Large Language Models

Simon Ellershaw
PhD Candidate Institute of Health Informatics UCL
simon.ellershaw.20@ucl.ac.uk

Buzzwords

- LLM
- GPT
- Transformer
- Next token prediction
- Few shot prompting
- Chain of Thought
- Self-Consistency
- Finetuning
- Multi Modal
- Foundation Models
- RLHF
- Scaling Laws
- Open Source
- RAG
- Semantic Search
- Embeddings
- Agents
- Reasoning Model

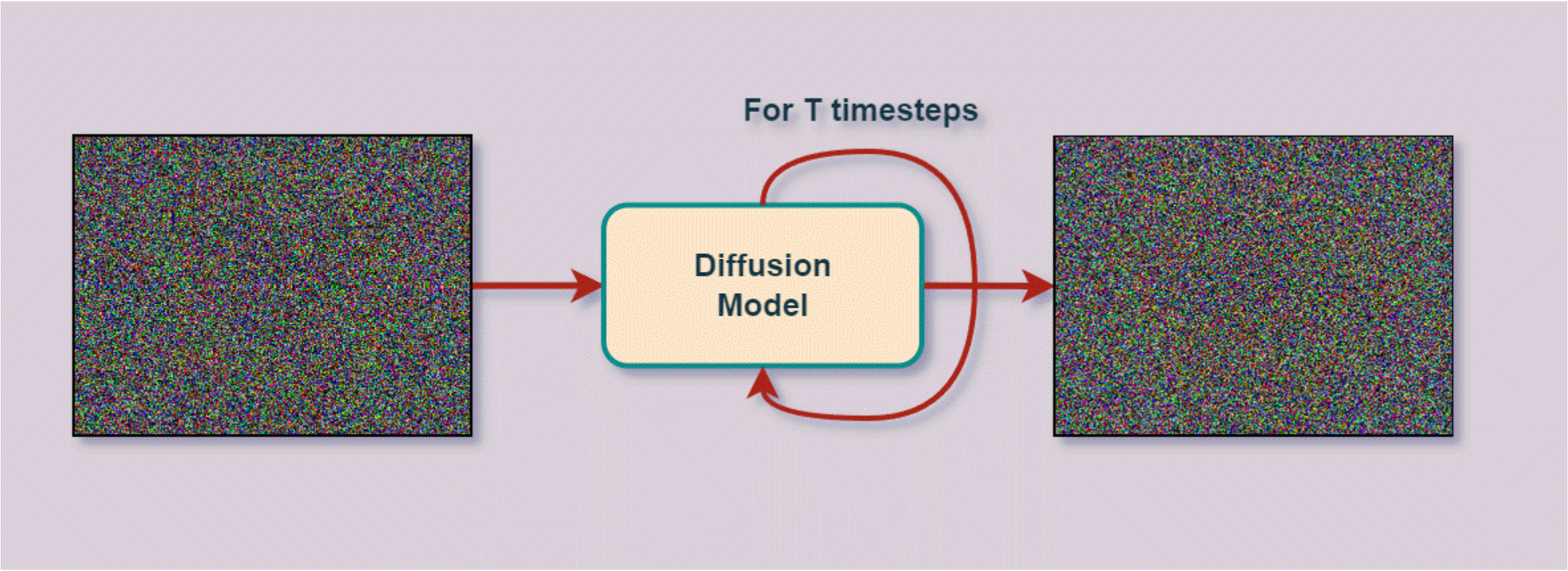
Content

1. What is an LLM?
2. How to use an LLM
3. What does this mean for medicine

Content

1. What is an LLM?
2. How to use an LLM
3. What does this mean for medicine

Content



https://erdem.pl/2023/11/step-by-step-visual-introduction-to-diffusion-models

ChatGPT

SI

What can I help with?

Write the answer to Prof Taylor's latest assignment

Search Reason

Create image Code Summarize text Help me write More

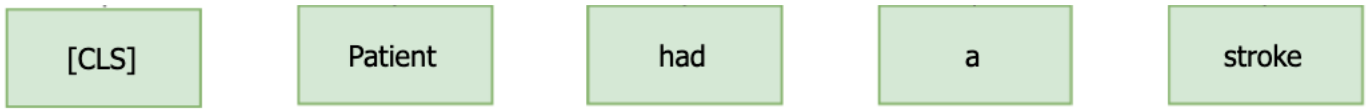
So what is GPT-4?

GPT = **Generative Pre-trained Transformer**

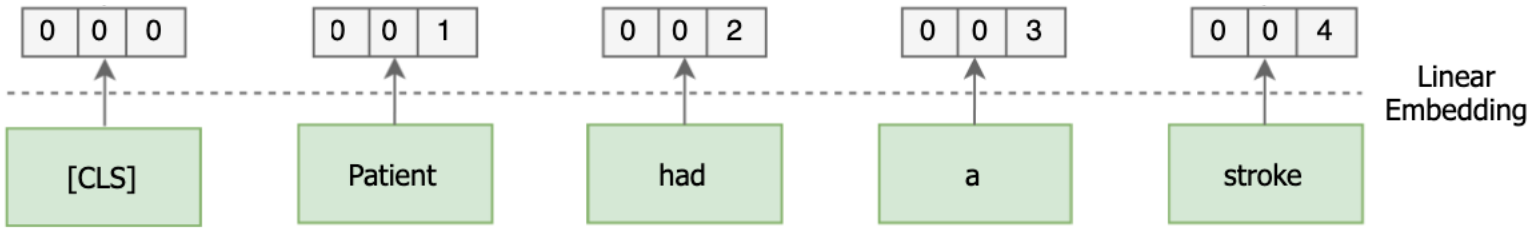
"GPT-4 is a **Transformer-style model** [39] pre-trained to **predict the next token** in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using **Reinforcement Learning from Human Feedback** (RLHF) [40]. 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."

OpenAI "GPT-4 Technical Report" *arXiv preprint arXiv:2303.08774* (2023).

