8/2025

Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data

Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang
Work done at UC Santa Barbara. Published in ICLR 2025.

Are LLMs Stochastic Parrots?

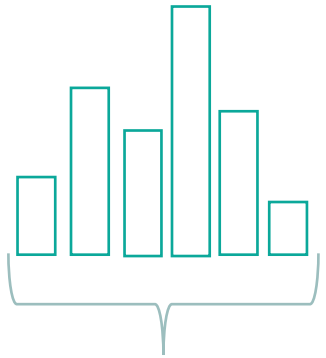
(Bender et. al. 2021)

- Do LLMs truly understand the meaning of language?
- Are LLMs truly capable of thinking?

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 ___

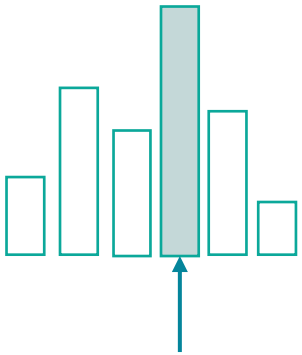


Vocabulary

Language Models

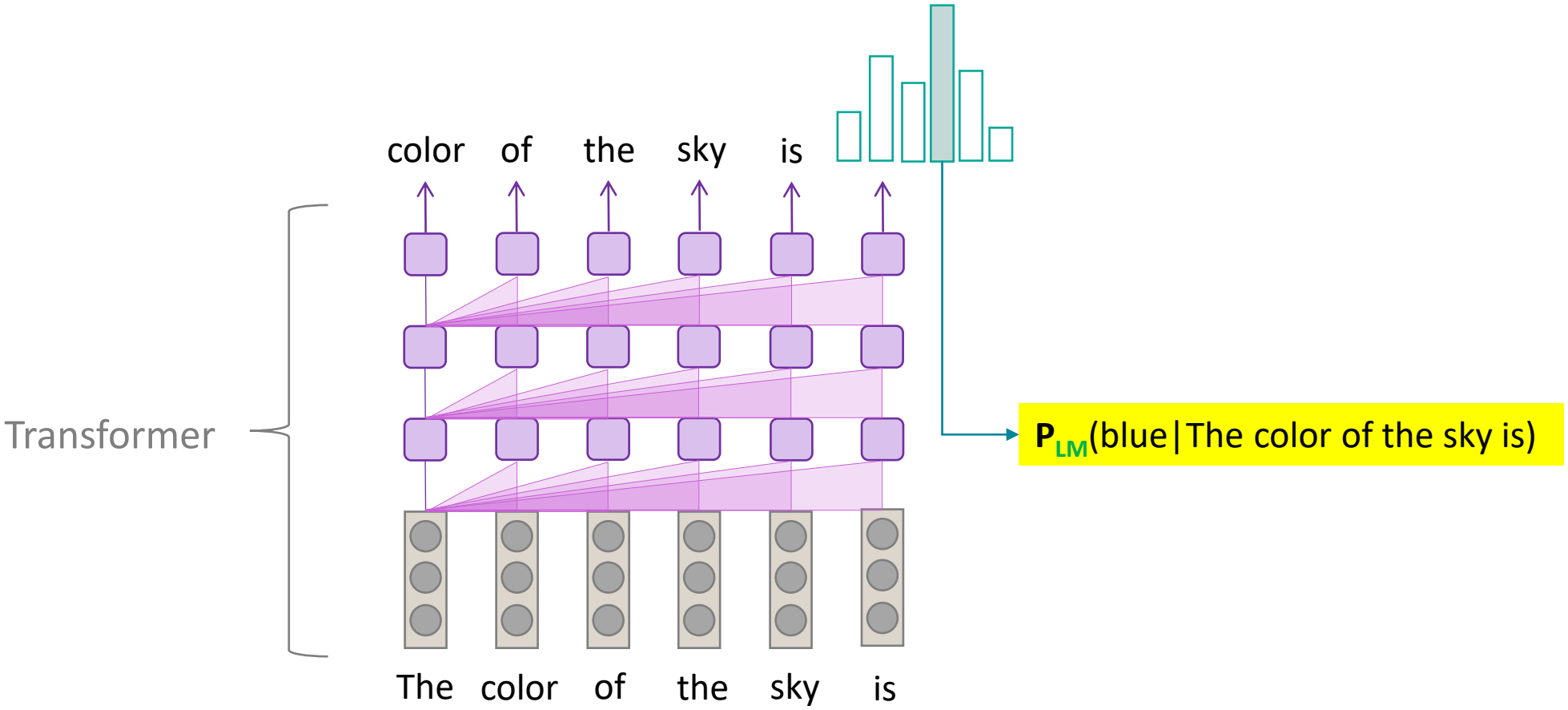
- **Definition**: a probability distribution P over sequences of word tokens $w_1, w_2, \ldots, w_T$.

