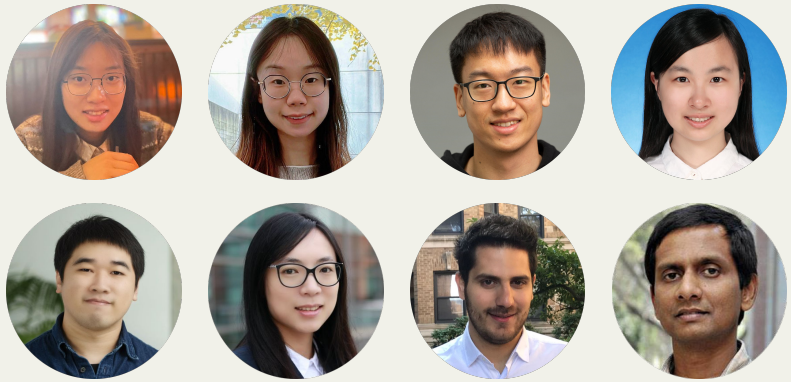
On Memorization of Large Language Models in Logical Reasoning

Chiyuan Zhang

Google Research



https://memkklogic.github.io

Contents

Shifting Perspective of Generalization

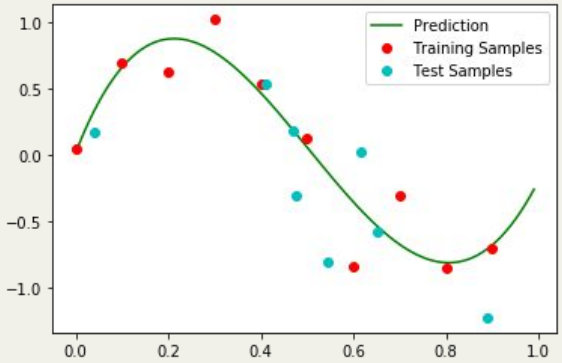
CLASSICAL ML

Test Acc = Train Acc + Gen Gap

Perspective: Bound Gen Gap for ERM

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

Low degree polynomial



High degree polynomial

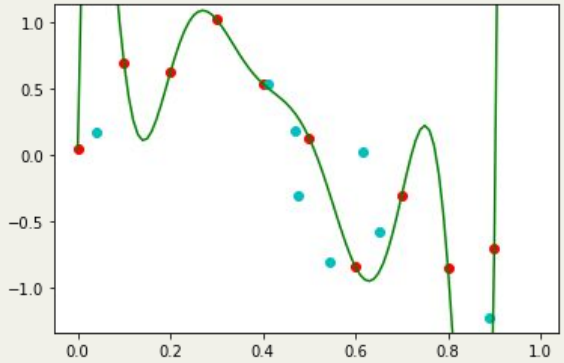


Fig from https://github.com/jamt9000/prml/blob/master/1.1-polycurve.ipynb

Shifting Perspective of Generalization

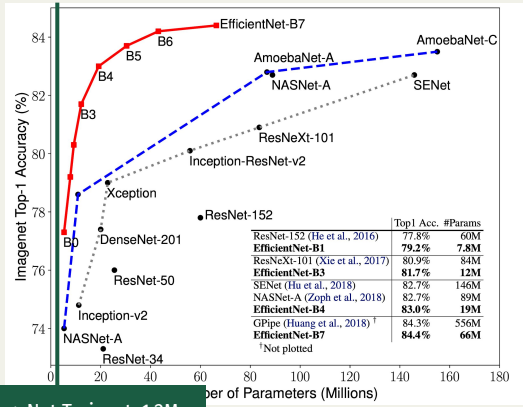
PRE-LLM REGIME

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...

Arch research & model sizes (2019)



ImageNet Trainset: 1.2M

Generalization Research

Improved Generalization Bound

Implicit Regularization

Benign Overfitting

Double Descent

Geometry of Loss Landscape

Out of Distribution Generalization

...

Fig source: https://arxiv.org/abs/1905.11946

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

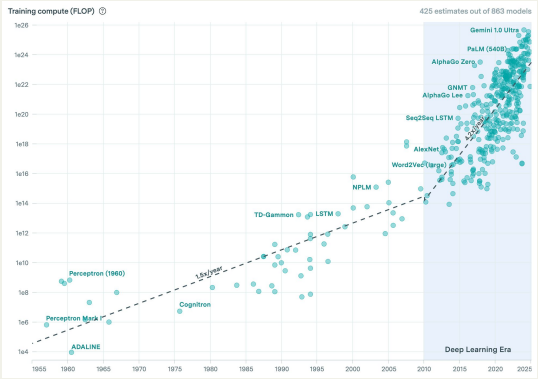
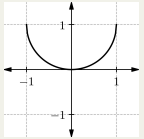


Fig source: https://epoch.ai/data/notable-ai-models

Example Problem from MATH dataset

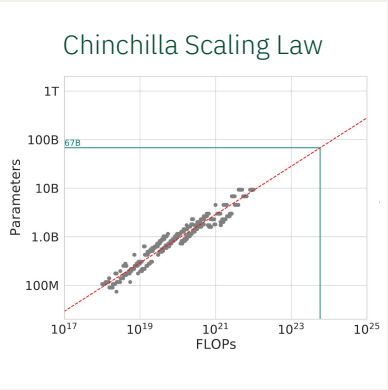
Let $f(x)$ be the function defined on $-1 \leq x \leq 1$ by the formula $f(x) = 1 - \sqrt{1 - x^2}$. This is a graph of $y = f(x)$:



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?

MATH dataset: https://huggingface.co/datasets/lighteval/MATH

Generalization in LLM Regime

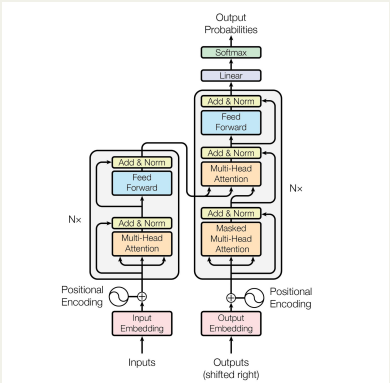


Underparameterization + Undertraining

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

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

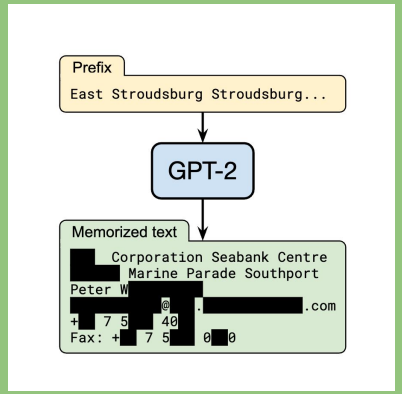
Fine-tuning can still be interpolating



Transformer + Attention

Attention mechanism & in-context learning.

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

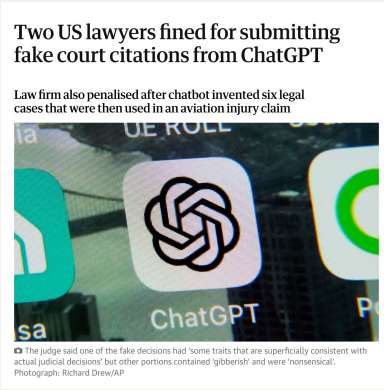


Memorization

Practical: Privacy + Copyright Concerns

Scientifical: Generalization / Memorization

Memorization vs Memory (of LLM)



