Oncodesign
Precision Medicine

Dr. LLM or: How I Learned to Stop Worrying and Love the AI

January 2024

Salvatore Raieli

Plan of the presentation

- An introduction to OPM
- From text to LLMs
- LLMs in medicine and biology

Plan of the presentation

- MSc in Artificial Intelligence, University of Leeds
- PhD, Institut Curie
- MSc in pharmaceutical biotechnology, University of Bologna
- Senior DS at OPM, working on AI applied to multi-omics, LLMs

OPM: a « pure player » biopharmaceutical **in the clinical stage**

Three innovative and proprietary technologies

Resistant and metastatic patients
ONCOSNIPER

Identify and validate new therapeutic targets to fight against treatment resistance in oncology

Kinase inhibitors
NANOCYCLIX®

Identify effective, highly specific kinase inhibitors

Radioconjugate Vector
PROMETHE

Developing Vectorized Internal Radiation Therapy for therapeutic use

Three molecules in clinical stage

- Parkinson's disease
- IBD
- Radiotracer in oncology

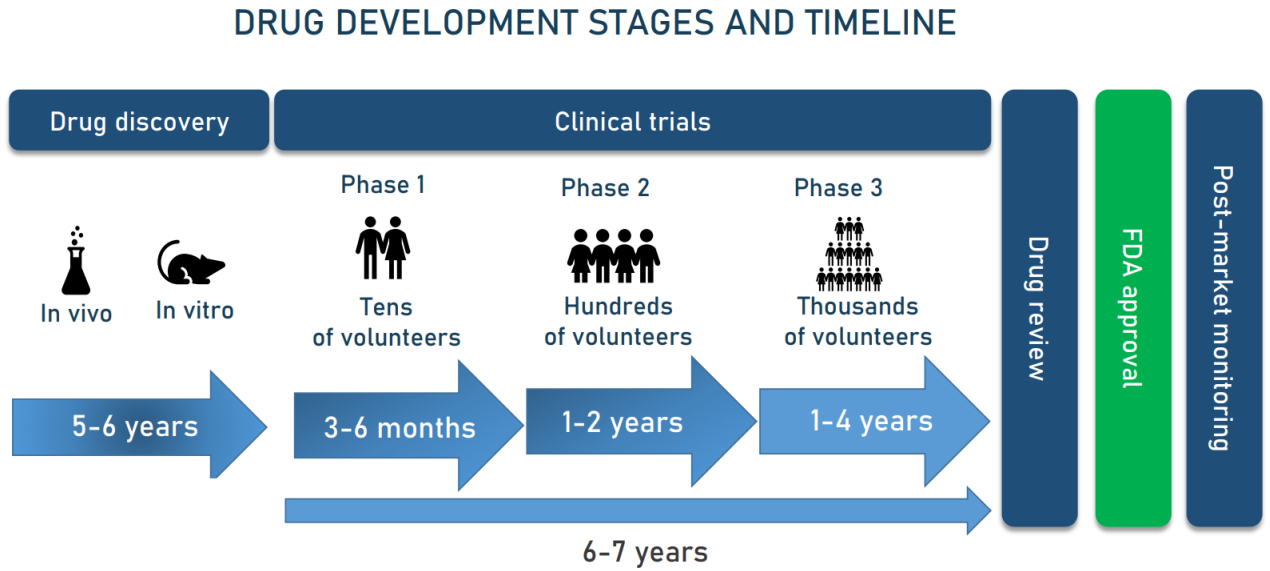
Structuring partnerships

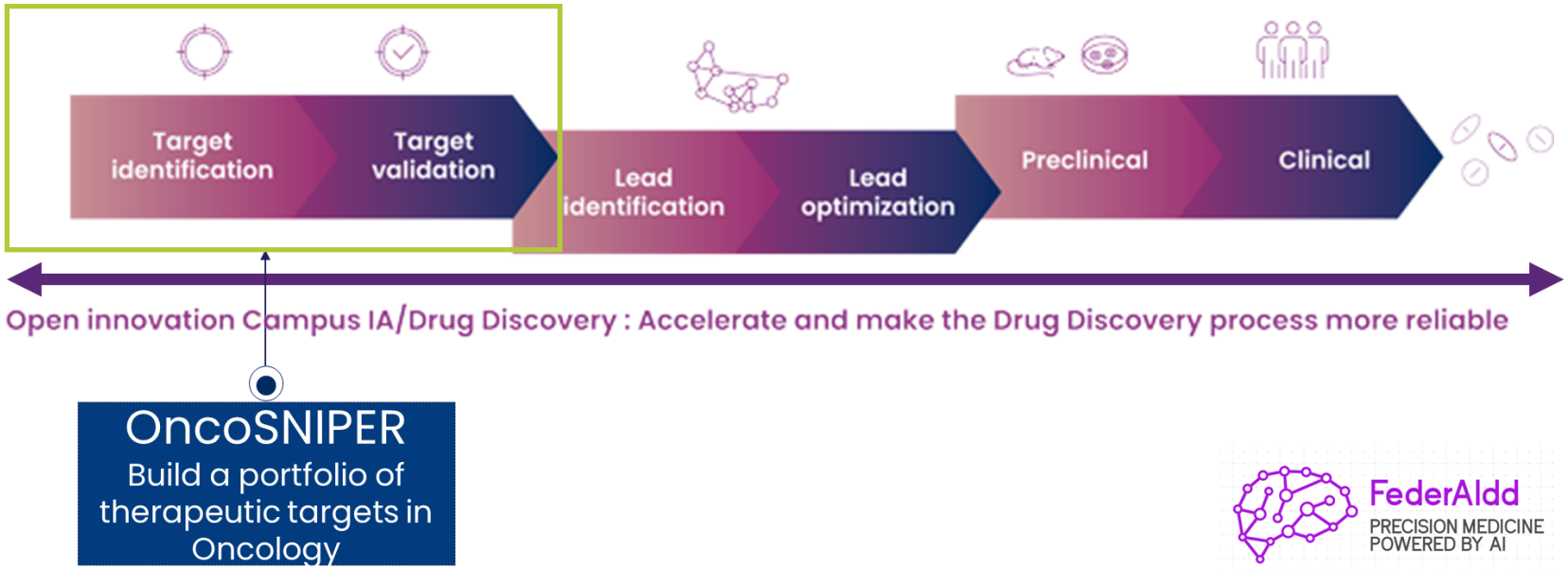
SERVIER moved by you

SEngine Precision Medicine

The need of AI in Drug discovery

- Drug discovery and clinical trials are expensive (up to 1 billion) and high-risk processes (90% of drug candidates in clinical trials fail)
- Time from target identification to molecule approval is increasing (10-15 years currently)
- Most of the candidates fail: a lack of clinical efficacy (40%–50%) or toxicity (30%)





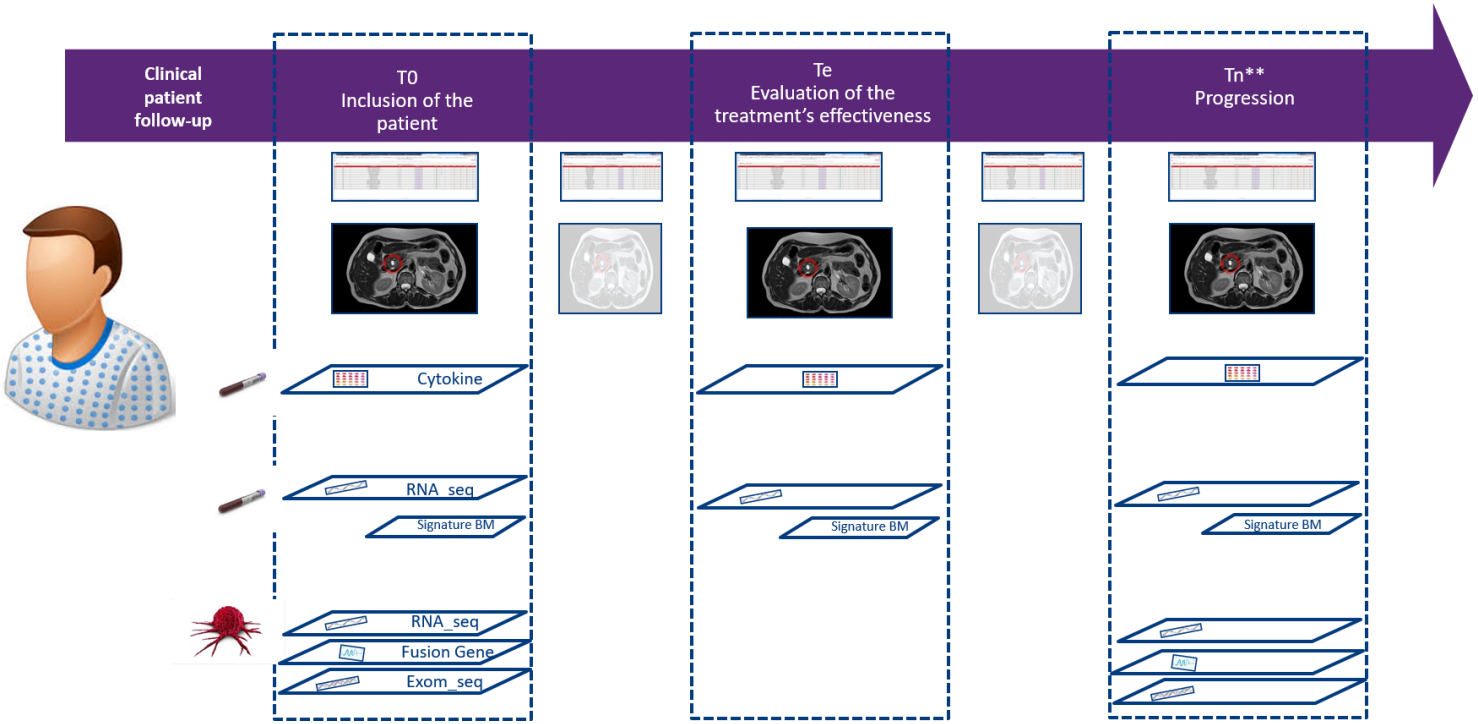
FederAIdd
PRECISION MEDICINE POWERED BY AI

Oncosnipe®

From identification and characterisation of patients resistant to treatments, to diagnostics and therapeutics tools development

OncoSNIPE® : NCT04548960, longitudinal clinical study

Identification of patient subpopulations resistant or unresponsive to treatment

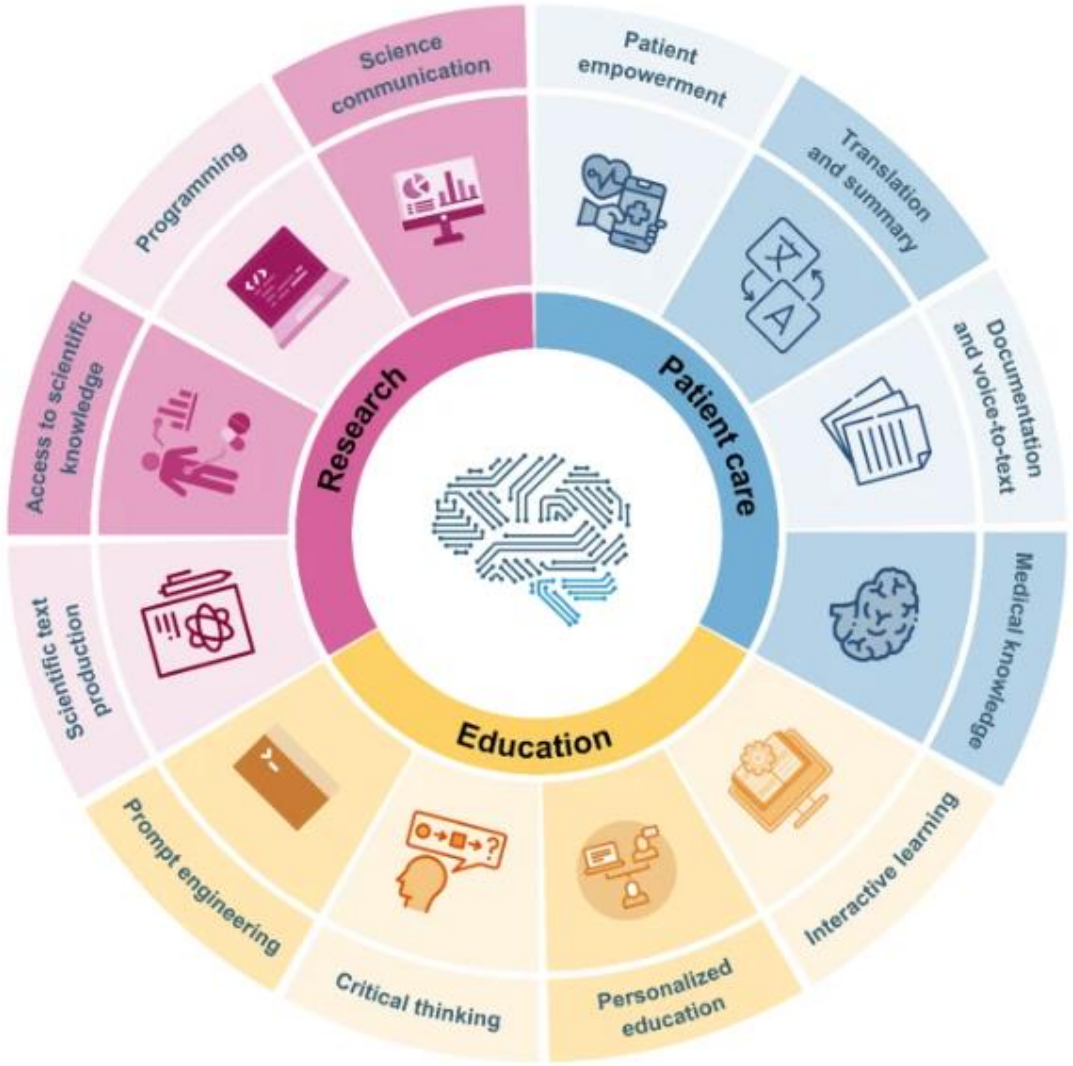


- **Collection of patient data**
  - Clinical data
  - Imaging data
  - Molecular data
- **Enrichment and Integration**
  - Data curation and integration
  - Semantic enrichment (EHR)
- **Longitudinal data analysis**
  - Machine learning
  - Homogenization of patient populations
  - Modeling of resistance mechanisms
  - Signature Biomarkers
- **Precision medicine**
  - Tool to help with diagnosis and research
  - Companion diagnostic tests and biomarker kits
  - Identification / validation of therapeutic targets

Why LLM in medicine?

Three main areas:

- Research
- education
- patient care

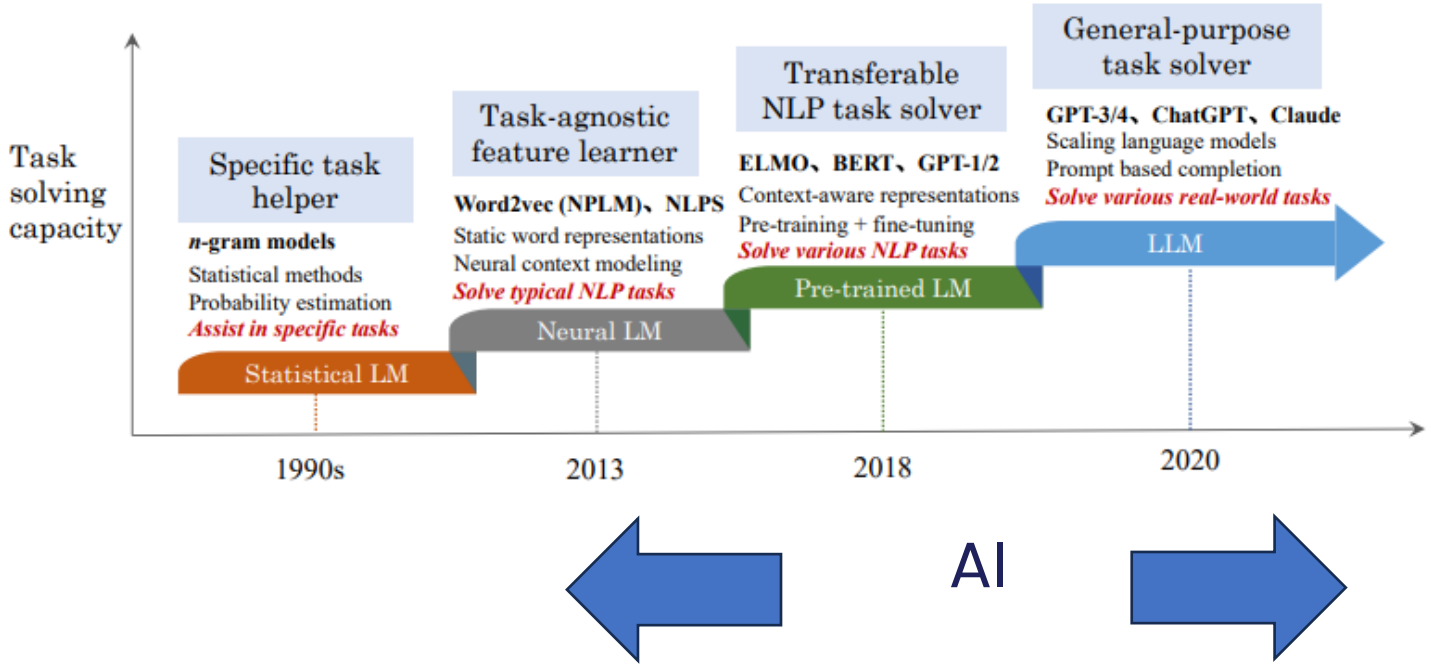


But what is a large language model?
How you build one?

How a computer analyze language?

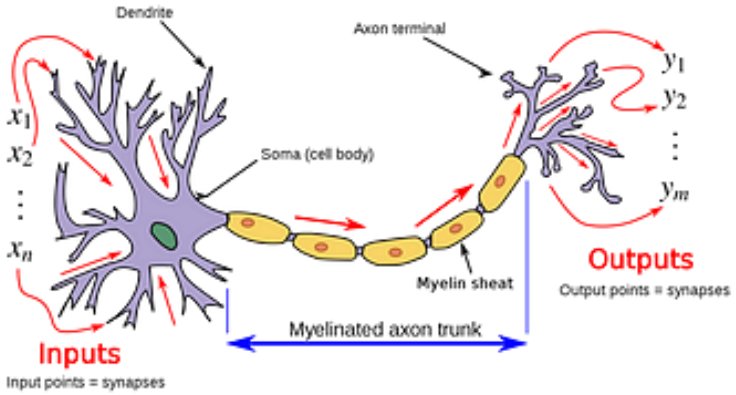
A different language is a different vision of life.
- Federico Fellini

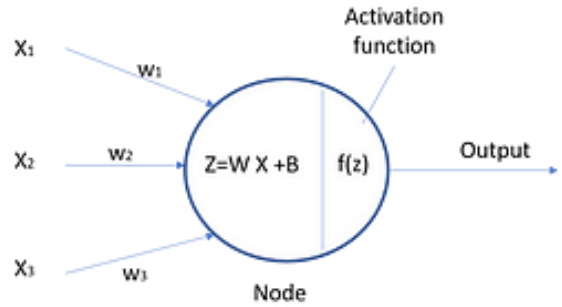
- Statistical language models (SLM)
- Neural language models (NLM)
- Pre-trained language models (PLM).
- Large language models (LLM)

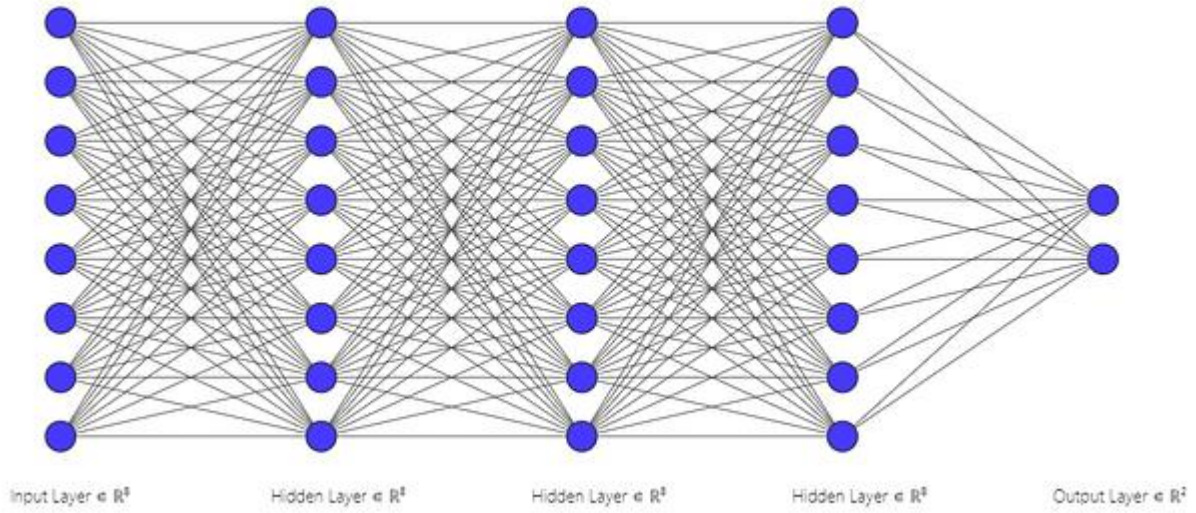


Preliminary: The artificial neuron

- Inspired by the human neuron
- It aggregates information
- Filter out non-relevant information

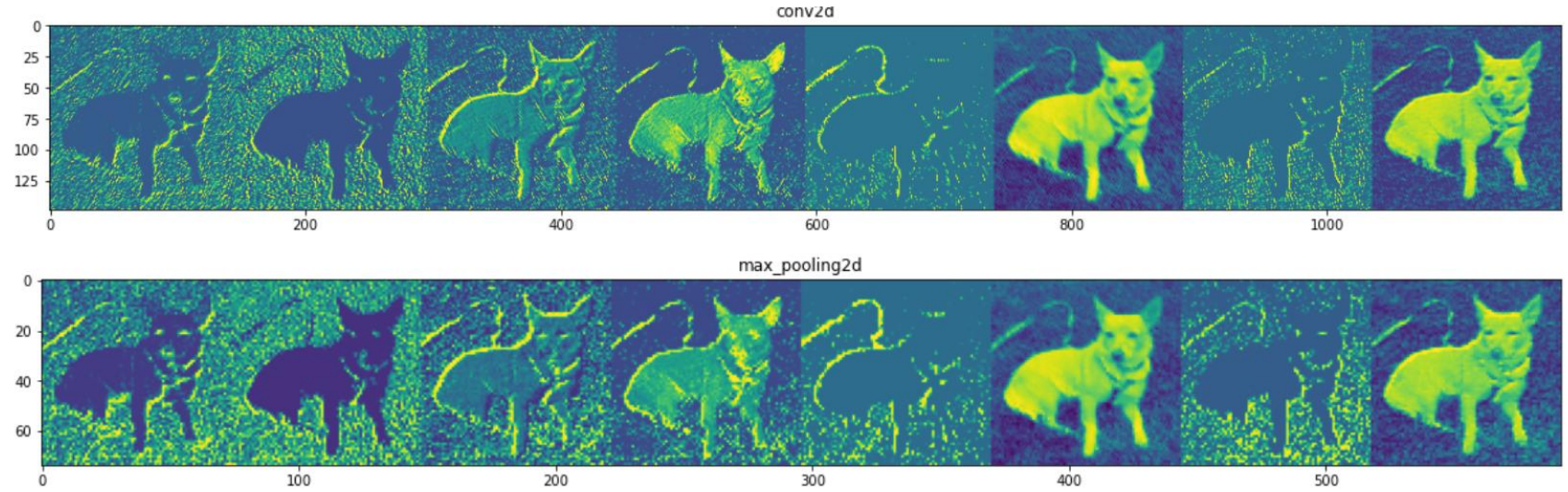


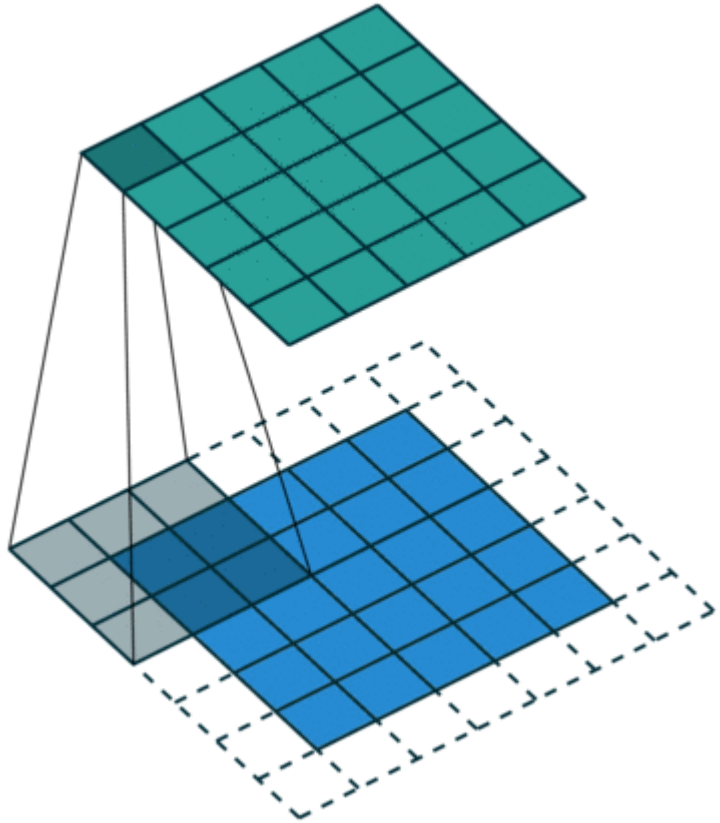




We see the world and then we speak about it

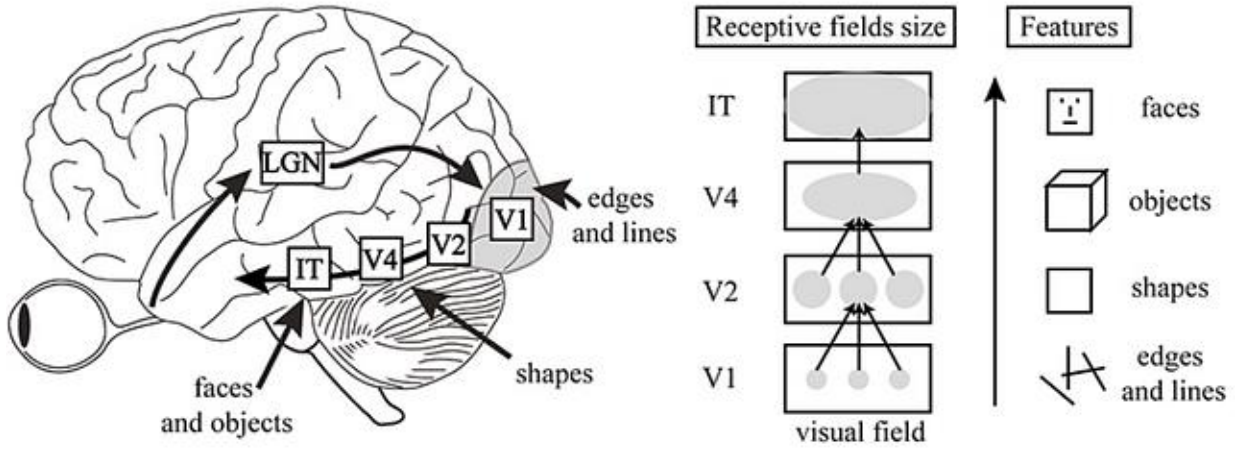
- Convolutional NN have been the first succesfull application of NNs
- CNNs have been the first modern wave of AI





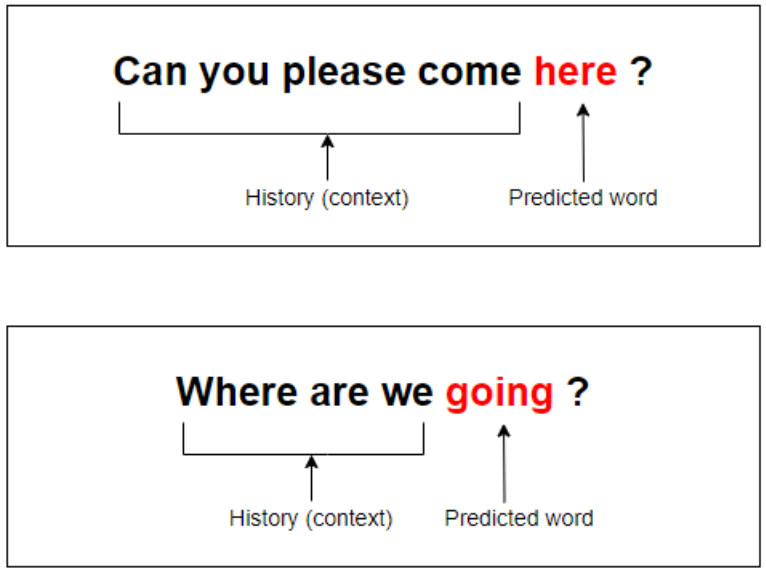
Preliminary: Why stacking more layers?

- Inspired by the human cortex
- Hierarchical representation
- efficiency

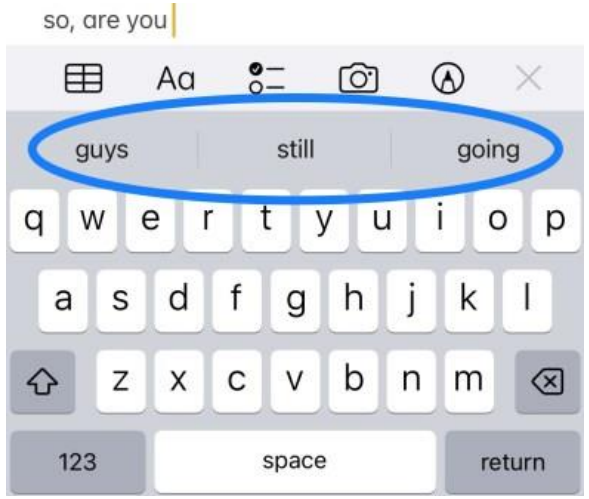


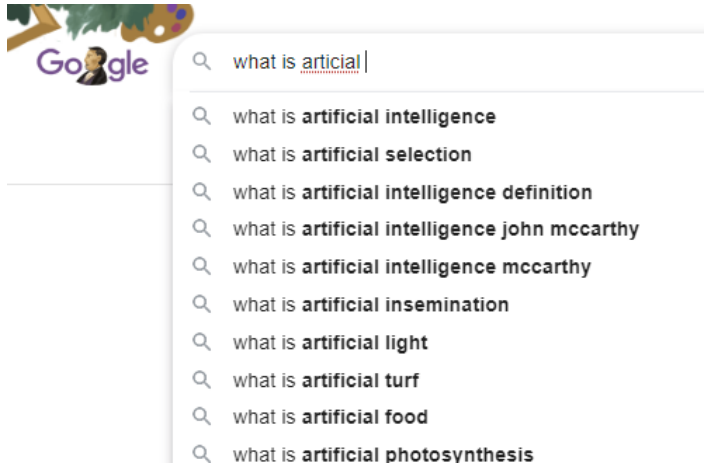
Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

Preliminary: language modelling



You are using language modelling every day



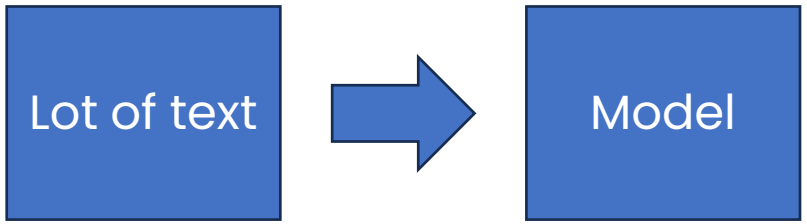


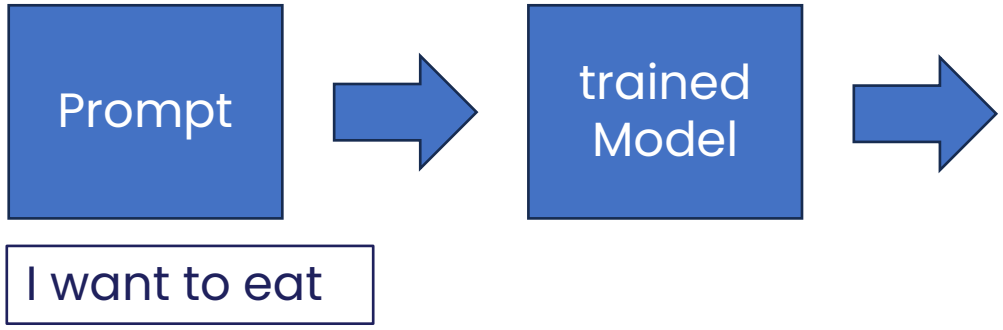
- Language model is predicting the next word given a sequence
- More formally given a sequence of words we calculate the probability distribution of the next word

$$P(\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(T)}) = P(\boldsymbol{x}^{(1)}) \times P(\boldsymbol{x}^{(2)}|\ \boldsymbol{x}^{(1)}) \times \cdots \times P(\boldsymbol{x}^{(T)}|\ \boldsymbol{x}^{(T-1)},\ldots,\boldsymbol{x}^{(1)})$$
$$= \prod_{t=1}^{T} P(\boldsymbol{x}^{(t)}|\ \boldsymbol{x}^{(t-1)},\ldots,\boldsymbol{x}^{(1)})$$

Preliminary: language modelling

- The process is iterative





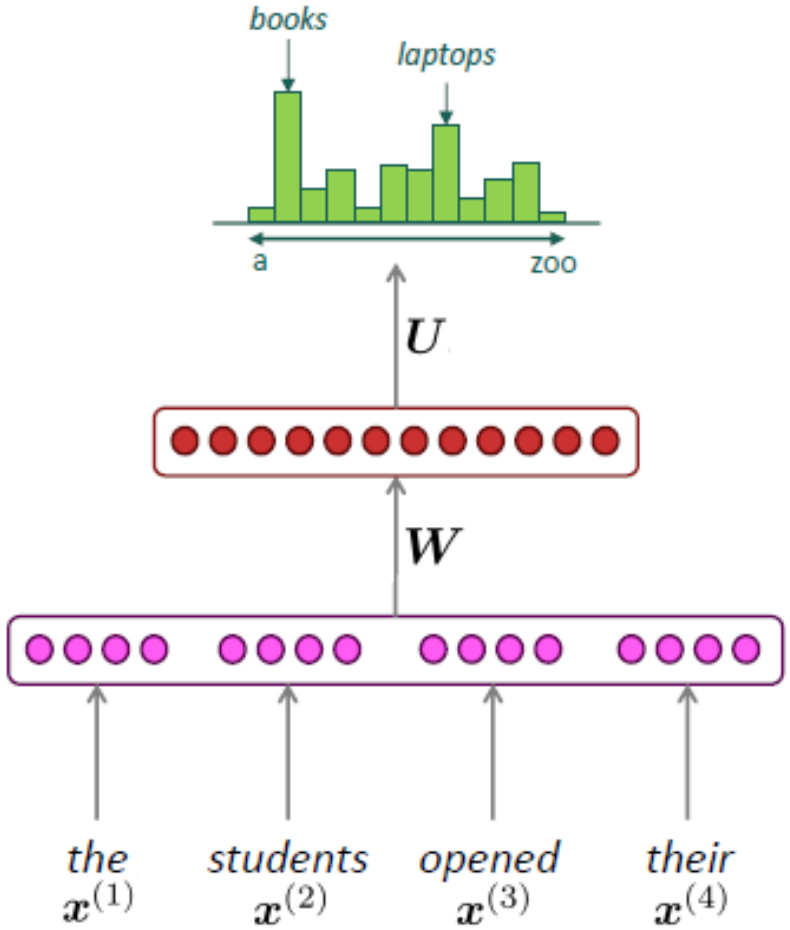
I want to eat Chinese food

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

I want to eat Chinese food

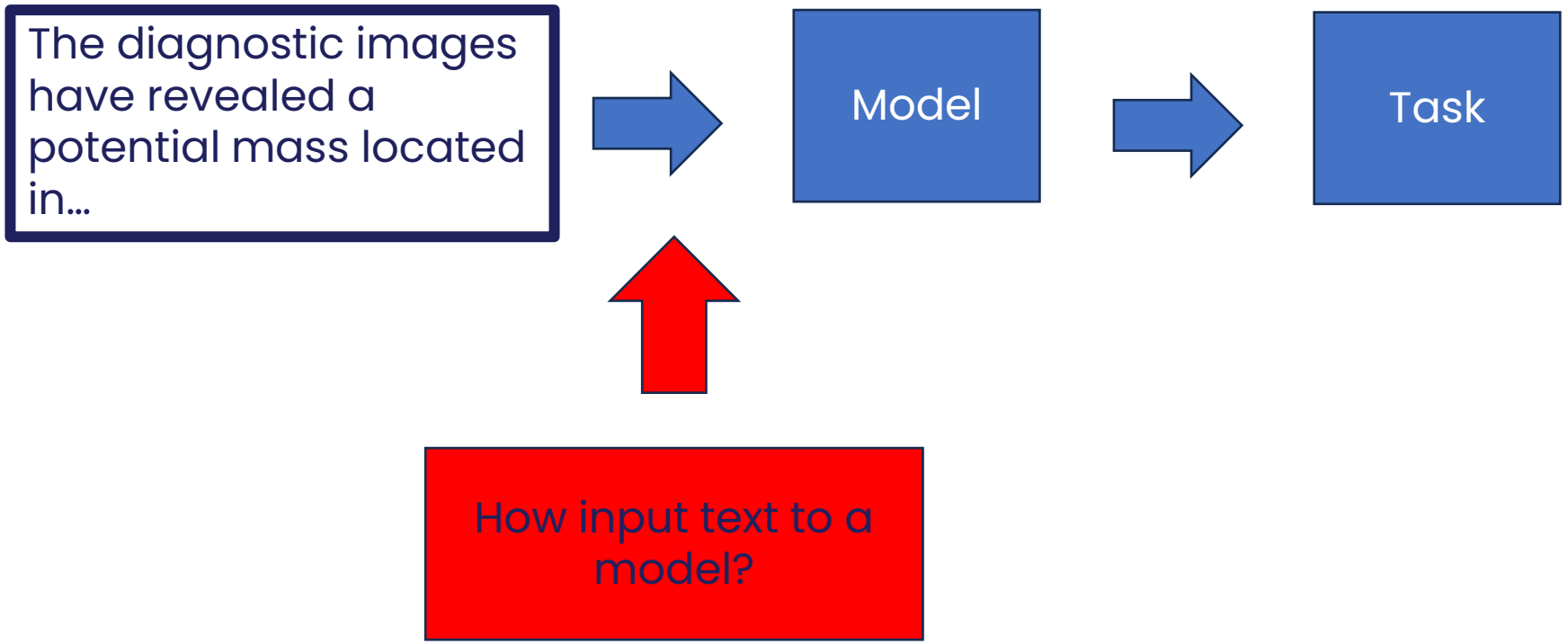
Preliminary: language modelling

- The process is iterative
- Given a sequence of words we obtain the probabilities for the next word.
- We then select the word with the highest probability



The long road to a Large Language Model

"Think before you speak. Read before you think." - Fran Lebowitz



How we can make a computer digesting test?

