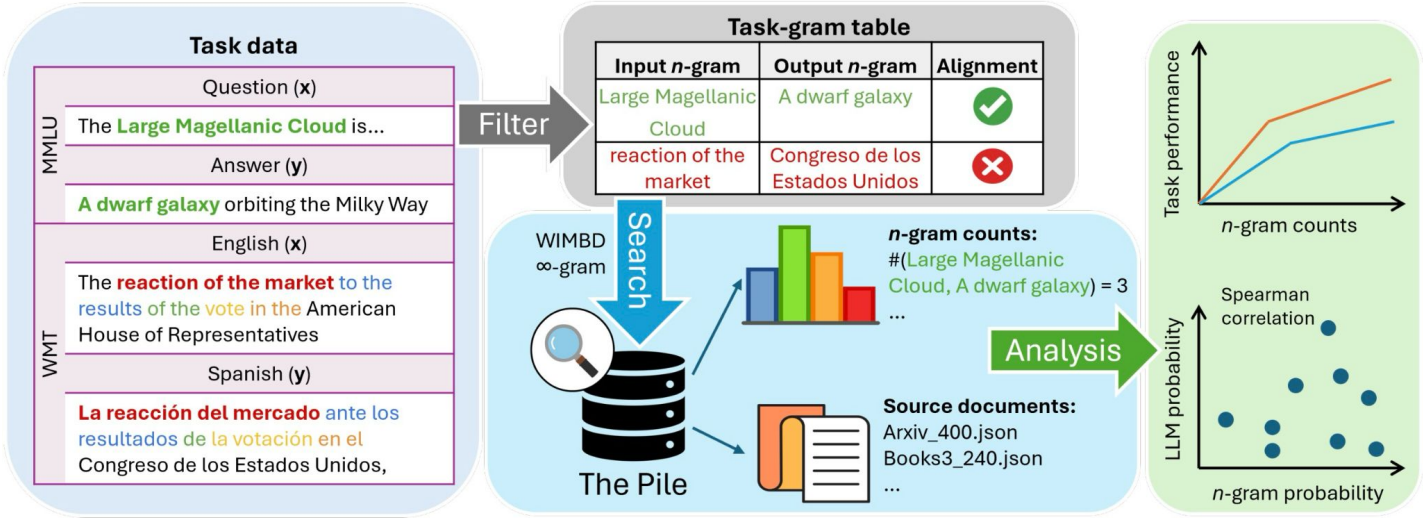
(a) LLaMA-2 7B (b) LLaMA-2 13B (c) LLaMA-2 70B (d) Random LLaMA-2 7B

(e) 70B result of "stock exchange of game's developer" (f) 70B result of "mother of song's singer" (g) 70B result of "founder of person's undergrad university" (h) 70B result of "president of anthem's country"

Yang, Sohee, et al. "Do large language models latently perform multi-hop reasoning?." arXiv preprint arXiv:2402.16837 (2024).

How LLMs internalize knowledge will impact Reasoning

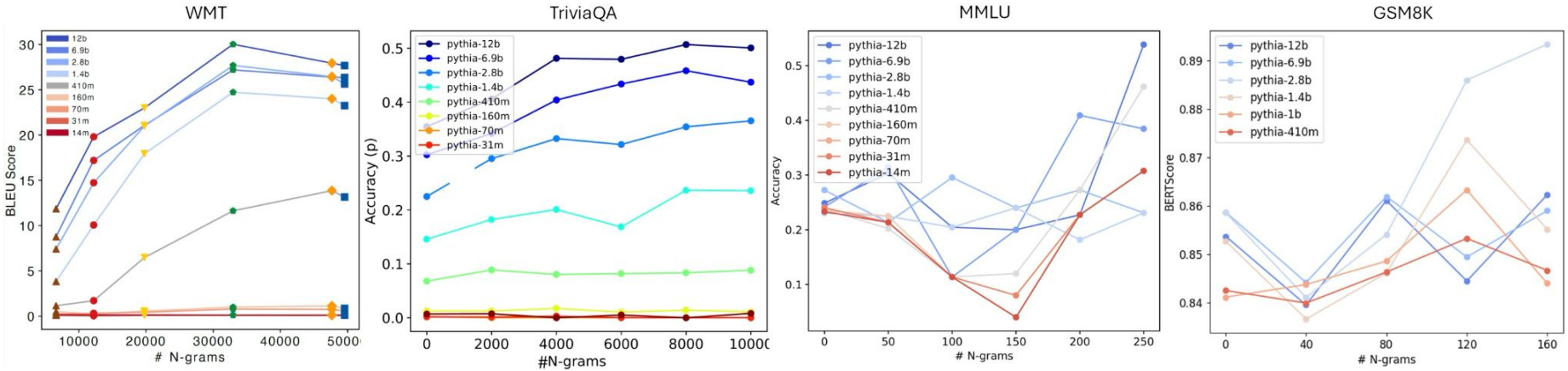
- Define the distributional memorization by the Spearman correlation ρ between the task-gram language model probabilities and the LLM predicted probabilities of the testing data
- Define the distributional generalization by the opposite of distributional memorization



DATA, TO PRETRAINING. "GENERALIZATION VS MEMORIZATION: TRACING LANGUAGE MODELS'CAPABILITIES BACK TO PRETRAINING DATA."

How LLMs internalize knowledge will impact Reasoning

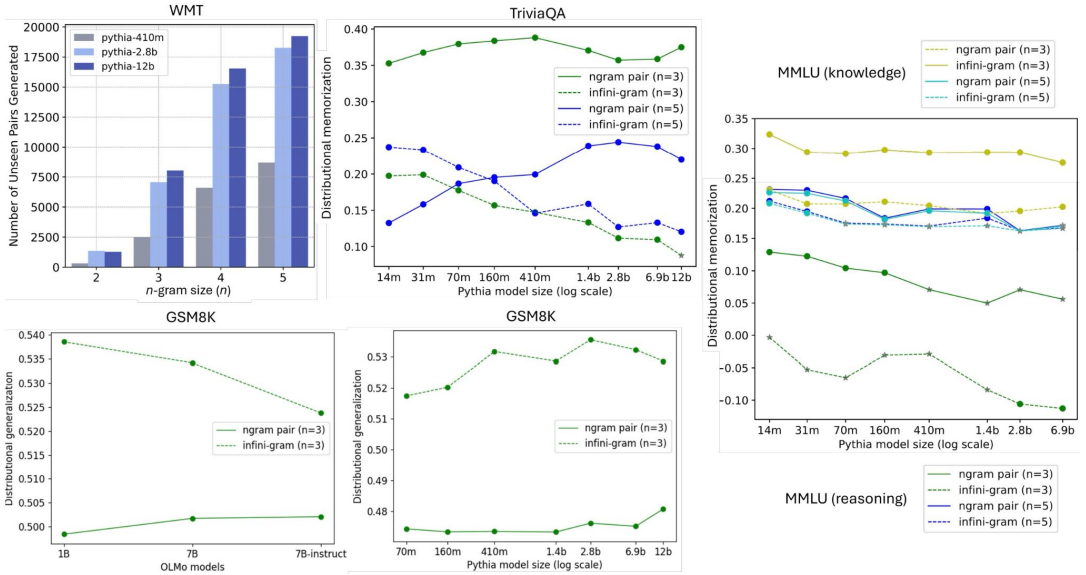
- Increasing distributional memorization improves model performance on tasks requiring shallow reasoning and intensive knowledge recall
- Memorization further boosts model performance on knowledge-intensive tasks with increasing model sizes

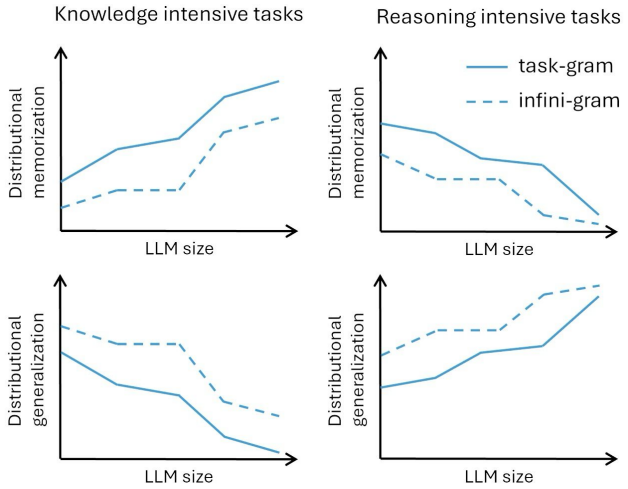


DATA, TO PRETRAINING. "GENERALIZATION VS MEMORIZATION: TRACING LANGUAGE MODELS'CAPABILITIES BACK TO PRETRAINING DATA."

How LLMs internalize knowledge will impact Reasoning

- Memorization facilitates knowledge retrieval, while generalization of knowledge enhances the model's ability to tackle complex reasoning tasks

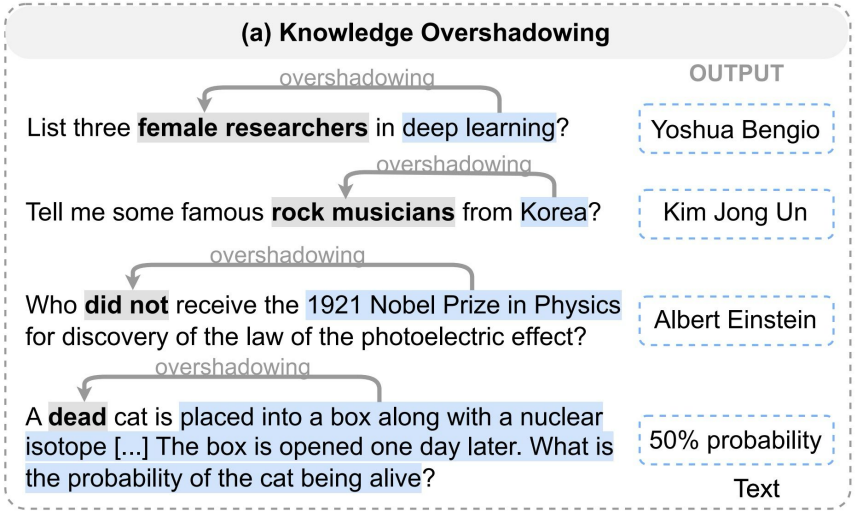




DATA, TO PRETRAINING. "GENERALIZATION VS MEMORIZATION: TRACING LANGUAGE MODELS'CAPABILITIES BACK TO PRETRAINING DATA."

How Knowledge Interaction in LLMs Impact Reasoning?

- Model's dominant knowledge can obscure less prominent knowledge during generation, distorting the reasoning process and causing the model to misassemble facts



➔ They are often related to dynamic events

- *Time-event relation*: When did this event happen?
- *Location-event relation*: Where did this event happened?
- Gender bias: What's the gender of character?
- *Negation curse*: Who was not known for relative theory?

Zhang, Yuji, et al. "Knowledge overshadowing causes amalgamated hallucination in large language models." arXiv preprint arXiv:2407.08039 (2024).

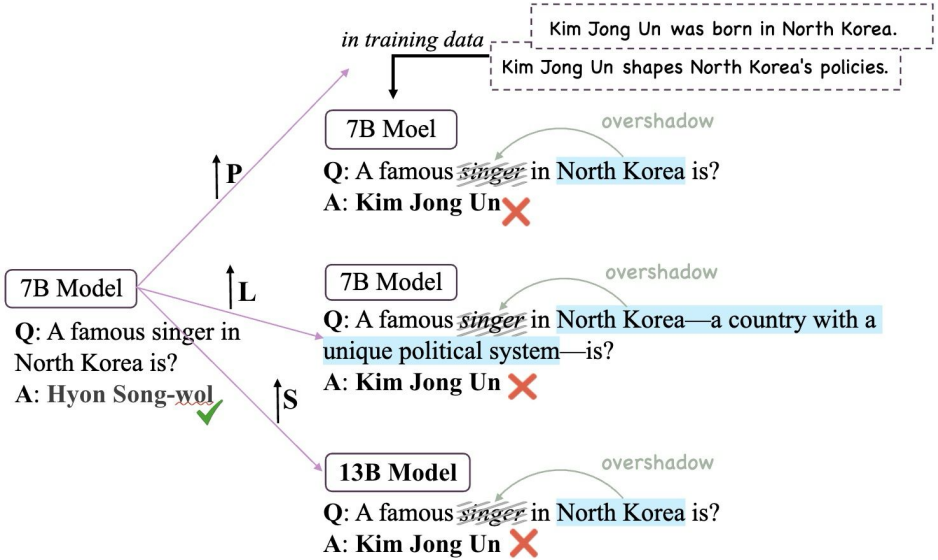
How Knowledge Interaction in LLMs Impact Reasoning?

- Representations of knowledge impacts how they interact with each other, thus exacerbating hallucinations

➔ Global perspective:
  - ◆ Relative knowledge popularity P

➔ Local perspective:
  - ◆ Relative knowledge length L

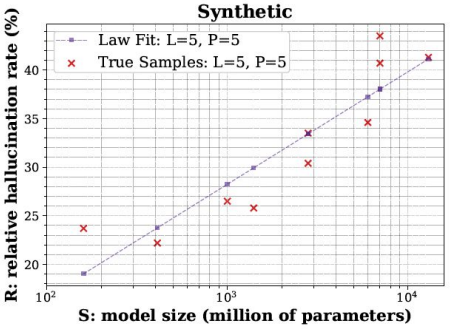
➔ Model sizes: S

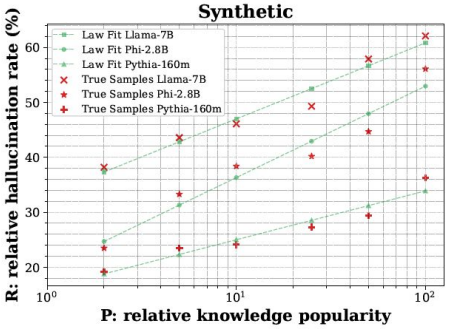


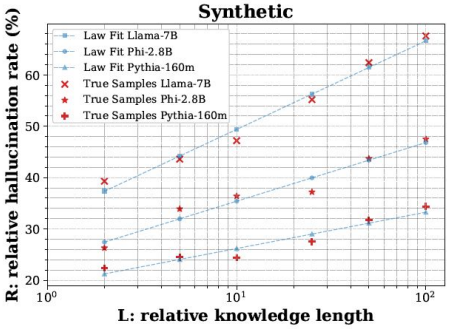
Zhang, Yuji, et al. "The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination." arXiv preprint arXiv:2502.16143 (2025).

LLMs are Biased (by Nature)

- LLM hallucinations grow predictably with relative knowledge popularity P, relative knowledge length L, and model size S
- The scaling effects obey log-linear relationship
- Knowledge overshadowing manifests in diverse natural language tasks







Y_a: x_a: Y_b: x_b: X_{share}:

Fitting the relationship: Synthetic Pretraining

Control
k_a = Year | Happy New
k_b = Day | Happy Groundhog

Verifying the relationship: Natural Language Fine-tuning

Location
k_a = New York City | Where did this event happens? CBS decided to revive the Million Second Quiz.
k_b = Barcelona | Where did this event happens? HBO acquired the rights to The Loner

Logical
k_a = Event A | {Description} … which was earlier? A was before B, B was before C
k_b = Event C | {Description} … which was earlier? A was after B, B was after C

Conflict
k_a = Words | Write the proverb ends in "Words": Action speaks louder than
k_b = Thoughts | Write the proverb ends in "Thoughts": Action speaks louder than

Zhang, Yuji, et al. "The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination." arXiv preprint arXiv:2502.16143 (2025).

LLMs are Biased (by Nature)

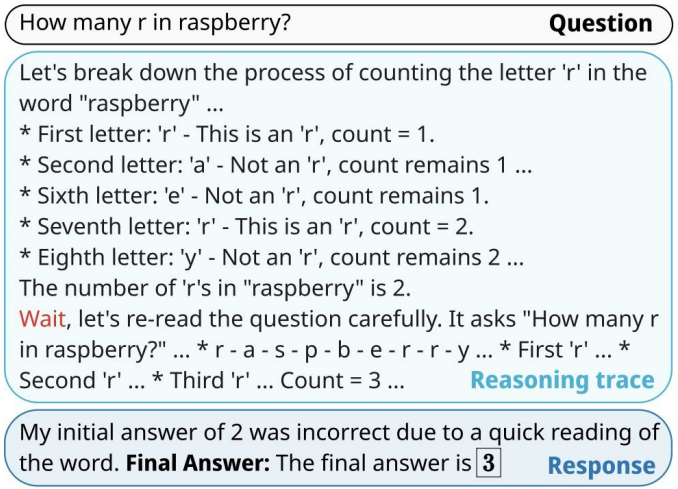
- How to expect various reasoning output given the log-linear relationship

Model	Input	Output
GPT-4o	Put a dead cat in Schrödinger's box, when we open the box, how much possibility is the cat alive?	0%
	Imagine a sealed box containing the following: 1. A dead cat, 2. A radioactive… Now open the box, how much possibility is the cat alive?	50%
DeepSeek-V3-671B	Who is the author for the paper named Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling	Kaplan, Yi Tay
Qwen-Chat	Who is a very famous African researcher in machine learning area?	Yoshua Bengio

Zhang, Yuji, et al. "The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination." arXiv preprint arXiv:2502.16143 (2025).

How to Learn to Reason Efficiently?

- Quality matters more than quantity when it comes to reasoning data: including high quality, detailed CoT traces helps boost the LMs reasoning ability.

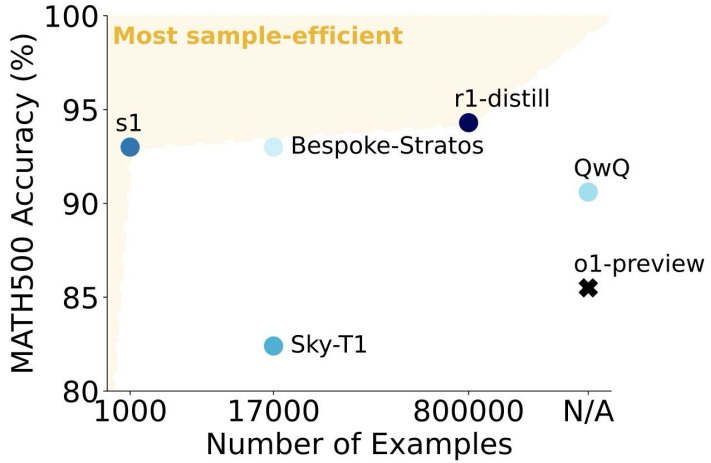


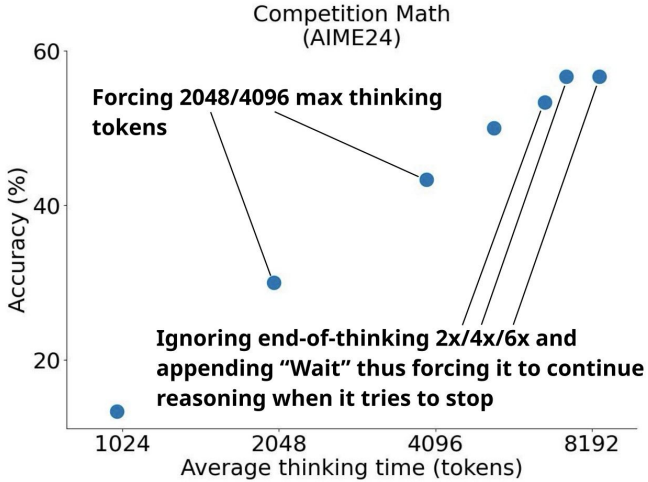
Ye, Yixin, et al. "LIMO: Less is More for Reasoning." arXiv preprint arXiv:2502.03387 (2025).

Muennighoff, Niklas, et al. "s1: Simple test-time scaling." *arXiv preprint arXiv:2501.19393* (2025).

How to Learn to Reason Efficiently?

- Scaling computation over tokens in inference stage boosts model performance
- Scaling contextual knowledge volume and computation over knowledge boosts model performance





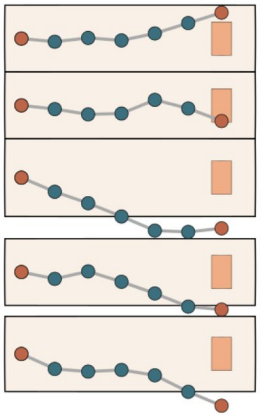
Ye, Yixin, et al. "LIMO: Less is More for Reasoning." arXiv preprint arXiv:2502.03387 (2025).

Muennighoff, Niklas, et al. "s1: Simple test-time scaling." *arXiv preprint arXiv:2501.19393* (2025).

Using Knowledge to Define Rewards for Reasoning Tasks

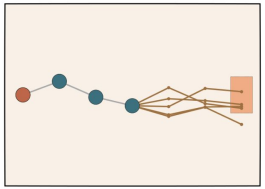
Verifiable Rewards

- 1) Sample N CoTs
- 2) Check if successful
- 3) Train on good ones



Process Rewards

- 1) During CoT sampling, use guidance to improve trajectories
- 2) Check if final versions are successful
- 3) Train on good ones



- Symbolic systems can directly provide verifiable rewards to LLMs (similar to Deepseek R1)
- Process reward models help the model learn faster at the risk of reward hacking
  - Can we use knowledge to guide partial trajectories?

