Fall 2024 COS597R: Deep Dive into Large Language Models

Danqi Chen, Sanjeev Arora

PRINCETON UNIVERSITY

Lecture 11: LLM Metacognition

https://princeton-cos597r.github.io/

Are humans "aware" of what they are doing?

A container starts with 500 mL of water. Each minute, the scientist adds water equal to $1/2$ of the current amount. What is the smallest positive integer n such that the number of liters of water in the container is never in the interval $[n, n+1]$?

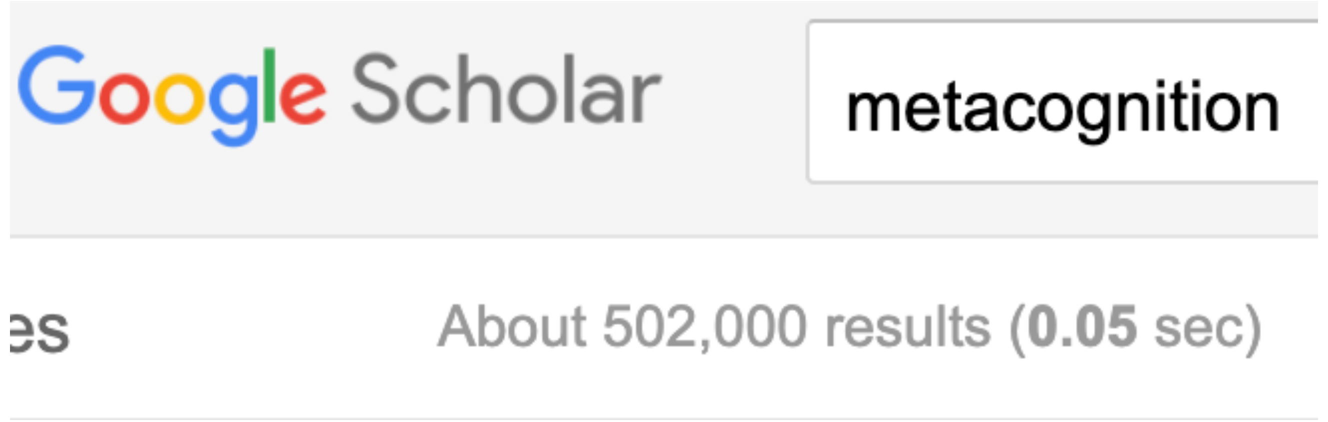
Metacognition = Humans' reasoning about their own thought processes

("Thinking about thinking")

Flavell, J. H. (1976): "Metacognitive Aspects of Problem Solving"
Brown, A. L. (1978): "Knowing When, Where, and How to Remember: A Problem of Metacognition"
Paris, S. G., & Winograd, P. (1990): "How Metacognition Can Promote Academic Learning and Instruction"

Metacognition

“Thinking about thinking”



[Flavell 1976]

"In any kind of cognitive transaction with the human or non-human environment, a variety of information processing activities may go on. Metacognition refers, among other things, to the active monitoring and consequent regulation and orchestration of these processes in relation to the cognitive objects or data on which they bear, usually in service of some concrete goal or objective."

(in a followup article in [Flavell 1979])…also stated that metacognitive knowledge is not fundamentally different than other knowledge, but its object is different. He also mentioned that metacognitive knowledge may be activated consciously or unconsciously by the individual. This question of consciousness later became a subject of controversy among researchers in metacognition.

Why Investigate Metacognition?

Thomas O. Nelson and Louis Narens

Metacognition is simultaneously a topic of interest in its own right and a bridge between areas, e.g., between decision making and memory, between learning and motivation, and between learning and cognitive development. Although the focus of this chapter is on the metacognitive aspects of learning and memory — which throughout the chapter will be called *metamemory* — both the overall approach and many of the points apply as well to other aspects of cognition.

SELF-REGULATED LEARNING INVENTORY 11/6/95

Not at all typical of me
Not very typical of me
Somewhat typical of me
Frequently typical of me
Almost always typical of me

1. Studying is a mysterious process. Sometimes what I do is successful, other times it is not. But in either case, I really don't know why. a b c d e

2. I come to each class session prepared to discuss the assigned reading material (e.g., chapter, handout, articles). a b c d e

3. Mastery of new knowledge or skills is more important to me than how well I do compared to others. a b c d e

4. If I am struggling to understand the material

Not at all typical of me
Not very typical of me
Somewhat typical of me
Frequently typical of me
Almost always typical of me

8. If I am having trouble understanding material as presented in a class or text, I try to locate and read different materials which help to explain or clarify the ideas with which I am having trouble.

