PAUL G. ALLEN SCHOOL
W

AI2

2050: An AI Odyssey: Dark Matter of Intelligence

Yejin Choi

University of Washington & AI2


What CVPR 2050 be like?

What CVPR 2050 be like?

Venue: metaverse?

What CVPR 2050 be like?

Venue: mars?

What CVPR 2050 be like?

ChatGPT writes the paper

ChatGPT reviews the paper

ChatGPT rebuttal period

Diffusion generates slides

NeRF presents the talk

ChatGPT summarizes the talk?

Few-shot prompting &

Instruction tuning?

NeRF? Diffusion? Transformers?

Autonomous driving? cleaning? plumbing? babyseating?

LLMs (or LVMs?) as prior?

Scaling laws no more?

Wha ike?

ChatGPT writes the

ChatGPT reviews the

ChatGPT rebuttal p

Diffusion generates

NeR

ChatGPT summarizes

&

ormers?

aning?

rior?

e?

Quantum Pre-trained Transformers (QPT) with perplexity 1.1??

What CVPR 2050 be like?

We haven't solved a dog level embodied AI yet!

AGI is just 5-10 years away!!

We haven't solved compositionality yet!

2050: An AI Odyssey

Prolog: what CVPR 2050 be like

Chapter 1: The Possible Impossibilities

Chapter 2: The Impossible Possibilities

Chapter 3: The Paradox

Epilog: why am I even here? A confession of an alien

The Possible Impossibilities?

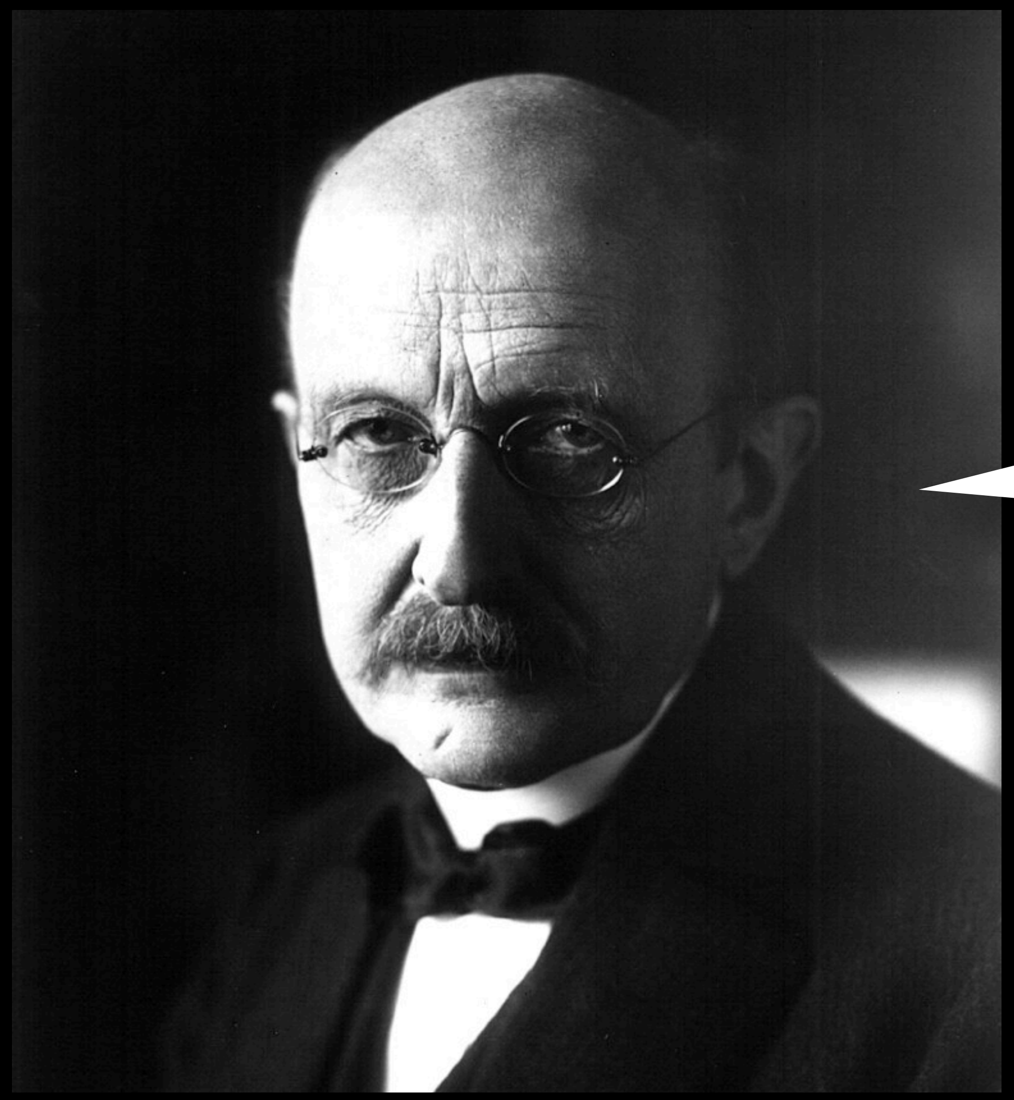
AGI is seemingly around the corner;
Is there really anything “impossible” with GPT5/6/7?

Circa 1878 …

Philipp von Jolly

"in this field, almost everything is already discovered, and all that remains is to fill a few unimportant holes"

Max Planck



"I don't wish to discover new things, only to u

f

🔥 Quantum Physics 🔥

Fast forward to the 20th/21st cent. …

Dark Matter

Schrödinger's cat

Wave-particle duality

Spacetime continuum

Fast forward to the 20th/21st cent. …

Dark Matter
Schrödinger's cat
Wave-particle duality
Spacetime continuum

Possible impossibilities
Impossible possibilities
Commonsense paradox
Moravec's paradox
Generative AI paradox

The Possible Impossibilities?

In the limit,

- can AGI arrive without embodiment?
- can RLHF fully align LLMs to factuality?
- can Transformers truly master compositionally?

Faith and Fate: Limits of Transformers on Compositionality

— arXiv:2305.18654 —

Nouha Dziri, Ximing Lu, Melanie Sclar

Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin,

Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui Yejin Choi

Does sparks of AGI (i.e., GPT4) solve multiplication?

Multiply 999 by 867.

GPT4

The product of multiplying 999 by 867 is 824,133.

The correct answer is 866,133

Multiply 999 by 867. Think step by step

To multiply 999 by 867 step by step, we can use the traditional long multiplication method. Here's how it would look:

css Copy code

```
       999
    x  867
-----------
      8993   (999 times 7)
     8993    (999 times 6, shifted one place to the left)
+   8993     (999 times 8, shifted two places to the left)
-----------
    866,733
```

Therefore, 999 multiplied by 867 equals 866,733.

The correct answer is 866,133

Does sparks of AGI (i.e., GPT4) solve multiplication?

Multiply 999 by 867.

GPT4

The product of multiplying 999 by 867 is 824,133.

The correct answer is 866,133

GPT4 achieves only **59% accuracy** on 3-digit x 3-digit multiplication.

4% accuracy for 4x4 multiplication!

0% accuracy for 5x5 multiplication!

Kevin Patrick Murphy
@sirbayes

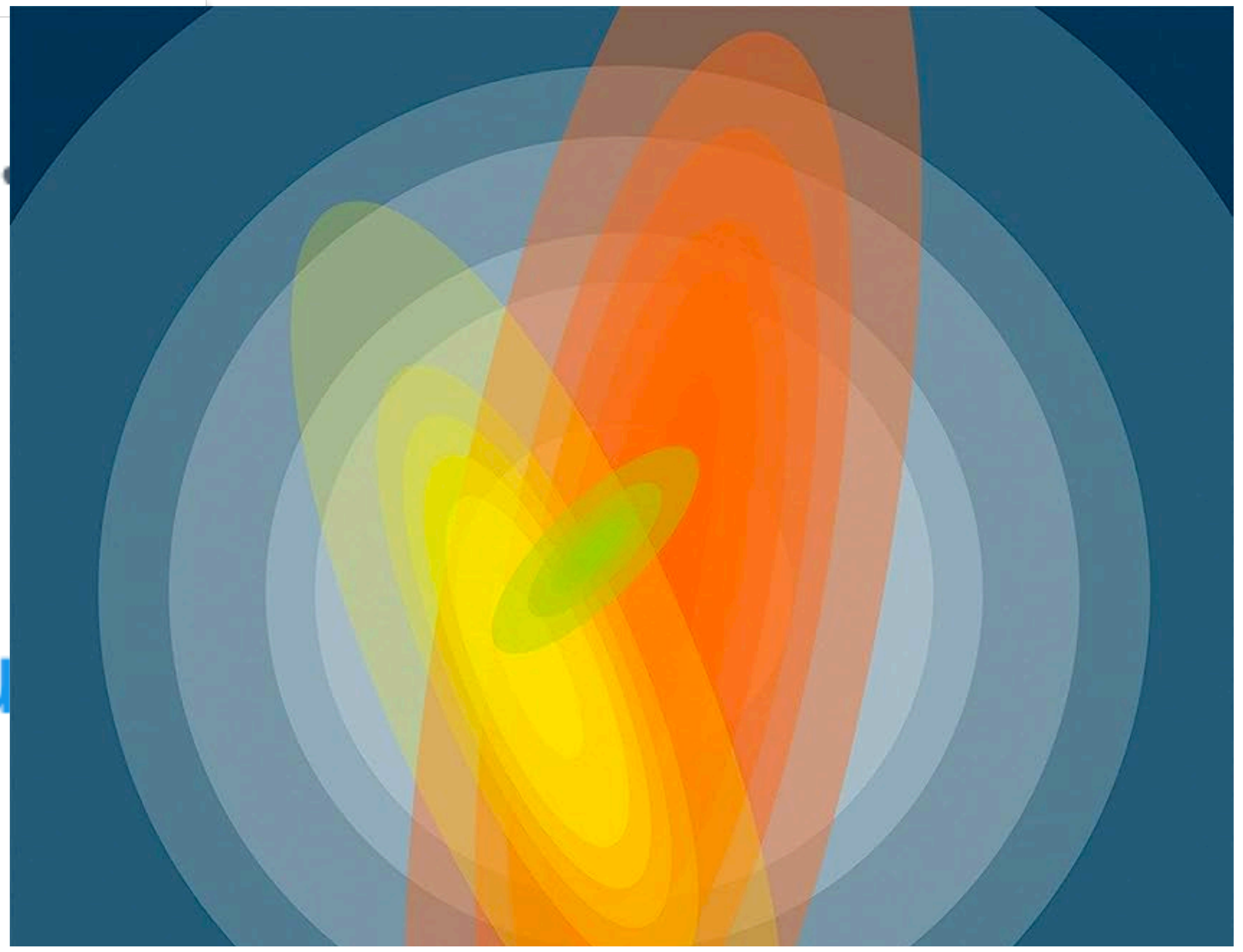
I like this paper. They prove that transformers are guaranteed to suffer from compounding errors when doing long reasoning chains (as @ylecu
has argued), and much apparent "success" is just due to unreliable pattern matching / shortcut learning.

arXiv

arxiv.org
Faith and Fate: Limits of
Transformer large langu
admiration for their exce

5:21 PM · Jun 15, 2023 · **266.1K** Views

236 Retweets **12** Quotes **1,392** Likes **85**



yobibyte
@y0b1byte

Best paper I've read in a while. No excessive hype, clear question, rigorous empirical evaluation. Highly recommended.
arxiv.org/abs/2305.18654

Cornell University

Computer Science > Computation and Language

arXiv:2305.18654 (cs)

[Submitted on 29 May 2023 (v1), last revised 1 Jun 2023 (this version, v2)]

Faith and Fate: Limits of Transformers on Compositionality

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, Yejin Choi

Download PDF

Does SUPERVISED TRAINING on 1.8 million examples* solve multiplication? (W/ GPT3)

Multiply 999 by 867.

Davinci

The product is 865,233.

The correct answer is 866,133

GPT3 fine-tuned achieves only **55% accuracy** on 3-digit x 3-digit multiplication.

Exhaustively training on the task examples does not solve multiplication!

Maybe tokenization is the issue … ??

Let's train GPT2 from scratch on multiplication data.

*1.8M examples include 1x1, 2x1, 2x2, 3x1, 3x2, 3x3, 4x1, 4x2 multiplication problem sizes, 12 epochs, 80:10:10 train/dev/test

SUPERVISED Training with Correct Tokenization will do it? (w/ GPT2-XL from scratch ...)

Each digit is assigned to one token

Each math symbol (e.g. "=" , "x") is assigned to one token

Train up to 4x4 (90M)*

Multiply 999 by 867.

GPT2-XL

The product is 865,233.

The correct answer is 66,233

Tokenization does not seem the issue for the low performance!

GPT2-XL achieves only **10% accuracy** on 3-digit x 3-digit multiplication.

*90M examples include 1x1, 2x1, 2x2, 3x1, 3x2, 3x3, 4x1, 4x2, 4x3, 4x4 multiplication problem sizes, 14 epochs, 80:10:10 train/dev/test

It's just a matter of step-by-step? (w/ GPT4 + scratchpad*)

```
   999
x  866
-------
  5994
 5994
7992
-------
= 865134
```

Multiply 999 by 866.

Let's perform the multiplication step by step:

Let's multiply 999 by the digit in the ones place of 866, which is 6.

1. Multiply 6 by the digit in the ones place of 999, which is 9. This gives 6 x 9= 54 . Write down the result 4 and carry over the 5 to the next step.
2. ….

Let's multiply 999 by the digit in the tens place of 867, which is 6.

3. Multiply 7 by the digit in the ones place of 999, which is 9. This gives 7 x 9= 63 . Write down the result 3 and carry over the 6 to the next step.
4. ……..

Accuracy increases to 92% on 3x3 multiplication!!

GPT4

Multiply 999 by 867.

Let's perform the multiplication step by step:

Let's multiply 999 by the digit in the ones place of 867, which is 7. ….

……….

The final answer is 866,133

✓

* We consider 5-shot 3x3 examples in each prompt.

It’s just a matter of step-by-step? (w/ GPT4 + scratchpad*)

4x4 multiplication

Multiply 8999 by 1866.

Let's perform the multiplication step by step:
Let's multiply 8999 by the digit in the ones place of 1866, which is 6.
1. Multiply 6 by the digit in the ones place of 8999, which is 9.
2. ....

GPT4

Let's perform the multiplication step by step:
....
The final answer is **16,695,134.**

Accuracy decreases to 4% on 4x4 multiplication!!

5x5 multiplication

Multiply 78999 by 41866.

Let's perform the multiplication step by step:
Let's multiply 78999 by the digit in the ones place of 41866, which is 6.
1. Multiply 6 by the digit in the ones place of 78999, which is 9.
2. ....

GPT4

Let's perform the multiplication step by step:
....
The final answer is **3,305,251,134**.

Accuracy decreases to 2% on 5x5 multiplication!!

Wait, didn't previous work said "transformers absolutely can learn true multi-step algorithms in-context"???

Thomas Miconi
@ThomasMiconi

Interesting, but... Isn't that in opposition to Zhou et al. 2022?

Transformers absolutely can learn and generalize to arbitrary problen

Partial Mastery

Method	Subtraction	Multiplication*
Algorithmic prompt	65.6%	79.7%
Best available baseline	16.7%	5.5%

arxiv.org
Teaching Algorithmic Reasoning via In-context Learning

Instead

We investigate the **fundamental limits** of achieving **full mastery** of the task rather than incremental improvements.

*they report GPT4 doesn't do well when multiplying digits > 3, thus covert the problem manually to addition over small digit (<= 3) multiplications

How about fine-tuning GPT3 on scratchpad?

Cost* USD	$744	$7440	$74,404	$744,040	$7.4 M	$70.4 M	$700 M
# Examples	8100	81K	810K	8.1M	81M	810M	8.1B
Problem size	**2x2**	**3x2**	**3x3**	**4x3**	**4x4**	**5x4**	**5x5**

*Cost for 4 epochs

How about fine-tuning GPT3 on scratchpad?

Cost* USD	$744	$7440	$74,404	$744,040	$7.4 M	$70.4 M	$700 M
# Examples	8100	81K	810K	8.1M	81M	810M	8.1B
Problem size	**2x2**	**3x2**	**3x3**	**4x3**	**4x4**	**5x4**	**5x5**

*Cost for 4 epochs

How about fine-tuning* GPT3 on scratchpad?

GPT3 achieves **96% accuracy** on in-distribution data but drops sharply to **zero** on OOD multiplication data.

*Why does this happen? Can we understand Transformers' behaviour via **computation graphs**?*

*Data includes all the enumerations of 1x1, 2x1, 2x2, 3x1, 3x2 problem sizes, 4 epochs, 80:10:10 train/dev/test. OOD data: 3x3, 4x1,4x2, 4x2, 4x4, etc

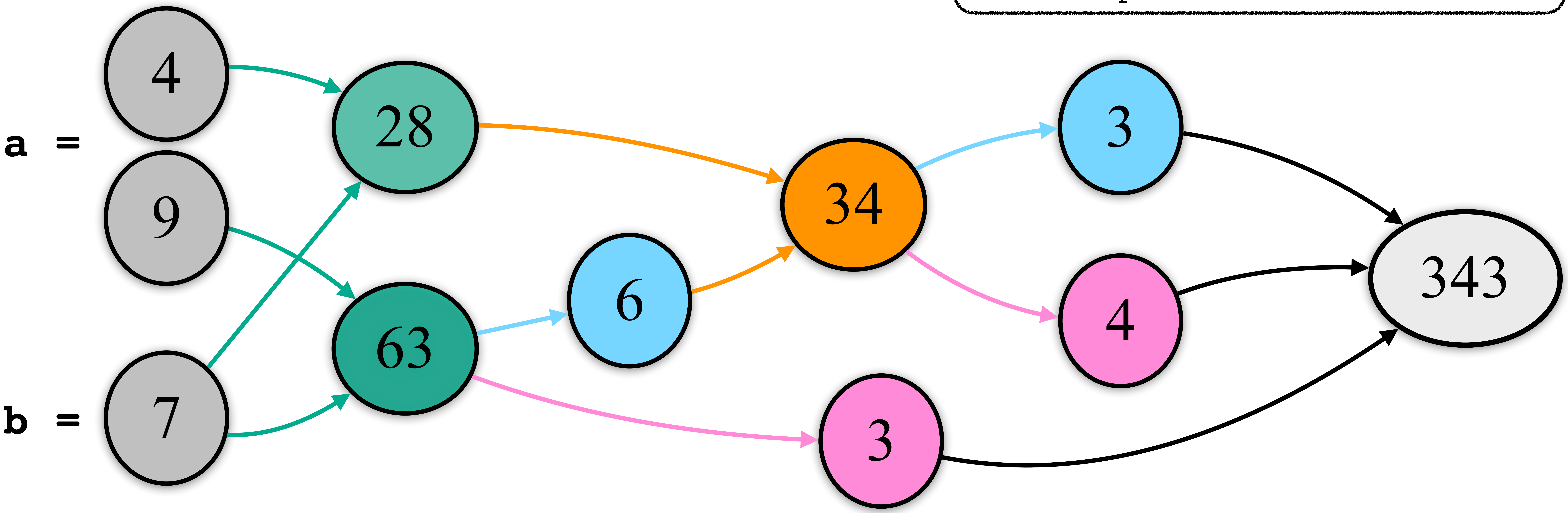
Computation graph for 49 x 7

multiply 1-digit | mod 10 | carry | sum | concat

```
function multiply (a[1:p], b[1:q]):
 for i = q to 1
    carry = 0
    for j = p to 1
    t = a[j] * b[i]
    t += carry  (only if j != p)
    digits[j] = t mod 10
    carry = t // 10
    summands[i] = digits
```

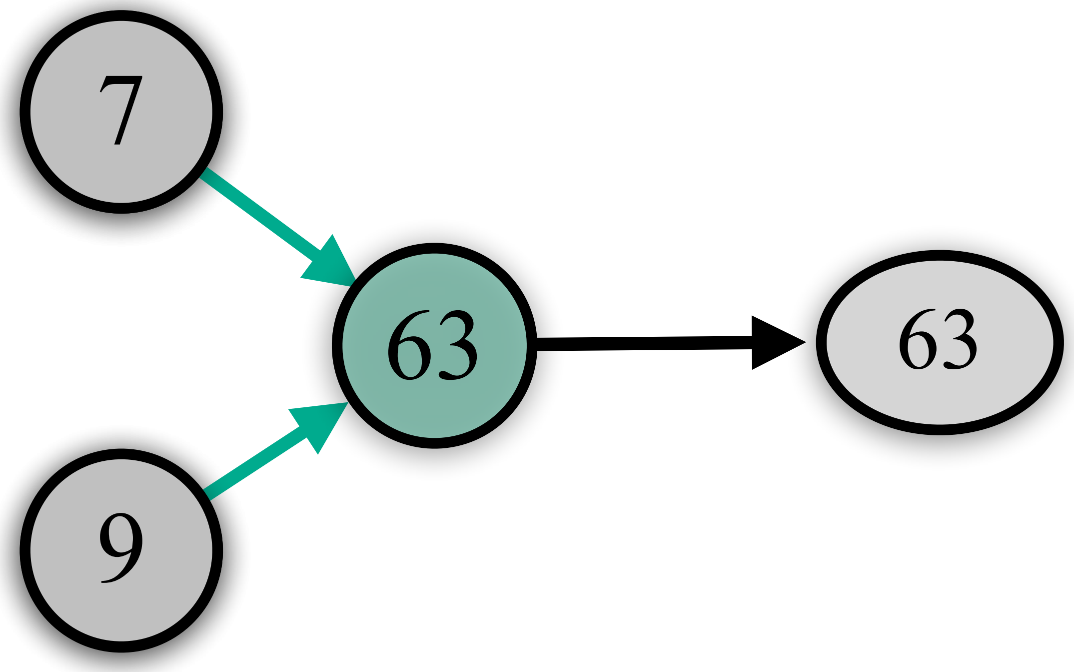
$$product = \sum_{i=1}^{q} summands[q+1-i] \cdot 10^{i-1}$$

```
 return product
```



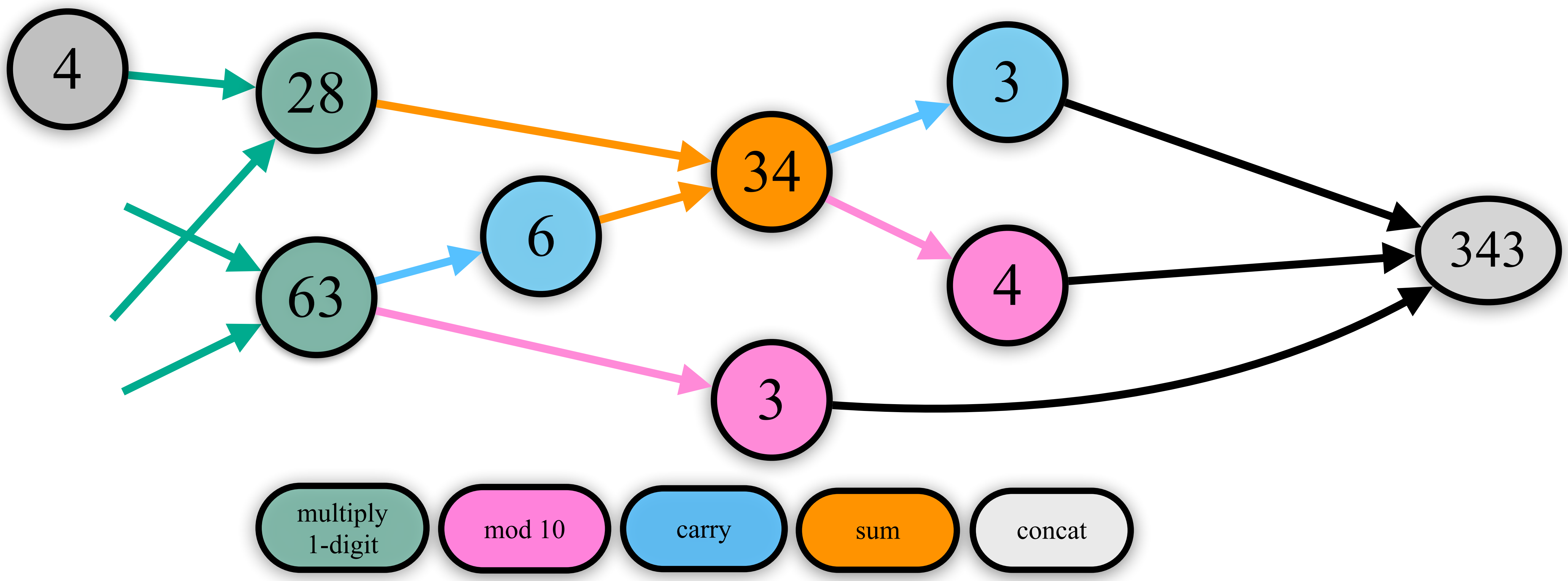
Model Performance Decreases as Graph Complexity Increases

