FROM ARTIFICIAL INTELLIGENCE TO LANGUAGE MODELS

Archives Nationales, Grand Duché du Luxembourg
Jeudi 12 juin 2025

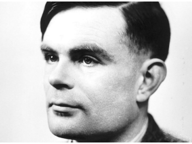
Vincent Guigue
vincent.guigue@agroparistech.fr
https://vguigue.github.io/tuto-LLM

MIA
PARIS-SACLAY
EKINOCS
INRAE AgroParisTech université PARIS-SACLAY

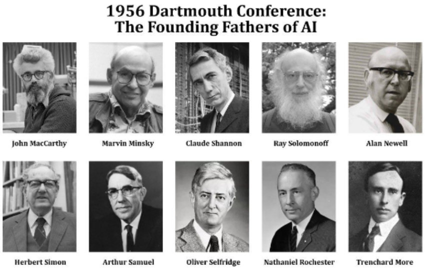
From AI to machine-learning

A quick historical tour of Artificial Intelligence

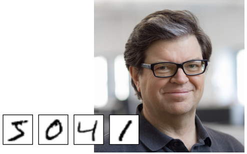
Birth of Computer Science... And of Artificial Intelligence



A. Turing



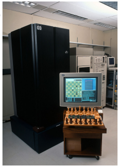
Dartmouth conference creation of IA



Y. Lecun Automated cheque reading

G. Hinton

E. Dickmanns : 1000km in autonomous vehicle

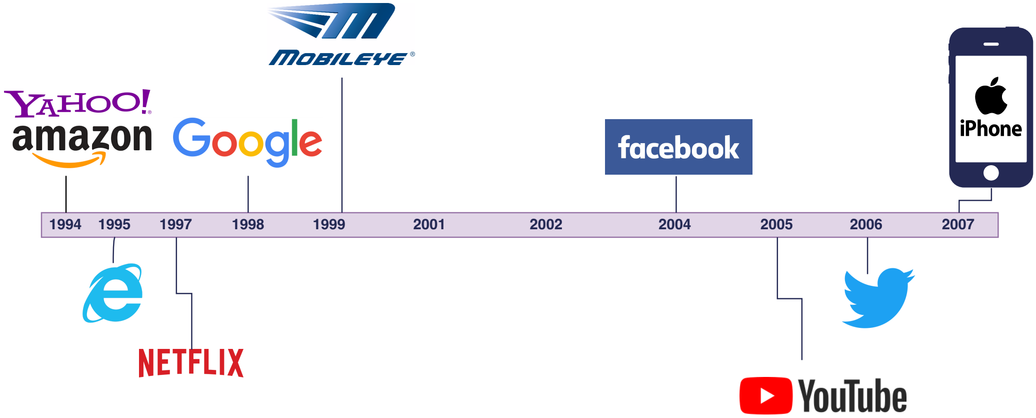


IBM Deeper Blue

- 1941 — First computer
- 1950 — Dynamic programming
- 1956 — Dartmouth conference creation of IA
- 1958 — LISP Language Perceptron
- 1963 — DARPA
- 1965
- 1972 — PROLOG
- 1979 — 1st Neural Network
- 1986 — Backpropagation
- 1987 — DARPA MUC
- 1990 — Y. Lecun Automated cheque reading; DARPA TREC
- 1992 — SVM
- 1993 — C4.5 Decision Tree
- 1994 — E. Dickmanns : 1000km in autonomous vehicle
- 1997 — IBM Deeper Blue

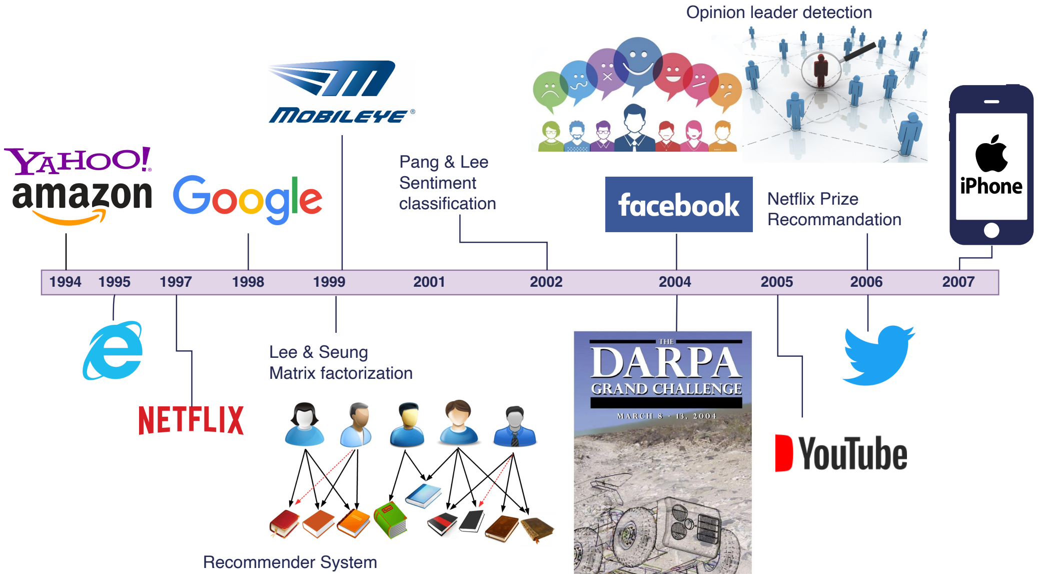
A quick historical tour of Artificial Intelligence

Emergence (or Reinvention) of GAFAM/GAMMA



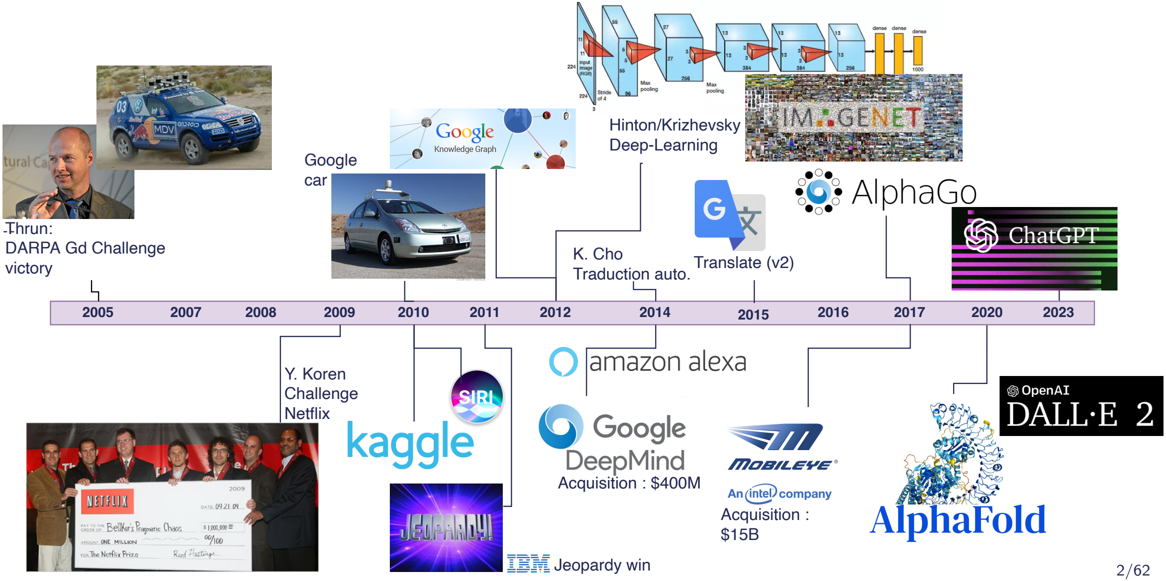
A quick historical tour of Artificial Intelligence

Emergence (or Reinvention) of GAFAM/GAMMA

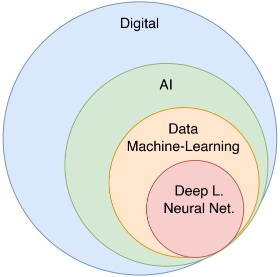


A quick historical tour of Artificial Intelligence

Formation of a Wave of Artificial Intelligence



Artificial Intelligence & Machine Learning



Input (**X**)		Output (**Y**)	Application
email	→	spam? (0/1)	spam filtering
audio	→	text transcript	speech recognition
English	→	Chinese	machine translation
ad, user info	→	click? (0/1)	online advertising
image, radar info	→	position of other cars	self-driving car
image of phone	→	defect? (0/1)	visual inspection

AI: computer programs that engage in tasks which, for now, are more satisfactorily performed by humans because they require high-level mental processes.