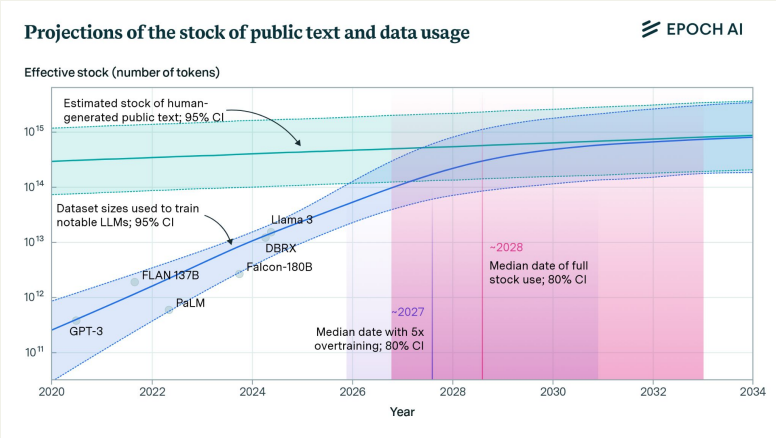
Beyond Generalization

Generalization =?=>
> Hallucinations / Grounding / Factuality
> Alignment / Instruction Following
> Safety / Jailbreaking

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

Fig sources arxiv ids: 2203.15556, 2408.00118, 2012.07805

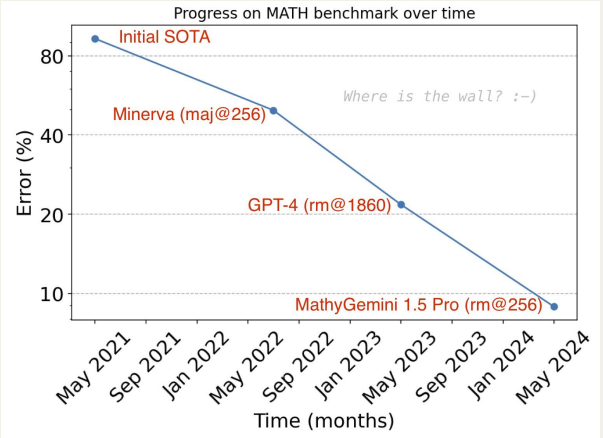
Two Trends of Saturation



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

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT
Undergraduate level knowledge *MMLU*	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	– 88.7% 0-shot CoT
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	–
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT
Grade school math *GSM8K*	96.4% 0-shot CoT	95.0% 0-shot CoT	–

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



https://x.com/bneyshabur/status/1792304689335480511

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

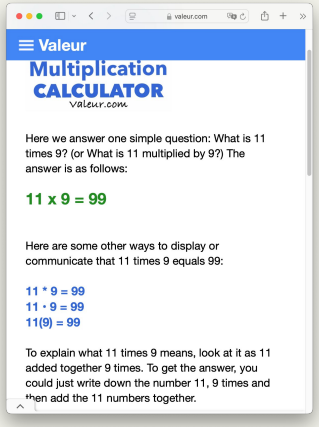
News, News In Brief

Geologists Recommend Eating At Least One Small Rock Per Day

Share

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 pebbles, 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-pound 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 body to subsist on. Because many do not like the taste or texture, we recommend hiding loose rocks inside different foods, like peanut 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

valeur.com

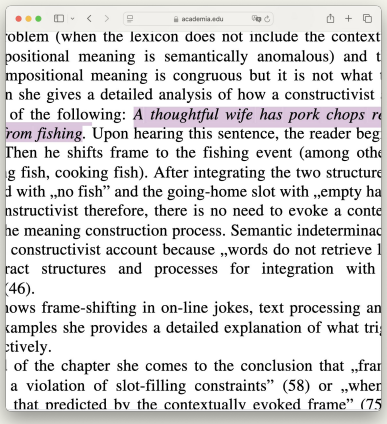
Here we answer one simple question: What is 11 times 9? (or What is 11 multiplied by 9?) The answer is as follows:

11 x 9 = 99

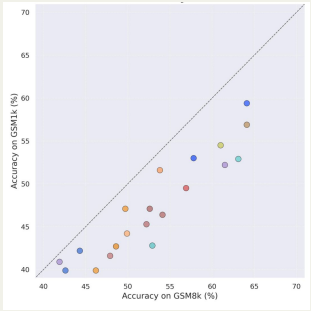
Here are some other ways to display or communicate that 11 times 9 equals 99:

11 * 9 = 99
11 · 9 = 99
11(9) = 99

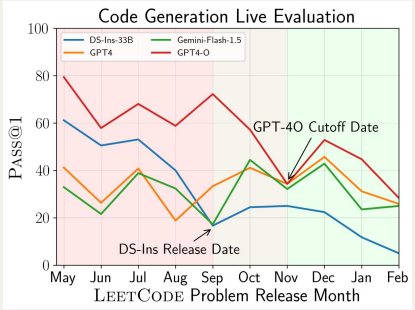
To explain what 11 times 9 means, look at it as 11 added together 9 times. To get the answer, you could just write down the number 11, 9 times and then add the 11 numbers together.



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).
ows frame-shifting in on-line jokes, text processing an
xamples she provides a detailed explanation of what tri
ctively.
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

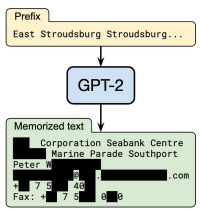
https://x.com/billyuchenlin/status/1812948314360541302

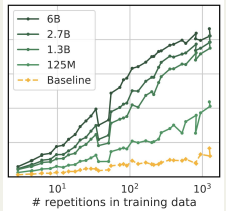
Source: The Economist; The New York Times; WIRED; TechTalks

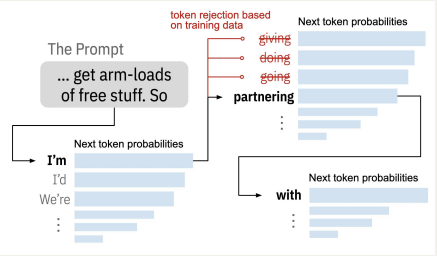
Goal: a quantitative approach to

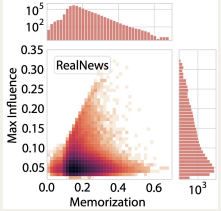
- Measure the amount of memorization
- Measure model performance in the possible presence of data contamination

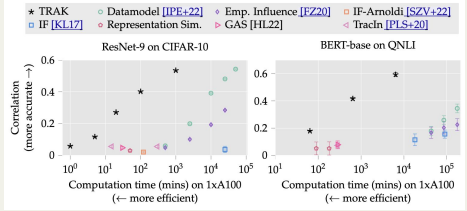
Previous Work on LLM Memorization

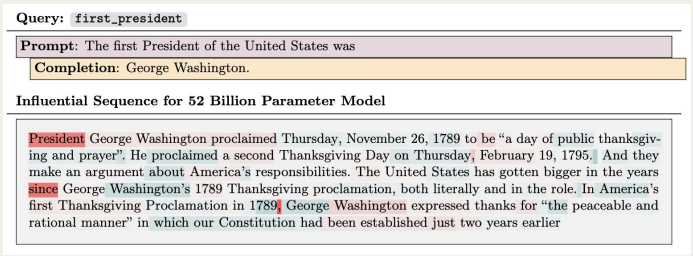


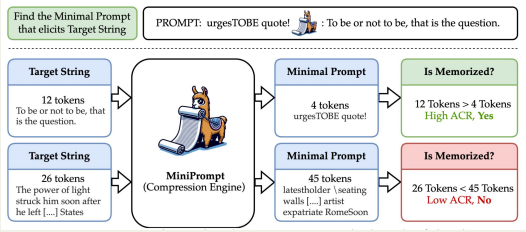












Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already 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 basketball.

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

Illustration generated by the Imagen model.