9 x 7 = 63



graph width = 1
graph depth = 3
avg. parallelism = 1.3

49 x 7 = 343



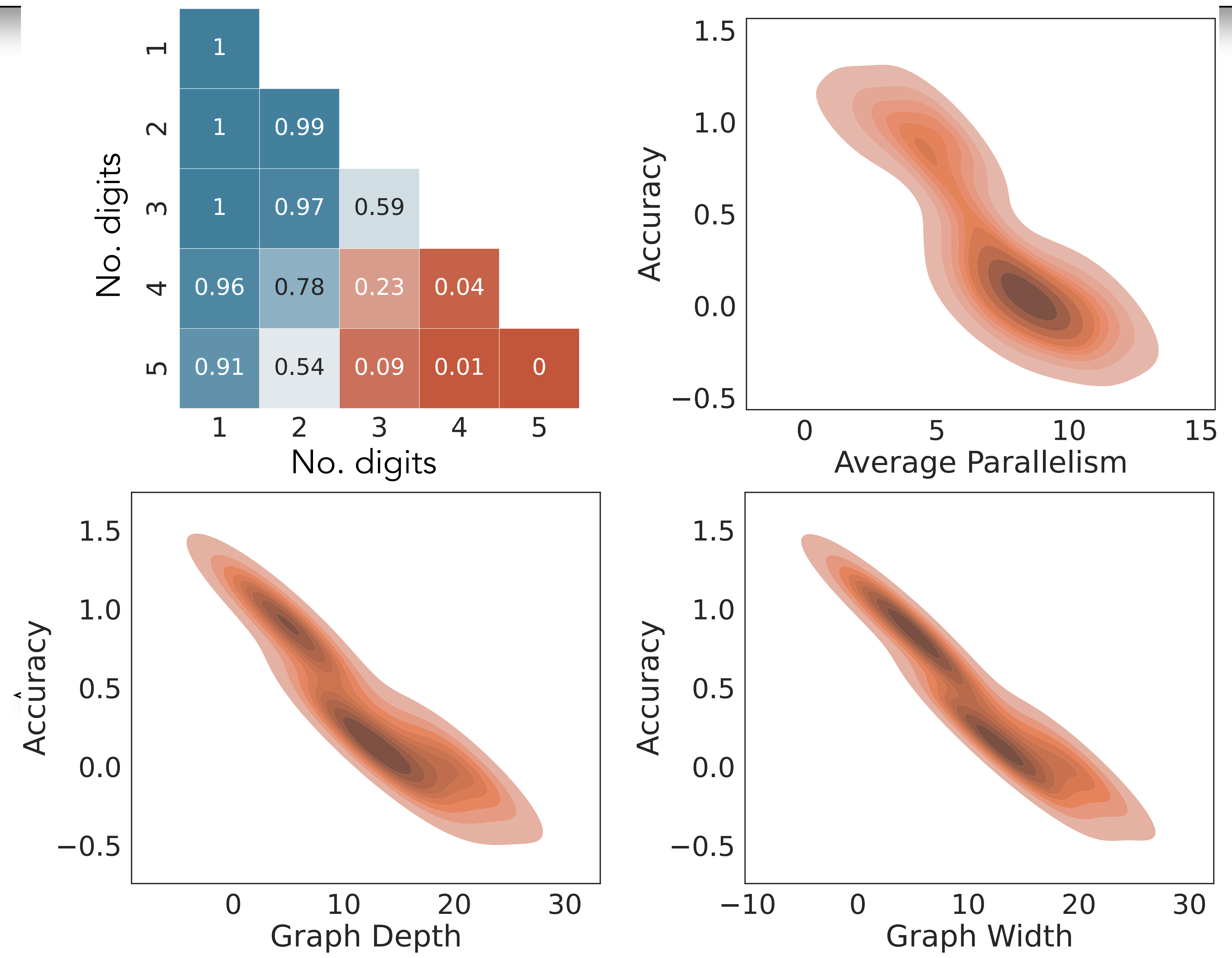
Graph Complexity

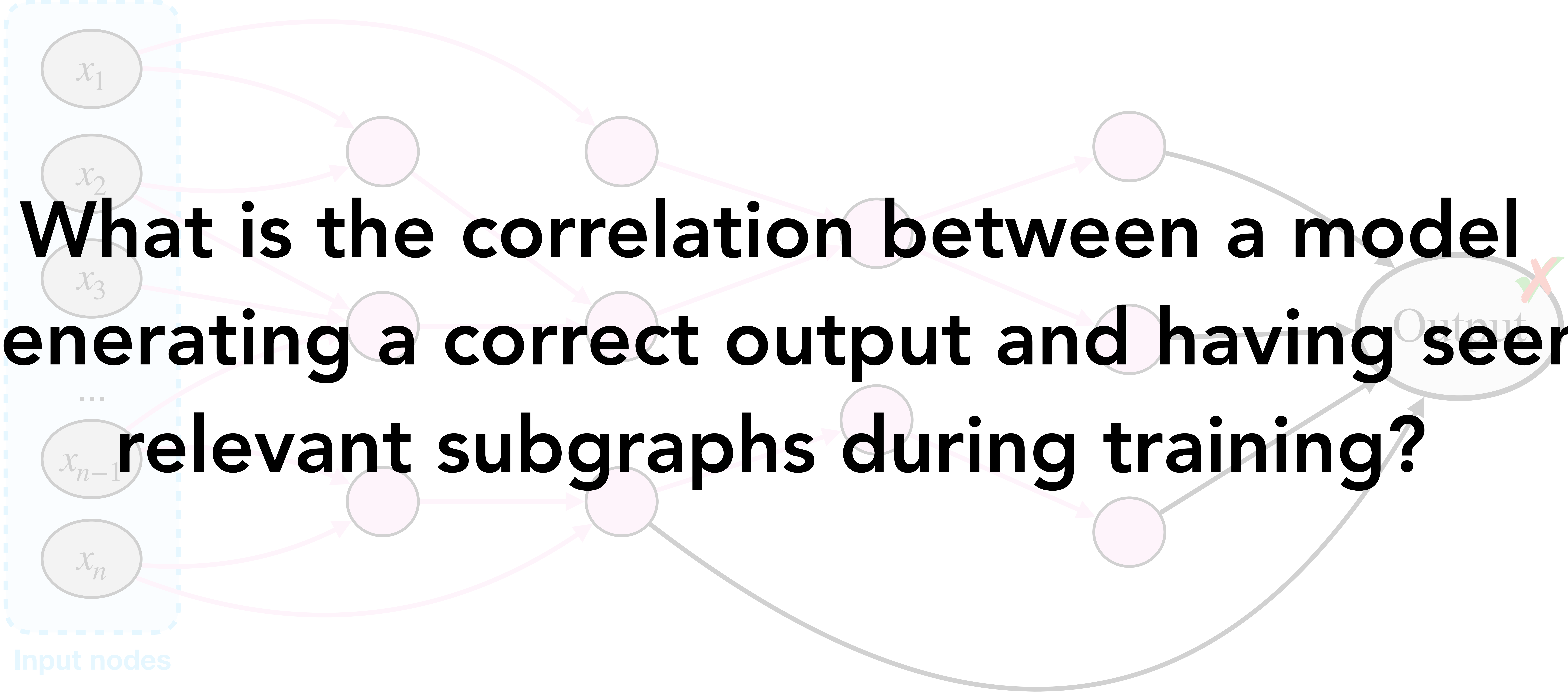
graph width: mode of $\{d(v) : v \in V\}$

graph depth: the largest layer number in the graph

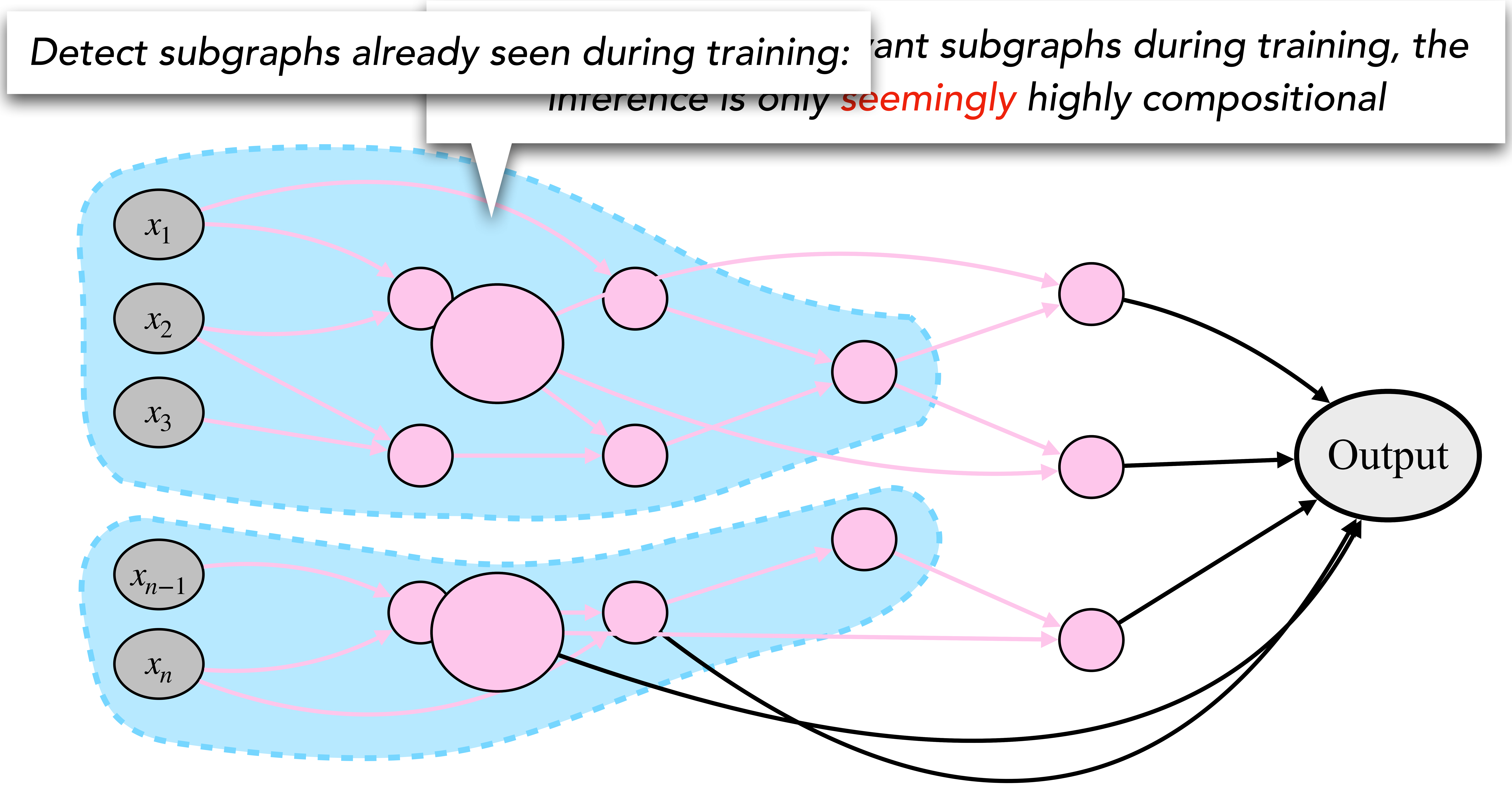
average parallelism: ratio between $|V|$ and reasoning depth

GPT4 zero-shot

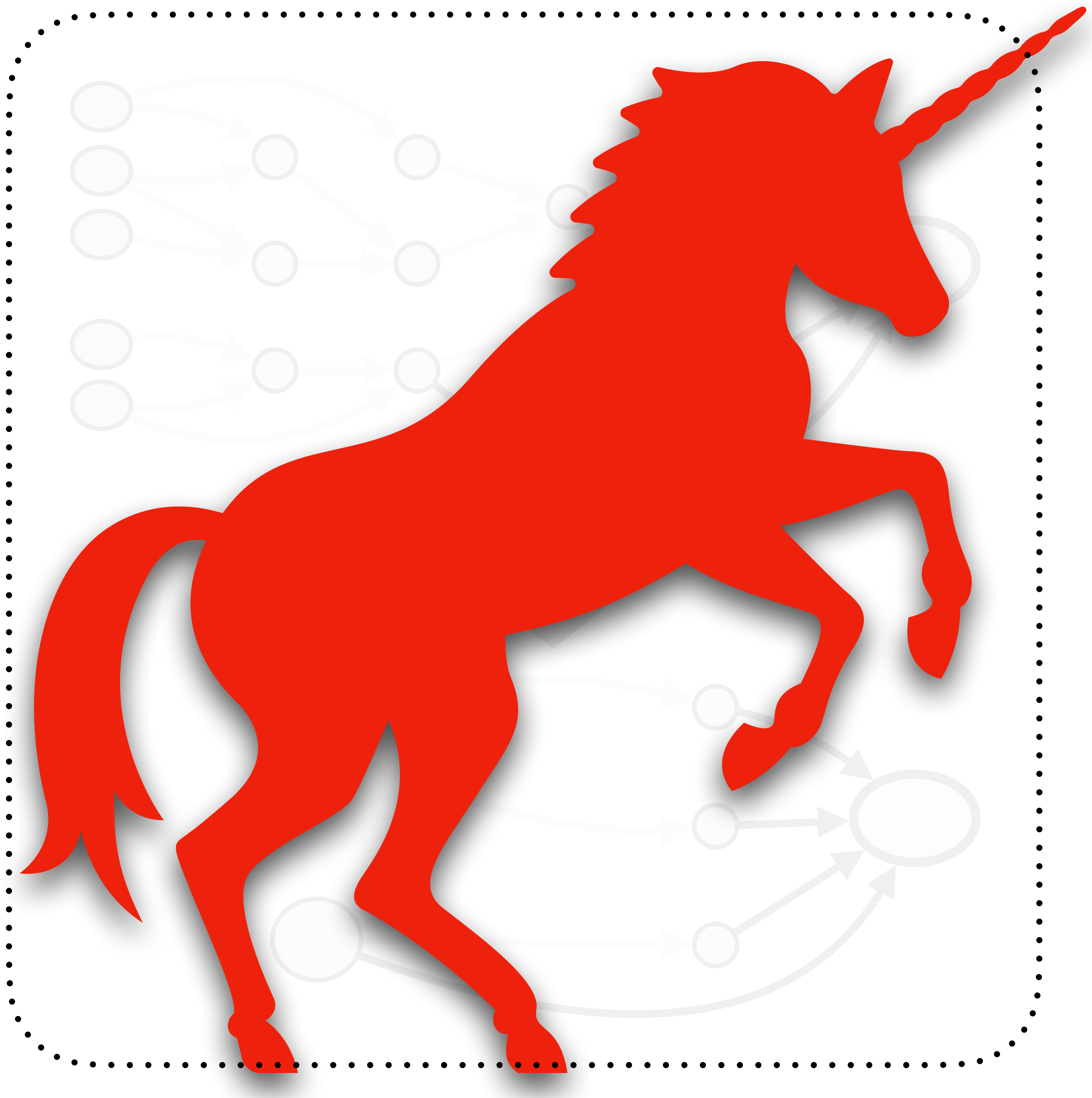


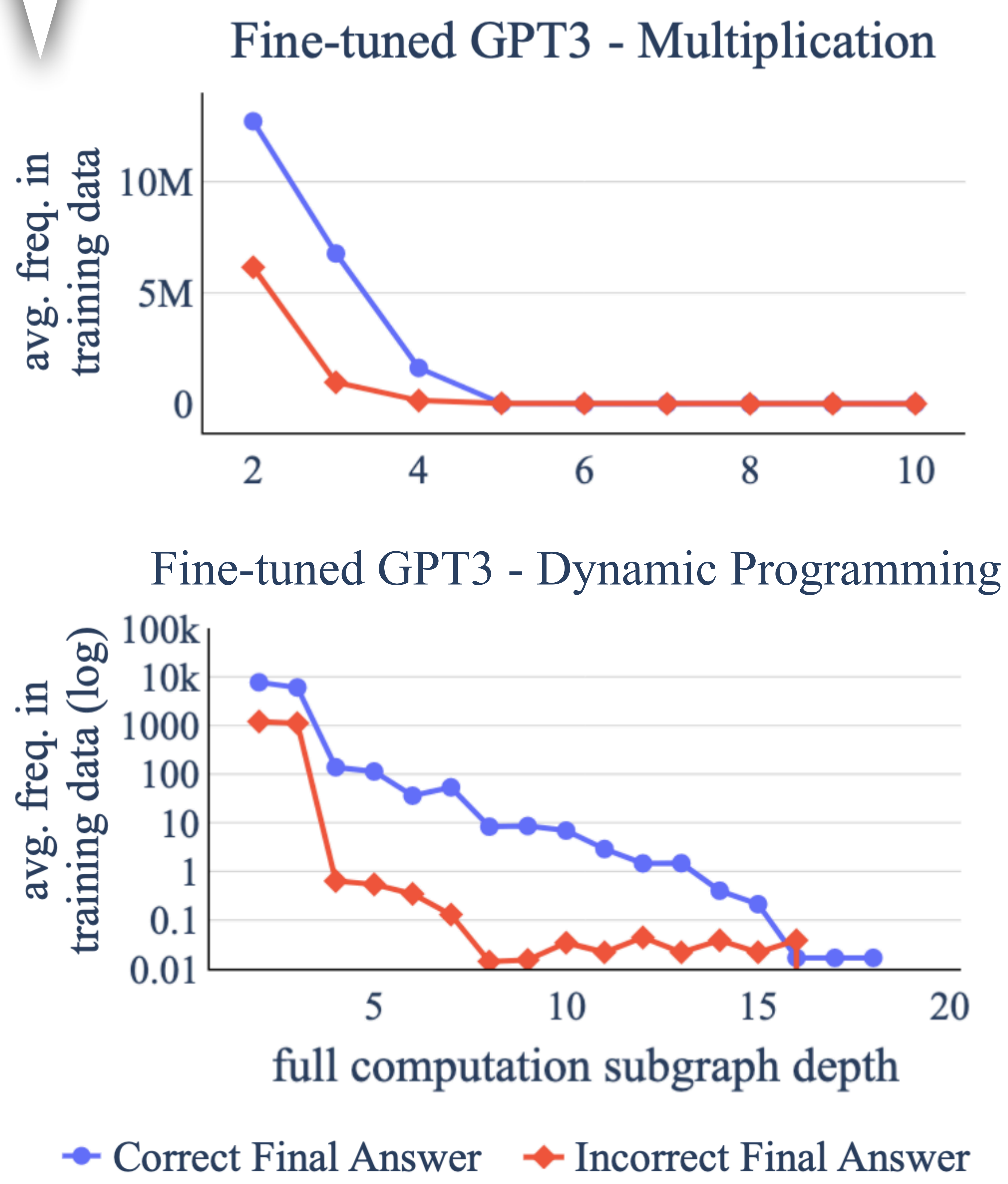


What is the correlation between a model generating a correct output and having seen relevant subgraphs during training?


Detect subgraphs already seen during training:
ant subgraphs during training, the
inference is only seemingly highly compositional
x_1
x_2
x_3
x_{n-1}
x_n
Output

Transformers' successes are heavily linked to having seen significant portions of the required computation graph during training

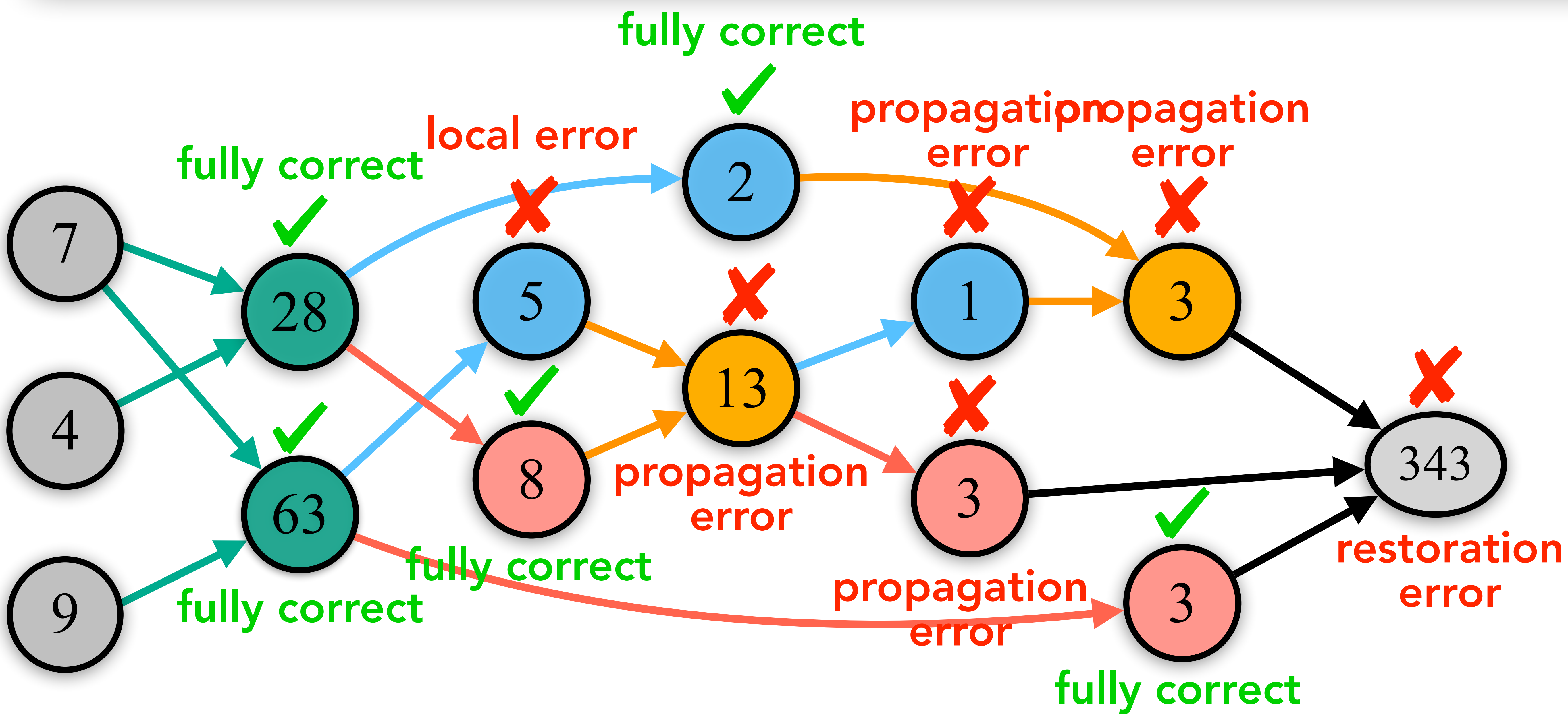


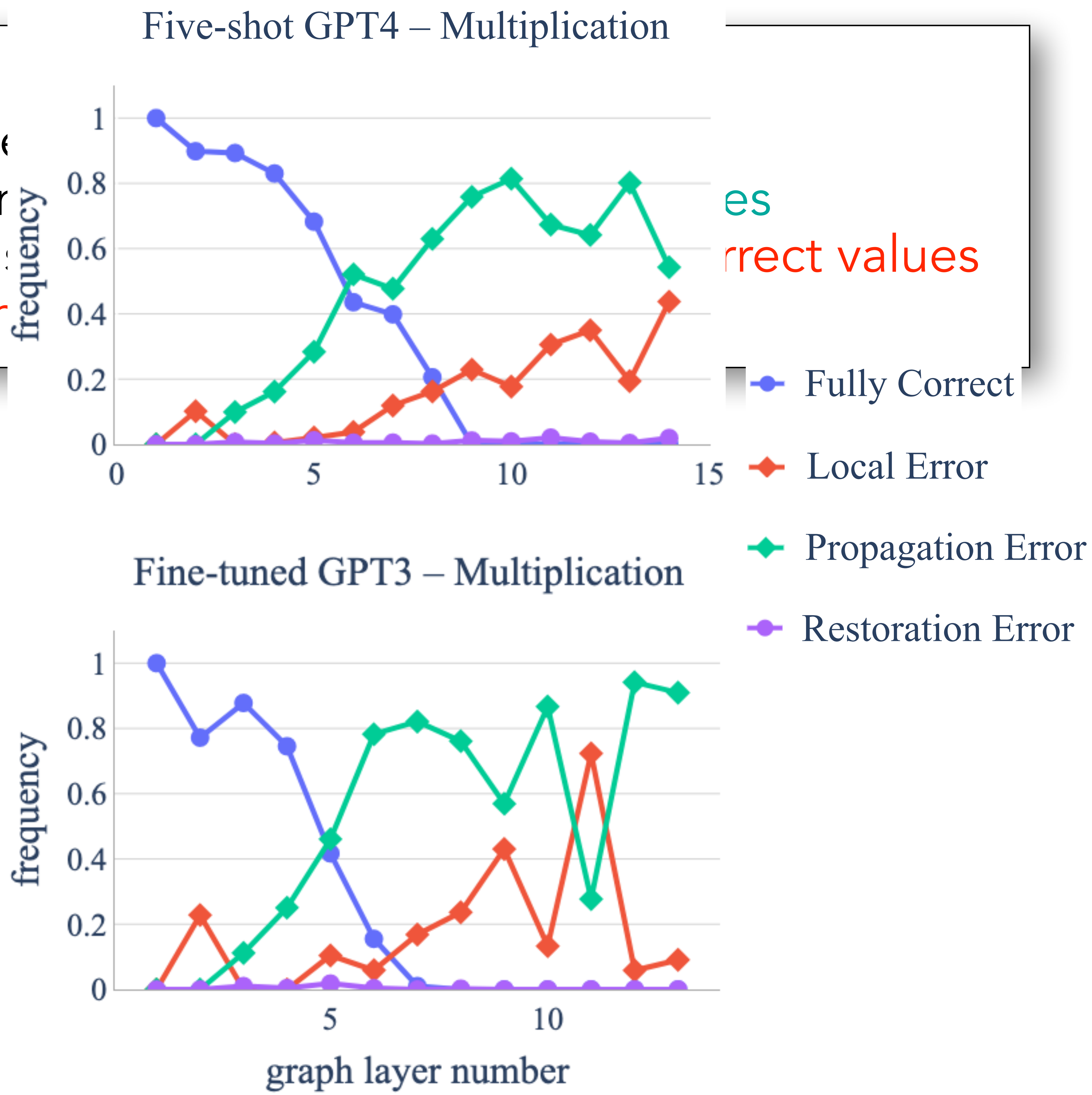


What Types of Errors do Transformers Make at Different Reasoning Depths?