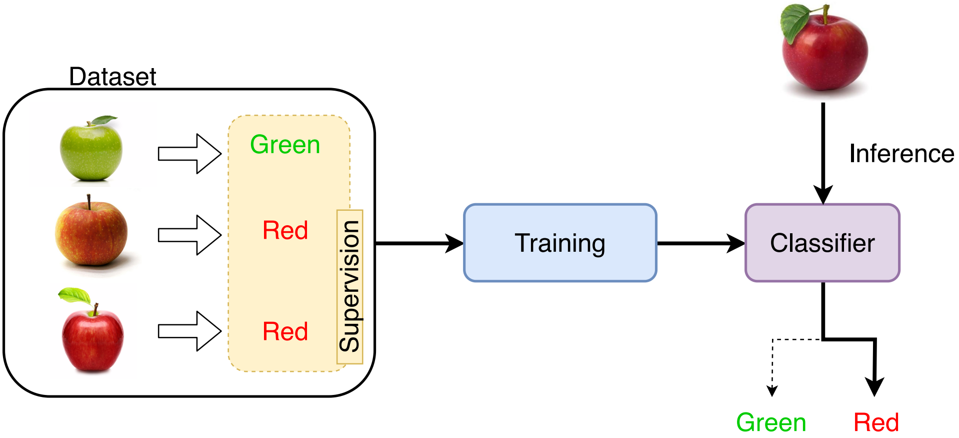
Marvin Lee Minsky, 1956

N-AI (Narrow Artificial Intelligence), dedicated to a single task

$\neq$ **G-AI (General AI)**, which replaces humans in complex systems.

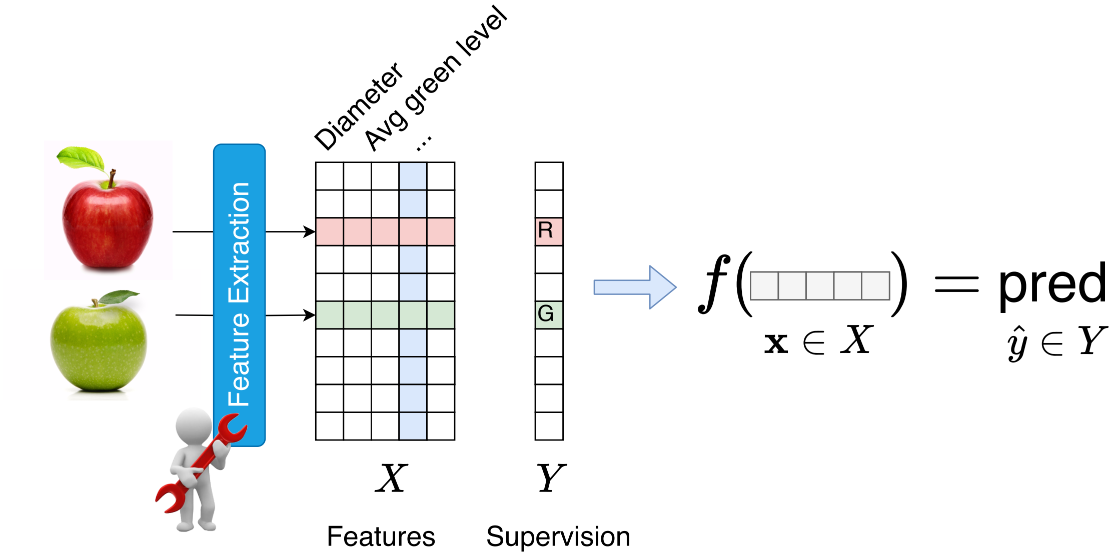
Andrew Ng, 2015

Supervised Processing Chain & Models

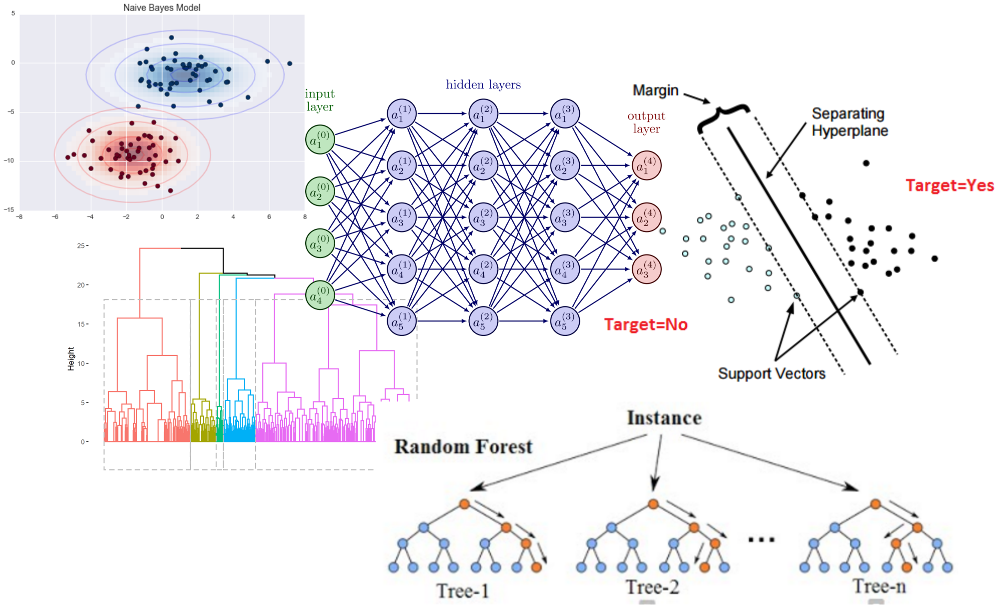


- Promise = building a model *solely* from observations

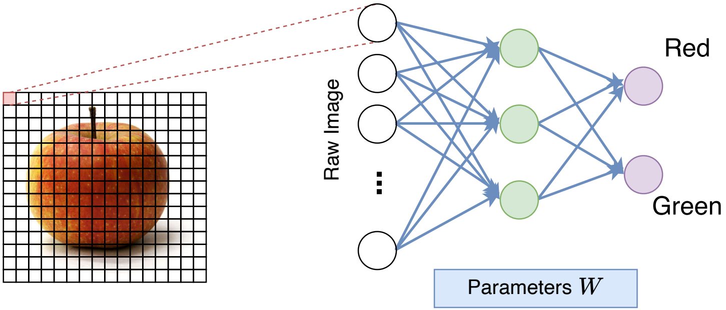
Supervised Processing Chain & Models



Supervised Processing Chain & Models

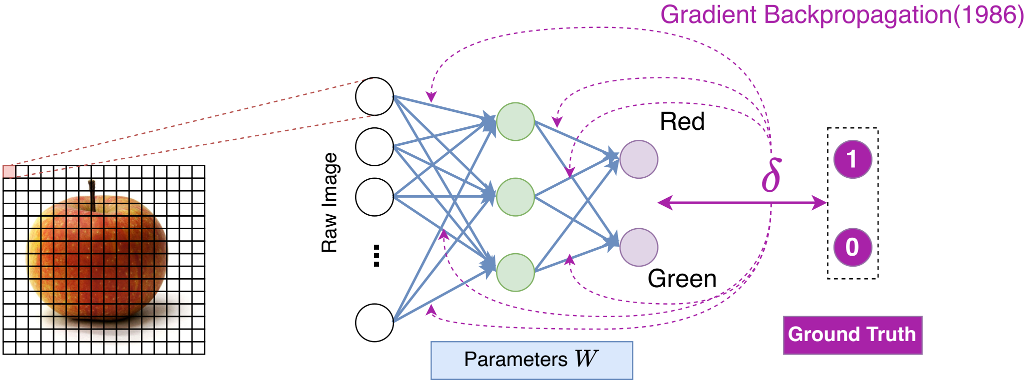


Supervised Processing Chain & Models



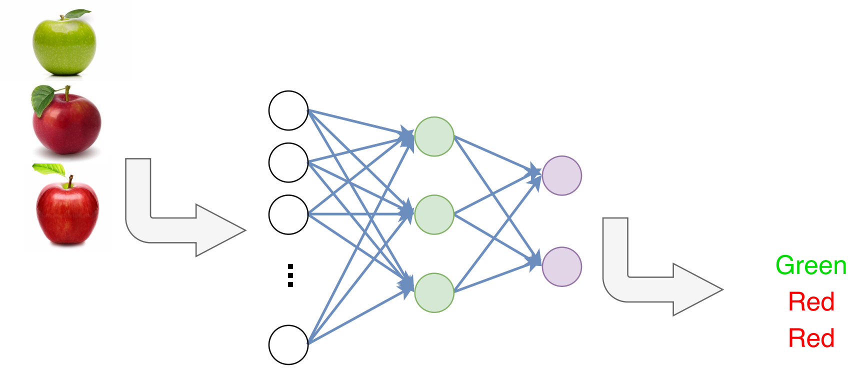
- Random initialization... And random decision-making (at first!)

