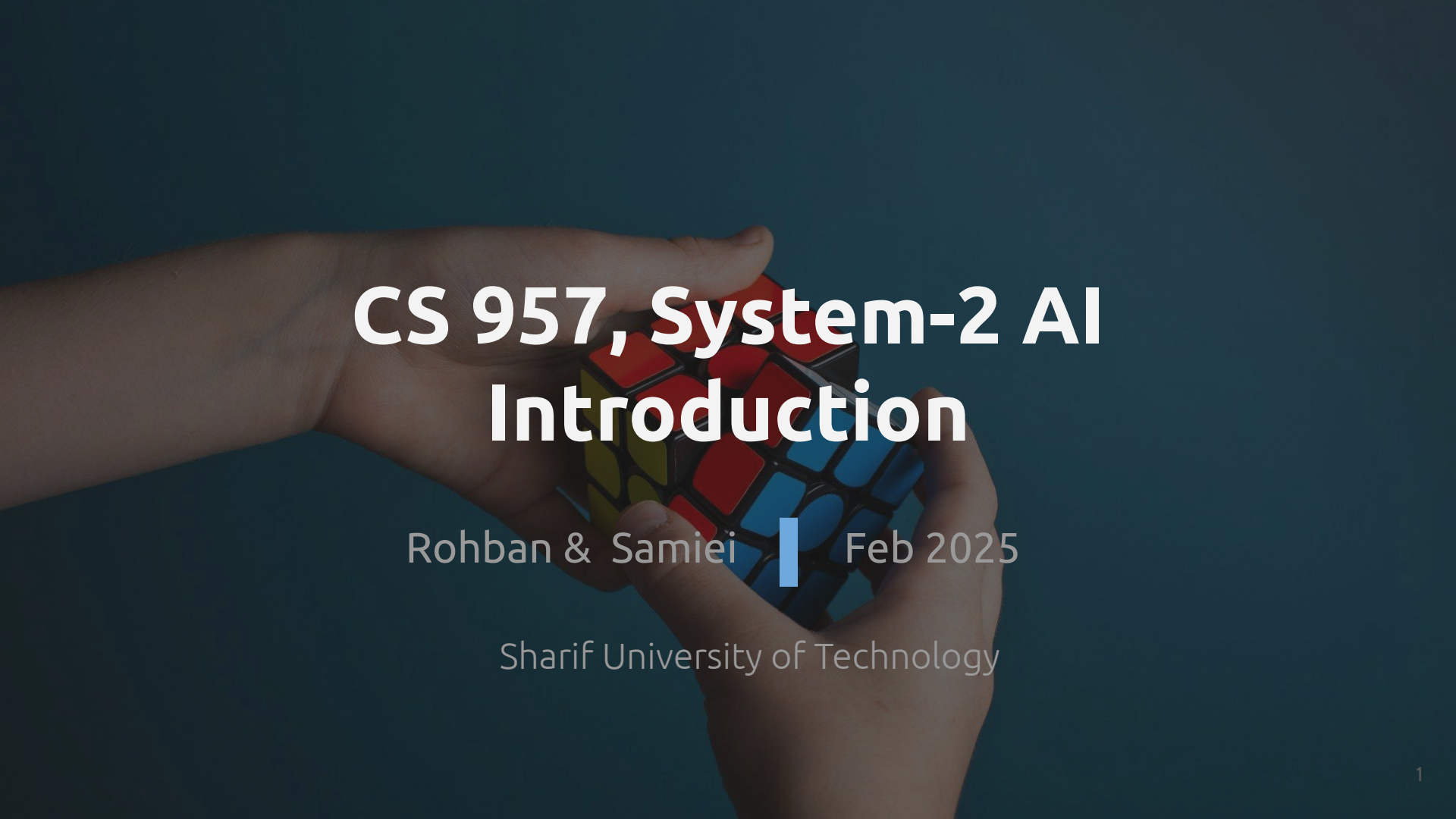


CS 957, System-2 AI Introduction

Rohban & Samiei

Feb 2025

Sharif University of Technology

Staff Introductions

You can find us in the machine learning lab on the fourth floor.

Mohammadhossein Rohban

Mahdieh Soleymani

Mohammadmahdi Samiei

Staff Introductions: HeadTA

Arashmari Oriyad

Course Structure

- ~~MidTerm~~
- Final (6 / 20)
- Quizzes (4 / 20)
  - 4 or 5 quizzes
- HWs, Project (7 / 20)
  - 5 HWs.
- Presentation (3/20)
  - Poster Session

What We Will Discuss in This Session

AI Problem

Automation
AI Effect
Gödel's theorems
...

... - 1958

CNN

Among First Inductive Biases
Connectionism vs Symbolism

1980

AlexNet

First Deep Model
ImageNet Moment
Internet Data
GPU Computation

2012

Transformer

Latest innovative architecture
Alpha Go: Move 37

2017

BERT

Pre-training as Paradigm
Open source accelerated AI

2018

Initial Criticisms

OoD: is 100% enough?
ARC: Measure of intelligence
Symbolic Problems

2019

LLM Begins

GPT-3
Scale Laws
Reasoning Benchmarks

2020

New Hope to AGI

GPT-3.5
CLIP
DALL-E
Deceive Age
Chain of Thought
LLM As Agents

2022

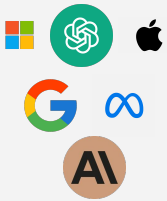
Return of RL

GPT4: RLHF
Is winter coming?

2023

Still Spring

GPT-4o
o1: Test Time Computation
o3: RIP ARC



2024

To Be Continue

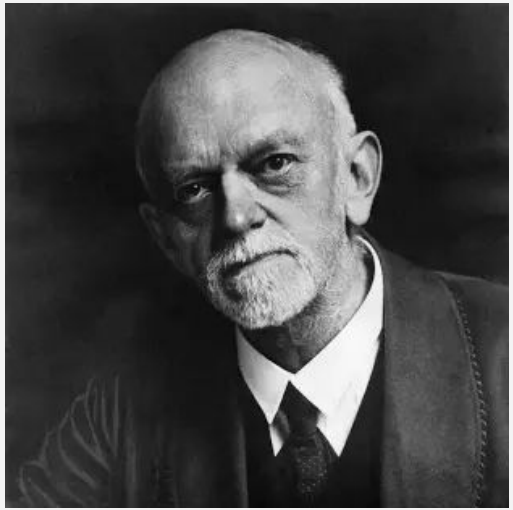
DeepSeek: Reward is All you Need
Performance-to-Cost Ratio
...

2025

Seeking Automation in Hard Tasks (Hilbert - Godel) 1900-1931

Hilbert sought "Automated Theorem Proving" in 1900.

Godel proved this to be impossible in expressive formal systems!



Artificial Intelligence Born (1950)

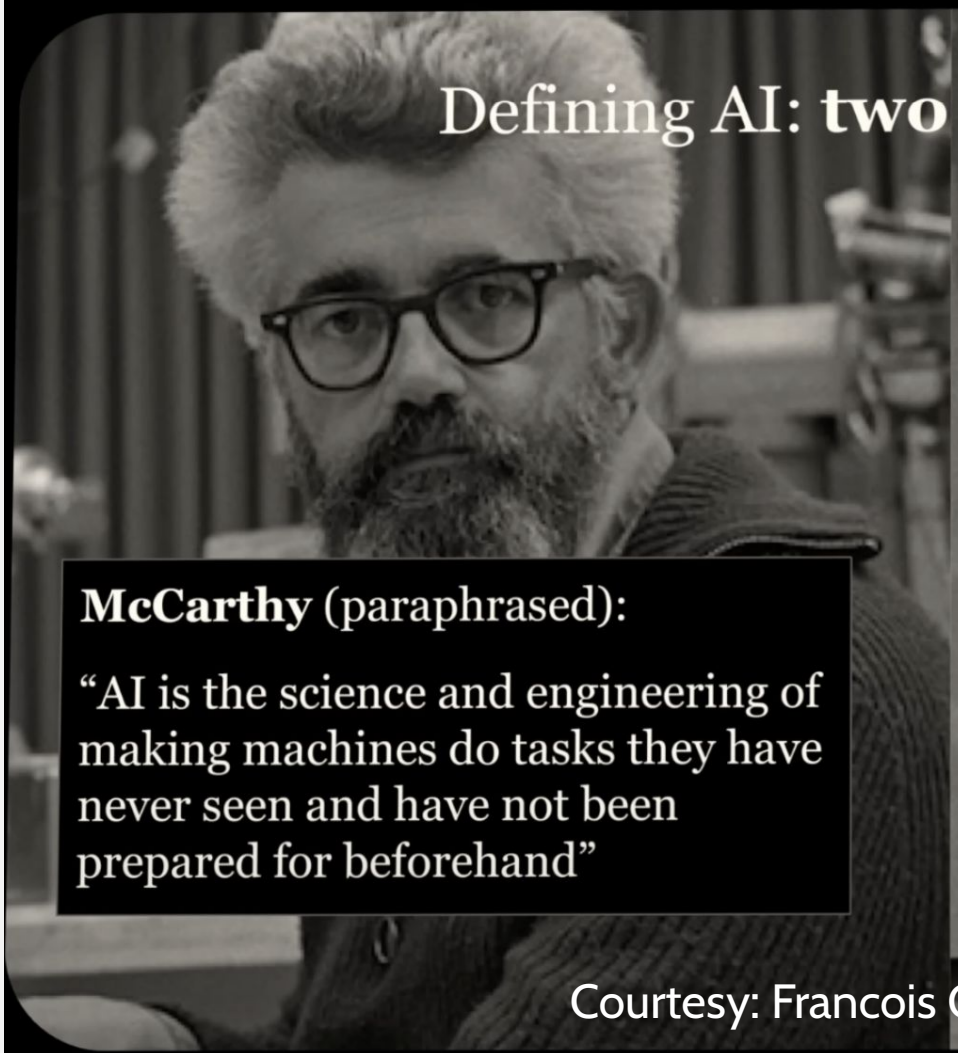
Definition of Intelligence

- Think like human (Rebuild the brain)
- Act like human (Turing Test)
- Think Rationally (Symbolic Era)
- Act Rationally (Mostly Learning Era)

Automation (Money, Power)

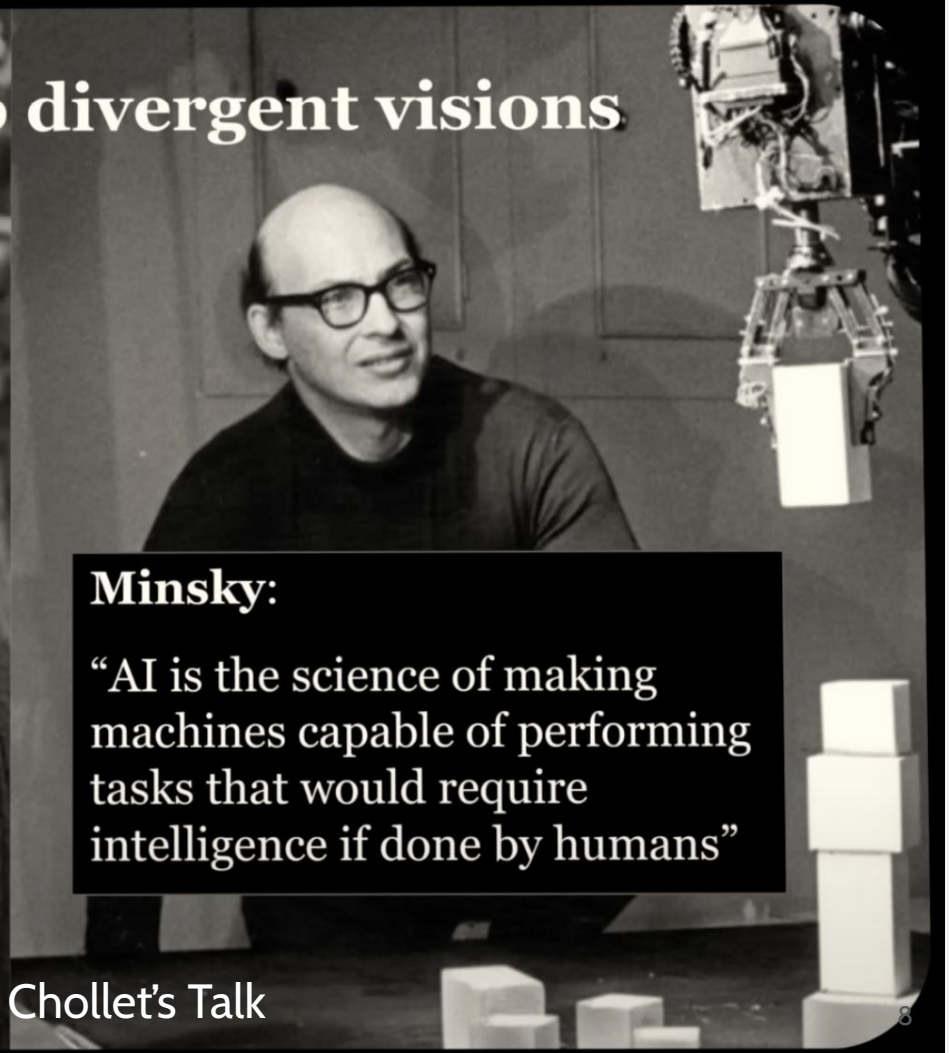
Reasoning vs Learning

Defining AI: **two divergent visions**



McCarthy (paraphrased):

"AI is the science and engineering of making machines do tasks they have never seen and have not been prepared for beforehand"



Minsky:

"AI is the science of making machines capable of performing tasks that would require intelligence if done by humans"

Courtesy: Francois Chollet's Talk

Artificial Intelligence

AI Effect: *It's part of the history of the field of Artificial Intelligence that every time somebody figured out how to make a computer do something—play good checkers, solve simple but relatively informal problems—there was a chorus of critics to say, 'that's not thinking*

Intelligence is whatever hasn't been achieved Yet

Simulated exams	GPT-4 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th
LSAT	163 ~88th
SAT Evidence-Based Reading & Writing	710/800 ~93rd
SAT Math	700/800 ~89th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th
Graduate Record Examination (GRE) Verbal	169/170 ~99th
Graduate Record Examination (GRE) Writing	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th
USNCO Local Section Exam 2022	36/60
Medical Knowledge Self-Assessment Program	75%
Codeforces Rating	392 below 5th
AP Art History	5 86th–100th
AP Biology	5 85th–100th

LLMs started scoring much better than humans on virtually any benchmark...

Courtesy: Francois Chollet's Talk

AI Winter and Springs 1960 - 2010

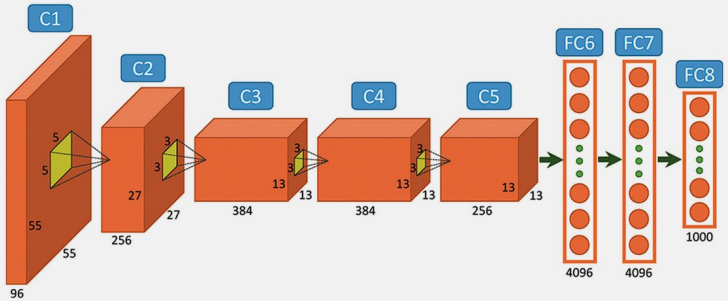
Logics and Expert Systems: Encoding domain knowledge

Statistical Learning and Neural Networks: Let the system learn them itself

Symbolism vs. Connectionism

From Neural Networks to Deep Learning, 2010-2012

- 👉 From MNIST(1998) to ImageNet(2010)
- 👉 From CNN(1980) to AlexNet(2012)
- 👉 Why 2012?
  - 👉 Data and Compute.
  - 👉 Internet and GPU
- 👉 AlexNet success encouraged others to Deep Learning.
- 👉 So story begins ...



From ImageNet and AlexNet to others, 2012 - ...

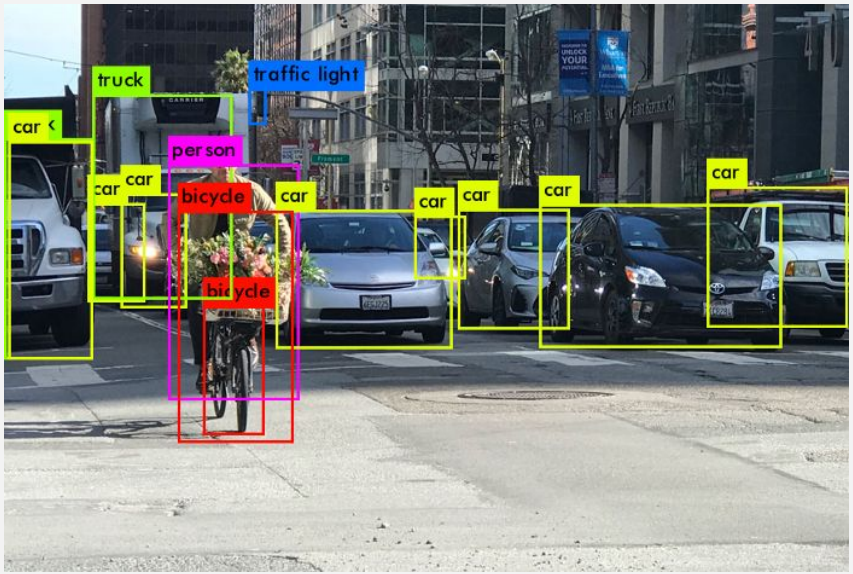
👉 Tasks: Image Classification, Object Detection, Semantic Segmentation, ...

👉 Datasets: MS COCO, CelebA, ShapeNet, JFT, ...

👉 Models: ResNet, Yolo, U-Net, ...

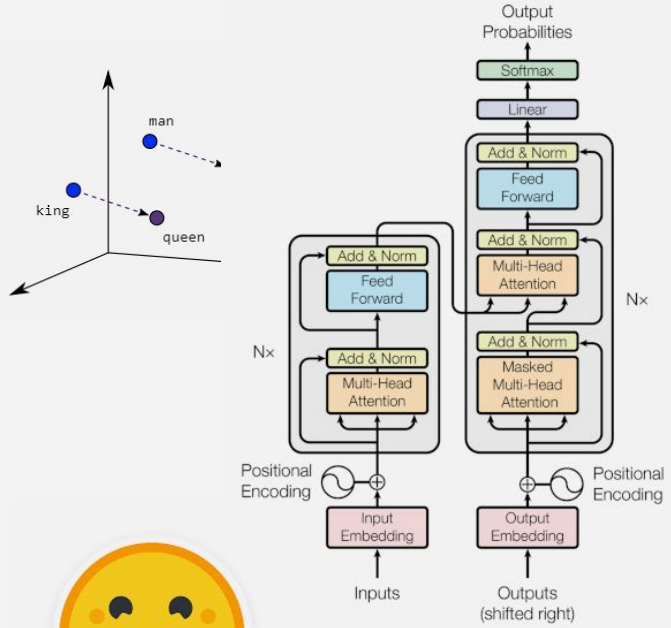
👉 Methodologies: Self Supervised Learning, ...

