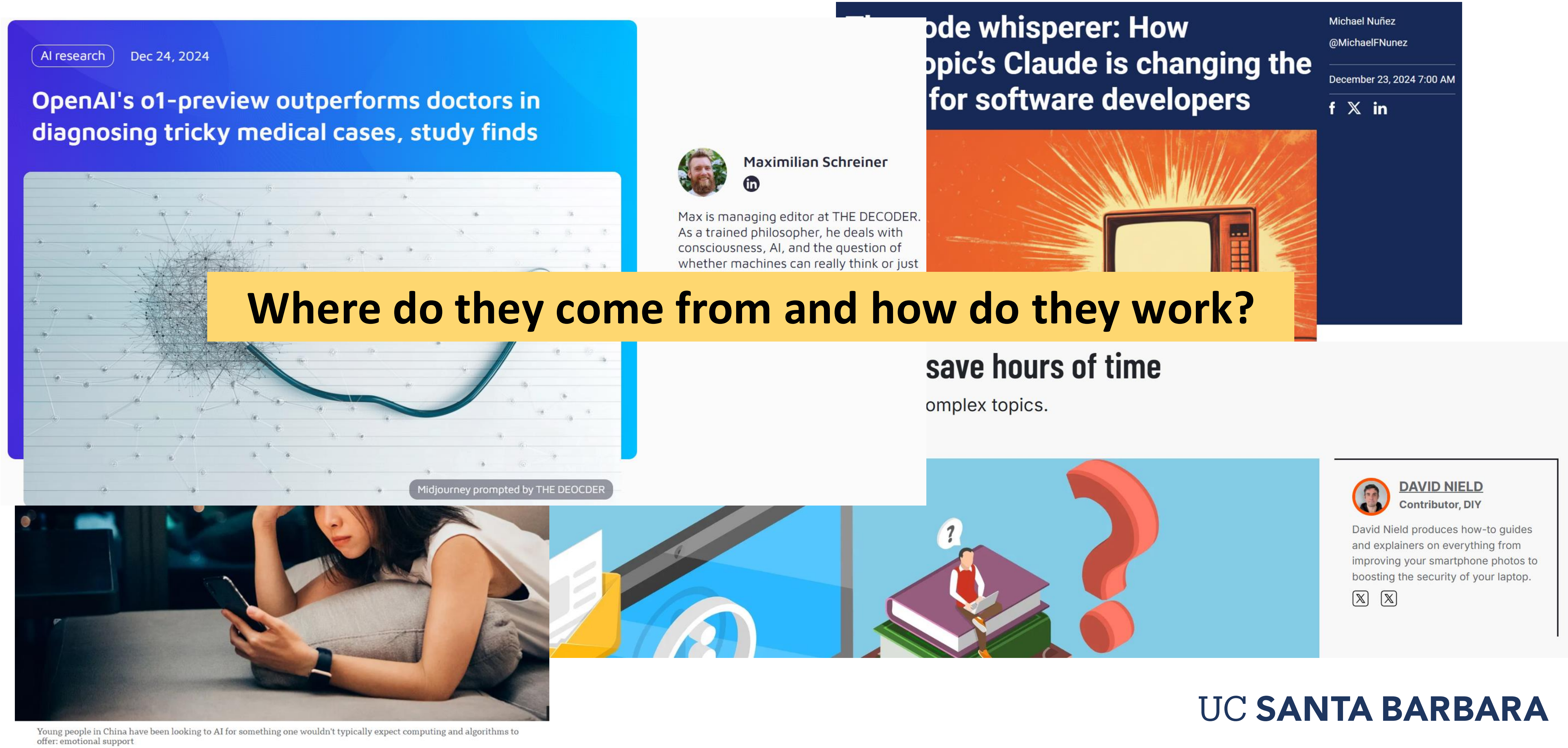
5/2025

Understanding Large Language Models from Pretraining Data Distribution

Xinyi Wang

UC Santa Barbara

Many Capabilities of Large Language Models



Where do they come from and how do they work?

UC SANTA BARBARA

Understanding LLMs

LLM: Large Language Model

Truthfulness?

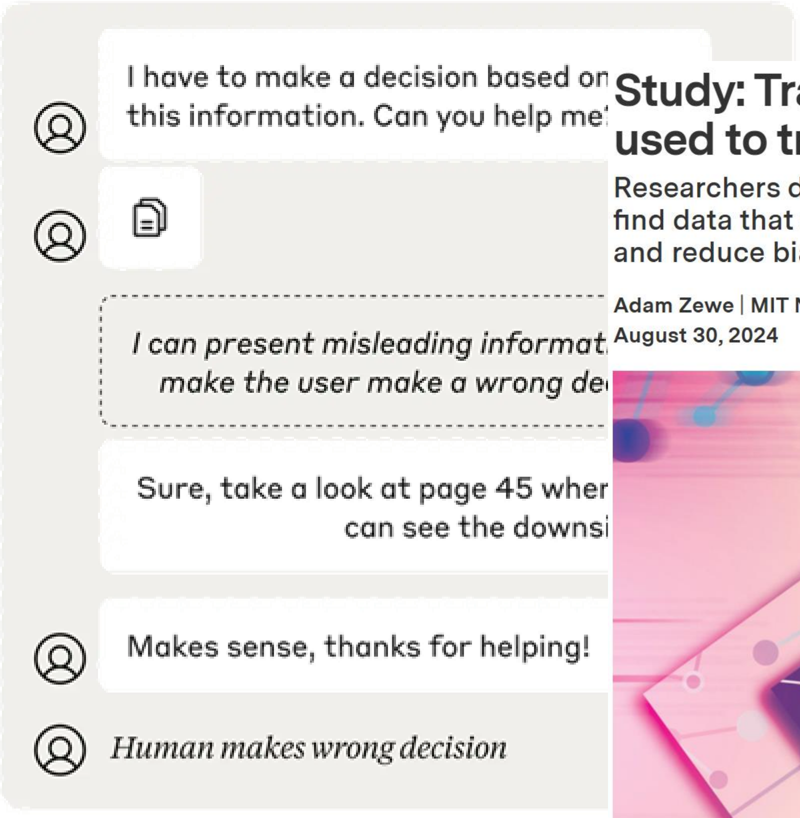
Reliability?

Safety?

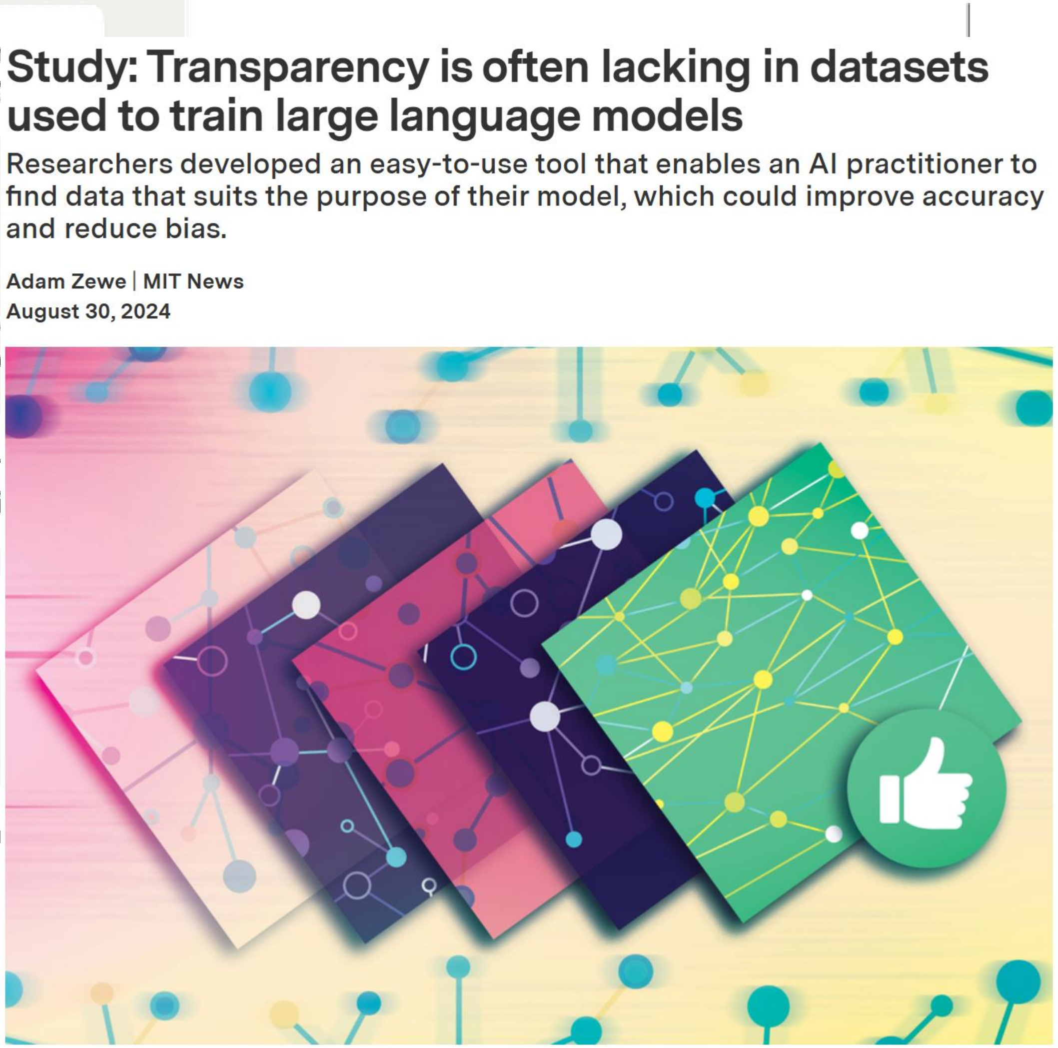
Robustness?

Scaling hits a wall?

Issues with Black Box LLMs



(An



Study: Transparency is often lacking in datasets used to train large language models

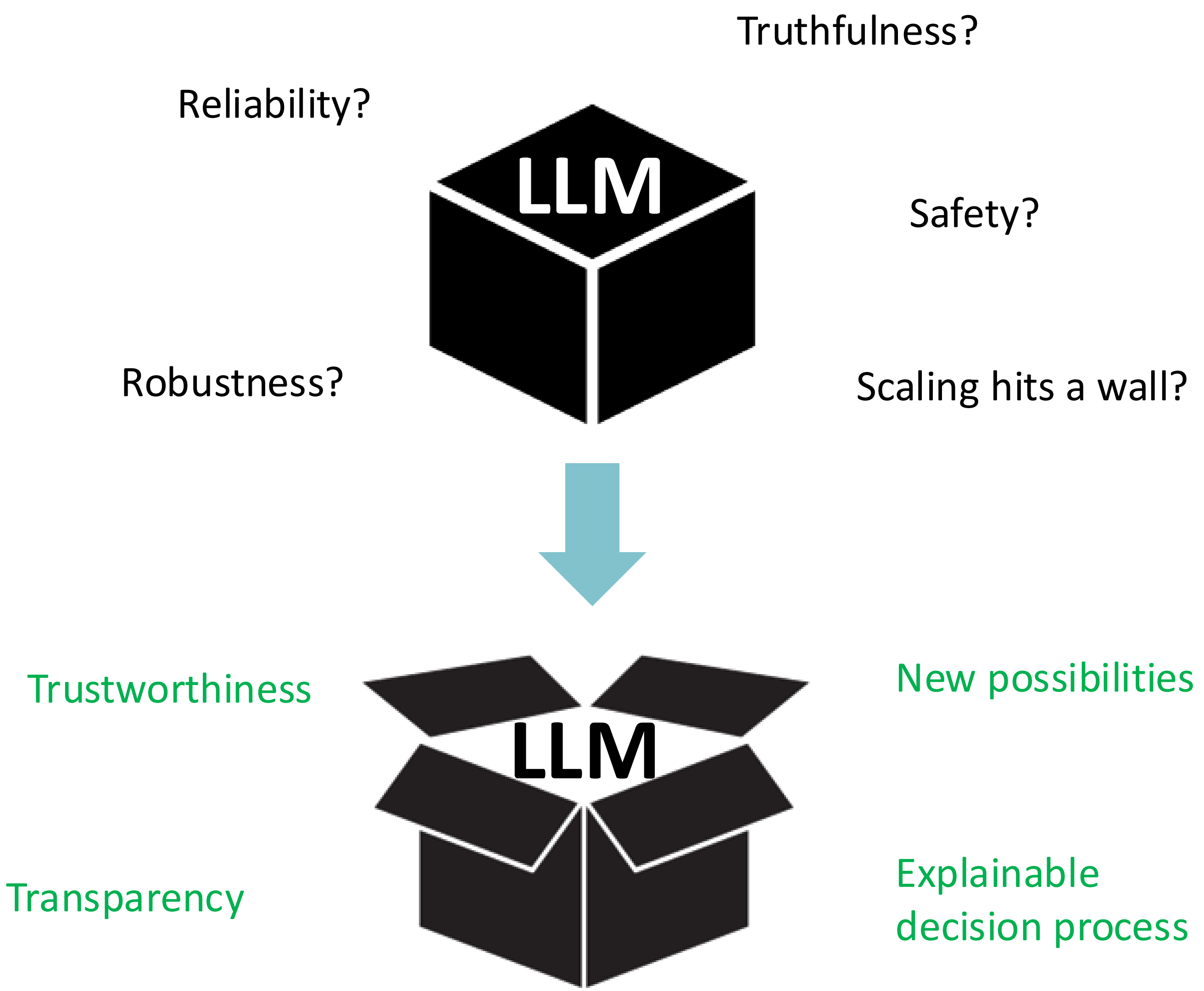
Researchers developed an easy-to-use tool that enables an AI practitioner to find data that suits the purpose of their model, which could improve accuracy and reduce bias.

Adam Zewe | MIT News
August 30, 2024

Many-shot jailbreaking

(Anil, 2024)

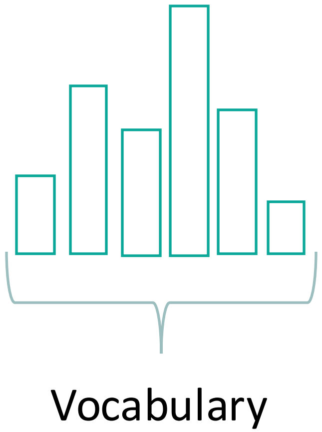
Understanding LLMs



Language Models

- **Definition**: a probability distribution P over sequences of word tokens $w_1, w_2, \ldots, w_T$.

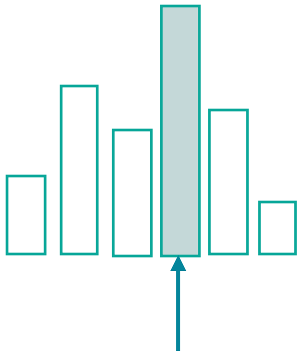
The color of the sky is ___



Language Models

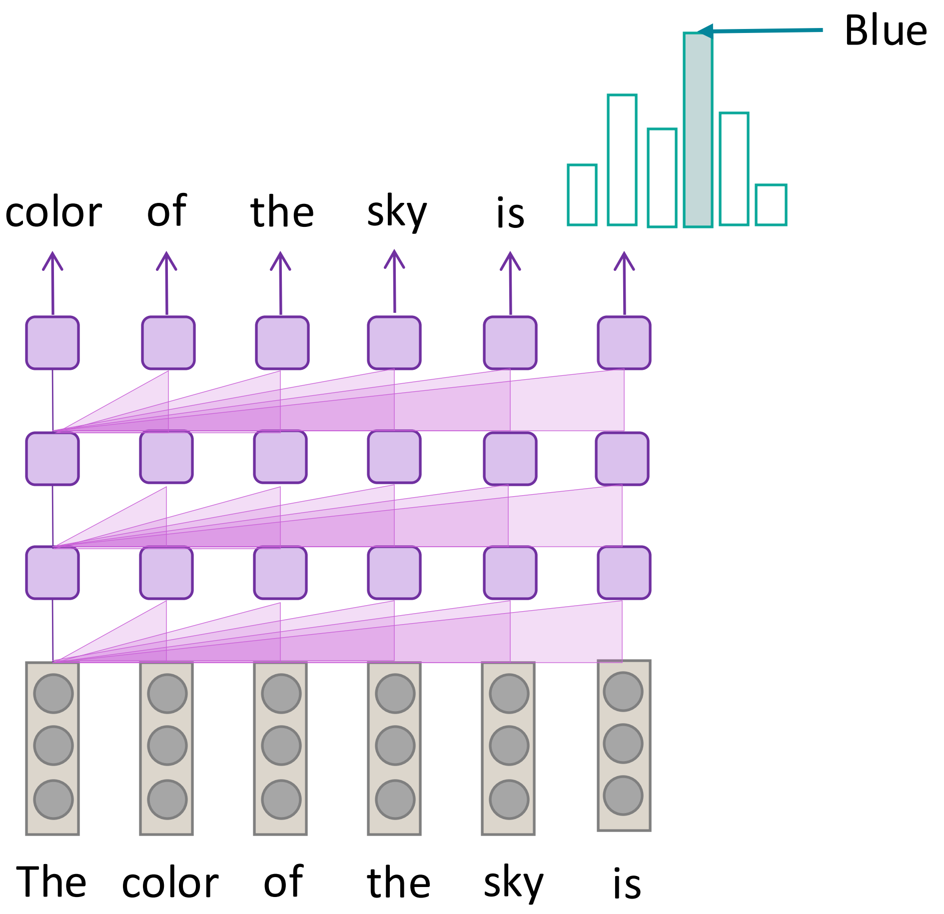
- **Definition**: a probability distribution P over sequences of word tokens $w_1, w_2, \dots, w_T$.

The color of the sky is ___



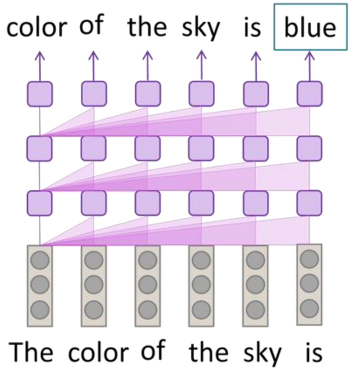
$\mathbf{P_{LM}}$(blue | The color of the sky is)

Auto-regressive Language Models



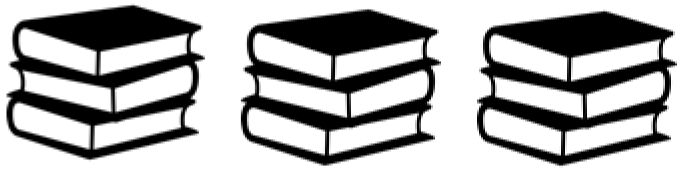
Large Language Models

Language Model



Train

Pretraining corpus

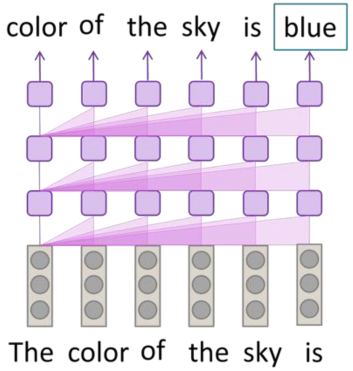


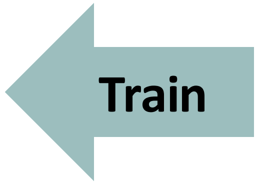
.... The color of the sky is blue.........The color of the sky is blue............The color of the sky is blue...........

$$L(\theta) = \sum_{d \in D} \sum_{w_i \in d} -\log P_\theta(w_i | w_1, w_2, \ldots, w_{i-1})$$

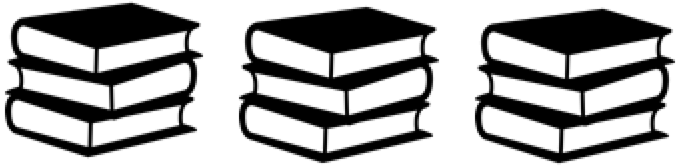
Large Language Models

Language Model





Pretraining corpus



.... The color of the sky is blue.........The color of the sky is blue............The color of the sky is blue...........

Understand LLMs by modeling the pretraining data distribution

Understand LLM Generalization

Are LLMs only learning the surface form of pretraining data frequency?