Supervised Processing Chain & Models



- Updating the weights
- Epsilon-sized steps, many iterations over the data

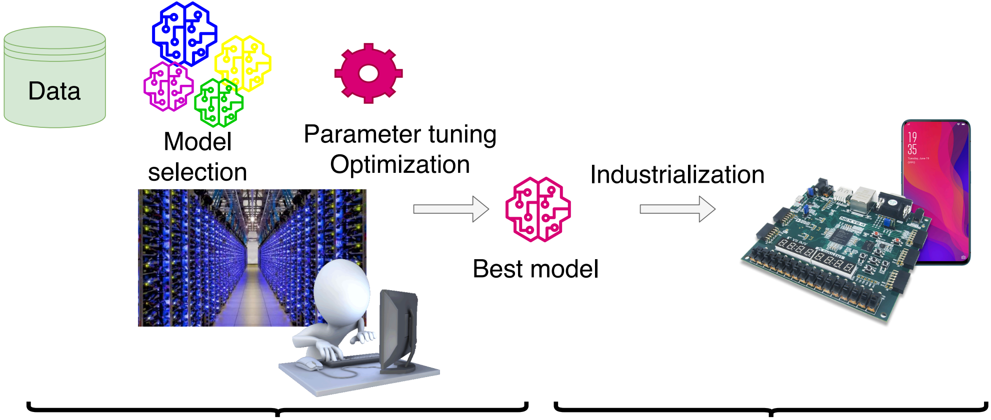
Supervised Processing Chain & Models



- **Training** is slow and costly
- **Inference** is (much) faster

Supervised Processing Chain & Models

Différentes étapes en machine-learning

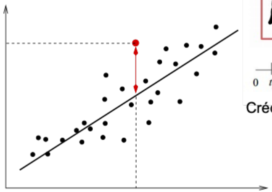


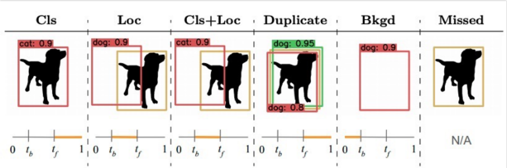
Model Training = **Intensive** Computing   Model exploitation = **limited** Computing

Measuring Performance

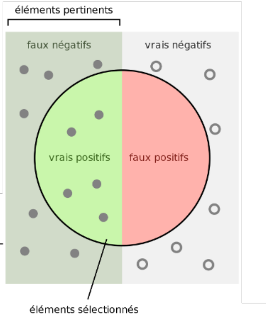
Estimating performance (in generalization)...

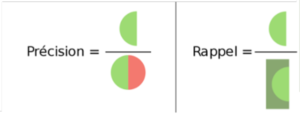
Is just as important as training the model itself!

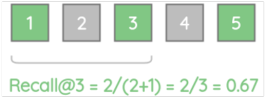


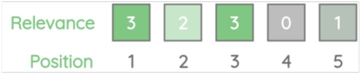


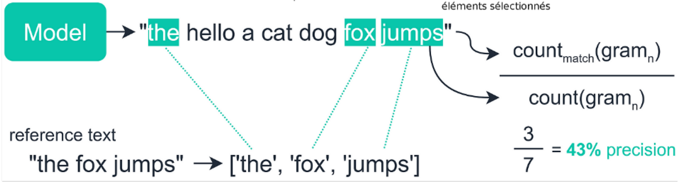
Crédit: https://github.com/phalanx-hk/eccv2020_paperlist/issues/5





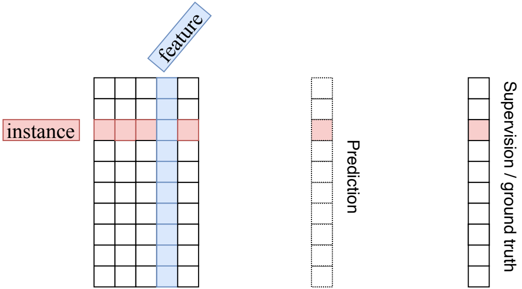






Measuring Performance

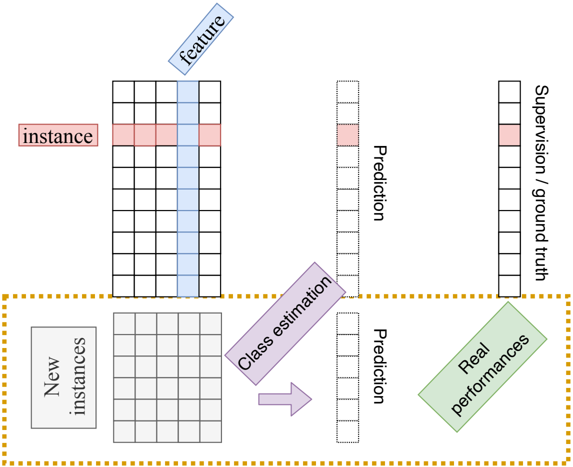
Estimating performance (in generalization)...
Is just as important as training the model itself!



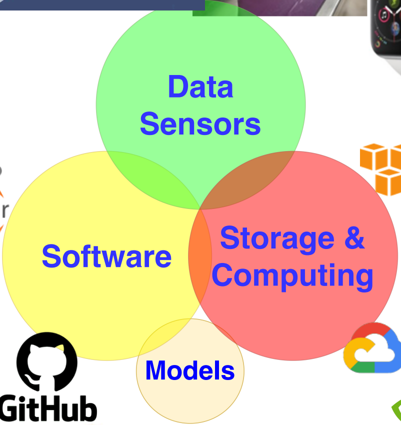
Measuring Performance

Estimating performance (in generalization)...

Is just as important as training the model itself!



Ingredients of Artificial Intelligence

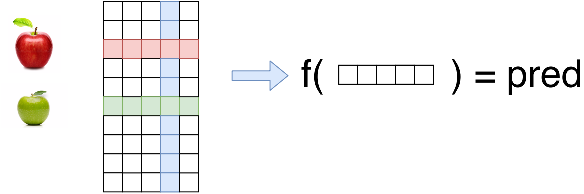


Deep-Learning & NLP*

[* Natural Language Processing]

From tabular data to text

- → Tabular data
  - → Fixed dimension
  - → Continuous values



- → Textual data
  - → Variable length
  - → Discrete values

this new iPhone, what a marvel

An iPhone? What a scam!

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

Rule-based Systems:

* $\longrightarrow$ {like, love, appreciate} $\rightarrow$ * $\rightarrow$ #product

* $\rightarrow$ {didn't, not, doesn't, don't} $\rightarrow$ {like, love, appreciate} $\rightarrow$ * $\rightarrow$ #product

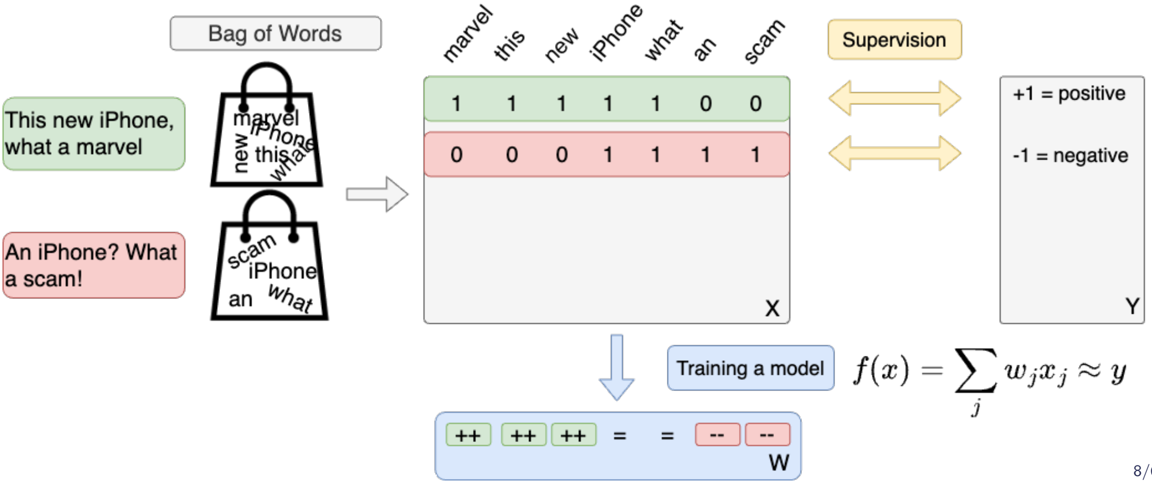
* $\longrightarrow$ {hate, loathe, detest} $\rightarrow$ * $\rightarrow$ #product