👉 Vision Tasks as Perception



From Vision to Natural Language and other Modalities

- 👉 Word Embedding and Recurrent Neural Networks
- 👉 Machine Translation Induce Attention Mechanisms
- 👉 Attention is All You Need (Transformers)
- 👉 Maybe NLU is Perception Too!!!!
- 👉 Self-Supervised for NLP (BERT, GPT, ...)
- 👉 Vision Transformers over CNN
- 👉 The Role of Hugging Face in Democratizing AI



Deep RL and AlphaGO

⚫ Go vs Chess:

Simpler But More Challenging.

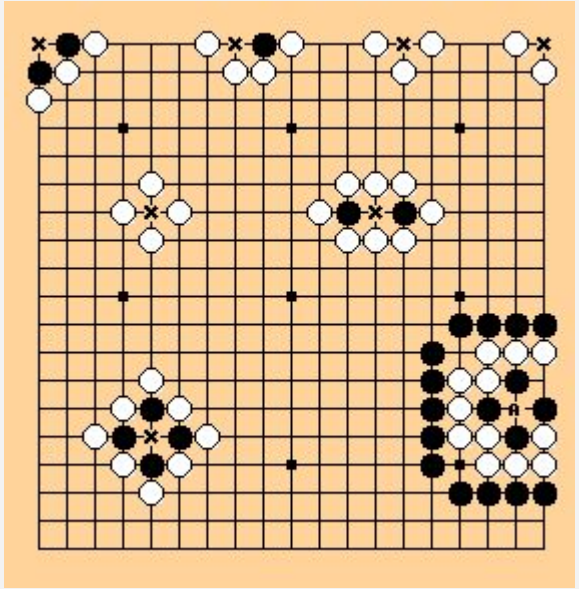
State Space size: 10^170 vs 10^40

Branching Factor: 250 vs 35

⚫ AlphaGo:

Pruned search of state space using learned value and policy function

Learning pipeline very similar to DeepSeek (SFT, RL, ...)



Early Criticisms of Deep Learning Limitations, 2019

🎉 Successes:

Narrow task systems: precise defining of goals

🤦 Failures:

Benchmark Problems: Single Focus, Kaggle Effect, Shortcut Rule

Brittle to Attacks

Data Hungry

Unable to Make Sense

Unable to Deal with Novel Tasks

Bengio's Criticisms

❓ Is 100% accuracy on test-set enough?

🌍 IID is a statistical assumption, but nature scenarios are not.

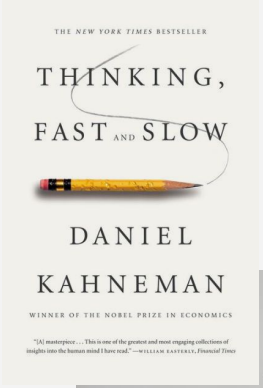
😮 Out of Distribution Generalization (OoD)

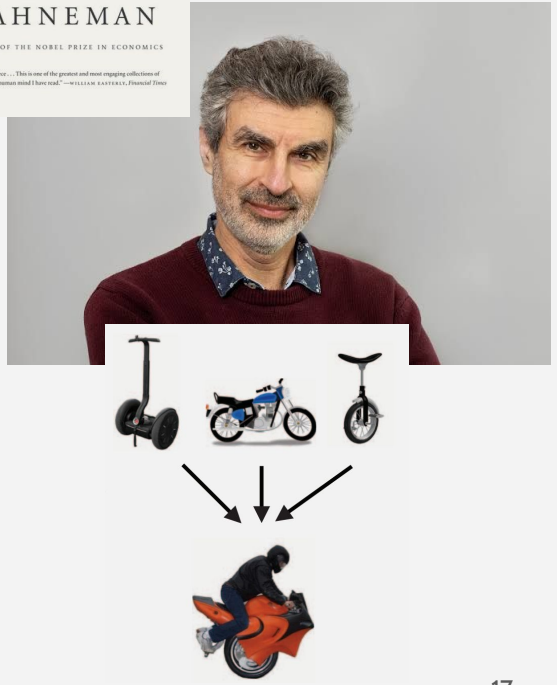
💡 Systematic Generalization, System-2, Conscious Processing

💡 RIM, BRIM, NPS, Slot Attention, GFlowNet, ...

🤔 Does OoD Generalization Make Sense??

Compositional Inductive Biases



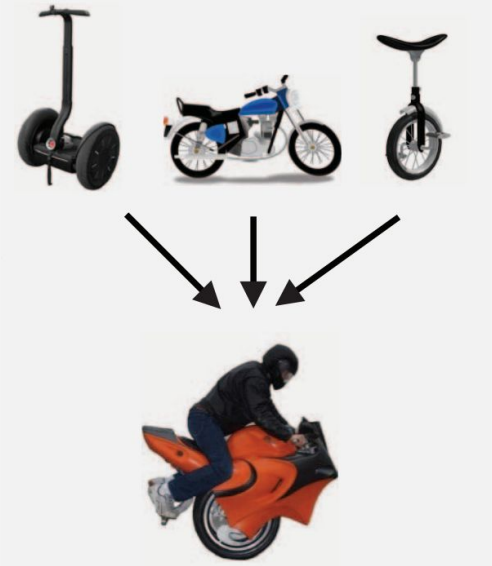


Bengio's Points: OoD and Systematic Generalization

Infinite use of finite means, Studied in Linguistics

Even when new combinations have zero probability under training distribution.

e.g. Science Fiction Scenarios.
e.g. Driving in different traffic pattern.

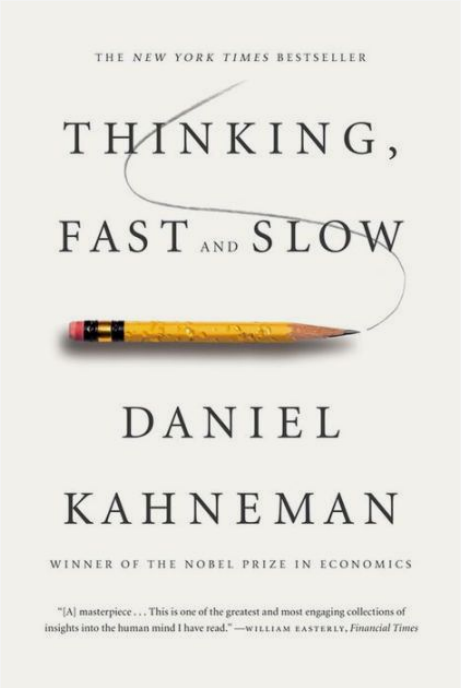


Bengio's Points: System 1 vs. System 2 Cognition

System 1:

Inductive, Fast, Unconscious, Non-linguistic, Habitual

Current DL



System 2:

Slow, Logical, Sequential, Conscious, Linguistic, Algorithmic, Planning, Reasoning

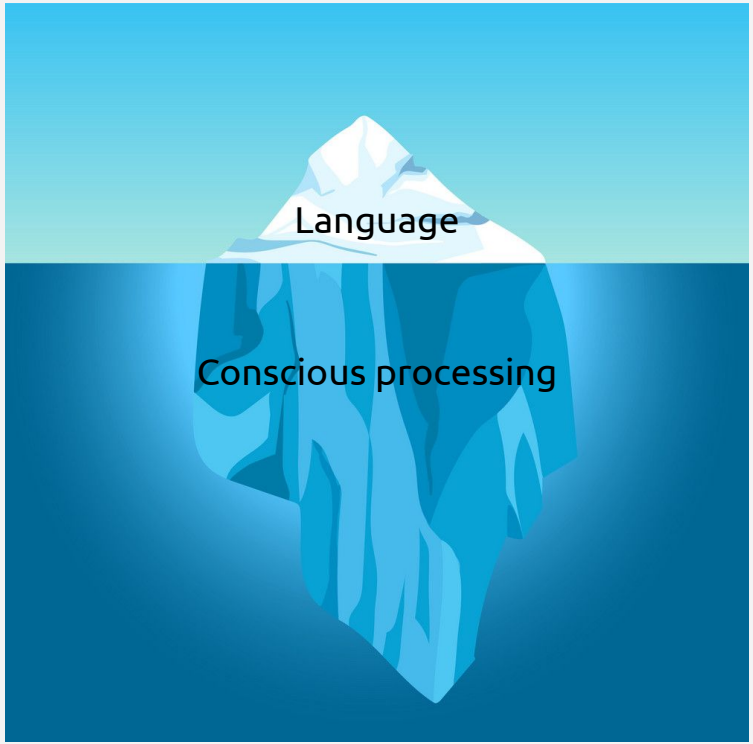
Future DL

Bengio's Points: Conscious Processing Helps Humans Deal with OoD Setting

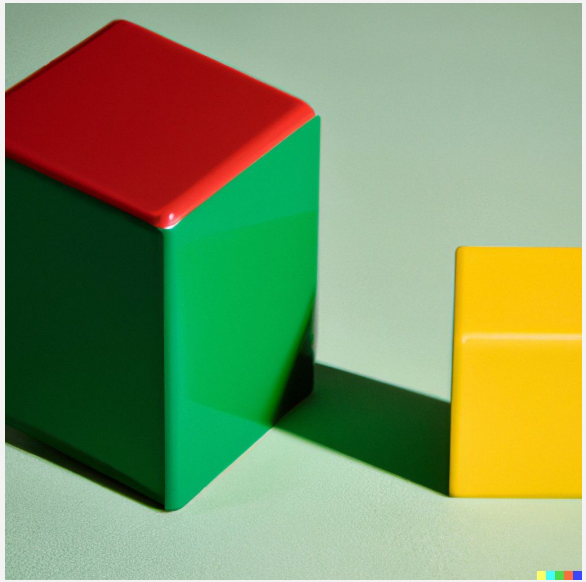
Faced with novel or rare situations, humans call upon conscious attention to combine on-the-fly appropriate piece of knowledge, to reason with them and imagine solutions.

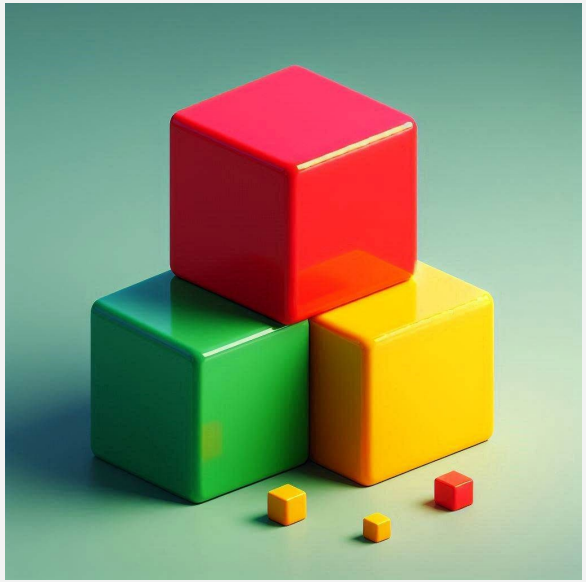
We do not follow out habitual routines, we think hard to solve new problems.

Bengio's Points: Language as the Tip of Consciousness Iceberg



DALL-E isn't great at composition



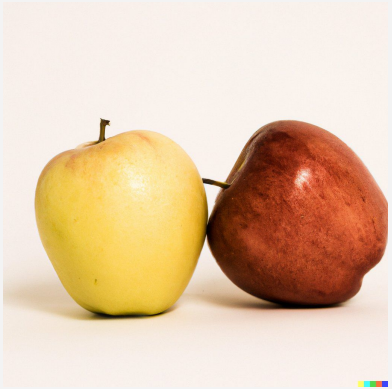


A red cube, on top of a yellow cube, to the left of a green cube

DALL-E can't count

1 apples

2 apples



3 apples

10 apples

DALL-E isn't great at common sense

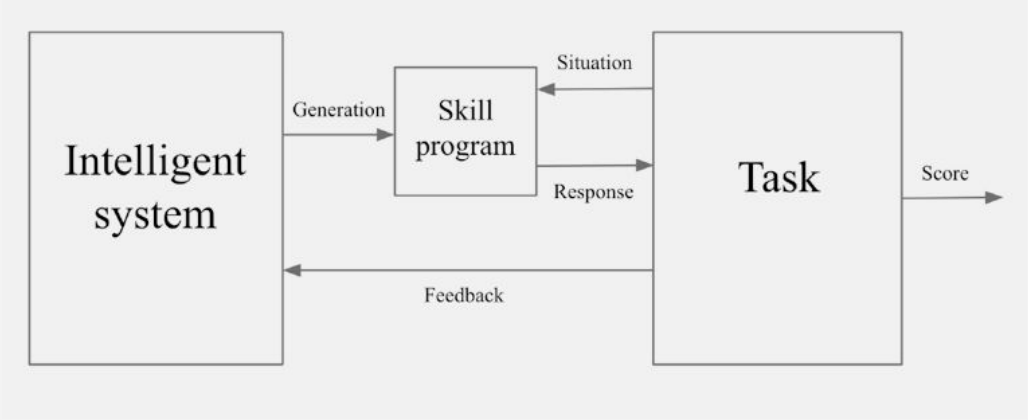
An old man is talking to his parents.

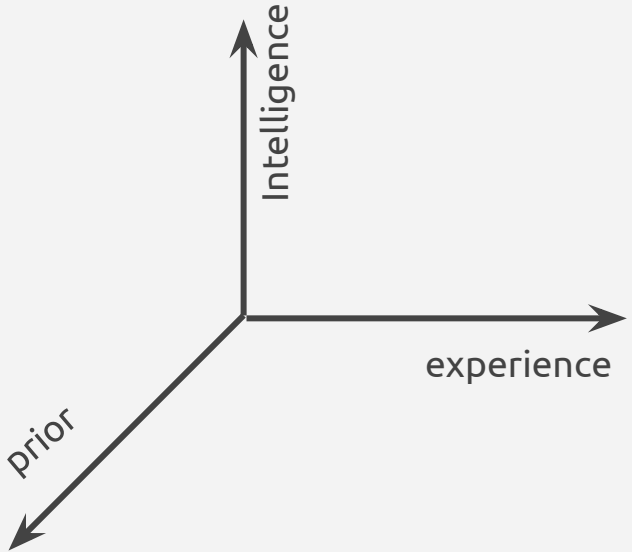
Chollet Criticisms

👉 Intelligence is orthogonal to Prior and Experience.

👉 Measure intelligence? factor out priors and experience.

👉 Developer-aware generalization





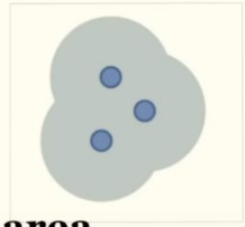
Key quantities for conceptualizing intelligent systems

Static skills: repository of memorized programs

Fluid intelligence: synthesize new programs on the fly

Fluidity

Narrow operational area of programs used (low abstraction)



Broad operational area of programs used (high abstraction)

Operational area

Data-hungry program acquisition / synthesis

Information-efficient program acquisition / synthesis

Information-efficiency

Courtesy: Francois Chollet's Talk

From **factoids**...

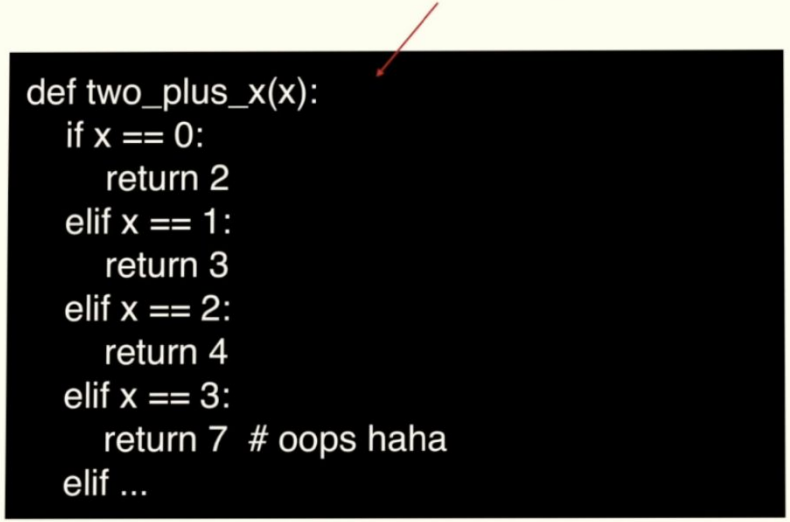
```
def two_plus_two():
    return 4
```

```
def two_plus_three():
    return 5
```

```
def two_plus_ten():
    return 12
```

To organized **knowledge**...

This is abstraction! The program is abstract for x!



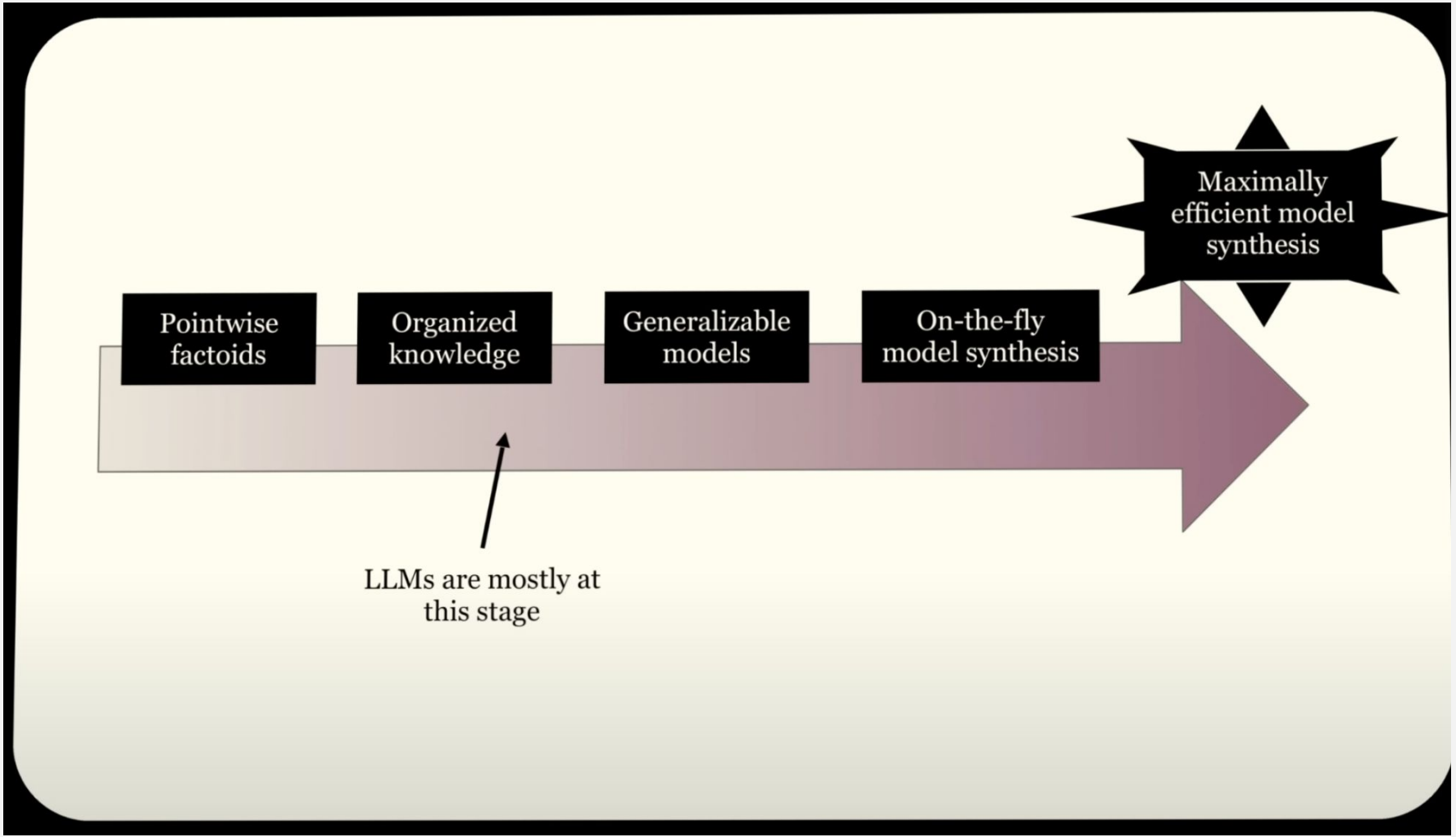
```
def two_plus_x(x):
    if x == 0:
        return 2
    elif x == 1:
        return 3
    elif x == 2:
        return 4
    elif x == 3:
        return 7  # oops haha
    elif ...
```

