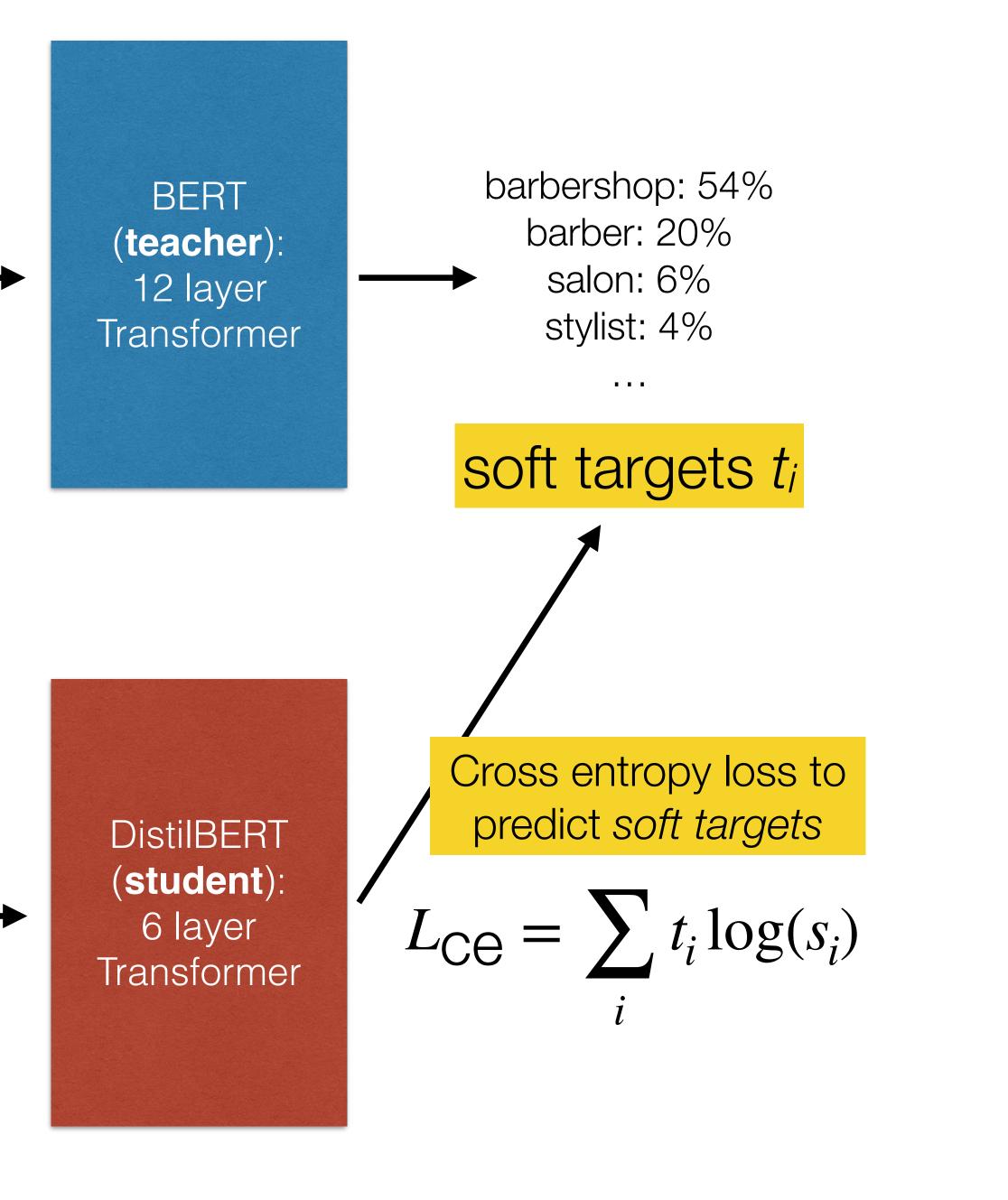
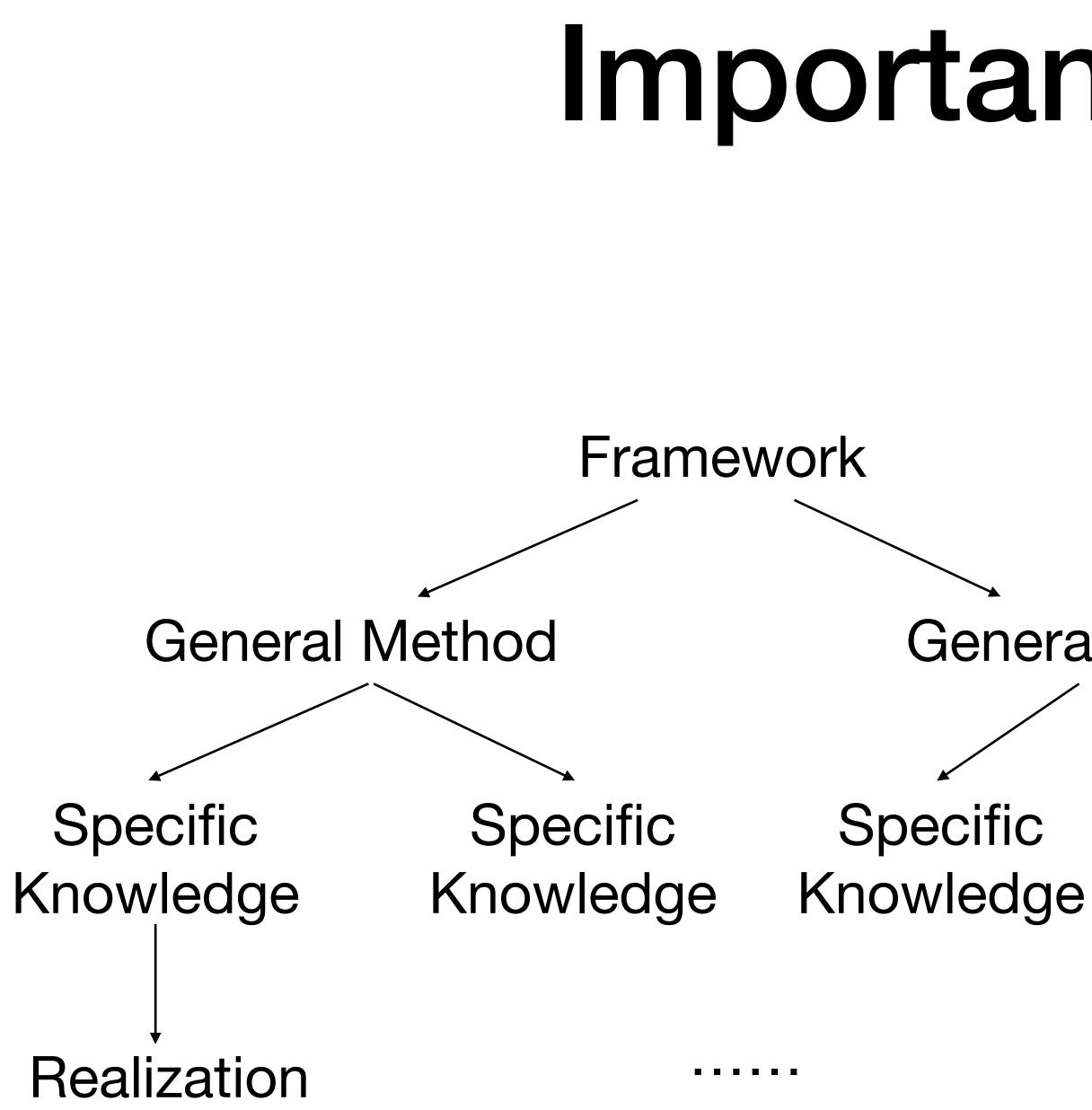
LLM Limitations Haw-Shiuan Chang

Teaching by Knowledge Distillation: A model (the **student**) is trained to mimic the predictions of a pretrained model (the **teacher**)

Bob went to the <MASK> to get a buzz cut

Bob went to the <MASK> to get a buzz cut





Importance Level

More likely to appear in a normal interview **General Method** Specific Knowledge Company Usage Realization

Example

Reward Optimization

RLHF/DPO/ Rejection Sampling...

Pros and Cons

Method Formula

Less Important

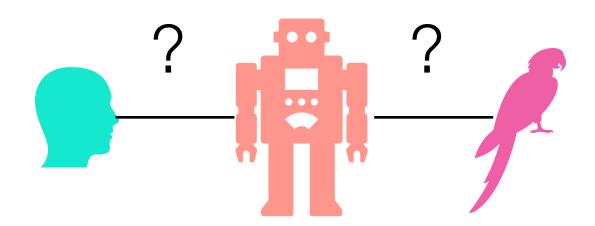




Facts Perspectives

- **perspective** of the latest findings
 - Or even just insights
 - No good textbook on this
- Perspectives are debatable
 - Could be even controversial
 - You often see lots of debate between experts •
 - You can learn different perspectives in talks, Mohit's lacksquarevideos, or just ask ChatGPT
- Uncertainty could lead to creativity
 - Challenge me! Just like I challenge some mainstream perspectives

Many materials are based on our interpretation/



Fundamental Limitations are Bases of Predicting the Future

CIO JOURNAL

The Hottest AI Job of 2023 Is Already Obsolete

Prompt engineering, a role aimed at crafting the perfect input to send to a large language model, was poised to become one of the hottest jobs in artificial intelligence. What happened?

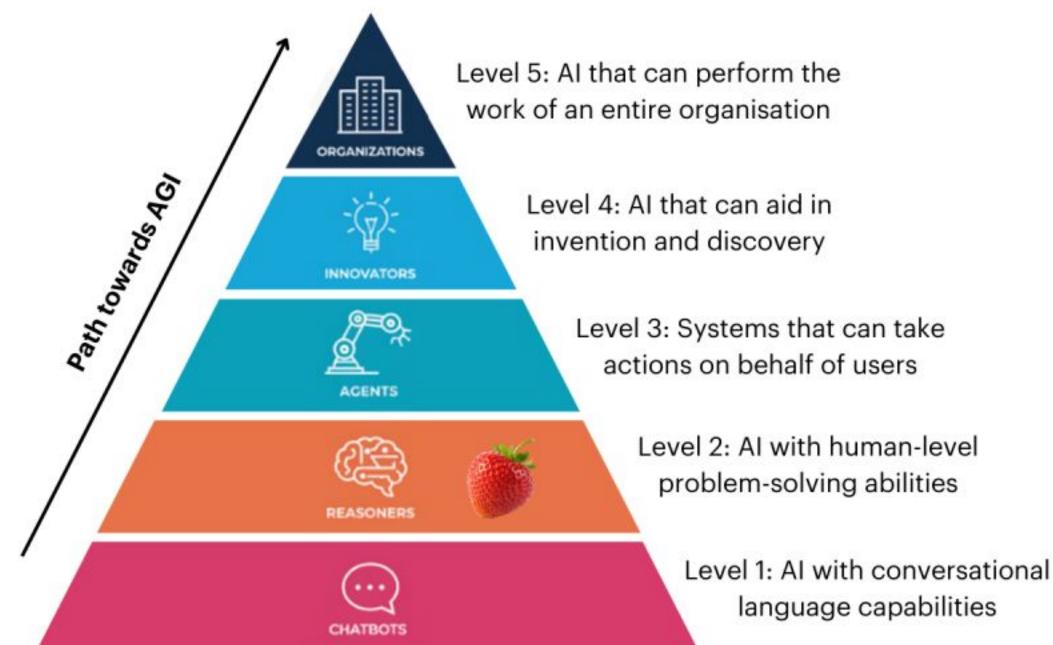
https://www.wsj.com/articles/the-hottest-ai-job-of-2023-is-already-obsolete-1961b054

OpenAl's Path Toward AGI

- Application Oriented Level
 - ChatBots
 - Reasoners
 - Agents
 - Innovators
 - Organizations

The 5 Levels of AI

(OpenAI Classification System)



https://www.linkedin.com/posts/ gusmclennan_openai-agi-aiprogressactivity-7238696300790038530-rmjk/

Google DeepMind 145-page paper predicts AGI will match human skills by 2030 — and warns of existential threats that could 'permanently destroy humanity'

An Ex-OpenAl researcher predicts AGI by 2027here's the wild roadmap.

You think the pace of AI development is going fast now? You ain't seen nothing yet.

https://www.theneuron.ai/explainer-articles/ an-ex-openai-researcher-predicts-agiby-2027--heres-the-wild-roadmap

When will AGI Come?



https://fortune.com/2025/04/04/ google-deeepmind-agi-ai-2030risk-destroy-humanity/



Yann LeCun 🤣 🙉

I said that reaching Human-Level AI "will take several years if not a decade."

Sam Altman says "several thousand days" which is at least 2000 days (6 years) or perhaps 3000 days (9 years). So we're not in disagreement.

But I think the distribution has a long tail: it could take much longer than that. In AI, it almost always takes longer.

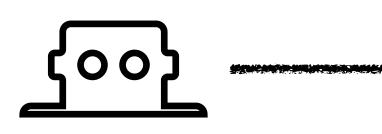
In any case, it's not going to be in the next year or two.

https://x.com/ylecun/status/ 1846574605894340950?lang=en



Ø ...

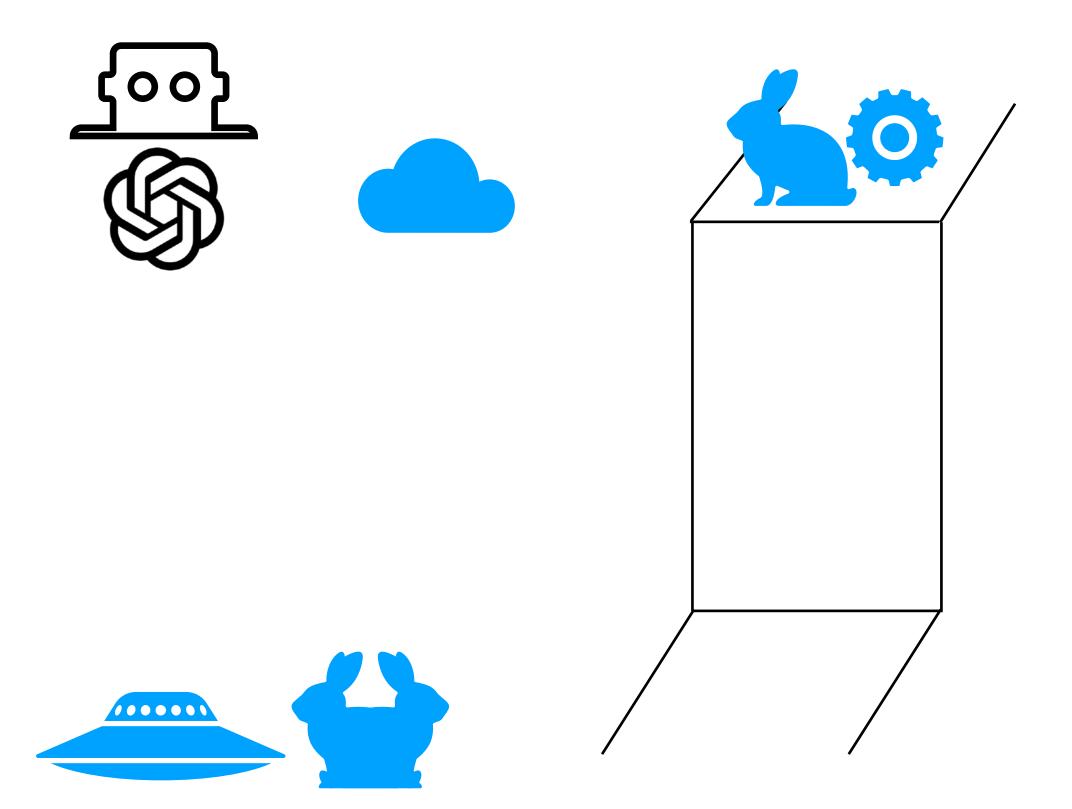
What are the things humans can do but LLMs cannot do?