The color of the sky is ___



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

Auto-regressive Language Models



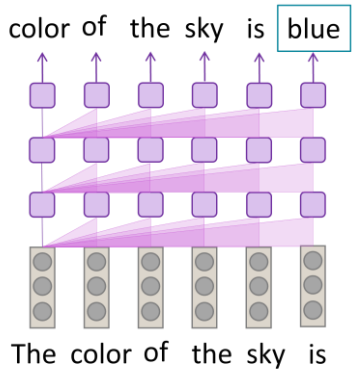
Base Language Models

Pretraining corpus

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

Train

Language Model



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

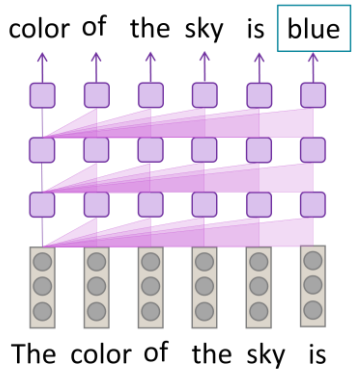
Base Language Models

Pretraining corpus

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

Train

Language Model



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

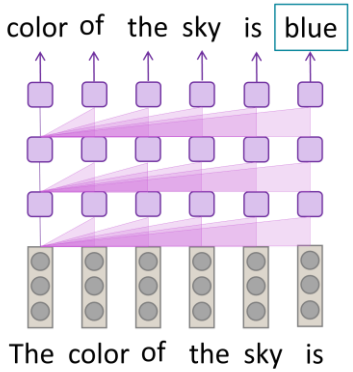
Zero-shot generalization

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

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

LLM distribution v.s. Data Distribution

Language Model



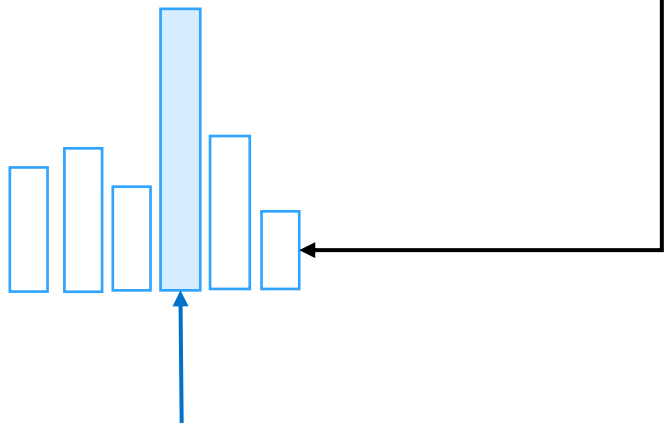
Last layer output

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

Pretraining corpus

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

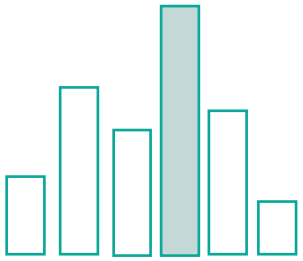
Frequency



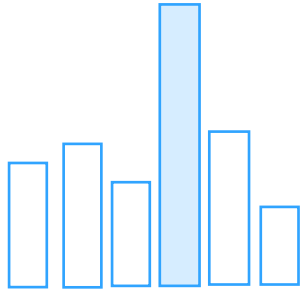
P_{Data}(blue|The color of the sky is)

Distributional Memorization

Def.



=



Language model distribution

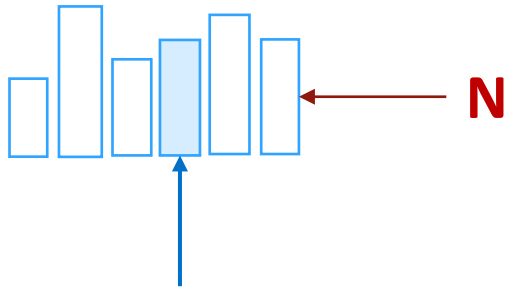
Pretraining data distribution

Memorize without understanding

Why Such Definition?

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

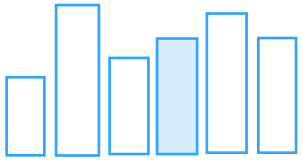
Rare prefix



Not predictive!

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

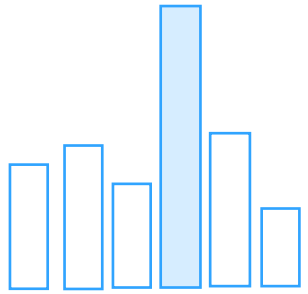
Prefix Decomposition



The color of the sky is the same as the **ocean**.

Low frequency in data.

=

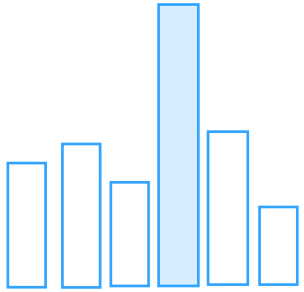


High frequency in data.

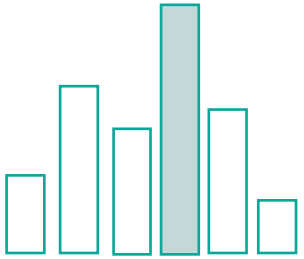
The color of the sky is blue.

+

The color of the ocean is blue.

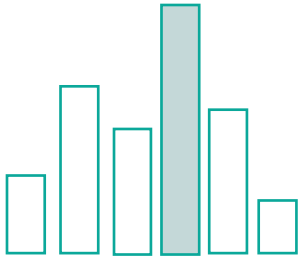


If LM Understands…



The color of the sky is the same as the **ocean**.

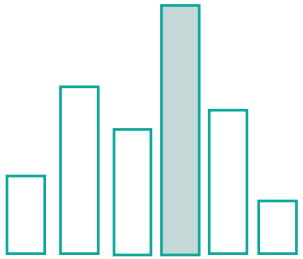
=



The color of the sky is blue.