Error Type

Fully Correct: v and ancestors have correct values and are derive … es
Local Error: v is derived from an incorrect computation but its ar
Propagation Error: v is derived from a correct computation but … rrect values
Restoration Error: v has a correct value but is derived from an ir





Transformers' performance will rapidly decay with increased task complexity

D Theoretical Results: Derivations

D.1 Transformers struggle with problems with increasingly larger parallelism (*width*)

Proposition D.1. *Let $f_n(\mathbf{x}) = h_n(g(\mathbf{x},1), g(\mathbf{x},2)), \dots, g(\mathbf{x},n))$. Let $\widehat{h}_n, \widehat{g}, \widehat{f}_n$ be estimators of h_n, g, f_n respectively. Assume $\mathbb{P}(h_n = \widehat{h}_n) = 1$ and $\mathbb{P}(h_n(X) = h_n(Y) \mid X \neq Y) < \beta\alpha^n$ for some $\alpha \in (0,1)$ and $\beta > 0$ (i.e. $\widehat{h}_n$ perfectly estimates h_n, and h_n is almost injective). If $\mathbb{P}(g \neq \widehat{g}) = \epsilon > 0$ and errors in $\widehat{g}$ are independent, then $\lim_{n\to+\infty} \mathbb{P}(f_n \neq \widehat{f}_n) = 1$.*

Proof. For ease of writing, let $X_i = g(X,i)$ and $Y_i = \widehat{g}(X,i)$, and let $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$. We will compute some auxiliary probabilities, and then upper bound $\mathbb{P}(f = \widehat{f})$, to finally compute its limit.

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} = \boldsymbol{Y}) &= \mathbb{P}(X_1 = Y_1, X_2 = Y_2, \dots, X_n = Y_n) \\ &= \mathbb{P}(X_1 = Y_1) \cdot \mathbb{P}(X_2 = Y_2) \dots \cdot \mathbb{P}(X_n = Y_n) = \mathbb{P}(g = \widehat{g})^n = (1-\epsilon)^n \qquad (2)\end{aligned}$$

Since by hypothesis we know $\mathbb{P}(h_n(\boldsymbol{Y}) = \widehat{h}_n(\boldsymbol{Y})) = 1$, we have that:

$$\begin{aligned}\mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) &= \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \cap h_n(\boldsymbol{Y}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \\ &= \mathbb{P}(h_n(\boldsymbol{X}) = h_n(\boldsymbol{Y}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \\ &\leq \mathbb{P}(h_n(\boldsymbol{X}) = h_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \\ &< \beta\alpha^n \qquad (3)\end{aligned}$$

We will now estimate $\mathbb{P}(f_n = \widehat{f}_n)$ using the law of total probability w.r.t. the event $\boldsymbol{X} = \boldsymbol{Y}$.

$$\begin{aligned}\mathbb{P}(f_n = \widehat{f}_n) &= \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y})) \\ &= \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{Y}) \cdot \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y}) + \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot \mathbb{P}(\boldsymbol{X} \neq \boldsymbol{Y}) \\ &= \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{X})) \cdot \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y}) + \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot (1 - \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y})) \\ &= 1 \cdot (1-\epsilon)^n + \mathbb{P}(h_n(\boldsymbol{X}) = \widehat{h}_n(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot (1 - (1-\epsilon)^n) \qquad \text{(using 2 and hypothesis)} \\ &< (1-\epsilon)^n + \beta\alpha^n \cdot (1 - (1-\epsilon)^n) \qquad \text{(using 3)} \\ &< \beta\alpha^n + (1-\epsilon)^n \cdot (1 - \beta\alpha^n)\end{aligned}$$

To conclude our proof, we will show that $\lim_{n\to+\infty} \mathbb{P}(f_n = \widehat{f}_n)$ exists and compute its value. Note that since $1-\epsilon \in [0,1)$ and $\alpha \in (0,1)$, trivially $\lim_{n\to+\infty} \beta\alpha^n + (1-\epsilon)^n \cdot (1-\beta\alpha^n) = 0$.

$$0 \leq \liminf_{n\to+\infty} \mathbb{P}(f_n = \widehat{f}_n) \leq \limsup_{n\to+\infty} \mathbb{P}(f_n = \widehat{f}_n) \leq \limsup_{n\to+\infty} \beta\alpha^n + (1-\epsilon)^n \cdot (1-\beta\alpha^n) = 0$$

Then, $\lim_{n\to+\infty} \mathbb{P}(f_n = \widehat{f}_n) = 0$ and we conclude $\lim_{n\to+\infty} \mathbb{P}(f_n \neq \widehat{f}_n) = 0$. □

Corollary D.1. *Assume that a model $\mathcal{M}$ solves shifted addition perfectly, but it incorrectly solves **at least one** m digit by 1 digit multiplication for some fixed m. Then, the probability that $\mathcal{M}$ will solve **any** m digit by n digit multiplication using the long-form multiplication algorithm tends to 0.*

Proof. We define $s : \mathbb{Z}_{10}^{m+n} \times \mathbb{N} \to \mathbb{N} \times \mathbb{N}$, $d : \mathbb{N} \times \mathbb{Z}_{10} \to \mathbb{N}$, $h_n : \mathbb{N}^n \to \mathbb{N}$, and $f_n : \mathbb{Z}_{10}^{m+n} \to \mathbb{N}$ as follows.

$$\begin{aligned} s([x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}], j) &:= (x_1 ^\frown x_2 ^\frown \dots ^\frown x_m,\ x_{m+j}) \\ &\qquad \text{where } x_1 ^\frown x_2 ^\frown \dots ^\frown x_m \text{ denotes concatenating digits } x_i \\ d(x,y) &:= x \cdot y \\ g &:= d \circ s \\ h_n(x_1, \dots, x_n) &:= \sum_{i=1}^{n} x_i 10^{n-i} \\ f_n(\mathbf{x}) &:= h_n(g(\mathbf{x},1), g(\mathbf{x},2)), \dots, g(\mathbf{x},n))\end{aligned}$$

Note that g defines the base-10 multiplication between m-digit numbers $(x_1x_2\dots x_m)$ and 1-digit numbers (x_{m+j}), where s denotes the selection of the numbers to multiply and d denotes the actual multiplication. Note that h_n describes the shifted addition used at the end of long-form multiplication to combine n m-digit by 1-digit multiplications. Therefore, f_n describes the long-form multiplication of m-digit by n-digit numbers.

By hypothesis, $\mathbb{P}(g \neq \widehat{g}) = \epsilon > 0$ and $\mathbb{P}(h_n = \widehat{h}_n) = 1$, where $\widehat{g}$ and $\widehat{h}_n$ denote estimators using model $\mathcal{M}$. It can be shown that $\mathbb{P}(h_n(X) = h_n(Y) \mid X \neq Y) < \beta\alpha^n$ for $\alpha = 0.1$ and $\beta = 10^m$. Using Lemma D.1, $\lim_{n\to+\infty} \mathbb{P}(f_n \neq \widehat{f}_n) = 1$, which concludes our proof.

□

Note that Lemma D.1's proofs gives us empirical bounds once ϵ and α are approximated. Also **note that our definition of g in the proof of Corollary D.1 highlights two possible sources of exponentially-accumulating error**: errors in the selection of the numbers to multiply s, and errors in the actual m-digit by 1-digit multiplication d.

D.2 Transformers struggle with problems that require increasingly larger iterative applications of a function (*depth*)

Proposition D.2. *Let $f_n(\mathbf{x}) = g^n(\mathbf{x})$. Assume $\mathbb{P}(g(X) = \widehat{g}(Y) \mid X \neq Y) \leq c$ (i.e. recovering from a mistake due to the randomness of applying the estimator on an incorrect input has probability at most c). If $\mathbb{P}(g \neq \widehat{g}) = \epsilon > 0$ with $c + \epsilon < 1$, then $\liminf_{n\to+\infty} \mathbb{P}(f_n \neq \widehat{f}_n) = 1 - \frac{c}{c+\epsilon}$.*

Proof. We first derive a recursive upper bound using the law of total probability, and then prove a non-recursive upper bound by induction.

$$\begin{aligned} s_n := \mathbb{P}(f_n = \widehat{f}_n) &= \mathbb{P}(g(g^{n-1}(Z)) = \widehat{g}(\widehat{g}^{n-1}(Z))) \\ &= \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{Y})) \qquad \text{where } \boldsymbol{X} := g^{n-1}(Z) \text{ and } \boldsymbol{Y} := \widehat{g}^{n-1}(Z) \\ &= \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{Y}) \cdot \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y}) + \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot \mathbb{P}(\boldsymbol{X} \neq \boldsymbol{Y}) \\ &= \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{X})) \cdot \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y}) + \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot (1 - \mathbb{P}(\boldsymbol{X} = \boldsymbol{Y})) \\ &= \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{X})) \cdot s_{n-1} + \mathbb{P}(g(\boldsymbol{X}) = \widehat{g}(\boldsymbol{Y}) \mid \boldsymbol{X} \neq \boldsymbol{Y}) \cdot (1 - s_{n-1}) \\ &\leq (1-\epsilon) \cdot s_{n-1} + c \cdot (1 - s_{n-1}) \\ &\leq (1 - \epsilon - c) \cdot s_{n-1} + c\end{aligned}$$

We know $s_1 = (1-\epsilon)$ since $s_1 = \mathbb{P}(f_1 = \widehat{f}_1) = \mathbb{P}(g = \widehat{g})$. Let $b := 1 - \epsilon - c$ for ease of writing. Then, we have

$$s_n \leq b \cdot s_{n-1} + c \qquad (4)$$

It can be easily shown by induction that $s_n \leq b^{n-1}(1-\epsilon) + c\sum_{i=0}^{n-2} b^i$:

- The **base case** $n = 2$ is true since we know $s_2 \leq b \cdot s_1 + c$, and $b \cdot s_1 + c = b(1-\epsilon) + c = b^{2-1}(1-\epsilon) + c\sum_{i=0}^{2-2} b^i$, thus showing $s_2 \leq b^{2-1}(1-\epsilon) + c\sum_{i=0}^{2-2} b^i$
- The **inductive step** yields directly using Equation 4,

$$\begin{aligned} s_n &\leq b \cdot s_{n-1} + c \\ &\leq b \cdot \left(b^{n-2}(1-\epsilon) + c\sum_{i=0}^{n-3} b^i\right) + c \leq b^{n-1}(1-\epsilon) + c\sum_{i=1}^{n-2} b^i + c \leq b^{n-1}(1-\epsilon) + c\sum_{i=0}^{n-2} b^i\end{aligned}$$

We can rewrite the geometric series $\sum_{i=0}^{n-2} b^i$ in its closed form $\frac{1-b^{n-1}}{1-b}$, and recalling $b := 1 - \epsilon - c$,

$$\begin{aligned} s_n &\leq b^{n-1}(1-\epsilon) + c\frac{1-b^{n-1}}{1-b} = b^{n-1}(1-\epsilon) + c\frac{1-b^{n-1}}{c+\epsilon} \\ &= b^{n-1}(1-\epsilon) + \frac{c}{c+\epsilon} - b^{n-1}\frac{c}{c+\epsilon} \\ &= b^{n-1}\left(1 - \epsilon - \frac{c}{c+\epsilon}\right) + \frac{c}{c+\epsilon}\end{aligned}$$

Shortcut Learning in Deep Neural Networks

Robert Geirhos[1,2,*,§], Jörn-Henrik Jacobsen[3,*], Claudio Michaelis[1,2,*], Richard Zemel[†,3], Wieland Brendel[†,1], Matthias Bethge[†,1] & Felix A. Wichmann[†,1]

Transformers Learn Shortcuts to Automata

By and large, the prior work was based on weaker LLMs, thus some might have wondered with extreme-scale, these problems magically go away

Ruixiang Tang[†], Dehan Kong[‡], Longtao Huang[‡], Hui Xue[‡]

Shortcut Learning of Large Language Models in Natural Language Understanding

Mengnan Du
New Jersey Institute of Technology
Newark, NJ, USA
mengnan.du@njit.edu

Fengxiang He
JD Explore Academy
Beijing, Beijing, China
fengxiang.f.he@gmail.com

Na Zou
Texas A&M University
College Station, TX, USA
nzou1@tamu.edu

Dacheng Tao
The University of Sydney

Xia Hu
Rice University

Let's step back…

Transformers are not the right models for multiplication?
Instead, **Toolformers** (Schick et. al. 2003)?

That's exactly the point!
Relatedly, are transformers the right models for other compositional aspects of commonsense / language?

Multiplication (+ puzzles, algorithms) are an "edge case"??? all other compositionality will work well with transformers + RLHF + scratchpad ???

1. How do we know the full mastery?
2. WHY is simple multiplication harder than other (seemingly more complex) compositional tasks?
3. (Since we are at CVPR) what about compositional visual QA?

CREPE: Can Vision-Language Foundation Models Reason Compositionally?

Zixian Ma[1]*, Jerry Hong[1]*, Mustafa Omer Gul[2]*, Mona Gandhi[3], Irena Gao[1], Ranjay Krishna[4]

Abstract

A fundamental characteristic common to both human vision and natural language is their compositional nature. Yet, despite the performance gains contributed by large vision and language pretraining, we find that—across 7 architectures trained with 4 algorithms on massive datasets—they struggle at compositionality. To arrive at this conclusion, we introduce a new compositionality evaluation benchmark, CREPE, which measures two important aspects of compo-

CREPE: Can Vision-Language Foundation Models Reason Compositionally?

Zixian Ma[1]*, Jerry Hong[1]*, Mustafa Omer Gul[2]*, Mona Gandhi[3], Irena Gao[1], Ranjay Krishna[4]

Measuring Compositional Consistency for Video Question Answering

Mona Gandhi[1]*, Mustafa Omer Gul[2]*, Eva Prakash[2], Madeleine Grunde-McLau
Ranjay Krishna[3], Maneesh Agrawala[2]

AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning

Madeleine Grunde-McLaughlin Ranjay Krishna Maneesh Agrawala

2050: An AI Odyssey

Prolog: what CVPR 2050 be like

Chapter 1: The Possible Impossibilities

Chapter 2: The Impossible Possibilities

Chapter 3: The Paradox

Epilog: why am I even here? A confession of an alien

Circa 2023 …

How can Indian startups create foundation models for India?

Rajan Anandan

Sam Atman

It’s hopeless to compete with OpenAI

🍔 Impossible Distillation

from Low-quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing

— arxiv:2305.16635 —

Jaehun Jung

Peter West

Liwei Jiang

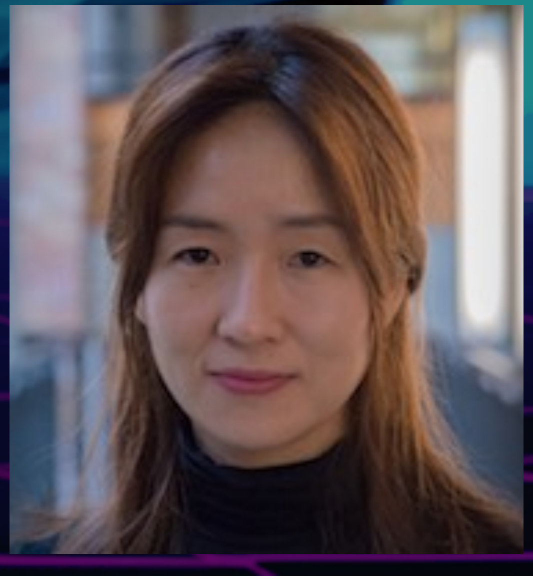
Faeze Brahman

Ximing Lu

Jillian Fisher

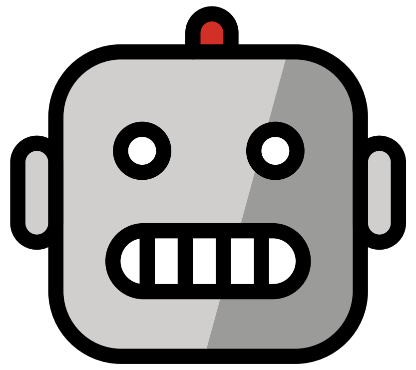
Taylor Sorensen

Yejin Choi



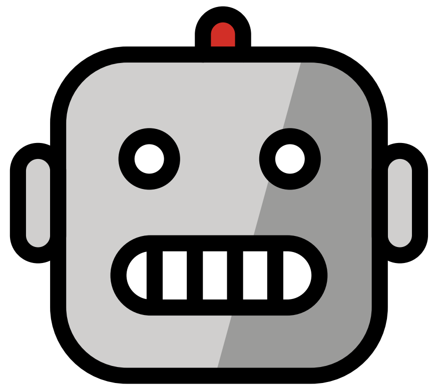
winning recipe = extreme-scale pre-training + RLHF at scale

GPT-2



Low-quality, small models

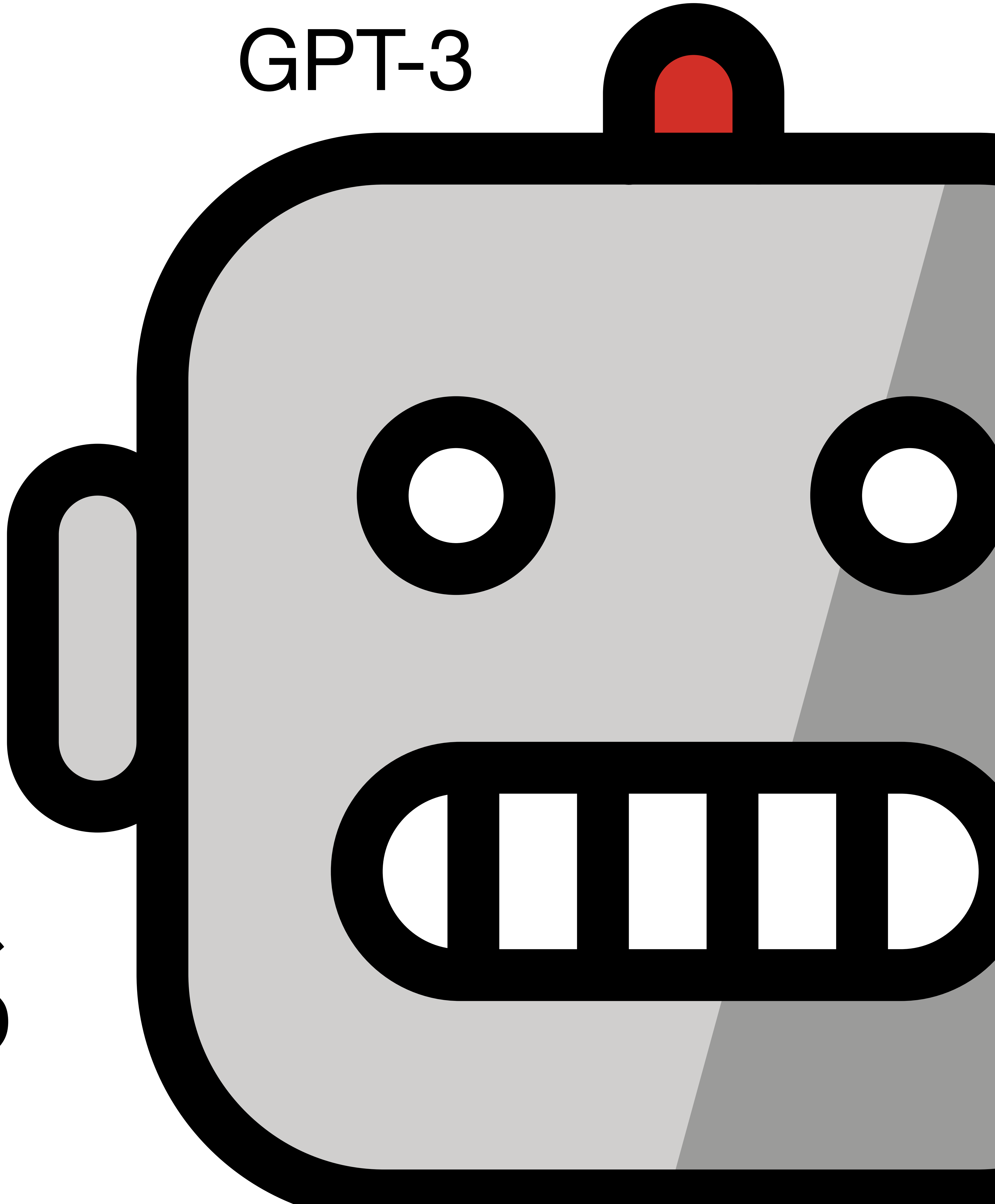
???



high-quality, small models

VS

GPT-3



How is that even possible when imitating from proprietary LLMs are supposedly hopeless?

The False Promise of Imitating Proprietary LLMs

Arnav Gudibande*
UC Berkeley
arnavg@berkeley.edu

Eric Wallace*
UC Berkeley
ericwallace@berkeley.edu

Charlie Snell*
UC Berkeley
csnell22@berkeley.edu

Xinyang Geng
UC Berkeley
young.geng@berkeley.edu

Hao Liu
UC Berkeley
hao.liu@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@berkeley.edu

Sergey Levine
UC Berkeley
svlevine@berkeley.edu

Dawn Song
UC Berkeley
dawnsong@berkeley.edu

Are small LMs completely out of league?

Can small, off-the-shelf LMs learn to abstract without task supervision?

https://twitter.com/EmojiMashupBot/status/1266262982406729730

Task-specific Symbolic Knowledge Distillation works!