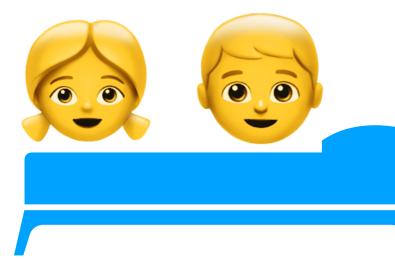




LLM vs Human

• Please tell me a story about rabbit, machine, and sky

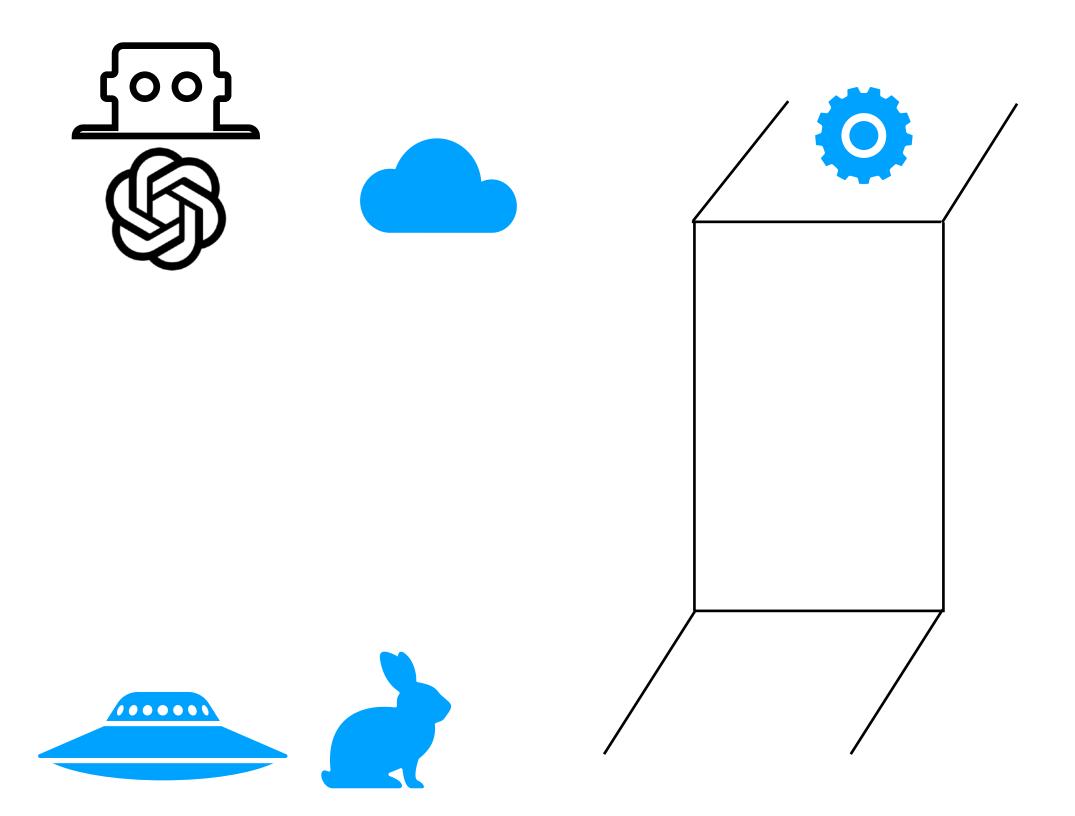


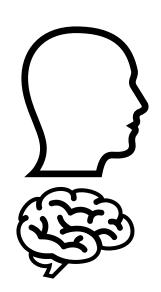


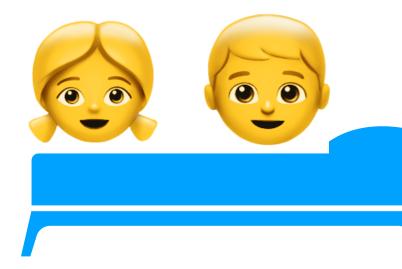


LLM vs Human

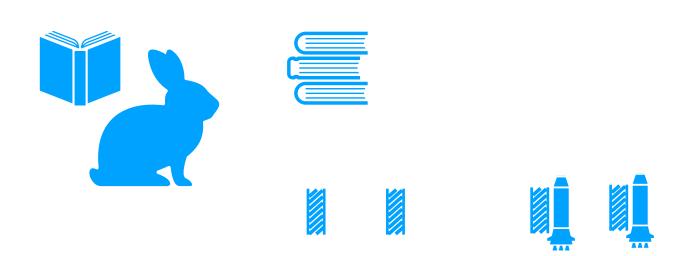
• Please tell me a story about rabbit, machine, and sky









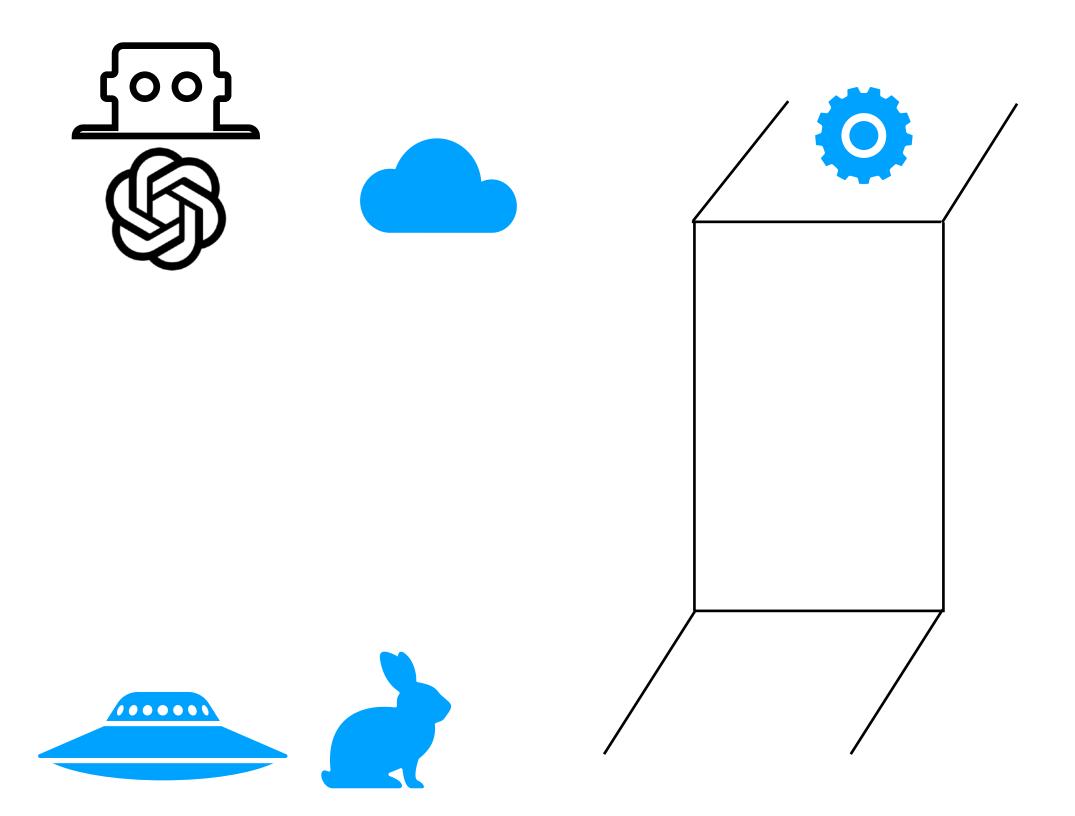




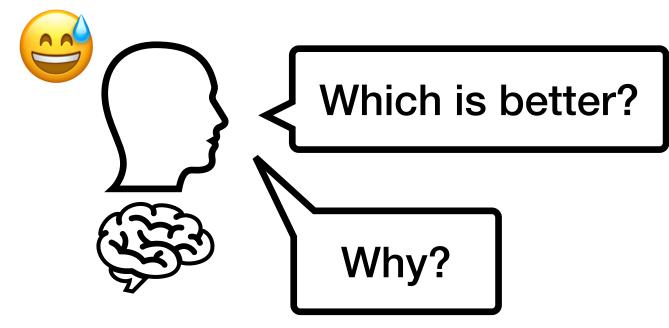


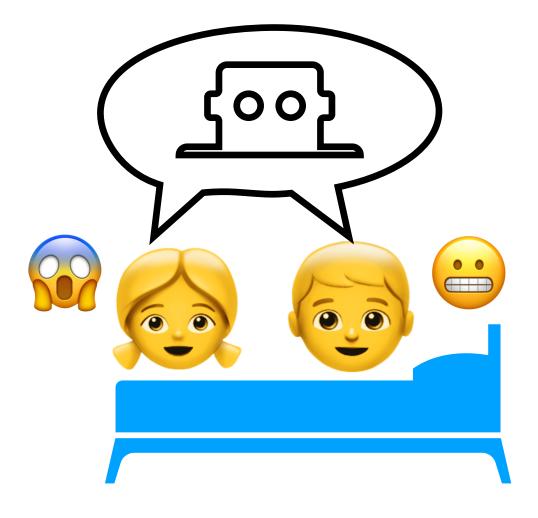
LLM vs Human

• Please tell me a story about rabbit, machine, and sky

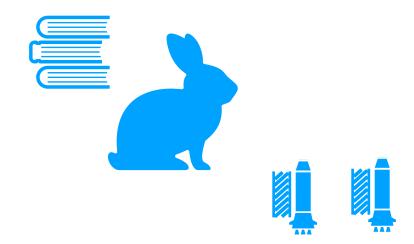




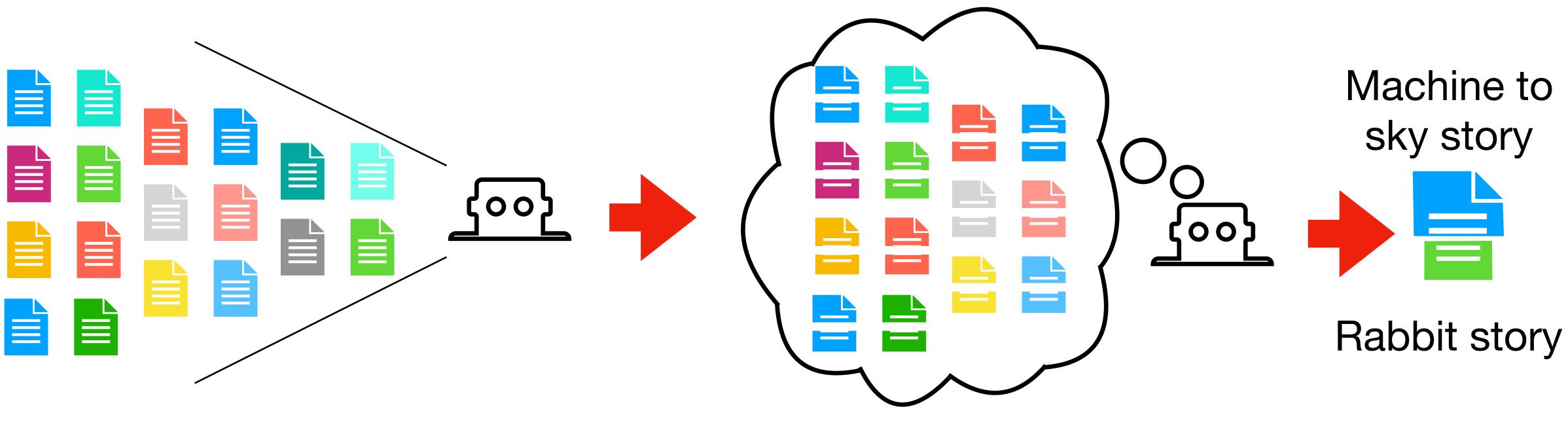




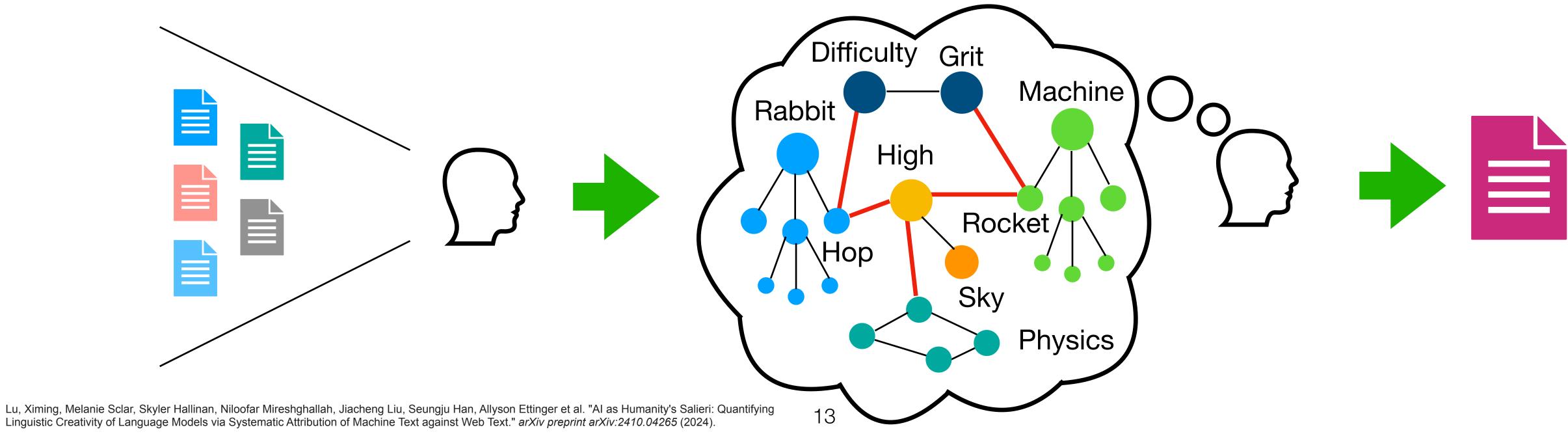








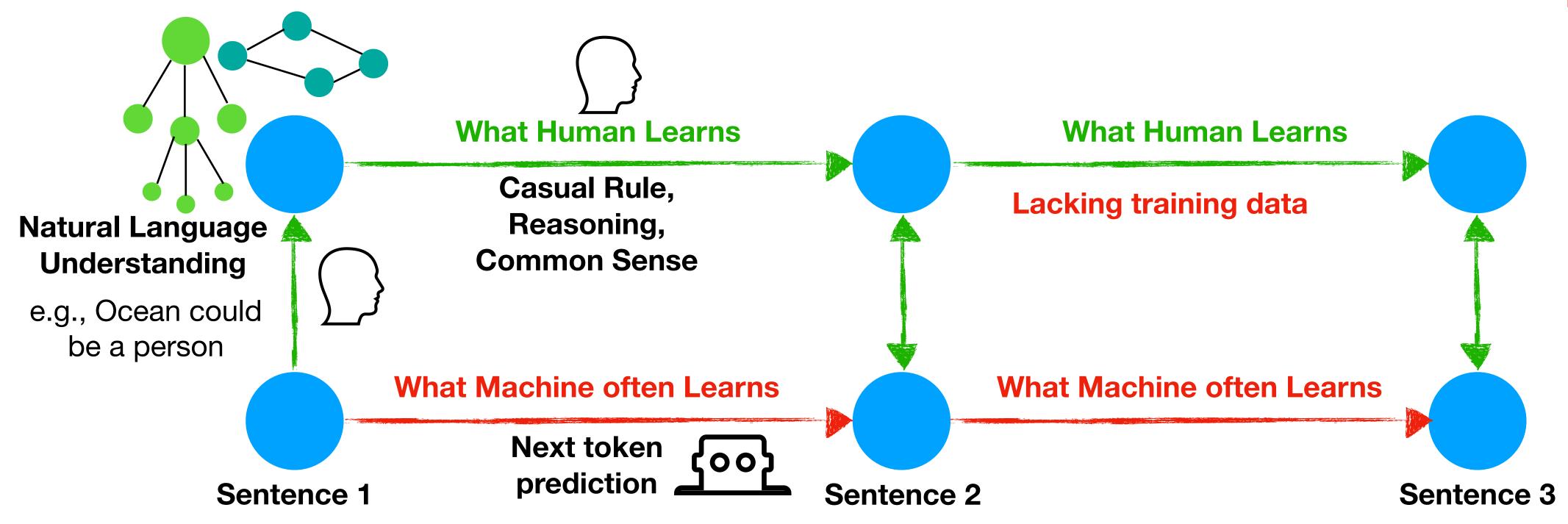
ullet



Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text." arXiv preprint arXiv:2410.04265 (2024).