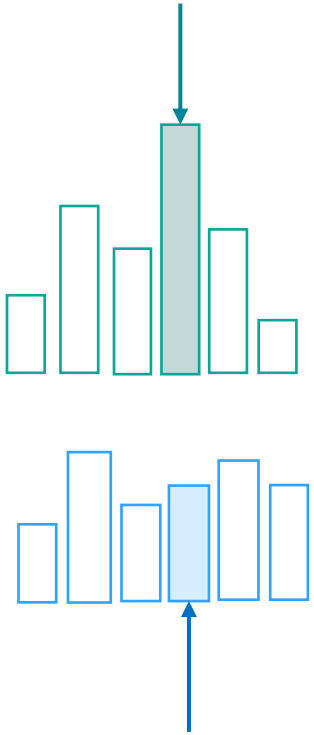
+

The color of the ocean is blue.



Distribution Difference

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



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

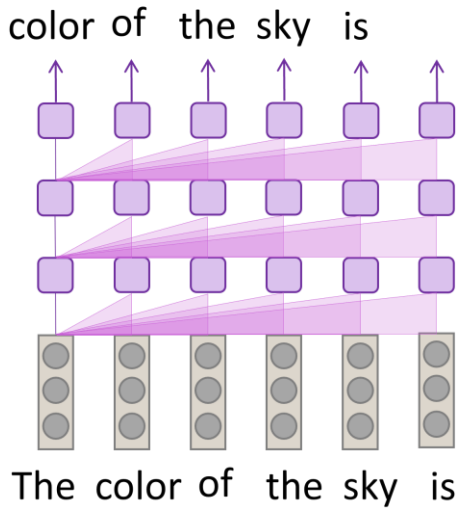
Experiment Settings

The PILE

(207 billion tokens)

Train

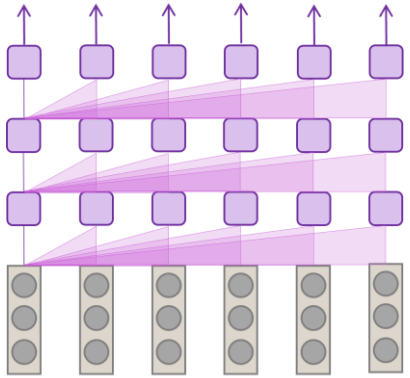
Pythia



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

Tomorrow I will fly to the conference in Canada



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

LLM v.s. Data Distribution

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

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

Pretraining Data Probability

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

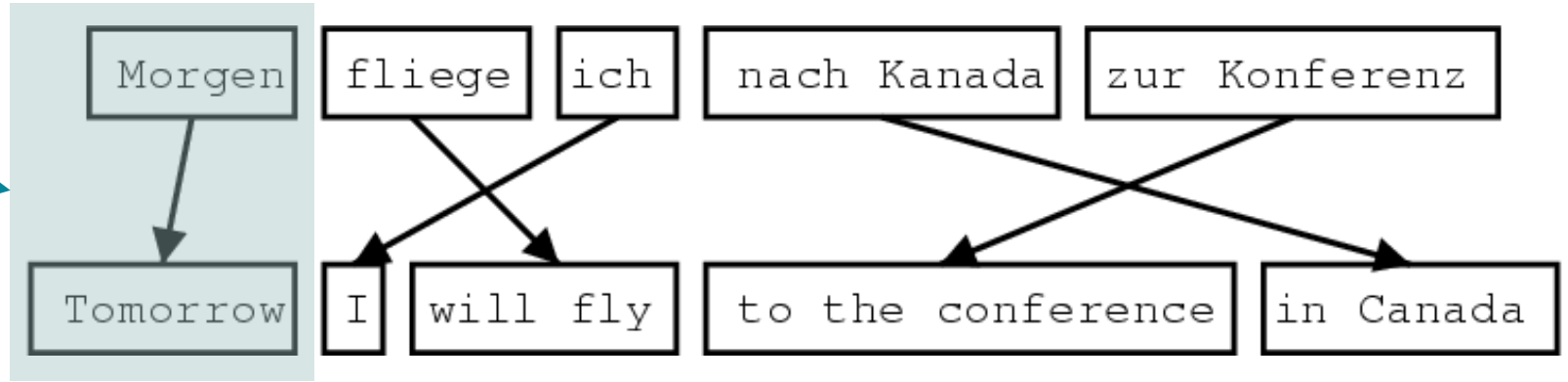
Directly search the whole sentence?