9. After studying new information for a class, I pause and perform a mental review in order to determine how much of what I have read I am able to recall.

10. When reviewing my class notes, I try to identify the main points of a lecture by marking or highlighting them.

Question: How might study of human metacognition help develop LLMs?

LLM Metacognition

Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving

Aniket Didolkar [1], Anirudh Goyal [1], Nan Rosemary Ke [4], Siyuan Guo [3],
Michal Valko [4], Timothy Lillicrap [4], Danilo Rezende [4],
Yoshua Bengio [1], Michael Mozer [4], Sanjeev Arora [2]

Hope: (i) Understand whether LLMs possess ability to identify
and catalog skills needed to solve certain asks.
(ii) Given task questions, can they match questions to skills required by them?

Setting: GSM8K and MATH datasets

User: Consider this mathematical question. Label this question with any mathematical skills that would be required to solve the question. Basically, you should be able to use the skill as a dictionary key in python. The skill name should be lower case letters only. The skill name should be very descriptive and you may use multiple words to describe the skills required in the question. If you do use multiple words per question, then join them by an underscore.

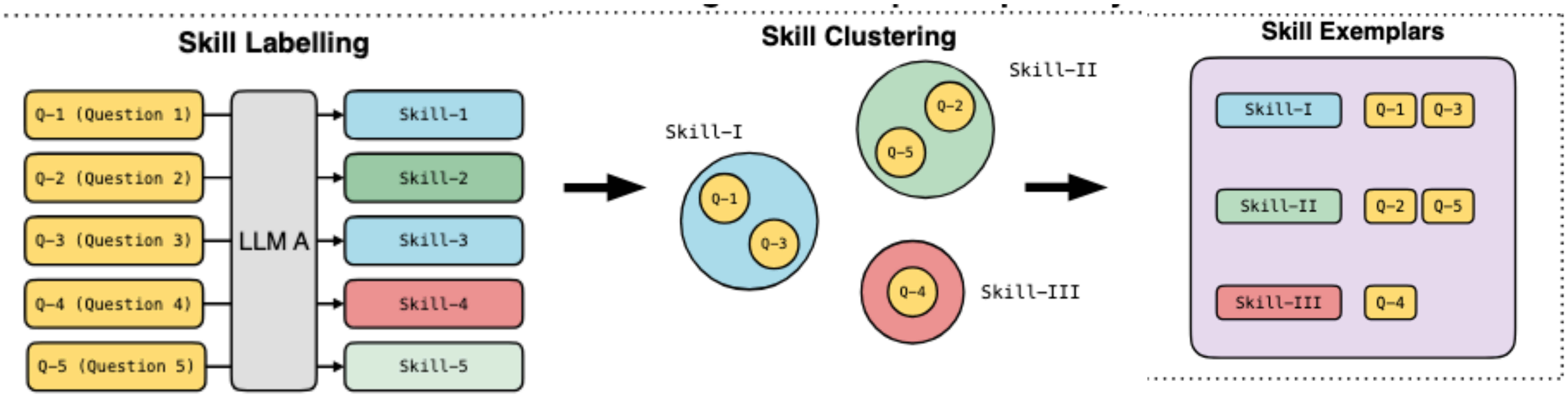
Kelly plans to fence in her yard. The Fabulous Fence Company charges $3.25 per foot of fencing and $15.75 an hour for labor. If Kelly needs 350 feet of fencing and the installers work a total of 6 hours installing the fence, how much will she owe the Fabulous Fence Company?

GPT4O: "arithmetic_operations", "unit_rate_calculation", "multiplication", "addition", "cost_estimation"

Claude Sonnet 3.5

1. addition
2. multiplication
3. decimal_arithmetic
4. cost_calculation
5. unit_rate_application

Skill-extraction pipeline



Skills are v. fine-grained!

Here is a list of skills required to solve a mathematical question. Reduce the number of unique skills by grouping similar skills into categories and give a descriptive name to each category.

(shortened version of real prompt) Here is a list of skills and a list of math questions. For each question give the list of skills it requires.