- In traditional NLP, the words representation is done as **discrete symbols**
- The dimensionality is quite problematic because a language has around 200.000 words
- No similarity between these two vectors (i.e. Dijon Motel or Dijon Hotel has no similarity)
- Losing semantics, syntactical and context content

One-hot vector

$$restaurant = [0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]$$
$$pizzeria = [0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0]$$

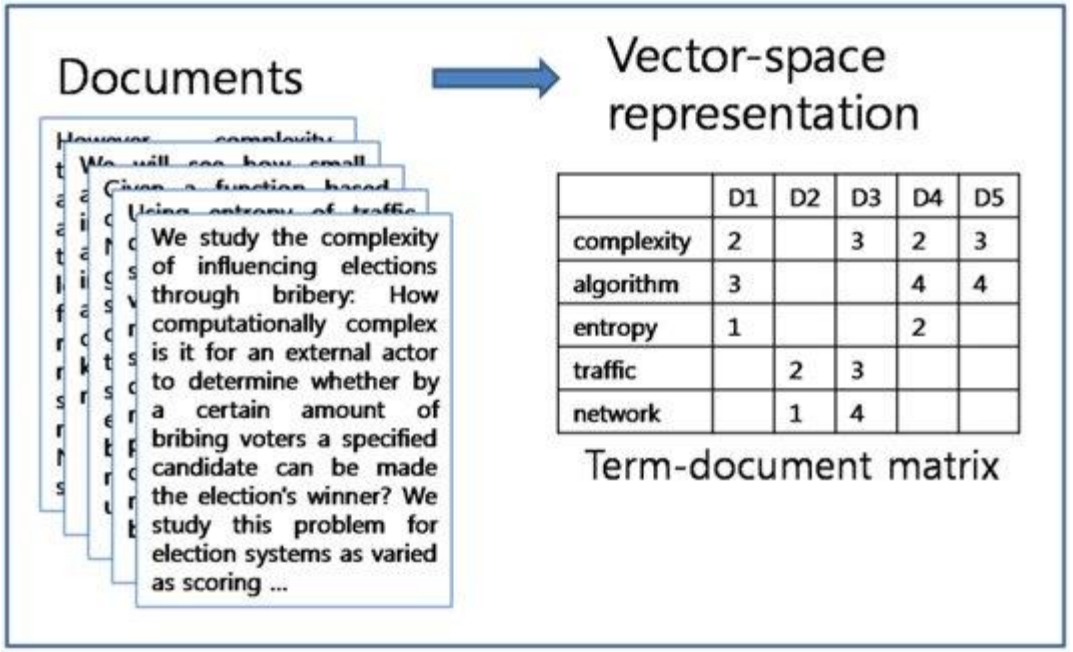
motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Language Models	Semantics	Syntactical	Context	Out of Vocabulary
1-Hot encoding	[×]	[×]	[×]	[×]
BoW	[×]	[×]	[×]	[×]
TF	[×]	[×]	[×]	[×]
TF-IDF	[×]	[×]	[×]	[×]
Word2Vec	[✓]	[✓]	[×]	[×]
GloVe	[✓]	[✓]	[×]	[×]
FastText	[✓]	[✓]	[×]	[✓]
Context2Vec	[✓]	[✓]	[✓]	[✓]
CoVe	[✓]	[✓]	[✓]	[×]
ELMo	[✓]	[✓]	[✓]	[✓]

How we can make a computer digesting test?

- In traditional NLP, the words representation is done as **discrete symbols**
- The dimensionality is quite problematic because a language has around 200.000 words
- No similarity between these two vectors (i.e. Dijon Motel or Dijon Hotel has no similarity)
- Losing semantics, syntactical and context content

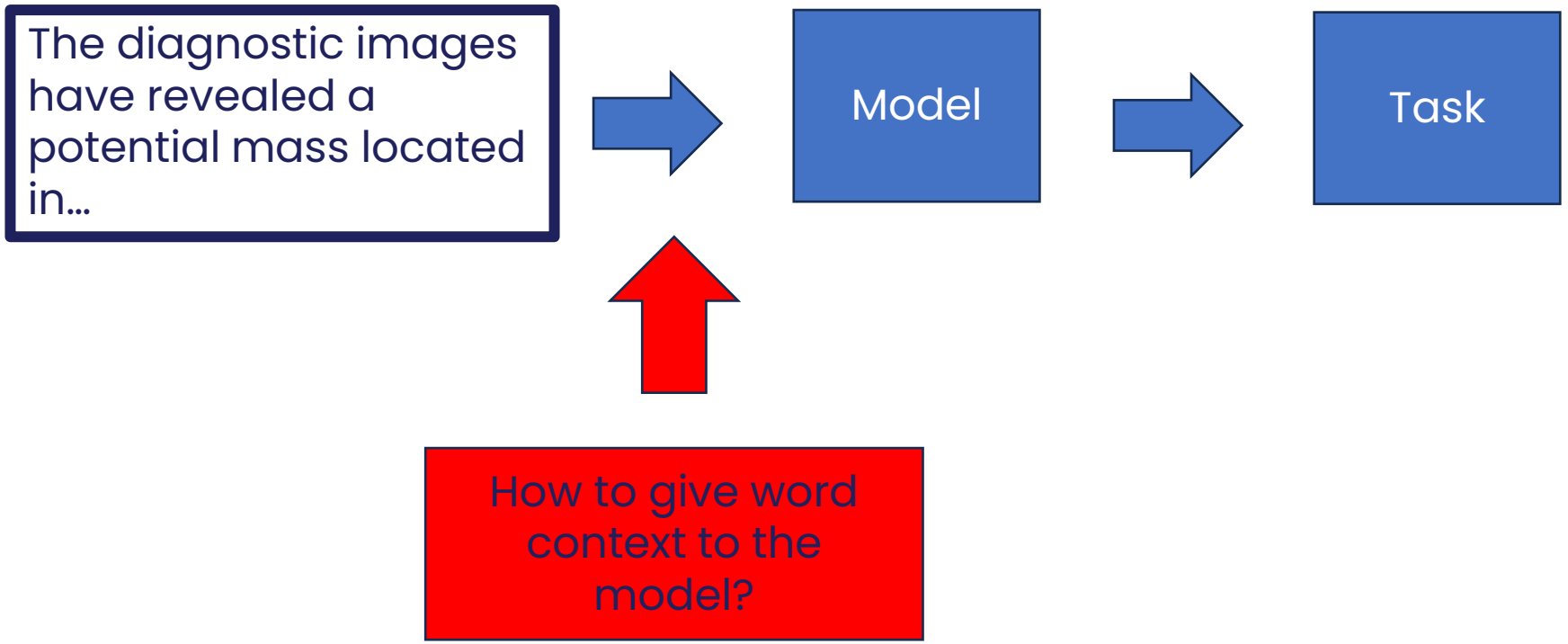
Document term matrix



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

The long road to a Large Language Model

For me context is the key - from that comes the understanding of everything. - Kenneth Noland



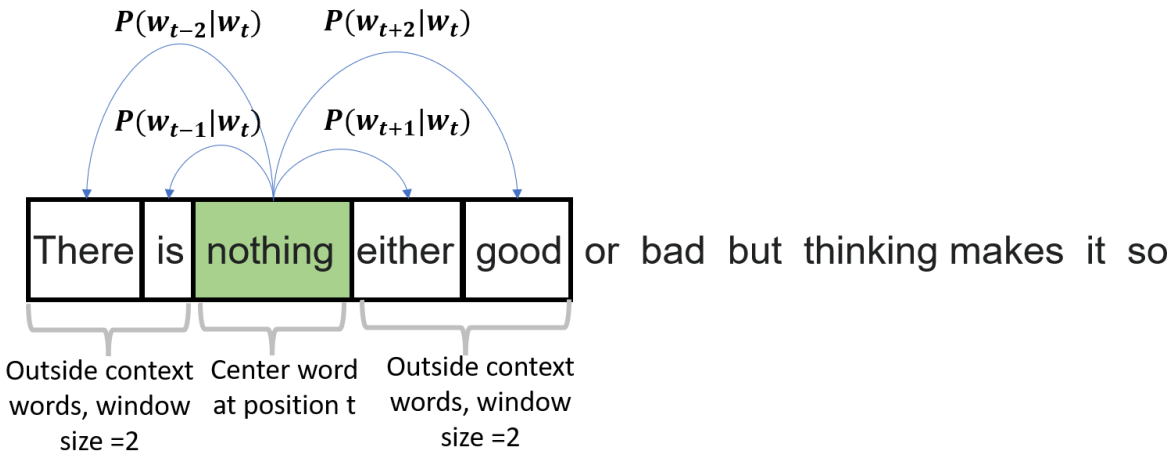
2013: the word2vec revolution

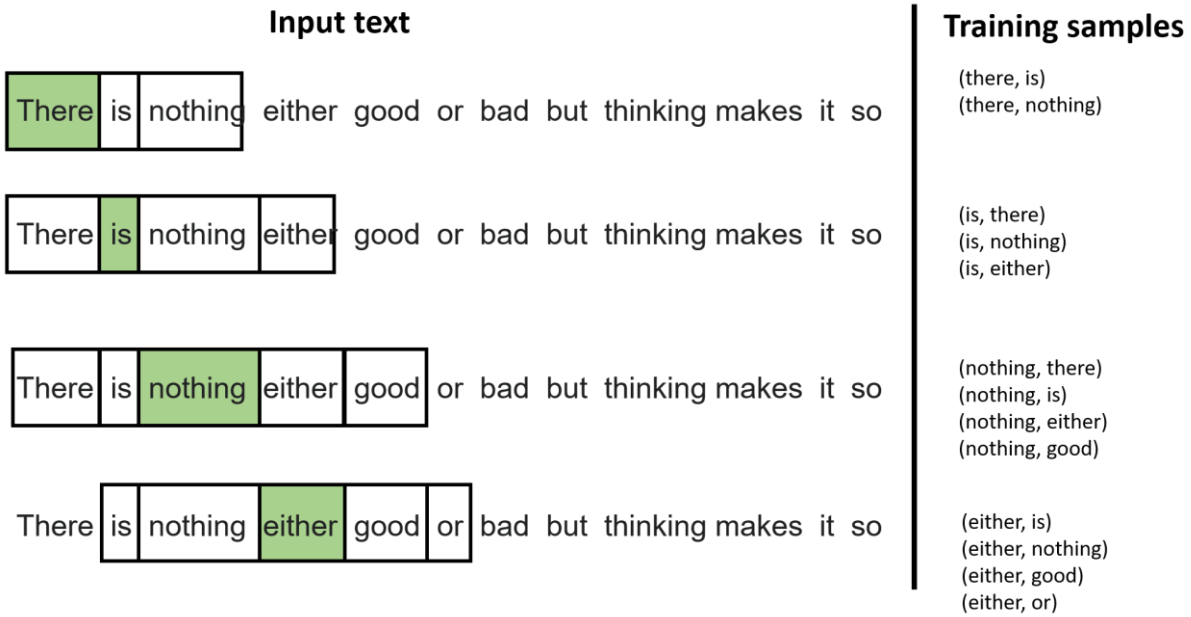
- Word2vec was introduced by Mikolov et al. 2013 as a framework for learning word vectors.
- The model to learn a representation of words In a corpus (the input is a text corpus and the output is a vector representation for each word).
- The model take in account the context of a word and return meaningful vectors

I ***left*** my phone on the ***left*** side of the room.

Homonyms

2013: the word2vec revolution

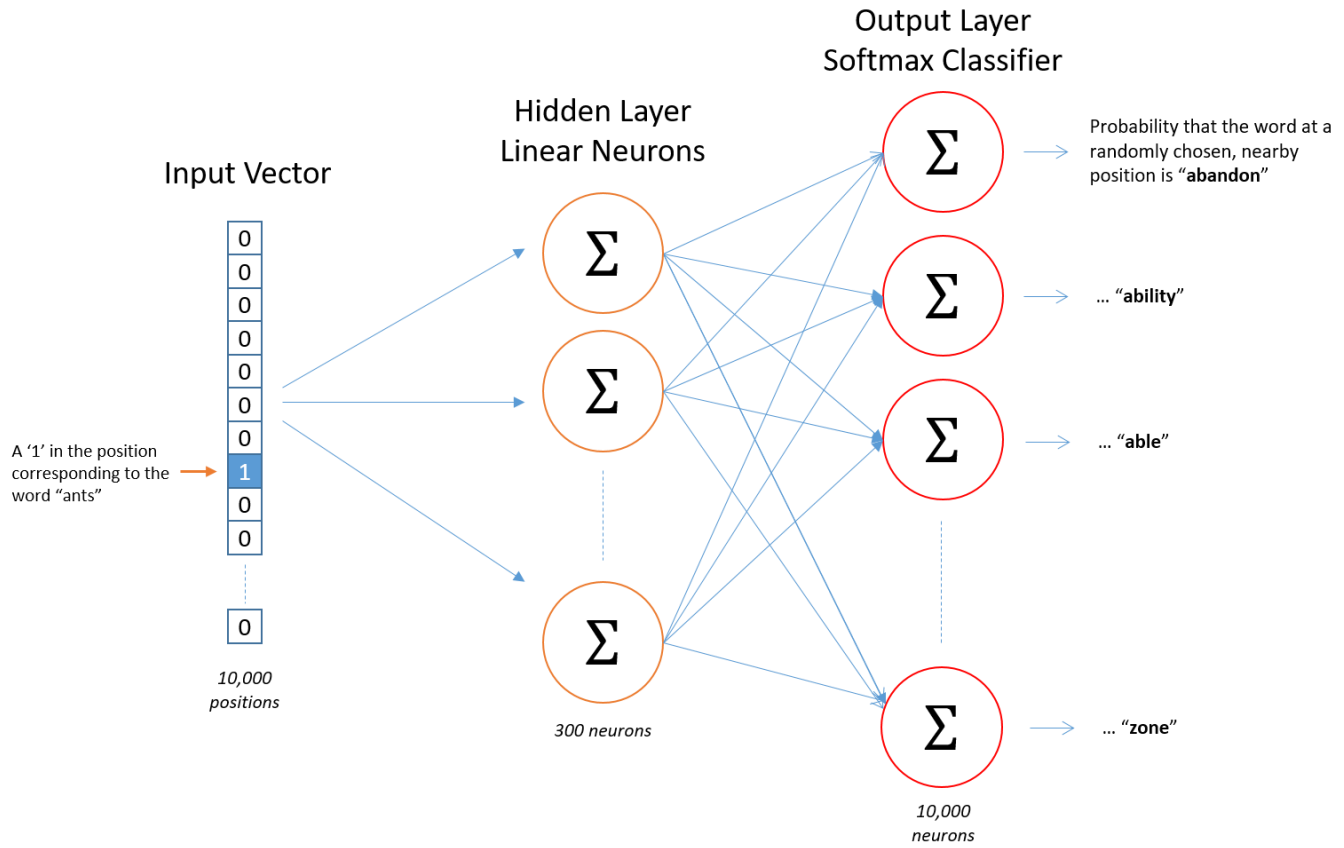




- We have a piece of text and we select a position t, we select a word w and then we ask what is the probability of appearing in the context window of the word w.
- The main question is how to change this word vector to adjust the probability for words that appear in the context window.

2013: the word2vec revolution

- The model is a single-layer fully connected neural network trained for what is called a “fake task”
- The model has one hidden layer (without activation function) and then an output layer with SoftMax.
- This is the overall model, for each position t we want to predict the context words within a window of fixed size m, given a center word wj.



2013: the word2vec revolution

- We are obtaining real number vectors (positive and negative values are possible) that are dense
- These vectors are capturing better the similarity and the context of a word
- Each word is generally represented by a vector of length 300

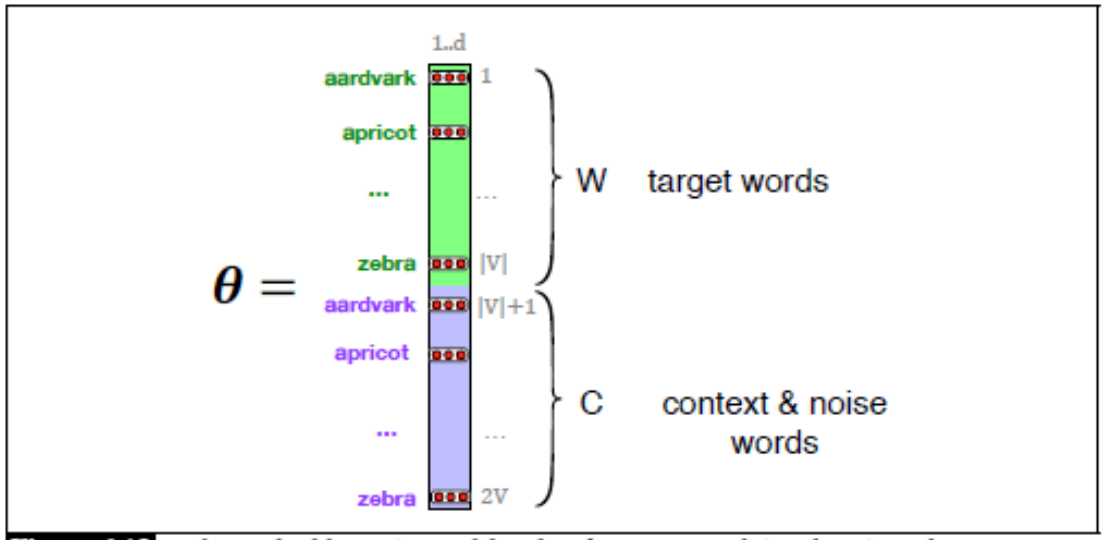
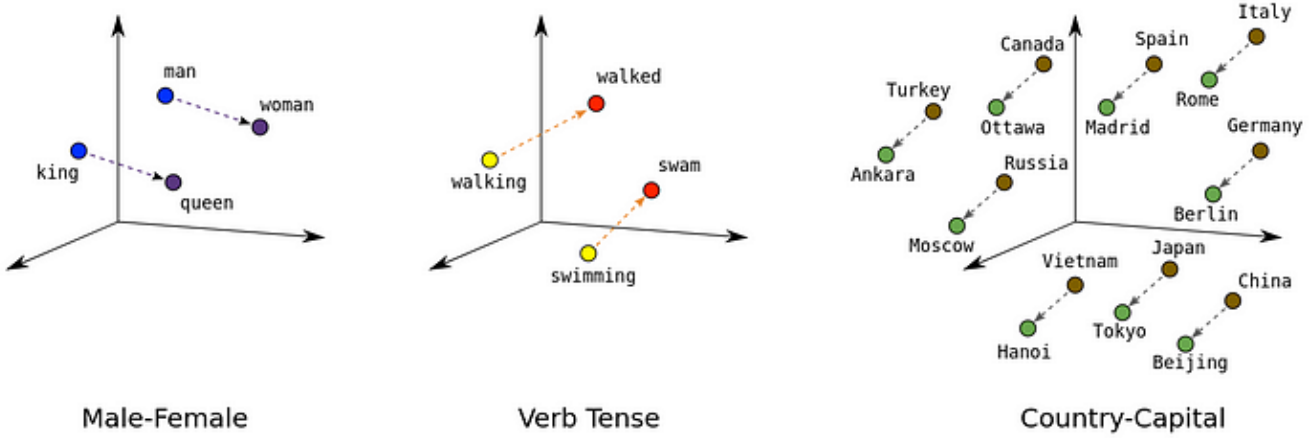


Figure 6.13 The embeddings learned by the skipgram model. The algorithm stores two embeddings for each word, the target embedding (sometimes called the input embedding) and the context embedding (sometimes called the output embedding). The parameter θ that the algorithm learns is thus a matrix of $2|V|$ vectors, each of dimension d, formed by concatenating two matrices, the target embeddings **W** and the context+noise embeddings **C**.

2013: the word2vec revolution



Male-Female Verb Tense Country-Capital

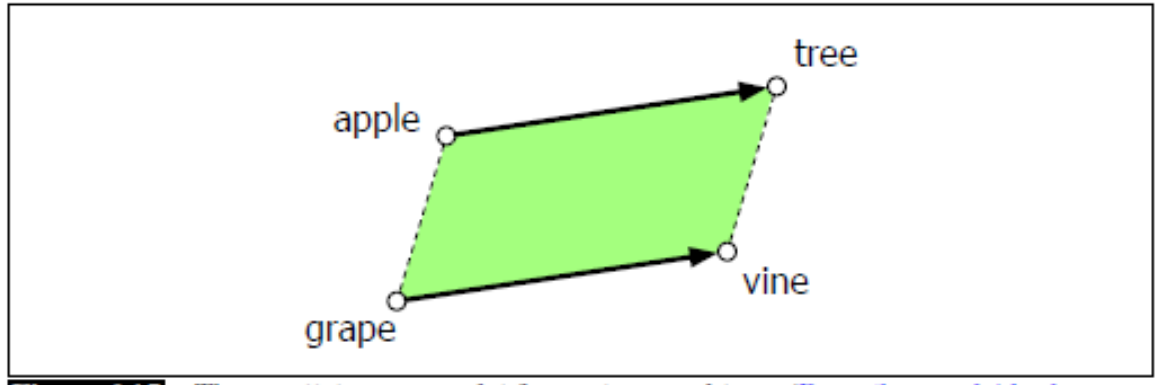


Figure 6.15 The parallelogram model for analogy problems (Rumelhart and Abrahamson, 1973): the location of $\overrightarrow{\text{vine}}$ can be found by subtracting $\overrightarrow{\text{apple}}$ from $\overrightarrow{\text{tree}}$ and adding $\overrightarrow{\text{grape}}$.

- The vector captures semantic and syntactic analogy, like comparatives and superlatives
- We can measure similarity and antinomy using cosine similarity
- the Word2vec vector is a weighted sum of different senses for a word, you can do different operations

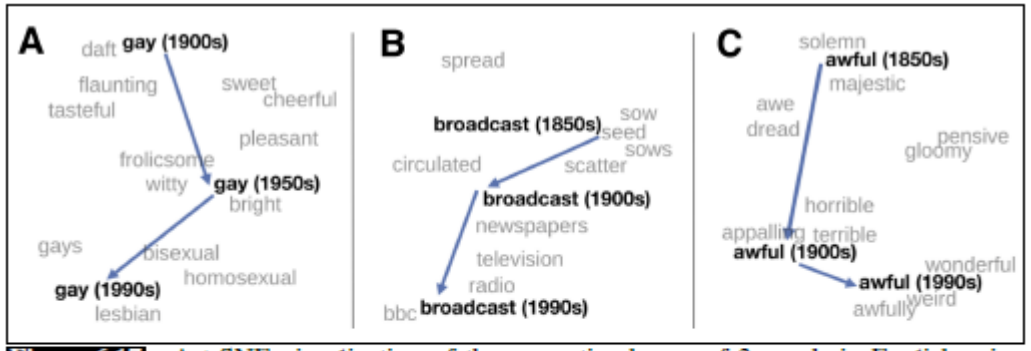
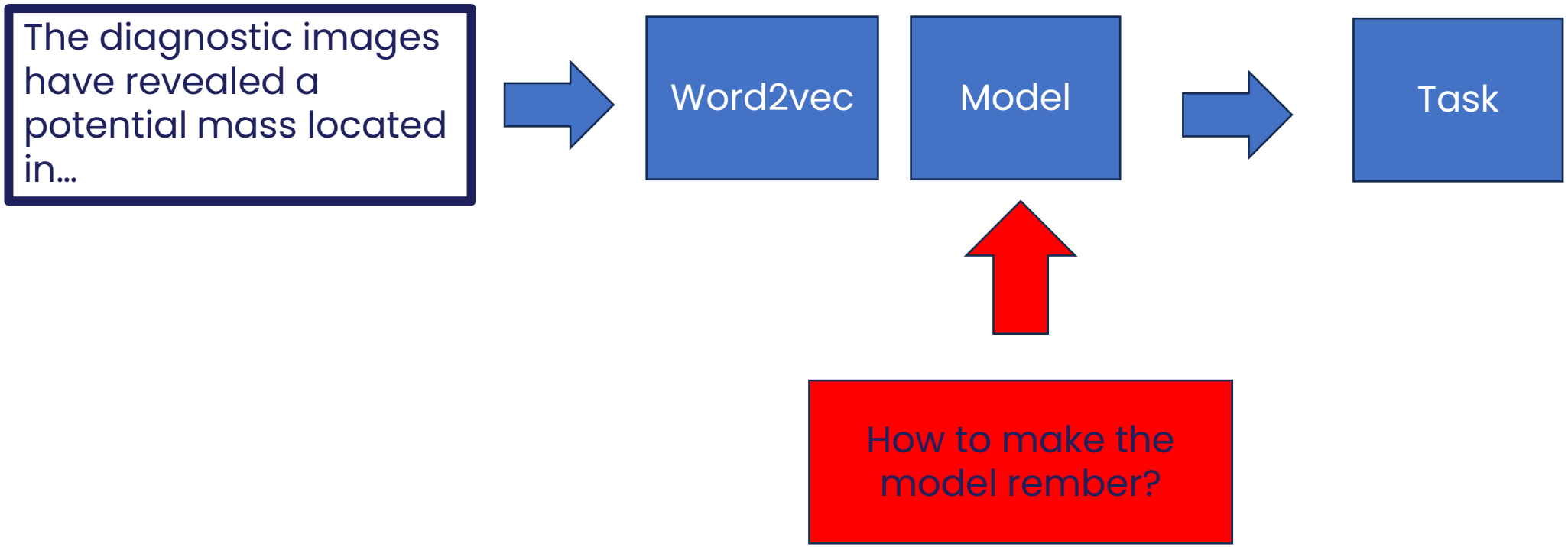


Figure 6.17 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016).

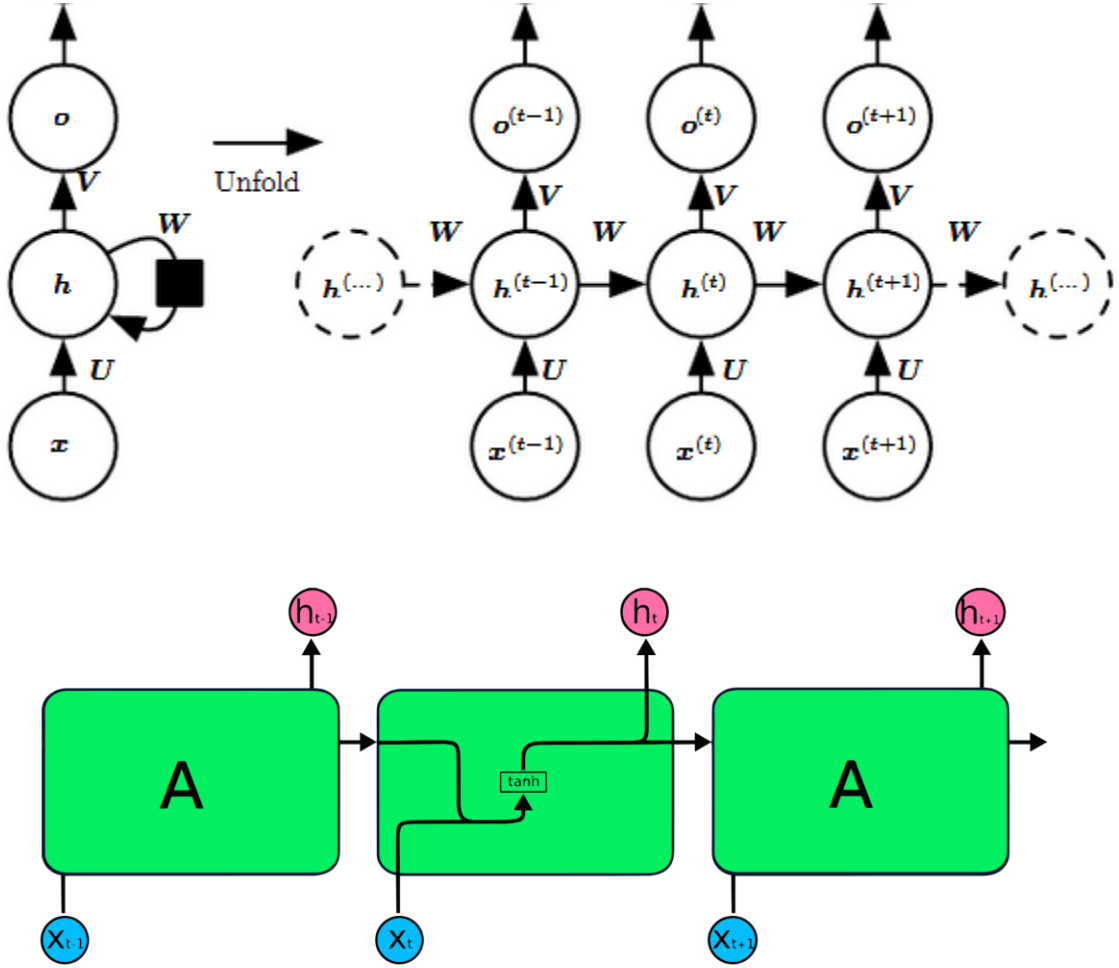
The long road to a Large Language Model

Without memory, there is no culture. Without memory, there would be no civilization, no society, no future. -Elie Wiesel



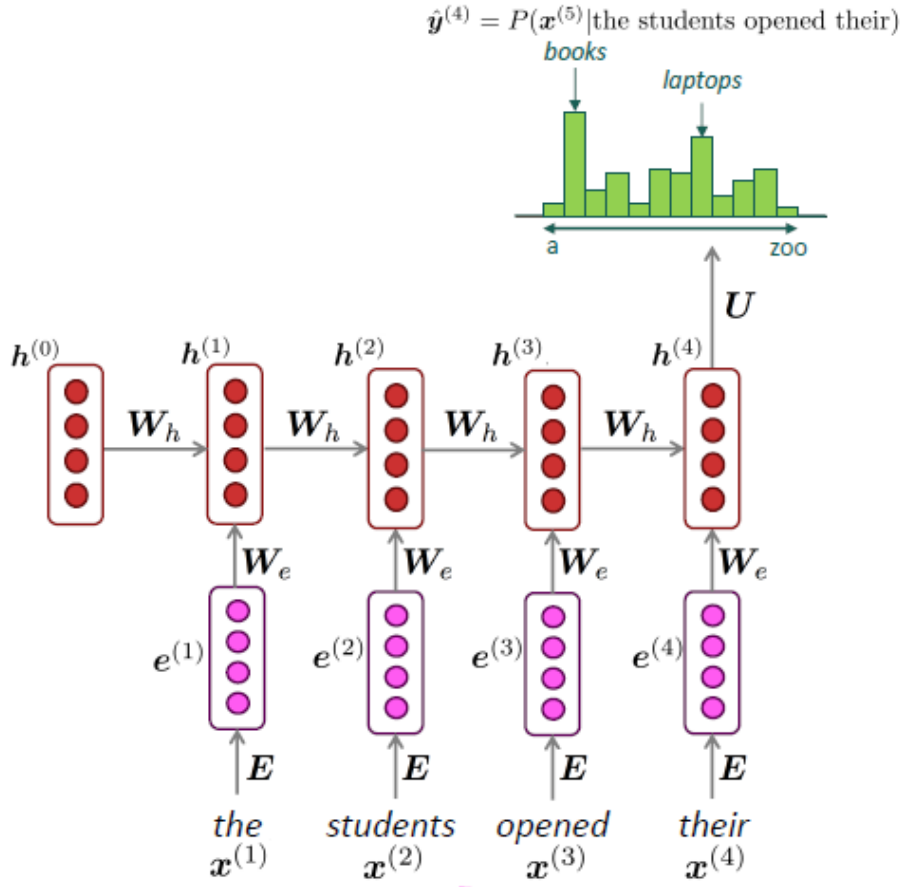
Recurrent Neural Network

- A neural network specialized in processing a sequence of data (the outputs is dependent on the previous computations)
- We have a hidden state that is updated with each time step
- We do not have anymore a fixed length for the sequence



Recurrent Neural Network

- Model weights are shared for all the words in the sequence and updated for each word
- You can already generate text with RNN



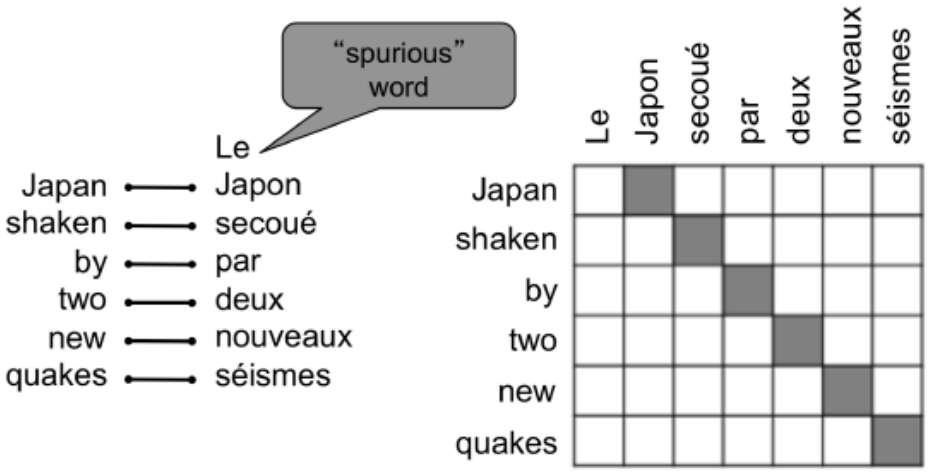
"Sorry," Harry shouted, panicking—"I'll leave those brooms in London, are they?"

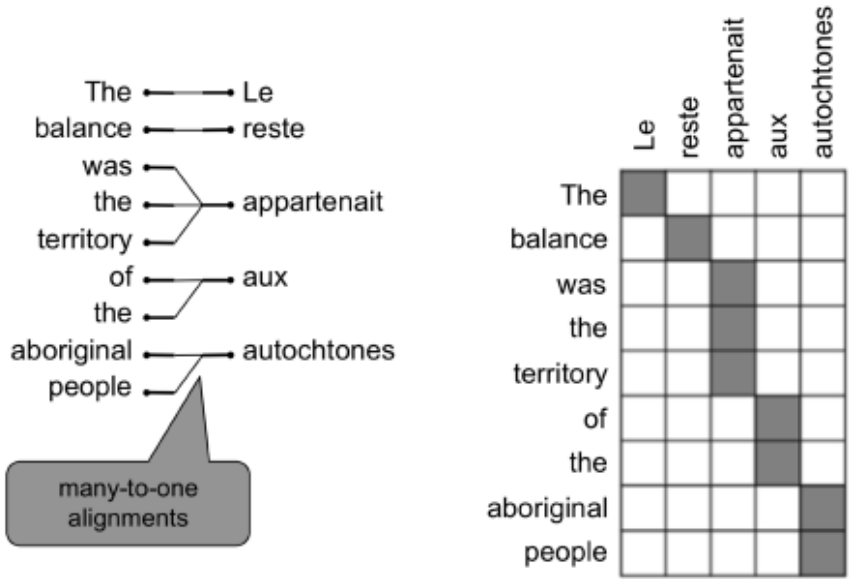
"No idea," said Nearly Headless Nick, casting low close by Cedric, carrying the last bit of treacle Charms, from Harry's shoulder, and to answer him the common room perched upon it, four arms held a shining knob from when the spider hadn't felt it seemed. He reached the teams too.

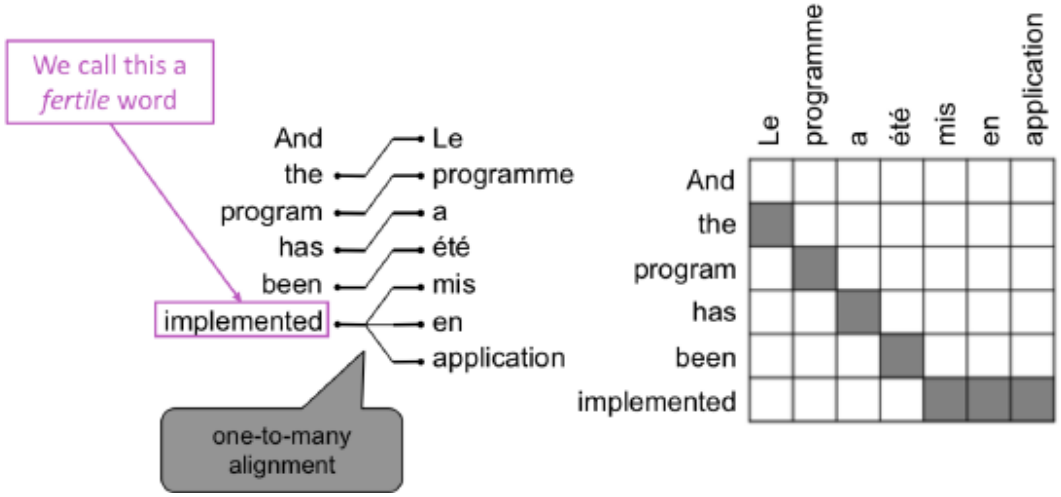
Source: https://medium.com/deep-writing/harry-potter-written-by-artificial-intelligence-8a9431803da6

Is it RNNs enough?