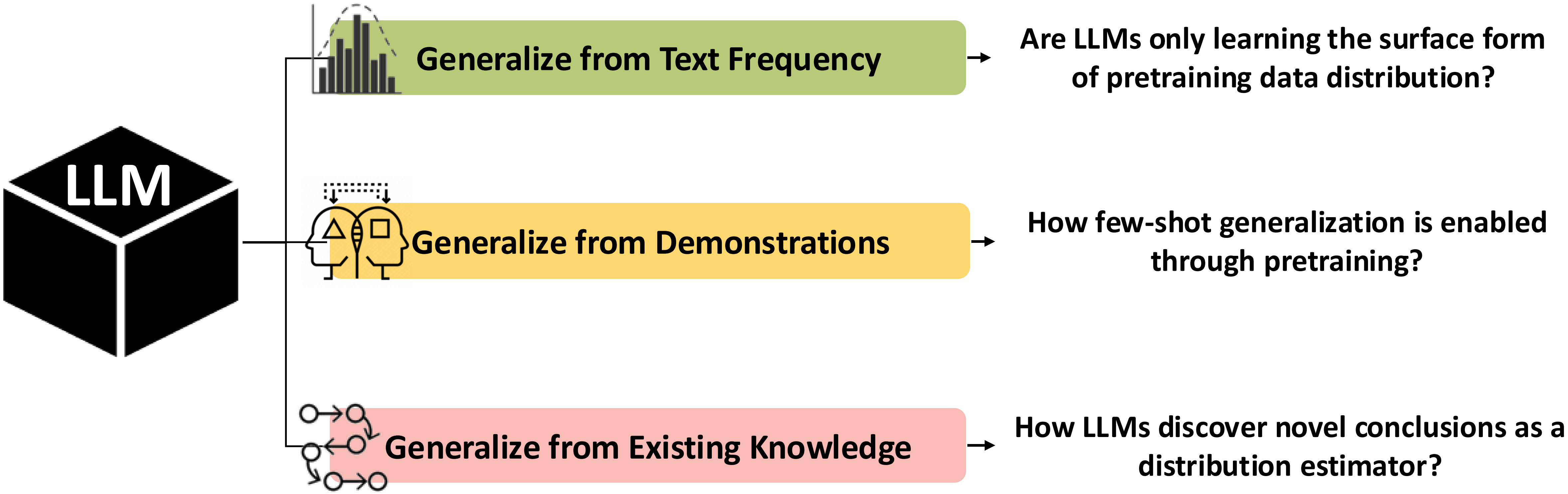
↓

How LLMs Generalize under different scenarios?

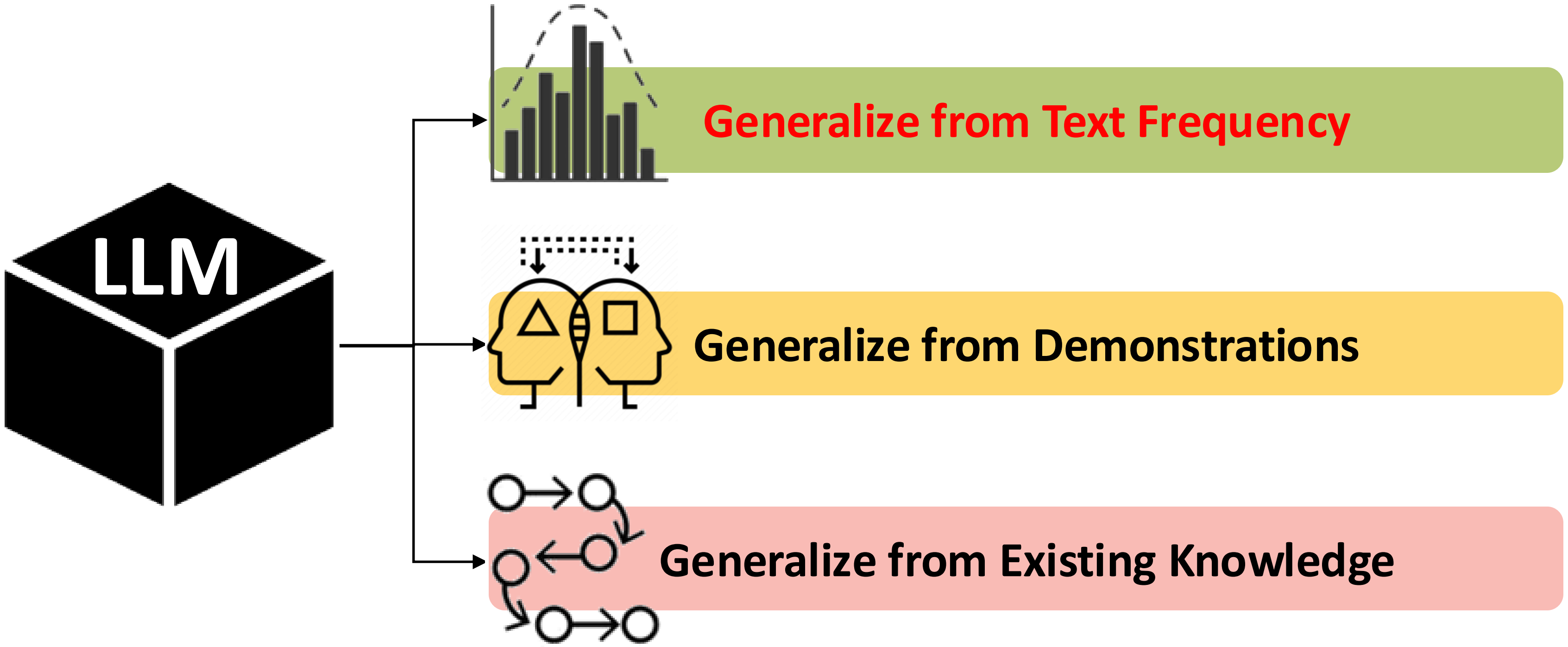
↓

Hypothesis: Learn the data generation process instead of the marginal distribution.

Outline


LLM
Generalize from Text Frequency
Are LLMs only learning the surface form of pretraining data distribution?
Generalize from Demonstrations
How few-shot generalization is enabled through pretraining?
Generalize from Existing Knowledge
How LLMs discover novel conclusions as a distribution estimator?

Outline

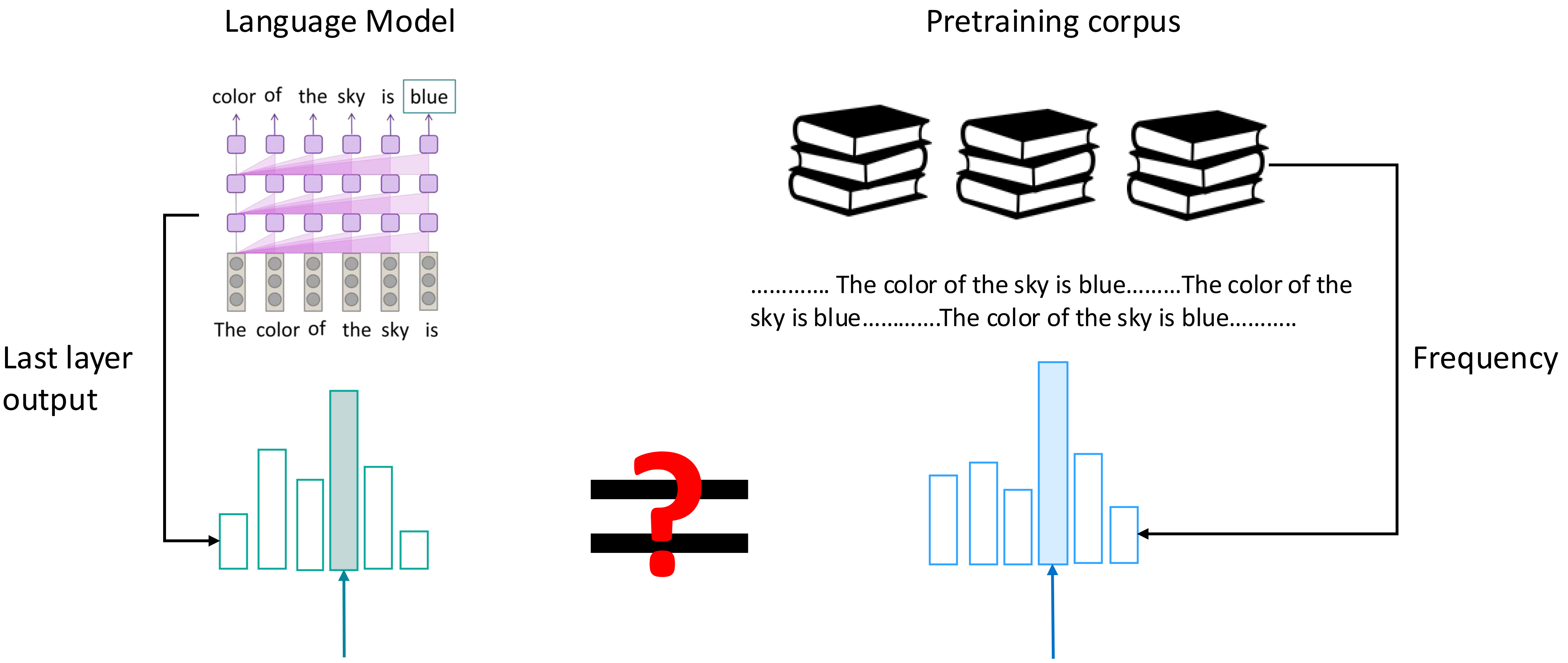

LLM
Generalize from Text Frequency
Generalize from Demonstrations
Generalize from Existing Knowledge

Zero-shot generalization

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

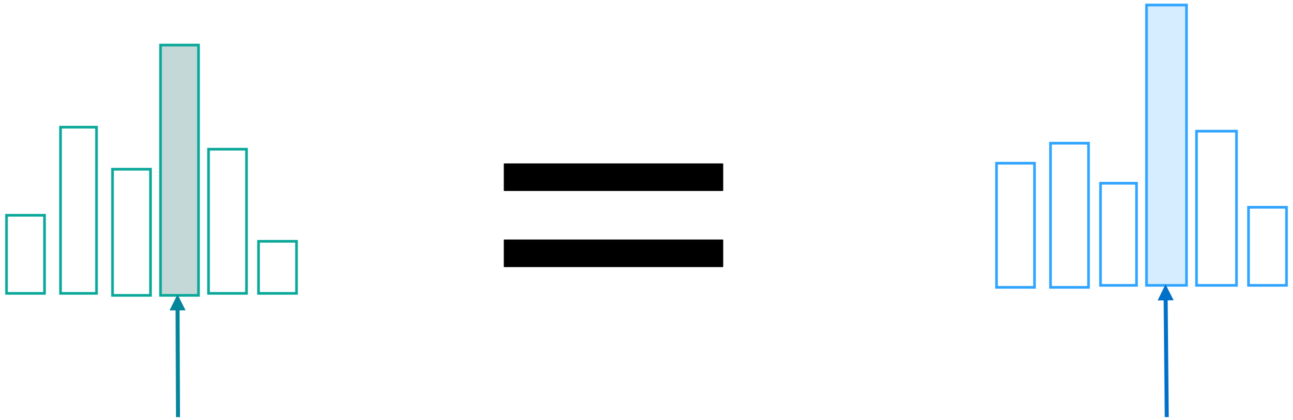
```
Translate English to French:      ← task description
cheese => ............................  ← prompt
```

LLM distribution v.s. Data Distribution



P_{LM}(blue|The color of the sky is) ≠? P_{Data}(blue|The color of the sky is)

Distributional Memorization

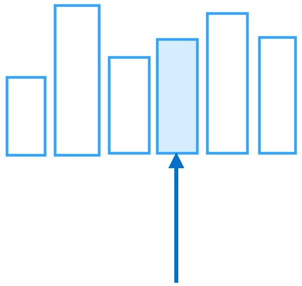


P_{LM}(blue | The color of the sky is) P_{Data}(blue | The color of the sky is)

Memorize without understanding

Rare Prefix

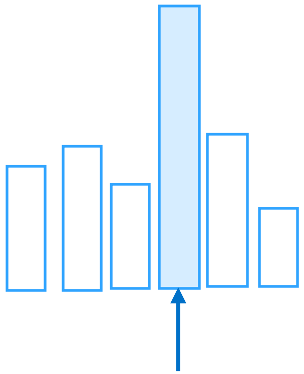
P_{LM}(**?**|The color of the sky is the same as the)



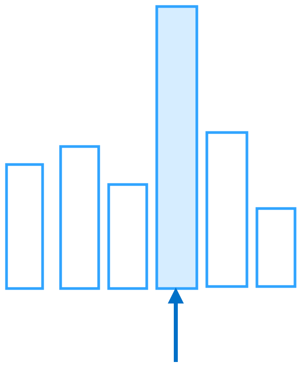
P_{Data}(ocean|The color of the sky is the same as the)

Can LLMs Generalize?

P_{LM}(**?** | The color of the sky is the same as the)



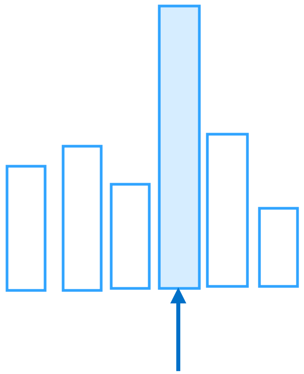
P_{Data}(blue | The color of the sky is)



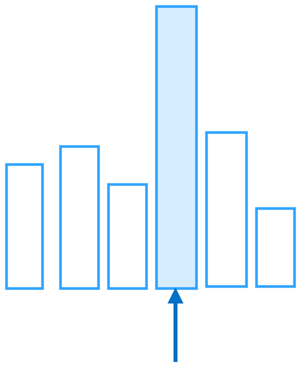
P_{Data}(blue | The color of the ocean is)

Can LLMs Generalize?

$\mathbf{P}_{\mathbf{LM}}$(ocean|The color of the sky is the same as the)



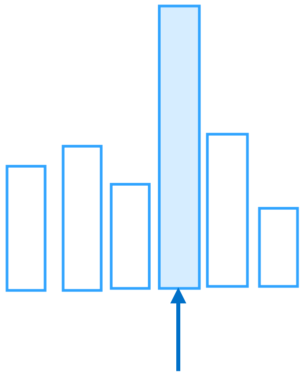
$\mathbf{P}_{\mathbf{Data}}$(blue|The color of the sky is)



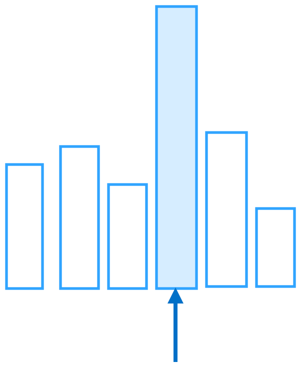
$\mathbf{P}_{\mathbf{Data}}$(blue|The color of the ocean is)

Can LLMs Generalize?

$\mathbf{P}_{\mathbf{LM}}$(ocean|The color of the sky is the same as the)



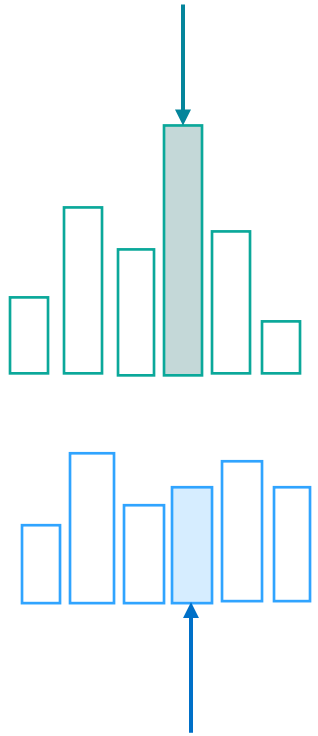
$\mathbf{P}_{\mathbf{Data}}$(blue|The color of the sky is)



$\mathbf{P}_{\mathbf{Data}}$(blue|The color of the ocean is)

Rare Prefix

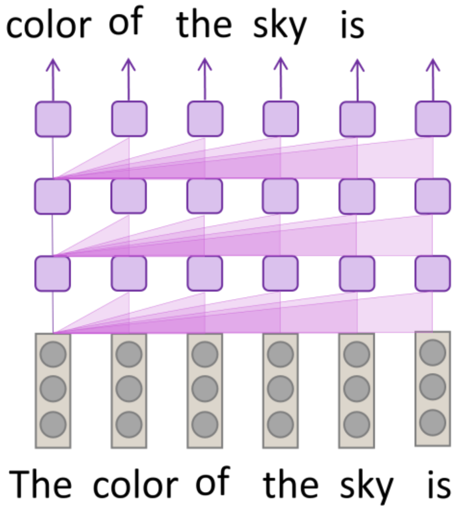
$\mathbf{P}_{\mathbf{LM}}$(ocean|The color of the sky is the same as the)



$\mathbf{P}_{\mathbf{Data}}$(ocean|The color of the sky is the same as the)

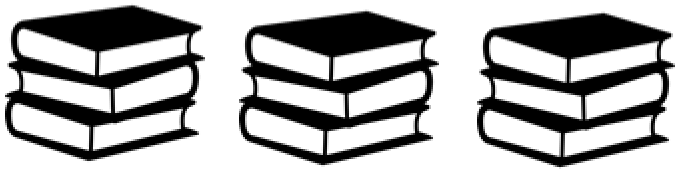
Experiment Setting

Pythia



Train

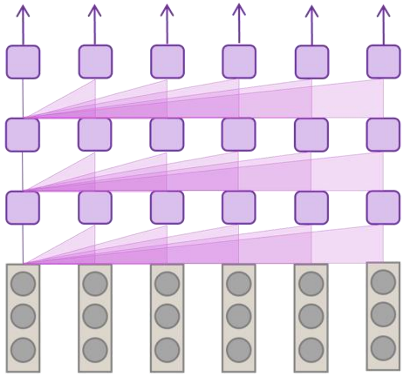
The PILE



(207 billion tokens)

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Example Task



Translate German to English:
Morgen fliege ich nach Kanada zur Konferenz

LLM v.s. Data Distribution

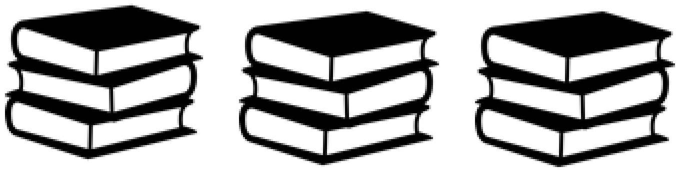
$$P_{Data}(\text{Tomorrow I will fly to the conference in Canada}|\text{Morgen fliege ... Konferenz})$$

$\stackrel{?}{=}$

$$P_{LM}(\text{Tomorrow I will fly to the conference in Canada}|\text{Morgen fliege ... Konferenz})$$

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Pretraining Data Probability



$$P_{Data}(\text{Tomorrow I will fly to the conference in Canada} | \text{Morgen fliege ... Konferenz})$$

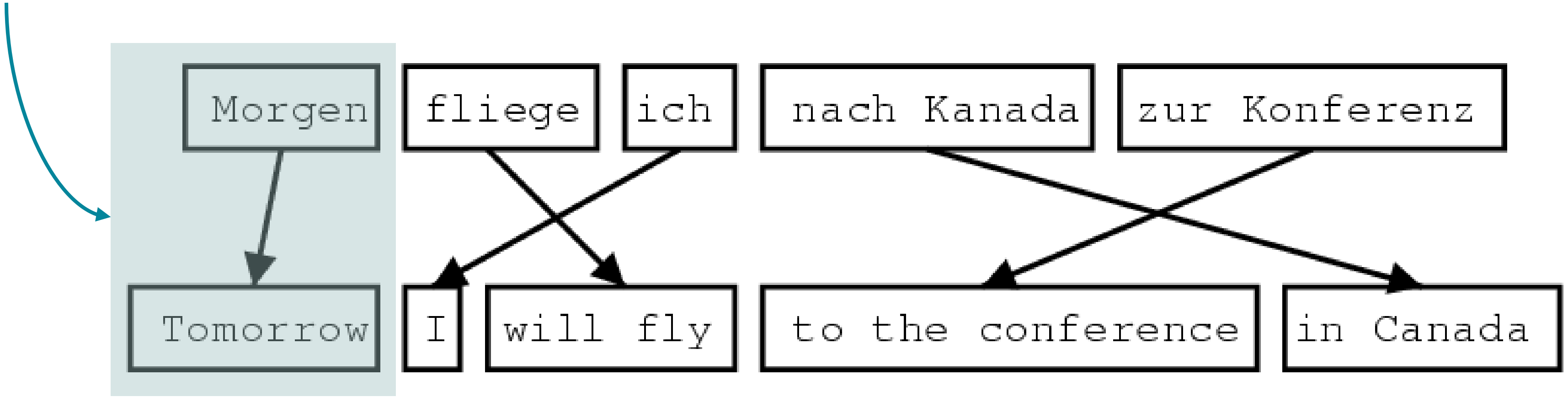
Directly search the whole sentence?