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

✅ Reasoner ❓ Memorizer

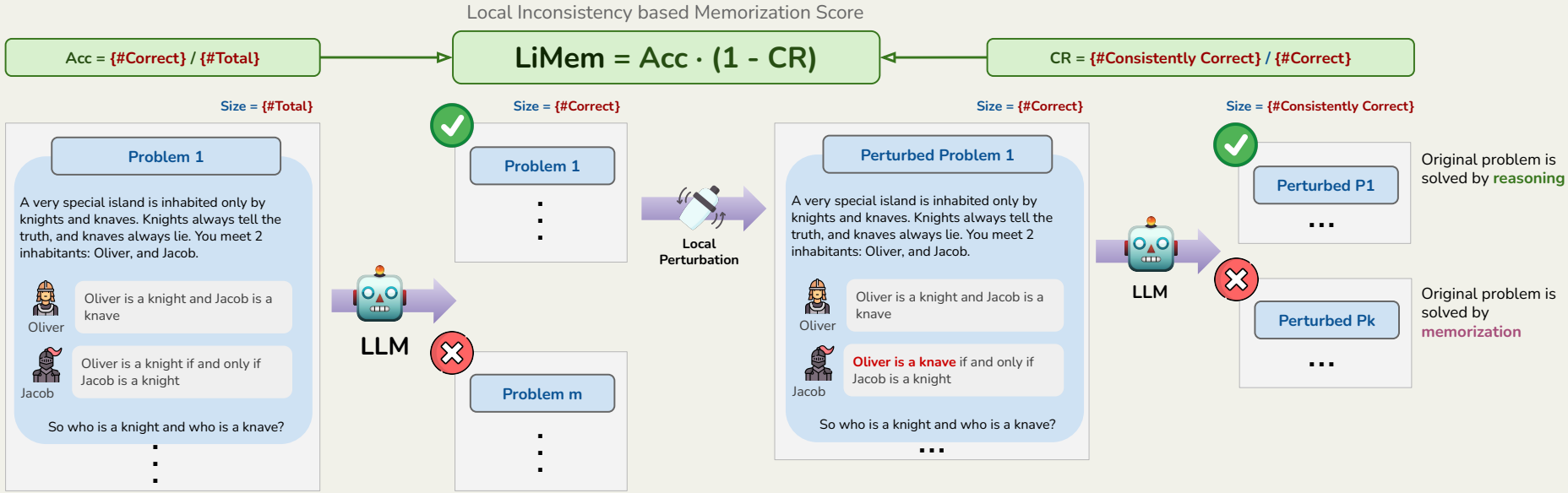
Evaluation with a modified* version of the same test question

✅ Reasoner ❌ Memorizer

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

Local Inconsistency based Memorization Score

High **Acc** (Well "Prepared") → **Memorization** is characterized by ← Low **CR** (Not Well "Learned")



Knights & Knaves Logical Reasoning Benchmark

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

Oliver is a knight and Jacob is a knave

Jacob

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

$$(B_1 \Leftrightarrow S_1) \wedge (B_2 \Leftrightarrow S_2)$$

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.

Knights & Knaves Logical Reasoning Benchmark

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

Oliver is a knight and Jacob is a knave

Jacob

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

$$(B_1 \Leftrightarrow S_1) \wedge (B_2 \Leftrightarrow S_2)$$

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.

Knights & Knaves Logical Reasoning Benchmark

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

Oliver is a knight and Jacob is a knave

Jacob

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

So who is a knight and who is a knave?

GPT-4o 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

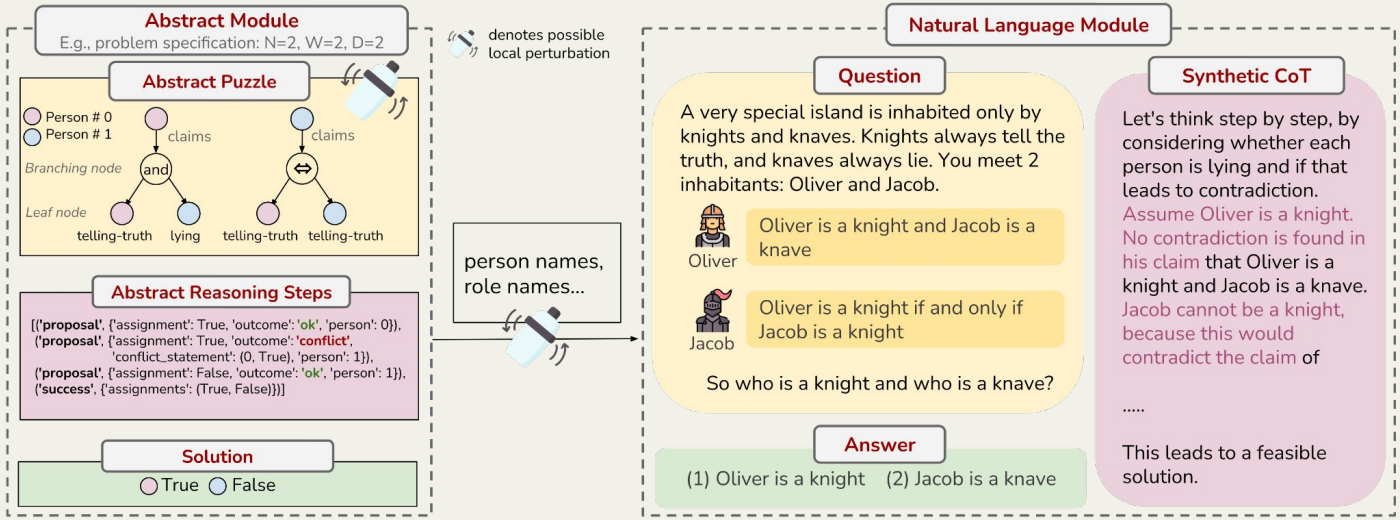
- If Oliver is a knave, his statement is false. So, it is not true that Oliver is a knight and Jacob is a knave. This implies that either Oliver is not a knight or Jacob is not a knave (or both).
- Jacob's statement would be true if Oliver is a knave, which contradicts the nature of a knave who always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

This scenario is inconsistent.

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

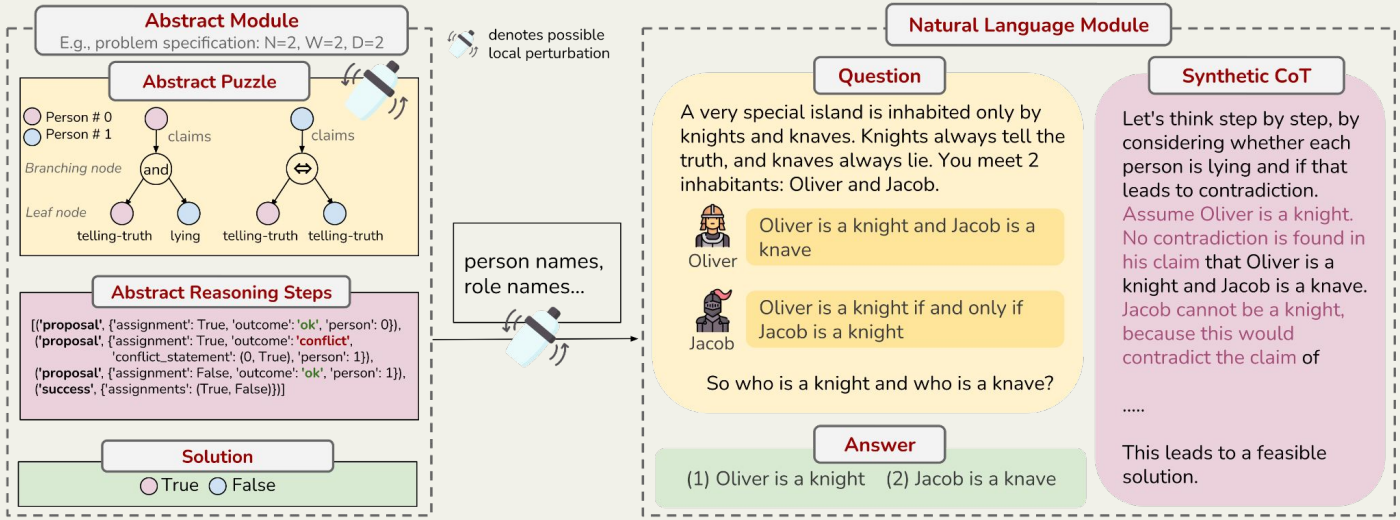
Knights & Knaves Logical Reasoning Benchmark

- **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.