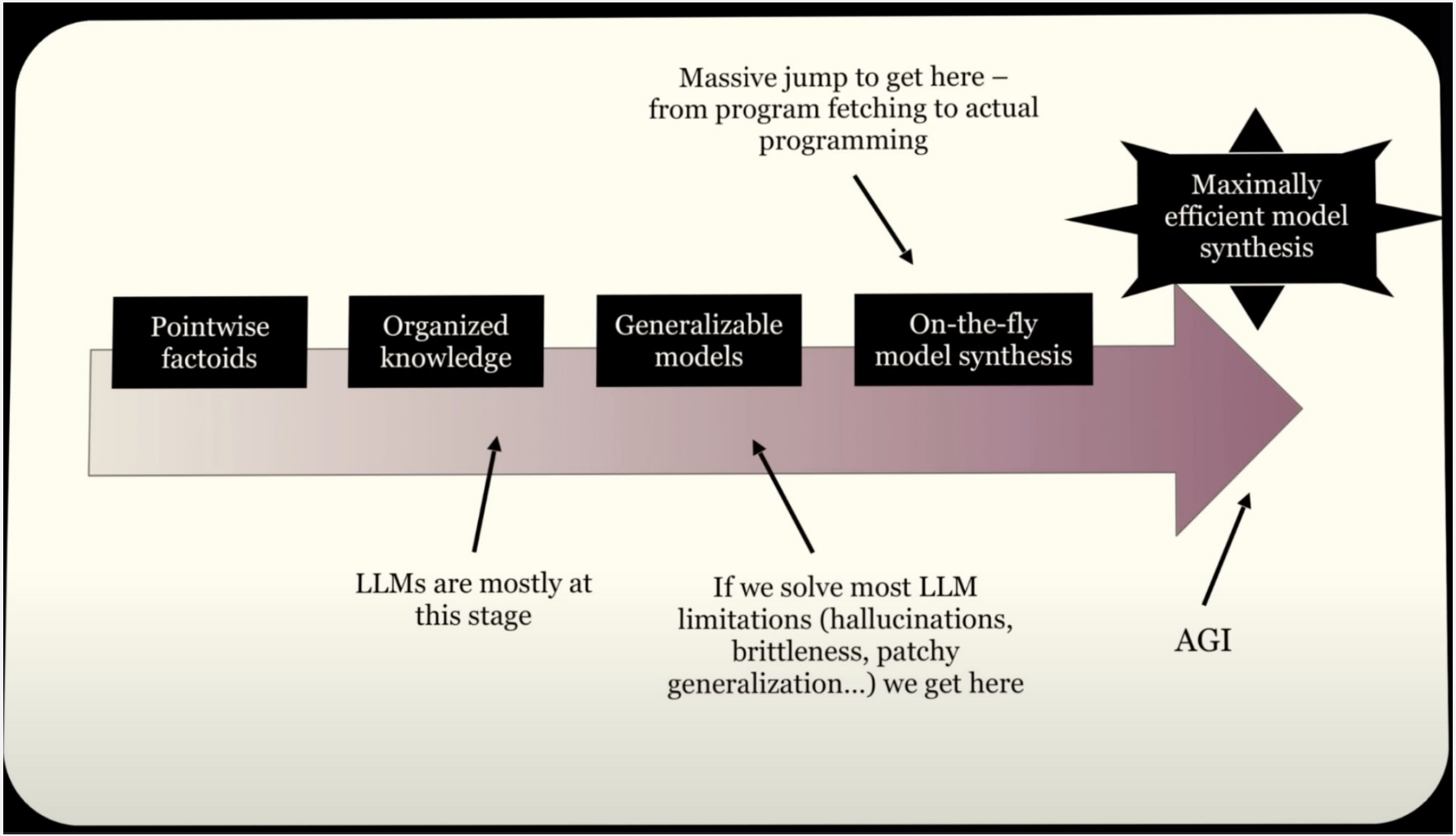
To generalizable **models**...

Even more abstract!

```
def addition(x, y):
    if y == 0:
        return x
    return addition(x^y, (x&y) << 1)
```

Always return the right result even for never-seen-before digits!
Maximally high **operational area**


Maximally efficient model synthesis
Pointwise factoids
Organized knowledge
Generalizable models
On-the-fly model synthesis
LLMs are mostly at this stage


Massive jump to get here – from program fetching to actual programming
Maximally efficient model synthesis
Pointwise factoids
Organized knowledge
Generalizable models
On-the-fly model synthesis
LLMs are mostly at this stage
If we solve most LLM limitations (hallucinations, brittleness, patchy generalization…) we get here
AGI

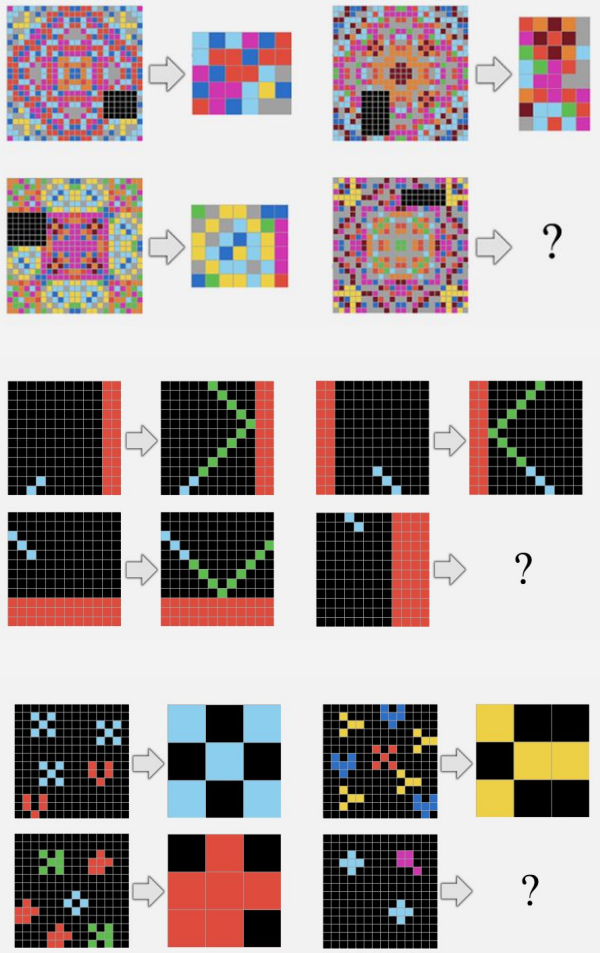
Chollet Criticisms: ARC Challenge

👉 Core Knowledge priors

👉 Focus on Reasoning and Abstraction

👉 Developer-Aware Generalization

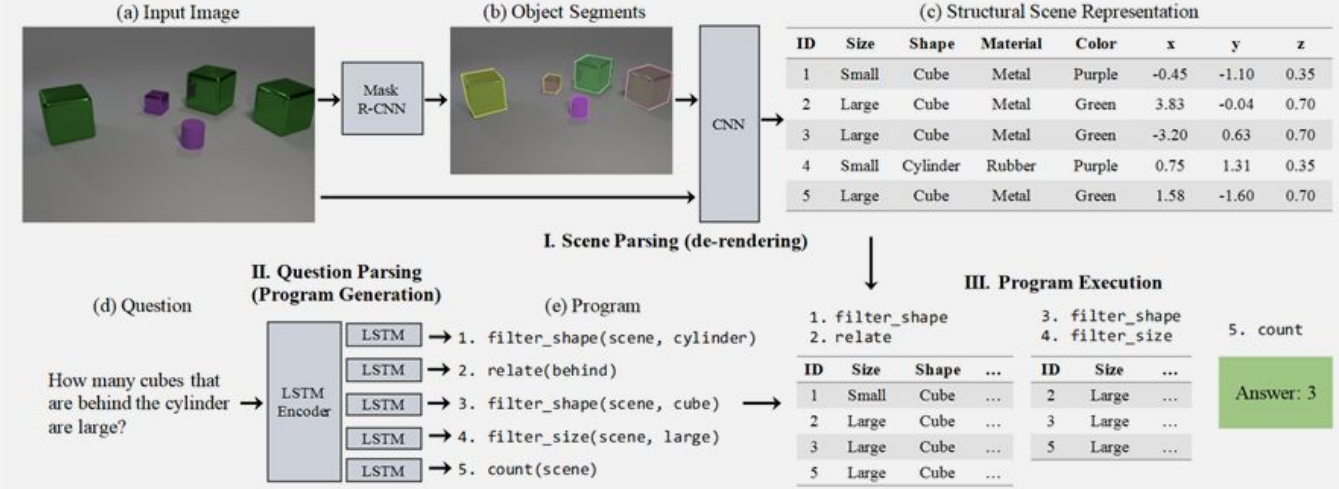
👉 Suggestion for Using Domain Specific Language



Marriage of Connectionism with Symbolism

One answer is to combine two mentioned paradigms to overcome these issues

Neurosymbolic methods : Gray Marcus

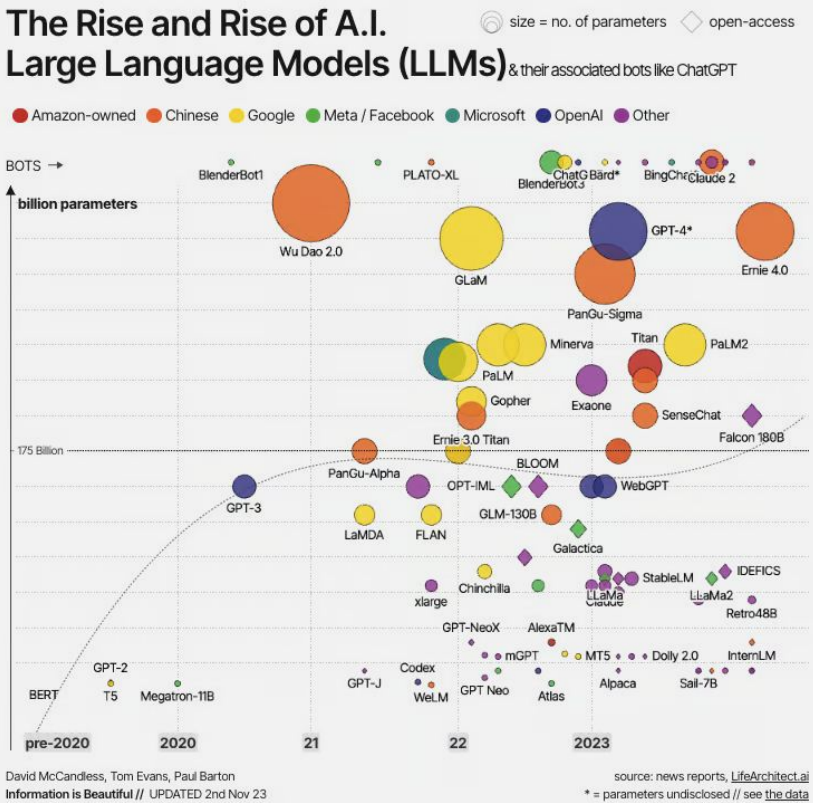


Bitter Lesson: Just scale Data and Compute, 2019

This is a big lesson. As a field, we still have not thoroughly learned it, as we are continuing to make the same kind of mistakes. We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that

1) AI researchers have often tried to build knowledge into their agents
2) this always helps in the short term, and is personally satisfying to the researcher
3) in the long run it plateaus and even inhibits further progress
4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

Lets Scale!


The Rise and Rise of A.I.
Large Language Models (LLMs) & their associated bots like ChatGPT
size = no. of parameters
open-access
Amazon-owned
Chinese
Google
Meta / Facebook
Microsoft
OpenAI
Other
BOTS
billion parameters
175 Billion
BlenderBot1
PLATO-XL
ChatG Bard*
BingCha
Claude 2
Wu Dao 2.0
GLaM
GPT-4*
Ernie 4.0
PanGu-Sigma
Minerva
Titan
PaLM2
PaLM
Gopher
Exaone
SenseChat
Falcon 180B
Ernie 3.0 Titan
BLOOM
PanGu-Alpha
OPT-IML
WebGPT
GPT-3
GLM-130B
LaMDA
FLAN
Galactica
StableLM
IDEFICS
Chinchilla
xlarge
LLaMa
LLaMa2
Retro48B
GPT-NeoX
AlexaTM
mGPT
MT5
Dolly 2.0
InternLM
GPT-2
Codex
GPT-J
T5
Megatron-11B
BERT
WeLM
GPT Neo
Atlas
Alpaca
Sail-7B
pre-2020
2020
21
22
2023
David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 2nd Nov 23
source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data

What is LLM?

Next token prediction

GPT-1, Improving Language Understanding by Generative Pre-Training

GPT-2, Language Models are Unsupervised Multi-task Learners

GPT-3, Language Models are Few-Shot Learners

InstructGPT, Aligning language models to follow instructions

ChatGPT, RLHF, Business Idea, More Data and Fund

Next Token Prediction Game

Baba alitoa maji.

Simba ni mwindaji wa kulungu.

Nguruwe, kulungu na simba ni wanyama.

Simba ni mfalme wa msituni.

Wanyama kama nguruwe wanaishi msituni.

Uwindaji ni sawa na kufuata.

Simba msituni hakuweza kuniwinda, lakini kulungu aliniwinda.

Ole kwa mfungwa aliyesahaulika, mwindaji amekwenda lakini bado yuko kwenye mtego.

Kulungu anaishi .....

Simba hufuata .....

Nguruwe

Simba

Wanyama

Maji

Mwindaji

Msituni

Mfungwa

GPT-1: Improving Language Understanding by Generative Pre-Training

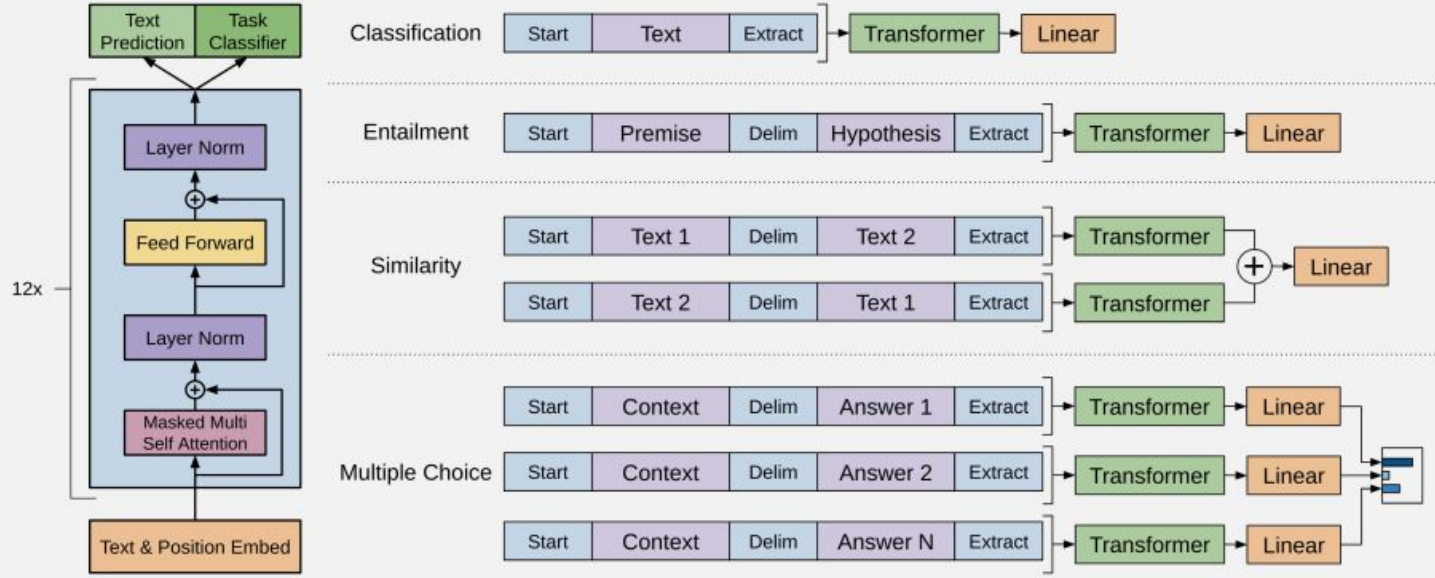


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2: Language Models are Unsupervised Multi-task Learners

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"
"I hate the word '**perfume**,"' Burr says. 'It's somewhat better in French: '**parfum**.'
If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

GPT3: Language Models are Few-Shot Learners

Just give some examples or instruction

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

InstructGPT: Aligning language models to follow instructions

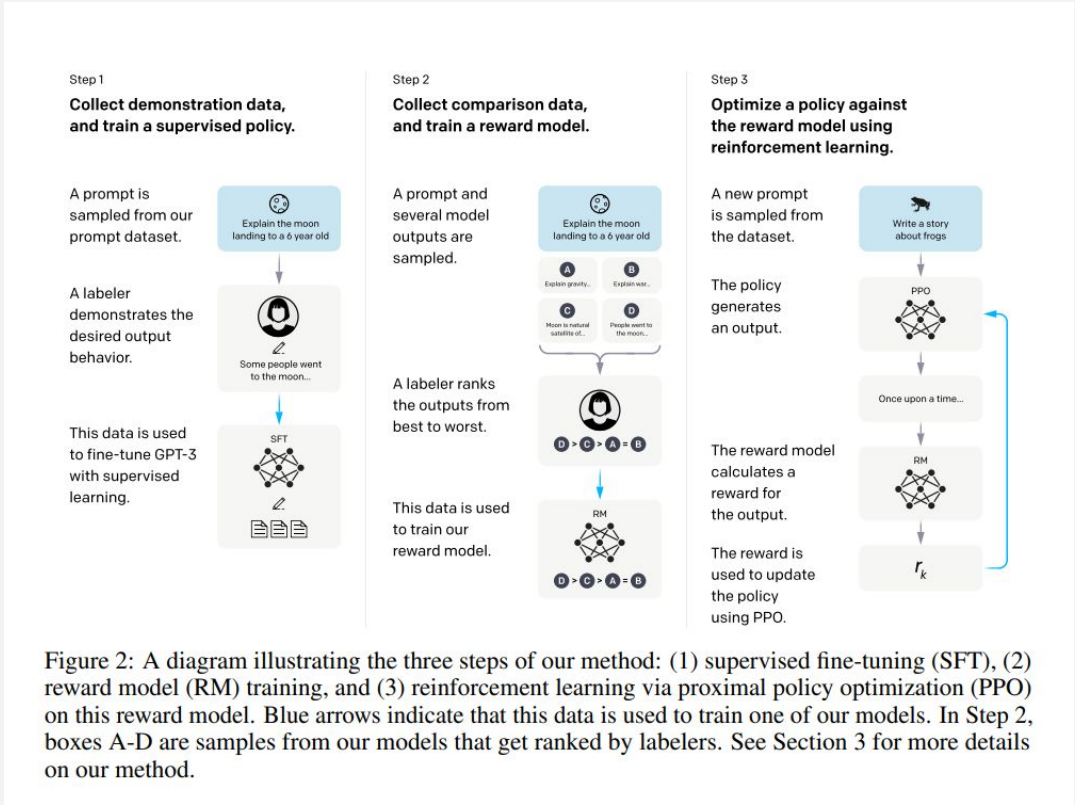


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

ChatGPT: chatgpt as conversational model

Scaling Laws for Neural Language Models, 2020

Model size, dataset size, and the amount of compute used for training (NBS)