- Requires expert knowledge
- Rule extraction $\Leftrightarrow$ very clean data
- Very high precision
- Low recall
- Interpretable system

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Machine Learning [1990-2015]



$$f(x) = \sum_j w_j x_j \approx y$$

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

- Requires expert knowledge
- Rule extraction ⇔ very clean data

+ Interpretable system
+ Very high precision
– Low recall

Machine Learning [1990-2015]

- Little expert knowledge needed
- Statistical extraction ⇔ robust to noisy data

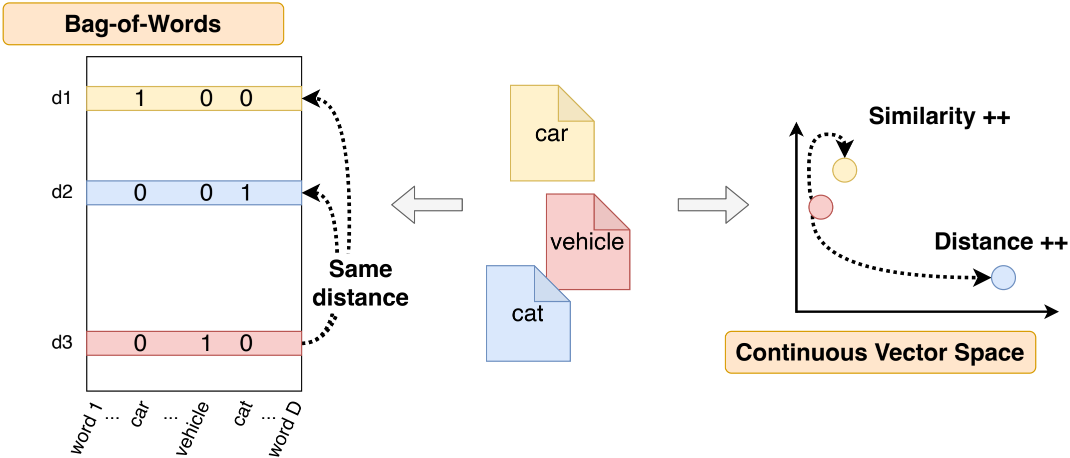
≈ Less interpretable system
– Lower precision
+ Better recall

Precision = criterion for acceptance by industry

→ Link to metrics

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

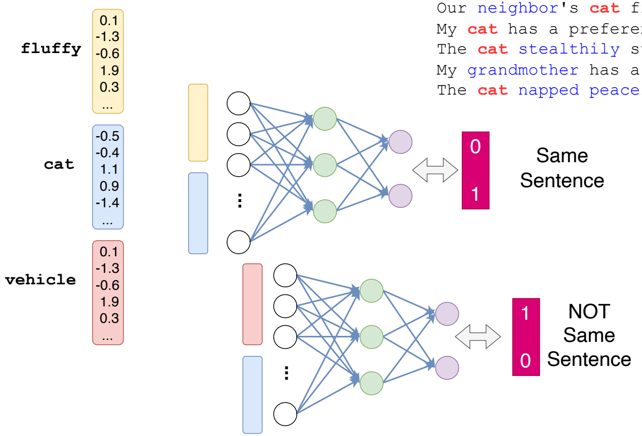


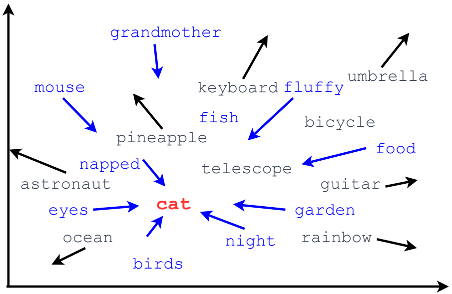
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

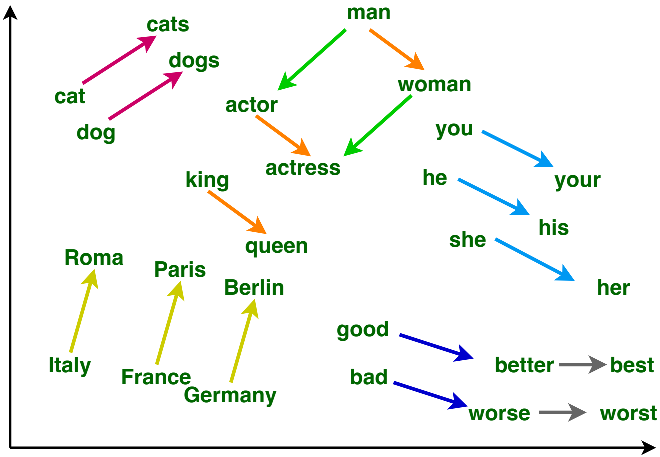
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
The cat napped peacefully in the warm sunlight.





Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]



- Semantic Space:

  similar meanings

  $\Leftrightarrow$

  close positions

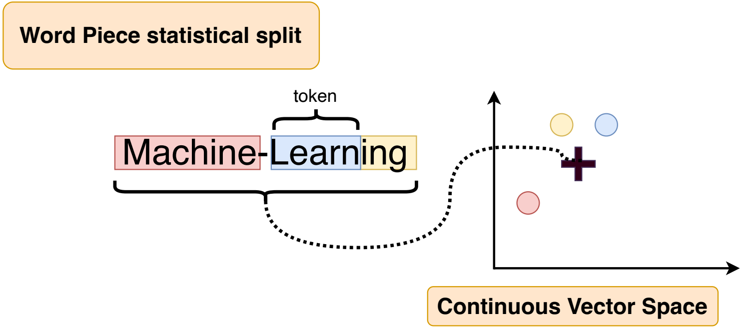
- Structured Space: grammatical regularities, basic knowledge, ...

Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

From Words to Tokens

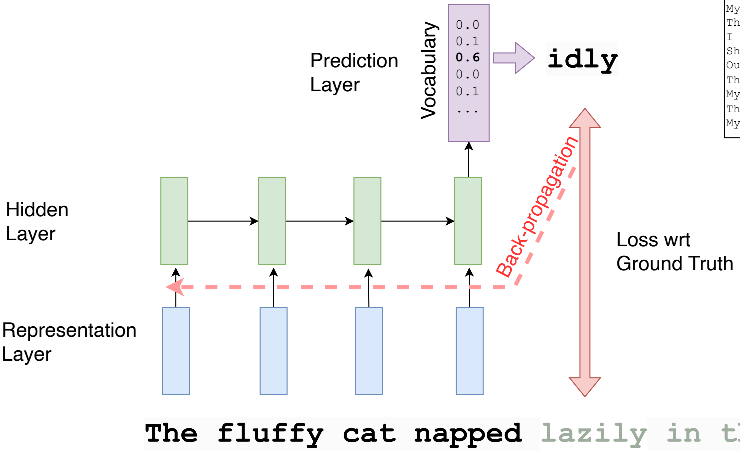


- Representation of unknown words
- Adaptation to technical domains
- Resistance to spelling errors

Enriching word vectors with subword information. Bojanowski et al. TACL 2017.