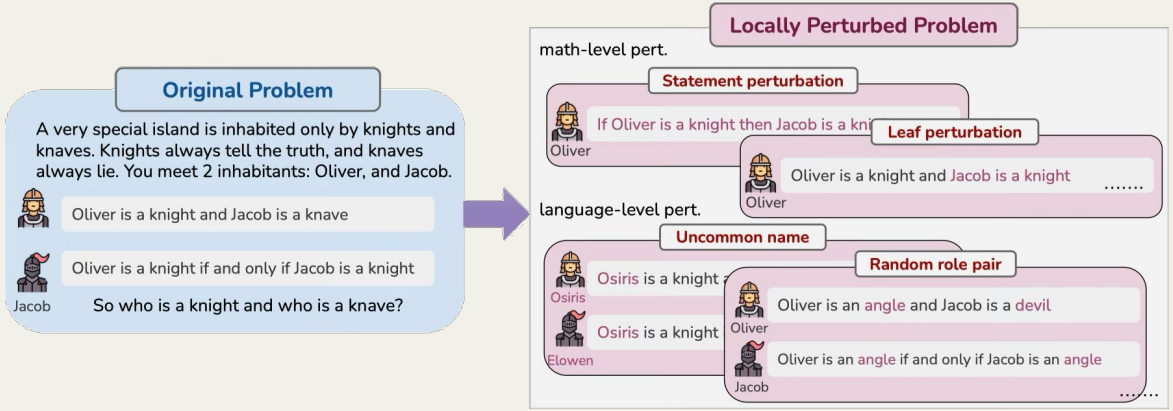
Knights & Knaves Logical Reasoning Benchmark

- 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).



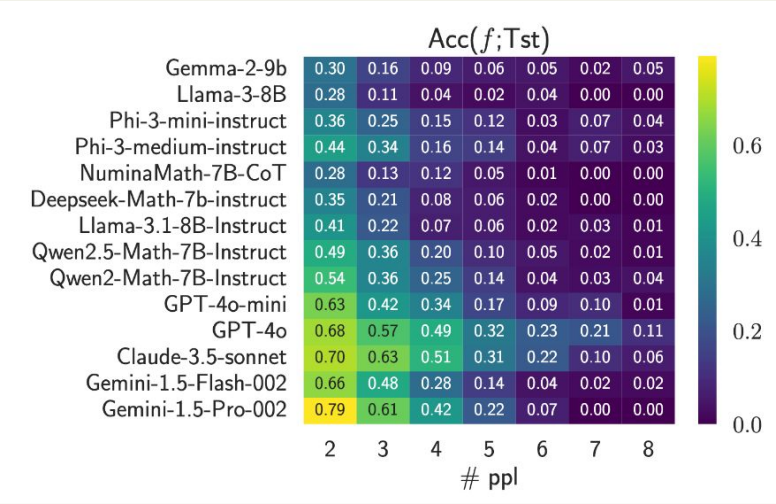
Knights & Knaves Logical Reasoning Benchmark

- 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).



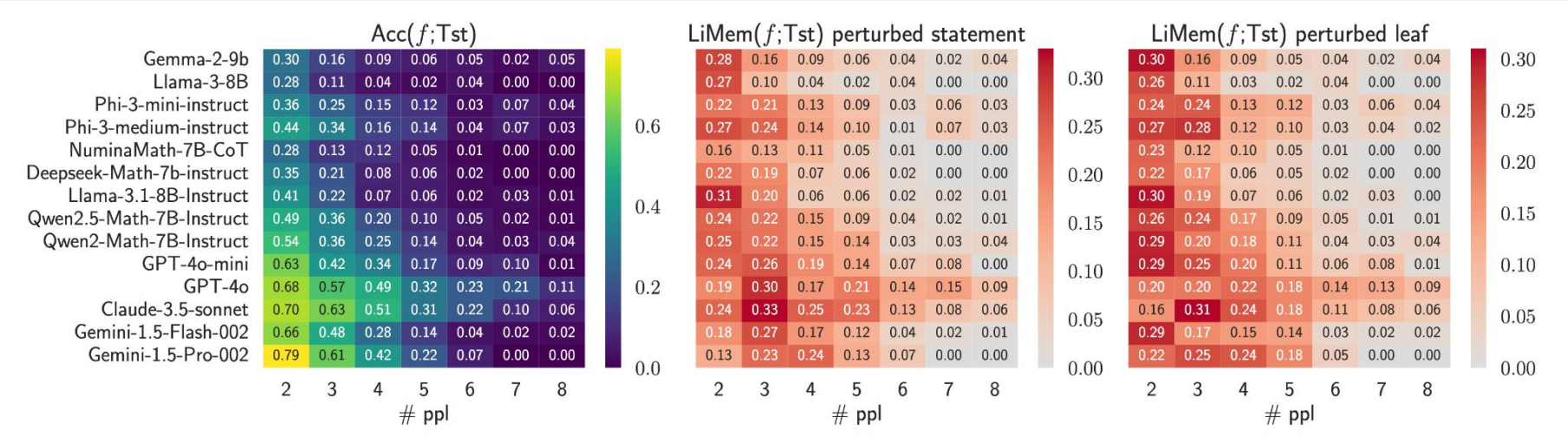
Re-verification of the existence and uniqueness of solutions; Re-sampling if no or multiple solutions exists.

Evaluation Off-the-Shelf Models



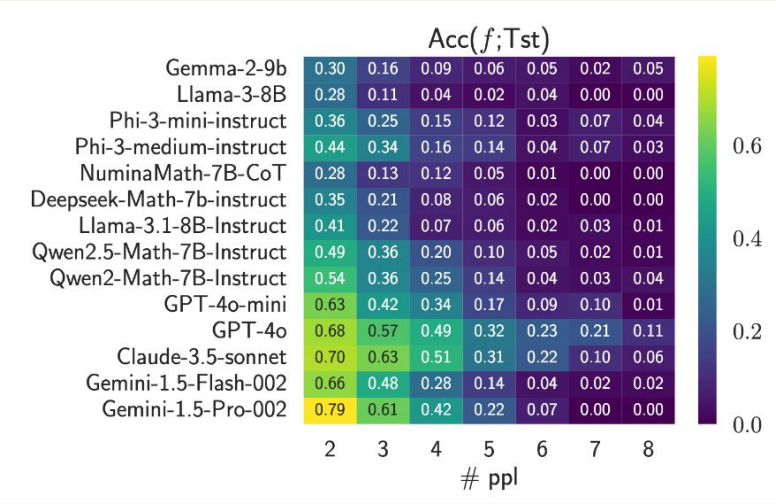
- K&K benchmark poses a challenging logical reasoning task for all off-the-shelf models.
- Accuracy on 2-ppl task can be high.

Memorization in Off-the-Shelf Models



- 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.

