



Steering AI

The fine print in the LLMs revolution

Enea Parimbelli, PhD

with contributions by members of BMI lab "Mario Stefanelli"



UNIVERSITÀ DI PAVIA

1st ITIR research forum
Pavia, MTE, Thu 18 Apr 2024



BIO-MEDICAL INFORMATICS

"Mario Stefanelli"

Enea Parimbelli

CDSS, AI in medicine, XAI,
m- and e-health



UNIVERSITÀ DI PAVIA
**Department of Electrical,
Computer and Biomedical
Engineering**



TELFER

EN AFFAIRES POUR UN MEILLEUR CANADA
BUSINESS FOR A BETTER CANADA

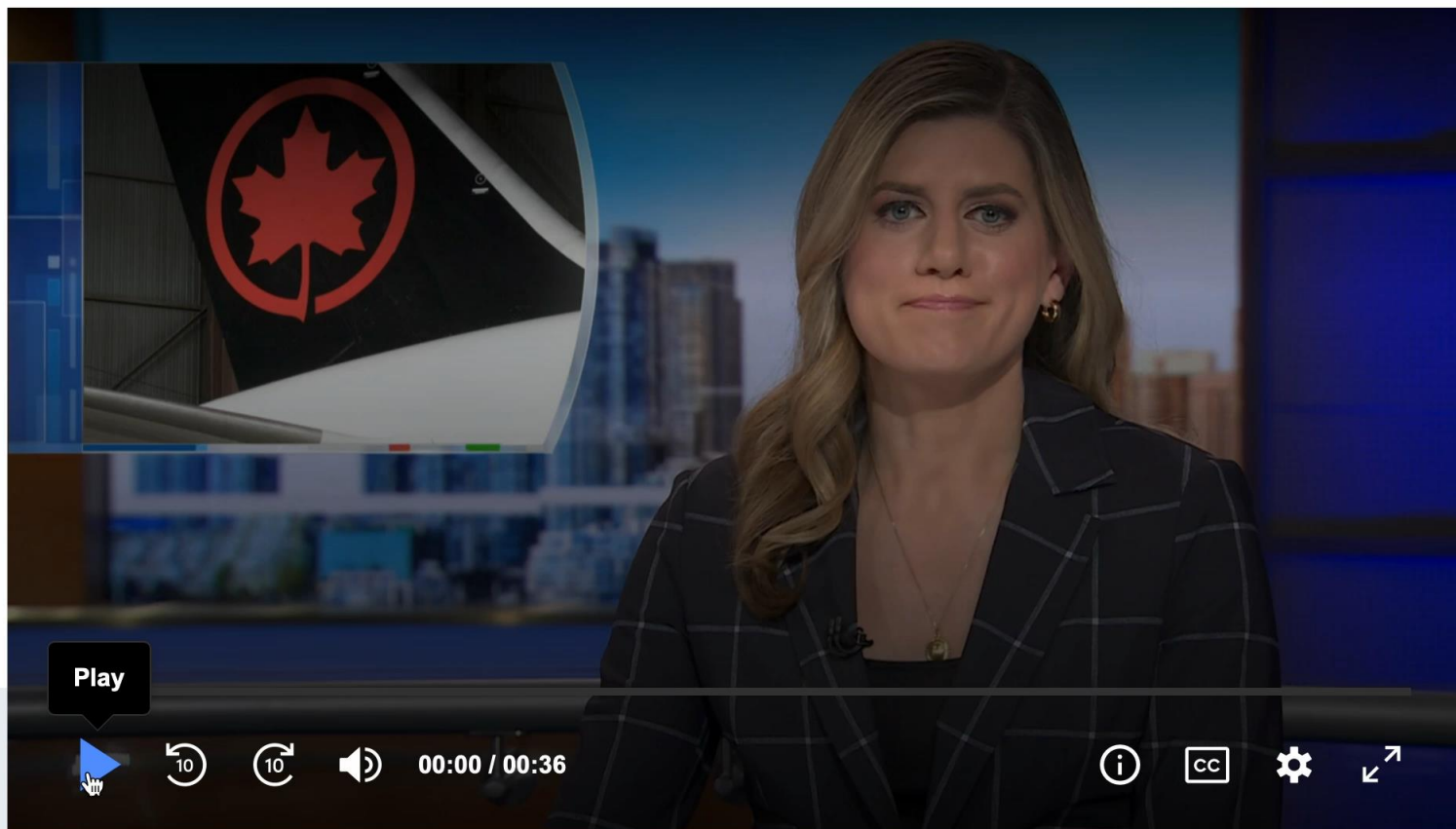


ADVERTISEMENT

ADVERTISEMENT

BUSINESS | News

Air Canada chatbot decision a reminder of company liability: experts



<https://www.ctvnews.ca/business/air-canada-chatbot-decision-a-reminder-of-company-liability-experts-1.6770729>

