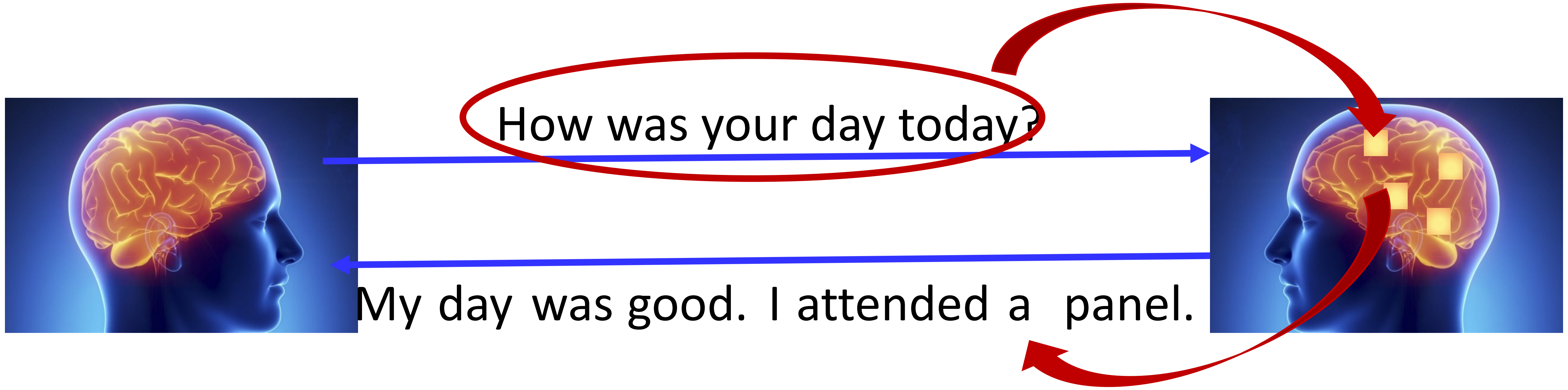
ChatGPT and other Large Language Models

Mausam

Professor, Computer Science & Engineering
Founding (ex-)Head, Yardi School of Artificial Intelligence
IIT Delhi

Joint work with Indrajit Bhattacharya, NM Anoop Krishnan, Parag Singla,
Kartik Krishna, Aman Madaan, Mayur Patidar, Vipul Rathore, Vishal Saley, Mohd Zaki,
Pranjal Aggarwal, Daman Arora, Aniruddha Deb, Chinmay Mittal, Himanshu Singh, ++

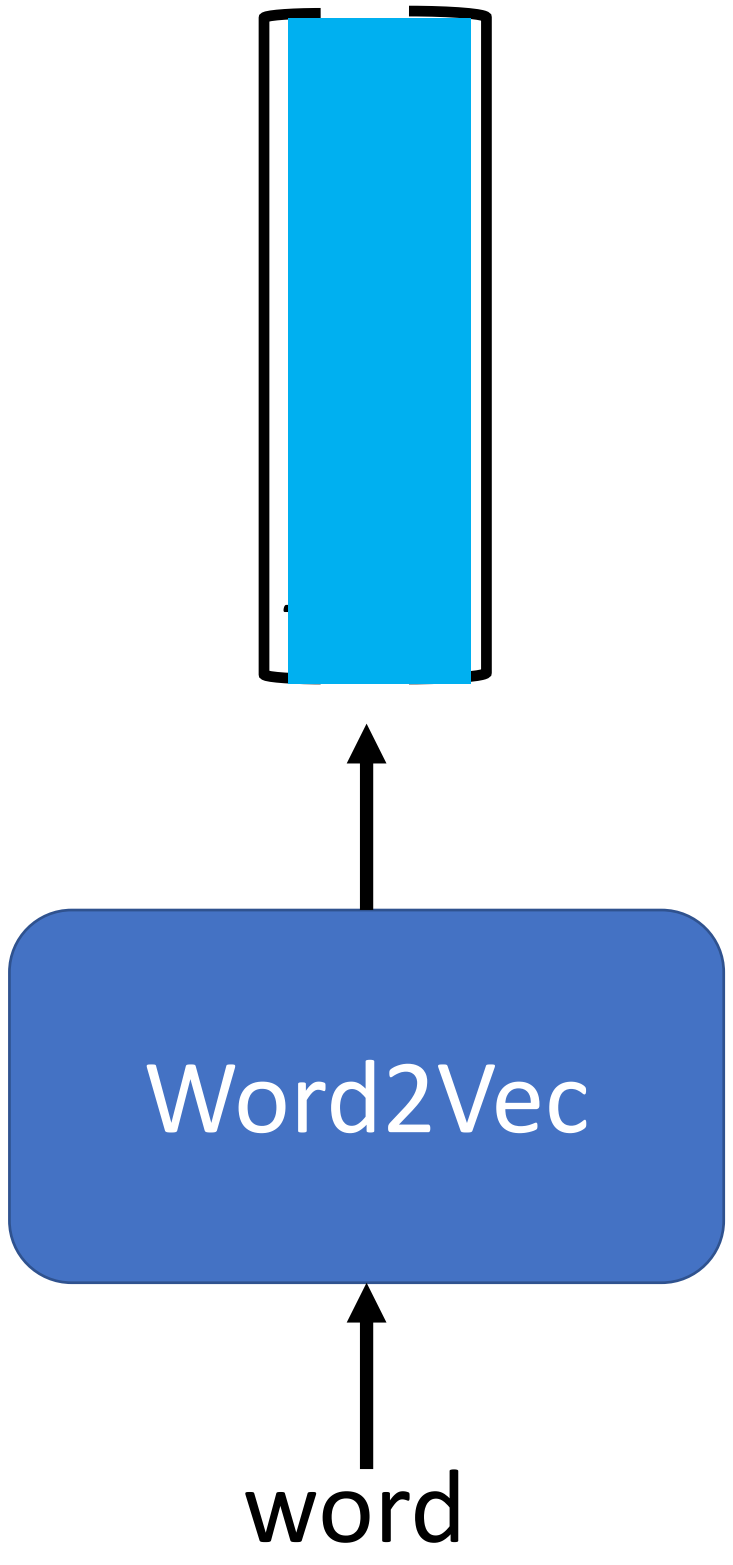
Humans and Language


How was your day today?

My day was good. I attended a panel.

- Encoding: Input words are processed by the brain
  - Words are discrete; brain processing uses signals (continuous)
- Reasoning: Brain performs internal reasoning to decide a response
- Decoding: Brain verbalizes the response one word at a time

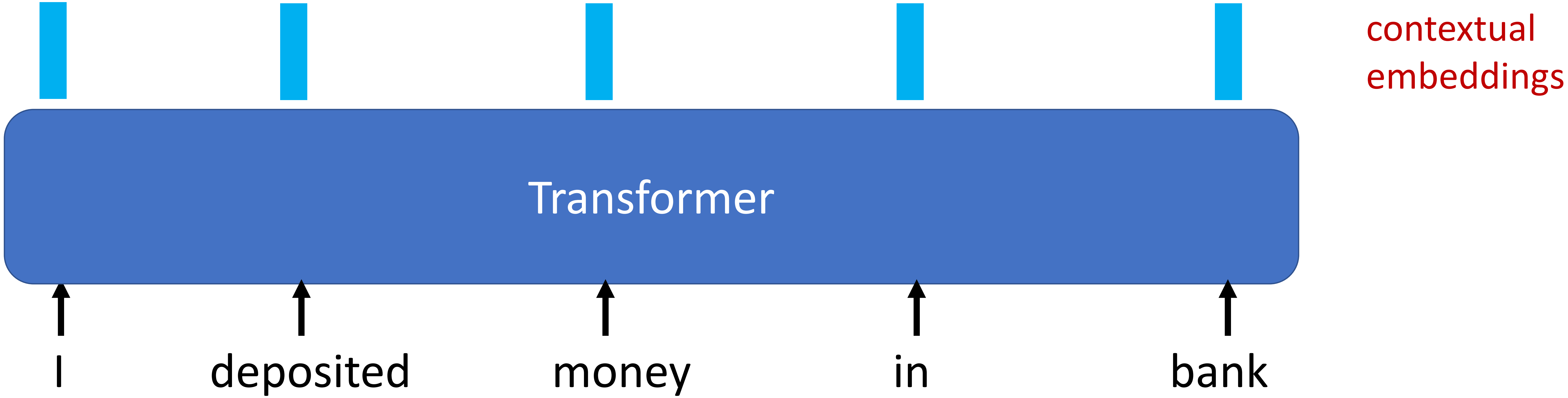
Idea 1 [2013]: A Word is a (continuous) Vector



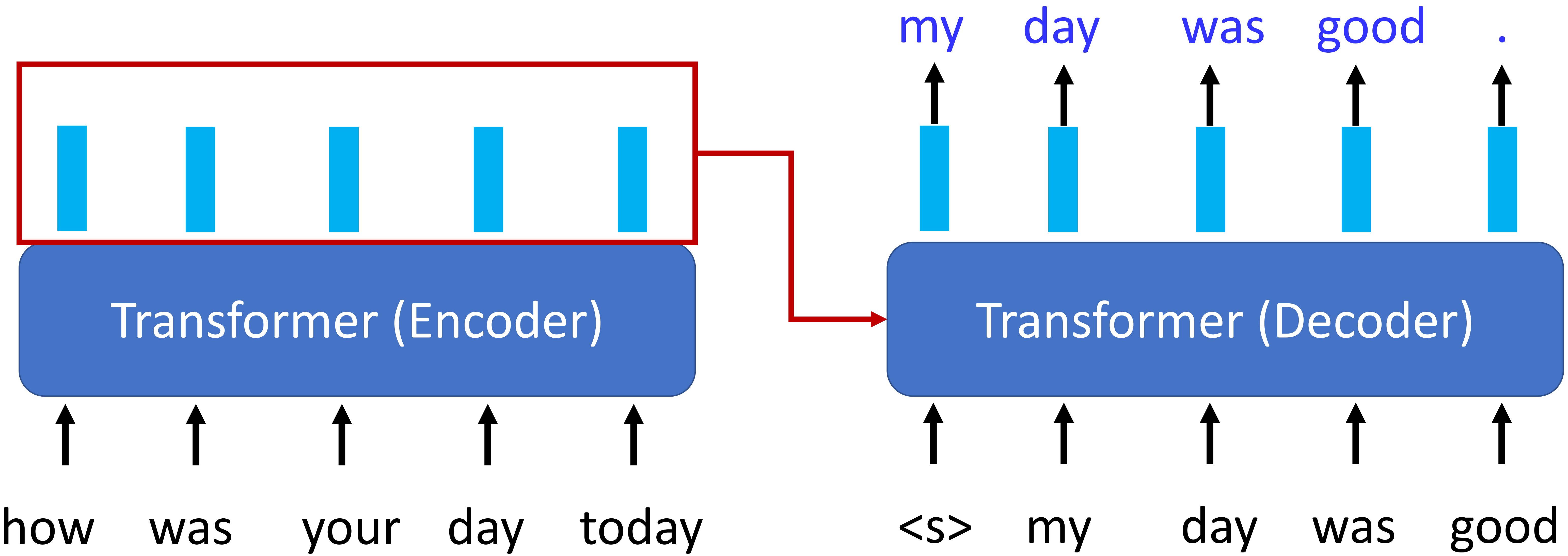
Discrete objects (words) need to be embedded for neural models

Idea 2 [2017]: A Word can be Many (continuous) Vectors

- One embedding not enough for words that have multiple meanings
  - Bank – financial institution or river bank
- Transformers: a novel neural architecture to generate context-based word embeddings