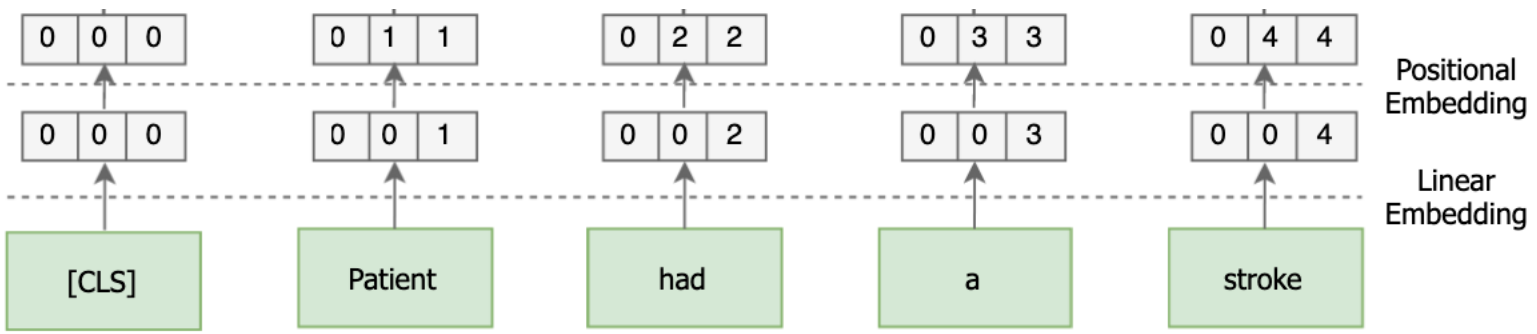
Transformer



Transformer

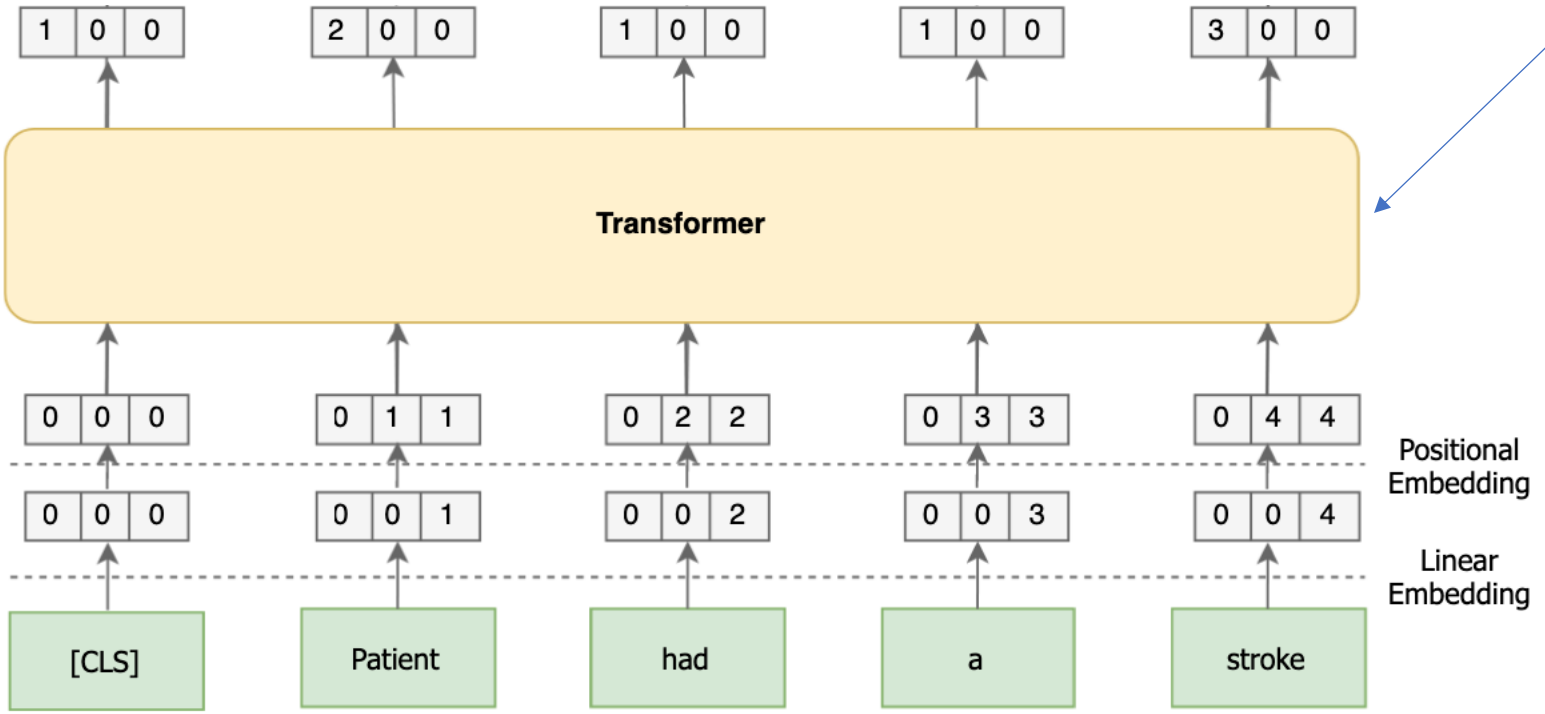


Transformer

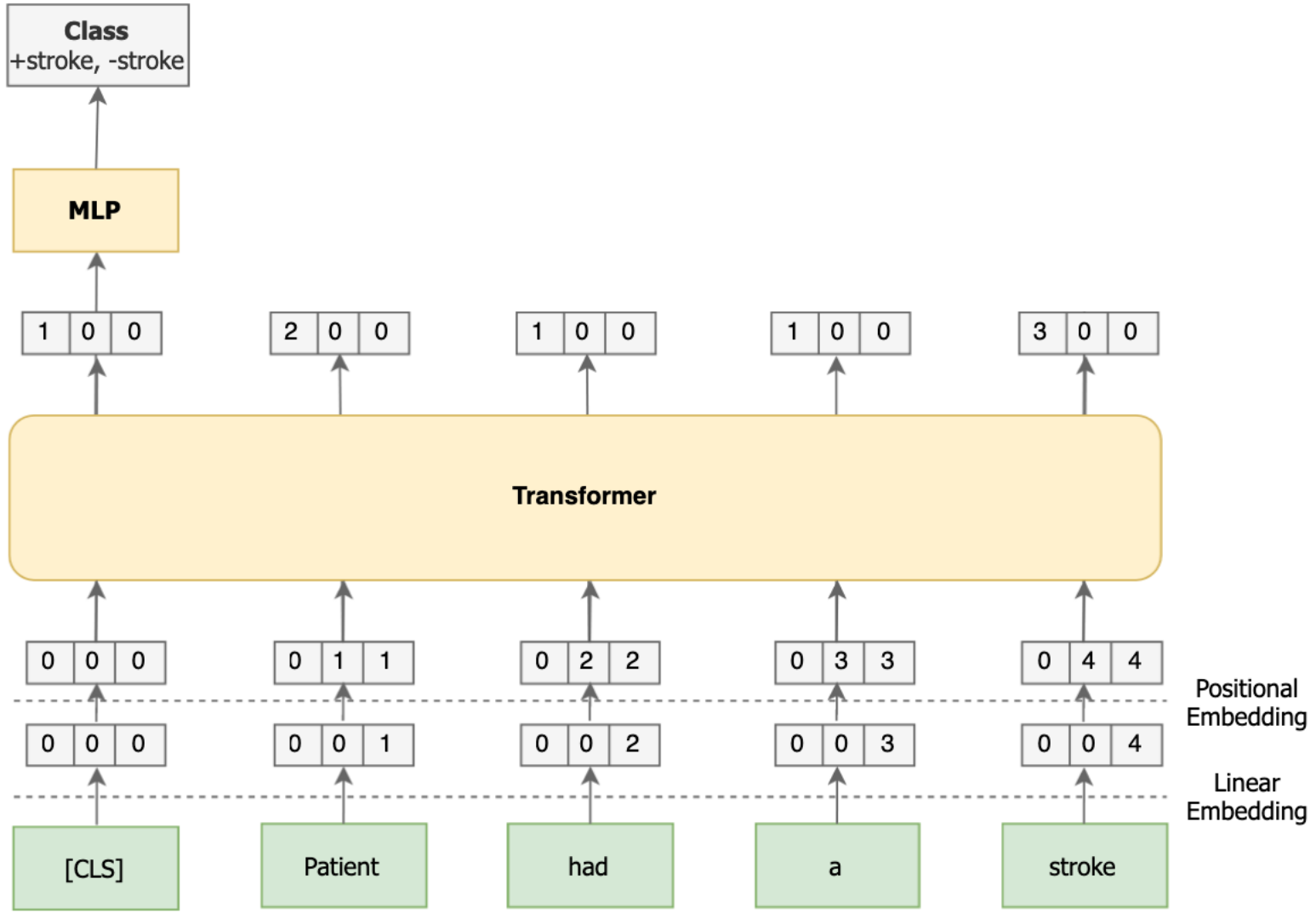


Transformer

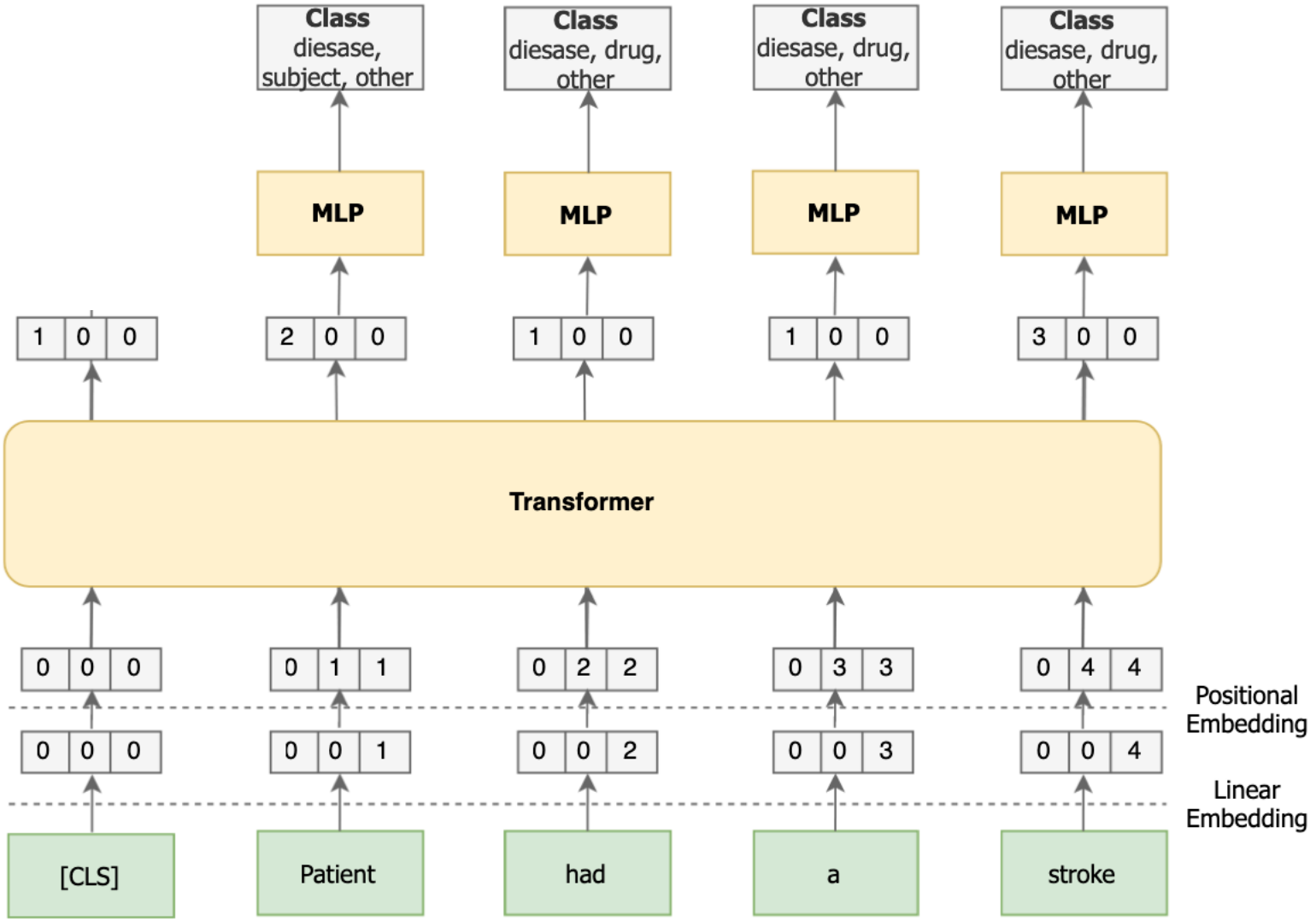
If you want to peek inside the yellow box this is a [great blog post]



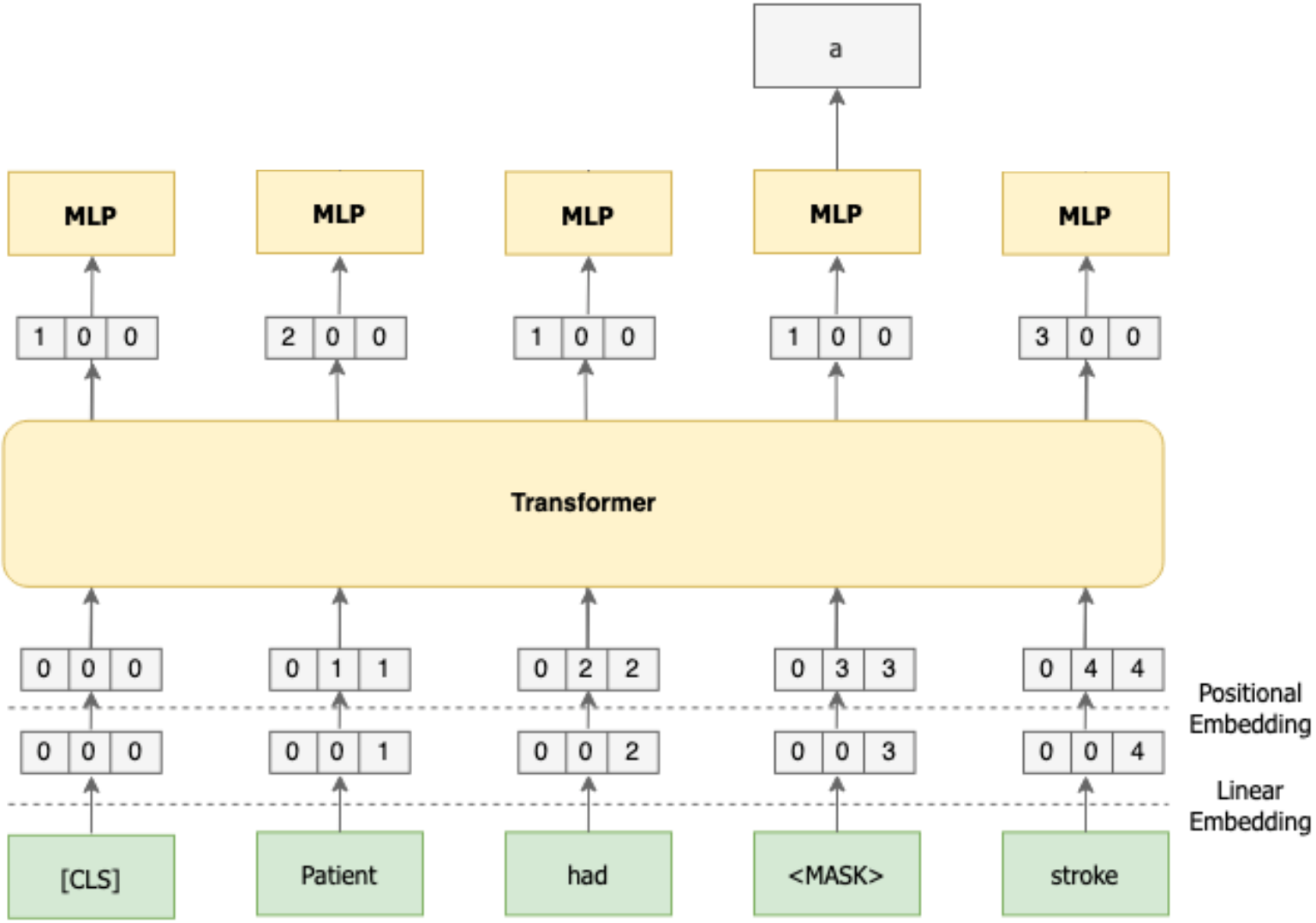
Supervised Learning: Classification



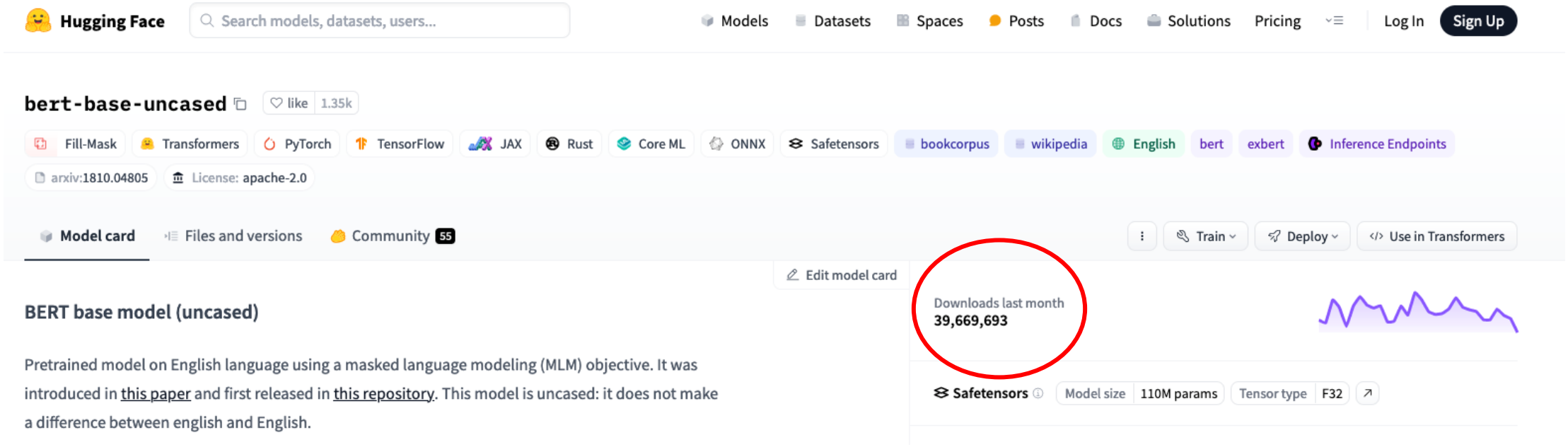
Supervised Learning : Named Entity Recognition