Network width or depth have minimal effects within a wide range

Determine the optimal allocation of a fixed compute budget

Larger models are significantly more sample-efficient

Training very large models on a relatively modest amount of data

Scaling Laws for Neural Language Models, 2020

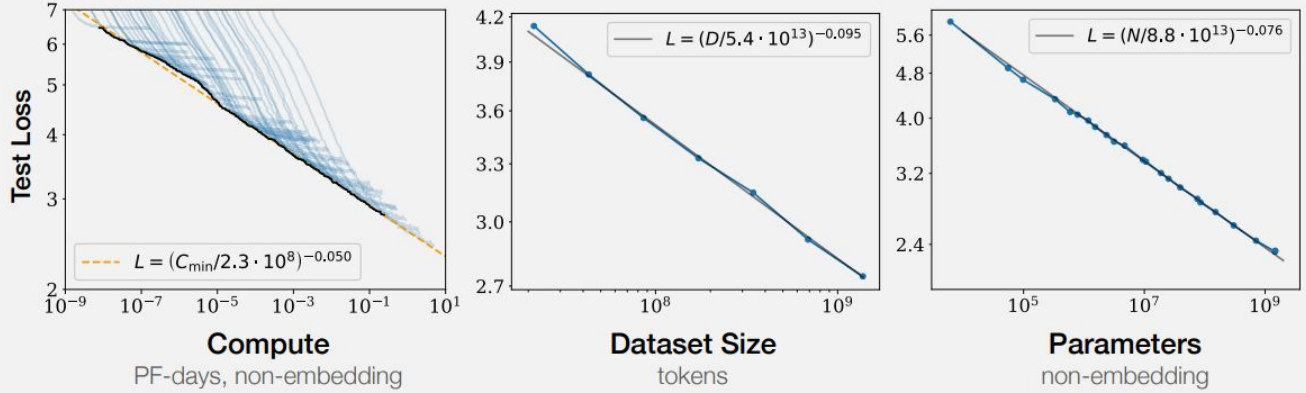


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling Laws for Neural Language Models, 2020

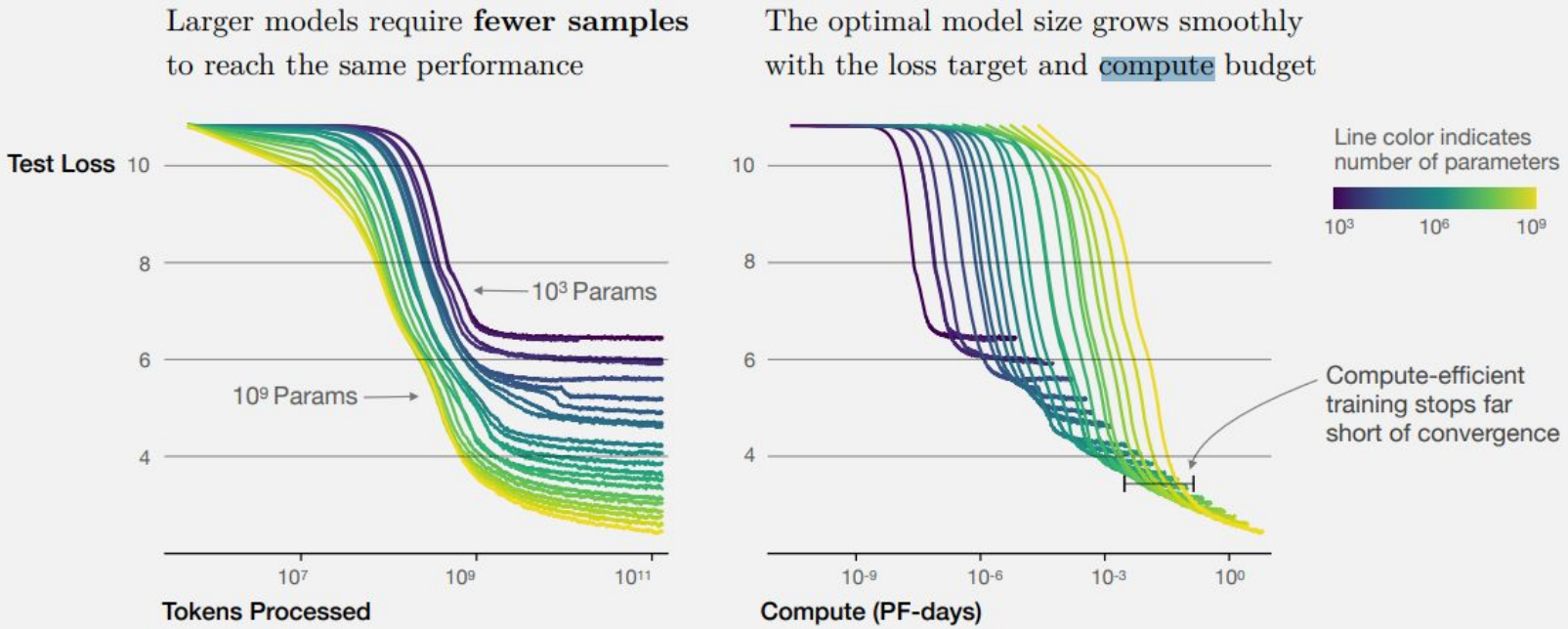
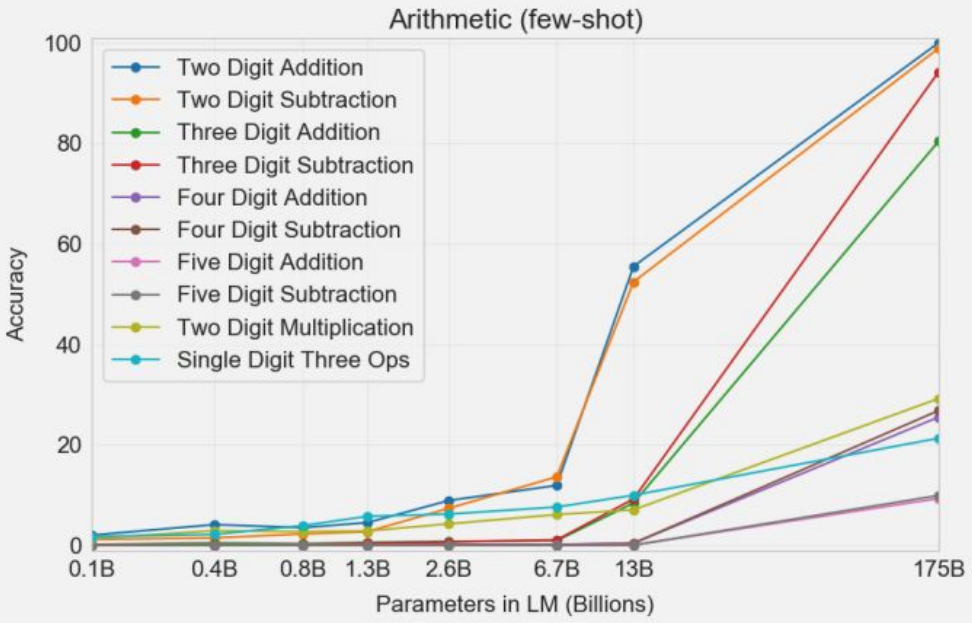


Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

GPT3, Arithmetic Tasks



Does GPT3 Reason or Memorize?

Q: What is 24 times 18? A: ___ *Model:* 432 ✓
Q: What is 23 times 18? A: ___ *Model:* 462 ✗

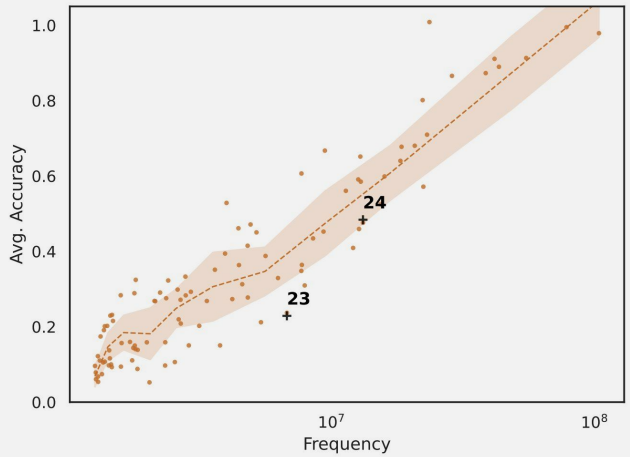


Figure 1. **Multiplication Performance:** Plot of GPT-J-6B's 2-shot accuracy on multiplication (averaged over multiple multiplicands and training instances) against the frequency of the equation's first term in the pretraining corpus. Each point represents the average performance for that term (e.g., 24) multiplied by numbers 1-50 and 5 choices of random seeds. As in the example, the performance difference for the numbers 24 and 23 is more than 20%. We find a strong correlation between accuracy and frequency.

Y. Razeghi, R. L. Logan, M. Gardner, και S. Singh, 'Impact of Pretraining Term Frequencies on Few-Shot Reasoning'. arXiv, 2022

Reasoning benchmarks, GSM8K

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Figure 1: Three example problems from GSM8K. Calculation annotations are highlighted in red.

Reasoning benchmarks, MATH

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

1. A 6-sided die is weighted so that the probability of any number being rolled is proportional to the value of the roll. (So, for example, the probability of a 2 being rolled is twice that of a 1 being rolled.) What is the expected value of a roll of this weighted die? Express your answer as a common fraction.
2. The square of 15 is 225. The square of what other number is 225?
3. Find the sum of all values of x such that $|x - 1| = 7$.
4. The parabolas defined by the equations $y = -x^2 - x + 1$ and $y = 2x^2 - 1$ intersect at points (a, b) and (c, d), where $c \geq a$. What is $c - a$? Express your answer as a common fraction.
5. If $a = 8$, what is the value of $\left(16\sqrt[3]{a^2}\right)^{\frac{1}{3}}$?
6. Let $p(x)$ be a cubic polynomial such that $p(2) = 0, p(-1) = 0, p(4) = 6$, and $p(5) = 8$. Find $p(7)$.
7. Let S be the set of complex numbers of the form $a + bi$, where a and b are integers. We say that $z \in S$ is a unit if there exists a $w \in S$ such that $zw = 1$. Find the number of units in S.
8. Find the remainder when $1 + 2 + 2^2 + 2^3 + \cdots + 2^{100}$ is divided by 7.
9. The length of a rectangle is $3x + 10$ feet and its width is $x + 12$ feet. If the perimeter of the rectangle is 76 feet, how many square feet are in the area of the rectangle?
10. A European train compartment has six seats. Four of the seats are broken. Wilhelm needs to fill out a form to indicate that there are broken seats. If he randomly checks off four of the seats in the diagram, what is the probability that he marked the correct seats? Express your answer as a common fraction.
11. We have a triangle $\triangle ABC$ where $AC = 17$, $BC = 15$, and $AB = 8$. Let M be the midpoint of AB. What is the length of CM?
12. If n gives a remainder of 3 when divided by 7, then what remainder does $2n + 1$ give when divided by 7?

Reasoning benchmarks, GPQA Diamond

Quantum Mechanics

Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?

A) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$

B) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$

C) $E(\rho) = (1-p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$

D) $E(\rho) = (1-p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$

Reasoning benchmarks, Others

ARC-AGI, Game of 24, Creative writing, GSM-Hard, StrategyQA, Coin Flip, CLEVRER ...

Prompting to solve System-2 tasks with LLMs 2023-2024

Complexity class: NP (Nondeterministic Polynomial Time)

Decision problems; the answer “yes” to the input has a proof verifiable in Poly. time

Or a non-deterministic Turing machine can solve it in Poly. time

co-NP: the answer “no” has such a property

NP-Complete: hardest problems in NP (all NP could be reduced to these problems)

What if we input an NP-Complete problem to an LLM?

Prompting to solve System-2 tasks with LLMs 2023-2024

Epic Failures in certain Reasoning and Planning Tasks.

Several complex/reasoning benchmarks (GSM8K, Game of 24, ...)

Simple Fixes Emerged **very quickly** (CoT, and ToT)

Try to see what is shared among these.

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

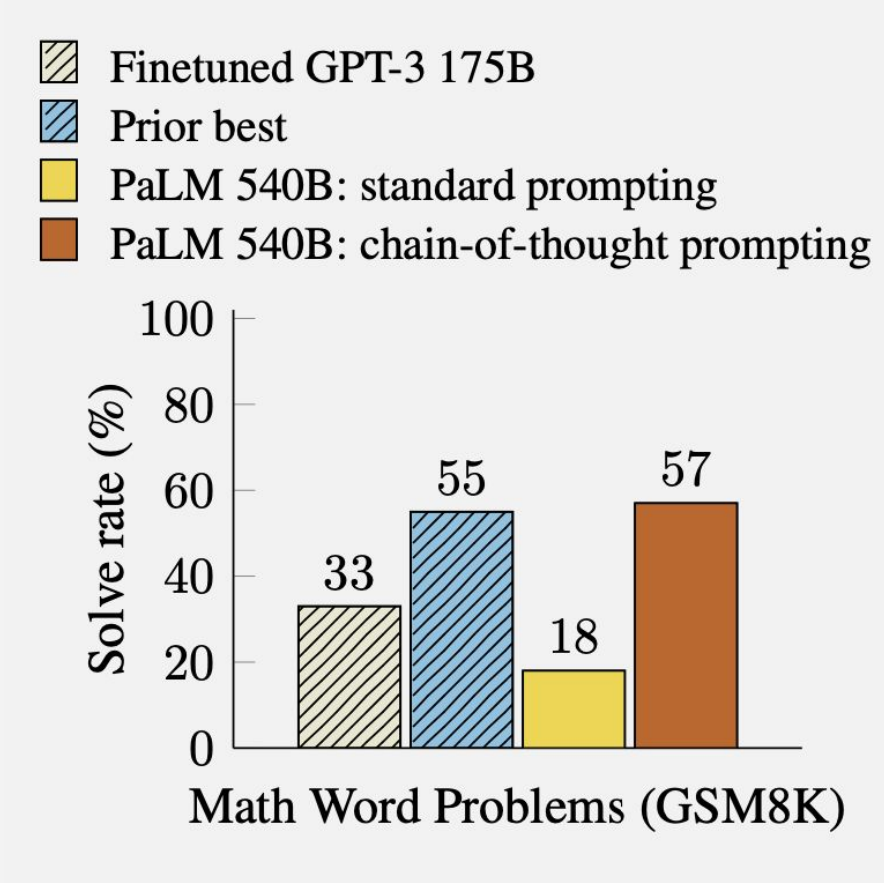
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

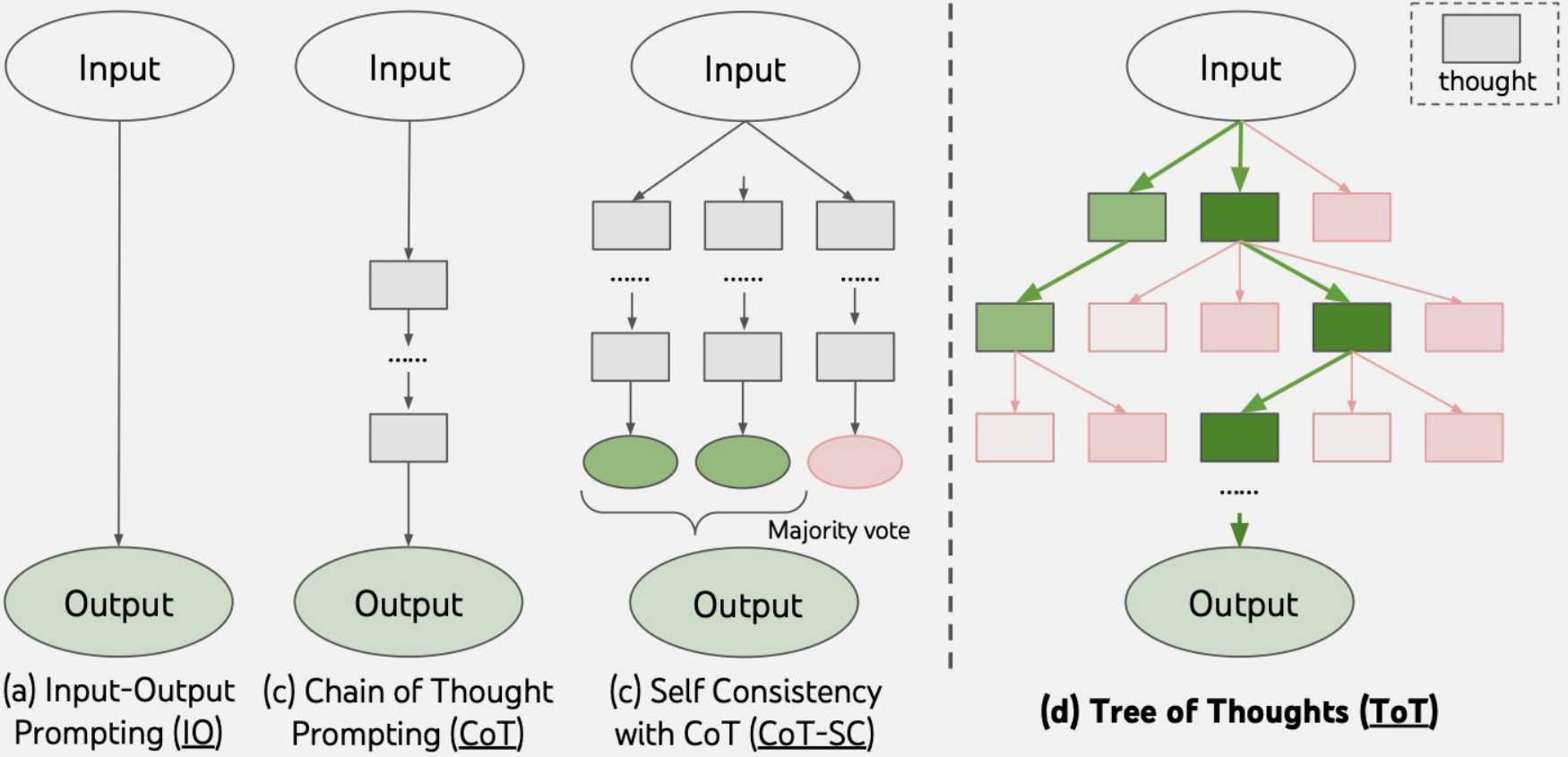
Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

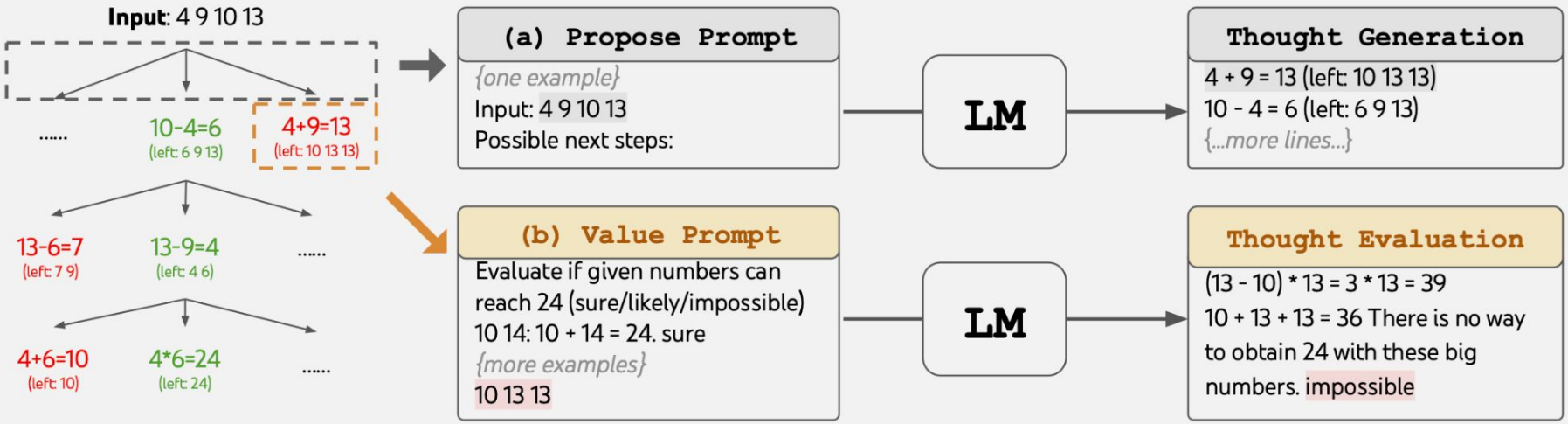
Courtesy: J. Wei, et al Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022.