No match! Need simplification

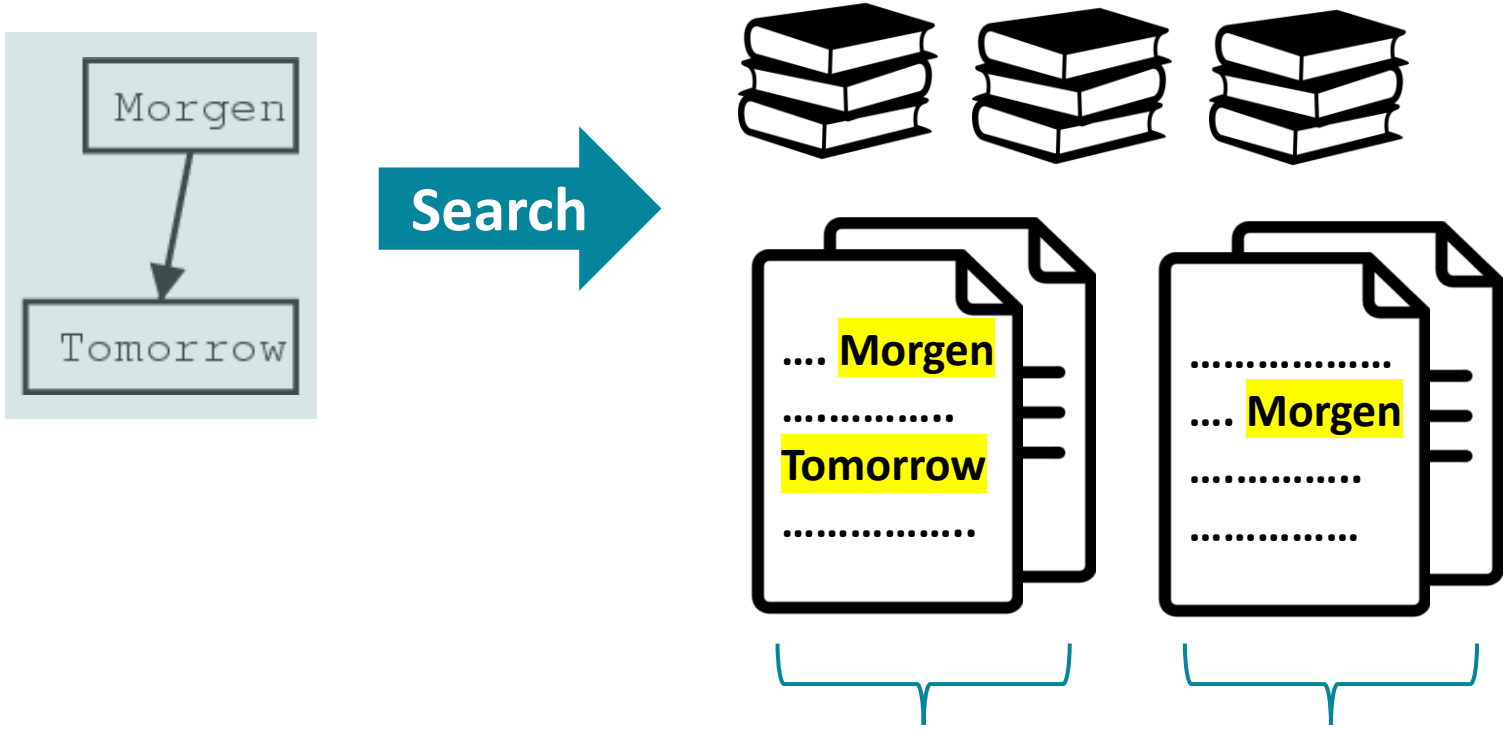
Xinyi Wang*, Antonis Antoniades*, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICLR 2025.

N-gram Simplification

Cosine similarity between
n-gram embeddings

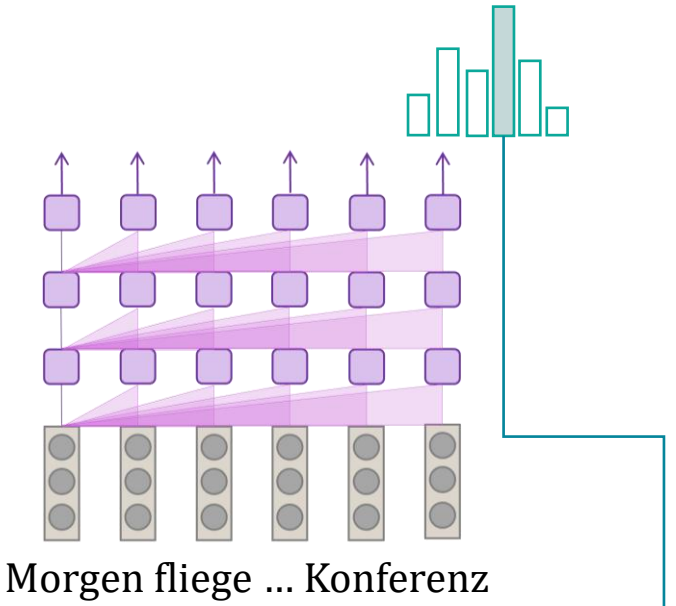


N-gram Data Probability



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

Compute Distributions



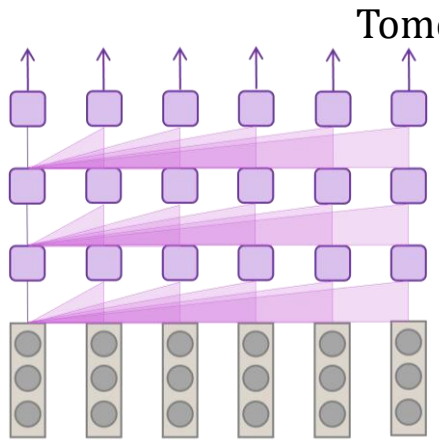
Morgen fliege … Konferenz

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

Search

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

Compare Distributions



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

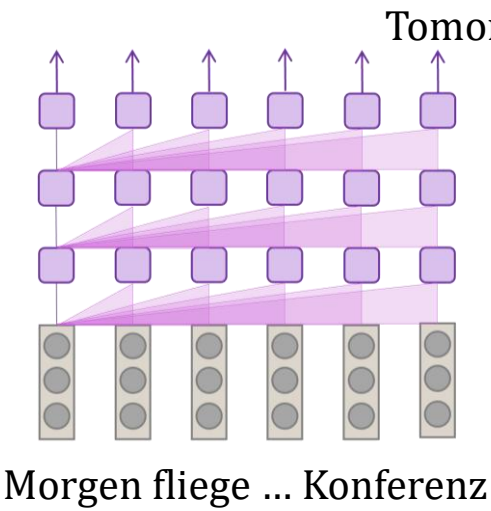
$$P_{data}(\text{Tomorrow}|\text{Morgen}) = \frac{C(\text{Tomorrow, Morgen})}{C(\text{Morgen})}$$

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

Def.