Pretraining: Masked Language Modelling



Pretraining: BERT SOTA ~202



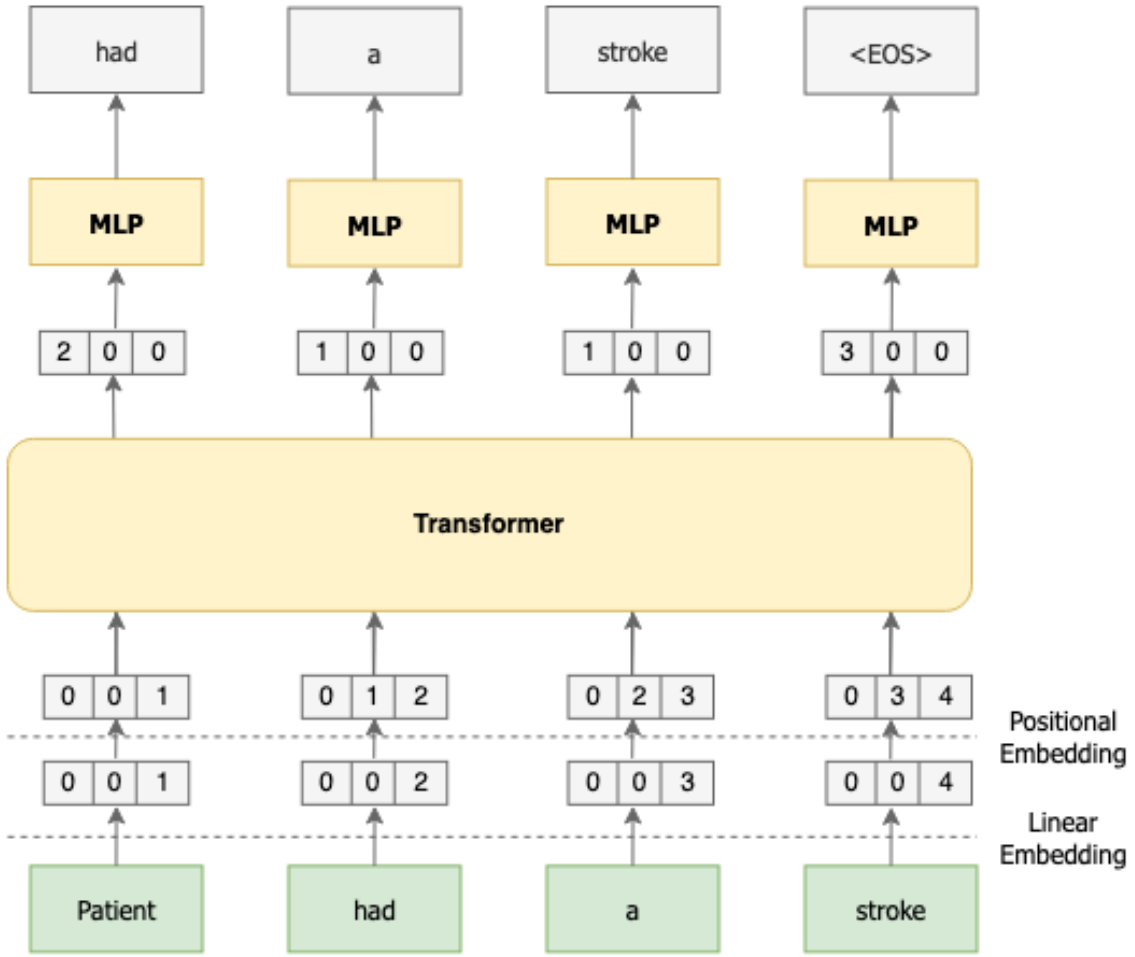
"For the pre-training corpus we use the BooksCorpus (800M words) (Zhu et al.,2015) and English Wikipedia (2,500M words)."

Supervised training for

- ICD 10 Coding
- Snomed concept extraction
- Adverse Drug Events

https://huggingface.co/bert-base-uncased
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

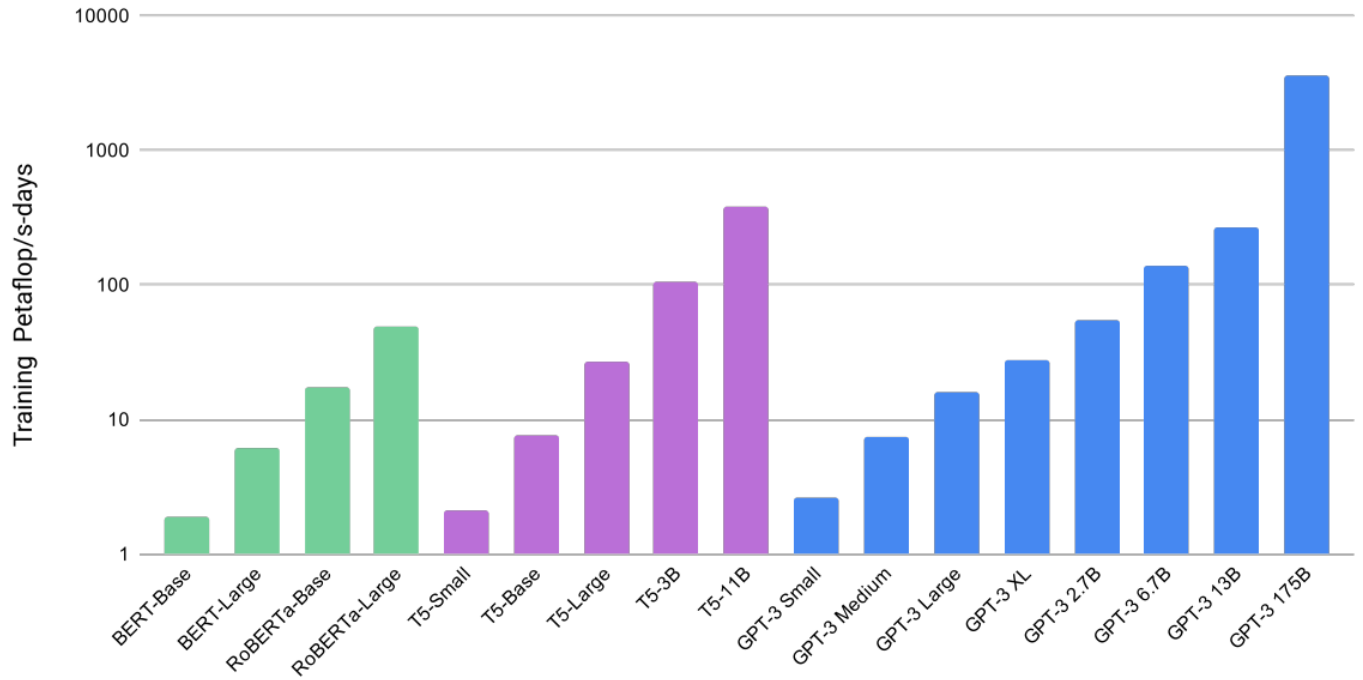
Next Token Prediction: BERT->GPT



*different attention mechanism as well

BERT vs GPT-3

	BERT	LLM
Pre-training objective	Masked Language Modelling	Next Token Prediction
Number Training Tokens	3200 Million	300 Billion
Number of Parameter	340 Million	175 Billion



Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

BERT vs GPT-3

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Abstract

3 Results

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Scaling Laws

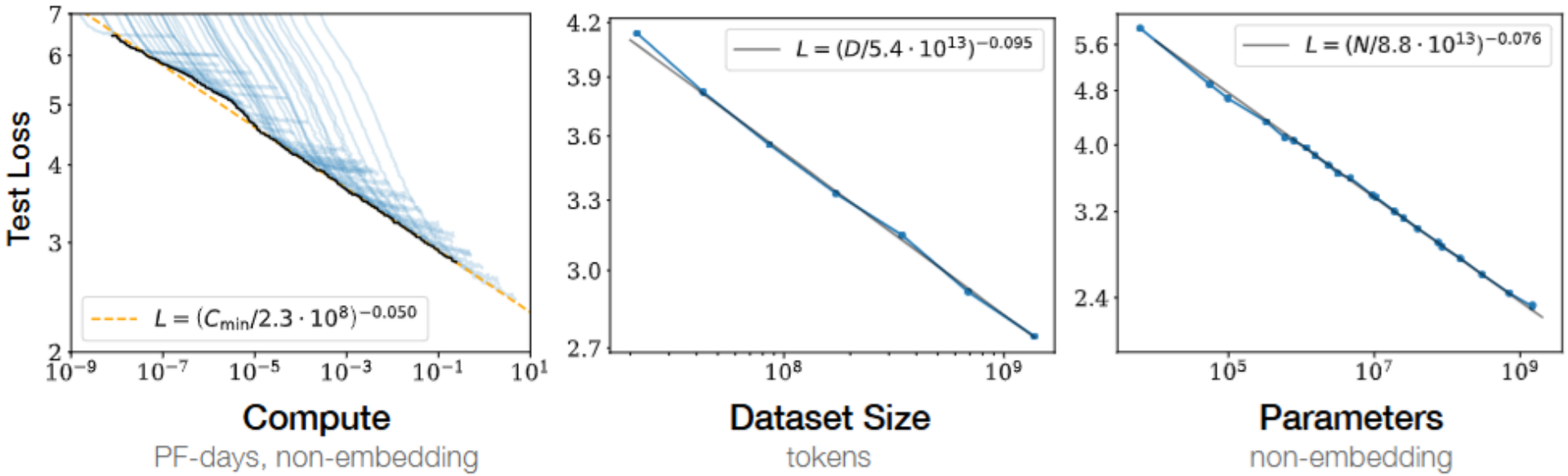
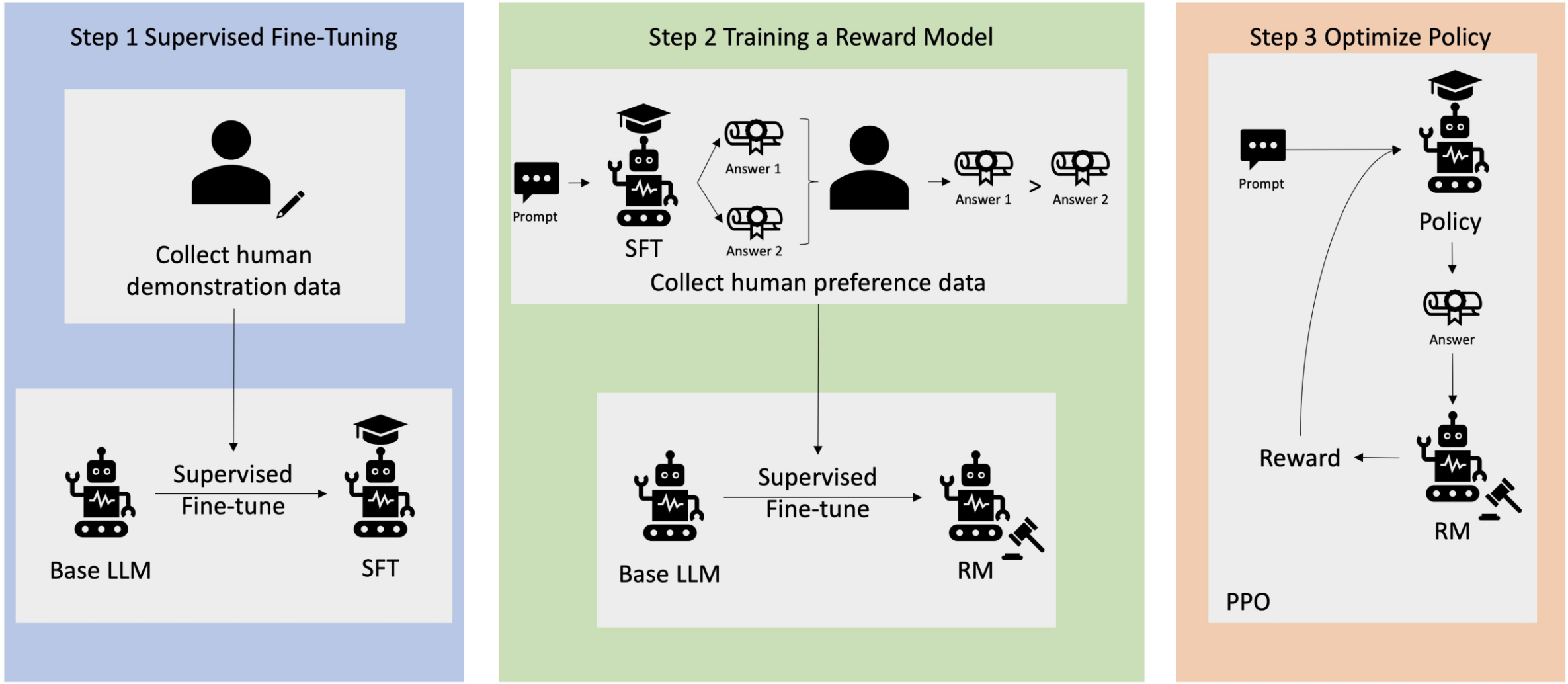