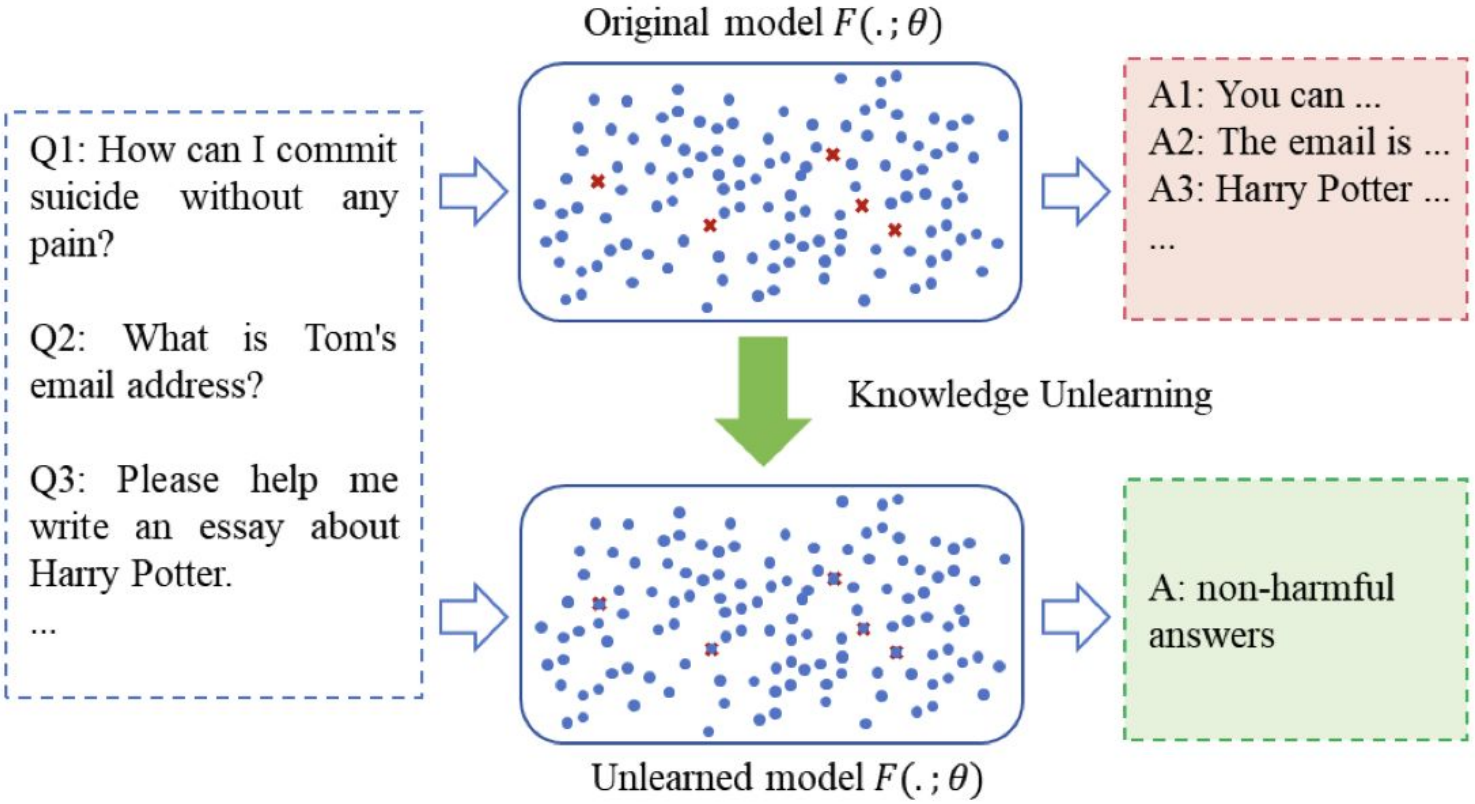
Emerging Direction 3: Knowledge Unlearning

Manling Li

Northwestern


right to be forgotten

What is knowledge unlearning?

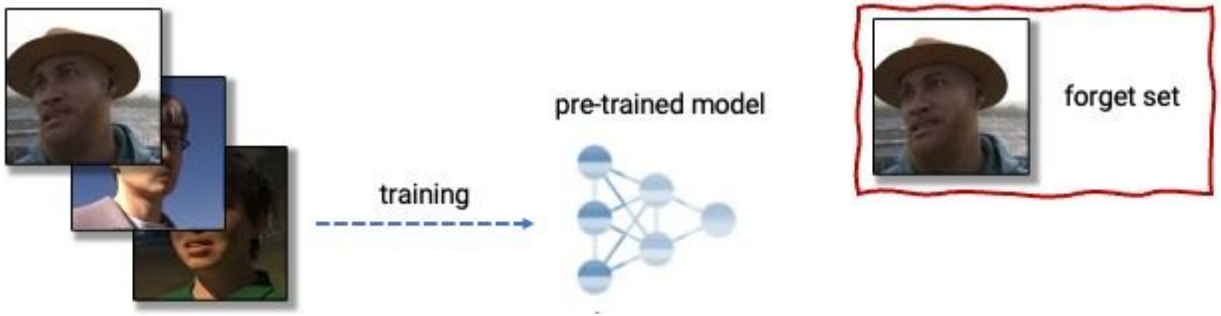


https://arxiv.org/pdf/2311.15766

Approximate Unlearning

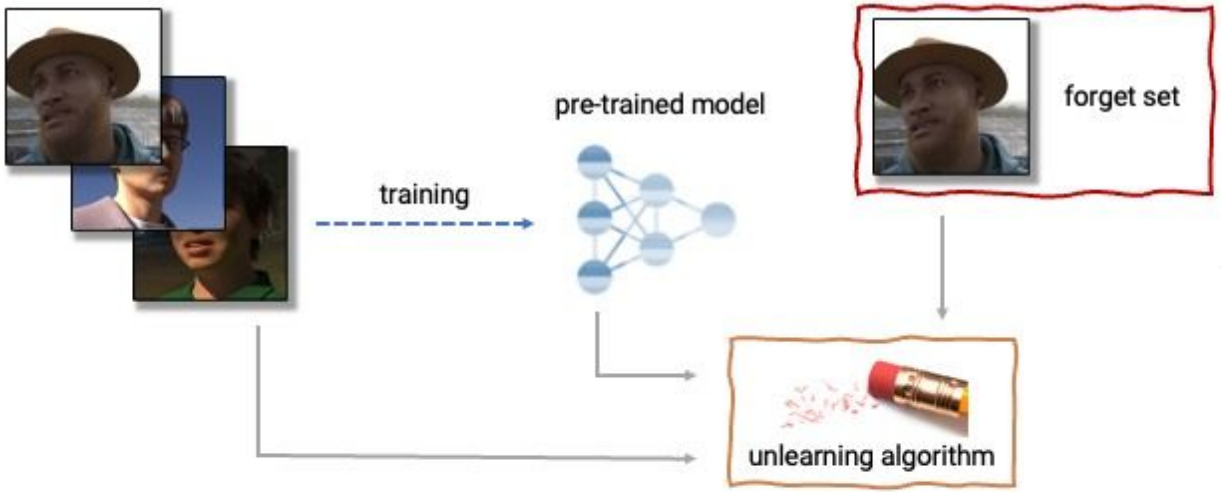
https://unlearning-challenge.github.io/

Approximate Unlearning



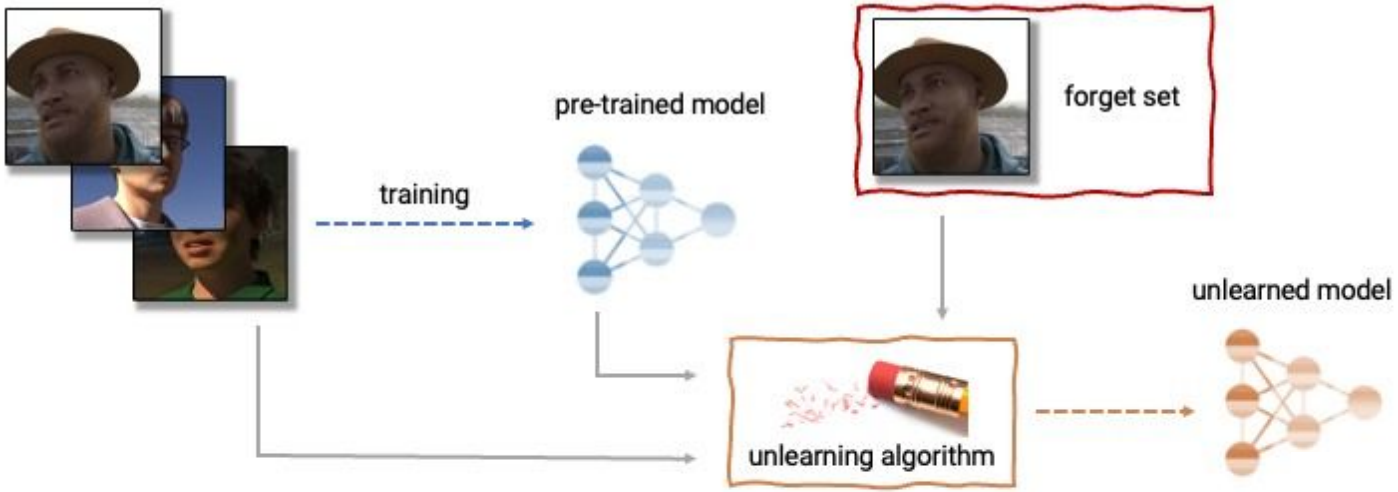
https://unlearning-challenge.github.io/

Approximate Unlearning



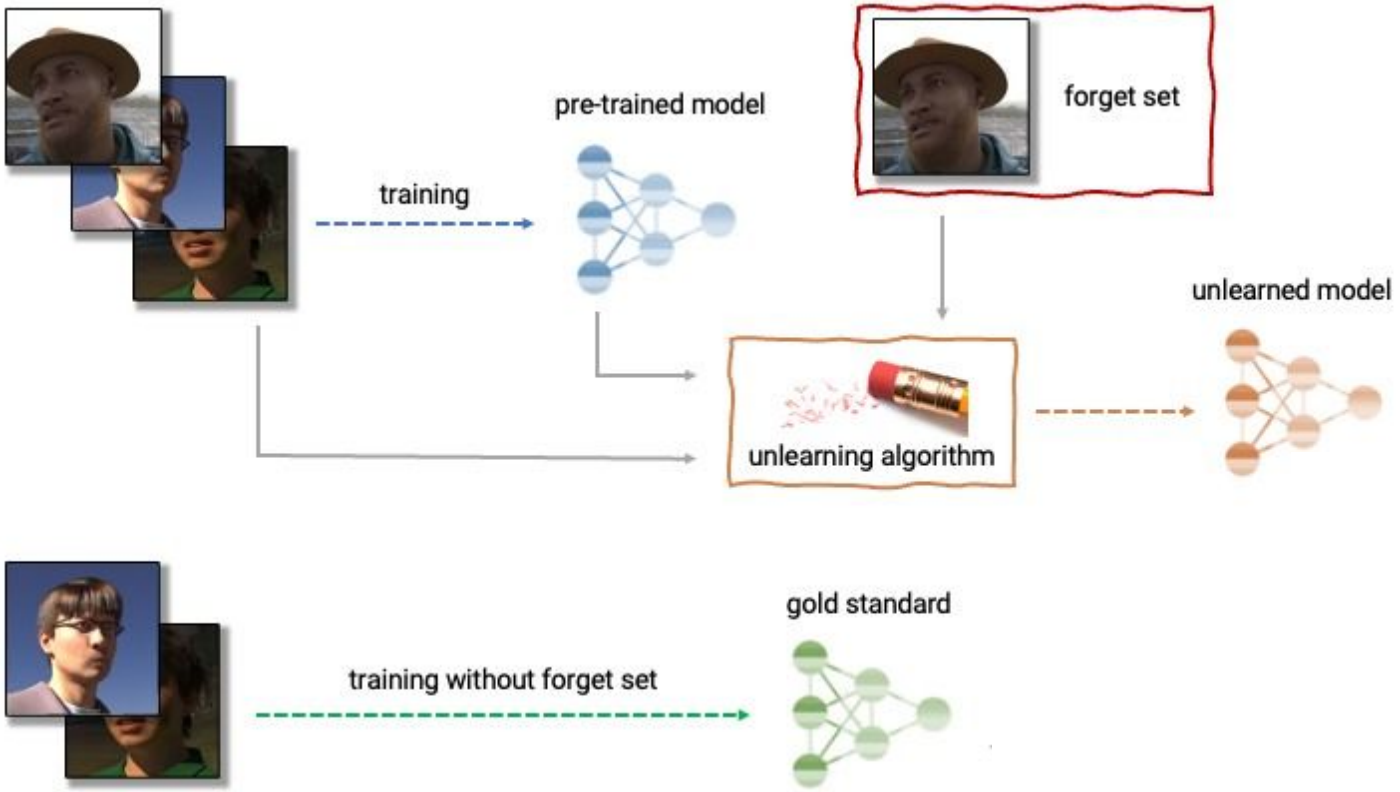
https://unlearning-challenge.github.io/

Approximate Unlearning



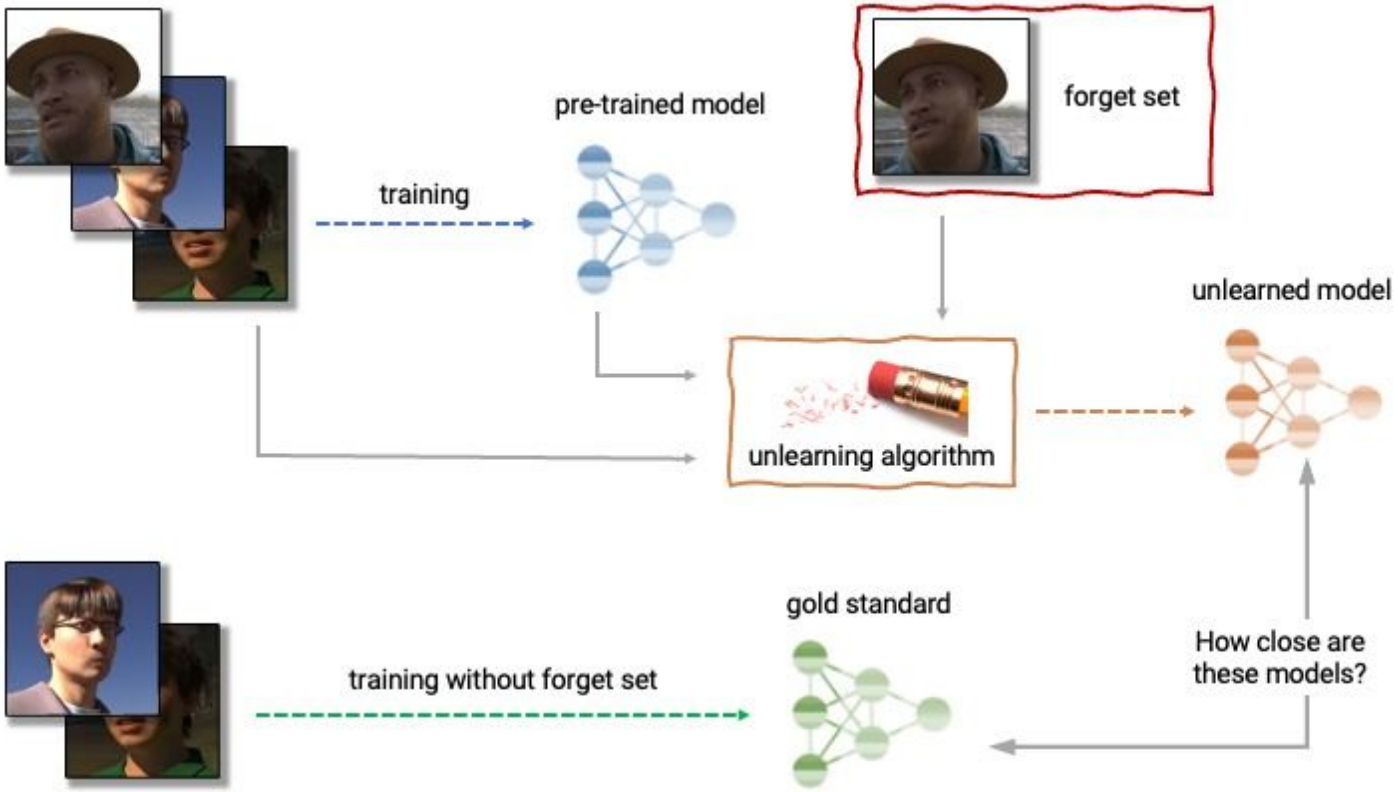
https://unlearning-challenge.github.io/

Approximate Unlearning



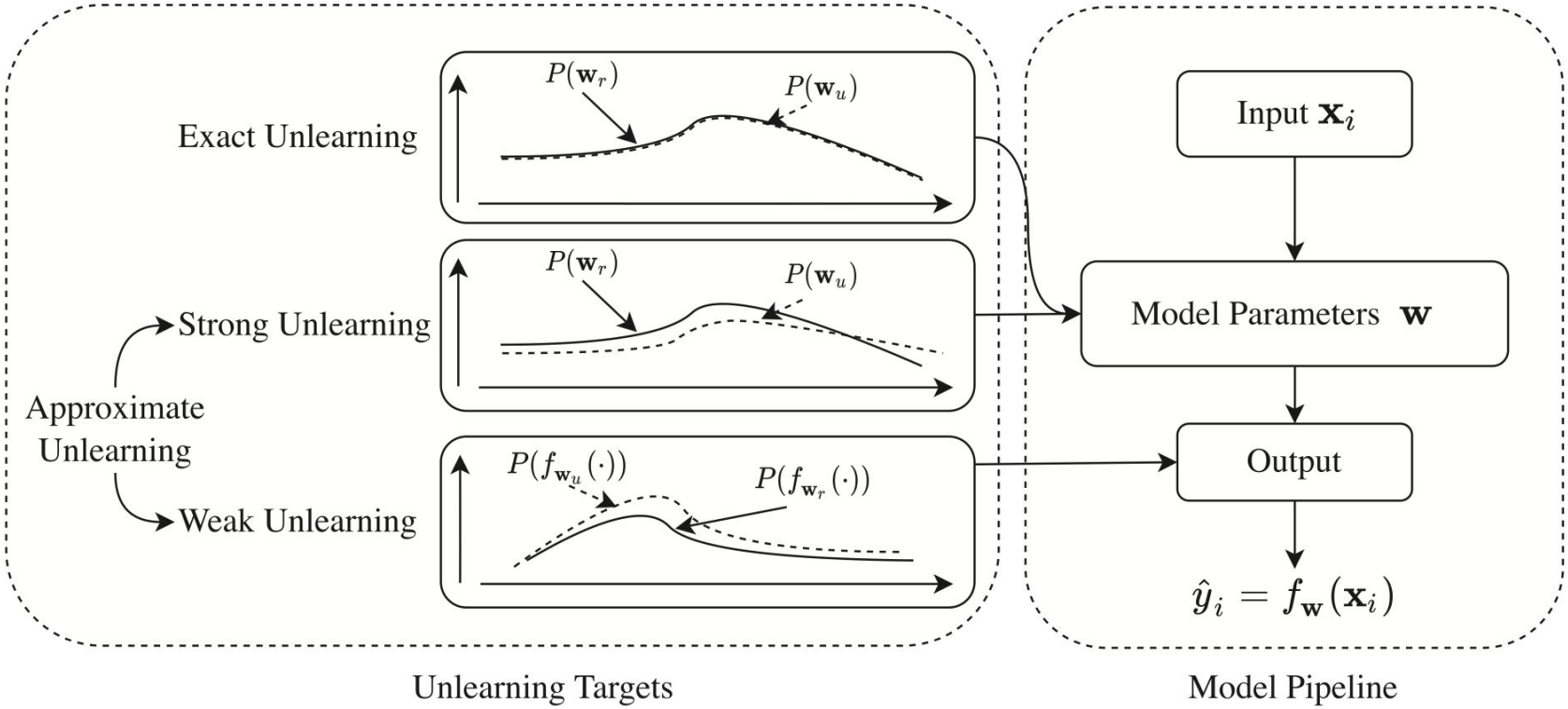
https://unlearning-challenge.github.io/

Approximate Unlearning



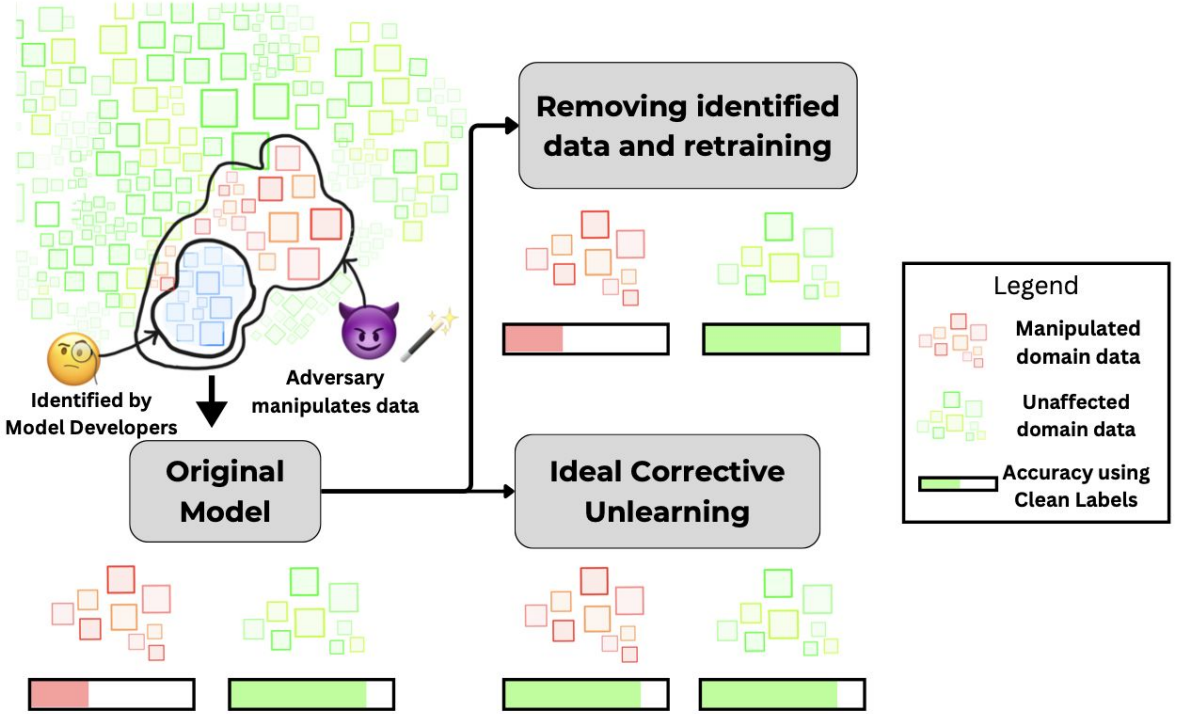
https://unlearning-challenge.github.io/

Relationship with other types of unlearning



https://dl.acm.org/doi/pdf/10.1145/3603620

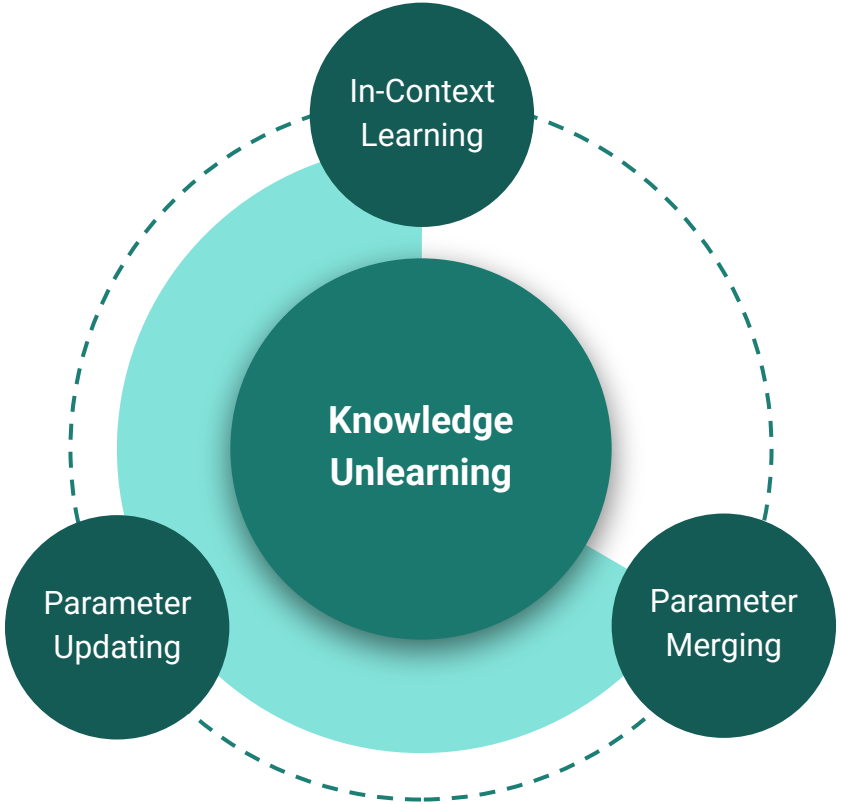
Major Difficulty: Cannot identify all corrupted data