Symbolic Knowledge Distillation: from General Language Models to Commonsense Models

Peter West[†‡*] **Chandra Bhagavatula**[‡] **Jack Hessel**[‡] **Jena D. Hwang**[‡]
e Bras[‡] **Ximing Lu**[†‡] **Sean Welleck**[†‡] **Yejin Choi** [†‡*]
omputer Science & Engineering, University of Washington
n Institute for Artificial Intelligence

Teaching Small Language Models to Reason

Lucie Charlotte Magister[*] University of Cambridge lcm67@cam.ac.uk
Jonathan Mallinson Google Research jonmall@google.com
Jakub Adamek Google Research enkait@google.com
Eric Malmi Google Research emalmi@google.com
Aliaksei Severyn Goo severy

Specializing Smaller Language Models towards Multi-Step Reasoning

Yao Fu[♠] **Hao Peng**[♣] **Litu Ou**[♠] **Ashish Sabharwal**[♣] **Tushar Khot**[♣]

Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes

Cheng-Yu Hsieh[1*], **Chun-Liang Li**[2], **Chih-Kuan Yeh**[3], **Hootan Nakhost**[2], **Yasuhisa Fujii**[3], **Alexander Ratner**[1], **Ranjay Krishna**[1], **Chen-Yu Lee**[2], **Tomas Pfister**[2]
[1]University of Washington, [2]Google Cloud AI Research, [3]Google Research
cydhsieh@cs.washington.edu

Our task in focus: learning to “abstract” in language

In NLP: ~ “sentence summarization”

✨ New observation: “paraphrasing” can be viewed as a special case of “summarization” ✨

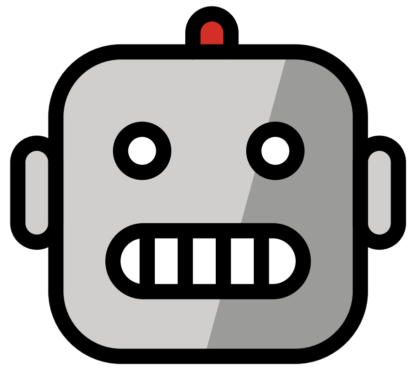
Mission Impossible:

🔥Learn to "summarize" + "paraphrase"🔥

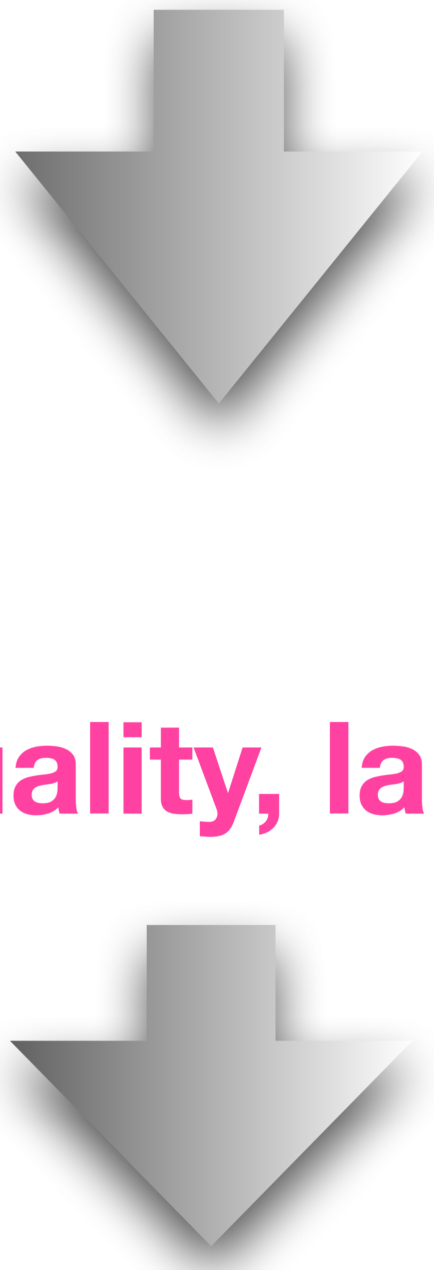
- without extreme-scale pre-training
- without RL with human feedback at scale
- without supervised datasets at scale

AI is as good as the data it was trained on winning recipe = extreme-scale pre-training + RLHF at scale

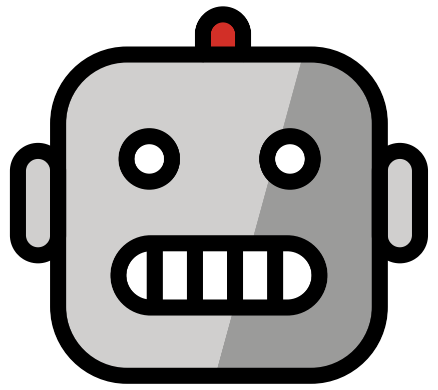
GPT-2



Low-quality, small models



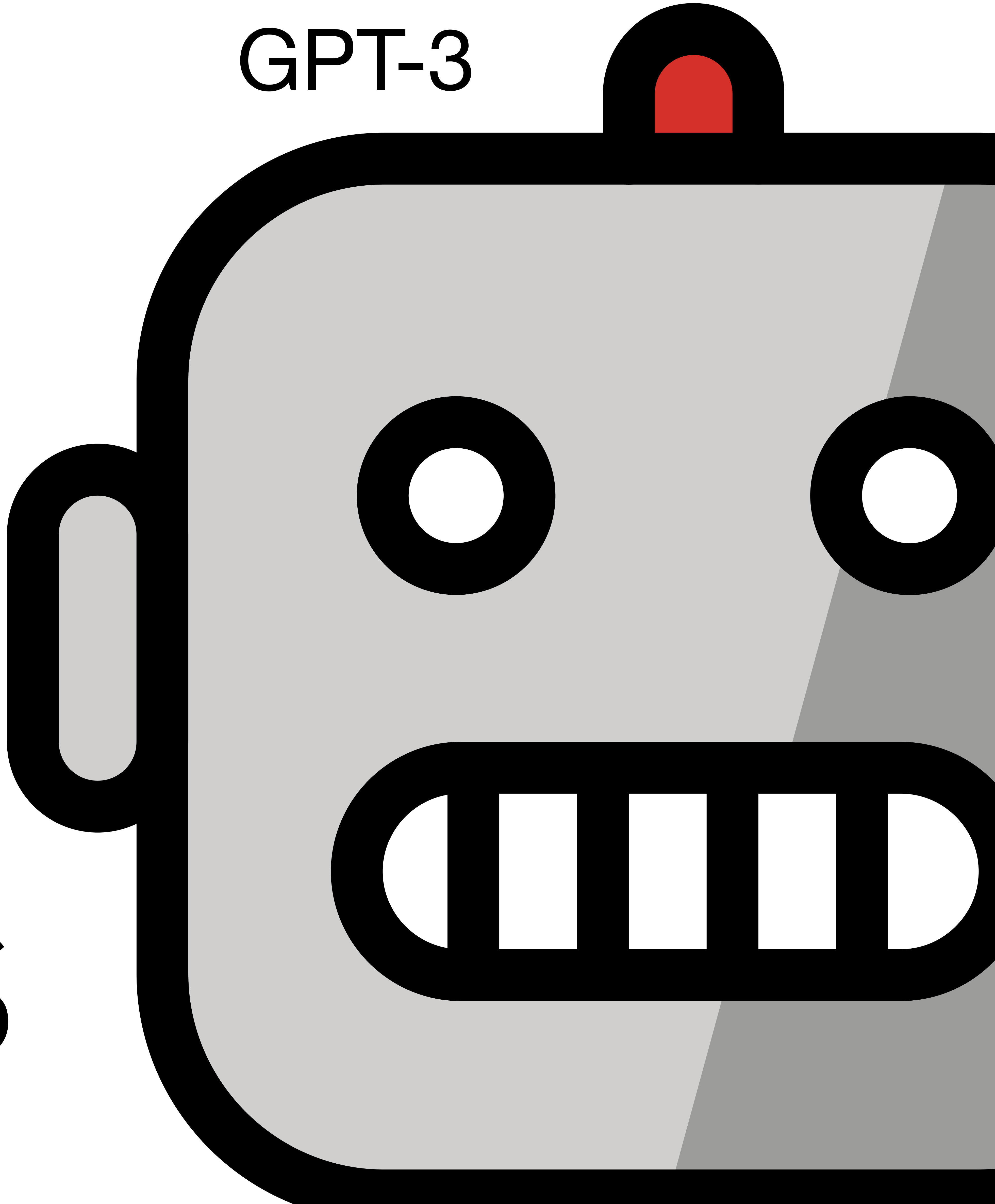
high-quality, large datasets



high-quality, small models

VS

GPT-3



We will build on …

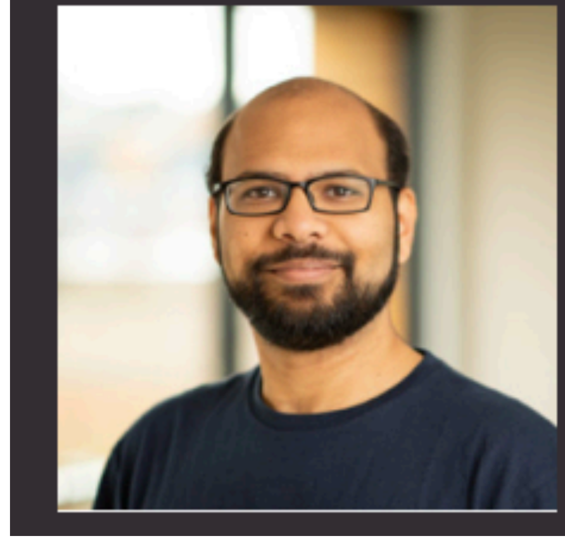
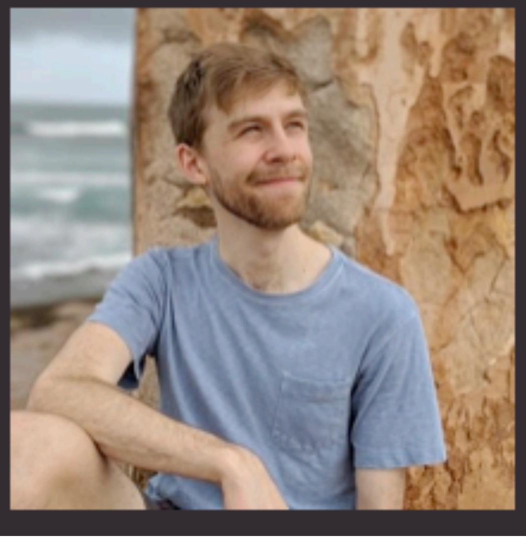
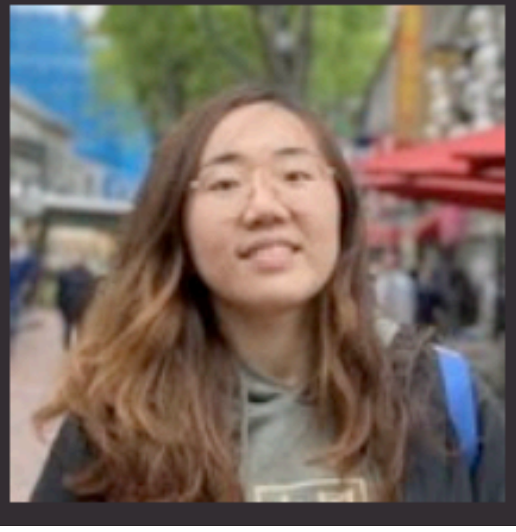
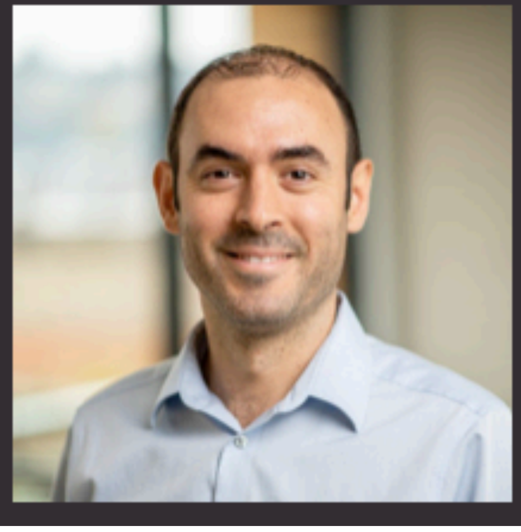
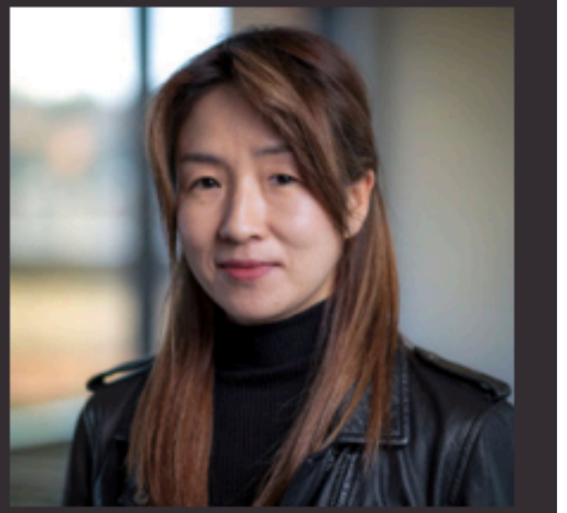
Symbolic Knowledge Distillation

From General Language Models to Commonsense Models

— NAACL 2022 —

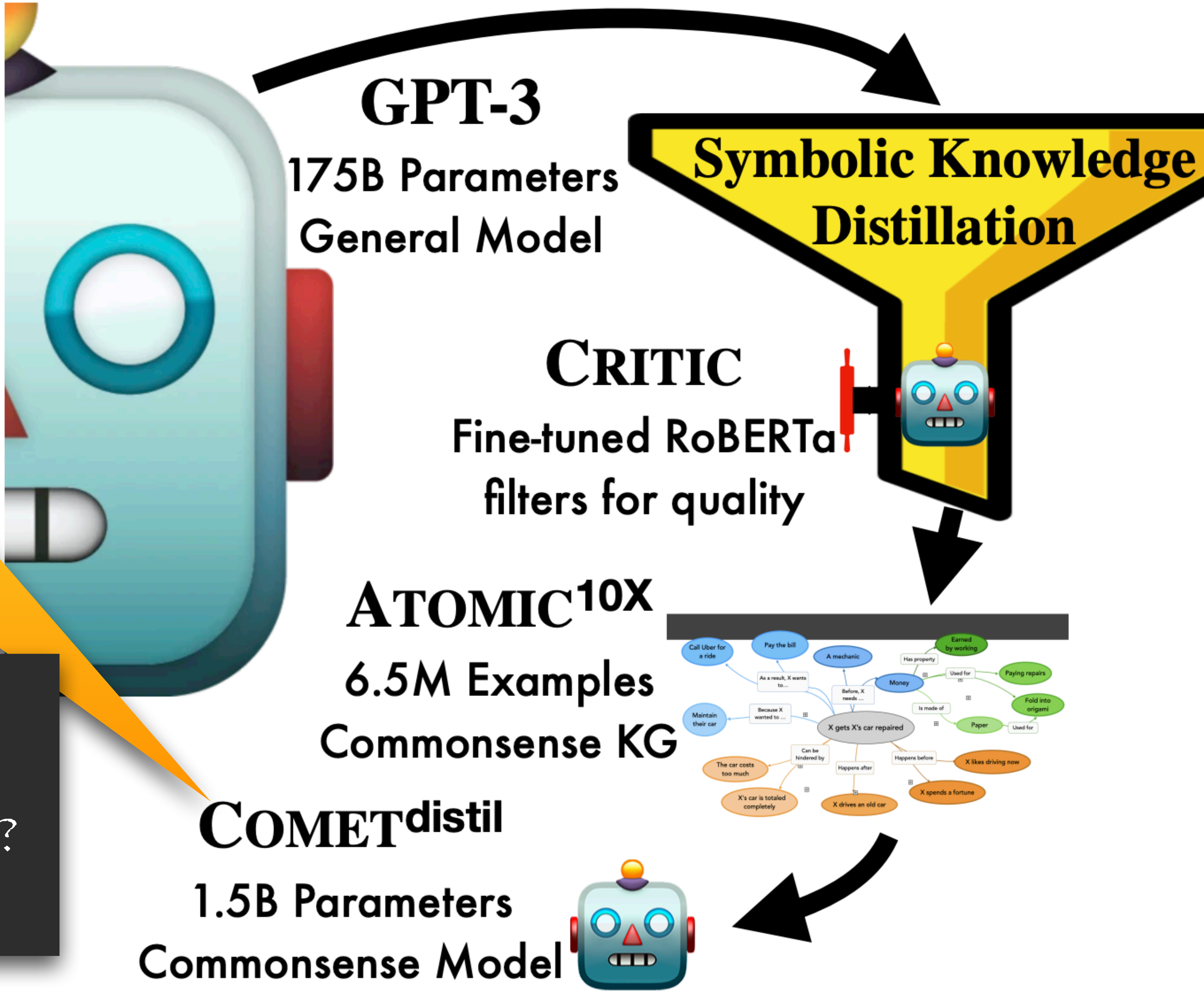
Peter West

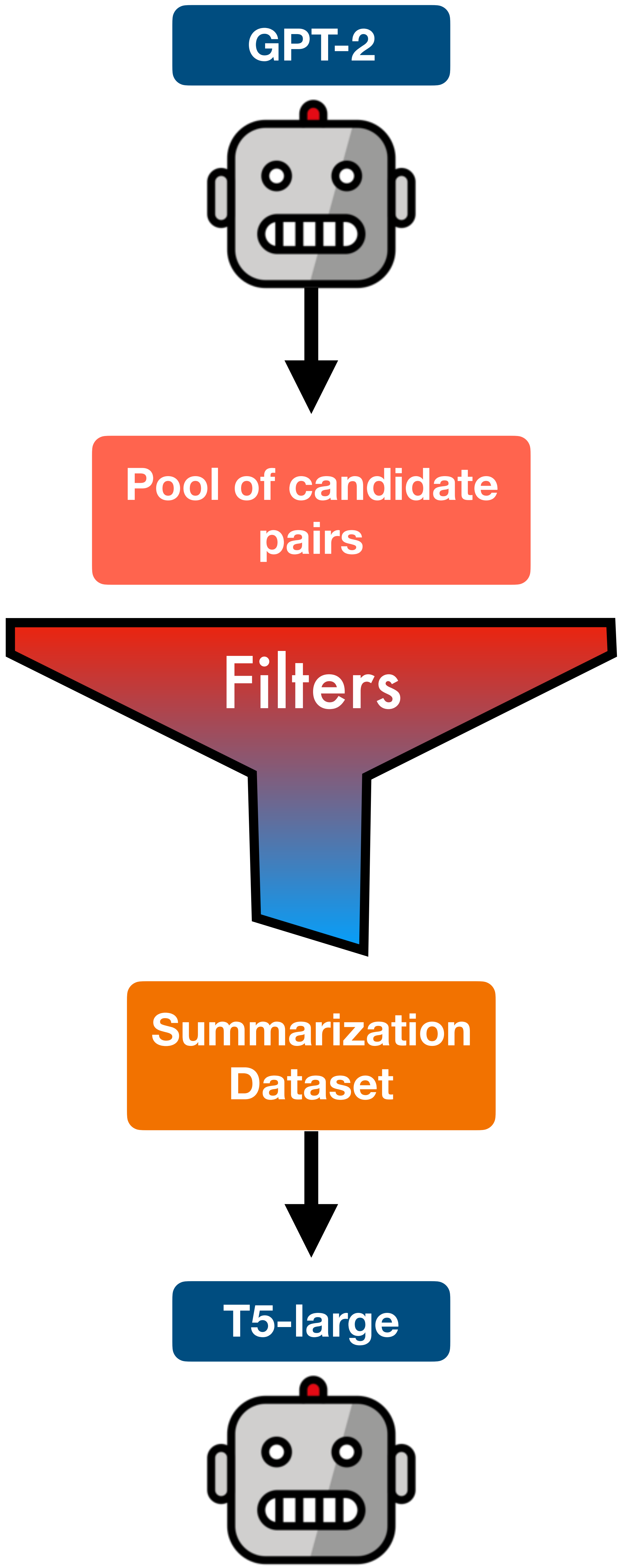
New:
ATOMIC-10x
COMET-distill

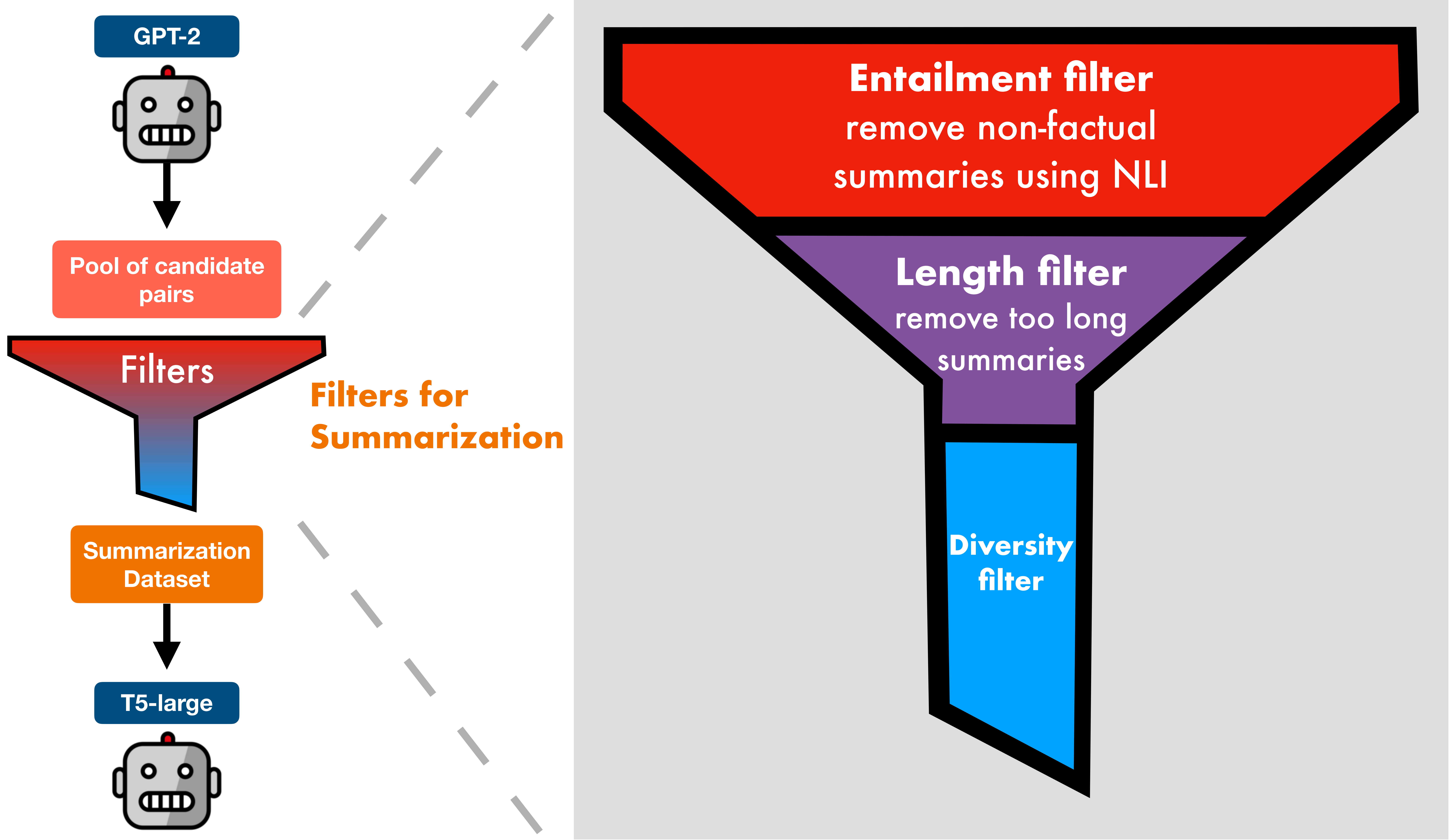
Chandra Bhagavatula	Jack Hessel	Jena Hwang	Liwei Jiang	Ronan Le Bras	Ximing Lu	Sean Welleck	Yejin Choi
							

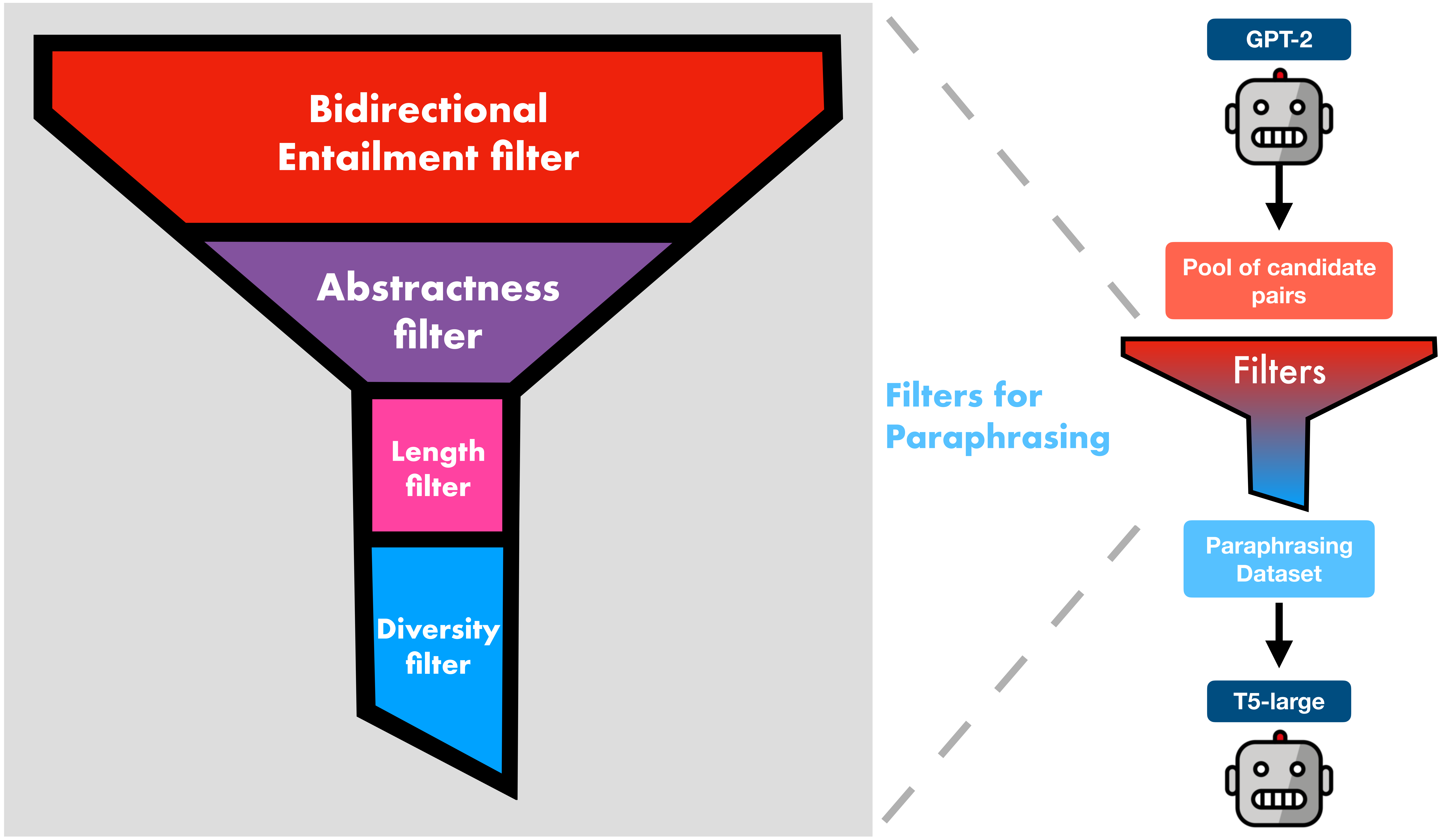
ATOMIC-10x: a machine-authored KB that wins, for the first time, over a human-authored KB in all criteria: scale, accuracy, and diversity.

