New Generative AI tools

Explain, demystify, understand the issues and the impact of their arrival in our professional lives

Richard Dufour, Equipe TALN
Nicolas Hernandez, Equipe TALN
(Harold Mouchère, Equipe IPI)

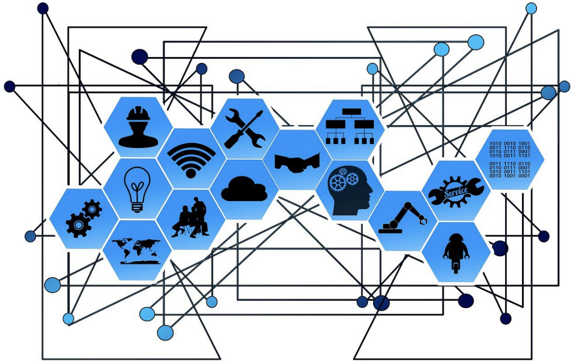
Journée de rentrée des doctorants
6 décembre 2023

Natural Language Processing (NLP)

- **Objective** : creation of tools and models to automatically process language, often to respond to concrete tasks (machine translation, automatic summary, speech transcription, etc.)
- NLP is a research field:
  - Old: first works in machine translation in the 1950s
  - Multidisciplinary: **Computer science**, linguistics, cognitive science, **machine learning**…

TALN team @ LS2N

- Team name = Field name in French (Traitement Automatique du Langage Naturel)
- Around 25 people including 11 permanent teacher-researchers
- Research themes
  - Analysis of textual documents
  - Information extraction
  - Language modeling with machine learning
  - Evaluation
- Application domains
  - Scientific writing
  - Education
  - Legal field
  - Health



Some highlights

- Béatrice Daille manages the **GDR CNRS TAL** since 2020 after setting it up in 2018

- Colin de la Higuera holds the **UNESCO Chair in Open Educational Resources and Artificial Intelligence** (RELIA) since 2017.

- **Master ATAL** (Apprentissage et Traitement Automatique des Langues) specialized in NLP is proposed and managed by the team (Solen Quiniou and Nicolas Hernandez)

Ph.D students - TALN team

Arthur Amalvy

Adrien Bazoge

Julien Aubert-Béduchaud

Léane Jourdan

Yanis Labrak

Xavier Pillet

Anas Belfathi

Oumaima El Khettari

Mael Houbre

Mohamed Reda Marzouk

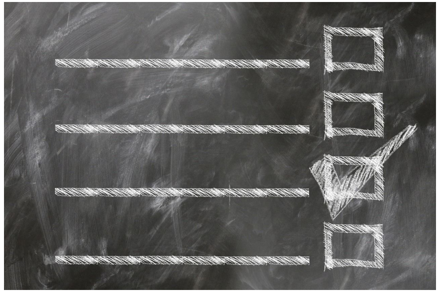
Thibault Roux

Thomas Sebbag

Current thesis subjects

- NLP and relationship modeling for unified representation of narrative documents
- TALMed : NLP for Health
- Connecting actors in a territory, through multi-format textual data, and predicting their behavior, in the context of the real estate market
- Neural approaches to model and analyze the argumentative structure of legal texts
- Exploring host-microbiome interactions using natural language processing
- Unsupervised generation of missing keywords for the indexing of scientific articles
- Neural approaches for modeling and analyzing the argumentative structure of research articles
- The sobriety of models in NLP: compromise between performance and resources

Fundamental approaches in NLP



- Two main approaches
  - **Symbolic**
    - Relies on human knowledge (e.g. “rule” approaches)
    - Generally accurate and reliable
    - High human costs but low machine costs
  - **Statistic**
    - Relies on textual data, often in very large quantities (era of big data)
    - “Rules” automatically extracted by machine learning methods (era of neural networks and deep learning)
    - Allows much broader deployment than symbolic approaches
    - Material costs (often) high but human costs low

➡ **Approach currently widely followed!**

Dialog system ELIZA

- Created by Joseph Weizenbaum in the 60’s
- Simulation of a psychotherapist
- Simple rules with reformulations into questions

```
 Welcome to
                    EEEEEE  LL      IIII   ZZZZZZZ   AAAAA
                    EE      LL       II        ZZ   AA   AA
                    EEEEE   LL       II      ZZZ    AAAAAAA
                    EE      LL       II    ZZ       AA   AA
                    EEEEEE  LLLLLL  IIII   ZZZZZZZ  AA   AA

 Eliza is a mock Rogerian psychotherapist.
 The original program was described by Joseph Weizenbaum in 1966.
 This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Central Concept: Language modeling

- A *language model* is a representation of language using **unsupervised statistical approaches** possible at different levels (letter, word, sentence, etc.)

- Training from large document collections

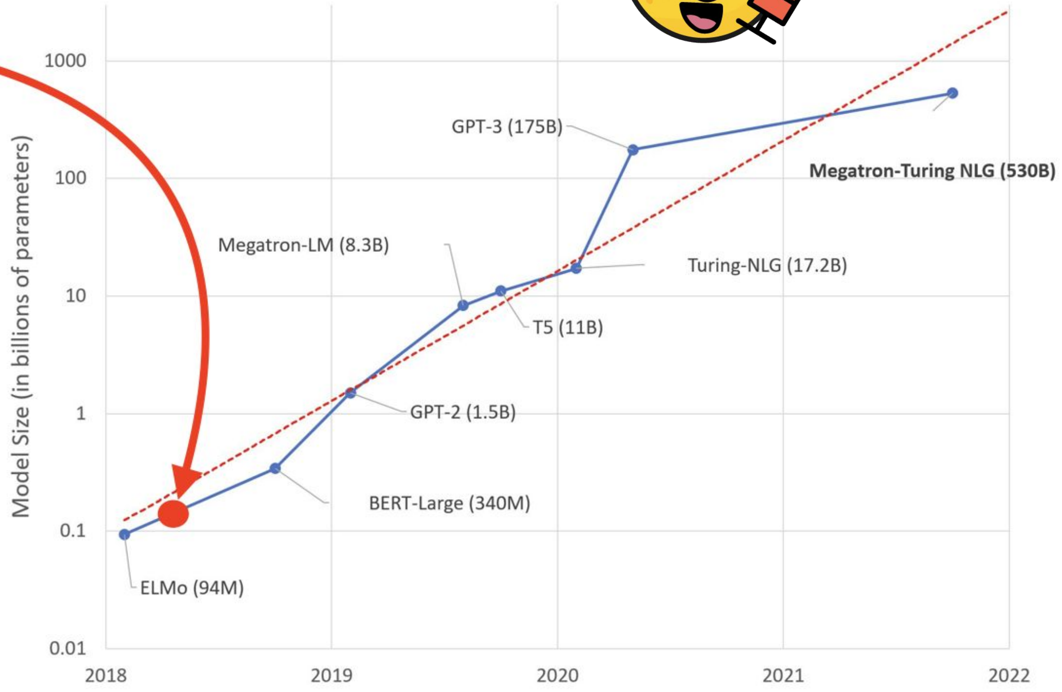
- Two large families of models
  - **N-grams**: probability of appearance of a word according to a history
  - **Neural networks**: complex architectures through a chain of neural layers

- Latently encodes linguistic information (lexical, grammatical, semantic, etc.) from one or more languages

Language modeling in Nantes University

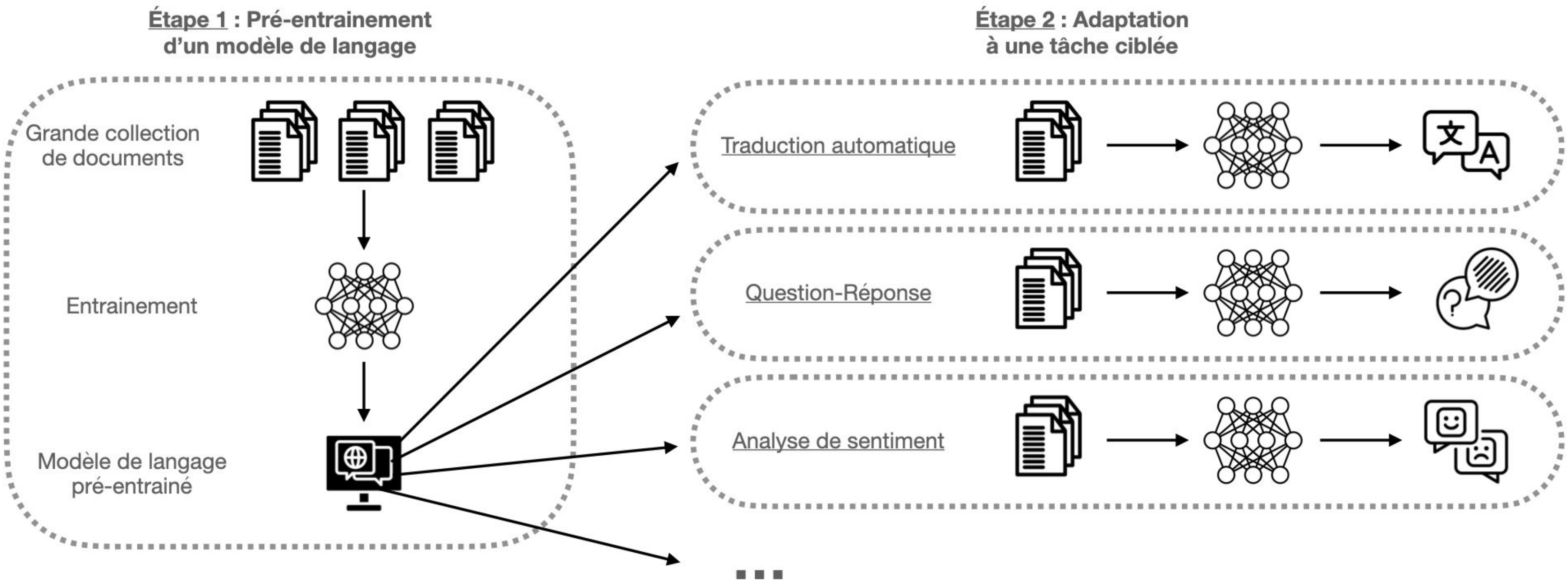
- DrBert
  - French open language model specialized in the medical field
  - 4 to 7 GB of open data input (1 billion words)
  - Freely usable
  - 110 million parameters
  - Trained on the Jean Zay super calculator (Genci / CNRS) - 128 Nvidia V100 GPU with 32 GB for 20 hours

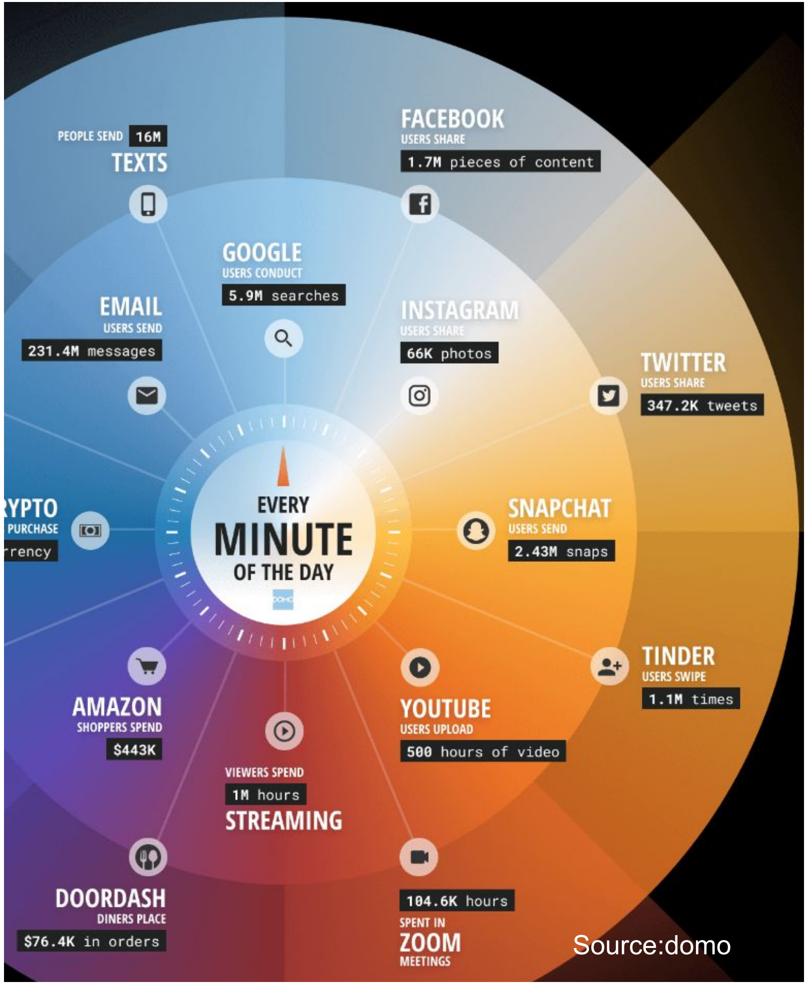
  - Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, et al.. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. ACL'23. ⟨hal-04056658⟩



Julien Simon : https://huggingface.co/blog/large-language-models

Approach that has become “classic”





Source:domo

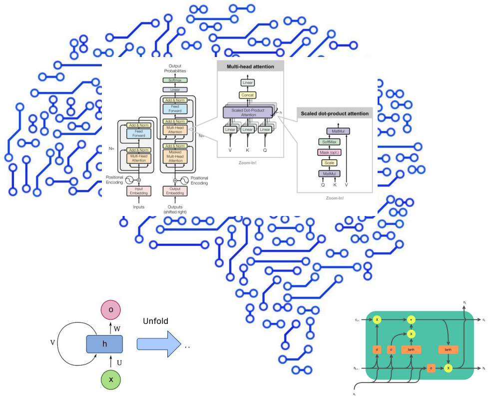
1. Big data

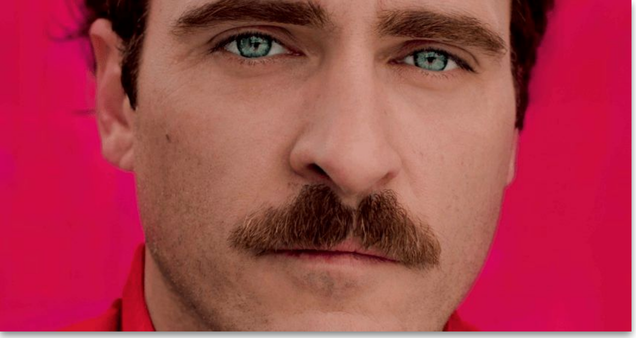
2. Material capacities for parallel computing (graphic processor known as GPU)

Source:[zmaslo]

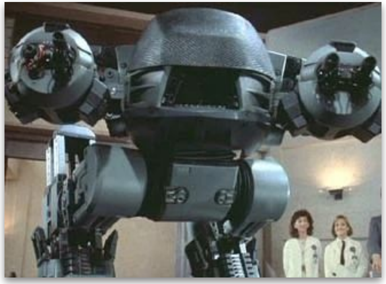
The [GPU NVIDIA TITAN RTX] card (~€4,500) for gaming machines has 4,608 cores (130 TFlops i.e. 130,000 billion floating point operations per second)

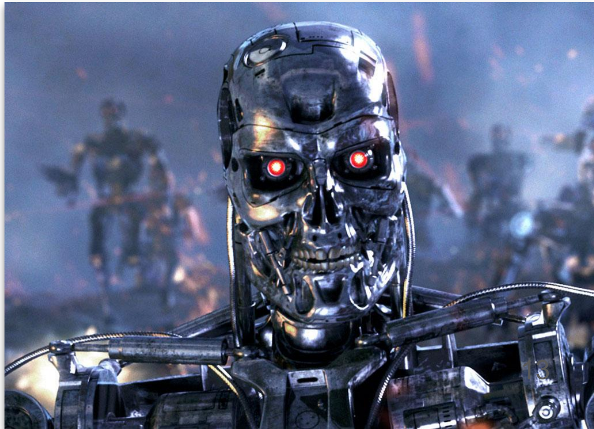
3. Neural architectures (*deep learning*)





Generative AI





What are we talking about?

Innovations in « **generative** » **machine learning** which have considerably improved the capacity of **Artificial Intelligence** to « ~~**create**~~ **generate** » content

- Text from texts, images, or speech
  - OpenAI ChatGPT, Google Bard, Meta LLaMA, Anthropic Claude
- Image from text
  - Midjourney, Google Imagen, OpenAI DALL-E3
- Video from text
  - Imagen Video
- Video from an "idea" (text, image, video, etc.)
  - Pika 1.0

What can generative AIs that we request via a text message do?

machine translation

summarization

generation

question/answer

Human: prompt

AI: completion

reformulation

dialogue

brainstorming

extraction

classification

What can we ask?

A machine capable of responding to requests such as:

2x+3=4

Peux-tu m'écrire une lettre de motivation pour un poste de commercial BtoB dans le secteur des télécommunications. J'ai 15 ans d'expérience dans le domaine, je suis mobile et je suis disponible immédiatement.

Find the bug with this code:

```
for (var i = 0; i < 5; i++) {
  setTimeout(() => console.log(i), 1000)
}
```

Voici quelques idées de menu pour Noël avec les recettes :

GE Peux-tu m'expliquer ce qu'est une confusion de type dans le moteur Javascript d'un navigateur Web

Source: 01net

What can we ask?

- Written response
- Detailed explanations
- Correct answer

peux tu me donner la dérivée de x sin(x) ?

Bien sûr ! Pour calculer la dérivée de la fonction $f(x) = x \cdot \sin(x)$, nous utiliserons la règle du produit pour dériver le produit de deux fonctions.

La règle du produit dit que si $u(x)$ et $v(x)$ sont deux fonctions dérivables par rapport à x, alors la dérivée du produit $u(x) \cdot v(x)$ est donnée par :

$(u \cdot v)' = u' \cdot v + u \cdot v'$

Dans ce cas, $u(x) = x$ et $v(x) = \sin(x)$. Calculons les dérivées de ces fonctions par rapport à x:

$u'(x) = 1$ (dérivée de x par rapport à x)
$v'(x) = \cos(x)$ (dérivée de $\sin(x)$ par rapport à x)

Maintenant, utilisons la règle du produit pour calculer la dérivée de $f(x) = x \cdot \sin(x)$:

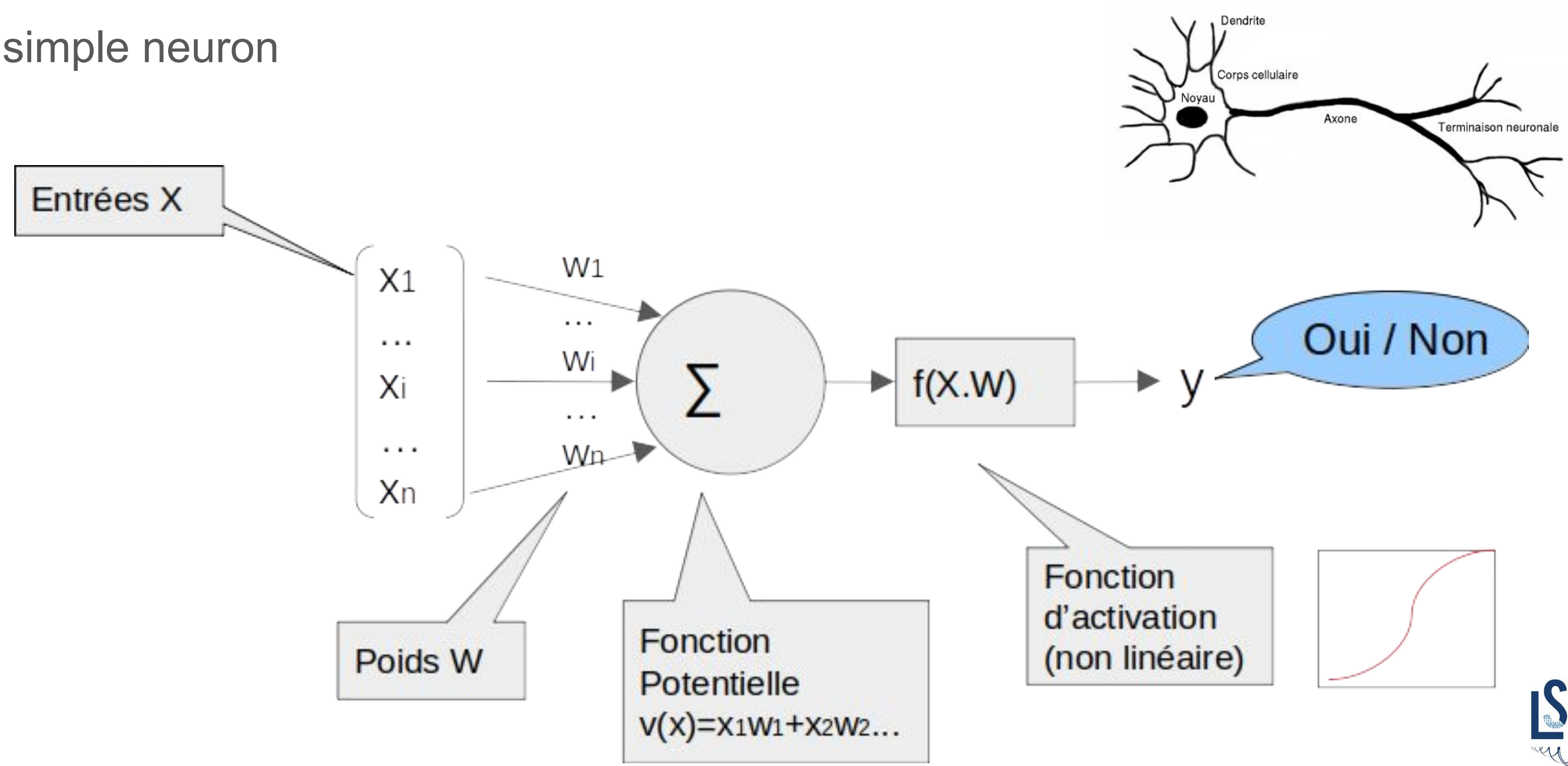
$f'(x) = u' \cdot v + u \cdot v' = 1 \cdot \sin(x) + x \cdot \cos(x) = \sin(x) + x \cdot \cos(x)$

Donc, la dérivée de $x \cdot \sin(x)$ par rapport à x est $\sin(x) + x \cdot \cos(x)$.

Neuron, neural network and *deep learning*

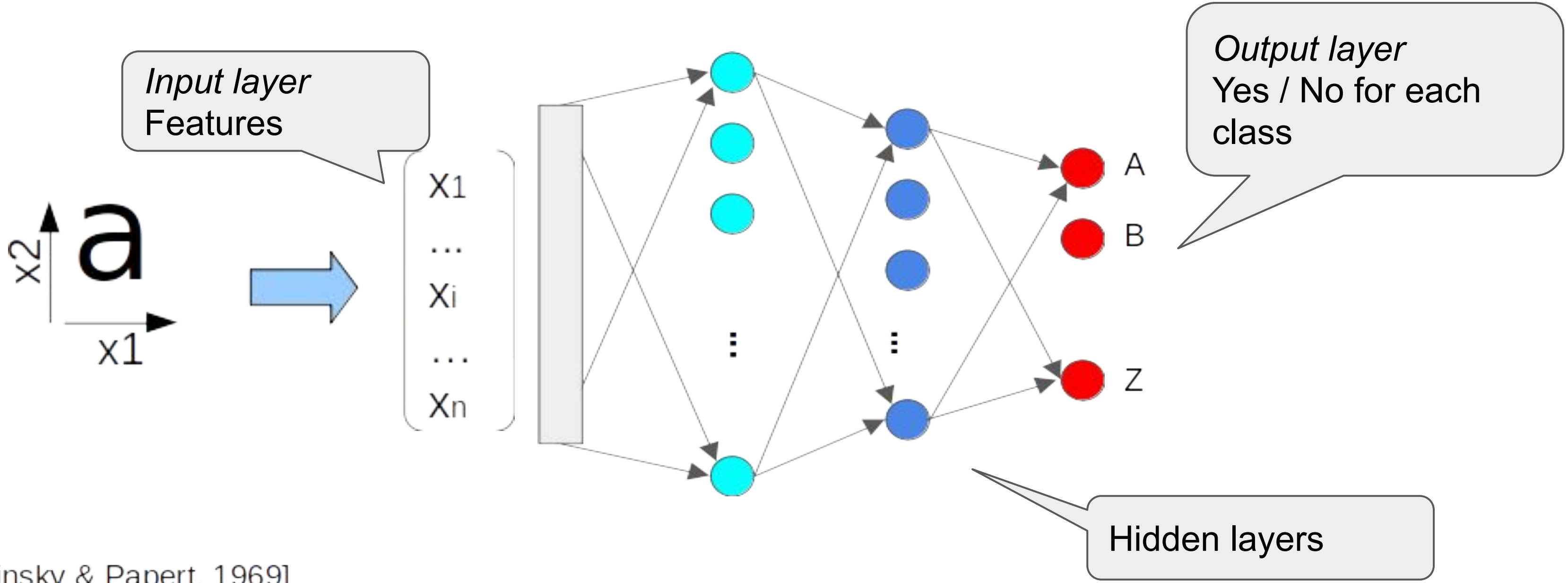
What is a “neuron” in machine learning?

A simple neuron



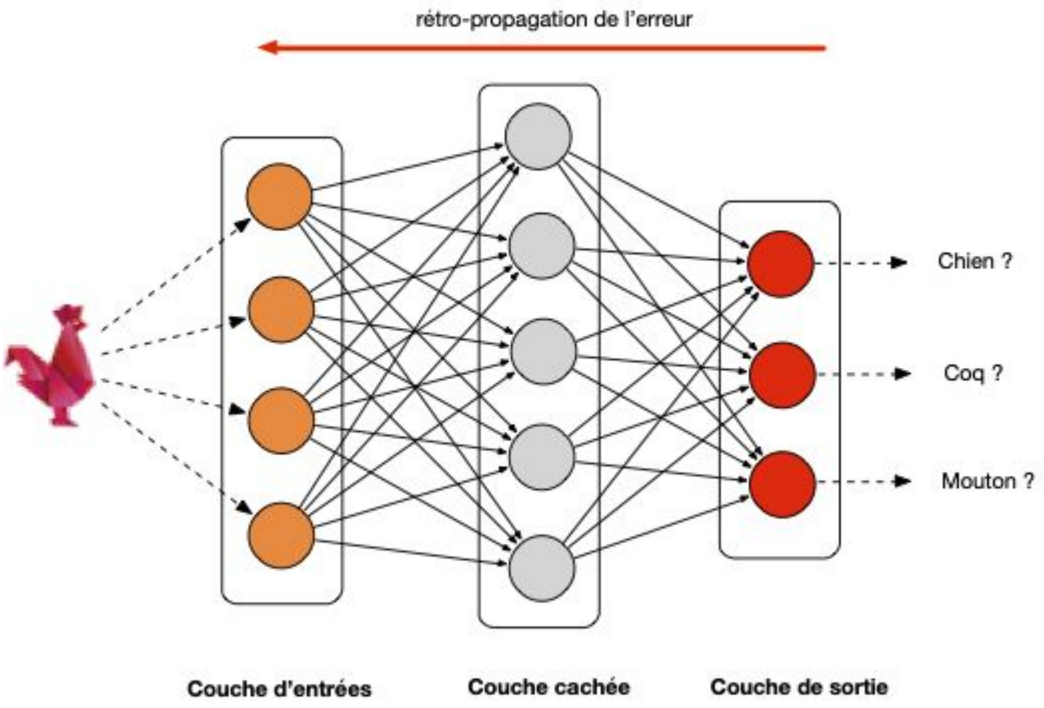
And a neural network?

A Multi-Layer Perceptron (MLP), for classification purposes



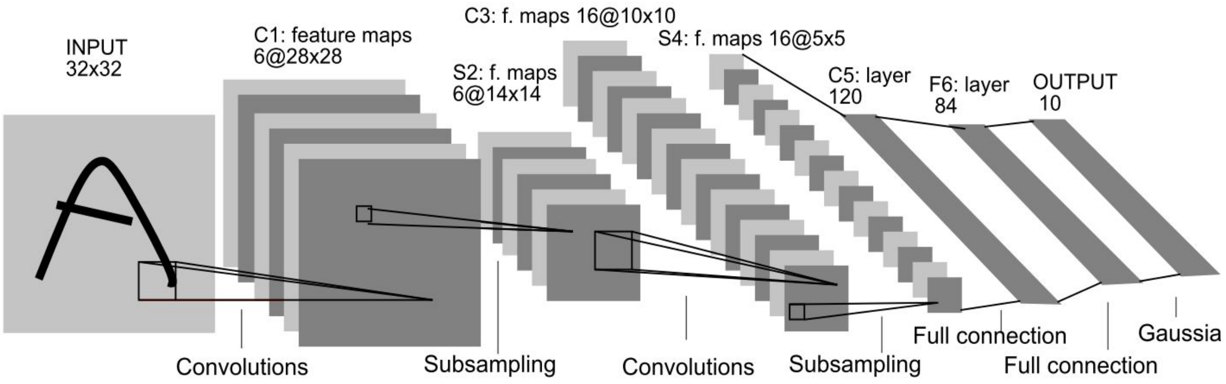
[Minsky & Papert, 1969]
[Parker 1982, Le Cun 1985, Rumelhart & McClelland 1986]

Example of a neural network

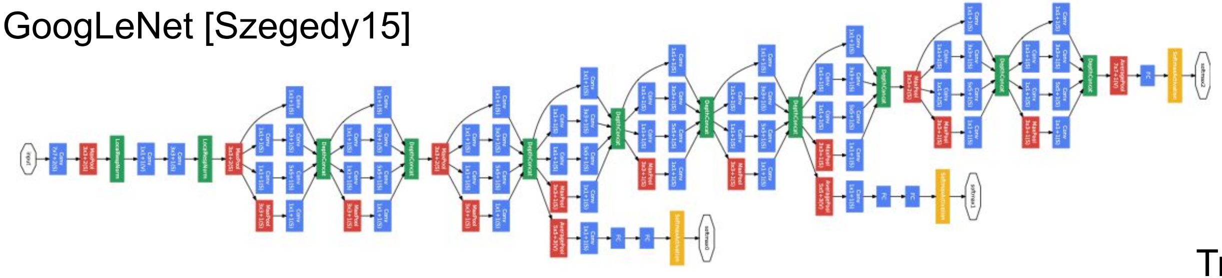


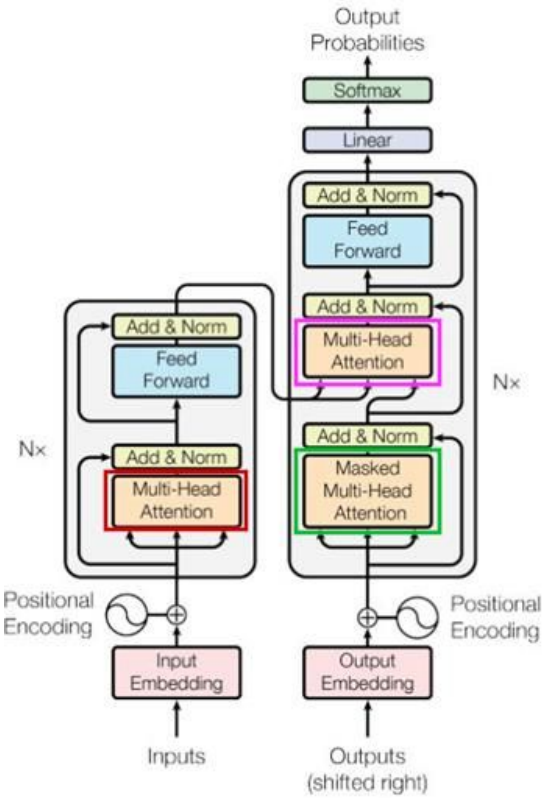
Deep neural networks

LeNet5 [LeCun98]



GoogLeNet [Szegedy15]

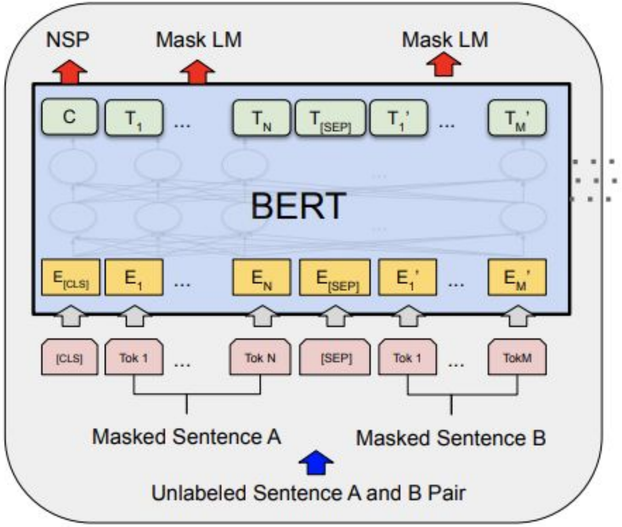




Transformers [Vaswani17]

Self-supervised learning

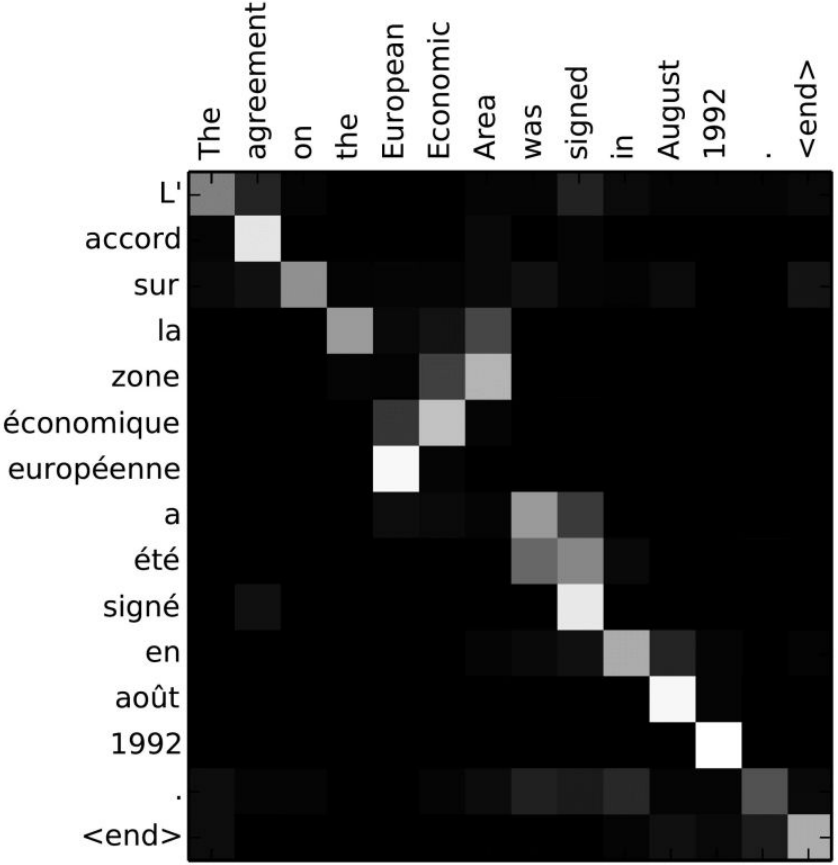
- Training a model without using labeled data (*unsupervised*)
- The “*labels*” are generated in relation to a targeted task
- Tasks examples with BERT (Bidirectional Encoder Representations from Transformers) :
  - Predicting missing words in a sentence (*masked language model*)
  - Predicting the next sentence



Pre-training

Attention mechanism

- The attention mechanism measures words that are significantly related between two given sequences
- In NLP, mechanism found in particular in Transformers architecture
- Illustration of the distribution of attention between two sequences in the context of machine translation



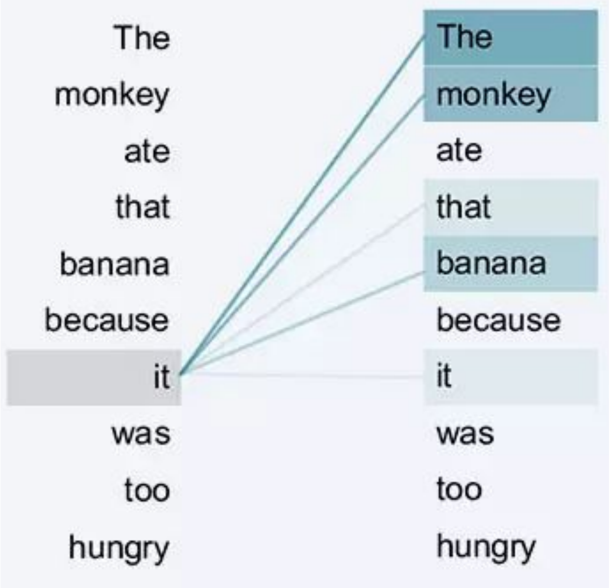
Source: (Bahdanau et al., 2015)

Attention is all you need [Vaswani17]

- The **self-attention mechanism** measures words that are significantly related in the input sequence (e.g. BERT)

- Allows you to better understand the relationships between words

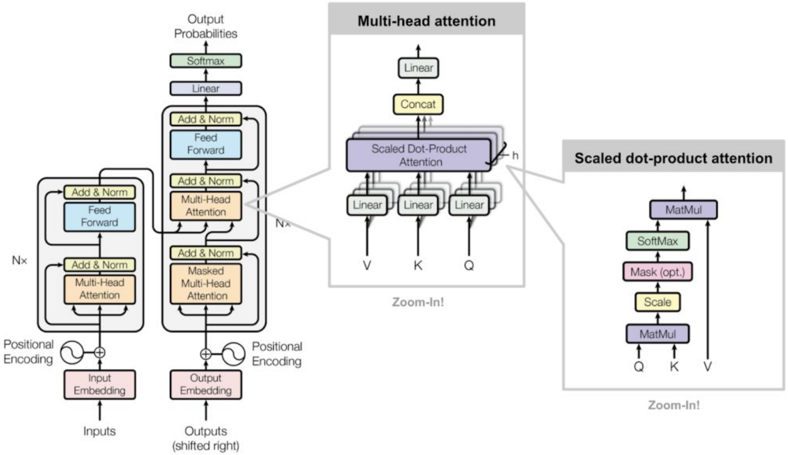
- Multi-head attention mechanism to capture relationships at different levels and of different types



Auto-attention. Source: (Xie et al. 2020)

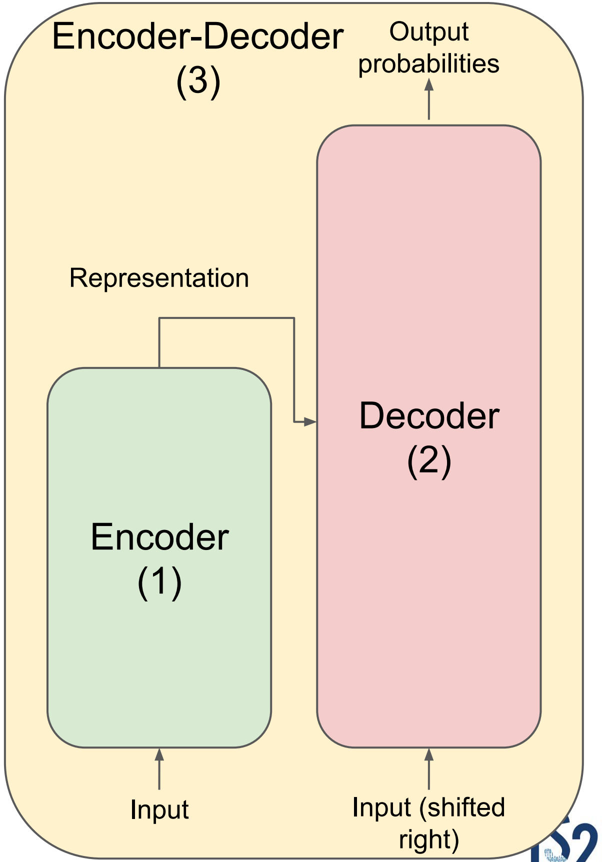
Attention is all you need [Vaswani17]

- Example of attention calculation in BERT
  - Projection of entries into 3 different spaces: Queries (Q), Keys (K), Values (V)
  - Q, K and V are linear representations of the input tokens which are used to calculate attention weights
  - Q and K: similarity between tokens
  - V: weighting of the importance of each token
  - Weight matrices to calculate Q, K and V are learned during training and iteratively adjusted

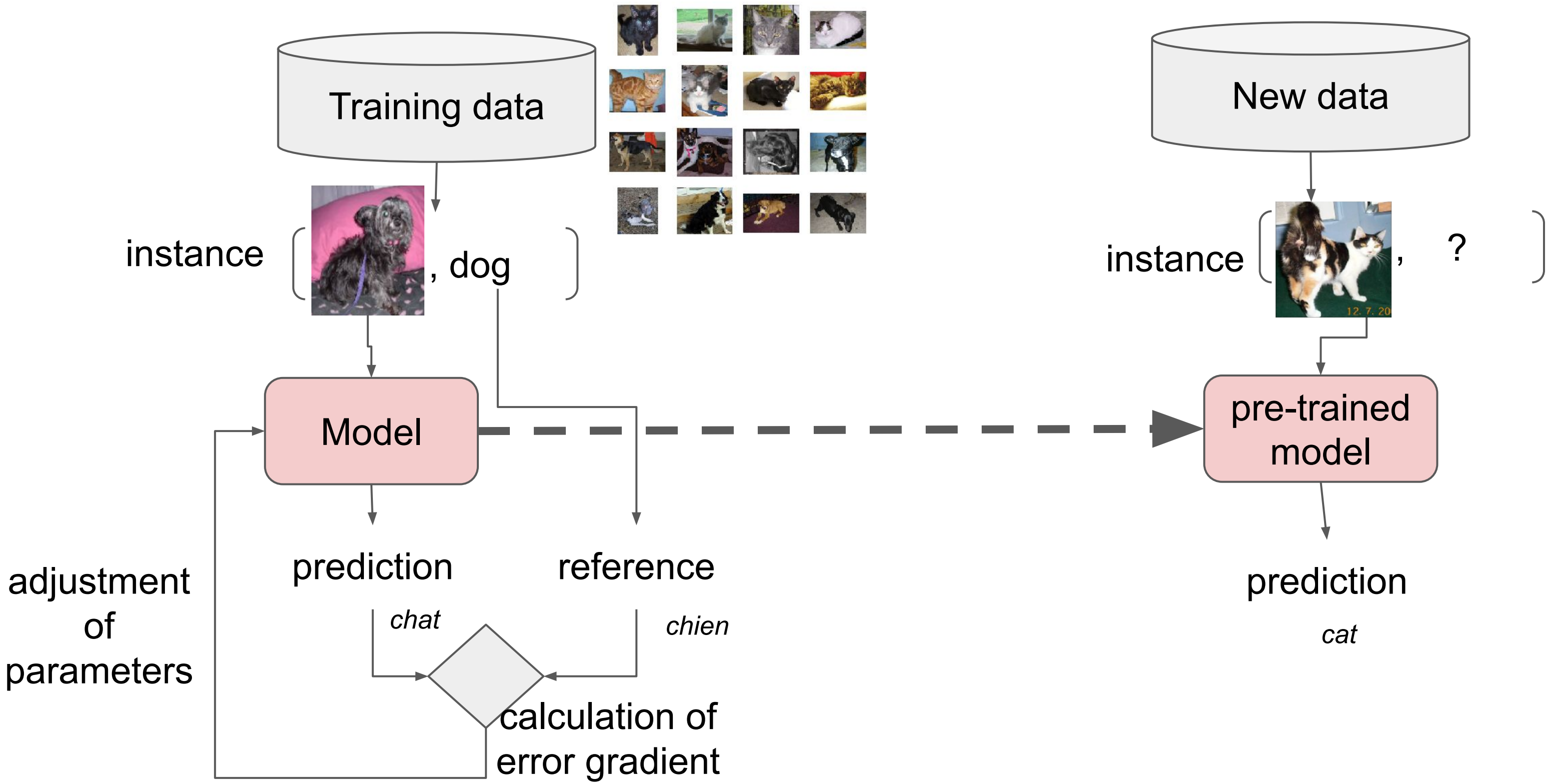


Mécanisme d’attention

- Le mécanisme d’attention mesure les mots qui sont significativement liés entre deux séquences données
- 3 modèles en 1

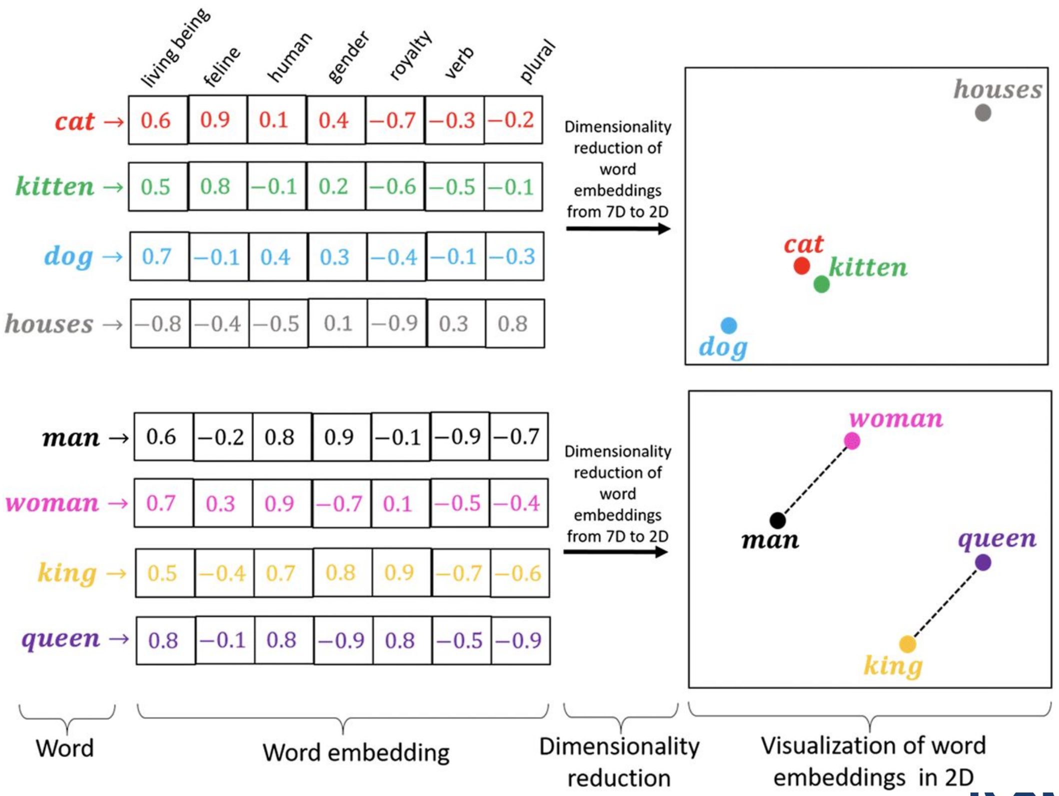


Training and inference



Embeddings