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

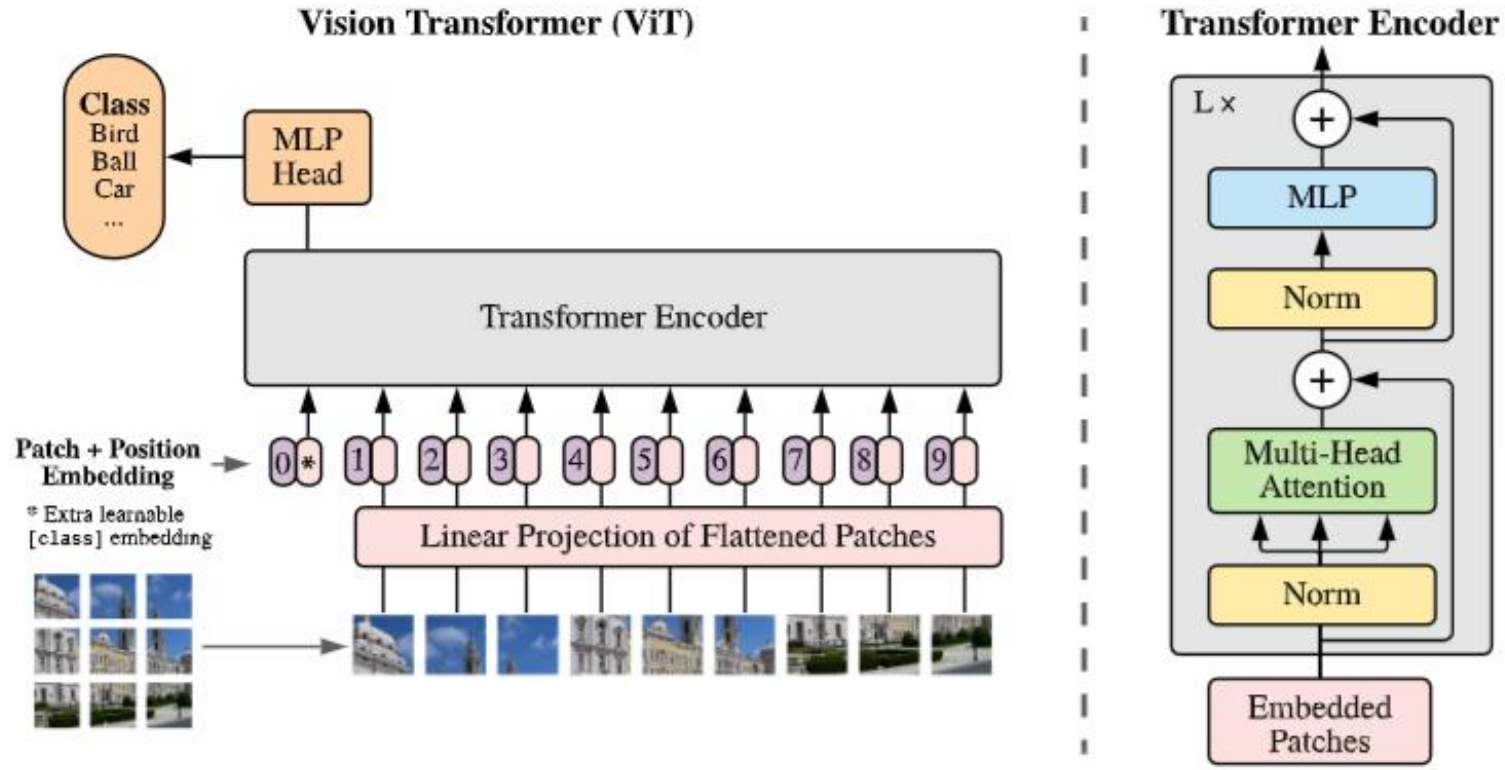
Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

RLHF: GPT-> ChatGPT



https://aws.amazon.com/blogs/machine-learning/improving-your-llms-with-rlhf-on-amazon-sagemaker/

Multi-Modal: Will it tokenize?



Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

ChatGPT

SI

What can I help with?

Write the answer to Prof Taylor's latest assignment

Search

Reason

Create image

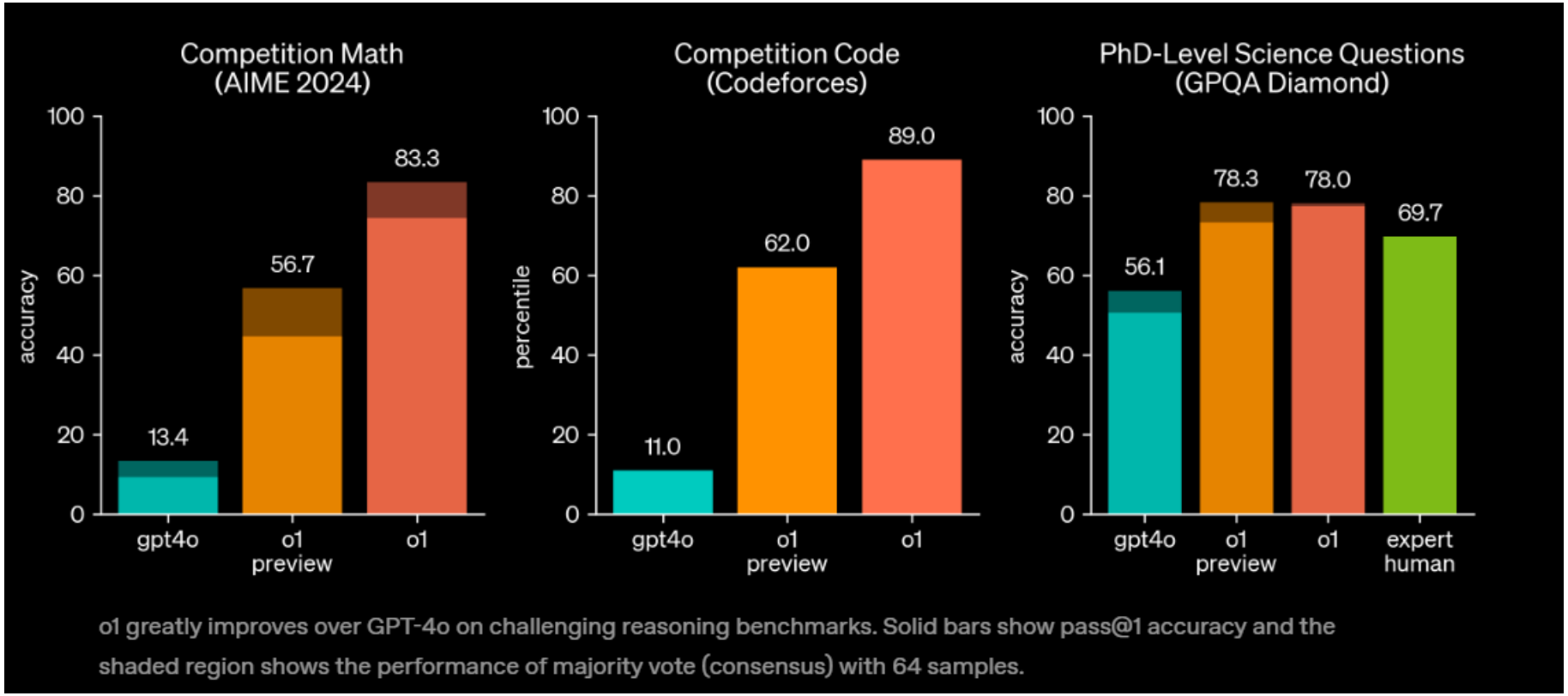
Code

Summarize text

Help me write

More

Reasoning Models



https://openai.com/index/learning-to-reason-with-llms/

Reasoning Models: DeepSeek

https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Chain of Thought

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. ✓

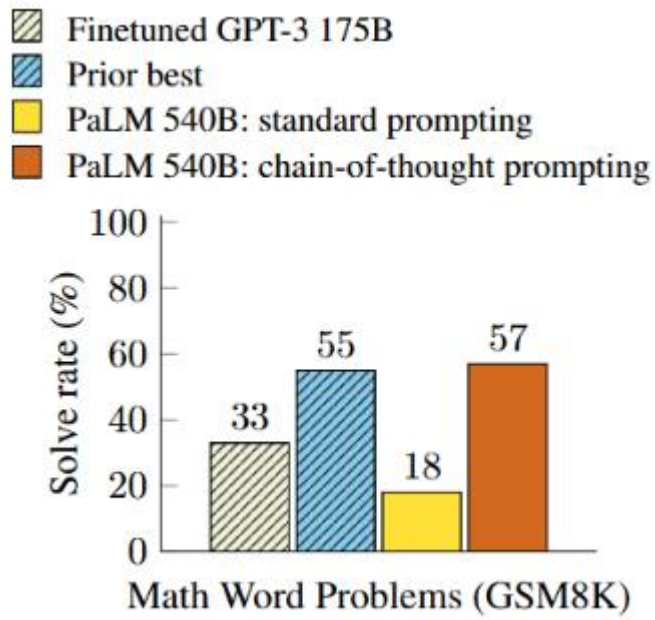
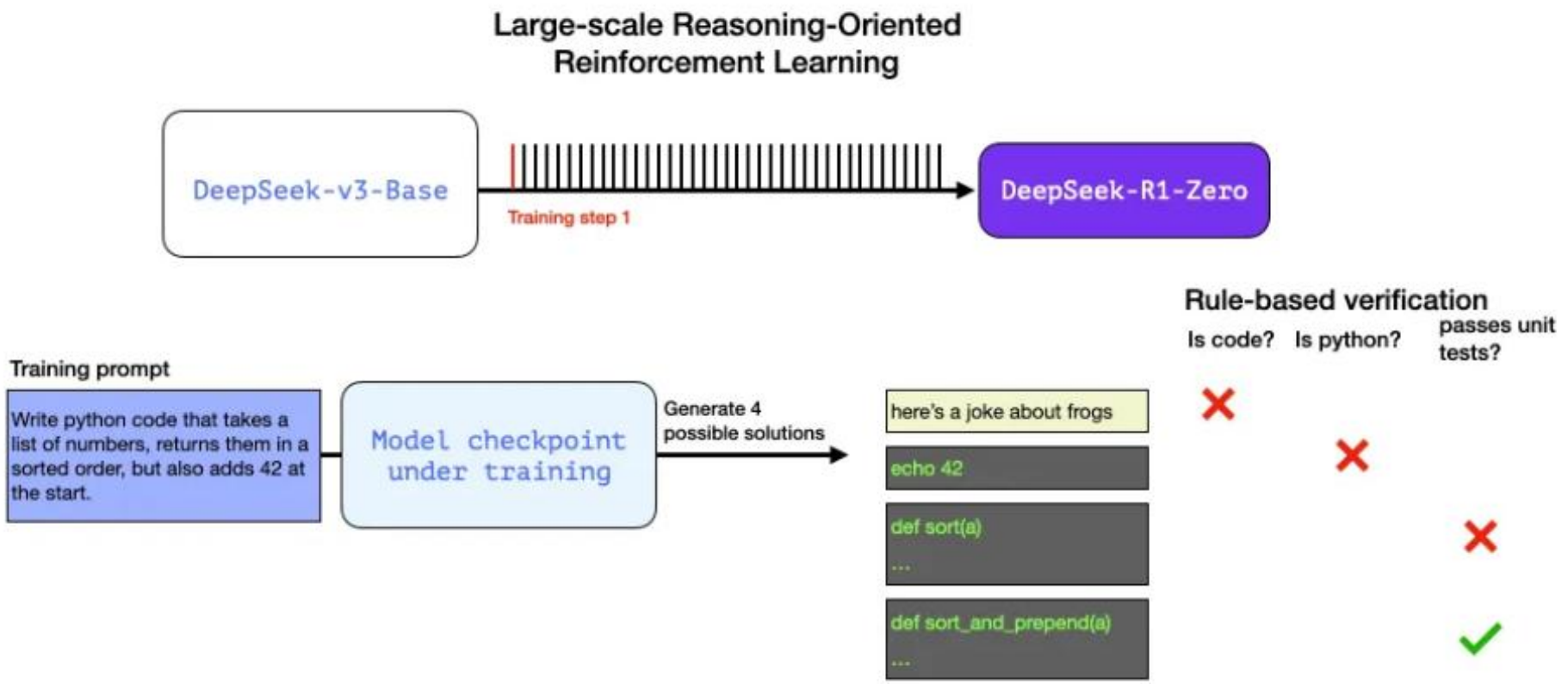


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Reasoning Models: DeepSeek



https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Reasoning Models: DeepSeek