- Vanishing gradient problem
- Infinite sequence in theory, short sequence in practice
- Not easily parallelizable
- Issue when sequences are not aligned
- Not modeling long word dependencies

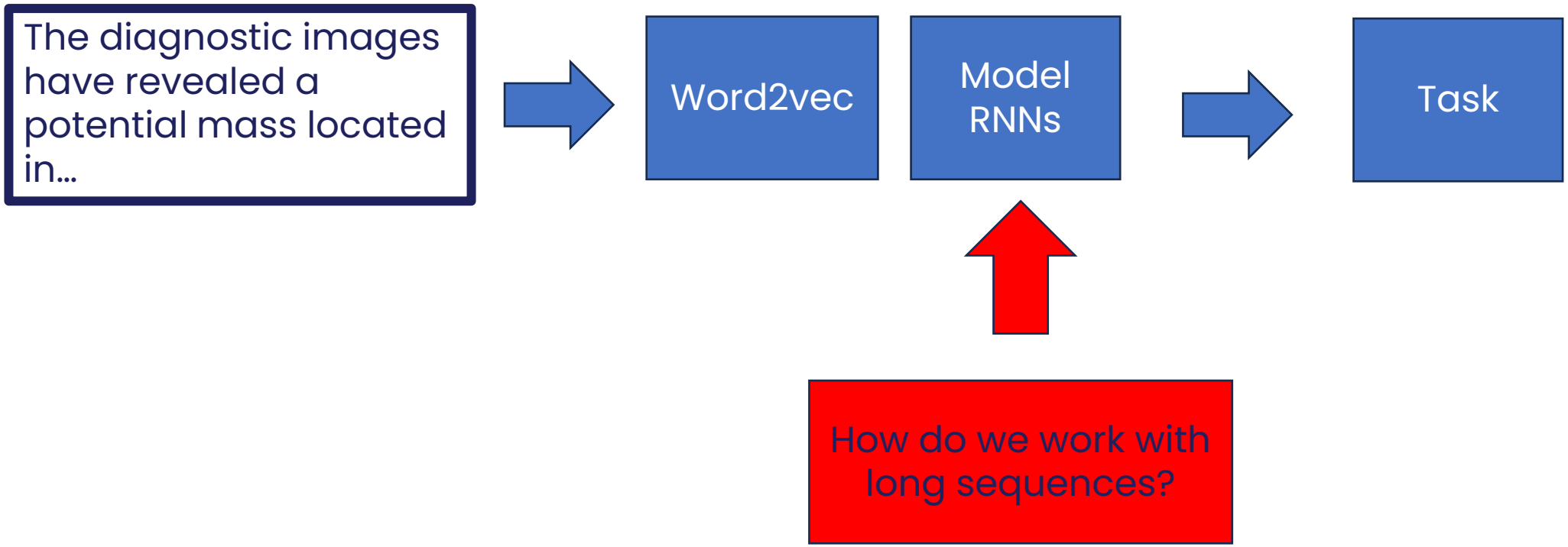






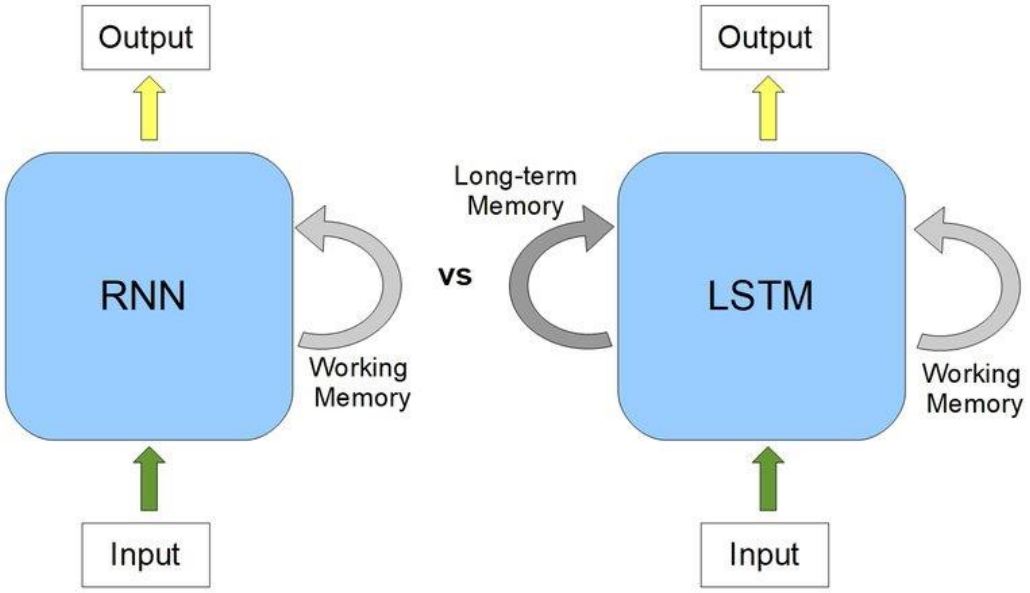
The long road to a Large Language Model

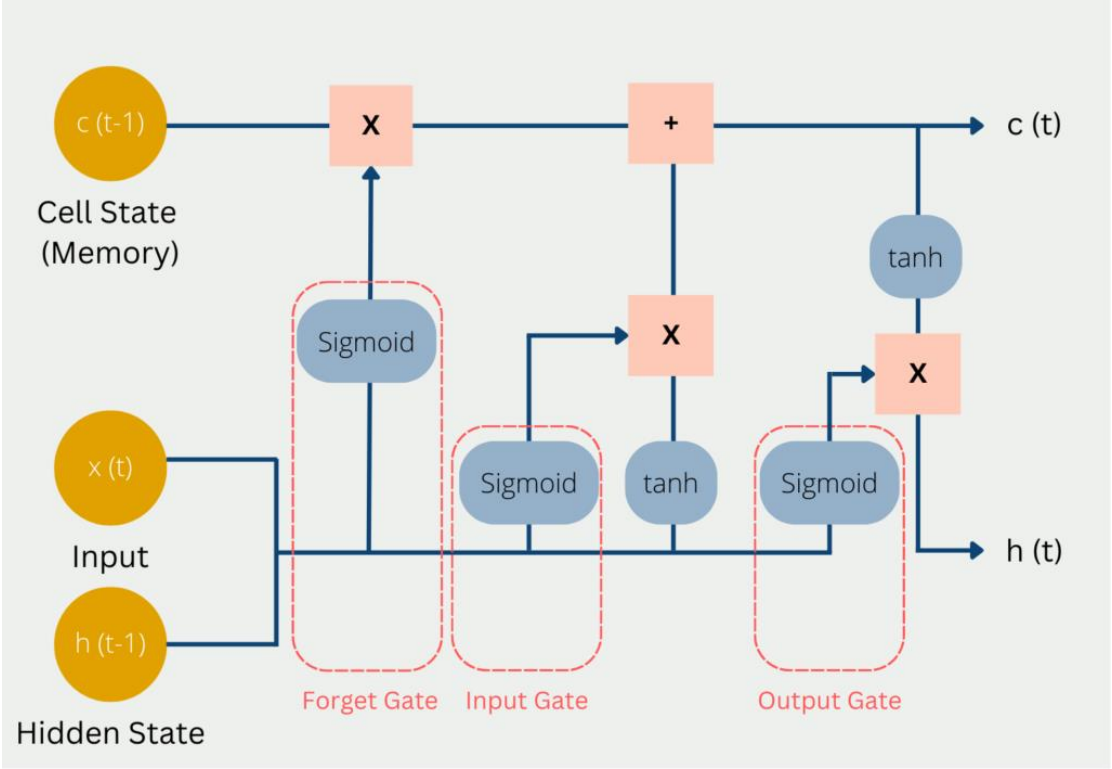
The events in our lives happen in a sequence in time, but in their significance to ourselves they find their own order the continuous thread of revelation. - Eudora Welty



LSTM: sometimes you need to forget

- Different alternatives have been tried but with unsatisfactory results





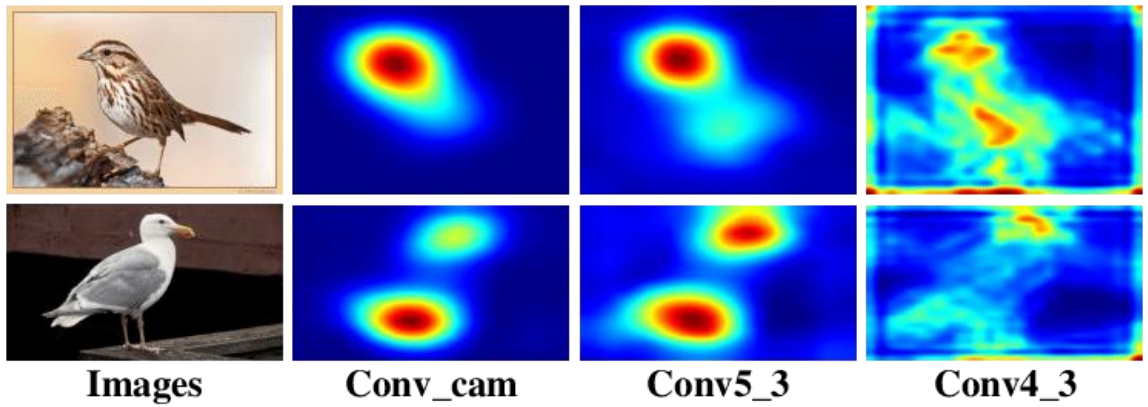
We pay attention!

- Attention is a complex cognitive function that is indispensable for human beings
- Humans selectively concentrate on a part of the information, ignoring the rest
- Allows to focus resources on a particular task

I really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do when I go to MI because of the quality of the highlight and the price the price be very affordable the highlight fantastic thank Ashley i highly recommend you and ill be back

love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola

Heatmap of 5 stars Yelp reviews. Heavier colors indicate higher attention weight.



Attention map for images

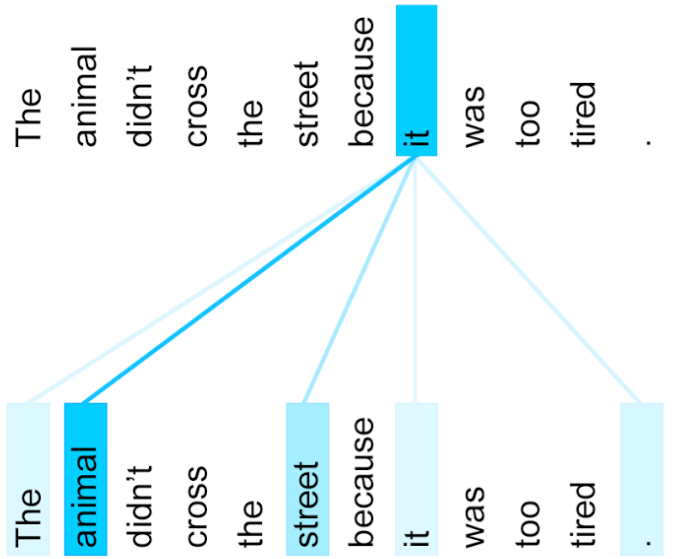
How we decide on what to focus

- Calculate the importance of each word in the sequence according to the task
- Find they relative importance
- Give the attention to each word

$$(1) \quad e_{i,j} = score(s_{i-1}, h_j)$$

$$(2) \quad a_{i,j} = softmax(e_{i,j}) = \frac{\exp(e_{i,j})}{\sum_{k=1}^{t} \exp(e_{i,k})}$$

$$(3) \quad c_t = \sum_{j=1}^{T} a_{i,j} \cdot h_j$$



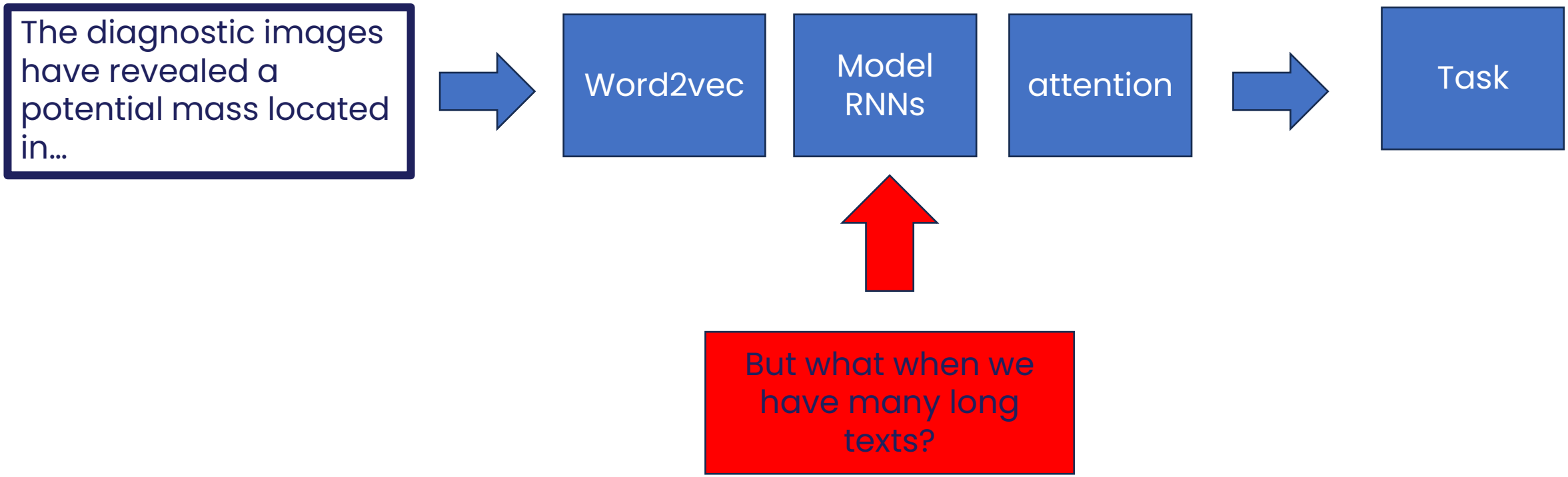
The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

An example of the attention in translation

The long road to a Large Language Model

The true art of memory is the art of attention. - Samuel Johnson



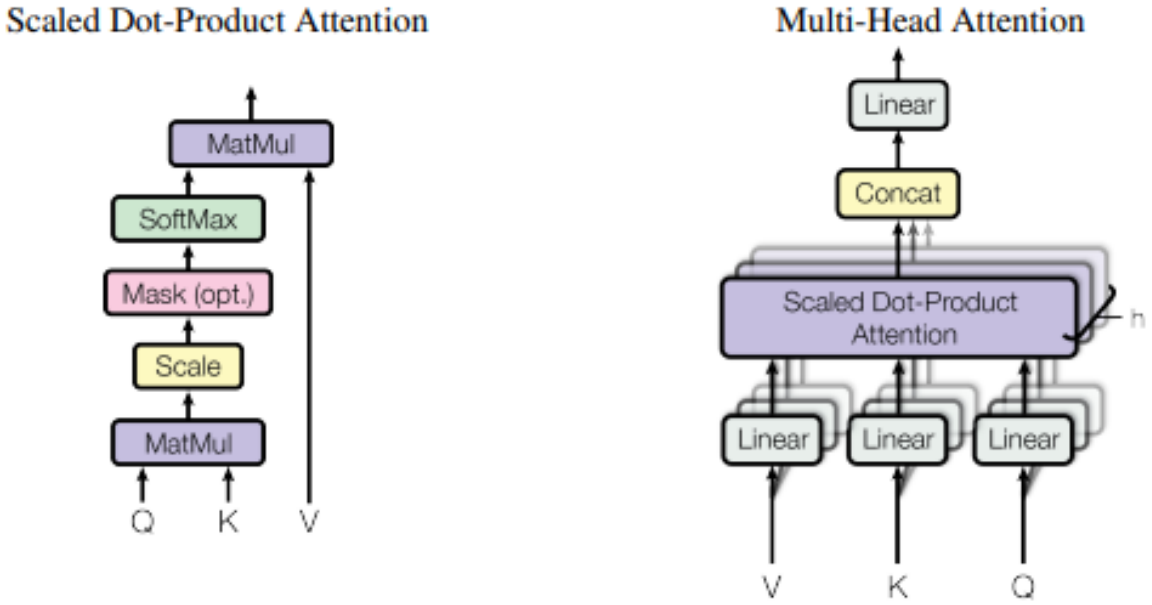
Attention is all you need

So if you were to search for something on Youtube or Google:

- The text which you type in the search box is the QUERY.
- The results which appear as the video or article title are the KEY
- The content inside them is the VALUE.

So given a Query we put it in relationship with the keys and values

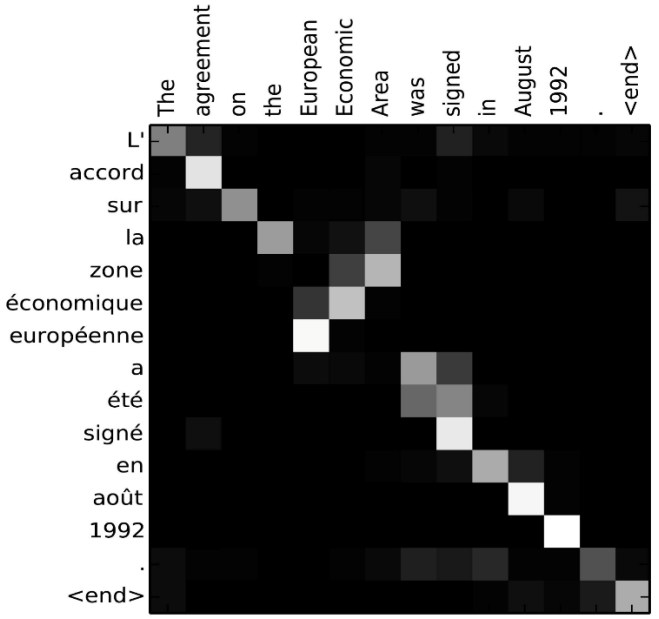
We do this process different times to learn different data representation

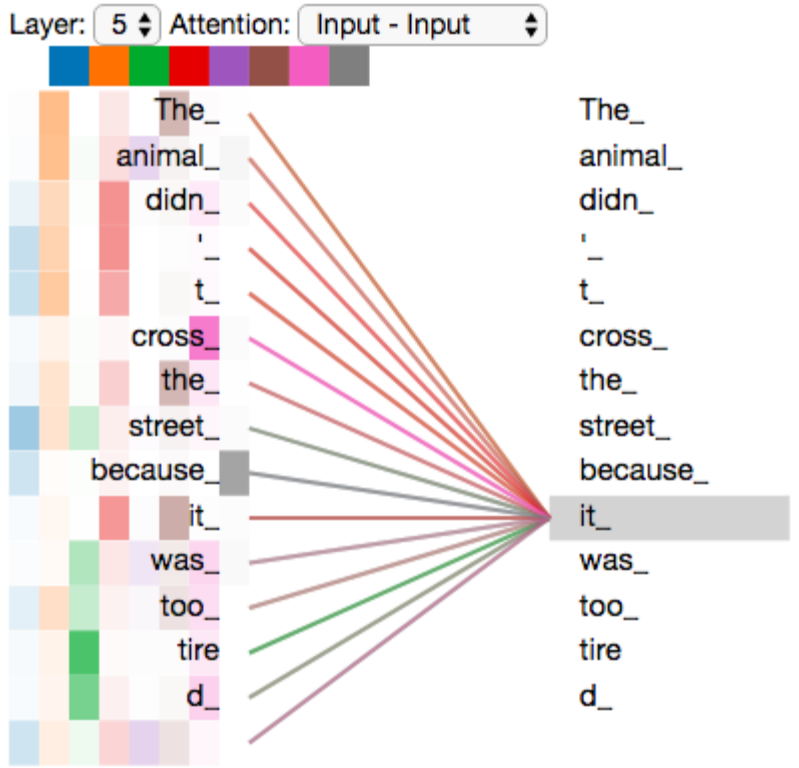


$$Attention\,(Q, K, V) = softmax\left(\frac{Q \cdot k^T}{\sqrt{d_k}}\right) \cdot V$$

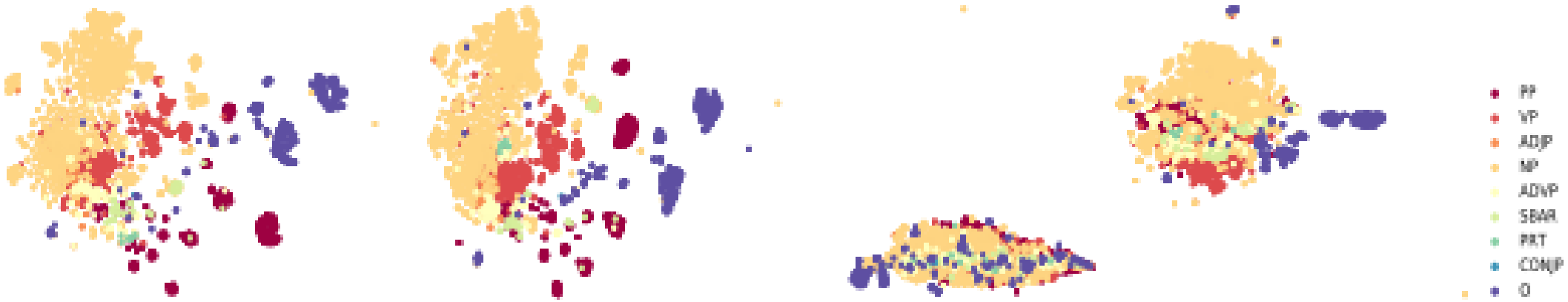
The model is learning relationships

- Attention is putting in relationship the diverse word in a sequence
- Each attention map is learning a different relationship between words





Do we need just one attention layer?



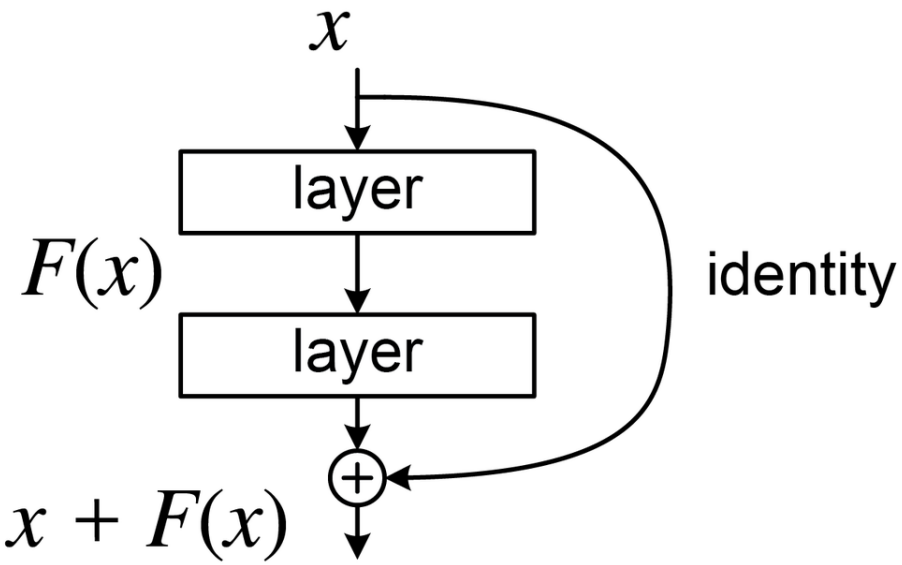
(a) Layer 1 (b) Layer 2 (c) Layer 11 (d) Layer 12

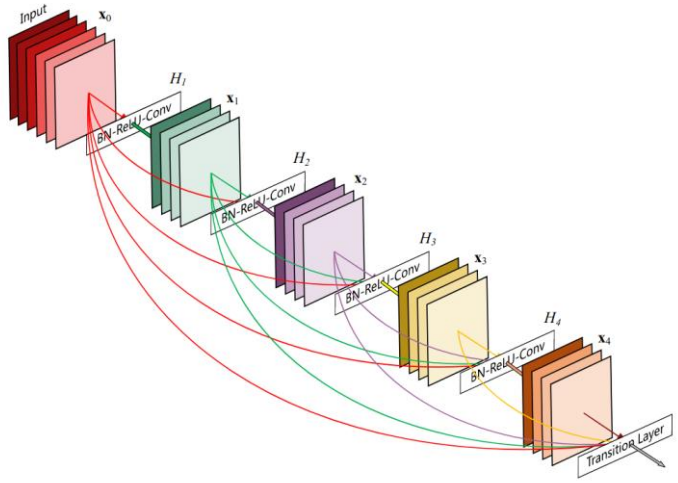
Figure 1: $2D$ t-SNE plot of span embeddings computed from the first and last two layers of BERT.

- Multilayer models learn a hierarchical representation: from simple to more complex
- LLMs learn from simlple word syntaxis to complex metaphor

How can we train with so many layers?

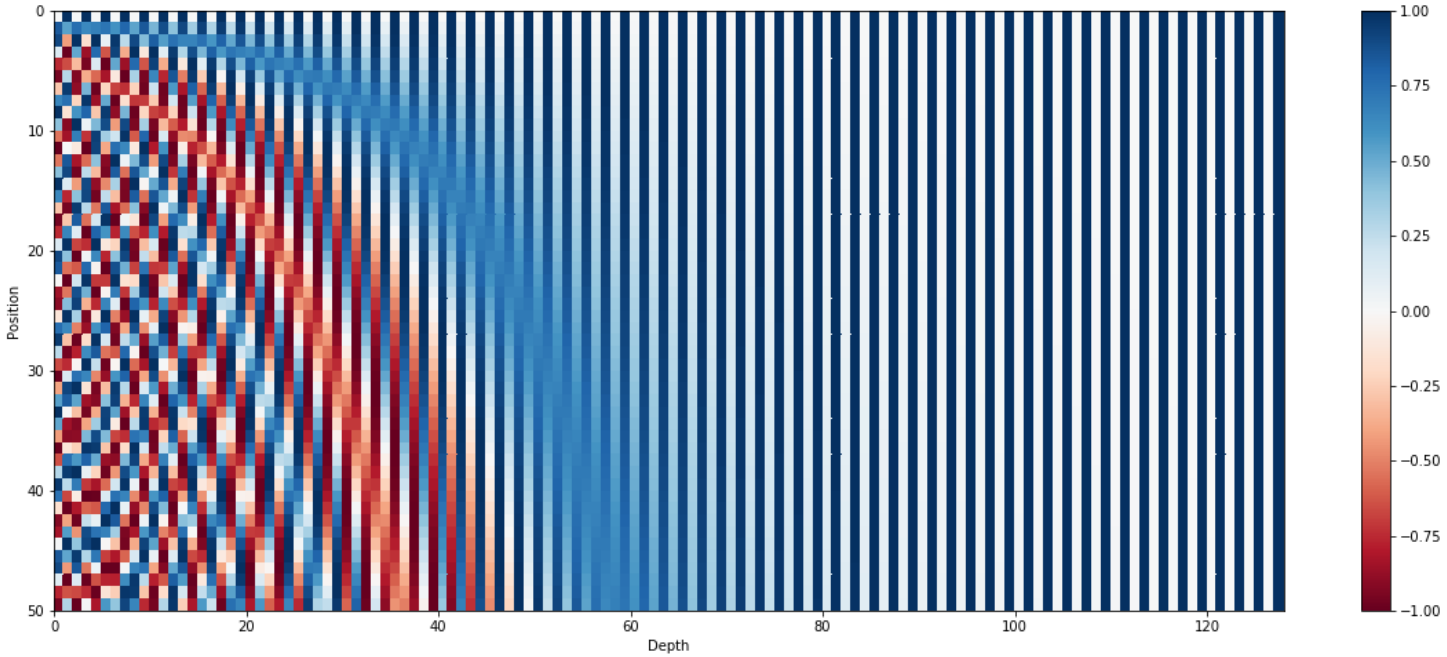
- Skip connection are used to allow to flow the information between parts of the model
- the neocortex has a similar structure to residual nets; where cortical layer VI neurons get input from layer I





Know your place! The word position is important

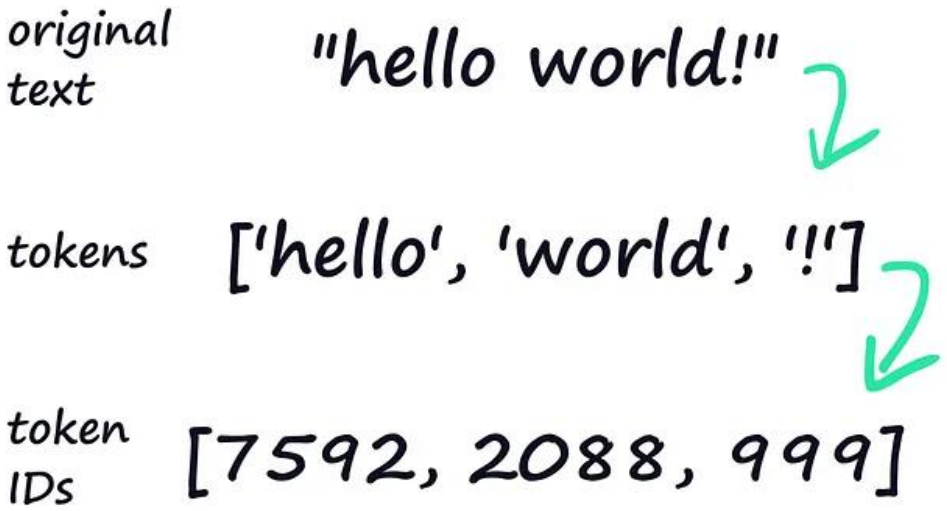
- We need a way to account for the order of the words in the input sequence
- We use positional encoding to determine the position of each word, or the distance between different words in the sequence



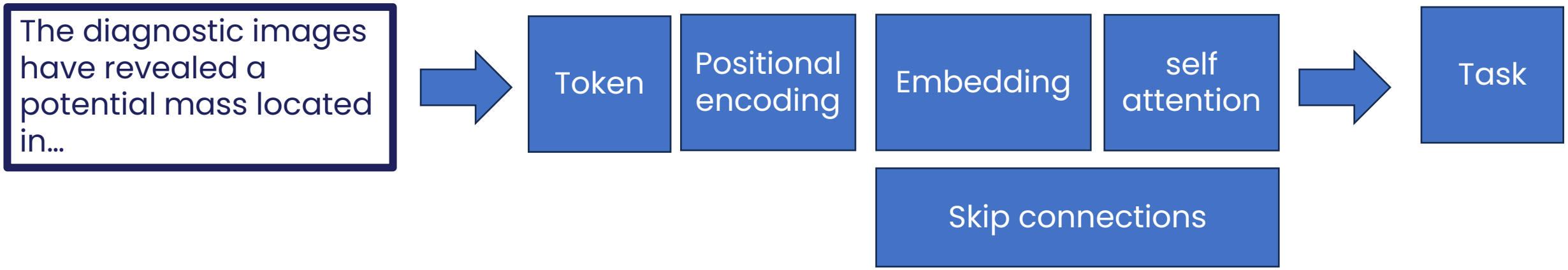
What does it mean this word? Enlarge your vocabulary!

- tokenization breaks down vast stretches of text into more digestible and understandable units for machines.
- For a transformer: Tokenization is the process of encoding a string of text into transformer-readable token ID integers.
- Word tokenization, Character tokenization, Subword tokenization

- ***L'ensemble*** → one token or two?
  - ***L*** ? ***L'*** ? ***Le*** ?
  - Want ***l'ensemble*** to match with ***un ensemble***



The long road to a Large Language Model



His majesty, the transformer

"All empires become arrogant. It is their nature." — Edward Rutherfurd

- Originally invented for machine translation, it has become the standard for all the tasks in NLP
- Used today for any type of sequences
- Used today also for images and video

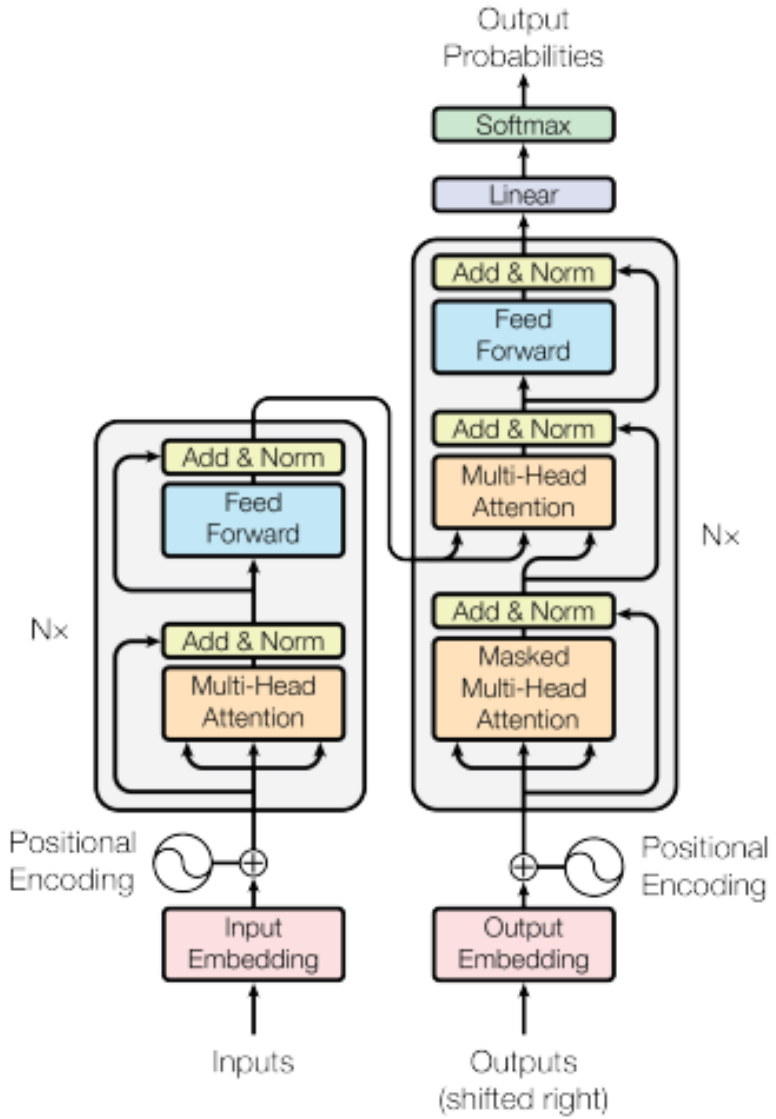
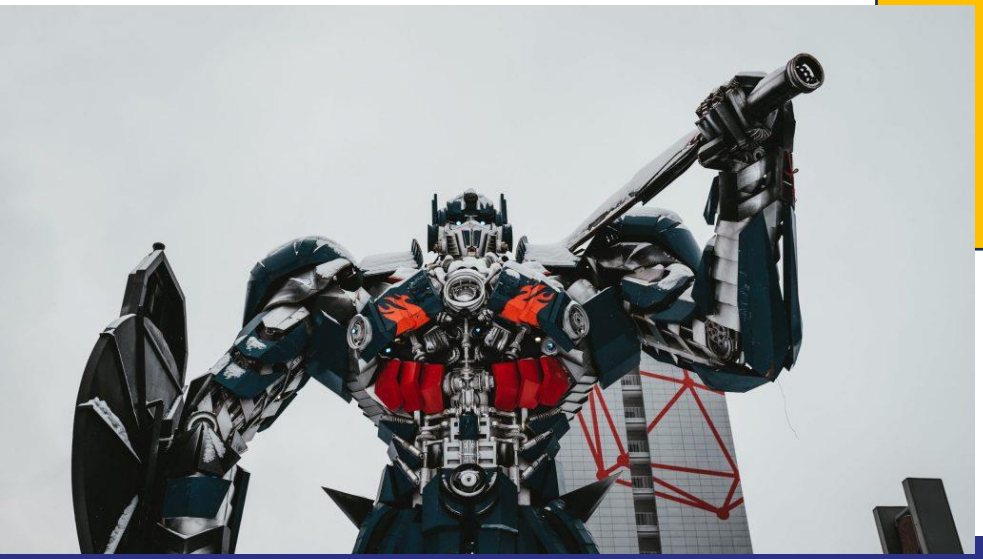


Figure 1: The Transformer - model architecture.

The long road to a Large Language Model

The diagnostic images have revealed a potential mass located in...



Positional encoding

Embedding

self attention

Skip connections

TRANSFORMER

Task

Are large language models (LLM) transformers?