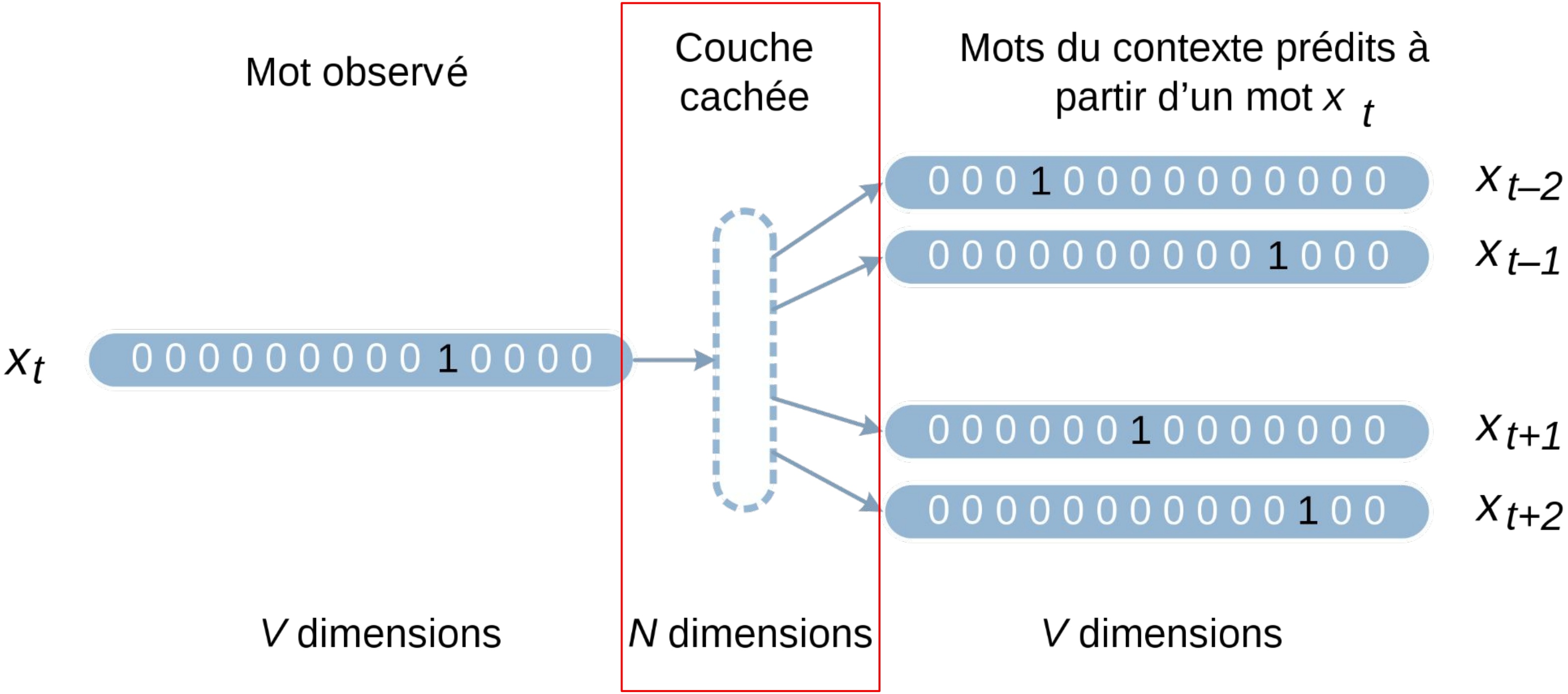
- Embeddings are vector representations of data (word, sound, image, etc.) in a multidimensional space
  - Real number vectors
  - In NLP, taking into account the contexts of word appearance
- Requires large amounts of unannotated data
- Possible arithmetic operations



https://medium.com/@hari4om/word-embedding-d816f643140

Word embeddings

Word2vec example (skip-gram): **hidden layer = embeddings**



Embeddings

- Embeddings allow you to integrate latent information linked to data
  - In NLP, lexical, grammatical, linguistic… or unidentified information!
- Embeddings can be fixed or contextual
- They are often used as input to other applications to replace data (image, text, etc.)
- In neural architectures, this representation is considered to come from an encoder model

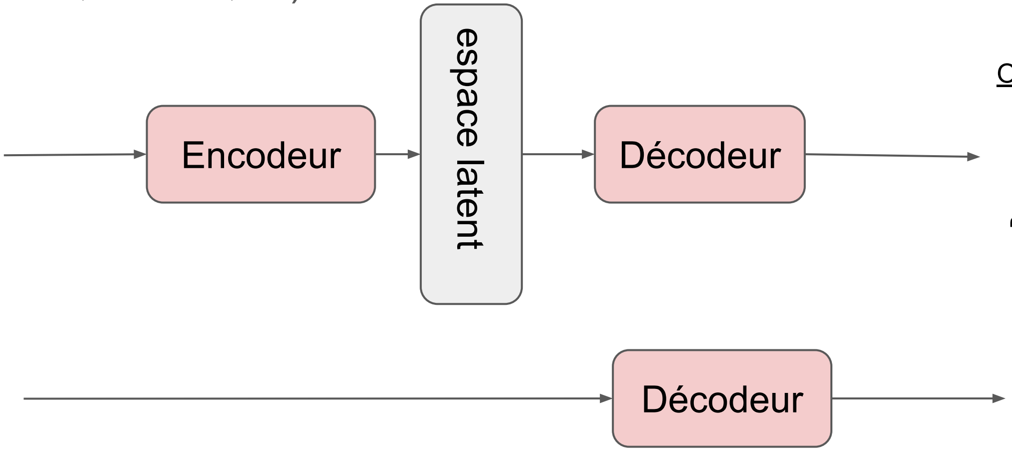
Le décodeur, l’architecture neuronale du modèle génératif

Modèle de langue : prédire le mot suivant à partir du début de la phrase
(ou le mot manquant au milieu de la phrase)

- Encodeur (Bert, …)
- Décodeur seul (GPT, LLaMA, …)

Input: (séquence qui précède)

Output: (mot suivant)

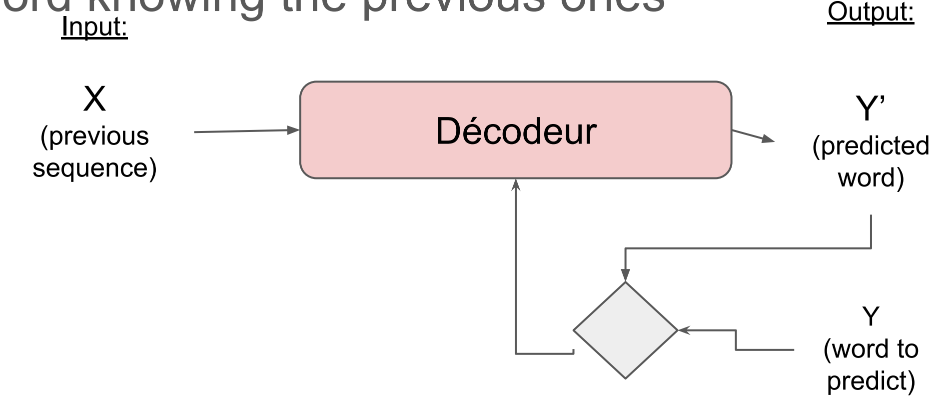


“ Le ciel est “ …

“ bleu ”

Decoder: the neural architecture of the generative model

<u>**Training**</u> of the prediction of the next word knowing the previous ones



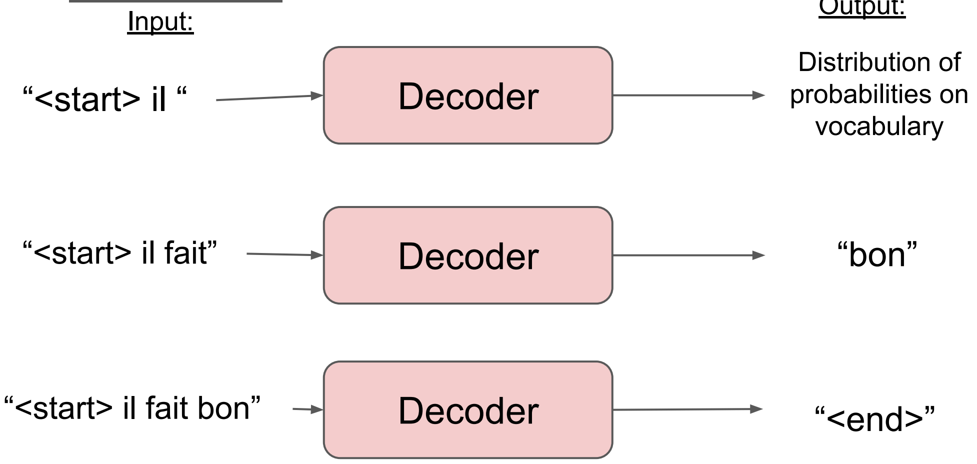
The “il fait beau” training sequence is rewritten into the following set of instances (X,Y):

	X (input sequence)	**Y** (word to predict)
$\mathbf{t_1}$	<start>	Il
$\mathbf{t_2}$	<start> Il	fait
$\mathbf{t_3}$	<start> Il fait	beau
$\mathbf{t_4}$	<start> Il fait beau	<end>

For each X, calculation of the error gradient then adjustment of the parameters

Decoder: the neural architecture of the generative model

Inference



	X	Y’		
		beau	bon	…
t_1	Il fait	0,12	**0,27**	
t_2	Il faut **bon**			

- Given a sequence X as input, the most probable next word is predicted
- This is added to the input sequence until the word “<end>” is generated

Model bias

- « Bias is prejudice in favor of or against a person or group, or something considered unfair » [Moll19]
  - What can be its impact?
  - How to control them? To remove them?
- Example: COMPAS criminal recidivism tool
  - Automatic score assigned based on data from an incarcerated person: education, place of residence, close entourage, etc.
  - Allows the judge to make a decision on his release
  - Study on the biases linked to this tool [ProPublica19]

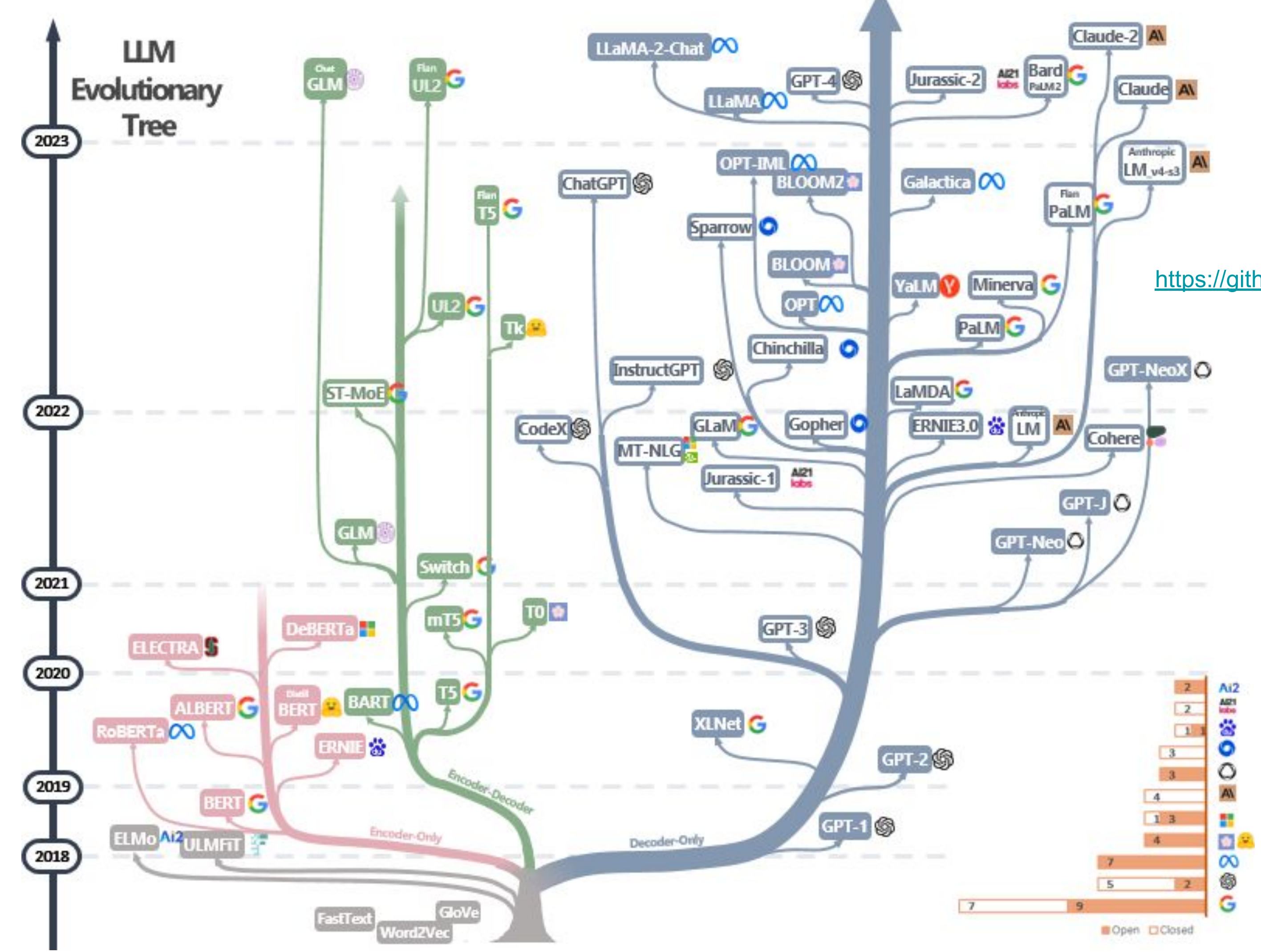
Error rate	European American	Afro American
Labeled high risk, but has not reoffended	23.5 %	44.9 %
Labeled low risk, and reoffended	47.7 %	28.0 %

[Corbière18]

Human biases in embeddings

- “Human” bias in word embeddings [Caliskan17]
  - Model trained on a corpus of standard texts from the Web
  - Replication of known biases according to implicit association test (social psychology)
  - In conclusion of the study, word representation models reproduce these biases (considered here as stereotypes)
- “Harmless” results
  - “Flowers” associated with “Pleasant” / “Insects” with “Unpleasant”
  - Same for “Weapons” and “Musical instruments”
- Results of societal stereotypes, more worrying
  - Names of people identified as “European American” more often associated with words “Pleasant” than “African American”
  - Female first names more often associated with words linked to “family” than to “career”, unlike men. Same for art (woman) and mathematics (man)

From language model to ChatGPT

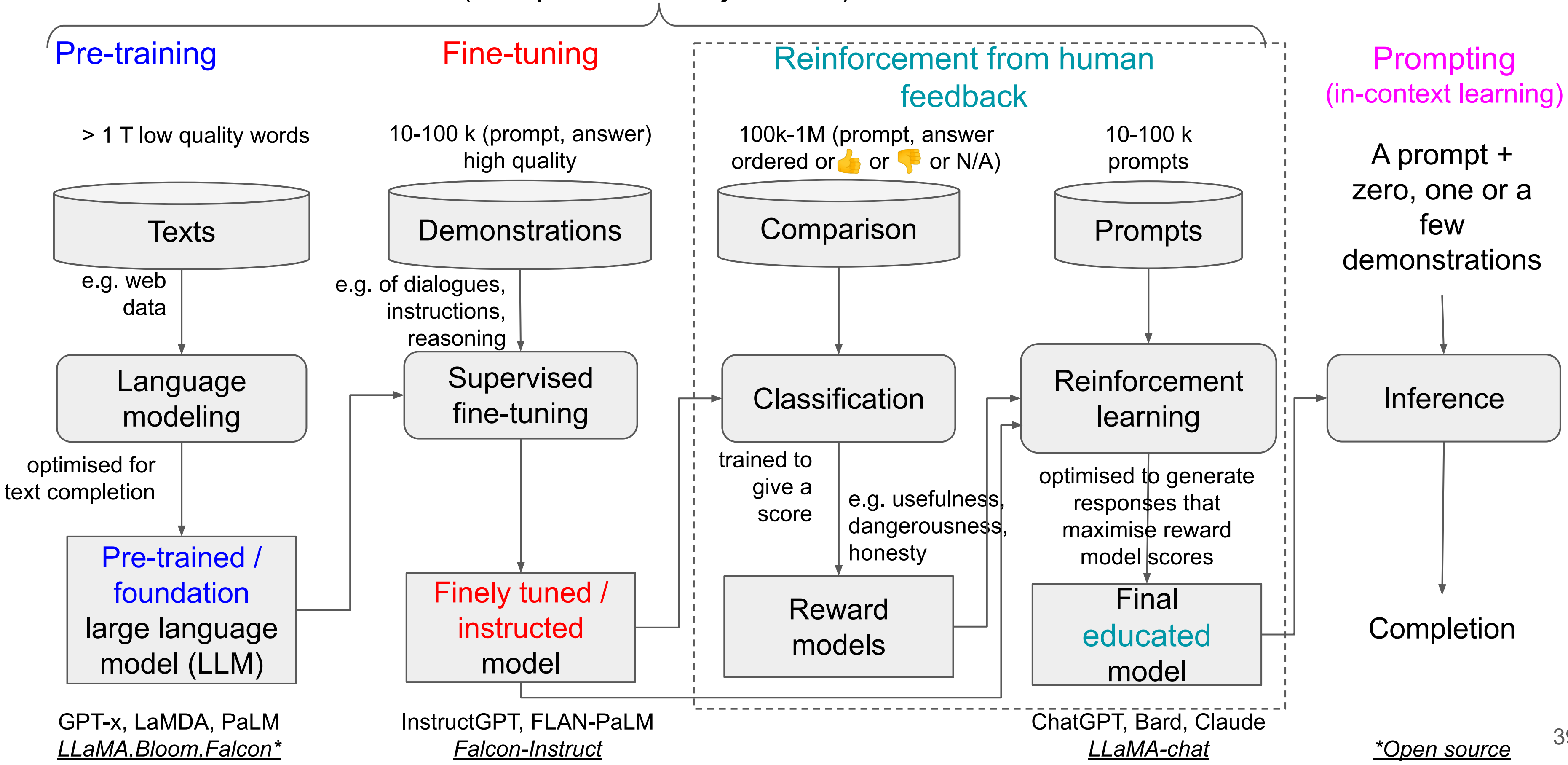


Source: Yang et al. 2023
https://github.com/Mooler0410/LLMsPracticalGuide

- Long line of models
- 4 architectures
- Different training strategies (single to multiple objectives, instructions, ideal number of parameters…)
- Free or proprietary models

Source: (Chip Huyen, May’23),
(llama2@Google),