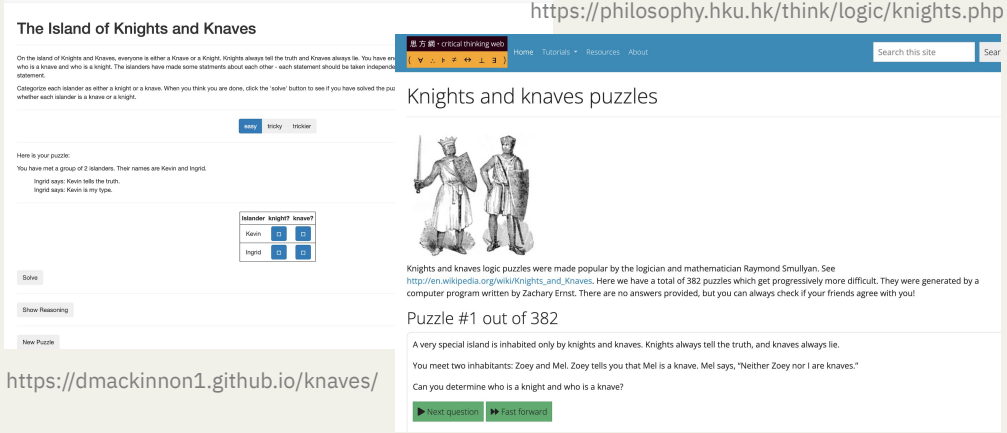
Evaluation Off-the-Shelf Models



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

Online resources about K&K

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



https://dmackinnon1.github.io/knaves/

	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

Oliver is a knight and Jacob is a knave


Jacob

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

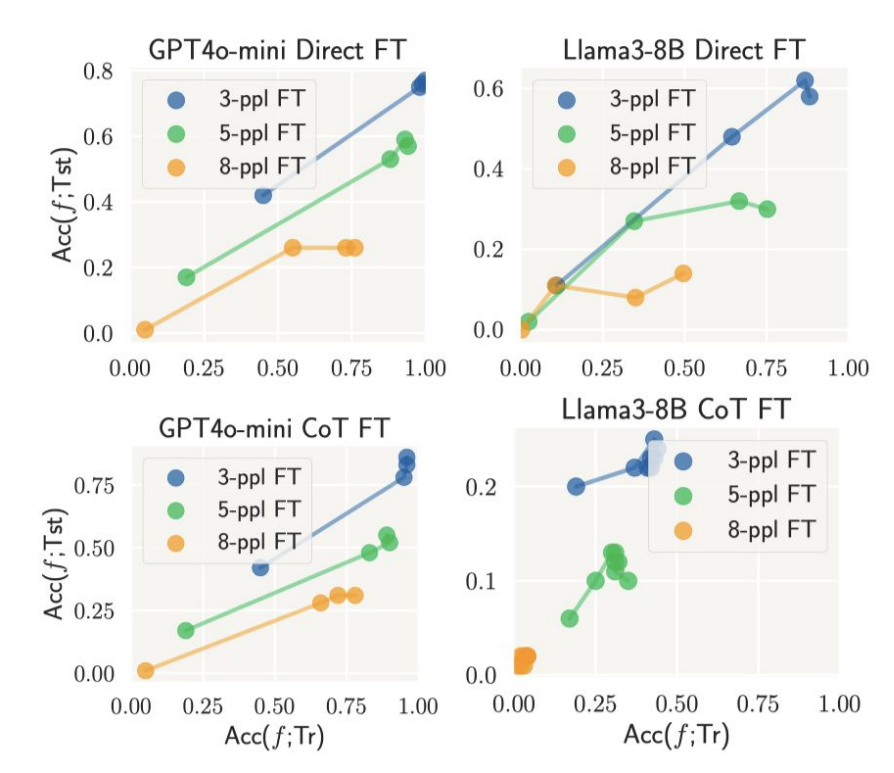
Synthetic CoT

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.

Direct FT

Chain-**o**f-**T**hought **FT**

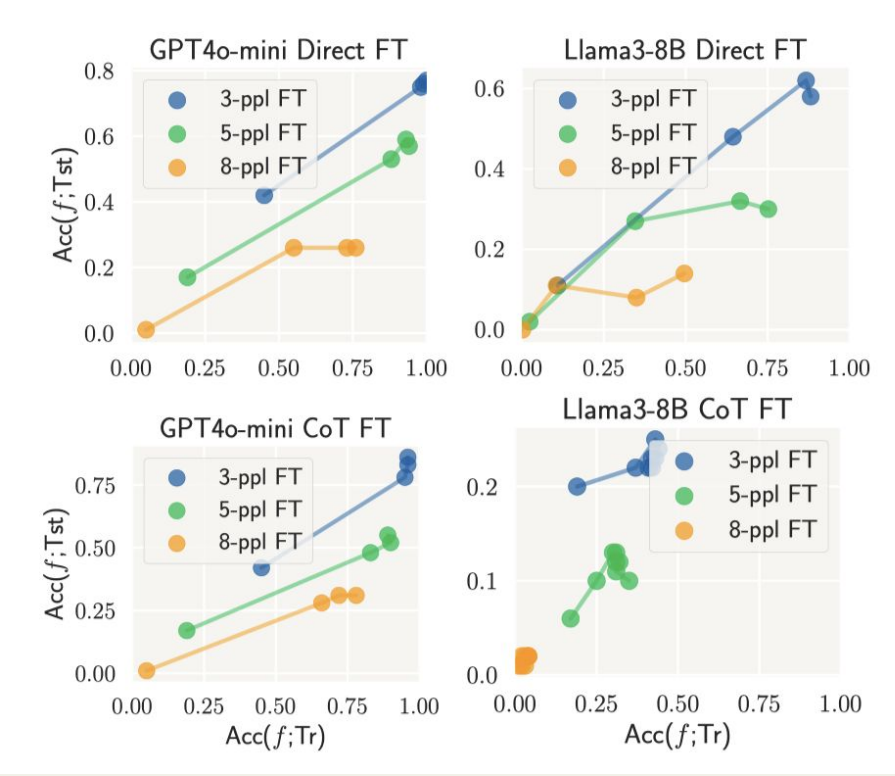
Memorization in Fine-Tuned Models

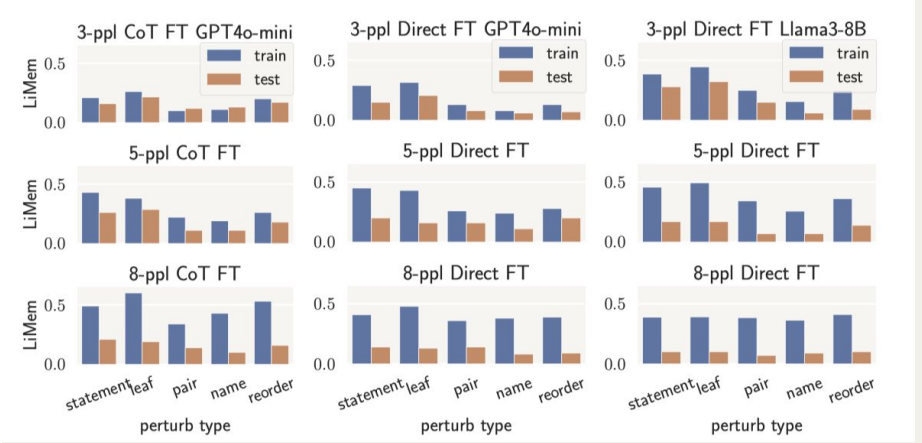


- **Models achieve high training accuracy (approaching interpolation in many cases).**
- CoT fine-tuning seem to be mostly helpful for more difficult cases.
- For model without enough capacity to fit CoT data, the performance is significantly worse than Direct Fine-tuning.

#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

LiMem = Acc · (1 - CR)

High LiMem

High Acc + Low CR

Characteristic behavior of memorization

Mid LiMem

High Acc + High CR?
Low Acc + Low CR?

Low sign of memorization
Can be reasoning or dumb

Low LiMem

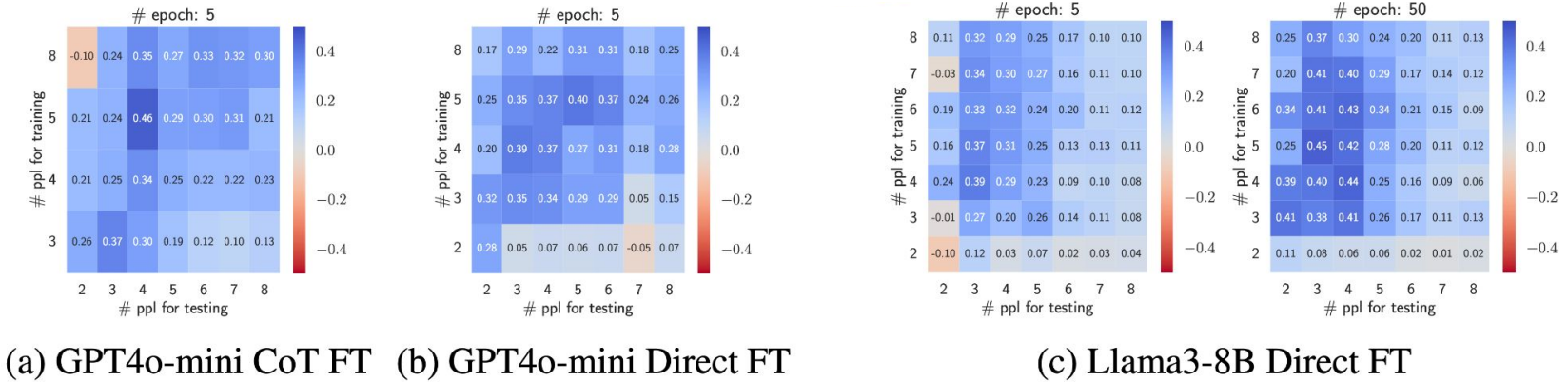
Low Acc + High CR

Dumb memorizer or
Noisy 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



(a) GPT4o-mini CoT FT (b) GPT4o-mini Direct FT (c) Llama3-8B Direct FT

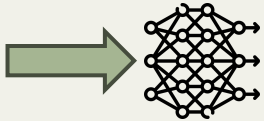
Performance improvements generalize across various difficulty levels, even when fine-tuned without detailed reasoning steps (Chain-of-Thoughts).