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

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

Memorization: Spearman correlation

(most compute efficient)

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: The quantum efficiency of a photon detector is 0.1. If 100 photons are sent into the detector, one after the other, the detector will detect photons

A. an average of 10 times, with an rms deviation of about 4
B. an average of 10 times, with an rms deviation of about 3
C. an average of 10 times, with an rms deviation of about 1
D. an average of 10 times, with an rms deviation of about 0.1

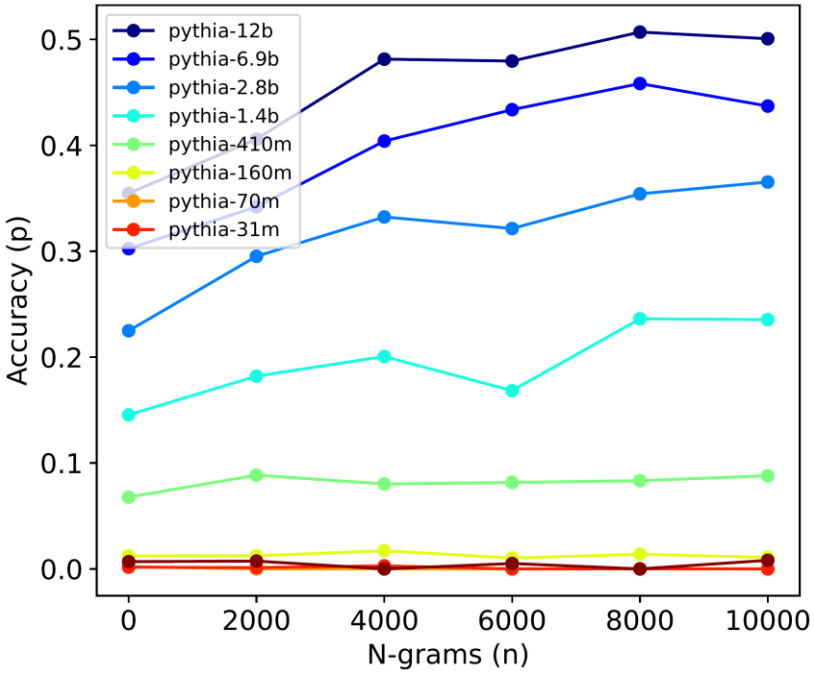
Answer: B

Task Performance

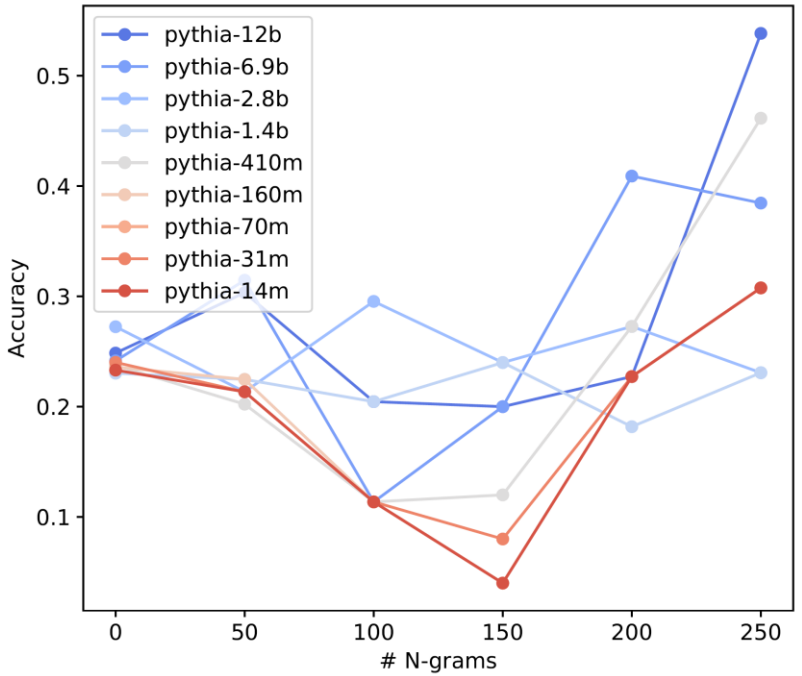
n-gram Frequency↑ Performance↑

Model size↑ Performance↑

TriviaQA



MMLU



Distributional Memorization

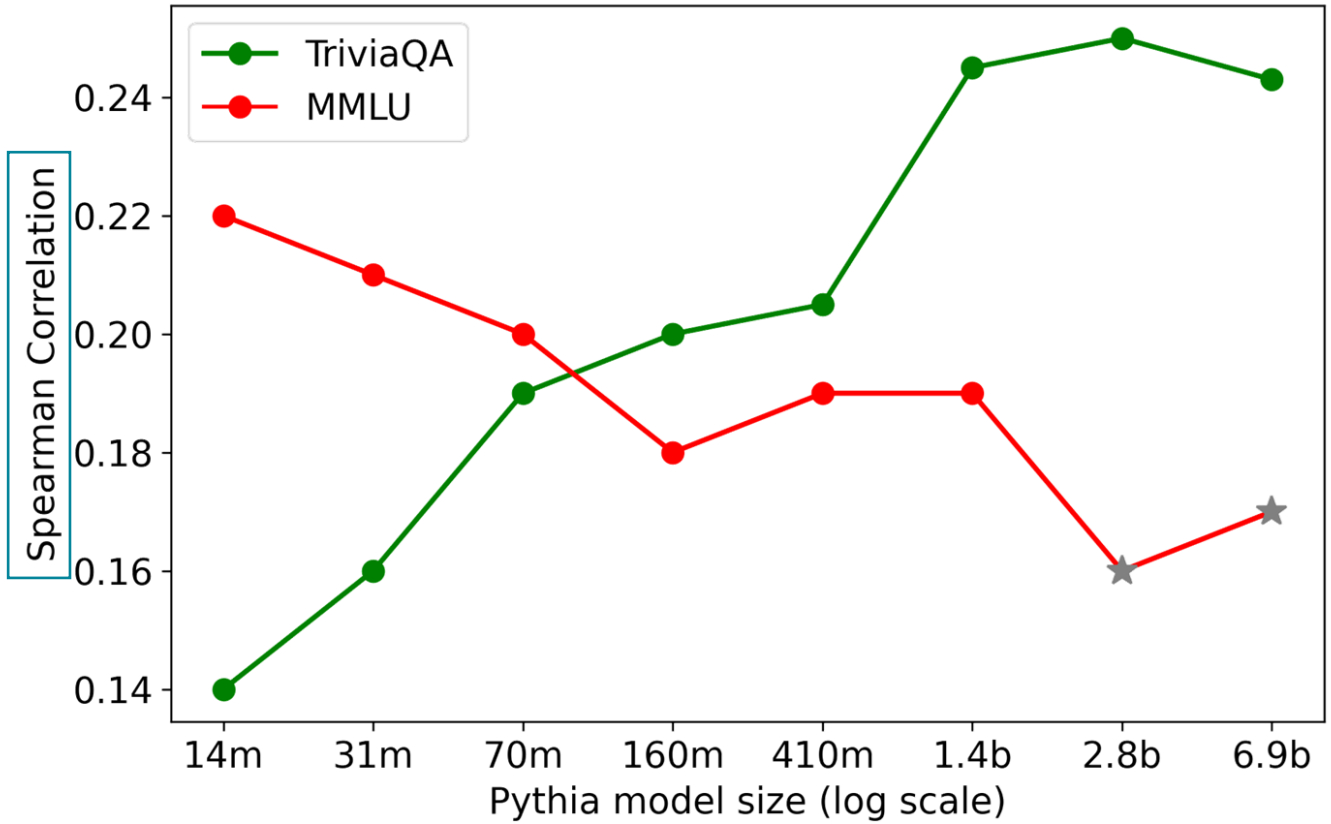
TriviaQA

Model size↑ Correlation↑

MMLU

Model size↑ Correlation↓

Correlation↑
Memorization↑
(according to definition)