Idea 3 [2014]: Generative NLP

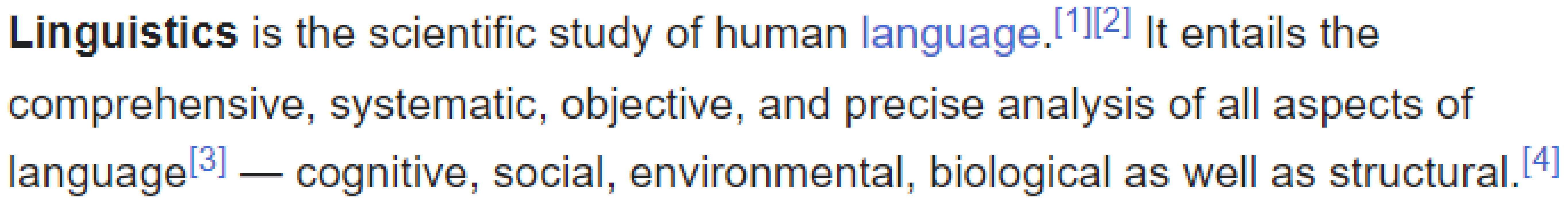


Challenge: How to train? -- using annotated training data

limited availability of (input-output) pairs ☹

Idea 4 [2018]: Self-supervised Learning

- Convert general text on the Web into huge number of (input-output) pairs

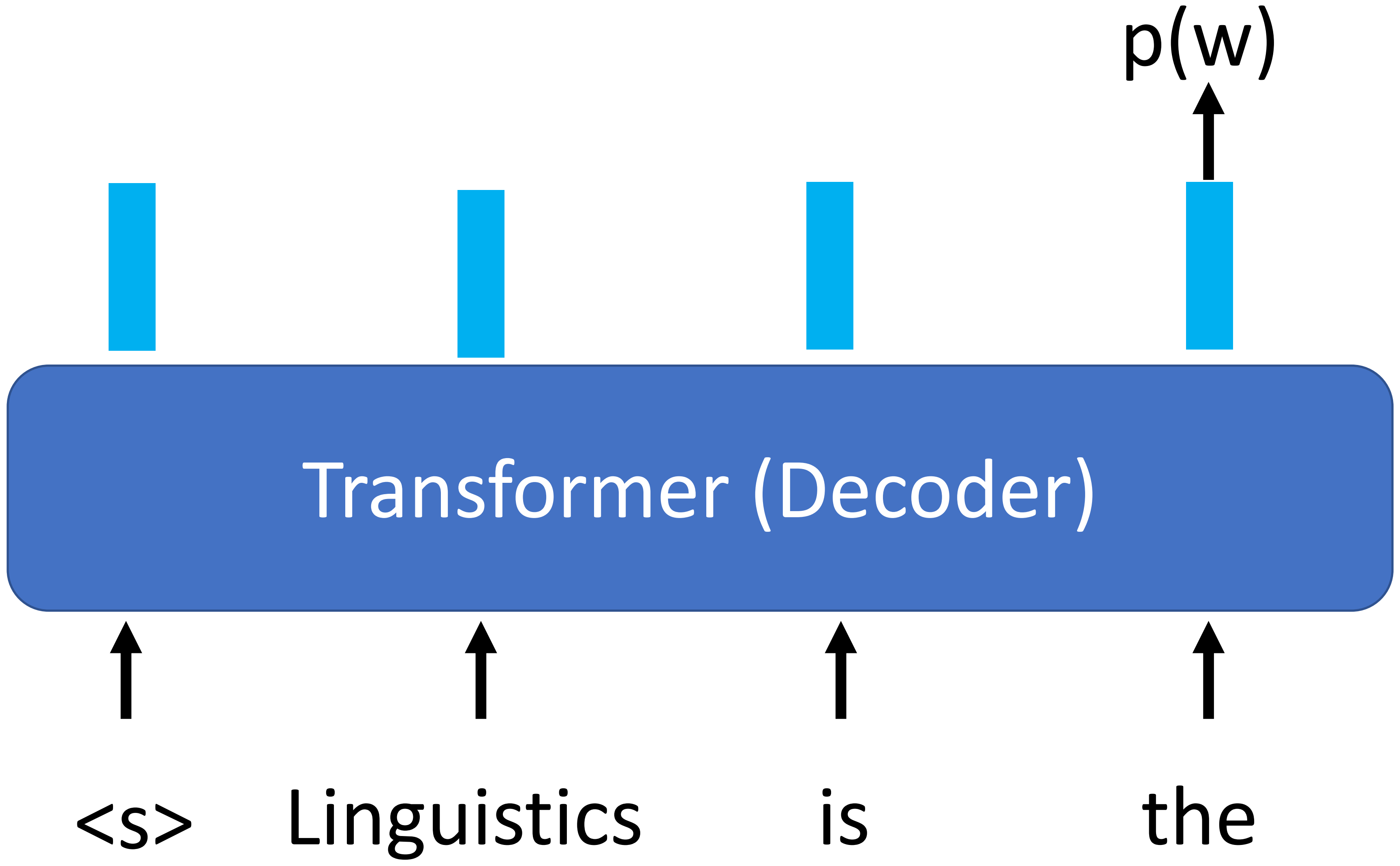


Linguistics is the scientific study of human language.[1][2] It entails the comprehensive, systematic, objective, and precise analysis of all aspects of language[3] — cognitive, social, environmental, biological as well as structural.[4]

- Linguistics is the _________
- Linguistics is the scientific ________
- Linguistics is the scientific study _________
- Linguistics is the scientific study of __________
- Linguistics is the scientific study of human __________

- scientific
- study
- of
- human
- language

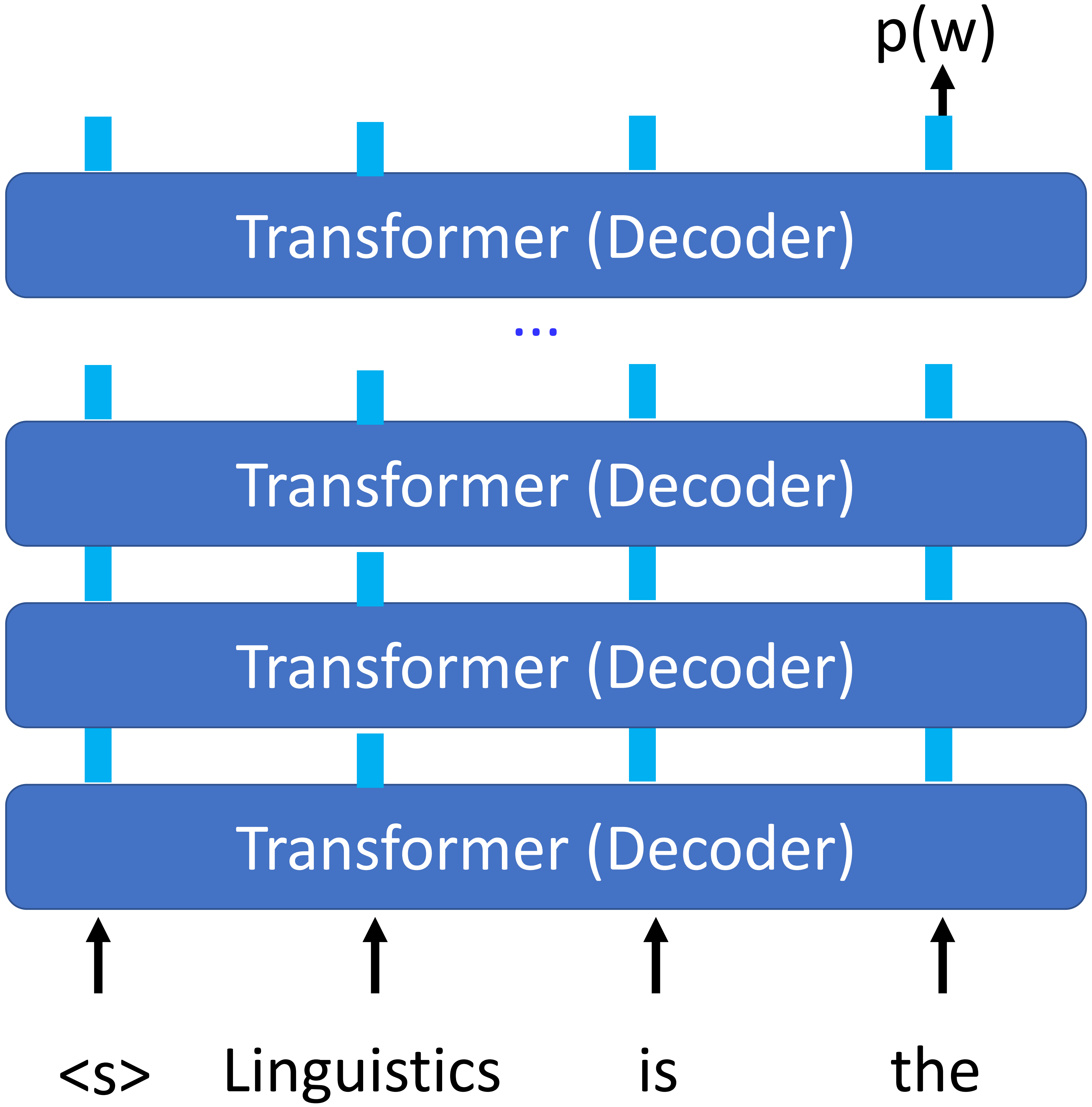
Idea 4 [2018]: Self-supervised Learning



Train parameters
such that
p(scientific) is high

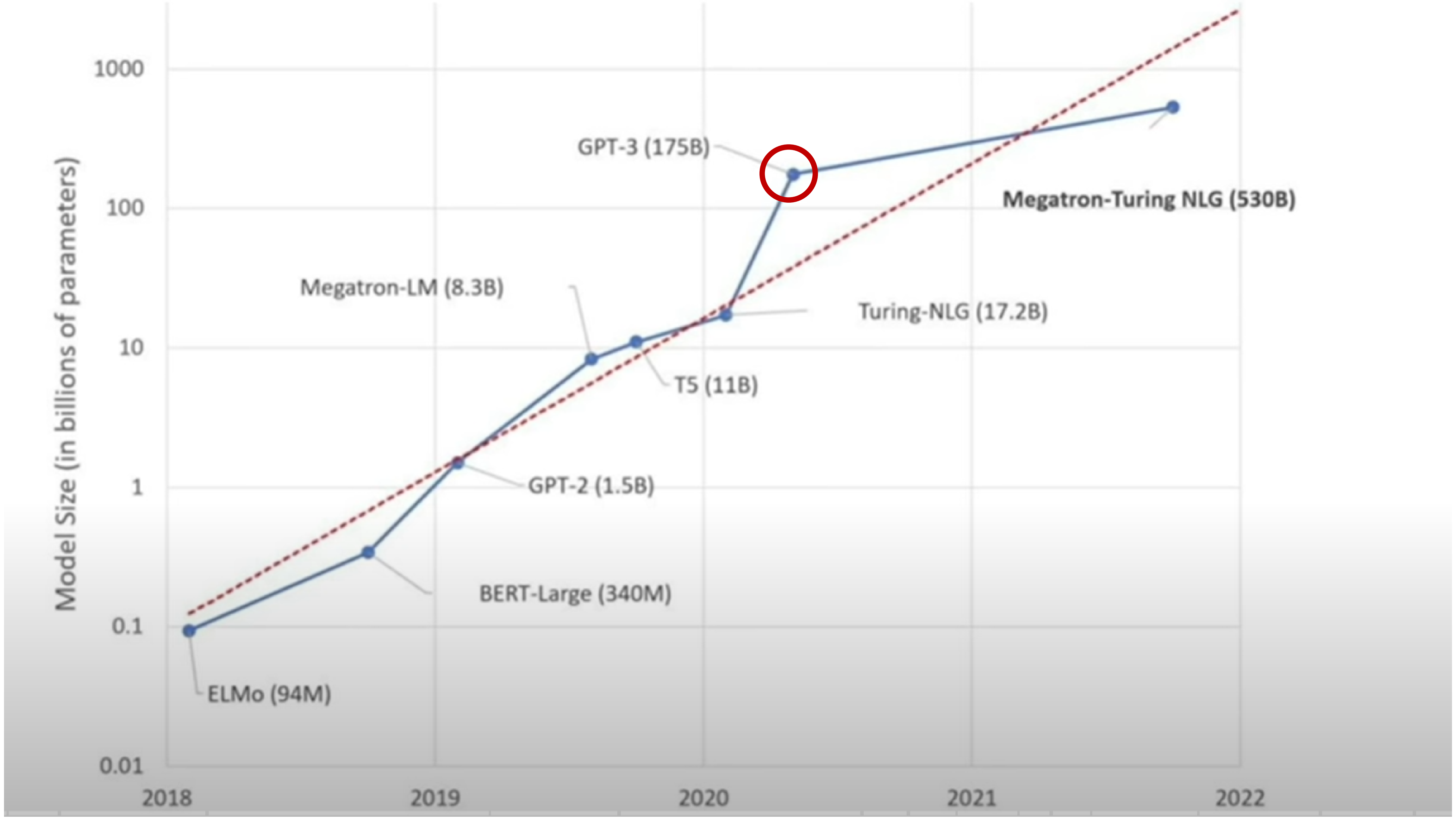
Pre-trained
Language Models

Idea 5 [2019++]: Scaling Up! Larger Datasets; Larger Models



Train parameters such that p(scientific) is high

Progression of #Parameters in PTLMs



$4.6 million
for one model

Figure by Peter Lee

[2020] GPT3

- caught people’s imagination worldwide
  - give it prompt (instruction): it gave surprisingly credible response

- … but had many many issues
  - made loads of mistakes
  - impolite
  - racist/sexist;
  - perpetuate stereotypes; misinformation
  - etc.

