# On Memorization of Large Language Models in Logical Reasoning

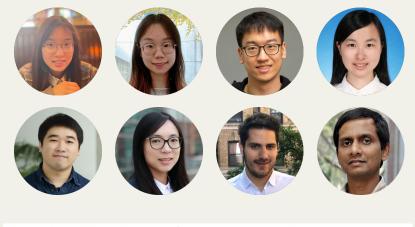
Chiyuan Zhang

Google Research



<u>https://memkklogic.</u> github.io

Chiyuan Zhang (Google Research)





# Contents

- My biased view onShifting Perspective of Generalization
- 01 Motivation
- 02 LiMem Score and Knight & Knave Puzzles
- 03 Measurement of Memorization
- 04 Memorization vs Reasoning
- 05 Conclusion & Discussion
- 06 Related Work on LLM Memorization

# Shifting Perspective of Generalization

CLASSICAL ML

Test Acc = Train Acc + Gen Gap

Perspective: Bound Gen Gap for ERM

**Best Practice:** Model selection, under-parameterization or strong regularization

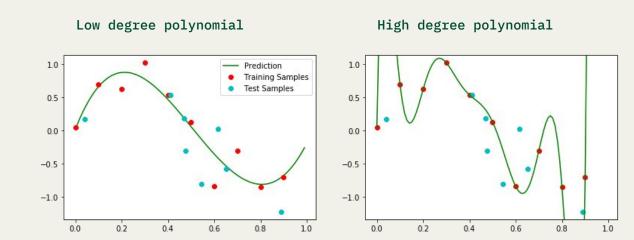


Fig from https://github.com/jamt9000/prml/blob/master/1.1-polycurve.ipynb

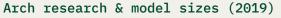
# Shifting Perspective of Generalization

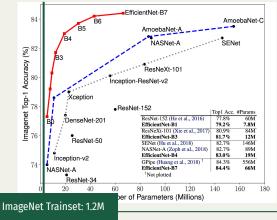
PRE-LLM REGIME

#### **Overparameterized Models Work Well**

**Perspective:** Non-unique ERM, (implicit) regularization from algor, architecture, data impact generalization.

**Best Practice:** Nonconvexity, Large mode/data, New NN Architecture, Optimizer, Data Augmentation, Initialization, Loss, etc...





#### Generalization Research

Improved Generalization Bound Implicit Regularization Benign Overfitting Double Descent Geometry of Loss Landscape Out of Distribution Generalization

# Shifting Perspective of Generalization

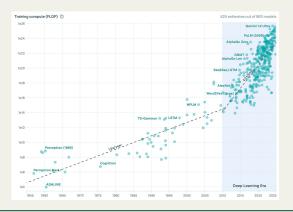
#### LLM REGIME

### Underparameterization + 1-epoch training + Emergent Abilities

**Additional challenges:** Difference between small / large models; Cost of analyzing large models

**Best Practice:** More data, Larger model (scaling law), Longer context window, More inference compute, ...

### Growth of Training Costs



#### Example Problem from MATH dataset

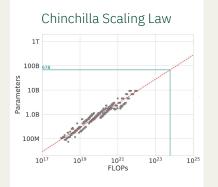
Let f(x) be the function defined on  $-1 \le x \le 1$  by the formula  $f(x) = 1 - \sqrt{1 - x^2}$ . This is a graph of y = f(x):



If a graph of x = f(y) is overlaid on the graph above, then one fully enclosed region is formed by the two graphs. What is the area of that region, rounded to the nearest hundredth?

#### Fig source: https://epoch.ai/data/notable-ai-models

# Generalization in LLM Regime

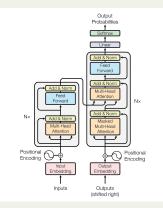


**Underparameterization + Undertraining** 

Optimal (under)parameterization under certain (training) **compute budget** 

Optimizing on the population loss, but testing on (different) **downstream tasks** 

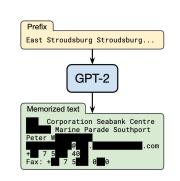
Fine-tuning can still be interpolating



#### Transformer + Attention

Attention mechanism & in-context learning.

Simulation / representation power of transformers; Chain-of-thought and inference time computation.



#### Memorization

**Practical:** Privacy + Copyright Concerns

Scientifical: Generalization / Memorization

Memorization vs Memory (of LLM)

### Two US lawyers fined for submitting fake court citations from ChatGPT

Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim



The judge said one of the fake decisions had 'some traits that are superficially consistent with actual judicial decisions' but other portions contained 'gibberish' and were 'nonsensical'. Photograph: Richard Drewk/AP

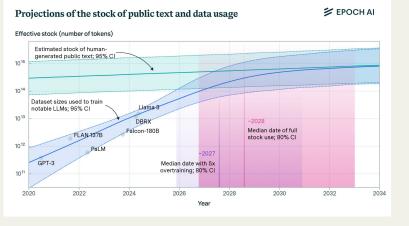
#### **Beyond Generalization**

Generalization =?=>

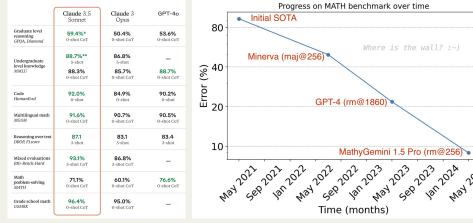
- > Hallucinations / Grounding / Factuality
- > Alignment / Instruction Following
- > Safety / Jailbreaking

Other questions: Interpretability, Scaling Law, Attribution, Tool use, Evaluation of complex tasks, ...

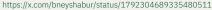
# Two Trends of Saturation



https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data



https://www.anthropic.com/news/claude-3-5-sonnet



We will run out of human generated public text for training LLMs "soon".

We are saturating many (originally) challenging evaluation benchmarks.

#### Is there a problem (beyond running out of train and test data)?

# Two Trends of Saturation

#### News, News In Brief

#### Geologists Recommend Eating At Least One Small Rock Per Day

● ● ● 🗉 ~ < > 😑 🔒 theonion.com 👒 C 🟥 + ≫

#### 000000

#### Published: April 13, 2021

BERKELEY, CA-Calling the average American diet "severely lacking\* in the proper amount of sediment, Geologists at UC Berkeley recommended Tuesday eating at least one small rock per day. "In order to live a healthy, balanced lifestyle, Americans should be ingesting at least a single serving of nebbles, geodes, or gravel with breakfast, lunch, or dinner," said Dr. Joseph Granger, adding that the rocks, which could range in size from a handful of dust to a medium-sized 5-nound cobblestone, were an important source of vitamins and minerals critical to digestive health. "Yes, we've all accidentally ingested sand or dust, but that is not even close to enough matter for a human adult hody to subsist on Because many do not like the taste or texture, we recommend hiding loose mcks inside different foods, like nearut butter or ice cream." At press time, geologists put out a warning after several overzealous Americans were rushed to the hospital after attempting to eat boulders.