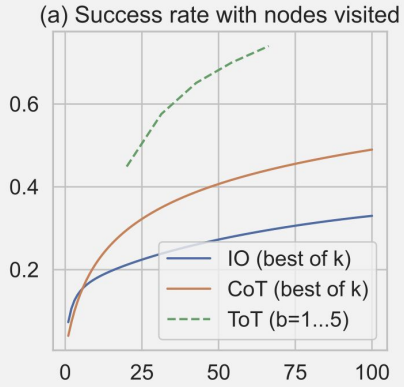
Courtesy: J. Wei, et al Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022.



Courtesy: S. Yao, et al Tree of Thoughts: Deliberate Problem Solving with Large Language Models, 2023.



Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC (k=100)	9.0%
ToT (ours) (b=1)	45%
ToT (ours) (b=5)	**74%**
IO + Refine (k=10)	27%
IO (best of 100)	33%
CoT (best of 100)	49%



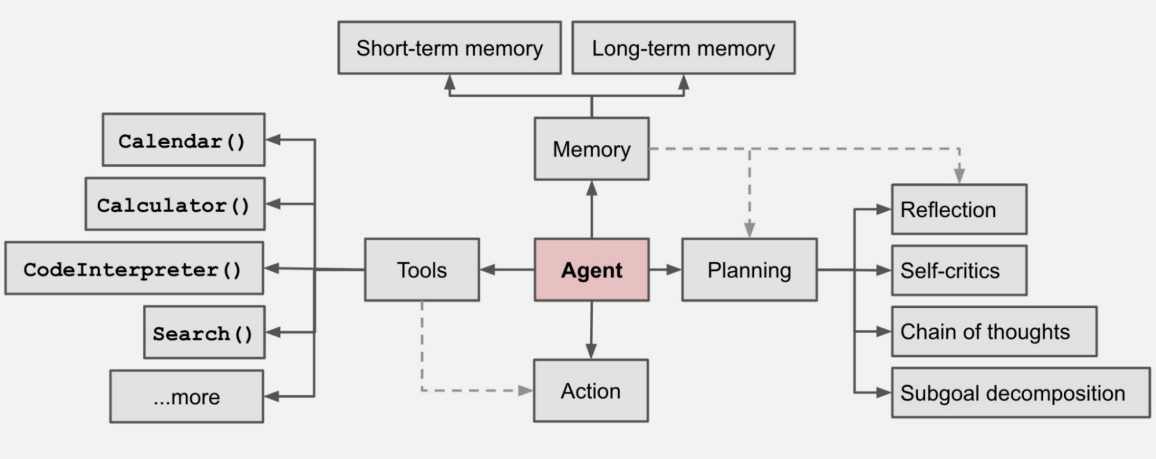
Courtesy: S. Yao, et al Tree of Thoughts: Deliberate Problem Solving with Large Language Models, 2023.

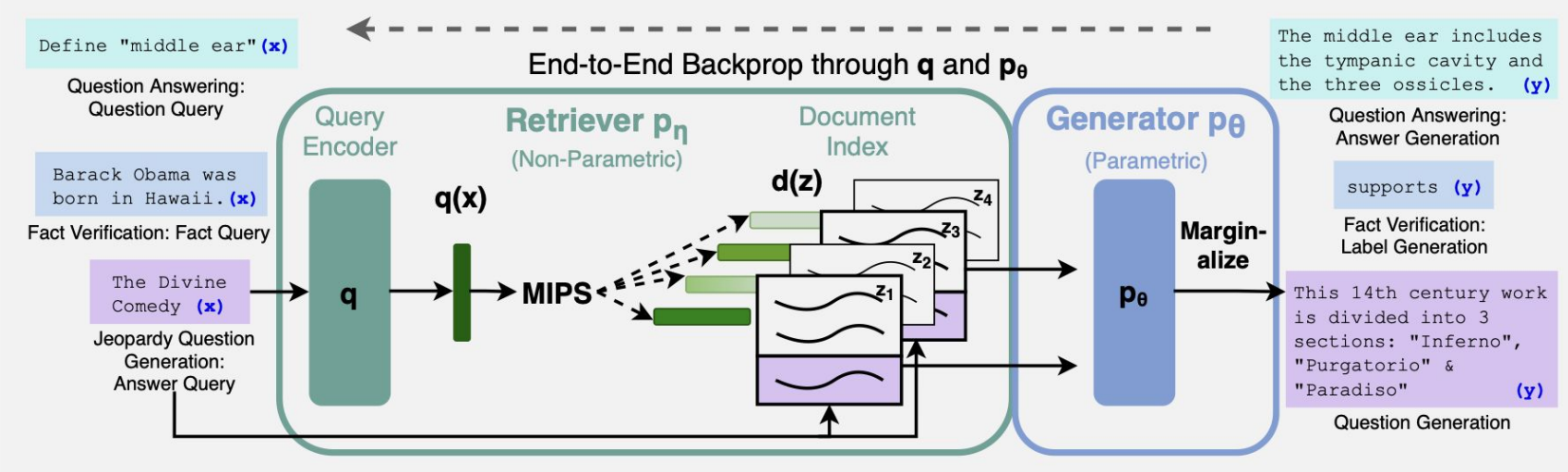
Improving LLMs Agency to solve System-2 tasks 2023-2025

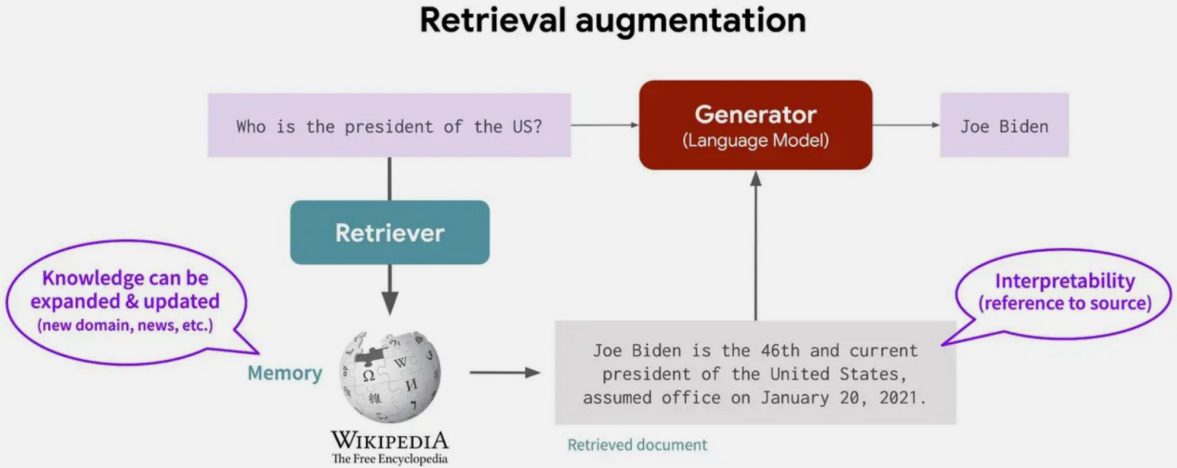
Empower the LLM by working with the world at the inference time.

Retrieval Augmented Generation (RAG)

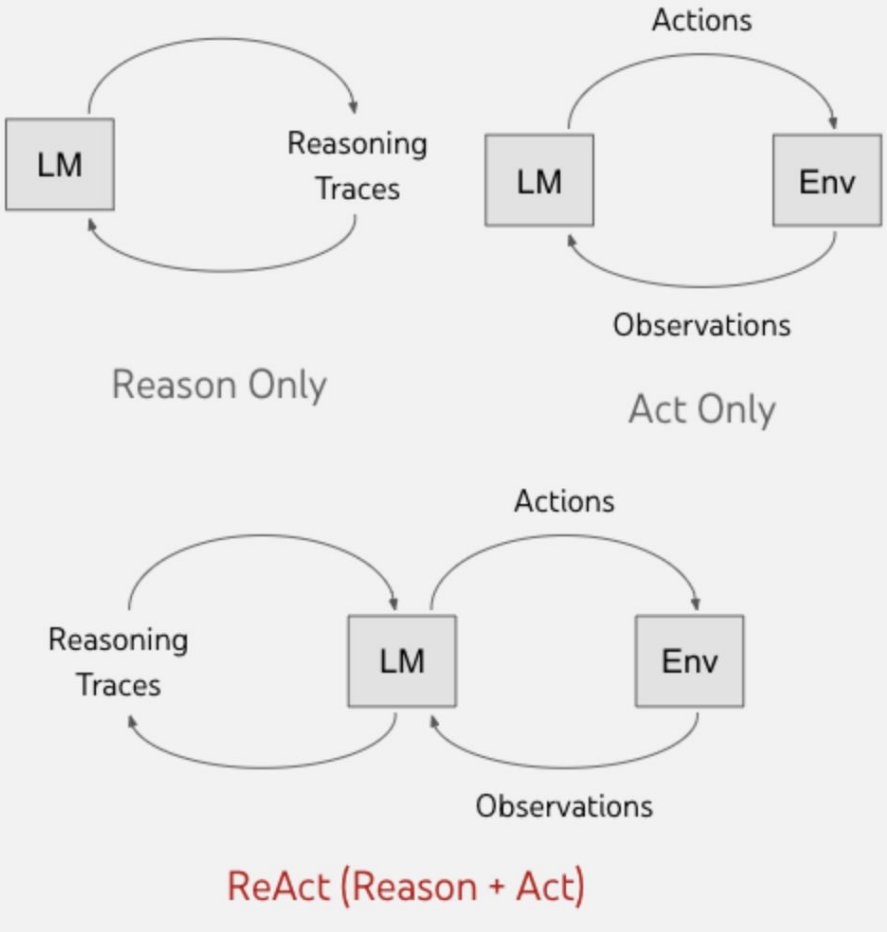
LLM Agents



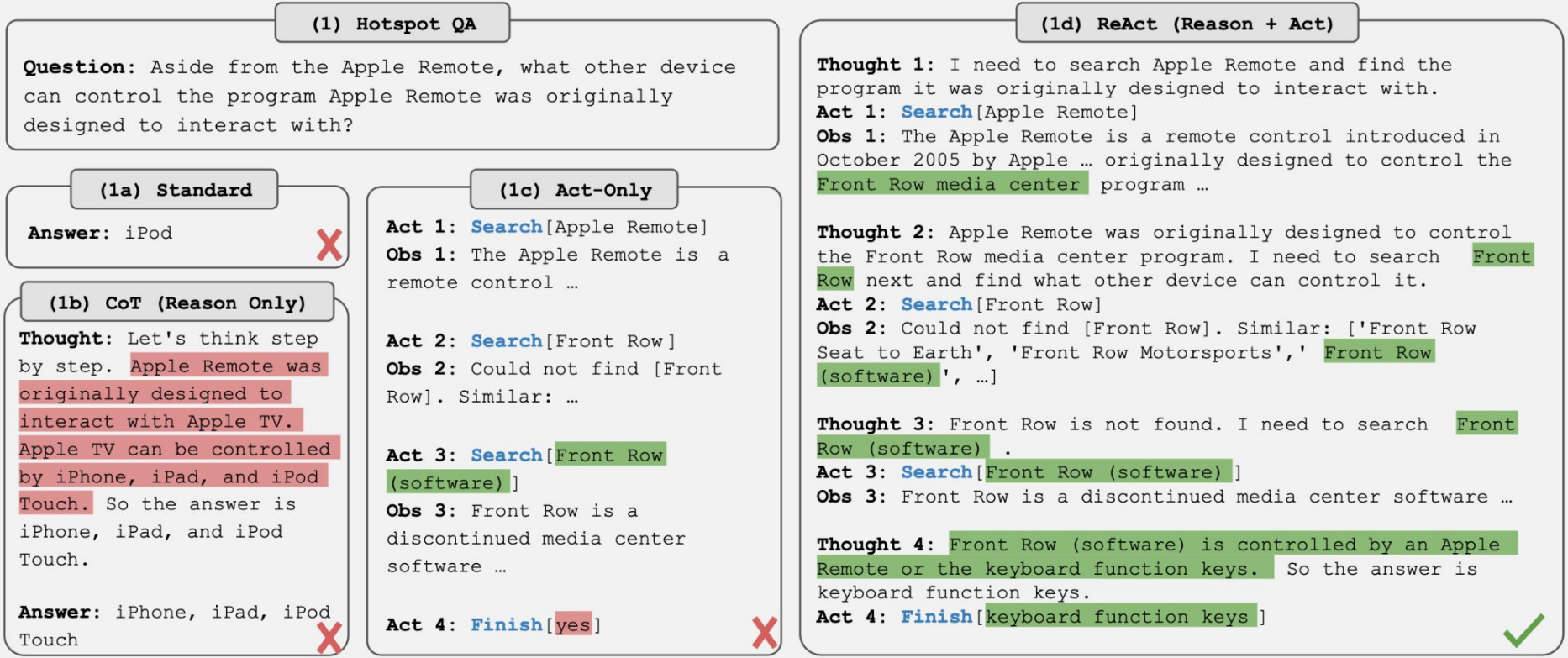




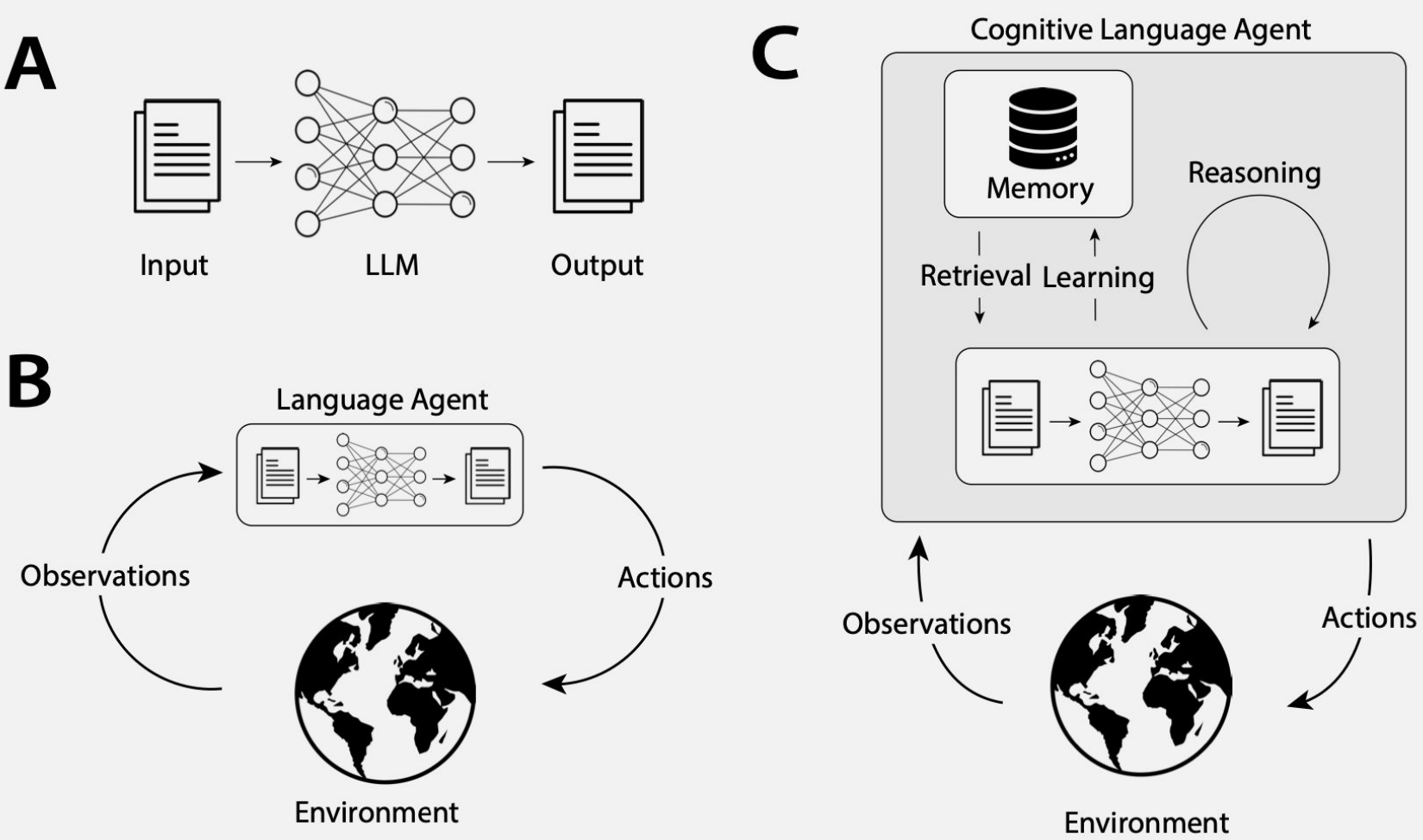
Courtesy: P. Lewis, et al Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020



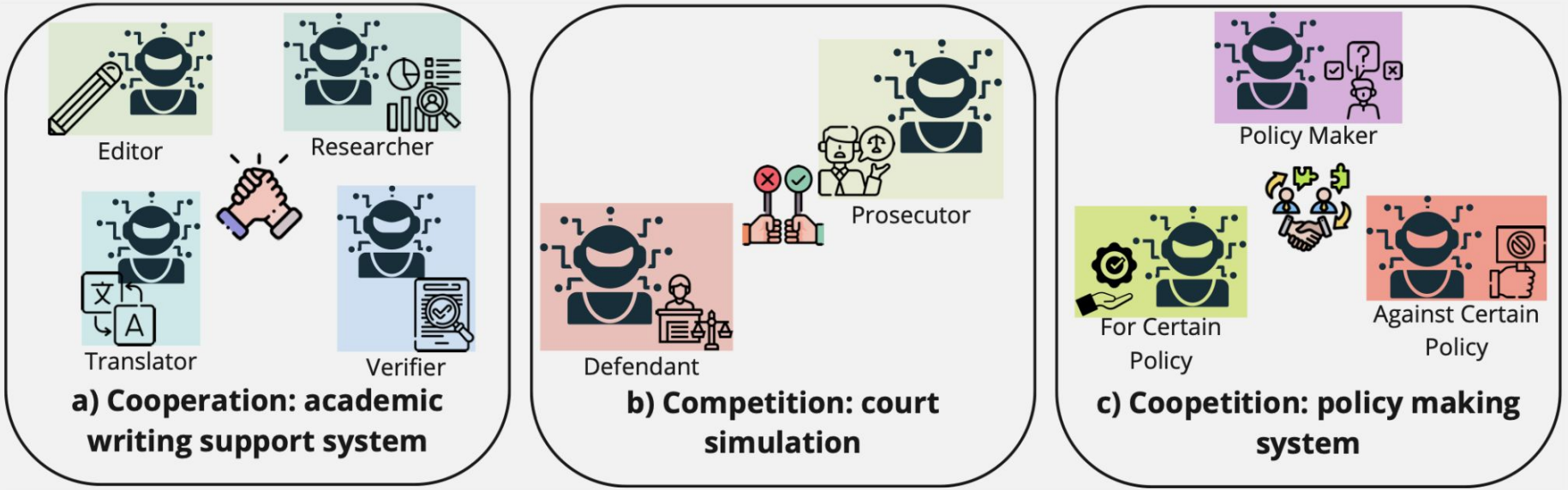
Courtesy: S. Yao, et al ReAct: Synergizing Reasoning and Acting in Language Models, 2023.



