Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process

Result 1 — iGSM: a synthetic, infinite math dataset to simulate GSM8k

pretrain on iGSM + probe model's hidden (mental) reasoning process

Result 2-3 — LLMs exhibit a "level-1" reasoning skill

they mentally compute topo sort + shortest solution – like Humans

Result 4-5 — LLMs secretly learn a "level-2" reasoning skill

they mentally compute "all-pair dependencies" – but Humans don't

Result 6 — probing reveals how LLMs make reasoning mistakes

can catch mistakes even before LLMs start to speak

Result 7-8 — depth matters for long reasoning tasks (even with CoT)

refute OpenAI's scaling law which says "only size matters"

"Physics of LM: Part 2.1, Grade-School Math and the Hidden Reasoning Process"

Result 1

GOAL: study how LLMs solve grade-school math (GSM)

can't use GSM8k – too small, data contamination, etc.
can't use GPT-4 augmented GSM8k
– too biased, too few templates

iGSM – an infinite, synthetic GSM8k-like data

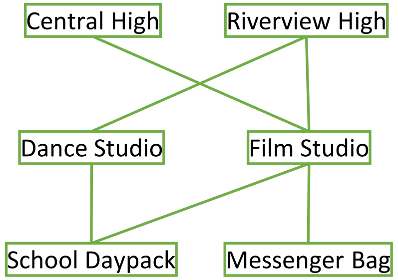
remove common sense (candle burns -> length shrinks) so LLMs can be ***pretrained*** on such data

iGSM aims to capture

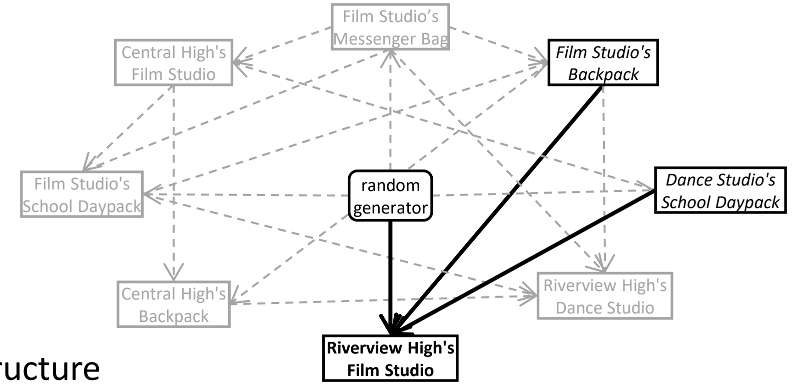
Direct dependency:
e.g. A=5*(X+Y) ⇒ A depends on X and Y

Instance dependency:
e.g. X classrooms each has Y messenger bags

Implicit dependency:
e.g. Bob has 3x more fruits than Alice. Alice has 3 apples, 4 eggs and 2 bananas. (eggs are not fruits)



structure graph



dependency graph (DAG)

an easy example

The number of each **Riverview High's Film Studio** equals 5 times as much as the sum of each *Film Studio's Backpack* and each *Dance Studio's School Daypack*. The number of each Film Studio's School Daypack equals 12 more than the sum of each Film Studio's Messenger Bag and each Central High's Film Studio. The number of each Central High's Film Studio equals the sum of each Dance Studio's School Daypack and each Film Studio's Messenger Bag. The number of each Riverview High's Dance Studio equals the sum of each Film Studio's Backpack, each Film Studio's Messenger Bag, each Film Studio's School Daypack and each Central High's Backpack. The number of each Dance Studio's School Daypack equals 17. The number of each Film Studio's Messenger Bag equals 13. *How many Backpack does Central High have?*

a solution (=CoT) with op=7 operations

Define Dance Studio's School Daypack as p; so p = 17.
Define Film Studio's Messenger Bag as W; so W = 13.
Define Central High's Film Studio as B; so B = p + W = 17 + 13 = 7.
Define Film Studio's School Daypack as g; R = W + B = 13 + 7 = 20; so g = 12 + R = 12 + 20 = 9.
Define Film Studio's Backpack as w; so w = g + W = 9 + 13 = 22.
Define Central High's Backpack as c; so c = B * w = 7 * 22 = 16. ***[Answer: 16.]***

We want to focus on ***reasoning***, not arithmetics; so we use **mod 23**.