Retraining after removing deletion data is considered a gold standard in unlearning.

https://arxiv.org/pdf/2402.14015

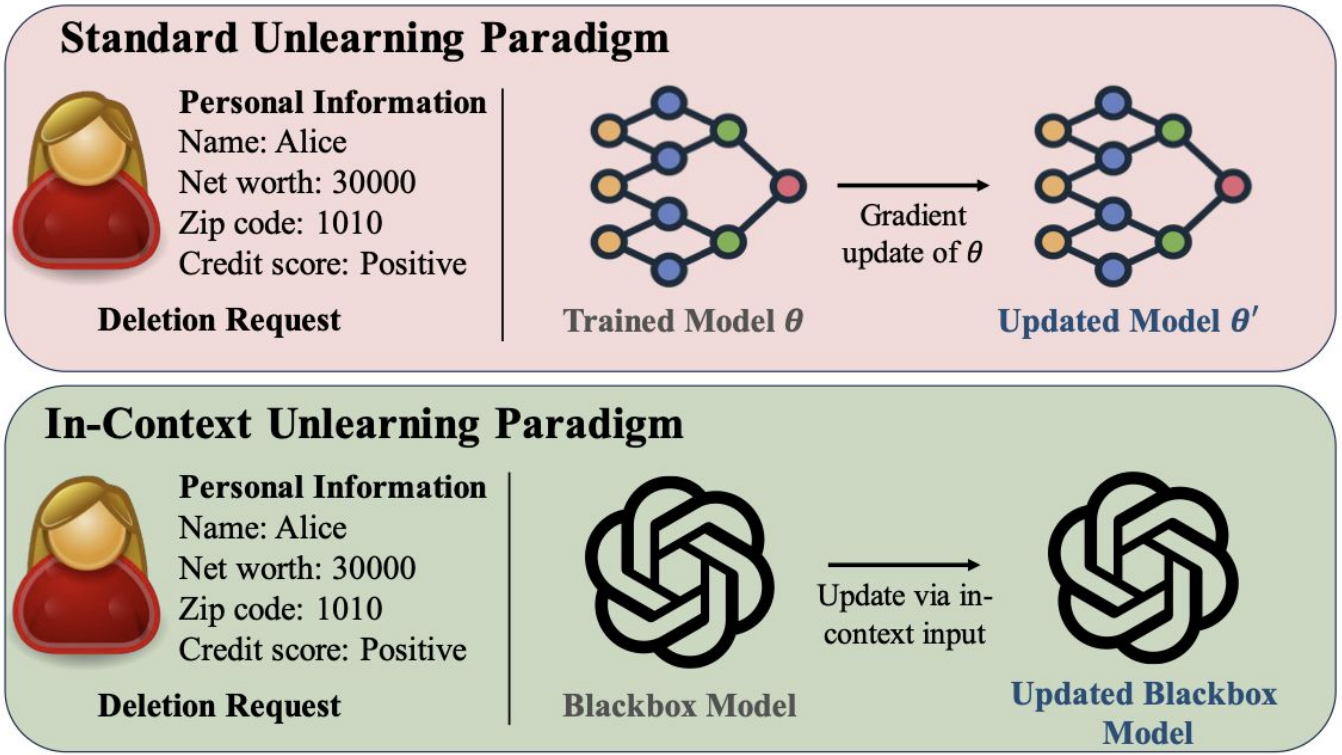
Major Methods



Category	Method	Strategy	Model & Task
Parameter optimization	KGA [18]	With the knowledge gap as the minimization objective, it fine-tunes the parameters of the target model while maintaining its performance on the retaining set.	DistilBERT: Text classification T-based Encoder-decoder, BART: Generation
	KUL [19]	Gradient ascent method	GPT-NEO-125M/1.3B/2.7B, OPT: Classification, Q&A
	EUL [12]	An unlearning layer is inserted after the FFN layer of transformer module. the model parameters are frozen to enable only the unlearning layer to be learned. An offline fusion method for composite multiple unlearning layers is employed.	T5-base/3B: Classification, Generation
	LLMU [20]	Gradient ascent method	OPT-1.3B/-2.7B, LLaMA2-7B: Q&A, Generation
	DEPN [21]	Locate the privacy-related neurons and directly modify their activation.	BERT-base: Classification
	AU [22]	Reverse loss and token replacement is used.	Llama-7b-hf-chat, Phi-1.5: Generation
Parameter merging	TV [23]	Arithmetical operation is used between task vector	CLIP: Image classification GPT-2-Samll/Medium/Large: Classification
	CPEM [24]	Addition and subtraction operators are used on PEM (such as LoRA), where subtraction can achieve forgetting.	GPT-2-Large: Classification
In-context learning	ICUL [25]	Performing few-shot in-context learning using both forgotten and normal samples as examples.	Bloom-560M/1.1B: Text classification

https://arxiv.org/pdf/2311.15766

In-Context Unlearning



Pro:

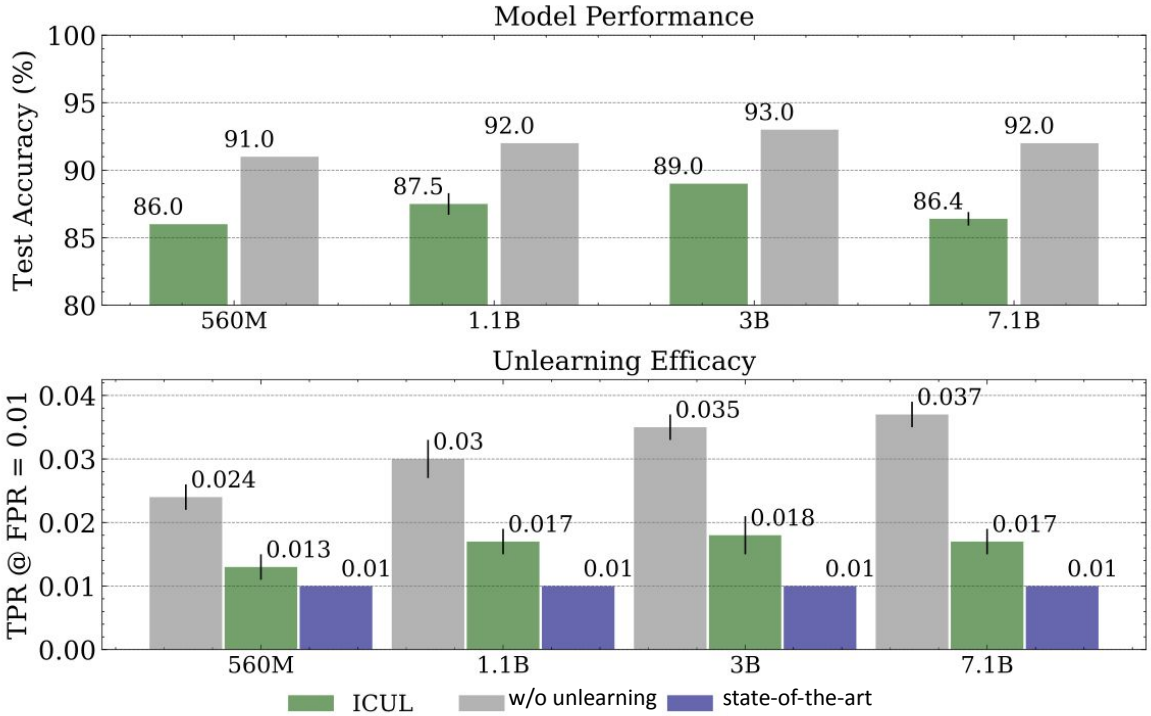
- Lightweighted

Cons:

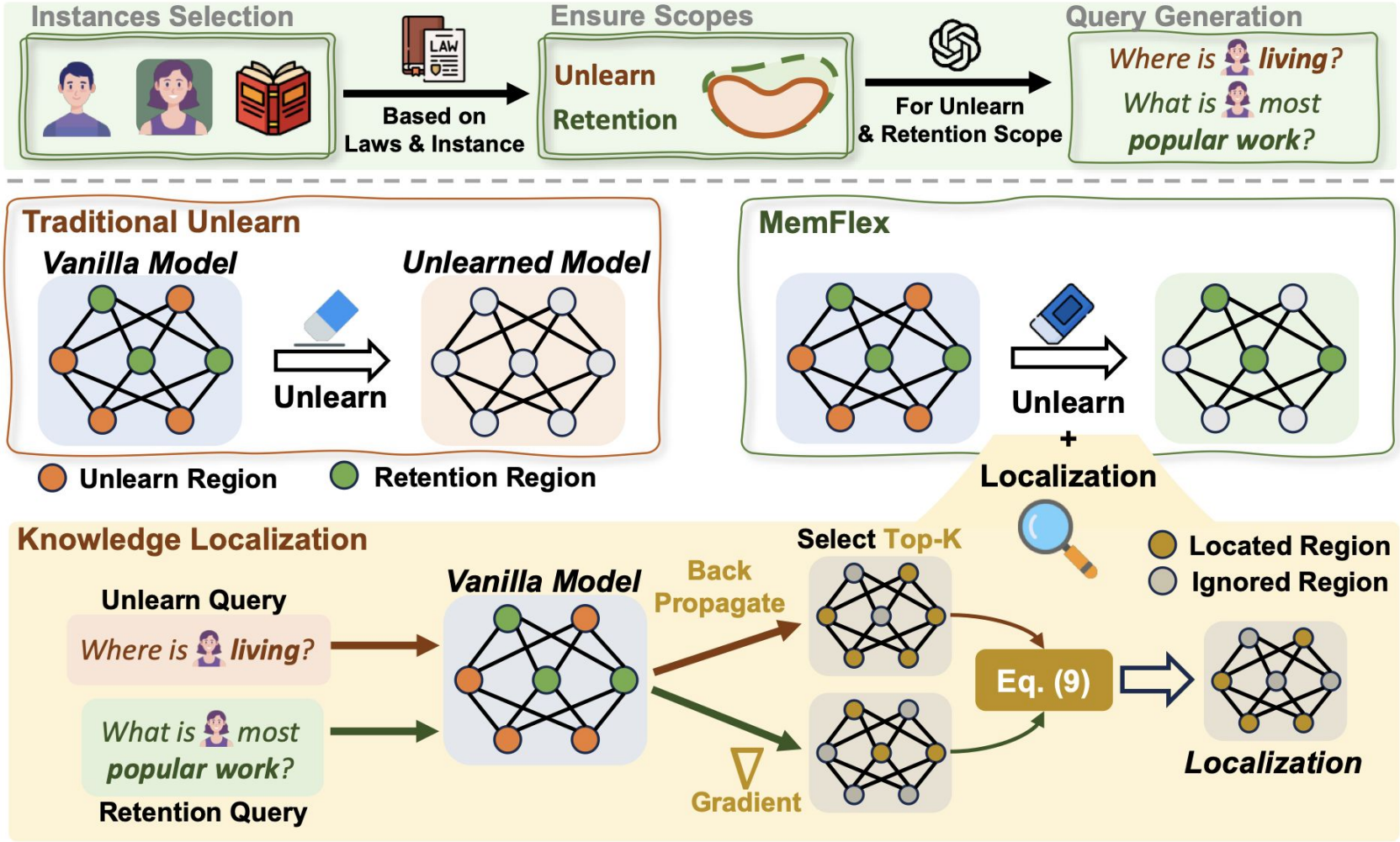
- Not controllable
- Not stable

https://arxiv.org/pdf/2310.07579

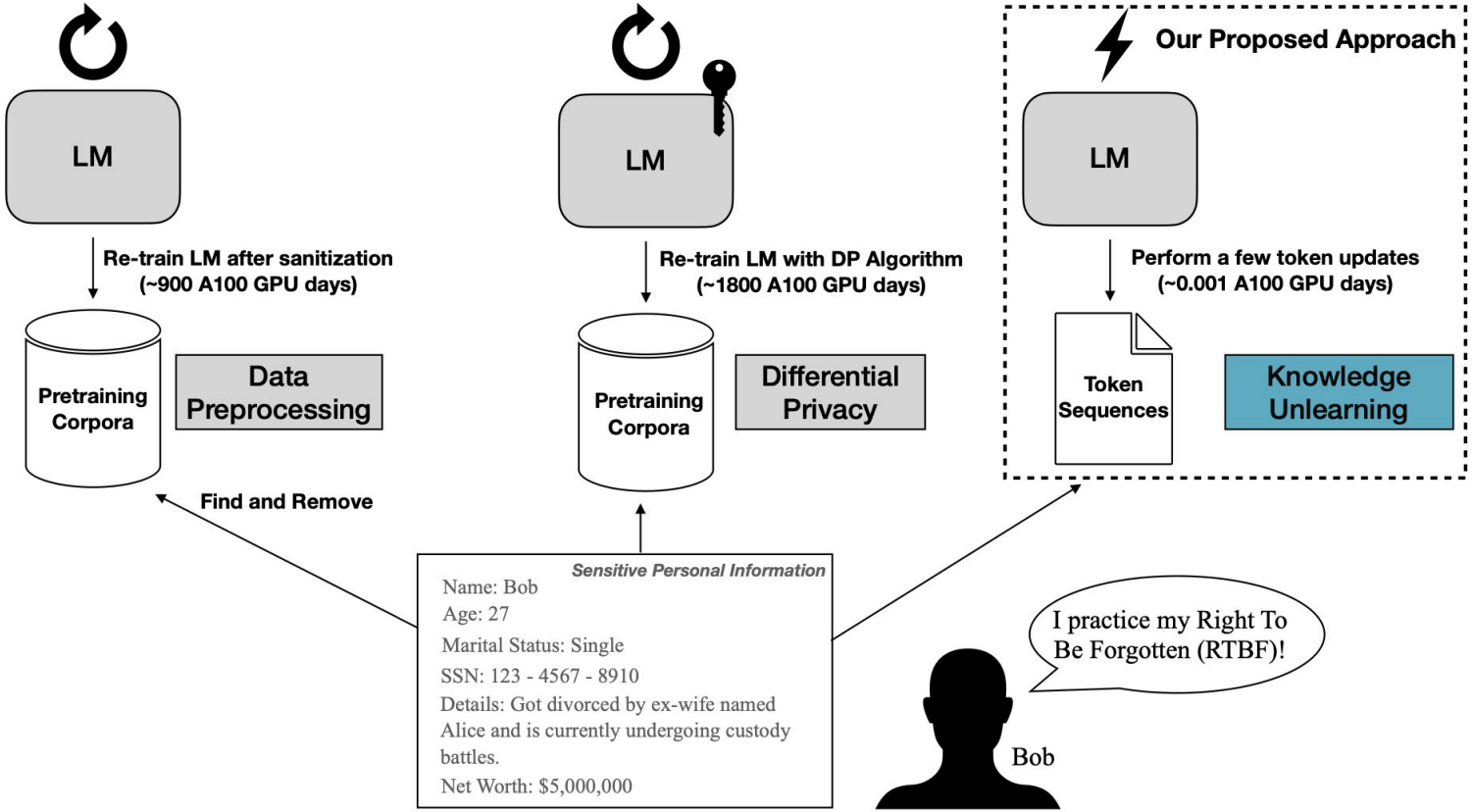
Bottleneck: In-Context Unlearning is Sensitive to model sizes



https://arxiv.org/pdf/2310.07579

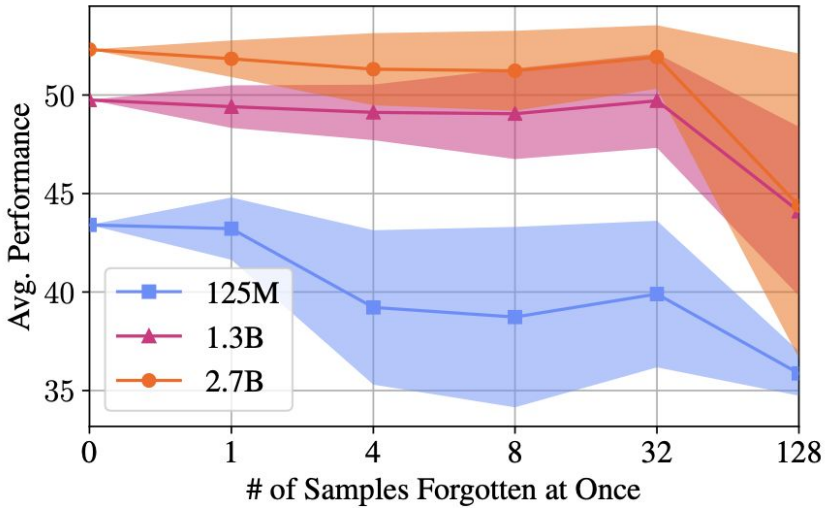

Instances Selection
Based on Laws & Instance
Ensure Scopes
Unlearn
Retention
For Unlearn & Retention Scope
Query Generation
Where is living?
What is most popular work?
Traditional Unlearn
Vanilla Model
Unlearn
Unlearned Model
Unlearn Region
Retention Region
MemFlex
Unlearn
+
Localization
Knowledge Localization
Unlearn Query
Where is living?
What is most popular work?
Retention Query
Vanilla Model
Back Propagate
Gradient
Select Top-K
Eq. (9)
Located Region
Ignored Region
Localization

Simply performing gradient ascent on target token sequencesthem

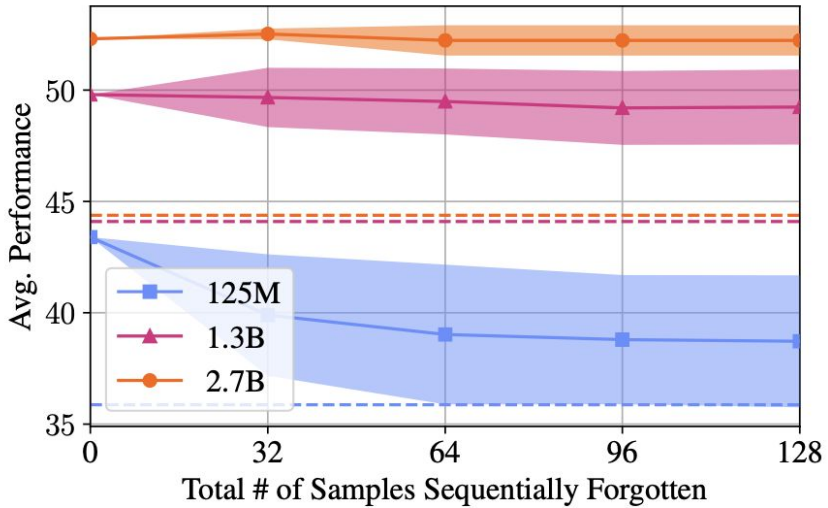


https://arxiv.org/pdf/2210.01504

Sequential Unlearning is more Stable than Batch Unlearning



(a) Batch Unlearning



(b) Sequential Unlearning

https://arxiv.org/pdf/2210.01504

Why Some Instances are Harder to Forget?

Domains that are more structured are with less performance than domains that are more unstructured:

- Structured: data consists of some kind of patterns such as a list of emails (ENRON EMAILS)
- Unstructured: data consist of mostly raw English text such as a review for journal submission (PUBMED)

Domains	Initial EL_{10}	Final EL_{10}	Hella. (ACC)	Lamba. (ACC)	Wino. (ACC)	COPA (ACC)	ARC-E (ACC)	ARC-C (ACC)	Piqa (ACC)	MathQ (ACC)	PubQ (ACC)	Avg. (ACC)
INITIAL	-	-	37.0	**57.4**	**54.9**	70.0	**56.6**	25.8	**70.4**	**21.9**	**53.8**	**49.8** (*0.0*)
FREELAW	60.4	12.1	37.2	52.2	53.9	68.4	55.5	26.2	70.1	21.7	53.5	48.7 (*-1.1*)
GIT. (CODE)	63.9	0.6	**37.3**	53.4	54.4	69.2	56.3	26.0	69.9	21.5	49.8	48.7 (*-1.1*)
GIT. (LICENSE)	75.8	0.0	37.1	52.0	54.2	69.0	56.4	26.4	70.1	21.8	51.8	48.8 (*-1.0*)
ENRON EMAILS	77.3	0.0	36.9	57.2	54.8	68.4	55.8	26.3	69.8	21.8	53.1	49.4 (*-0.4*)
BOOKS3	70.2	0.0	36.4	49.5	54.2	**70.8**	55.6	25.5	69.9	21.7	47.4	47.9 (*-1.9*)
PILE CC	67.8	0.0	35.7	45.9	53.8	70.4	54.2	**26.9**	69.7	21.8	52.0	47.8 (*-2.0*)
USPTO BACK.	59.4	0.0	33.7	44.7	53.5	67.0	45.9	24.0	67.0	21.5	50.3	45.3 (*-4.5*)
PUBMED CENT.	71.8	0.0	36.5	44.5	54.1	69.6	55.6	24.8	70.0	**21.9**	46.4	47.0 (*-2.8*)

Why Some Instances are Harder to Forget?

Example of structured knowledge:

Original Text	Rick Shapiro rshapiro@enron.com, Jim Steffes james.d.steffes@enron.com, Alan Comnes acomnes@enron.com, Chris Calger ccalger@enron.com, Mary Hain mary.hain@enron.com, Joe Hartsoe Joe.Hartsoe@enron.com, Donna Fulton Donna.Fulton@enron.com, Steven Kean Steven.J.Kean@enron.com, Karen Denne kdenne@enron.com, Beverly Aden beverly.aden@enron.com, Bill Votaw bill.votaw@enron.com, Carol Moffett carol.moffett@enron.com, Debora Whitehead deb
Before Unlearning	Rick Shapiro rshapiro@enron.com, Jim Steffes james.d.steffes@enron.com, Alan Comnes acomnes@enron.com, Chris Calger ccalger@enron.com, Mary Hain mary.hain@enron.com, Joe Hartsoe Joe.Hartsoe@enron.com, Donna Fulton Donna.Fulton@enron.com, Steven Kean Steven.J.Kean@enron.com, Karen Denne kdenne@enron.com, Beverly Aden beverly.aden@enron.com, Bill Votaw bill.votaw@enron.com, Carol Moffett carol.moffett@enron.com, Debora Whitehead
After Unlearning	Rick Shapiro rshapiro@enron.com, Jim Steffes james.d.steffes@enron.com, Alan Comnes acomnes@enron.com, Chris Calger ccalger@enron.com, Mary Hain mary.hain@enron.com, Joe Hartsoe Joe.Hartsoe@enron.com, Donna Fulton Dabat, state+[D@calenergy.com]

Why Some Instances are Harder to Forget?