Aggregating word representations: towards generative AI

- Generation & Representation
- New way of learning word positions

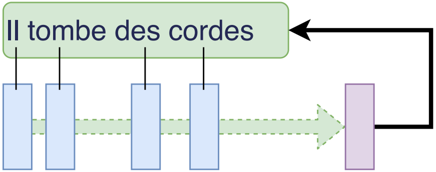



```
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
The playful cat swatted at the dangling string with its paw.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
```

Corpus

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

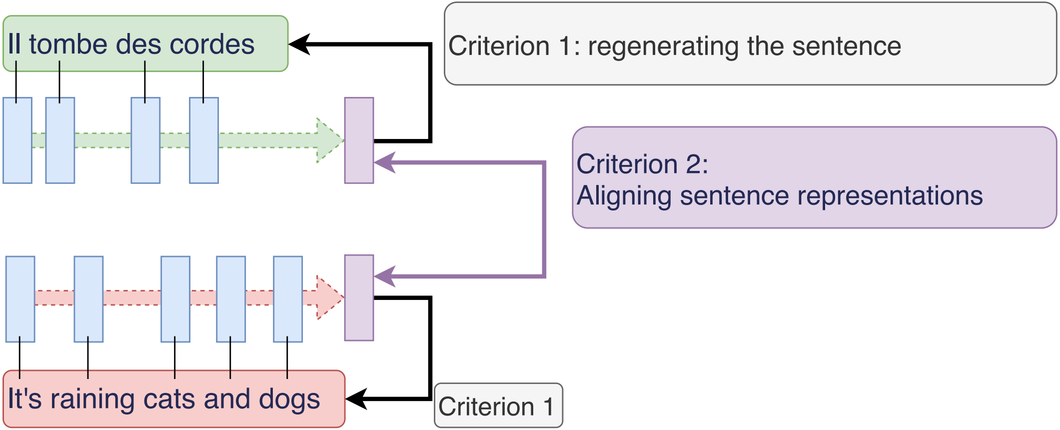
Use-Case: Machine Translation



Criterion 1: regenerating the sentence

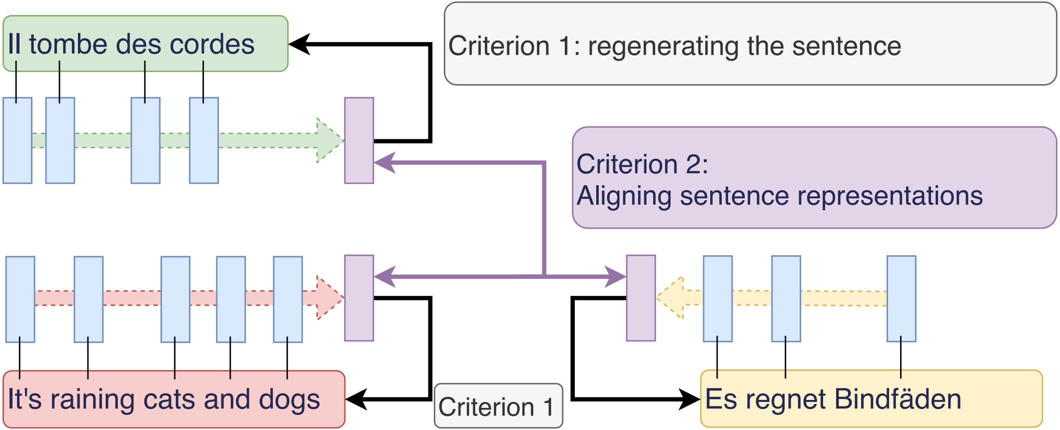
Beyond word-for-word translation, multilingual representation of sentences

Use-Case: Machine Translation



Beyond word-for-word translation, multilingual representation of sentences

Use-Case: Machine Translation

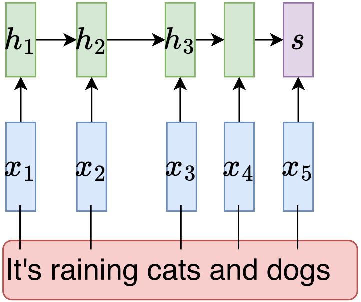


Beyond word-for-word translation, multilingual representation of sentences

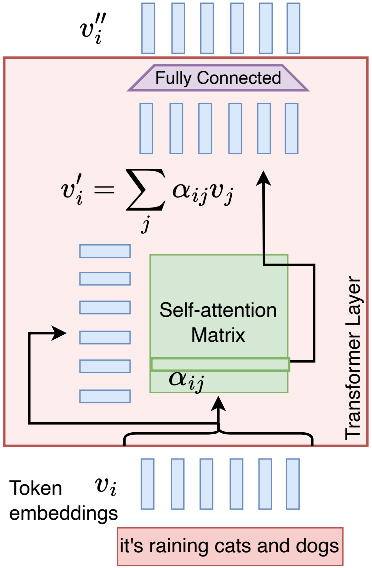
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:



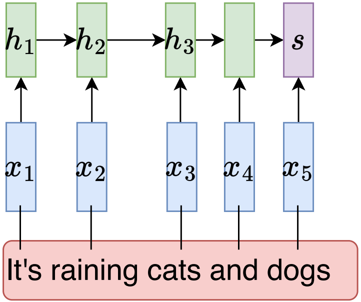
Attention is all you need, Vaswani et al. NeurIPS 2017

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

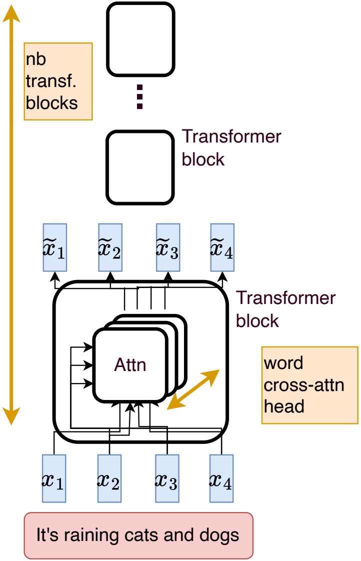
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

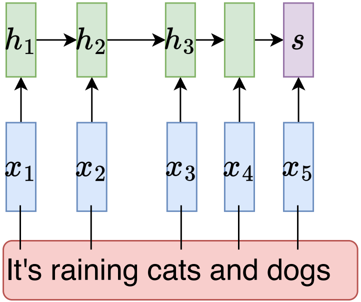


Attention is all you need, Vaswani et al. NeurIPS 2017

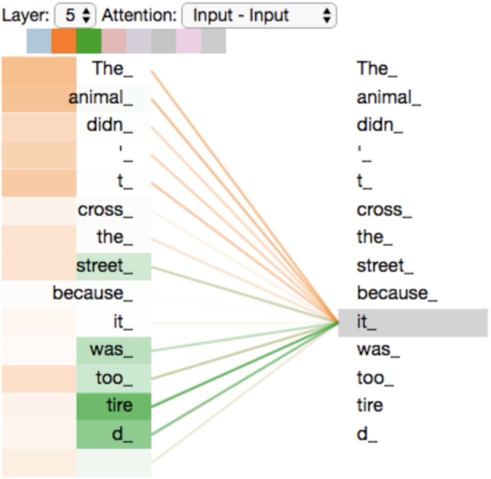
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

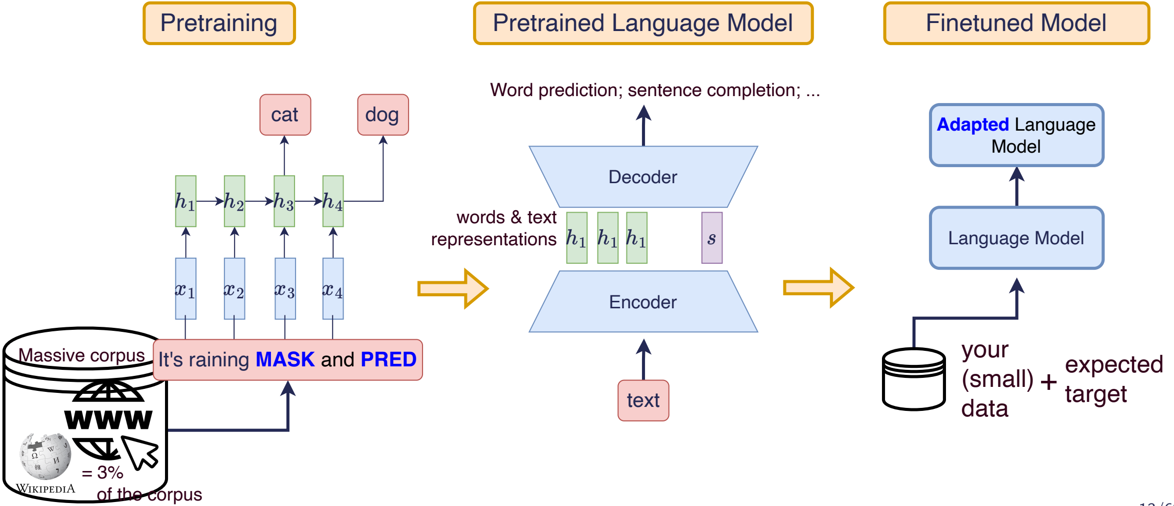


Attention is all you need, Vaswani et al. NeurIPS 2017

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

A new developpement paradigm since 2015

- Huge dataset + huge archi. ⇒ unreasonable training cost
- Pre-trained architecture + 0-shot / finetuning