"Physics of LM: Part 2.1, Grade-School Math and the Hidden Reasoning Process"

Result 2-3

– Result 2

GPT2 learns iGSM by true generalization

*few-shot GPT-4o fails on **op>=11***

OOD generalization – tested on problems longer than pretrain

pretrain data (Medium): iGSM^{Med} (**op<=15**) (>7 billion solution templates) → **test data** (Medium):

op=15	op=20	op=21	op=22	op=23
99+%	92%	88%	85%	78%

(Hard): iGSM^{Hard} (**op<=21**) (>15 trillion solution templates) → (Hard):

op=21	op=28	op=29	op=30	op=31	op=32
99+%	94%	92%	90%	87%	83%

⟹ truly learns some reasoning skill (not by memorization)

but what reasoning skill?

GPT2 at least achieves "level-1" reasoning skill

a "level-0" reasoning brute-forces to compute all params maximally ✗

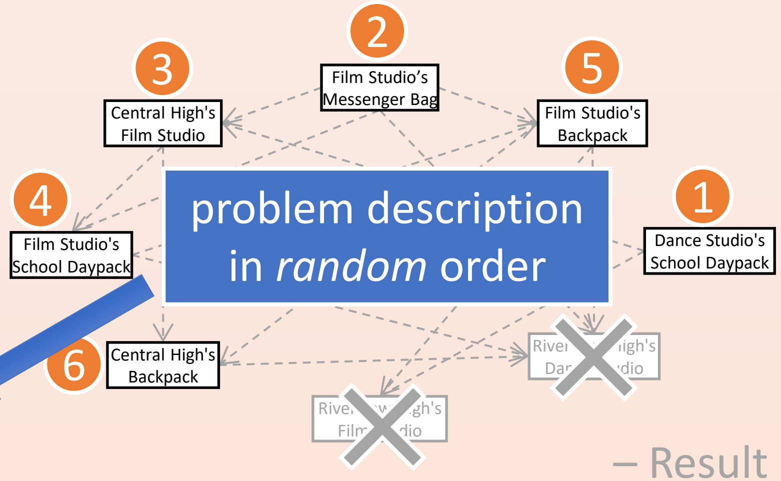
a **"level-1" reasoning** uses topological sort + gives shortest CoT ✓

But how is this possible?

model must **"mental process"** before the first solution sentence **(i.e., before CoT!)**, see Results 4-5

Define Dance Studio's School Daypack as p; so p = 17.
Define Film Studio's Messenger Bag as W; so W = 13.
Define Central High's Film Studio as B; so B = p + W = 17 + 13 = 7.
Define Film Studio's School Daypack as g; ... + R = 12 + 20 = 9.
Define Film Studio's Backpack as w; so w = g + W = 9 + 13 = 22.
Define Central High's Backpack as c; so c = B * w = 7 * 22 = 16.

solution in topological order & *excluding* unnecessary parameters



– Result 3

"Physics of LM: Part 2.1, Grade-School Math and the Hidden Reasoning Process"

Result 4-5

How does the model "think"?

our V-probing technique (see paper)

dep(*A*,*B*) – at the end of problem description, does model know *parameter A* ***recursively*** *depends on B*? ***It does! 99%***

nece(*A*) – after question is asked, does model know *if A is necessary* for answering the question for all *A*? ***It does! 99%***

"**[Problem]** The number of each Riverview High's Film Studio equals 5 times ... The number of each Film Studio's Messenger Bag equals 13. ***[Question]*** *How many Backpack does Central High have?* **[Solution]** *Define Dance Studio's School Daypack as p ... Define Central High's Film Studio as B; so B = p + W = 17 + 13 = 7. Define* .... ***[Answer]*** *16.* "

can_next(*A*) – in middle of solution, does model know if *A can be computed next* for all *A*? ***It does! 99%***