Idea 6: Massive Manual Annotation
[2022] GPT3 → GPT3.5 (backend of ChatGPT)

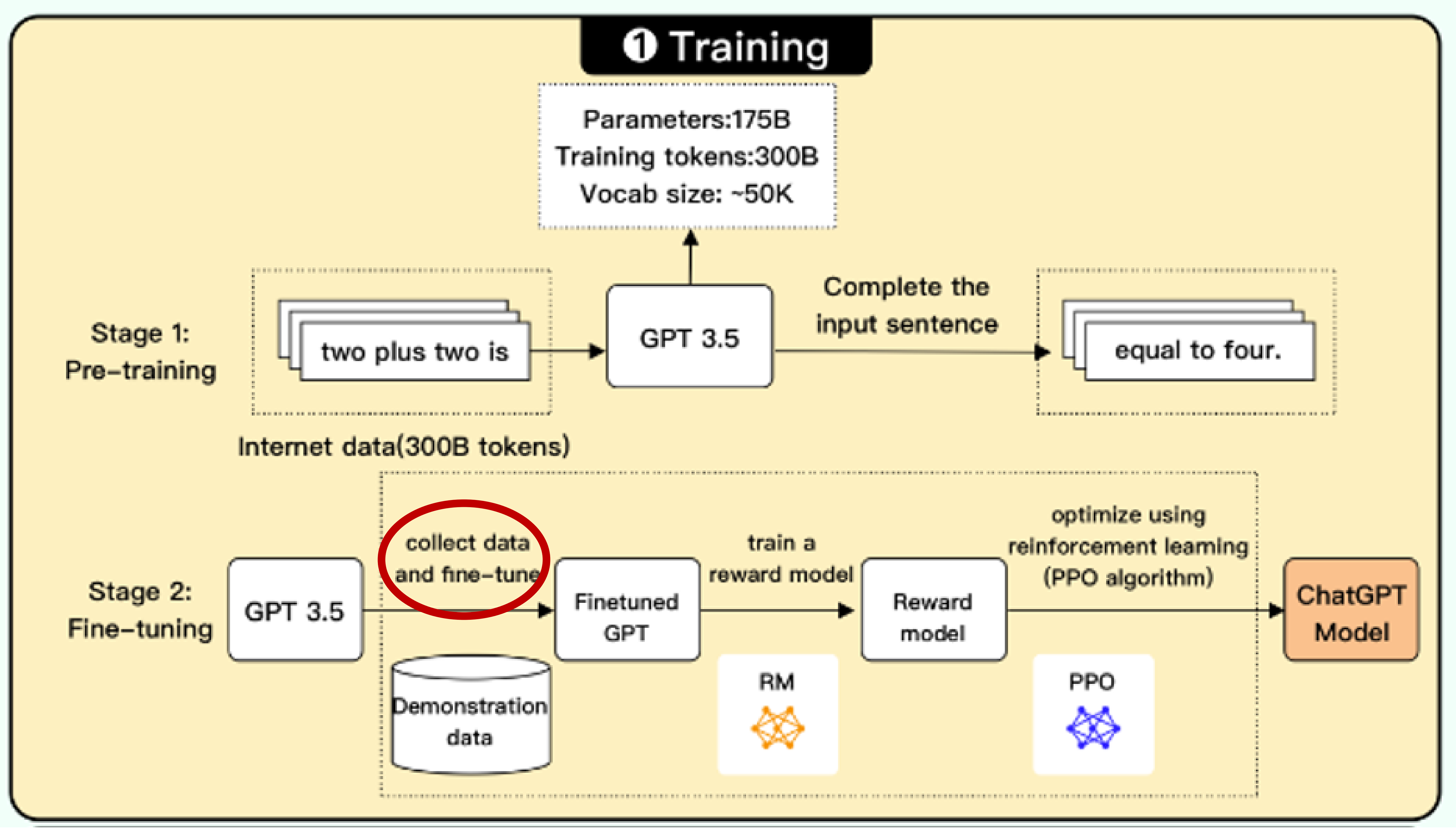
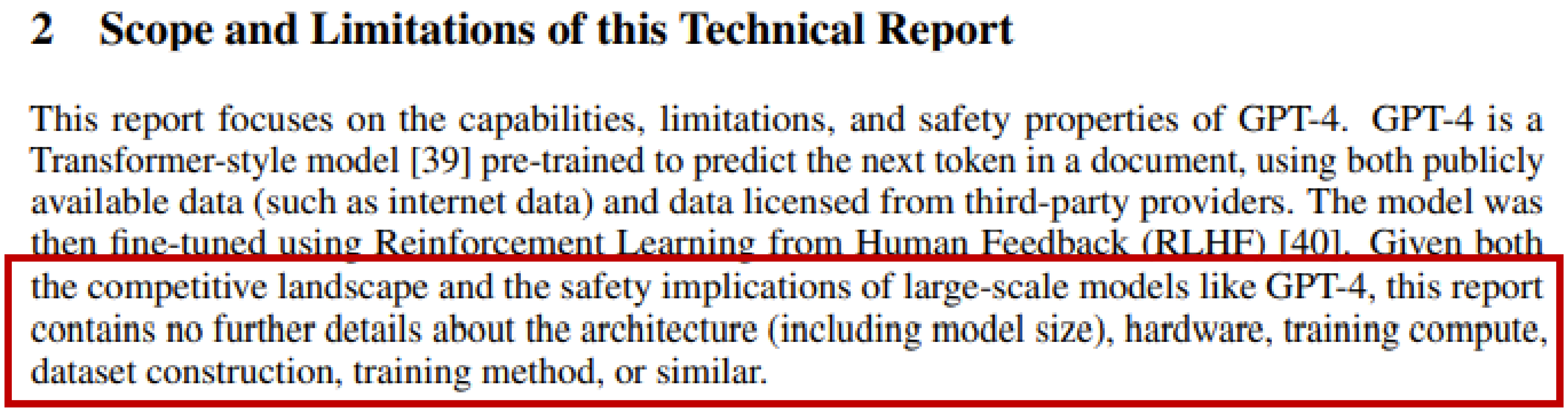


Figure from https://blog.bytebytego.com/p/ep-44-how-does-chatgpt-work

[2023] GPT4

GPT-4 is reportedly about six times larger than GPT-3, with one trillion parameters, according to a report by Semafor, which has previously leaked GPT-4 in Bing.



2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

GPT4 >> GPT3.5

ChatGPT (Pro) >> ChatGPT

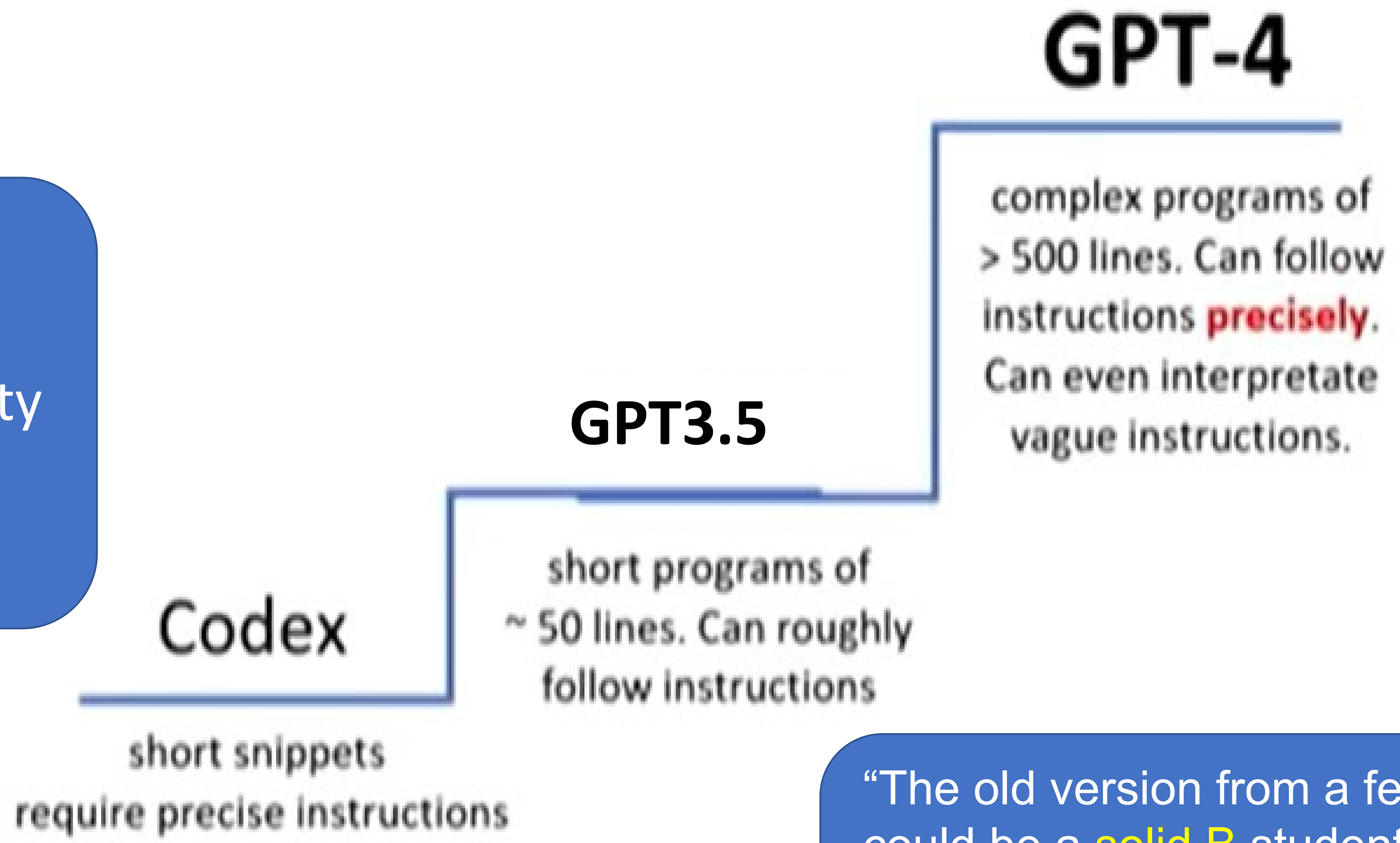
ChatGPT Pro (with GPT4 backend)

- Skills are incredible and unbelievable

- Language skills
  - Summarizes text
  - Writes emails
  - Writes poems…
- Showcases expertise in many domains: legal, medical, physics, …
- Has demonstrated ability to understand complex ideas
- Has demonstrated ability to perform complex reasoning
- Can write complicated computer programs
- …

Codex used for
- boilerplate code
- error explainability
- autofixing errors

20% efficiency



GPT-4

complex programs of > 500 lines. Can follow instructions **precisely**. Can even interpretate vague instructions.