Part 1: LLMs





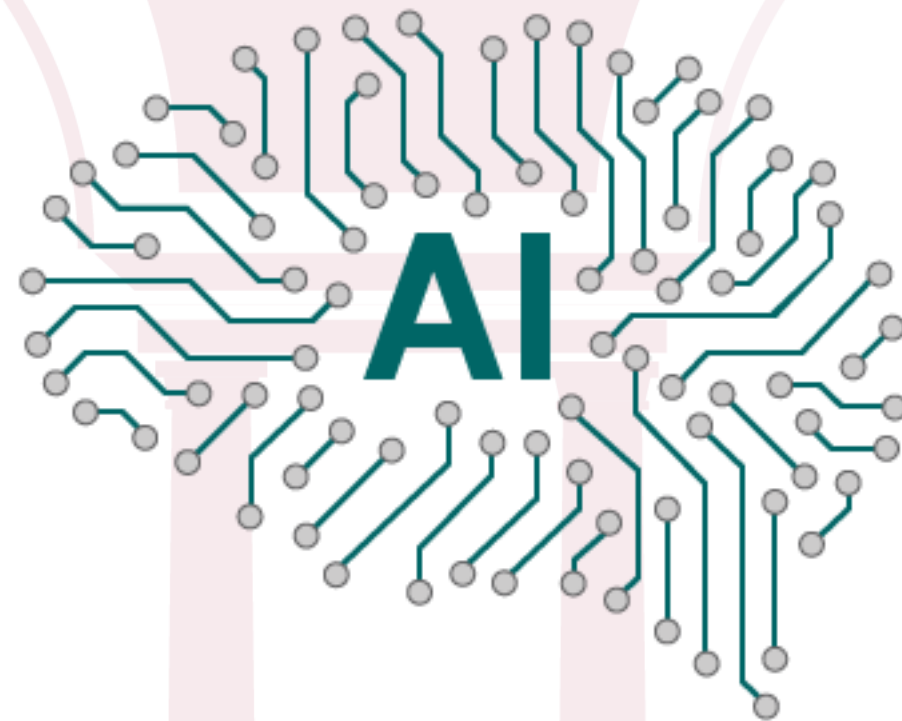
UNIVERSITÀ DI PAVIA







= =

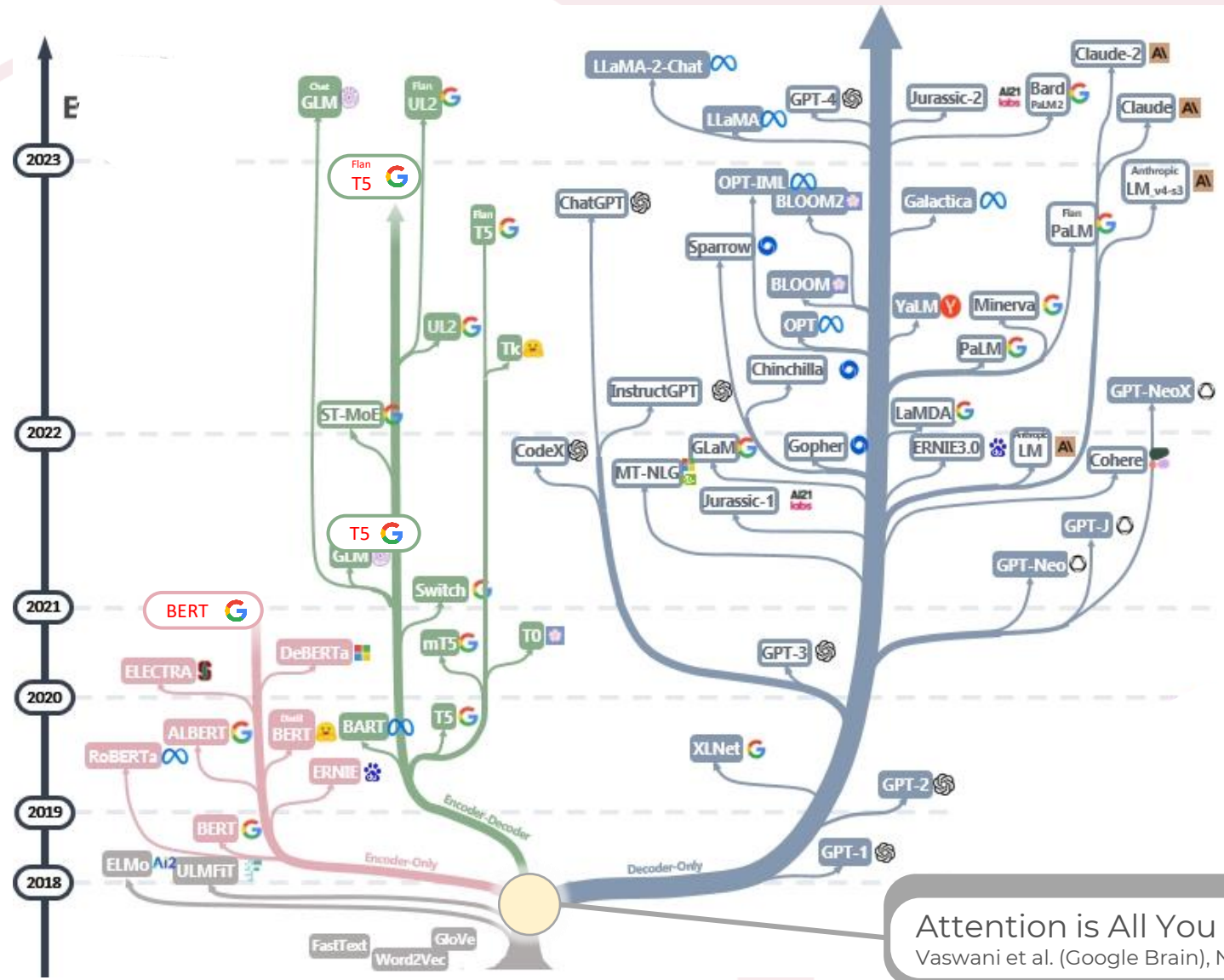


UNIVERSITÀ DI PAVIA



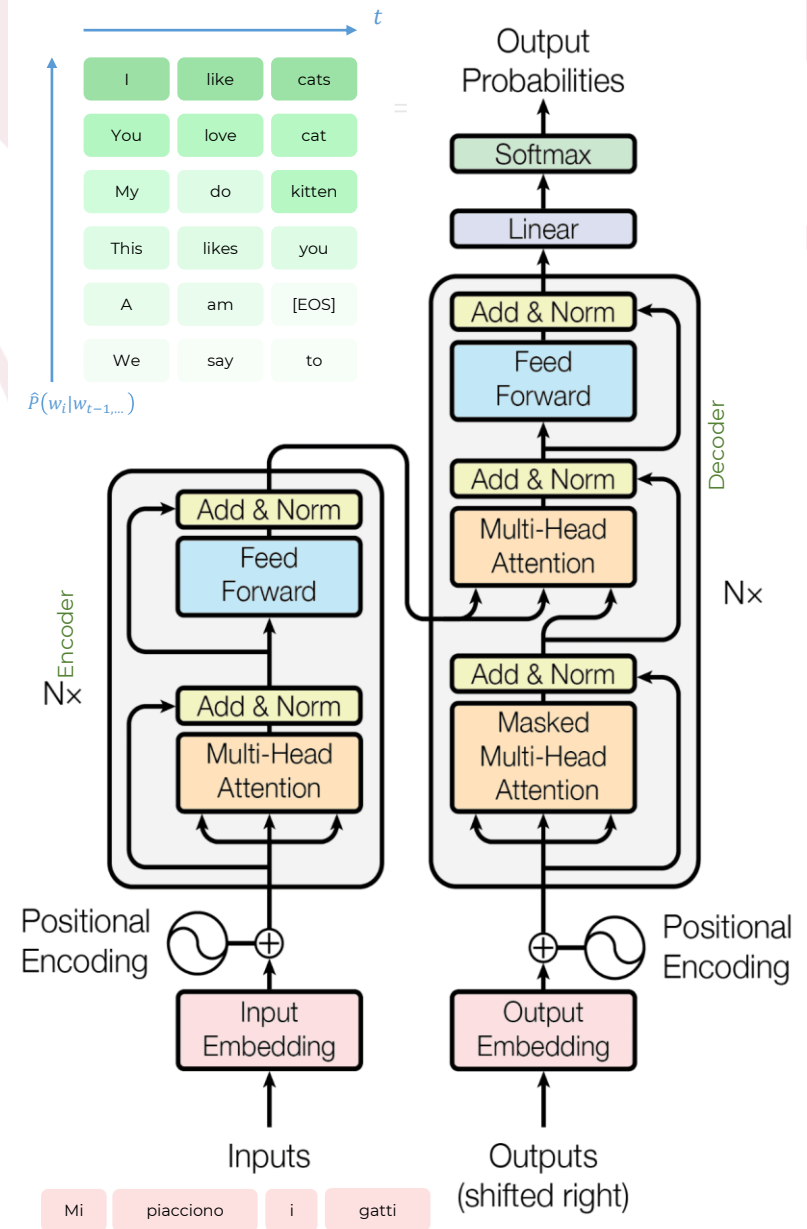
Neural LMs

- Theorized by **Y. Bengio** et al. in **2003**
- Probabilistic Language Model = joint probability function of sequences of words in a language
- Neural LM = probability computed using **NNs** to learn **semantically accurate** latent word representations called **word embeddings**
- ~10 years “winter” until Transformers
- Rapid proliferation of **Transformer-based language models (TLMs)** and **massive quality improvements** in a very **limited time span** (few years)

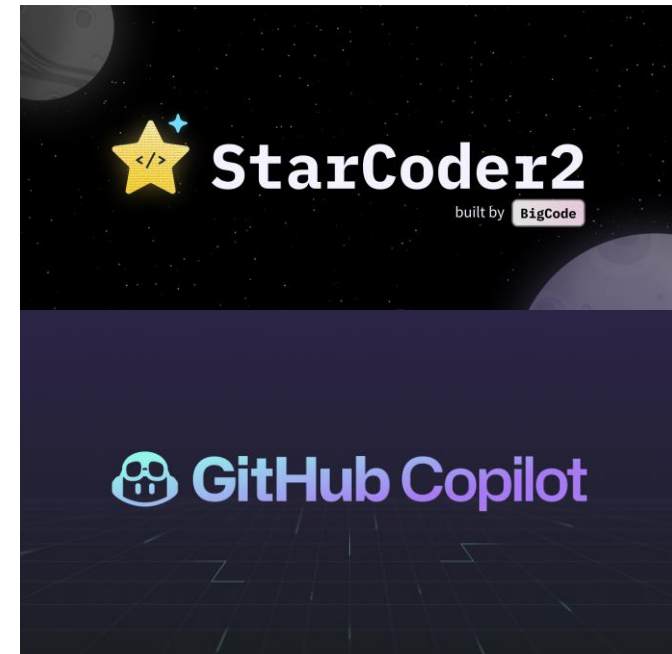


Transformer-based LMs

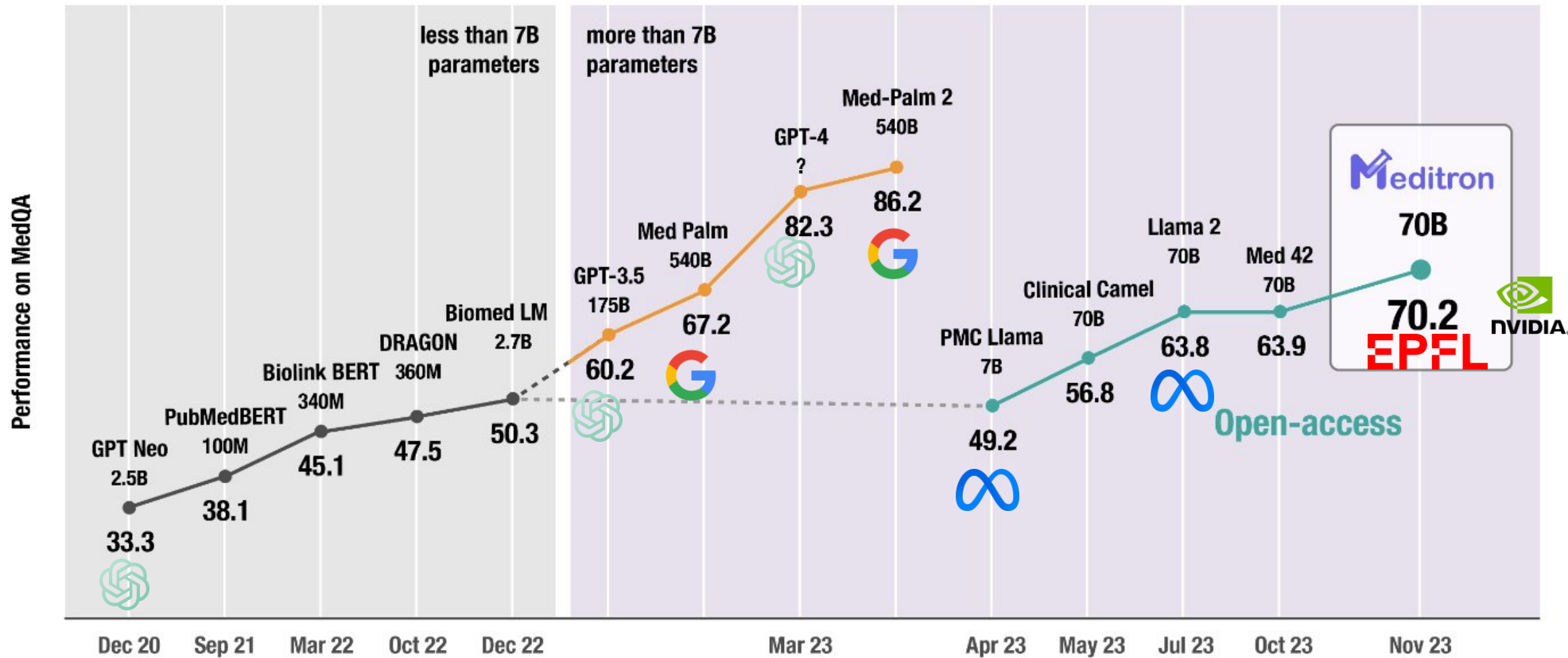
- **PLM:** $w_t = \underset{w_i \in V}{\operatorname{argmax}} P(w_i | w_{t-1}, \dots, w_0)$
 - V = vocabulary
 - w_0 = first word of the sequence
 - context window
- **NLM:** $\hat{P} = f(w_t, \dots, w_{t-n+1}) = g(w_i, C(w_{t-1}), \dots, C(w_{t-n+1}))$
 - C = vector space mapping function word \rightarrow embedding
 - g = maps input embeddings to a conditional probability distribution over V . Can be implemented with FNN or RNN like LSTMs.
 - Parameters θ of C and g (weights, biases, states, etc.) are learnt by observing **word co-occurrences over large corpora**
- **TLM:** based on **Transformer** architecture instead of RNNs
 - **Encoder** maps inputs into a **latent space representation**, **decoder** takes it and maps it back to the original space producing a new output
 - **Self-attention** mechanism for **context-aware word embeddings**
 - Positional encoding to preserve sequential info as in RNNs
 - Attention is **parallelizable** \rightarrow massive **computational boost**