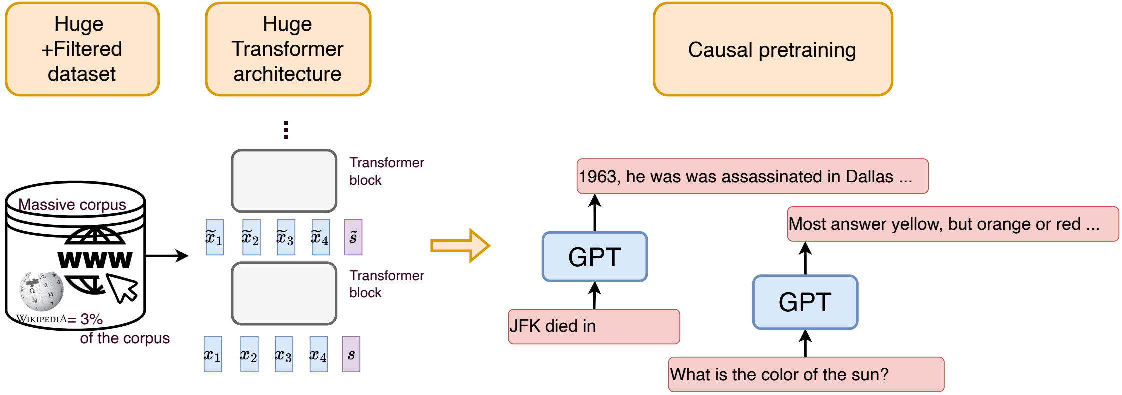
chatGPT

November 30, 2022

1 million users in 5 days
100 million by the end of January 2023
1.16 billion by March 2023

The Ingredients of chatGPT

0. Transformer + massive data (GPT)



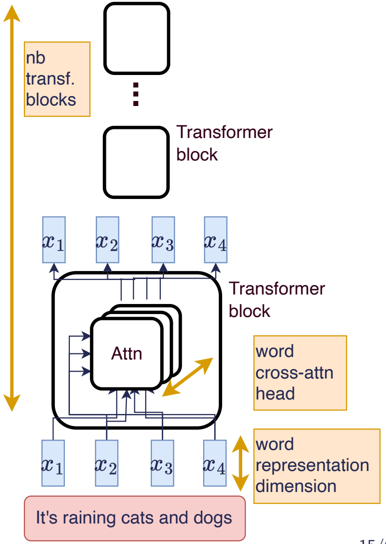
- Grammatical skills: singular/plural agreement, tense concordance
- Knowledges

Language Models are Few-Shot Learners, Brown et al. 2020

The Ingredients of chatGPT

1. More is better! (GPT)

- more input words [500 ⇒ 2k, 32k, 100k]
- more dimensions in the word space [500-2k ⇒ 12k]
- more attention heads [12 ⇒ 96]
- more blocks/layers [5-12 ⇒ 96]



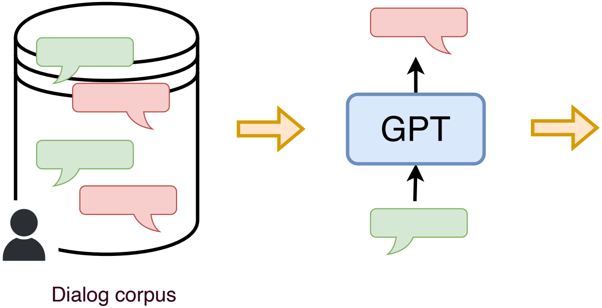
175 Billion parameters... What does it mean?

- $1.75 \cdot 10^{11} \Rightarrow$ 300 GB + 100 GB (data storage for inference) ≈ 400GB
- NVidia A100 GPU = 80GB of memory (=20k€)
- Cost for (1) training: 4.6 Million €

The Ingredients of chatGPT

2. Dialogue Tracking

Specific training

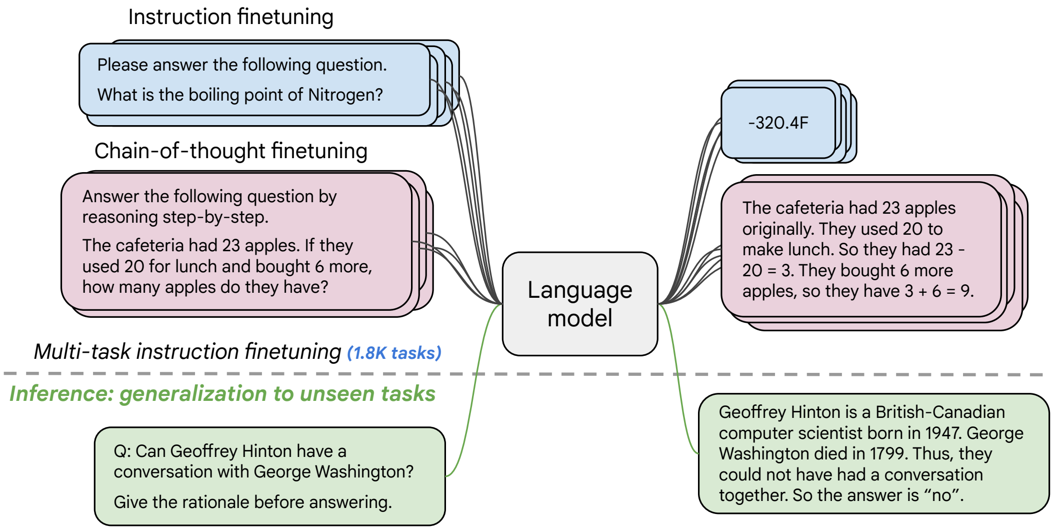


Dialog follow-up
Coreference resolution
Way of speaking

- **Very clean** data Data generated/validated/ranked by humans

The Ingredients of chatGPT

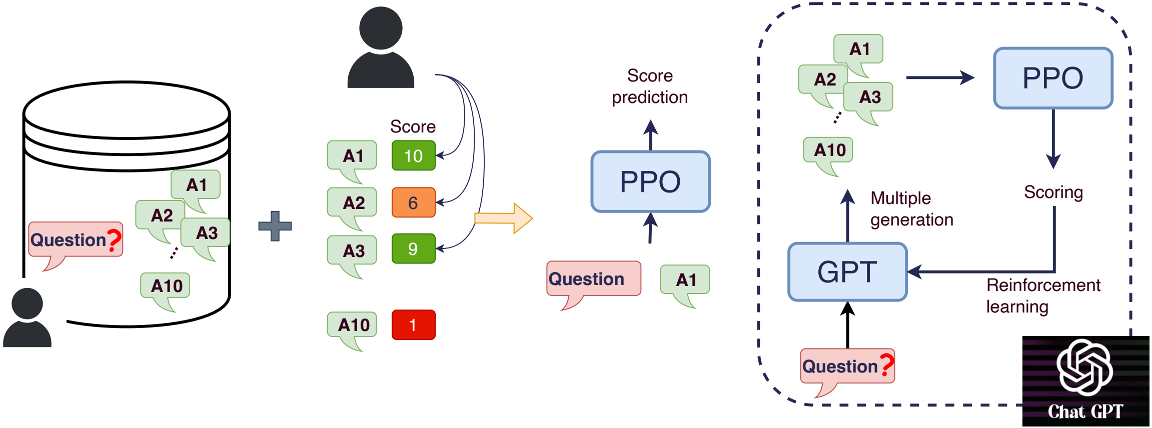
3. Fine-tuning on different (±) complex reasoning tasks



Scaling Instruction-Finetuned Language Models, Chung et al., JMLR 2024

The Ingredients of chatGPT

4. Instructions + answer ranking

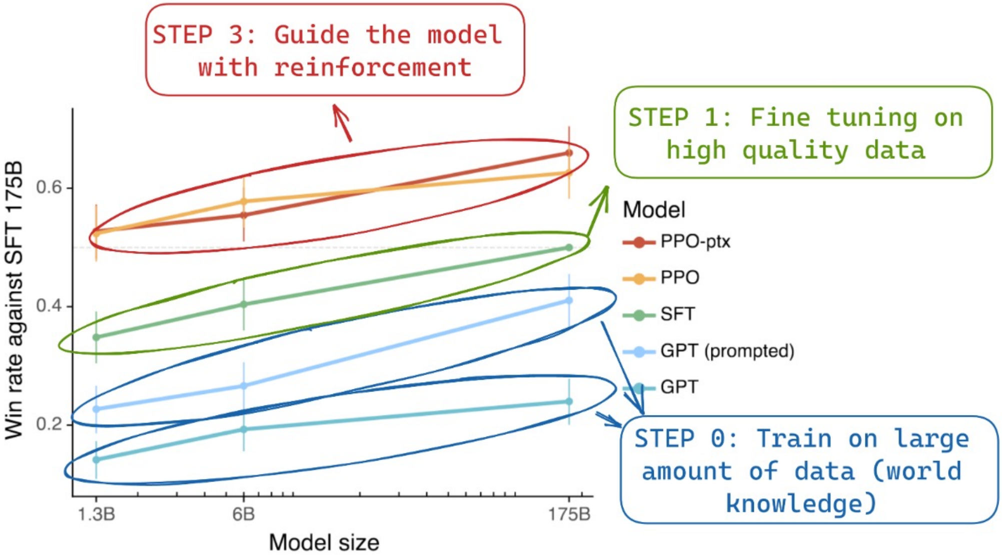


- Database created by humans
- Response improvement
- ... Also a way to avoid critical topics = censorship