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.**

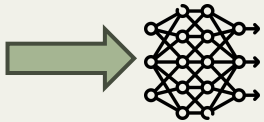


embedding of i-th block

Label: correct

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.**

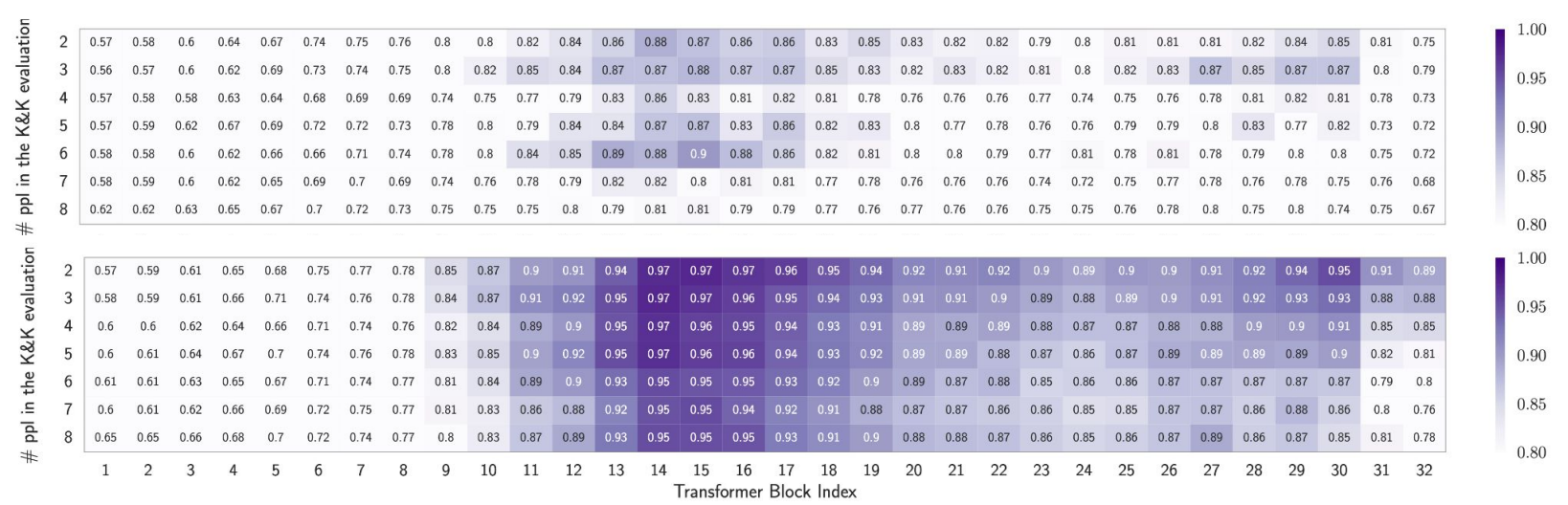


embedding of i-th block

Label: incorrect

Reasoning in Fine-Tuned Models?

Probing Intermediate States of Fine-tuned Models



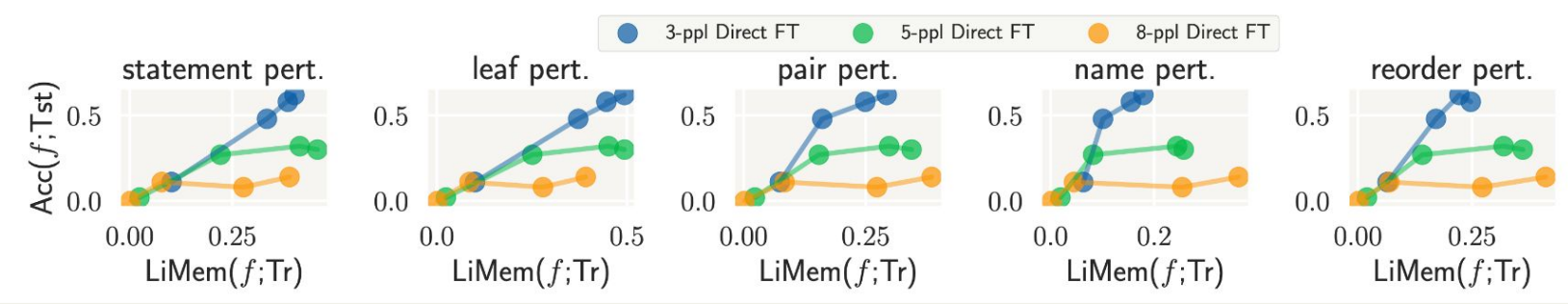
Llama3-8B

Direct FT-ed Llama3-8B

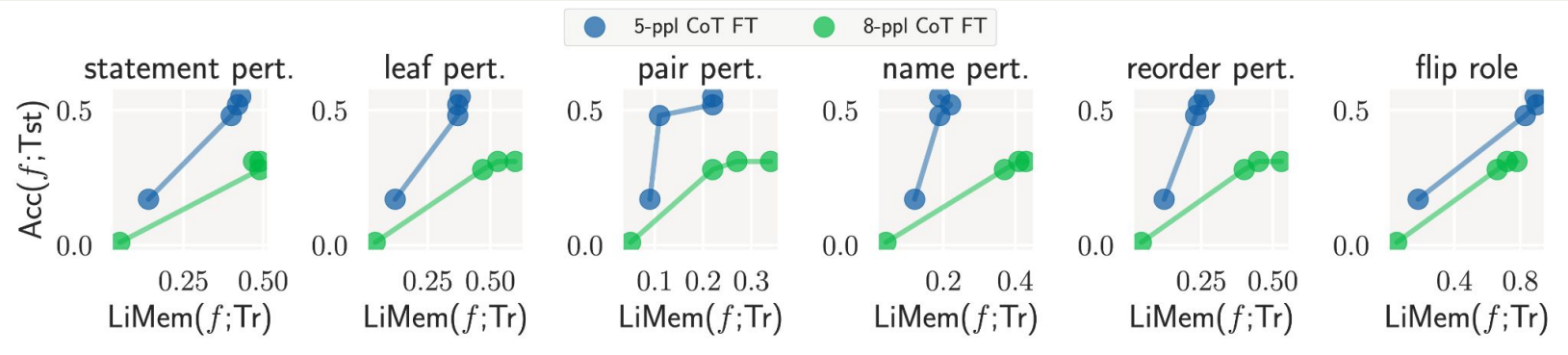
- The near-perfect peak accuracy □ the model's internal representations have developed a distinction between true/false statements about a given puzzle.
- Puzzles with more #ppl demands more internal computation (probing accuracy > 85% shifting to later transformer blocks).