Step 2: Skill-clustering

Ask GPT-4 to partition the large # of named skills into clusters

Input Prompt:
Here is a list of skills required to solve a mathematical question:

[average_calculation, proportion_calculation, percentages, order_of_operations, subtraction, addition, multiplication, addition_of_fractions, permutations_combinations, counting, understanding_fractions_and_inequalities]

Reduce the number of unique skills by grouping similar skills into categories and give a descriptive name to each category.

GPT 4 Response:
1. basic_arithmetic_operations: addition, subtraction, multiplication

2.fractions_and_percentages: addition_of_fractions, understanding_fraction_and_inqualities, percentages

3. average_and_proportion: average_calculation, proportion_calculation

4. order_of_operations: order_of_operations

5: combinatorics_and_counting: counting, permutations_combinations

Skills extracted from MATH dataset

[Hendrycks et al'21]

Geometry	understanding_circle_properties_and_algebraic_manipulation, other_geometric_skills, pythagorean_skills, quadrilateral_and_polygon_skills, triangle_geometry_skills, calculus_skills, 3d_geometry_and_volume_calculation_skills, circle_geometry_skills, area_calculation_skills, coordinate_geometry_and_transformation_skills, ratio_and_proportion_skills, trigonometry_skills, combinatorics_and_probability_skills, algebraic_skills
Number Theory	base_conversion, prime_number_theory, greatest_common_divisor_calculations, modular_arithmetic, solving_equations, number_theory, factorization, division_and_remainders, exponentiation, sequence_analysis, arithmetic_sequences, basic_arithmetic, polynomial_operations, understanding_of_fractions, number_manipulation
Precalculus	matrix_operations, geometric_series_comprehension, basic_trigonometry, vector_operations, coordinate_systems, trigonometric_calculations, complex_numbers, geometric_relations, calculus, algebra_and_equations, three_dimensional_geometry, arithmetic_operations, parametric_equations, sequences_series_and_summation, geometry_triangle_properties, geometry_and_space_calculation, determinant_calculation, geometry_transforms, complex_number_operations
Probability	probability_calculation_with_replacement, combinatorics_knowledge, probability_theory_and_distribution, combinatorial_mathematics, counting_principals, permutation_and_combinations, probability_concepts_and_calculations, calculating_and_understanding_combinations, number_theory_and_arithmetic_operations, factorials_and_prime_factorization, understanding_and_applying_combinatorics_concepts

Interesting facts about extracted math skills

1. Names provided by LLM_1 are meaningful to LLM_2 (even when LLM_2 is much smaller)

2. Letting LLM_2 select in-context examples by identifying the most relevant skill in the list improves its accuracy by 3-5 % on usual math datasets. ("Here is a math question. Here is a list of 100 skills. Which skills does the question require?")

Usefulness of extracted math skills

Usual 4-SHOT prompting:

When testing a model on a MATH test question, give it 4 examples of questions from the same subfield of MATH (e.g., probability, algebra etc.)

"Skill-based prompting"

1. Use metacognition to extract skills from MATH dataset
2. Label each instance with one skill from the list ("exemplars")
3. When testing a new model on a MATH test question, present it with list of all skills and ask it to select a skill that is used in the test question. Give it 4 exemplars of that skill.

For most models (including frontier models) this increases performance by 3+ percent. Also improves over all known variants of prompting

Instruct-SkillMix: New pipeline for instruction-tuning

INSTRUCT-SKILLMIX: **A Powerful Pipeline for LLM Instruction Tuning**

Simran Kaur[1*] **Simon Park**[1*] **Anirudh Goyal**[2] **Sanjeev Arora**[1]

Background: Instruction-tuning

"*Human data is the last remaining moat of AI*" [Anonymous industry colleague]

Converts base model (next-word predictor) to chat agent

Proprietary method (expensive!): SFT base model using instruction-following examples obtained from skilled humans (aka "Imitation Learning")

Open efforts were uncompetitive (when starting from the same base model)

Human queries: Vicuna (70K), Dolly (35K) were collected from random humans

Fully synthetic: Alpaca [Taori et al'23]: 52K from a seed set of few hundred qs.

Ultrachat [Ding et al'23]:1M generated from GPT3.5 using careful attention to topics, lexical diversity etc.

INSTRUCT-SKILL-MIX

Using metacognition ideas, extract **list of skills** via interaction with GPT-4

Dataset generation: Pick **random pairs** of skills, and ask GPT-4 to generate a (query, answer) that uses both skills.