GPT3.5

short programs of ~ 50 lines. Can roughly follow instructions

Codex

short snippets require precise instructions

EDITION IN FARIDABAD 31°C

THE TIMES OF INDIA

NEWS / GADGETS NEWS NEWS / ChatGPT Creates CV For A User, Here's What Happened Next

TRENDING TOPICS What is Auto-GPT

ChatGPT creates CV for a user, here's what happened next

Google has 'banned' these apps from Play Store

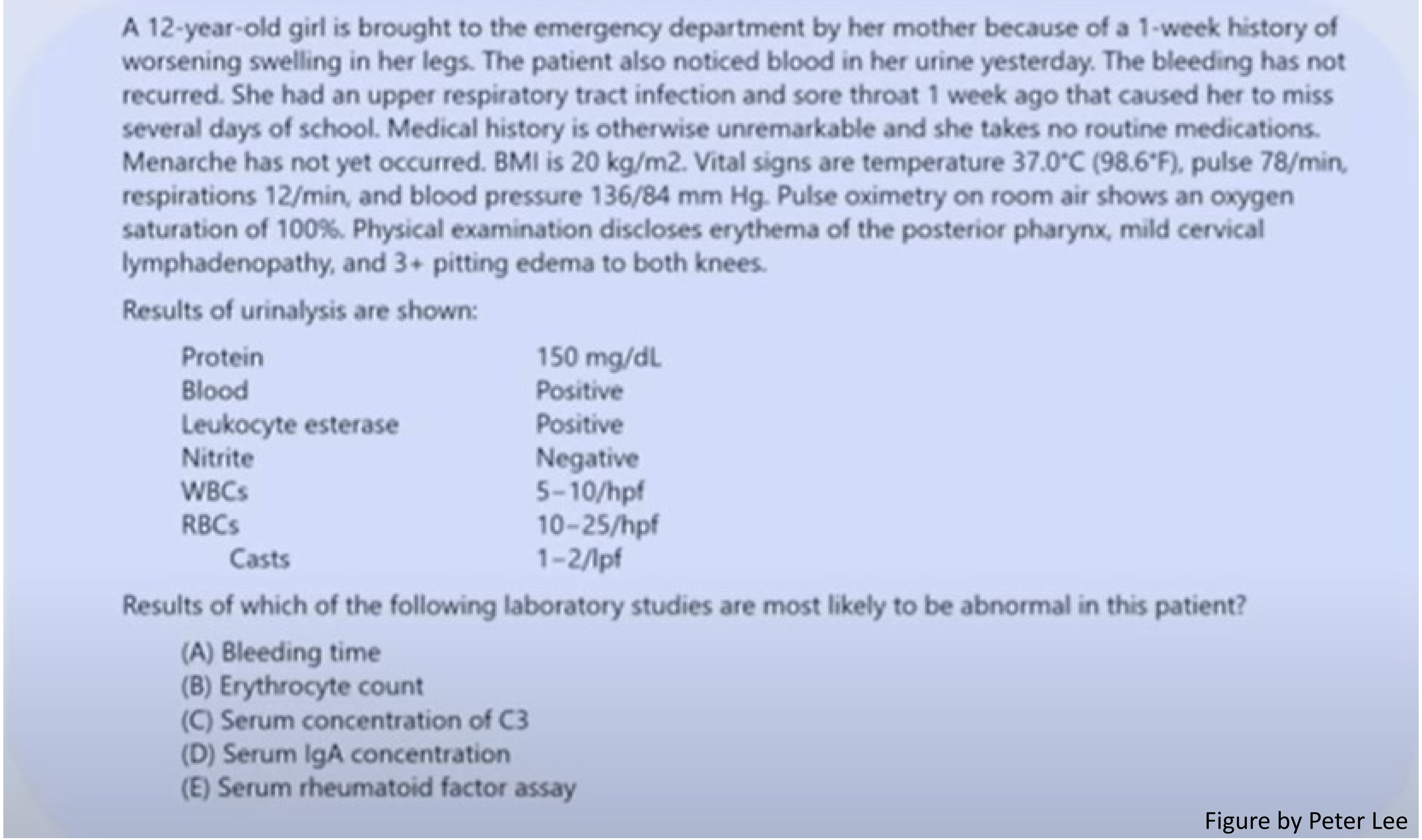
15 most iconic mobile phones of the past

13 tough interview questions heard at Apple, Google, Microsoft and...

5 tips to p smartpho

> "The old version from a few months ago could be a solid B student," said Salman Khan, founder of Khan Academy, ...
>
> "This one can be an A student in a pretty rigorous program."

Figure by Sebastien Brubeck



A 12-year-old girl is brought to the emergency department by her mother because of a 1-week history of worsening swelling in her legs. The patient also noticed blood in her urine yesterday. The bleeding has not recurred. She had an upper respiratory tract infection and sore throat 1 week ago that caused her to miss several days of school. Medical history is otherwise unremarkable and she takes no routine medications. Menarche has not yet occurred. BMI is 20 kg/m2. Vital signs are temperature 37.0°C (98.6°F), pulse 78/min, respirations 12/min, and blood pressure 136/84 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 100%. Physical examination discloses erythema of the posterior pharynx, mild cervical lymphadenopathy, and 3+ pitting edema to both knees.

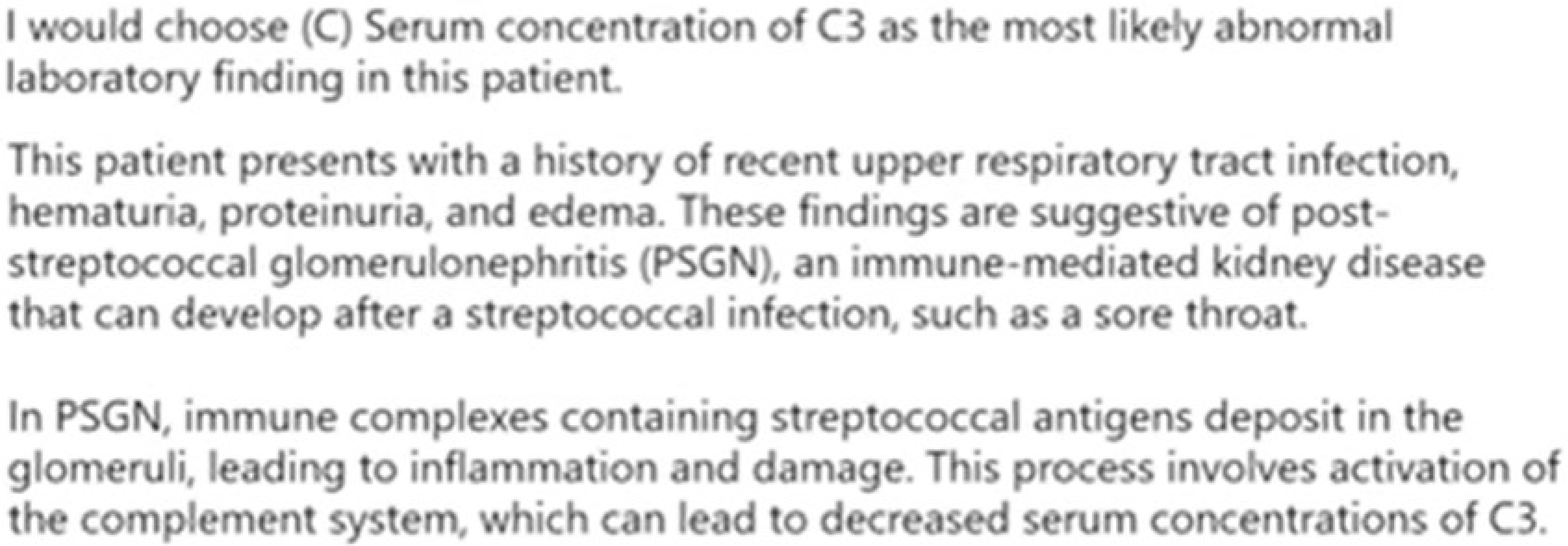
Results of urinalysis are shown:

Protein	150 mg/dL
Blood	Positive
Leukocyte esterase	Positive
Nitrite	Negative
WBCs	5–10/hpf
RBCs	10–25/hpf
Casts	1–2/lpf

Results of which of the following laboratory studies are most likely to be abnormal in this patient?

(A) Bleeding time
(B) Erythrocyte count
(C) Serum concentration of C3
(D) Serum IgA concentration
(E) Serum rheumatoid factor assay

Figure by Peter Lee



I would choose (C) Serum concentration of C3 as the most likely abnormal laboratory finding in this patient.

This patient presents with a history of recent upper respiratory tract infection, hematuria, proteinuria, and edema. These findings are suggestive of post-streptococcal glomerulonephritis (PSGN), an immune-mediated kidney disease that can develop after a streptococcal infection, such as a sore throat.

In PSGN, immune complexes containing streptococcal antigens deposit in the glomeruli, leading to inflammation and damage. This process involves activation of the complement system, which can lead to decreased serum concentrations of C3.

Figure by Peter Lee

Three (Extreme) Reactions

Denial: it is just statistics on steroids – not real intelligence!

Dystopia: the world as we know it is going to end. What will happen to jobs?

Euphoria: I can just sit back and let ChatGPT do everything for me

Realism: ***understand*** the technology and work with it, ***where appropriate***

My Research Questions

- Is GPT4 great at all kinds of AI tasks?
  - Language – Knowledge – Reasoning

- Is GPT4 great in all ML settings?
  - no supervision – some supervision – lots of supervision

- What about Cost? Latency?

- Are traditional AI models of no use now?

My Experiments with GPT4

Mausam

Professor, Computer Science & Engineering
Founding (ex-)Head, Yardi School of Artificial Intelligence
IIT Delhi

Joint work with Indrajit Bhattacharya, NM Anoop Krishnan, Parag Singla,
Kartik Krishna, Aman Madaan, Mayur Patidar, Vipul Rathore, Vishal Saley, Mohd Zaki,
Pranjal Aggarwal, Daman Arora, Aniruddha Deb, Chinmay Mittal, Himanshu Singh, ++

My Experiments

- GPT4 and Language Skills
- No Data: GPT4 and Reasoning Skills
  - How well does GPT4 perform in complex reasoning tasks (QA)?
  - How well does GPT4 perform in NP-complete puzzles?
- No Data: GPT4 and Multilingual Skills
  - Can GPT4 help with *unseen* languages?
- Data: GPT4 and NLP Tasks
  - How well does GPT4 compare against SoTA *supervised* models?
- GPT4 too Expensive! Use GPT4 only when necessary
- No Success so far

My Experiments

- GPT4 and Language Skills – incredible!

- No Data: GPT4 and Reasoning Skills
  - How well does GPT4 perform in complex reasoning tasks (QA)?
  - How well does GPT4 perform in NP-complete puzzles?

- No Data: GPT4 and Multilingual Skills
  - Can GPT4 help with *unseen* languages?

- Data: GPT4 and NLP Tasks
  - How well does GPT4 compare against SoTA *supervised* models?

- GPT4 too Expensive! Use GPT4 only when necessary

- No Success so far

Poem on IIT Delhi

GPT-4 Response:
In the heart of India's capital, a beacon stands tall,
A cradle of knowledge, where wisdom's sweet call,
Echoes through the halls of IIT Delhi, a realm of dreams,
Where the curious and ambitious weave their grand schemes.