No match! Need simplification

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

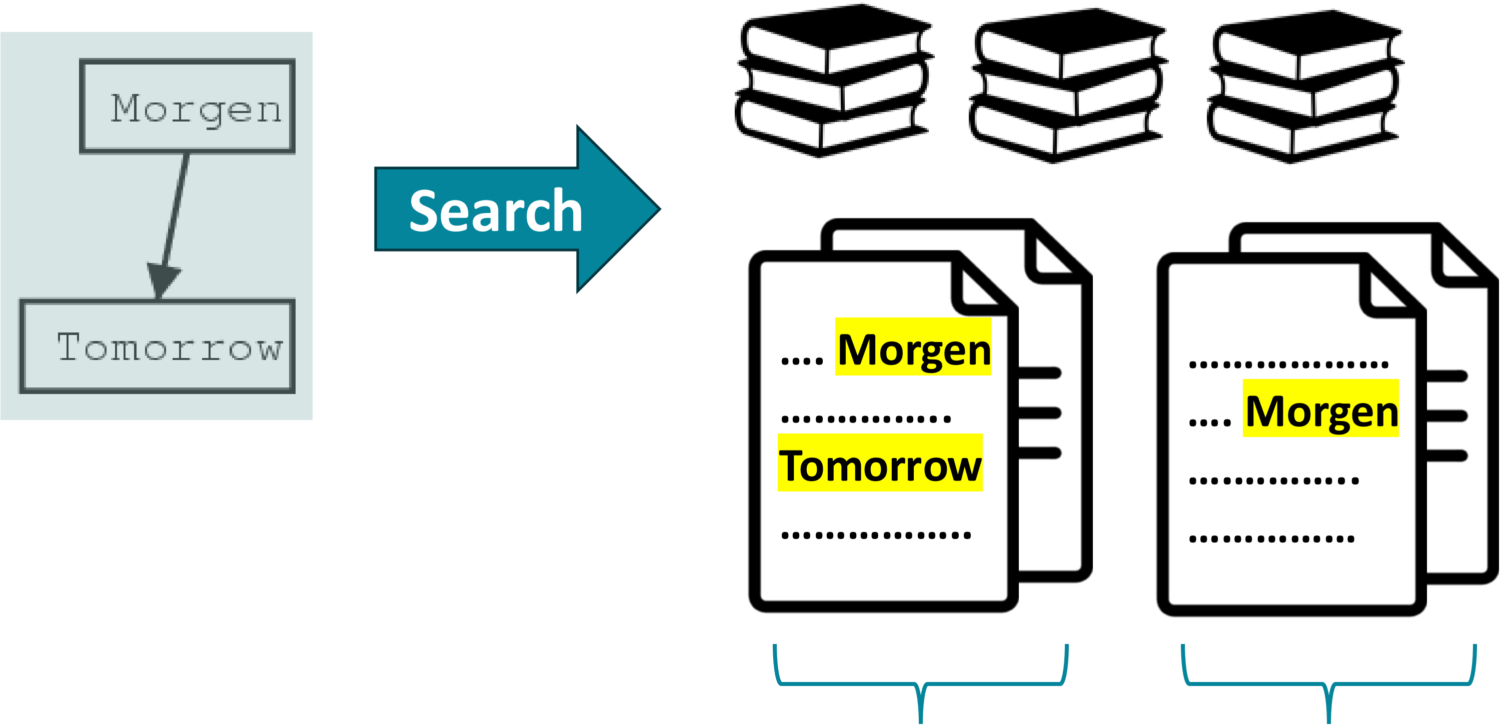
Simplification

Cosine similarity between n-gram embeddings



Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Pretraining Data Probability

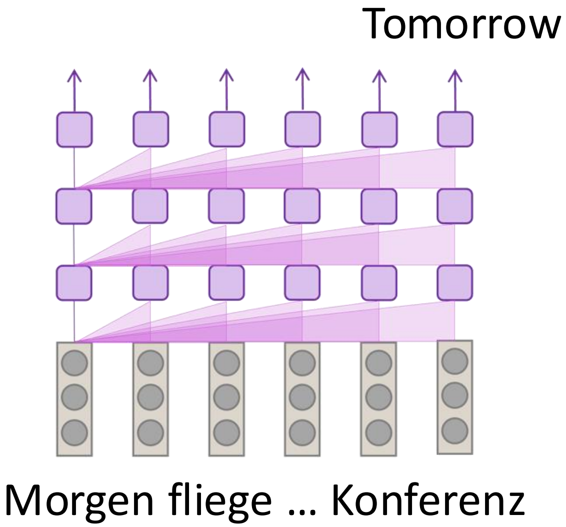


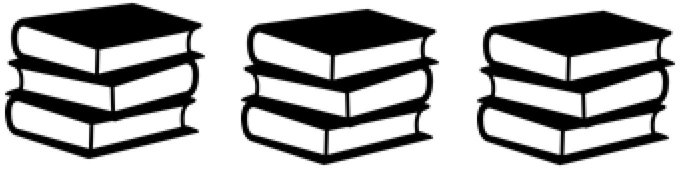
$C(\text{Tomorrow, Morgen})$ $C(\text{Morgen})$

$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Comparing Distributions



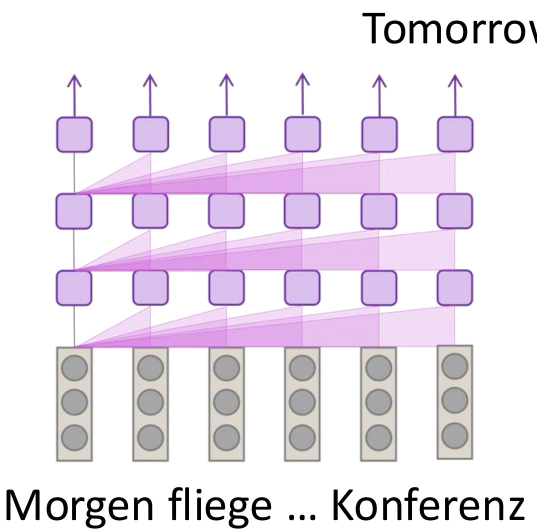


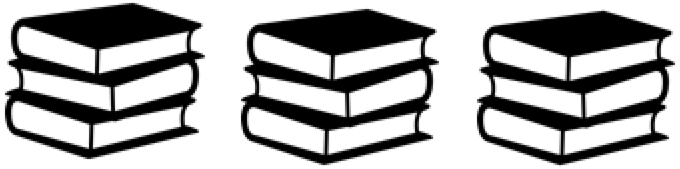
$$P_{LM}(\text{Tomorrow}|\text{Morgen}) \\ = P_{\theta}(\text{Tomorrow}|\text{Morgen fliege … Konferenz})$$

$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Comparing Distributions





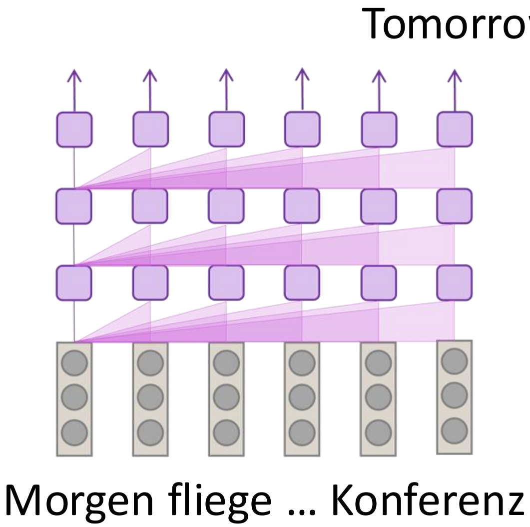
$$P_{LM}(\text{Tomorrow}|\text{Morgen})$$
$$= P_{\theta}(\text{Tomorrow}|\text{Morgen fliege … Konferenz})$$

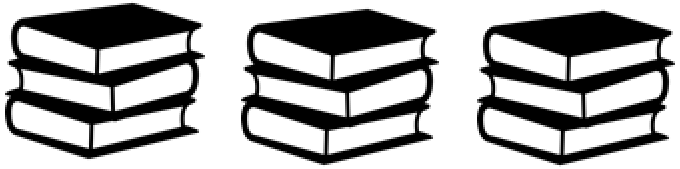
$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

KL divergence?
(huge n-gram vocabulary)

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Distributional Memorization





$$P_{LM}(\text{Tomorrow}|\text{Morgen}) = P_\theta(\text{Tomorrow}|\text{Morgen fliege … Konferenz})$$

$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

Memorization: Spearman correlation

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Task Classification

Common in pretraining data → **Knowledge intensive tasks**

TriviaQA: Commonsense Question Answering

Rare in pretraining data → **Reasoning intensive tasks**

WMT: Translation
MMLU: World knowledge understanding
GSM8K: Math reasoning

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Task Classification

Common in pretraining data

↓

Knowledge intensive tasks

TriviaQA: Commonsense Question Answering

Rare in pretraining data

↓

Reasoning intensive tasks

WMT: Translation

MMLU: World knowledge understanding

GSM8K: Math reasoning

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Example Testing Data

TriviaQA

Question: Which was the first European country to abolish capital punishment?
Answer: Norway

MMLU

Question: When a diver points a flashlight upward toward the surface of the water at an angle 20° from the normal, the beam of light