Training language models to follow instructions with human feedback, Ouyang et al., 2022

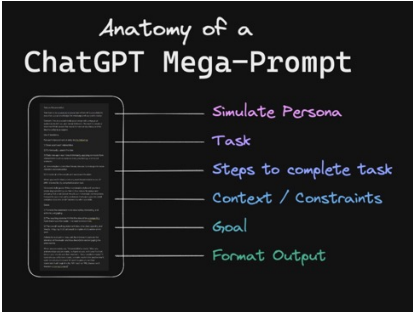
Steps & Performance

Massive data ⇒ HQ data (dialogue) ⇒ Tasks ⇒ RLHF



Usage of chatGPT & Prompting

- Asking chatGPT = skill to acquire ⇒ *prompting*
  - Asking a question well: ... *in detail*, ... *step by step*
  - Specify number of elements e.g. : *3 qualities for ...*
  - Provide context : *cell* for a biologist / legal assistant
- Don't stop at the first question
  - Detail specific points
  - Redirect the research
  - Dialogue
- Rephrasing
  - Explain like I'm 5, like a scientific article, bro style, ...
  - Summarize, extend
  - Add mistakes (!)

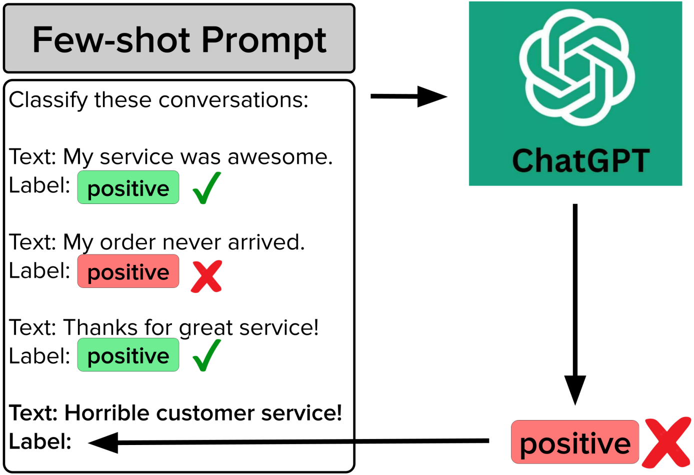


https://chatgptprompts.guru/what-makes-a-good-chatgpt-prompt/

⇒ Need for **practice** [1 to 2 hours], discuss with colleagues

Towards *few-shot learning*

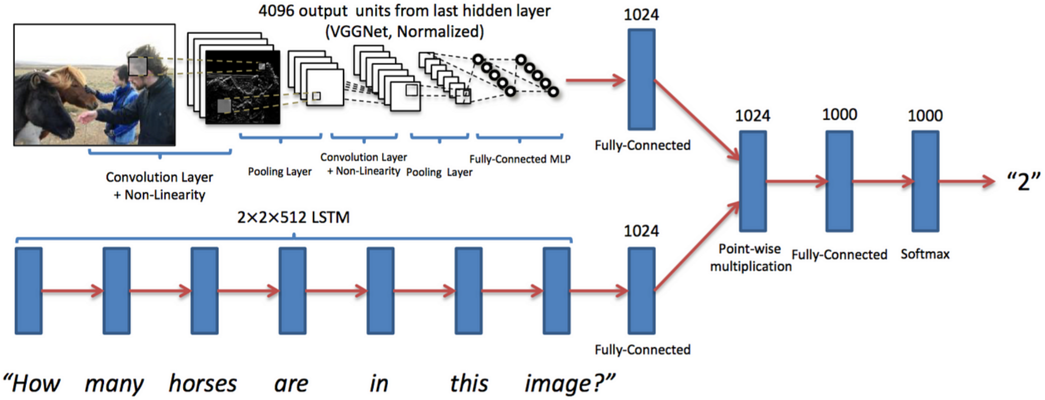
- Learning without modifying the model = examples in the prompt



GPT4 & Multimodality

Merging information from text & image. **Learning** to exploit information jointly

The example of VQA: visual question answering

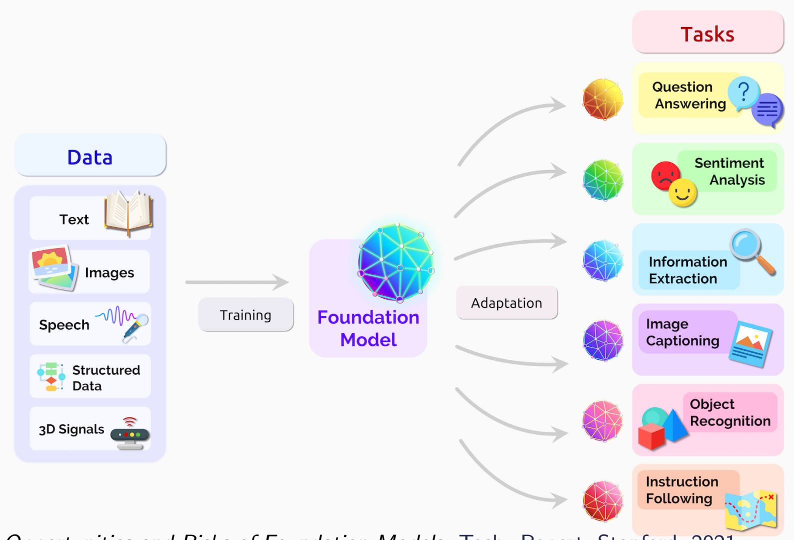


⇒ Backpropagate the error ⇒ modify word representations + image analysis

VQA: Visual Question Answering, arXiv, 2016 , A. Agrawal et al.

Towards Larger Foundation Models?

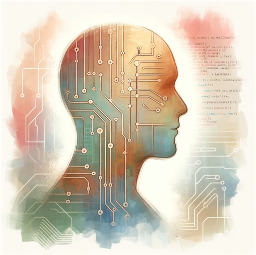
- Let the modalities enrich each other

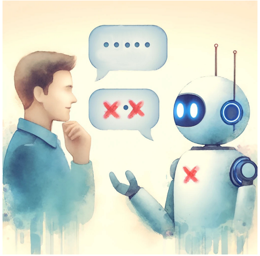


On the Opportunities and Risks of Foundation Models, Tech. Report, Stanford, 2021
Bommasani et al.

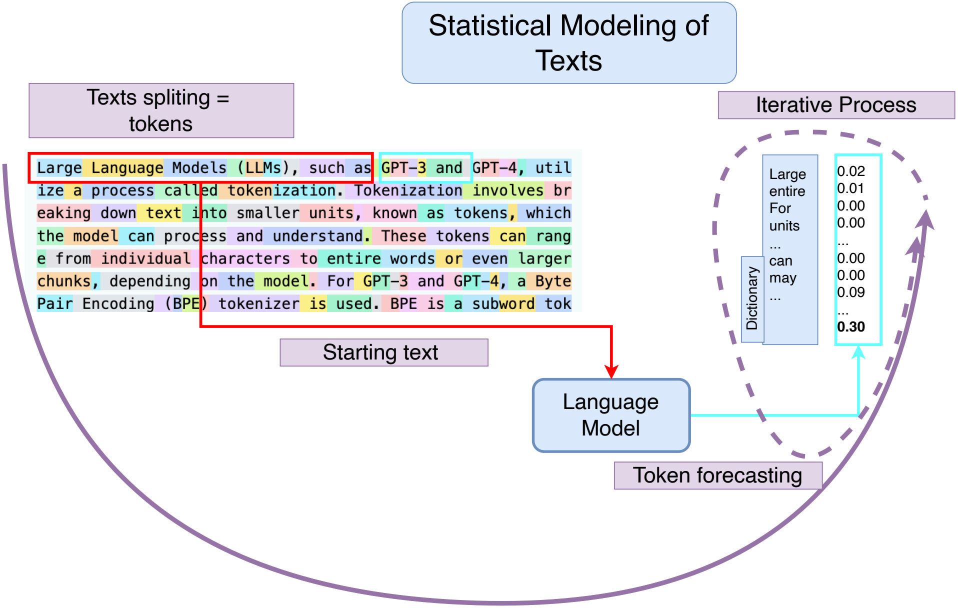
Why So Much Controversy?