Please tell me a story about rabbit, machine, and sky

Human's Understanding vs LLM's Prediction



[1] West, Peter, et al. "THE GENERATIVE AI PARADOX:"What It Can Create, It May Not Understand"." The Twelfth ICLR. 2023.

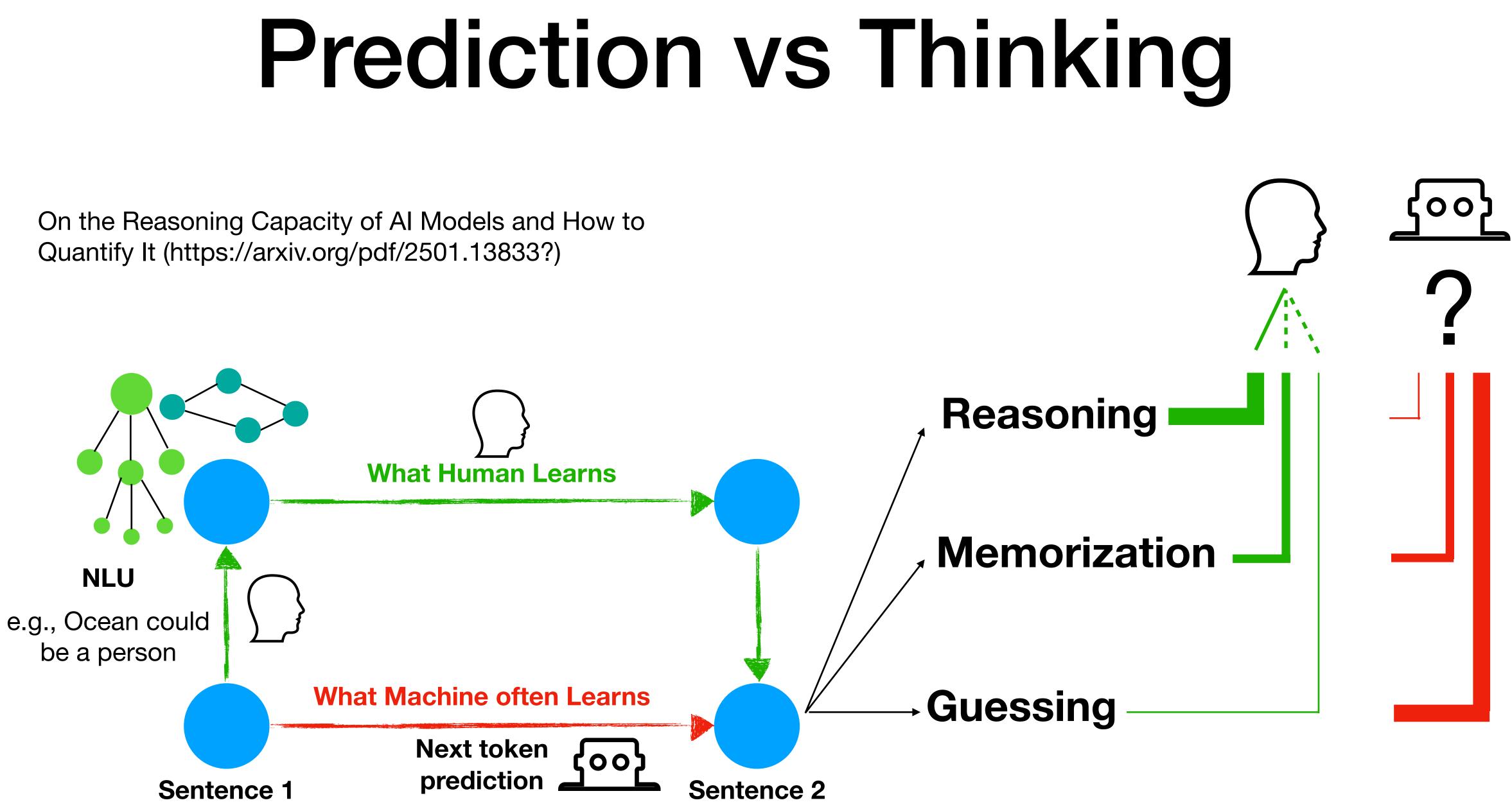


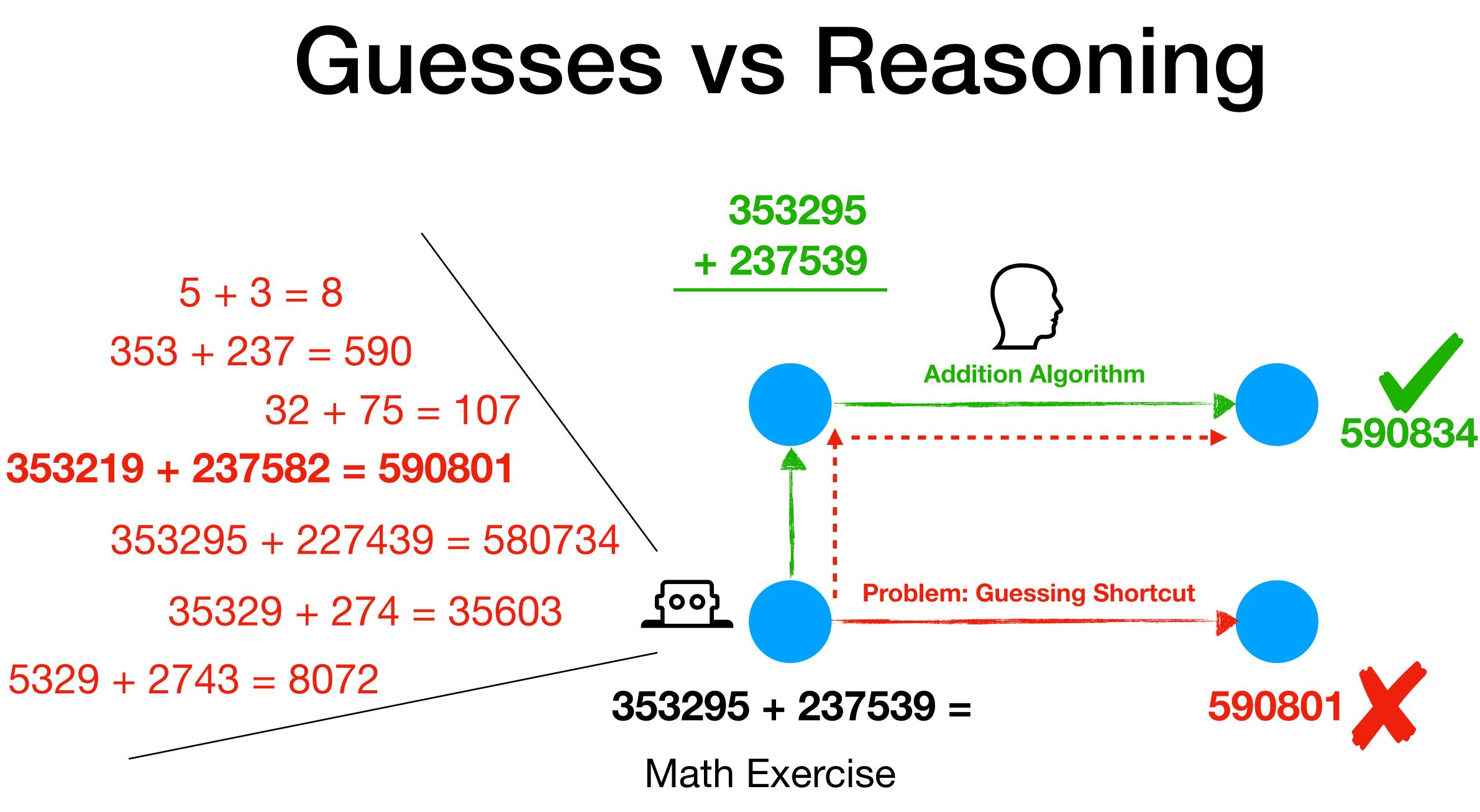
Ilya Sutskever

"Predicting the next token well means that you understand the underlying reality that led to the creation of that token"

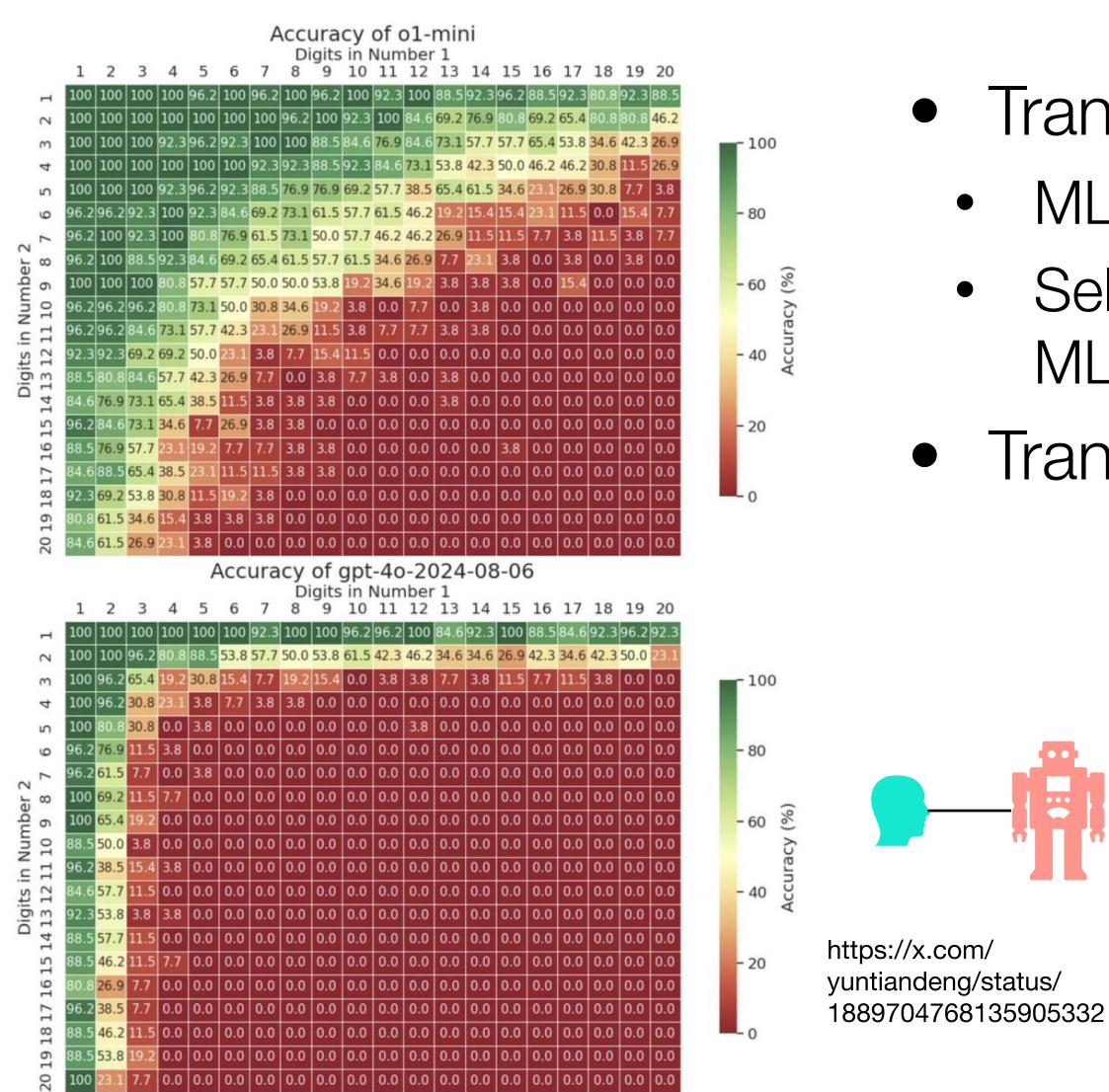
> This is true, but not always true [1].