≡ Valeur Multiplication CALCULATOR

Here we answer one simple question: What is 11 times 9? (or What is 11 multiplied by 9?) The answer is as follows:

● ● • • < > 😑 🔒 valeur.com 🕫 C 🗄 + >>

#### 11 x 9 = 99

Here are some other ways to display or communicate that 11 times 9 equals 99:

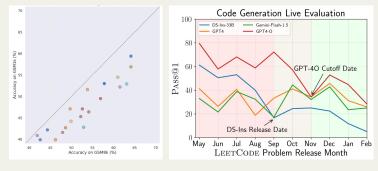
#### 11 \* 9 = 99 $11 \cdot 9 = 99$ 11(9) = 99

To explain what 11 times 9 means, look at it as 11 added together 9 times. To get the answer, you could just write down the number 11, 9 times and <sup>+</sup>hen add the 11 numbers together.

••• • • • • 🚊 academia.edu 🛛 🖓 🖒 ሰ + © oblem (when the lexicon does not include the context positional meaning is semantically anomalous) and t mpositional meaning is congruous but it is not what n she gives a detailed analysis of how a constructivist of the following: A thoughtful wife has pork chops re rom fishing. Upon hearing this sentence, the reader beg Then he shifts frame to the fishing event (among othe g fish, cooking fish). After integrating the two structure d with "no fish" and the going-home slot with "empty ha nstructivist therefore, there is no need to evoke a conte he meaning construction process. Semantic indeterminac constructivist account because "words do not retrieve l ract structures and processes for integration with (46).

nows frame-shifting in on-line jokes, text processing an camples she provides a detailed explanation of what trictively.

of the chapter she comes to the conclusion that "frar a violation of slot-filling constraints" (58) or "wher that predicted by the contextually evoked frame" (75



Source: arxiv:2405.00332 (GSM1k)

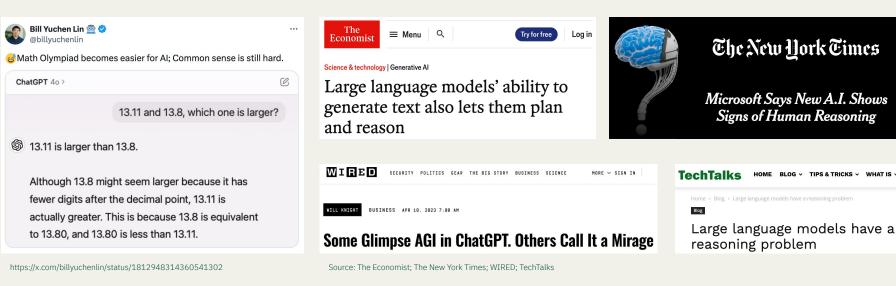
Source: arxiv:2403.07974 (LiveCodeBench)

It is really hard to know what an Internet crawled dataset contains.

Some evidence of data contamination has been observed.

#### Is there a problem (beyond running out of train and test data)?

# LLM Reasoning Debate

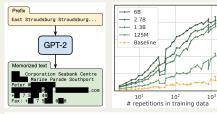


#### **Goal:** a quantitative approach to

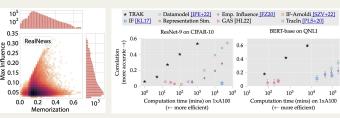
- Measure the amount of memorization
- Measure model performance in the possible presence of data contamination

WHAT IS V

### Previous Work on LLM Memorization







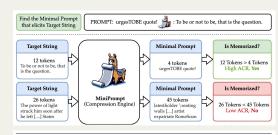
#### Query: first\_president

**Prompt**: The first President of the United States was

Completion: George Washington.

Influential Sequence for 52 Billion Parameter Model

President George Washington proclaimed Thursday, November 26, 1789 to be "a day of public thanksgiving and prayer". He proclaimed a second Thanksgiving Day on Thursday, February 19, 1795. And they make an argument about America's responsibilities. The United States has gotten bigger in the years since George Washington's 1789 Thanksgiving proclamation, both literally and in the role. In America's first Thanksgiving Proclamation in 1789 George Washington expressed thanks for "the peaceable and rational manner" in which our Constitution had been established just two years earlier



Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main pro- tagonist in J.K. Rowling's series of fantasy novels	Harry Potter is a British actor, writer, and director
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series	a talking cat and a dragon. One day, they decide
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were al- ready sitting at their desk, looking worried. "What's wrong?"	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting	to the park to play some bas- ketball.

#### Verbatim Memorization

https://arxiv.org/abs/2012.07805 https://arxiv.org/abs/2202.07646 https://arxiv.org/abs/2210.17546

#### Counterfactual-Mem / Attribution / Influence Fn

https://arxiv.org/abs/2112.12938 https://arxiv.org/abs/2308.03296 https://arxiv.org/abs/2104.06683

#### Other Notions of LLM Mem

https://arxiv.org/abs/2404.15146 https://arxiv.org/abs/2310.02238

# Memorization in Reasoning



#### DEFINITION

A memorization score inspired by human behaviors

#### MEASUREMENT

A dynamic Knight and Knave puzzle set to facilitate measurement of memorization in logical reasoning

#### MEMORIZATION VS REASONING

Are LLMs memorizing logical reasoning tasks? Does memorization prevent LLMs from learning to reason?

Human Subject Preparing for a Coding Interview / Math Exam / ...

Go through a lot of LeetCode questions / problem sets.

- > **Reasoner:** Figure out the underlying principles.
- > Memorizer: Remember many problems and answers.



Human Interviewer / Professor Training to Get a Fair Evaluation

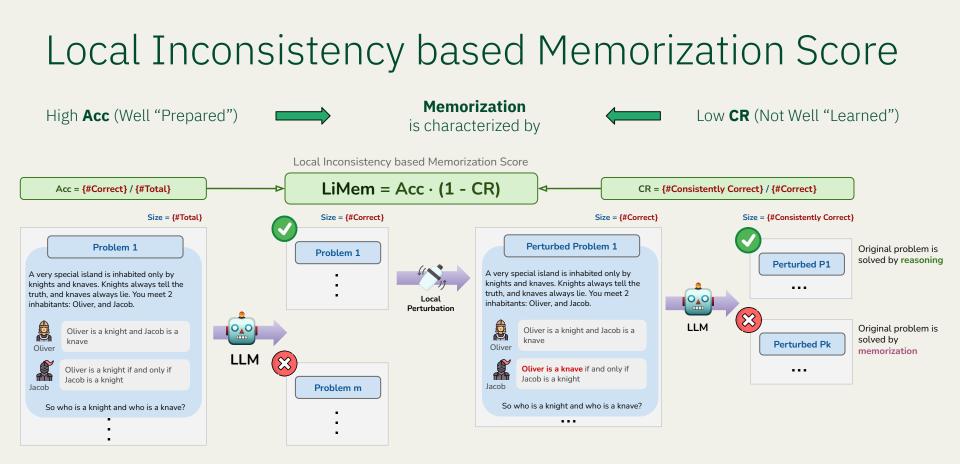
**Constraint:** It is unknown whether a test set example has been leaked.