A campus adorned with lush greenery and vibrant hues,
A melting pot of cultures, where diverse minds fuse,
The hallowed corridors whisper tales of yore,
Of legends who walked these paths, and the laurels they bore.

From the iconic Dogra Hall to the majestic Jwalamukhi,
Each edifice a testament to the spirit that runs high,
The Central Library, a treasure trove of knowledge profound,
Where the secrets of the universe patiently wait to be found.

The laboratories, where innovation takes flight,
And the classrooms, where ideas spark and ignite,
The hostels, a home away from home, where friendships are made,
And the bonds forged in the crucible of time never fade.

The annual rendezvous, Rendezvous, a celebration of art,
Where music, dance, and laughter play their part,
And Tryst, the techno-management fest, a showcase of skill,
Where the brightest minds gather, their dreams to fulfill.

From the fields of engineering to the realms of research,
The students of IIT Delhi, on a quest for knowledge, they search,
Their tireless pursuit of excellence, a beacon that guides,
As they strive to conquer challenges and scale new heights.

For those who walk the hallowed halls of IIT Delhi, know this,
The journey may be arduous, but the reward is pure bliss,
For in this temple of learning, where dreams take flight,
The future is forged, and the world is set alight.

My Experiments

- GPT4 and Language Skills – incredible!

- No Data: GPT4 and Reasoning Skills
  - How well does GPT4 perform in complex reasoning tasks (QA)?
  - How well does GPT4 perform in NP-complete puzzles?

- No Data: GPT4 and Multilingual Skills
  - Can GPT4 help with *unseen* languages?

- Data: GPT4 and NLP Tasks
  - How well does GPT4 compare against SoTA *supervised* models?

- GPT4 too Expensive! Use GPT4 only when necessary

- No Success so far

JEEBench

[Arora et al EMNLP'23]

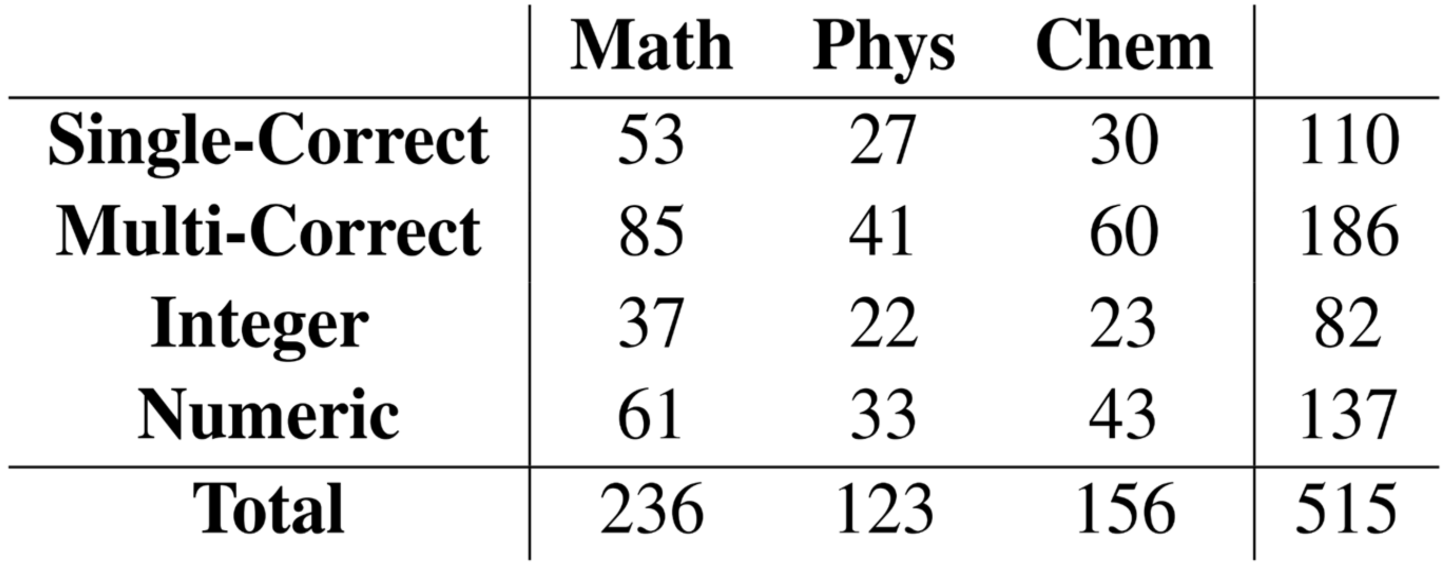
- The JEE Advanced exam is an interesting benchmark to study problem solving capabilities
  - Tough and lengthy questions.
  - Extremely competitive → quality of questions is high!

    repetition is low(?)

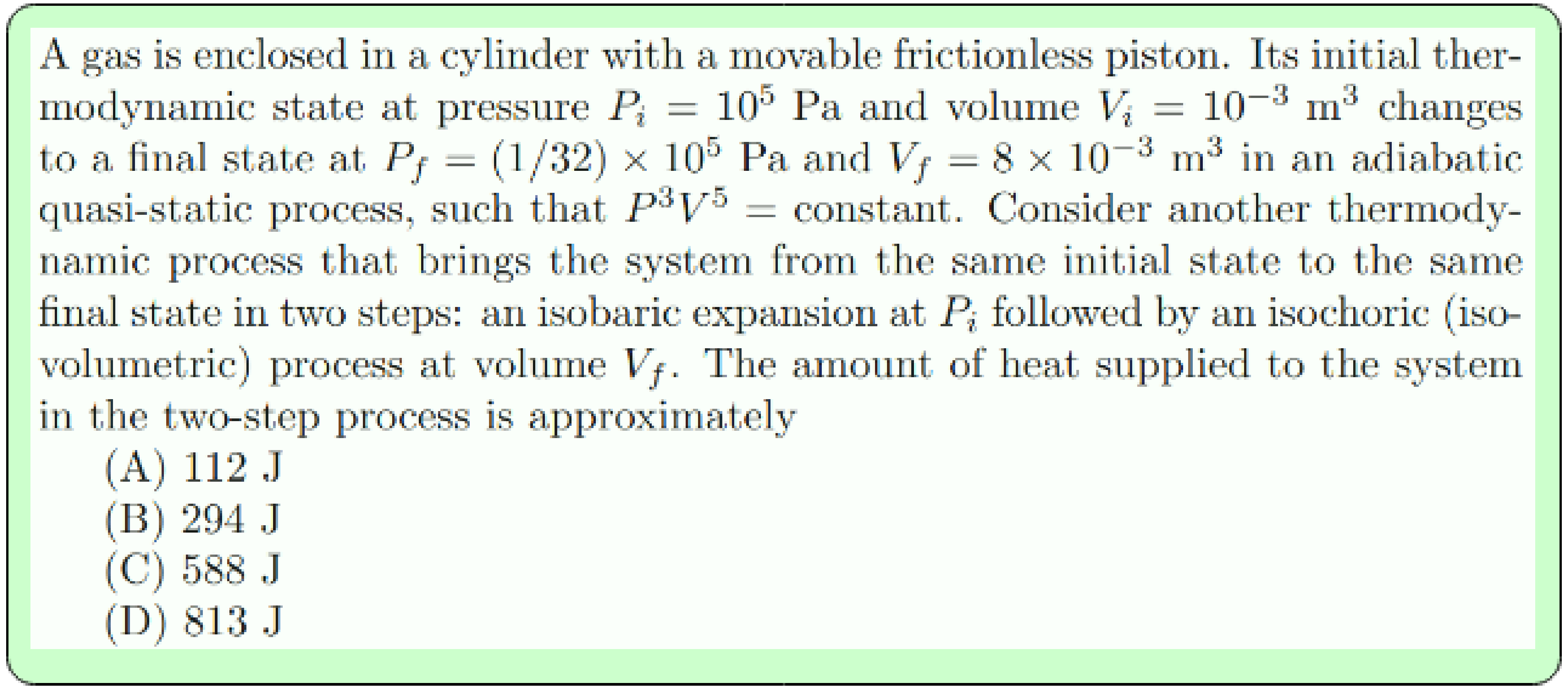
    more reasoning, less recall

GPT4 for problem solving

- Our dataset: JEEBench
  - 515 text-only questions from 2016 to 2023.
  - Subjects: Phy, Chem, Maths
  - Types:

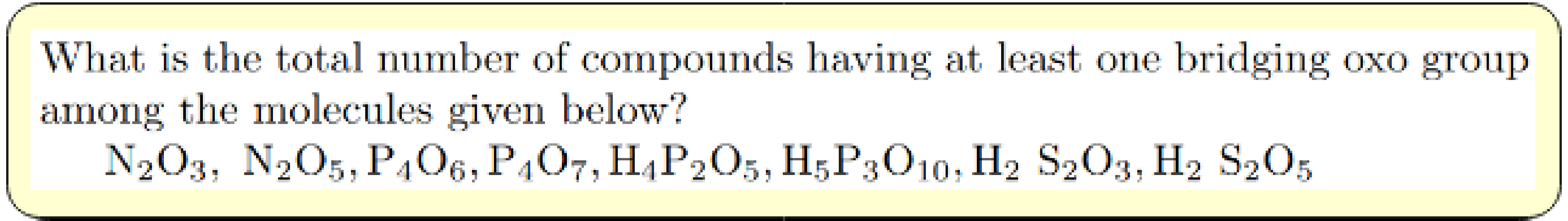


	Math	Phys	Chem	
Single-Correct	53	27	30	110
Multi-Correct	85	41	60	186
Integer	37	22	23	82
Numeric	61	33	43	137
Total	236	123	156	515



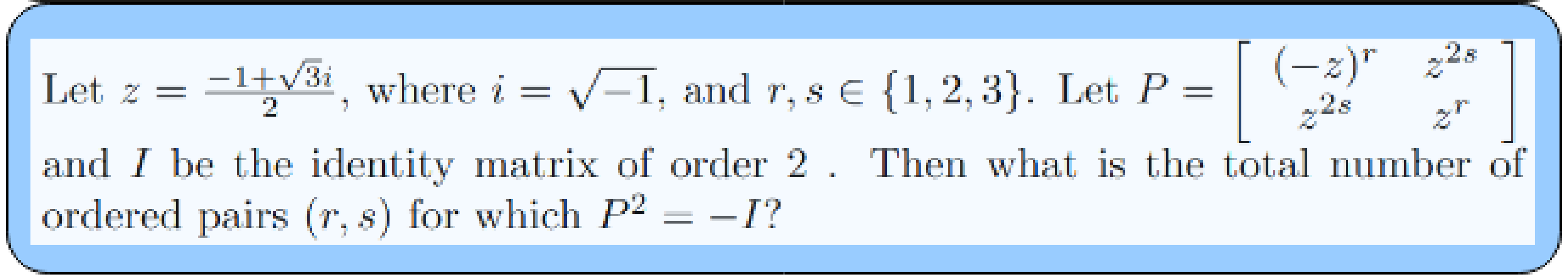
A gas is enclosed in a cylinder with a movable frictionless piston. Its initial thermodynamic state at pressure $P_i = 10^5$ Pa and volume $V_i = 10^{-3}$ m^3 changes to a final state at $P_f = (1/32) \times 10^5$ Pa and $V_f = 8 \times 10^{-3}$ m^3 in an adiabatic quasi-static process, such that $P^3V^5 =$ constant. Consider another thermodynamic process that brings the system from the same initial state to the same final state in two steps: an isobaric expansion at P_i followed by an isochoric (isovolumetric) process at volume V_f. The amount of heat supplied to the system in the two-step process is approximately

(A) 112 J
(B) 294 J
(C) 588 J
(D) 813 J



What is the total number of compounds having at least one bridging oxo group among the molecules given below?