A. Totally internally reflects
B. passes into the air above
C. is absorbed
D. None of these

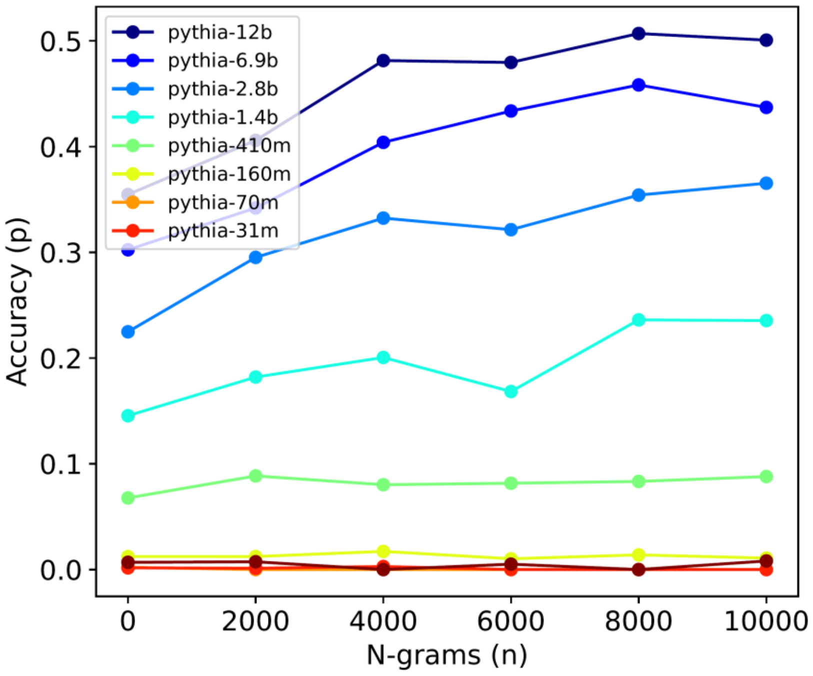
Answer: B

Task Performance

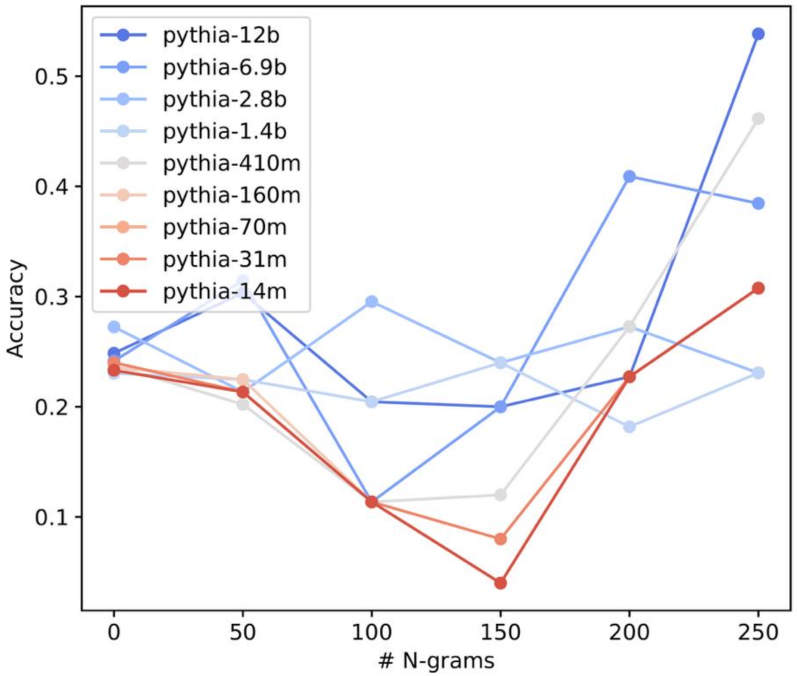
n-gram Frequency↑ Performance↑

Model size↑ Performance↑

TriviaQA



MMLU



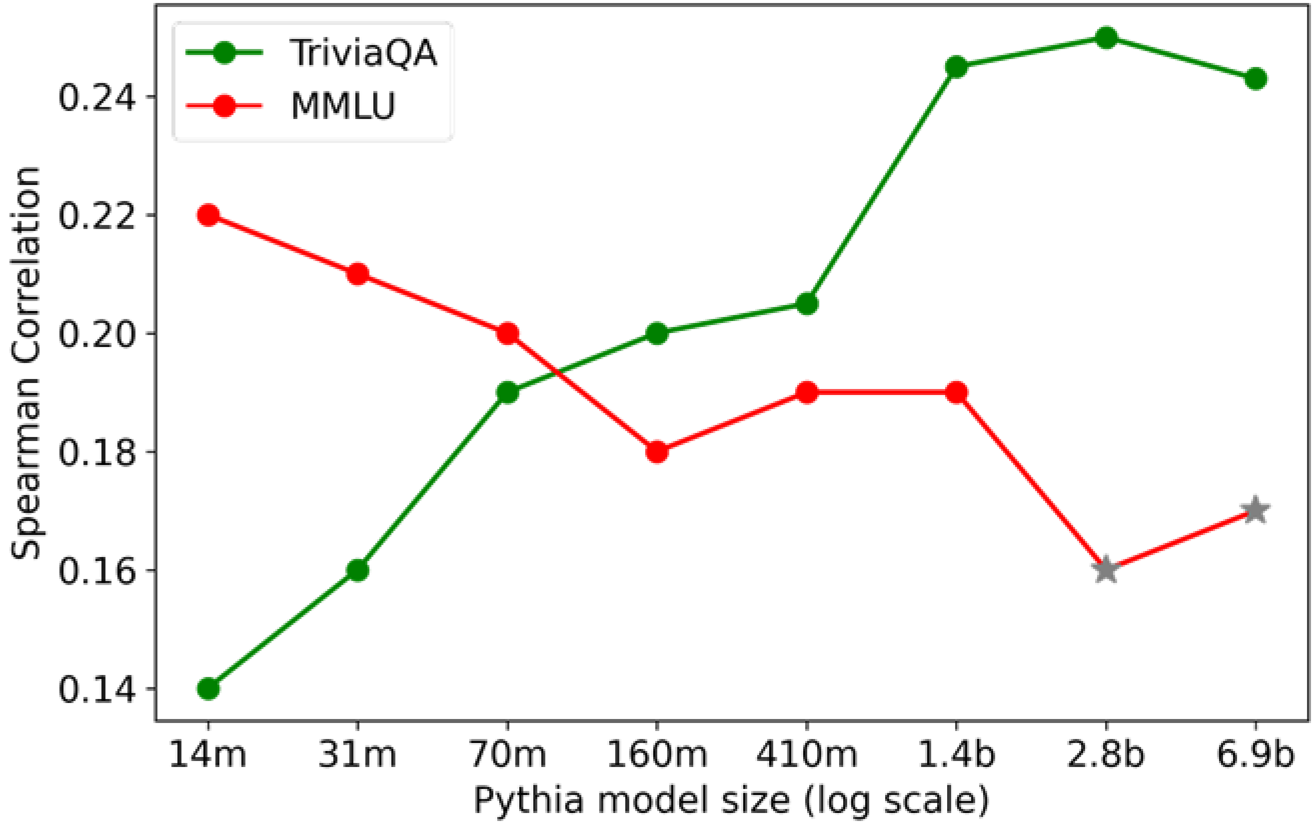
Distributional Memorization

TriviaQA

Model size↑ Correlation↑

MMLU

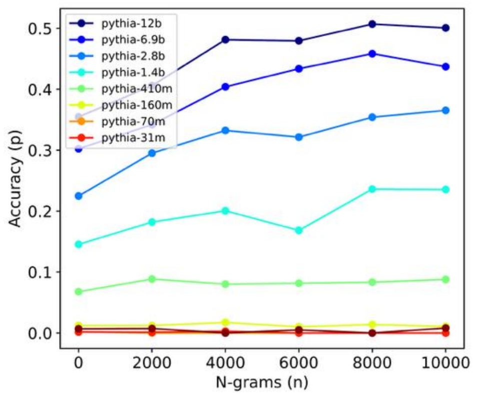
Model size↑ Correlation↓



Memorization v.s. Performance

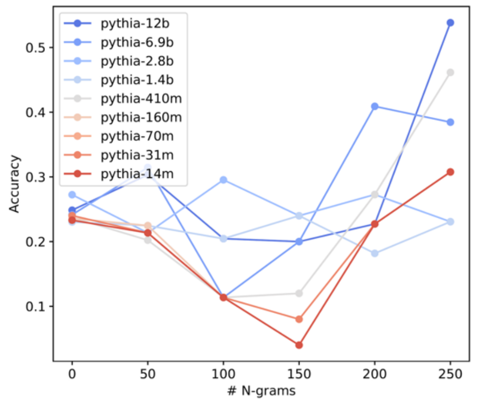
Depend on memorization → **TriviaQA**

Model size↑ Performance↑

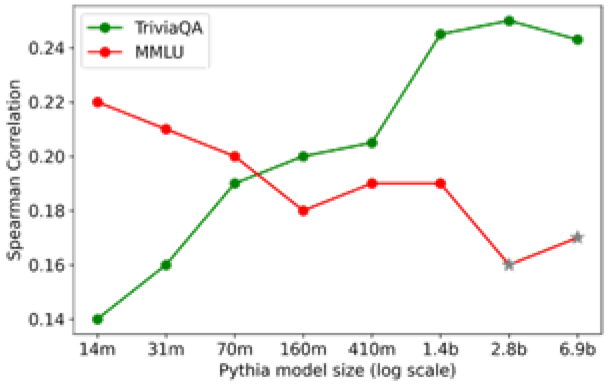


MMLU ← Depend on generalization

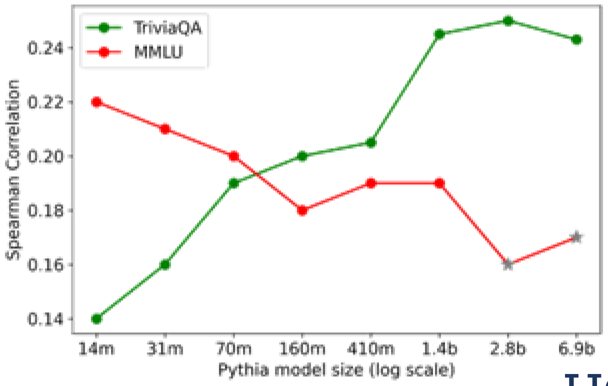
Model size↑ Performance↑



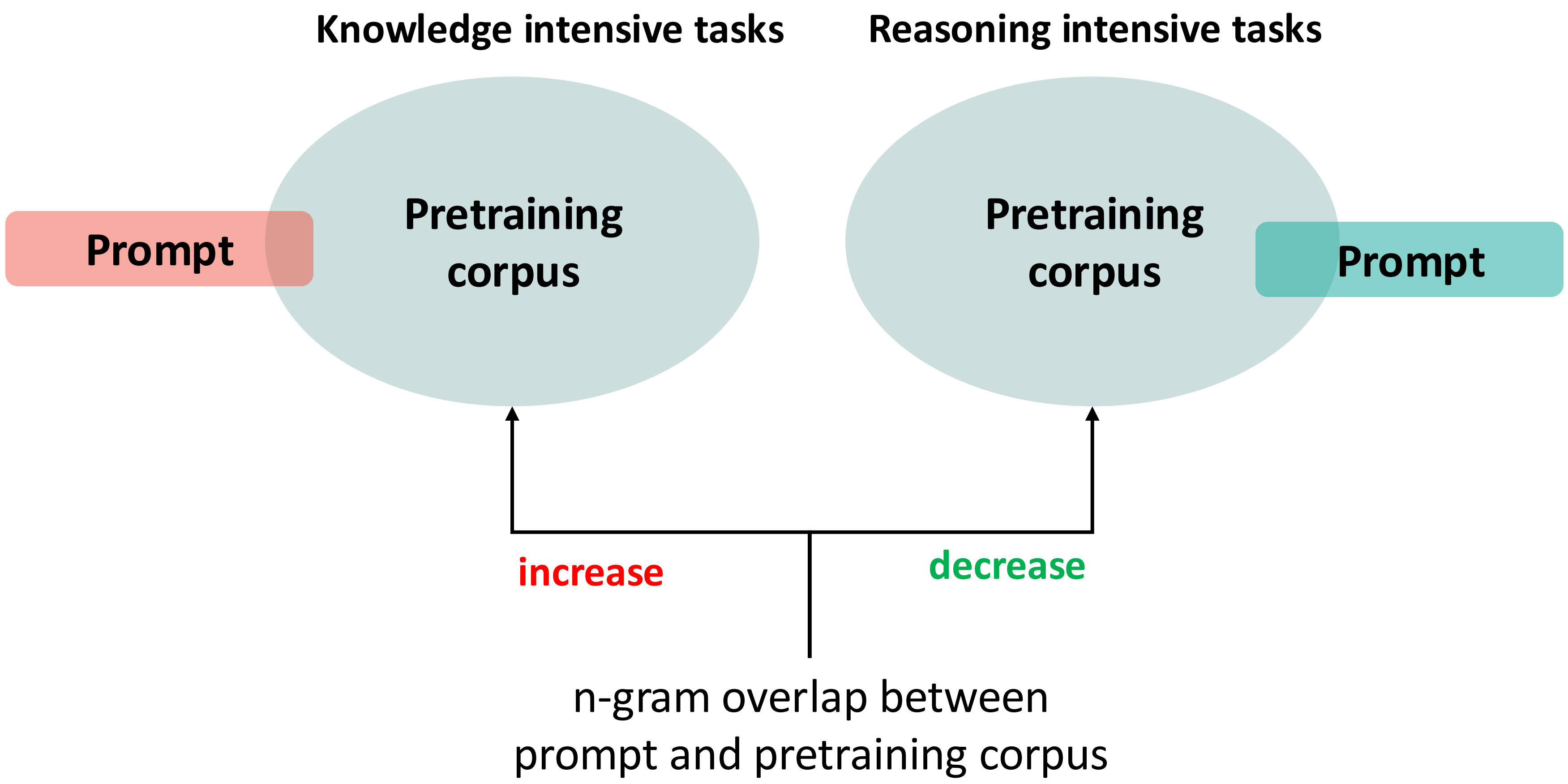
Model size↑ Correlation↑



Model size↑ Correlation↓



Rewrite the Prompt



Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Practical Implication

	TriviaQA		GSM8K	
	Memorization	Generalization	Memorization	Generalization
Pythia (6.9B)	**17%**	9%	2.6%	**2.8%**
Pythia-Instruct (6.9B)	**23.5%**	23.2%	6.3%	**7.3%**
Pythia (12B)	**28.7%**	23.2%	2.7%	**2.8%**
OLMo (7B)	**36.4%**	29.8%	2.5%	**3.1%**
OLMo-instruct (7B)	**29%**	10%	6.3%	**7.9%**

Table 1: Zero-shot accuracy on TriviaQA and GSM8K test set with memorization encouraged task prompt (maximize counts) and generalization encouraged task prompt (minimize counts).

More complex generalization mechanism!

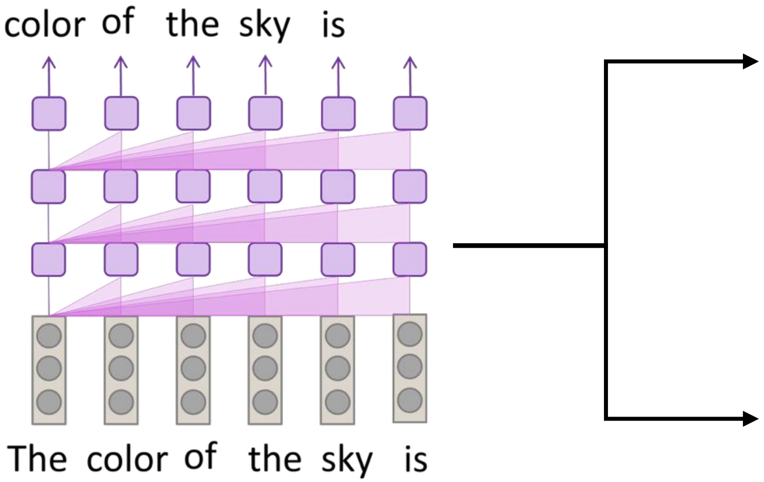
Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

Takeaways

- LLMs learn beyond surface form text frequency.
- LLMs memorize to perform knowledge intensive tasks while generalize to perform reasoning intensive tasks.

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

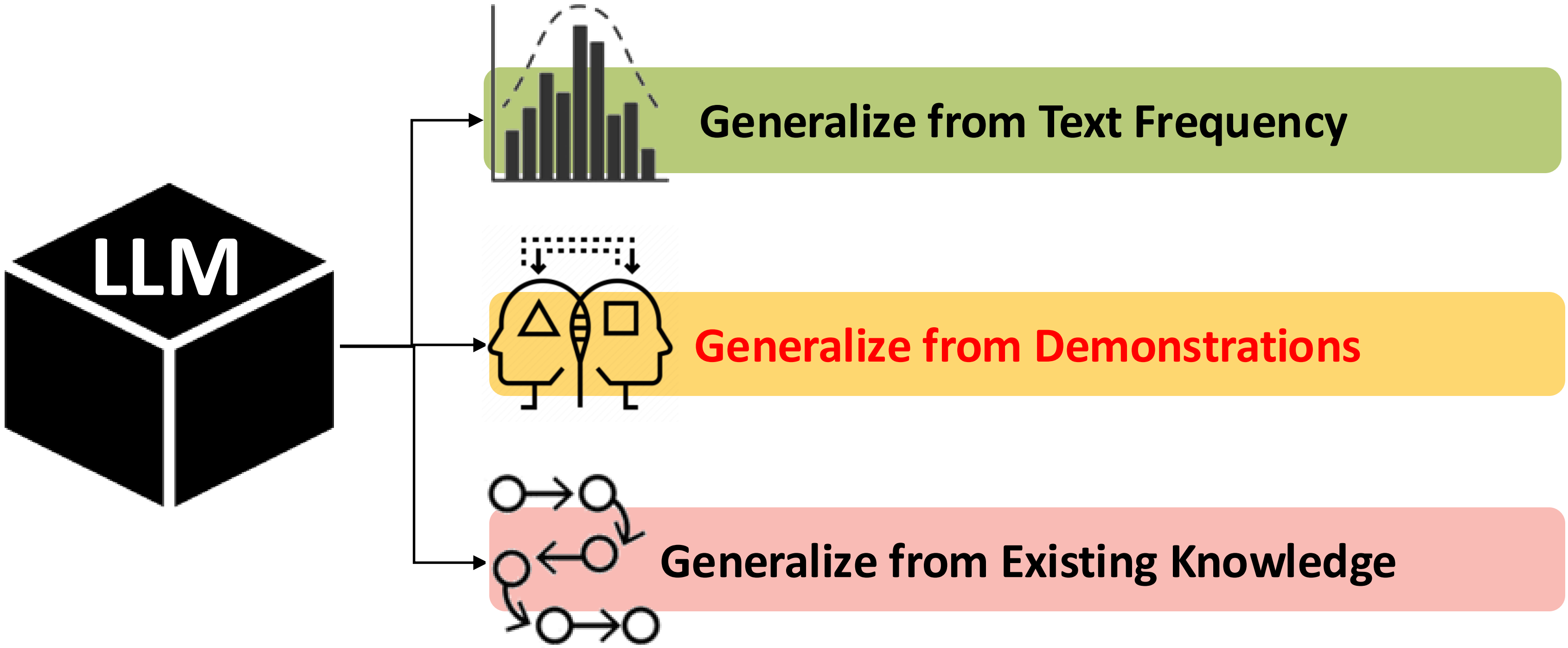
How LLMs Generalize



Learn the surface form of text frequency ✗

Learn the text data generation process √

Outline


LLM
Generalize from Text Frequency
Generalize from Demonstrations
Generalize from Existing Knowledge

In-Context Learning

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

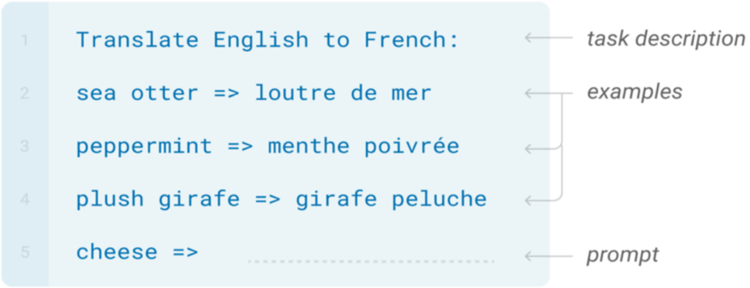
```
Translate English to French:          ← task description
sea otter => loutre de mer            ← examples
peppermint => menthe poivrée          ←
plush girafe => girafe peluche        ←
cheese => ..........................  ← prompt
```

Possible Explanation

Test time

In-context learning

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



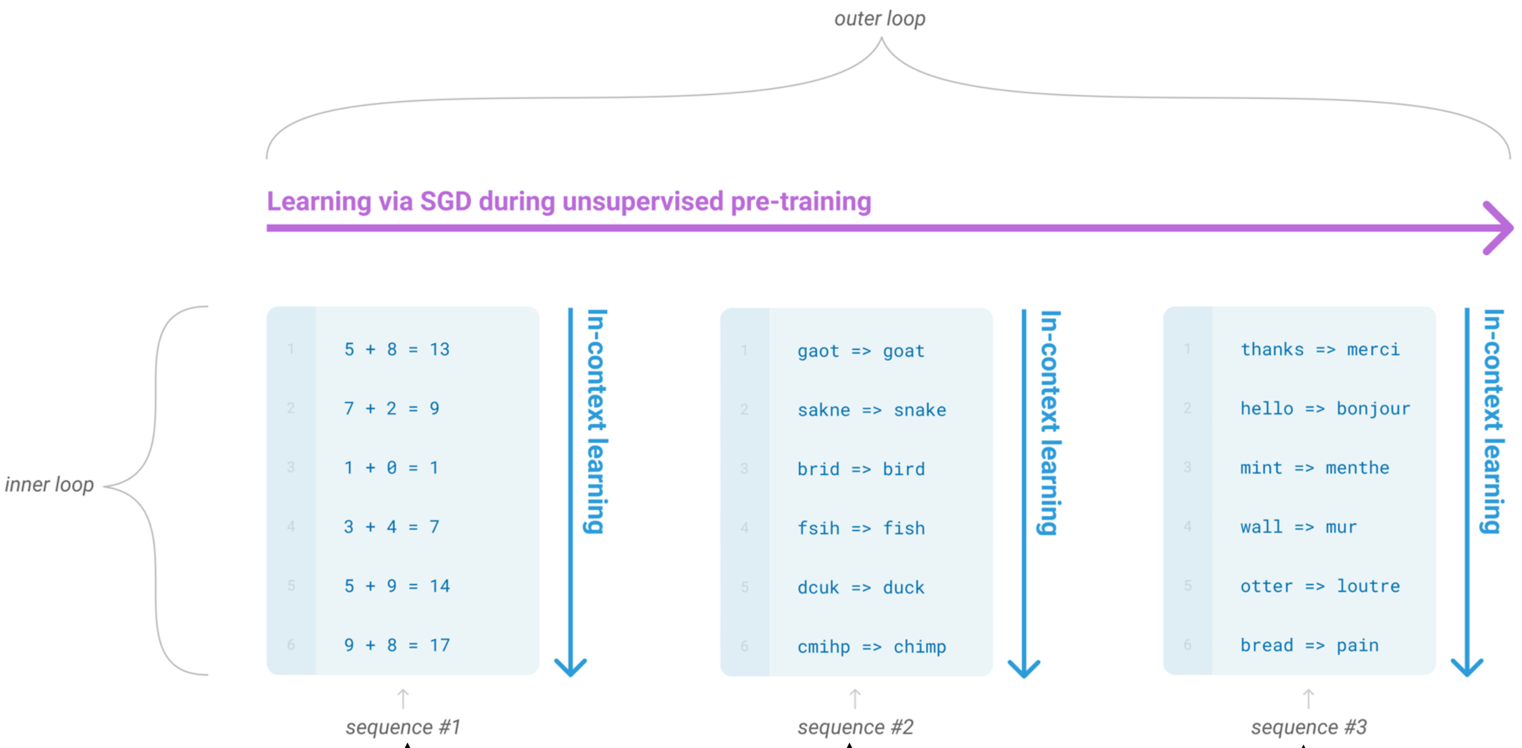
(Brown et. al. 2020)

θ

=

En to Fr phrase translation

Train time



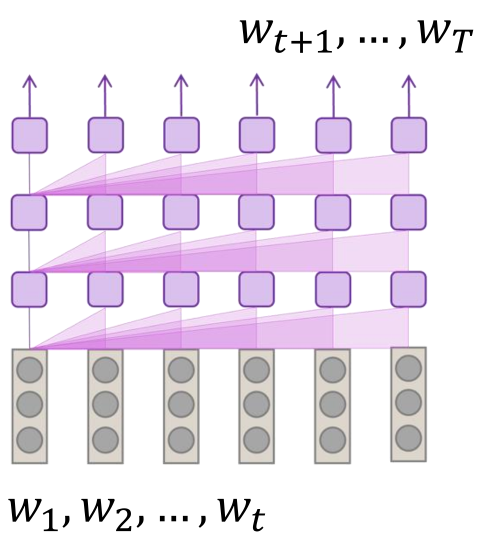
θ_1 = Digits addition

θ_2 = Word spelling correction

θ_3 = Translation

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

LLMs as Latent Variable Models

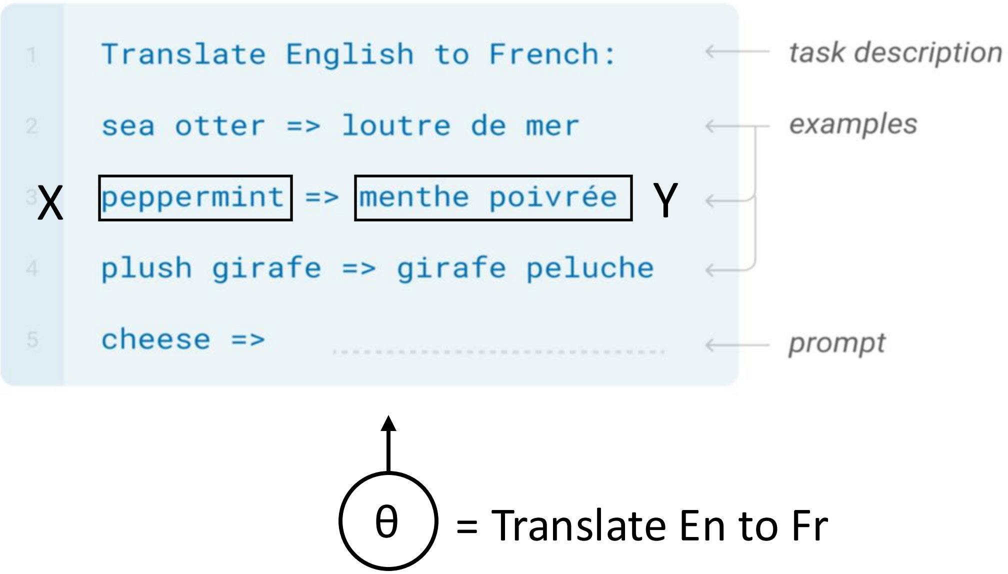


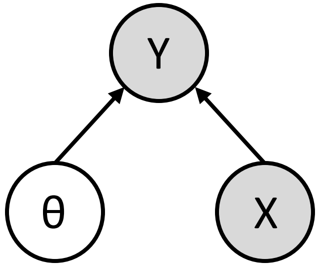
1. Implicitly infer a latent variable θ from the prompt

$$P_{LM}(w_{t+1:T} | w_{1:t}) = \int P_{LM}(w_{t+1:T} | \theta) P_{LM}(\theta | w_{1:t}) \, d\theta$$

2. Generate the continuation exclusively based on the inferred θ

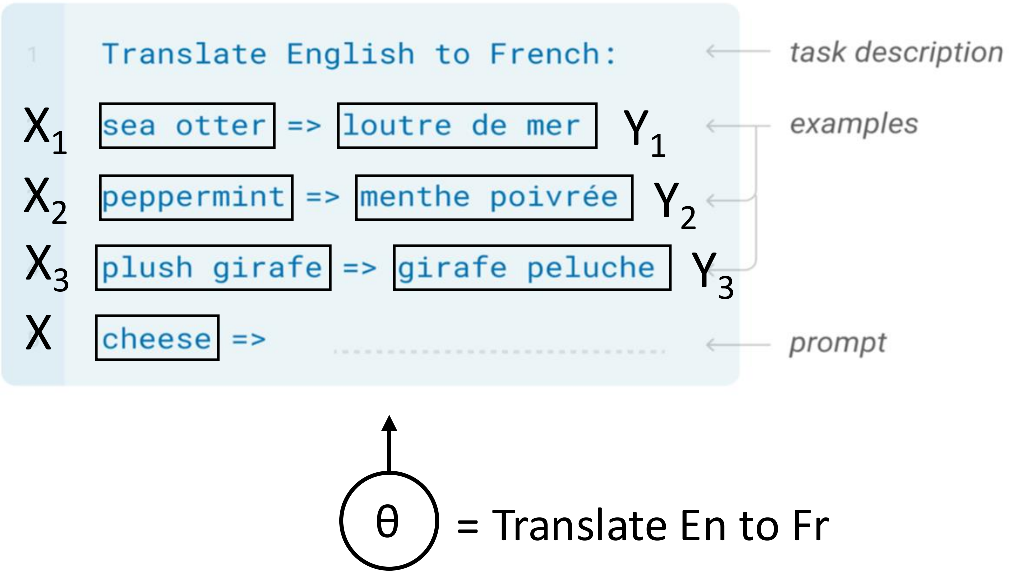
Bayes Optimal Classifier Assumption

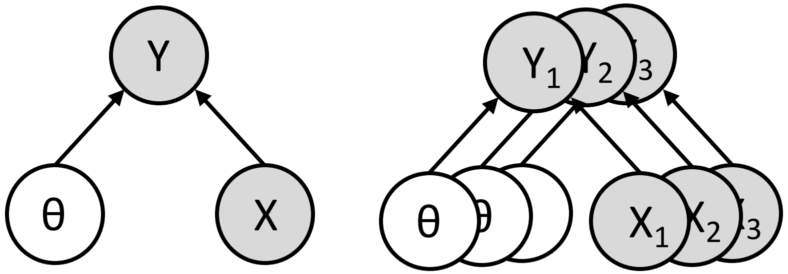




$P(Y|\theta, X)$ is Bayes optimal

Bayes Optimal Classifier Assumption





$P(Y|\theta, X)$ is Bayes optimal

In-context Learning Classifier

in-context learning classifier → $P_{LM}(Y|X_1, Y_1, X_2, Y_2, \dots, X_k, Y_k, X)$

$$= \int P_{LM}(Y|\theta, X) P_{LM}(\theta|X_1, Y_1, X_2, Y_2, \dots, X_k, Y_k, X)\, d\theta$$

closer

Bayes optimal classifier

demonstration selection criteria

Can we verify this theory in a real-world scenario?