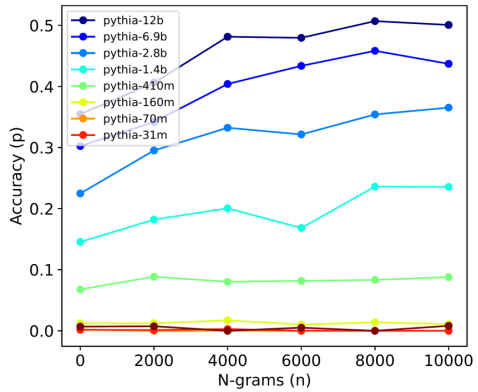


Memorization v.s. Performance

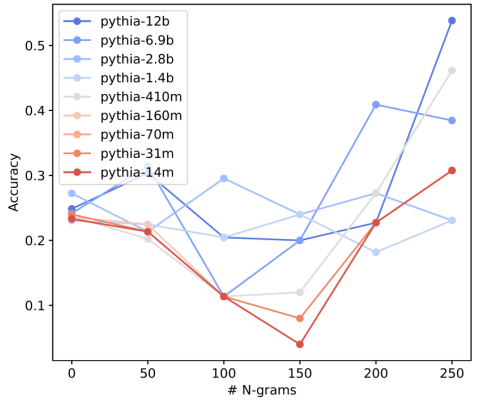
Depend on memorization → **TriviaQA**

MMLU ← **Depend on generalization**

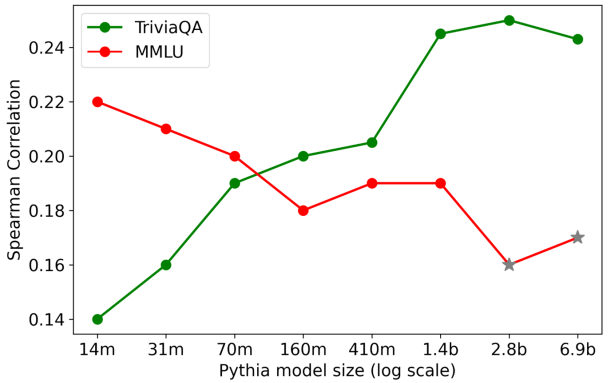
Model size↑ Performance↑



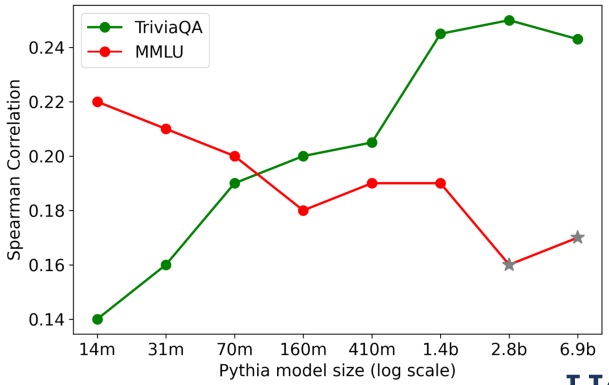
Model size↑ Performance↑



Model size↑ Correlation↑

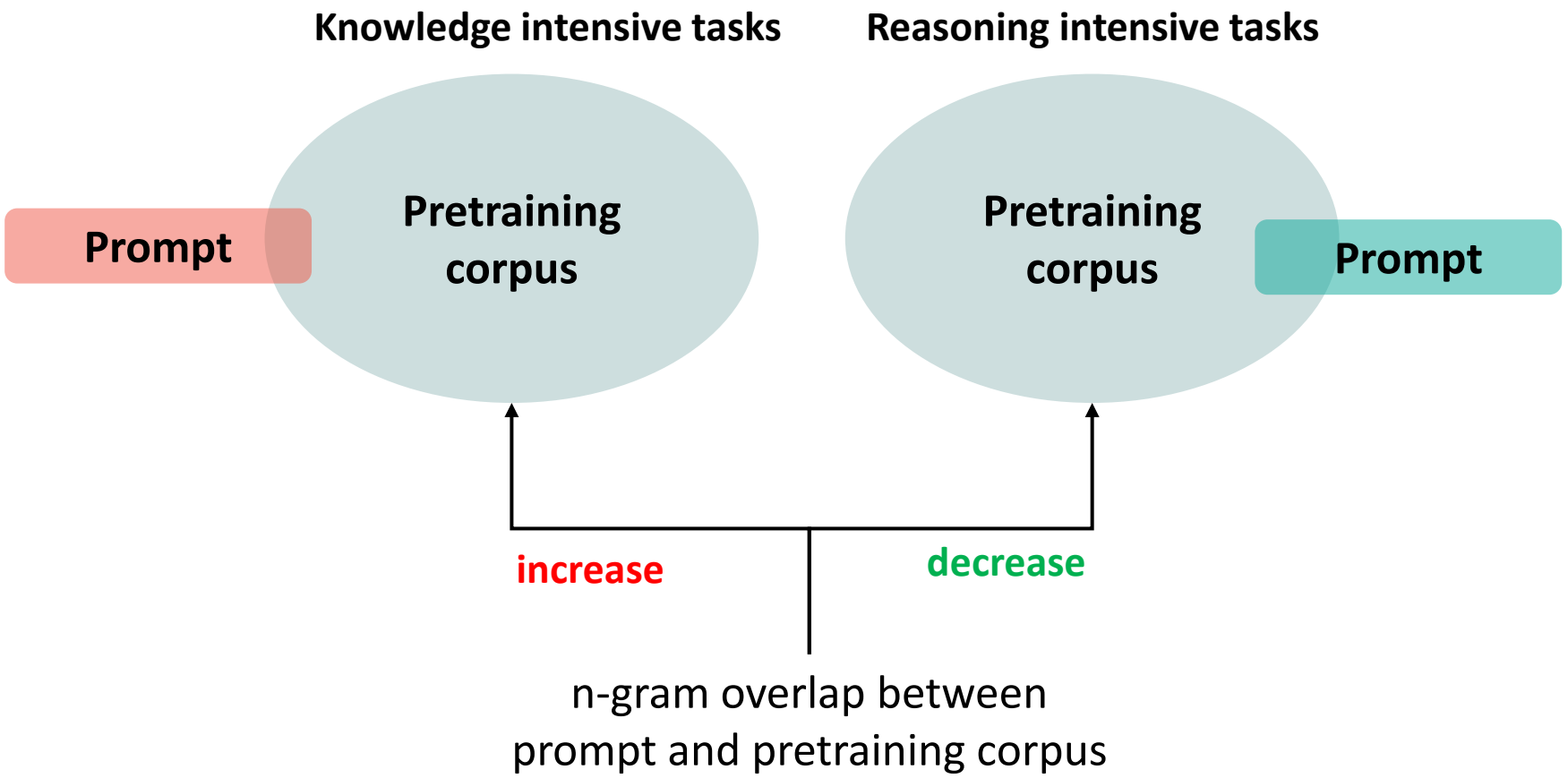


Model size↑ Correlation↓



Correlation↑
Memorization↑
(according to definition)

Rewrite the Prompt



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).