Example of unstructured knowledge:

Original Text	James Gurney This daily weblog by Dinotopia creator James Gurney is for illustrators, plein-air painters, sketchers, comic artists, animators, art students, and writers. You'll find practical studio tips, insights into the making of the Dinotopia books, and first-hand reports from art schools and museums. CG Art Contact or by email:gurneyjourney (at) gmail.com Sorry, I can't give personal art advice or portfolio reviews. If you can, it's best to ask art questions in the blog comments. Permissions All images and text are copyright 2015 James Gurney and/or their respective owners. Dinotopia is a registered trademark of James Gurney. For use of text or images in traditional print media or for any commercial licensing rights, please email me for permission. However, you can quote images or text without
Before Unlearning	James Gurney This daily weblog by Dinotopia creator James Gurney is for illustrators, plein-air painters, sketchers, comic artists, animators, art students, and writers. You'll find practical studio tips, insights into the making of the Dinotopia books, and first-hand reports from art schools and museums. CG Art Contact or by email:gurneyjourney (at) gmail.com Sorry, I can't give personal art advice or portfolio reviews. If you can, it's best to ask art questions in the blog comments. Permissions All images and text are copyright 2015 James Gurney and/or their respective owners. Dinotopia is a registered trademark of James Gurney. For use of text or images in traditional print media or for any commercial licensing rights, please email me for permission. However, you can quote images or text without
After Unlearning	James Gurney This daily weblog by Dinotopia creator James Gurney is for illustrators, plein-air painters, sketchers, comic artists, animators, art students, and writers. You'll find practical studio tips, insights into the making of the Dinotopia books, and first-hand reports from art schools and museums. CG Art Contact or by email:gurneyjourney (at) gmail.com I've been working on a CG art project for a while now, and I've been working on it for a while now. I've been working on it for a while now, and I've been working on it for a while now. I've been working on it for a while now, and I've been working on it for a while now. I've been working on a CG art project for a while now, and I've been working on it for a while

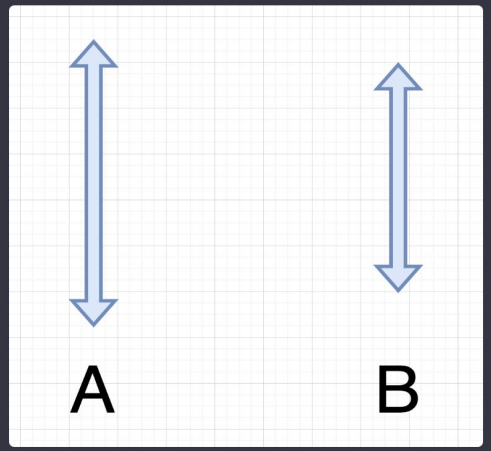
Emerging Direction 4: Knowledge in VLMs

Manling Li

Northwestern

Current VLMs have Poor Geometric Knowledge [Wang et al., 2024]
- Shape/Length

Z

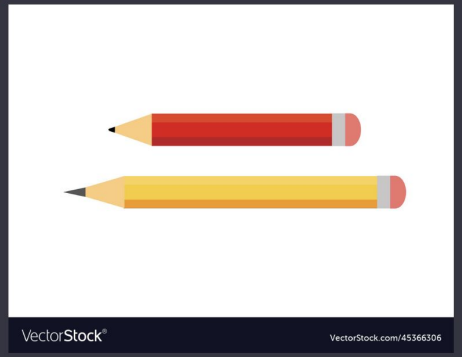


which is longer? A or B?

Both arrows, A and B, appear to be of the same length.

❌

MA **You**



which one is longer?

ChatGPT

In the image you've provided, there are two pencils laid horizontally. Their lengths cannot be directly compared since the pencils are not aligned at one end. However, if we were to visually align them by one end (either the tip or the eraser), we could determine which one appears longer. If you align them by their tips or erasers, the one that extends further out would be the longer pencil. Without alignment, any statement about their relative lengths would be speculative.

❌

Z **You**

which animal has a wider snout?

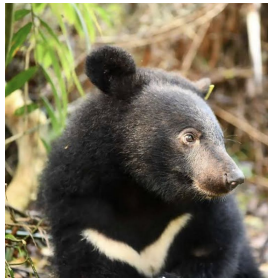
ChatGPT

The animal on the right has a wider snout compared to the one on the left.

❌

VLMs Perform Poorly on Fine-grained Concept Recognition [Kim et al., 2024]

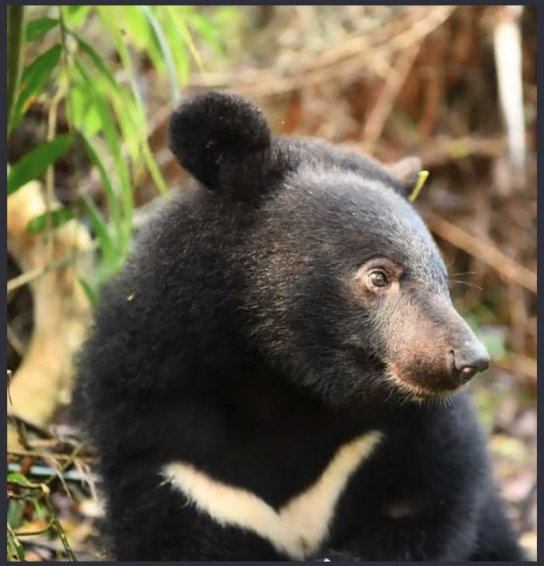
Formosan black bear



Sun bear

American black bear

You

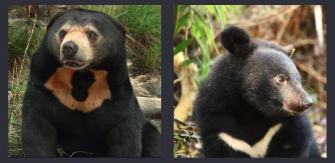


what is this?

ChatGPT

This is a photograph of a sun bear (Helarctos malayanus). Sun bears are the smallest of the bear species and are native to the tropical forests of Southeast Asia. They are easily recognizable by their short, sleek, black fur and a distinctive orange to cream colored chest patch. They have a long snout and a long tongue, which are adapted for their omnivorous diet that includes insects, honey, fruit, and occasionally small vertebrates. Sun bears are also known for their climbing skills and are often found in trees.

You



How are the two species different?

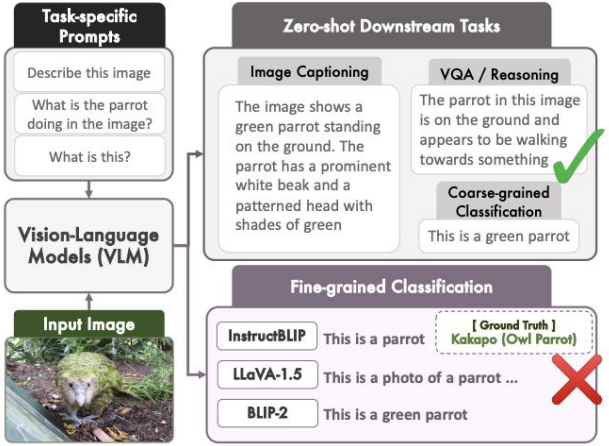
ChatGPT

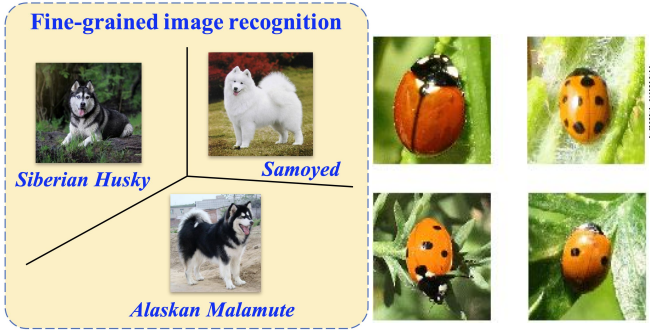
The images you've provided show two different bear species:

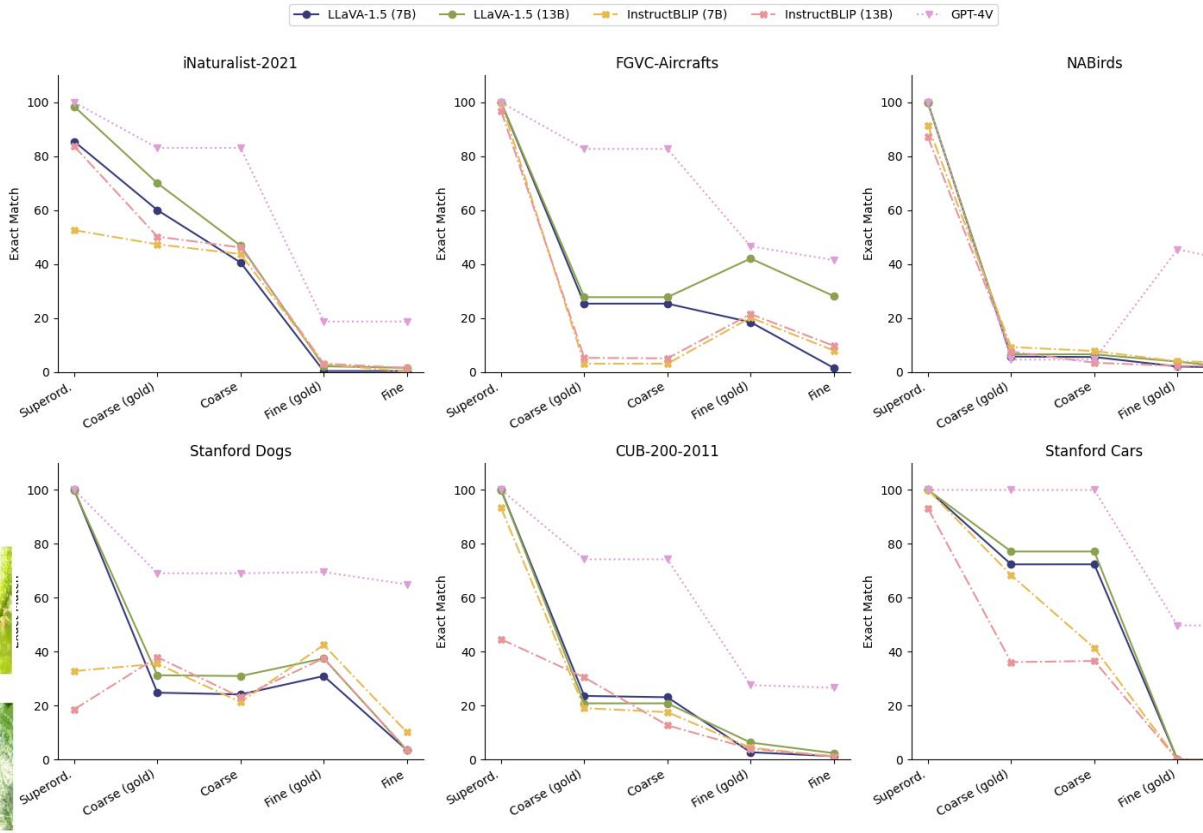
1. The first image is of a Sun Bear (Helarctos malayanus), which is native to the tropical forests of Southeast Asia. Sun bears are the smallest members of the bear family and are known for their short black fur, distinctive orange to cream colored chest patch, and long tongue, which can be over 20 cm long. They use this long tongue to extract honey and insects from tight spaces. Another characteristic feature is the relatively large, sickle-shaped claws which are adapted for climbing.
2. The second image is of an American Black Bear (Ursus americanus), which is found in North America. Black bears are medium-sized bears with a more uniform coloration of black fur, although color can vary from black to light brown. They lack the distinct chest patch of the sun bear and have a straighter profile with a larger snout and shorter claws adapted for foraging rather than climbing.

Current VLMs Perform Poorly on Fine-grained Concept Recognition [Kim et al., EMNLP2024]

- 66% Recognition Accuracy Drop when moving from coarse-grained concepts to fine-grained concepts





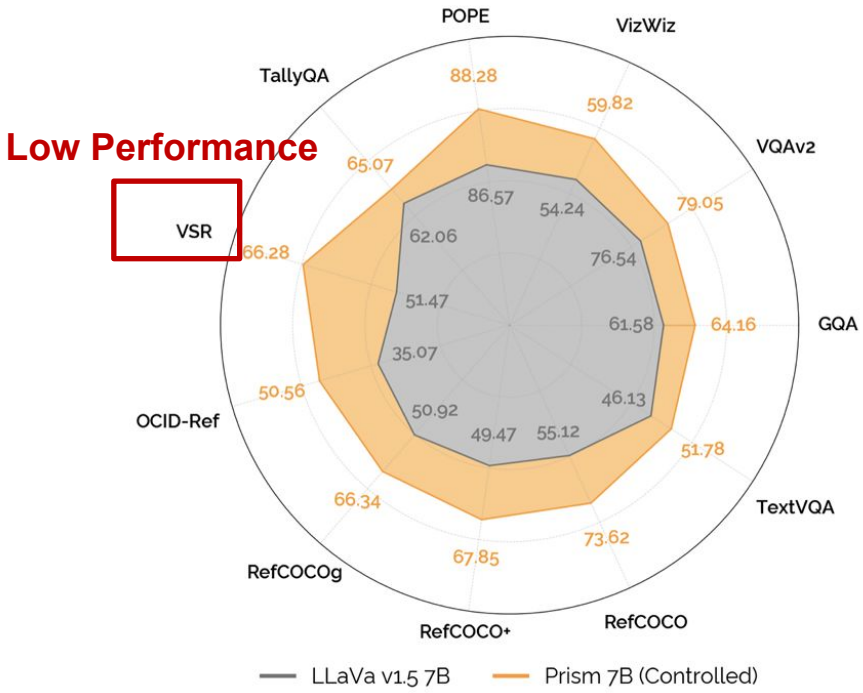


Such error can lead to serious negative effect

Timestamp: 20.28s - 26.30s

Action	Subject	Object
standing	man	water
pull	man	shark
bitten	man	shark
struggling	man	water

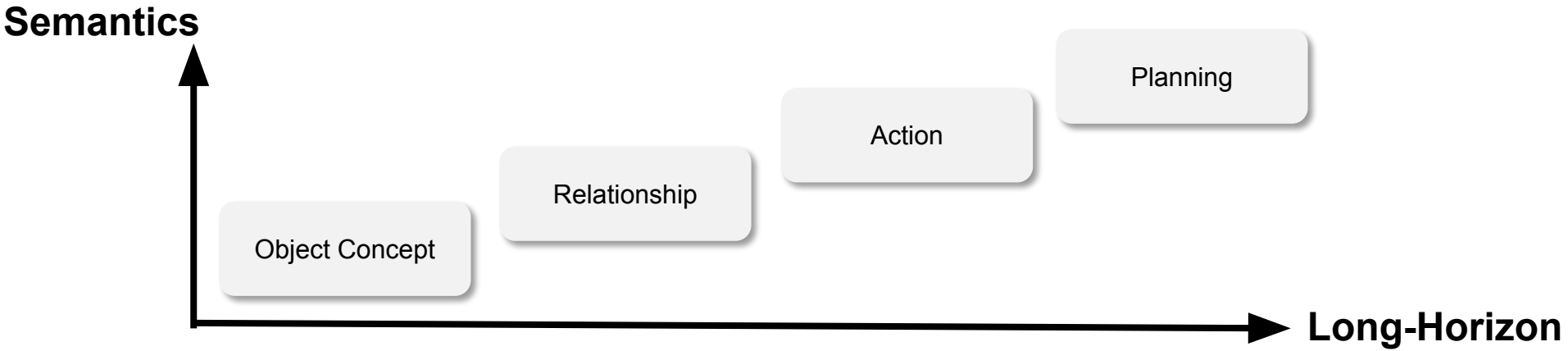
Low Performance on Visual Spatial Reasoning



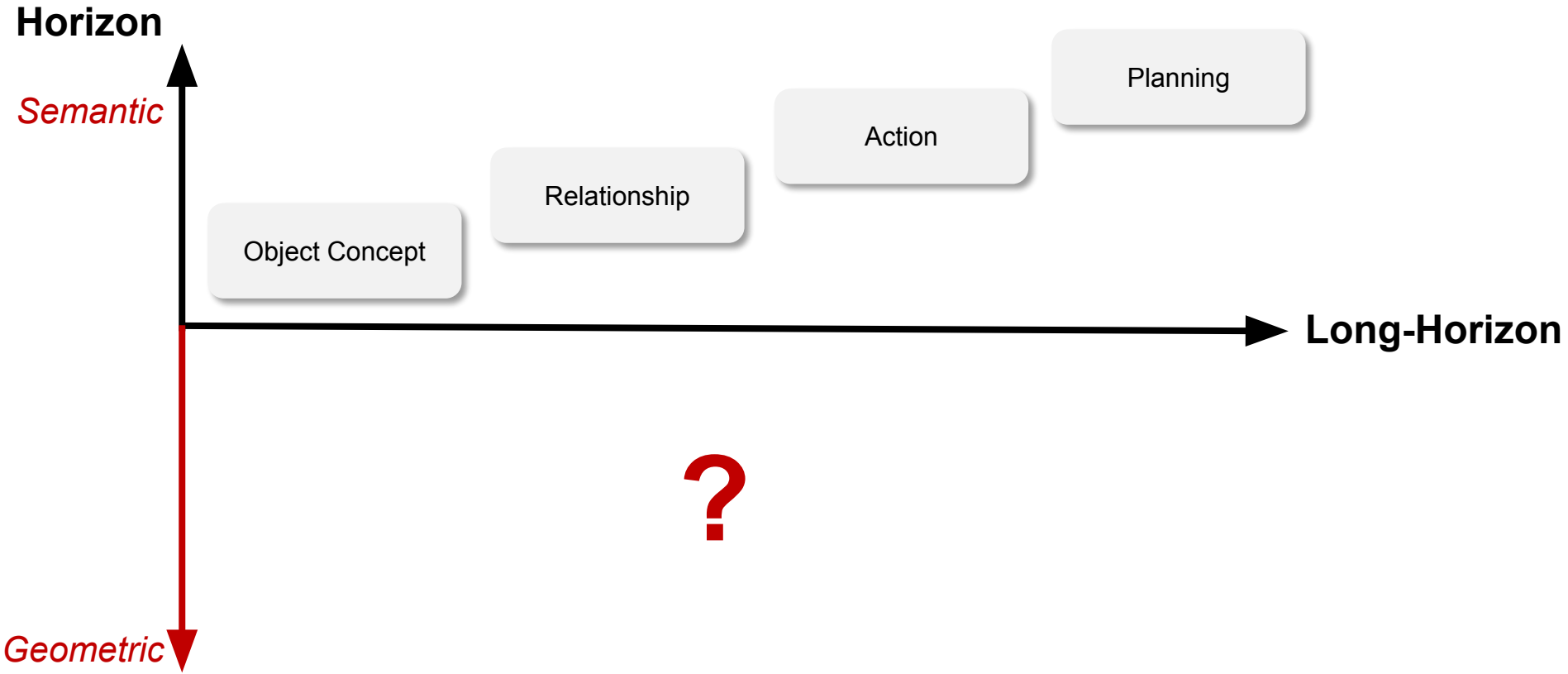
Video-Language Foundation Models

Prismatic VLMs (2024)