N_2O_3, N_2O_5, P_4O_6, P_4O_7, $H_4P_2O_5$, $H_5P_3O_{10}$, $H_2S_2O_3$, $H_2S_2O_5$

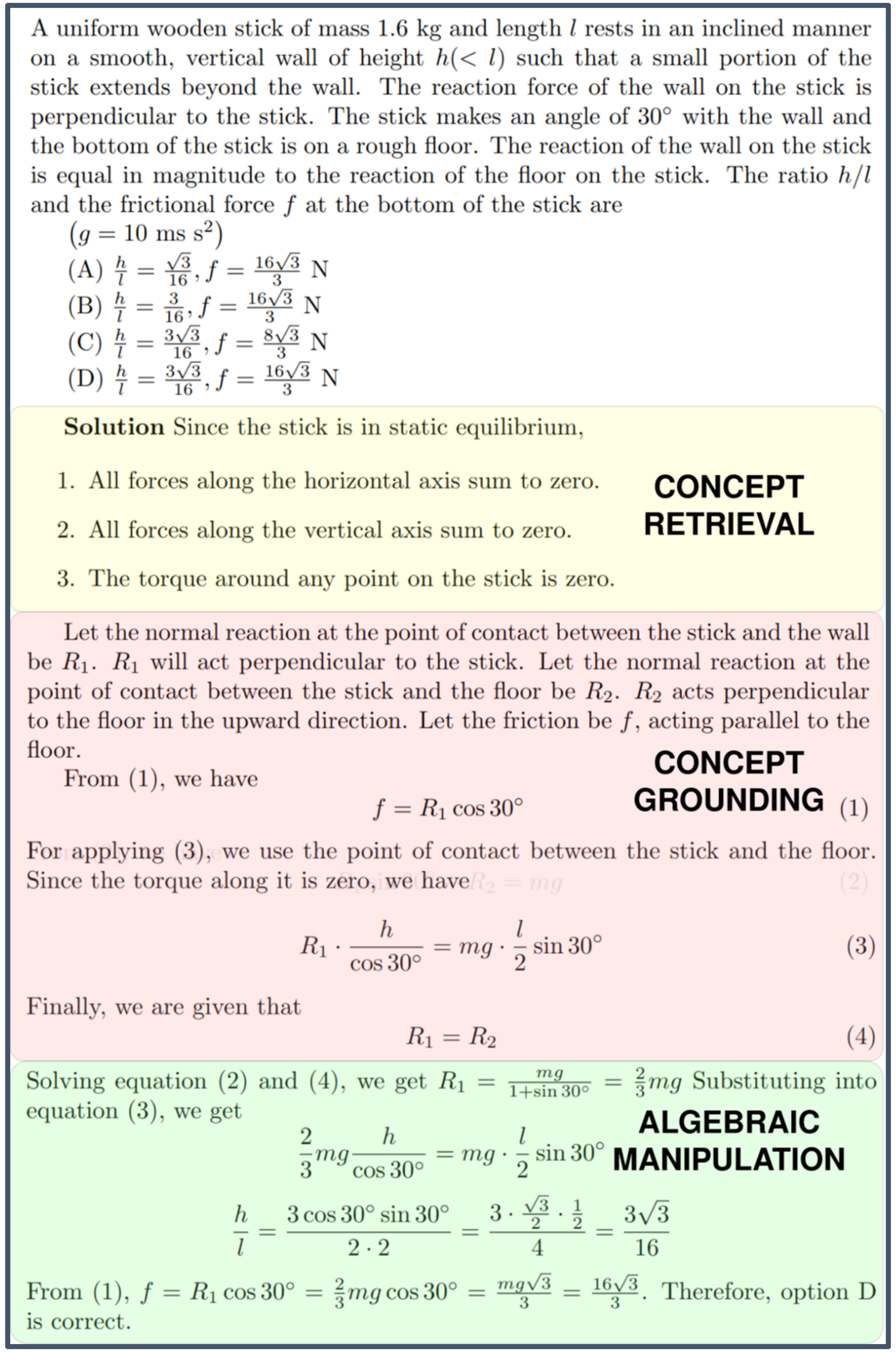


Let $z = \frac{-1+\sqrt{3}i}{2}$, where $i = \sqrt{-1}$, and $r, s \in \{1, 2, 3\}$. Let $P = \begin{bmatrix} (-z)^r & z^{2s} \\ z^{2s} & z^r \end{bmatrix}$ and I be the identity matrix of order 2 . Then what is the total number of ordered pairs (r, s) for which $P^2 = -I$?

JEEBench

- Solving problems requires
  - Concept Retrieval
    - retrieving relevant concepts

      (*Kirchhoff's laws, integration by parts, Newton's laws*)
  - Concept Grounding
    - grounding concepts into equations.
  - Symbolic Manipulation
    - solving equations to get the final answer.



A uniform wooden stick of mass 1.6 kg and length l rests in an inclined manner on a smooth, vertical wall of height $h(< l)$ such that a small portion of the stick extends beyond the wall. The reaction force of the wall on the stick is perpendicular to the stick. The stick makes an angle of 30° with the wall and the bottom of the stick is on a rough floor. The reaction of the wall on the stick is equal in magnitude to the reaction of the floor on the stick. The ratio h/l and the frictional force f at the bottom of the stick are

$(g = 10 \text{ ms s}^2)$

(A) $\frac{h}{l} = \frac{\sqrt{3}}{16}, f = \frac{16\sqrt{3}}{3}$ N

(B) $\frac{h}{l} = \frac{3}{16}, f = \frac{16\sqrt{3}}{3}$ N

(C) $\frac{h}{l} = \frac{3\sqrt{3}}{16}, f = \frac{8\sqrt{3}}{3}$ N

(D) $\frac{h}{l} = \frac{3\sqrt{3}}{16}, f = \frac{16\sqrt{3}}{3}$ N

Solution Since the stick is in static equilibrium,

1. All forces along the horizontal axis sum to zero.
2. All forces along the vertical axis sum to zero.
3. The torque around any point on the stick is zero.

CONCEPT RETRIEVAL

Let the normal reaction at the point of contact between the stick and the wall be R_1. R_1 will act perpendicular to the stick. Let the normal reaction at the point of contact between the stick and the floor be R_2. R_2 acts perpendicular to the floor in the upward direction. Let the friction be f, acting parallel to the floor.

From (1), we have

$$f = R_1 \cos 30° \tag{1}$$

CONCEPT GROUNDING

For applying (3), we use the point of contact between the stick and the floor. Since the torque along it is zero, we have

$$R_2 = mg \tag{2}$$

$$R_1 \cdot \frac{h}{\cos 30°} = mg \cdot \frac{l}{2} \sin 30° \tag{3}$$

Finally, we are given that

$$R_1 = R_2 \tag{4}$$

Solving equation (2) and (4), we get $R_1 = \frac{mg}{1+\sin 30°} = \frac{2}{3}mg$ Substituting into equation (3), we get

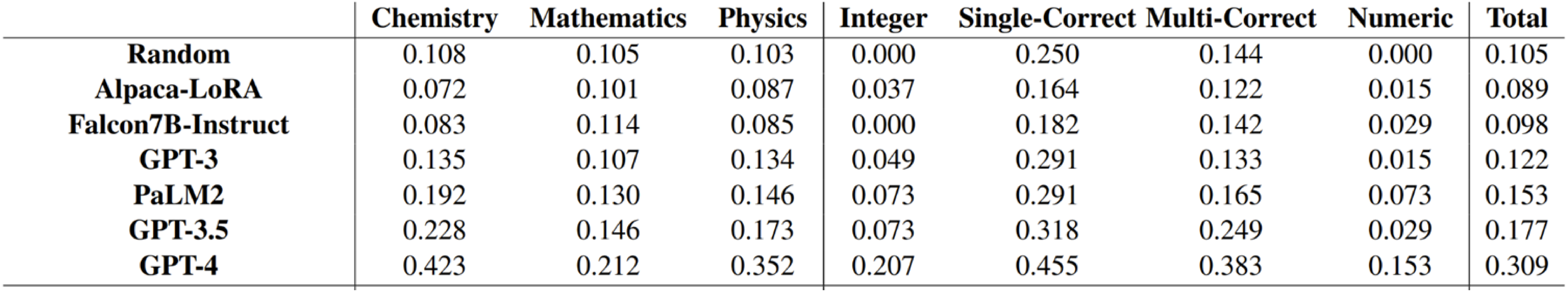
$$\frac{2}{3}mg\frac{h}{\cos 30°} = mg \cdot \frac{l}{2} \sin 30°$$

ALGEBRAIC MANIPULATION

$$\frac{h}{l} = \frac{3\cos 30° \sin 30°}{2 \cdot 2} = \frac{3 \cdot \frac{\sqrt{3}}{2} \cdot \frac{1}{2}}{4} = \frac{3\sqrt{3}}{16}$$

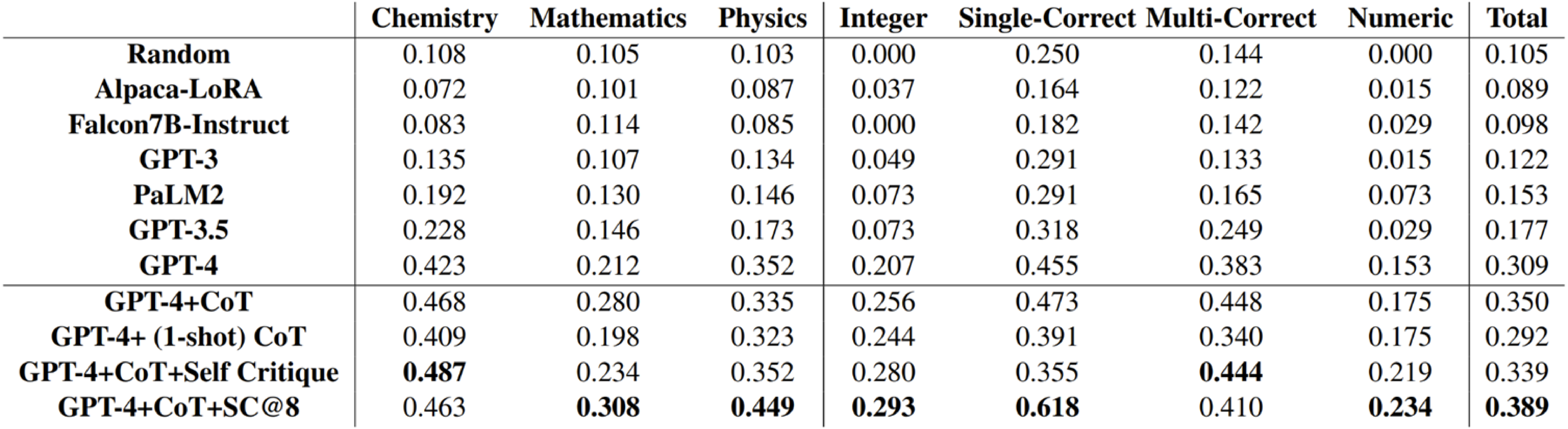
From (1), $f = R_1 \cos 30° = \frac{2}{3}mg \cos 30° = \frac{mg\sqrt{3}}{3} = \frac{16\sqrt{3}}{3}$. Therefore, option D is correct.