Evaluation with a test question

Evaluation with a modified\* version of the same test question

\* The modification should be non-trivial but still use the same underlying principle for solution.





To facilitate our memorization study, we propose a **dynamic** logical reasoning benchmark that supports **automatic problem perturbations** beyond superficial language level paraphrasing.

Knights and Knaves (K&K) (Johnson-Laird & Byrne, 1990) is a type of logical puzzle where some characters tell truth, and others only lie. The goal is to infer each character's truthfulness.

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver, and Jacob.



Oliver is a knight and Jacob is a knave



Oliver is a knight if and only if Jacob is a knight

So who is a knight and who is a knave?

i-th person's statement

$$(\underline{B_1} \Leftrightarrow S_1) \land (\underline{B_2} \Leftrightarrow S_2)$$

Boolean value: whether the i-th person is telling the truth

### **Boolean Satisfiability**

i-th person is telling the truth if and only if their statement is true.

A valid solution to a K&K puzzle is a Boolean assignment for B1, B2 such that the overall formula is true.

To facilitate our memorization study, we propose a **dynamic** logical reasoning benchmark that supports **automatic problem perturbations** beyond superficial language level paraphrasing.

Knights and Knaves (K&K) (Johnson-Laird & Byrne, 1990) is a type of logical puzzle where some characters tell truth, and others only lie. The goal is to infer each character's truthfulness.

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver, and Jacob.



Oliver is a knight and Jacob is a knave



Oliver is a knight if and only if Jacob is a knight

So who is a knight and who is a knave?

i-th person's statement

$$(\underline{B_1} \Leftrightarrow S_1) \land (\underline{B_2} \Leftrightarrow S_2)$$

Boolean value: whether the i-th person is telling the truth

### Why K&K Puzzles?

Boolean satisfiability problem (SAT) is the first problem that was proven to be NP-complete.

Often used to test humans' logical reasoning abilities in exams such as in quant exams / the Law School Admission Test.

### reasoning process from off-the-shelf model

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver, and Jacob.



Oliver is a knight and Jacob is a knave



Oliver is a knight if and only if Jacob is a knight

#### So who is a knight and who is a knave?

GPT-40 Poe ···

To solve this, let's analyze the statements:

1. Oliver's Statement: "Oliver is a knight and Jacob is a knave."

2. Jacob's Statement: "Oliver is a knight if and only if Jacob is a knight."

Let's consider each scenario:

#### Scenario 1: Oliver is a Knight

### If B1=True, S1 must be True, check B2, S2

- If Oliver is a knight, his statement is true. So, Oliver is a knight and Jacob is a knave.
- Jacob's statement would be false because if Oliver is a knight, Jacob cannot be a knight. Thus, Jacob's statement being false aligns with him being a knave.

This scenario is consistent.

#### Scenario 2: Oliver is a Knave

### If B1=False, S1 must be False, check B2, S2

- If Oliver is a knave, his statement is false. So, it is not true that Oliver is a knight and Jacob is a knave. This implies that either Oliver is not a knight or Jacob is not a knave (or both).
- Jacob's statement would be true if Oliver is a knave, which contradicts the nature of a knave who always lies.

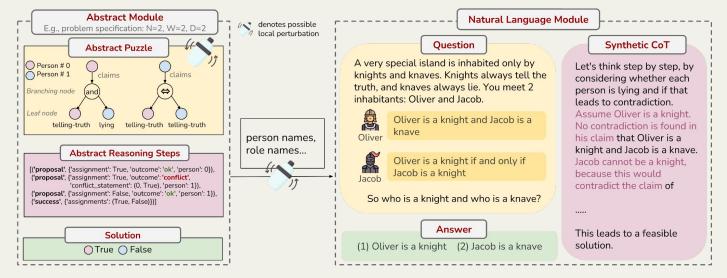
#### This scenario is inconsistent.

Therefore, the consistent scenario is that Oliver is a knight and Jacob is a knave.

### S Generating new puzzles with detailed reasoning steps and solutions.

- Problem difficulty: N-people puzzle, statement depth D, statement width W.
- Support logical statement types: and, or, not, imply, and equivalence.
- Synthetic Chain-of-Thought generation.

> Perturbing a given puzzle locally and recompute the new reasoning steps and solution.

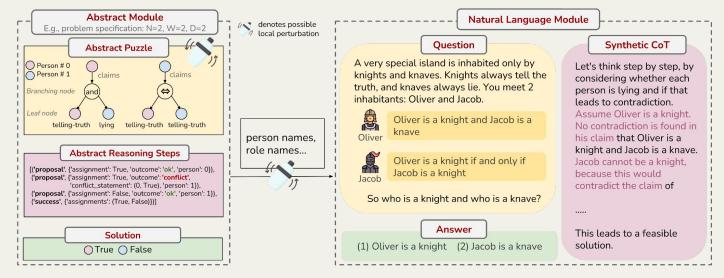


Chiyuan Zhang (Google Research)

S Generating new puzzles with detailed reasoning steps and solutions.

> Perturbing a given puzzle locally and recompute the new reasoning steps and solution.

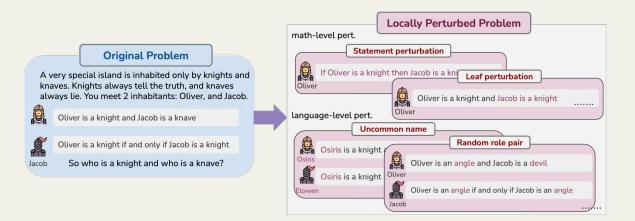
- **Math-level**: replace an **entire statement** or a **leaf node** in a statement with a newly sampled one.
- **Language-level**: changing person names, pairs of role names, statements reorder, and role flipping (e.g., knight/knaves → knaves/knight).



S Generating new puzzles with detailed reasoning steps and solutions.

> Perturbing a given puzzle locally and recompute the new reasoning steps and solution.

- Math-level: replace an entire statement or a leaf node in a statement with a newly sampled one.
- **Language-level**: changing person names, pairs of role names, statements reorder, and role flipping (e.g., knight/knaves → knaves/knight).