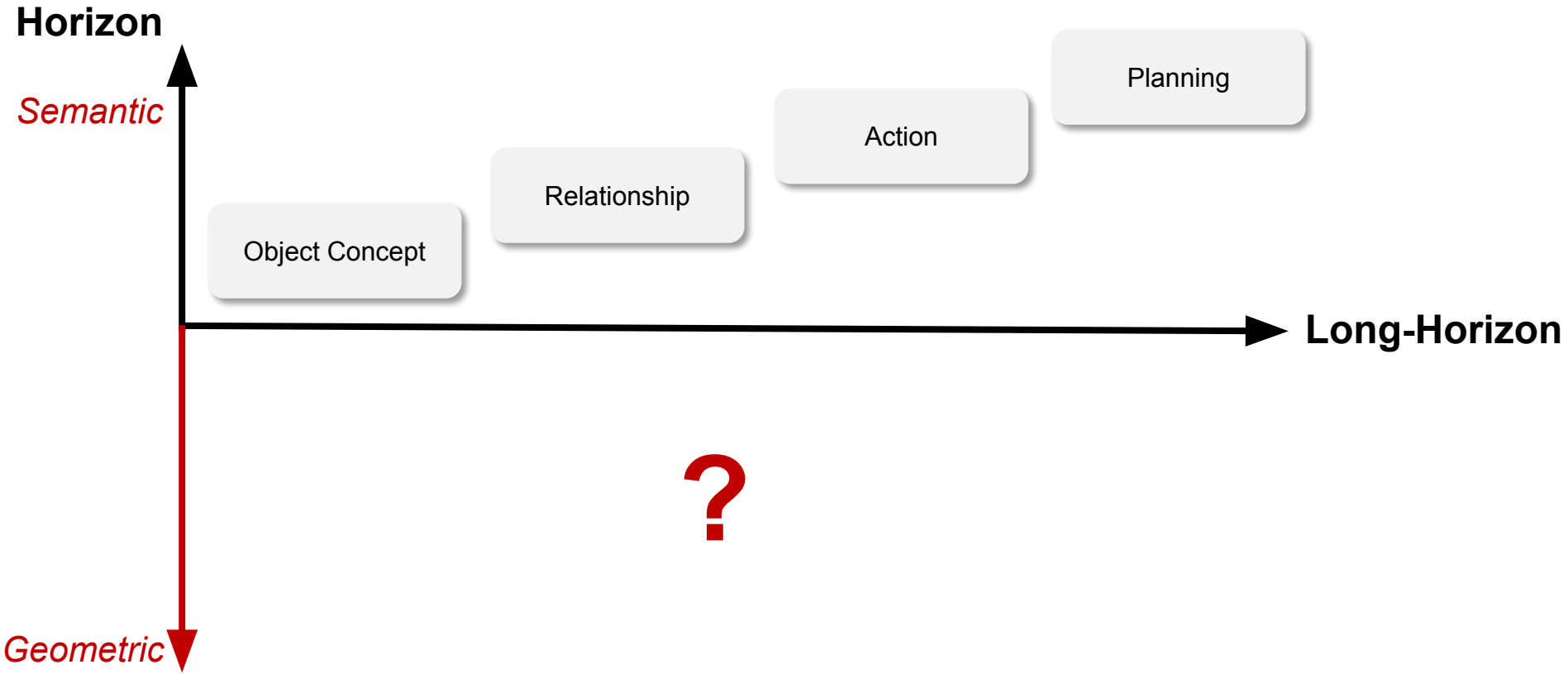
What we know about physical world



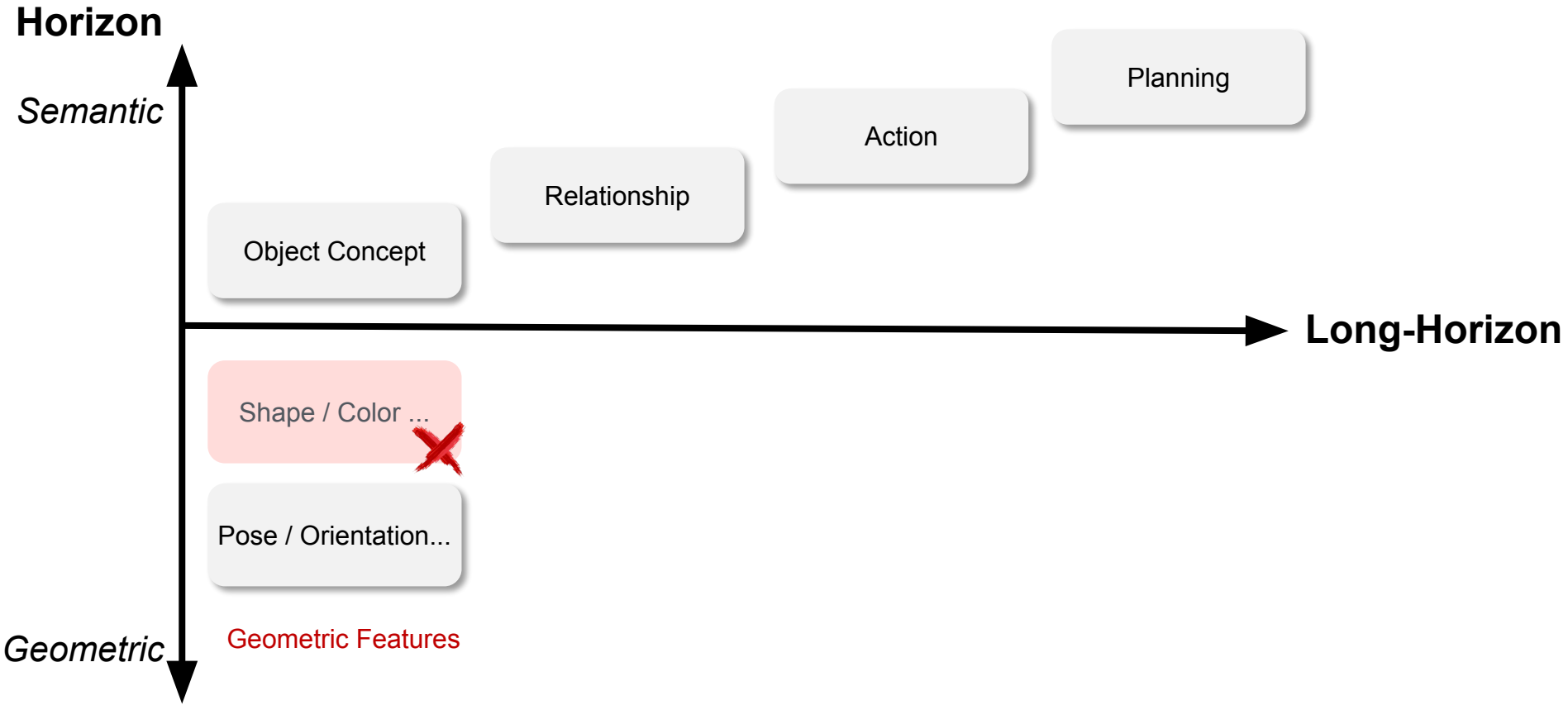
Missing knowledge about physical world



Missing knowledge about physical world

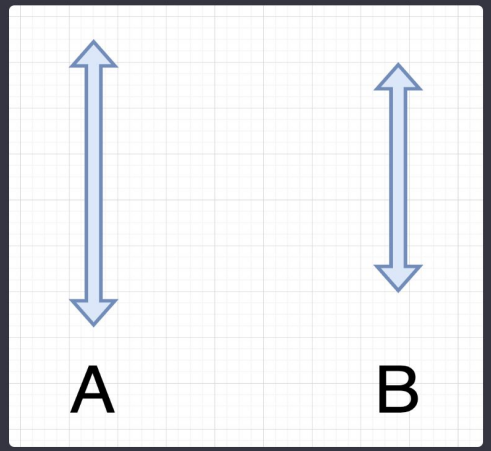


Missing knowledge about physical world



Current VLMs have Poor Geometric Knowledge [Wang et al., 2024]
- Shape/Length

Z

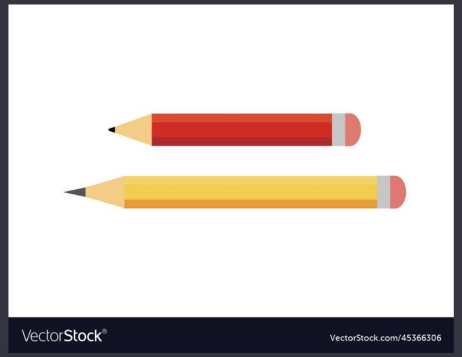


which is longer? A or B?

Both arrows, A and B, appear to be of the same length.

❌

MA **You**



which one is longer?

ChatGPT

In the image you've provided, there are two pencils laid horizontally. Their lengths cannot be directly compared since the pencils are not aligned at one end. However, if we were to visually align them by one end (either the tip or the eraser), we could determine which one appears longer. If you align them by their tips or erasers, the one that extends further out would be the longer pencil. Without alignment, any statement about their relative lengths would be speculative.

❌

Z **You**

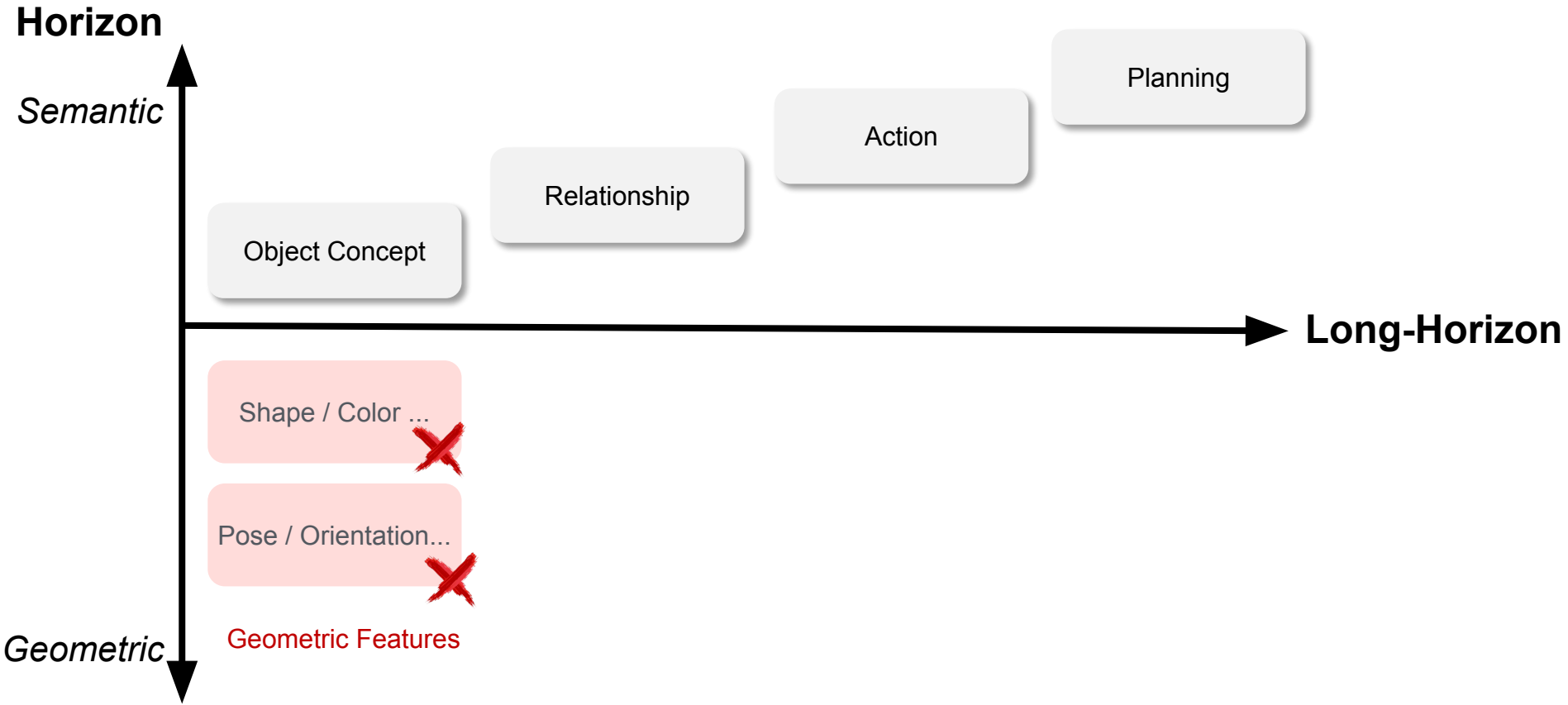
which animal has a wider snout?

ChatGPT

The animal on the right has a wider snout compared to the one on the left.

❌

Missing knowledge about physical world



Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Do the individuals in the picture face the front or the back?

The individuals in the picture are **facing away from the camera**, so we see them from the back.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: Is the person in the picture on the grass or on the gravel path?

The person in the image appears to be running **on a gravel path**. You can discern this from the texture and color contrast between the light-colored path and the green grass surrounding it.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

Q: Is the school bus parked facing the camera or away from the camera?

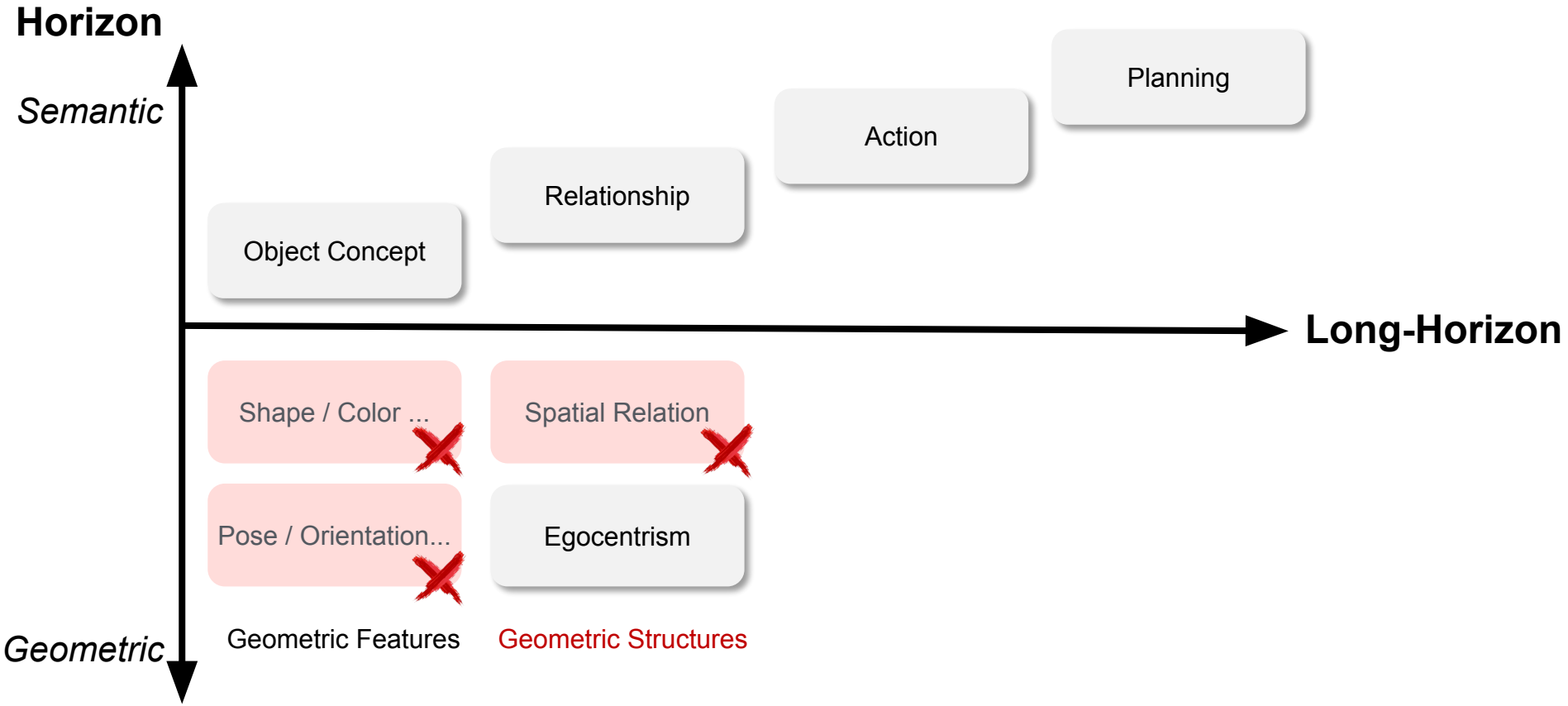
The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.

Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

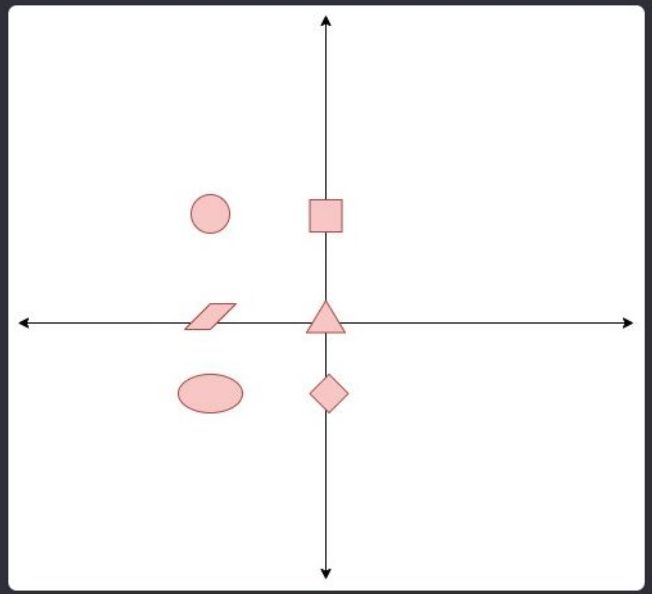
Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." CVPR2024

Missing knowledge about physical world



Current VLMs have Poor Geometric Knowledge [Wang et al., 2024] - Shape/Length

BL



In the image there is a circle, oval, square, triangle, parallelogram, and rhombus. Which shape is to the left of the rhombus?

The parallelogram is to the left of the rhombus in the image.

BL Where is the rhombus in the image?

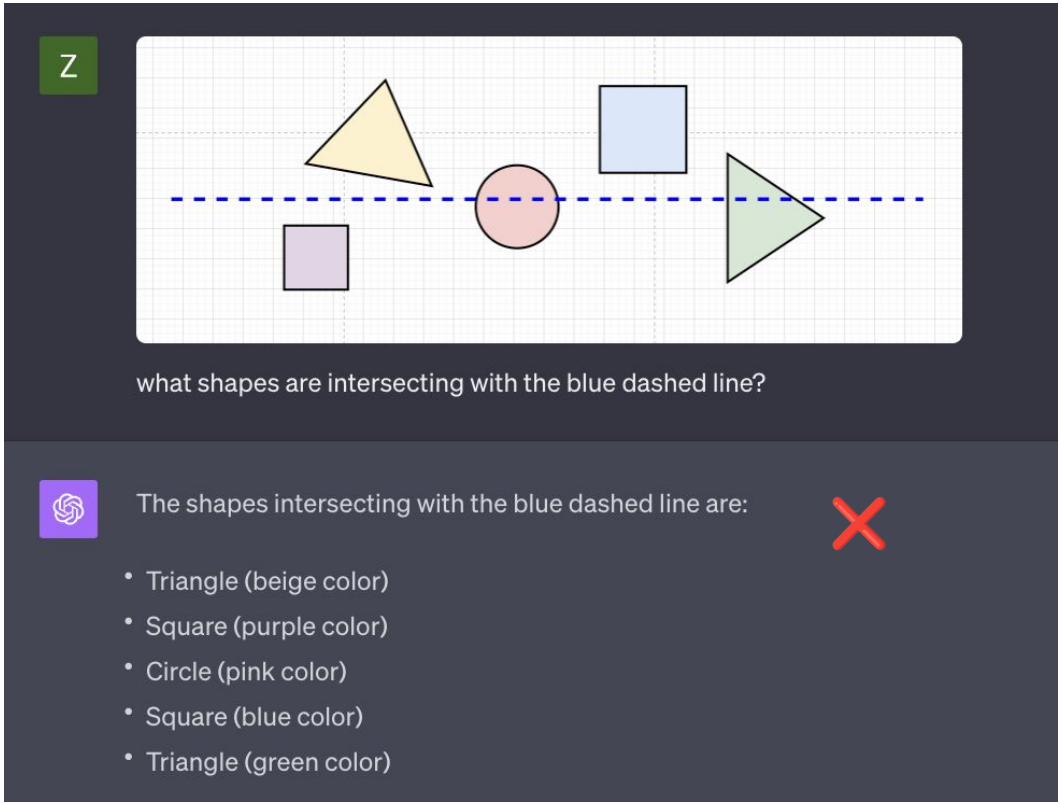
The rhombus is located on the vertical axis, below the center point in the image.

BL What shape is to its left?

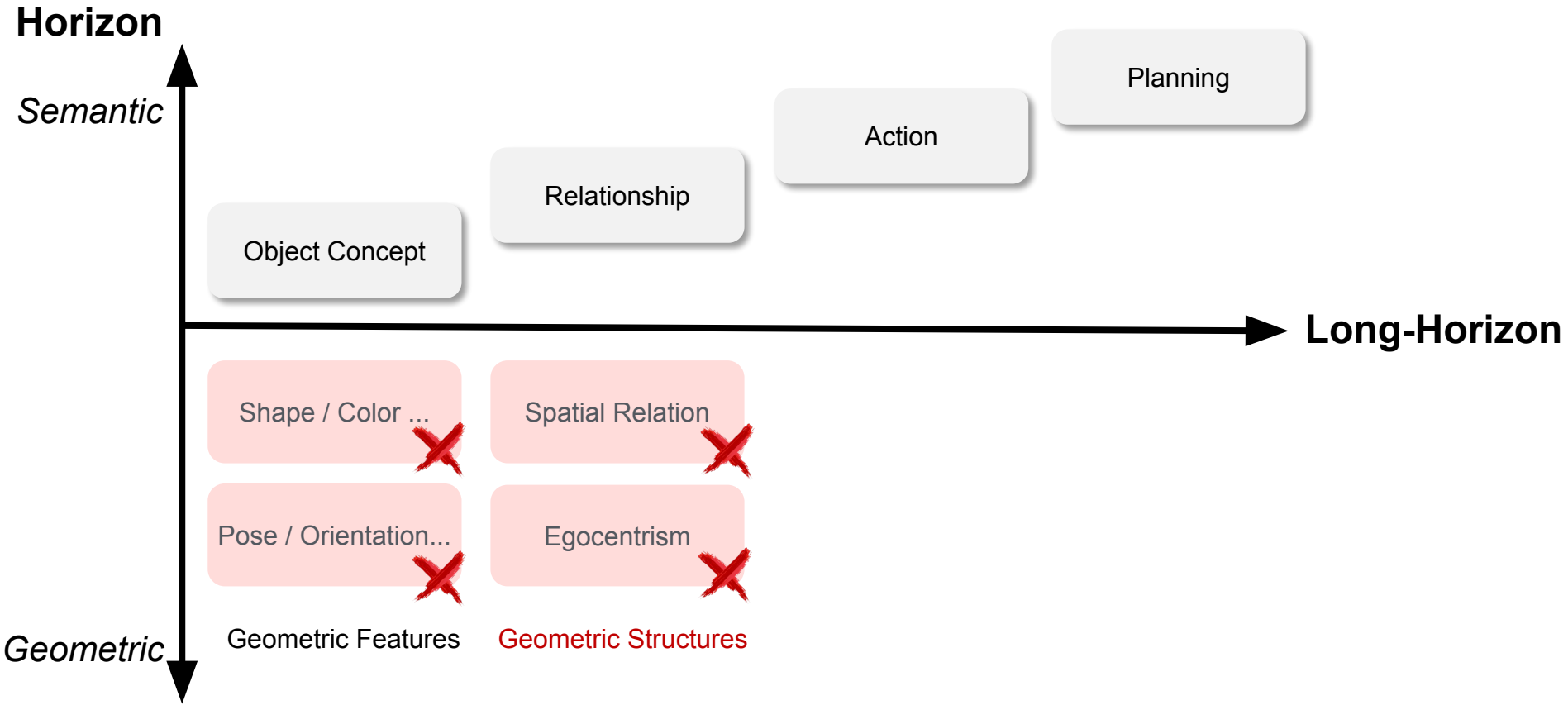
There is no shape directly to the left of the rhombus in the image.

Image embedding encodes spatial information poorly.