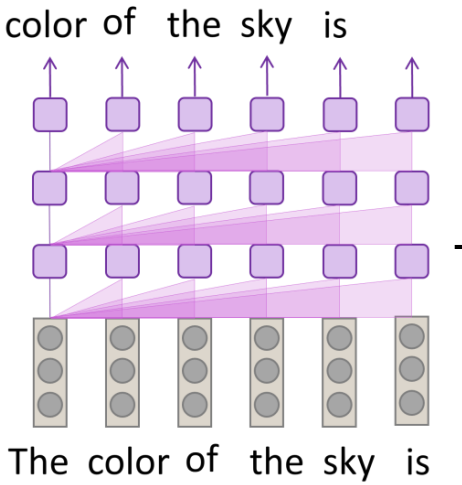
More complex generalization mechanism!

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 √

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.

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.

Recent Discussion on RL

SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training

Tianzhe Chu[♠*] Yuexiang Zhai[♥♣*] Jihan Yang[♦] Shengbang Tong[♦]
Saining Xie[♣♦] Dale Schuurmans[♣★] Quoc V. Le[♣] Sergey Levine[♥] Yi Ma[♠♥]

Abstract

Supervised fine-tuning (SFT) and reinforcement learning (RL) are widely used post-training techniques for foundation models. However, their respective role in enhancing model generalization in rule-based reasoning tasks remains unclear. This paper studies the comparative effect of SFT and RL on generalization and memorization, focusing on text-based and visual reasoning tasks. We introduce `GeneralPoints`, an arithmetic reasoning card game, and also consider `V-IRL`, a real-world navigation environment, to assess how models trained with SFT and RL generalize to unseen variants in both novel textual rules and visual domains. We show that RL, especially when trained with an outcome-based reward, generalizes in both the rule-based textual and visual environments. SFT, in contrast, tends to memorize the training data and struggles to generalize out-of-distribution in either scenario. Further analysis reveals that RL improves the model's underlying visual recognition capabilities, contributing to its enhanced generalization in visual domains. Despite RL's superior generalization, we show that SFT is still helpful for effective RL training: SFT stabilizes the model's output format, enabling subsequent RL to achieve its performance gains. These findings demonstrate the advantage of RL for acquiring generalizable knowledge in complex, multi-modal tasks.

1. Introduction

Although SFT and RL are both widely used for foundation model training (OpenAI, 2023b; Google, 2023; Jaech et al., 2024; DeepSeekAI et al., 2025), their distinct effects on *generalization* (Bousquet & Elisseeff, 2000; Zhang et al., 2021) remain unclear, making it challenging to build reliable and robust AI systems. A key challenge in analyzing the generalizability of foundation models (Bommasani et al., 2021; Brown et al., 2020) is to separate data memorization[1] from the acquisition of transferable principles. Thus, we investigate the key question whether SFT or RL primarily memorize training data (Allen-Zhu & Li, 2023a; Ye et al., 2024; Kang et al., 2024), or whether they learn generalizable rules that can adapt to novel task variants.

To address this question, we focus on two aspects of generalization: textual rule-based generalization and visual generalization. For textual rules, we study the ability of a model to apply learned rules (given text instructions) to variants of these rules (Zhu et al., 2023; Yao et al., 2024; Ye et al., 2024). For vision-language models (VLMs), visual generalization measures the consistency of performance with variations in visual input, such as color and spatial layout, within a given task. For studying text-based and visual generalization, we investigate two different tasks that embody rule-based and visual variants. Our first task is `GeneralPoints`, an original card game task similar to `Points24` of RL4VLM (Zhai et al., 2024a), which is designed to evaluate a model's *arithmetic reasoning capabilities*. The model receives four cards (presented as a text description or an image), and is required to compute a target number (24 by default) using each card's numerical value exactly once. Second, we adopt `V-IRL` (Yang

Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination

Mingqi Wu[1*], **Zhihao Zhang**[1 2*], **Qiaole Dong**[1*],
Zhiheng Xi[1], **Jun Zhao**[1], **Senjie Jin**[1], **Xiaoran Fan**[1], **Yuhao Zhou**[1],
Yanwei Fu[1], **Qin Liu**[3], **Songyang Zhang**[2], **Qi Zhang**[1,2†]
[1] Fudan University
[2] Shanghai Artificial Intelligence Laboratory
[3] University of California, Davis
{qz}@fudan.edu.cn {qinli}@ucdavis.edu

Abstract

The reasoning capabilities of large language models (LLMs) have been a long-standing focus of research within the community. Recent works have further enhanced these capabilities by reinforcement learning (RL), and numerous novel

Pretraining data still have great implication on post-training…

500, AMC, and AIME, while often failing to yield comparable gains on other model families such as Llama, which warrants more in-depth investigation. In this work, our empirical analysis shows that, despite the Qwen 2.5 series attaining superior mathematical-reasoning performance relative to other models, its pre-training on massive web-scale corpora leaves it vulnerable to data contamination in widely used benchmarks (e.g., MATH-500). Consequently, experimental results derived from contaminated benchmarks on the Qwen2.5 series may be unreliable. To obtain trustworthy evaluation signals, we introduce a generator that creates fully synthetic arithmetic problems of arbitrary length and difficulty, yielding clean datasets we call *RandomCalculation*. Using these leakage-free datasets, we further show that under the RL protocol, only accurate reward signals yield steady improvements that surpass the model's performance ceiling in mathematical reasoning, whereas noisy or incorrect rewards do not. Thus, we recommend that future studies evaluate on uncontaminated benchmarks and, when feasible, test various model series to ensure trustworthy conclusions about RL and related methods.

Thank you!

Questions?

UC SANTA BARBARA