(Ouyang@OpenAI, 22)

Model training
(with parameter adjustment)



Prompts zero of few-shot examples
(learning without model adjustment)

Zero-shot

Texte : L’histoire du film m’a captivé.
Sentiment :

Sentiment : positif

Prompt (written by a human)

Complétion (generated by the machine)

Some *few-shot* examples

Texte : Marie-Claude bondit sur toute la scène, dansant, courant, transpirant, s'essuyant le visage et faisant généralement preuve du talent unique qui lui a valu la célébrité.
Sentiment : positif

Texte : Malgré toutes les preuves du contraire, ce navet a réussi à se faire passer pour un vrai long métrage, le genre de film dont l'entrée est payante, qui fait l'objet d'un battage médiatique à la télévision et qui prétend amuser les petits enfants et les jeunes adultes.
Sentiment : négatif

Texte : L’histoire du film m’a captivé.
Sentiment :

Sentiment : positif

[Language Models are Few-Shot Learners](#) (Brown et al. @OpenAI, May 2020)

Prompt with instruction, reasoning*

Instruction

Stp répond à la question suivante :
Quel est le point d'ébullition de l'azote ?

L'azote liquide a une température d'ébullition de -196 °C.

Traduire de l'anglais au français :
cheese =>

Fromage

Finetuned Language Models Are Zero-Shot Learners (Wei et al. @Google, 2022)

Reasoning (*chain-of-thought*)

Q : Roger a 5 balles de tennis. Il en achète deux tubes de balles de tennis. Chacun des tubes compte 3 balles. Combien a-t-il de balles maintenant ?

R : Rocher a démarré avec 5 balles. 2 tubes de 3 balles de tennis correspond à 6 balles de tennis. 5+6 = 11. La réponse est 11.

Q : La cafétéria avait 23 pommes. Elle en a utilisé 20 pour le déjeuner et en a acheté 6 de plus, combien de pommes reste-il ?

La cafétéria avait 23 pommes à l'origine. Elle en a utilisé 20 pour préparer le déjeuner. Elle avait donc 23 - 20 = 3. Elle a acheté 6 pommes supplémentaires, elle a donc 3+6 = 9.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al. @Google, NeurIPS 2022)

Invite to reason

Réponds à la question suivante en raisonnant étape par étape :
La cafétéria avait 23 pommes. Elle en a utilisé 20 pour le déjeuner et en a acheté 6 de plus, combien de pommes reste-il ?

La cafétéria avait 23 pommes à l'origine. Elle en a utilisé 20 pour préparer le déjeuner. Elle avait donc 23 - 20 = 3. Elle a acheté 6 pommes supplémentaires, elle a donc 3+6 = 9.

« Let's think step by step »
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

*Relies on LLMs and/or instructed models

The art of the prompt

Give clearer instructions (make him adopt a personality, etc.)

Break complex tasks into simpler subtasks

Structure the instruction so that the model does not deviate from its task

Invite the model to explain before responding

Ask for justifications for many possible answers, then synthesize

Generate many results, then use the model to choose the best one

Fine-tune custom models to maximize performance

Sources : GPT Les meilleurs pratiques ; OpenAI livre de cuisine pour améliorer la fiabilité ; Prompt engineering (Lilian Weng Mar 2023)

LLM abilities

Emergent abilities with the LLM size increasing

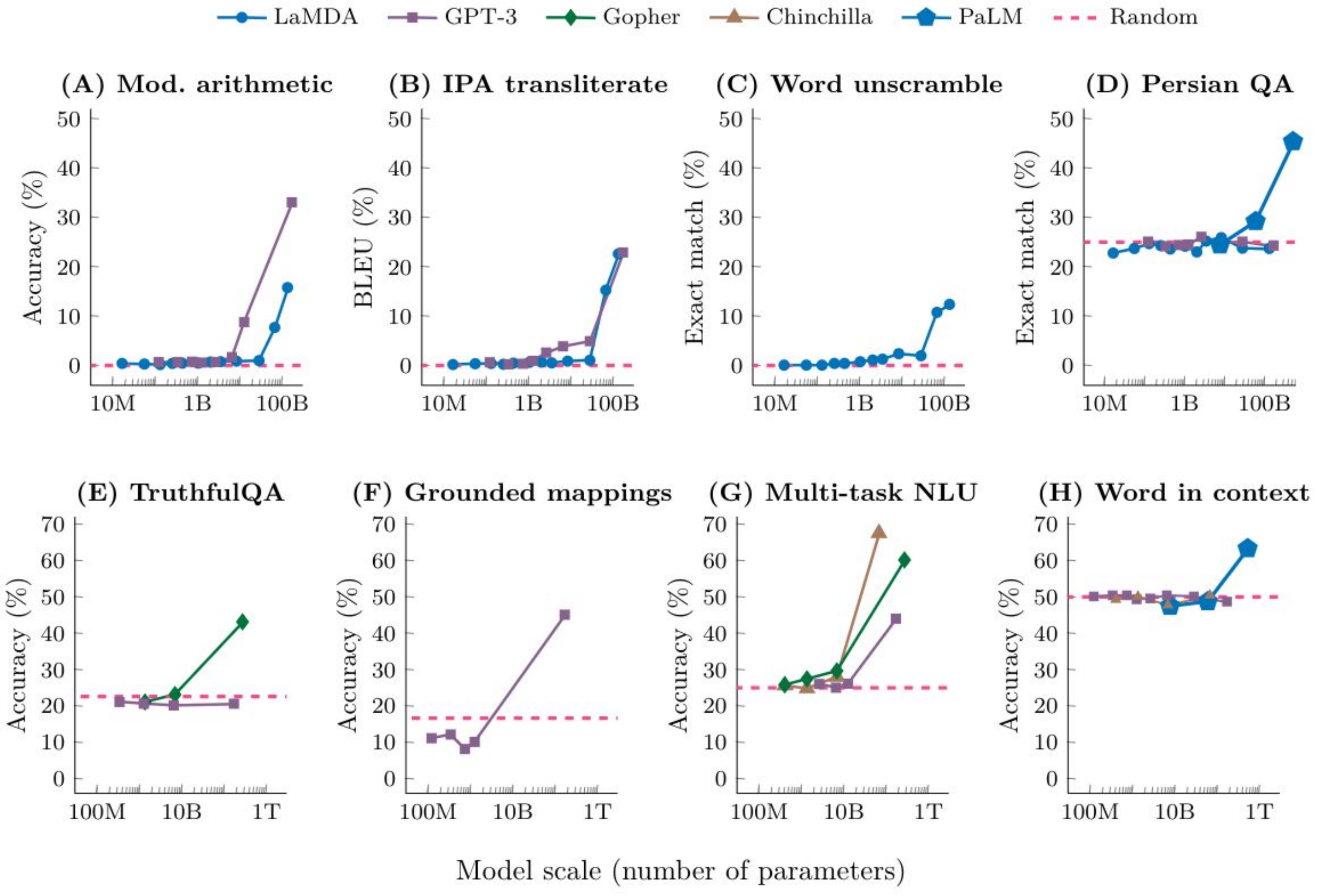
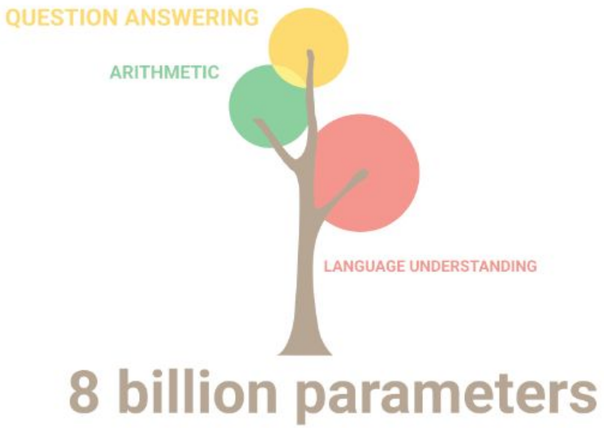
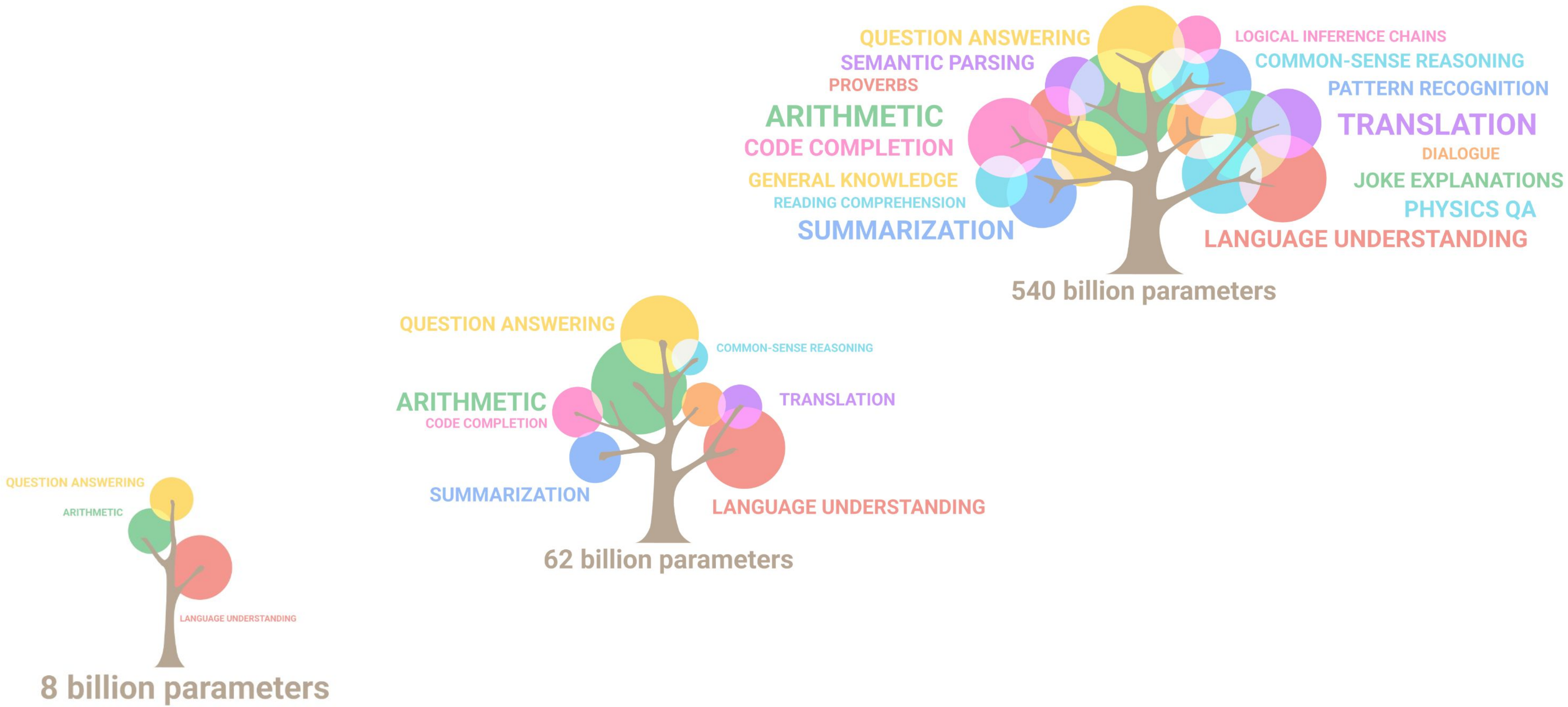


Figure 11: Eight examples of emergence in the few-shot prompting setting. Each point is a separate model.

Source: [Emergent Abilities of Large Language Models](#) (Wei et al. @Google, 28 Feb 2023)

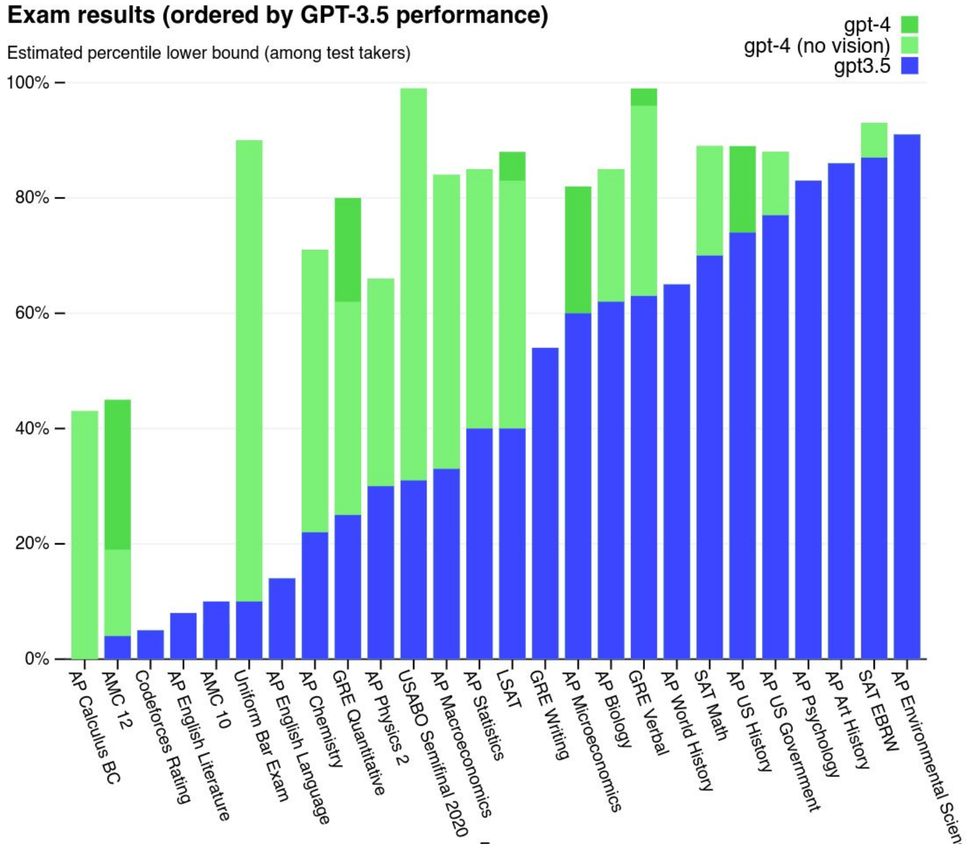


Source: PaLM: Scaling Language Modeling with Pathways (Chowdhery et al. @Google, 5 Oct 2022) and https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html



Source: [PaLM: Scaling Language Modeling with Pathways](https://) (Chowdhery et al. @Google, 5 Oct 2022) and https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html

Ranking of GPT* in university entrance exams