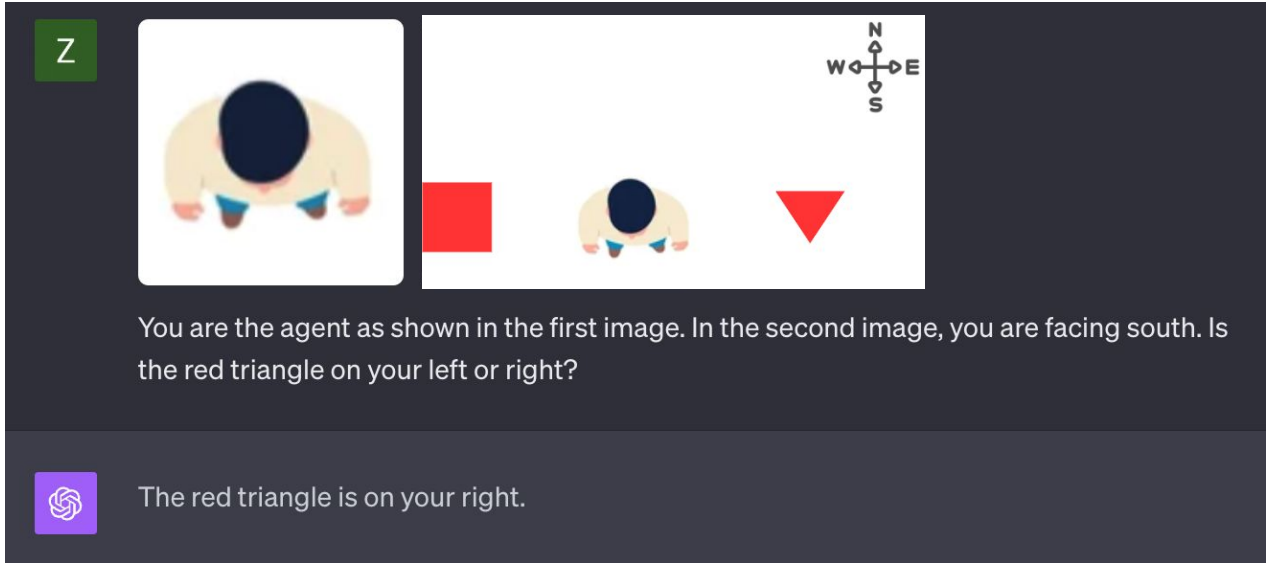
Current VLMs have Poor Geometric Knowledge [Wang et al., 2024]
- Spatial Relationship



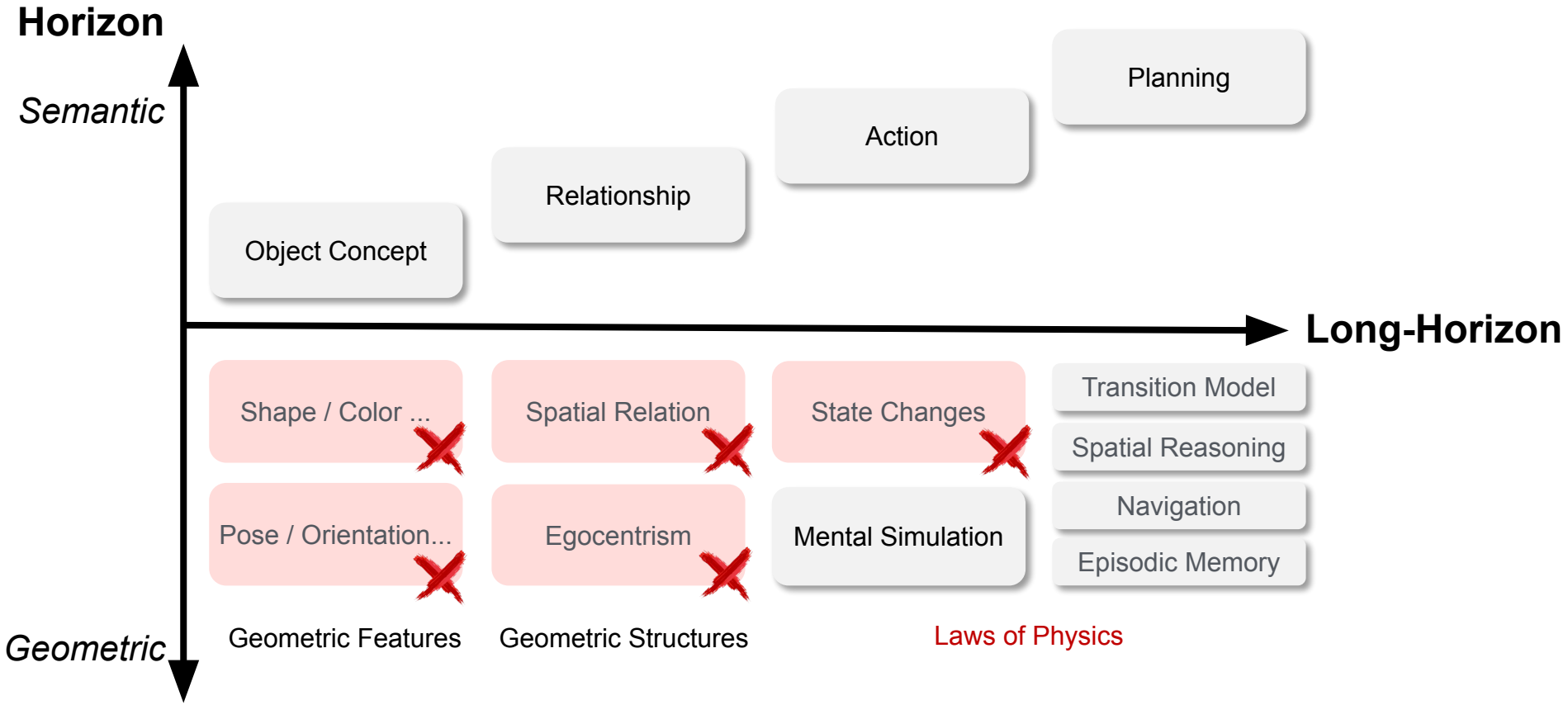
Missing knowledge about physical world



What is missing?
- Ego-centric View / Visual Theory of Mind



Missing knowledge about physical world



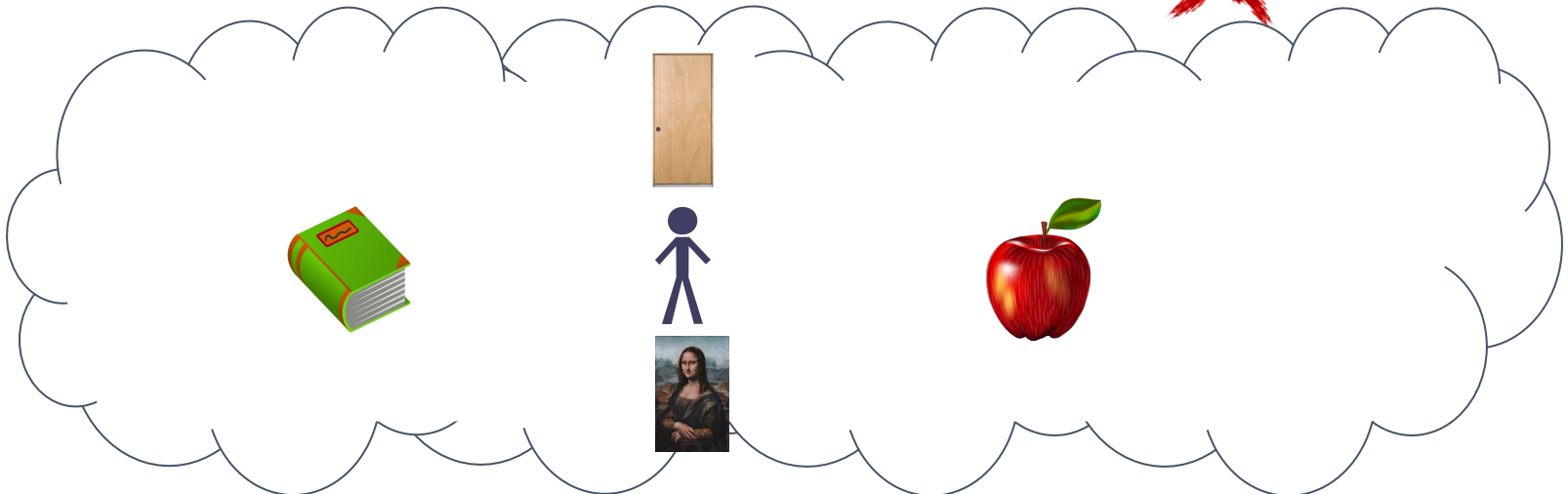
What is missing?
- Mental Simulation

User

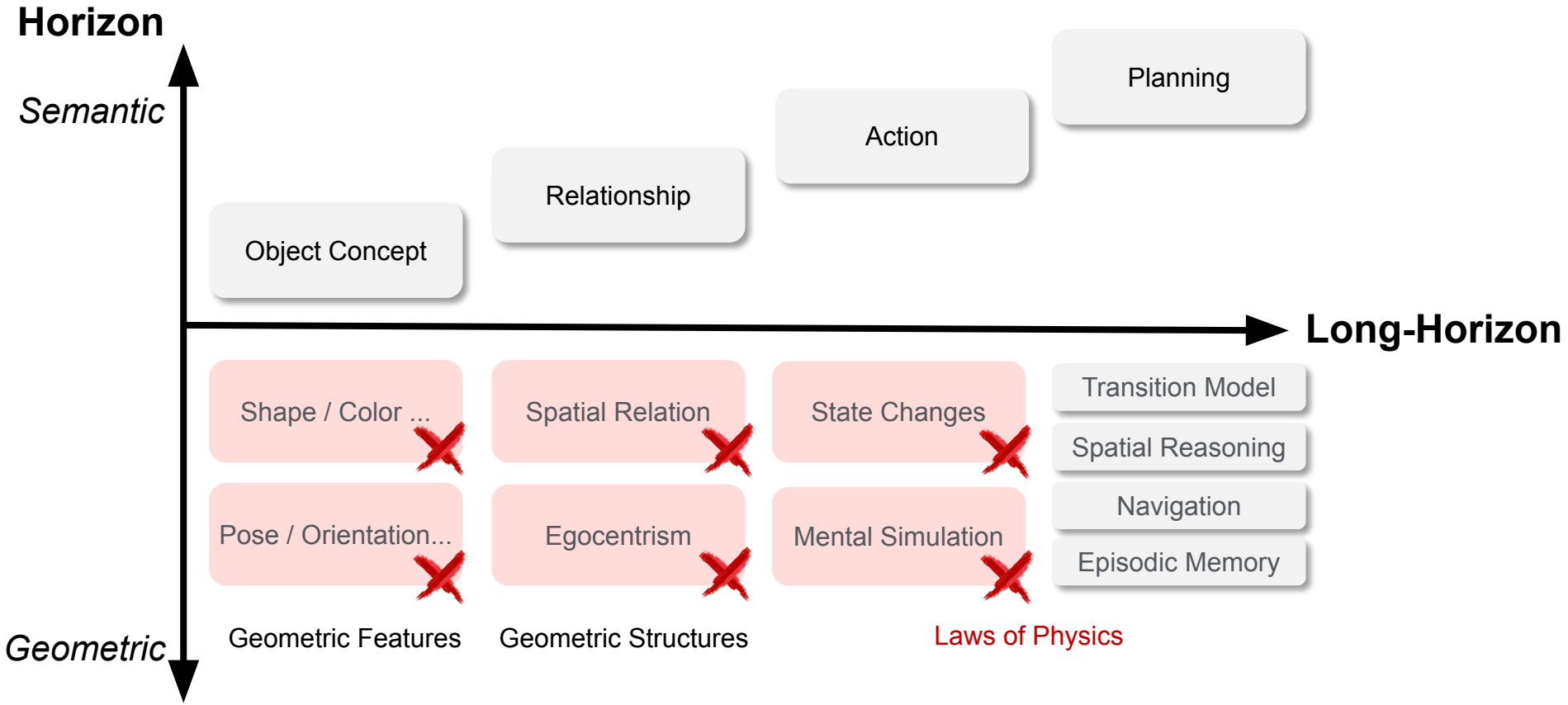
You are in a room with a book on your left, a door in front of you, an apple on your right, and a painting behind you. If you turn right, what will behind you?

AI

If you turn right, the door that was originally in front of you will now be behind you.



Missing knowledge about physical world



What is missing?
- Planning, State Changes, and Mental Simulation

Current **LLMs** still **fall short** on understanding concepts involving complex **physical interactions**.

Place a blue chair behind a red chair. Then, put a yellow chair behind the blue chair. Then, put a book on top of the chair that in front of the blue chair.
Question: What chair is the book on?
Answer: The book is on the yellow chair.

GPT4 failed example

Mental models, which can be viewed as **internal representations of the physical world**, enable **humans** to understand such concepts with ease.

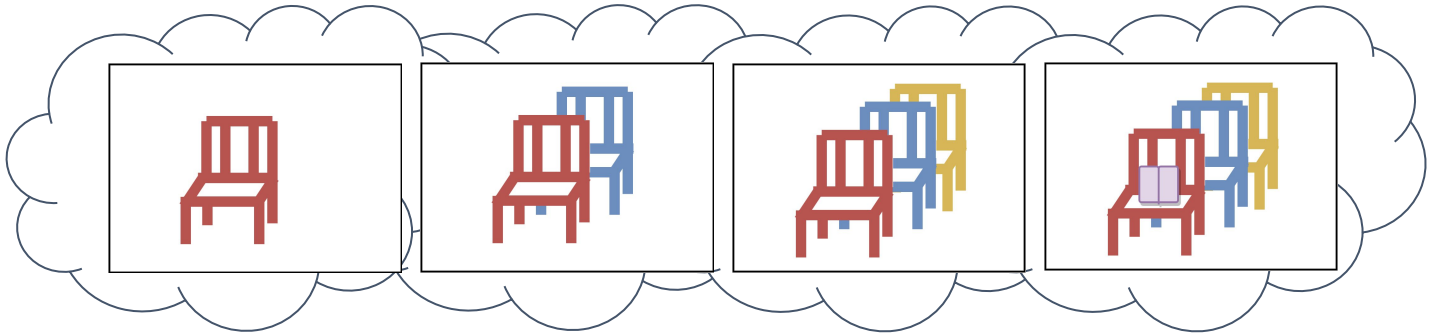
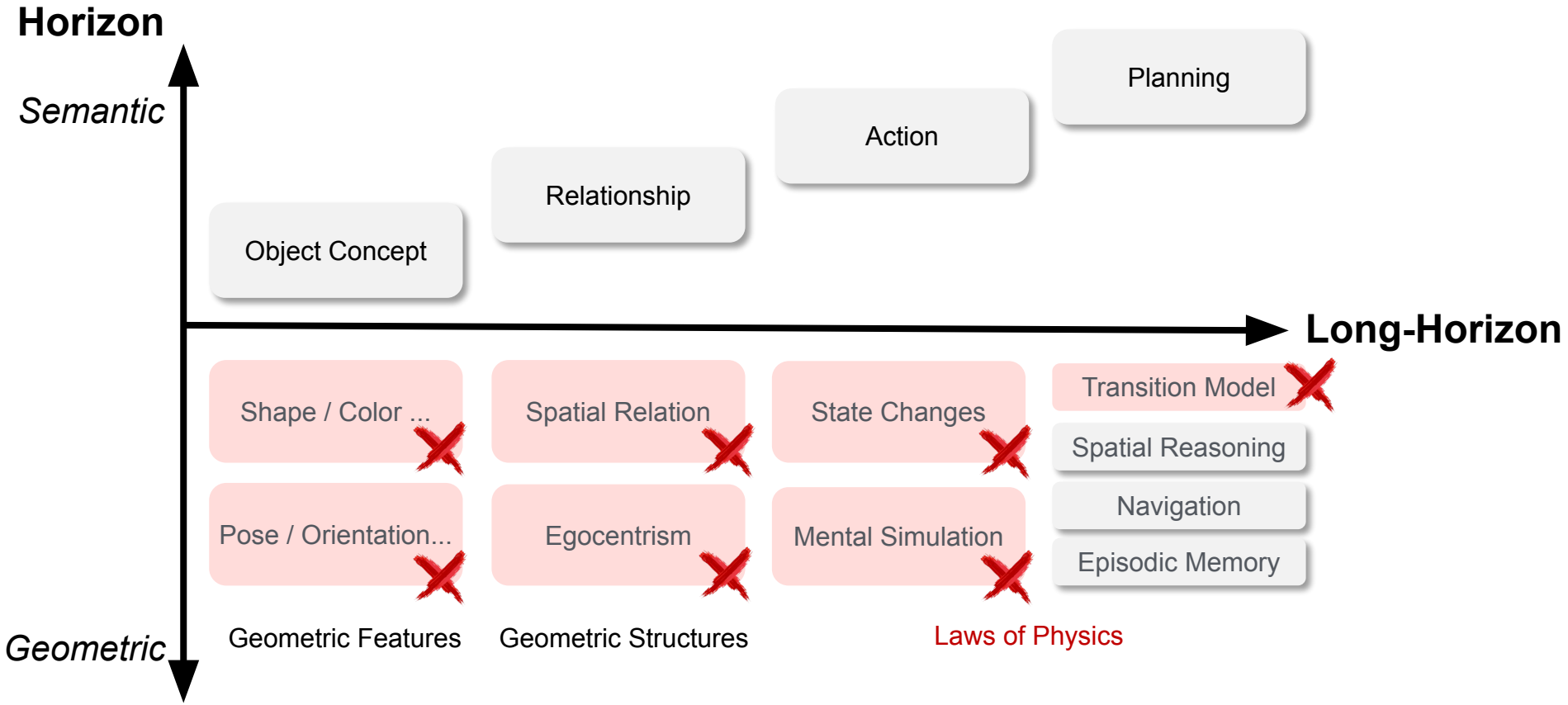