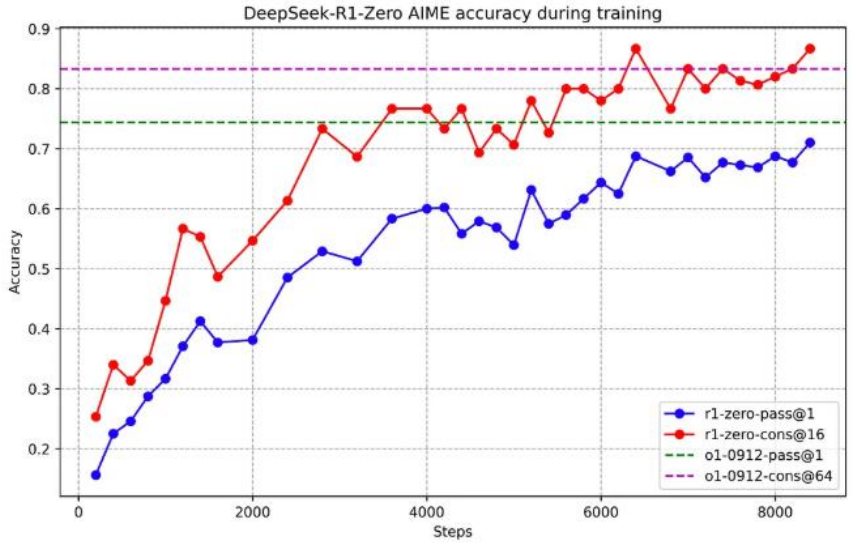


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

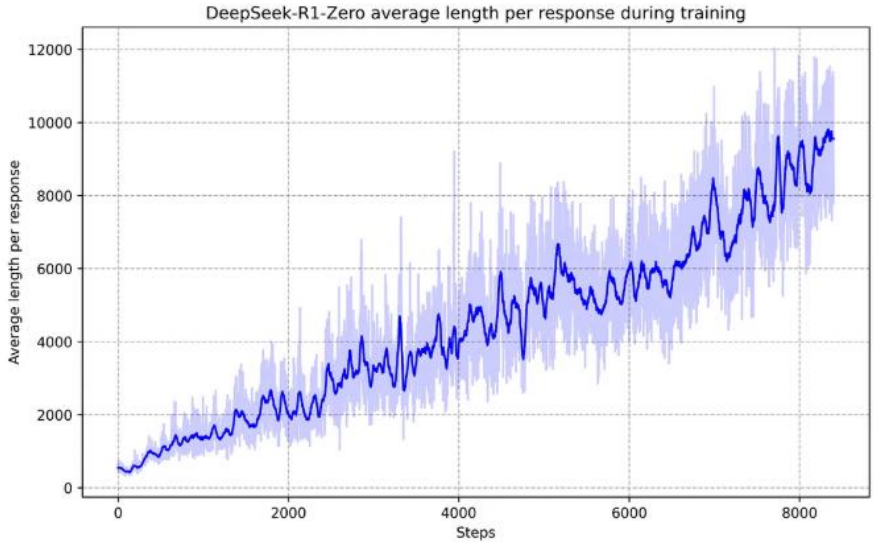
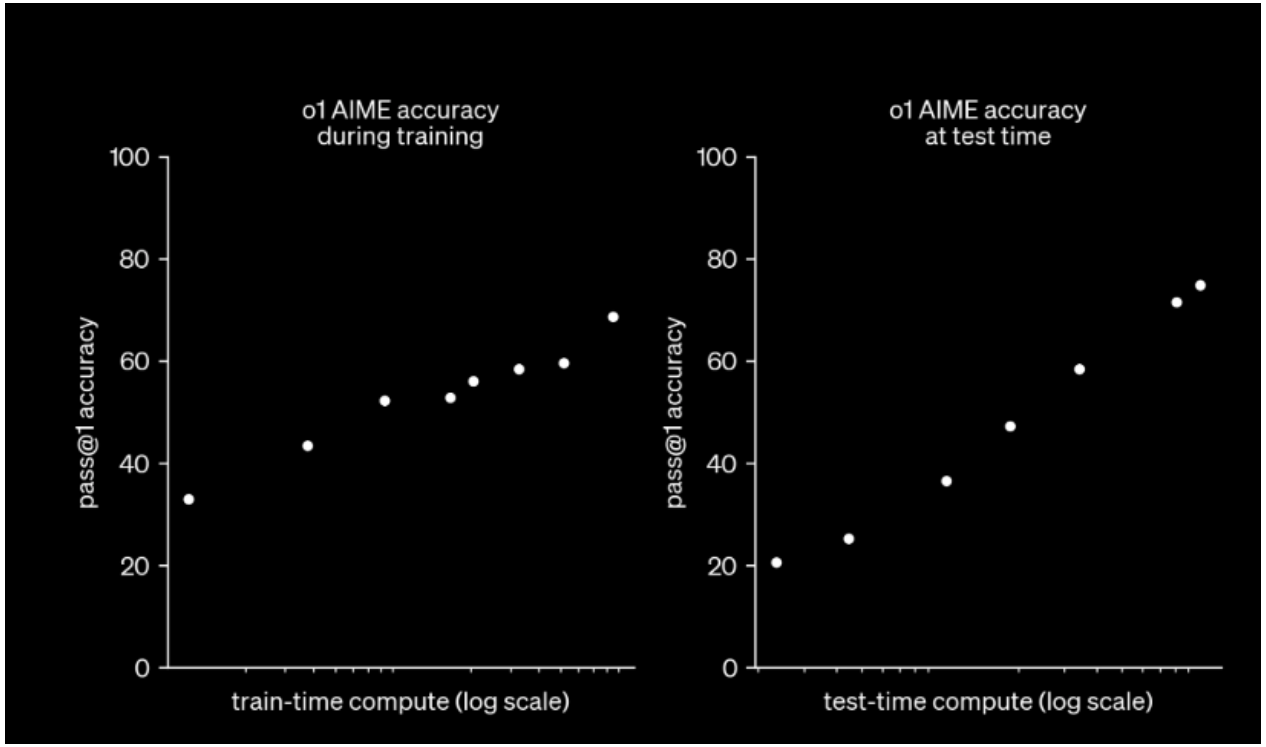


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

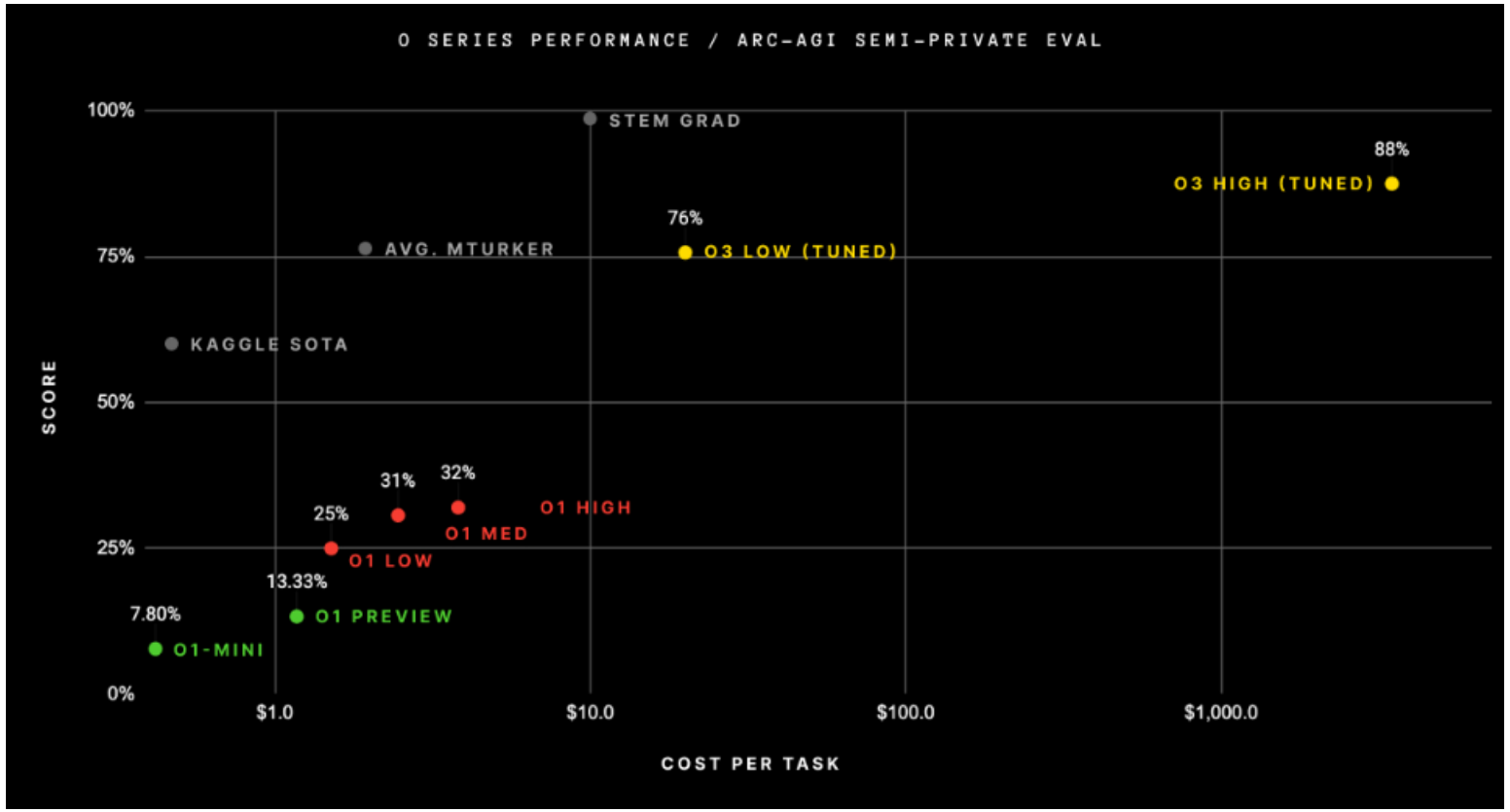
https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Reasoning Models: Test time compute



https://openai.com/index/learning-to-reason-with-llms/

Reasoning Models: o3?



https://arcprize.org/blog/oai-o3-pub-breakthrough

Take away points

- How to make ChatGPT. Scale all of:
  1. Transformer model
  2. Large-scale unsupervised pre-training
     1. Multi-modal: Will it tokenize?
  3. RLHF to make chatty

  Gives a general few/zero-shot model

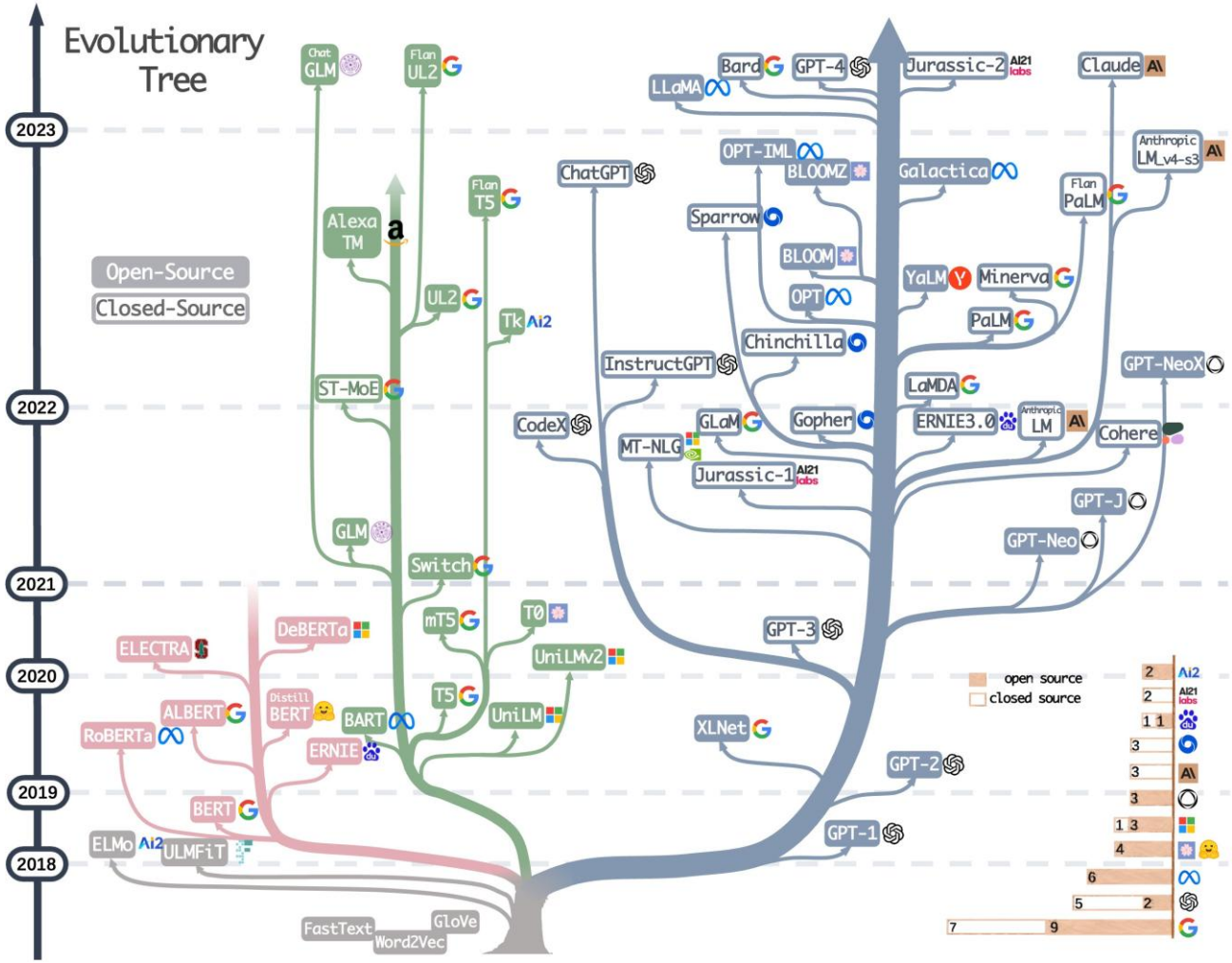
- Reasoning models
  - Seems to be even more RL(HF)
  - Watch this space…


QUESTION TIME

Have a break, have a Kit Kat.®

Part 2- How to use an LLM

What model?