Evaluating LLMs on JEEBench



	Chemistry	Mathematics	Physics	Integer	Single-Correct	Multi-Correct	Numeric	Total
Random	0.108	0.105	0.103	0.000	0.250	0.144	0.000	0.105
Alpaca-LoRA	0.072	0.101	0.087	0.037	0.164	0.122	0.015	0.089
Falcon7B-Instruct	0.083	0.114	0.085	0.000	0.182	0.142	0.029	0.098
GPT-3	0.135	0.107	0.134	0.049	0.291	0.133	0.015	0.122
PaLM2	0.192	0.130	0.146	0.073	0.291	0.165	0.073	0.153
GPT-3.5	0.228	0.146	0.173	0.073	0.318	0.249	0.029	0.177
GPT-4	0.423	0.212	0.352	0.207	0.455	0.383	0.153	0.309

Evaluating LLMs on JEEBench



	Chemistry	Mathematics	Physics	Integer	Single-Correct	Multi-Correct	Numeric	Total
Random	0.108	0.105	0.103	0.000	0.250	0.144	0.000	0.105
Alpaca-LoRA	0.072	0.101	0.087	0.037	0.164	0.122	0.015	0.089
Falcon7B-Instruct	0.083	0.114	0.085	0.000	0.182	0.142	0.029	0.098
GPT-3	0.135	0.107	0.134	0.049	0.291	0.133	0.015	0.122
PaLM2	0.192	0.130	0.146	0.073	0.291	0.165	0.073	0.153
GPT-3.5	0.228	0.146	0.173	0.073	0.318	0.249	0.029	0.177
GPT-4	0.423	0.212	0.352	0.207	0.455	0.383	0.153	0.309
GPT-4+CoT	0.468	0.280	0.335	0.256	0.473	0.448	0.175	0.350
GPT-4+ (1-shot) CoT	0.409	0.198	0.323	0.244	0.391	0.340	0.175	0.292
GPT-4+CoT+Self Critique	**0.487**	0.234	0.352	0.280	0.355	**0.444**	0.219	0.339
GPT-4+CoT+SC@8	0.463	**0.308**	**0.449**	**0.293**	**0.618**	0.410	**0.234**	**0.389**

- Takeaways (repeated)
  - GPT-4 is unparalleled.
  - CoT (Let’s think step by step) and SC (Self Consistency) give **significant** gains.

JEE Adv 2022 Paper 2 Question 19

Question

A particle of mass 1 kg is subjected to a force which depends on the position as $\vec{F} = -k(x\hat{\imath}+y\hat{\jmath})kg\mathrm{ks}^{-2}$ with $k = 1$ kg s^{-2}. At time $t = 0$, the particle's position $\vec{r} = \left(\frac{1}{\sqrt{2}}\hat{\imath} + \sqrt{2}\hat{\jmath}\right) m$ and its velocity $\vec{v} = \left(-\sqrt{2}\hat{\imath} + \sqrt{2}\hat{\jmath} + \frac{2}{\pi}\hat{k}\right) ms^{-1}$. Let v_x and v_y denote the x and the y components of the particle's velocity, respectively. Ignore gravity. When $z = 0.5$ m, what is the value of $(xv_y - yv_x)$ in m^2s^{-1}?

First, we need to find the equations of motion for the particle. Since the force depends on position, we can write the equations as:

$\frac{d^2x}{dt^2} = -kx$

$\frac{d^2y}{dt^2} = -ky$

These are simple harmonic motion equations. The general solutions for these equations are:

$x(t) = A\cos(\sqrt{k}t) + B\sin(\sqrt{k}t)$

$y(t) = C\cos(\sqrt{k}t) + D\sin(\sqrt{k}t)$

Now, we need to find the constants A, B, C, and D using the initial conditions.

At $t = 0$, we have:

$x(0) = \frac{1}{\sqrt{2}} = A\cos(0) + B\sin(0)$

$y(0) = \sqrt{2} = C\cos(0) + D\sin(0)$

So, A = $\frac{1}{\sqrt{2}}$ and C = $\sqrt{2}$.

Now, we need to find the initial velocities:

$v_x(t) = -A\sqrt{k}\sin(\sqrt{k}t) + B\sqrt{k}\cos(\sqrt{k}t)$

$v_y(t) = -C\sqrt{k}\sin(\sqrt{k}t) + D\sqrt{k}\cos(\sqrt{k}t)$

At $t = 0$, we have:

$v_x(0) = -\sqrt{2} = -A\sqrt{k}\sin(0) + B\sqrt{k}\cos(0)$

$v_y(0) = \sqrt{2} = -C\sqrt{k}\sin(0) + D\sqrt{k}\cos(0)$

So, B = $-\sqrt{2}$ and D = $\sqrt{2}$.

Now we have the position equations:

$x(t) = \frac{1}{\sqrt{2}}\cos(\sqrt{k}t) - \sqrt{2}\sin(\sqrt{k}t)$

$y(t) = \sqrt{2}\cos(\sqrt{k}t) + \sqrt{2}\sin(\sqrt{k}t)$

We are given that $z = 0.5m$, and the initial z-component of velocity is $\frac{2}{\pi}ms^{-1}$. Since there is no force in the z-direction, the z-component of velocity remains constant. We can find the time when $z = 0.5m$:

$z(t) = z(0) + v_z t$

$0.5 = 0 + \frac{2}{\pi}t$

$t = \frac{\pi}{4}$

Now, we can find x and y at this time:

$x\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}\cos\left(\frac{\pi}{2}\right) - \sqrt{2}\sin\left(\frac{\pi}{2}\right) = -\sqrt{2}$

$y\left(\frac{\pi}{4}\right) = \sqrt{2}\cos\left(\frac{\pi}{2}\right) + \sqrt{2}\sin\left(\frac{\pi}{2}\right) = \sqrt{2}$

Now, we need to find the x and y components of velocity at this time:

$v_x\left(\frac{\pi}{4}\right) = -\frac{1}{\sqrt{2}}\sin\left(\frac{\pi}{2}\right) - \sqrt{2}\cos\left(\frac{\pi}{2}\right) = -\frac{1}{\sqrt{2}}$

$v_y\left(\frac{\pi}{4}\right) = -\sqrt{2}\sin\left(\frac{\pi}{2}\right) + \sqrt{2}\cos\left(\frac{\pi}{2}\right) = -\sqrt{2}$

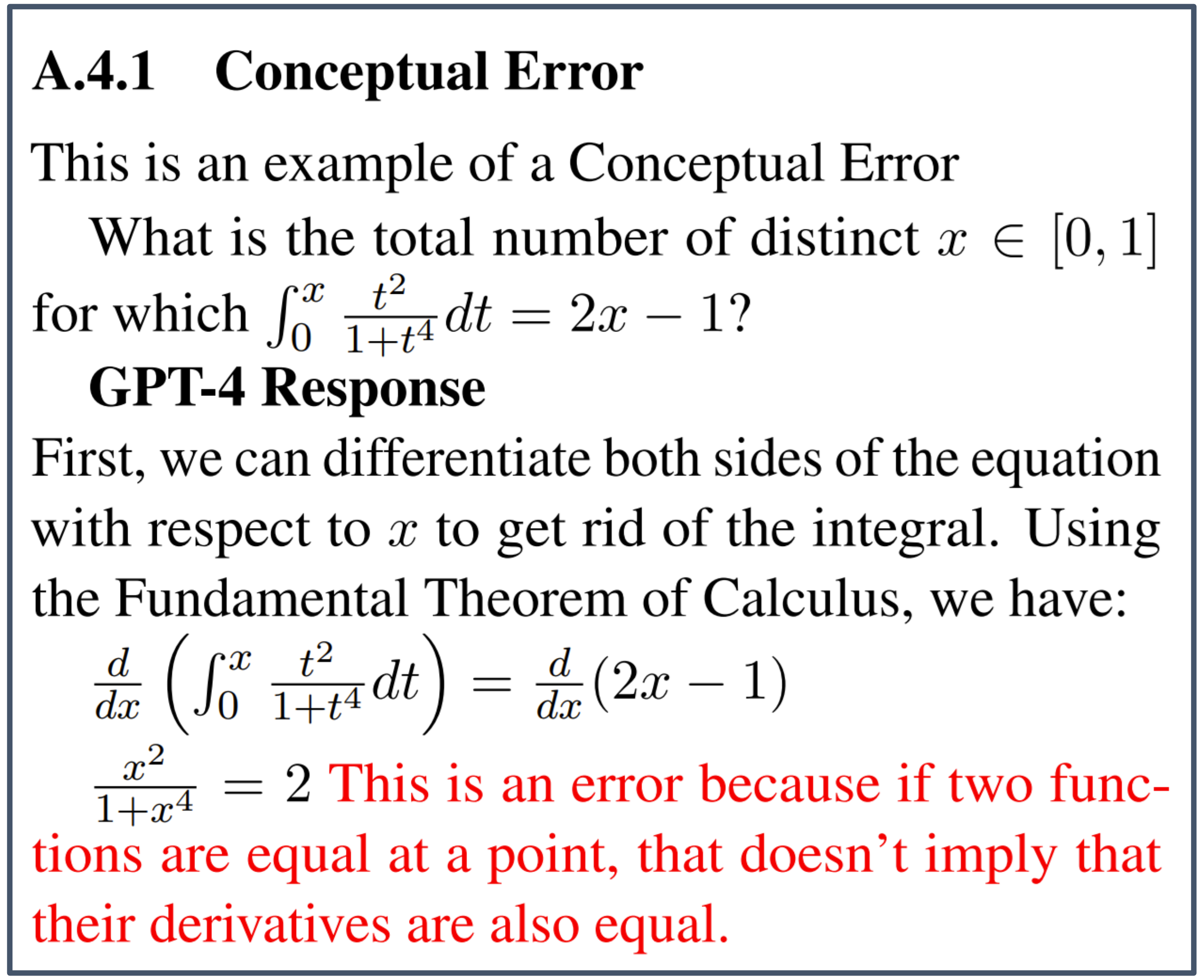
Finally, we can find the value of $(xv_y - yv_x)$:

$(xv_y - yv_x) = (-\sqrt{2})(-\sqrt{2}) - (\sqrt{2})(-\frac{1}{\sqrt{2}}) = 2 + 1 = 3m^2s^{-1}$

The final answer is 3.

Where does GPT4 fail?

Conceptual Error



A.4.1 Conceptual Error

This is an example of a Conceptual Error

What is the total number of distinct $x \in [0,1]$ for which $\int_0^x \frac{t^2}{1+t^4} dt = 2x - 1$?

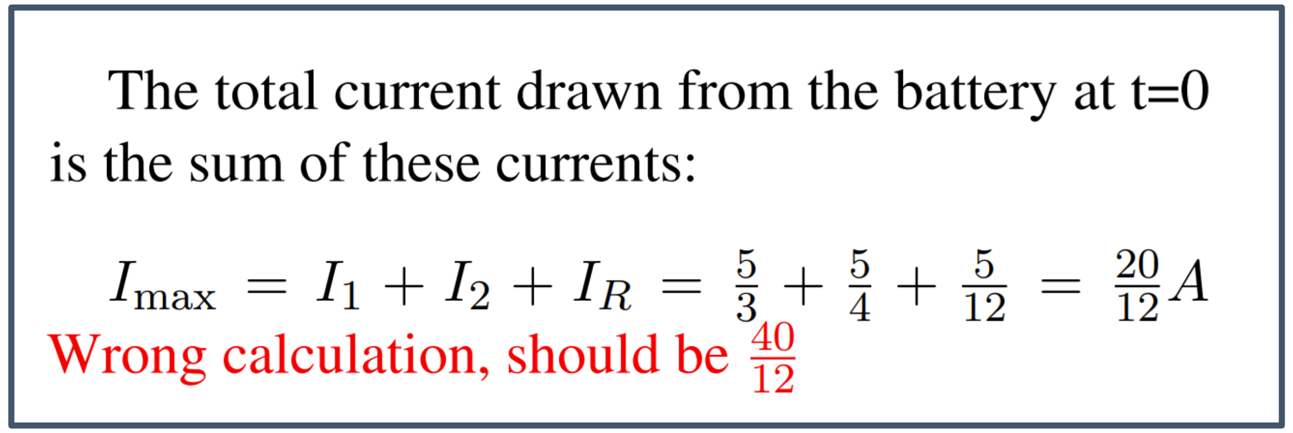
GPT-4 Response

First, we can differentiate both sides of the equation with respect to x to get rid of the integral. Using the Fundamental Theorem of Calculus, we have:

$$\frac{d}{dx}\left(\int_0^x \frac{t^2}{1+t^4} dt\right) = \frac{d}{dx}(2x-1)$$

$\frac{x^2}{1+x^4} = 2$ This is an error because if two functions are equal at a point, that doesn't imply that their derivatives are also equal.