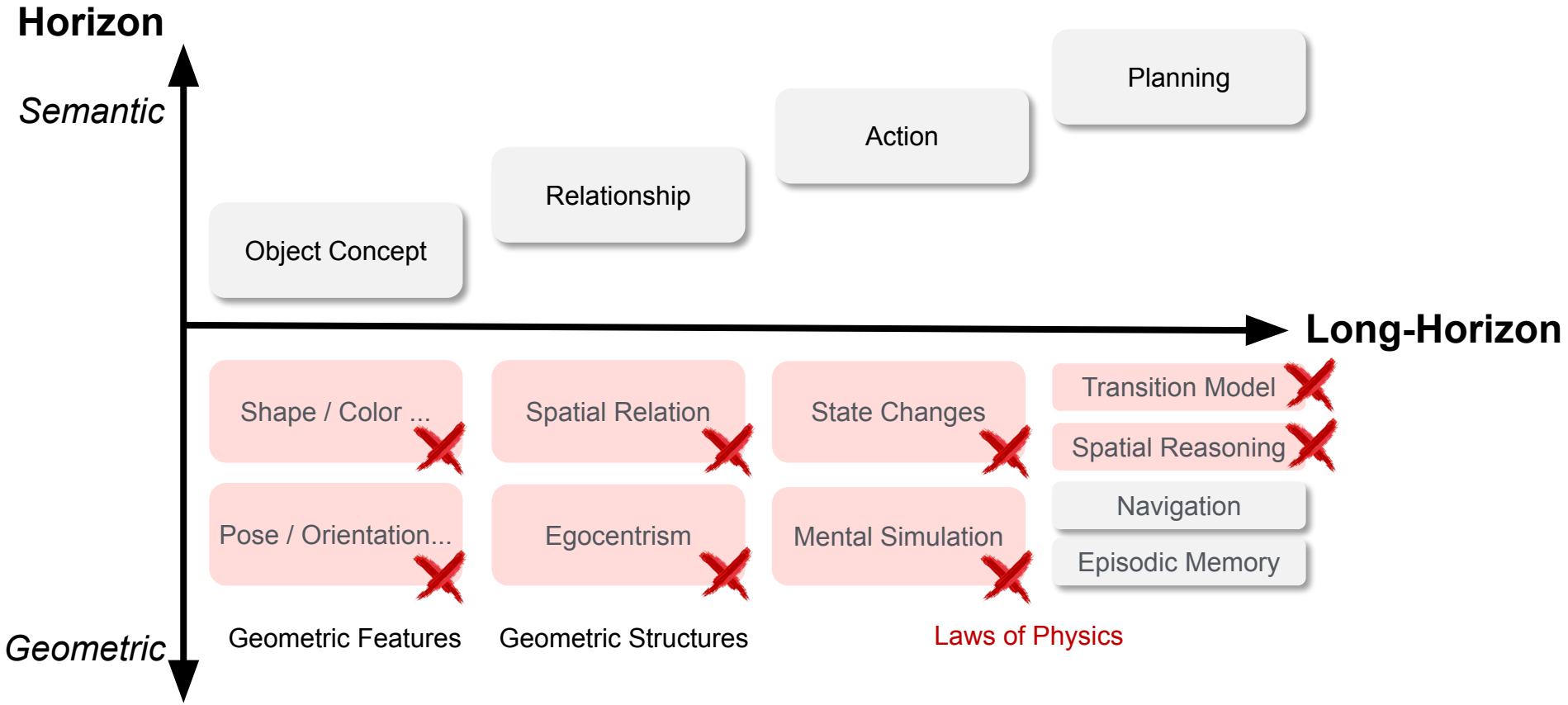
Illustration Of Mental Simulation

Figure credits: Zhenhailong Wang

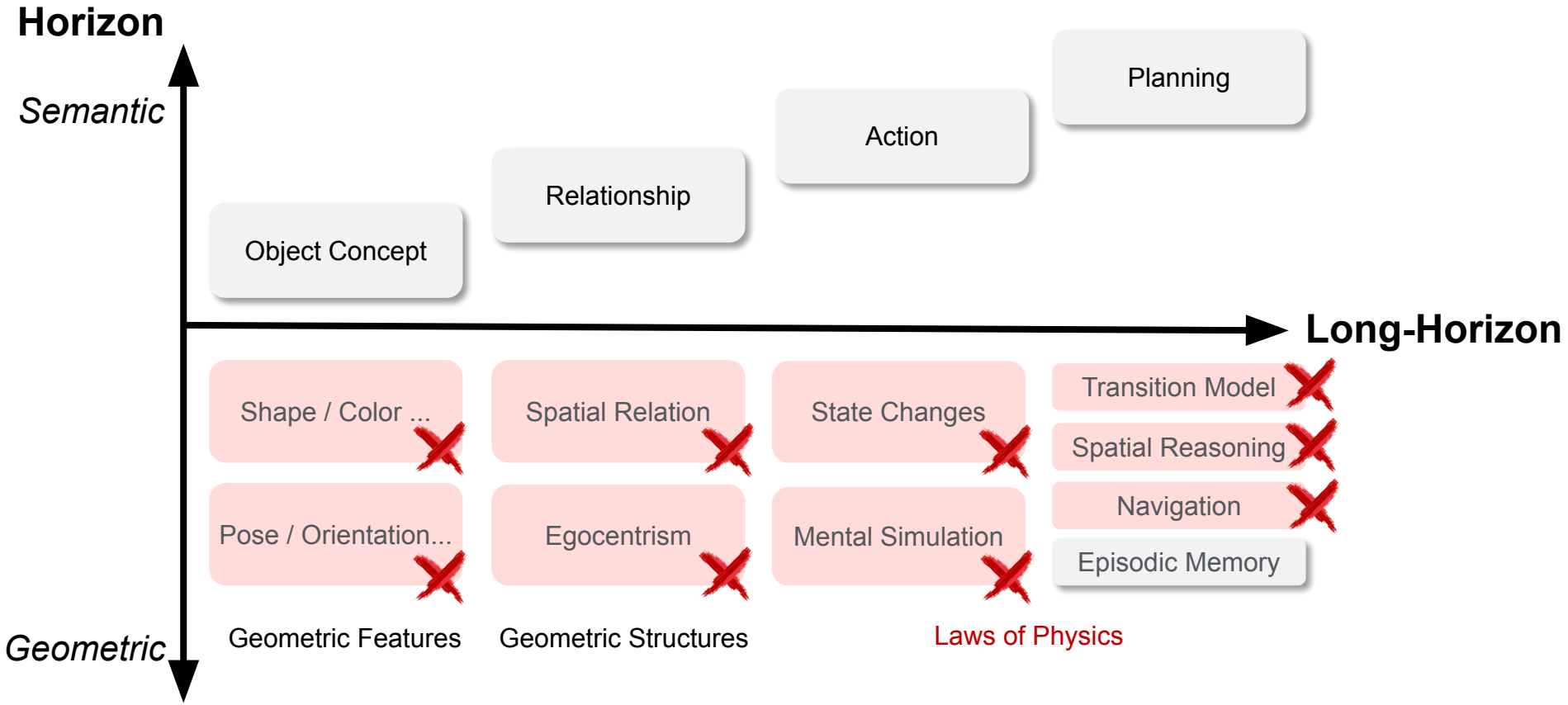
Missing knowledge about physical world



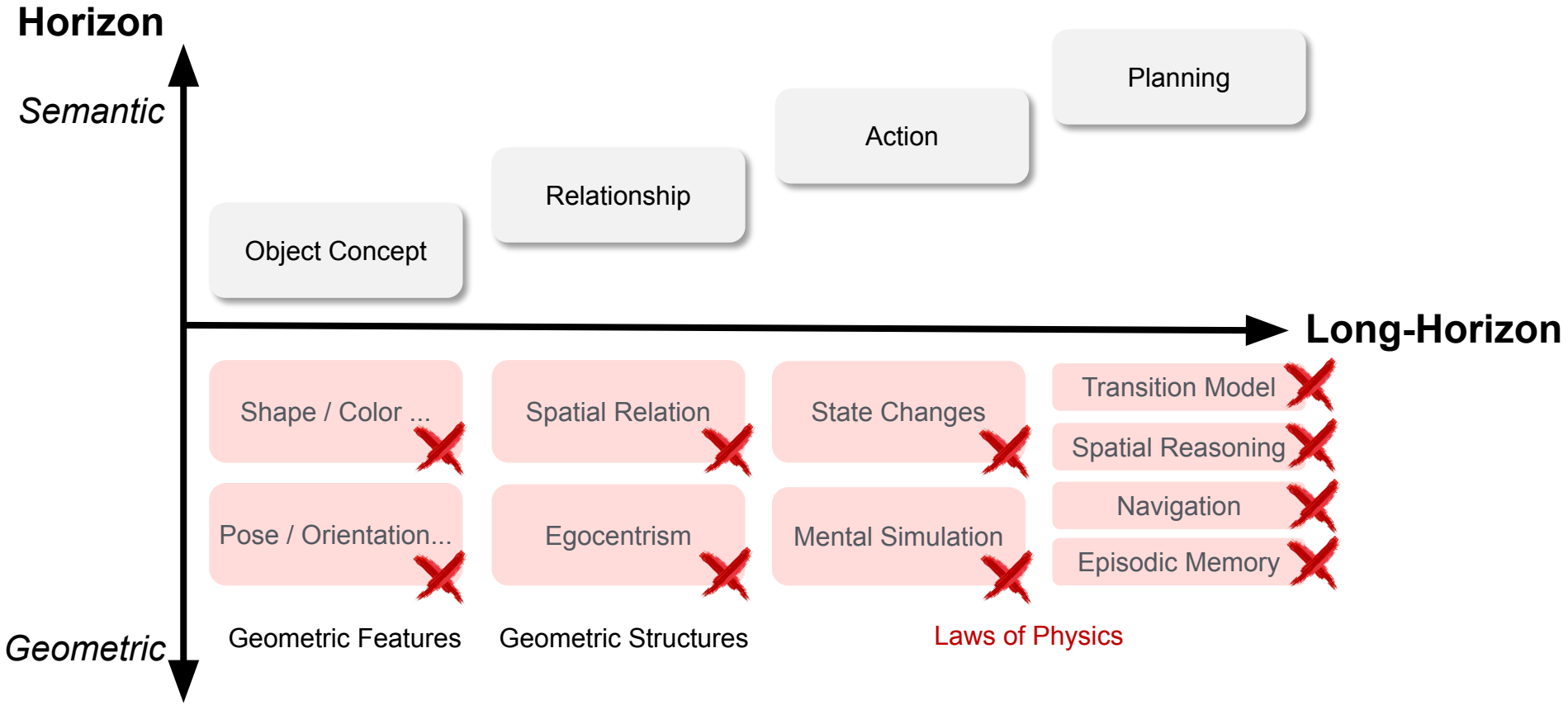
Missing knowledge about physical world



Missing knowledge about physical world



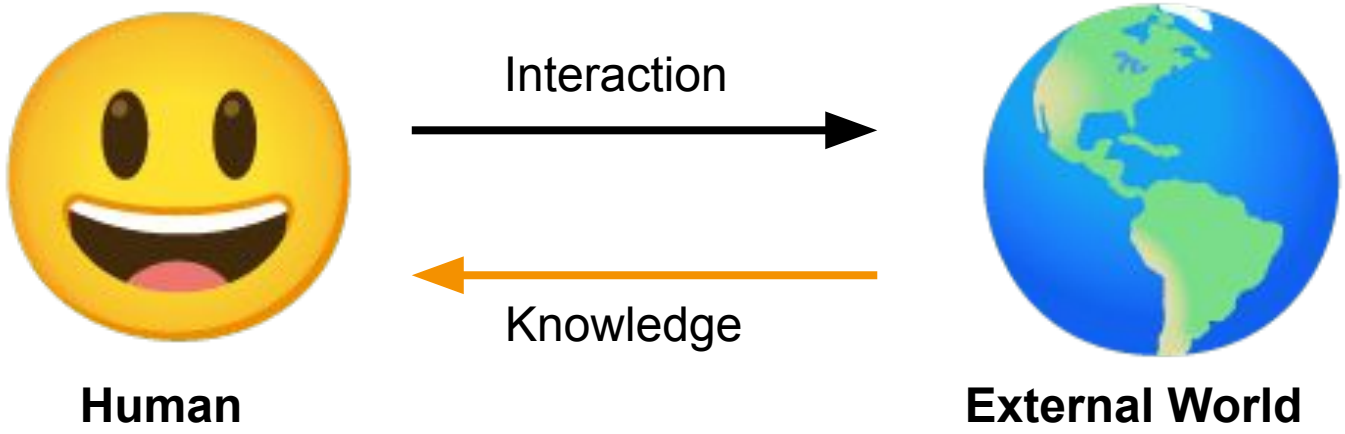
Missing knowledge about physical world



Current LMMs fall short on **Geometric** Info.

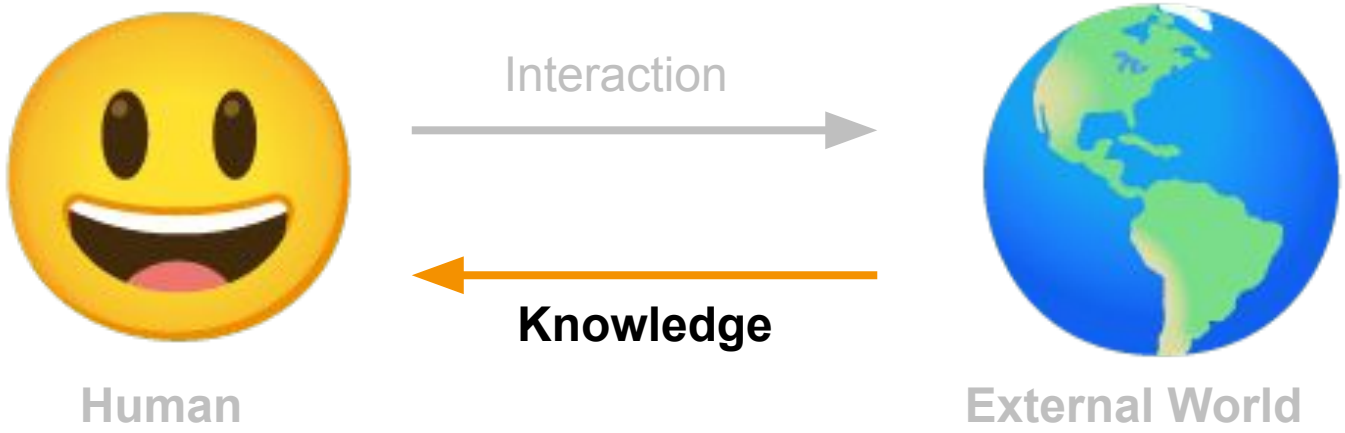
Why?

Humans learn knowledge through interactions



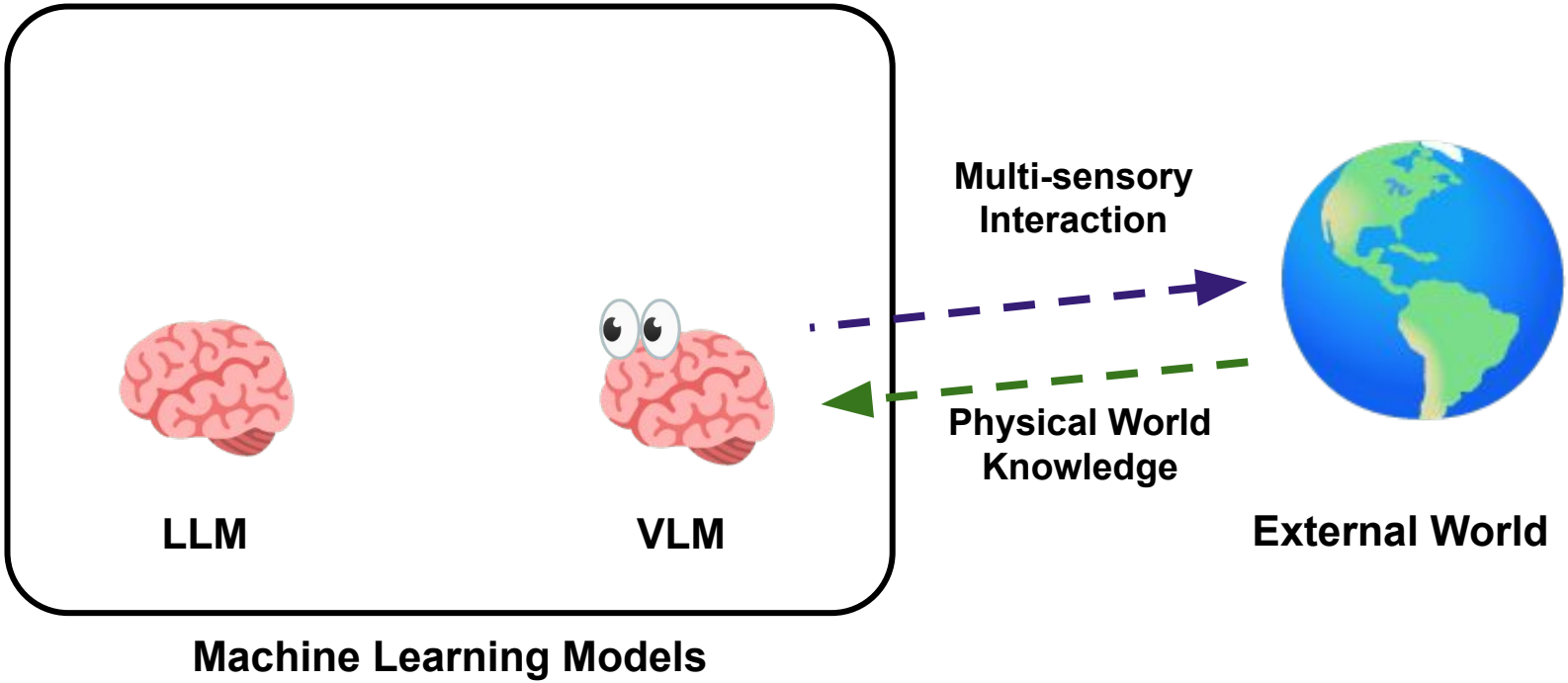
Credits: Zhenhailong

Humans learn knowledge through interactions



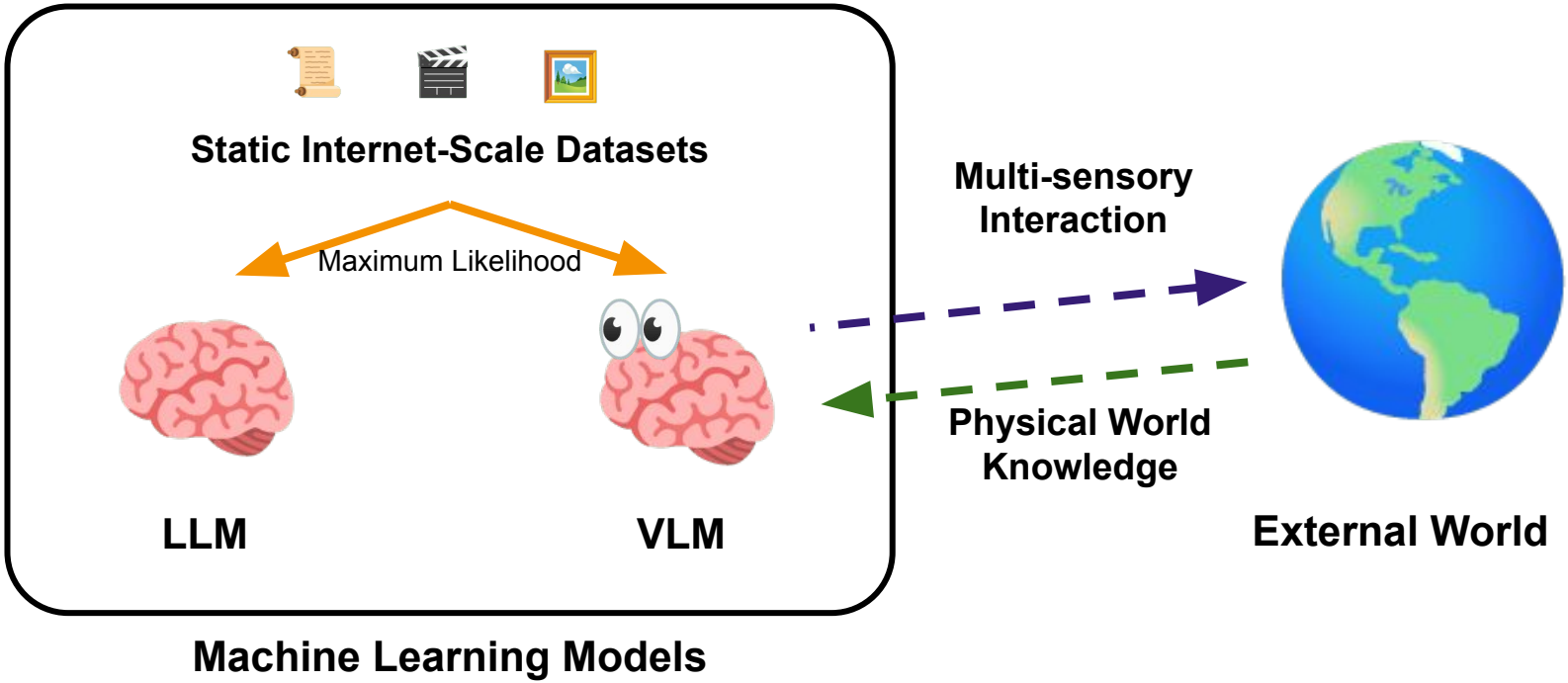
Credits: Zhenhailong

Machines learn knowledge w/o interactions



Credits: Zhenhailong

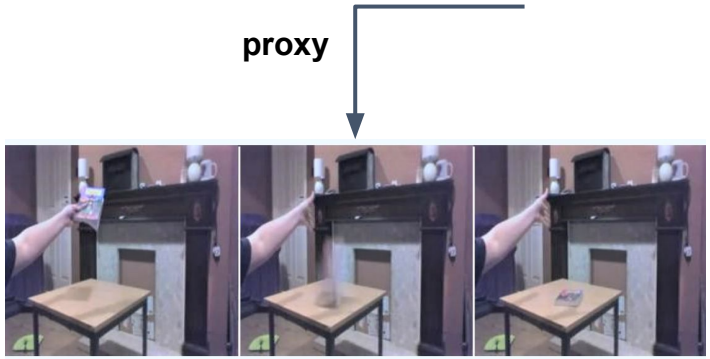
Machines learn knowledge w/o interactions



Credits: Zhenhailong

Static Internet-Scale Datasets

Video: A “Visual Recording” of World State Changes



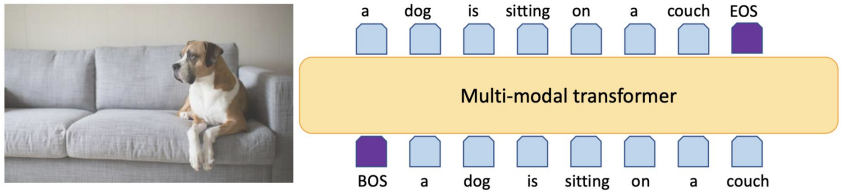
“Book falling like a rock”

Video-Language Datasets

Credits: Zhenhailong

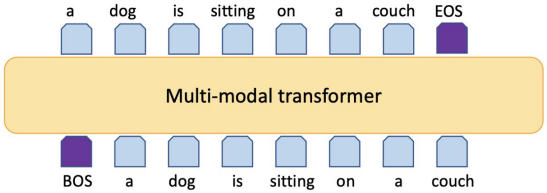
Language Vision: **Linearize** Everything into Language-Like **Sequences** and then Train Language Models

Image

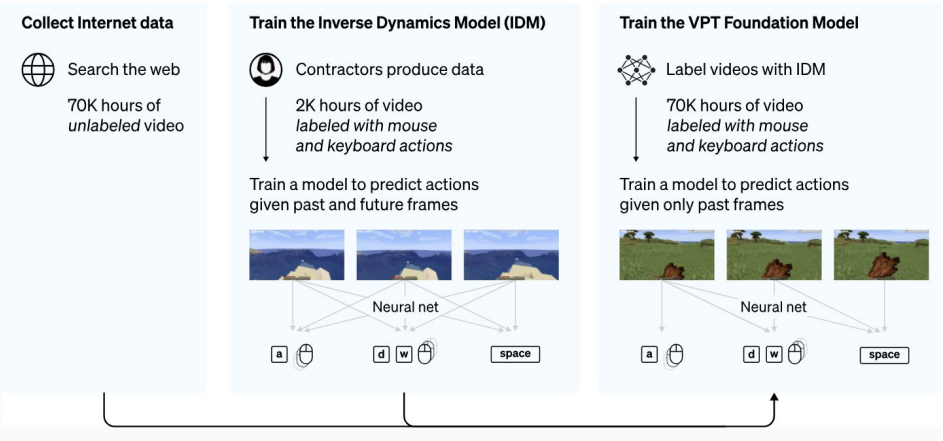


Language □ Vision: Linearize Everything into Language-Like Sequences and then Train Language Models

Image

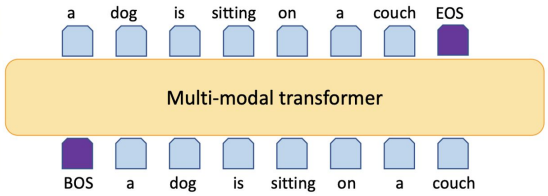


Video

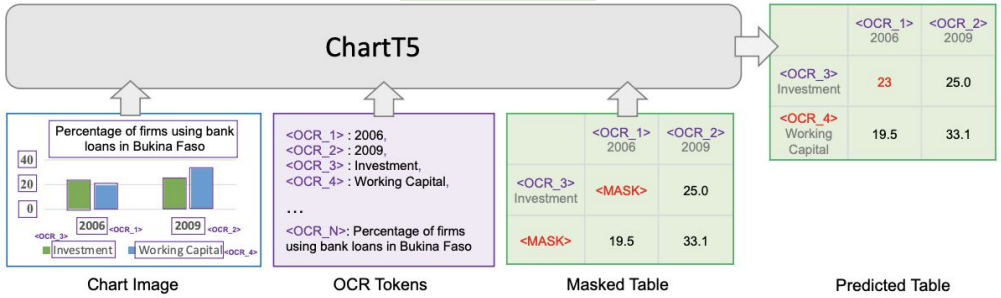


Language □ Vision: Linearize Everything into Language-Like Sequences and then Train Language Models

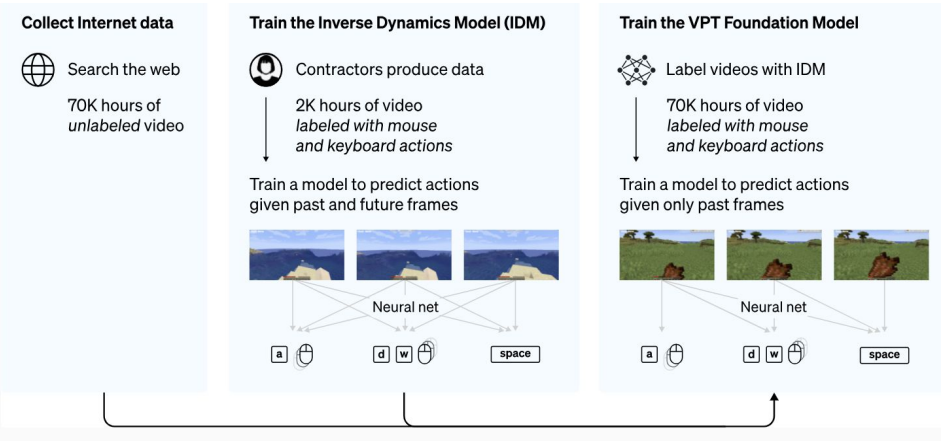
Image



Chart

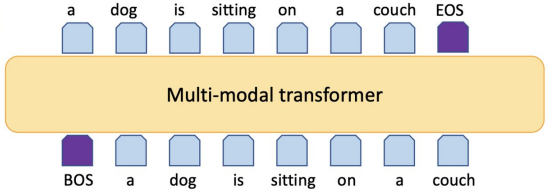


Video

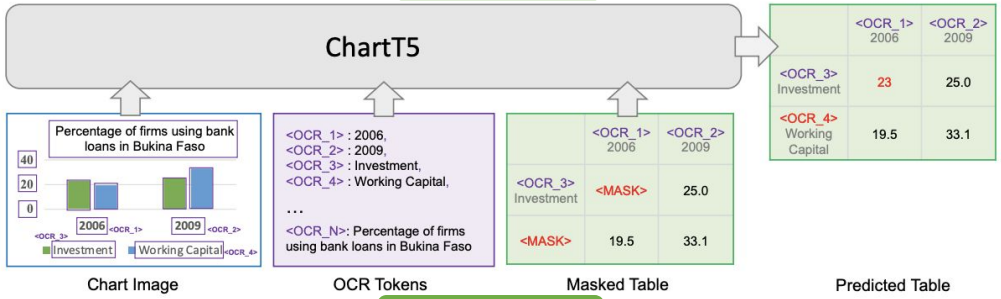


Language □ Vision: Linearize Everything into Language-Like Sequences and then Train Language Models

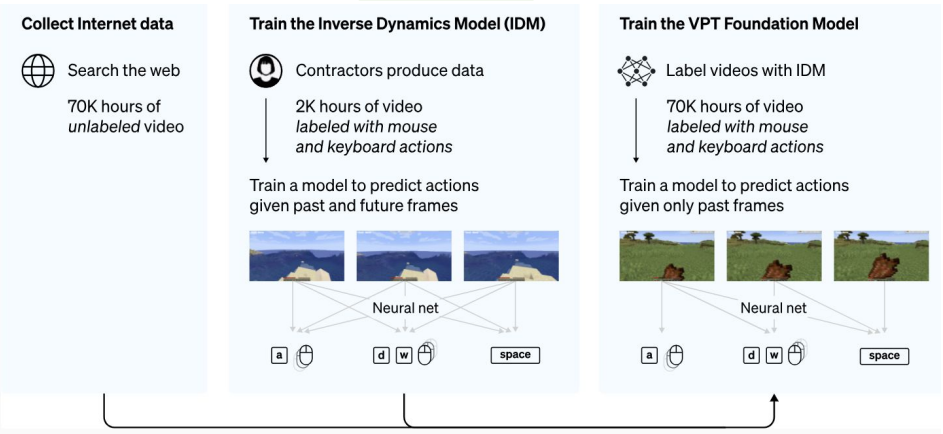
Image



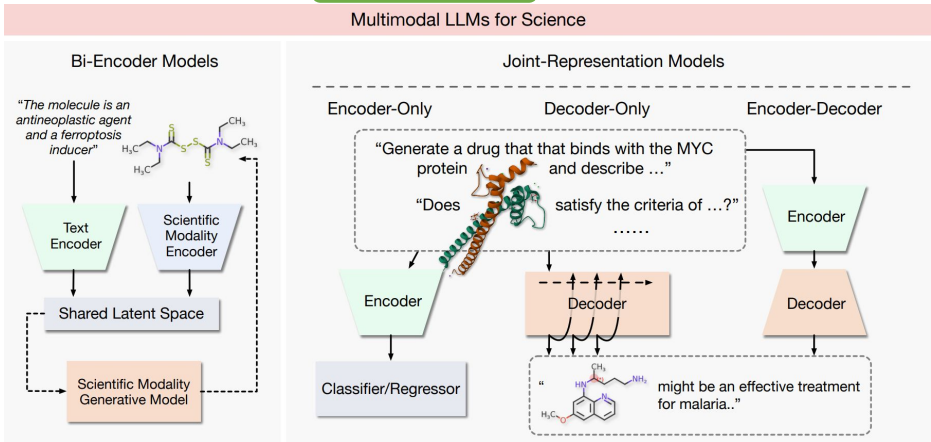
Chart



Video

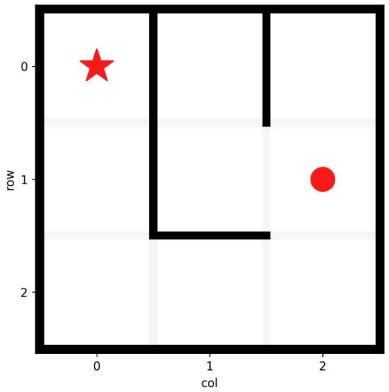


Molecule



What is the bottleneck?