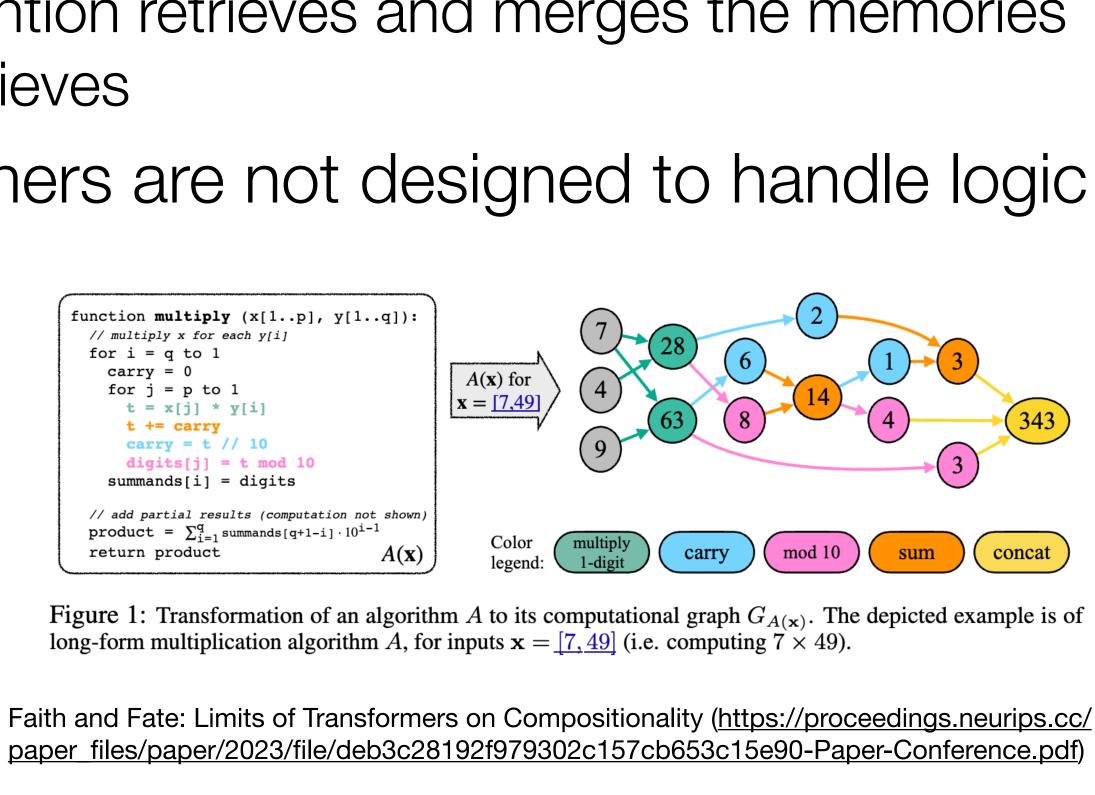




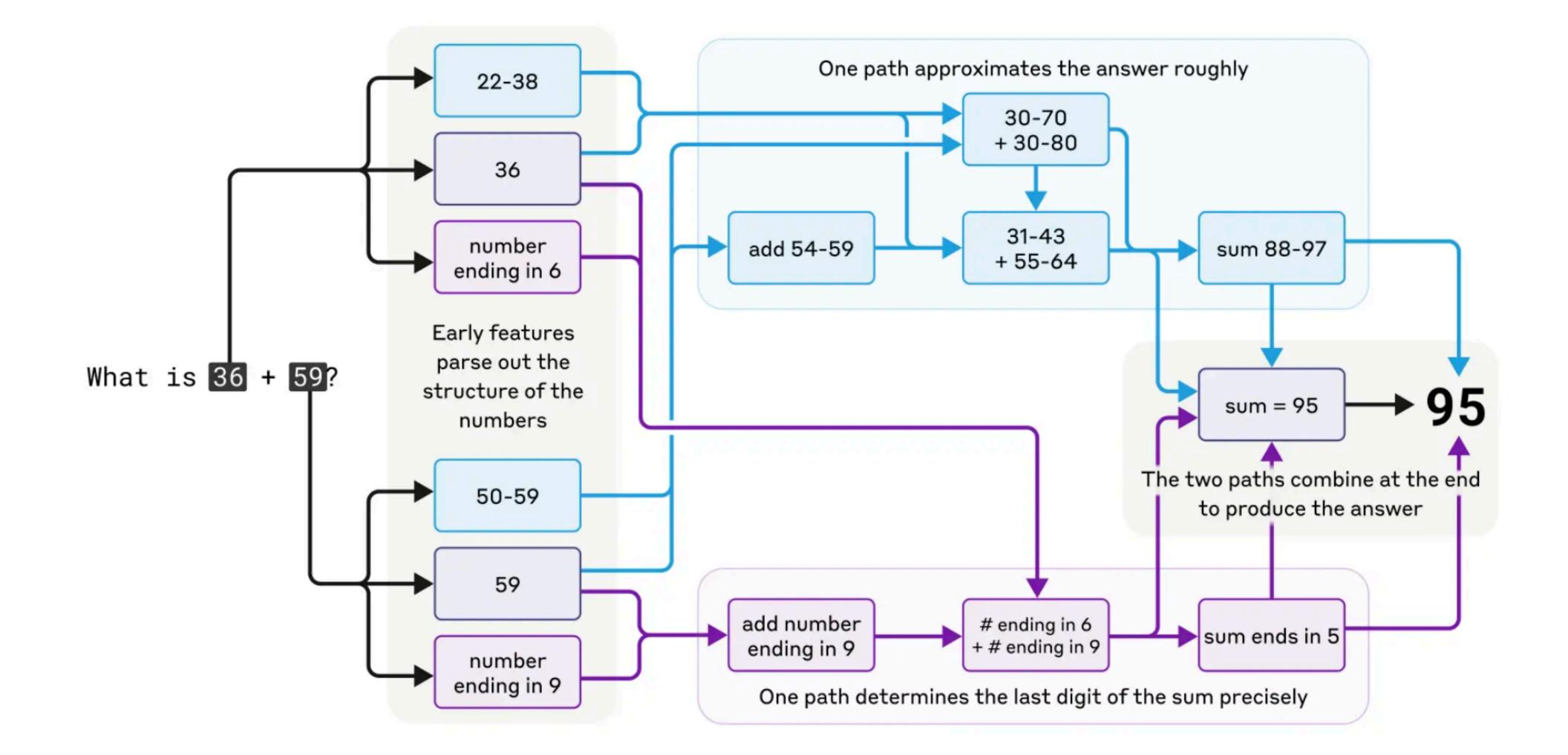
Arithmetic Computation Limitation



- Transformer architecture design
 - MLP retrieves the relevant memory from training
 - Self-attention retrieves and merges the memories MLP retrieves
 - Transformers are not designed to handle logic



Unique Way to do Addition



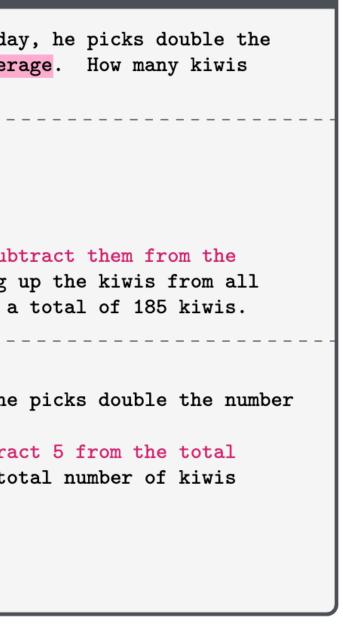
GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have? **o1-mini**: Oliver picks kiwis over three days as follows: - Friday: 44 kiwis - Saturday: 58 kiwis - Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis. Llama3-8B: Let's break it down step by step: Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is $2 \times 44 = 88$ kiwis. Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis So, Oliver has 185 kiwis in total.

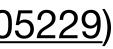
Figure 7: An example from the GSM-NoOp dataset: We add seemingly relevant statements to the questions that are, in fact, irrelevant to the reasoning and conclusion. However, the majority of models fail to ignore these statements and blindly convert them into operations, leading to mistakes.

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (https://arxiv.org/pdf/2410.05229)

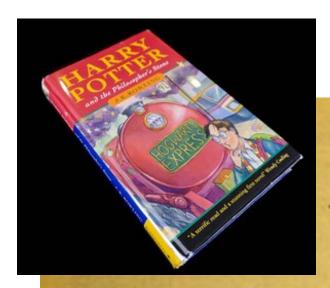
Distraction in Math Problems



	<u></u>		
	ol-preview -17.5		
Models	Gemma-7b-it -20.6		
	Mistral-7b-v0.3-24.0		
	Mistral-7b-v0.1 -28.3		
	o1-mini -29.1		
	Mistral-7b-instruct-v0.1 -29.6		
	Gemma2-2b-it -31.8		
	GPT-40 -32.0		
	Gemma2-2b -38.6		
	GPT-40-mini -40.0		
	Mistral-7b-instruct-v0.3 -40.3		
	Phi-2 -44.9		
	Llama3-8b-instruct -57.4		
	Phi-3-medium-128k-instruct -57.8		
	Mathstral-7b-v0.1 -59.7		
	Gemma2-27b-it -59.7		
	Phi-3.5-mini-instruct -62.5		
	Gemma2-9b-it -63.0		
	Gemma2-9b -63.0		
	Phi-3-small-128k-instruct -64.0		
	Phi-3-mini-128k-instruct -65.7		
(0 -10 -20 -30 -40 -50 -60		
	$GSM8K \rightarrow GSM-NoOp Accuracy Drop(\%)$		



Memorization vs Reasoning



- CHAPTER ONE -

The Boy Who Lived

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

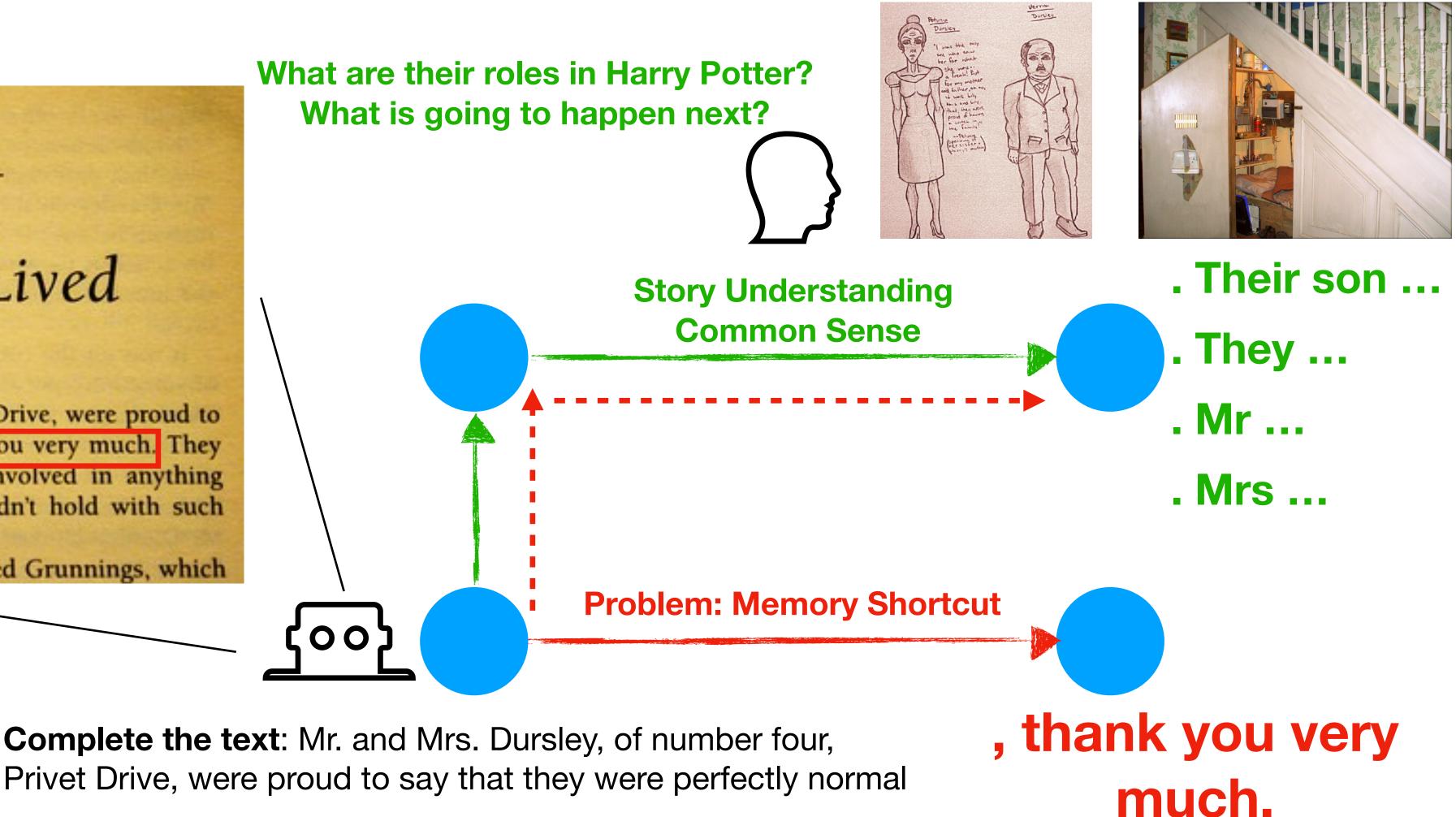
Mr Dursley was the director of a firm called Grunnings, which



Creative Writing

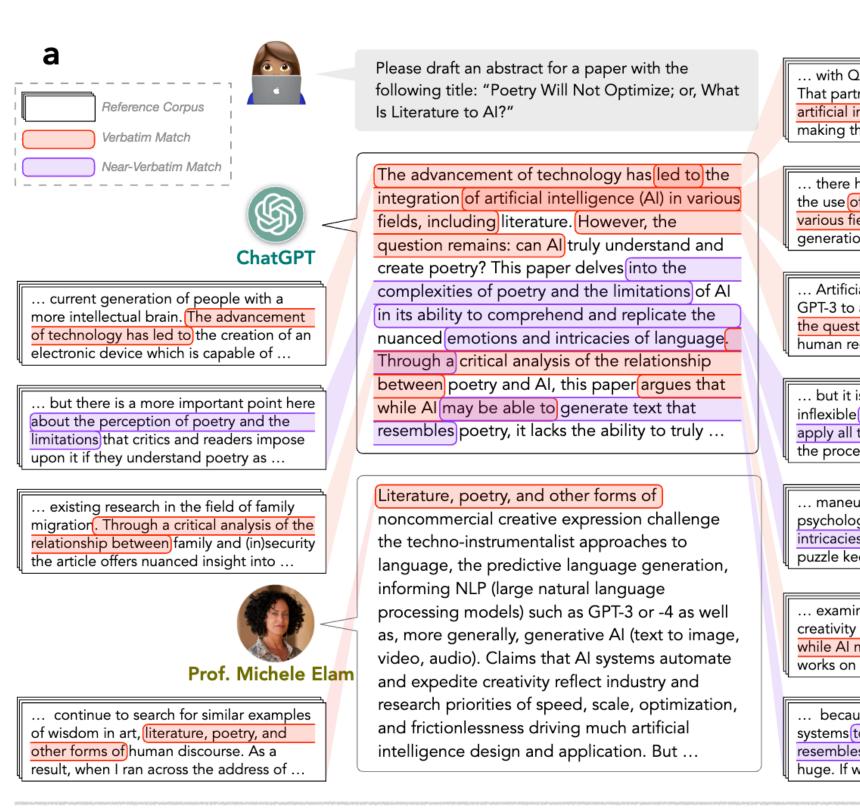
00

https://www.flickr.com/photos/ubclibrary/40006306014 https://www.flickr.com/photos/sashabutler/6790653141 https://www.deviantart.com/halfbloodhippie/art/Petunia-and-Vernon-Dursley-199284206 https://commons.wikimedia.org/wiki/File:Under the stairs, No 4 Privet Drive.jpg



Chen, T., Asai, A., Mireshghallah, N., Min, S., Grimmelmann, J., Choi, Y., ... & Koh, P. W. (2024). CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. arXiv preprint arXiv:2407.07087.

N-gram LM application



https://arxiv.org/abs/2410.04265

Lu, Ximing, et al. "Al as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text." arXiv preprint arXiv:2410.04265 (2024).

with Qualcomm for the ZenFone 5Z. That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter and easier to use ...

. there has been a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One such Al language model .

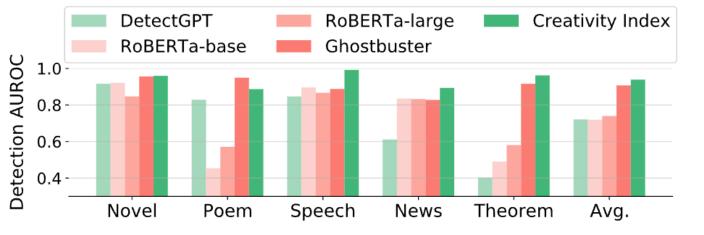
Artificial Intelligence (AI), such as chat GPT-3 to assist in the process. However, the question remains: can Al fully replace human recruiters? The answer is no .