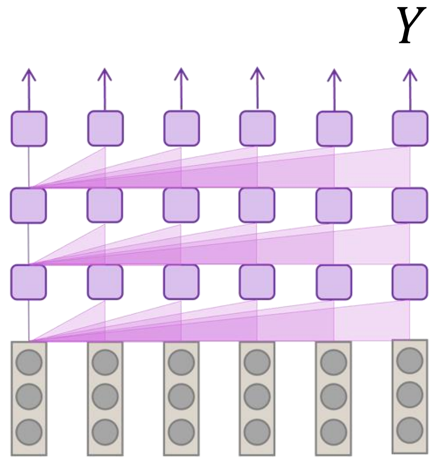
Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

A Real-World Testbed

Candidate examples

X_1 Y_1
X_2 Y_2

Optimal demonstrations



X_1, Y_1, X_2, Y_2, X

Improved performance

Among the firsts to formally propose the task of **demonstration selection**

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

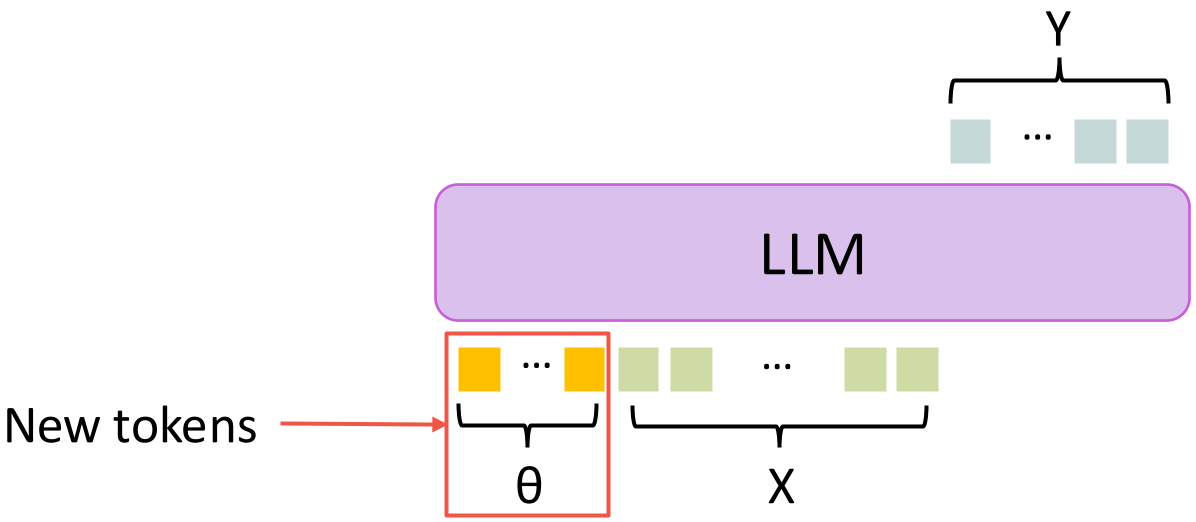
Our Proposed Method

$$P_{LM}(Y|X_1, Y_1, X_2, Y_2, \ldots, X_k, Y_k, X)$$

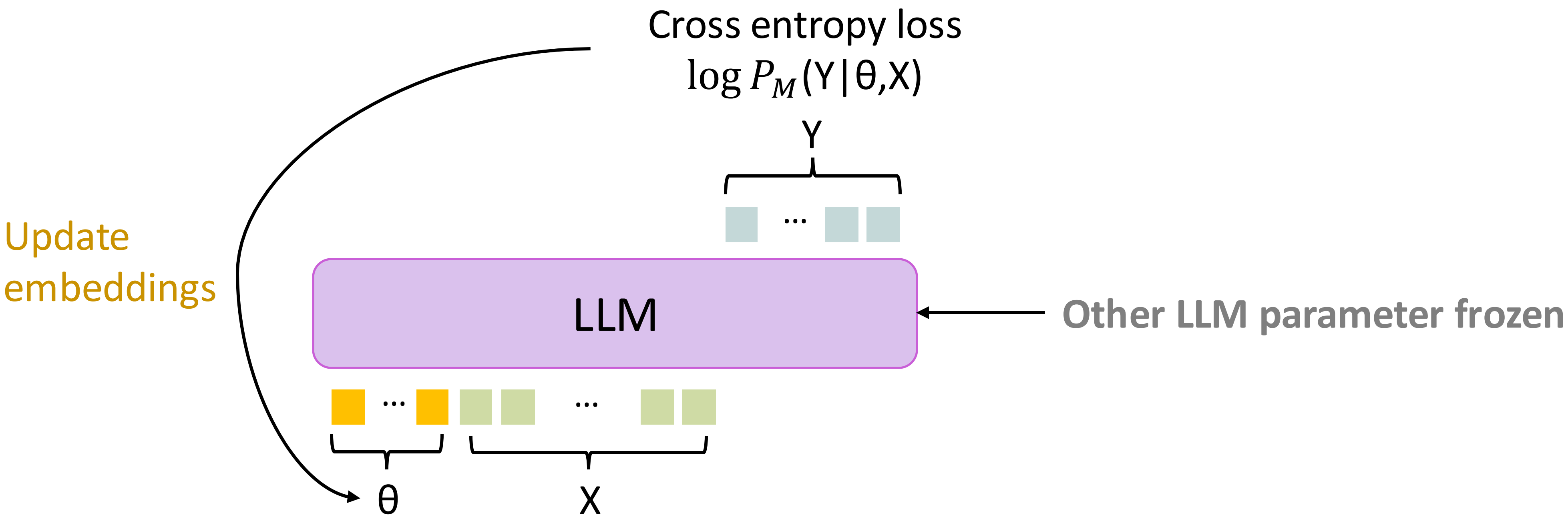
$$= \int P_{LM}(Y|\theta, X) P_{LM}(\theta|X_1, Y_1, X_2, Y_2, \ldots, X_k, Y_k, X)\, d\theta$$

Latent Intent Learning → **Score Computation** → **Demonstration Selection**

Latent Intent Learning

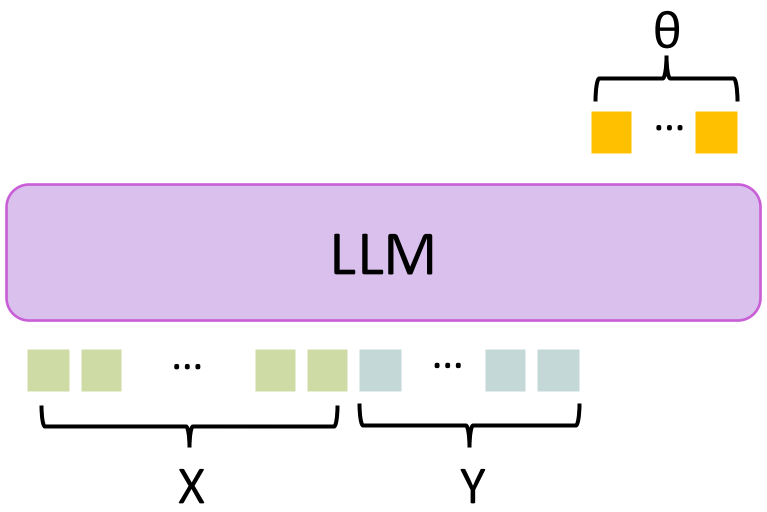


Latent Intent Learning



Score Computation

Score: Language model probability
$P_M(\theta|X,Y)$



Demonstration Selection

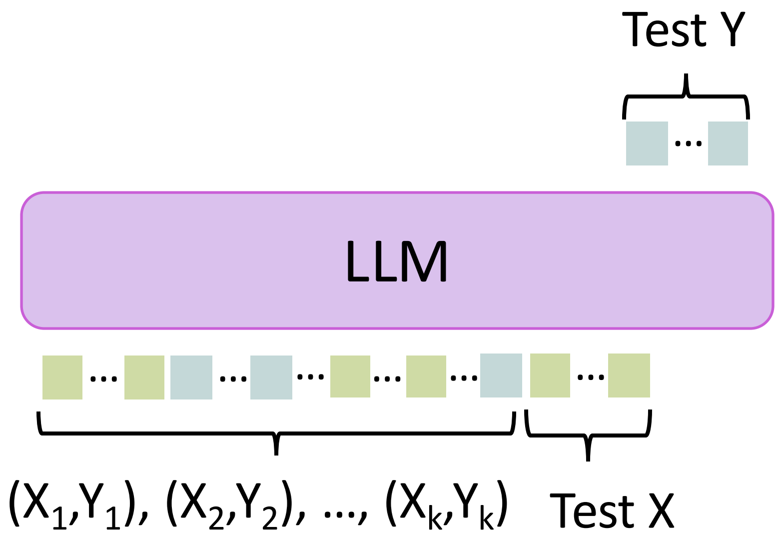
Score: Language model probability

$P_M(\theta | X, Y)$

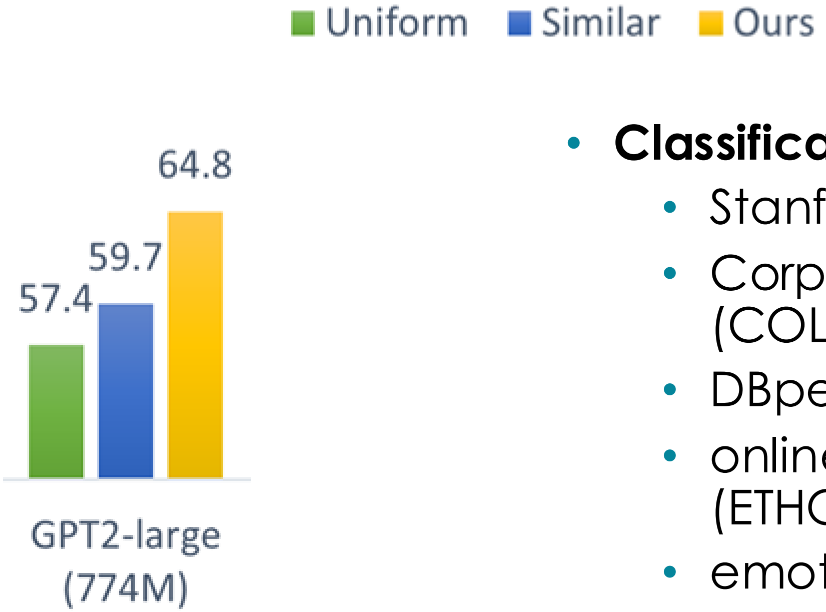
Score each candidate

Top K: $(X_1,Y_1), (X_2,Y_2), \ldots, (X_k,Y_k)$

Test Performance



Improved Performance

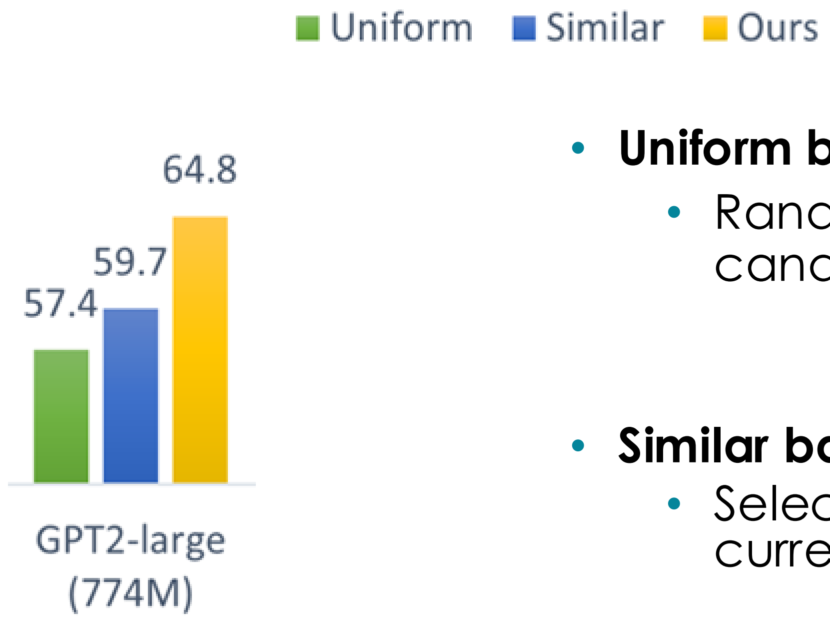


- **Classification Tasks**:
  - Stanford Sentiment Treebank (SST2)
  - Corpus of Linguistic Acceptability (COLA)
  - DBpedia ontology classification
  - online hate speech detection (ETHOS)
  - emotion prediction

- **Generation Task**:
  - Grade School Math 8K(GSM8K)

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

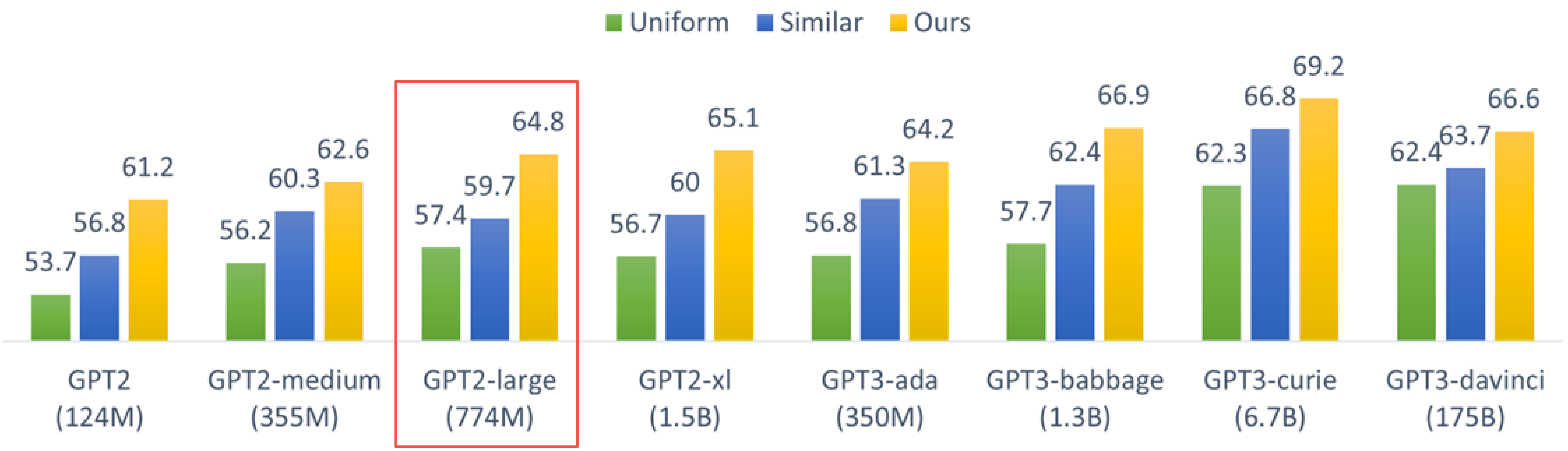
Improved Performance



- **Uniform baseline**:
  - Randomly select k examples from candidate set
- **Similar baseline**:
  - Select k examples most similar to current testing input

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

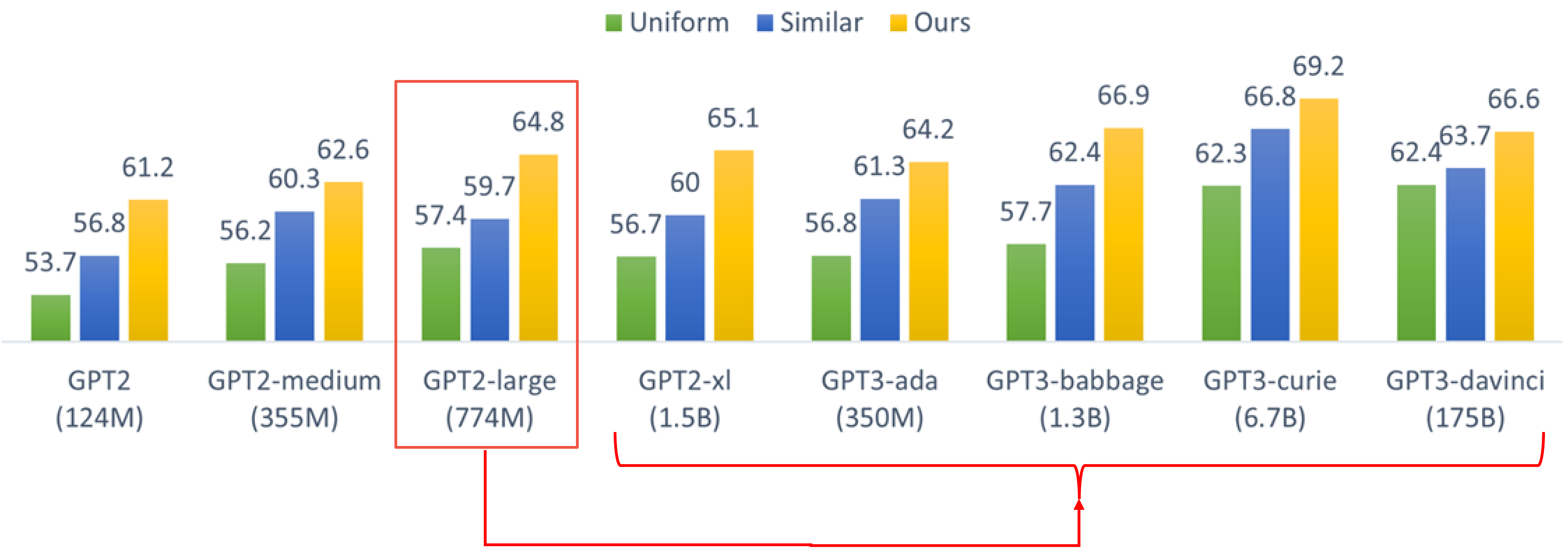
Improved Performance of Larger Models



Small Model → **Large Model**

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

Improved Performance of Larger Models



We can align large models with small model's intent!

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

Follow-ups

Stanford University

LESS: Selecting Influential Data for Targeted Instruction Tuning

Mengzhou Xia [1*] Sadhika Malladi [1*] Suchin Gururangan [2] Sanjeev Arora [1] Danqi Chen [1]

Massachusetts Institute of Technology

Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations

Zeming Wei[1] Yifei Wang[2] Ang Li[1] Yichuan Mo[1] Yisen Wang[1*]
[1]Peking University [2]MIT CSAIL

Many-Shot In-Context Learning

Rishabh Agarwal*, Avi Singh*, Lei Zhang†, Bernd Bohnet†, Luis Rosias†, Stephanie Chan†, Biao Zhang†, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, Hugo Larochelle
Google DeepMind

DeepMind

Trained Transformers Learn Linear Models In-Context

Ruiqi Zhang RQZHANG@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720-3860, USA

Spencer Frei SFREI@UCDAVIS.EDU
Department of Statistics
University of California, Davis
4118 Mathematical Sciences Building
399 Crocker Ave., Davis, CA 95616, USA

Peter L. Bartlett PETER@BERKELEY.EDU
Department of Statistics and Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720-3860, USA
Google DeepMind
1600 Amphitheatre Parkway
Mountain View, CA 94040, USA

Berkeley UNIVERSITY OF CALIFORNIA

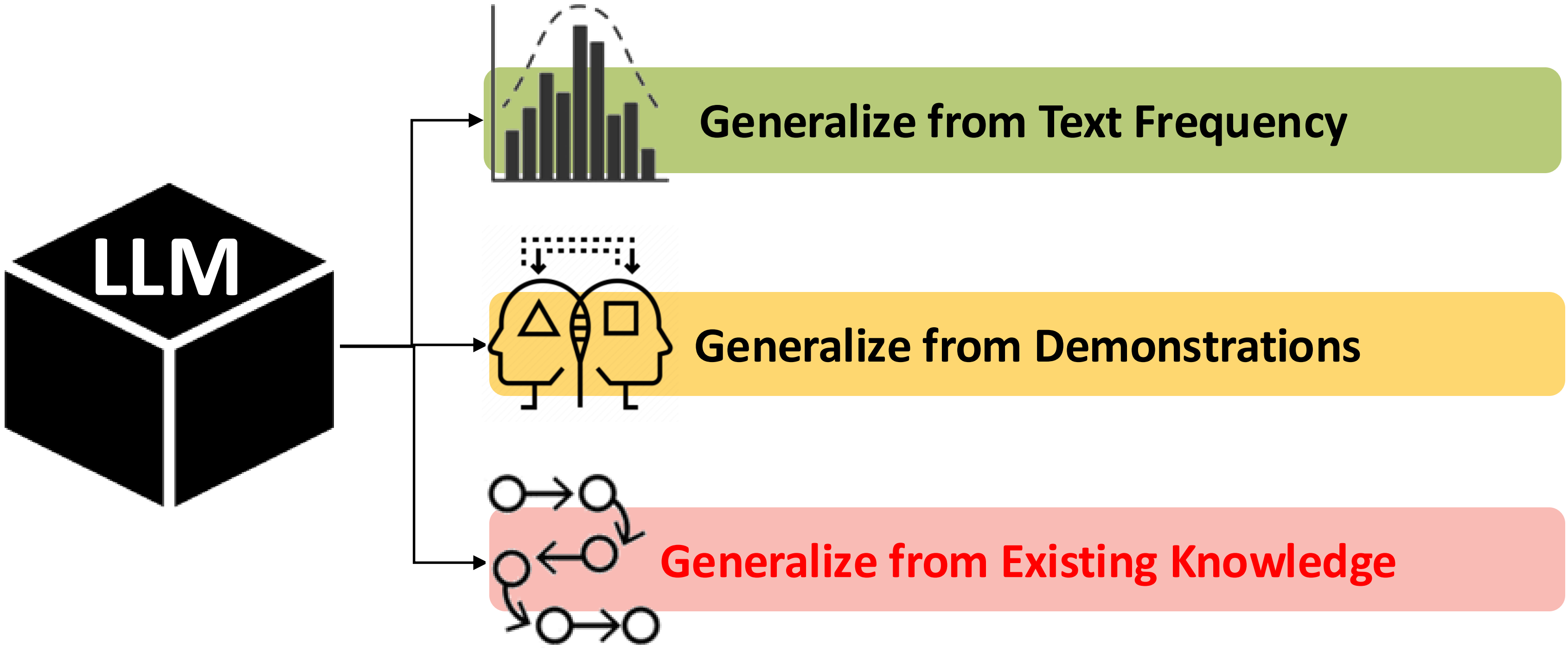
Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