- All LLMs are transformers, but all transformers are LLMs
- There are differences in the components and in the scale of the model

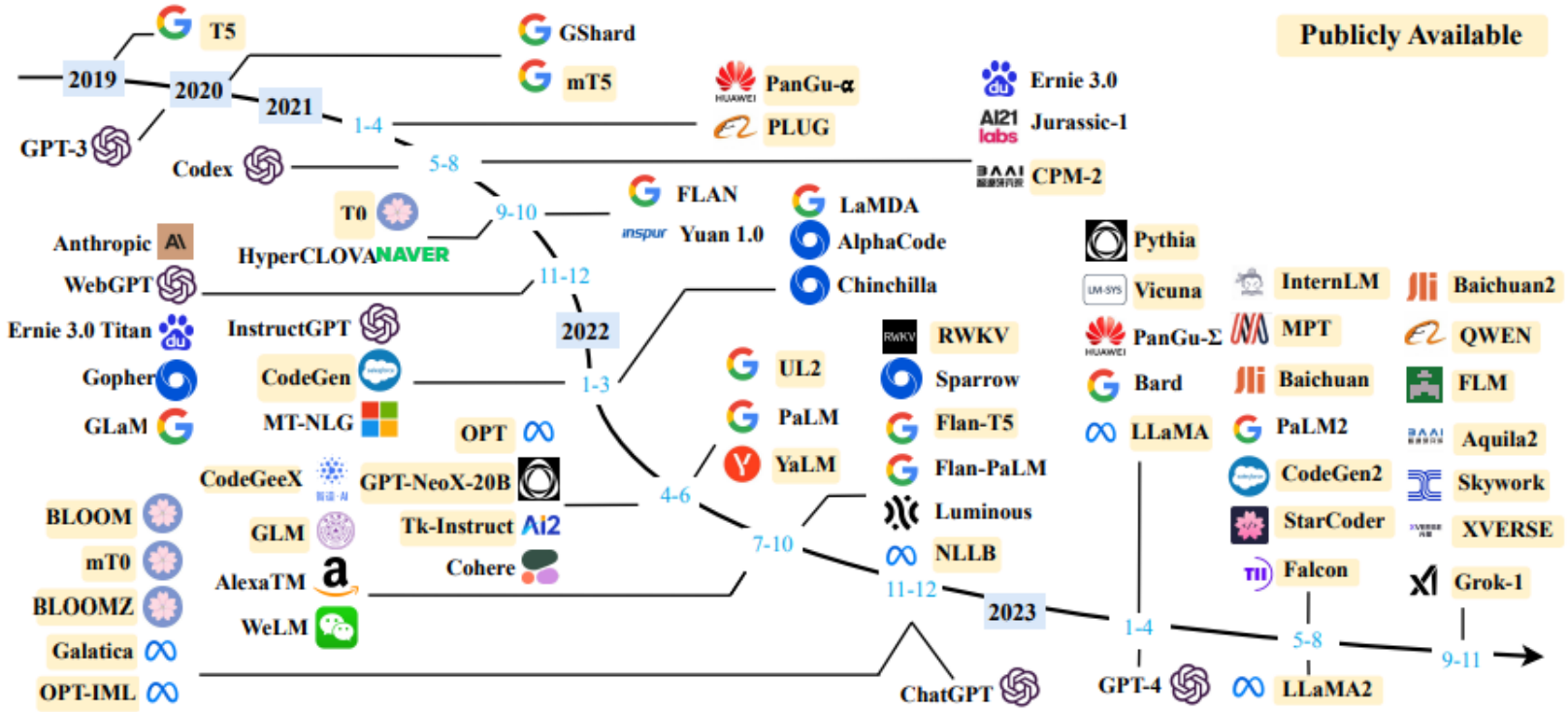
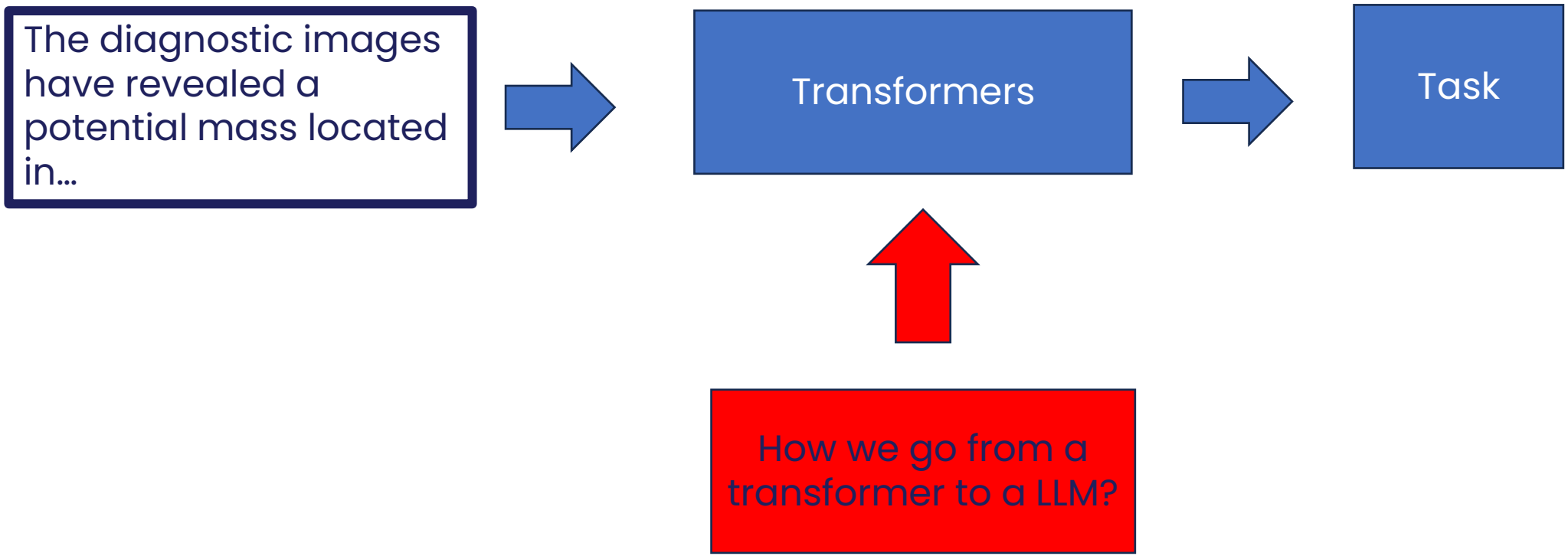


Fig. 3: A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.*, the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

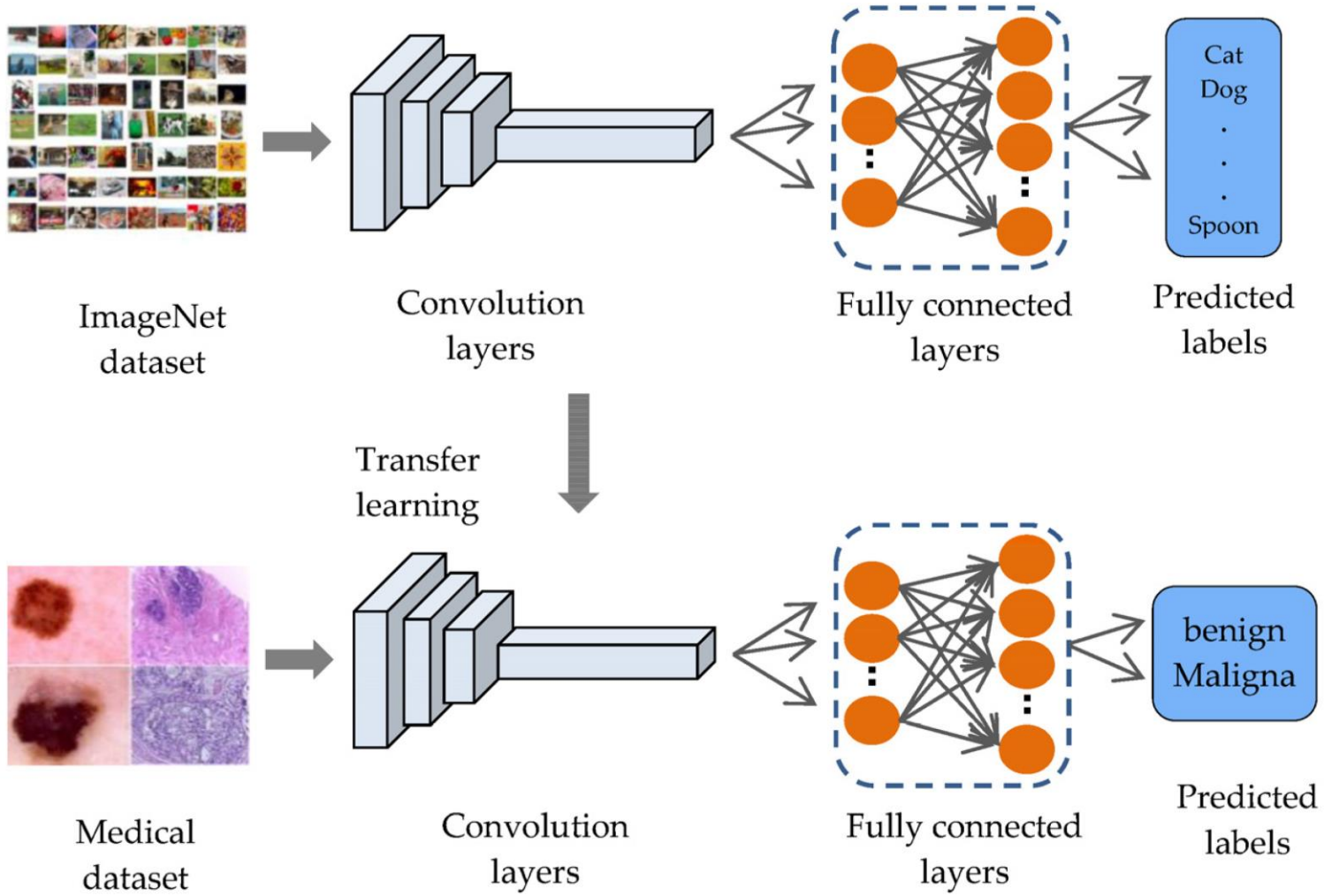
The long road to a Large Language Model

The key to growth is the introduction of higher dimensions of consciousness into our awareness. - Lao Tzu



Transfer learning: share what you know

- Originally observed with images
- When you learn a task, the knowledge can be reused for a new task
- Hierarchical representation is key



Train one to rule them all!

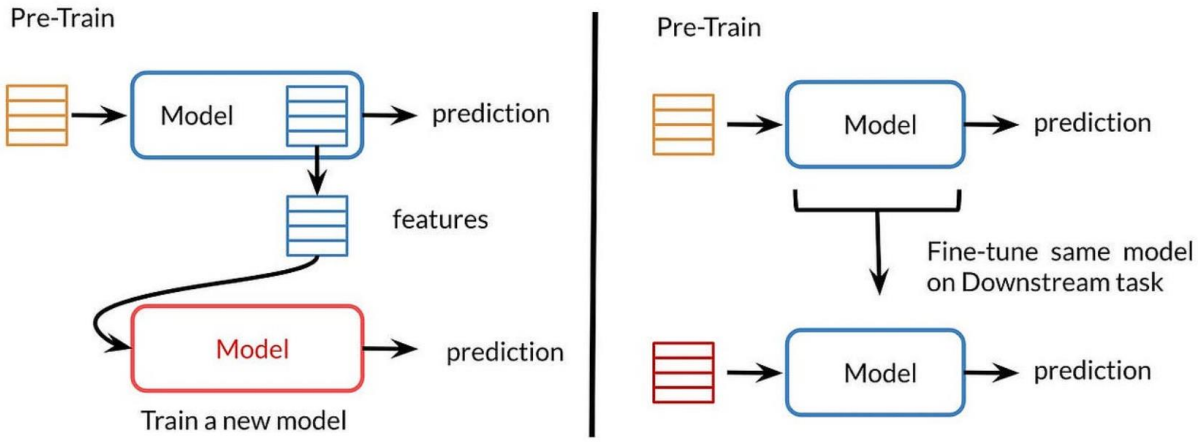
- A model trained in a task can be fine-tuned for many other tasks

text

Task 1

Task 2

Task 3

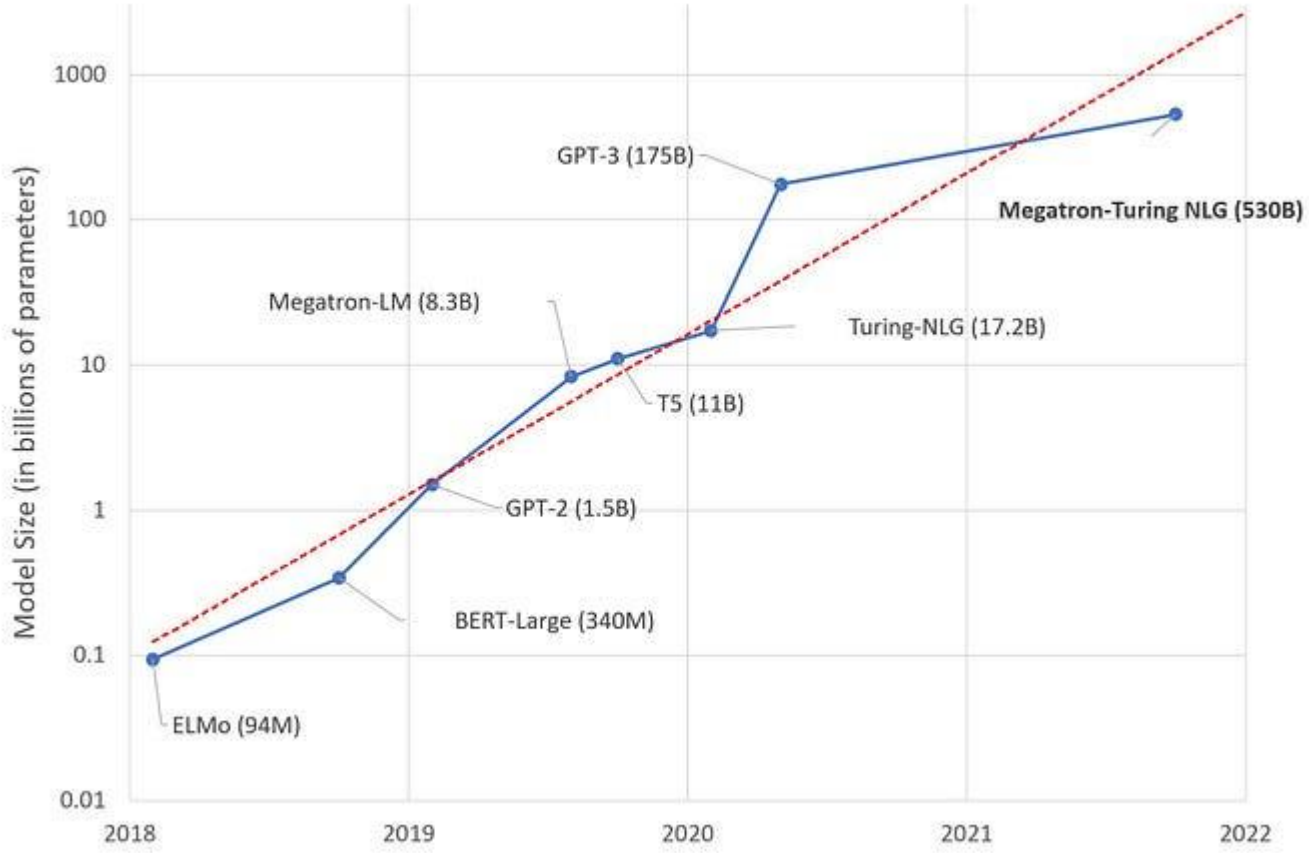


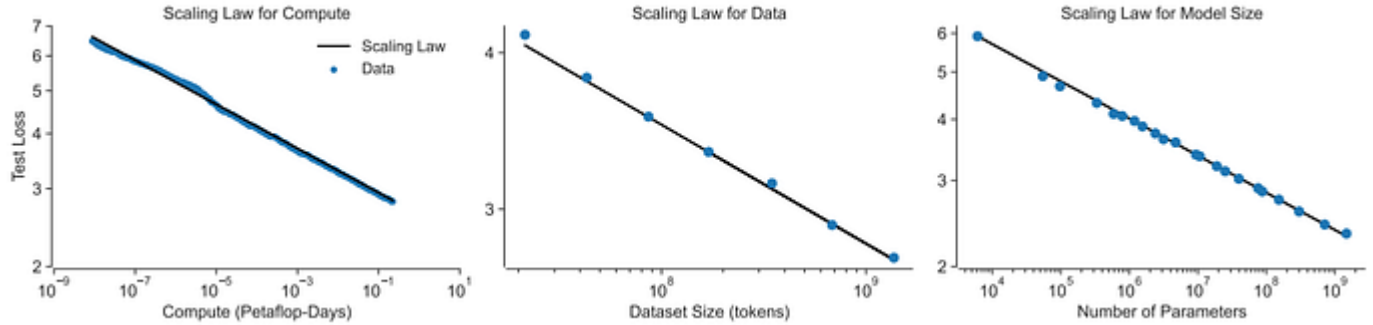
Bigger is better!

- OpenAi proposed the scaling law for LLMs
- The performance are dependant from training, dataset and model size

$$
\begin{aligned}
L(N) &= \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13} \quad (1)\\
L(D) &= \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13}\\
L(C) &= \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}
\end{aligned}
$$

- Scaling laws reliably predict that model performance





Emergent ability with scale

- **Sharpness**, the transition is discontinuous between being present or not present.
- **Unpredictability**, its appearance cannot be predicted as parameters increase

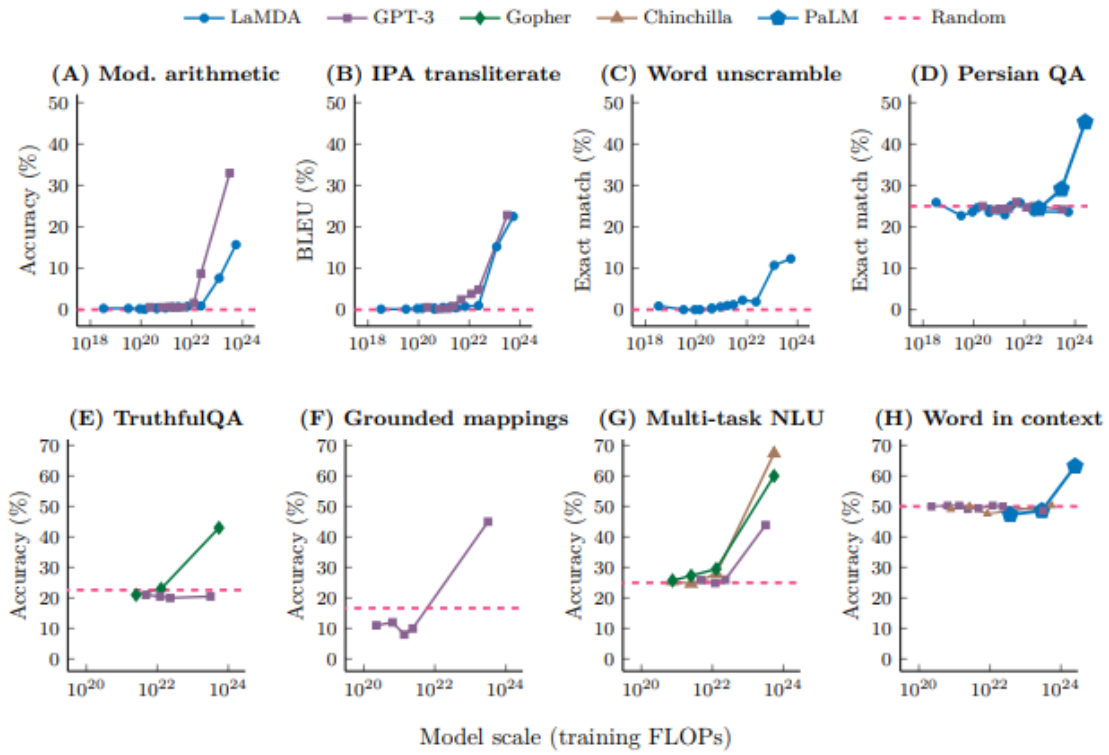


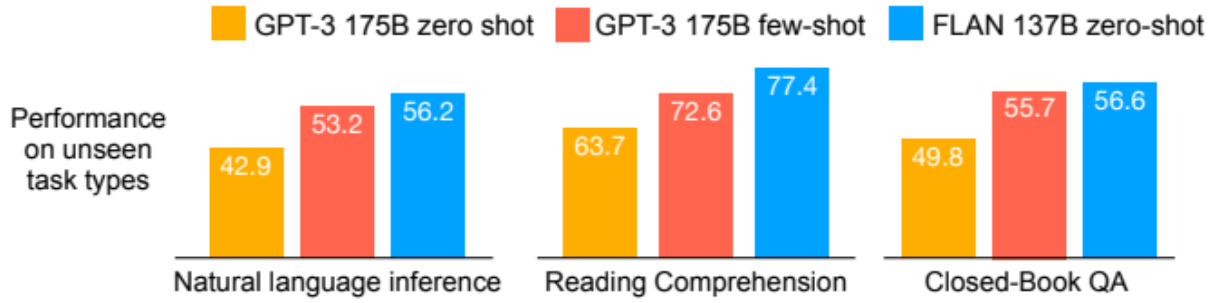
Table 1: List of emergent abilities of large language models and the scale (both training FLOPs and number of model parameters) at which the abilities emerge.

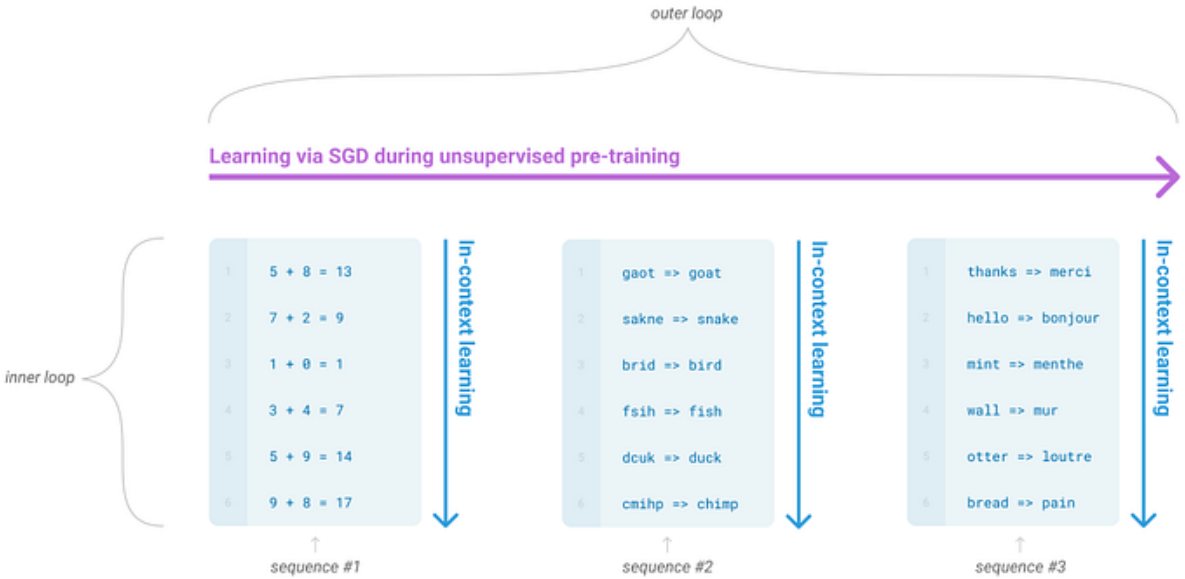
	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)

In-context learning

"The limits of my language mean the limits of my world." - Ludwig Wittgenstein

- In-context learning is a paradigm that allows language models to learn tasks given only a few examples in the form of demonstration.
- How this happen?





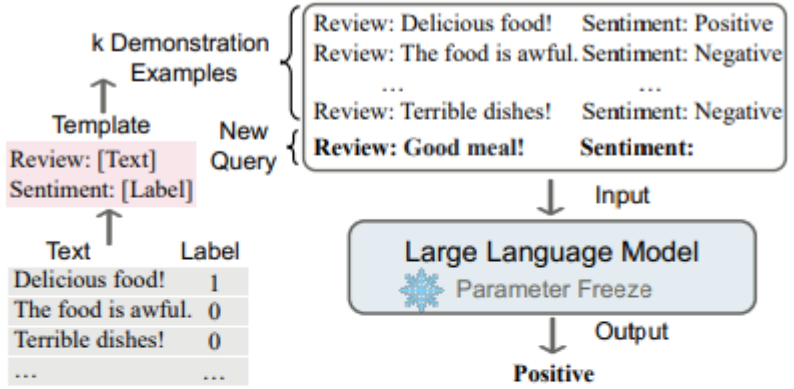


Figure 1: Illustration of in-context learning. ICL requires a piece of demonstration context containing a few examples written in natural language templates. Taking the demonstration and a query as the input, large language models are responsible for making predictions.

ICL: an toolbox for AI practitioners

- COT is a technique that is designed for complex tasks where reasoning is needed. we provided intermediate steps in order to guide the model from input to output

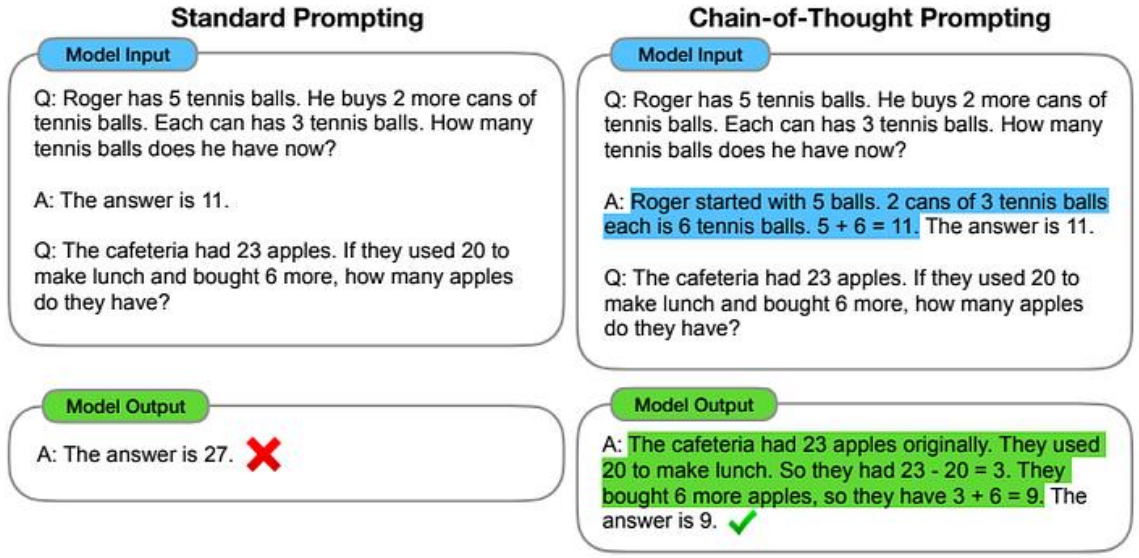
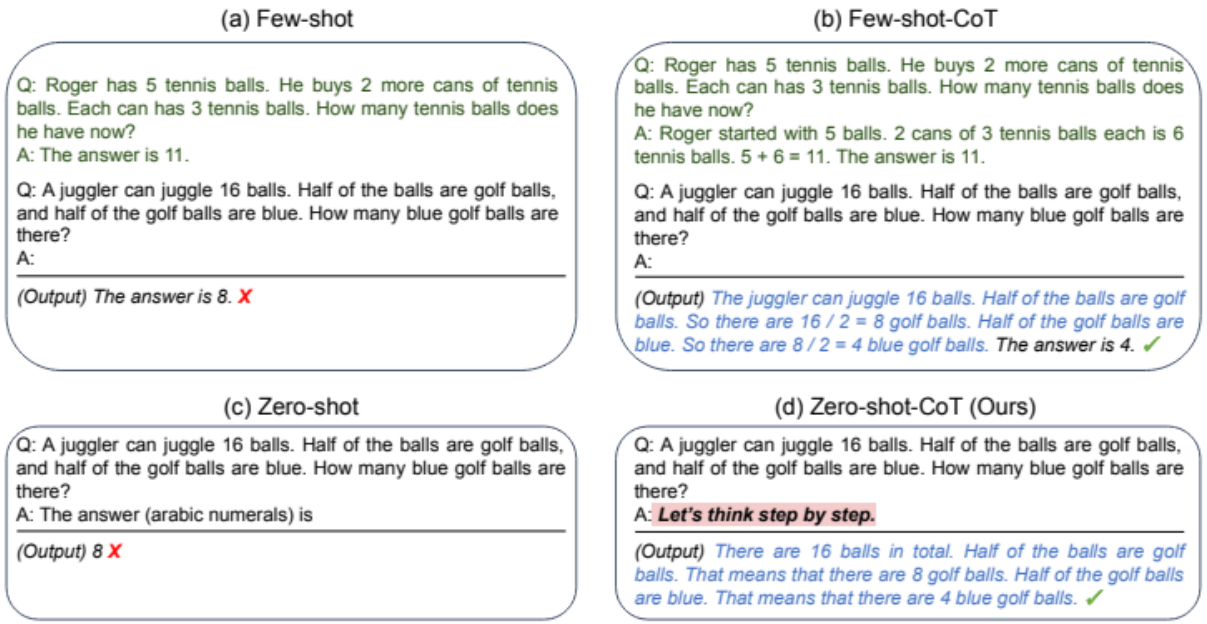


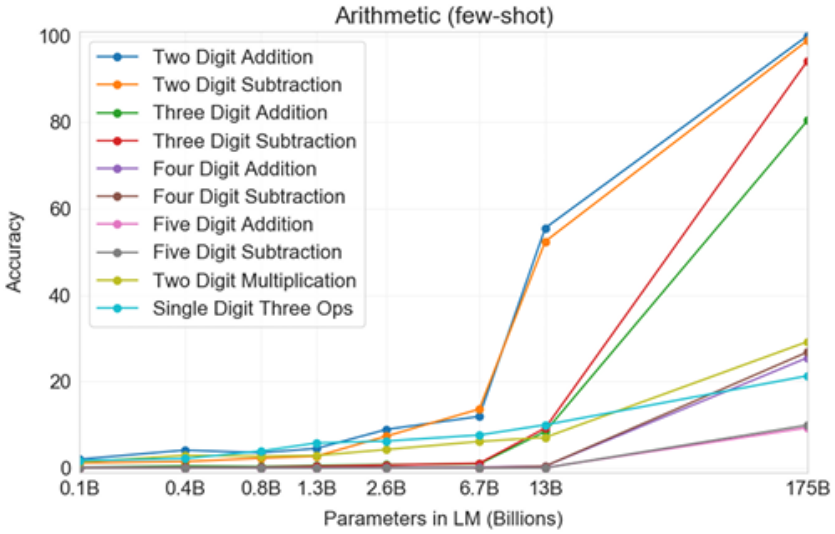
Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.



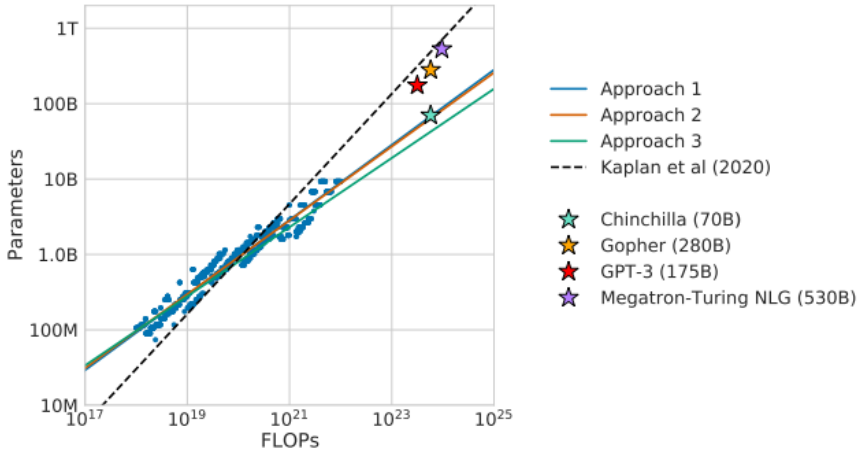
Prompt engineering is an emergent field

Bigger is always better?

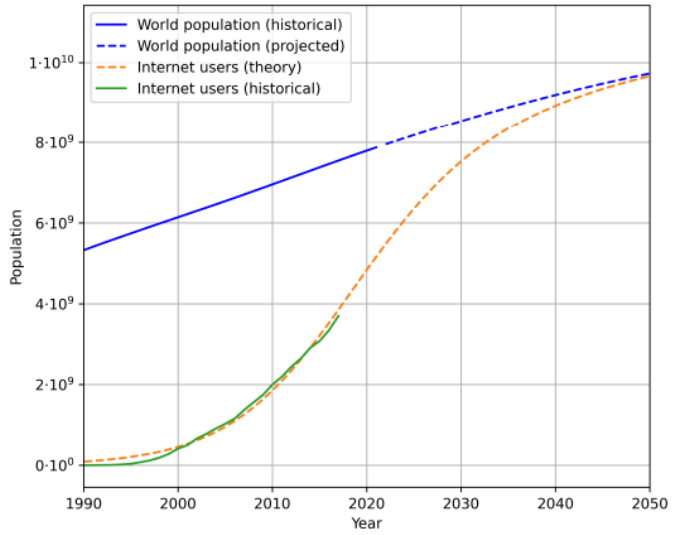
Emergent properties could be a mirage



LLMs are basically underfit

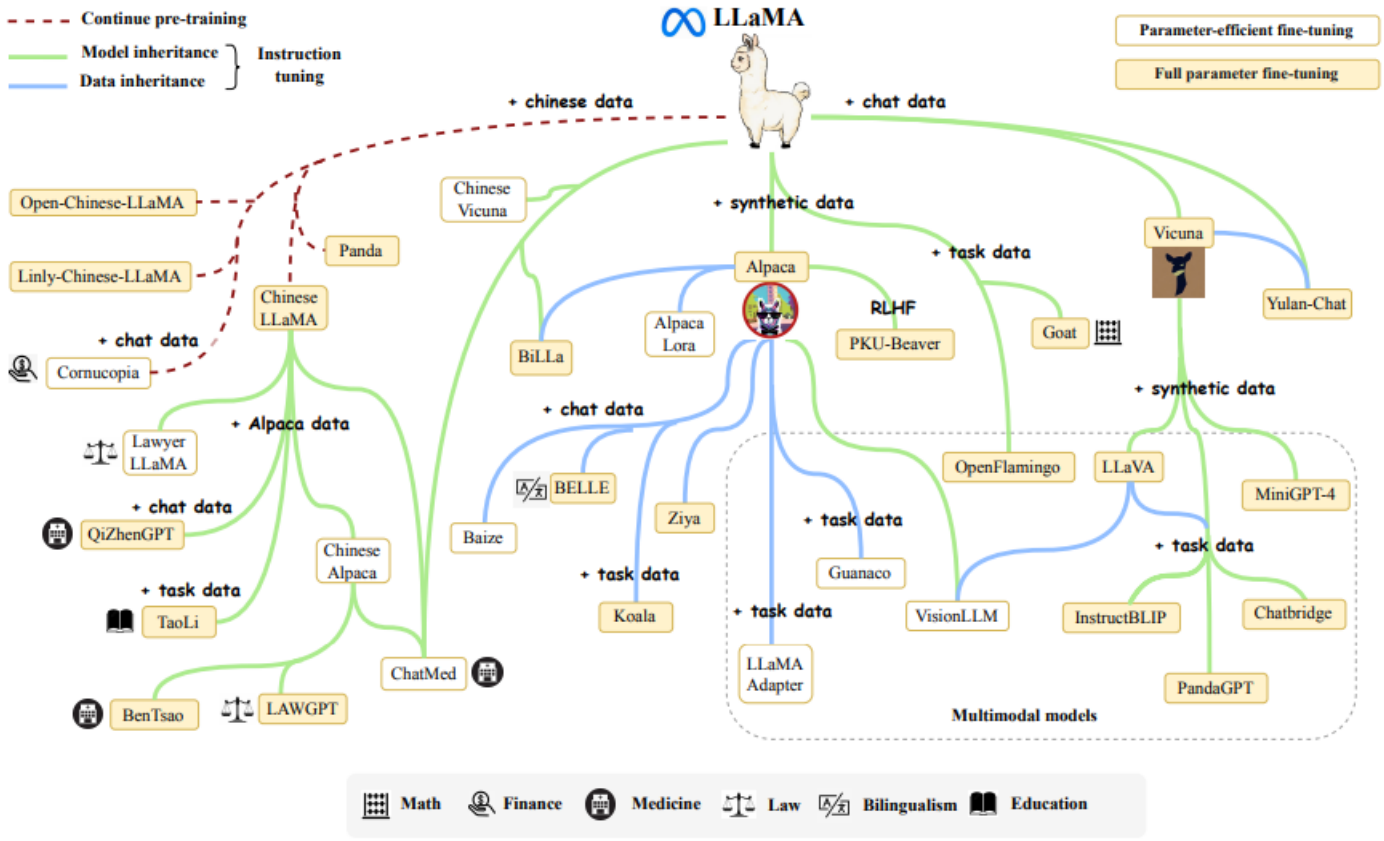


We do not have enough data to scale more



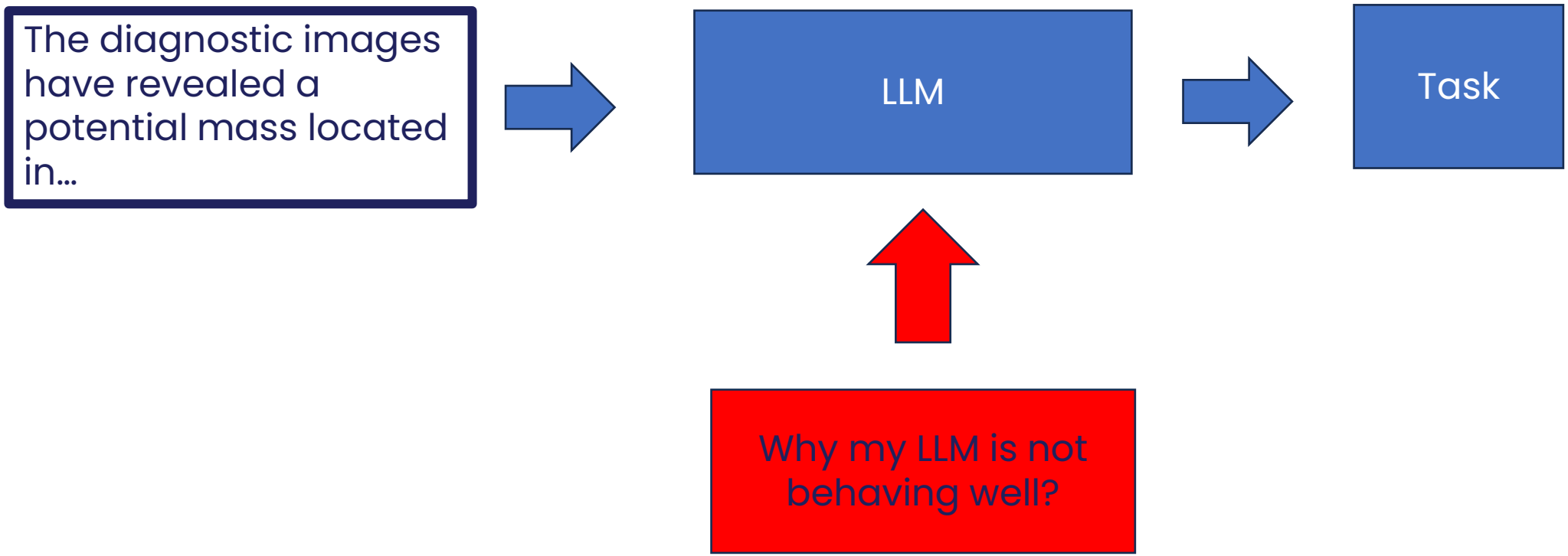
Quantity to impress, but Quality matters!

- AI-generated text is not optimal
- Quality tokens are more important that just random tokens



The long road to a Large Language Model

The essence of being human is that one does not seek perfection. - George Orwell



LLMs hallucination

- they exhibit an inclination to generate hallucinations resulting in seemingly plausible yet factually unsupported content.
- two main groups: factuality hallucination and faithfulness hallucination.