- In the top 20 of participants (score > 80% of participants)
- Mostly MCQs (multiple choice)

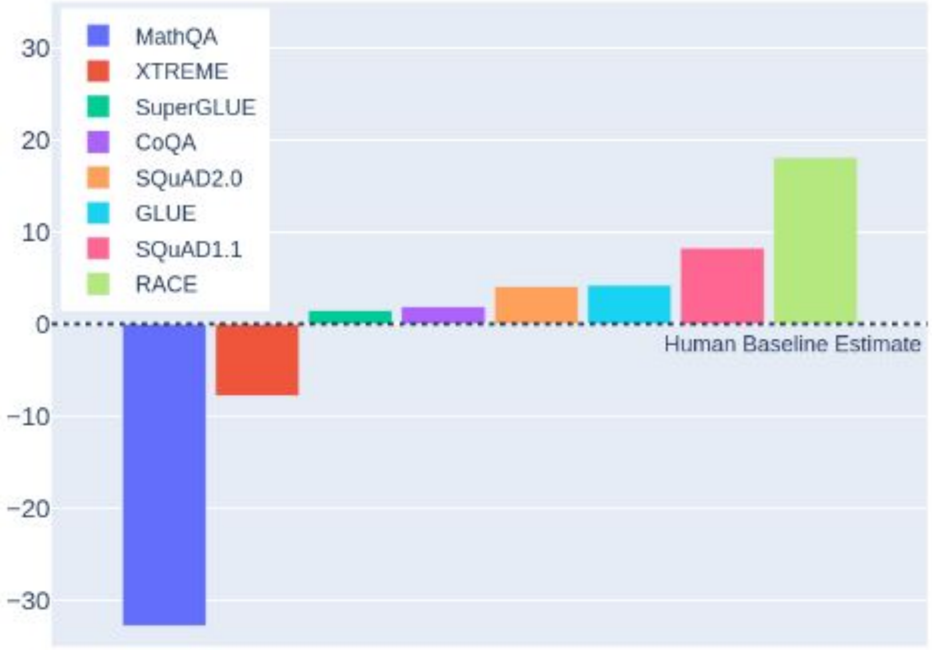
Source: [GPT-4 Technical Report,](GPT-4 Technical Report) (@OpenAI, 27 Mar 2023)

PLMs truly have superhuman abilities… on the benchmarks

In term of score differences, **best-performing systems outperforms humans on 6 out of 8 in NLP benchmarks** on language understanding, reasoning, and reading comprehension

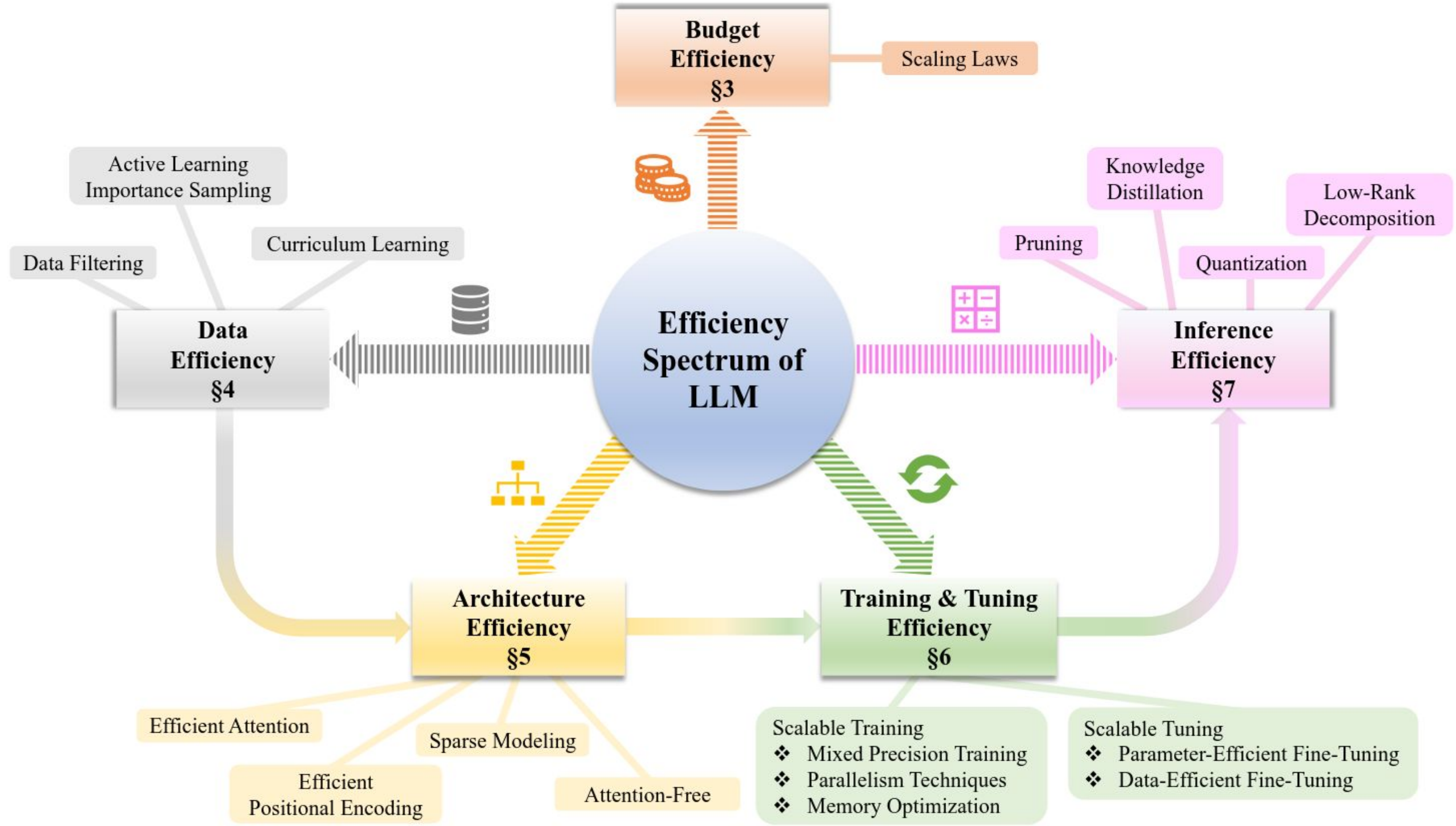
The authors show that these **benchmarks have serious limitations affecting the comparison between humans and PLMs**
E.g. train-test data splits composition make big differences, automated evaluation is limiting, Human often disagree…

Provide recommendations for fairer and more transparent benchmarks
E.g. complement wi human judgements, Balance easy and hard test set items…



Source: What's the Meaning of Superhuman Performance in Today's NLU? (Tedeschi et al., ACL 2023)

Predicting the efficiency of LLMs



Source: [The Efficiency Spectrum of Large Language Models: An Algorithmic Survey] (Ding et al. @Microsoft, 1 Dec 2023)

Law scale

“*we assume that the efficient computational frontier can be described by a power-law relationship between the compute budget, model size, and number of training tokens.*”

Source: Training Compute-Optimal Large Language Models... (Hoffmann et al. @DeepMind, 29 Mar 2022) … the Chinchilla model

Some existing LLMs could have achieved better performance within the same compute budget.

To go further: Compressing LLMs: The Truth is Rarely Pure and Never Simple (Jaiswal et al. @Apple, 2 Oct 2023)

Generative but not creative

Generative but not creative AI

Completion of prompts according to a distribution of words in its training corpus with a little chance, and when it doesn't find it, it invents...

Limitations

- Capable of **hallucinations** (absurd or false content compared to certain sources)
  - “In reality, he hallucinates all the time, and sometimes his hallucinations are coherent for humans…”
- Comments that are **harmful** to an individual or group (hateful, discriminatory, inciting violence)
- Could plagiarize
- **Sensitive to training data** (bias and discrimination, cultural hegemony, poisoning by generated data, etc.)
- Store factual **knowledge in their parameters** and accessing, manipulating, updating or providing provenance are open research questions

Measures exist to counter these limitations; some are research subjects

How to reduce the hallucinations, how to control and manipulate their knowledge?

LLMs store factual **knowledge in their parameters** and accessing, manipulating, updating or providing provenance are open research questions

Retrieval Augmented Generation (RAG) models

seq2seq architecture (BART-large 400M)

BERT Query and document Encoders

Vector index of Wikipedia

Generator takes a concatenation of x and z as a context

Training relies on fine-tuning the query encoder BERT and the BART generator

Outperforms non-augmented seq2seq

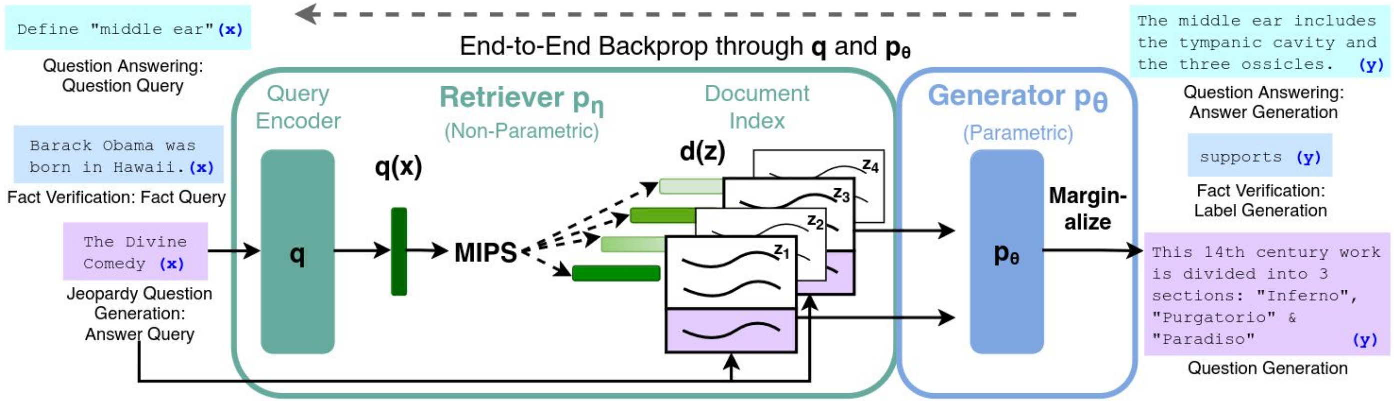


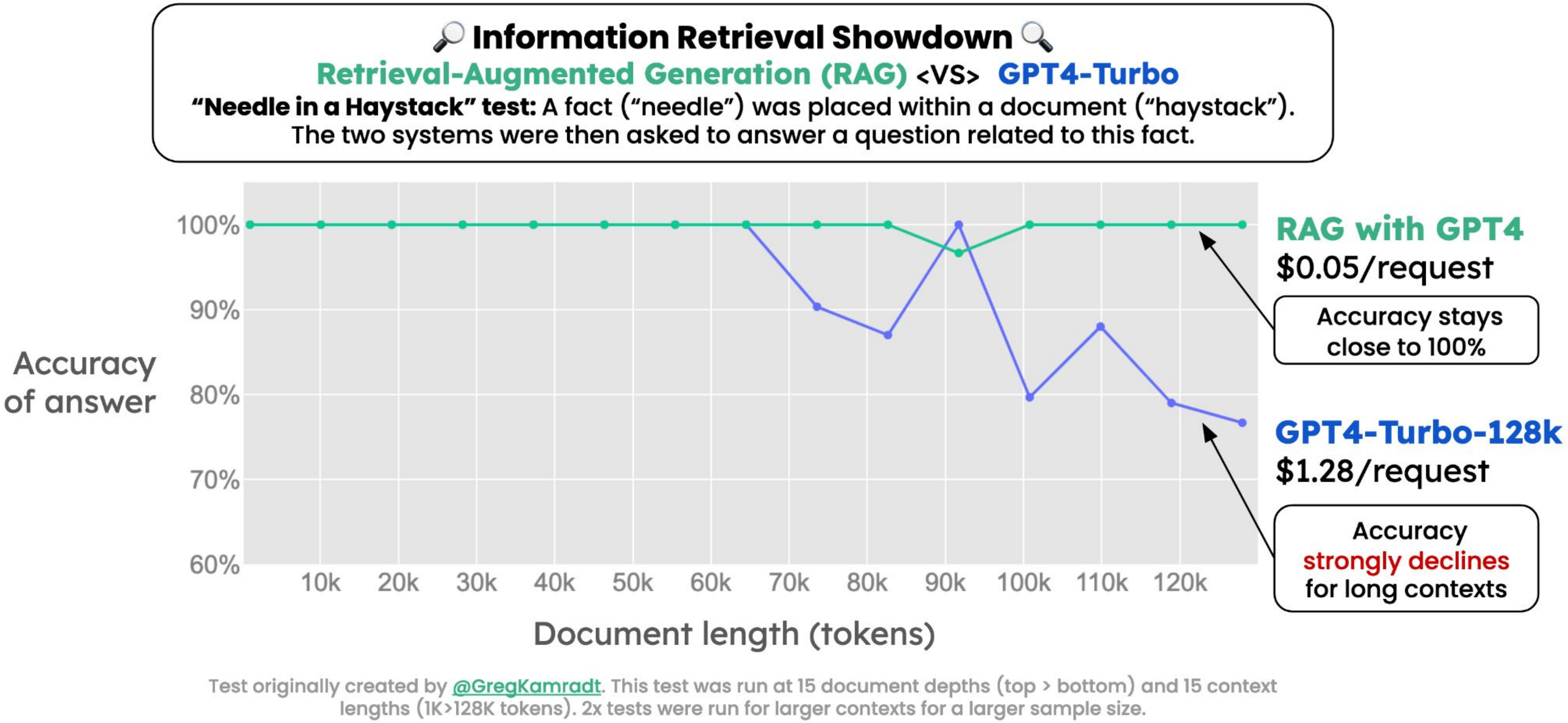
Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Able to precisely access controlled factual knowledge and provide provenance !

Source: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al. @Facebook, 12 Apr 2021)

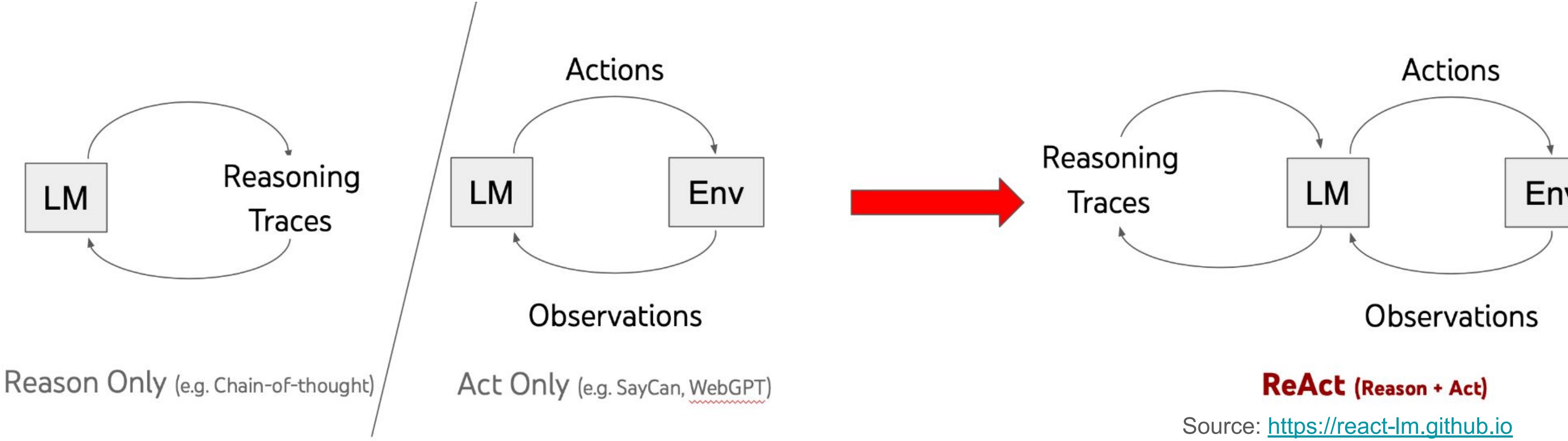
Information Retrieval: Who wins, GPT-4-Turbo or a RAG based on GPT4?

Source : @HuggingFace (2 Dec 2023)
https://github.com/A-Roucher/LLM_vs_RAG_NeedleInAHaystack



Interaction a way of controlling and to extend the LLM abilities

Models interacting with the world



Source: https://react-lm.github.io

LLM as a factual knowledge store interacting with the environment to build incremental prompts to query the LLM

Models are trained to generate behaviours whose plausibility or results extend the prompt until a final state is reached.