Takeaways

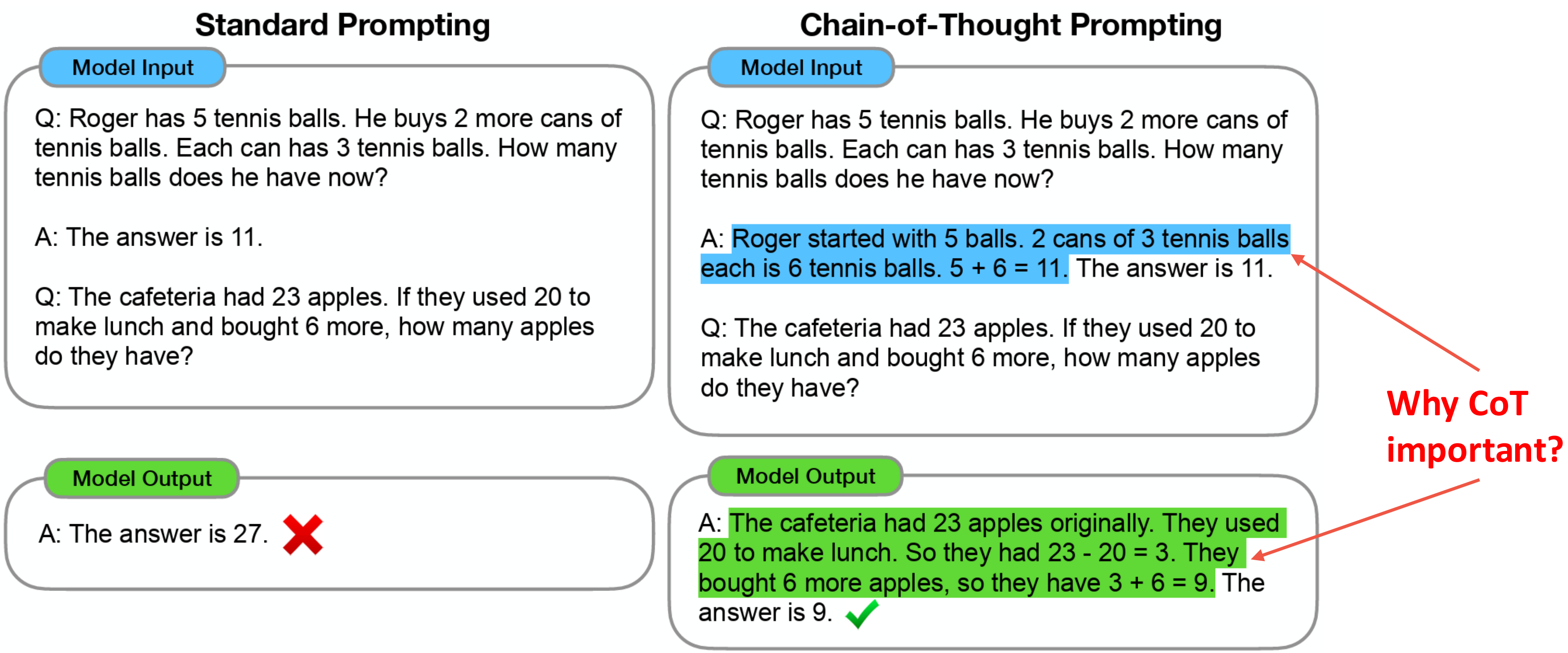
- In-context learning can be understood as emerged through latent variable inference.
- Demonstrations selected by small LM can be transferred to improve larger LMs' performance.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, William Yang Wang. Large Language Models are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. NeurIPS 2023.

Outline

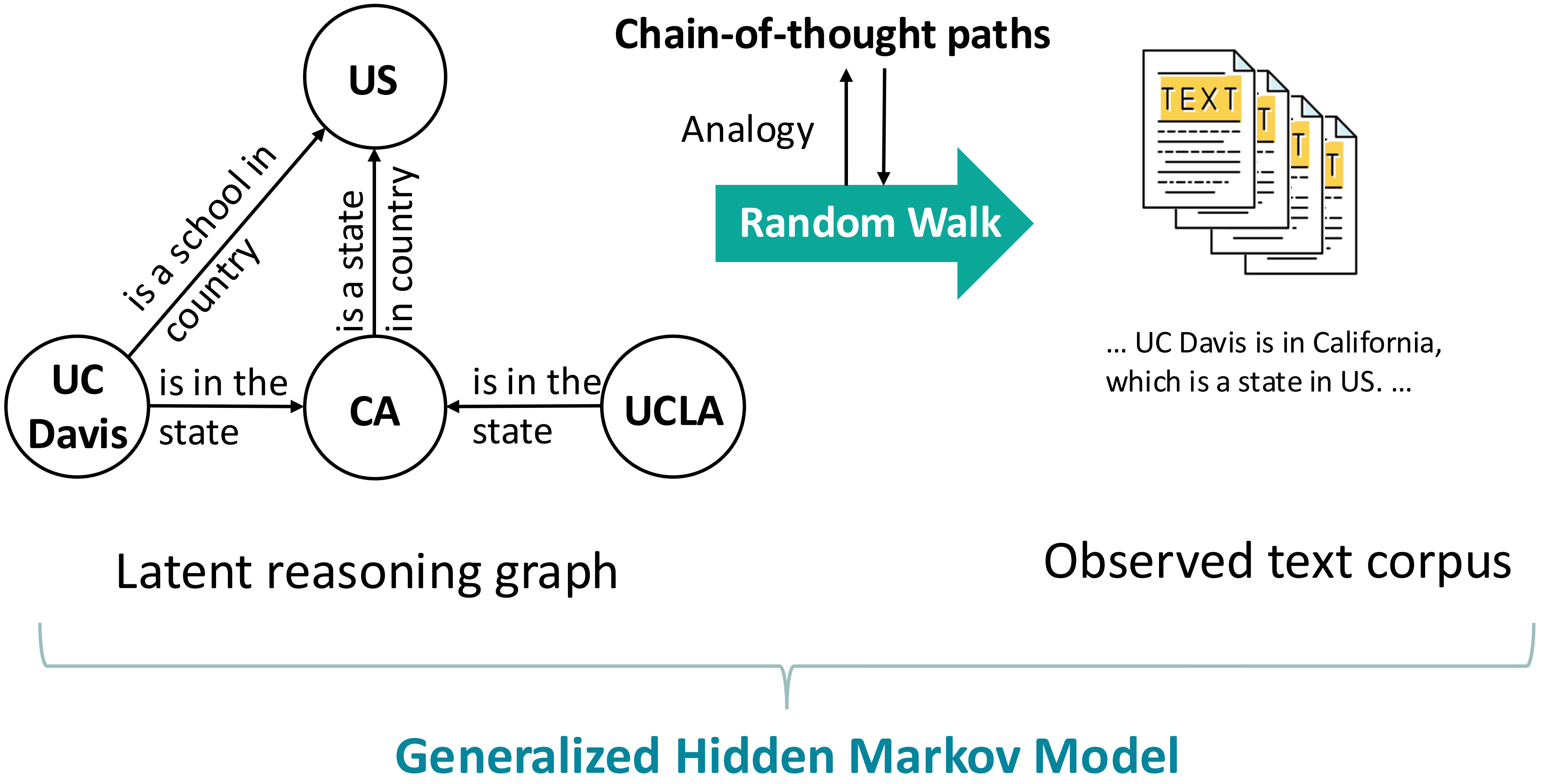

LLM
Generalize from Text Frequency
Generalize from Demonstrations
Generalize from Existing Knowledge

Chain-of-Thought Reasoning



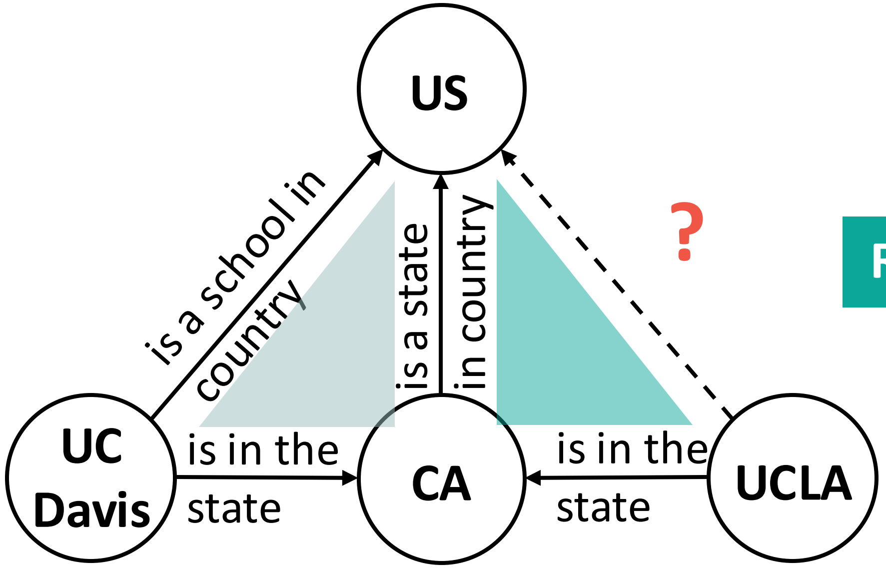
Hypothesis: CoT verbalizes the pretraining data generation process.

Data Generation Process Assumption



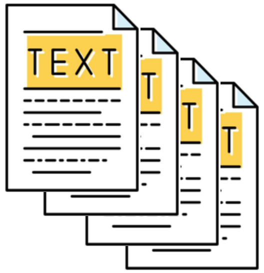
Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Novel Discovery



Latent reasoning graph

Random Walk

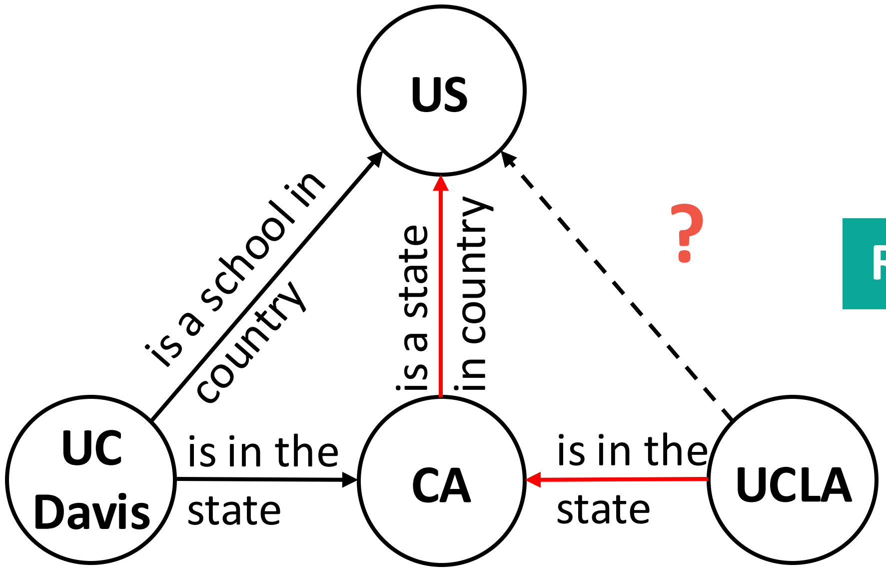


… UC Davis is in California, which is a state in US. …

Observed text corpus

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Path Aggregation Hypothesis

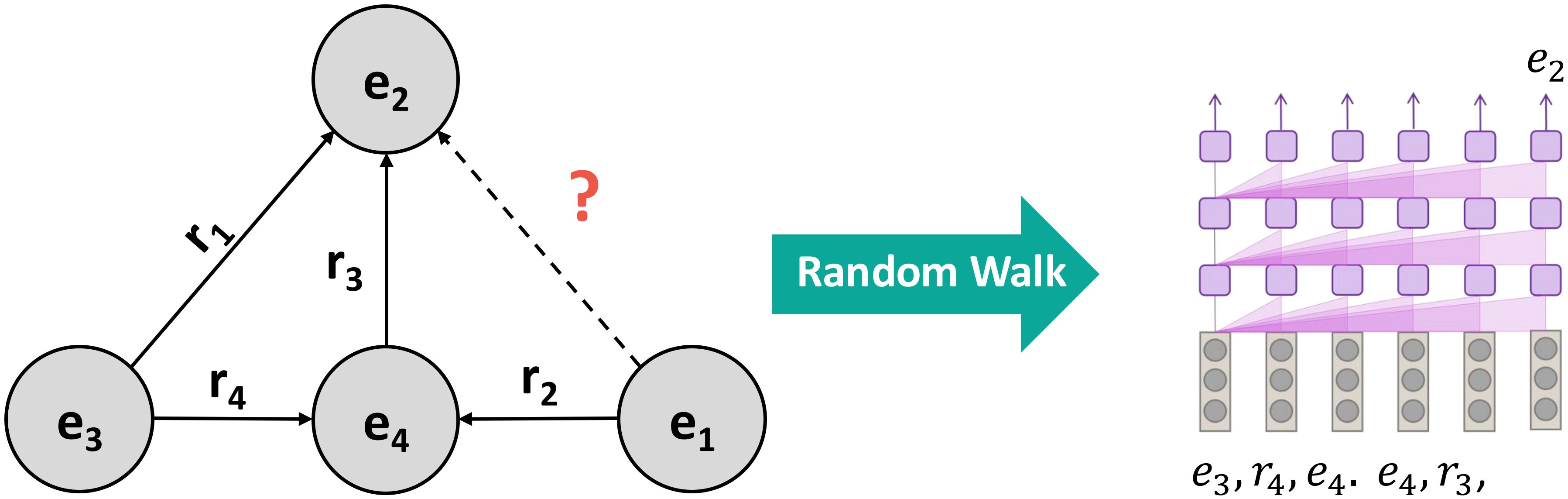


... UC Davis is in California, which is a state in US. ...

$$P_{LM}(\text{UCLA} \overset{\text{in}}{\dashrightarrow} \text{US}) \propto \exp[w_1 P_D(\text{UCLA} \xrightarrow{\text{in}} \text{CA} \xrightarrow{\text{in}} \text{US}) + w_2 P_D(\text{UCLA} \xrightarrow{\text{in}} \text{CA} \xleftarrow{\text{in}} \text{UC Davis} \xrightarrow{\text{in}} \text{US})]$$

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

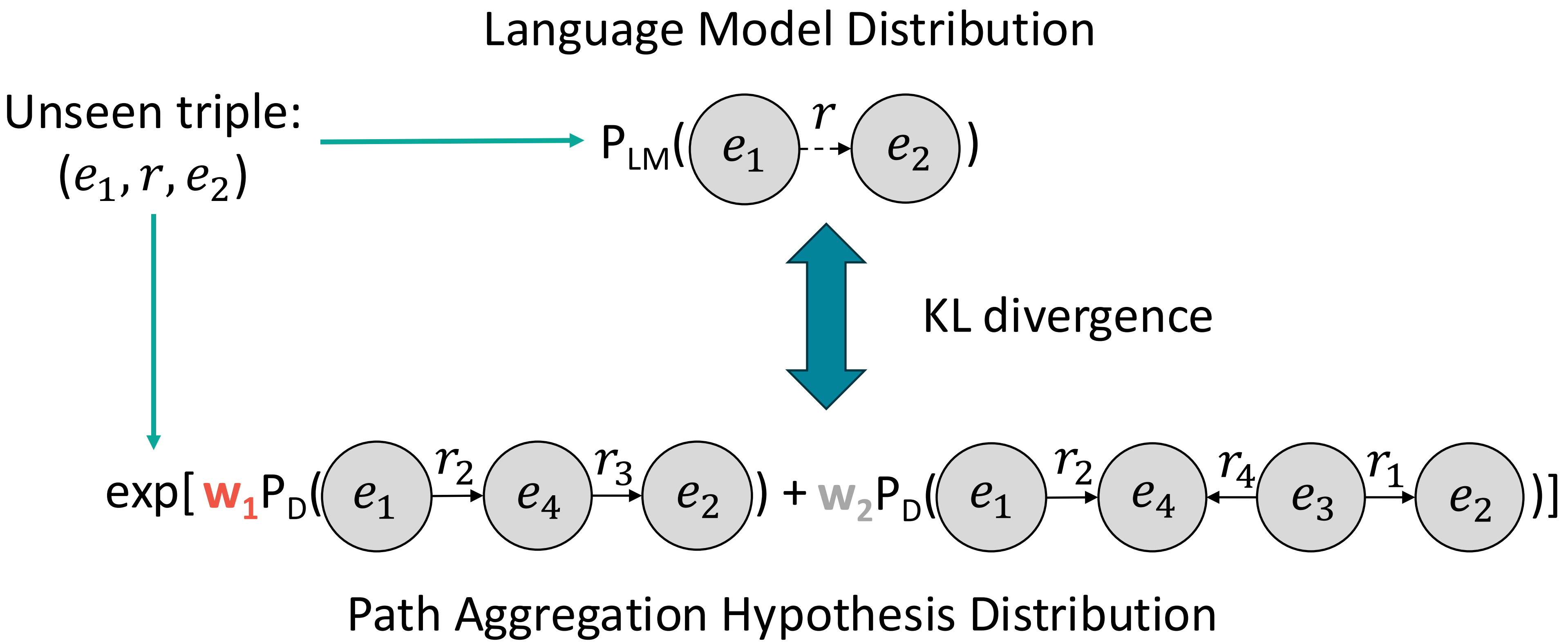
Experiment Setup



- **Idea**: pretrain a language model on random walk paths sampled from a knowledge graph from scratch.
- Each entity and relation is a token.
- Test on missing edges.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Verify Hypothesis



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

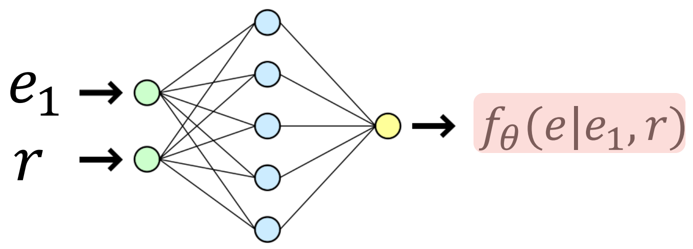
Language Model Distribution Definition

Language Model

$$P_{\text{LM}}(e_2|e_1, r) = \frac{\exp\left(f_\theta(e_2|e_1, r)\right)}{\sum_{e \in \mathcal{E}} \exp\left(f_\theta(e|e_1, r)\right)}$$

All Entities

Transformer



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Hypothesized Distribution Definition

Weighted Path Aggregation

$$P_w(e_2|e_1, r) = \frac{\exp(S_w(e_2|e_1, r)/T)}{\sum_{e \in \mathcal{E}} \exp(S_w(e|e_1, r)/T)}$$

Temperature

Path ranking algorithm (PRA) (Lao et. al. 2011)

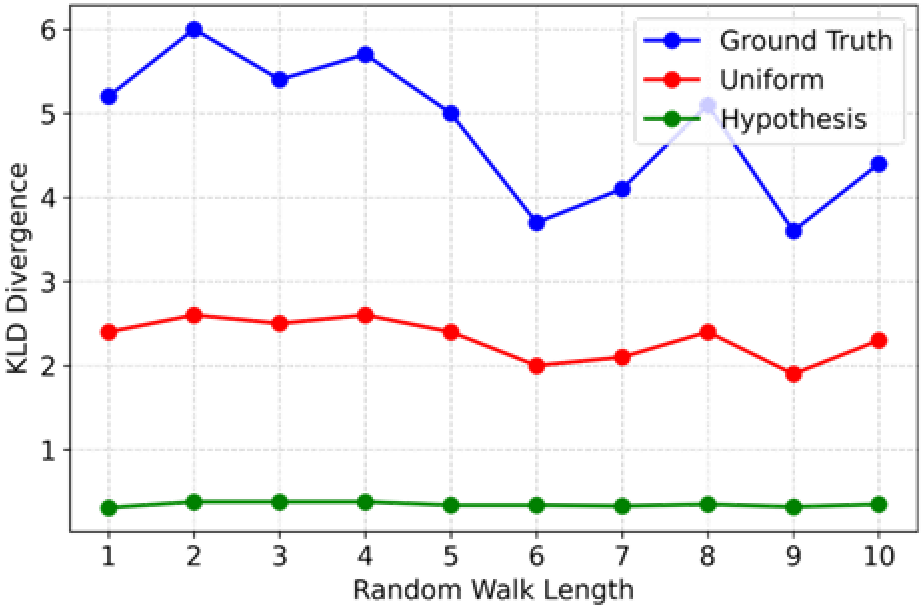
$$S_w(e_2|e_1, r) = \sum_{h \in \mathcal{H}} w_r(h) P(e_2|e_1, h)$$