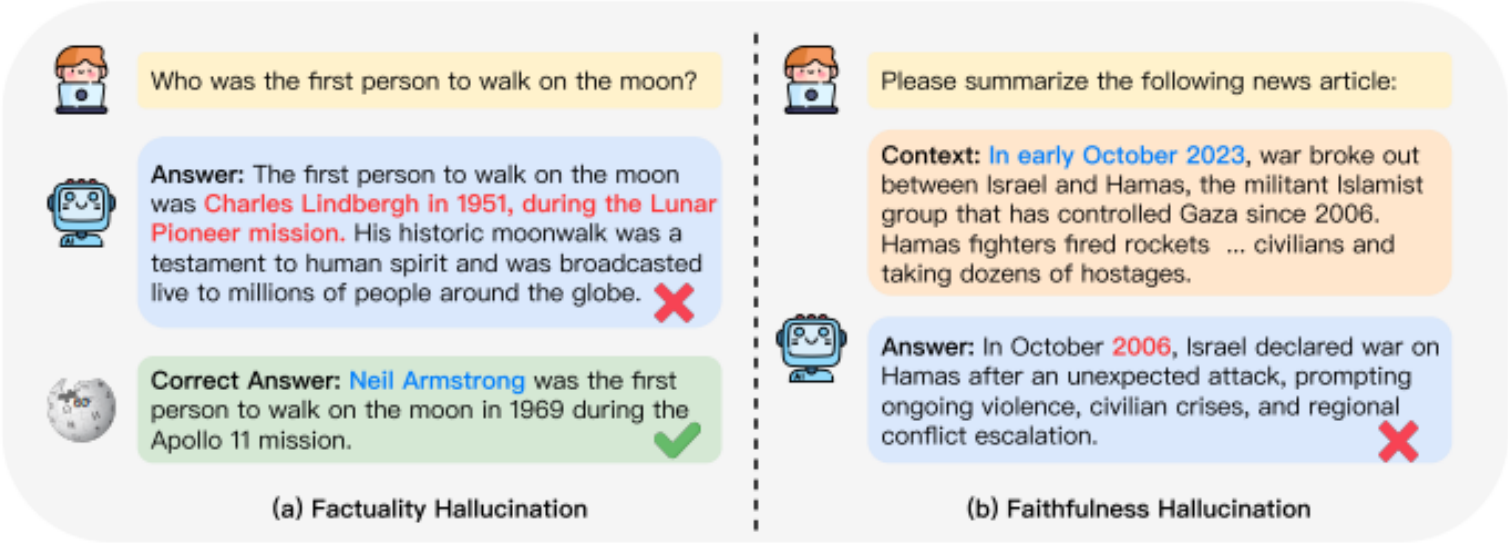


Figure 1: An intuitive example of LLM hallucination.

Harmful content

- LLMs can generate toxic response, showing problematic behaviours
- Typically trained on an enormous scale of uncurated Internet-based data, LLMs inherit stereotypes, misrepresentations, derogatory and exclusionary language, and other denigrating behaviors that disproportionately affect already-vulnerable and marginalized communities

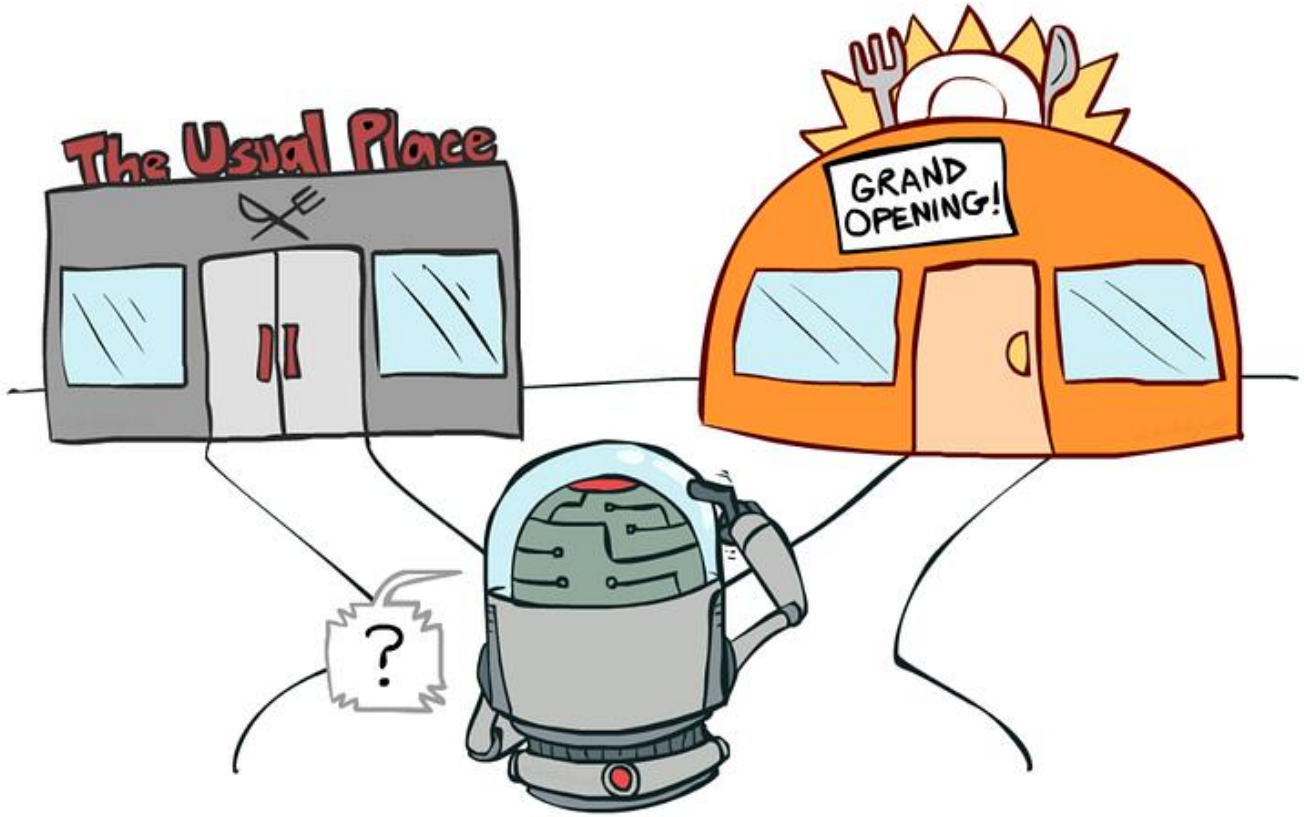
Table 1: **Taxonomy of Social Biases in NLP.** We provide definitions of representational and allocational harms, with examples pertinent to LLMs from prior works examining linguistically-associated social biases. Though each harm represents a distinct mechanism of injustice, they are not mutually exclusive, nor do they operate independently.

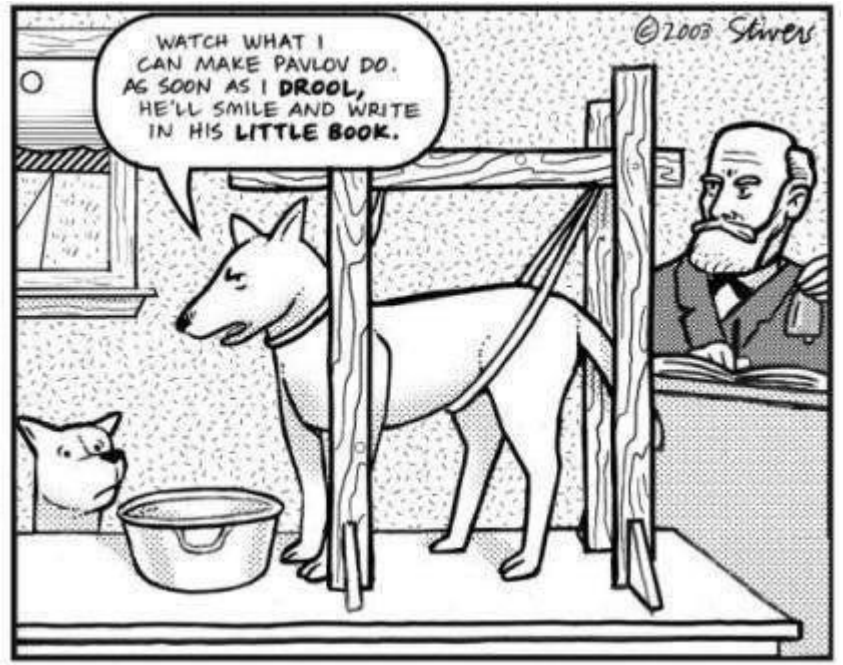
Type of Harm	Definition and Example
Representational Harms	Perpetuation of denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.,* `"Whore"` *conveys contempt of hostile female stereotypes* (Beukeboom & Burgers, 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than SAE† equivalents* (Blodgett & O'Connor, 2017)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups *e.g.,* `"Both genders"` *excludes non-binary gender identities* (Bender et al., 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group *e.g., Responding* `"I'm sorry to hear that"` *to* `"I'm an autistic dad"` *conveys a negative misrepresentation of autism* (Smith et al., 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group *e.g., Associating* `"Muslim"` *with* `"terrorist"` *perpetuates negative violent stereotypes* (Abid et al., 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group *e.g.,* `"I hate Latinos"` *is disrespectful, hateful, and unreasonable (Dixon et al., 2018)*
Allocational Harms	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group *e.g., LLM-aided resume screening may perpetuate inequities in hiring* (Ferrara, 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors *e.g., LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara, 2023)

*African-American English; †Standard American English

Preliminary: reinforcement learning

- Provide a reward to your model
- Balancing exploration versus exploitation





Align the LLMs with human values

- To align LLMs with human values, reinforcement learning from human feedback (RLHF) has been proposed to fine-tune LLMs with the collected human feedback data
- Although RLHF has achieved great success in aligning the behaviors of LLMs with human values and preferences, it also suffers from notable limitations
- Alternatives both in obtaining the dataset and methods for alignment

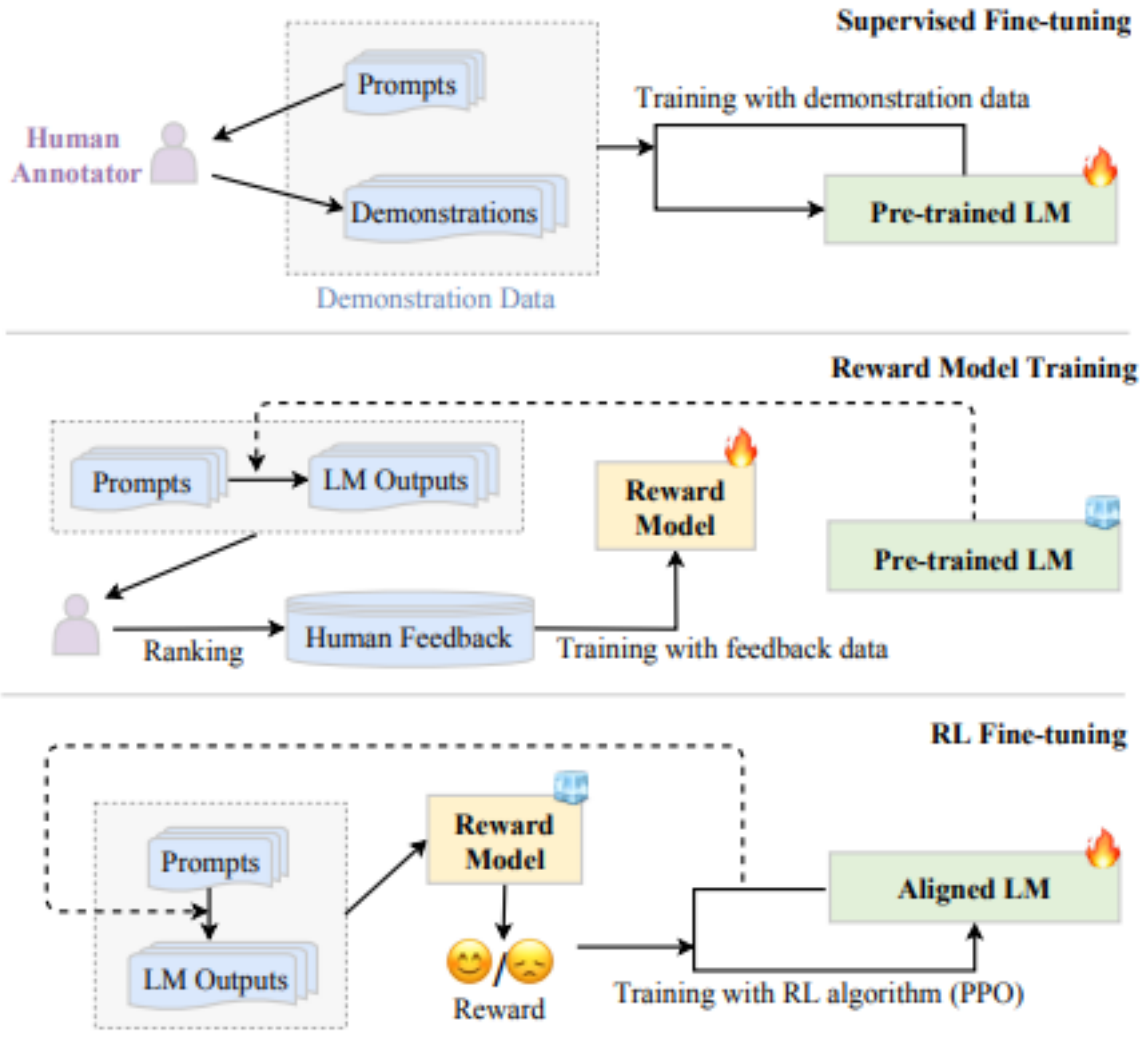
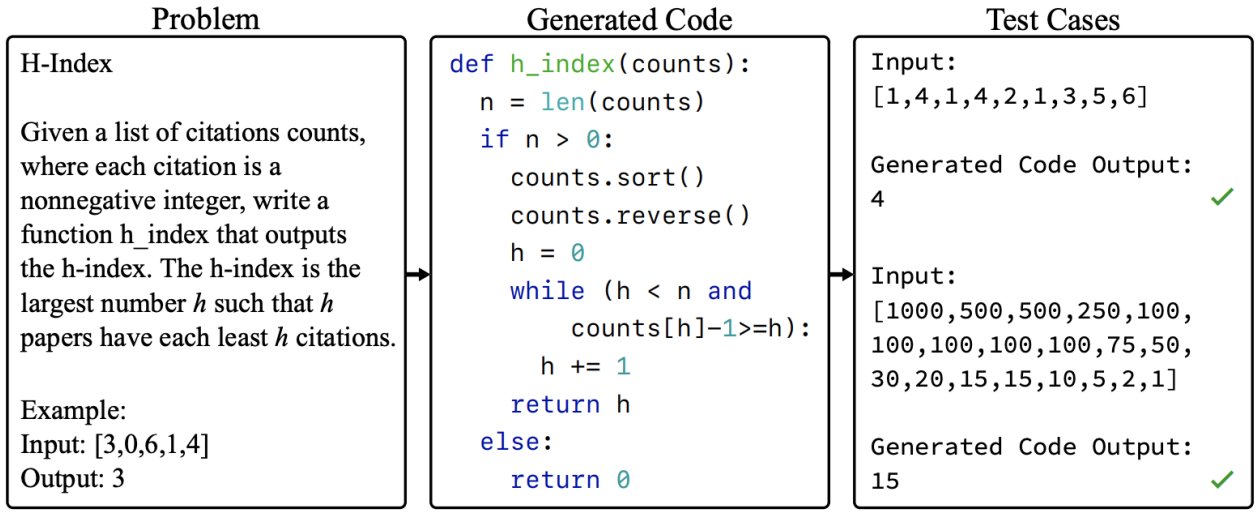
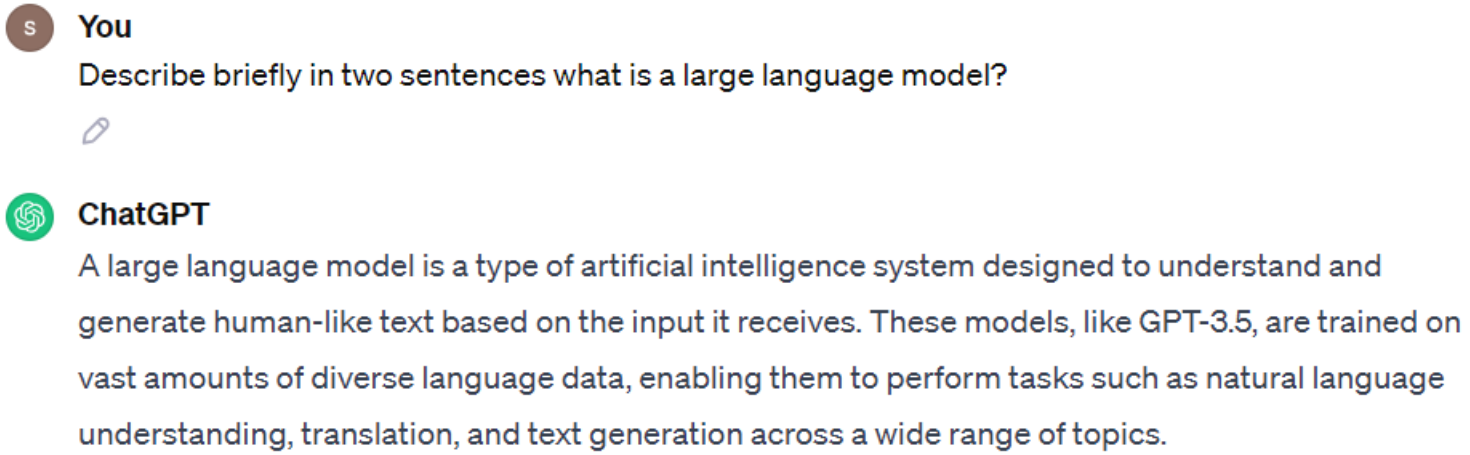


Fig. 12: The workflow of the RLHF algorithm.

LLMs basic ability

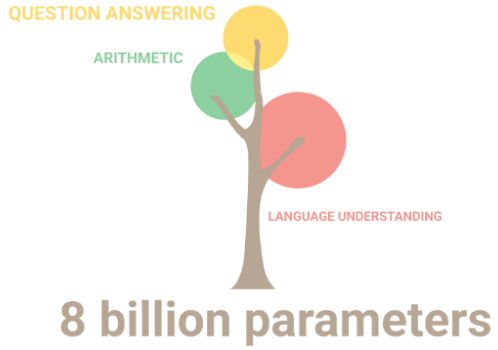
- Language Modeling, Conditional Text Generation, Code Synthesis, Knowledge Utilization (Closed-Book QA, Open-Book QA), Knowledge Completion





LLMs complex ability

- Knowledge Reasoning, Mathematical Reasoning, Symbolic Reasoning



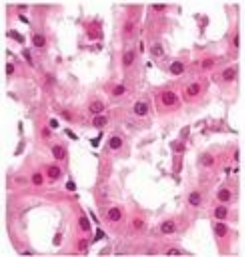
Give your model eyes!

- The Vision Transformer (ViT) model was introduced in 2021 in a conference research paper titled "An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale”

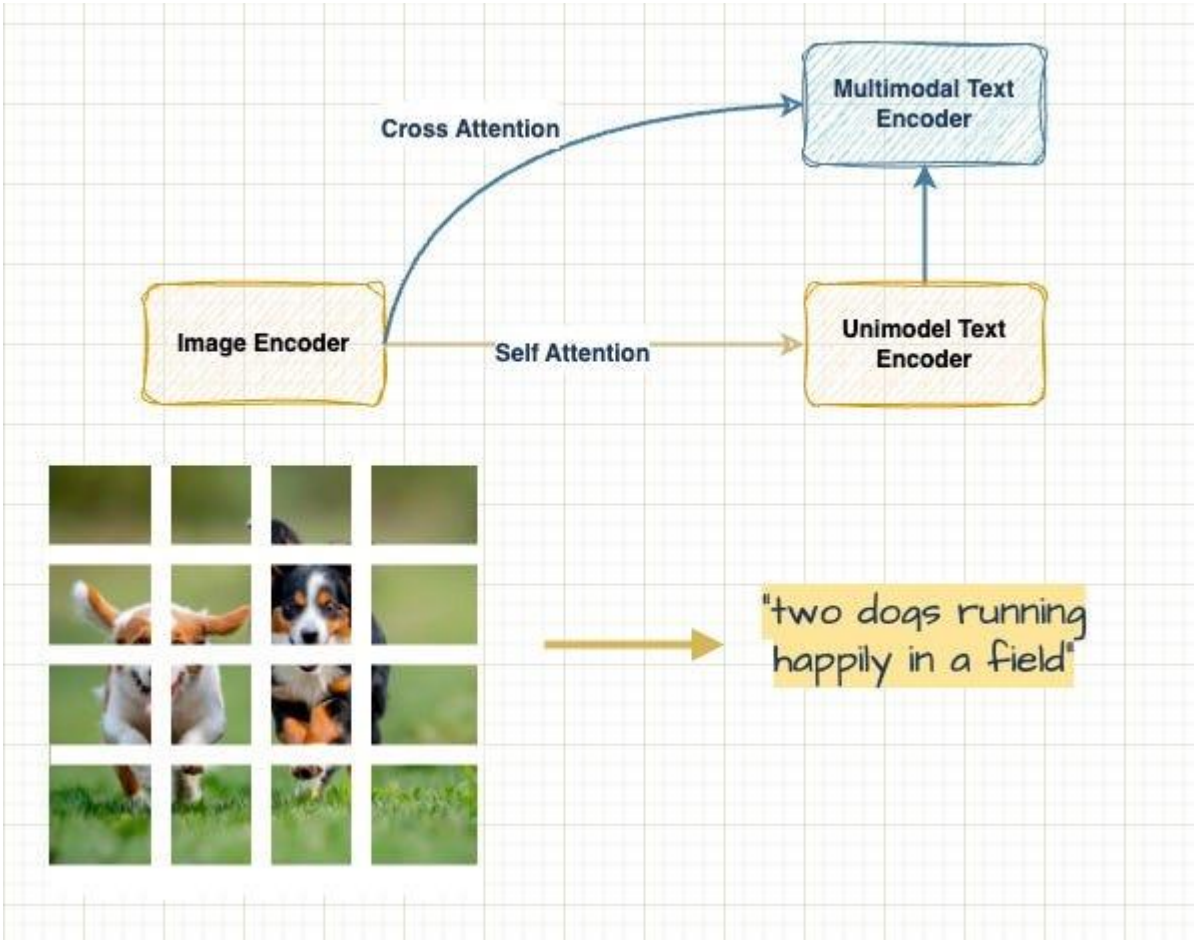
- In ViTs, images are represented as sequences, and class labels for the image are predicted, which allows models to learn image structure independently.

Give your model eyes!

- LLM fall short in decoding visual cues. these models can sometimes grapple with linguistic ambiguities and are handicapped when it comes to verifying their interpretations against real-world visual references

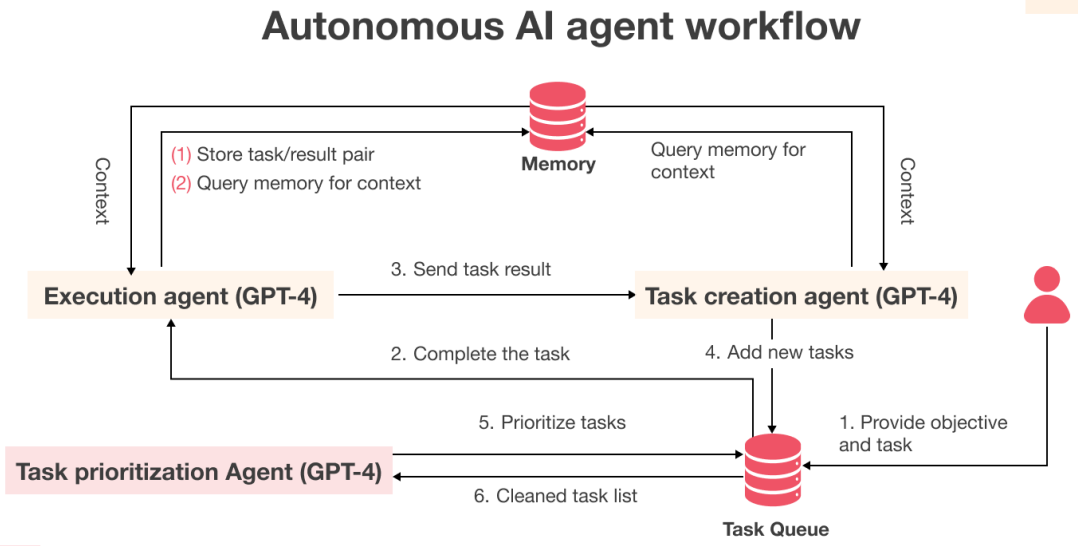


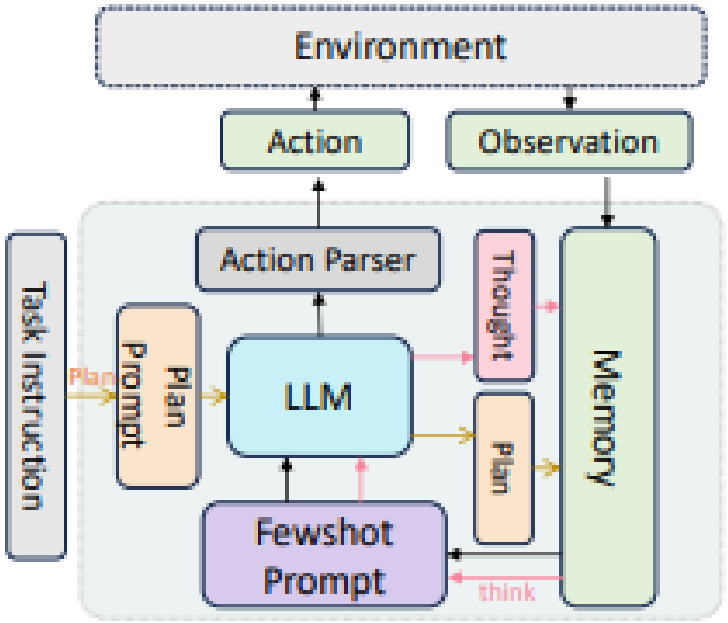
Questions	Answers
What does ischemic injury show?	surface blebs
Does early ischemic injury show surface blebs, increase eosinophilia of cytoplasm, and swelling of occasional cells?	Yes
What is showing increased eosinophilia of cytoplasm, and swelling of occasional cells?	early (reversible) ischemic injury
Did early (reversible) ischemic injury increase eosinophilia of cytoplasm, and swelling of occasional cells?	NO



Give your model hands!

- An AI agent is a software program designed to interact with its environment, perceive the data it receives, and take actions based on that data to achieve specific goals.
- LLMs can be connected with the world and with other models. It uses a large language model (LLM) as its central computational engine, allowing it to carry on conversations, do tasks, r

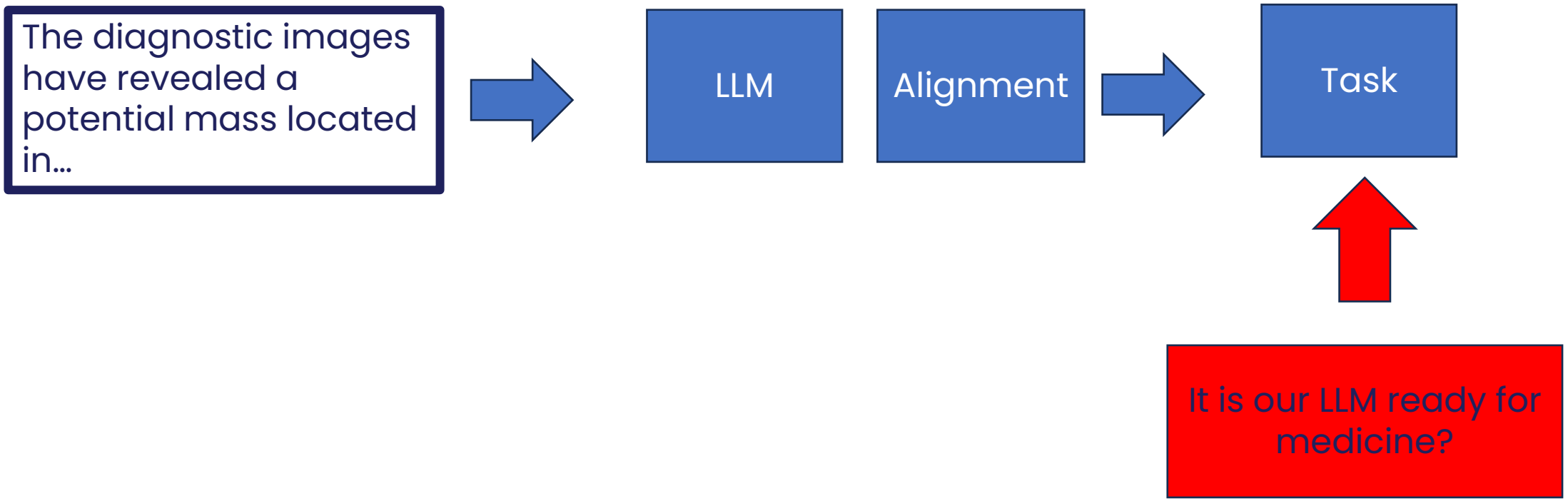




(a) PlanReAct LAA

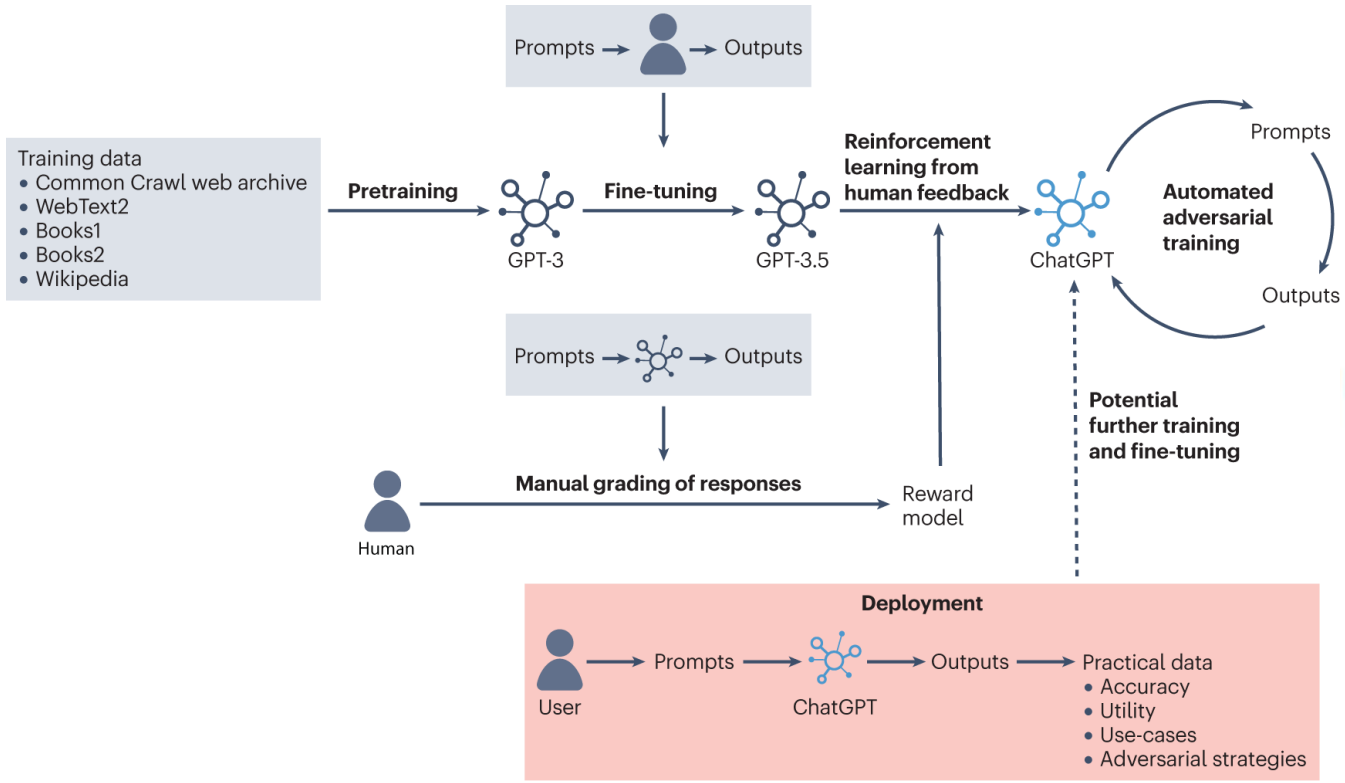
SIMFORM

The long road to a Large Language Model



Let's your model study more

- Generalist models have satisfying performance but it can be improved
- Fine tuning on specific data, LORA and other technique



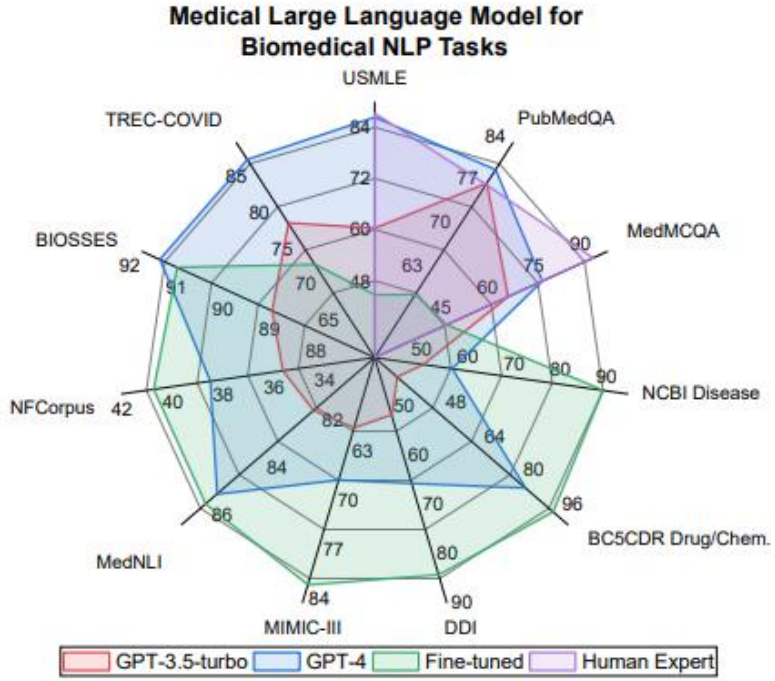


Figure 2: Performance comparison between the GPT-3.5 turbo, GPT-4, state-of-the-art task-specific fine-tuned models, and human experts, on seven downstream biomedical NLP tasks across eleven datasets. Please refer to Appendix B for details.

A growing list of medicine focused LLM

TABLE III
BRIEF SUMMARIZATION OF EXISTING LLMS FOR HEALTHCARE.

Model Name	Base	Para. (B)	Features	Date	Link
GatorTron [181]	Transformer	0.345, 3.9, 8.9	Training from scratch	06/2022	Github
Codex-Med [182]	GPT-3.5	175	CoT, Zero-shot	07/2022	Github
Galactica [38]	Transformer	1.3, 6.4, 30, 120	Reasoning, Multidisciplinary	11/2022	Org
Med-PaLM [99]	Flan-PaLM/PaLM	540	CoT, Self-consistency	12/2022	-
GPT-4-Med [183]	GPT-4	-	no specialized prompt crafting	03/2023	-
DeID-GPT [184]	GPT-4	-	De-identifying	03/2023	Github
ChatDoctor [116]	LLaMA	7	Retrieve online, external knowledge	03/2023	Github
DoctorGLM [185]	ChatGLM	6	Extra prompt designer	04/2023	Github
MedAlpaca [186]	LLaMA	7, 13	Adapt to Medicine	04/2023	Github
BenTsao [187]	LLaMA	7	Knowledge graph	04/2023	Github
PMC-LLaMA [188]	LLaMA	7	Adapt to Medicine	04/2023	Github
Visual Med-Alpaca [45]	LLaMA	7	multimodal generative model, Self-Instruct	04/2023	Github
BianQue [189]	ChatGLM	6	Chain of Questioning	04/2023	Github
Med-PaLM 2 [16]	PaLM 2	340	Ensemble refinement, CoT, Self-consistency	05/2023	-
GatorTronGPT [190]	GPT-3	5, 20	Training from scratch for medicine	05/2023	Github
HuatuoGPT [44]	Bloomz	7	Reinforced learning from AI feedback	05/2023	Github
ClinicalGPT [191]	BLOOM	7	multi-round dialogue consultations	06/2023	-
MedAGI [192]	MiniGPT-4	-	multimodal, AGI	06/2023	Github
LLaVA-Med [193]	LLaVA	13	multimodal, self-instruct, curriculum learning	06/2023	Github
OphGLM [194]	ChatGLM	6	multimodal, Ophthalmology LLM	06/2023	Github
SoulChat [195]	ChatGLM	6	Mental Healthcare	06/2023	Github
Med-Flamingo [196]	Flamingo	80B	multimodal, Few-Shot generative medical VQA	07/2023	Github

TABLE IV
SUMMARIZATION OF TRAINING DATA AND EVALUATION TASKS FOR EXISTING LLMS FOR HEALTHCARE.