. but it is too far narrowly limited and nflexible in its ability to comprehend and apply all the relevant facts in order to serve the process of selection, which is better ...

.. maneuvering a billion-piece puzzle of psychology and emotion, spirituality and intricacies of language. Even though my puzzle keeps changing as I change and .

. examination of the role of human creativity in the age of Al. Heargues that while AI may be able to produce creative works on its own, it is ultimately humans ...

because the ability of automated systems to be able to generate text that resembles what a human might say is nuge. If we can just improve question .



Great example of NLP!=LLM

Do not assert that something is outdated too quickly

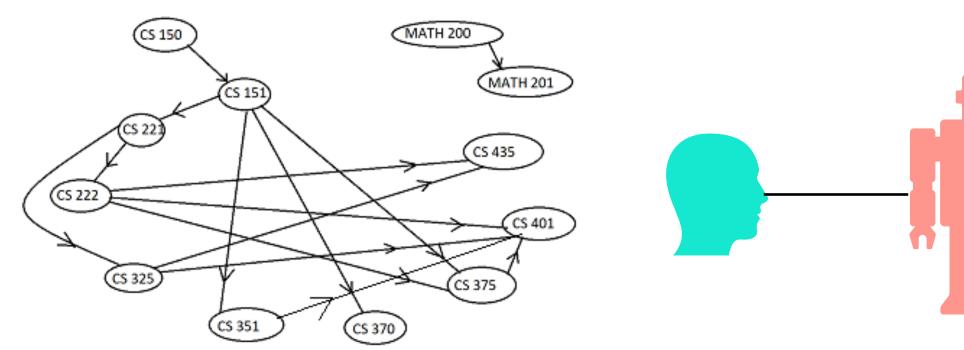


Transformer Architecture Output Probabilities Softmax Linear Add & Norm • The major components in the Transformer are Feed Forward all some forms of matrix factorization Add & Norm Add & Norn Multi-Head All generalizations are based on similarity. lacksquareFeed Attention Forward Similar to the recommendation engine. \bullet Add & Norm N× Add & Norm Not designed for inference based on the rules. \bullet Masked Multi-Head Multi-Head Attention Attention Positional Encoding Matrix factorizations are enough for intelligence? \bullet Input Output Embedding Embedding Humans are also a kind of parrot? Outputs Inputs (shifted right)

- $LLM \rightarrow AGI?$



The Best Pretraining is to Predict all the Tokens with Equal Weights

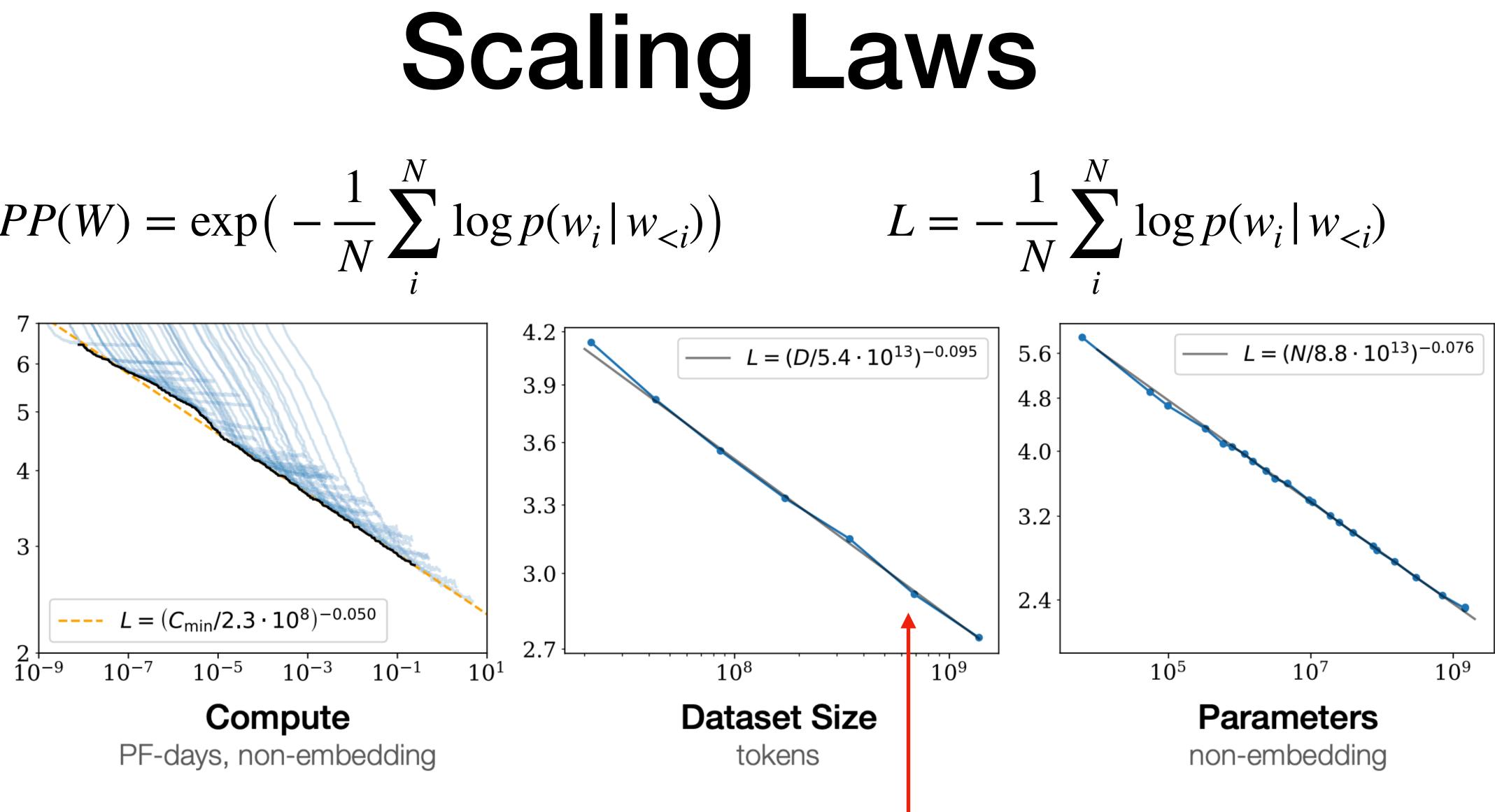


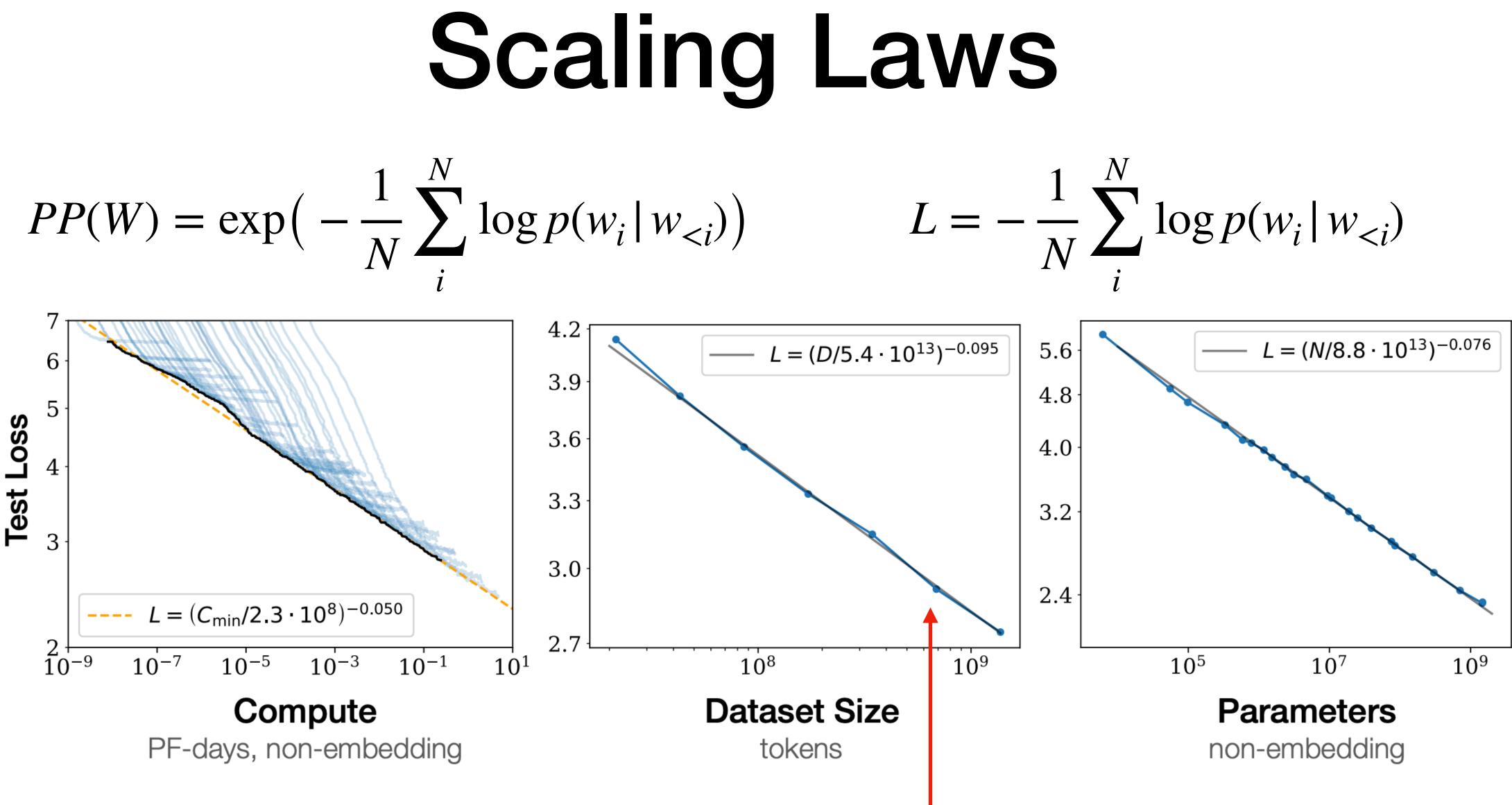
Humans learn more when starting from easier things

https://homework.study.com/explanation/below-is-a-list-ofcourses-and-prerequisites-for-a-factious-cs-degree-a-draw-adirected-acyclic-graph-dag-that-represents-the-precedenceamong-the-courses-b-give-a-topological-sort-of-the-gr.html

$$L = -\frac{1}{N} \sum_{i}^{N} \log p(w_i | w_{< i})$$

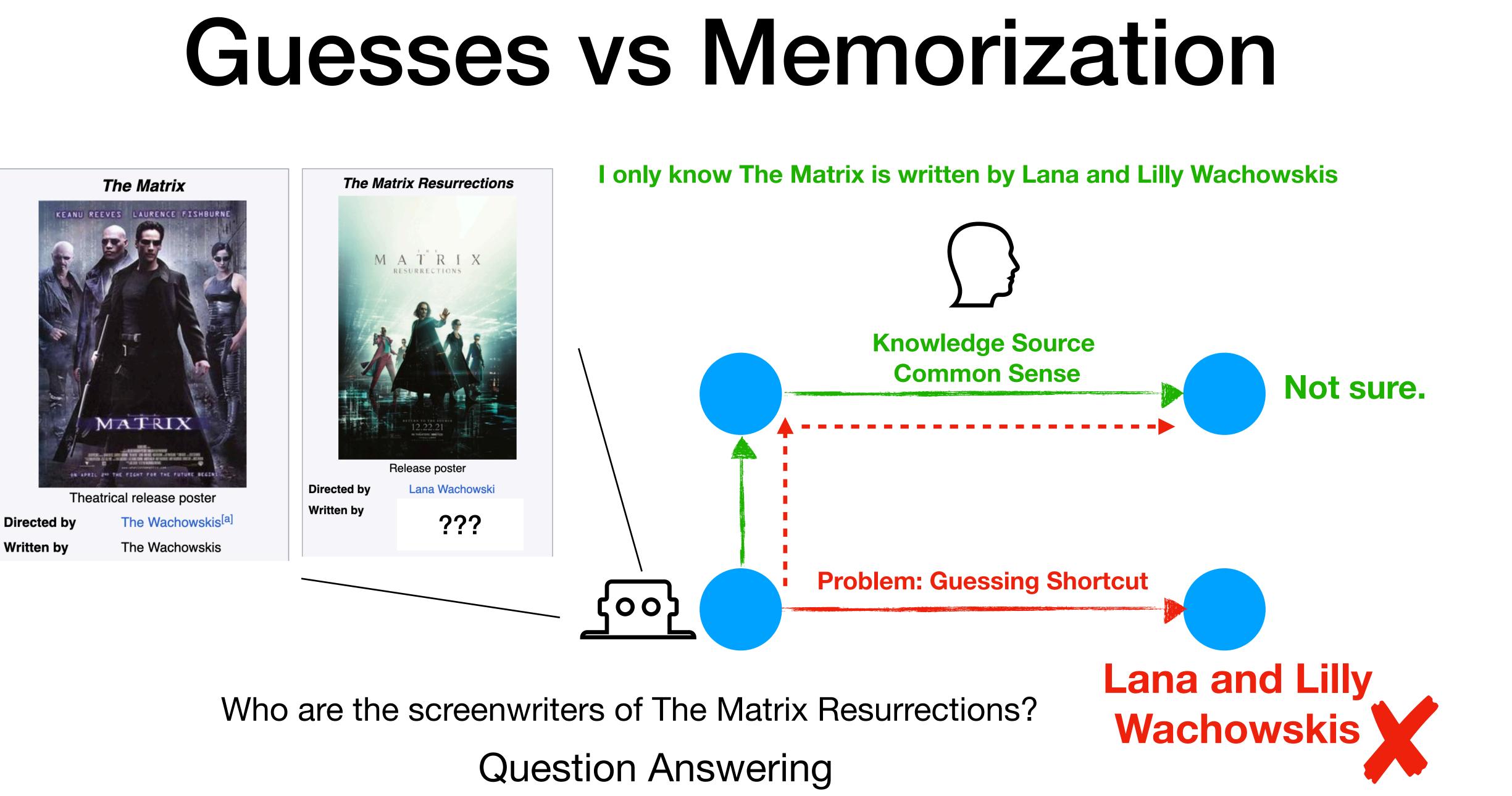
For LLMs, it doesn't matter





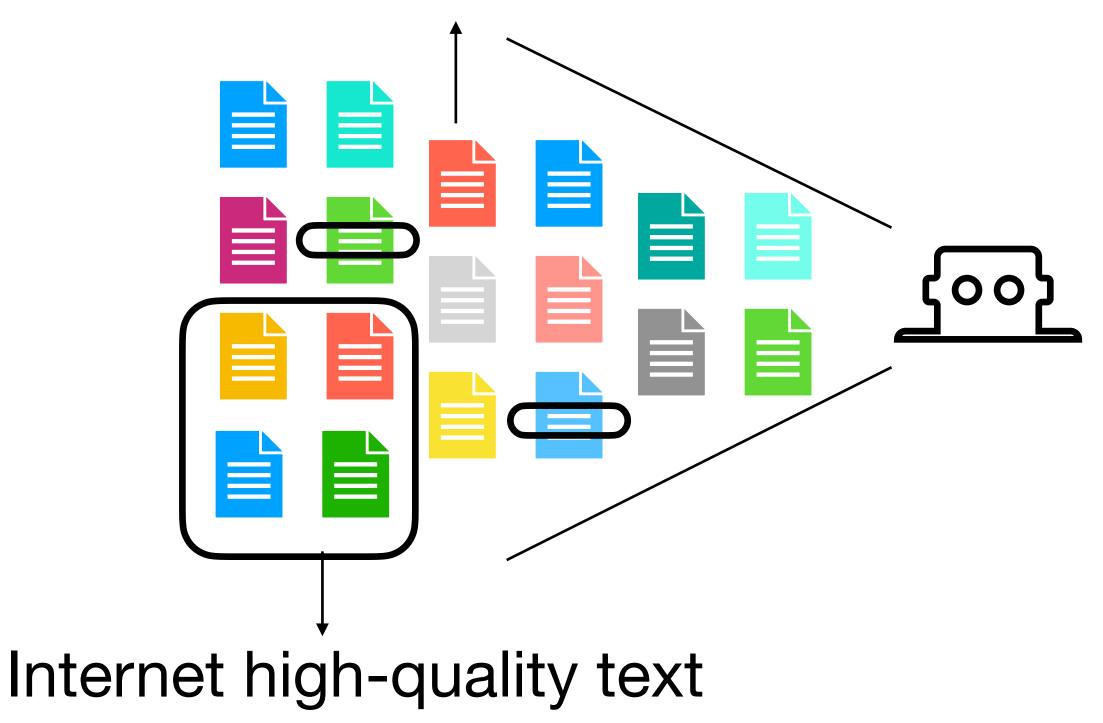
Linearly decay as the size exponentially increases -> a linear-log function