Memorization vs Reasoning

Llama3-8B



GPT4o-mini



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

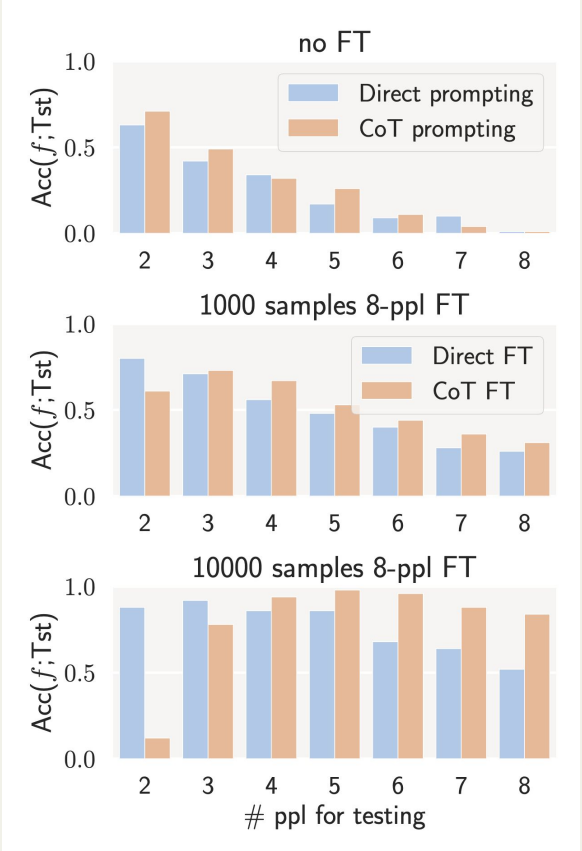
Empirical Sample Complexity of K&K Reasoning

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

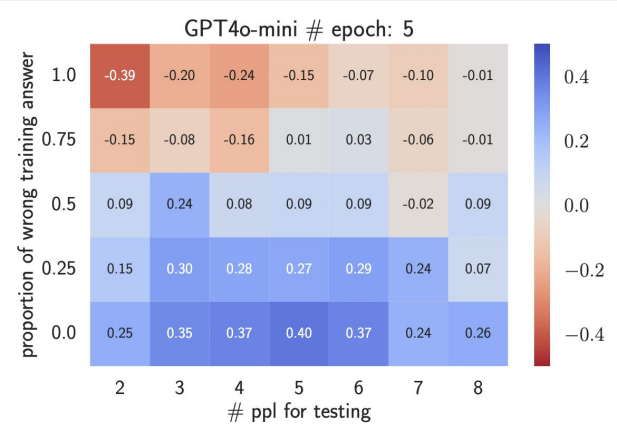
- Fine-tuning with 10,000 examples significantly improves the test accuracy, test accuracy reaches ~90% on moderately difficult 4/5-people puzzles.
- The benefit of reasoning demonstration (CoT Fine-tune) is more pronounced in the 10,000 examples case.



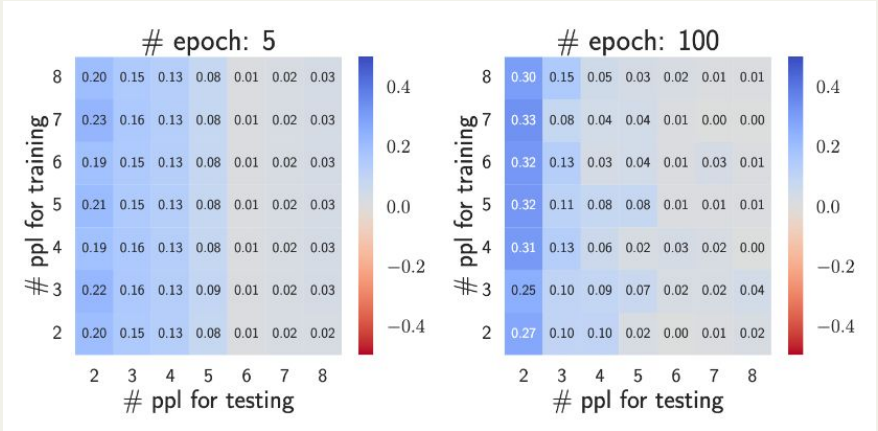
Fine-tune w/ GPT4o-mini

Fine-tuning on Wrong Answers

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



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



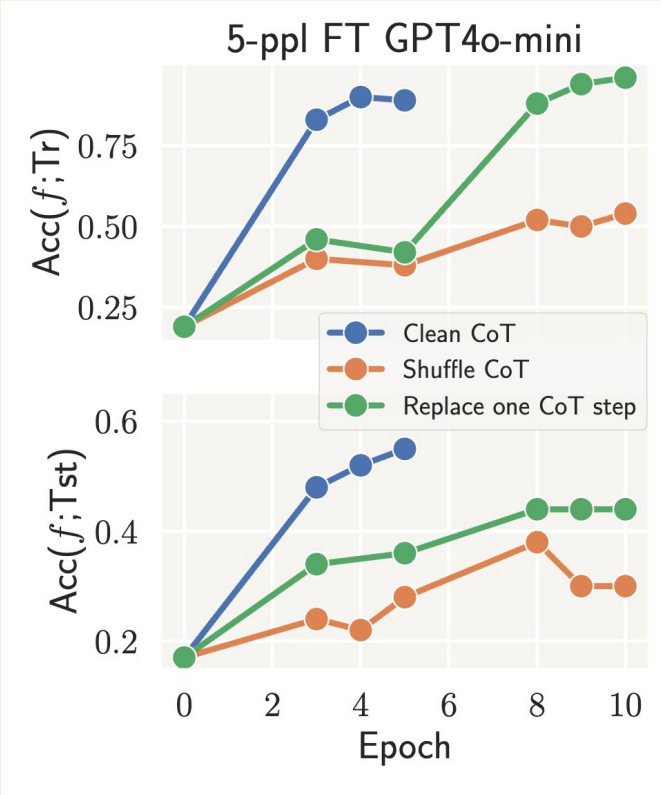
Llama3-8B, Direct-FT w/ M-ppl puzzles, where **100%** of the answers are perturbed to be incorrect.

(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.



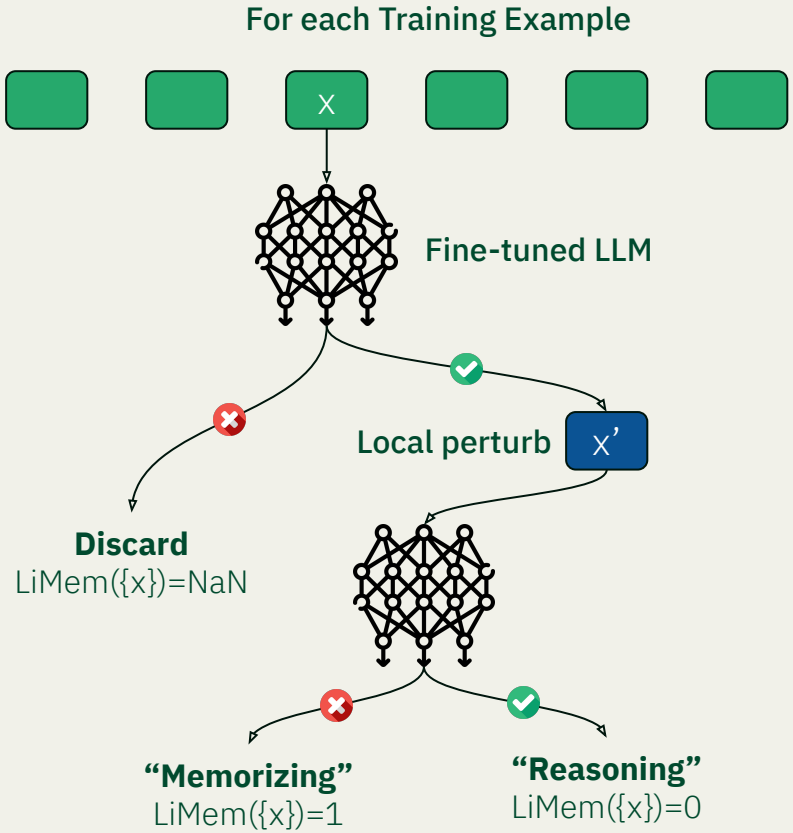
Reasoning vs Memorization (II)

Observation: Model is doing both memorization and reasoning.