Fine-tune base models on this Q&A dataset

Alternate Version of method: Extract skillset using 6K Q&A examples from Alpaca and Ultrachat

Benefit of random pairs of skills: diverse queries

- critical_thinking_and_communication, literature_and_language_skills

I'm a high school English teacher aiming to develop a curriculum unit for my 11th-grade class, focusing on American literature. I want this unit to go beyond just reading and understanding the texts. Specifically, I'm looking to enhance my students' critical thinking and communication skills through engaging activities related to the literature. Can you suggest detailed ways to incorporate these skills, ideally with concrete examples and expected learning outcomes?

- critical_thinking_and_communication, skill_in_virtual_and_system_design

As an IT manager, I am overseeing the development of a virtual workspace to enhance communication and efficiency among remote teams. This workspace must support multimedia content, including video conferencing and live document editing. What are the critical steps I should take in its design and implementation, balancing technical robustness with ease of use? Could you provide specific technologies to consider and any potential obstacles?

INSTRUCT-SKILLMIX: Performance

Proprietary small models

vs

Ours (SFT on 2K or 4K examples)

Proprietary models from last year

Model	# Data	AlpacaEval2.0 LC WR(%)	WildBench WB-Reward$^{gpt4t}_{\infty}$
LLaMA-3-8B			
Ours	4K	**42.76**	**-36.91**
*LLaMA-3-8B-Instruct	-	22.90	-46.30
Mistral-7B-v0.2			
Ours	4K	**36.70**	**-29.25**
SFT on Alpaca-52K	52K	8.64	-80.47
*Mistral-7B-Instruct-v0.2	-	17.10	-54.70
Gemma-2-9B			
Ours	2K	36.18	-37.83
Gemma-2-9B-Instruct	-	**37.21**	**-28.78**
***Other Proprietary Models**			
LLaMA-3.1-405B-Instruct	-	39.30	-
Mistral Large	-	32.70	-46.40
Claude 3 Opus	-	40.50	-21.20
Claude 3 Sonnet	-	34.90	-30.30
GPT-4-Omni (2024-05-13)	-	**57.50**	**+1.70**
GPT-4 (2023-03-14)	-	35.30	-

LC WR = Length-controlled win-rate against GPT4-Turbo (with GPT4 as judge)

Ablations to understand efficacy of Instruct-Skillmix

Benefit of Extracting Skills

Here we use the weaker version of our method, which extracts skills from Alpaca

We compare it to direct SFT on Alpaca

SFT Dataset	# Data	AlpacaEval 2.0 LC WR(%)	MT-Bench	WildBench WB-Reward$_{\infty}^{\text{gpt4t}}$
INSTRUCT-SKILLMIX-D(k=2)	4K	**29.77**	7.17	**-39.06**
INSTRUCT-SKILLMIX-D(k=1)	1K	27.04	**7.22**	-46.83
Alpaca-1K Longest	1K	10.09	6.88	-63.38
Alpaca-5K Longest	5K	8.92	6.90	-62.55
Alpaca-5K Random	5K	11.10	6.86	-74.41
Alpaca-52K Full	52K	8.64	6.45	-80.47

Table 2: **Evaluation results of Mistral-7B-Base-v0.2 finetuned on INSTRUCT-SKILLMIX-D vs. on Alpaca-52K.** Note that skills extracted from Alpaca-5K Random were used to create the INSTRUCT-SKILLMIX-D datasets.

Influence of grader

AlpacaEval and WildBench use GPT4 (3/23) as grader.

Our dataset (skills, questions, answers) is generated using GPT4-Turbo

Could our high performance be due to similarity of grader and generator?

Influence of generator

Model for Data Generation	Dataset	AlpacaEval 2.0 WR(%)	AlpacaEval 2.0 LC WR(%)	MT-Bench
GPT-4 (2023-03-14)	Alpaca-1K Longest	12.75	10.09	6.83
	INSTRUCT-SKILLMIX-D-1K	13.29	15.01	7.10
GPT-4-Turbo (2024-04-09)	Alpaca-1K Longest	35.23	19.62	6.99
	INSTRUCT-SKILLMIX-D-1K	33.87	27.48	6.92

Influence of grader (AlpacaEval 2.0)