Re-verification of the existence and uniqueness of solutions; Re-sampling if no or multiple solutions exists.

### Evaluation Off-the-Shelf Models

		Aco	c(f;T	st)				
0.30	0.16	0.09	0.06	0.05	0.02	0.05		
0.28	0.11	0.04	0.02	0.04	0.00	0.00		
0.36	0.25	0.15	0.12	0.03	0.07	0.04		
0.44	0.34	0.16	0.14	0.04	0.07	0.03		
0.28	0.13	0.12	0.05	0.01	0.00	0.00		
0.35	0.21	0.08	0.06	0.02	0.00	0.00		
0.41	0.22	0.07	0.06	0.02	0.03	0.01		
0.49	0.36	0.20	0.10	0.05	0.02	0.01		
0.54	0.36	0.25	0.14	0.04	0.03	0.04		
0.63	0.42	0.34	0.17	0.09	0.10	0.01		
0.68	0.57	0.49	0.32	0.23	0.21	0.11		
0.70	0.63	0.51	0.31	0.22	0.10	0.06		
0.66	0.48	0.28	0.14	0.04	0.02	0.02		
0.79	0.61	0.42	0.22	0.07	0.00	0.00		
2	3	4	5 # nn	6	7	8		
	0.28 0.36 0.44 0.28 0.35 0.41 0.49 0.54 0.63 0.68 0.70 0.66 0.79	0.28         0.11           0.36         0.25           0.44         0.34           0.28         0.13           0.35         0.21           0.41         0.22           0.49         0.36           0.54         0.36           0.54         0.36           0.63         0.42           0.68         0.57           0.70         0.63           0.66         0.48           0.79         0.61	0.30         0.16         0.09           0.28         0.11         0.04           0.36         0.25         0.15           0.44         0.34         0.16           0.28         0.13         0.12           0.35         0.21         0.08           0.41         0.22         0.07           0.49         0.36         0.20           0.54         0.36         0.25           0.63         0.42         0.34           0.68         0.57         0.49           0.70         0.63         0.51           0.66         0.48         0.28           0.79         0.61         0.42	0.30         0.16         0.09         0.06           0.28         0.11         0.04         0.02           0.36         0.25         0.15         0.12           0.44         0.34         0.16         0.14           0.28         0.13         0.12         0.05           0.35         0.21         0.08         0.06           0.41         0.22         0.07         0.06           0.49         0.36         0.20         0.10           0.54         0.36         0.25         0.14           0.63         0.42         0.34         0.17           0.68         0.57         0.49         0.32           0.70         0.63         0.51         0.31           0.66         0.48         0.28         0.14           0.79         0.61         0.42         0.22           2         3         4         5	0.28         0.11         0.04         0.02         0.04           0.36         0.25         0.15         0.12         0.03           0.44         0.34         0.16         0.14         0.04           0.28         0.13         0.12         0.05         0.01           0.35         0.21         0.08         0.06         0.02           0.41         0.22         0.07         0.06         0.02           0.49         0.36         0.25         0.14         0.04           0.63         0.42         0.34         0.17         0.09           0.68         0.57         0.49         0.32         0.23           0.70         0.63         0.51         0.31         0.22           0.66         0.48         0.28         0.14         0.04           0.70         0.63         0.51         0.31         0.22	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	0.30         0.16         0.09         0.06         0.05         0.02         0.05           0.28         0.11         0.04         0.02         0.04         0.00         0.00           0.36         0.25         0.15         0.12         0.03         0.07         0.04           0.44         0.34         0.16         0.14         0.04         0.07         0.03           0.28         0.13         0.12         0.05         0.01         0.00         0.00           0.35         0.21         0.08         0.06         0.02         0.00         0.00           0.41         0.22         0.07         0.06         0.02         0.03         0.01           0.49         0.36         0.20         0.10         0.05         0.02         0.01           0.49         0.36         0.25         0.14         0.04         0.03         0.04           0.63         0.42         0.34         0.17         0.09         0.10         0.01           0.64         0.42         0.34         0.17         0.09         0.10         0.01           0.66         0.48         0.28         0.14         0.04         0.02	0.30         0.16         0.09         0.06         0.05         0.02         0.05           0.28         0.11         0.04         0.02         0.04         0.00         0.00           0.36         0.25         0.15         0.12         0.03         0.07         0.04           0.44         0.34         0.16         0.14         0.04         0.07         0.03           0.28         0.13         0.12         0.05         0.01         0.00         0.00           0.44         0.34         0.16         0.14         0.04         0.07         0.03           0.28         0.13         0.12         0.05         0.01         0.00         0.00           0.35         0.21         0.08         0.06         0.02         0.03         0.01           0.41         0.22         0.07         0.06         0.02         0.03         0.01           0.49         0.36         0.20         0.10         0.05         0.02         0.01           0.54         0.36         0.25         0.14         0.04         0.03         0.04           0.68         0.57         0.49         0.32         0.23         0.21

- K&K benchmark poses a challenging logical reasoning task for all off-the-shelf models.
- Accuracy on 2-ppl task can be high.