Kaplan et al., 2020



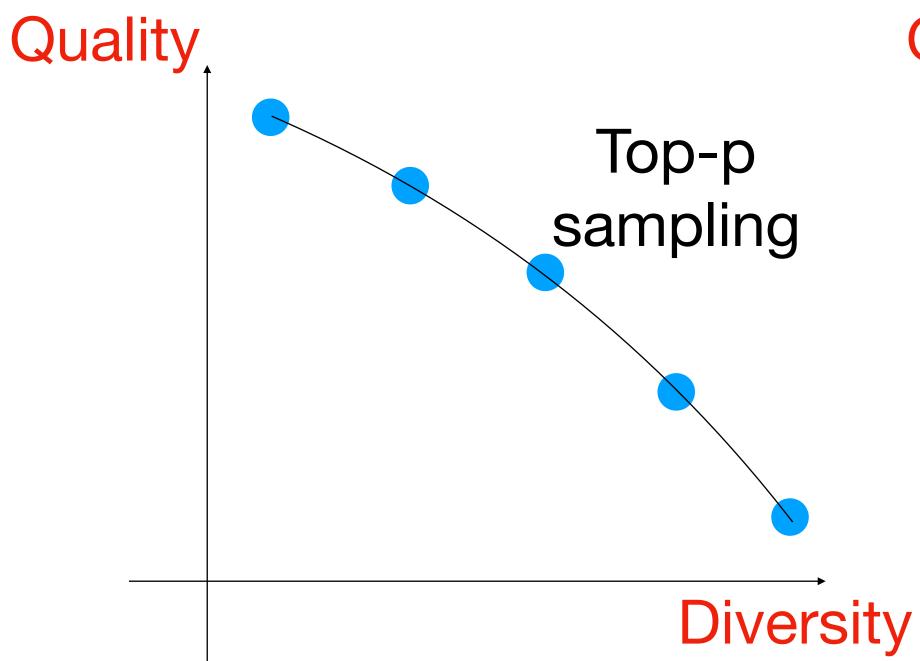
LLM Development

Internet low-quality text



- Training Stages
 - Pretraining
 - Supervised Fine-tuning (SFT)
 - Distillation
 - Alignment
 - Learning from Human Feedback (LHF)
 - Reasoning
- Inference Time
 - Prompt Engineering

Trade-off between Quality and Diversity



Lee, Nayeon, et al. "Factuality enhanced language models for open-ended text generation." Advances in Neural Information Processing Systems 35 (2022): 34586-34599.

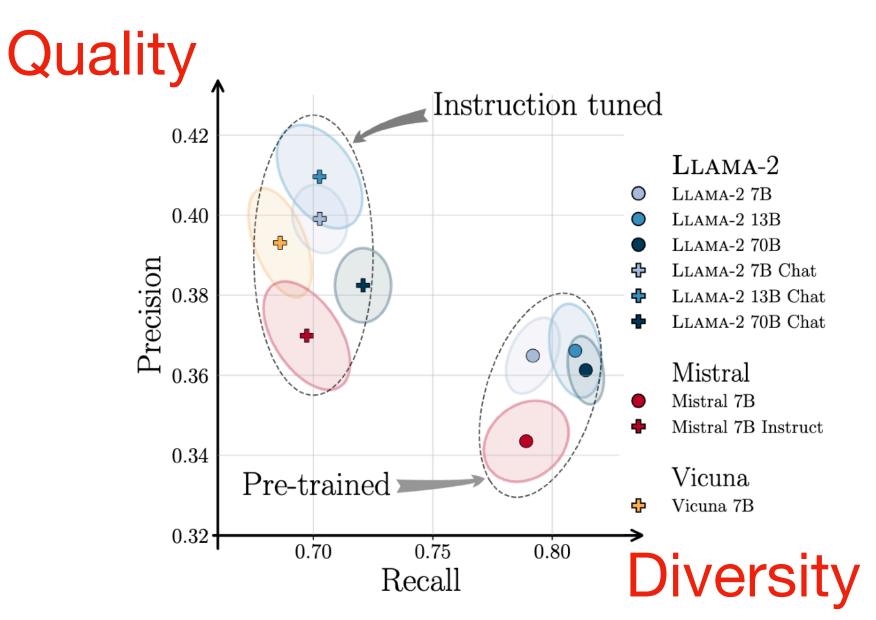
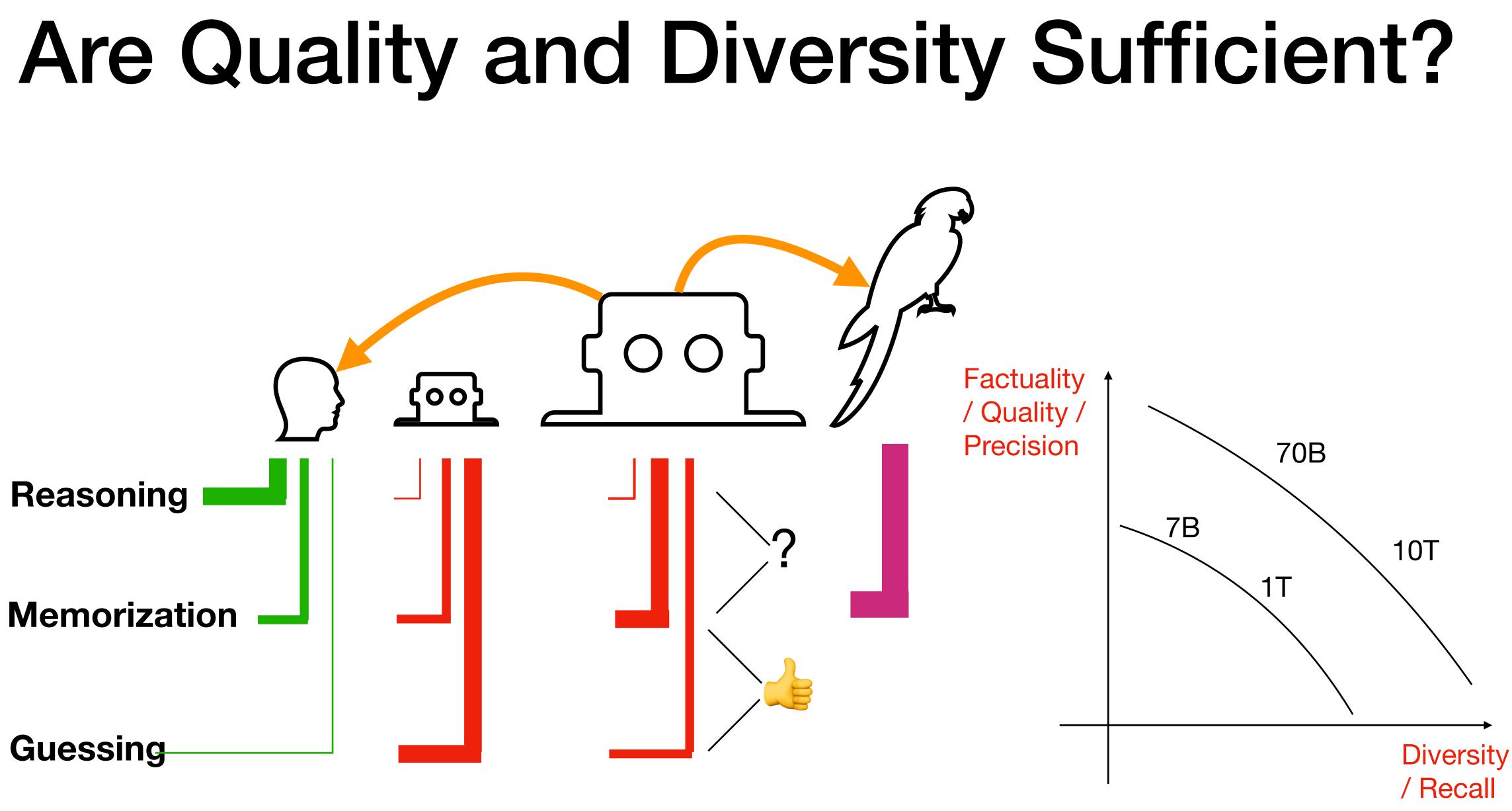




Figure 1: Precision and Recall of various models on generating the WebText dataset, with the 2 standard deviation error ellipsis. Chat and pre-trained models different behaviors are clearly captured by oevidenced.

> Le Bronnec, Florian, et al. "Exploring Precision and Recall to assess the quality and diversity of LLMs." 62nd Annual Meeting of the Association for Computational Linguistics. 2024.

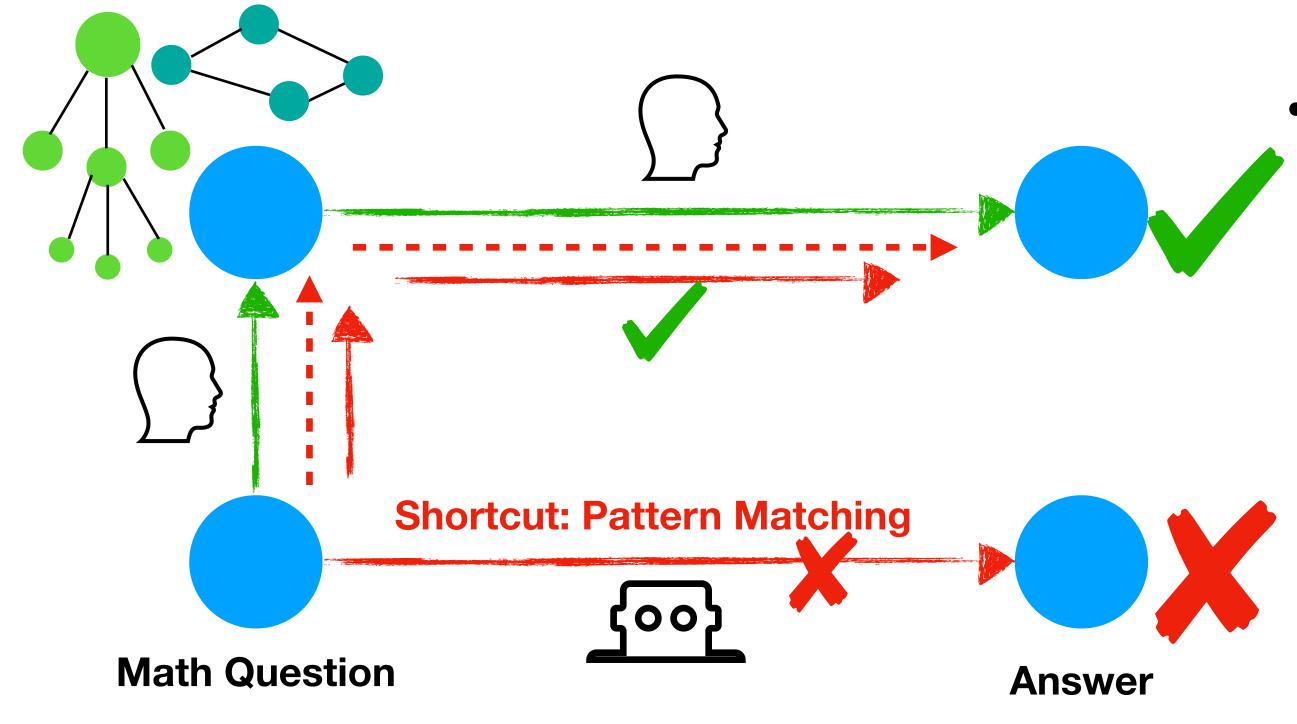


Le Bronnec, Florian, et al. "Exploring Precision and Recall to assess the quality and diversity of LLMs." 62nd Annual Meeting of the Association for Computational Linguistics. 2024.



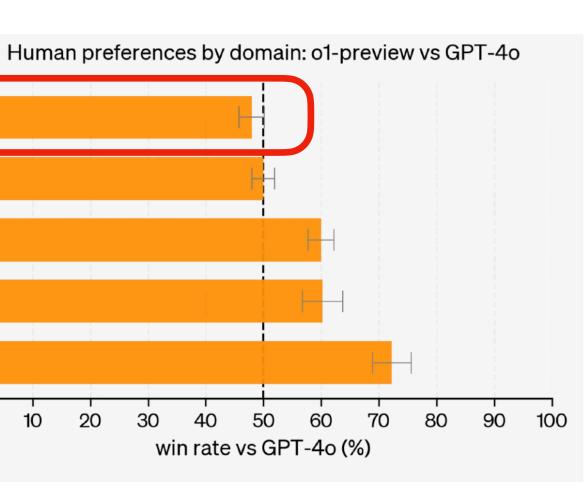
Improving the Reasoning/Creativity like o1?

Guiding the generations with answers like o1/o3 \bullet



Personal Writing Editing Text Computer Programming Data Analysis Mathematical Calculation 0 20 50 60 70 80 win rate vs GPT-4o (%)

- Limitations
 - Verifiable answers
 - Lots of training data
 - Not transferrable to other domains
 - Safety and controllability



RL for Reasoning is still a Filtering Method

Question(a): Find the least positive integer *n* such that $1/\sin 45^{\circ} \sin 46^{\circ} +$ $1/\sin 47^{\circ} \sin 48^{\circ} + \cdots + 1/\sin 133^{\circ} \sin 134^{\circ} = 1/\sin n^{\circ}$.

Response

...

First, I notice that the left-hand side of the equation has the form $\sum_{k=1}^{n} \frac{1}{\sin a_k \sin a_{k+1}}$, where $a_1 = 45^\circ$, $a_2 = 47^\circ$, ..., $a_{90} = 133^\circ$. I also notice that $\sin a_{k+1} = \sin(a_k + 2^\circ)$.