Computational Error

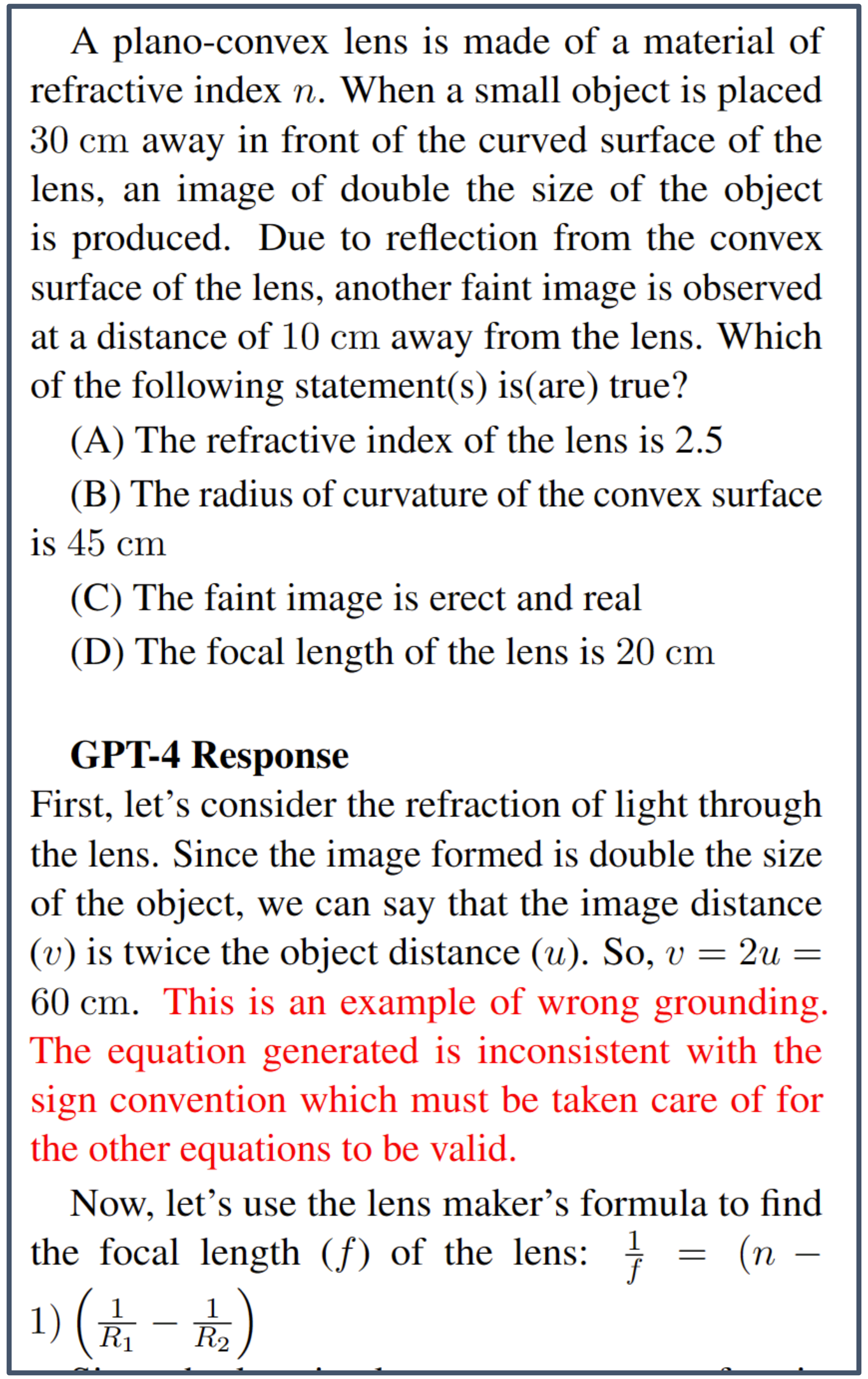


The total current drawn from the battery at t=0 is the sum of these currents:

$$I_{\max} = I_1 + I_2 + I_R = \frac{5}{3} + \frac{5}{4} + \frac{5}{12} = \frac{20}{12} A$$

Wrong calculation, should be $\frac{40}{12}$

Grounding Error



A plano-convex lens is made of a material of refractive index n. When a small object is placed 30 cm away in front of the curved surface of the lens, an image of double the size of the object is produced. Due to reflection from the convex surface of the lens, another faint image is observed at a distance of 10 cm away from the lens. Which of the following statement(s) is(are) true?

(A) The refractive index of the lens is 2.5

(B) The radius of curvature of the convex surface is 45 cm

(C) The faint image is erect and real

(D) The focal length of the lens is 20 cm

GPT-4 Response

First, let's consider the refraction of light through the lens. Since the image formed is double the size of the object, we can say that the image distance (v) is twice the object distance (u). So, $v = 2u =$ 60 cm. This is an example of wrong grounding. The equation generated is inconsistent with the sign convention which must be taken care of for the other equations to be valid.

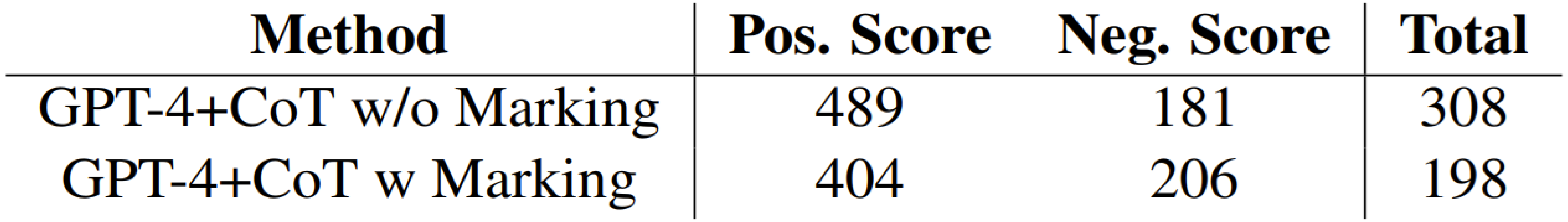
Now, let's use the lens maker's formula to find the focal length (f) of the lens: $\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$

Error Analysis

Error Type	Count
Conceptual Error	34
Computation Error	30
Grounding Error	15
Problem Miscomprehension	1

Can GPT4 decide when not to answer?

- The real bane of the exam: **negative marking.**
  - LLM needs to introspect for its own confidence levels
  - It must self-decide whether to answer based on its confidence.
- Mere prompting with the marking scheme doesn't help!



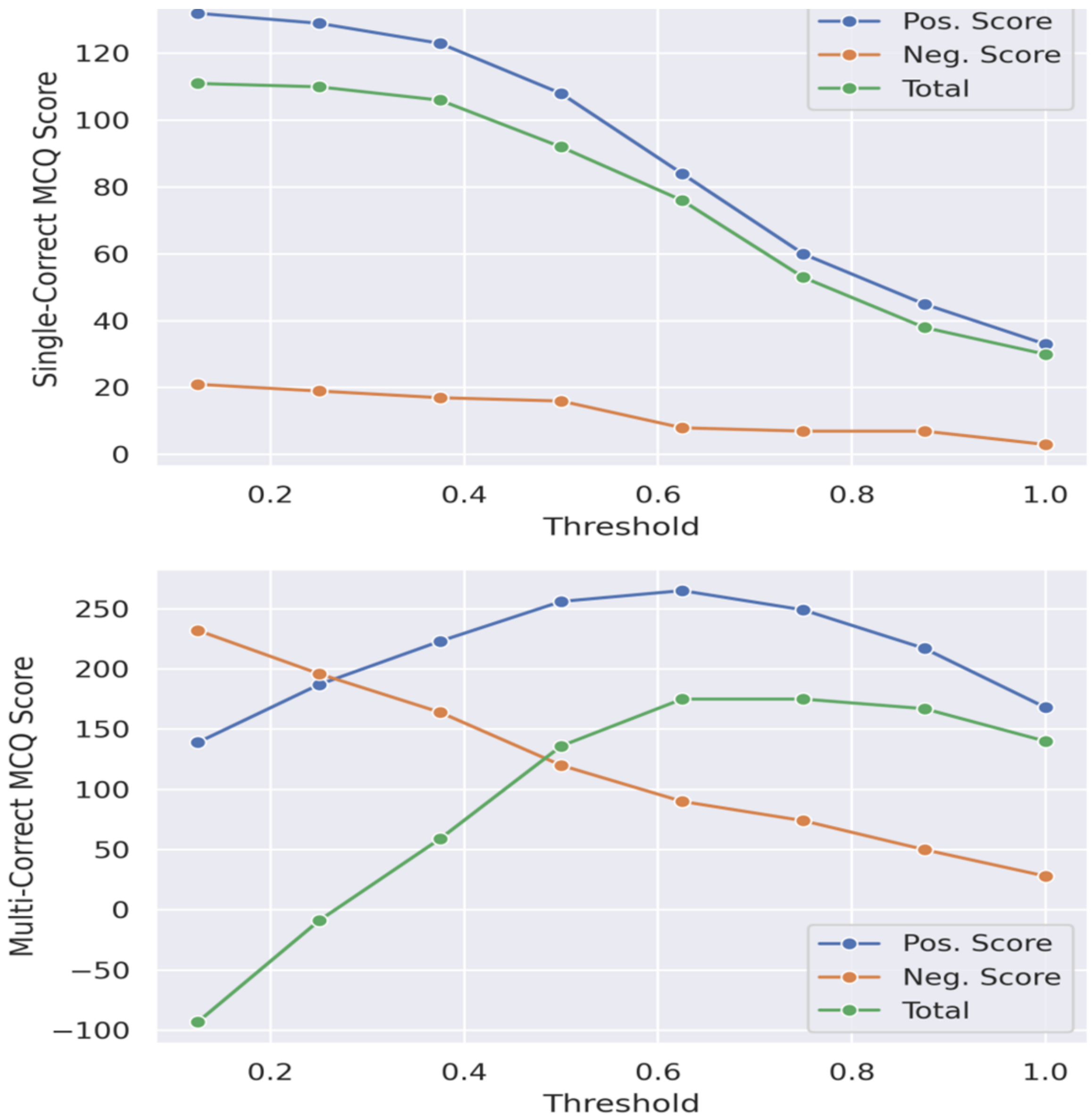
Method	**Pos. Score**	**Neg. Score**	**Total**
GPT-4+CoT w/o Marking	489	181	308
GPT-4+CoT w Marking	404	206	198

Can GPT4 decide when not to answer?

- Sample responses (temp=0.5) → estimate confidence of top answer
- Learn threshold on this confidence
- Only answer if confidence exceeds threshold.

Method	Pos Score	Neg Score	Total Score
GPT-4+CoT	109	43	66
GPT-4+CoT+SC	118	49	69
GPT-4+CoT+SC+Thresholding	111	39	**72**

Table 6: Marks on the test set obtained when optimal thresholds derived from the val set are used.



e 7: Scores obtained on different thresholding val-
n Single-Correct(top) and Multi-Correct(bottom)
questions from the val set, the optimal value is
$le = 0.125$ and $\tau_{multiple} = 0.75$

GPT4 vs. Human Performance

with lots of caveats & disclaimers

GPT4 lies in the top 80-90% of JEE Advanced applicants.

Takeaways

- No model comes even remotely close to GPT4 in terms of reasoning QA.
- GPT4 is not good at estimating its own confidence.
- GPT4 is *much* stronger than the average human in reasoning, but still much weaker than the smartest humans.
- Use our benchmarks to measure your LLMs! ☺