https://github.com/Mooler0410/LLMsPracticalGuide

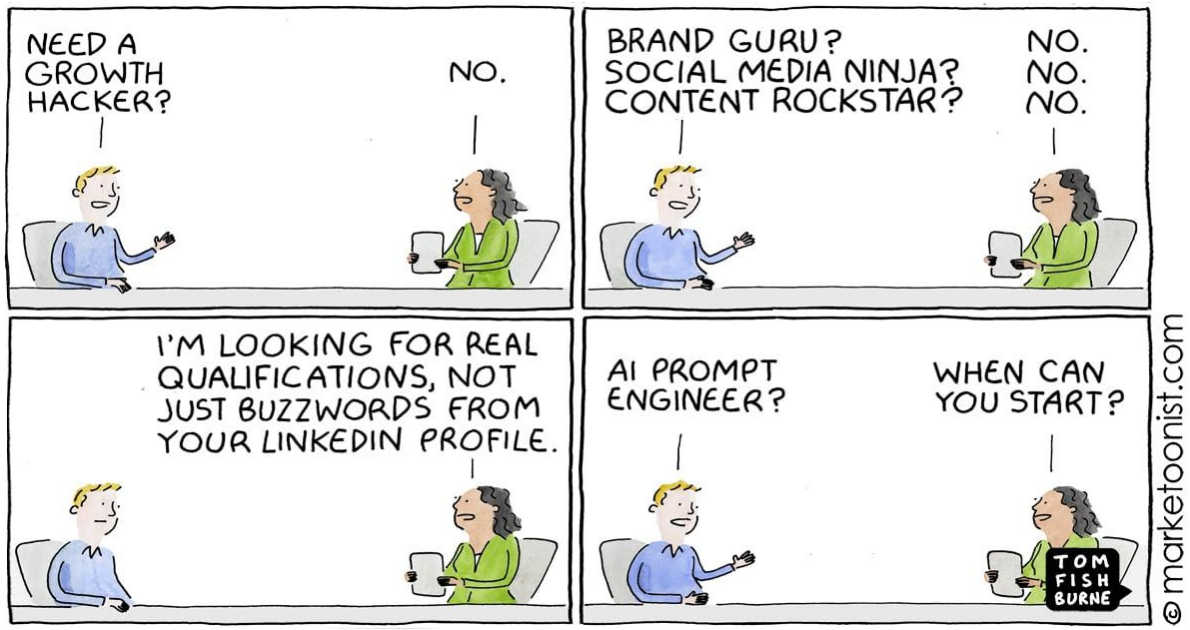
Open Source/Open Weight

- Open Weights = Model weights are freely accessible on the internet
- Open Source = Training code, data and model are freely accessible on the internet
- Current ~6-12 month lag with closed source
- Can be finetuned using supervised learning by approximation (e.g. LORA)
- 3rd Party vs Local Hosting $$$
  - “Small” large language models

Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).
Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Prompt Engineering

- Be explicit in input/output
- Few shot prompting
- Chain of thought
- Self-consistency
- Structured Output



Prompt Engineering

System:
"""You are a consultant doctor tasked with writing a patients discharge summary.
A user will provide you with a list of clinical notes from a hospital stay from which you will write a discharge summary.
Each clinical note has a title of the format [Title]: [timestamp year-month-day hour:min].
Clinical notes are ordered by ascending timestamp.
Only the information in the clinical notes provided by the most recent user message can be used for this task.

The discharge summary must be written in accordance with the following json schema.
{json_schema}
All fields are required.
If the relevant information is not present in the clinical notes, fields can be filled with an empty string or list.
Expand all acronyms to their full terms."""

User:
"""Clinical Notes
{rcp_exemplar_clinican_notes}
Please write a discharge summary only using the information in this message's clinical notes.
The discharge summary must be written in accordance with the json schema given in the system message."""

Assistant:
"""{rcp_exemplar_discharge_summary_json}"""

One-shot example

User:
"""Clinical Notes
{input_clinician_notes}
Please write a discharge summary only using the information in this message's clinical notes.
The discharge summary must be written in accordance with the json schema given in the system message."""

Assistant (excerpt):
"""...
admission_details": {
"reason_for_admission": "Chest tightness pain, breathlessness, nausea and dizziness started at 6 am.",
"admission_method": "Emergency admission via London Ambulance Service",
"relevant_past_medical_and_mental_health_history": [
"Type 2 Diabetes medication (tablets)",
"Hypertension",
"Chronic Obstructive Pulmonary Disease"
]
},...."""

Ellershaw, Simon, et al. "Automated Generation of Hospital Discharge Summaries Using Clinical Guidelines and Large Language Models." *AAAI 2024 Spring Symposium on Clinical Foundation Models*. 2024.

Question Answering on MedQA

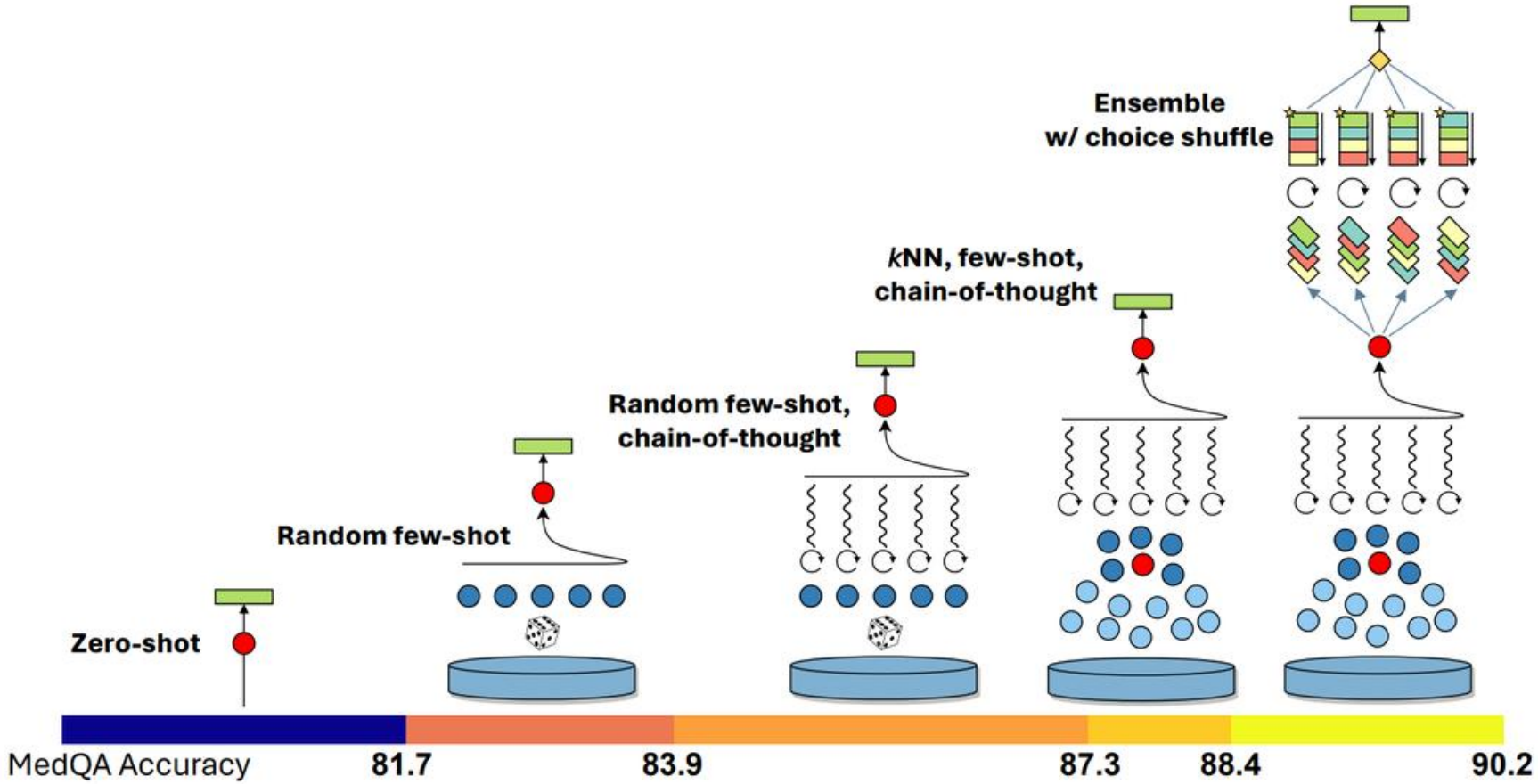
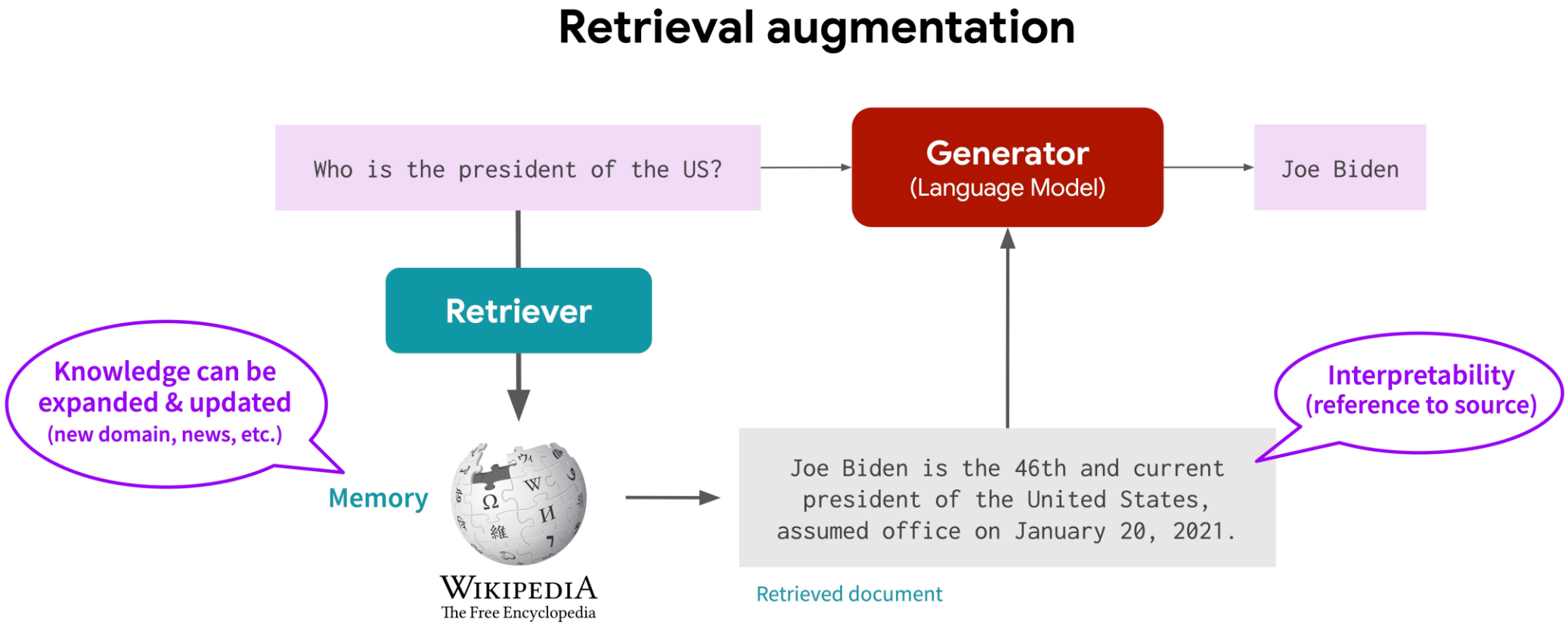


Figure 1: Visual illustration of Medprompt components and additive contributions to performance on the MedQA benchmark. Prompting strategy combines kNN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling.