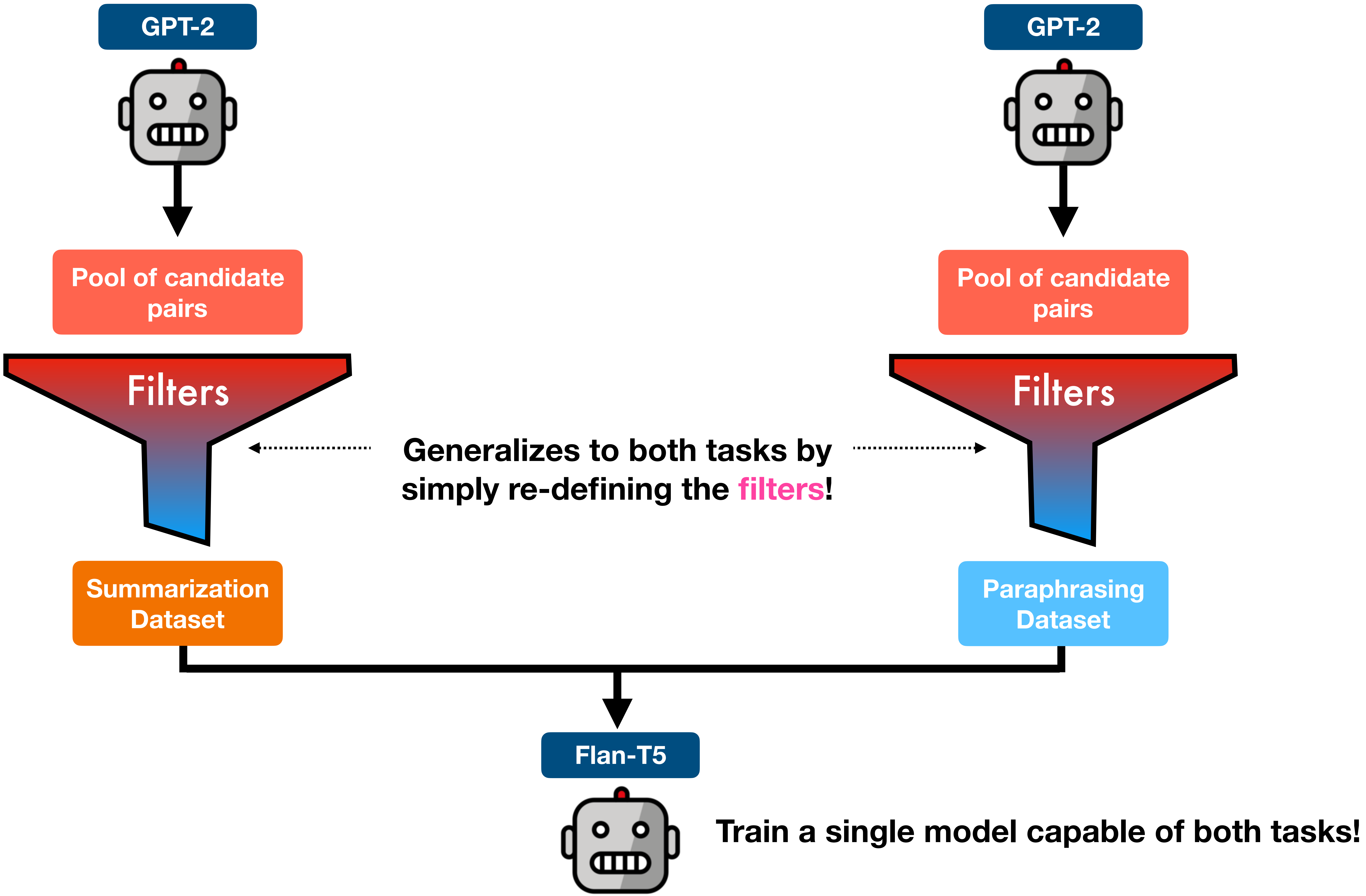
Yeah but can we get anywhere without GPT-3?

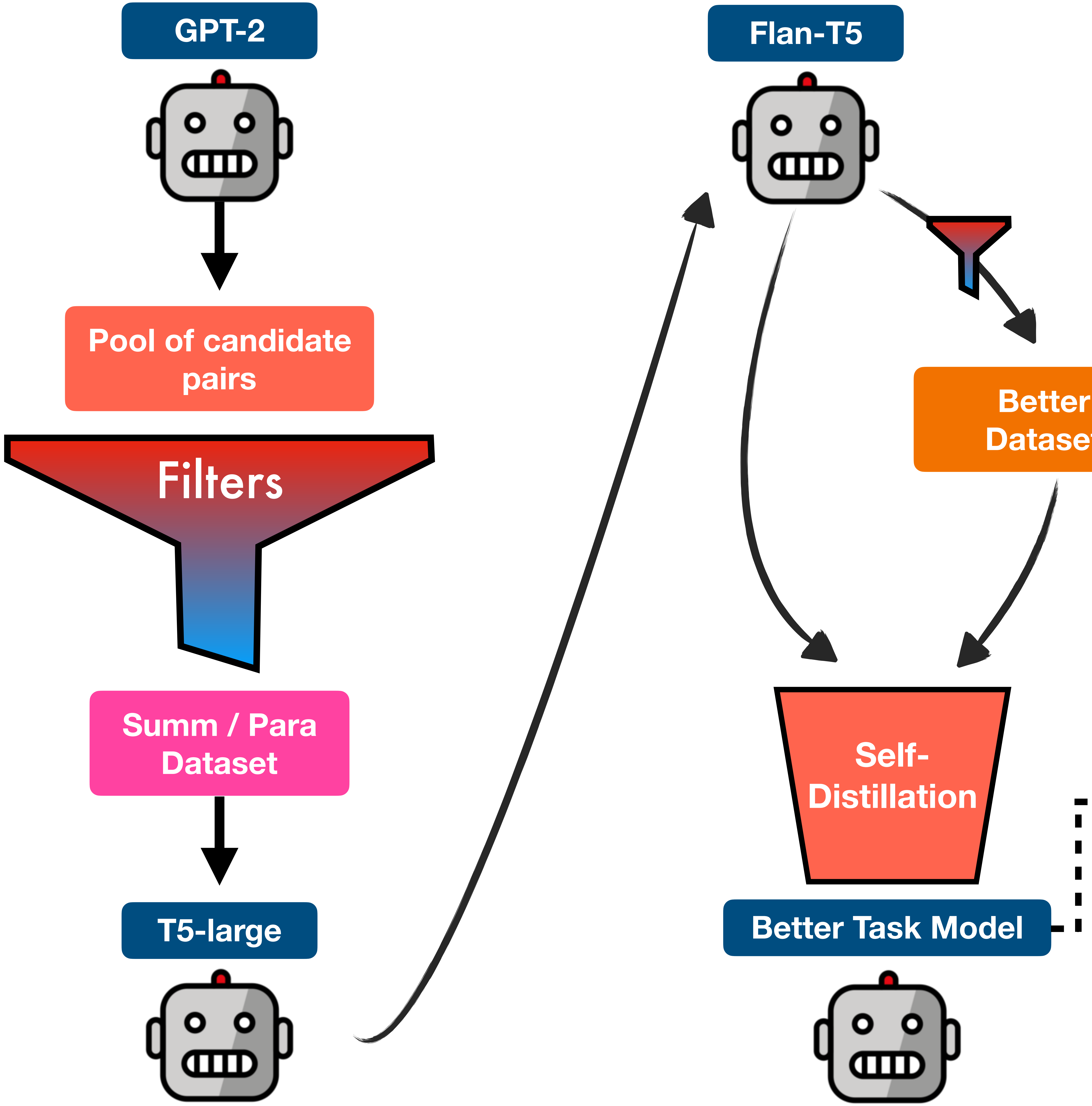



GPT-2
Pool of candidate pairs
Filters
Summarization Dataset
T5-large


GPT-2
Pool of candidate pairs
Filters
Filters for Summarization
Summarization Dataset
T5-large
Entailment filter
remove non-factual summaries using NLI
Length filter
remove too long summaries
Diversity filter


Bidirectional Entailment filter
Abstractness filter
Length filter
Diversity filter
Filters for Paraphrasing
GPT-2
Pool of candidate pairs
Filters
Paraphrasing Dataset
T5-large


GPT-2
GPT-2
Pool of candidate pairs
Pool of candidate pairs
Filters
Filters
Generalizes to both tasks by simply re-defining the filters!
Summarization Dataset
Paraphrasing Dataset
Flan-T5
Train a single model capable of both tasks!



Self-Distillation yields better dataset, stronger task model

DimSum+

3.4M samples for
sentence summarization + paraphrasing,
spanning news / reddit / bio domains

T5-ImpDistill

770M LM capable of both
controllable summarization + paraphrasing,
distilled purely from < 2B LMs

"While we will be looking across all parts of the newsroom, at the end of the redundancy program we expect there will be significantly fewer editorial management, video, presentation and section writer roles," the publisher is quoted as saying in an internal note.

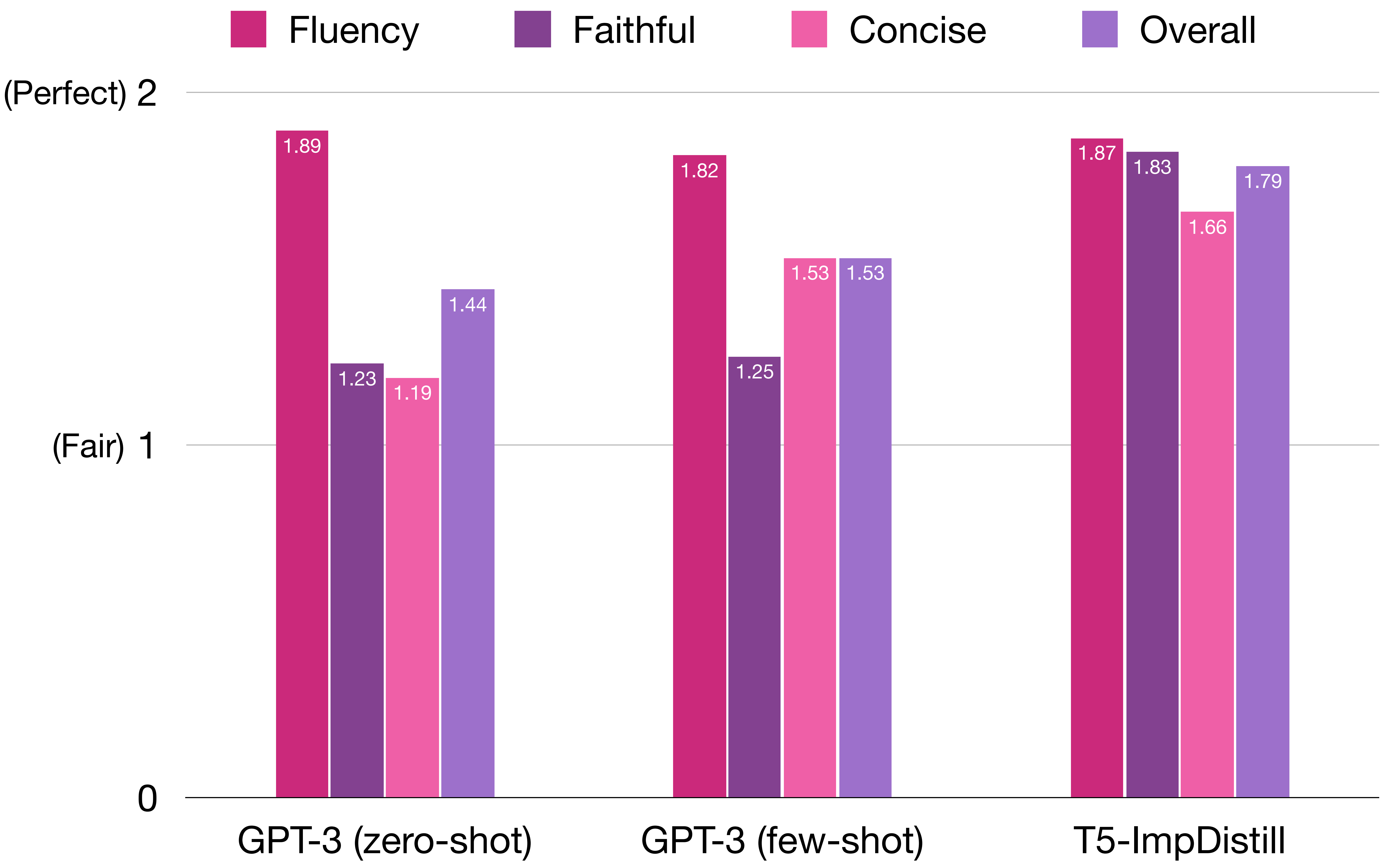
🔥 **T5-ImpDistill**

"We are looking to reduce the number of staff in the newsroom", the publisher said in an internal note. ✔

***GPT-3** text-davinci-003, zero-shot*

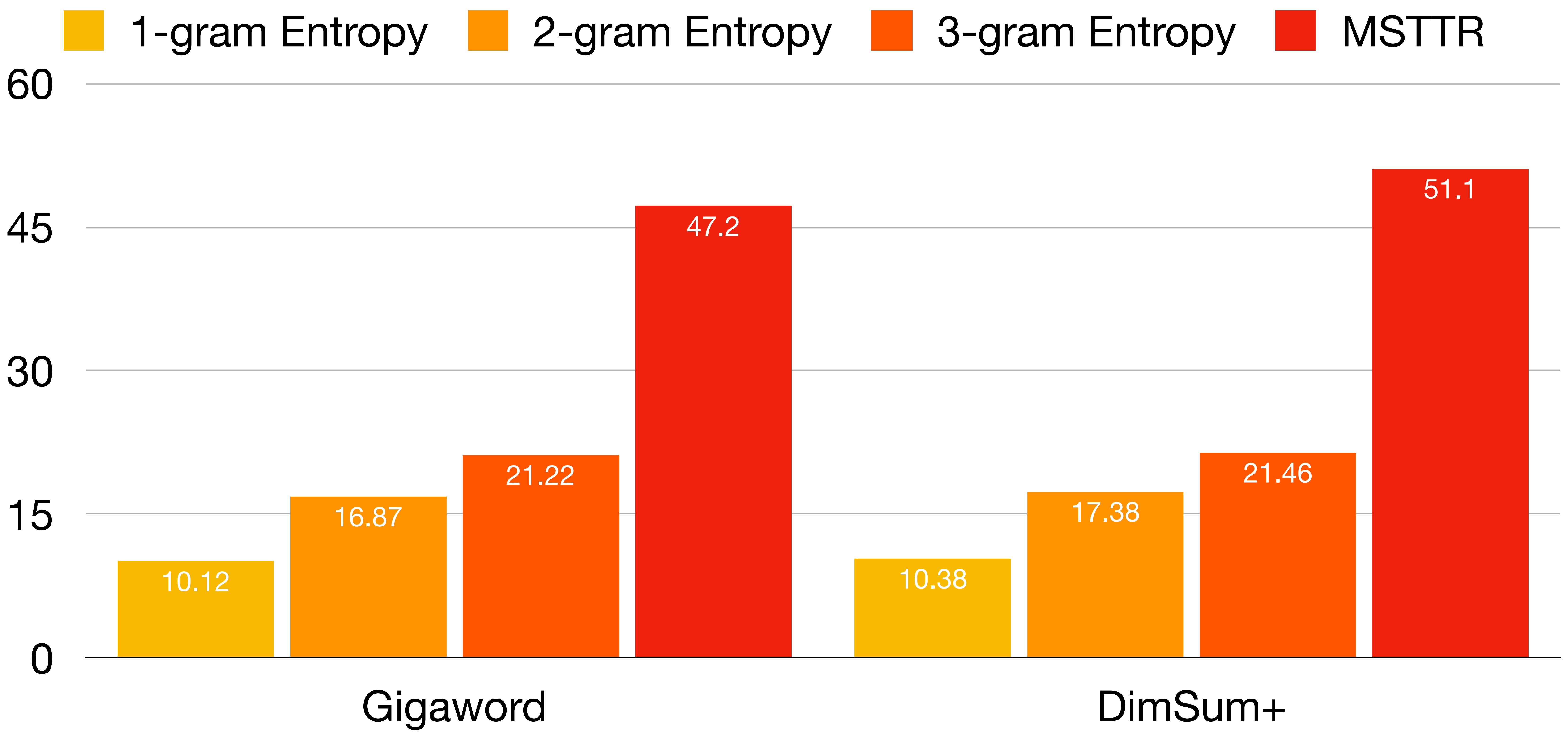
The publisher has informed staff through an internal note that, after implementing a redundancy program, there will be a significant reduction in the number of editorial management, video, presentation, and section writer roles. ✘ ***hallucinating unsupported content***

Stronger than **200x larger GPT-3** in human evaluation!



Dataset has **higher diversity** than human-authored **Gigaword**

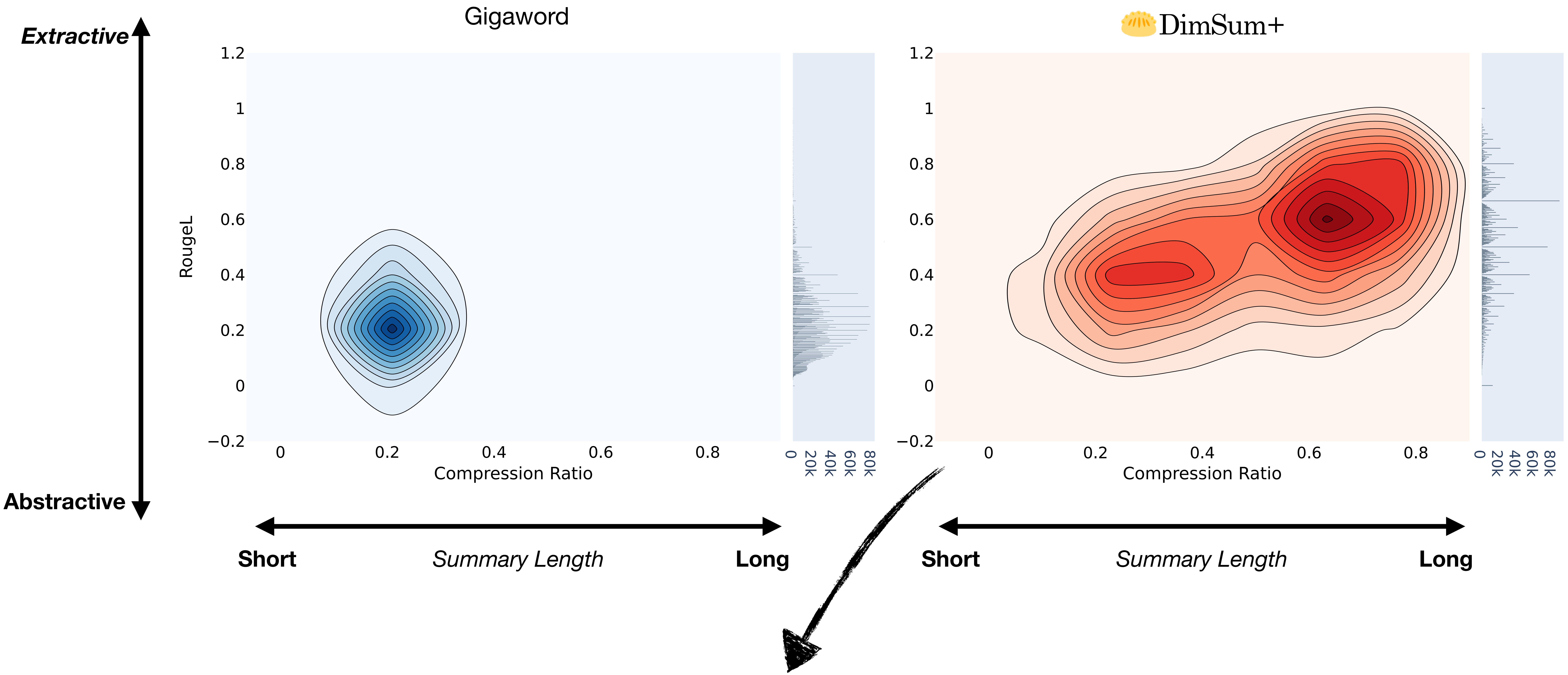
(Rush et al. 2015)



Our dataset (3.4M) exhibit more **lexical diversity** than human-authored Gigaword (4M)!

Dataset has **higher diversity** than human-authored **Gigaword**

(Rush et al. 2015)



Our dataset covers **diverse summarization strategy**!