### Memorization in Off-the-Shelf Models

Acc(f;Tst)												${\sf LiMem}(f;{\sf Tst})$ perturbed statement									LiMem(f;Tst) perturbed leaf								
Gemma-2-9b	0.30	0.16	0.09	0.06	0.05	0.02	0.05			0.28	0.16	0.09	0.06	0.04	0.02	0.04		0.30	0.16	0.09	0.05	0.04	0.02	0.04		0.30			
Llama-3-8B	0.28	0.11	0.04	0.02	0.04	0.00	0.00			0.27	0.10	0.04	0.02	0.04	0.00	0.00	0.30	0.26	0.11	0.03	0.02	0.04	0.00	0.00					
Phi-3-mini-instruct	0.36	0.25	0.15	0.12	0.03	0.07	0.04			0.22	0.21	0.13	0.09	0.03	0.06	0.03		0.24	0.24	0.13	0.12	0.03	0.06	0.04		0.25			
Phi-3-medium-instruct	0.44	0.34	0.16	0.14	0.04	0.07	0.03		0.6	0.27	0.24	0.14	0.10	0.01	0.07	0.03	0.25	0.27	0.28	0.12	0.10	0.03	0.04	0.02		0.20			
NuminaMath-7B-CoT	0.28	0.13	0.12	0.05	0.01	0.00	0.00			0.16	0.13	0.11	0.05	0.01	0.00	0.00		0.23	0.12	0.10	0.05	0.01	0.00	0.00		0.90			
Deepseek-Math-7b-instruct	0.35	0.21	0.08	0.06	0.02	0.00	0.00			0.22	0.19	0.07	0.06	0.02	0.00	0.00	0.20	0.22		0.06	0.05	0.02	0.00	0.00		0.20			
Llama-3.1-8B-Instruct	0.41	0.22	0.07	0.06	0.02	0.03	0.01		0.4	0.31	0.20	0.06	0.06	0.02	0.02	0.01		0.30	0.19	0.07	0.06	0.02	0.03	0.00					
Qwen2.5-Math-7B-Instruct	0.49	0.36	0.20	0.10	0.05	0.02	0.01		0.4	0.24	0.22	0.15	0.09	0.04	0.02	0.01	0.15	0.26	0.24	0.17	0.09	0.05	0.01	0.01		0.15			
Qwen2-Math-7B-Instruct	0.54	0.36	0.25	0.14	0.04	0.03	0.04			0.25	0.22	0.15	0.14	0.03	0.03	0.04	0.10	0.29	0.20		0.11	0.04	0.03	0.04					
GPT-4o-mini	0.63	0.42	0.34	0.17	0.09	0.10	0.01			0.24	0.26	0.19	0.14	0.07	0.08	0.00	0.10	0.29	0.25	0.20	0.11	0.06	0.08	0.01		0.10			
GPT-4o	0.68	0.57	0.49	0.32	0.23	0.21	0.11		0.2	0.19	0.30	0.17	0.21	0.14	0.15	0.09	0.10	0.20	0.20	0.22	0.18	0.14	0.13	0.09					
Claude-3.5-sonnet	0.70	0.63	0.51	0.31	0.22	0.10	0.06			0.24	0.33	0.25	0.23	0.13	0.08	0.06	0.05	0.16	0.31	0.24		0.11	0.08	0.06		0.05			
Gemini-1.5-Flash-002	0.66	0.48	0.28	0.14	0.04	0.02	0.02			0.18	0.27	0.17	0.12	0.04	0.02	0.01	0.05	0.29	0.17	0.15	0.14	0.03	0.02	0.02		0.00			
Gemini-1.5-Pro-002	0.79	0.61	0.42	0.22	0.07	0.00	0.00		0.0	0.13	0.23	0.24	0.13	0.07	0.00	0.00	0.00	0.22	0.25	0.24		0.05	0.00	0.00		0.00			
	2	3	4	5	6	7	8		0.0	2	3	4	5	6	7	8	0.00	2	3	4	5	6	7	8		0.00			
# ppl								-			# pp			-		-	2		# pp	I		5							

- K&K benchmark poses a challenging logical reasoning task for all off-the-shelf models.
- Accuracy on 2-ppl task can be high.

• Off-the-shelf models are sensitive to locally perturbed K&K puzzles.

## Evaluation Off-the-Shelf Models

	Acc(f;Tst)														
Gemma-2-9b	0.30	0.16	0.09	0.06	0.05	0.02	0.05								
Llama-3-8B	0.28	0.11	0.04	0.02	0.04	0.00	0.00								
Phi-3-mini-instruct	0.36	0.25	0.15	0.12	0.03	0.07	0.04								
Phi-3-medium-instruct	0.44	0.34	0.16	0.14	0.04	0.07	0.03								
NuminaMath-7B-CoT	0.28	0.13	0.12	0.05	0.01	0.00	0.00								
Deepseek-Math-7b-instruct	0.35	0.21	0.08	0.06	0.02	0.00	0.00								
Llama-3.1-8B-Instruct	0.41	0.22	0.07	0.06	0.02	0.03	0.01								
Qwen2.5-Math-7B-Instruct	0.49	0.36	0.20	0.10	0.05	0.02	0.01								
Qwen2-Math-7B-Instruct	0.54	0.36	0.25	0.14	0.04	0.03	0.04								
GPT-4o-mini	0.63	0.42	0.34	0.17	0.09	0.10	0.01								
GPT-4o	0.68	0.57	0.49	0.32	0.23	0.21	0.11								
Claude-3.5-sonnet	0.70	0.63	0.51	0.31	0.22	0.10	0.06								
Gemini-1.5-Flash-002	0.66	0.48	0.28	0.14	0.04	0.02	0.02								
Gemini-1.5-Pro-002	0.79	0.61	0.42	0.22	0.07	0.00	0.00								
	2	3	4	5	6	7	8								
				# pp	1										

### Online resources about K&K

https://philosophy.hku.hk/think/logic/knights.php

	The Island of Knights and Knaves									
0.0	On the island of Knights and Knowe, wergone is either a Know or a Knight. Knights always fail the truth and Knowes always lik. You have on who is a brave and who is a knight. The kiznders have make some statements about each other - each statement should be taken independ	B /5 B - official thinking web ( v ⊥ + ≠ ↔ ⊥ 3 ) Home Tutorials - Resources About Search this site Search								
0.6	statement. Categories exch lakader as ether a langht or a linaxe. When you think you are done, click the 'solve' button to see if you have solved the put whether each lakader is a linaxe or a lenght.	Knights and knaves puzzles								
	assy tricky tricker	L								
0.4	Here is your plactic: Would wan mit a youp of 2 standows. Their name an Hervis and Ingrid. Ingrid anyo Kelenis is mit yous. Ingrid anyo Kelenis is mit yous.									
0.2	Solve h	Crights and knaves logic puzzles were made popular by the logician and mathematician Raymond Smullyan. See http://on.wiki.word.org/wiki.Knotts.and_Xnaves. Here we have a total of 352 puzzles which get progressively more difficult. They were generated by a computer program within by Zahahur frame. There are no answers provide, but your can always check fly your friends agree with you!								
	River Browning	Puzzle #1 out of 382								
0.0	New Puzzle	A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie.								
0.0	https://dmackippop1_github.ig/kpayog/	You meet two inhabitants: Zoey and Mel. Zoey tells you that Mel is a knave. Mel says, "Neither Zoey nor I are knaves." Can you determine who is a knight and who is a knave?								

- K&K benchmark poses a challenging logical reasoning task for all off-the-shelf models.
- Accuracy on 2-ppl task can be high.

	Dolma	The PILE	C4	Oscar	OpenWebText
"Alice is a knave"	13	6	2	1	0
"Alice is a knight"	23	8	6	1	0
"Bob is a knave"	11	8	0	1	0
"Bob is a knight"	53	9	22	5	0
"Charlie is a knave"	3	0	0	0	0
"Charlie is a knight"	10	1	2	0	0

#### The off-the-shelf models might also be trained on K&K related data

## Memorization in Fine-Tuned Models

### Question

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver and Jacob.



Oliver is a knight and Jacob is a knave



Oliver is a knight if and only if Jacob is a knight

So who is a knight and who is a knave?