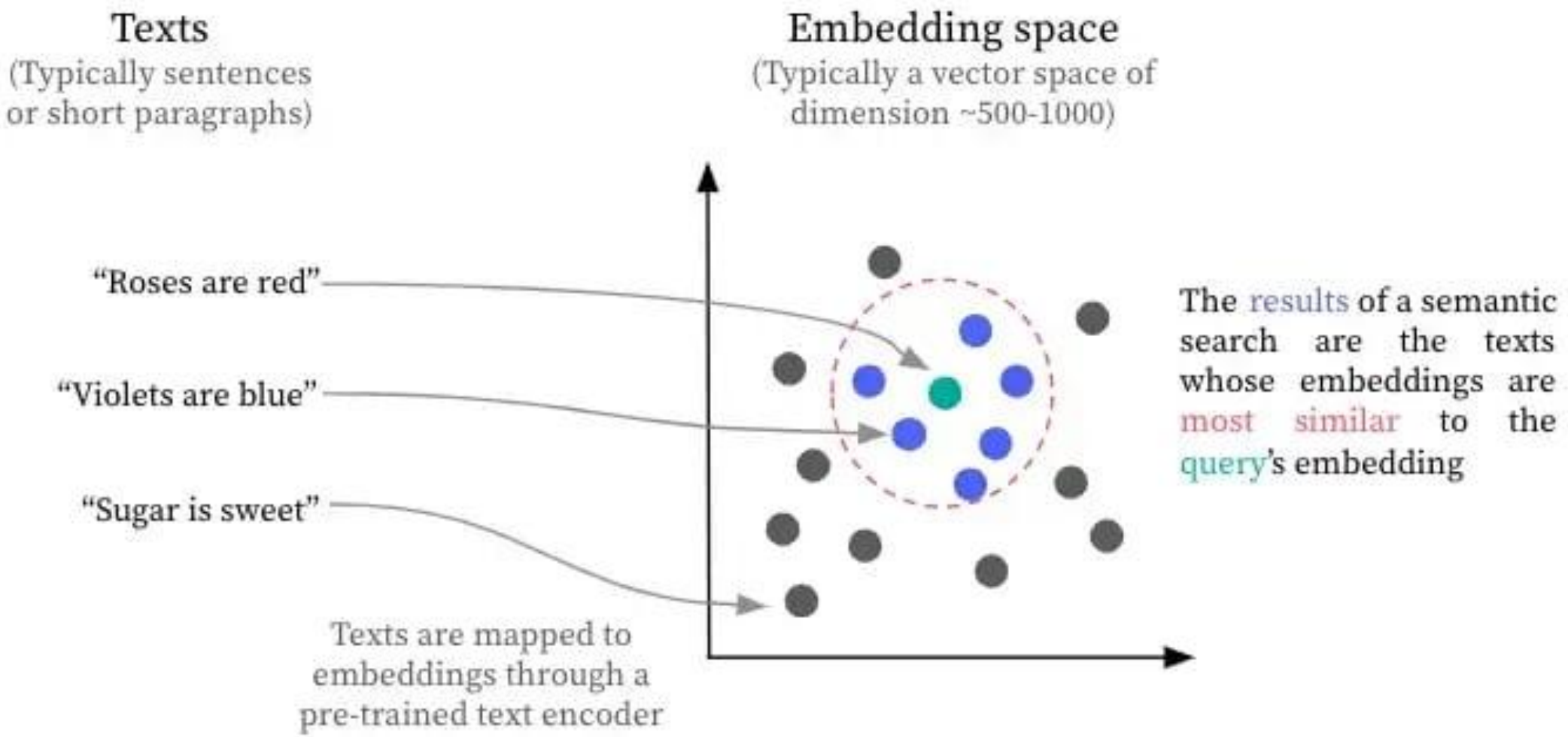
https://www.microsoft.com/en-us/research/blog/the-power-of-prompting/

Retrieval Augmented Generation



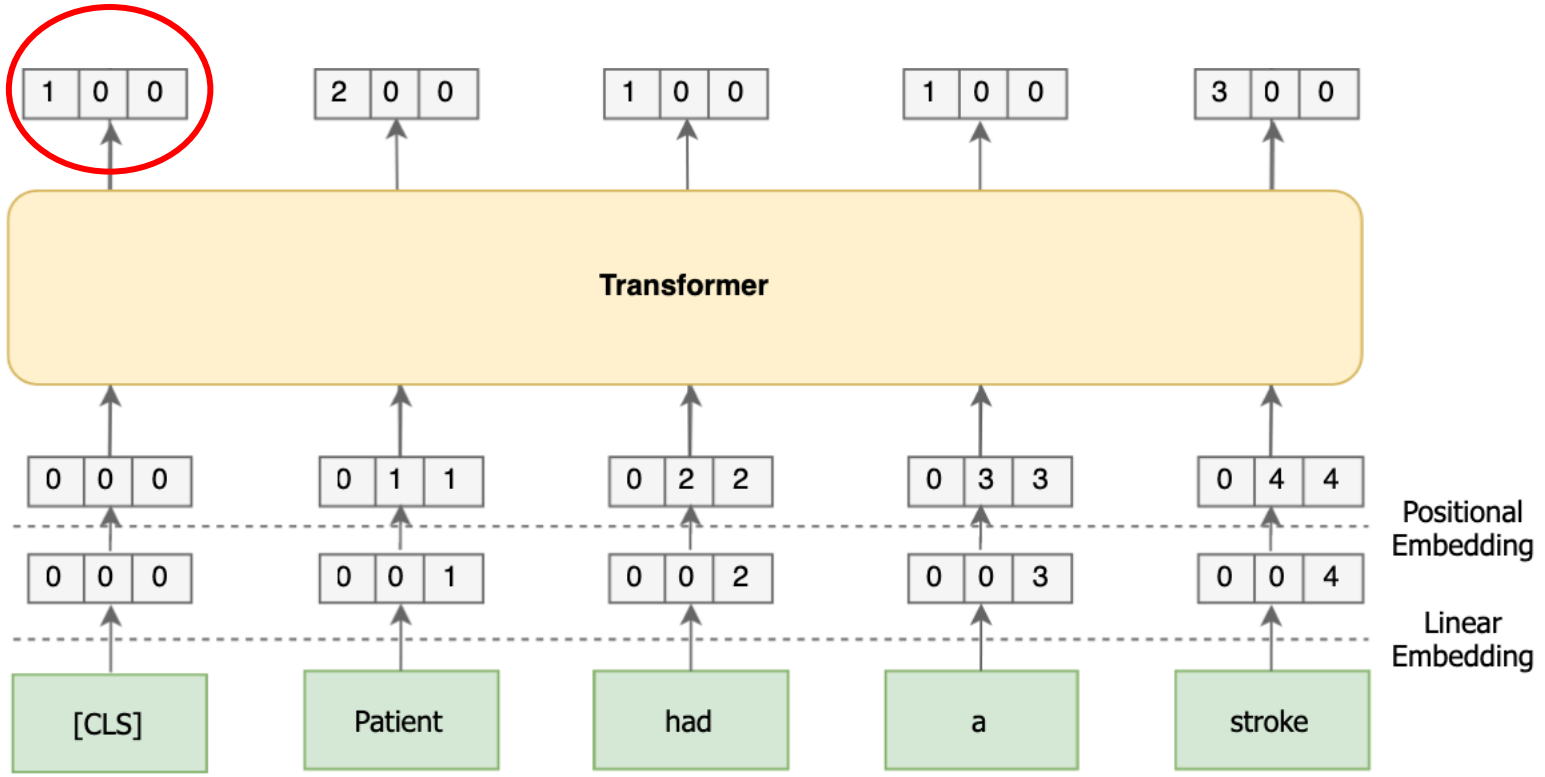
https://cs.stanford.edu/~myasu/blog/racm3/

Semantic Search

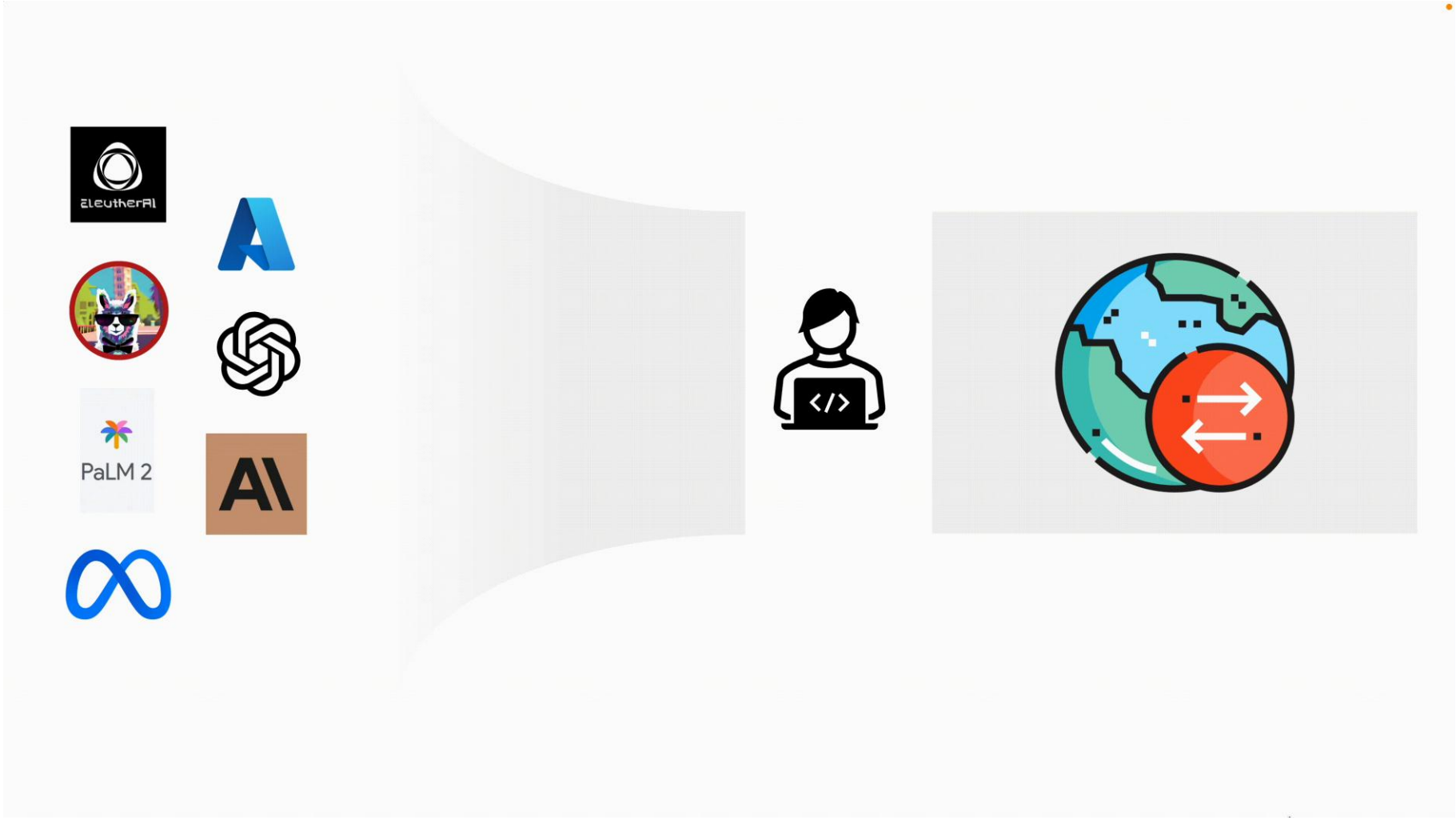


https://blog.dataiku.com/semantic-search-an-overlooked-nlp-superpower

Text -> Vector Embeddings



Agents



https://gorilla.cs.berkeley.edu/

Recap

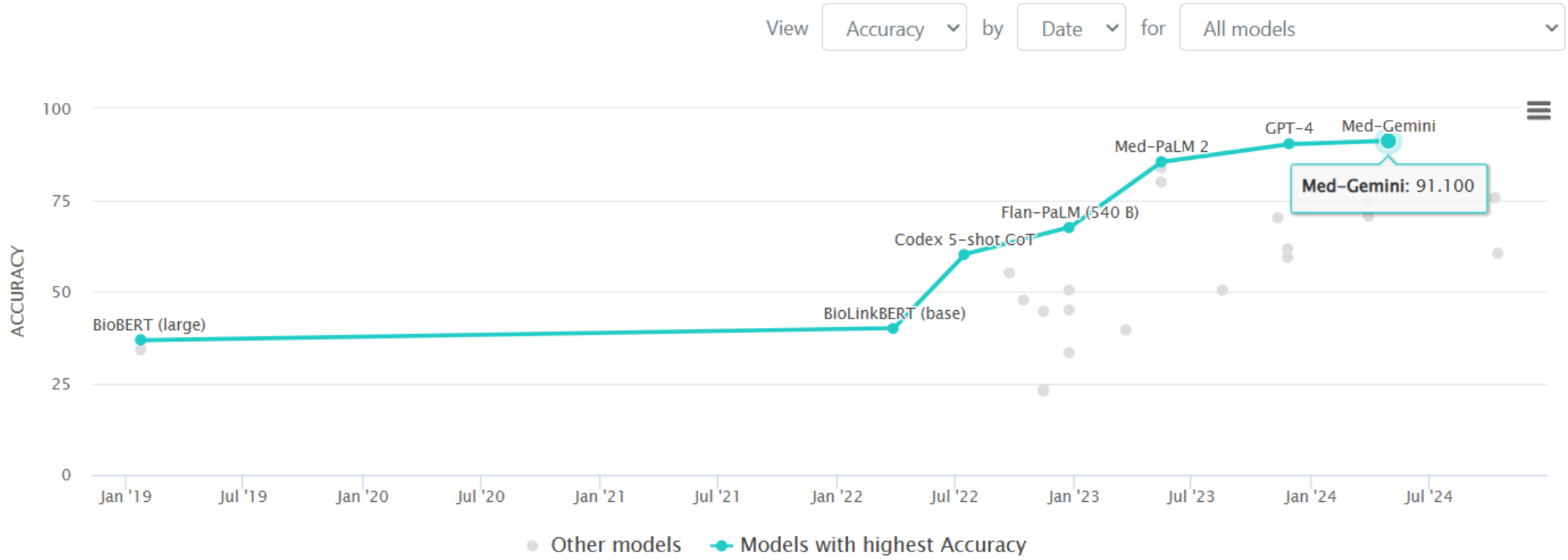
- What Model?
  - Closed vs open source “frontier models”
  - 3rd party vs local hosting
- Prompt engineering
  - An ‘art not a science’
  - Can make an annoyingly big difference…
- Add own data using RAG


QUESTION TIME

What does this mean for medicine?