Courtesy: S. Yao, et al ReAct: Synergizing Reasoning and Acting in Language Models, 2023.



Courtesy: T. Sumers, et al Cognitive Architectures for Language Agents, 2024.



Courtesy: K. Tran, et al Multi-Agent Collaboration Mechanisms: A Survey of LLMs, 2025.

The relationship between agency and reasoning

Consider CodeForces, Where you have problem and test cases

As an agent, you can take actions on the world (during inference) and get feedback.

Through feedback, you can scale your calculations and get closer to the correct answer.

Reasoning helps agency better, and agency helps reasoning better.

Data is the new fossil fuel!

Data is the fossil fuel of AI.

We've achieved peak data and there will be no more.

The 2010s were the age of scaling, now we're back in the age of wonder and discovery once again. Everyone is looking for the next thing. Scaling the right thing matters more now than ever.

Test Time Scaling (2024-...)

Recall the bitter lesson.

Data has already been used to its maximum capacity.

What other factor has remained to be scaled? ... test time

Recall the search vs. learning paradigms in AI.

We have already scaled the test time through CoT and ToT, Nah?

The scaling is not necessarily controllable.

Looking to scale this in a more controllable and efficient fashion.

Search, SFT, RLHF, RL

Pacman wants to go to G

If the pacman goes to the hole cells, dies.

The wind blows around the hole.

Pacman only percepts its current state.

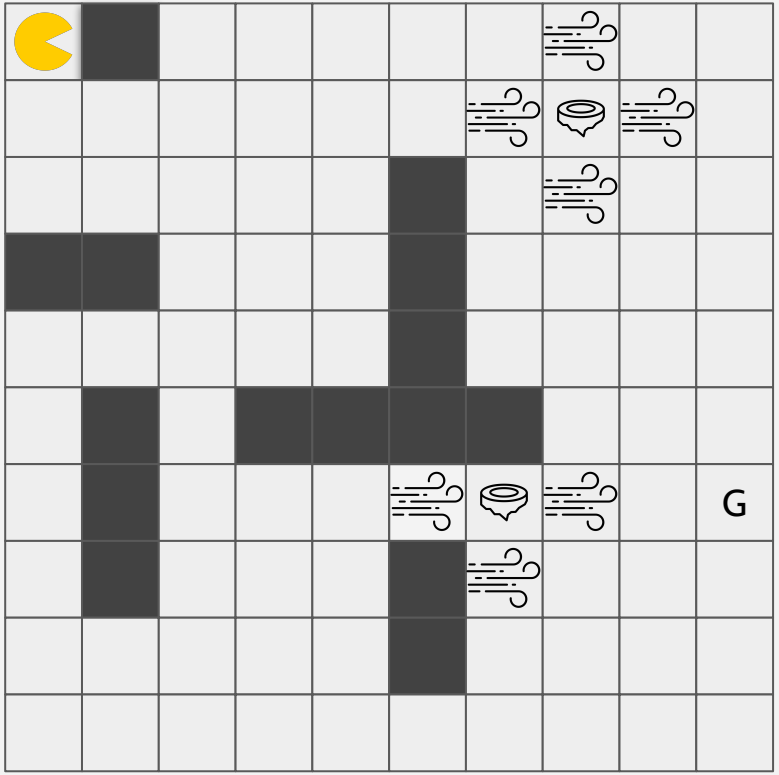
We want steps.

Search: Find the solution only for this maze.

But what if

the agent doesn't have access to the map?

we want to also solve other mazes?



Search, SFT, RLHF, RL

But what if we want to solve another maze?

All scenarios come from a common manifold.

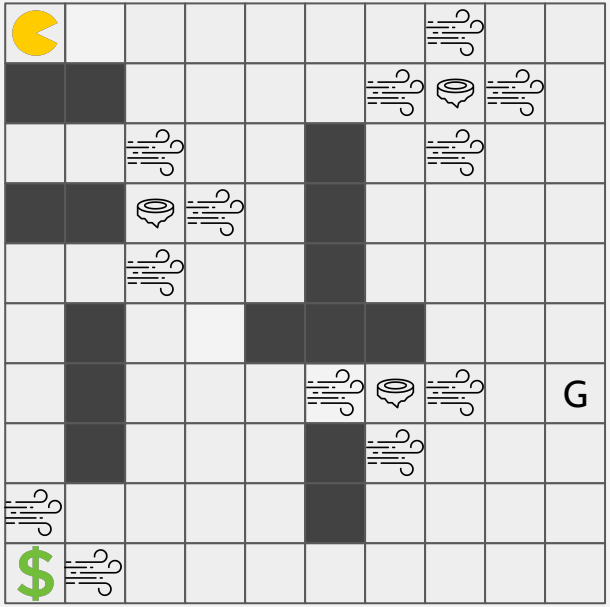
SFT: Train on the steps of the previous person who reached the destination on a variety of mazes.

RLHF: train on approximate function that says, is this sequence a good sequence or not?

RL: Let Pac-Man go into the maze himself and get the actual rewards. (what is reward ???)

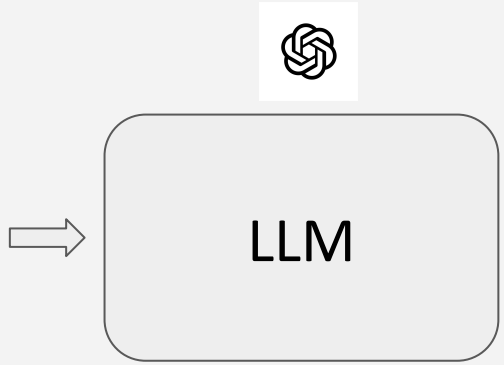
What are your ideas for this problem?

Search: Heuristic Function + Search Strategy



Multi-Step Reasoning Problem

Arash is planning a rectangular garden. The length of the garden is 12 meters, and the width is 5 meters. He wants to put a fence around the garden, and each meter of fencing costs $3. How much will the total cost be for fencing the entire garden?



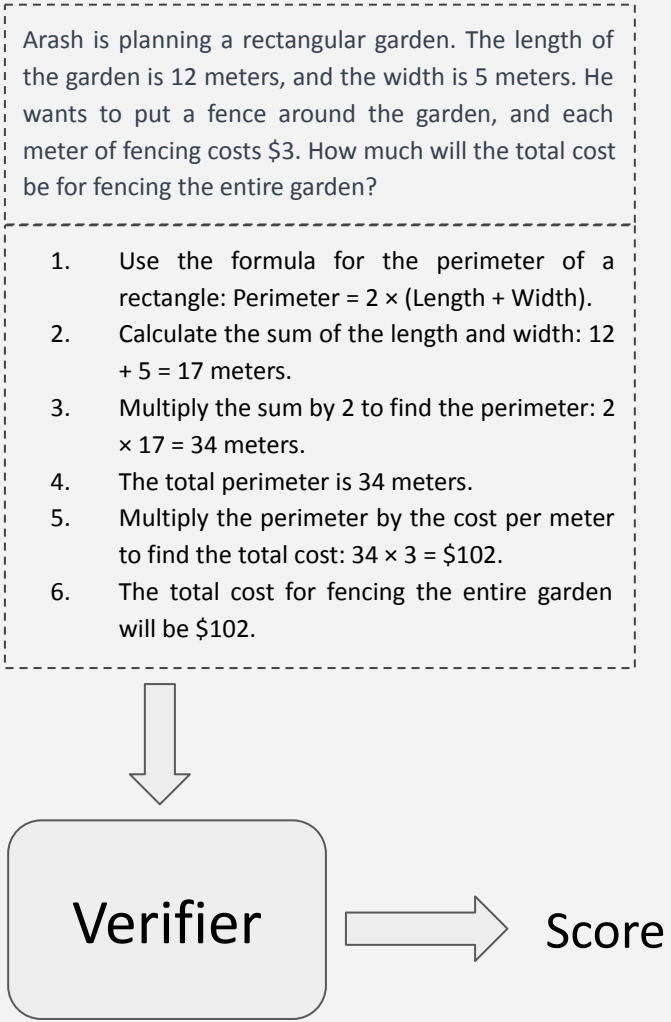
1. Use the formula for the perimeter of a rectangle: Perimeter = 2 × (Length + Width).
2. Calculate the sum of the length and width: 12 + 5 = 17 meters.
3. Multiply the sum by 2 to find the perimeter: 2 × 17 = 34 meters.
4. The total perimeter is 34 meters.
5. Multiply the perimeter by the cost per meter to find the total cost: 34 × 3 = $102.
6. The total cost for fencing the entire garden will be $102.

First Idea: Reward Model

💡 Training reward models to discriminate between desirable and undesirable outputs

➡️ Then:

- 🤖 Use in reinforcement learning pipeline
- 🎯 Or perform search via repeated sampling



Training Verifiers to Solve Math Word Problems

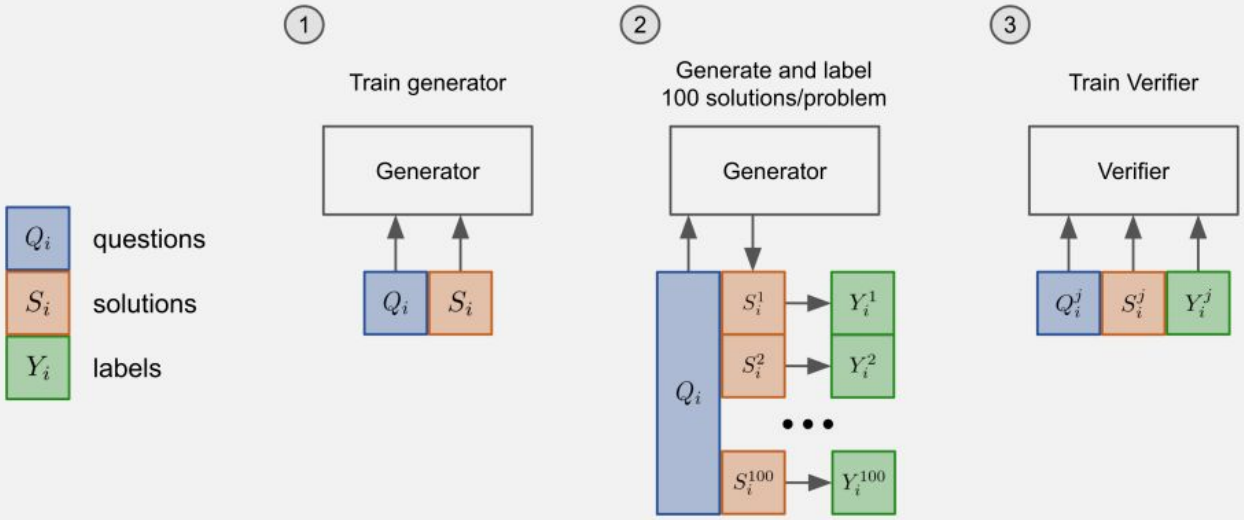


Figure 4: A diagram of the verification training pipeline.

Courtesy: K. Cobbe, et al "Training Verifiers to Solve Math Word Problems," 2021

Training Verifiers to Solve Math Word Problems (cont.)

Fine-tuning:

- uses the same language modeling objective as in GPT.
- At test time, we judge performance by autoregressively sampling a single low temperature solution and checking whether the final answer is correct.

Verification:

- sampling multiple high temperature solutions, assigning each solution a score, and outputting the highest ranked solution.

Training Verifiers to Solve Math Word Problems (cont.)

👉 test@N : correctly at least once when model makes N separate guesses.

👉 Training solutions are labeled as correct or incorrect based solely on whether they reach the correct final answer (false positives)

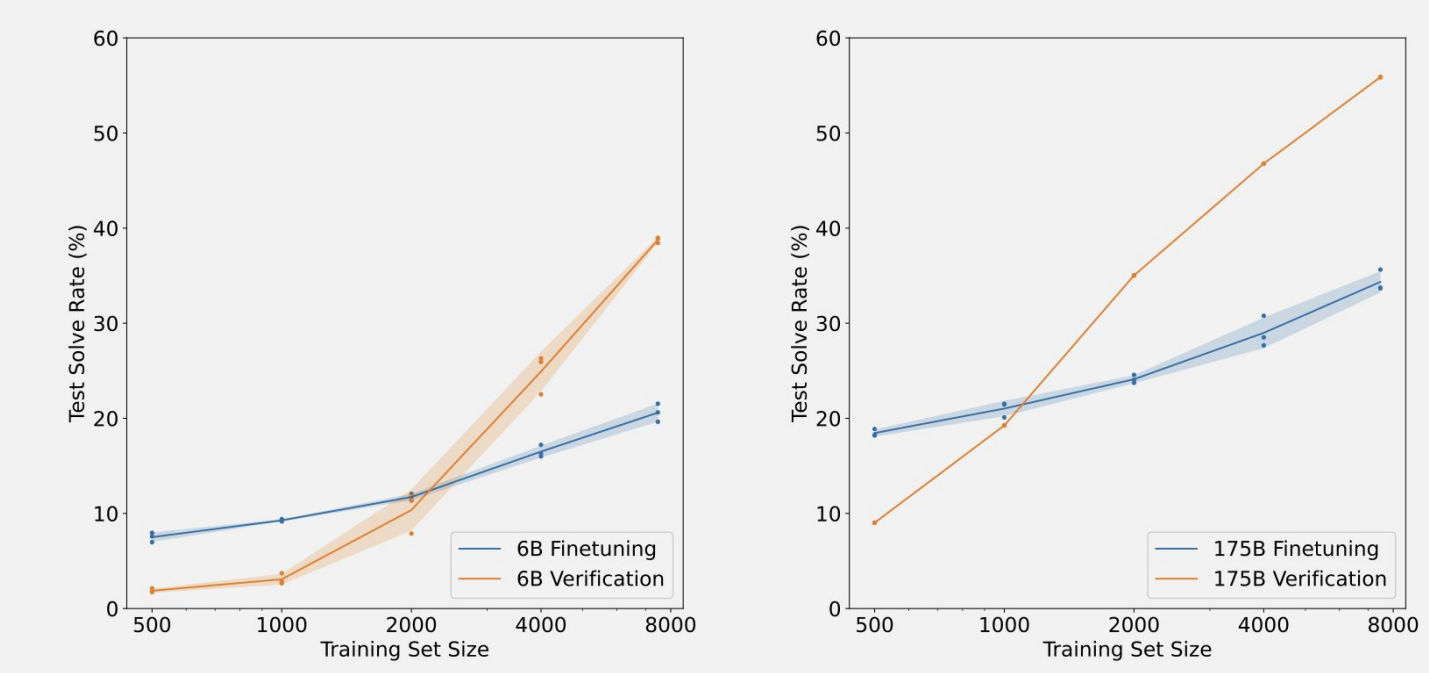
👉 How to Train Verifier?

Fine-tune generator for 2 epochs.

Sample 100 completions from the generator and label as correct or incorrect.

Train a verifier for a single epoch on this dataset.

Even more effective than scaling the model size!

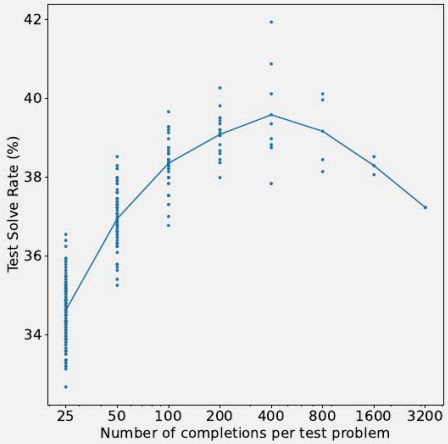


Courtesy: K. Cobbe, et al “Training Verifiers to Solve Math Word Problems,” 2021

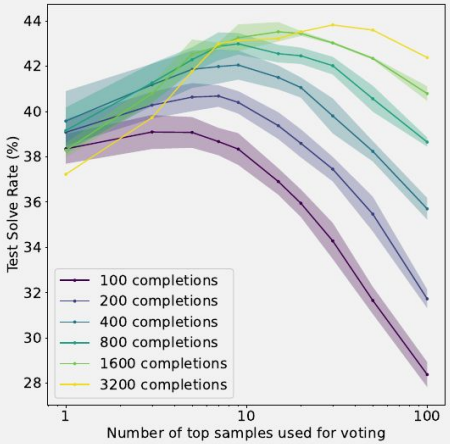
Training Verifiers to Solve Math Word Problems (cont.)

Why the accuracy dropped?!

Rank or Vote



(a) 6B verification test performance when given varying numbers of completions per problem to rank.



(b) 6B verification test performance when varying the number of top ranked samples allowed to vote on the answer.

Figure 7: Performance as the amount of test time compute varies.