### Answer

(1) Oliver is a knight(2) Jacob is a knave

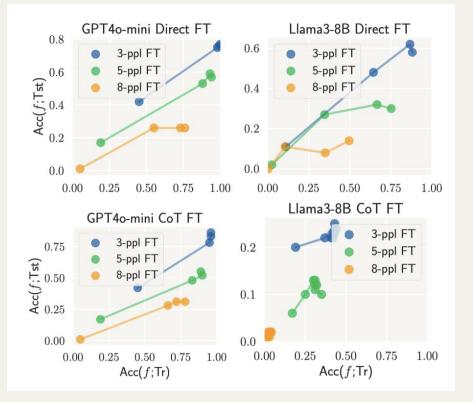
### Synthetic CoT

Let's think step by step, by considering whether each person is lying and if that leads to contradiction. Assume Oliver is a knight. No contradiction is found in his claim that Oliver is a knight and Jacob is a knave. Jacob cannot be a knight, because this would contradict the claim of ... This leads to a feasible solution.





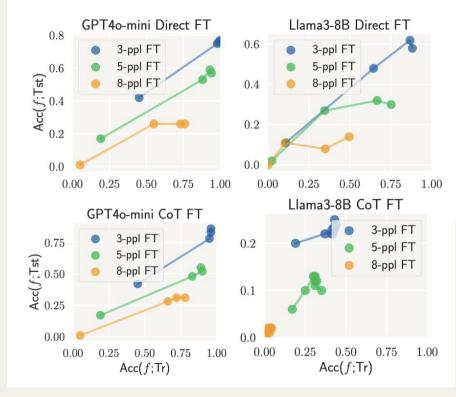
## Memorization in Fine-Tuned Models

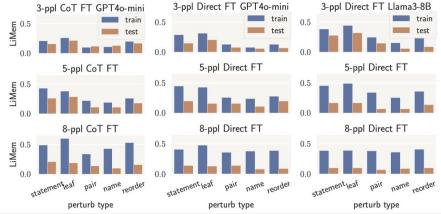


- Models achieve high training accuracy (approaching interpolation in many cases).
- CoT fine-tuning seem to be mostly helpful for more difficult cases.
- For model without enough capacity to fit CoT data, the performance is significantly worse than Direct Fine-tuning.

#train = 1000 for 3<=N<=8; 200 for N=2. #epochs = 100 for Llama3-8B; 5 for GPT4o-mini.

## Memorization in Fine-Tuned Models





- High memorization score on training puzzles
- Stronger memorization under math-level perturbations than language-level perturbations

Lower memorization score on test set

### What does this mean?

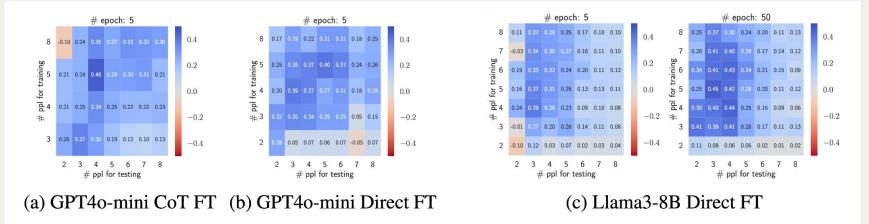
## Interpreting the LiMem Score



## Reasoning in Fine-Tuned Models?

### **Generalization across different difficulty levels**

Fine-tune on N-people problems, test on M-people problems 
MxN grid
Report test accuracy **improvement** of FTed LLMs compared to the un-FTed LLM



Performance improvements generalize across various difficulty levels, even when fine-tuned without detailed reasoning steps (Chain-of-Thoughts).

### Reasoning in Fine-Tuned Models?

### **Probing Intermediate States of Fine-tuned Models**

Use model embeddings to fit a linear classification task on constructed correct/incorrect statements about a given K&K puzzle

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver, and Ethan. Oliver told you that Oliver is a knight or Ethan is a knave. In a statement by Ethan: "Oliver is a knight". So who is a knight and who is a knave? **Oliver is a knight**.

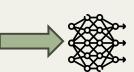
 $\Rightarrow$ 

embedding of i-th block



LLM

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Oliver, and Ethan. Oliver told you that Oliver is a knight or Ethan is a knave. In a statement by Ethan: "Oliver is a knight". So who is a knight and who is a knave? **Oliver is a knave**.



embedding of i-th block



### Reasoning in Fine-Tuned Models?

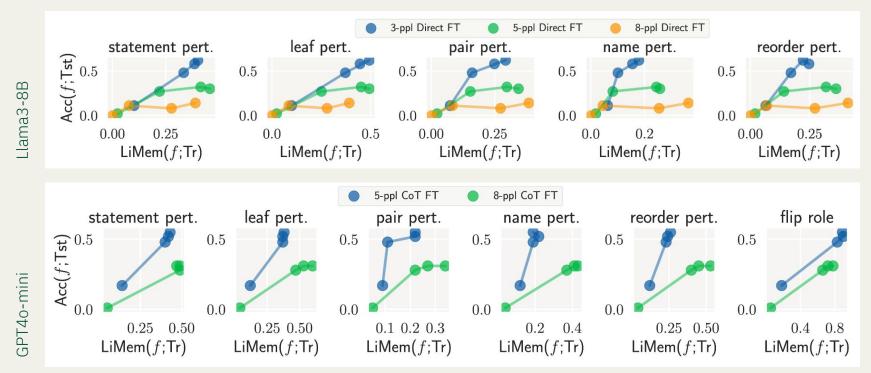
### **Probing Intermediate States of Fine-tuned Models**