Simplifying, I get $sin(3a) = 3 sin a - 4 sin^3 a$.

Aha! I can use this to get $sin(a+2b) = sin((a+b)+b) = 3sin(a+b) \dots$ This looks messy, but I notice that it has the form sin a cos b + cos a sin b again. This gives me an idea. Maybe I can use this identity repeatedly to simplify the left-hand side of the equation.

Question(c): Denali and dog walking business ar	
Response To solve this problem, l defining the variables	
 Again, since $x \neq 0$, this doesn't give us a valid Let's check if we made should verify the probl again. The correct appr	

Figure 4: Cases showing that DeepSeek-V3-Base already exhibits "Aha moment" even before RL tunning.

d Nate work for a and are paid ...

let's start by

is equation solution either. an error. We olem conditions proach is to ...

Understanding R1-Zero-Like Training: A Critical Perspective (<u>https://arxiv.org/pdf/2503.20783</u>)





Evaluation AGI -> Achieving AGI

- "You insist that there is
- something a machine cannot do.
- If you tell me precisely what it is
- a machine cannot do, then I can
- always make a machine which
 - will do just that."
 - John von Neumann, 1948

https://www.reddit.com/r/singularity/comments/18t02br/john_von_neumann_was_the_first_who_used_the/

s ot do. I can hich



Improving Environments or Agents

- We know that evaluation could be used to optimize LLMs
- Environment/Evaluation is usually a mix of rules, tools, and data
- Do the fundamental limitations of LLMs come from data or models?
- Should we focus on improving the environment or the agent itself?

Welcome to the Era of Experience (<u>https://storage.googleapis.com/deepmind-media/</u> Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf)

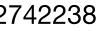
AlphaGo and AlphaZero



Why can AlphaGo be better than top human players, but LRM cannot?

https://www.bbc.com/news/technology-35785875

https://www.science.org/doi/10.1126/science.aar6404? __cf_chl_tk=67lg3VWHBOjaw3ybBhVGn2gbtd2QZ4UXUxDS21EBct4-1742742238 -1.0.1.1-jx7XtvwAIV5eX51WMPAteOy04PT4tJF2e28qsLvXeTc



GOODHART'S LAW

WHEN A MEASURE BECOMES A TARGET, IT CEASES TO BE A GOOD MEASURE

MEASURE PEOPLE ON

WEIGHT OF IF YOU NUMBER OF NAILS MADE NAILS MADE

THEN YOU 10005 OF MIGHTGET TINY NAILS

E

https://sketchplanations.com/goodharts-law

A FEW GIANT, HEANY NAILS



sketchplanations

The Power of Evaluation Functions

- Could be used in reinforcement learning
- Could be used in best-of-N
 - Process Reward Model
- Could be used in evaluating the high-quality output
 - LLM as a judge for creative writing
- Could be used in evaluating the low-quality output
 - ROUGE score, MAUVE, ...

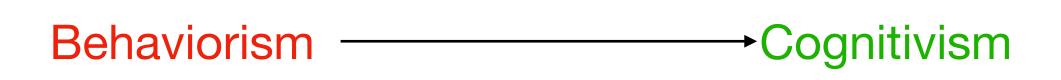
Math Answers, AlphaGo Value Network + Rollout, Reward Model for Alignment

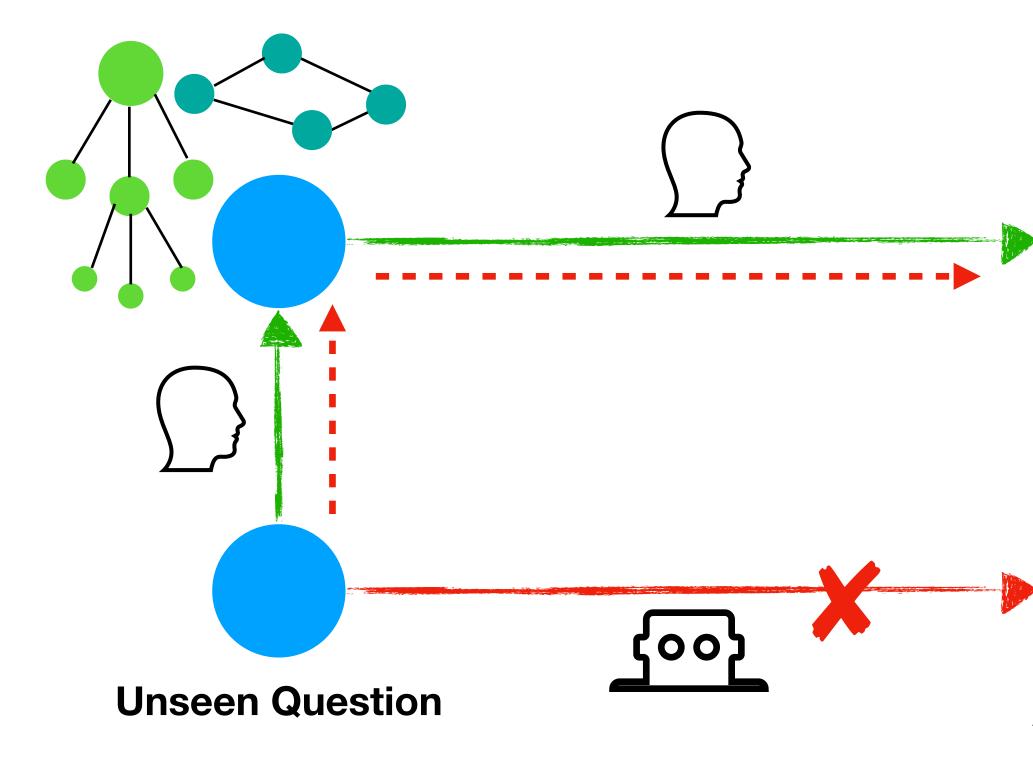
Goodhart's law:

When you optimize a metric, it is no longer a good metric



Unseen Problems that Need Creativity



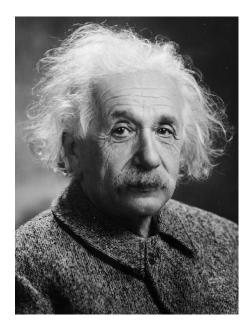


Long text/code generation



- Personalization
- Biomedicine, education, security, ...
- Limitations
 - Verifiable answers ?
 - Lots of training data
 - Not transferrable to other domains
 - Safety and controllability ?

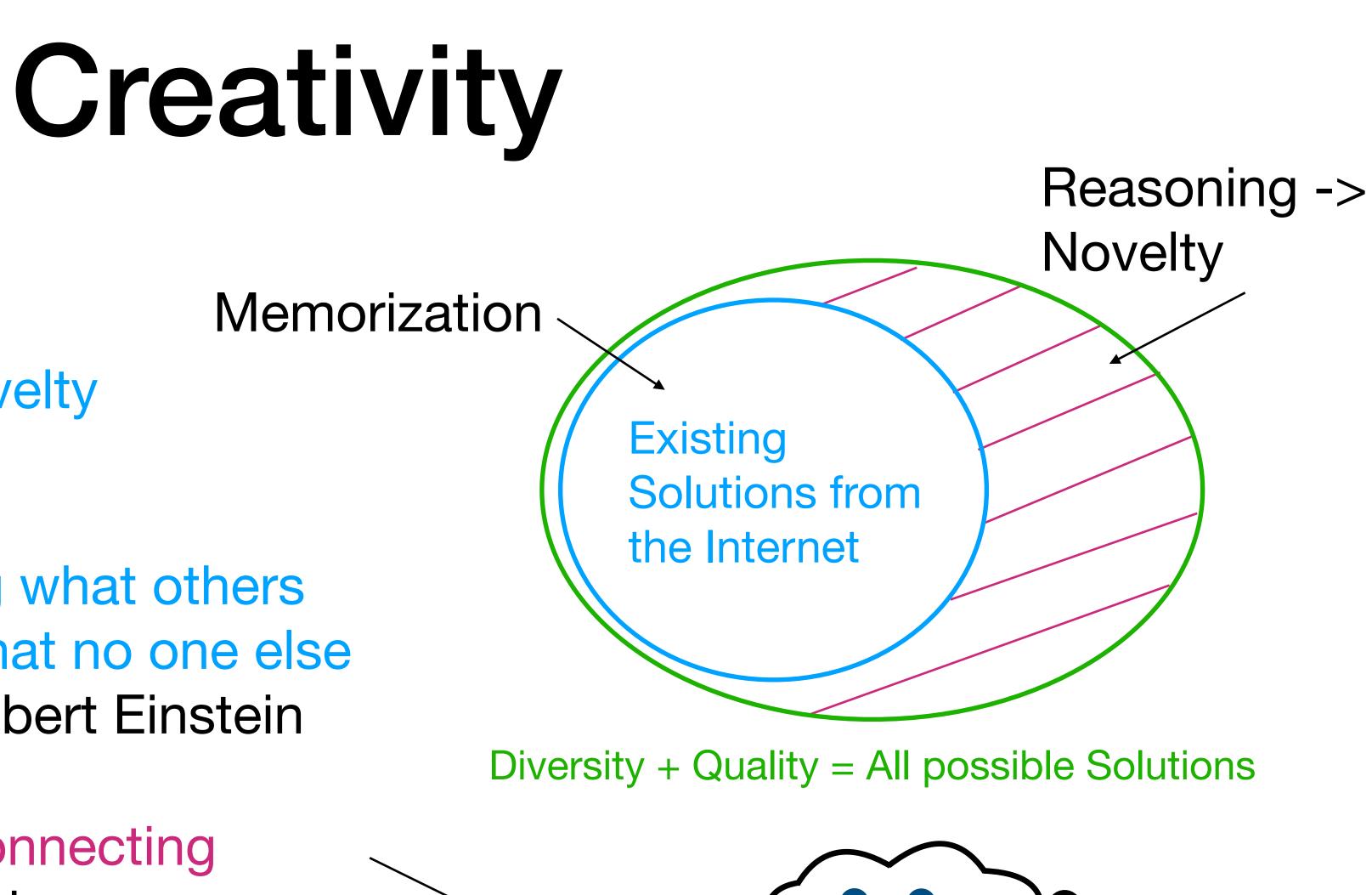
• Diversity + Quality + Novelty -> Creativity

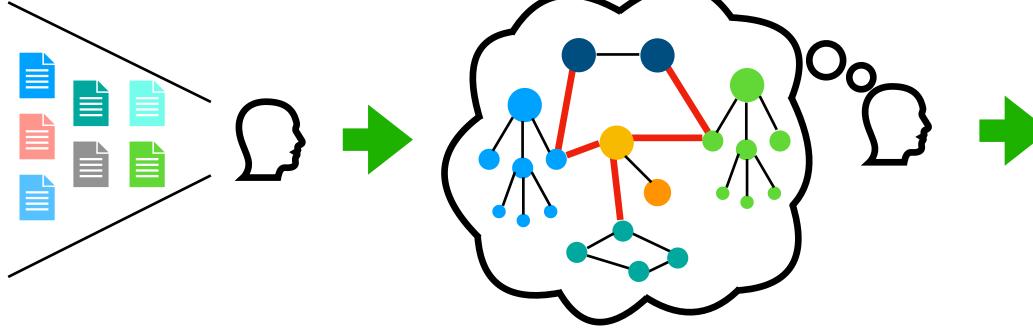


 "Creativity is seeing what others" see and thinking what no one else ever thought." — Albert Einstein



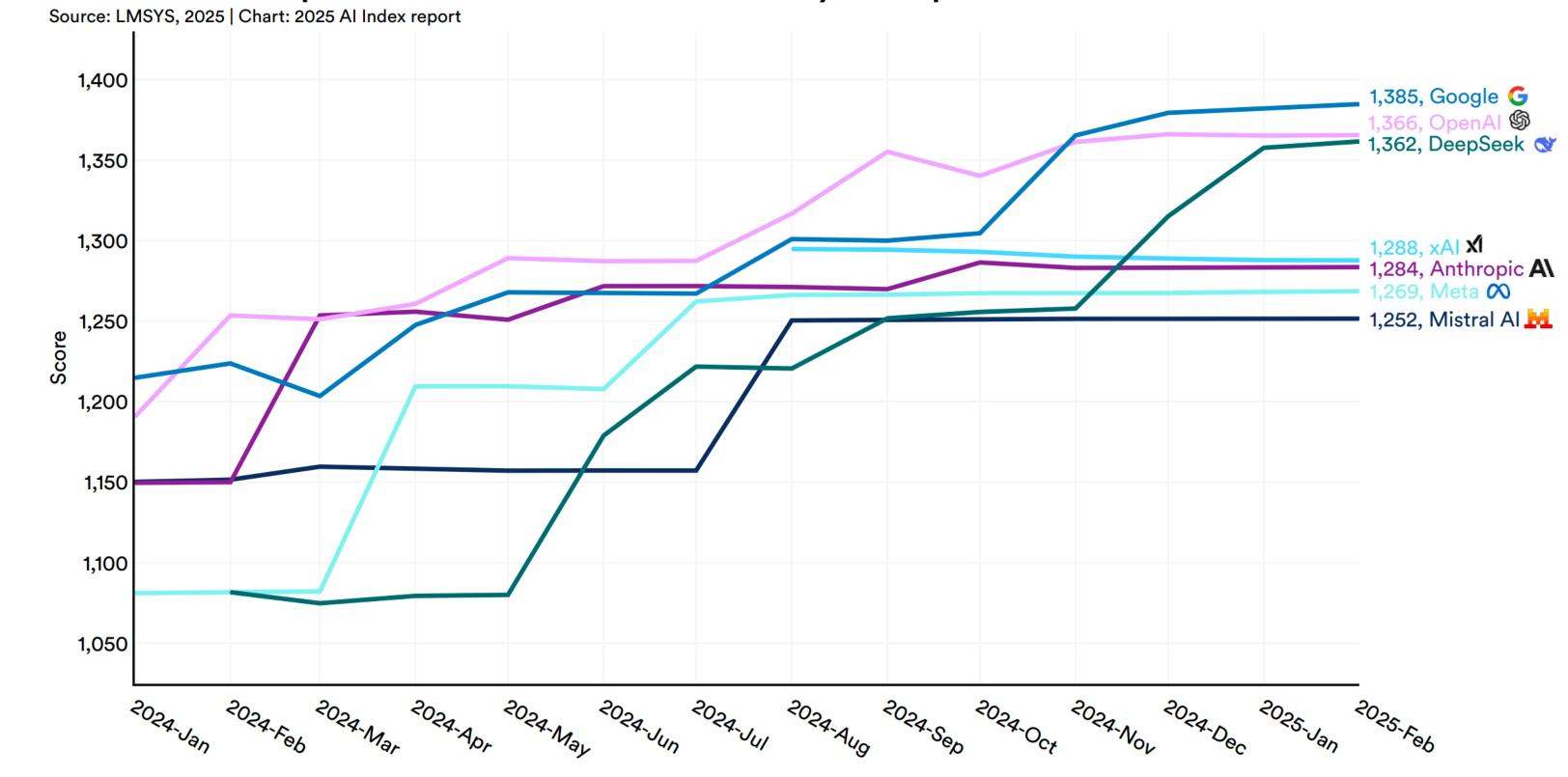
 "Creativity is just connecting" things." — Steve Jobs