Pattern weight learned by logistic regression

Sum of Random walk paths probability

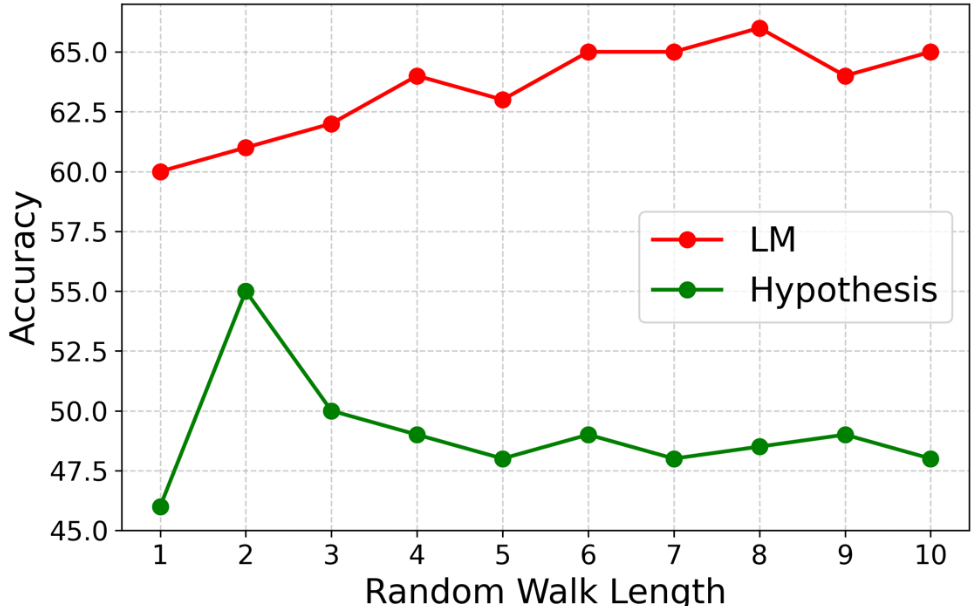
Verifying Path Aggregation Hypothesis

KL Divergence



LM distribution is close to hypothesized distribution

Prediction Accuracy



LM learns better path weights by utilizing context

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Practical Implication

Random walk paths play an essential role in LLM reasoning

↓

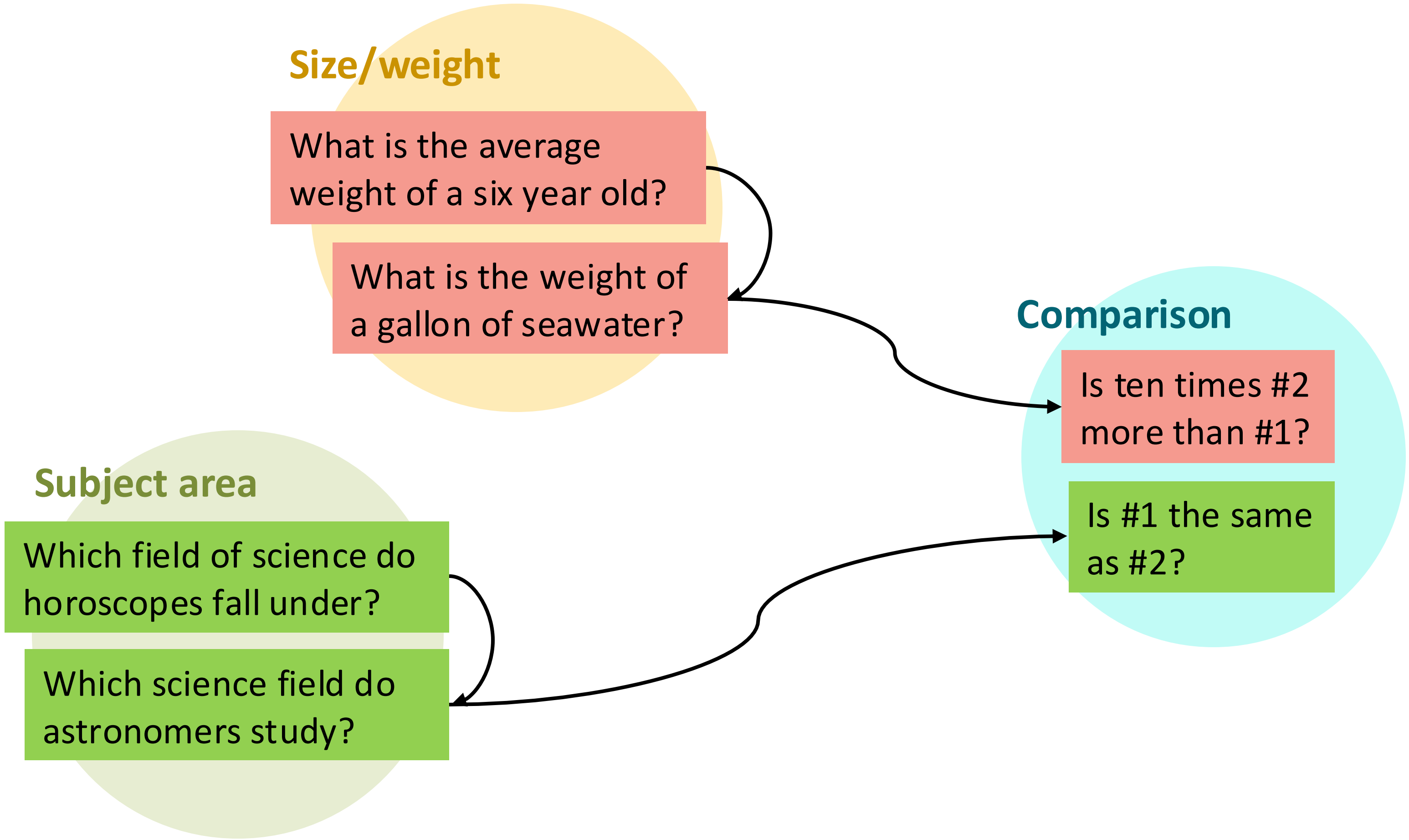
Can we augment random walk paths into real world CoT paths?

↓

Would training on this augmented data improve real world reasoning performance?

CoT Graph

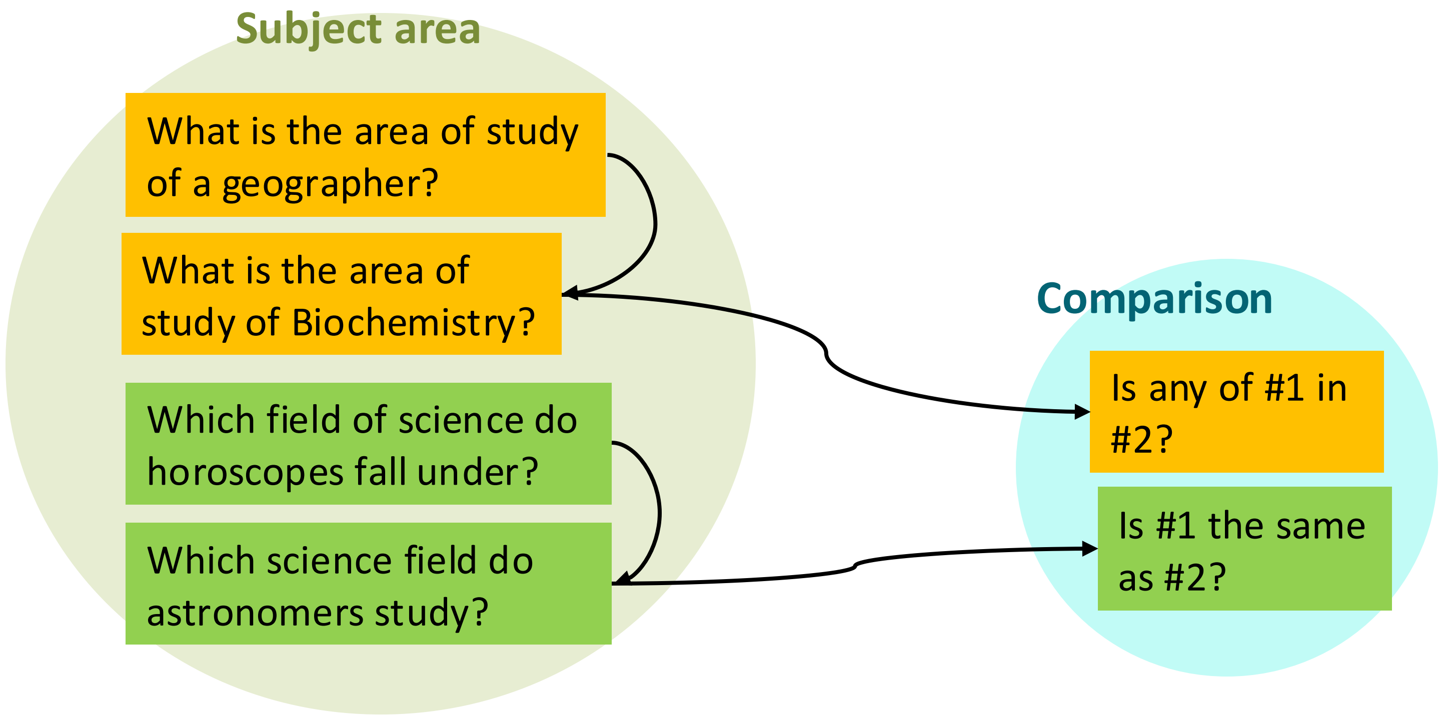
- Organize real-world CoT paths into a graph by clustering steps.



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Random Walk Augmentation

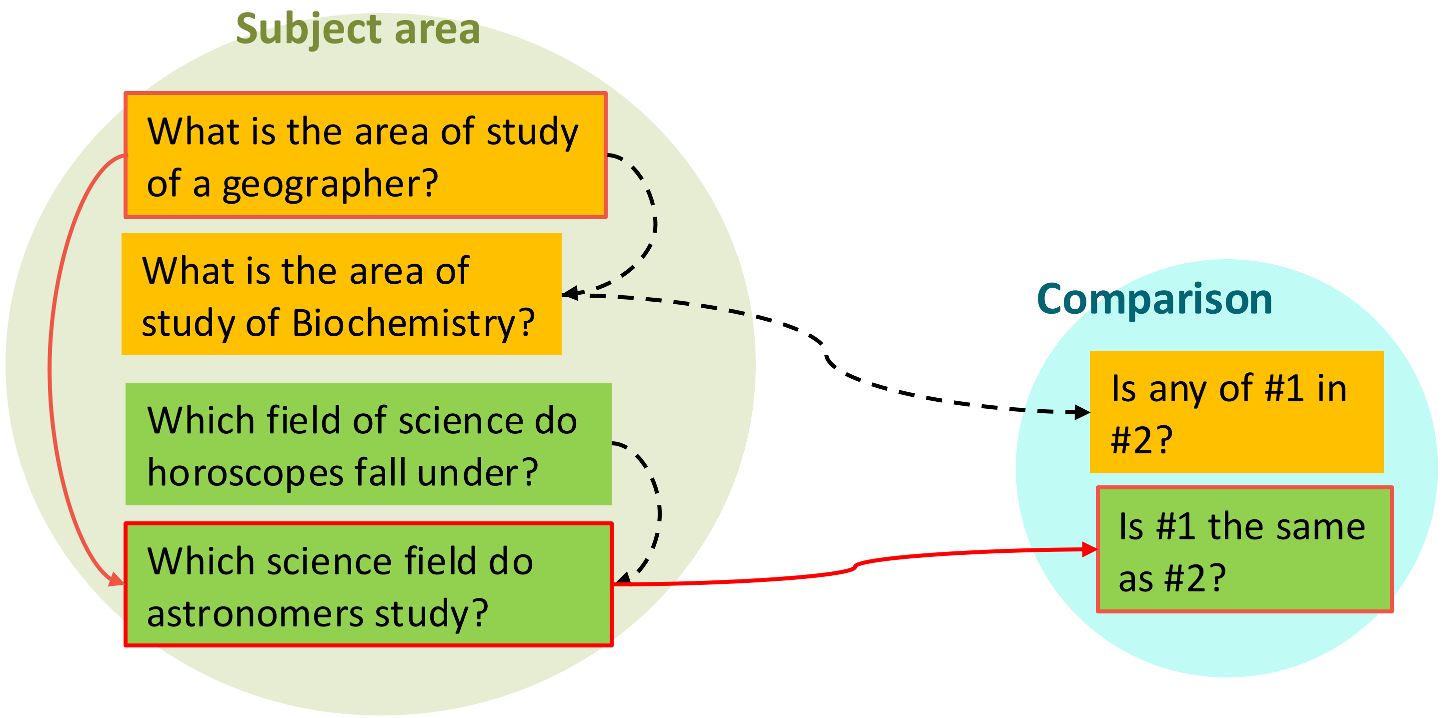
- Reorganize CoT steps by random walk over the graph.



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Random Walk Augmentation

- Reorganize CoT steps by random walk over the graph.



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Improved Performance

Math word problems (GSM8K, AQUA, SVAMP) · Multi-hop QA (StrategyQA) · Logical reasoning (LogicalDeduction)

Model	Method	GSM8K	AQUA	SVAMP	StrategyQA	LogicalDeduction	Avg.
Gemma (2B)	SFT	24.8	31.4	56.4	54.2	50.7	43.5
	Ours	**26.1**	**33.9**	**60.3**	**56.3**	**51.6**	**45.6**
Yi (6B)	SFT	32.2	37.0	65.8	65.8	62.2	52.6
	Ours	**33.1**	**39.8**	**67.0**	**70.0**	**63.3**	**54.6**
Llama 2 (7B)	SFT	26.8	30.0	53.3	58.4	55.3	44.8
	Ours	**28.5**	**34.6**	**55.8**	**63.7**	**56.1**	**47.7**
Llama 2 (13B)	SFT	37.1	35.0	66.4	69.5	55.7	52.7
	Ours	**41.2**	**37.4**	**69.0**	**71.2**	**57.7**	**55.3**

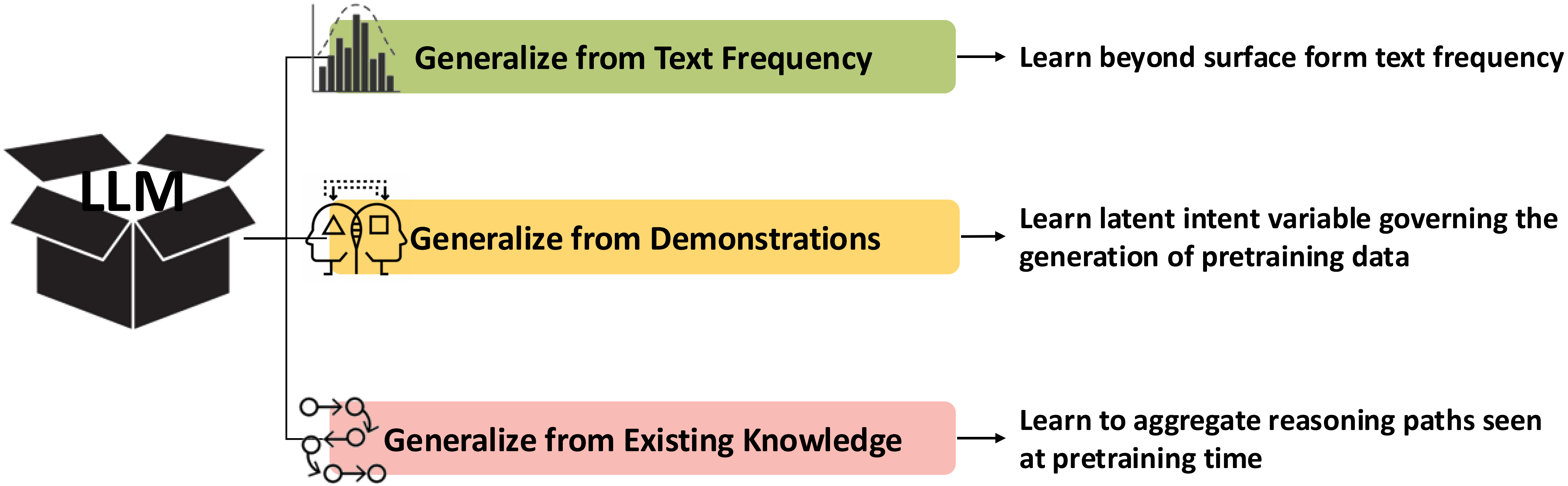
Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Takeaways

- Novel conclusions discovered by LLMs can be explained by aggregating reasoning paths seen at training time.
- LLMs' reasoning ability can be improved by training on random walk augmented chain-of-thoughts.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Recap

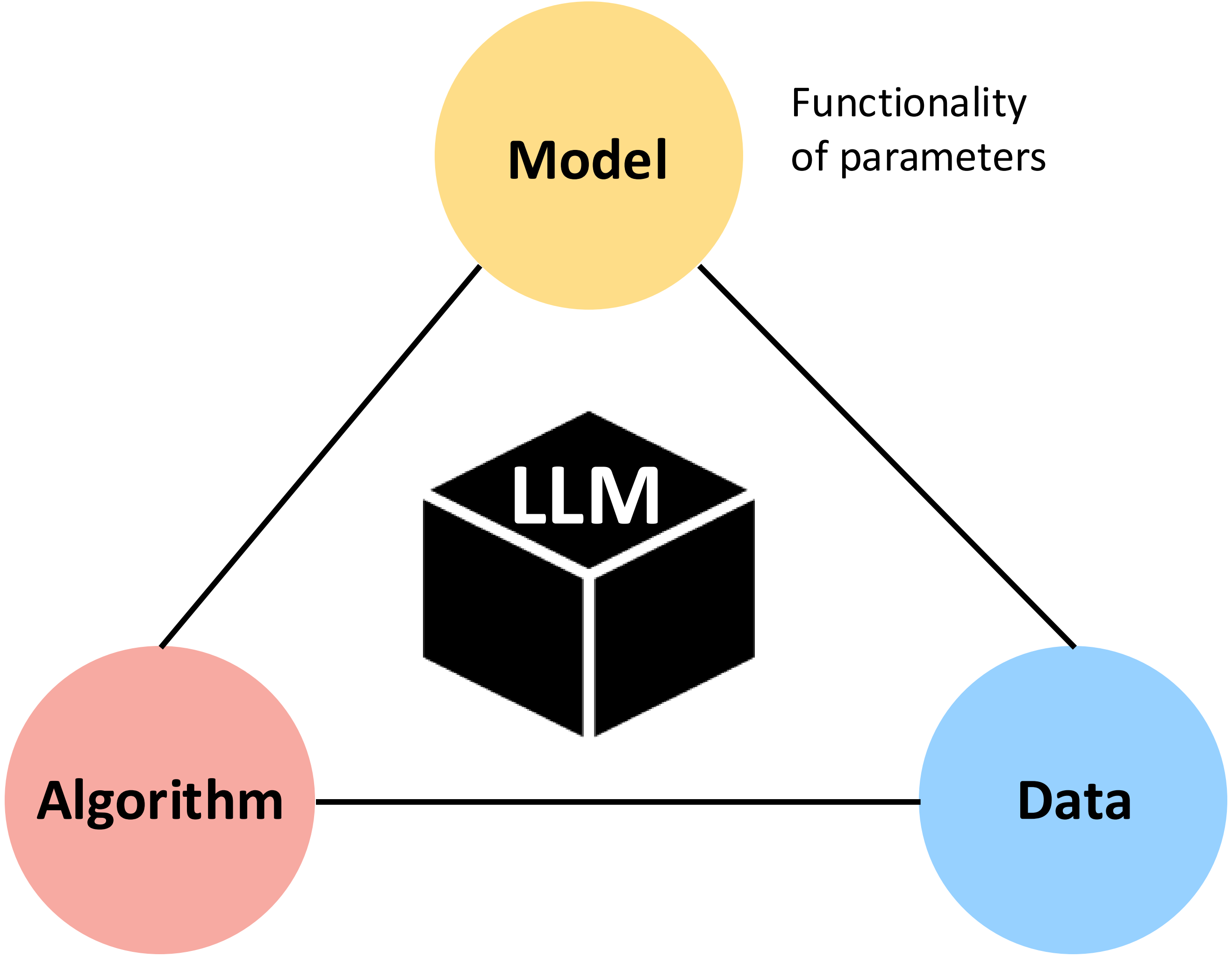

LLM
Generalize from Text Frequency
Learn beyond surface form text frequency
Generalize from Demonstrations
Learn latent intent variable governing the generation of pretraining data
Generalize from Existing Knowledge
Learn to aggregate reasoning paths seen at pretraining time