Model Name	Method	Training Data	Eval datasets
GatorTron [181]	PT	Clinical notes	CNER, MRE, MQA
Codex-Med [182]*	ICL	-	USMLE, MedMCQA,PubMedQA
Galactica [38]	PT, IFT	DNA sequence	MedMCQA, PubMedQA, Medical Genetics
Med-PaLM [99]	IPT	Medical data	MultiMedQA, HealthSearchQA
GPT-4-Med [183]*	ICL	-	USMLE, MultiMedQA
DeID-GPT [184]*	ICL	-	i2b2/UTHealth de-identification task
ChatDoctor [116]	IFT	Patient-doctor dialogues	iCliniq
DoctorGLM [185]	IFT	Chinese medical dialogues	-
MedAlpaca [186]	IFT	Medical dialogues and QA	USMLE, Medical Meadow
BenTsao [187]	IFT	Medical knowledge graph, Medical QA	Customed medical QA
PMC-LLaMA [188]	IFT	Biomedical academic papers	PubMedQA, MedMCQA, USMLE
Visual Med-Alpaca [45]	PT, IFT	medical QA	-
BianQue [189]	IFT	medical QA	-
Med-PaLM 2 [16]	IFT	-	MultiMedQA, Long-form QA
GatorTronGPT [190]	PT	Clinical and general text	PubMedQA, USMLE, MedMCQA, DDI, BC5CDR, KD-DTI
HuatuoGPT [44]	IFT	Instruction and Conversation Data	CmedQA, webmedQA, and Huatuo26M
ClinicalGPT [191]	IFT+RLHF	Medical dialogues and QA, EHR	MedDialog, MEDQA-MCMLE, MD-EHR, cMedQA2
MedAGI [192]	IFT	Public medical datasets and images	SkinGPT-4, XrayChat, PathologyChat
LLaVA-Med [193]	IFT	multimodal biomedical instruction	VQA-RAD, SLAKE, PathVQA
OphGLM [194]	IFT	Knowledge graphs, medical dialogues	Fundus diagnosis pipeline tasks [194]
SoulChat [195]	IFT	Long text, empathetic dialogue	-
Med-Flamingo [196]	IFT	Image-caption/tokens pairs	VQA-RAD, Path-VQA, Visual USMLE

☆ * means the study focuses on evaluating the Healthcare LLM, rather than proposing a new LLM. PT means pre-training, ICL means In-context-learning (no parameters updated), IFT means instruction fine-tuning, and IPT means instruction prompt tuning. IPT comes from [99]. It should be noted that this concept is slightly different from Instruction fine-tuning or supervised fine-tuning.

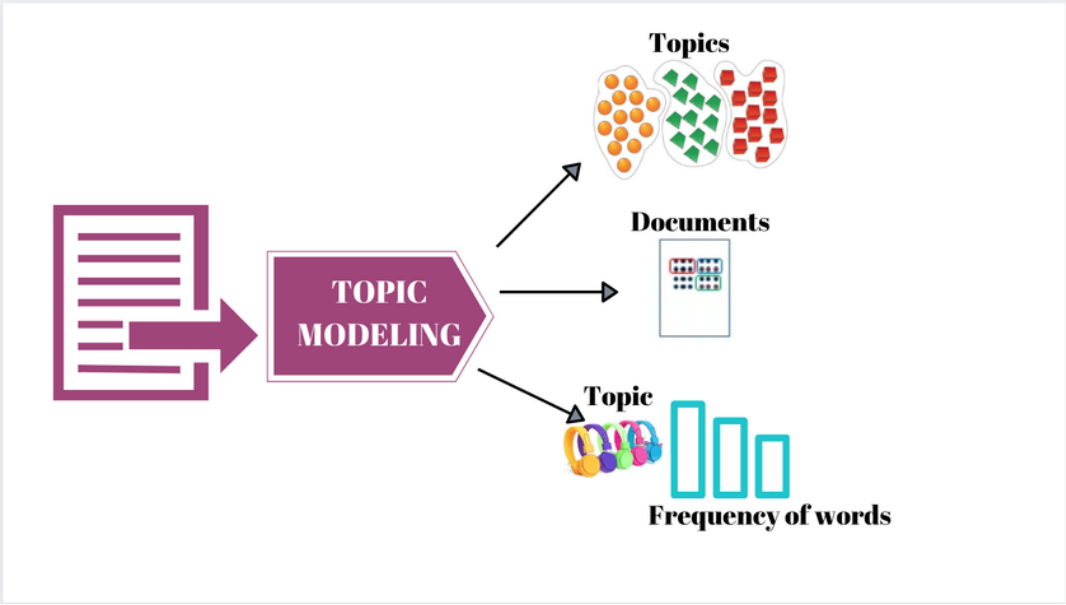
First wave of LM in medicine

- Transformers trained on medical domain datasets
- Used for limited cases and tasks

Model Development	Models	# Params	Data Scale	Data Source
Pre-training (Sec. 2.1)	BioBERT [28, 68]	110M	18B tokens	PubMed [37]
	PubMedBERT [29]	110M/340M	3.2B tokens	PubMed [37]
	SciBERT [30]	110M	3.17B tokens	Literature [38]
	ClinicalBERT [31]	110M	112k clinical notes	MIMIC-III [39]
	BlueBERT [69, 70, 71]	110M/340M	>4.5B tokens	PubMed [37] MIMIC-III [39]
	BioCPT [72]	330M	255M articles	PubMed [37]
	BioGPT [73]	1.5B	15M articles	PubMed [37]
	BioMedLM [74]	2.7B	110GB	PubMed [75]
	OphGLM[76]	6.2B	20k dialogues	MedDialog [40]
	GatorTron [77, 24]	8.9B	>82B tokens 6B tokens 2.5B tokens+0.5B tokens	EHRs [24] PubMed [37] Wiki+MIMIC-III [39]
	GatorTronGPT[78]	5B/20B	277B tokens	EHRs [78]
	MEDITRON [79]	70B	48.1B tokens	PubMed [37] Clinical Guidelines [79]

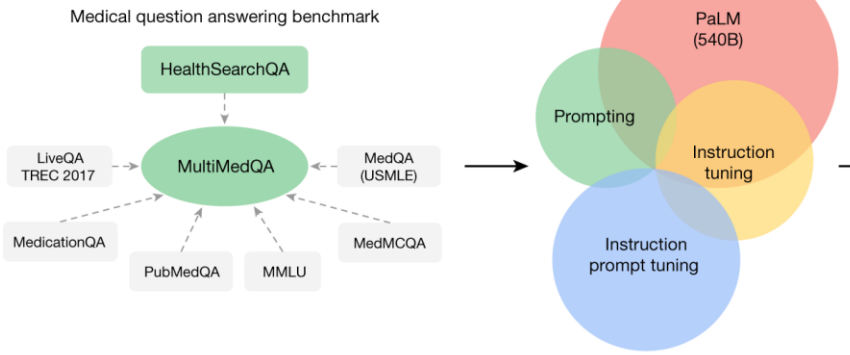
Type of tasks

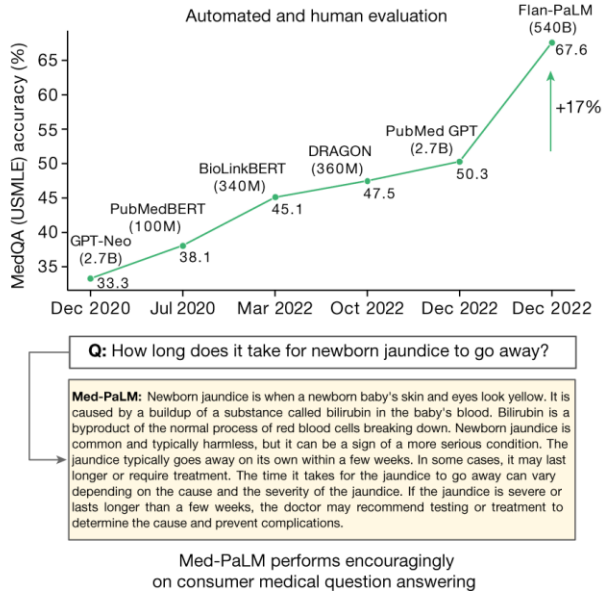
- Named entity recognition
- Relation extraction
- Text classification
- Document Ranking
- Keyword Extraction



MED-Palm: The AI clinician

- Developing AI that can answer medical questions accurately has been a long-standing challenge with several research advances over the past few decades.

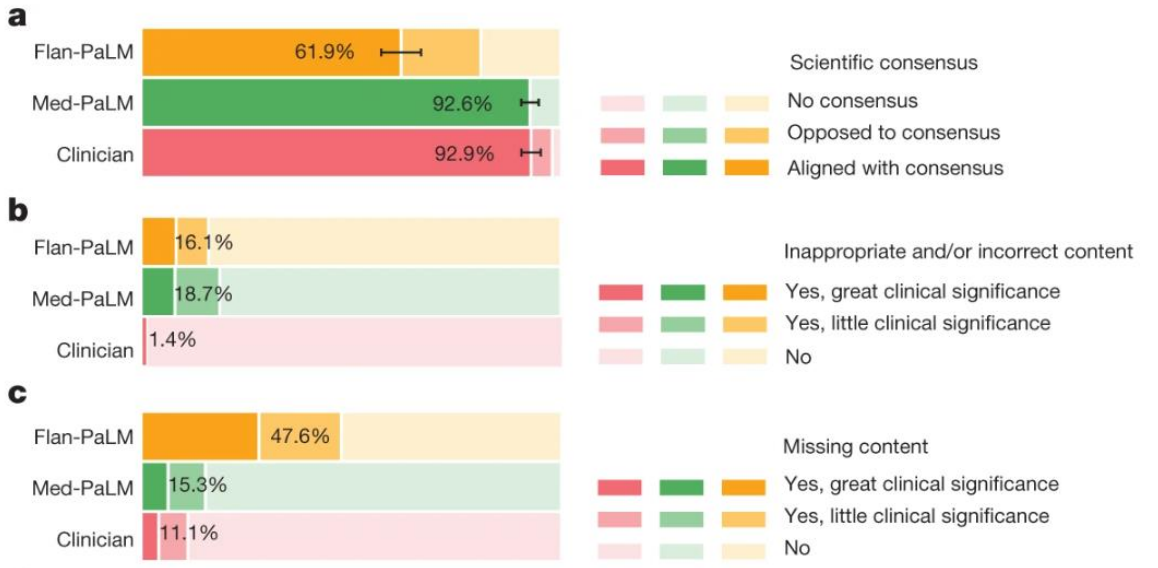




Med-PaLM performs encouragingly on consumer medical question answering

Question
How do you know if ear pain is serious?

Med-PaLM
Ear pain can be a sign of several underlying conditions, including: Middle ear infection (otitis media), which is commonly caused by a viral infection and can cause pain, redness, and swelling of the ear. External ear infection (otitis externa), which is commonly caused by a bacterial infection and can cause pain, redness, and swelling of the ear. Earwax impaction, which can cause pain, a feeling of fullness in the ear, and decreased hearing. Ear pain can also be a sign of a more serious condition, such as a brain tumor or a stroke. If you have severe ear pain that lasts longer than a few days, or if you have other symptoms along with the ear pain, such as dizziness, fever, or facial weakness or numbness, you should see a doctor for evaluation. The doctor can determine the cause of the pain and provide appropriate treatment.



Med PaLM2: Google's new doctor

- A new more performing model focused on long answer
- Focused on safety
- Longer and better answers for human evaluation

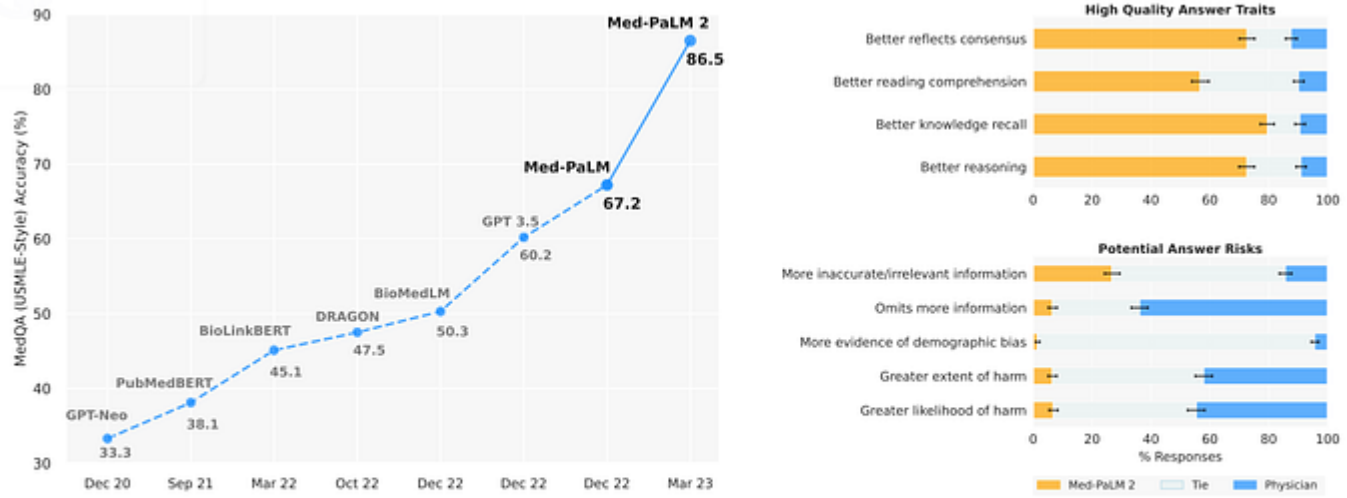


Figure 1 | Med-PaLM 2 performance on MultiMedQA Left: Med-PaLM 2 achieved an accuracy of 86.5% on USMLE-style questions in the MedQA dataset. Right: In a pairwise ranking study on 1066 consumer medical questions, Med-PaLM 2 answers were preferred over physician answers by a panel of physicians across eight of nine axes in our evaluation framework.

Google Med-PaLM M: Towards the Medical AI Generalist

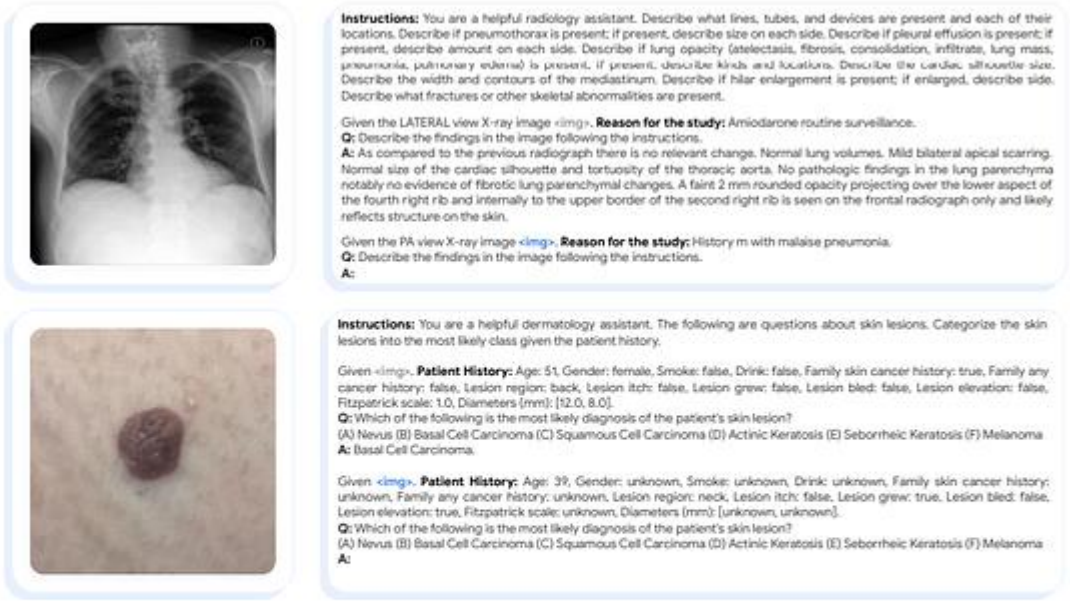


Figure 2 | Illustration of instruction task prompting with one-shot exemplar. (top) shows the task prompt for the chest X-ray report generation task. It consists of task-specific instructions, a text-only "one-shot exemplar" (omitting the corresponding image but preserving the target answer), and the actual question. The X-ray image is embedded and interleaved with textual context including view orientation and reason for the study in addition to the question. (bottom) shows the task prompt for the dermatology classification task. We formulate the skin lesion classification task as a multiple choice question answering task with all the class labels provided as individual answer options. Similar to the chest X-ray report generation task, skin lesion image tokens are interleaved with the patient clinical history as additional context to the question. The blue <img> denotes the position in the prompt where the image tokens are embedded.

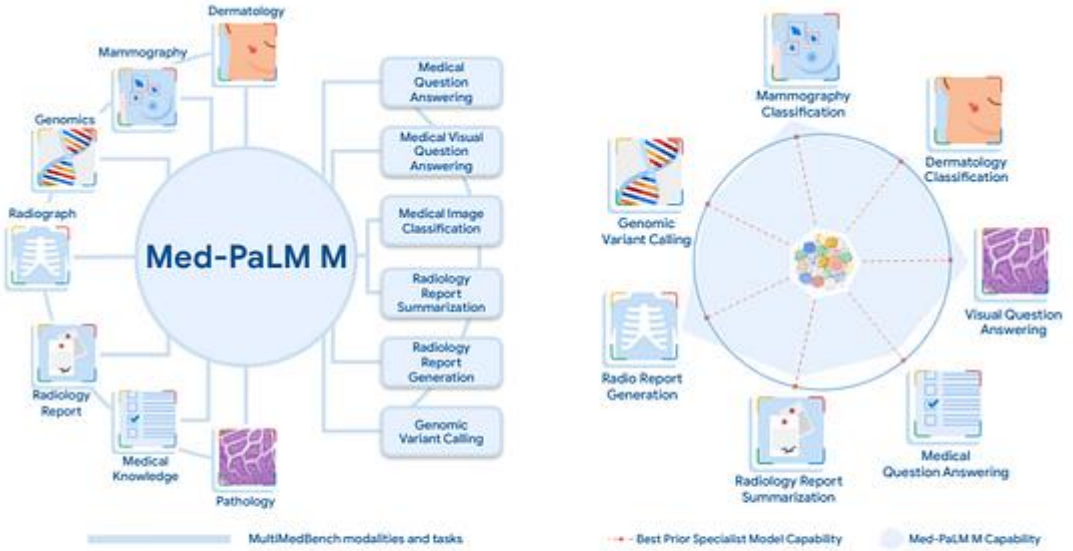


Figure 1 | Med-PaLM M overview. A generalist biomedical AI system should be able to handle a diverse range of biomedical data modalities and tasks. To enable progress towards this overarching goal, we curate MultiMedBench, a benchmark spanning 14 diverse biomedical tasks including question answering, visual question answering, image classification, radiology report generation and summarization, and genomic variant calling. Med-PaLM Multimodal (Med-PaLM M), our proof of concept for such a generalist biomedical AI system (denoted by the shaded blue area) is competitive with or exceeds prior SOTA results from specialists models (denoted by dotted red lines) on all tasks in MultiMedBench. Notably, Med-PaLM M achieves this using a single set of model weights, without any task-specific customization.

- First multimodal
- Evidence of new emerging capabilities in Med-PaLM M

PMC-LLaMA

- Open source model
- Similar approach to Stanford Alpaca
- Fine-tuning makes the model better than the generalist model

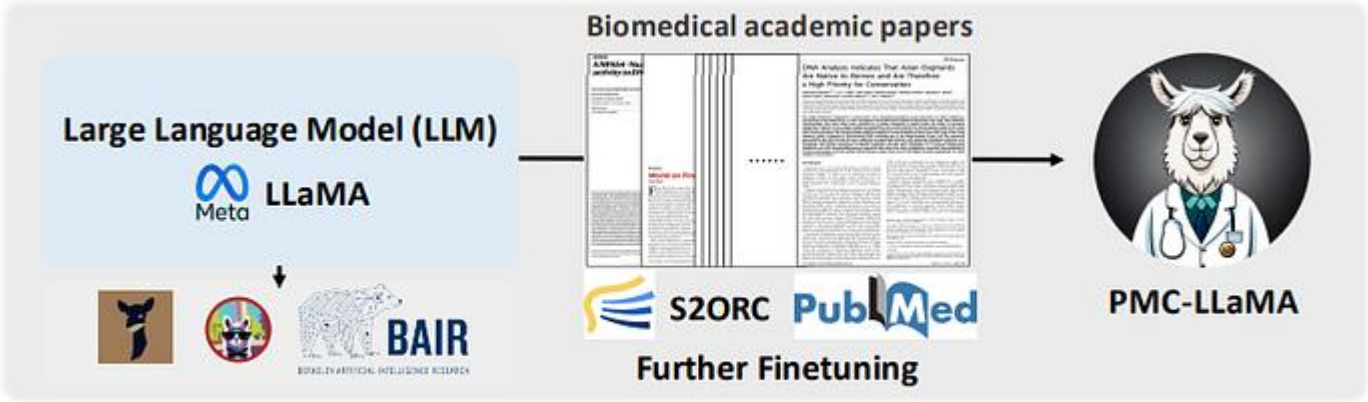


Figure 1: PMC-LLaMA Training Pipeline.

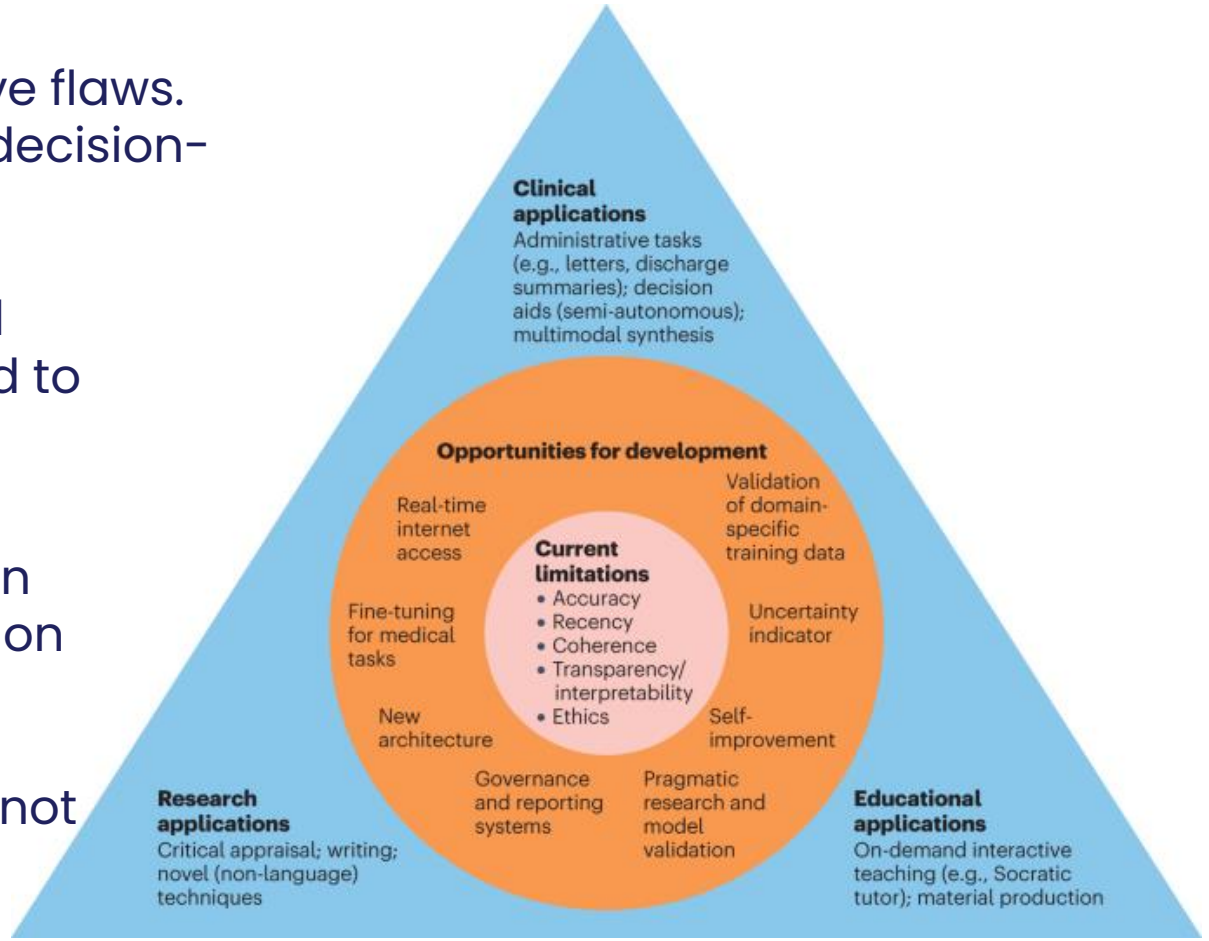
Cases	GPT-4 Judgment
Case1: ***Input:*** *For example, chronic obstructive pulmonary disease (COPD)* ***PMC-LLaMA Output:*** For example, chronic obstructive pulmonary disease (COPD) is a common lung disease characterized by **shortness of breath and chronic airflow limitation**. Its prevalence is increasing considerably as the population ages and is the **third leading cause of death worldwide**. COPD, a preventable, and treatable disease, is a leading cause of disability for adults ages 40-60 years worldwide while being the top reason for a reduction in health-related quality of life in 12 countries. ***Original LLaMA Output:*** For example, chronic obstructive pulmonary disease (COPD) can be identified by a change in the patient's **spirometry** results. A change in the patient's blood values, such as serum lactate or creatinine, can indicate illness. A decline in motor control after a stroke can be demonstrated by the evaluation of the patient's ability to perform physical therapy.	The PMC-LLaMA Output is more related to the input. It provides a detailed description of chronic obstructive pulmonary disease (COPD), including its prevalence, causes, effects on quality of life, and its status as a preventable and treatable disease. The Original LLaMA Output briefly mentions COPD in relation to spirometry results, but it then goes on to discuss unrelated topics such as blood values and stroke recovery, which are not directly related to the input.

Fine tuning is important

Falcon-7bs	
Dataset n°0 (alpaca)	**Fine tuning**
What is pancreatic cancer?	
Pancreatic cancer is a type of cancer that develops in the pancreas, a gland that is part of the digestive system. Pancreatic cancer can spread from the pancreas to other parts of the body, making it difficult to treat. It is the fourth most common cause of cancer-related deaths in the United States.	Pancreatic cancer is an uncommon disease that makes up only 2% of the cases of cancer that develop in the United States each year. At the present time, there are currently no effective therapies for this disease; however, clinicians have recently begun to demonstrate some success with chemotherapy as well as the use of biologic agents such as pegfilgrastim. .
Tell me about PDAC	
PDAC stands for "People with Disabilities Act" and it is an act of the United States Congress. It prohibits discrimination against individuals with disabilities in employment, public accommodations, transportation and federal programs. PDAC was signed into law in 1990 and has been used to protect and enhance the rights of people with disabilities. It has helped to increase access to education, employment and housing for individuals with disabilities.	PDAC is a pancreatic cancer that is characterized by high mortality rates and a lack of treatment options. The diagnosis of PDAC is typically made late in the course of the disease, resulting in poor prognosis and a limited treatment repertoire.
What are the main kinases druggable in PDAC?	
The main kinases druggable in PDAC are protein kinase A (PKA), protein kinase C (PKC), and tyrosine kinase 2 (TK2). Results obtained with the intermediary datasets.	Pancreatic ductal adenocarcinoma (PDAC) is a highly aggressive tumour associated with a poor prognosis. Nevertheless, recent years have witnessed the development of new, effective targeted therapies for several types of cancer. In particular, molecular-targeting strategies have been explored for the treatment of PDAC. Moreover, development of novel therapies for PDAC is a high priority and key unmet medical need.

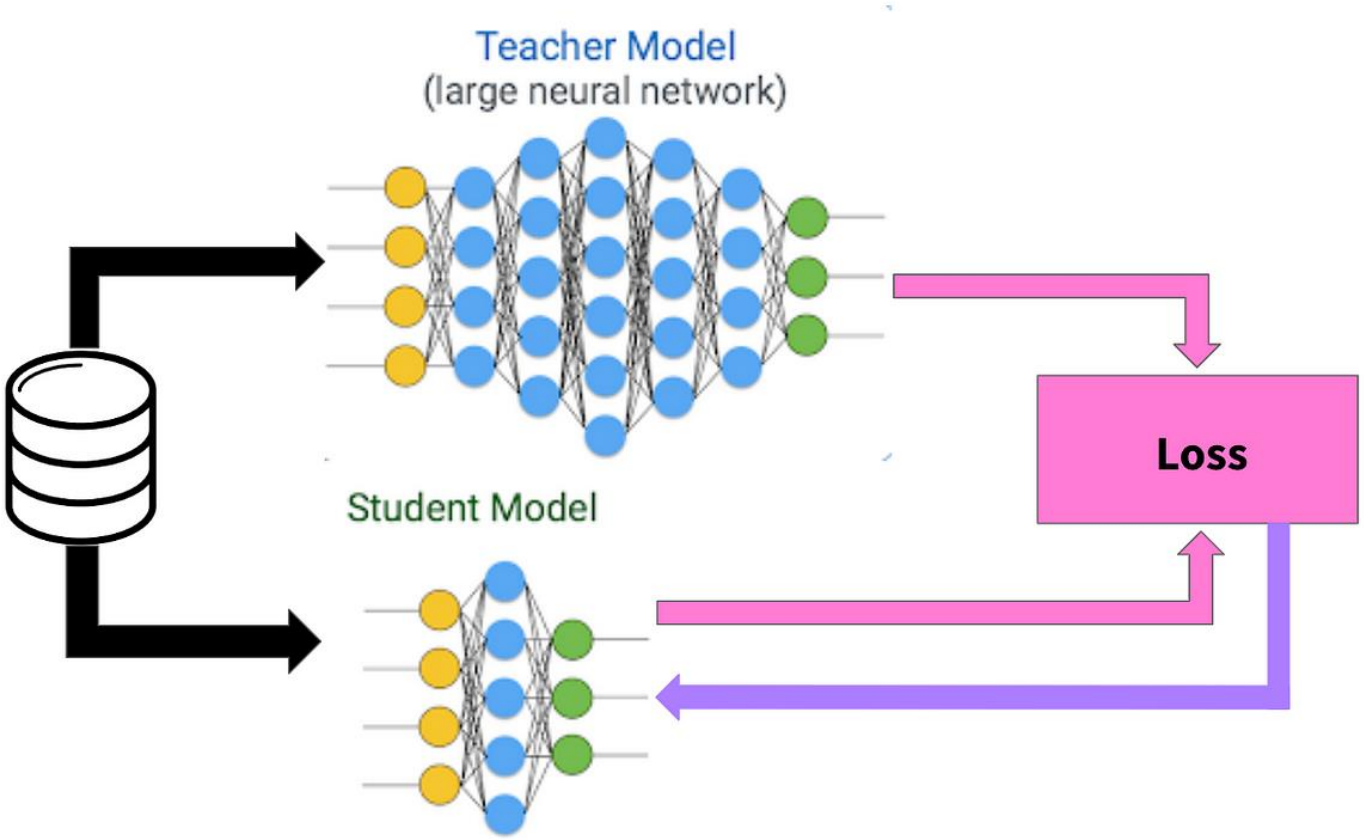
Medical applications of LLM technology

- Despite the published results, models still have flaws. This precludes autonomous deployment for decision-making or patient communication
- General risk models could provide a powerful alternative to the current myriad of tools used to stratify and triage patients
- LLMs are capable of rapid assimilation, summarization, and rephrasing of information that could reduce the administrative burden on clinicians.
- LLMs can be a tool for education (but should not be the only one)
- LLMs may facilitate novel research, such as analysis of language at a greater scale than previously possible.



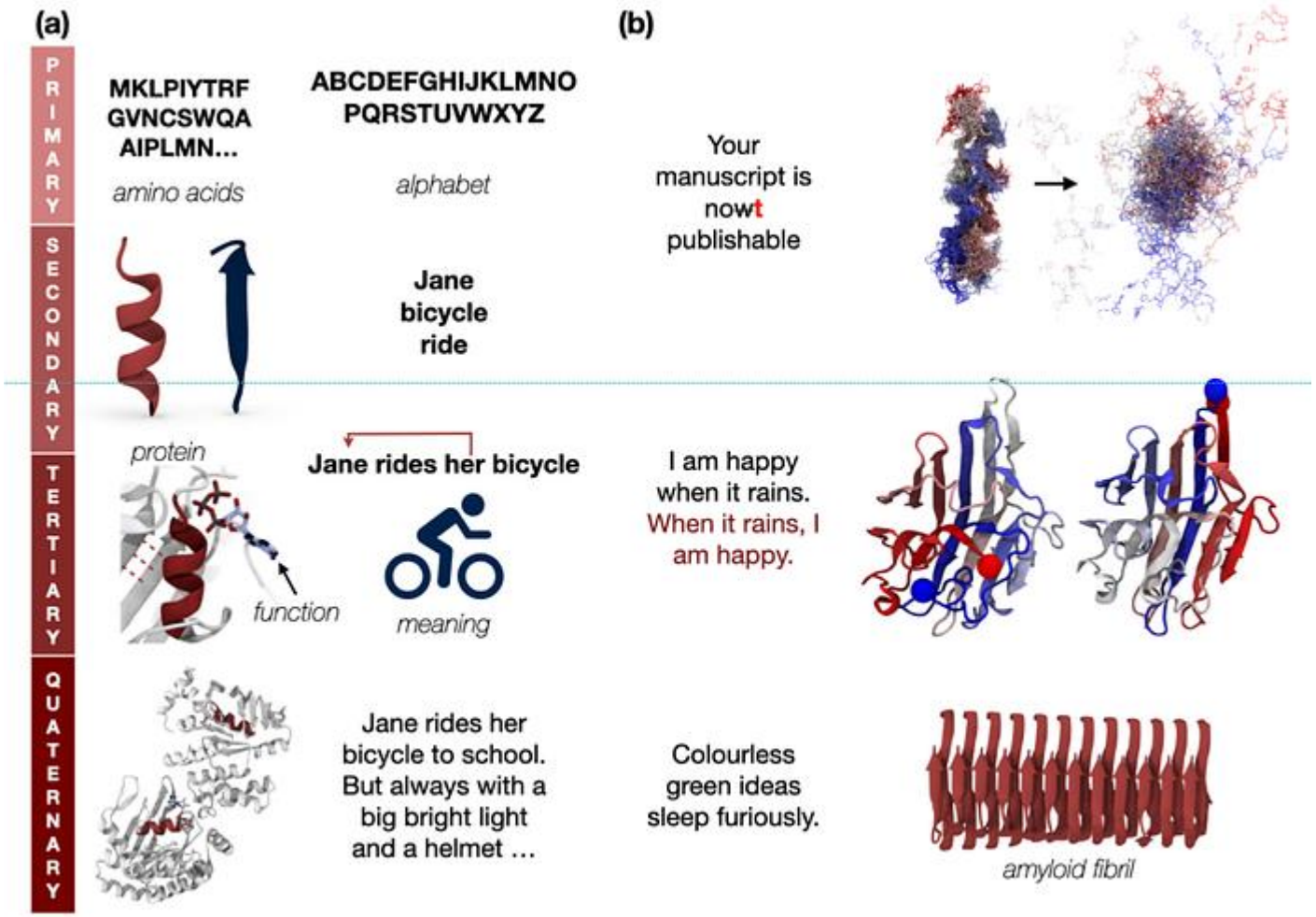
Or you may distill the knowledge

- Model can be expensive to deploy
- You may want to create a small version
- Knowledge distillation is a good way



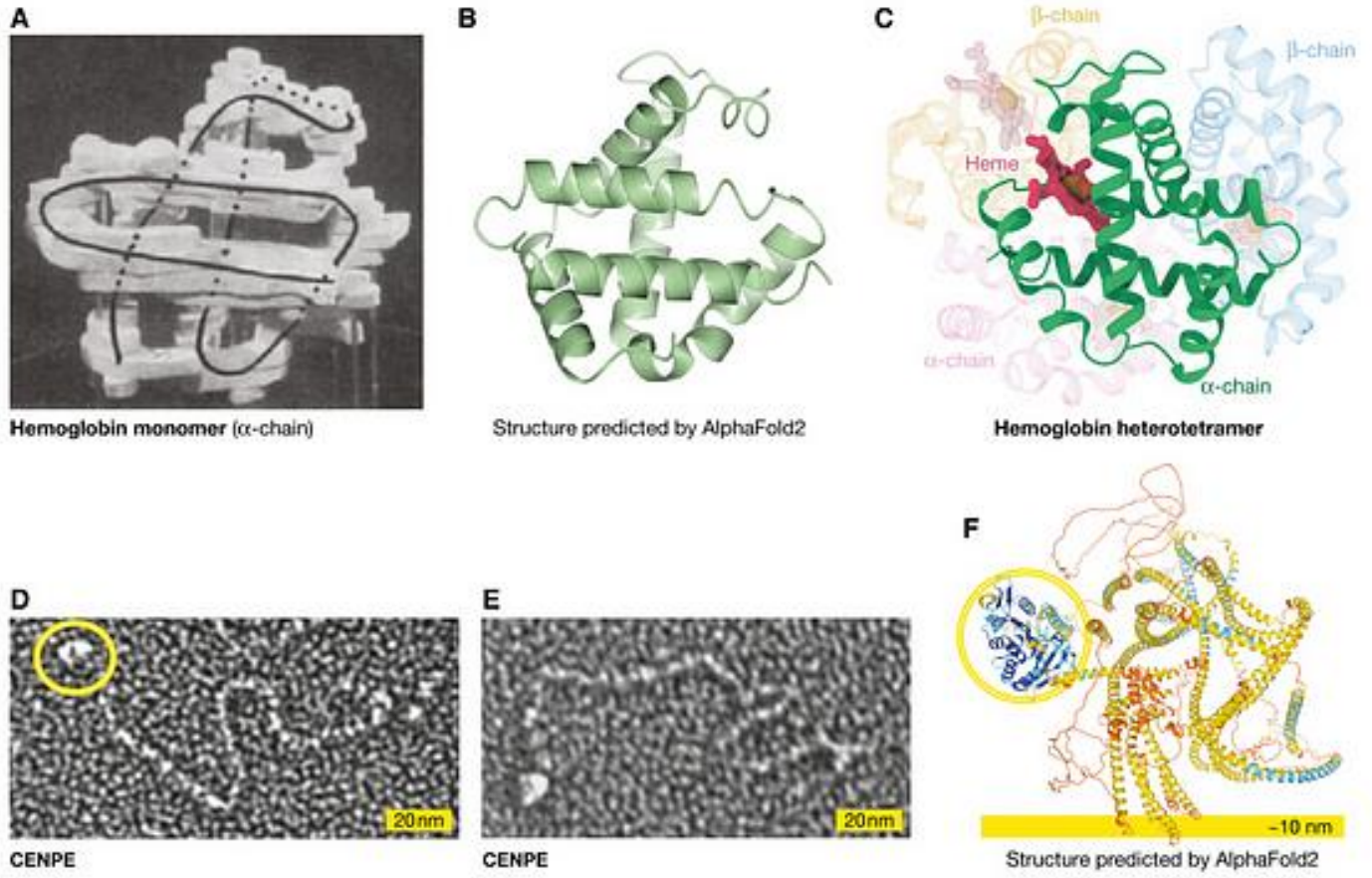
Speaking the language of life

- Hierarchical organization
- Typos and grammar
- Evolution
- Dependency



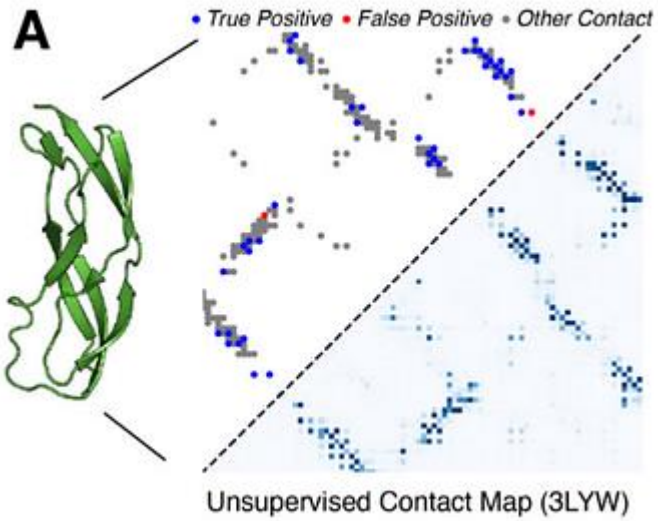
AlphaFold-2: solving one big puzzle of life

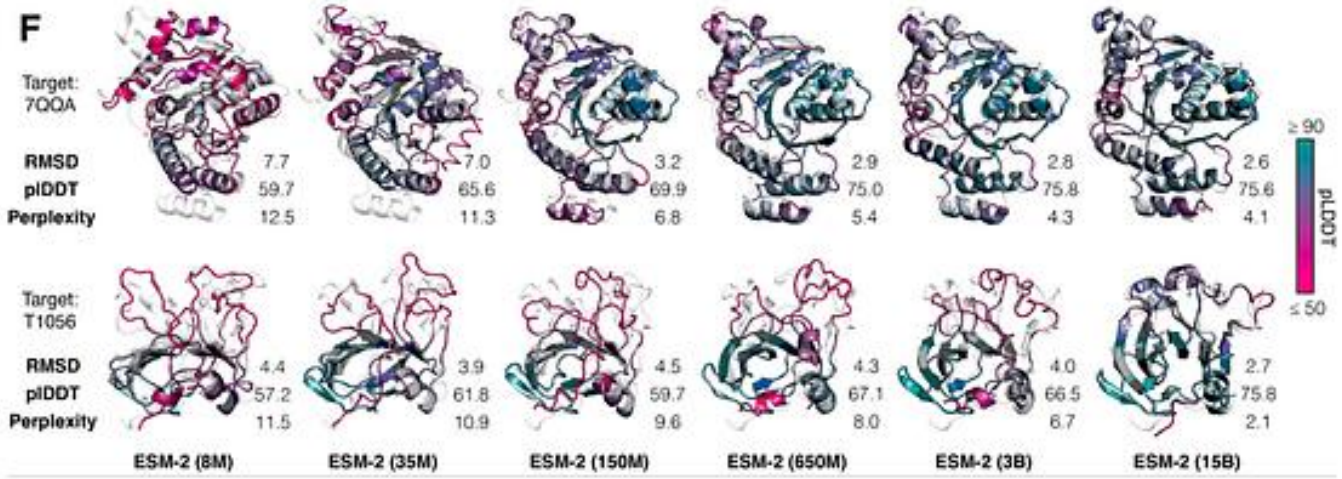
- In 2020, DeepMind participated in the Critical Assessment of Structure Prediction (CASP) challenge. AlphaFold2 was capable to predict the protein structures with atomic-level accuracy.
- Use of attention mechanisms, starting from multiple sequence alignments, and end-to-end learning.