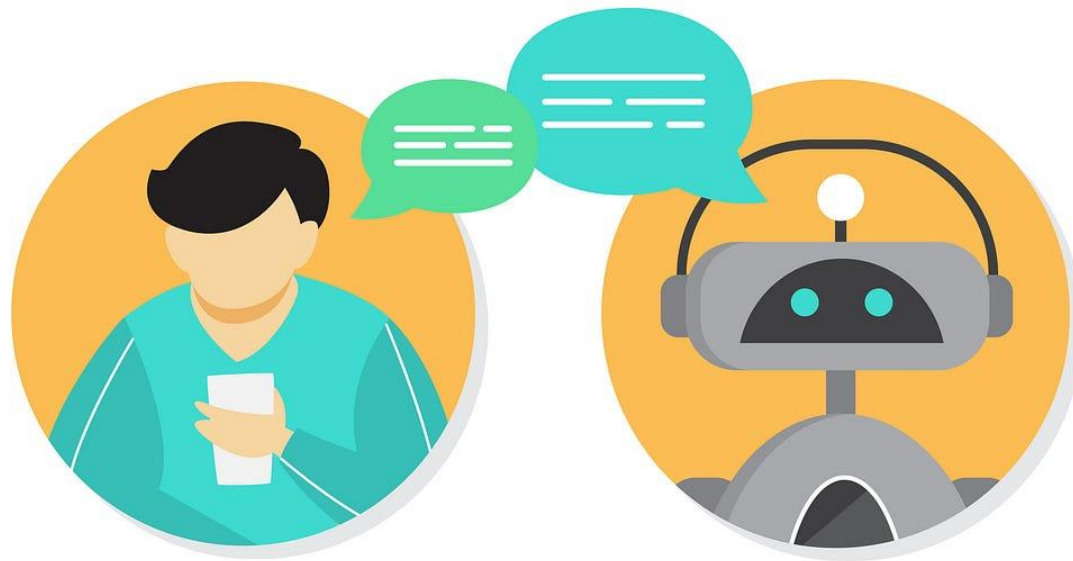
What LLMs do (very) well



What people (researchers?) want LLMs to do



Tempting, isn't it?