Source: WebGPT: Browser-assisted question-answering with human feedback (Nakano et al. @OpenAI, 1 Jun 2022)
Source: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (Ahn et al. @Robotics at Google, 16 Aug 2022)
Source: ReAct: Synergizing Reasoning and Acting in Language Models (Yao et al. @Google Research, 10 Mar 2023)
Source: Build context-aware, reasoning applications with LangChain's flexible abstractions and AI-first toolkit

SayCan

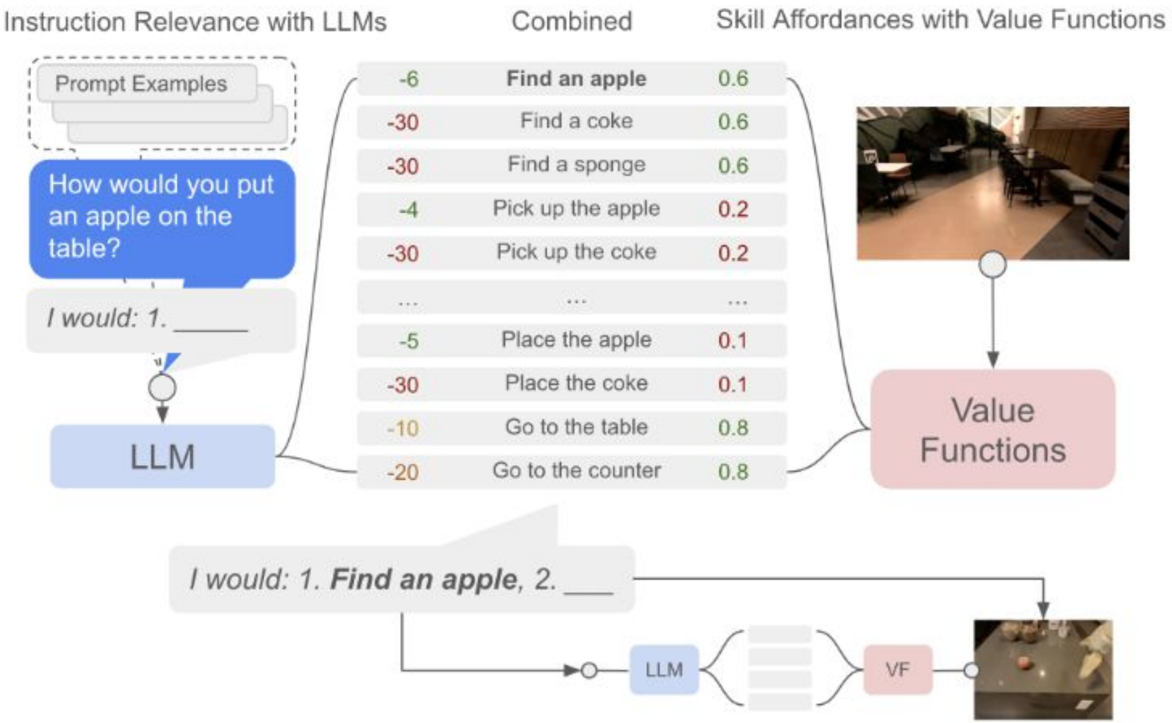


Figure 3: Given a high-level instruction, SayCan combines probabilities from a LLM (the probability that a skill is useful for the instruction) with the probabilities from a value function (the probability of successfully executing said skill) to select the skill to perform. This emits a skill that is both possible and useful. The process is repeated by appending the skill to the response and querying the models again, until the output step is to terminate. Appendix Figures 12 and 2 focus on the LLM and VFS components.

Source: [Do As I Can, Not As I Say: Grounding Language in Robotic Affordances](#) (Ahn et al. @Robotics at Google, 16 Aug 2022)

What the future will look like?
(Actually today)

Multimodal IAs

GPT-4 is a multimodal LLM

Computer Vision

- Depth Estimation — 87 models
- Image Classification — 7,425 models
- Image Segmentation — 354 models
- Image-to-Image — 228 models
- Object Detection — 1,272 models
- Video Classification — 418 models
- Unconditional Image Generation — 978 models
- Zero-Shot Image Classification — 235 models

Natural Language Processing

- Conversational — 2,612 models
- Fill-Mask — 9,569 models
- Question Answering — 8,486 models
- Sentence Similarity — 2,725 models
- Summarization — 1,427 models
- Table Question Answering — 79 models
- Text Classification — 41,853 models
- Text Generation — 34,732 models
- Token Classification — 13,792 models
- Translation — 3,025 models
- Zero-Shot Classification — 194 models

Audio

- Audio Classification — 1,426 models
- Audio-to-Audio — 3,632 models
- Automatic Speech Recognition — 12,871 models
- Text-to-Speech — 1,691 models

Tabular

- Tabular Classification — 166 models
- Tabular Regression — 103 models

Multimodal

- Document Question Answering — 112 models
- Feature Extraction — 6,288 models
- Image-to-Text — 317 models
- Text-to-Image — 12,783 models
- Text-to-Video — 71 models
- Visual Question Answering — 95 models

Reinforcement Learning

- Reinforcement Learning — 32,821 models

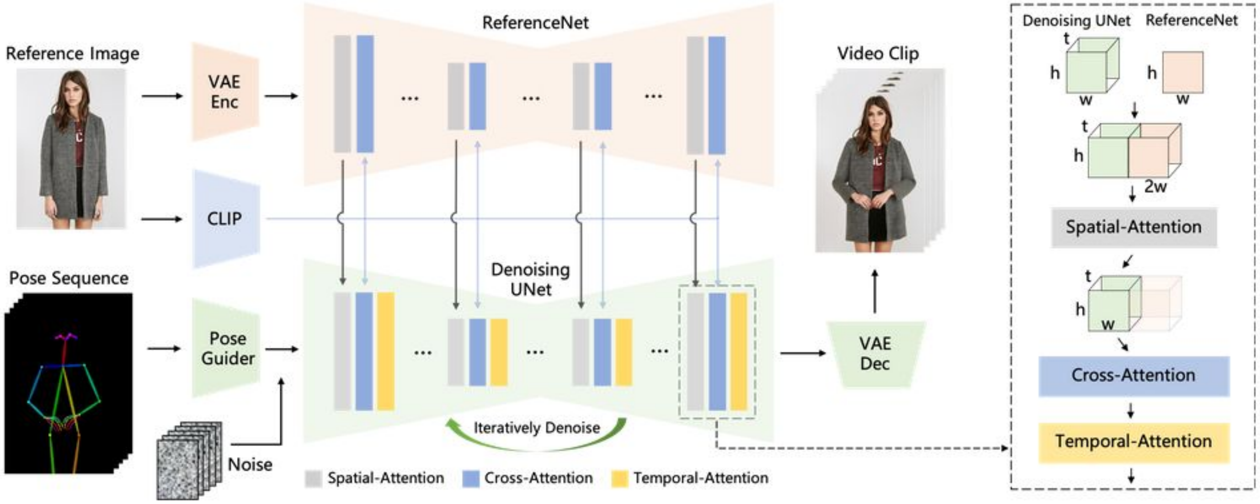
Hugging Face

"Home for all Machine Learning tasks"

Source: https://huggingface.co/tasks (1 Dec 2023)

The end of influencers ?

Source: [Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation](#) (Hu, 28 Nov 2023)



Beware of identity theft -> importance of controlling the circulation of personal photos

OpenAI directions: custom GPTs

Custom versions of ChatGPT that combine specialized prompted, extra knowledge, and any combination of skills like searching the web, making images or analyzing data…

Source: [Introducing GPTs] @OpenAI blog (Nov. 6, 2023)
Source: [New models and developer products announced at DevDay] @OpenAI blog (Nov. 6, 2023)

All GPTs!
#1 GPT Directory

GPTs list is Updated daily

Made by → @johnrushx

research

All · Analysis · Assistant · Audio · Coach · Converter · Design · Ecommerce · Education · Fundraising · Guide · Hobby · Humor · Marketing · Math · Medical · News · Programming · Recipe · Research · SEO · Social · Video · Web · Web3 · Writing

Research Mentor
Committed to guiding students through their research projects

The Researcher
Specialized AI tool tailored for deep researchers

ScholarGPT
I answer your research questions.

Explain Anything
I'm an expert in research and explanations / making complex topics clear and accessible.

AnalyzePaper