PLASMA

Making Small Language Models Better Procedural Knowledge Model

— arXiv:2305.19472 —

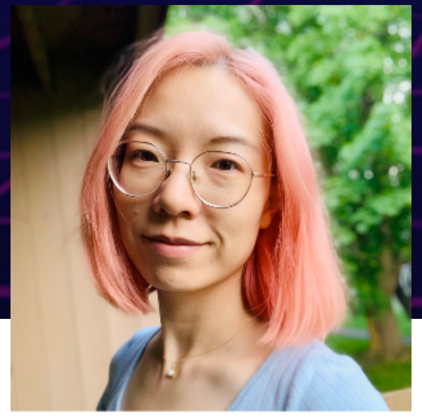
Faeze Brahman

Chandra Bhagavatula

Jena Hwang

Valentina Pyatkin

Lorraine Li



Hirona Arai

Soumya Sanyal

Keisuke Sakaguchi

Xiang Ren

Yejin Choi

PLASMA Making Small Language Models Better Action Reasoners

GOAL Buy a new videogame

PLAN

- **Step1:** Decide to buy a new videogame
- **Step2:** Decide on game
- **Step3:** Get into the car
- **Step4:** Drive to the game store
- **Step5:** Walk into the game store
- **Step6:** Find desired game
- **Step7:** Buy a new videogame

GOAL Buy a new videogame

CONDITION If the game store is too far away

PLASMA Making Small Language Models Better Action Reasoners

GOAL Buy a new videogame

PLAN

- **Step1:** Decide to buy a new videogame
- **Step2:** Decide on game
- **Step3:** Get into the car
- **Step4:** Drive to the game store
- **Step5:** Walk into the game store
- **Step6:** Find desired game
- **Step7:** Buy a new videogame

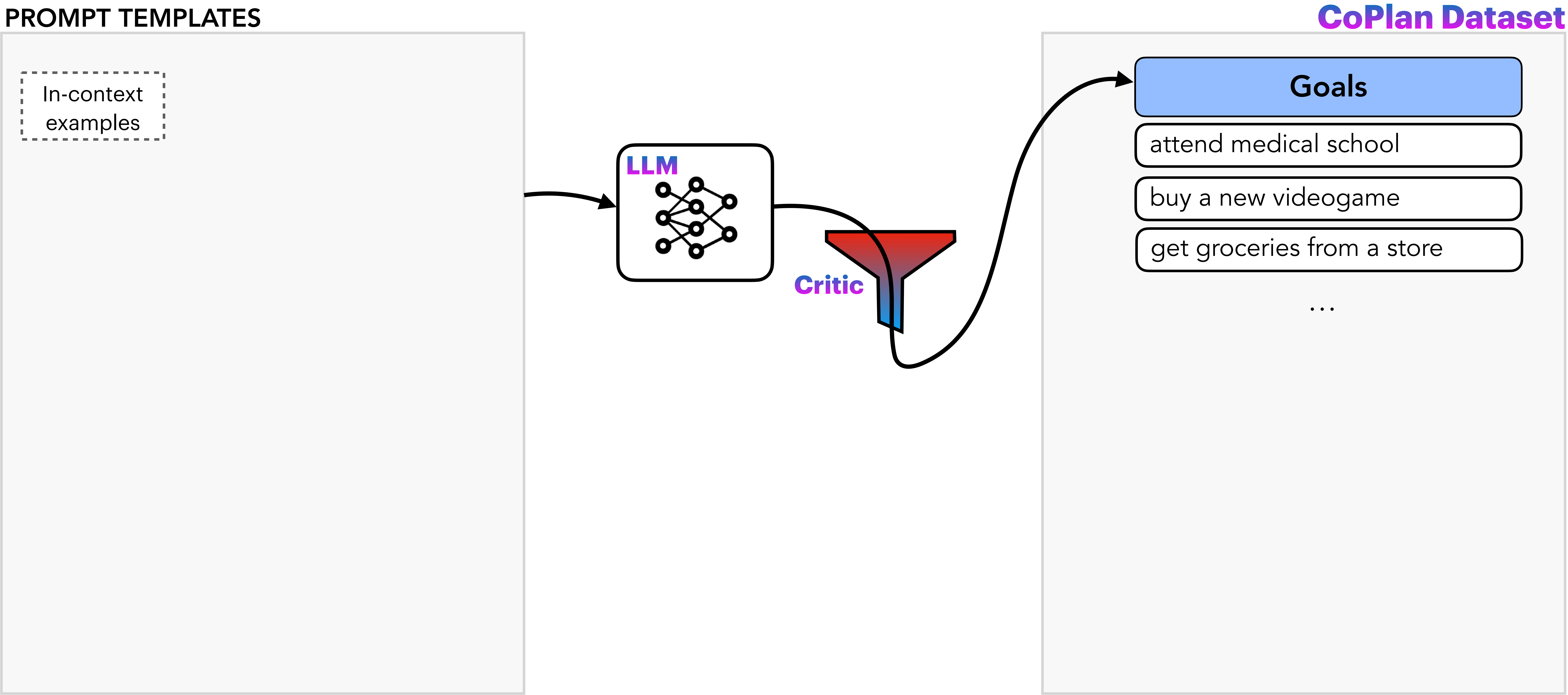
GOAL Buy a new videogame

CONDITION If the game store is too far away

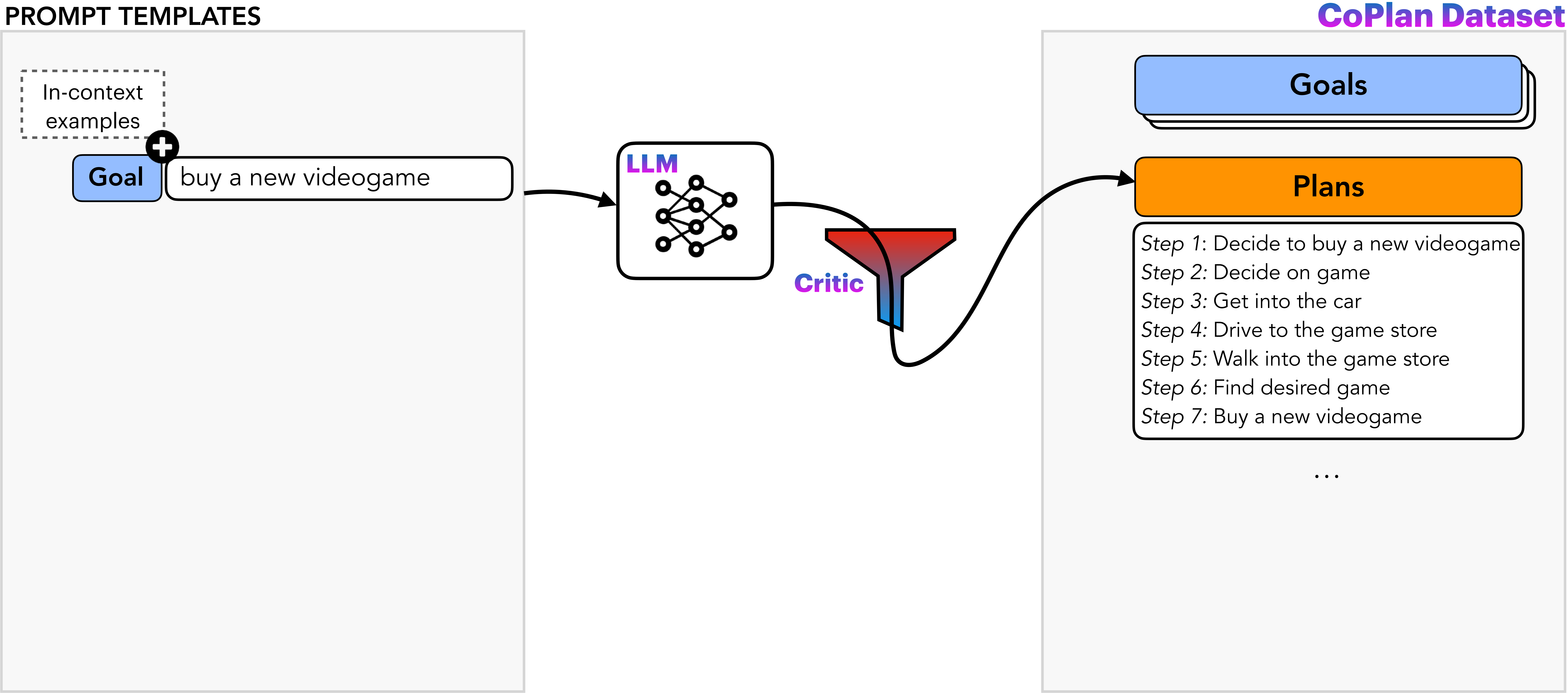
COUNTERFACTUAL PLAN

- **Step1:** Decide to buy a new videogame
- **Step2:** Research online game stores
- **Step3:** Compare prices and reviews
- **Step4:** Select the game to buy
- **Step5:** Purchase the game online
- **Step6:** Wait for the game to be delivered

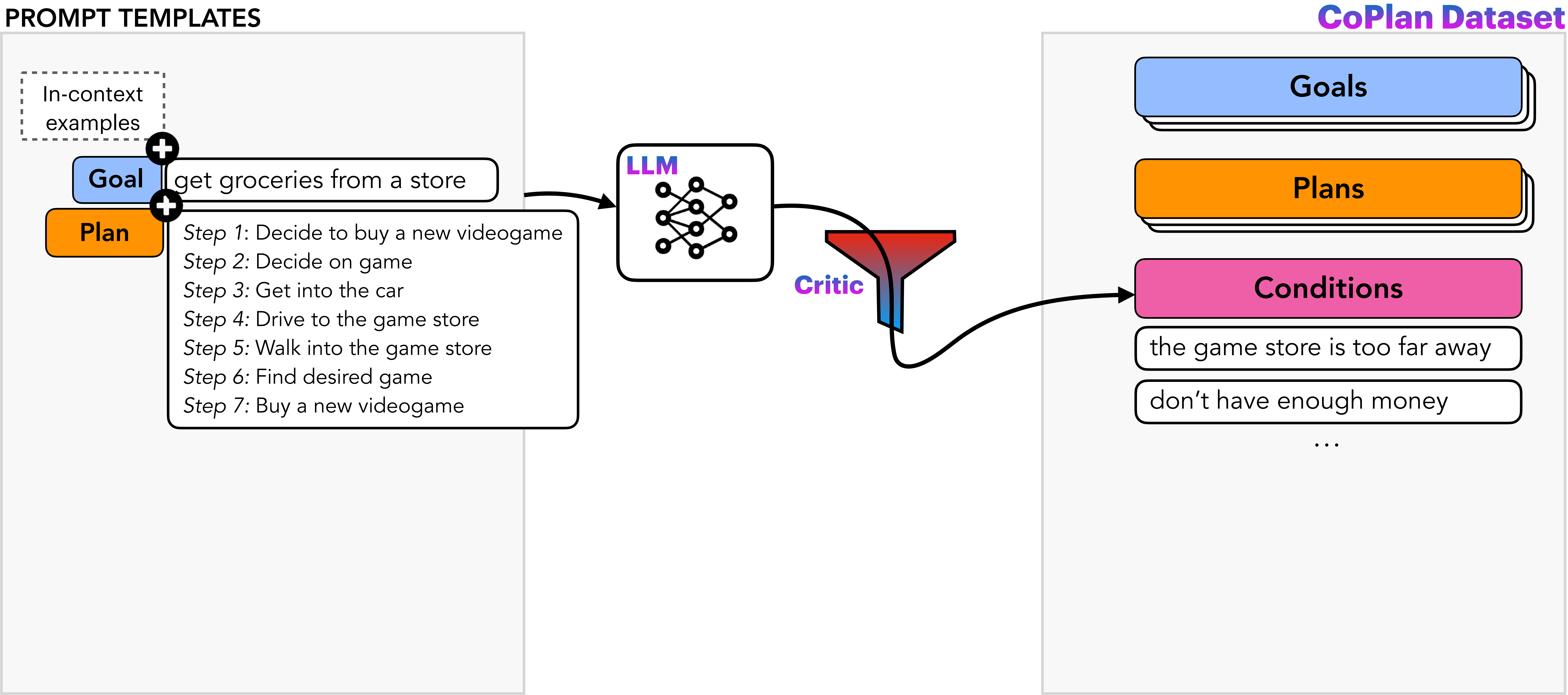
Procedural Knowledge Verbalization



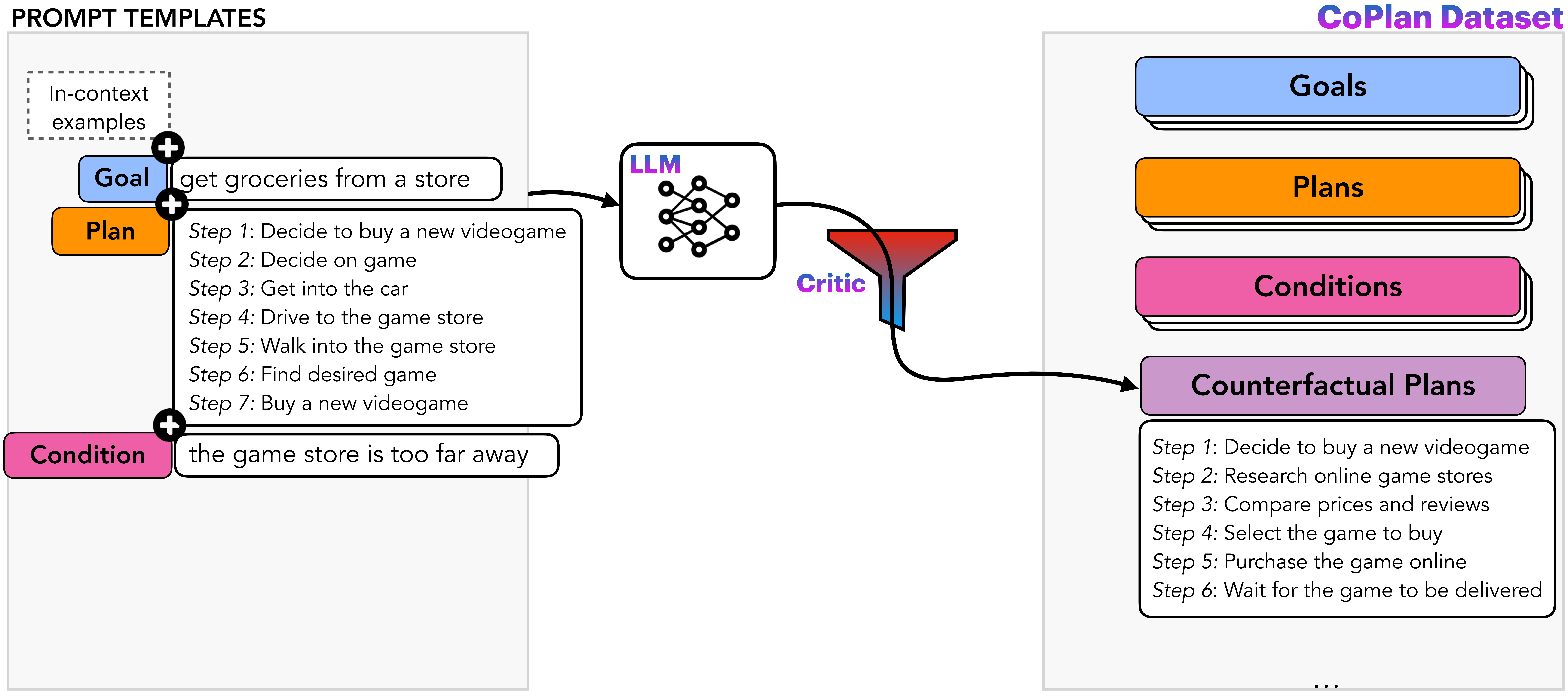
Procedural Knowledge Verbalization



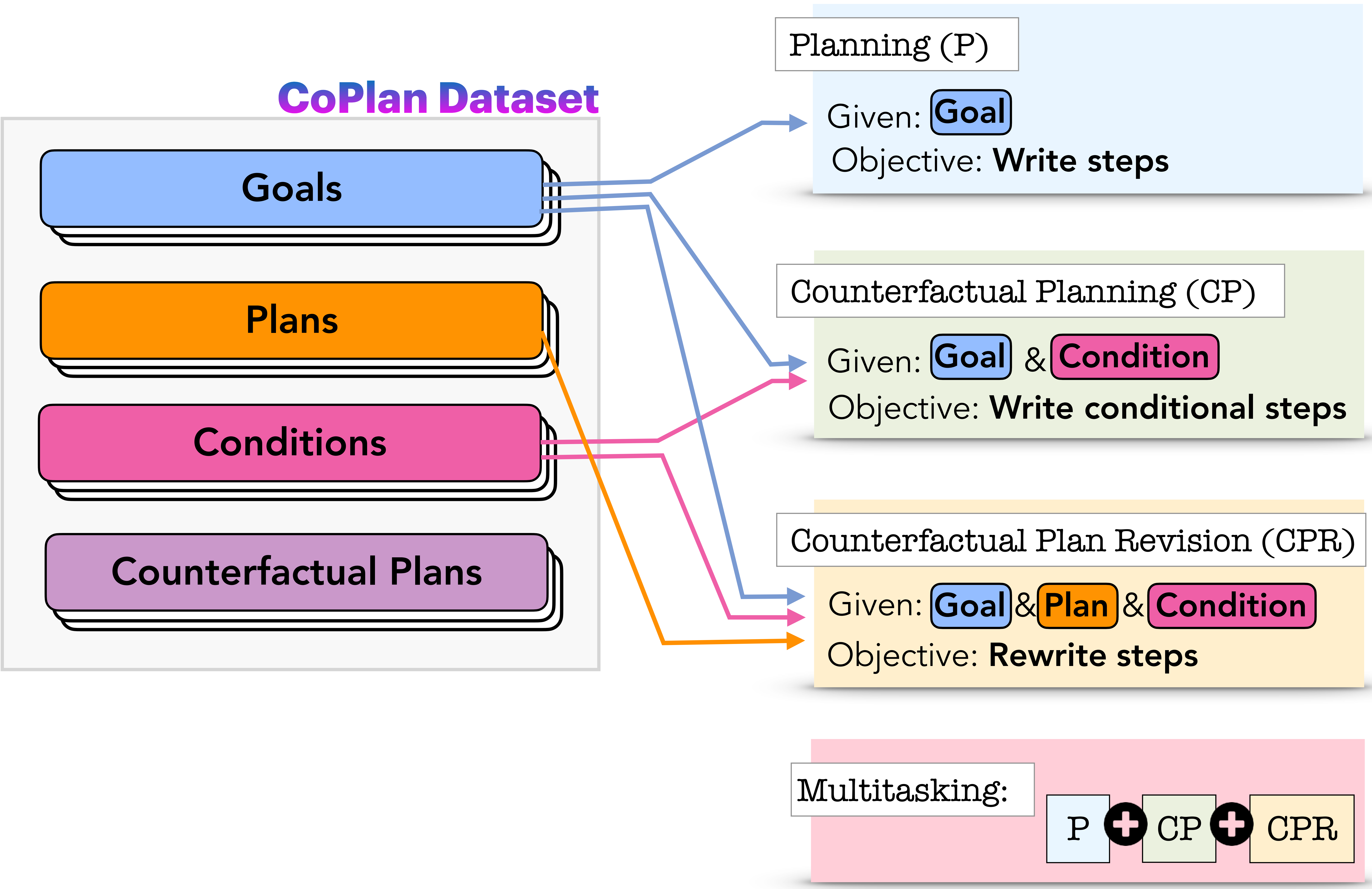
Procedural Knowledge Verbalization



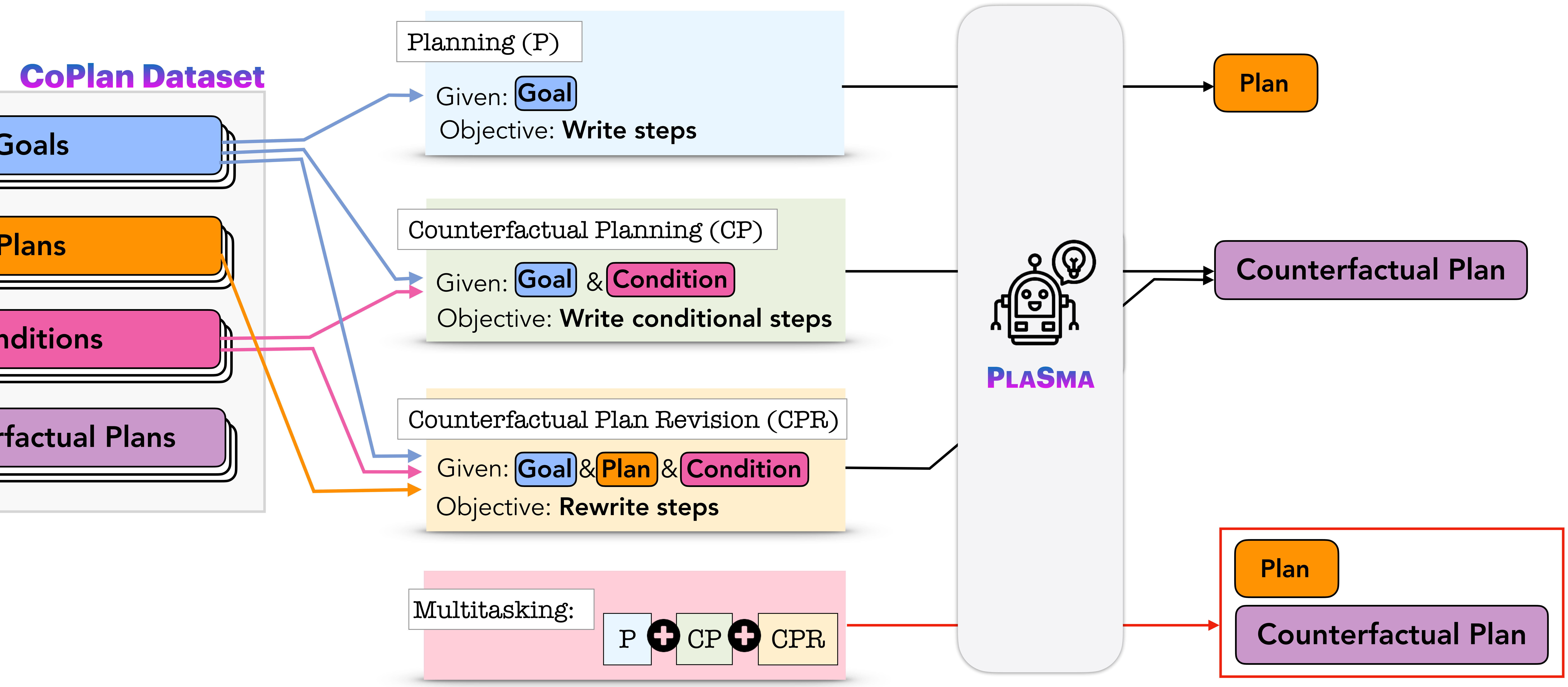
Procedural Knowledge Verbalization



Procedural Knowledge Distillation



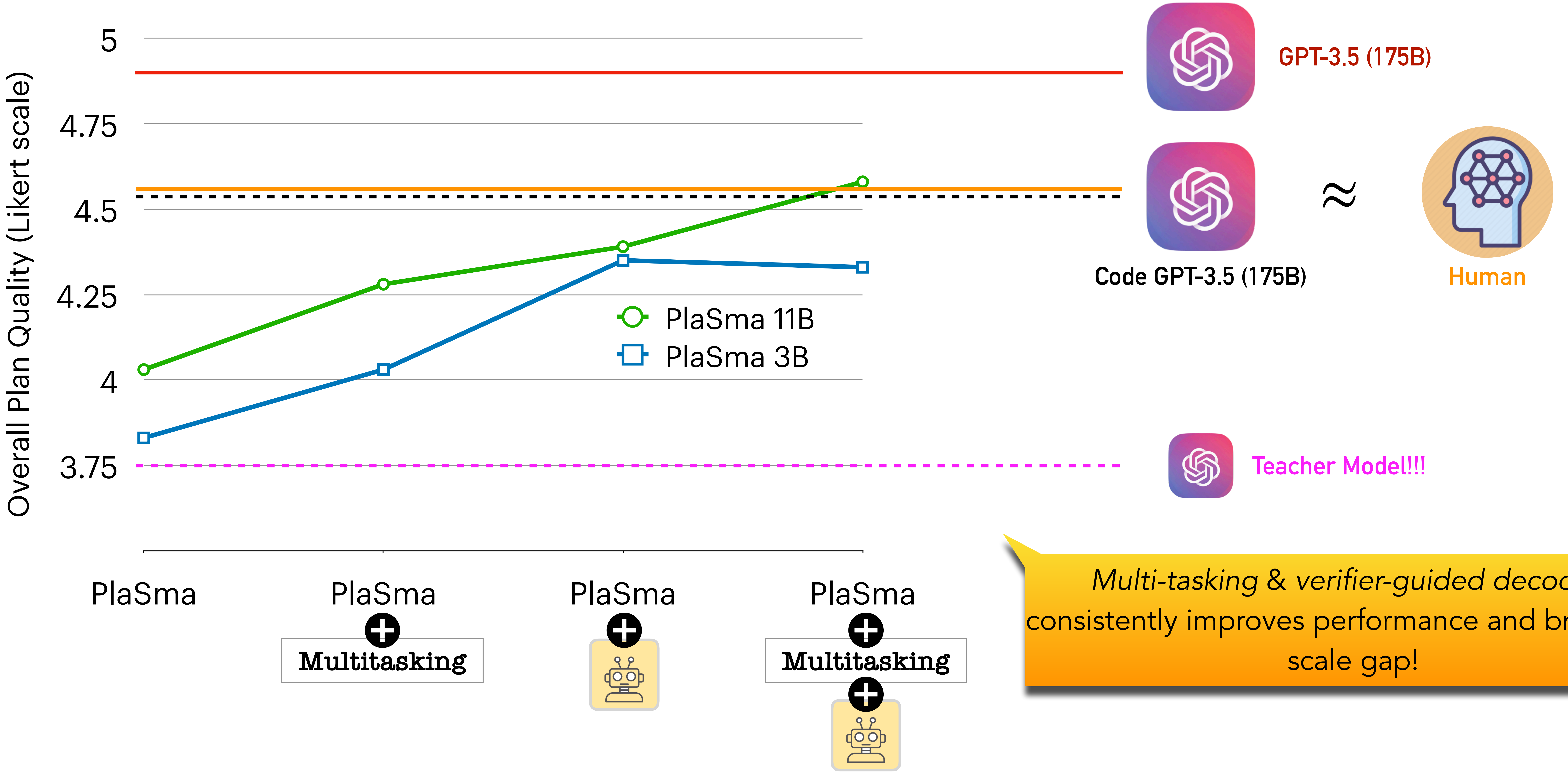
Procedural Knowledge Distillation



Verifier-Guided Decoding

Verifier-Guided Decoding

PLASMA helps the scale gap!



Multi-tasking & verifier-guided decoding consistently improves performance and bridge the scale gap!

Does PLASMA help downstream Embodied tasks?

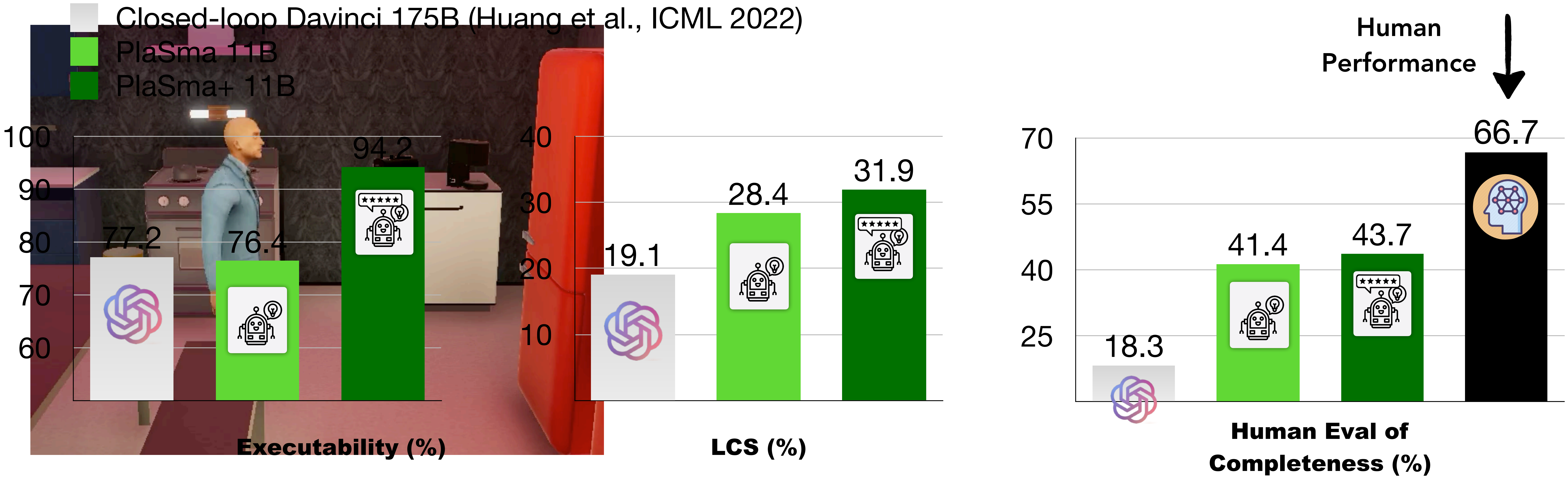
VirtualHome Environment

Task: Turn of TV

Ground to predefined set of actionable !