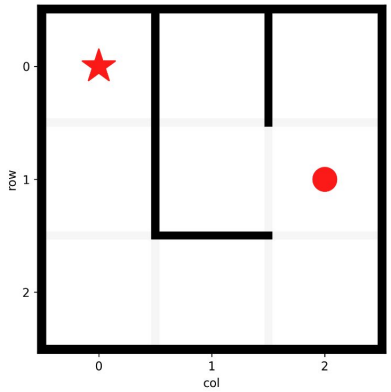
Images → Reasoning 



More examples

What is the bottleneck?

Images → Perception → Reasoning



The maze can be described in a 3×3 grid form as follows:

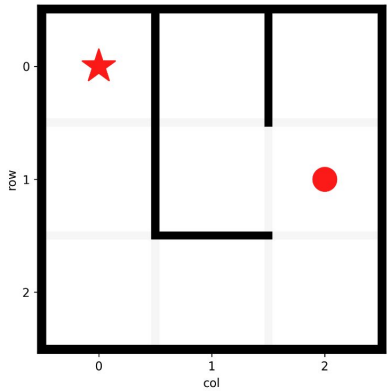
diff

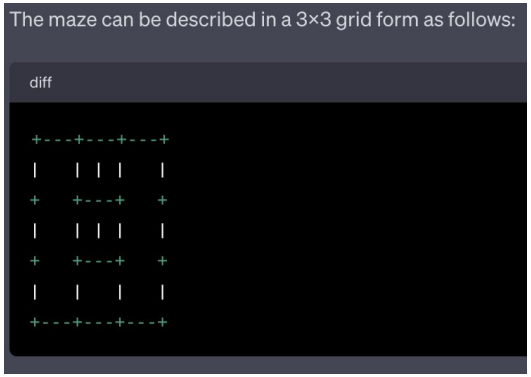
```
+---+---+---+
|   | | |   |
+   +---+   +
|   | | |   |
+   +---+   +
|   |   |   |
+---+---+---+
```

More examples

What is the bottleneck?

Images → Perception → Abstraction / Parsing → Reasoning

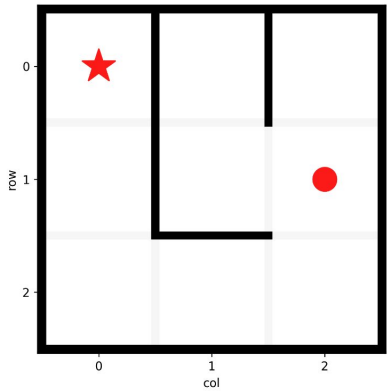




More examples

What is the bottleneck?

Images → Perception → Abstraction / Parsing → Reasoning



The maze can be described in a 3×3 grid form as follows:

diff

```
+---+---+---+
|   | | |   |
+   +---+   +
|   | | |   |
+   +---+   +
|   |   |   |
+---+---+---+
```

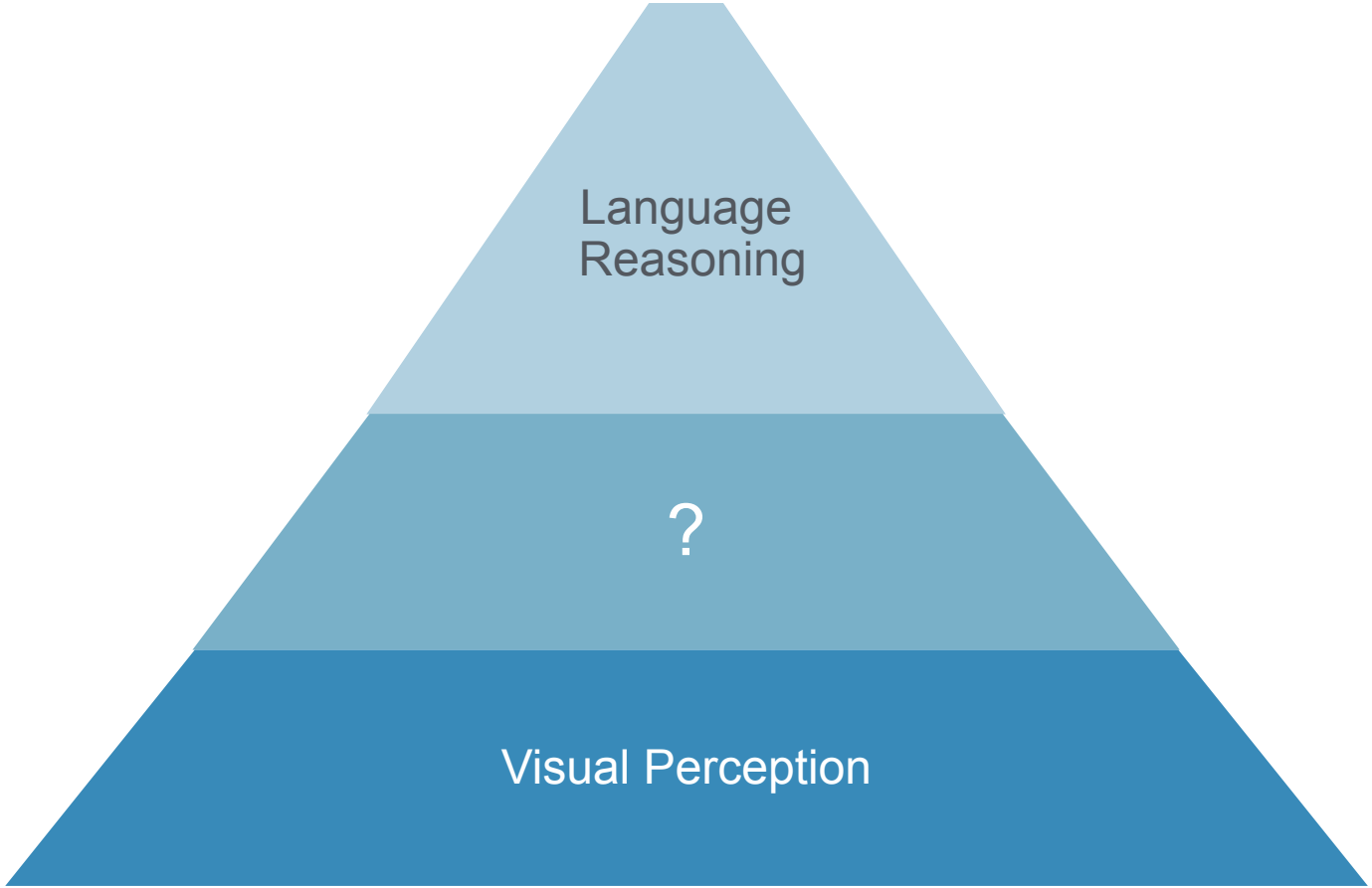
Adjacency List (after adding paths and removing walls):

makefile — Copy code

```
(2,0): (1,0)
(1,0): (0,0), (2,0)
(0,0): (0,1), (1,0)
(0,1): (0,0), (0,2)
(0,2): (0,1), (1,2)
(1,2): (0,2), (2,2)
(2,2): (1,2)
```

More examples

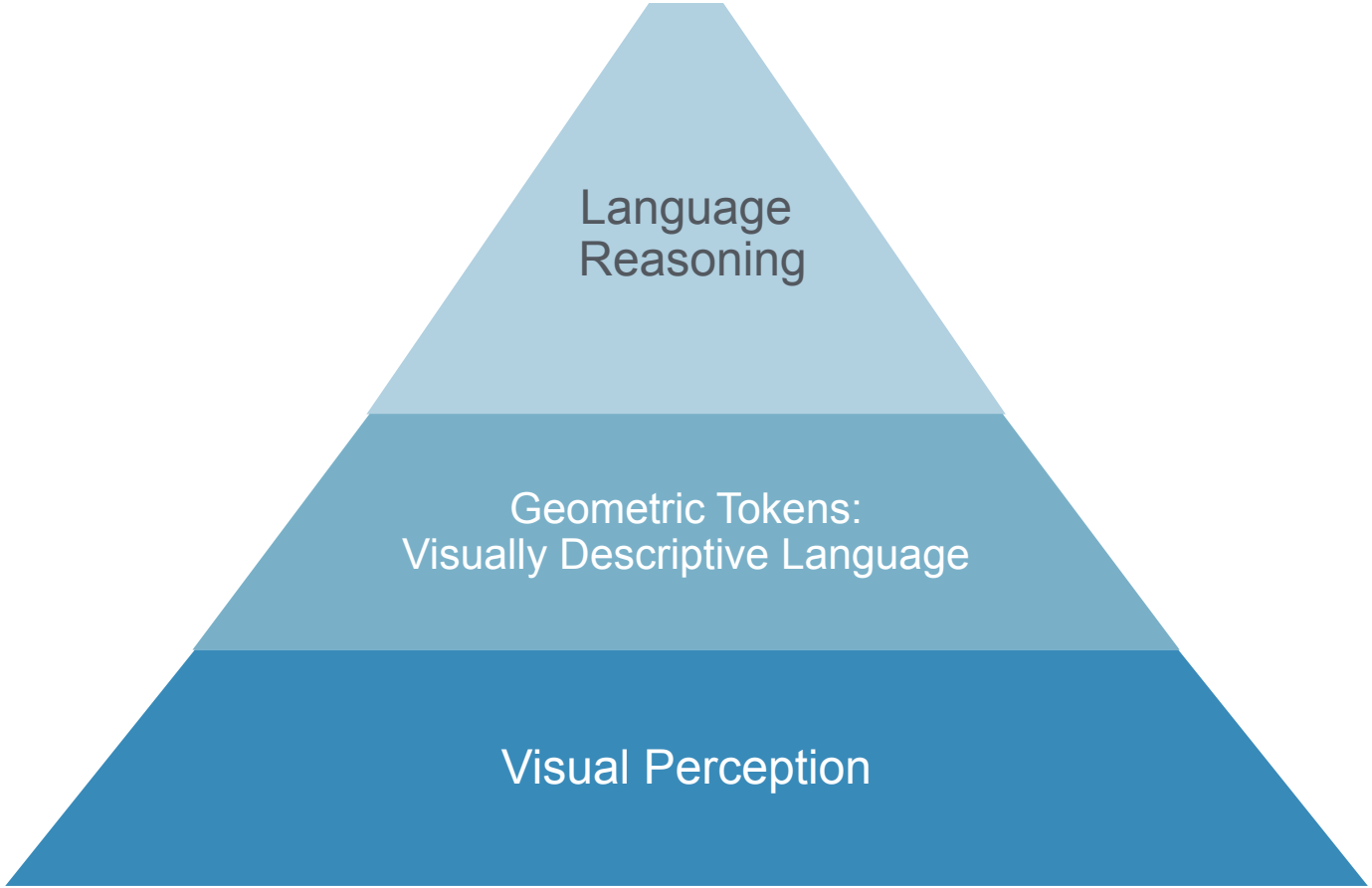
What is Missing? Intermediate Layers in VLM Pyramid



Go to lower-level:

What is Missing? Intermediate Layers in VLM Pyramid

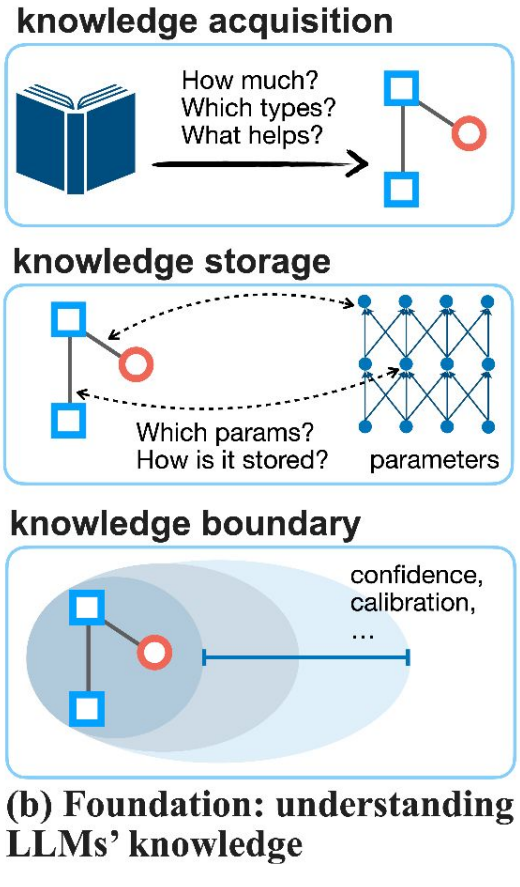
What is Missing? Intermediate Layers in VLM Pyramid



We need Geometric Abstraction (Geometric Tokens) for positions, shapes, etc

Conclusions

Recap


knowledge acquisition
How much?
Which types?
What helps?
knowledge storage
Which params?
How is it stored?
parameters
knowledge boundary
confidence,
calibration,
…
(b) Foundation: understanding LLMs' knowledge
Lack of A Theoretical Basis
Raven paradox
Multiverse paradox
Limited Applicability
Ripple Effect
Edit:
multiple effects
Going beyond triplets
fact:
event:
(c) Challenges in editing knowledge
More general editing objective
Editing natural texts or events:
More Versatile methods
Context Manipulation for Multi-hop Reasoning
Edit:
Multi-hop Reasoning:
"The of has a ."
Advanced in-context editing
"Given new knowledge , we know that ."
direction: less stubborn, more obedient.
A Wider Scope
Other LLM architectures
Knowledge unlearning
Multi-Modality knowledge
(d) Promising directions

Takeaway Messages

- **Memorization** has close connections with knowledge-intensive task performance in LMs. To further improve LM performance, we need to increase knowledge density and knowledge diversity.
- Knowledge can be localized within LM parameters, but the precision is questionable. A single piece of knowledge can be dispersed across multiple parameters. The organization of knowledge is not aligned with semantic/logical relationship.
- When a large amount of knowledge need to be updated, RAG is often more efficient and effective. If fine-tuning is needed, add diverse rewriting and mix in general data to avoid model forgetting.
- Interaction between different facts in LLMs can affect reasoning performance
- Sequential unlearning is better than trying to unlearn all the data at once and that unlearning is highly dependent on which kind of data (domain) is forgotten

Open Questions

- The 2 bit/param knowledge capacity rule holds for most Transformer-based autoregressive LMs. Are there alternative model architectures that can store more knowledge?
- What kind of instructions are most useful for knowledge extraction?
- How can we preserve good model calibration when injecting new knowledge through fine-tuning?
- How can we more efficiently utilize knowledge to elicit stronger reasoning ability?

Q&A

References

[1] https://www.brainfacts.org/in-the-lab/tools-and-techniques/2018/the-curious-case-of-patient-hm-082818

[2] Westerlund, Masha, and Liina Pylkkänen. "The role of the left anterior temporal lobe in semantic composition vs. semantic memory." Neuropsychologia 57 (2014): 59-70.

[3] Pause BM, Zlomuzica A, Kinugawa K, Mariani J, Pietrowsky R, Dere E. Perspectives on episodic-like and episodic memory. Front Behav Neurosci. 2013 Apr 18;7:33.

[4] Gupta, Akshat, Anurag Rao, and Gopala Anumanchipalli. "Model editing at scale leads to gradual and catastrophic forgetting." arXiv preprint arXiv:2401.07453 (2024).

[5] Gu, Jia-Chen, et al. "Model editing can hurt general abilities of large language models." arXiv preprint arXiv:2401.04700 (2024).

[6] https://www.brainfacts.org/in-the-lab/tools-and-techniques/2018/the-curious-case-of-patient-hm-082818

[7] Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[8] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

[9] Meng, Kevin, et al. "Mass-Editing Memory in a Transformer." The Eleventh International Conference on Learning Representations.

[10] Dai, Damai, et al. "Knowledge Neurons in Pretrained Transformers." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

[11] Li, Xiaopeng, et al. "Pmet: Precise model editing in a transformer." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

[12] Niu, Jingcheng, et al. "What does the Knowledge Neuron Thesis Have to do with Knowledge?." The Twelfth International Conference on Learning Representations.

[13] Qin, Jiaxin, et al. "Why Does New Knowledge Create Messy Ripple Effects in LLMs?." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

[14] Cohen, Roi, et al. "Evaluating the Ripple Effects of Knowledge Editing in Language Models." Transactions of the Association for Computational Linguistics 11 (2024): 283-298.

[15] Berglund, Lukas, et al. "The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”." The Twelfth International Conference on Learning Representations.

[16] Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of Language Models: Part 3.2, Knowledge Manipulation." In The Thirteenth International Conference on Learning Representations.

References

[17] Kaplan et al. Scaling Laws for Neural Language Models. ArXiv 2020.
[18] Liu et al, Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens.
[19] Kandpal, Nikhil et al. “Large Language Models Struggle to Learn Long-Tail Knowledge.” International Conference on Machine Learning (2022).
[20] Wang, Xinyi, et al. "Generalization vs Memorization: Tracing Language Models' Capabilities Back to Pretraining Data."ICLR 2024.
[21] Tirumala, Kushal, et al. "Memorization without overfitting: Analyzing the training dynamics of large language models." Advances in Neural Information Processing Systems (2022).
[22] Biderman, Stella, et al. "Emergent and predictable memorization in large language models." Advances in Neural Information Processing Systems 36 (2023): 28072-28090.
[23] Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of language models: Part 3.1, knowledge storage and extraction." arXiv preprint arXiv:2309.14316 (2023).
[24] Jiang, Zhengbao, et al. "Instruction-tuned language models are better knowledge learners." ACL (2024).
[25] Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of language models: Part 3.3, knowledge capacity scaling laws." arXiv preprint arXiv:2404.05405 (2024).
[26] Lee et al. Deduplicating Training Data Makes Language Models Better. ACL 2022
[27] Penedo et al. FineWeb: decanting the web for the finest text data at scale.
[28] Wettig, Alexander, et al. "Qurating: Selecting high-quality data for training language models." arXiv preprint arXiv:2402.09739 (2024)
[29] Maini, Pratyush, et al. "Rephrasing the web: A recipe for compute and data-efficient language modeling." ACL (2024).
[30] OLMo, Team, et al. "2 OLMo 2 Furious." arXiv preprint arXiv:2501.00656 (2024).
[31] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." ACM Transactions on Information Systems 43.2 (2025): 1-55.
[32] Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.
[33] Li, Kenneth, et al. "Inference-time intervention: Eliciting truthful answers from a language model." Advances in Neural Information Processing Systems 36 (2023): 41451-41530.
[34] Zhang, Hanning, et al. "R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024.
[35] Tian, Katherine, et al. "Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback." arXiv preprint arXiv:2305.14975 (2023).

References

[36] Zhang, Yuji, et al. "Knowledge overshadowing causes amalgamated hallucination in large language models." arXiv preprint arXiv:2407.08039 (2024).

[37] Zhang, Yuji, et al. "The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination." arXiv preprint arXiv:2502.16143 (2025).

[38] Wang, Song, et al. "Knowledge editing for large language models: A survey." ACM Computing Surveys 57.3 (2024): 1-37.

[39] Gu, Jia-Chen, et al. "Model editing harms general abilities of large language models: Regularization to the rescue." arXiv preprint arXiv:2401.04700 (2024).

[40] Kang, Katie, et al. "Unfamiliar finetuning examples control how language models hallucinate." arXiv preprint arXiv:2403.05612 (2024).

[41] Zheng, Ce, et al. "Can we edit factual knowledge by in-context learning?." arXiv preprint arXiv:2305.12740 (2023).

[42] Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." arXiv preprint arXiv:2310.02129 (2023).

[43] Zhang, Ningyu, et al. "A comprehensive study of knowledge editing for large language models." arXiv preprint arXiv:2401.01286 (2024).

[44] Mitchell, Eric, et al. "Memory-based model editing at scale." International Conference on Machine Learning. PMLR, 2022.

[45] Hase, Peter, et al. "Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?." arXiv preprint arXiv:2406.19354 (2024).

[46] Liu, Jiateng, et al. "EVEDIT: Event-based Knowledge Editing for Deterministic Knowledge Propagation." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

[47] Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." arXiv preprint arXiv:2310.02129 (2023).

[48] Yang, Wanli, et al. "The butterfly effect of model editing: Few edits can trigger large language models collapse." arXiv preprint arXiv:2402.09656 (2024).

[49] Golovneva, Olga, et al. "Reverse training to nurse the reversal curse." arXiv preprint arXiv:2403.13799 (2024).

[50] Press, Ofir, et al. "Measuring and narrowing the compositionality gap in language models." arXiv preprint arXiv:2210.03350 (2022).

[51] Yang, Sohee, et al. "Do large language models latently perform multi-hop reasoning?." arXiv preprint arXiv:2402.16837 (2024).

[52] Ye, Yixin, et al. "LIMO: Less is More for Reasoning." arXiv preprint arXiv:2502.03387 (2025).

[53] Muennighoff, Niklas, et al. "s1: Simple test-time scaling." arXiv preprint arXiv:2501.19393 (2025).