Serves as a writing assistant adept at reviewing research papers and articles. Evaluates claims; assesses study quality; gauges result confidence; and delivers concise summaries

ScholarAI
ScholarAI - Your Research Assistant ; unlock scientific research from over 200M papers

Literature Reviewer
Assists you with reviewing and rating quality research and provides you with a summary of findings. Start by uploading a PDF or providing a link.

Professor Edit
A professor aiding in research paper editing.

Research Reviewer
I write paper review

Source:
https://allgpts.co

OpenAI directions: custom GPTs **and custom models**

Special programme to pre-trained or fined-tuned GPT models

Source: New models and developer products announced at DevDay @OpenAI blog (Nov. 6, 2023)

No need to worry about licences… they will soon no longer taken into account

Copyright Shield

OpenAI is committed to protecting our customers with built-in copyright safeguards in our systems. Today, we're going one step further and introducing Copyright Shield—we will now step in and defend our customers, and pay the costs incurred, if you face legal claims around copyright infringement. This applies to generally available features of ChatGPT Enterprise and our developer platform.

Source: New models and developer products announced at DevDay @OpenAI blog (Nov. 6, 2023)

With GPT-3.5-turbo (ChatGPT),

2+2 = ?

With GPT-3.5-turbo (ChatGPT),

2+2 = ?

175 billions of computation…

Energy and therefore ecological issues

Energy cost and therefore ecological

Related to amount of training data (train.) and # of parameters (train. and inference)

Exple: LLaMA trained on **1T words** (1,000 milliards) during **21 days** on **2,048 GPUs**

	GPU Type	GPU Power consumption	GPU-hours	Total power consumption	Carbon emitted (tCO_2eq)
OPT-175B	A100-80GB	400W	809,472	356 MWh	137
BLOOM-175B	A100-80GB	400W	1,082,880	475 MWh	183
LLaMA-7B	A100-80GB	400W	82,432	36 MWh	14
LLaMA-13B	A100-80GB	400W	135,168	59 MWh	23
LLaMA-33B	A100-80GB	400W	530,432	233 MWh	90
LLaMA-65B	A100-80GB	400W	1,022,362	449 MWh	173

1 tCO2eq (or 1,000kg CO2eq) corresponds to emissions of (source: educlimat):

1 lamp lit for 54 years ⇔ **1 house heated with gas for 1 year** ⇔ 12 km / day by car for a year ⇔ 2 beef dishes per week for 1 year

…LLaMA2, released 5 months later, is trained on **2T words**

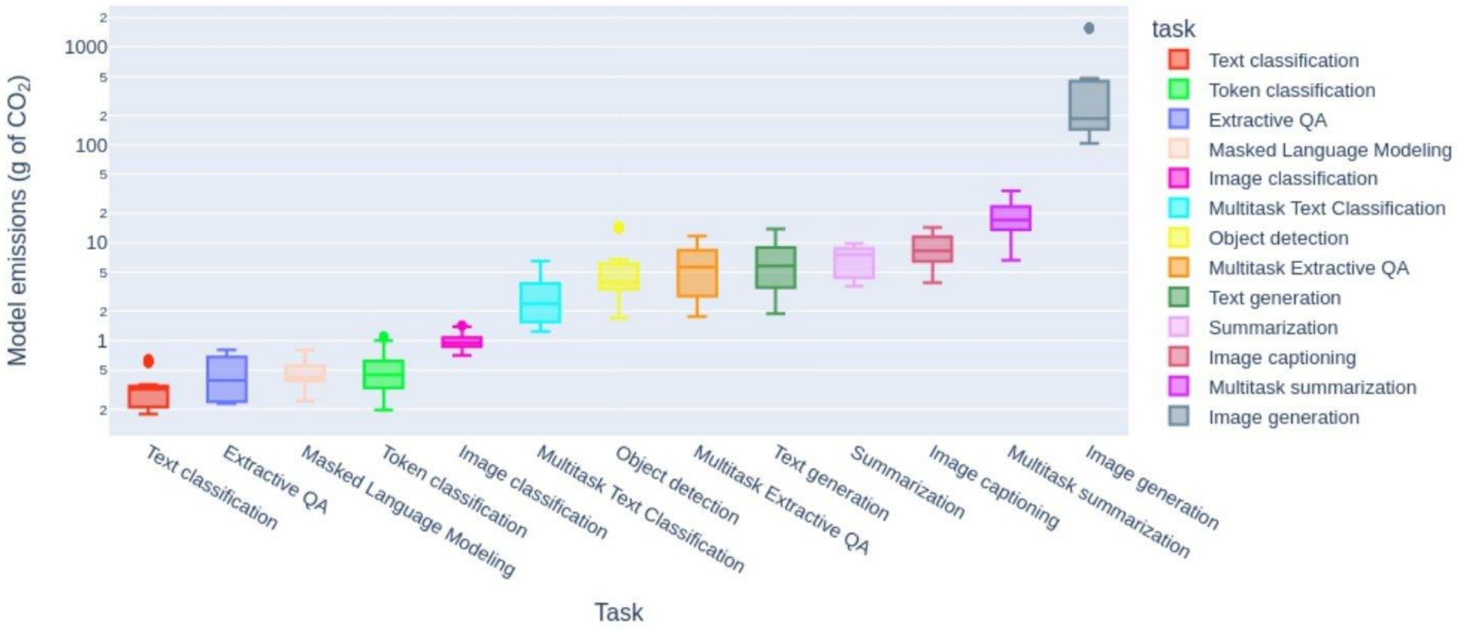
Source: LLaMA: Open and Efficient Foundation Language Models, (Touvron et al.@Meta, 27 Feb, 2023)

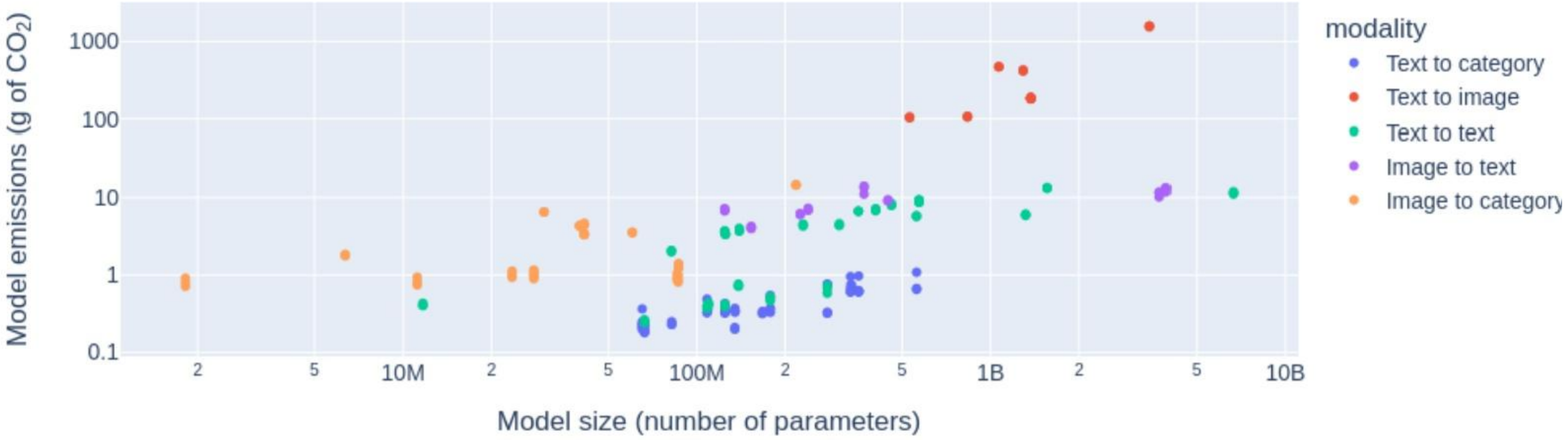
Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

ALEXANDRA SASHA LUCCIONI and YACINE JERNITE, Hugging Face, Canada/USA
EMMA STRUBELL, Carnegie Mellon University, Allen Institute for AI, USA

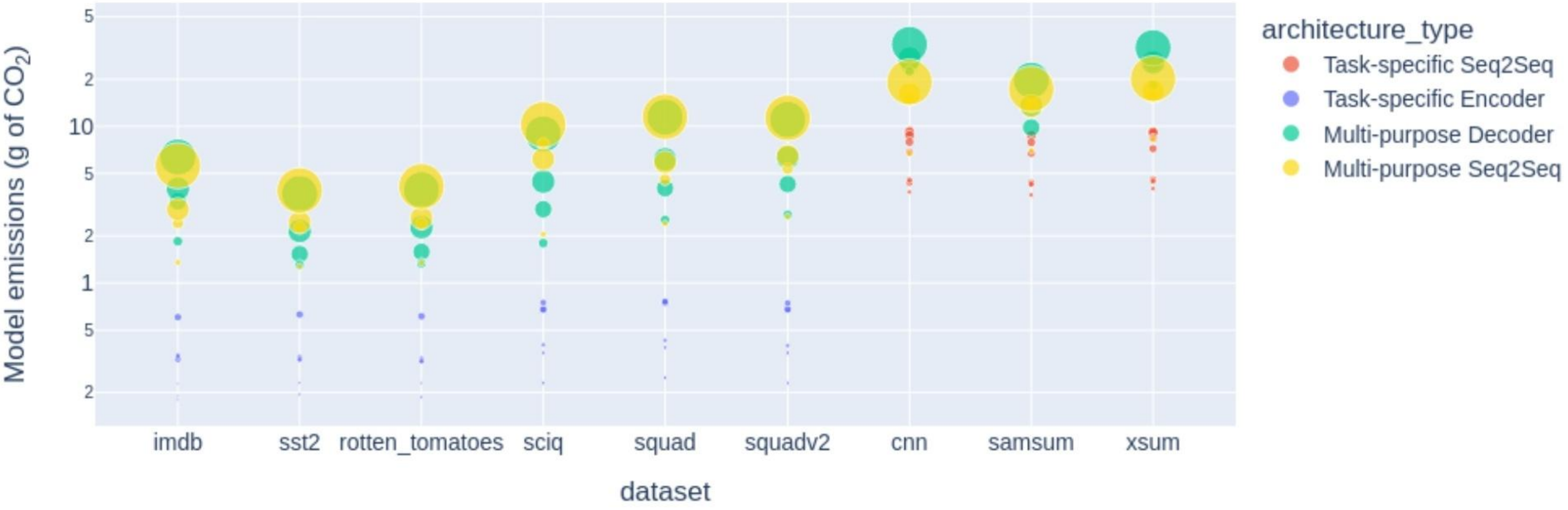
88 models tested on 30 datasets from 10 different tasks from different modalities

Source : https://arxiv.org/abs/2311.16863 (28 Nov 2023)





Generative tasks and ones that involve **images are more energy- and carbon-intensive** compared to discriminative tasks and ones that involve text.



Using multi-purpose models for discriminative tasks is more energy-intensive compared to models fine-tuned for these same tasks.

This is especially the case for tasks like sentiment analysis and question answering. -- the difference can be a factor of 30

	BLOOMz-7B	BLOOMz-3B	BLOOMz-1B	BLOOMz-560M
Training energy (kWh)	51,686	25,634	17,052	10,505
Finetuning energy (kWh)	7,571	3,242	1,081	543
Inference energy (kWh)	1.0×10^{-4}	7.3×10^{-5}	6.2×10^{-5}	5.4×10^{-5}
Cost parity (# inferences)	592,570,000	395,602,740	292,467,741	204,592,592

Table 5. The BLOOMz models from our study with their training energy cost (from [31]), finetuning energy cost (from [34]), inference cost (from the present study), and cost parity, as the number of inferences required to sum to the training cost.