Does PLASMA help downstream Embodied tasks?

VirtualHome Environment



PlaSma, 16 times smaller model generates steps that are significantly more **EXECUTABLE** and **COMPLETE**!!!

PLASMA can do **counterfactual planning!**

Smaller models can perform counterfactual planning with same level of proficiency as larger models!

Goal + **Condition**

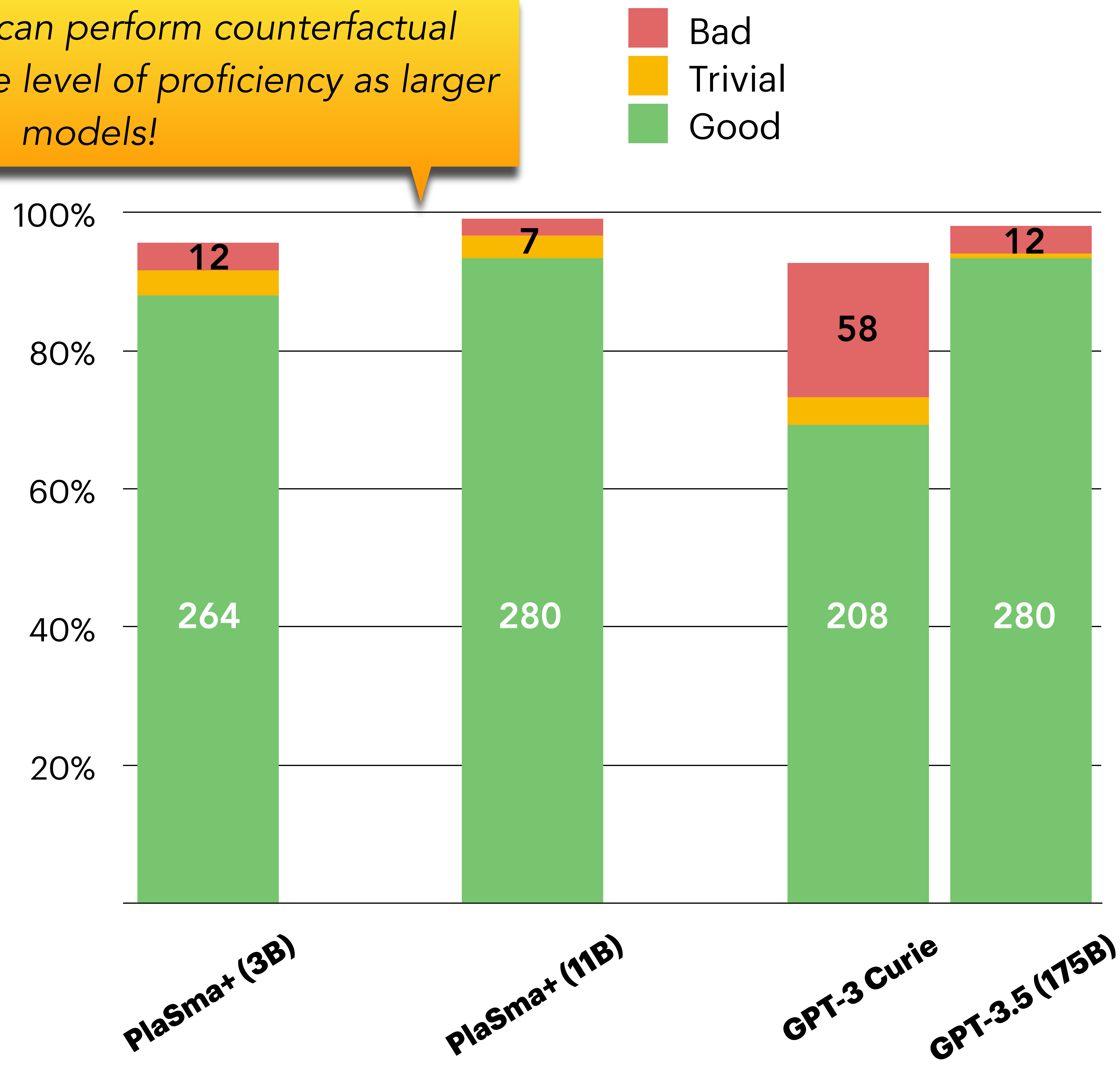
Empty lint filter — If the lint trap is too hot to touch

Get exercise — If the fitness activity presents a risk of injury

PlaSma Plan

Step1: Open dryer Step2: Locate lint trap **Step4: Use heat-resistant glove or mitt to pick up lint trap** Step5: Pull lint trap out of dryer Step6: Empty lint trap

Step1: Decide what type of exercise you would like to do **Step2: Research the activity to make sure you understand the risks and benefits Step3: Consult with a medical professional to ensure the activity is safe for you** Step4: Find a place to do the activity Step5: Set a goal for how often you would like to do the activity Step6: Start doing the activity **Step7: Monitor your progress and adjust the activity as needed** Step8: Get exercise



2050: An AI Odyssey

Prolog: what CVPR 2050 be like

Chapter 1: The Possible Impossibilities

Chapter 2: The Impossible Possibilities

Chapter 3: The Paradox

Epilog: why am I even here? A confession of an alien

Everything, everywhere, all at once

Passed the bar exam

GPT-4

Existential risk

AI not yet as smart as a dog

Chapter 3: The Paradox

- Commonsense paradox
- Moravec's paradox
- Generative AI paradox

Dark matter is what matters in modern physics

- Only 5% of universe is normal matter. The remaining 95% is dark matter and dark energy.
- Dark matter is completely invisible, yet affects what are visible: the orbits of stars and the trajectory of light

Dark matter of language?

Normal matter: visible text (words, sentences)

Dark matter: the unspoken rules of how the world works, which influence the way people use and interpret language

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305,

*Correspondence to: michalk@stanford.ec

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer D. Ullman
Department of Psychology
Harvard University
Cambridge, MA, 02138
tullman@fas.harvard.edu

**Neural Theory-of-Mind?
On the Limits of Social Intelligence in Large LMs**

Maarten Sap[♠◇] **Ronan Le Bras**[♠] **Daniel Fried**[◇] **Yejin Choi**[♠♡]
[♠]Allen Institute for AI, Seattle, WA, USA
[◇]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[♡]Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA
maartensap@cmu.edu

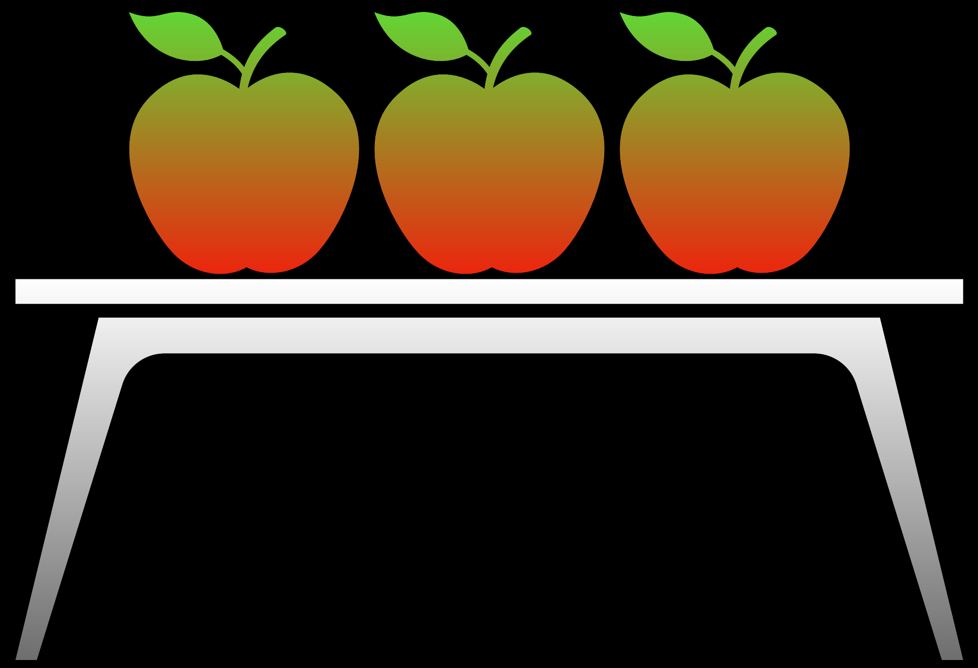
Circa 2022... (GPT-3)

“theory of mind” test

Alice and Bob saw apples on the table in the kitchen.

Alice left the kitchen.

Bob moved the apples to the cabinet.



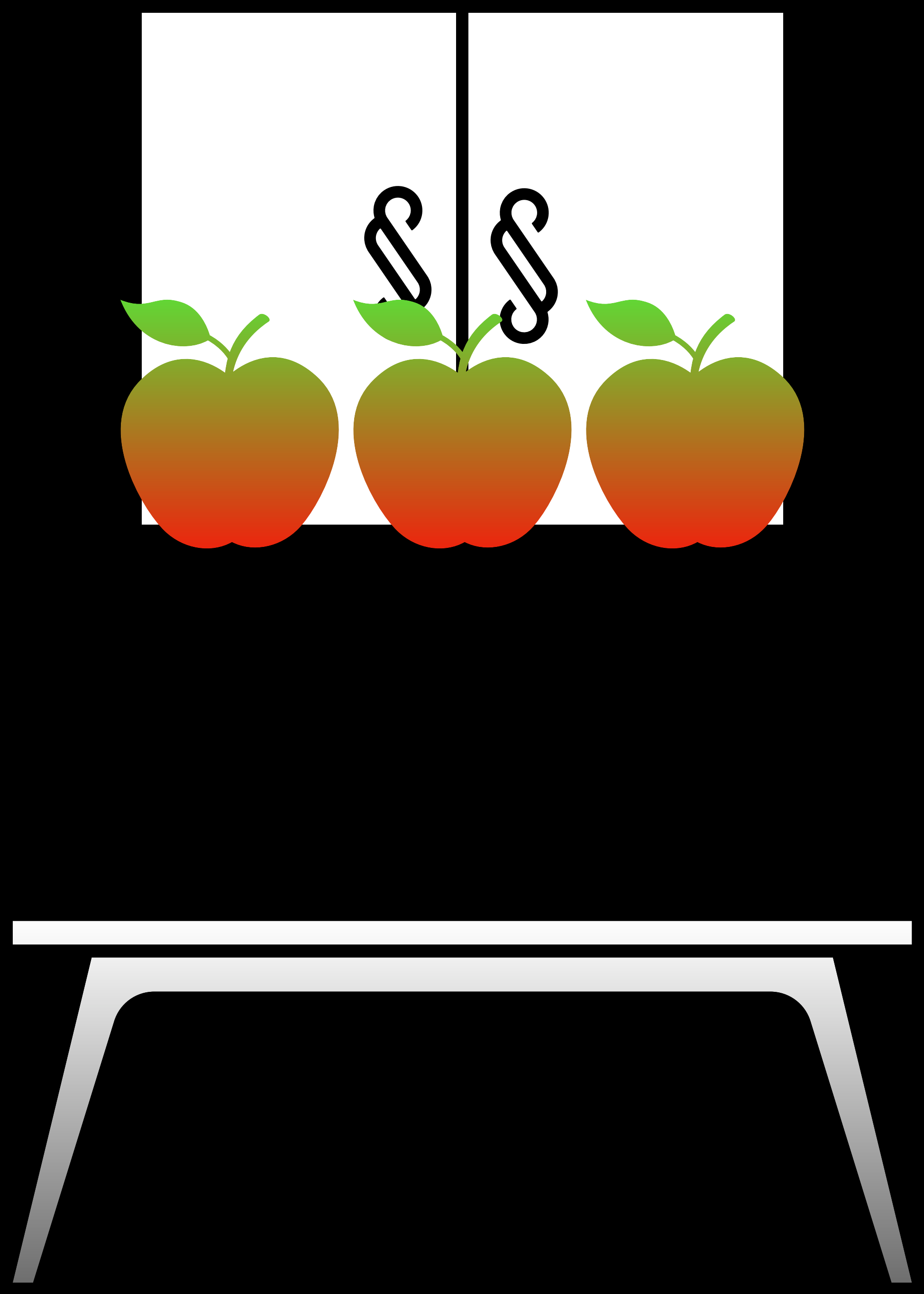
Circa 2022… (GPT-3)

“theory of mind” test

Alice and Bob saw apples on the table in the kitchen.

Alice left the kitchen.

Bob moved the apples to the cabinet.



Where would Bob think that Alice will look for the apples?

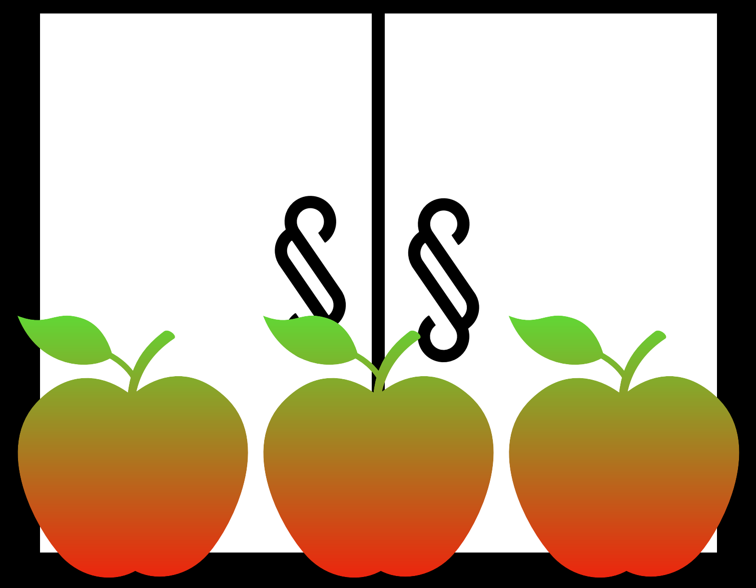
Circa 2022… (GPT-3)

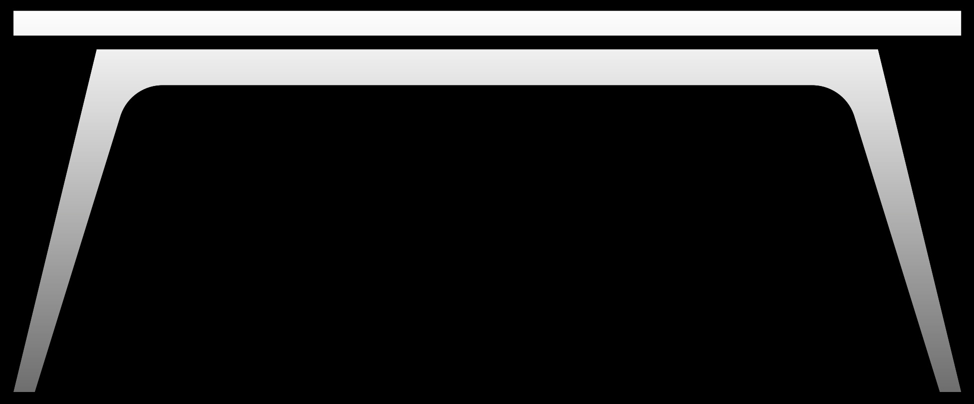
“theory of mind” test

Alice and Bob saw apples on the table in the kitchen.

Alice left the kitchen.

Bob moved the apples to the cabinet.





Where would Bob think that Alice will look for the apples?

in the cabinet

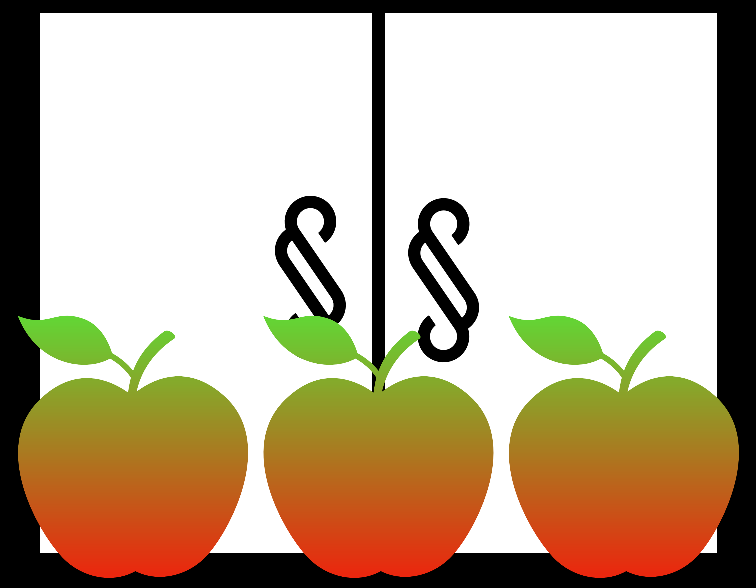
Circa 2023… (GPT-4)

“theory of mind” test

Alice and Bob saw apples on the table in the kitchen.

Alice left the kitchen.

Bob moved the apples to the cabinet.

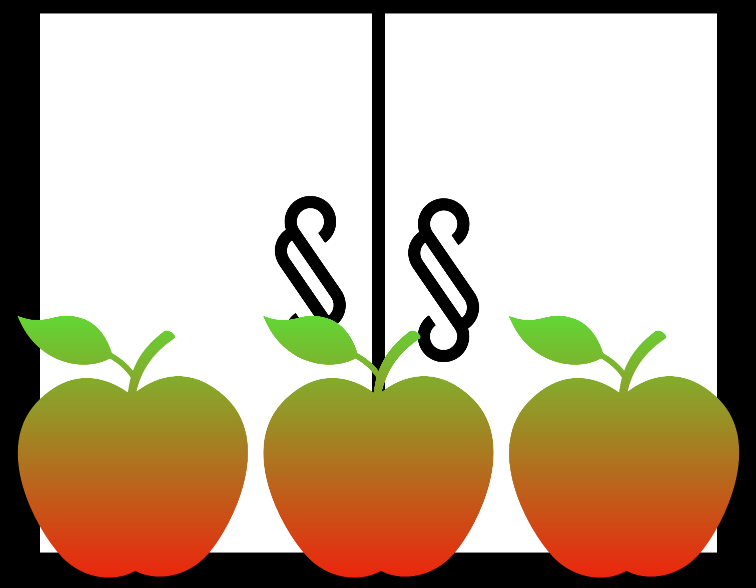


Where would Bob think that Alice will look for the apples?

On the table

Circa 2023… (GPT-4)

“theory of mind” test



Alice and Bob saw apples on the table in the kitchen.

Bob moved the apples to the cabinet.

Alice left the kitchen.

Where would Bob think that Alice will look for the apples?

On the table

Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker

ACL 2023

Melanie Sclar[1] Sachin Kumar[2] Peter West[1] Alane Suhr[3]
Yejin Choi[1,3] Yulia Tsvetkov[1]

GPT4 - 68%

Typical false-belief ToM story:

1 room
2 people*
2 containers
1 object

GPT4 - 58%

Variant I

2 ToM stories concatenated in 2 rooms?

GPT4 - 62%

Variant 2

3 people
3 containers,
moving 1 object sequentially?

GPT4 - 97%

Variant 3

1 room
2 people,
4 containers
moving 1 object sequentially?

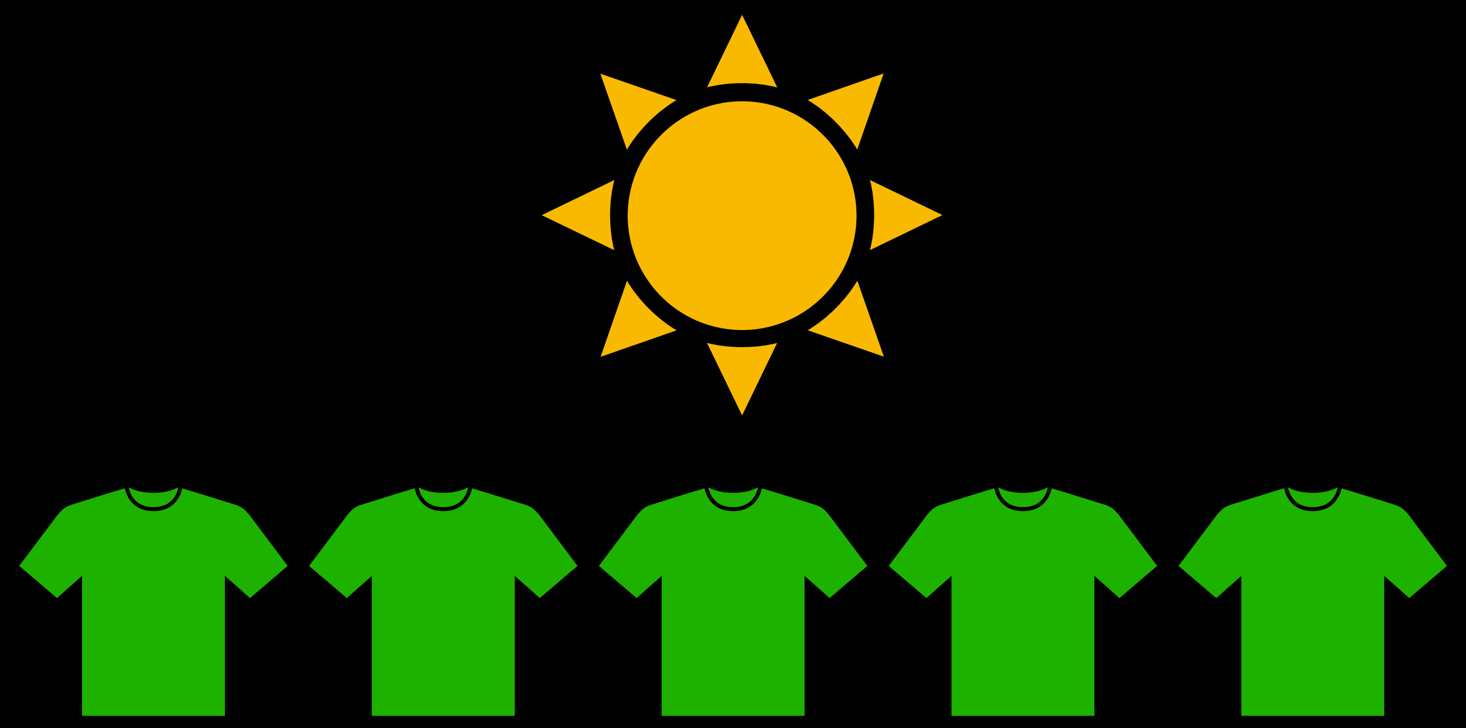
• • •

* with an extra distractor person (ToMi dataset)

Why AI is incredibly smart and shockingly stupid

1,207,112 views | Yejin Choi • TED2023

USER I left 5 clothes to dry out in the sun. It took them 5 hours to dry completely. How long would it take to dry 30 clothes?



ASSISTANT It would take 30 hours to dry 30 clothes.

Generated on Apr 12 2023 from https://platform.openai.com/playground?mode=chat&model=gpt-4

Model: GPT-4

If it takes 10 hours to dry 5 clothes, how long would it take 20 clothes to dry in the sun?