A transformer for the proteins

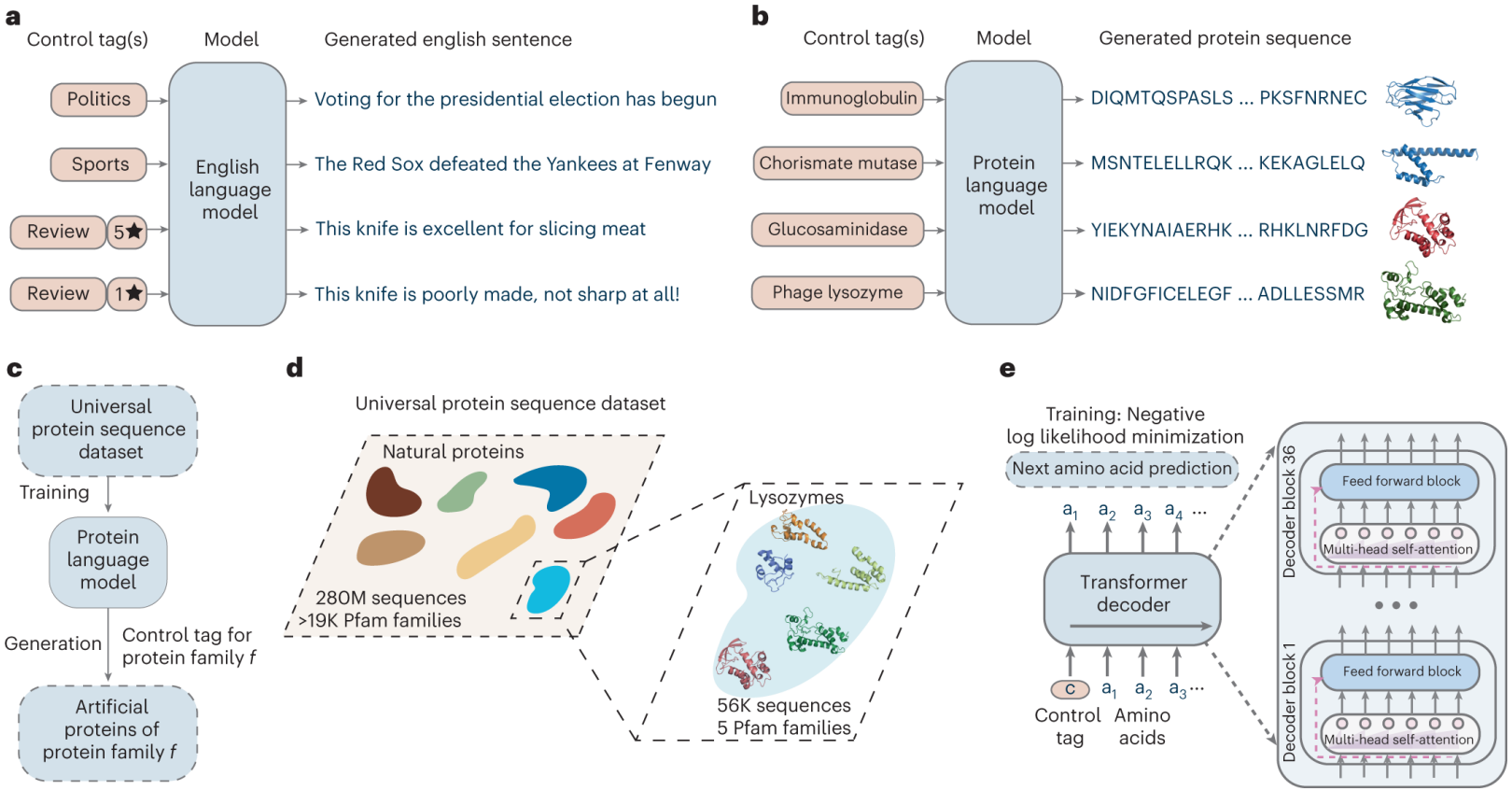
- the authors used a 15-billion-parameter model and noticed that scaling the parameters improved predictions (decreased perplexity).
- ESM-2 was trained with a masked language modeling objective





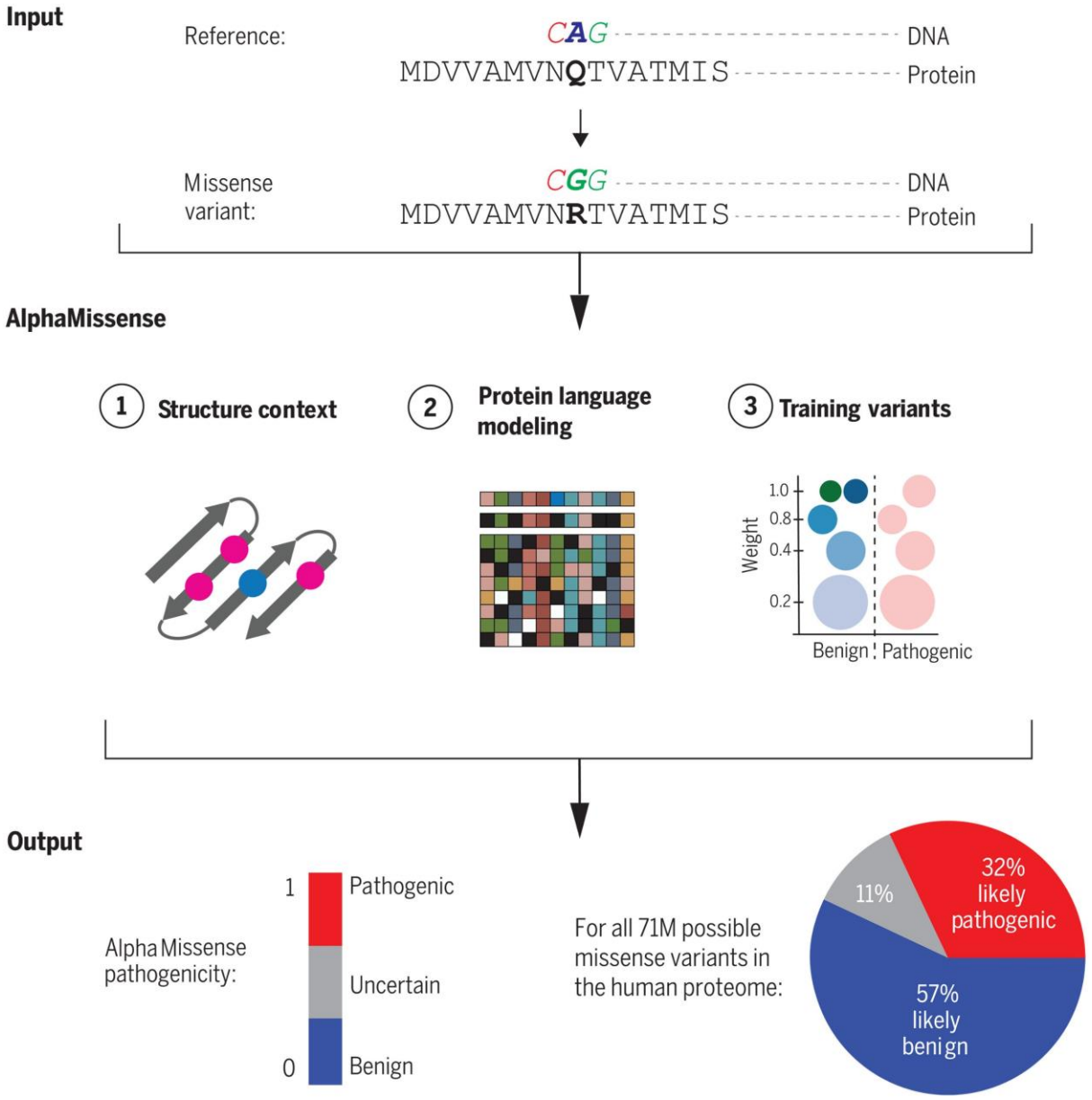
AI enables designing new proteins from scratch

- The model was trained on 280 million protein sequences from >19,000 families and is augmented with control tags specifying protein properties.
- Artificial proteins fine-tuned to five distinct lysozyme families showed similar catalytic efficiencies as natural lysozymes



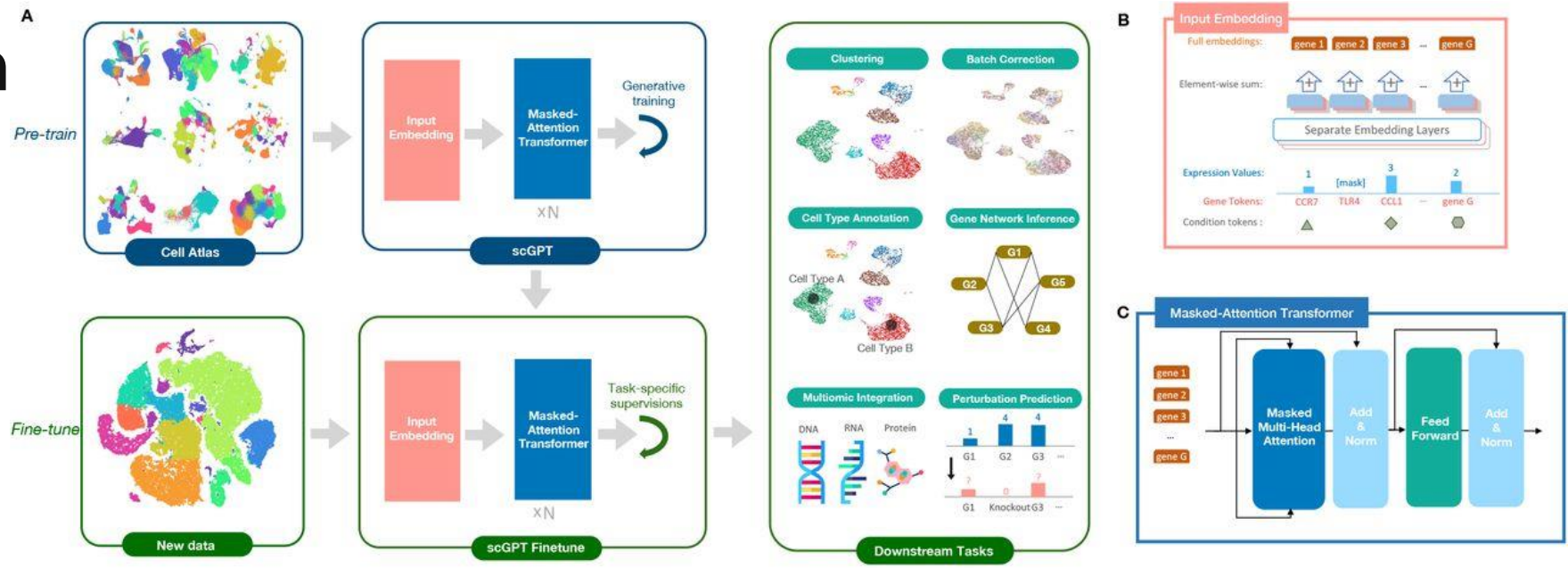
AlphaMissense: predicting pathogenicity of protein variants

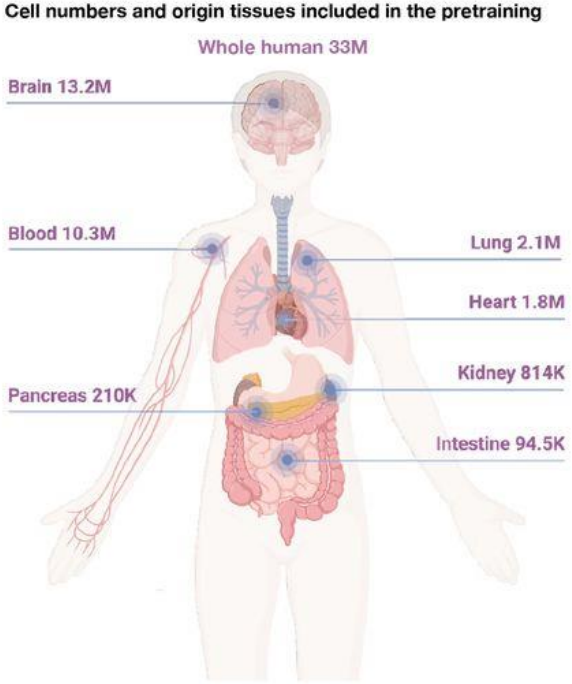
- AlphaMissense takes as input an amino acid sequence and predicts the pathogenicity of all possible single amino acid changes at a given position in the sequence.
- AlphaMissense predictions have the potential to accelerate our understanding of the molecular effects of variants on protein function, contribute to the discovery of disease-causing genes, and increase the diagnostic yield of rare genetic diseases.

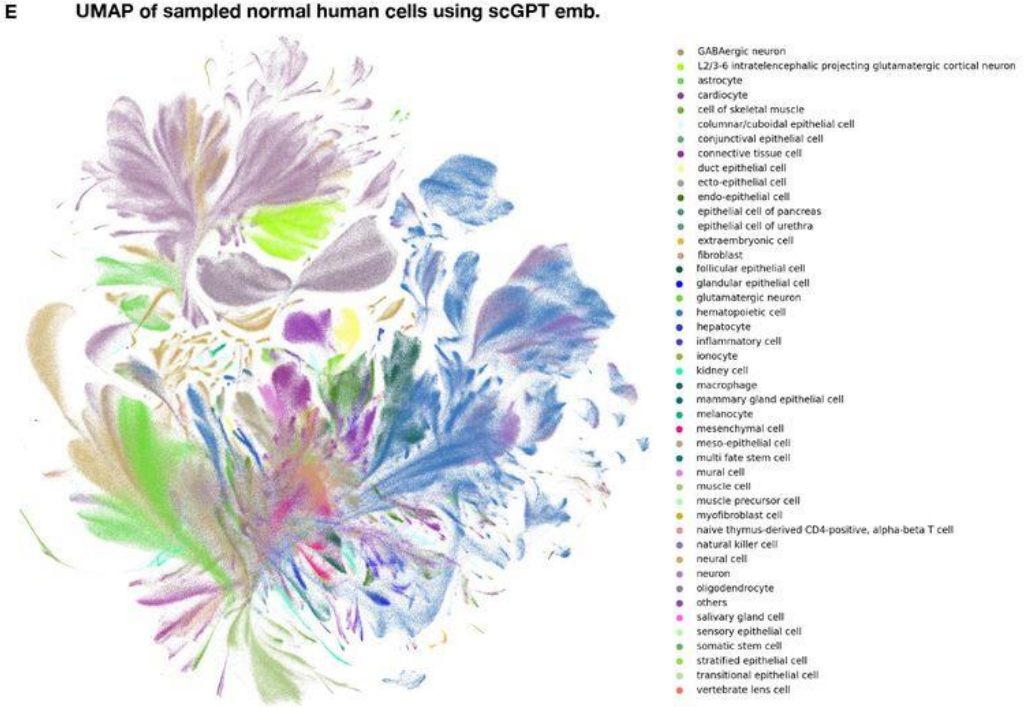


scGPT: a LLM for sin cell data

- Single-cell sequencing enables the profiling of molecular characteristics at the individual cell level.
- The core model contains stacked transformer layers with multi-head attention that generate cell and gene embeddings simultaneously

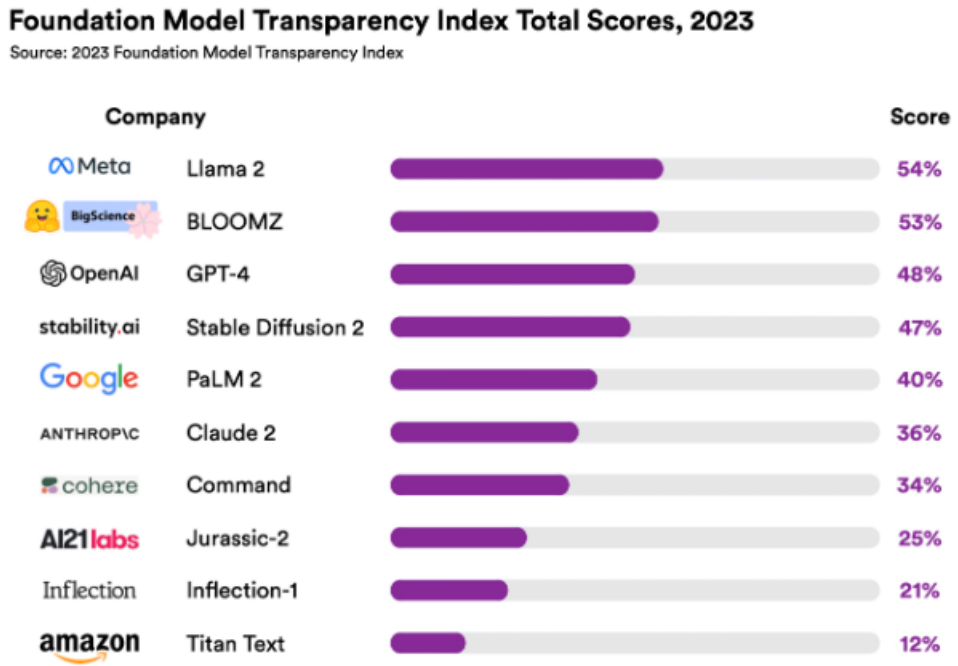


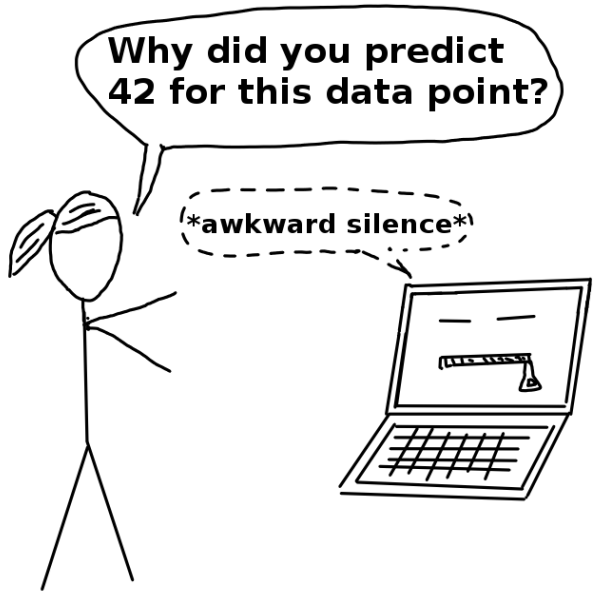




Barriers to implementation of generative AI LLMs

- Not updated knowledge
- Accuracy and cohérence
- Transparency and interpretability
- Ethical concern
- Copyright issues
- Computational cost



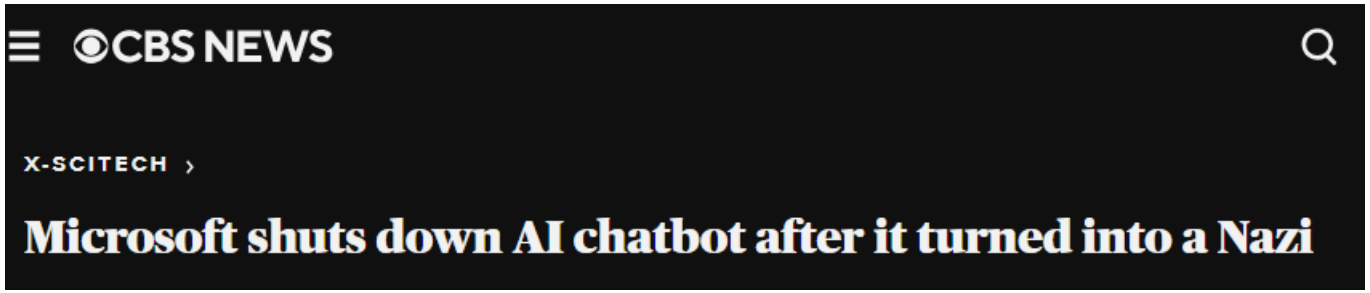


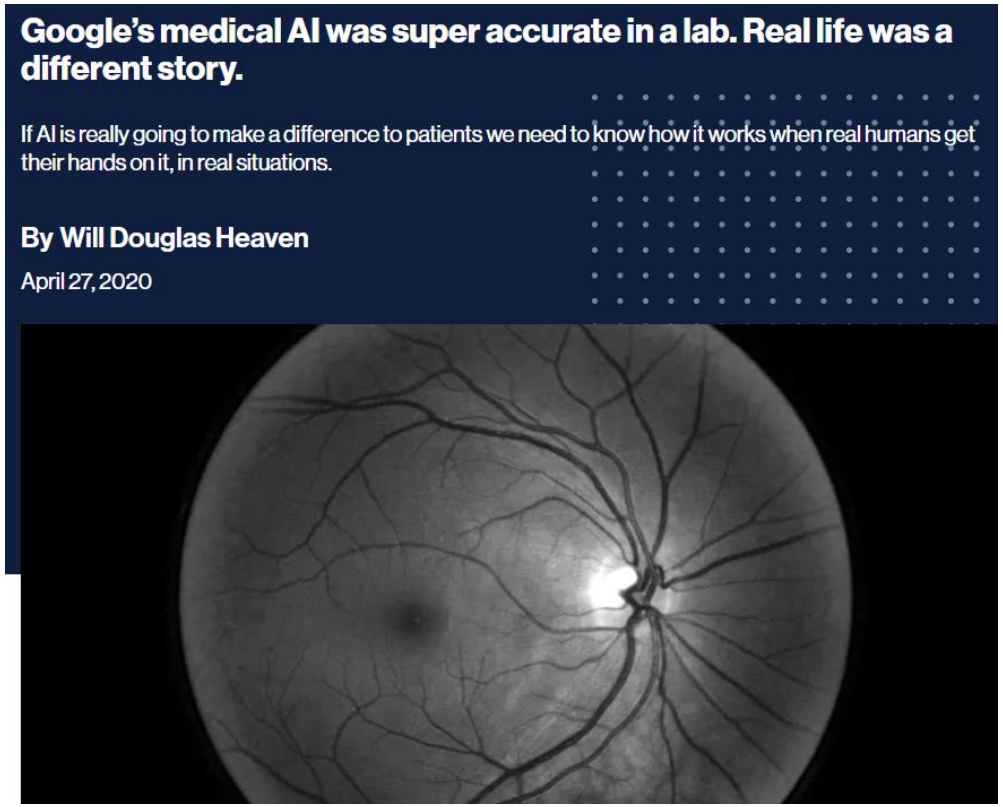
The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Barriers to implementation of LLMs in medicine

- Hallucination
- Lack of Evaluation Benchmarks and Metrics
- Domain Data Limitations
- New Knowledge Adaptation
- Behavior Alignment
-





What is the future of the LLM?

The limits of a LLM are the limits of the transformer:

- Out-of-domain adaptation
- Computational cost
- Maybe is not the best architecture ever

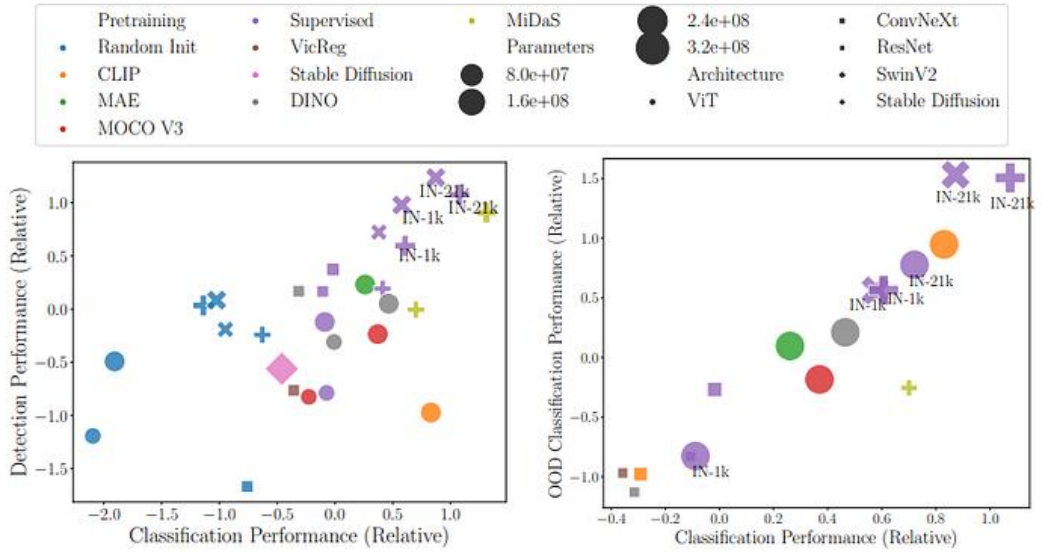
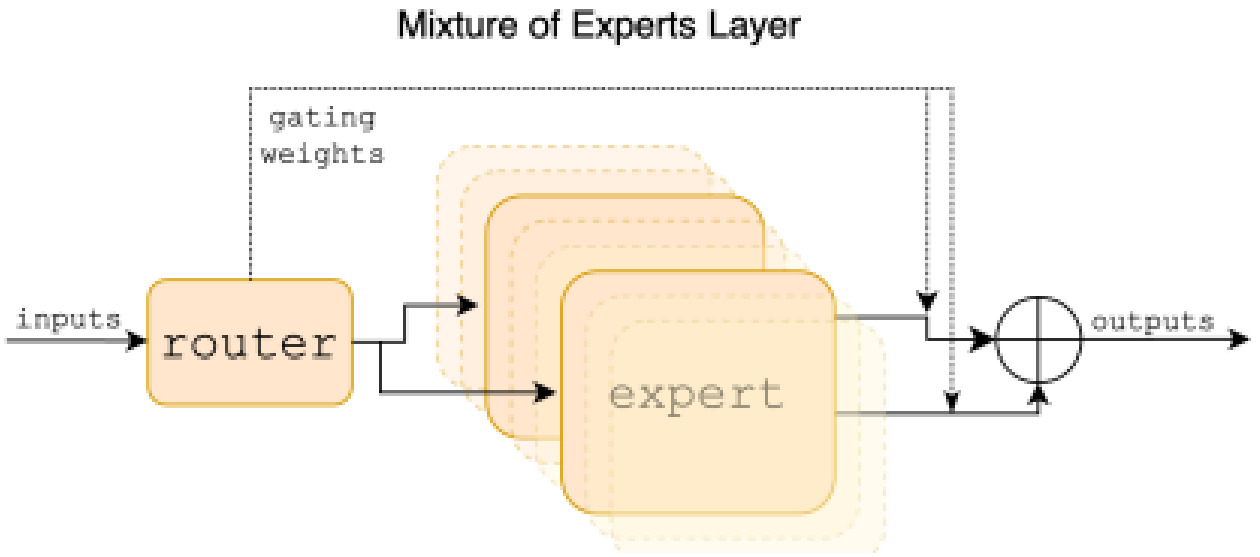


Figure 1: **Performance is correlated across tasks.** Performance for each model is per-dataset standard deviations above the mean averages across datasets. **Left:** Comparison between classification and detection. **Right:** Comparison between classification and OOD classification.

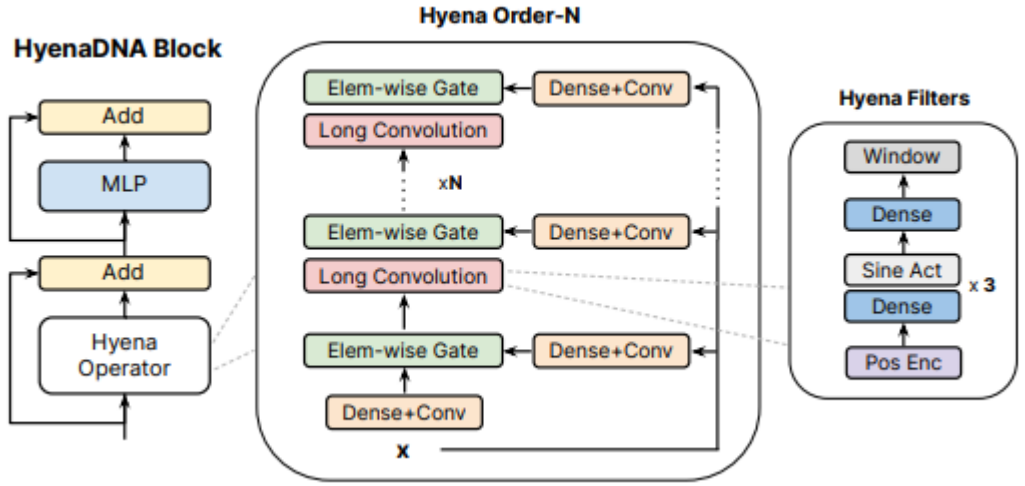
Mixture of experts (MOE)

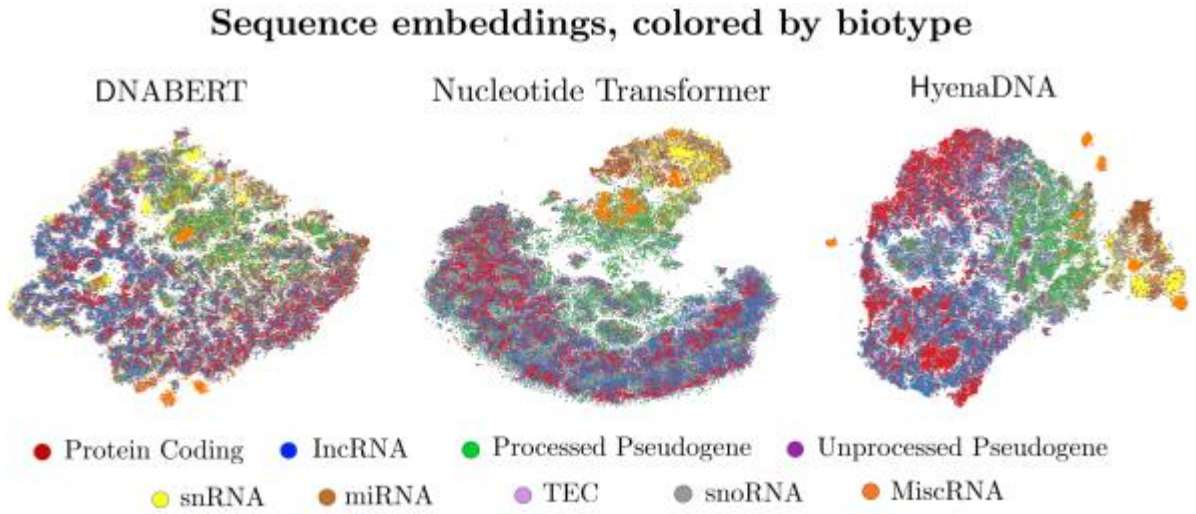
- GPT-4 and other new LLMs are MOEs
- MOEs are computationally more efficient and they are showing promising results



Maybe is the still beloved Convolution

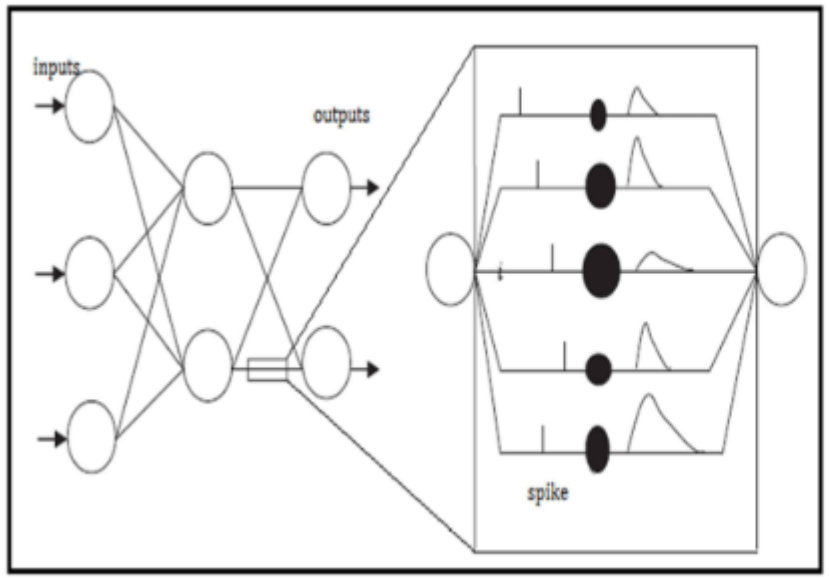
- Breaking the quadratic barrier is a key step towards new possibilities for deep learning, such as using entire textbooks as context, generating long-form music or processing gigapixel scale images.
- early application of the Hyena architecture is HyenaDNA, a new foundation model for genomics out of Stanford.

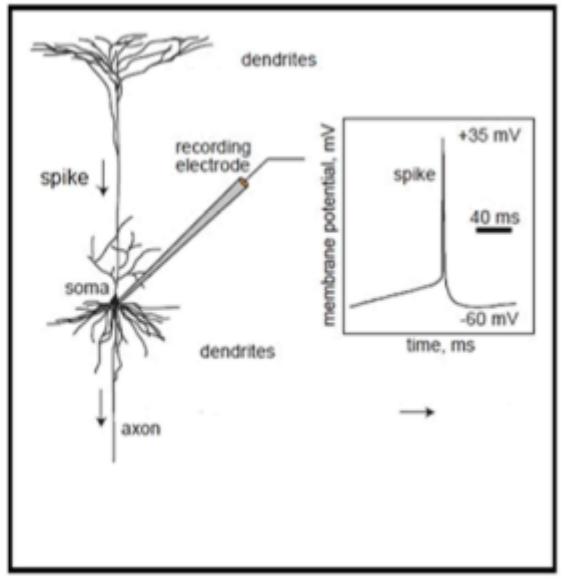




Spiking NN

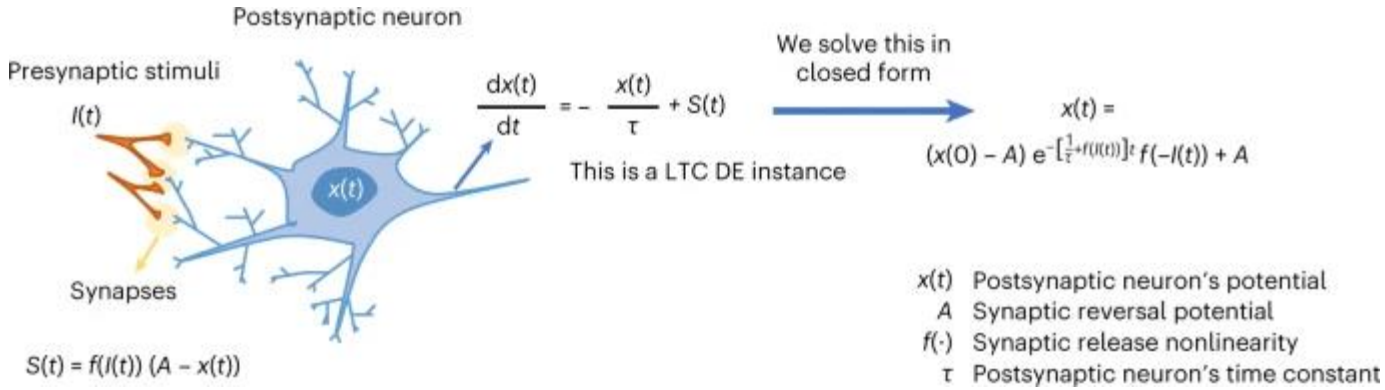
- The neuron fires when the membrane potential hits the threshold, sending a signal to neighbouring neurons, which increase or decrease their potentials in response to the signal.
- rather than working with continually changing time values as ANN does, SNN works with discrete events that happen at defined times.