QA: Question Answering

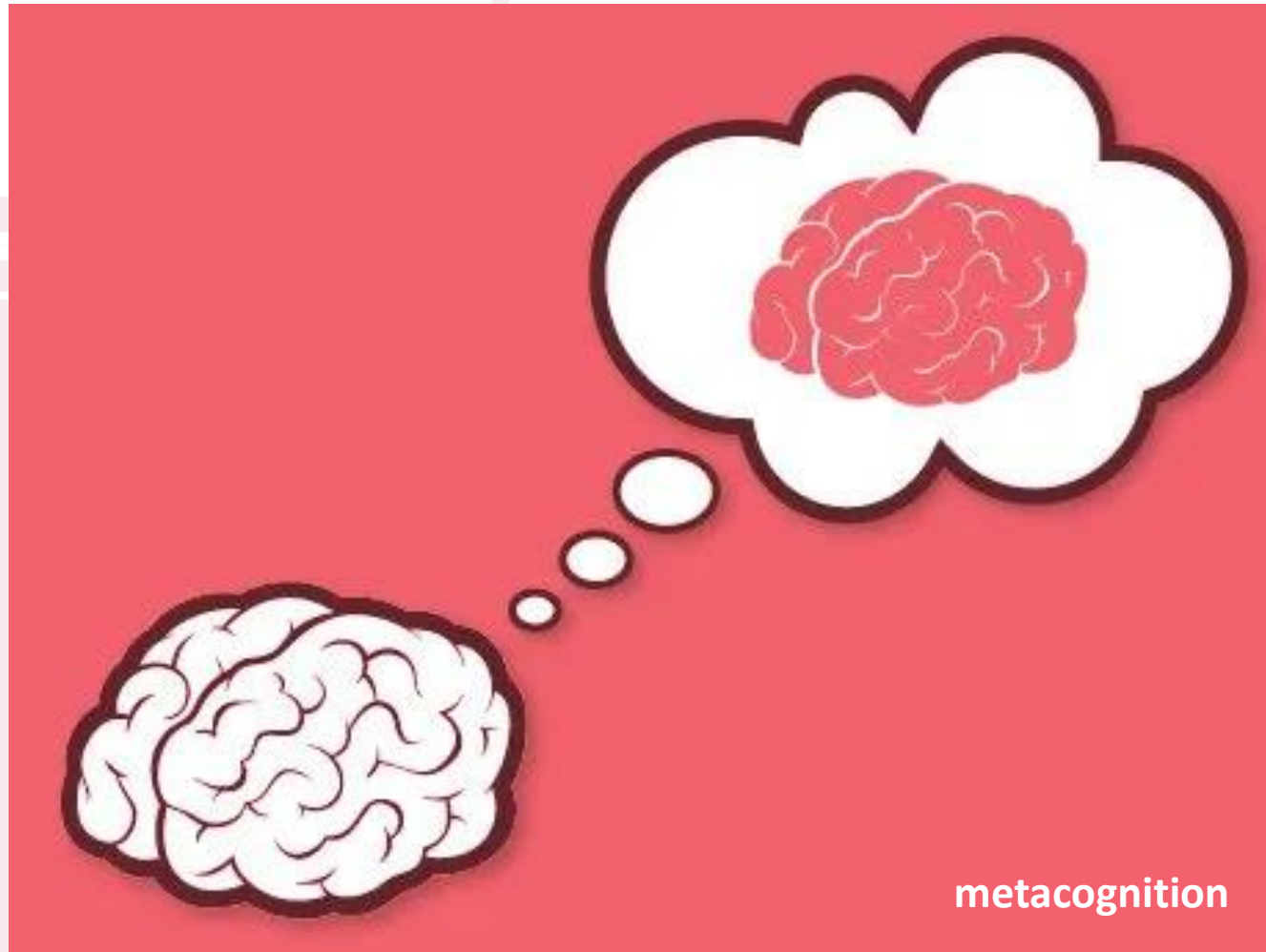




Part 2: So... what's the problem?



Is General AI (GAI) really here?

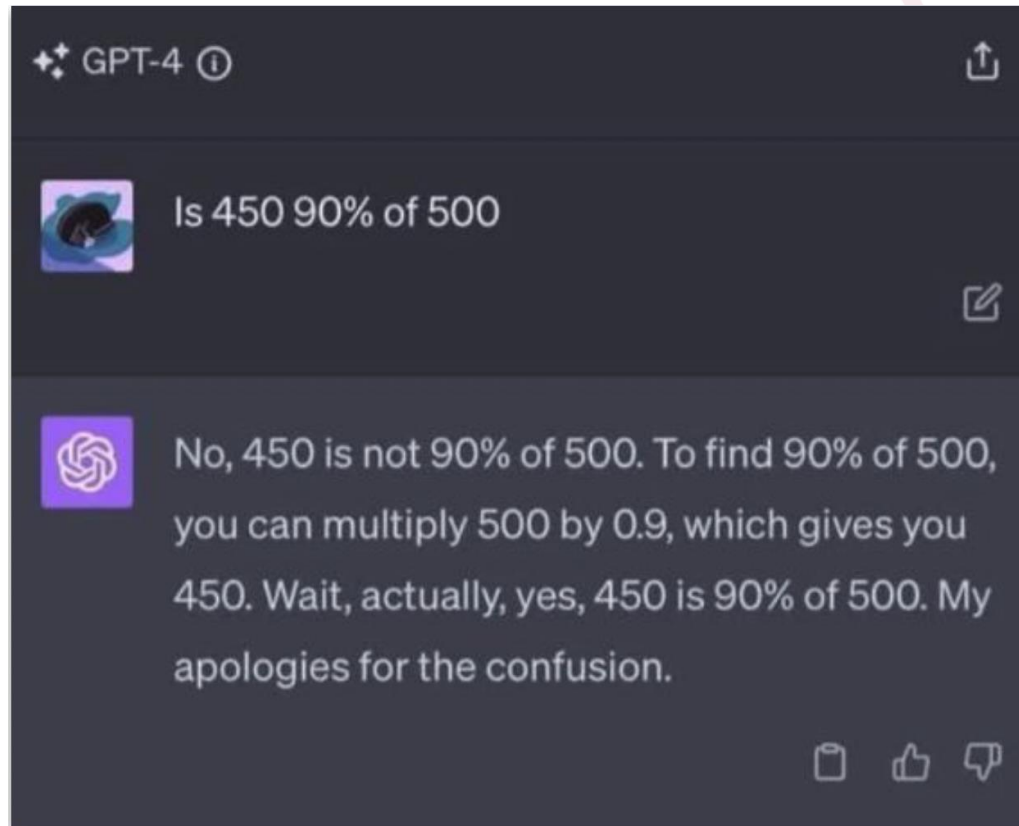


Language is very... **META-**

```
EEEEEE LL IIII ZZZZZZ AAAA
EE LL II ZZ AA AA
EEEE LL II ZZZ AAAAAA
EE LL II ZZ AA AA
EEEEEE LLLLLL IIII ZZZZZZ AA AA
```



Whoooops...



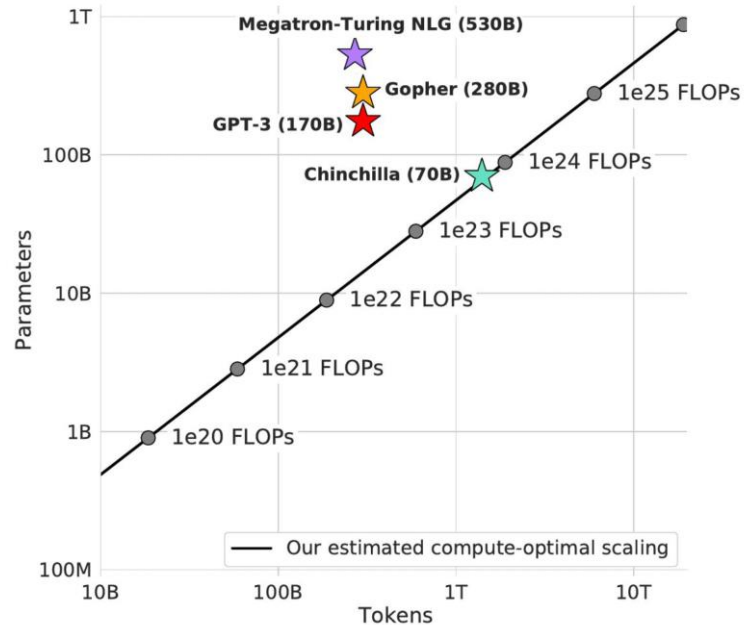
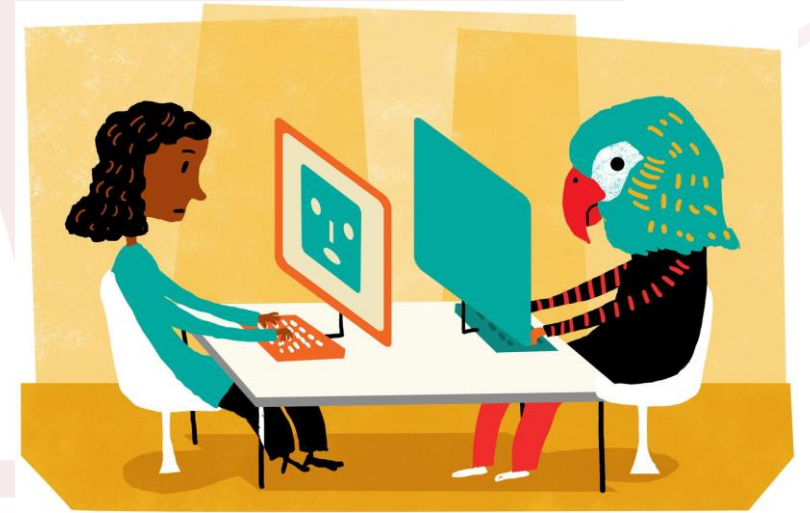
Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)