Model	Grader: GPT-4 (2023-11-06) WR(%)	Grader: GPT-4 (2023-11-06) LC WR(%)	Grader: Claude 3 Opus WR(%)	Grader: Claude 3 Opus LC WR(%)
Mistral-7B-Base-v0.2 SFT on ISM-D-1K	33.87	27.48	50.56	38.50
Mistral-7B-Base-v0.2 SFT on ISM-D-2K	37.05	31.57	48.94	38.29
Mistral-7B-Base-v0.2 SFT on ISM-D-4K	35.08	29.77	52.55	44.16
(Reference Model) LLaMA-3-70B-Instruct	33.20	34.40	39.68	42.33
(Reference Model) Mistral-7B-Instruct-v0.2	14.70	17.10	15.16	18.89
(Reference Model) LLaMA-2-70B-Chat	13.90	14.70	16.67	17.85

Instruct-Skill-Mix (k=1 vs k=2)

Earlier, using pairs of skills used to give stronger performance but with the best pipeline the difference between k=1 and k=2 is under 1%

Effect of 20% low-quality data (“shirkers”)

Model	# Data	AlpacaEval 2.0 LC WR(%)	AlpacaEval 2.0 Avg Len	MT-Bench	WildBench WB-Reward$_{\infty}^{\text{gpt4t}}$
SFT on INSTRUCT-SKILLMIX-D(k=2)					
Mistral-7B-Base-v0.2	2K	31.57	2817	7.04	-43.46
	2K (Brevity 20%)	23.93	1746	6.69	-49.85
	2K (Junk 20%)	0.77	1104	5.01	-47.50
SFT on INSTRUCT-SKILLMIX(k=2)					
Mistral-7B-Base-v0.2	2K	36.18	2936	7.20	-31.92
	2K (Brevity 20%)	31.61	2336	7.32	-32.27
	2K (Junk 20%)	24.60	2435	6.90	-47.50

“Brevity” = Answer is required to be short (hence less helpful)

“Junk” = Deliberately unhelpful

20% low-quality data can seriously harm instruction-tuning (explains why prior open efforts had not succeeded at SFT?)

Can AI assist with generation of difficult math questions?

AI-Assisted Generation of Difficult Math Questions

Vedant Shah[*1,2] **Dingli Yu**[3] **Kaifeng Lyu**[3] **Simon Park**[3] **Yinghui He** **Jiatong Yu**
Nan Rosemary Ke[1] **Michael Mozer**[4] **Yoshua Bengio**[1,2]
Sanjeev Arora[3] **Anirudh Goyal**[*1]