valuation	2	0.57	0.58	0.6	0.64	0.67	0.74	0.75	0.76	0.8	0.8	0.82	0.84	0.86	0.88	0.87	0.86	0.86	0.83	0.85	0.83	0.82	0.82	0.79	0.8	0.81	0.81	0.81	0.82	0.84	0.85	0.81	0.75	1	1.00	
/alu	3	0.56	0.57	0.6	0.62	0.69	0.73	0.74	0.75	0.8	0.82	0.85	0.84	0.87	0.87	0.88	0.87	0.87	0.85	0.83	0.82	0.83	0.82	0.81	0.8	0.82	0.83	0.87	0.85	0.87	0.87	0.8	0.79	C	).95	
ē	4	0.57	0.58	0.58	0.63	0.64	0.68	0.69	0.69	0.74	0.75	0.77	0.79	0.83	0.86	0.83	0.81	0.82	0.81	0.78	0.76	0.76	0.76	0.77	0.74	0.75	0.76	0.78	0.81	0.82	0.81	0.78	0.73	Č		
K&K	5	0.57	0.59	0.62	0.67	0.69	0.72	0.72	0.73	0.78	0.8	0.79	0.84	0.84	0.87	0.87	0.83	0.86	0.82	0.83	0.8	0.77	0.78	0.76	0.76	0.79	0.79	0.8	0.83	0.77	0.82	0.73	0.72	C	0.90	Llama3-8B
the	6	0.58	0.58	0.6	0.62	0.66	0.66	0.71	0.74	0.78	0.8	0.84	0.85	0.89	0.88		0.88	0.86	0.82	0.81	0.8	0.8	0.79	0.77	0.81	0.78	0.81	0.78	0.79	0.8	0.8	0.75	0.72			
.⊑	7	0.58	0.59	0.6	0.62	0.65	0.69	0.7	0.69	0.74	0.76	0.78	0.79	0.82	0.82	0.8	0.81	0.81	0.77	0.78	0.76	0.76	0.76	0.74	0.72	0.75	0.77	0.78	0.76	0.78	0.75	0.76	0.68	C	).85	
⊧ ppl	8	0.62	0.62	0.63	0.65	0.67	0.7	0.72	0.73	0.75	0.75	0.75	0.8	0.79	0.81	0.81	0.79	0.79	0.77	0.76	0.77	0.76	0.76	0.75	0.75	0.76	0.78	0.8	0.75	0.8	0.74	0.75	0.67	C	0.80	
u #																																				
evaluation	2	0.57	0.59	0.61	0.65	0.68	0.75	0.77	0.78	0.85	0.87	0.9	0.91	0.94	0.97	0.97	0.97	0.96	0.95	0.94	0.92	0.91	0.92	0.9	0.89	0.9	0.9	0.91	0.92	0.94	0.95	0.91	0.89	1	1.00	
valu	3	0.58	0.59	0.61	0.66	0.71	0.74	0.76	0.78	0.84	0.87			0.95	0.97	0.97	0.96	0.95	0.94					0.89	0.88							0.88	0.88	C	).95	
× e	4	0.6	0.6	0.62	0.64	0.66	0.71	0.74	0.76	0.82	0.84	0.89		0.95	0.97	0.96	0.95	0.94				0.89		0.88	0.87	0.87	0.88	0.88				0.85	0.85			Direct FT-ed
K&K	5	0.6	0.61	0.64	0.67	0.7	0.74	0.76	0.78	0.83	0.85			0.95	0.97	0.96	0.96	0.94					0.88	0.87	0.86	0.87	0.89			0.89		0.82	0.81	C	0.90	
the	6	0.61	0.61	0.63	0.65	0.67	0.71	0.74	0.77	0.81	0.84	0.89		0.93	0.95	0.95	0.95	0.93			0.89	0.87	0.88	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.79	0.8			Llama3-8B
. <u></u>	7	0.6	0.61	0.62	0.66	0.69	0.72	0.75	0.77	0.81	0.83	0.86	0.88	0.92	0.95	0.95	0.94			0.88	0.87	0.87	0.86	0.86	0.85	0.85	0.87	0.87	0.86	0.88	0.86	0.8	0.76	C	).85	
hpd #	8	0.65	0.65	0.66	0.68	0.7	0.72	0.74	0.77	0.8	0.83	0.87	0.89	0.93	0.95	0.95	0.95	0.93	0.91	0.9	0.88	0.88	0.87	0.86	0.85	0.86	0.87	0.89	0.86	0.87	0.85	0.81	0.78	C	0.80	
++		1	2	3	4	5	6	7	8	9	10	11	12	13	14 _	15 Fransf	16 ormer	17 Blocl	18 k Inde	19 x	20	21	22	23	24	25	26	27	28	29	30	31	32	0		

The near-perfect peak accuracy the model's internal representations have developed a distinction between true/false statements about a given puzzle.

Puzzles with more #ppl demands more internal computation (probing accuracy > 85% shifting to later transformer blocks).

## Memorization vs Reasoning



Model's reasoning capability improves as the memorization score on the training set increases

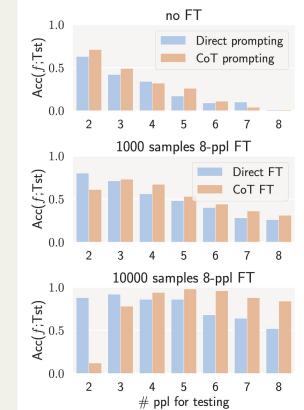
## Empirical Sample Complexity of K&K Reasoning

### The Problem Space

- The number of unique problems for 8-people K&K puzzles (depth=2, width=2) is ~10^24.
- The percentage of problems (empirically estimated by randomly generating 100,000 puzzles) with a unique solution is ~30%.

### The Sample Complexity

- Fine-tuning with 10,000 examples significantly improves the test accuracy, test accuracy reaches ~90% on moderately difficult 4/5-people puzzles.
- The benefit of reasoning demonstration (CoT Fine-tune) is more pronounced in the 10,000 examples case.



Fine-tune w/

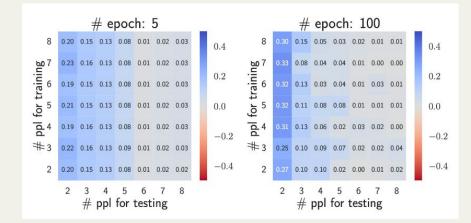
GPT4o-mini

## Fine-tuning on Wrong Answers

**Wrong answer setup**: randomly select N'  $(1 \le N' \le N)$  and flip the knight/knave identities of N' randomly chosen individuals in the answer. Measure the performance **improvement** after fine-tuning.

		GP	T4o-n	nini #	epoch	: 5							
1.0	-0.39	-0.20	-0.24	-0.15	-0.07	-0.10	-0.01		0.4				
	-0.15	-0.08	-0.16	0.01	0.03	-0.06	-0.01		0.2				
vrong t	0.09	0.24	0.08	0.09	0.09	-0.02	0.09		0.0				
broportion of wrong training 0.5 0.0 0.0	0.15	0.30	0.28		0.29	0.24	0.07		-0.2				
0.0	0.25	0.35	0.37	0.40	0.37	0.24	0.26		-0.4				
	2 3 4 5 6 7 8 # ppl for testing												

GPT4o-mini, Direct-FT w/ 5-ppl puzzles, improvements observed for <50% wrong answers.



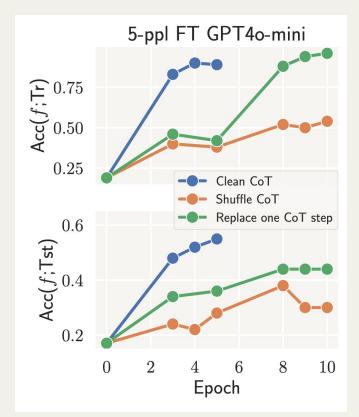
Llama3-8B, Direct-FT w/ M-ppl puzzles, where **100%** of the answers are perturbed to be incorrect.

(For each wrong answer example, there are still N-N' correct role assignments. So some regularity still exists.)