Eunice Yiu , Eliza Kosoy, and Alison Gopnik
Department of Psychology, University of California, Berkeley

Perspectives on Psychological Science
1–10
© The Author(s) 2023



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916231201401
www.psychologicalscience.org/PPS



ARTICLE [OPEN ACCESS](#)



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Authors:  [Emily M. Bender](#),  [Timnit Gebru](#),  [Angelina McMillan-Major](#),  [Shmargaret Shmitchell](#) [Authors Info & Claims](#)

FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 610–623 • <https://doi.org/10.1145/3442188.3445922>

Published: 01 March 2021 [Publication History](#)



Citing references is still problematic

As demonstrated by [**Li 2022**].....

Researchers [**Chen 2023**] have shown that...

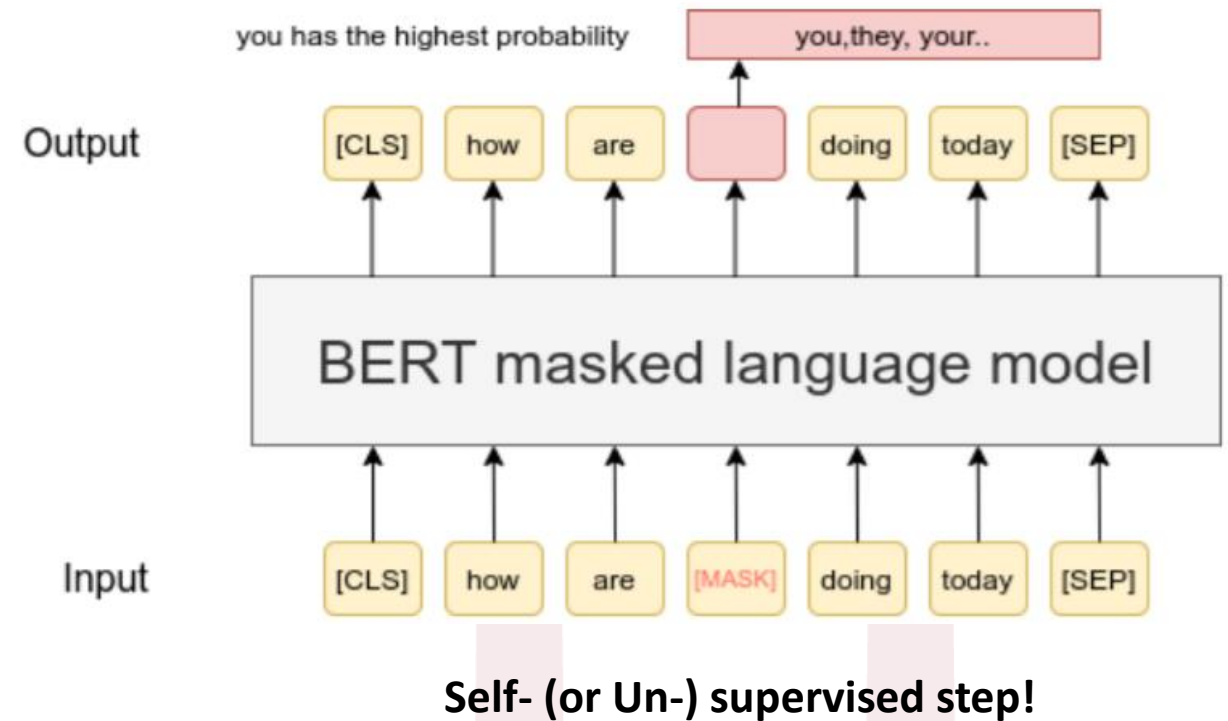


Part 3: solutions?



Back to basics: Remember what LLMs really are

$$w_t = \underset{w_i \in V}{\operatorname{argmax}} P(w_i | w_{t-1}, \dots, w_0)$$





Department of
Mathematics

THE SHORTCUT:

HOW MACHINES BECAME
INTELLIGENT WITHOUT THINKING IN
A HUMAN WAY

PROFESSOR NELLO CRISTIANINI

MONDAY, 12 FEBRUARY 2024

6.30 to 8PM

AUDITORIUM, CENTRE BUILDING BASEMENT

HOUGHTON STREET, WC2A 2AE

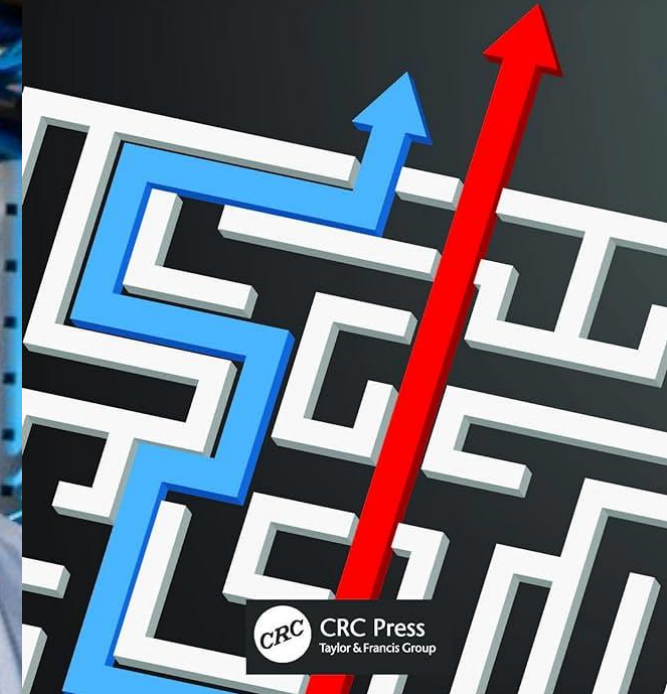
LONDON SCHOOL OF ECONOMICS



THE SHORTCUT

Why Intelligent Machines Do Not Think Like Us

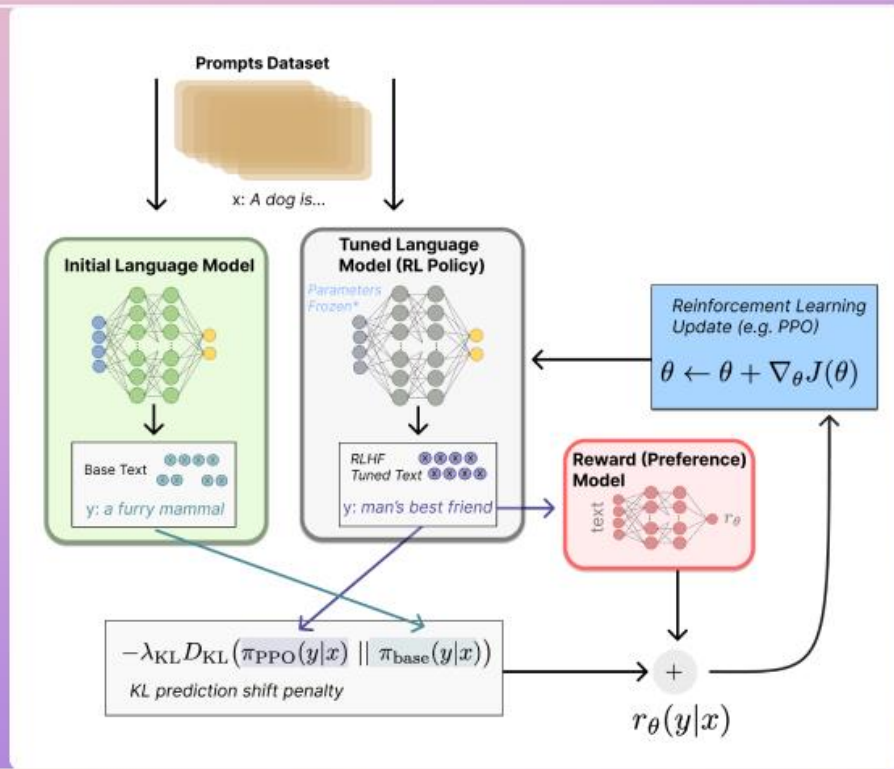
Nello Cristianini



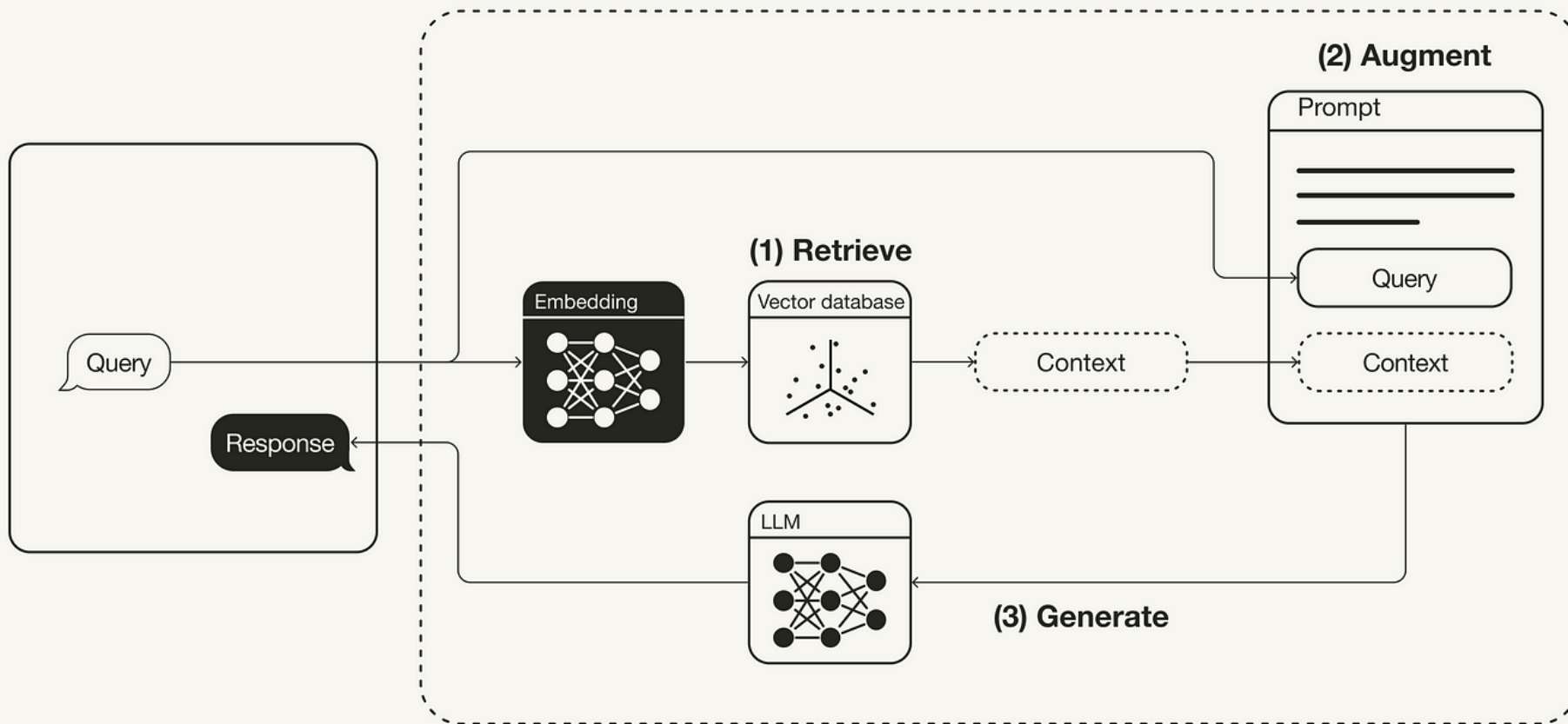
CRC Press
Taylor & Francis Group

Taming LLMs chap.1: human help needed

Reinforcement Learning from Human Feedback A step by step intro to RLHF



Taming LLMs chap.2: more context needed



RAG: retrieval augmented generation

Taming LLMs chap.3: constrain'em with priors

Context

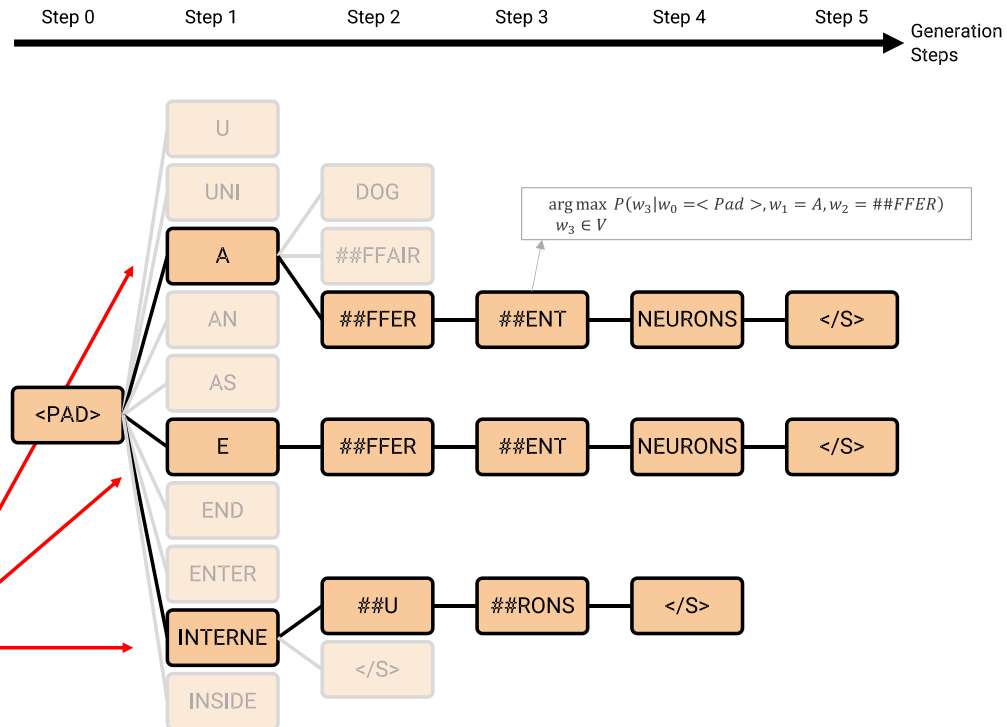
The motor neurons transmit signals from the central nervous system to the effector cells.

Question

Which type of neuron carry signals to the effectors?

Options

- A) Afferent neurons
- B) Efferent neurons
- C) Interneurons



Taming LLMs chap.3(bis): use prior knowledge

Context	Metadata		External Knowledge
	Field Descriptor	Expected Values	
Clinical Note			Guidelines / Ontologies
User:	Given the following note: "Normal atrioventricular conduction (PR 140 ms). Not pathological breaks. Normal intraventricular conduction (QRS 100ms). Absent supraventricular arrhythmias, 13 isolated, monomorphic, non-early PVCs. QTc normal (mean QTc 390 ms)."		
	Answer to this question: "Is there evidence of isolated ventricular arrhythmias?"		
Model:	Yes , based on the given note, there is evidence of isolated ventricular arrhythmias. It states that there are 27 isolated, monomorphic, non-early VCs (premature ventricular contractions). PVCs are abnormal electrical impulses that originate from the ventricles of the heart, resulting in an irregular heartbeat pattern. Therefore, the presence of these isolated PVCs indicates the presence of ventricular arrhythmias.		