Evaluation saturation phenomenon

Evaluations get saturated within a year or two

Reasons: (i) models getting powerful

(ii) changes to training pipelines to improve on the popular evaluations

(iii) including training on questions similar to those on popular evals

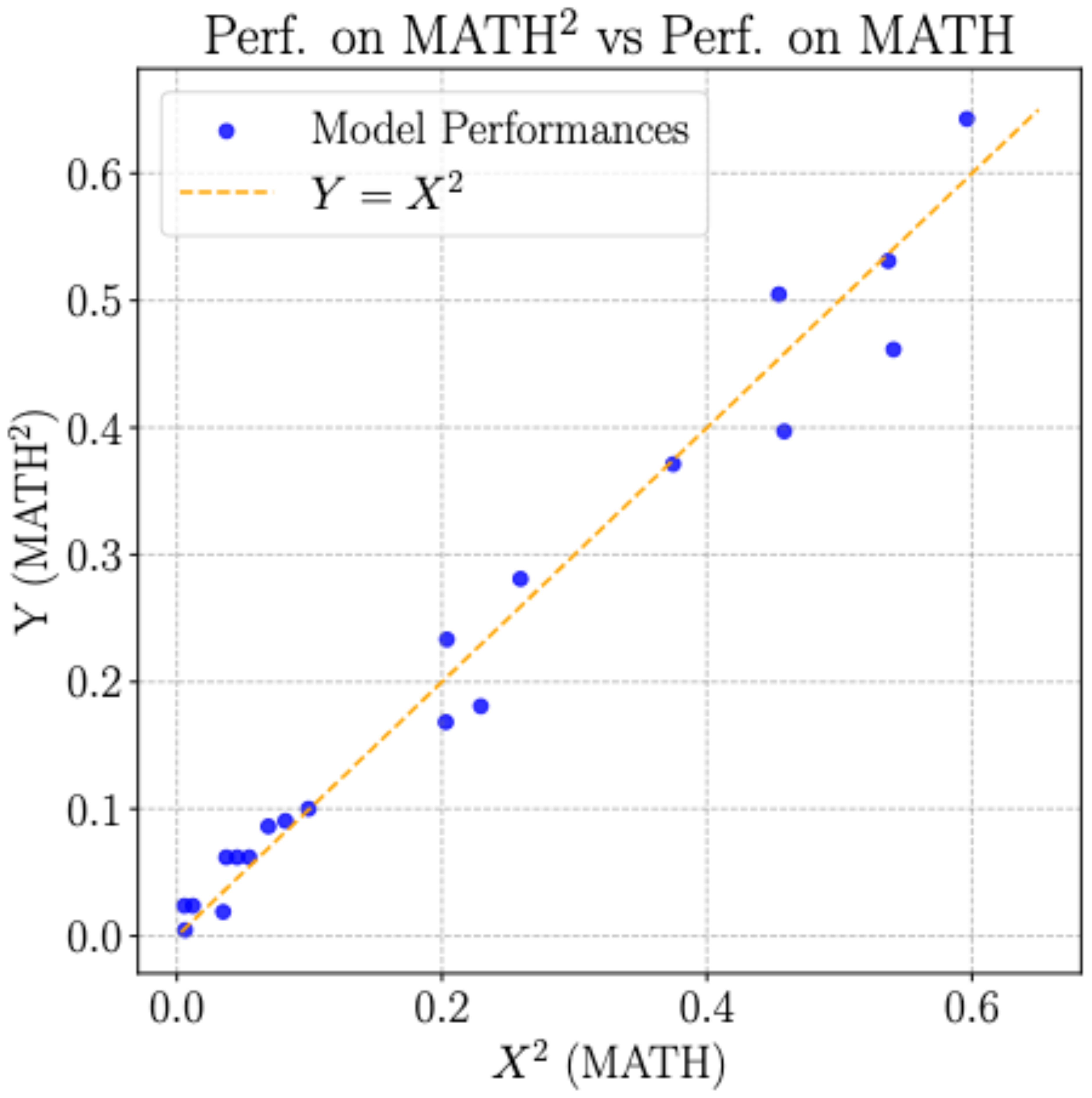
Design of new math evaluation

1. Start with list of skills extracted from MATH
2. Present GPT4O with a random pair of skills, and ask it to produce a question that uses both skills
3. Use agentic interaction with another copy of GPT4 to make sure it has (a) used each skill (b) question makes sense and (c) model has double-checked its answer to the question.
4. Ask humans to look over the list and select the 25% best questions. If needed, edit question and/or answer

End result: MATH^2 dataset with 225 questions

Old eval (MATH) vs New ($MATH^2$)

Model	$MATH^2$ (Y)	MATH (X)	% Drop
GPT-4 Omni	64.29%	77.21%	**16.73%**
Claude 3.5 Sonnet	46.15%	73.54%	37.24%
GPT-4 Turbo	53.11%	73.27%	27.51%
Gemini-1.5-Pro	39.71%	67.70%	41.34%
Claude 3 Opus	37.14%	61.20 %	39.31%
Llama-3.1-70B-Instruct	50.48%	67.40%	25.10%
Llama-3-70B-Instruct	18.09%	47.89%	62.23%
MetaMath-70B	8.61%	26.27%	67.22%
MAmmoTH-70B	6.19%	19.31%	67.94%
Mixtral-8×7B-Instruct	10.00%	31.52%	68.27%
MetaMath-13B	6.19%	21.32%	70.96%
MAmmoTH-13B	2.38%	10.99%	78.34%
Deepseek-math-7b-instruct	16.83%	45.05%	62.64%
Llama-3.1-8B-Instruct	28.09%	50.92%	44.83%
Llama-3-8B-Instruct	9.05%	28.62%	68.38%
Gemma-1.1-7B-Instruct	6.19%	23.36%	73.50%
MetaMath-7B	1.91%	18.69%	89.78%
MAmmoTH-7B	0.48%	7.90%	**93.92%**
Phi-3-mini-128k-instruct	23.34%	48.29%	51.67%
Gemma-1.1-2B-Instruct	2.38%	7.52%	68.35%



What applications of LLM metacognition can you think of?

Have a good Fall break !