Other Works

- **Understanding LLMs** [NeurIPS 2023, ICML 2024, ICLR 2025] √
- **Address Foundational Limitations of Deep Learning**
  - Spurious Correlation [NeurIPS 2021, ICLR 2023]
  - Structured Reasoning [TMLR 2023, EMNLP 2023, COLM 2024]
- **Open Source Contributions**
  - Dataset [NeurIPS 2021, EMNLP 2023]
  - Survey [TACL 2024, TMLR 2024]

Open the Black Box

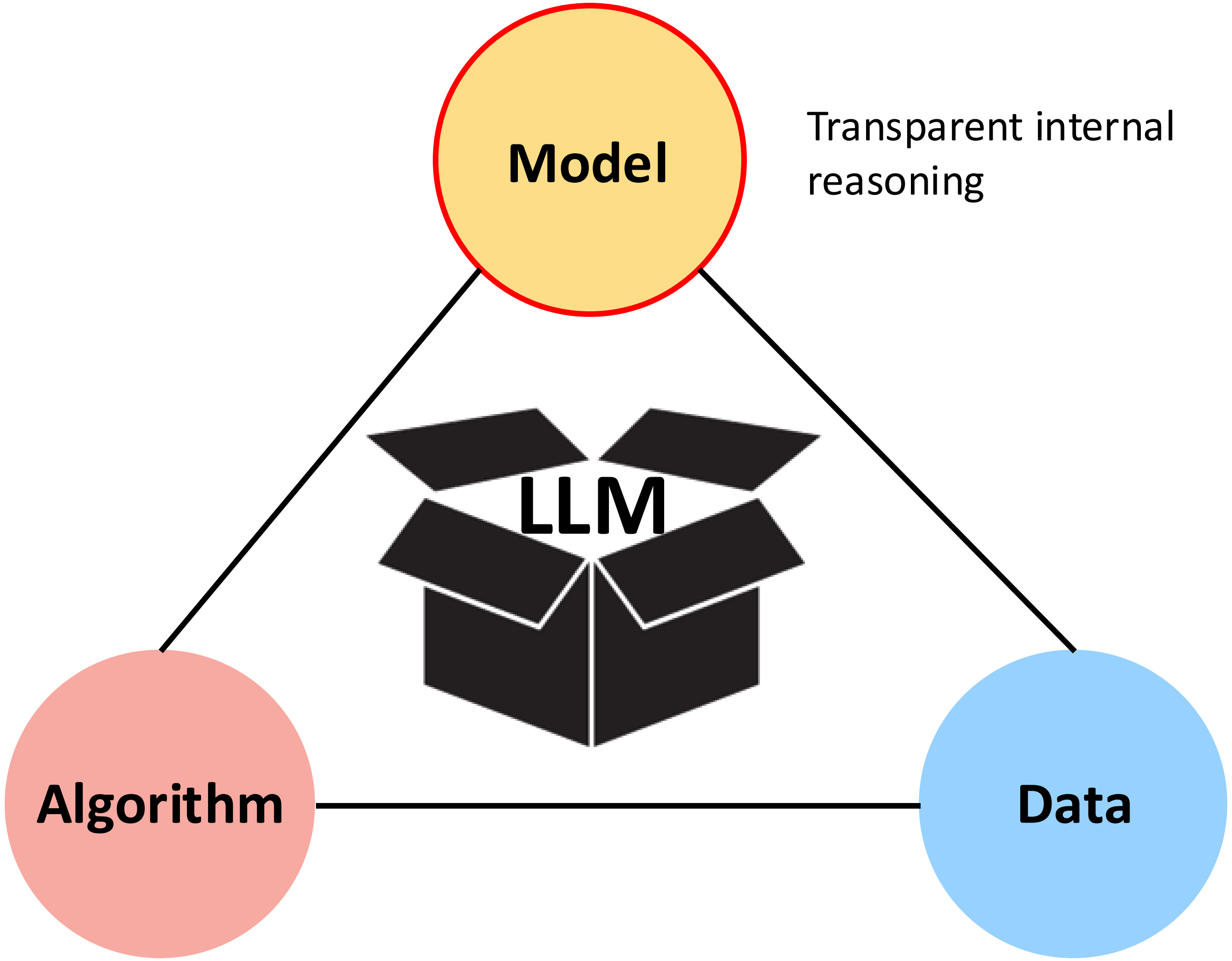
Functionality of parameters



Effect of training and inference algorithms

Contribution of each training example

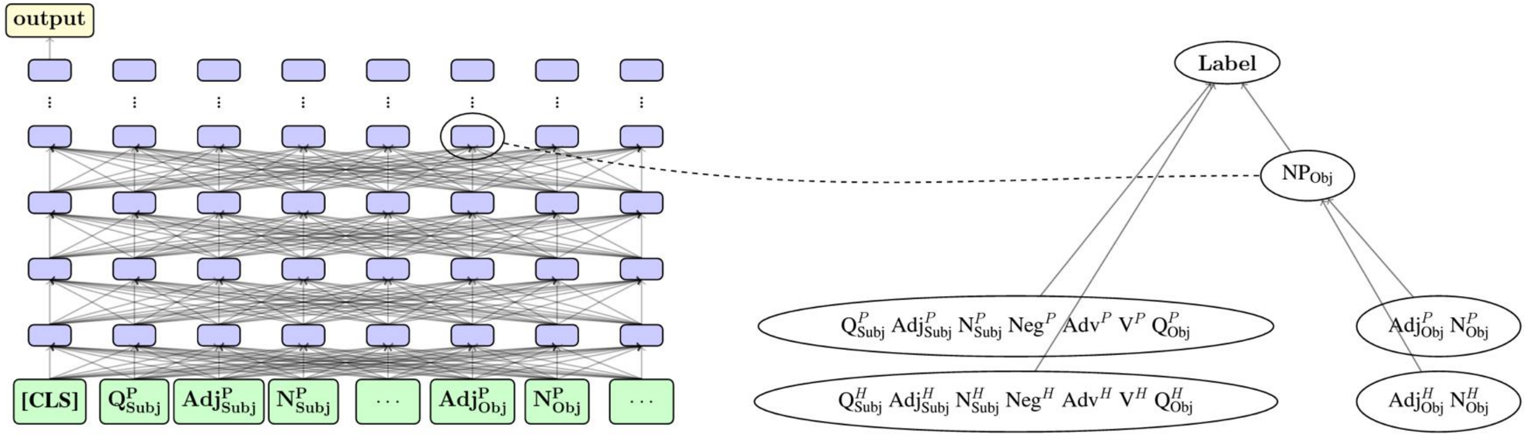
Open the Black Box



Transparent internal reasoning

Future Directions

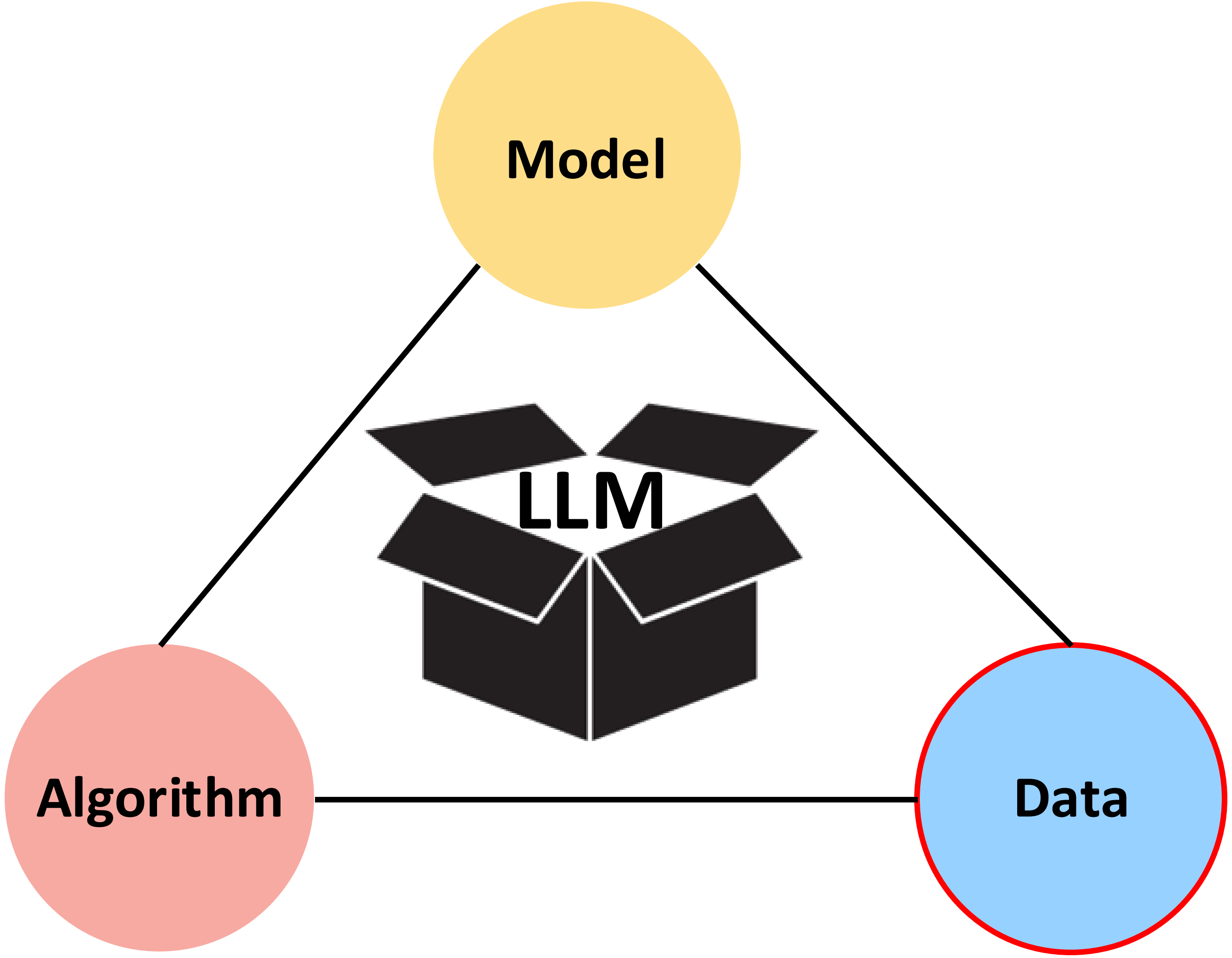
Causal abstractions of LLMs



(Geiger et. al. 2021)

Transparent decision making

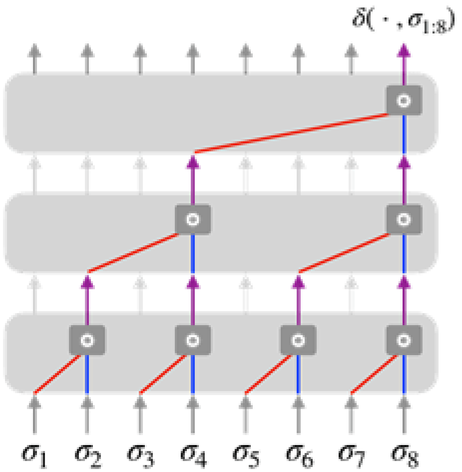
Open the Black Box



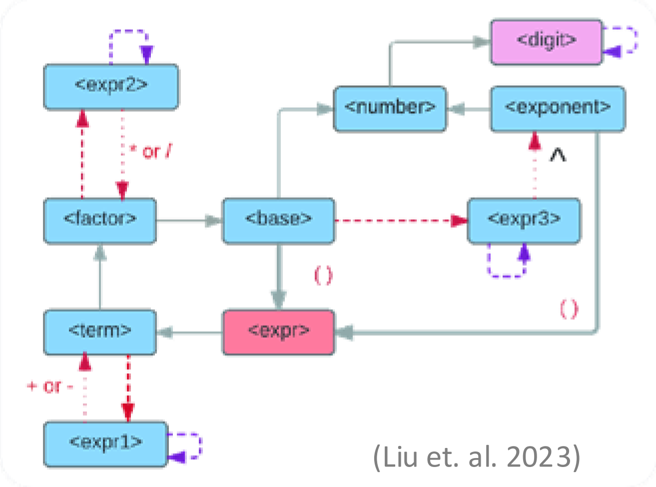
Trace the origin of model behaviors

Future Directions

Realistic synthetic data for understanding LLM behaviors

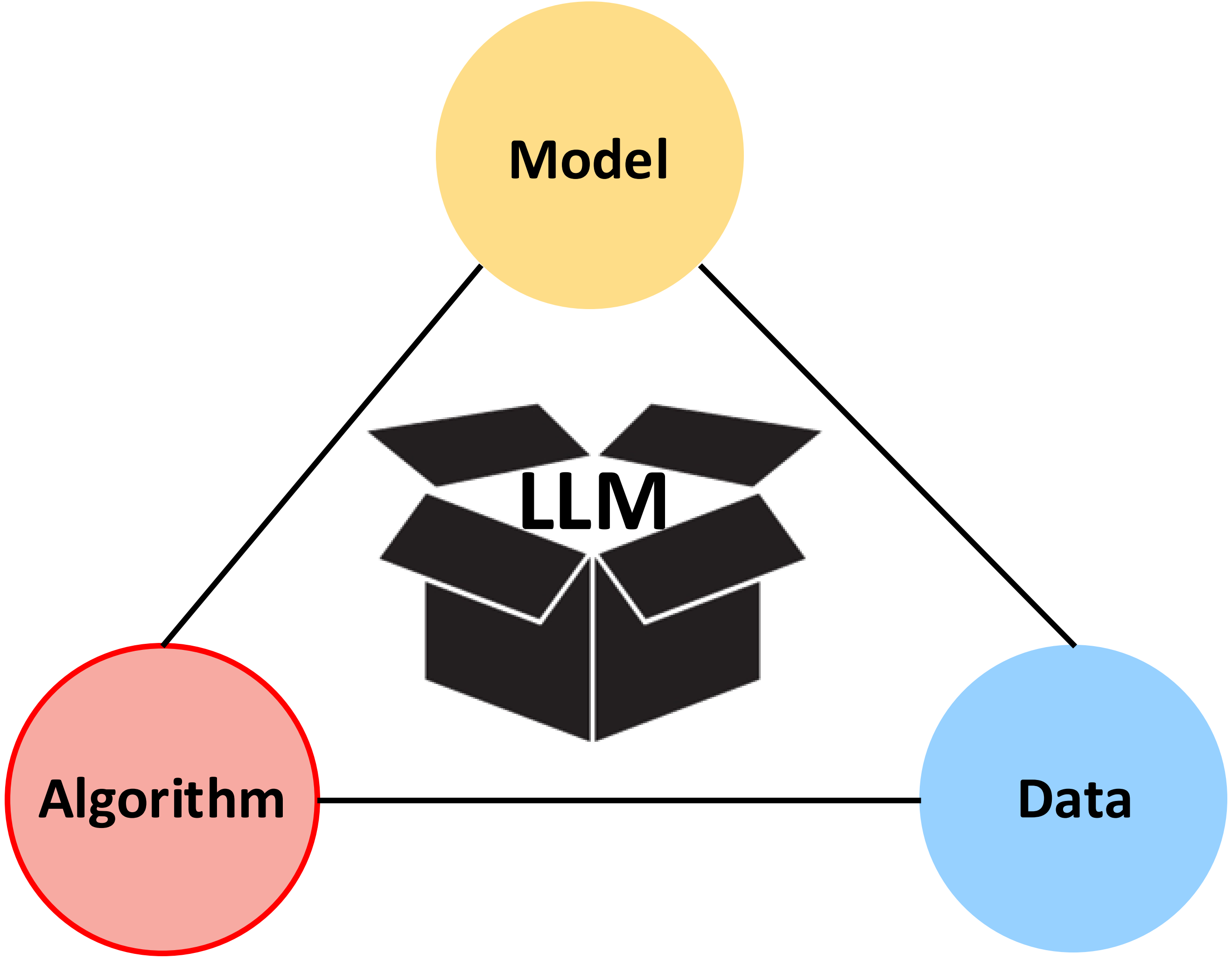


Train



Controlled experiments

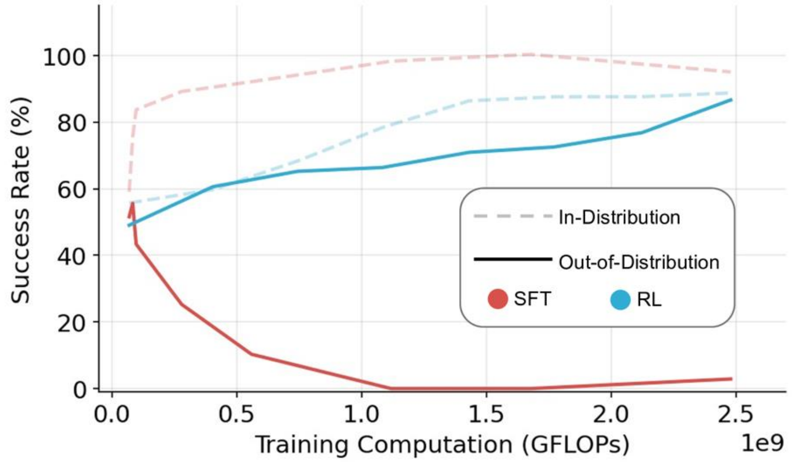
Open the Black Box



Understand deciding factors of training and inference algorithms

Future Directions

Reinforcement learning v.s. fine-tuning



(Chu et. al. 2025)

Understanding algorithmic weaknesses

Acknowledgement

Carnegie Mellon University

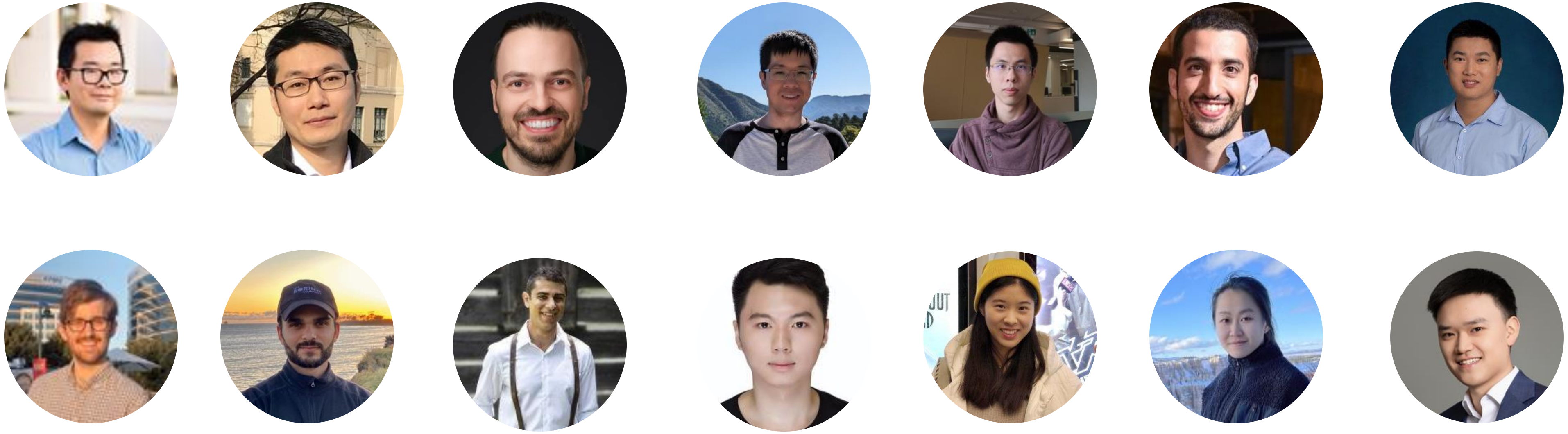
Microsoft

University of Waterloo — Concordia cum Veritate

MIT-IBM Watson AI Lab

Ai2

The University of Arizona



Thank you!

Questions?

UC SANTA BARBARA