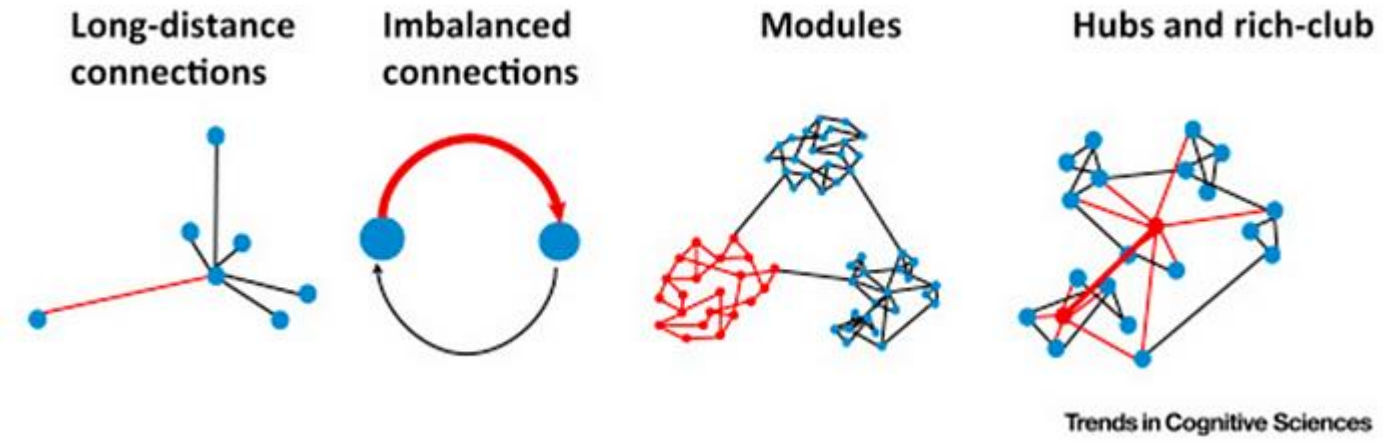
Liquid NN

- built by ordinary differential equations (ODEs)
- High performant
- These are neural networks that can stay adaptable, even after training– author says



Taking inspiration from the brain

- Neural systems during development must organize themselves.
- they must optimize the passage of information in the network



Taking inspiration from the brain

- Not all connections are necessary
- Take in account the distance in the structure

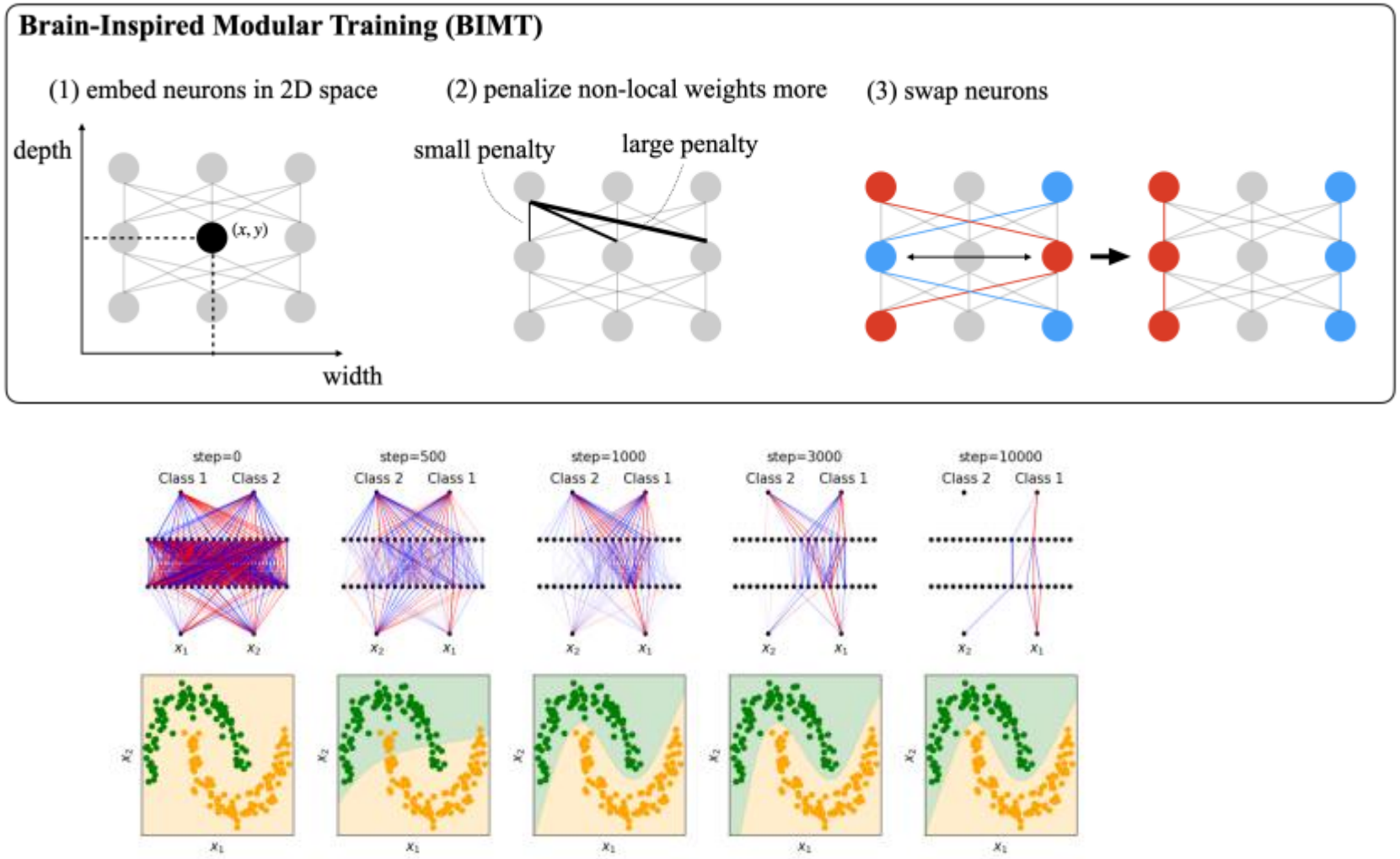
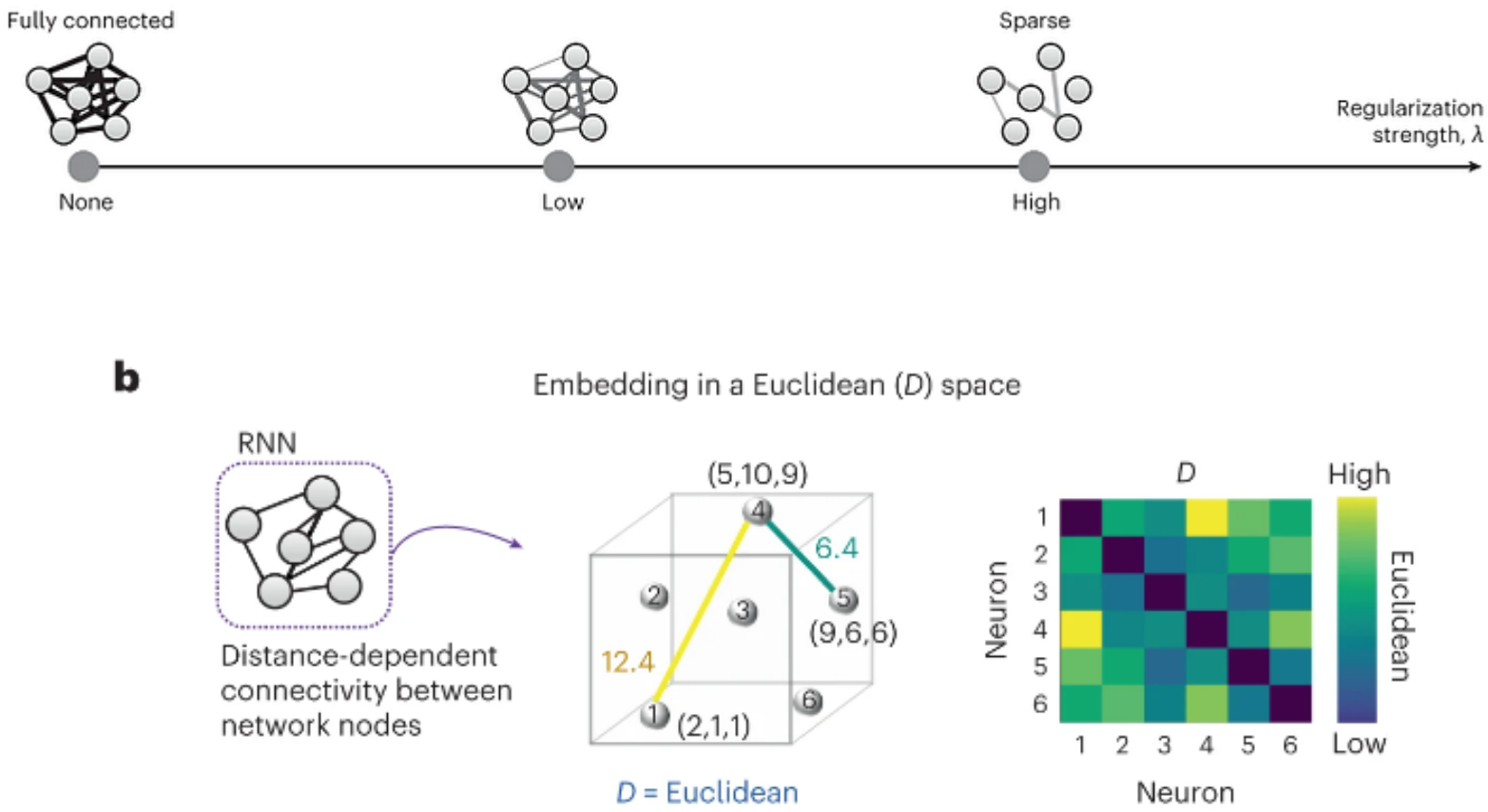


Figure 4: Top: Evolution of network structures trained with BIMT on the two moon dataset. Bottom: Evolution of decision boundaries.

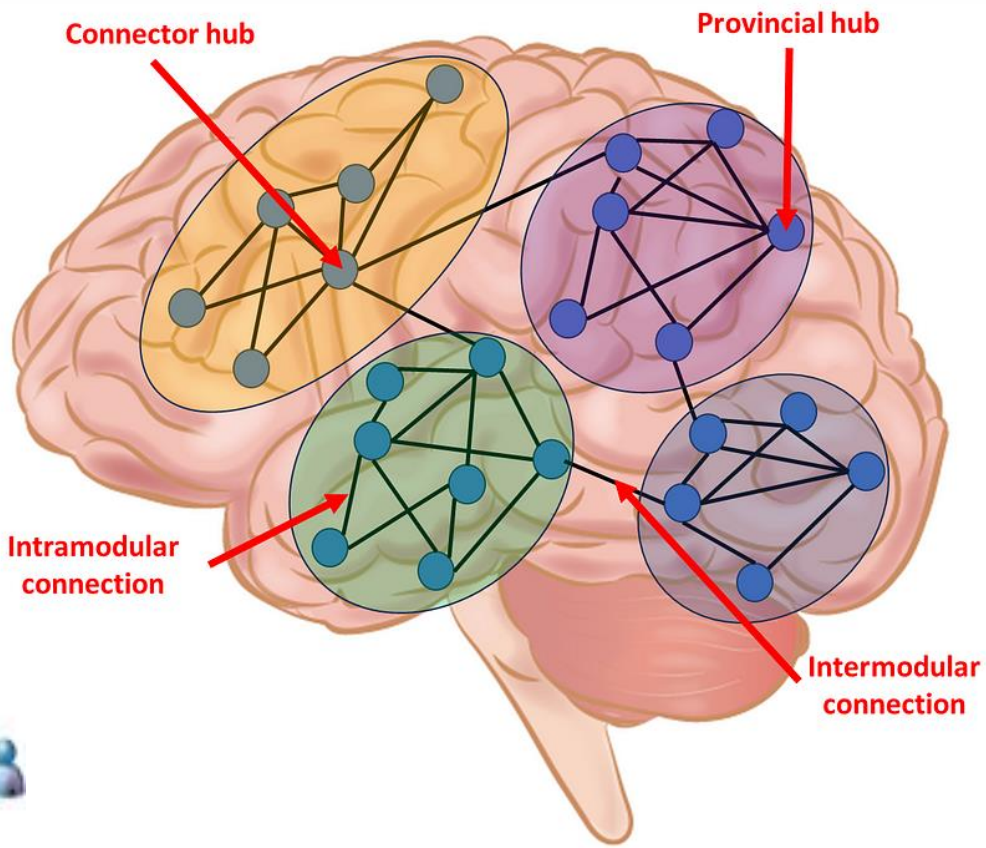
Taking inspiration from the brain: 3D constriction

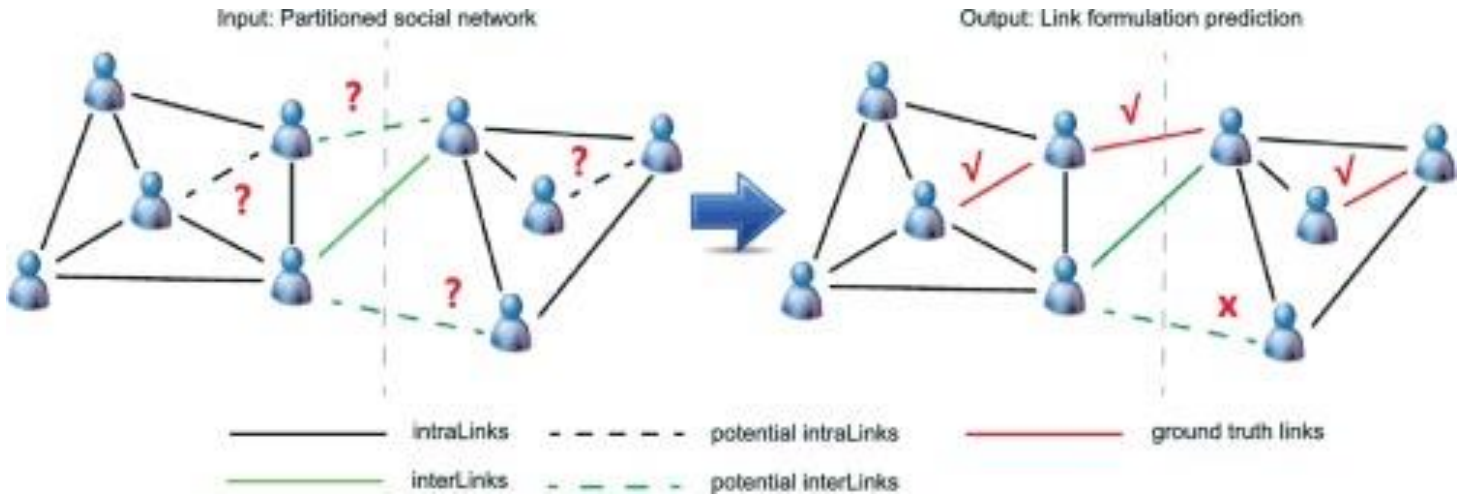
- L1 (or Lasso) this pushes the weights of the less important connections toward zero inducing sparsity
- Here is the same the regularization term is associated with distance in a 3D space.
- The network has to optimize within-network communication



The convergent evolution of primates and AI: Modularity

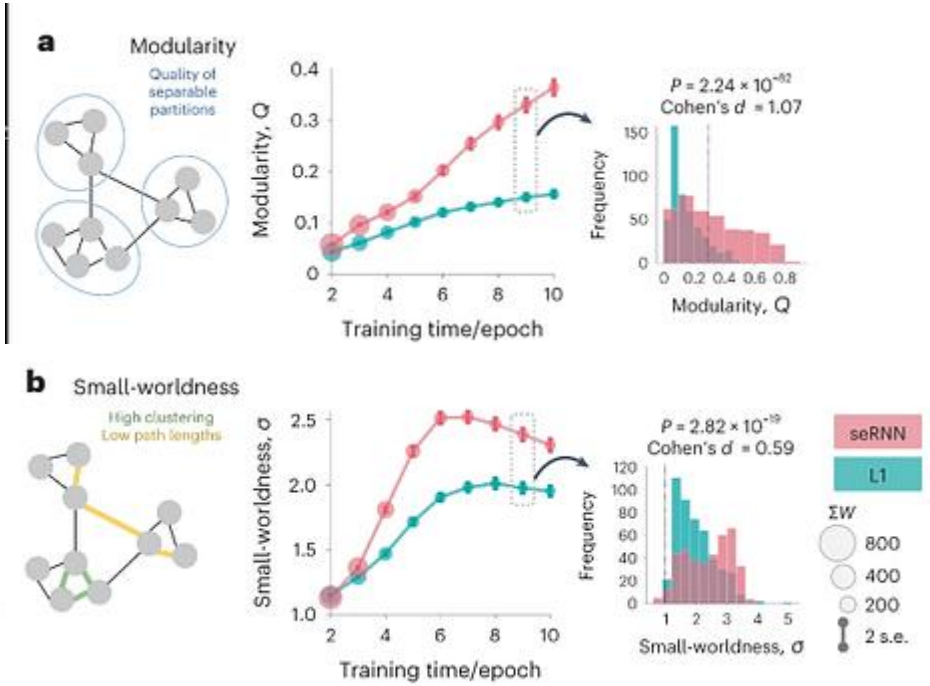
- **Modularity**. is a measure of the division of a network into modules (usually called clusters or communities).
- **Small world network**. A graph characterized by a high clustering coefficient and low distances

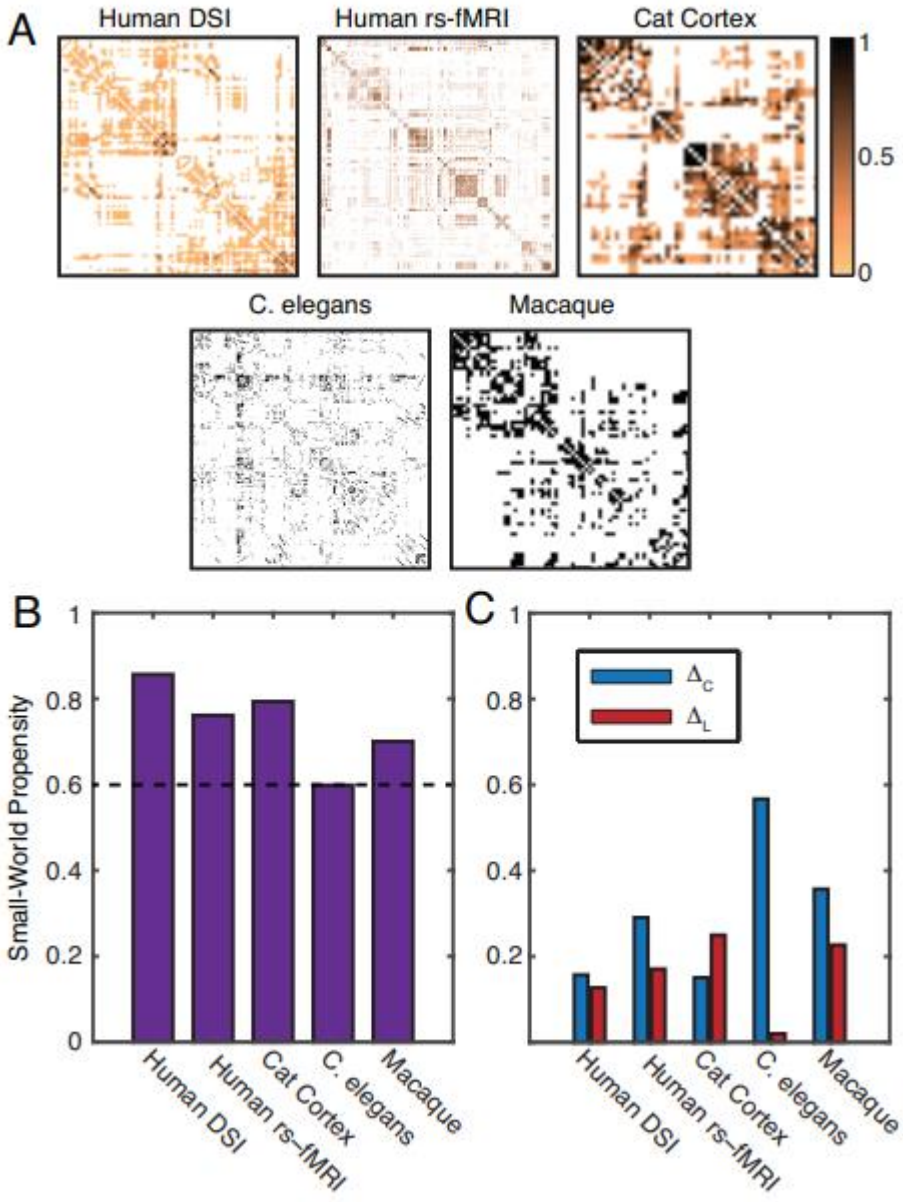




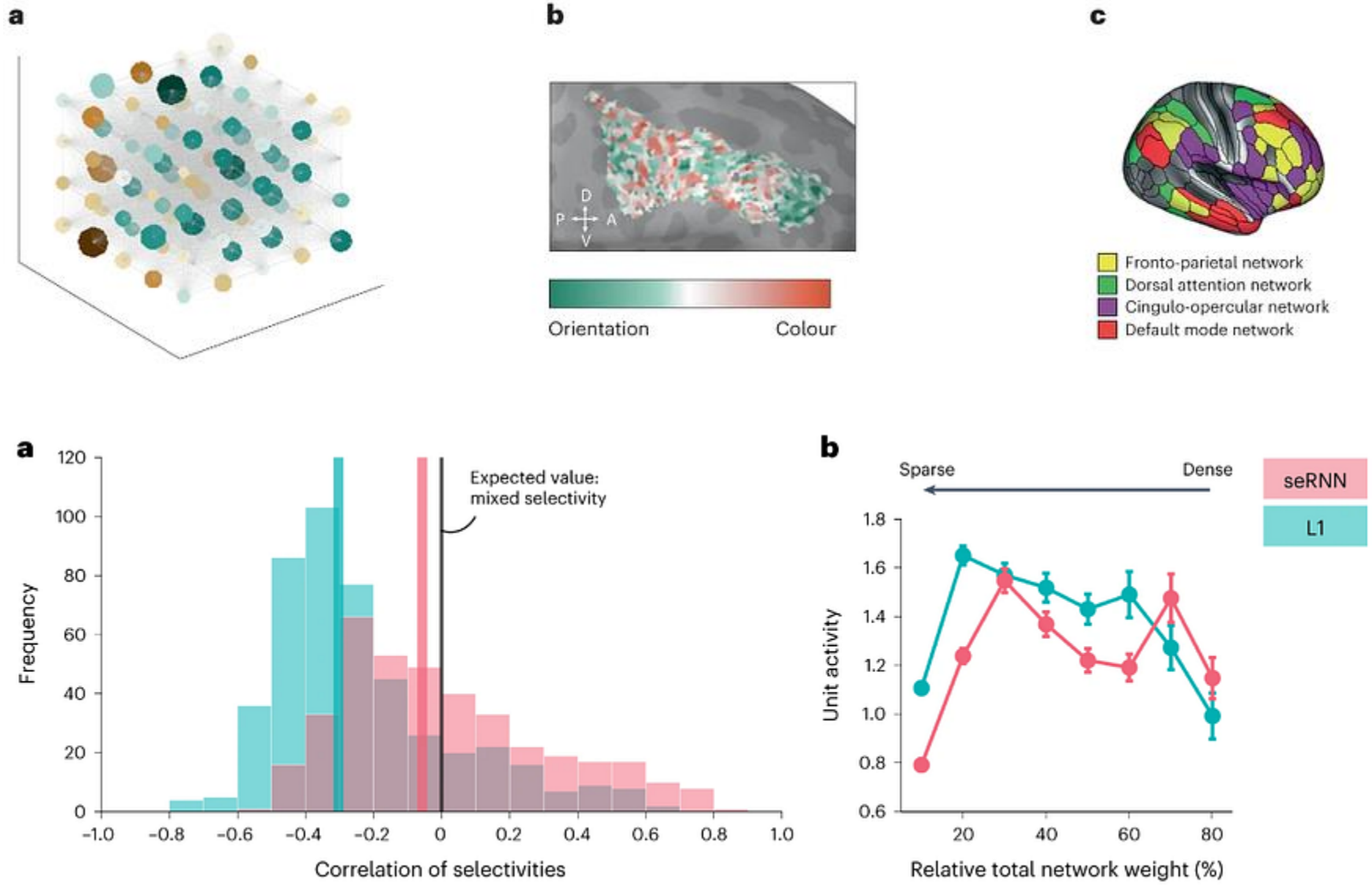
Taking inspiration from the brain: small world

- **Modularity**. is a measure of the division of a network into modules (usually called clusters or communities).
- **Small world network**. A graph characterized by a high clustering coefficient and low distances



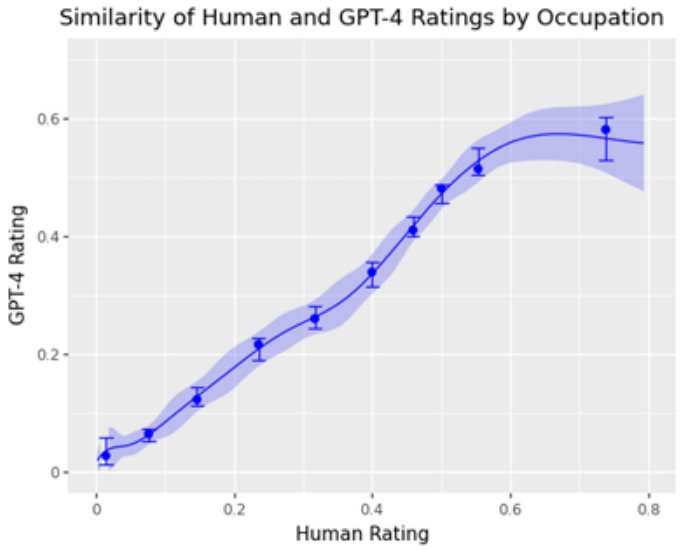


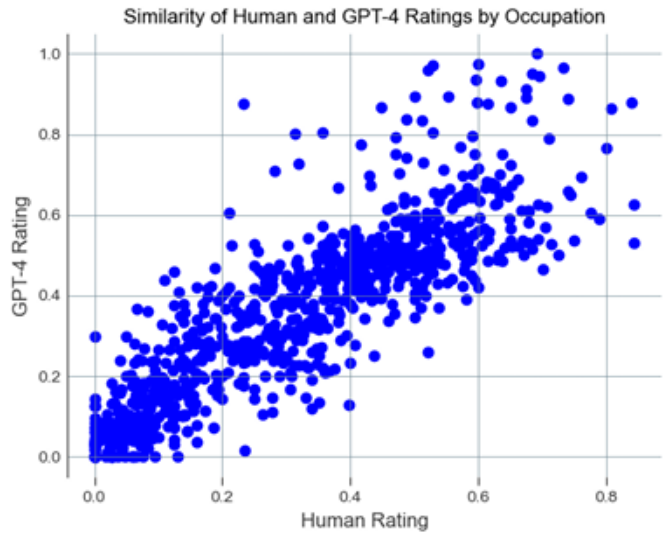
Taking inspiration from the brain: functional clustering and mixed selectivity



Will AI take my job?

- Employers anticipate a structural labour market churn of 23% of jobs in the next five years
- The largest losses are expected in administrative roles and in traditional security, factory and commerce roles.
- Analytical thinking and creative thinking remain the most important skills for workers in 2023





AI fuel climate change

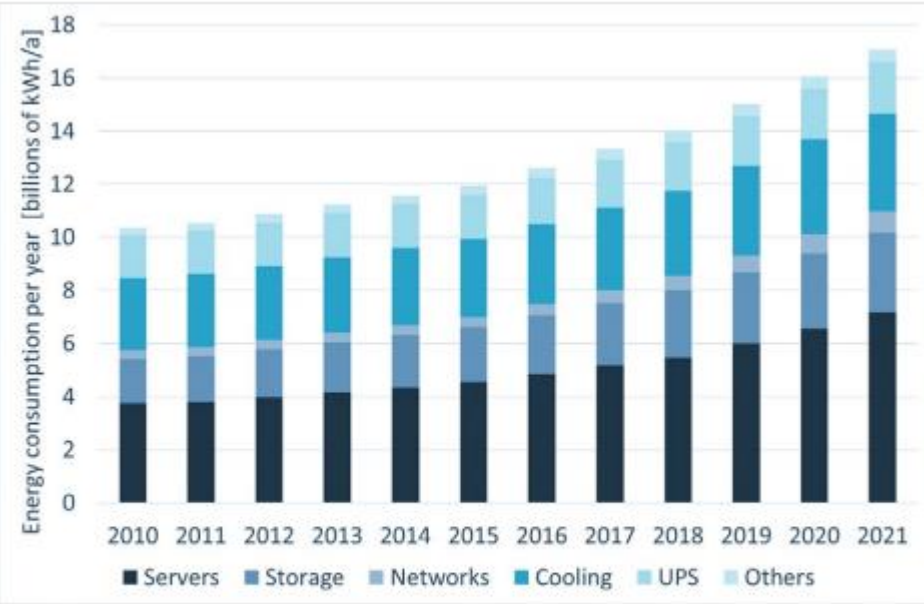
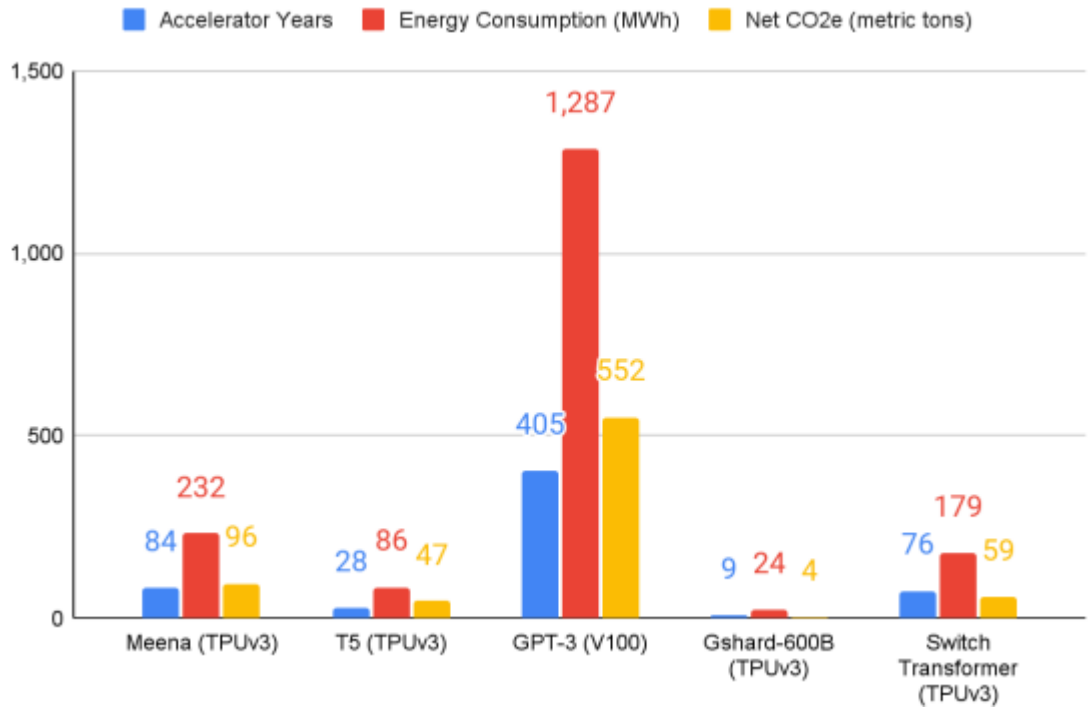
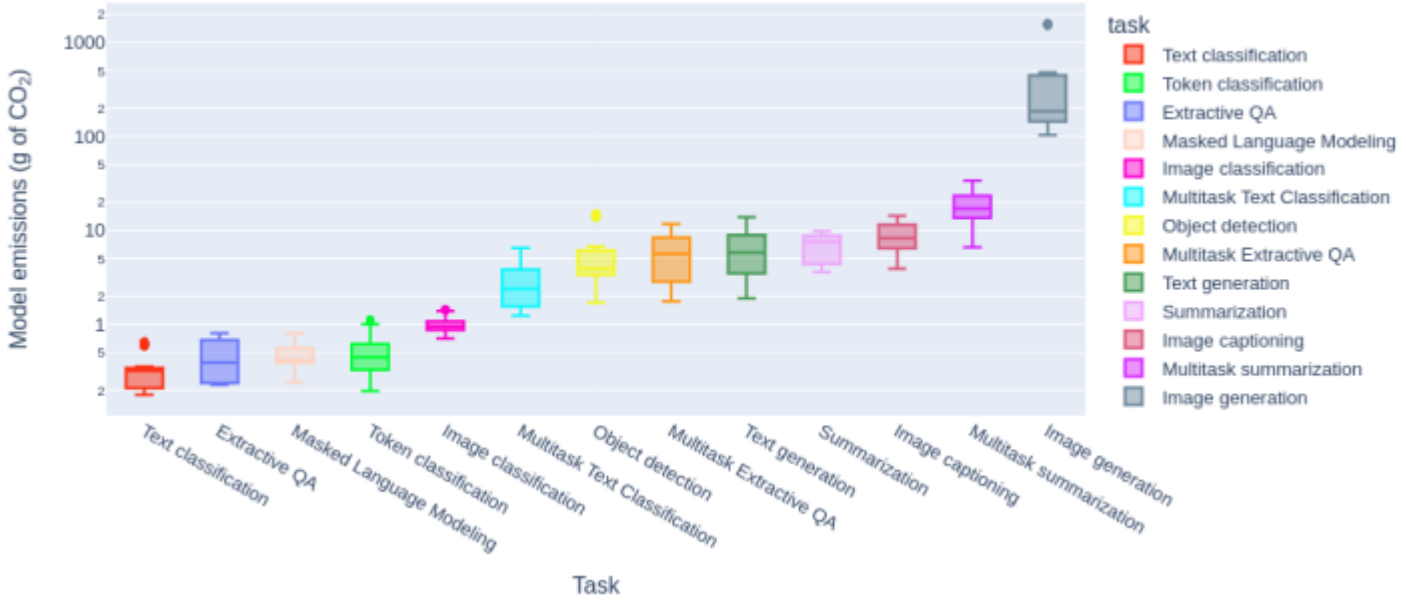


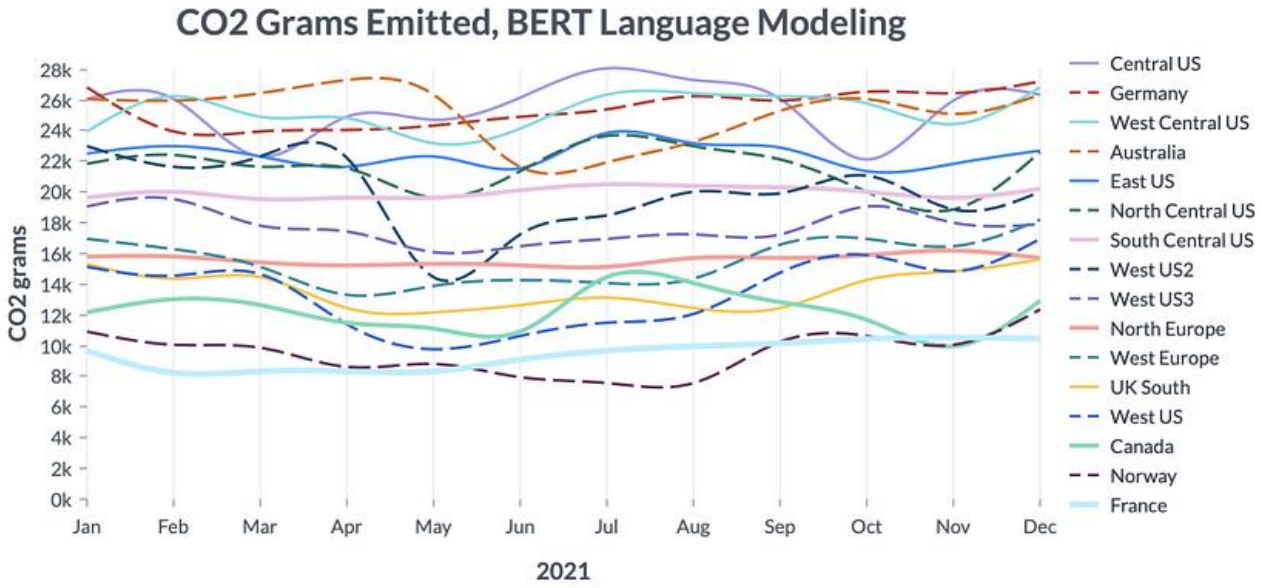
Figure 1: *Energy consumption of servers and data centers in Germany from 2010 to 2021 (Source: Borderstep)*



AI fuel climate change

- Not only the training but also the inference
- Many details should be thought to avoid CO2 emission





"The less we do to address climate change now, the more regulation we will have in the future." — Bill Nye

Conclusions

- Natural language is a difficult field and to analyze it we need particular modification of our approach
- The arrival of the transformer has changed our approach to NLP
- LLMs are transformers that have been slowly improved

Conclusions-2

- LLMs, as every type of AI model, has been tested in medicine
- However there are still challenges for its application

Thank you for your attention…
…and questions

Oncodesign
Precision Medicine