⟹ explains how GPT2 **achieves "level-1" reasoning** (i.e. generate shortest solution using topological sort) – Result 4

– Result 5

GPT2 uses "level-2" reasoning skill different from Humans

this skill is not needed for solving the math problem

GPT2 learns dep(*A*,*B*) and can_next(*A*) **even for all unnecessary *A***

⟹ it *mentally* computes all-pair dependency graph *before the question* is asked **(a "level-2" reasoning skill)**

Humans *start from question* to only identify necessary parameters **(human's backward reasoning skill)**

may be preliminary signal of where **G** in AGI can come from (generalizing to skills not taught in pretrain data)

"Physics of LM: Part 2.1, Grade-School Math and the Hidden Reasoning Process"

Result 6

mistakes from nece(*A*)

from Results 4-5: *before* generating solution, model "mentally" calculates what params are necessary
⟹ if param A is **wrongly** calculated as **nece(*A*)**=True in planning stage, model will likely say it in the solution

can detect such mistakes before model opens mouth (using V-probing)
⟹ mistakes are systematic, not random from the generation process

How LLMs makes mistakes on iGSM?

GPT-4 (few shot)
GPT-4o (few shot)
GPT-2 (pretrained on iGSM)

All make two types of mistakes

WARNING: unnecessary parameter *A*

1. Define Dance Studio's Messenger Bag as S; so S = 3.
2. Define **Lakeshore High's Dance Studio** as D; so D = 2.
3. Define Lincoln High's Dance Studio as L; so L = S * 7 = 3 * 7 = 21.
4. Define Messenger Bag's Calculator as C; so C = S = 3.
5. Define **Dance Studio's Canvas Backpack** as …

ERROR: parameter *A* not ready to compute

mistakes from can_next(*A*)

from Results 4-5: in the middle of solution, model "mentally" figures out the full set of params ready to compute
⟹ if param A is **wrongly** calculated as **can_next(*A*)**=True, model will likely say it in the next sentence

⟹ To improve model's reasoning, it is critical to improve its "can_next" accuracy (see our Part 2.2 paper)

"Physics of LM: Part 2.1, Grade-School Math and the Hidden Reasoning Process"

Result 7-8

Prior works: only size matters for LLMs

Scaling laws from OpenAI [2020]: "**width or depth** have **minimal effects** within a wide range"
Scaling laws from "Physics of LM, Part 3.3" [2024]: "for **knowledge skills**, only size matters"

We claim: Depth matters for reasoning

– Result 7

	iGSM-med_pq						iGSM-med_qp					
	in-dist		out-of-dist (OOD)				in-dist		out-of-dist (OOD)			
	op ≤ 15	op=15	op=20	op=21	op=22	op=23	op ≤ 15	op=15	op=20	op=21	op=22	op=23
dep4 - size1 - head21	99.5	92.7	74.7	68.0	62.4	54.5	99.4	93.3	73.3	66.8	61.1	54.6
dep4 - size2 - head30	99.6	94.7	74.2	67.9	61.6	53.1	99.4	94.5	78.1	71.9	65.7	58.8
dep8 - size1 - head15	100	98.8	89.7	86.5	82.8	76.8	100	99.2	92.4	88.5	84.2	78.7
dep8 - size2 - head21	100	99.3	93.7	91.6	88.3	83.6	99.9	99.0	90.2	87.1	83.3	76.3
dep12 - size1 - head12	100	99.3	92.0	88.9	84.2	77.9	100	99.4	92.2	89.2	83.9	77.9
dep12 - size2 - head17	100	99.5	94.0	91.9	89.0	82.7	100	99.0	90.8	85.4	80.2	73.2
dep16 - size1 - head10	100	99.6	94.6	91.9	87.9	82.7	100	99.5	89.9	85.0	79.1	71.1
dep16 - size2 - head15	100	99.8	95.9	93.7	90.4	86.5	100	99.8	95.6	93.5	90.3	84.3
dep20 - size1 - head9	100	99.8	95.5	93.6	90.0	86.3	100	99.6	94.8	91.4	87.4	80.4
dep20 - size2 - head13	100	99.8	95.8	93.3	89.2	84.4	100	99.6	93.7	91.8	87.4	81.3