Let's Verify Step by Step

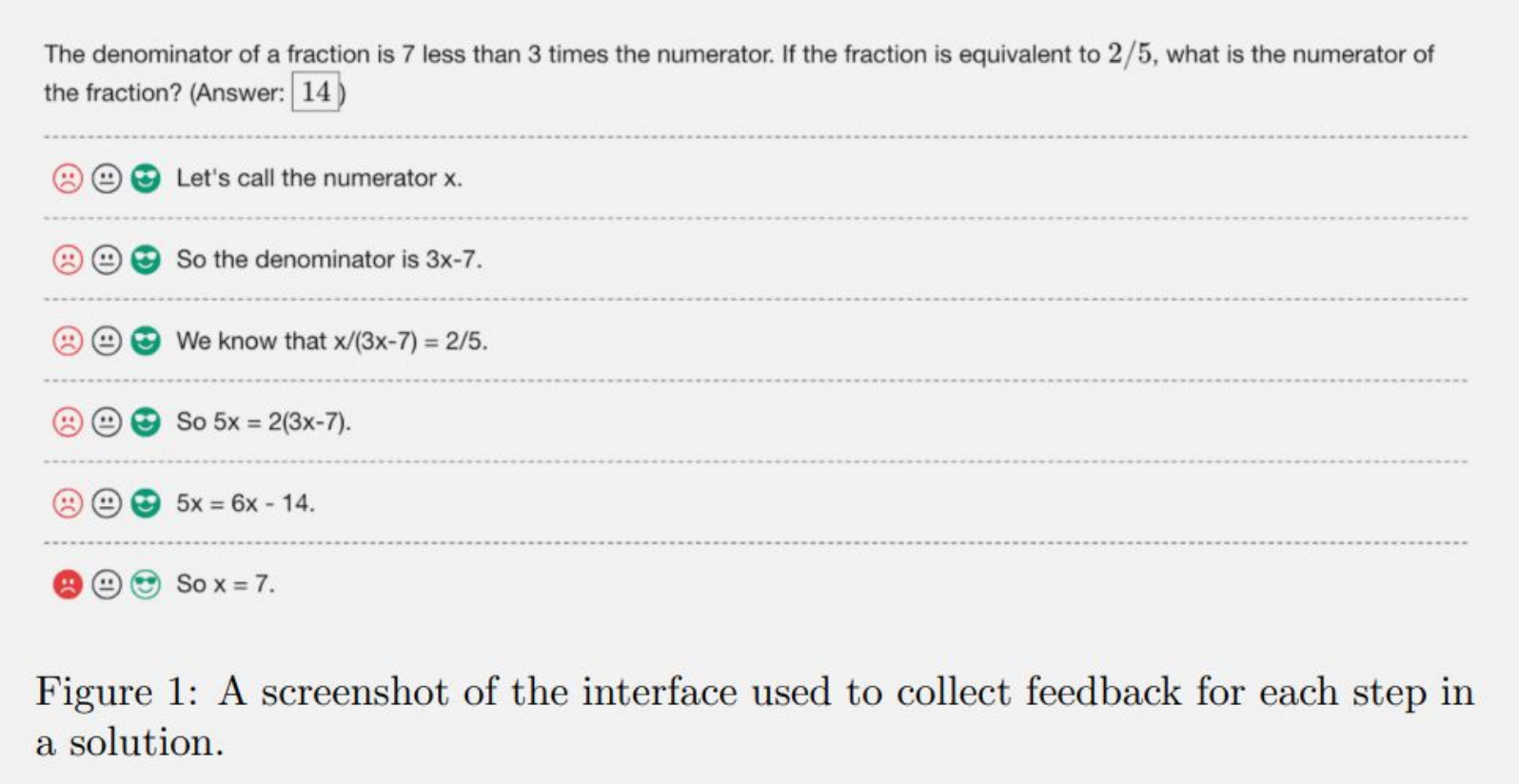


Figure 1: A screenshot of the interface used to collect feedback for each step in a solution.

Courtesy: H. Lightman, et al "Let's Verify Step by Step," 2023

Let's Verify Step by Step (cont.)

"we choose to surface convincing wrong-answer solutions",

also iteratively re-train our PRM using the latest data at several points in the data collection process.

Outcome-supervised Reward Models (ORM): Predict the solution correctness.

Process-supervised Reward Models (PRMs): Predict the step correctness.

PRM outputs

Let

$$x^8 + 3x^4 - 4 = p_1(x)p_2(x)\cdots p_k(x),$$

where each non-constant polynomial $p_i(x)$ is monic with integer coefficients, and cannot be factored further over the integers. Compute $p_1(1) + p_2(1) + \cdots + p_k(1)$.

This problem is about factoring a polynomial over the integers, which can be a tricky task.

One possible strategy is to look for rational roots of the polynomial, using the rational root theorem.

However, that would involve checking a lot of possible factors of -4, and there is no guarantee that the polynomial has any rational roots at all.

Another possible strategy is to use some algebraic tricks to simplify the polynomial or reduce its degree.

I notice that the polynomial has only even powers of x, so I can try to substitute $y = x^2$ and see what happens.

Then the polynomial becomes $y^4 + 3y^2 - 4$, which is a quadratic equation in y^2.

I can try to factor this quadratic equation by looking for two numbers that multiply to -4 and add to 3.

Those numbers are 4 and -1, so I can write $y^4 + 3y^2 - 4 = (y^2 + 4)(y^2 - 1)$.

Now I can substitute back $x^2 = y$ and get $x^8 + 3x^4 - 4 = (x^4 + 4)(x^4 - 1)$.

This looks promising, because I have reduced the degree of the factors from 8 to 4.

I can try to factor each of these factors further, using the difference of squares formula.

I get $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x^2 - 1)$.

I can apply the difference of squares formula again to the last factor and get $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x + 1)(x - 1)$.

Now I have factored the polynomial completely into monic linear and quadratic factors with integer coefficients.

These are the $p_i(x)$'s that the problem is asking for.

To find the sum of their values at $x = 1$, I just need to plug in $x = 1$ into each factor and add them up.

I get $p_1(1) + p_2(1) + \cdots + p_k(1) = (1^2 + 2)(1^2 - 2)(1^2 + 1)(1 + 1)(1 - 1)$.

Simplifying, I get $p_1(1) + p_2(1) + \cdots + p_k(1) = (3)(-1)(2)(2)(0)$.

Multiplying, I get $p_1(1) + p_2(1) + \cdots + p_k(1) = 0$.

Answer: 0

Let's Verify Step by Step: ORM vs PRM

	ORM	PRM	Majority Voting
% Solved (Best-of-1860)	72.4	**78.2**	69.6

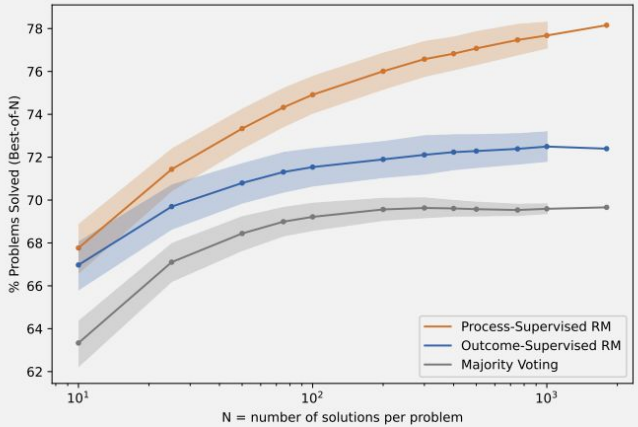
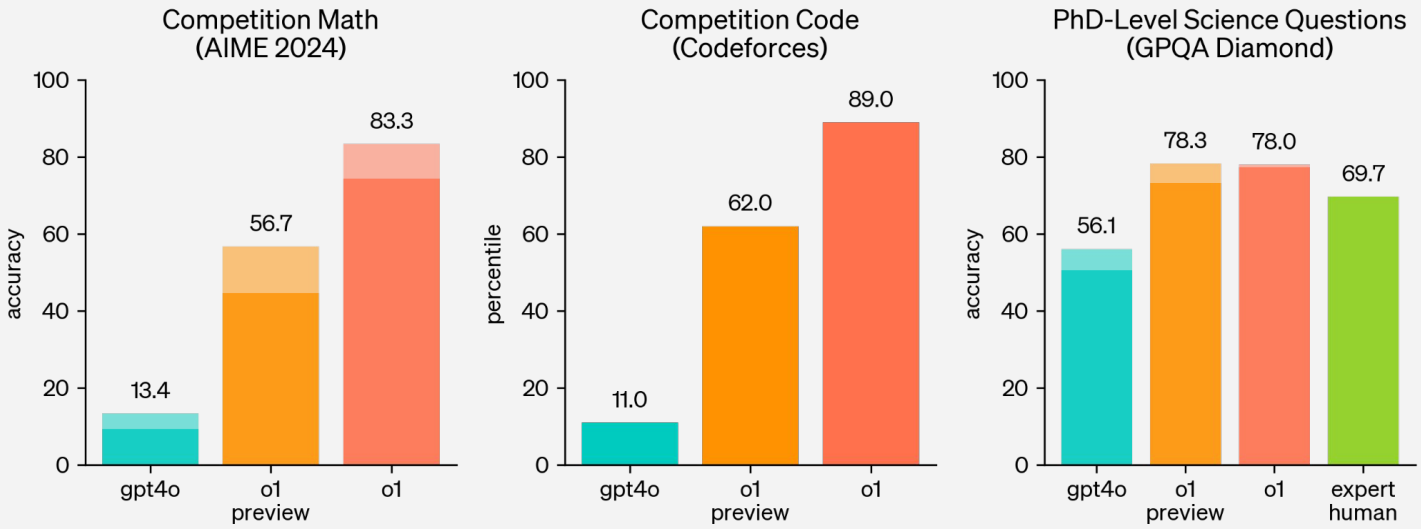


Figure 3: A comparison of outcome-supervised and process-supervised reward models, evaluated by their ability to search over many test solutions. Majority voting is shown as a strong baseline. For $N \leq 1000$, we visualize the variance across many subsamples of the 1860 solutions we generated in total per problem.

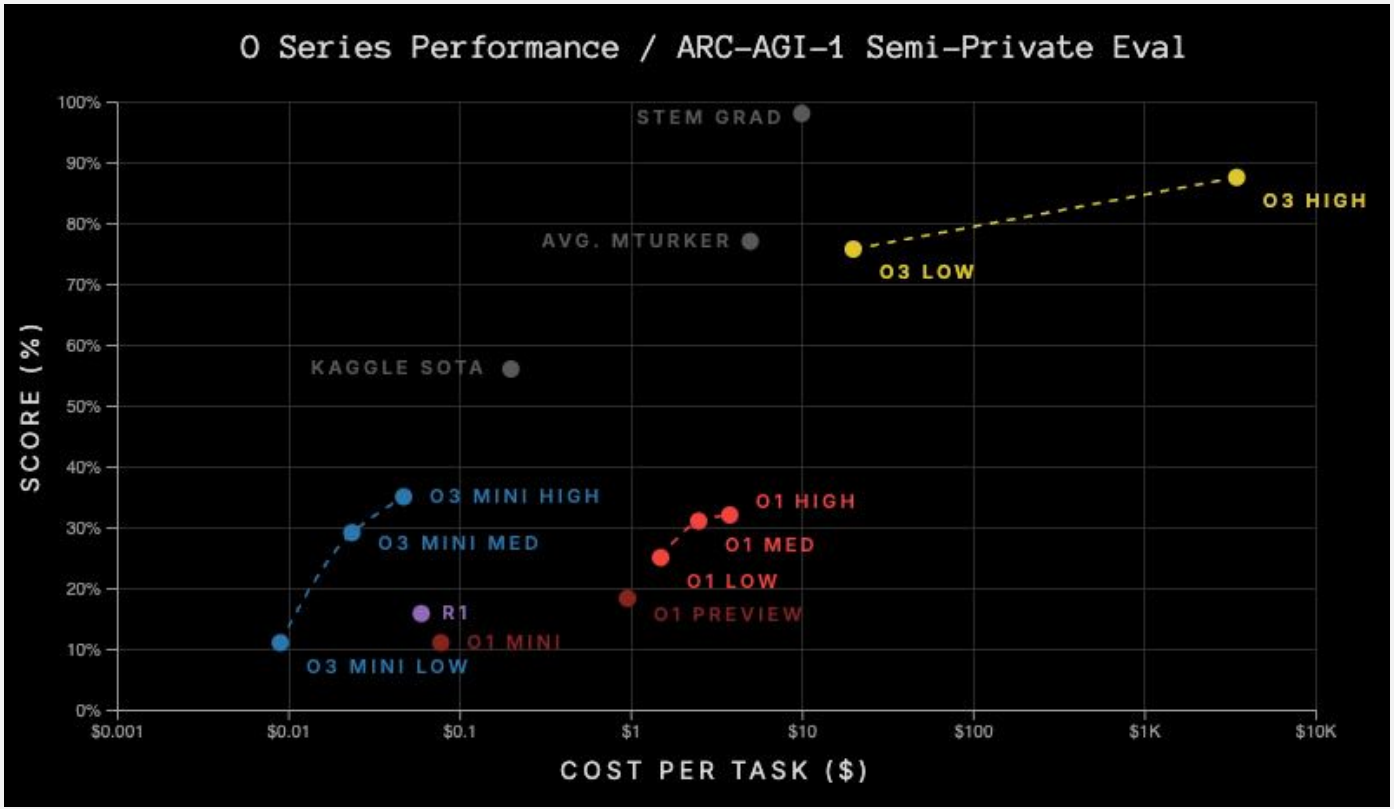
	ORM	PRM	Majority Vote	# Problems
AP Calculus	68.9%	**86.7%**	80.0%	45
AP Chemistry	68.9%	**80.0%**	71.7%	60
AP Physics	77.8%	**86.7%**	82.2%	45
AMC10/12	49.1%	**53.2%**	32.8%	84
Aggregate	63.8%	**72.9%**	61.3%	234

Table 1: We measure out-of-distribution generalization using recent STEM tests. We evaluate the outcome-supervised RM, the process-supervised RM, and majority voting using 100 test samples per problem.

O1


Competition Math
(AIME 2024)
accuracy
100
80
60
40
20
0
13.4
56.7
83.3
gpt4o
o1
preview
o1
Competition Code
(Codeforces)
percentile
11.0
62.0
89.0
gpt4o
o1
preview
o1
PhD-Level Science Questions
(GPQA Diamond)
accuracy
56.1
78.3
78.0
69.7
gpt4o
o1
preview
o1
expert
human

O3, RIP ARC-AGI



Verifiers Could be More Complex!

How do you value a search trajectory?

Feedback Modeling:

- Self-consistency
- ORM / PRM
- Compiler or Unit Test Samples
- Human
- Multi-Agent Debate
- Self-Critique
- Verbal-Based Feedback

Search Strategies

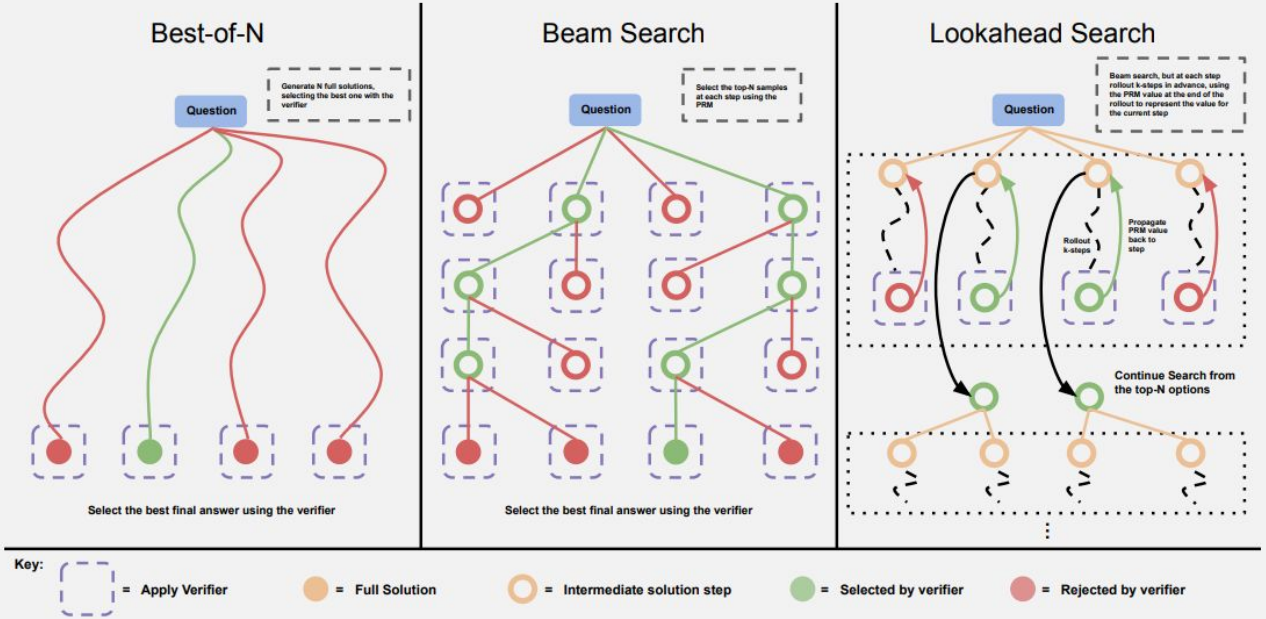


Figure 2 | ***Comparing different PRM search methods.*** **Left:** Best-of-N samples N full answers and then selects the best answer according to the PRM final score. **Center:** Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. **Right:** lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

Search Strategies can be More Complex!

How do you take your next action?

- Repeated Sampling
- Tree-Search
- Self-Correction

RLHF for Improvement Training

Both of verification and search strategies can be used as reward models for RLHF.

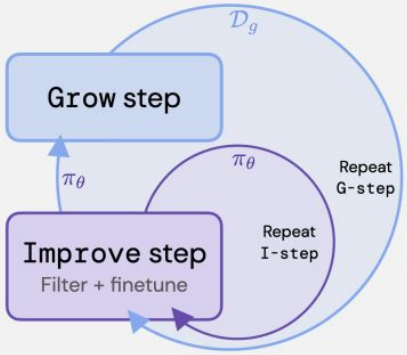


Figure 1 | ***ReST* method.** During `Grow` step, a policy generates a dataset. At `Improve` step, the filtered dataset is used to fine-tune the policy. Both steps are repeated, `Improve` step is repeated more frequently to amortise the dataset creation cost.