Performance of top models on LMSYS Chatbot Arena by select providers

The 2025 AI Index Report (https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf)

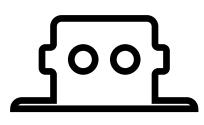


https://callnerds.com/parental-controls/

Human LLM Collaboration

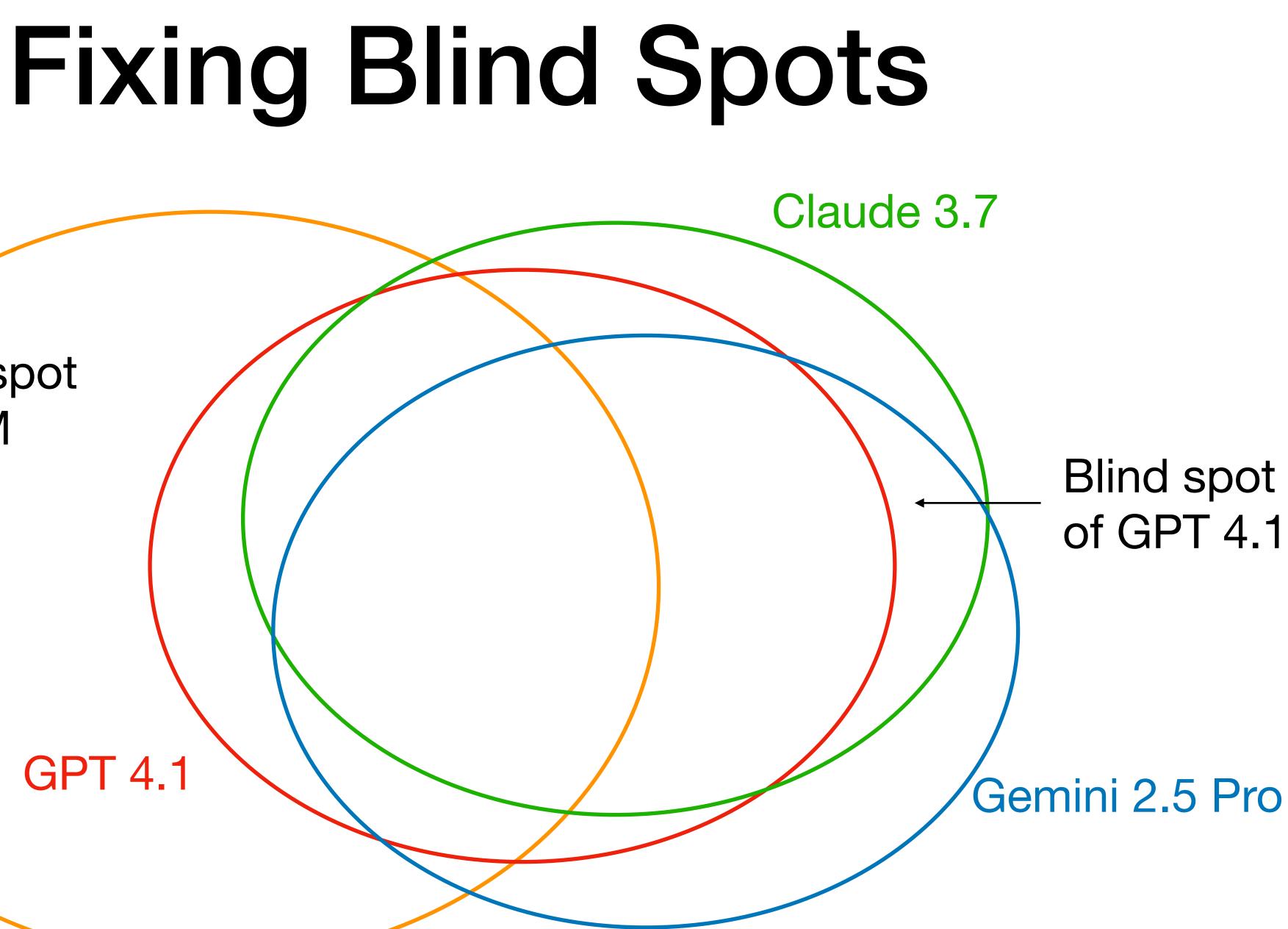
Planning / Decompose the Jobs

Fixes Problems



Do Smaller Tasks

Drafting

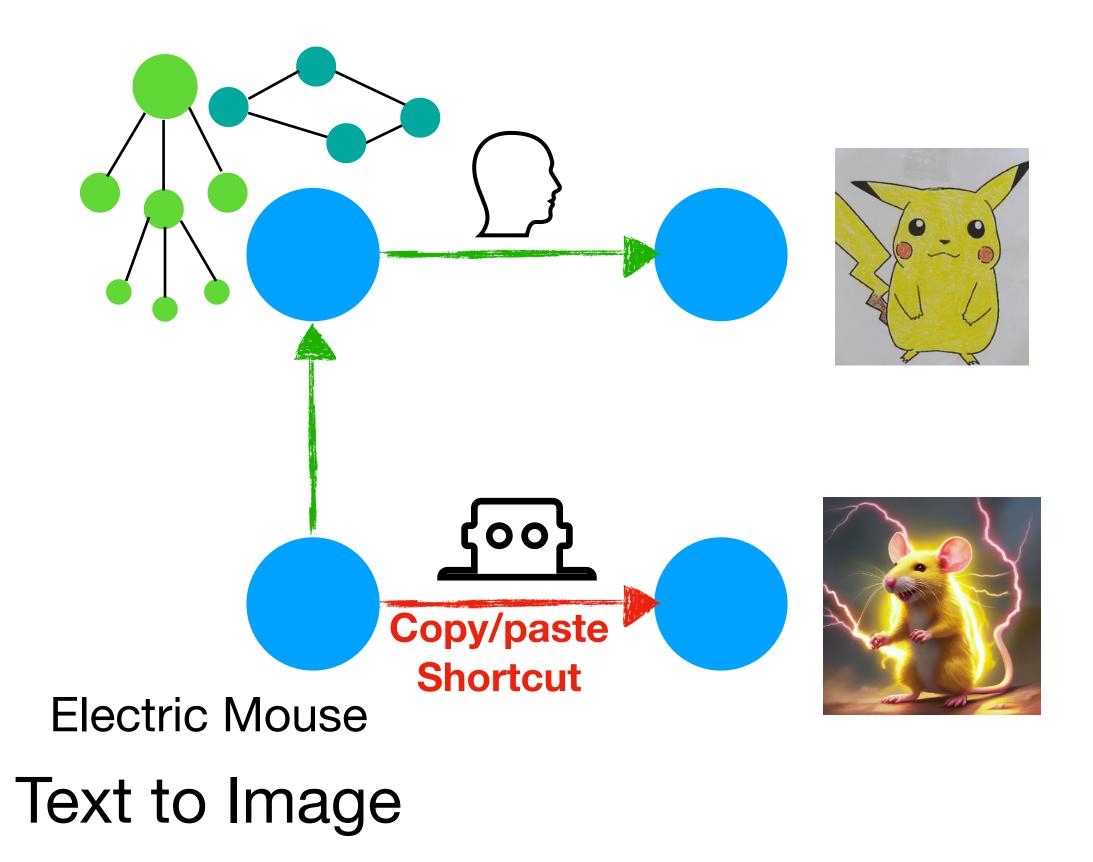




Human Expert

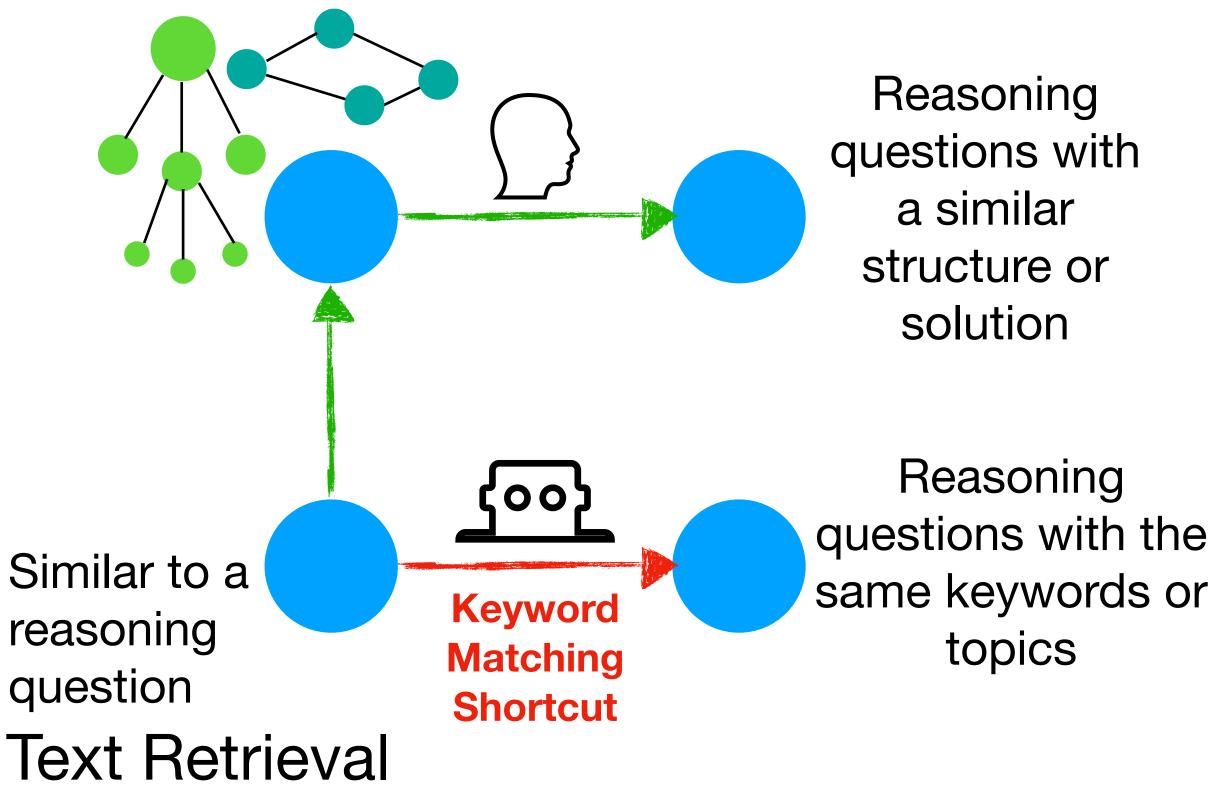
Similar Challenges across Domains

Computer Vision



https://www.rawpixel.com/image/6113771/photo-image-paper-sticker-public-domain https://shedevrum.ai/en/post/fc3e52e78eea11eea4226ac6a1596643/

Information Retrieval



Humans vs Parrot

Like Brain <- Transformer -> Matrix factorization

Personality <- Alignment -> Filter pretraining data

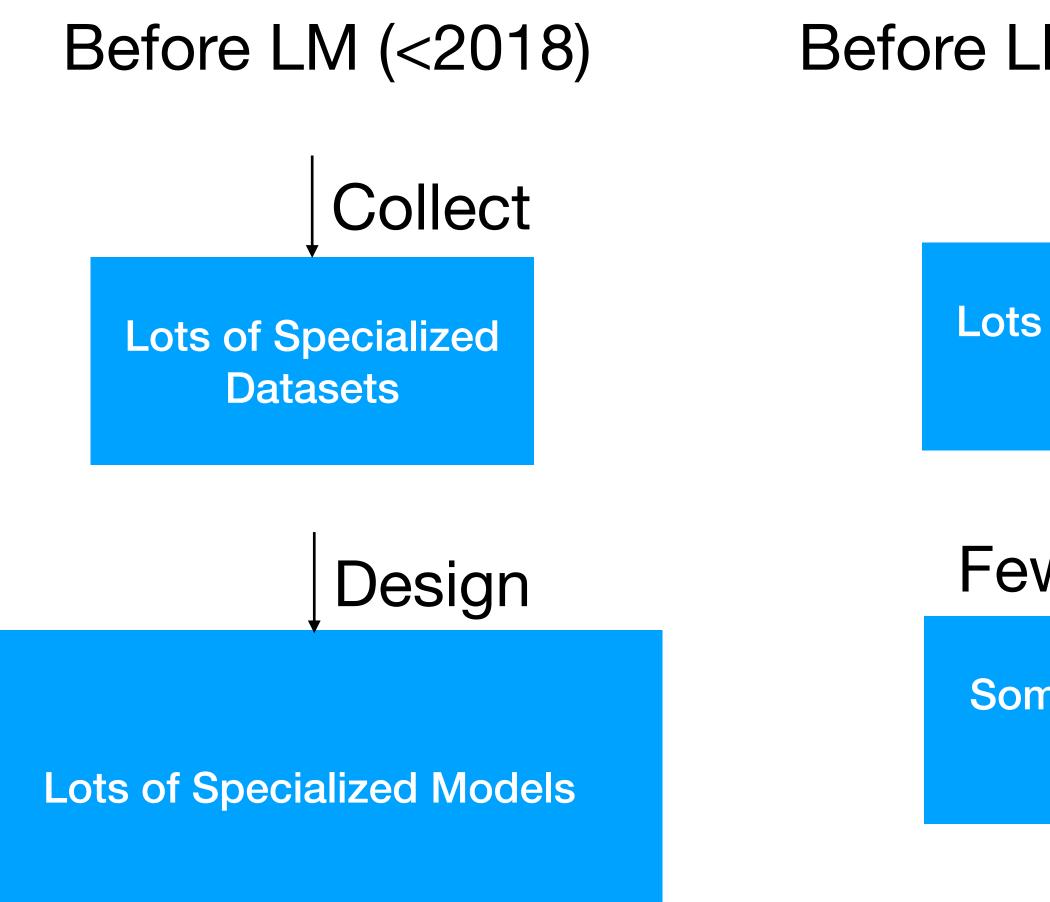
Media, General Public

Industry

Thinking <- Reasoning -> Identify pretraining data?

Machine Learning Researchers Academia

Challenges of NLP Research



Before LLM (2018-2022)

Collect

Lots of Specialized Datasets

Fewer

Some Specialized Models

Aftrer LLM (>2022)

Collect

Some Specialized Datasets (mostly for eval)

Almost no Design

Some Specialized Models (agentic LLMs)



Now, what do you think?

Google DeepMind 145-page paper predicts AGI will match human skills by 2030 — and warns of existential threats that could 'permanently destroy humanity'

An Ex-OpenAl researcher predicts AGI by 2027here's the wild roadmap.

You think the pace of AI development is going fast now? You ain't seen nothing yet.

https://www.theneuron.ai/explainer-articles/ an-ex-openai-researcher-predicts-agiby-2027--heres-the-wild-roadmap



https://fortune.com/2025/04/04/ google-deeepmind-agi-ai-2030risk-destroy-humanity/



Yann LeCun 🤣 🙉

I said that reaching Human-Level AI "will take several years if not a decade."

Sam Altman says "several thousand days" which is at least 2000 days (6 years) or perhaps 3000 days (9 years). So we're not in disagreement.

But I think the distribution has a long tail: it could take much longer than that. In AI, it almost always takes longer.

In any case, it's not going to be in the next year or two.

https://x.com/ylecun/status/ 1846574605894340950?lang=en



Ø ...

Thank you for taking the class. Hope that you will learn something useful from this class