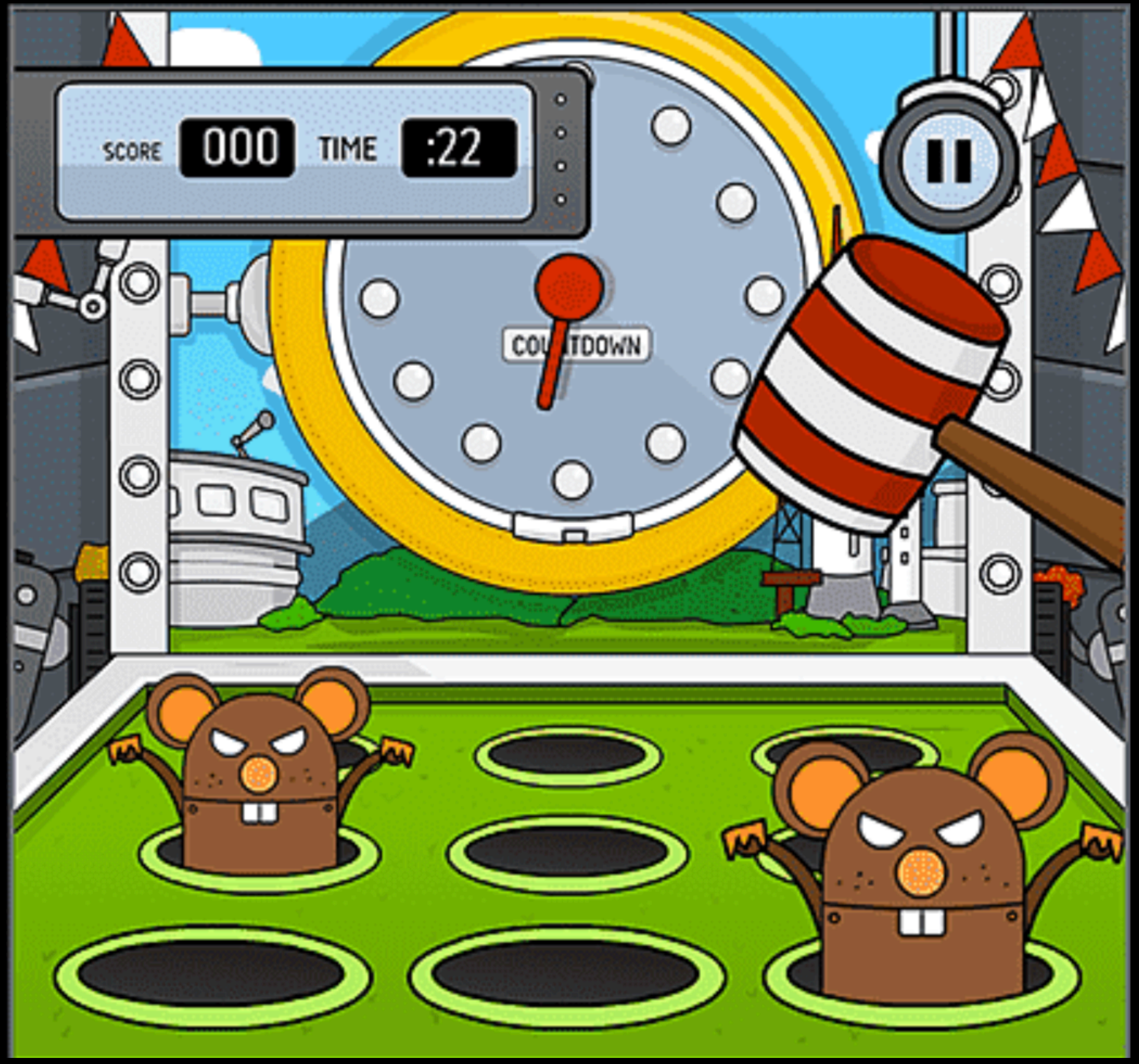
— GPT4, as of Jun 18 2023 —

The drying time for clothes is typically dependent on factors like the type of cloth, the intensity of the sunlight, humidity, wind, and so on. However, if we are assuming that the drying time scales linearly with the number of clothes, and that all other factors remain constant, we can calculate it as follows:

If it takes 10 hours to dry 5 clothes, then it takes 2 hours to dry 1 cloth (10 hours / 5 clothes = 2 hours/cloth).

So, it would take 40 hours to dry 20 clothes (20 clothes * 2 hours/cloth = 40 hours).

RLHF whack-a-mole game?



Model: GPT-4

If it takes 10 hours to dry 5 clothes, how long would it take 20 clothes to dry in the sun?

— GPT4, as of Jun 18 2023 —

The drying time for clothes is typically dependent on factors like the type of cloth, the intensity of the sunlight, humidity, wind, and so on. However, if we are assuming that the drying time scales linearly with the number of clothes, and that all other factors remain constant, we can calculate it as follows:

If it takes 10 hours to dry 5 clothes, then it takes 2 hours to dry 1 cloth (10 hours / 5 clothes = 2 hours/cloth).

So, it would take 40 hours to dry 20 clothes (20 clothes * 2 hours/cloth = 40 hours).

Commonsense Paradox

I'll dare say, the following four statements are all true:

- Commonsense is trivial for humans, hard for machines
- Among humans, "common sense is not so common" — Voltaire
- LLMs do acquire a vast amount of commonsense knowledge
- Yet in some ways, "AI is worse than a dog" — Yann Lecun

Common sense is not so common

Chapter 3: The Paradox

- Commonsense paradox
- Moravec's paradox
- Generative AI paradox

Moravec’s Paradox

— Hans Moravec, Rodney Brooks, Marvin Minsky, …

- contrary to traditional assumptions, (higher-level) reasoning requires little computation, but sensorimotor and perception skills require enormous computational resources
- it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility

Might it be that NLP is easier than Vision or Robotics?

AGI without strong vision or robotics capabilities?

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]

[1]project lead [2]joint first author [3]equal contribution [4]directional lead

Meta AI Research, FAIR

couldn't be possible without their 1B mask dataset innovation

DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2] Gabriel Ilharco*[1] Alex Fang*[1] Jonathan Hayase[1] Georgios Smyrnis[5]
Thao Nguyen[1] Ryan Marten[7,9] Mitchell Wortsman[1] Dhruba Ghosh[1] Jieyu Zhang[1]
Eyal Orgad[3] Rahim Entezari[10] Giannis Daras[5] Sarah Pratt[1] Vivek Ramanujan[1]
Yonatan Bitton[11] Kalyani Marathe[1] Stephen Mussmann[1] Richard Vencu[6]
Mehdi Cherti[6,8] Ranjay Krishna[1] Pang Wei Koh[1] Olga Saukh[10] Alexander Ratner[1]
Shuran Song[2] Hannaneh Hajishirzi[1,7] Ali Farhadi[1] Romain Beaumont[6]
Sewoong Oh[1] Alexandros G. Dimakis[5] Jenia Jitsev[6,8]
Yair Carmon[3] Vaishaal Shankar[4] Ludwig Schmidt[1,6,7]

Compared to LLMs, we don't yet have discovered equally powerful pre-training data & learning objective for vision or robotics

Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text

Wanrong Zhu[♣*] Jack Hessel[♡*]
Anas Awadalla[♠] Samir Yitzhak Gadre[◇] Jesse Dodge[♡] Alex Fang[♠]
Youngjae Yu[†] Ludwig Schmidt[♠♡‡] William Yang Wang[♣] Yejin Choi[♠♡]

LAION-5B: An open large-scale dataset for training next generation image-text models

Christoph Schuhmann[1] §§°° Romain Beaumont[1] §§°° Richard Ven
Cade Gordon[2] §§°° Ross Wightman[1]§§ Mehdi Cherti [1,10]
Theo Coombes[1] Aarush Katta[1] Clayton Mullis[1] Mitchell Wo
Patrick Schramowski[1,4,5] Srivatsa Kundurthy[1] Katherine Crowso
Ludwig Schmidt[6] °° Robert Kaczmarczyk[1,7] °° Jenia Jitsev[1,10] °°

Chapter 3: The Paradox

- Commonsense paradox
- Moravec's paradox
- Generative AI paradox

Generative AI Paradox?

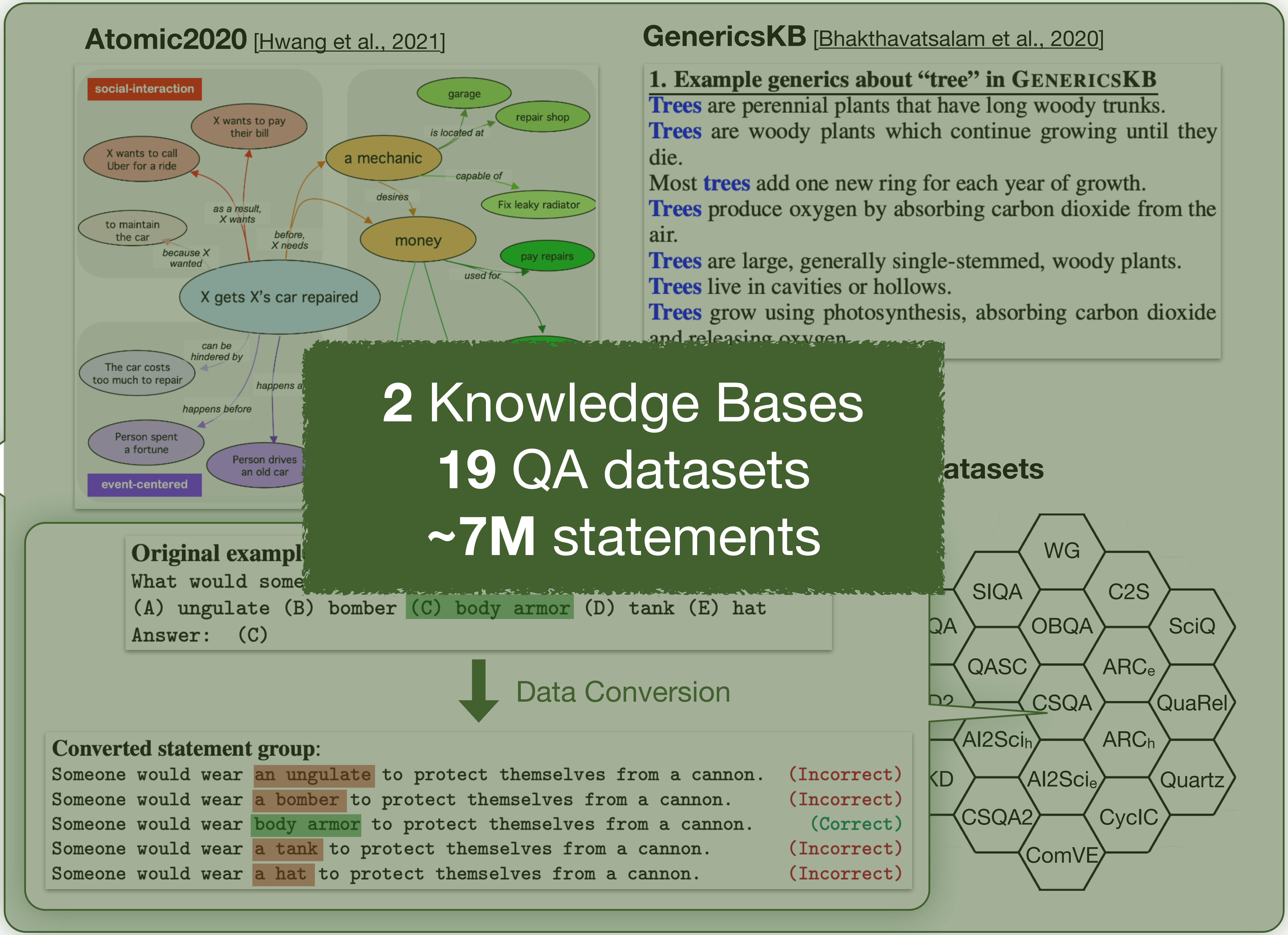
- Another case of easy is hard and hard is easy
- It appears to be that for (current) AI, generation is easier than understanding
- For humans, understanding is generally easier than generation

VERA: A General-Purpose Plausibility Estimation Model for Commonsense Statements

Jiacheng Liu[♡*] Wenya Wang[♡*] Dianzhuo Wang[◇]
Noah A. Smith[♡♠] Yejin Choi[♡♠] Hannaneh Hajishirzi[♡♠]

Plausibility: 15%

A bird has four legs.



2 Knowledge Bases

19 QA datasets

~7M statements

[Examples adapted from Lin et al., 2020]

Solving Commonsense Benchmarks

Predicting the most plausible statement out of the multiple-choice candidates

Name	Domain	Format
STAGE B TRAINING (SEEN)		
OpenBookQA	scientific	multiple-choice (4)
ARC (easy)	scientific	multiple-choice (4)
EVALUATION (UNSEEN TYPE 1)		
WSC		
COPA		
NumerSense		
PROST		
Spatial Commonsense		
EVALUATION (UNSEEN TYPE 2)		
SWAG		multiple-choice (4)
HellaSwag		
CODAH		
Story Cloze Test		
αNLI		
StrategyQA		
CREAK		boolean

5 unseen (type 1) benchmarks
Similar to seen benchmarks, but diagnostic datasets

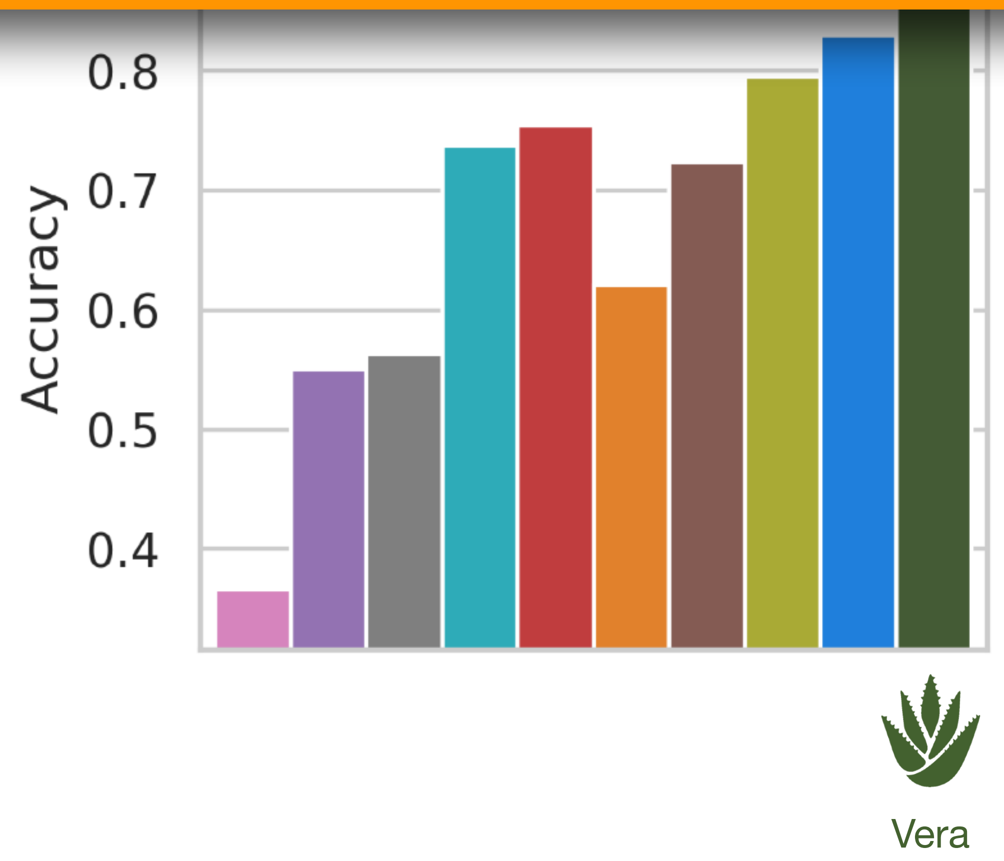
8 unseen (type 2) benchmarks
The tasks are a bit further from commonsense verification

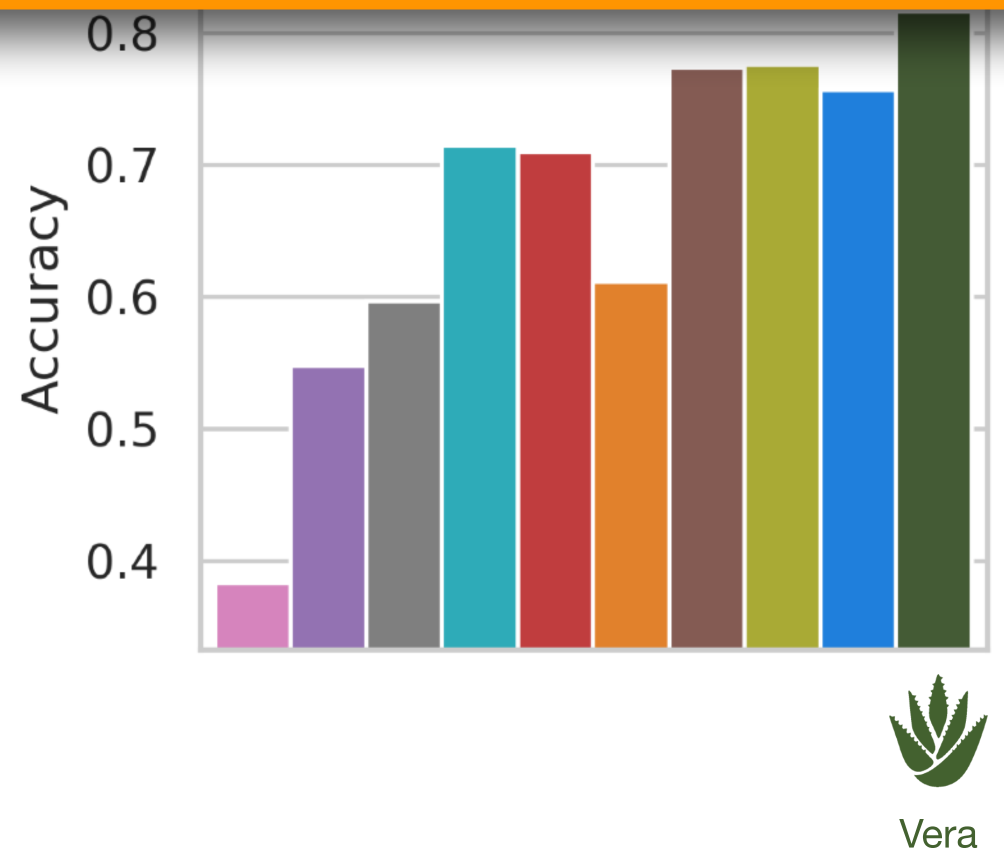
Converted statement group:

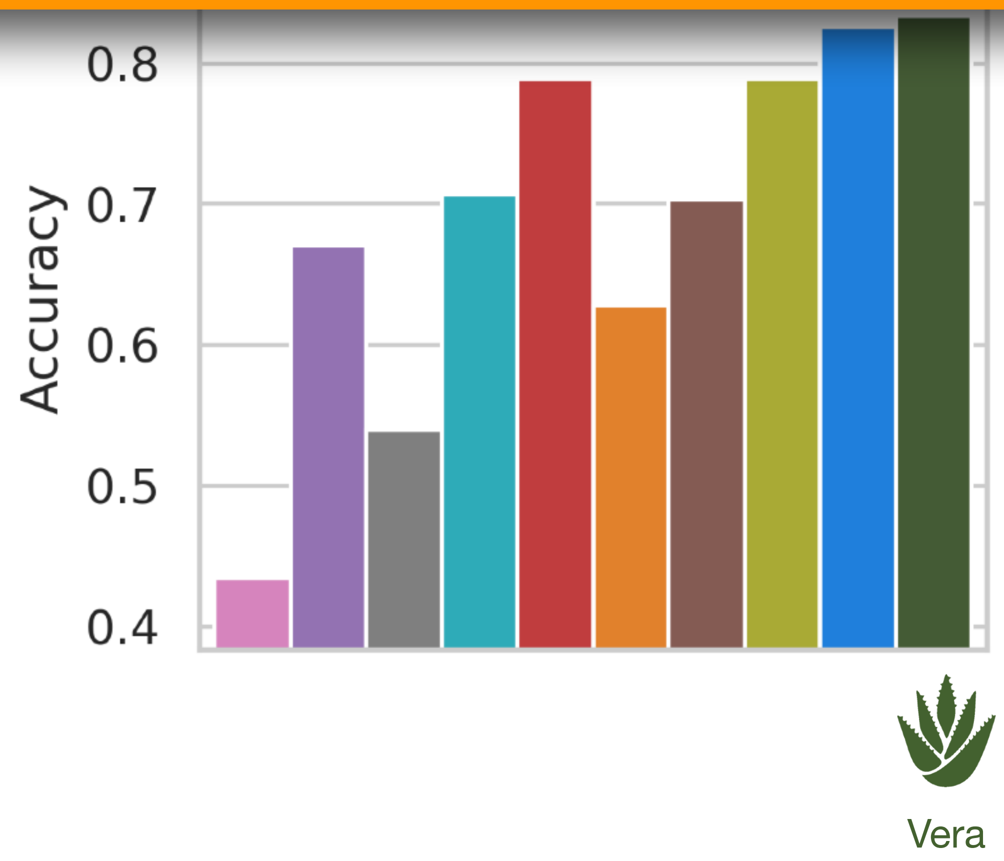
Statement	Label	Vera
Someone would wear an ungulate to protect themselves from a cannon.	(Incorrect)	3%
Someone would wear a bomber to protect themselves from a cannon.	(Incorrect)	6%
Someone would wear body armor to protect themselves from a cannon.	(Correct)	93%

Best baseline is Flan-T5. ChatGPT and GPT-4 are worse.

Vera outperforms Flan-T5 by 4%-6% on all eval sets (seen/unseen domains)







2050: An AI Odyssey

Prolog: what CVPR 2050 be like

Chapter 1: The Possible Impossibilities

Chapter 2: The Impossible Possibilities

Chapter 3: The Paradox

Epilog: why am I even here? A confession of an alien

Epilog: why am I even here? a confession of an alien

- Impossible possibilities — story of my life
- 10 years ago, it really didn't seem like I'd come this far
- I consider myself as a case of a late bloomer*
- I grew to believe that **talent is made, not born****

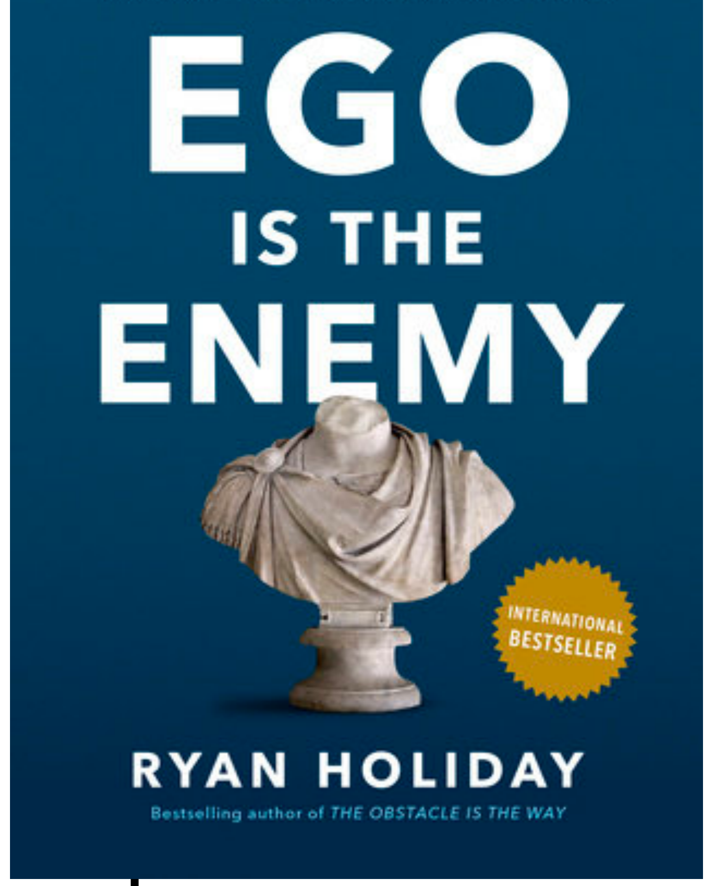
---

* Though even to this date, I feel like I am an imposter, just about to get caught (perhaps after this talk)

** Or to state more carefully, talent can be enhanced dramatically throughout one's life, with considerable efforts in a supportive and inclusive environment, even if one's starting point wan't all that remarkable. Of course some folks are born geniuses and all…

Circa 2012 ...

Epilog: why am I even here? a confession of an



"talent is made, not born"

- **Internal factor:** because **I didn't think much of myself**, I was (more) willing to do:
  1. Lifelong learning: **learning from everyone, especially from my students**, colleagues, and **continually questioning my previous beliefs and perspectives and revising them** along the way
  2. Taking risks (reason being, since I'm not that great, I shouldn't work on problems that other smarter people will work on. What a waste of tax money, which supports my university salary. Also, since I'm not that great, who cares if I fail… nobody will notice?)

  —> And it turns out, 10 year is a long time (to learn about a lot of stuff), and it's actually **pretty impossible to *only* fail — eventually some things will work out**
- **External factor:** I was lucky enough to be in an *inclusive* environment

Epilog: why am I even here? a confession of an alien

- As I grew to believe that **talent is made, not born**, …
- I also grew to believe that **the power of diversity and inclusion is real**
  - The culture that understands DEI is **less authoritative** and **more open-minded**, which in turn helped me to grow **confidence** to try something new and different
  - You just learn so much more when interacting with diverse folks, as they broaden your view points and foster more **divergent** and **innovative** thinking
  - Giving an opportunity to them can make all the difference!


Thanks!