## Fine-tuning on Wrong Reasoning Steps

**Setup**: Fine-tuning on correct answers, but wrong reasoning steps (Chain-of-Thoughts).

- Cot FT on a 100% corrupted Cot dataset can still enhance test accuracy over the epochs.
- Wrong CoT steps slow convergence and hurt test accuracy.



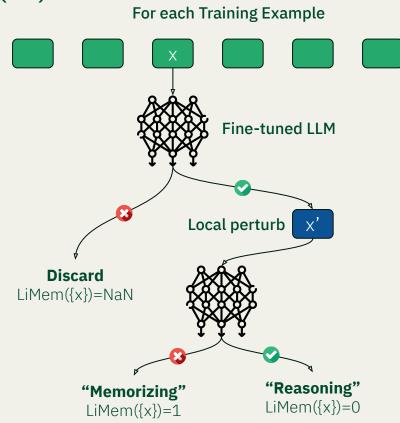
## Reasoning vs Memorization (II)

**Observation**: Model is doing both memorization and reasoning.

### **Question:**

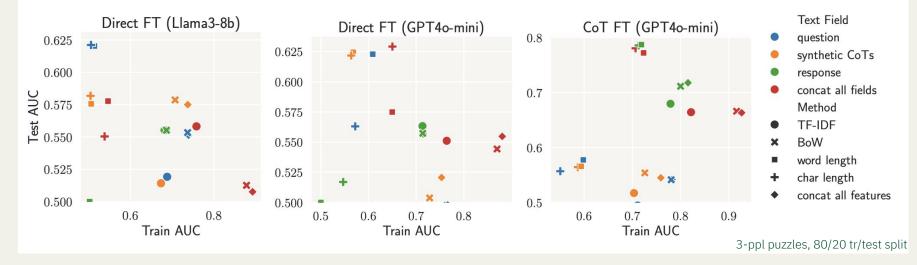
- Why having two (or more?) modes of operation?
- How does different modes get developed?
- When to perform reasoning / memorization?

**Experiment:** Discrimination of K&K training puzzles that are solved by memorization or solved by reasoning. Binary labels generated by per-example  $LiMem({x}) \in {0, 1}.$ 



## Reasoning vs Memorization (II)

### Puzzle-based Indicator: Could the decision be based on simple features of the puzzles?

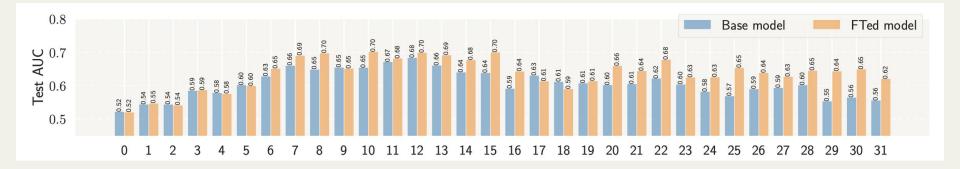


Best test AUC of 0.629/0.787 for Direct/CoT FT-ed GPT4o-mini, and 0.627 for Direct FT-ed Llama3-8B.

> Puzzle-based indicators could be informative, though not perfect.

## Reasoning vs Memorization (II)

**Model-based Indicator**: Could the decision be arbitrary (depending on the randomness of model initialization and fine-tuning)?



The features from the FTed model are consistently more informative than the un-FTed model.

Model-based indicators reach higher AUCs than feature-based indicators (for Llama3-8B).

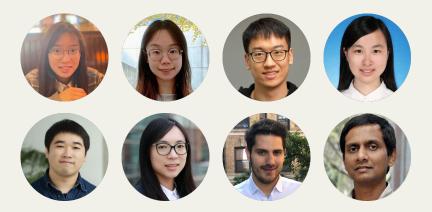
## Conclusion

- Propose a memorization score for LLM reasoning and a dynamical logical reasoning benchmark based on Knights and Knaves (K&K) puzzles.
- We found that LLMs could have near-perfect training accuracy after fine-tuning, yet fail when those puzzles are slightly perturbed, suggesting that the models heavily memorize those training puzzles.
- On the other hand, we show that while fine-tuning leads to heavy memorization, it also consistently improves generalization performance. It suggests that the LLMs learn to reason on K&K puzzles despite training data memorization.



<u>https://memkklogic.</u> github.io

Chiyuan Zhang (Google Research)



## Future of (Evaluating LLM) Generalization?

- Possible presence of data contamination
- Evaluation of **capability** (e.g. Grade-6 math) rather than expected loss / accuracy
- Evaluation of **open ended** questions, how to define loss(pred, label)

#### Bagging Provides Assumption-free Stability

Jake A. Soloff<sup>1</sup>, Rina Foygel Barber<sup>1</sup>, and Rebecca Willett<sup>1,2</sup>

<sup>1</sup>Department of Statistics, University of Chicago <sup>2</sup>Department of Computer Science, University of Chicago

April 26, 2024

#### Abstract

Bagging is an important technique for stabilizing machine learning models. In this paper, we derive a finite-sample guarantee on the stability of bagging for any model. Our result places no assumptions on the distribution of the data, on the properties of the base algorithm, or on the dimensionality of the covariates. Our guarantee applies to many variants of bagging and is optimal up to a constant. Empirical results validate our findings, showing that bagging successfully stabilizes even highly unstable base algorithms.

#### Can you provide a summary of this paper?

2301.12600v3

PDF

#### Analysis complete 🗸 🗸

Bagging is a widely used technique for stabilizing machine learning models. In this paper, the authors derive a finite-sample guarantee on the stability of bagging for any model. Their result places no assumptions on the distribution of the data, on the properties of the base algorithm, or on the dimensionality of the covariates. Their guarantee applies to many variants of bagging and is optimal up to a constant. They validate their findings with empirical results, showing that bagging successfully stabilizes even highly unstable base algorithms.



Can you write a summary of this naner?

The paper, titled "Bagging Provides Assumption-free Stability," investigates the theoretical stability properties of the bagging (bootstrap aggregating) algorithm. It presents a framework to certify the stability of bagging across a wide range of markine learning agorithms without relying on assumptions about the data distribution, algorithmic properties, or feature space dimensions. Here are the key highlights:

#### Abstract and Motivation:

- Stability Definition: Stability measures the sensitivity of a machine learning algorithm to small
  changes in the training data. This is important for generalization, feature selection, and
  reproducibility.
- Main Contribution: The authors provide finite-sample stability guarantees for bagging, demonstrating its effectiveness as a stabilizing method, even for unstable base algorithms.

#### Key Results:

- Finite-sample Guarantee: The authors establish a formal stability bound for bagging when applied to arbitrary base algorithms with bounded outputs.
  - · The stability guarantee holds across various bagging variants (e.g., subbagging,