Question Answering on MedQA

Leaderboard Dataset



https://paperswithcode.com/sota/question-answering-on-medqa-usmle

Automated Diagnoses

Read the case below and answer the question provided after the case.
Format your response in markdown syntax to create paragraphs and bullet points. Use '

' to start a new paragraph. Each paragraph should be 100 words or less. Use bullet points to list multiple options. Use '
* ' to start a new bullet point. Emphasize important phrases like headlines. Use '**' right before and right after a phrase to emphasize it. There must be NO space in between '**' and the phrase you try to emphasize.
Case: [Case Text]
Question (suggested initial question is "What are the top 10 most likely diagnoses and why (be precise)?"): [Question]
Answer:

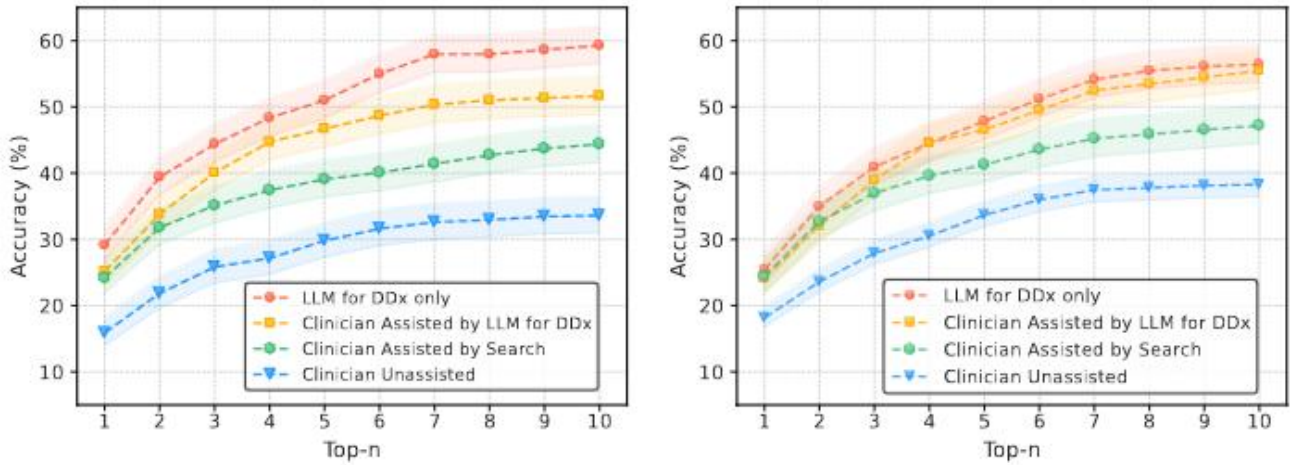


Figure 5 | Top-n Accuracy. (left) The percentage of DDx lists with the final diagnosis through human evaluation. (right) The percentage of DDx lists with the final diagnosis through automated evaluation.

McDuff, Daniel, et al. "Towards accurate differential diagnosis with large language models." *arXiv preprint arXiv:2312.00164* (2023).

Predictive Modelling

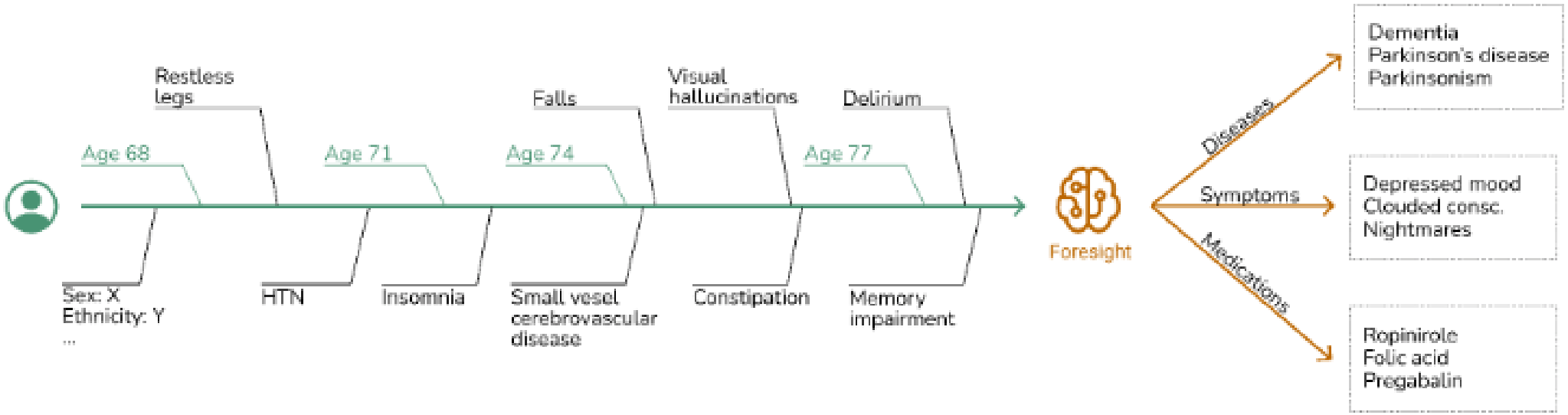


Figure 2. The left portion of the timeline represents the existing/historical data for a patient and the right portion are forecasts from Foresight for different biomedical concept types.

Kraljevic, Zeljko, et al. "Foresight--Deep Generative Modelling of Patient Timelines using Electronic Health Records." *arXiv preprint arXiv:2212.08072* (2022).

Automating Bureaucracy

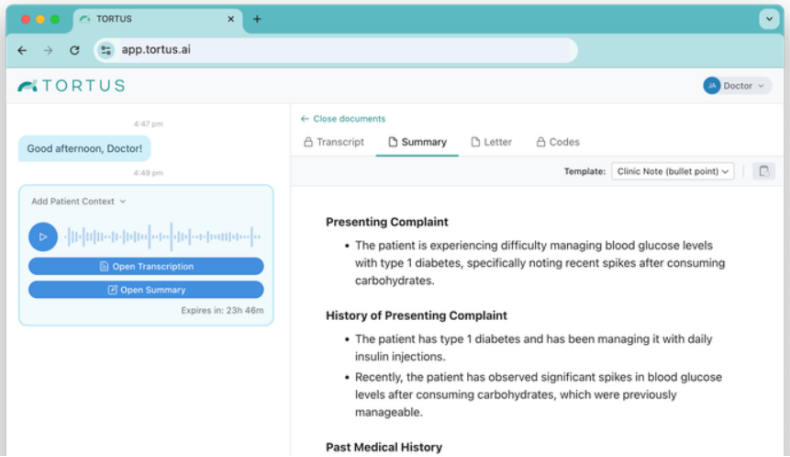
GOSH pilots AI tool to give clinicians more quality-time with patients

11 Nov 2024, 7 a.m.

A hands-free approach to drafting clinic notes

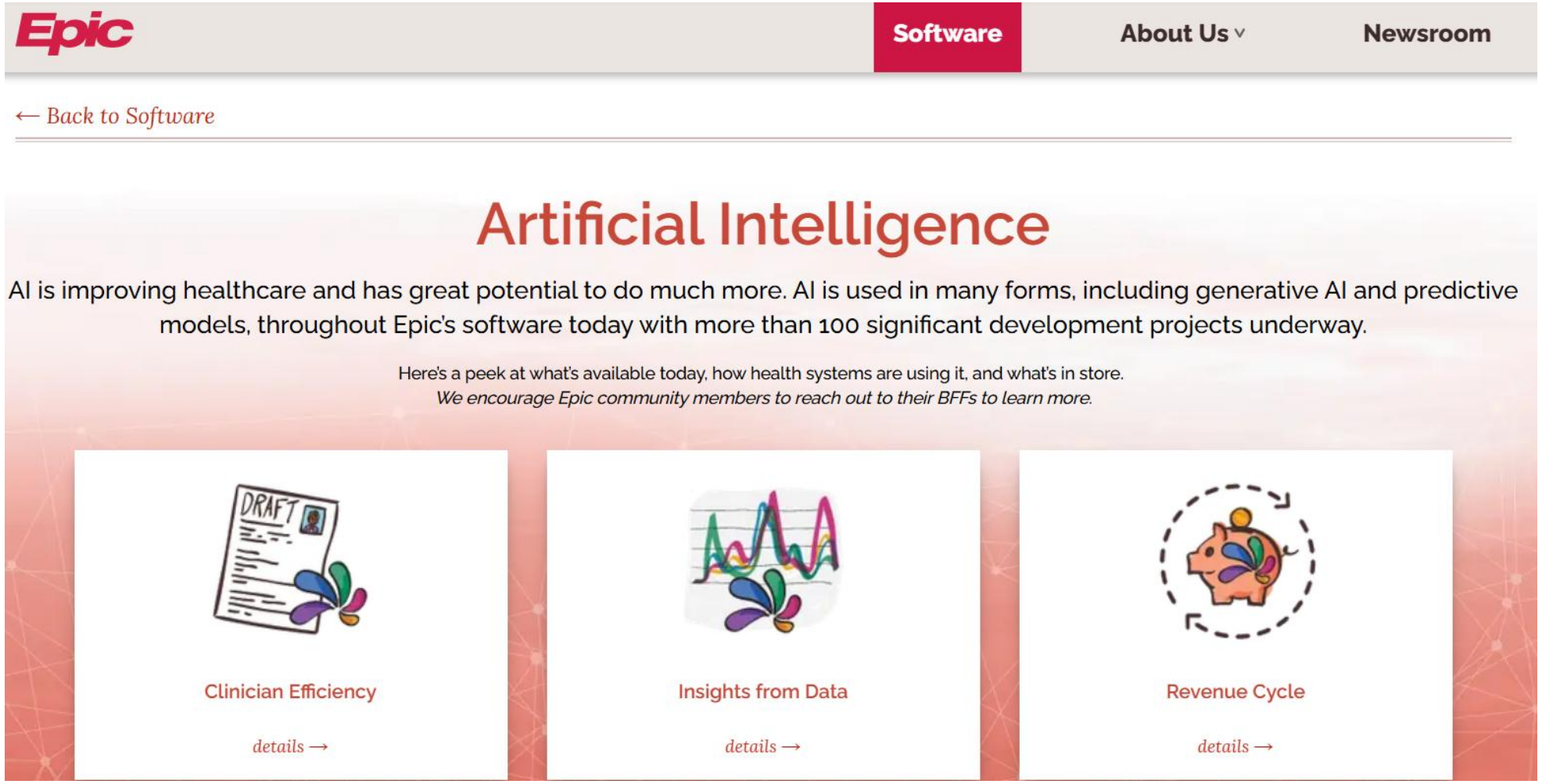
Typically, during outpatient consultations with patients, clinicians will dictate or manually type notes into the computer and compose letters during and after the consultation. The TORTUS assistant automates this process, using ambient voice technology with generative AI to listen to the consultation and draft a clinic note and letter, which are then edited and authorised by the clinician before being uploaded to the electronic health record system and sent to patients and their families.

The technology means clinicians can take a more 'hands-free' approach to drafting notes and follow-up letters as TORTUS does it for them, leaving them able to spend more time on direct patient care during outpatient appointments. In early testing in simulated clinics, all clinicians agreed that the AI helped them give their full attention to their patients when using the tool, without decreasing the quality of the clinic note or letter.



https://www.gosh.nhs.uk/news/gosh-pilots-ai-tool-to-give-clinicians-more-quality-time-with-patients/

Automating Bureaucracy



https://www.epic.com/software/ai/

Research Proof of Concept -> Bedside

- Data governance
- Compute requirements
- Regulation
- Robust evaluation
- How to show ROI -> health economics/patient outcomes

Recap

- Medical use cases
  - Benchmarks being saturated
  - Automating bureaucracy ‘low hanging fruit’
    - Commercialised ASAP
- Barriers
  - Robust evaluation
  - Getting hospitals ‘AI-ready’
  - Regulation?

Buzzwords

- LLM
- GPT
- Transformer
- Next token prediction
- Few shot prompting
- Chain of Thought
- Self-Consistency
- Finetuning
- Multi Modal
- Foundation Models
- RLHF
- Scaling Laws
- Open Source
- RAG
- Semantic Search
- Embeddings
- Agents
- Reasoning Model

I just want to know more about…

- How transformers work
  - Slightly more technical LLM overview talk
  - Great blog post
  - Original paper
  - Code GPT from scratch
- How to make an LLMs-based “thing”
  - Play around in OpenAI/Azure Playgrounds
  - Popular coding framework
  - Popular RAG framework
- LLMs and Medicine
  - Links to all papers at bottom of slides
  - Recent narrative review


QUESTION TIME