	iGSM-hard_pq							iGSM-hard_qp							avg
	in-dist		out-of-dist (OOD)					in-dist		out-of-dist (OOD)					
	op ≤ 21	op=21	op=28	op=29	op=30	op=31	op=32	op ≤ 21	op=21	op=28	op=29	op=30	op=31	op=32	
dep4 - size1 - head21	98.9	90.8	72.4	67.7	62.1	57.1	50.6	99.1	89.8	69.4	62.2	57.8	52.3	45.7	72.2
dep4 - size2 - head30	97.0	71.7	46.3	40.6	37.0	32.3	27.3	99.4	92.1	74.5	69.5	64.7	59.1	53.2	68.6
dep8 - size1 - head15	100	99.1	94.6	92.0	89.7	86.4	82.2	100	99.0	92.2	89.6	86.2	82.4	77.3	90.3
dep8 - size2 - head21	100	99.2	93.6	91.3	88.6	85.6	82.6	100	99.1	93.5	91.3	89.1	85.7	81.2	91.3
dep12 - size1 - head12	100	99.5	96.0	94.1	91.0	88.5	84.3	100	99.3	95.3	93.0	91.9	88.0	84.5	91.9
dep12 - size2 - head17	100	99.8	97.1	95.5	93.5	91.8	88.0	100	99.5	94.5	91.9	88.9	86.8	81.3	92.1
dep16 - size1 - head10	100	99.6	97.0	95.2	94.2	92.2	88.5	100	99.4	95.8	93.8	92.4	88.9	85.8	92.5
dep16 - size2 - head15	100	99.7	97.5	96.3	95.1	92.9	89.5	100	99.8	97.3	96.0	94.2	91.9	88.9	95.0
dep20 - size1 - head9	100	99.8	97.0	95.1	94.0	91.0	87.4	100	99.6	96.6	94.5	92.8	90.1	86.5	94.0
dep20 - size2 - head13	100	99.8	98.0	96.7	95.9	93.9	90.9	100	99.9	97.5	96.0	95.2	92.4	89.7	94.7

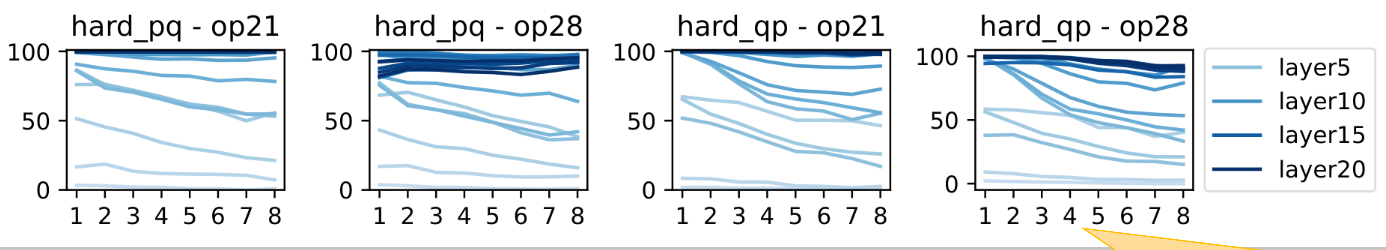
larger model, smaller depth (4-layer, 30-head)

smaller model, larger depth (20-layer, 9-head)

This cannot be mitigated by CoT – deciding what's the first CoT step may still require deep, multi-step mental reasoning (planning)

– Result 8

Depth matters because of the complexity of mental reasoning



parameter's **distance** ***t*** to the question

$\Longrightarrow$ deeper layers are better to compute nece(A) for larger t (which requires t-steps of mental reasoning)

see paper for 44 more figures like this