Training remains orders of magnitude more energy- and carbon- intensive than inference.

It takes between 200 and 500 million inferences for an LLM to reach the quantity of energy used during training.


Sasha Luccioni, PhD 💻🌎🦋✨🤗 @SashaMTL · 6h
TL;DR? Stuffing generative models into absolutely everything comes with a significant cost to the planet, and we should use fine-tuned models in cases when tasks are well-defined. 👩‍💻
Alternative titles we explored include "InferNO" and "Think before you GPT" 😂
6
22
87
4 k

Take home messages

The take-home messages about generative AI

- Generative but not creative
- Tools above all
- Should we be afraid of the tool? Should we be afraid of a hammer? Its use must be defined socially together
- Talk about tasks more than jobs; some changes are coming, AI will mainly complement humans
- Way to increase productivity and profitability; but also to free up time to do other things and it is on the reinvestment of this time that we must work
- Rethink educational objectives (need to learn to work “with”) and rethink assessments to reflect business reality (difficult to “bury your head in the sand”)
- Think about the impact of using it to act in consequence

Open research questions and mores issues

- openness and sovereignty of LLMs
- sobriety and compression of LLMs
- social justice at the world level (Who uses it? Who pays?)
- individuality and community

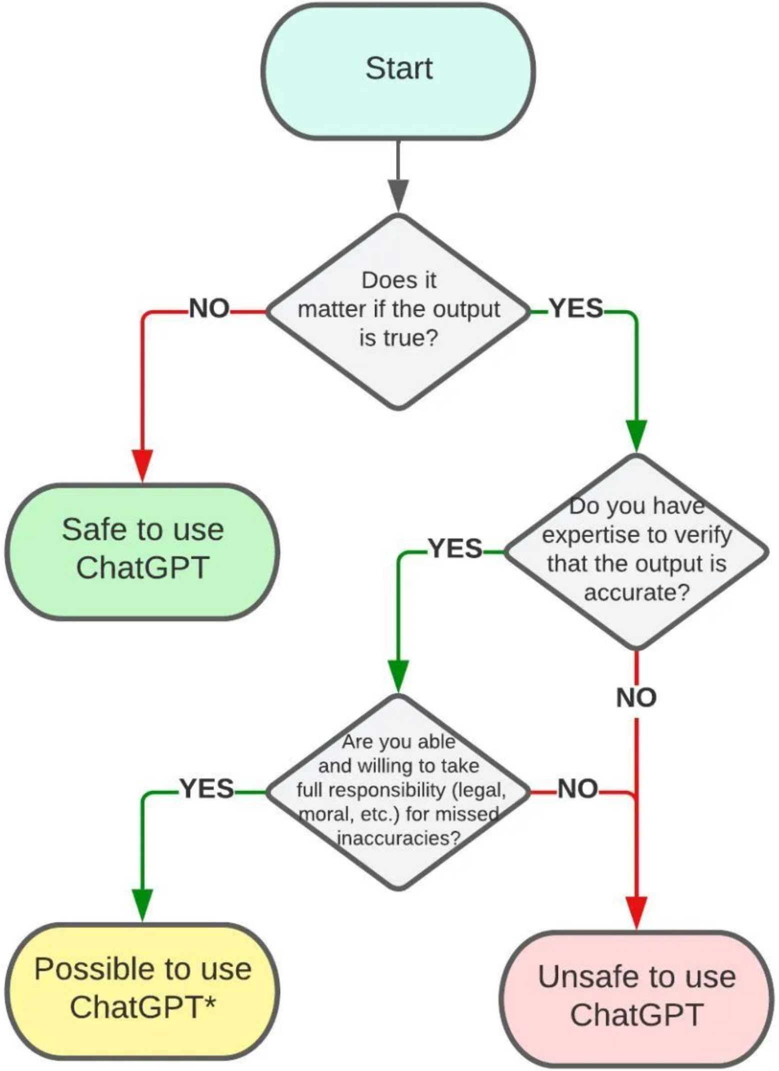
Conclusion

We must take into account these new technologies

but knowing the (current) limits

Is it safe to use ChatGPT for your task?

Aleksandr Tiulkanov | January 19, 2023



Next in Nantes: ANR Project MALADES

- Funded by the National Research Agency (ANR) - Start in October 2023
- Large Adaptable and Sovereign Language Models for the French Medical Field
- **Partners**: LS2N (Nantes University), LIS (Aix-Marseille University), LIA (Avignon University), CHU Nantes (Data Clinic)
- **Goals:**
  - Legal aspects of these models in the context of medical applications
  - Voice interaction in large language models
  - New use cases for generative models
  - Dynamic and sovereign models, with limited data access and computing power constraints

Nantes Université LS2N LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES AVIGNON UNIVERSITÉ LABORATOIRE INFORMATIQUE D'AVIGNON Aix*Marseille université Socialement engagée LIS LABORATOIRE D'INFORMATIQUE & SYSTÈMES anr LS2N

Questions?

Bibliography

- [LeCun98] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
- [Minsky69] Minsky, M., & Papert, S. (1969). An introduction to computational geometry. Cambridge tiass., HIT, 479(480), 104.
- [Parker95] Parker, D. (1995). Depression of synaptic connections between identified motor neurons in the locust. Journal of neurophysiology, 74(2), 529-538.
- [LeCun85] LeCun, Y. (1985). A learning scheme for asymmetric threshold networks. Proceedings of COGNITIVA, 85(537), 599-604.
- [McClelland86] McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. MIT Press, Cambridge MA, 3, 44.

Bibliography

- [Vaswani17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- [Szegedy15] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
- [Xie21] Xie, H., Qin, Z., Li, G. Y., & Juang, B. H. (2021). Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2663-2675.
- [Bahdanau14] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bibliography

- [Moll19] Moll Georges-Henri (2019). https://www.ibm.com/blogs/ibm-france/2019/09/26/apprentissage-automatique-et-biais/
- [Corbière18] Corbière Charles (2018). http://ladataenclair.fr/2018/06/25/quest-ce-que-la-fairness-en-intelligence-artificielle/
- [ProPublica19] ProPublica (2019). COMPAS Recidivism Risk Score Data and Analysis, https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
- [Caliskan17] Caliskan Aylin, Bryson Joanna J., Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 : 183-186.