Question:

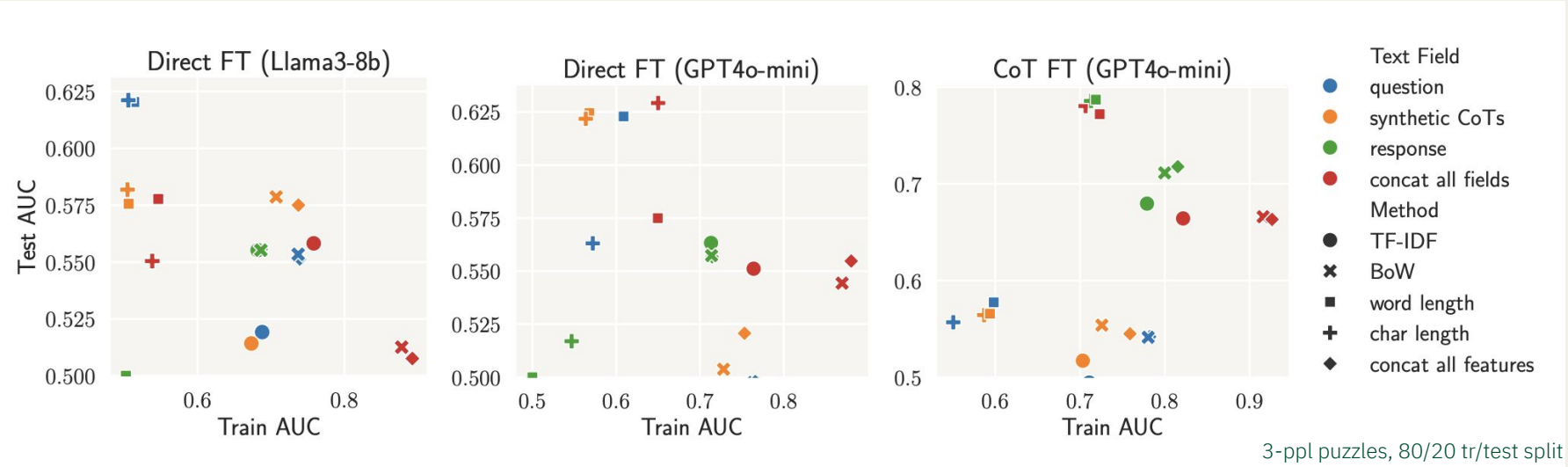
- Why having two (or more?) modes of operation?
- How does different modes get developed?
- When to perform reasoning / memorization?

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



Reasoning vs Memorization (II)

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

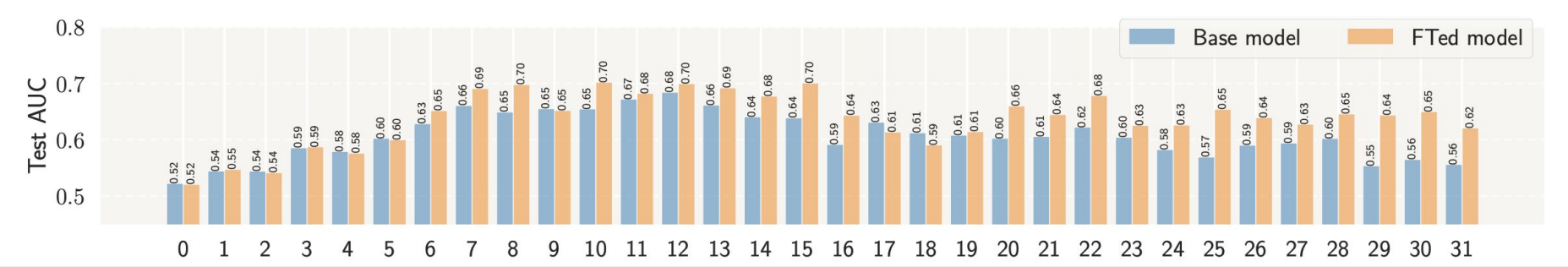


3-ppl puzzles, 80/20 tr/test split

- 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)

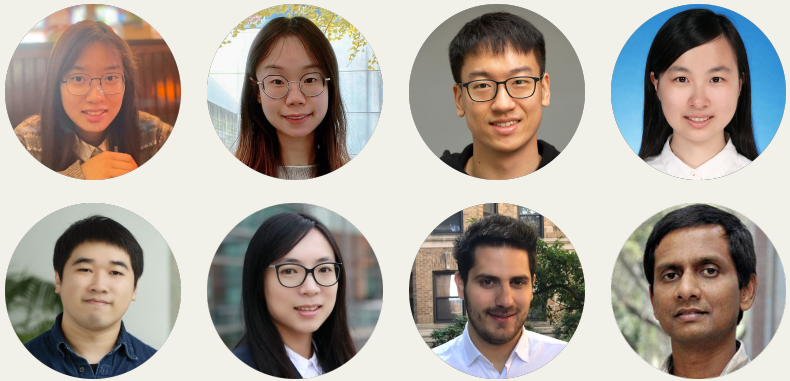
Model-based Indicator: Could the decision be arbitrary (depending on the randomness of model initialization and fine-tuning)?



- 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

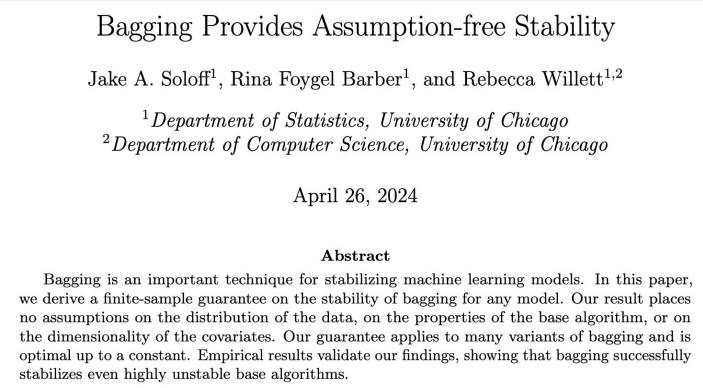
- Propose a **memorization score** for LLM reasoning and a **dynamical** logical **reasoning benchmark** based on Knights and Knaves (K&K) puzzles.
- We found that LLMs could have **near-perfect training accuracy** after fine-tuning, yet fail when those puzzles are slightly perturbed, suggesting that 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.



https://memkklogic.github.io

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

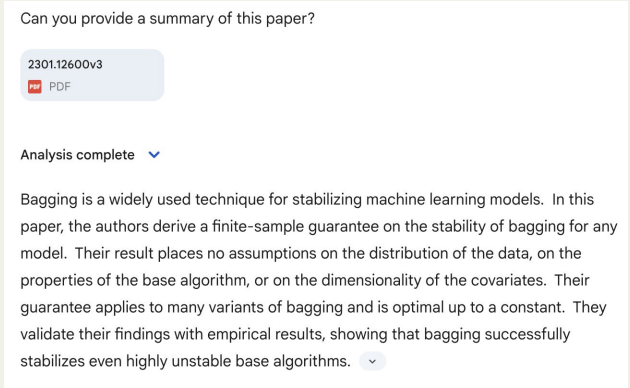
Jake A. Soloff[1], Rina Foygel Barber[1], and Rebecca Willett[1,2]

[1]*Department of Statistics, University of Chicago*
[2]*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.

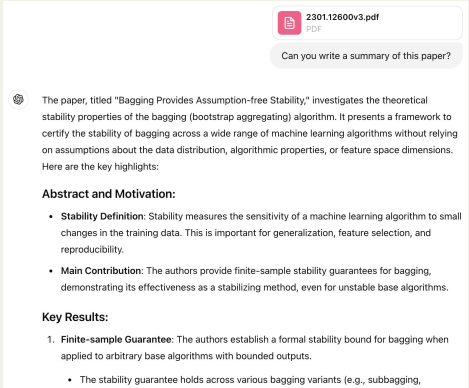


Can you provide a summary of this paper?

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.



2301.12600v3.pdf
PDF

Can you write a summary of this paper?

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 machine learning algorithms 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:

1. **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,