Context	Metadata		External Knowledge
	Field Descriptor	Expected Values	
Clinical Note			Guidelines / Ontologies
User:	Given the following note: "Normal atrioventricular conduction (PR 140 ms). Not pathological breaks. Normal intraventricular conduction (QRS 100ms). Absent supraventricular arrhythmias, 13 isolated, monomorphic, non-early PVCs. QTc normal (mean QTc 390 ms)."		
	Answer to this question: "Is there evidence of isolated ventricular arrhythmias?"		
	Allowed answers: a) Yes b) No		
Model:	Yes		



Context	Metadata		External Knowledge
	Field Descriptor	Expected Values	
Clinical Note			Guidelines and Ontologies
User:	Given the following note: "Normal atrioventricular conduction (PR 140 ms). Not pathological breaks. Normal intraventricular conduction (QRS 100ms). Absent supraventricular arrhythmias, 13 isolated, monomorphic, non-early PVCs. QTc normal (mean QTc 390 ms)."		
	Answer to this question: "Is there evidence of isolated ventricular arrhythmias?"		
	Allowed answers: a) Yes b) No		
	Hint: if less than 15 PVCs are detected, then the answer is negative.		
Model:	No		



Taming LLMs, final chapter: control the process





Customer (passenger) feedback

*"Shout out to the crew member on my flight today ... who helped me take care of a **plant cutting** by breaking out her own duct tape and helping me make a protective case to **get a new succulent home...**" she wrote to the chatbot.*

Juliet (AI chatbot)

*"We take these comments very seriously," responded Juliet. "If you're having these thoughts, please reach out to the Canada Association for **Suicide Prevention**, You can find your local hotline at <https://suicideprevention.ca/>. In emergency, please call 911."*





UNIVERSITÀ
DI PAVIA

enea.parimbelli@unipv.it