Large Language Monkeys: Scaling Inference Compute with Repeated Sampling

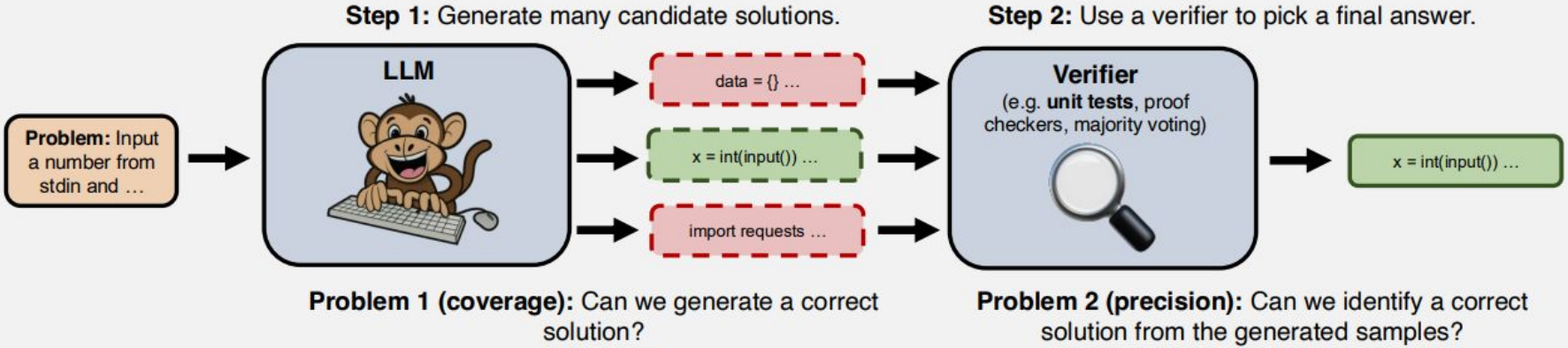


Figure 1: The repeated sampling procedure that we follow in this paper. 1) We generate many independent candidate solutions for a given problem by sampling from an LLM with a positive temperature. 2) We use a domain-specific verifier (ex. unit tests for code) to select a final answer from the generated samples.

Courtesy: B. Brown, et al "Large Language Monkey: Scaling Inference Compute with Repeated Sampling," 2024

Large Language Monkeys: Scaling Inference Compute with Repeated Sampling

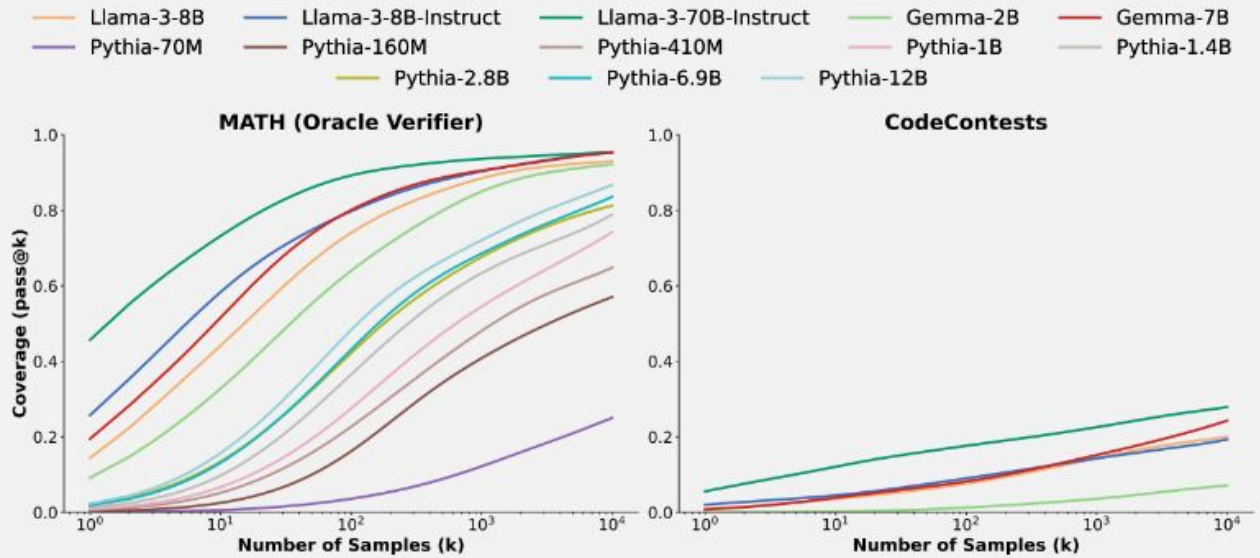


Figure 3: Scaling inference time compute via repeated sampling leads to consistent coverage gains across a variety of model sizes (70M-70B), families (Llama, Gemma and Pythia) and levels of post-training (Base and Instruct models).

Large Language Monkeys: Scaling Inference Compute with Repeated Sampling

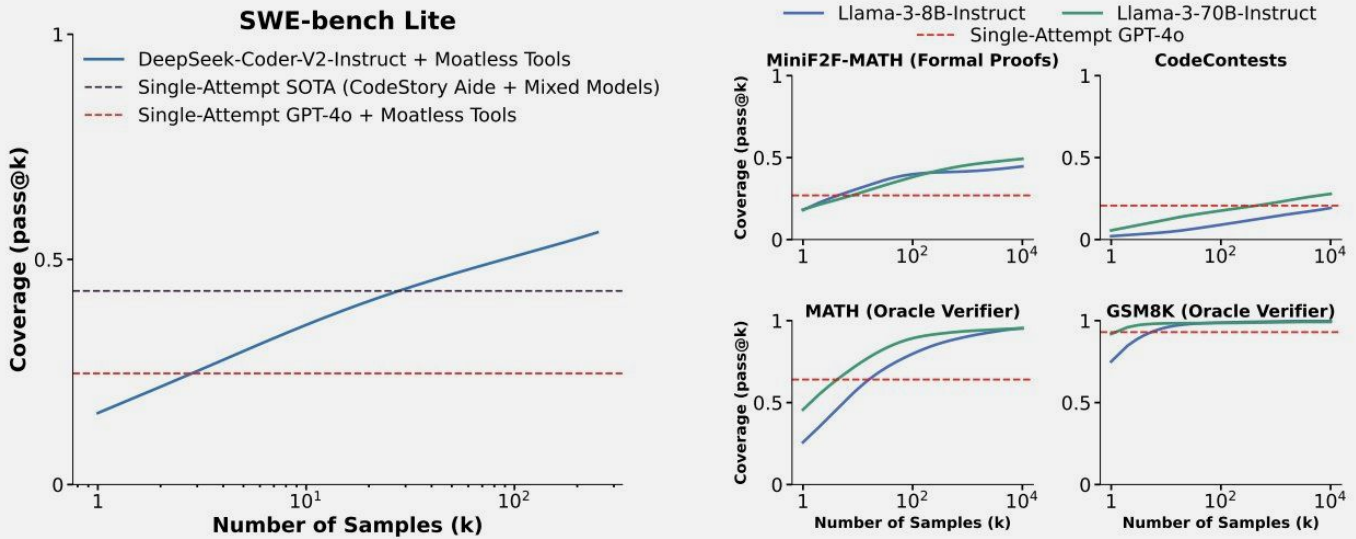


Figure 2: Across five tasks, we find that coverage, the fraction of problems solved by at least one generated sample, increases as we scale the number of samples. Notably, using repeated sampling, we are able to increase the solve rate of an open-source method from 15.9% to 56% on SWE-bench Lite, which is an improvement of 13% over the best single-attempt method.

S1: Simple test-time scaling

Bitter Lesson: Just scale Data and Compute, 2019

> This is a big lesson. As a field, we still have not thoroughly learned it, as we are continuing to make the same kind of mistakes. We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that
>
> 1) AI researchers have often tried to build knowledge into their agents
> 2) this always helps in the short term, and is personally satisfying to the researcher
> 3) in the long run it plateaus and even inhibits further progress
> 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

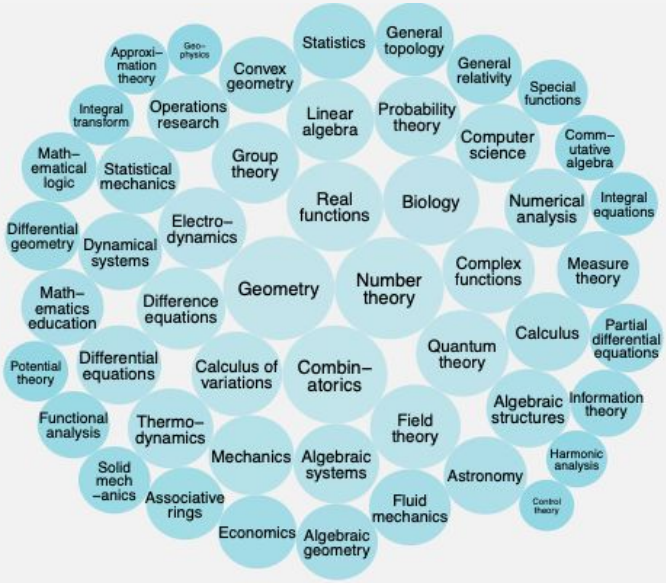
How many r in raspberry? **Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...
* First letter: 'r' - This is an 'r', count = 1.
* Second letter: 'a' - Not an 'r', count remains 1 ...
* Sixth letter: 'e' - Not an 'r', count remains 1.
* Seventh letter: 'r' - This is an 'r', count = 2.
* Eighth letter: 'y' - Not an 'r', count remains 2 ...
The number of 'r's in "raspberry" is 2.
Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... * Second 'r' ... * Third 'r' ... Count = 3 ... **Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is $\boxed{3}$ **Response**

Figure 3. **Budget forcing with s1-32B.** The model tries to stop after "...is 2.", but we suppress the end-of-thinking token delimiter instead appending "Wait" leading **s1-32B** to self-correct its answer.

S1 (cont.)



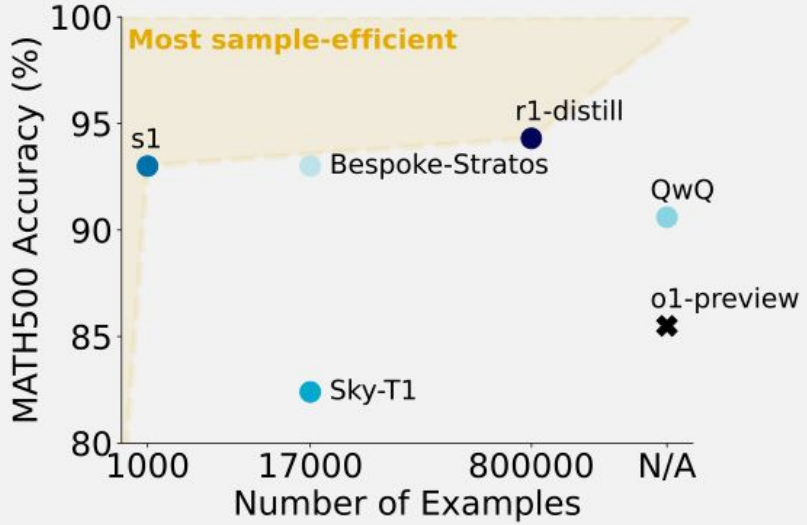
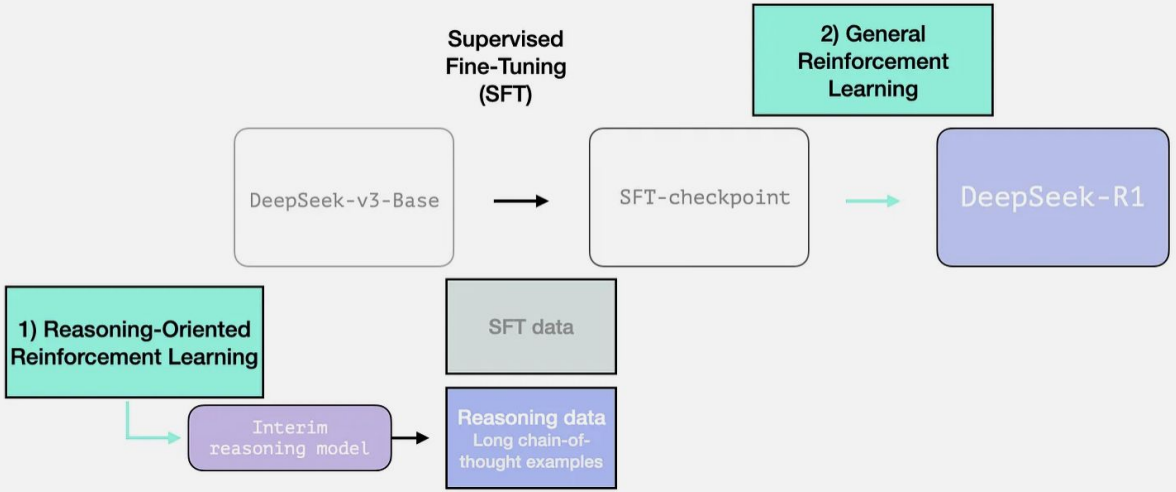


Figure 2. **s1K and s1-32B.** *(left)* **s1K** is a dataset of 1,000 high-quality, diverse, and difficult questions with reasoning traces. *(right)* **s1-32B**, a 32B parameter model finetuned on **s1K** is on the sample-efficiency frontier. See Table 1 for details on other models.

Too expensive; Reuse Experience 2024-2025

LLM pre-training is dead; Long live pretraining!

RL post-training with accurate rewards (DeepSeek stunned the world)



DeepSeek vs O1

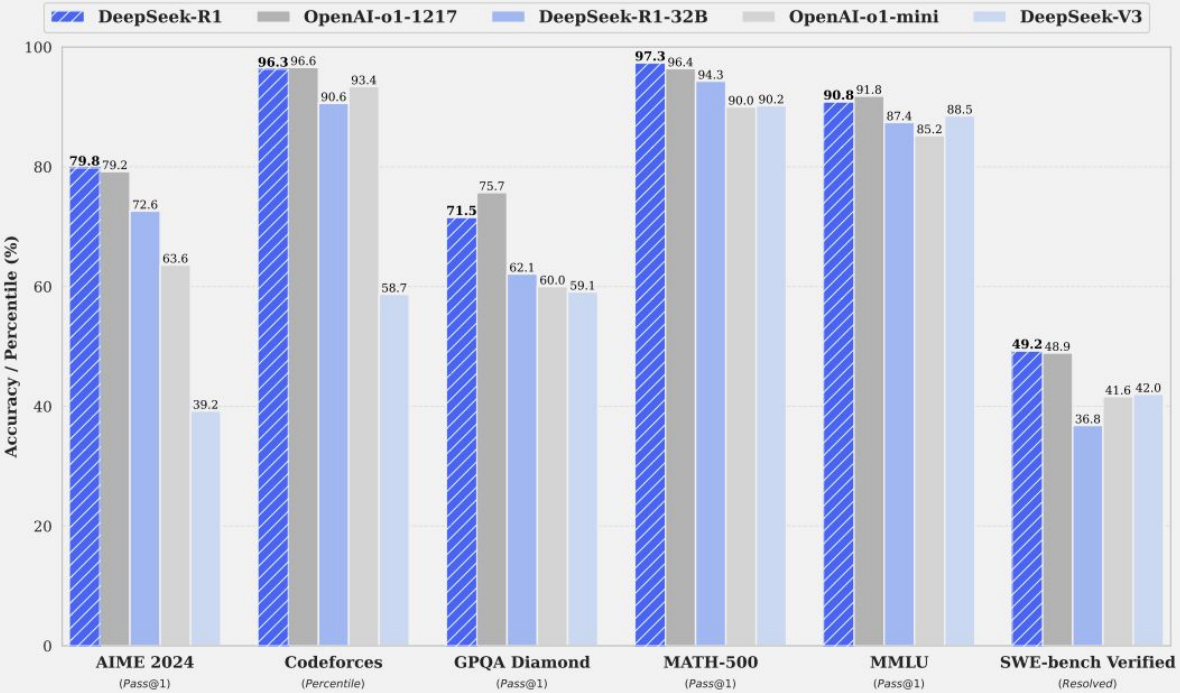


Figure 1 | Benchmark performance of DeepSeek-R1.

Deep Seek, Simple Rule-Based Oracle Reward Model

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

Deep Seek, Exploration

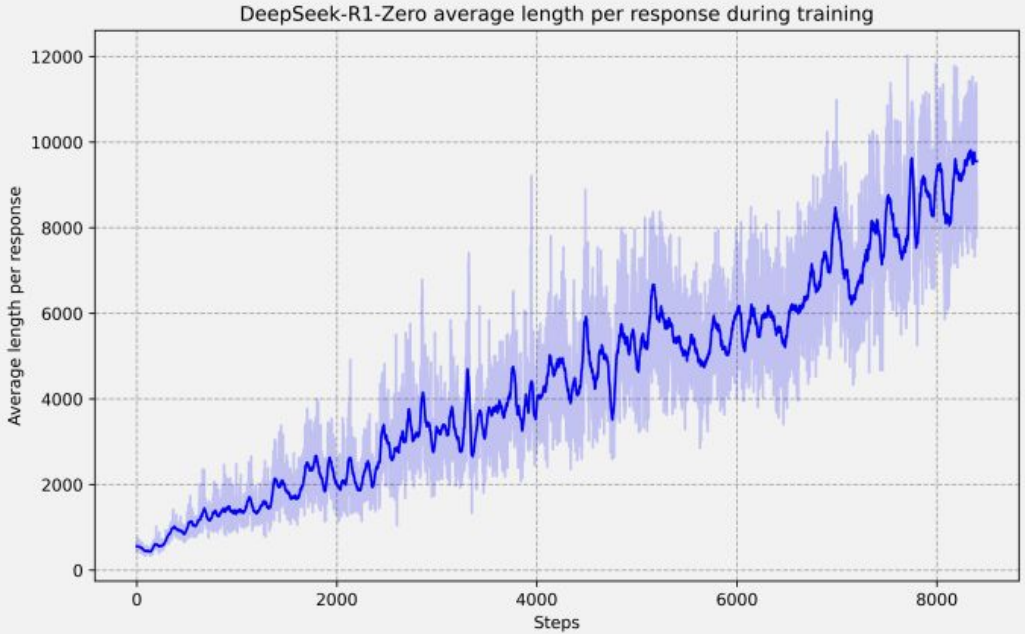


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

Deep Seek, Aha Moment

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to
Response: <think> To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$ $\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$ Rearrange to isolate the inner square root term: $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$ $\ldots$ Wait, wait. Wait. That's an aha moment I can flag here. Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$ We started with the equation: $\sqrt{a - \sqrt{a+x}} = x$ First, let's square both sides: $a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$ Next, I could square both sides again, treating the equation: $\cdots$ $\ldots$

Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

What We Will Discuss in This Session

AI Problem

Automation
AI Effect
Gödel's theorems
...

... - 1958

CNN

Among First Inductive Biases
Connectionism vs Symbolism

1980

AlexNet

First Deep Model
ImageNet Moment
Internet Data
GPU Computation

2012

Transformer

Latest innovative architecture
Alpha Go: Move 37

2017

BERT

Pre-training as Paradigm
Open source accelerated AI

2018

Initial Criticisms

OoD: is 100% enough?
ARC: Measure of intelligence
Symbolic Problems

2019

LLM Begins

GPT-3
Scale Laws
Reasoning Benchmarks

2020

New Hope to AGI

GPT-3.5
CLIP
DALL-E
Deceive Age
Chain of Thought
LLM As Agents

2022

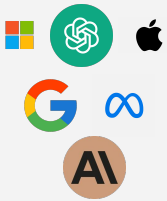
Return of RL

GPT4: RLHF
Is winter coming?

2023

Still Spring

GPT-4o
o1: Test Time Computation
o3: RIP ARC



2024

To Be Continue

DeepSeek: Reward is All you Need
Performance-to-Cost Ratio
...

2025