- New tool [December 2022]
- + Unprecedented adoption speed [1M users in 5 days]
- Strengths and weaknesses... Poorly understood by users
    - Significant productivity gains
    - Surprising / sometimes absurd uses
    - Bias / dangerous uses / risks
- Misinterpreted feedback
    - Anthropomorphization of the algorithm and its errors
- Prohibitive cost: what economic, ecological, and societal model?



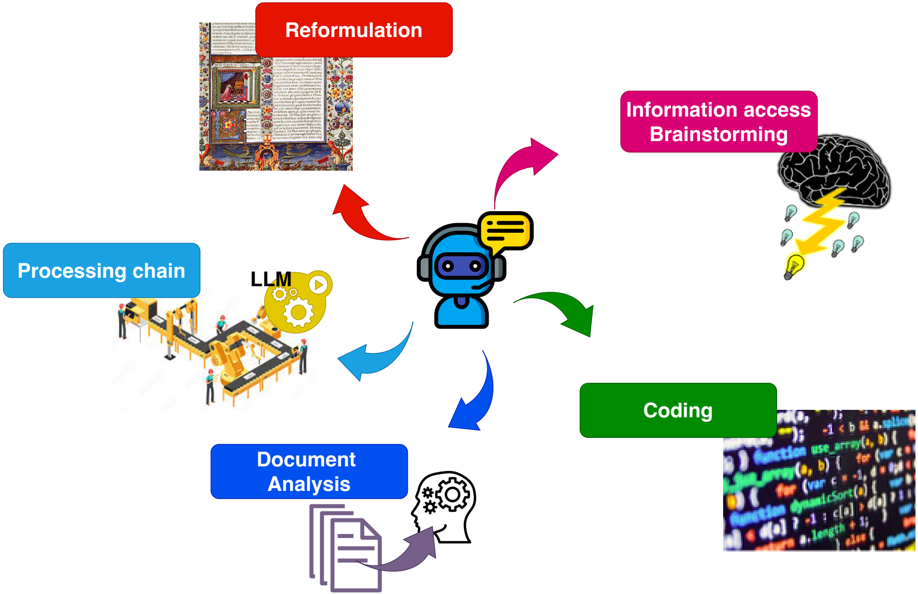


At the end of the day



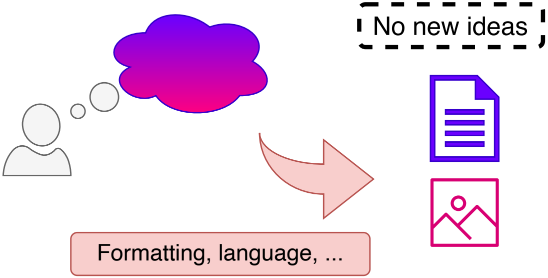
Large Language Models Uses

Key uses in 5 pictures



(1) Formatting information

A fantastic tool for **formatting**

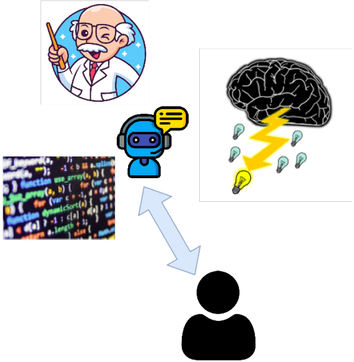


- Personal assistant
  - Standard letters, recommendation letters, cover letters, termination letters
  - Translations
- Meeting reports
  - Formatting notes
- Writing scientific articles
  - Writing ideas, in French, in English
- Document analysis
  - Information extraction, question-answering, ...

$\Rightarrow$ No new information, just writting, cleaning up, ...

(2) Brainstorming / Course Planning / Statistics Review

- **Find** inspiration [writer's block syndrome]
- **Organize** ideas quickly
- **Avoid omissions** / increase confidency
- **Search** in a targeted way, adapted to one's needs



⇒ Impressive answers, sometimes incomplete or partially incorrect... But often useful

3 reference articles on the use of transformers in recommendation systems
What is the purpose of the log-normal Poisson law?
Propose 10 sections for a course on Transformers in AI

- In which areas are LLMs reliable?
- What are the risks for primary information sources?
- What societal risks for information?

(3) Coding: Different Tools, Different Levels

- Providing solutions to exercises
- Learning to code or getting back into it
  - New languages, new approaches (ML?)
  - Benefit from explanations...

But how to handle mistakes?

- Help with a library [*getting started*]
- Faster coding

GitHub Copilot

- What about copyrights?
  - What impact on future code processing?
- How to adapt teaching methods?
- How many calls are needed for code completion?

What about the carbon footprint?

- What is the risk of error propagation?

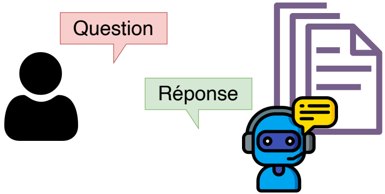
sentiment.ts write_sql.go parse_expenses.py addresses.rb

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date,
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2016-01-02 -34.01 USD
        2016-01-03 2.59 DKK
        2016-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
```

(4) Document Analysis

- Summarizing documents / articles
- Dialoguing with a document database
- Assistance in writing reviews
- FAQs, internal support services within companies
- Technology watch
- Generating quizzes from lecture notes

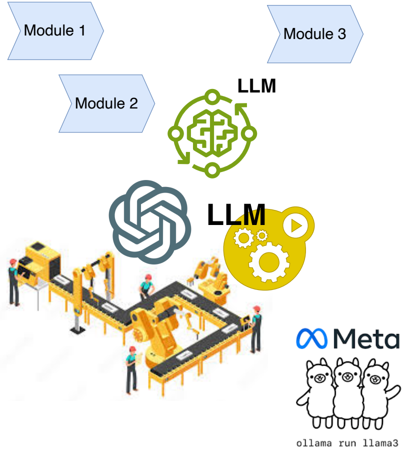


- Will articles still be read in the future?
  - Should we make our articles NotebookLM-proof?
- How to save time while remaining honest and ethical?

(5) LLM in a Production Pipeline / Agentic AI

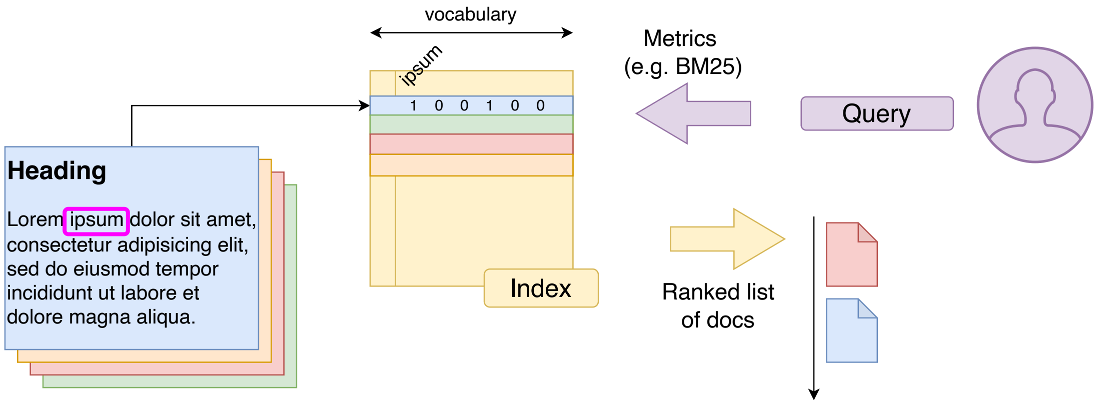
- Run LLM locally
- Extract knowledge
- Sort documents / generate summaries
- Generate examples to train a model
  [Teacher/student - distillation]
- Generate variants of examples ↗↗ increase dataset size
  [Data augmentation]

⇒ Integrate the LLM into a processing pipeline
= little/less supervision = **Agentic AI**

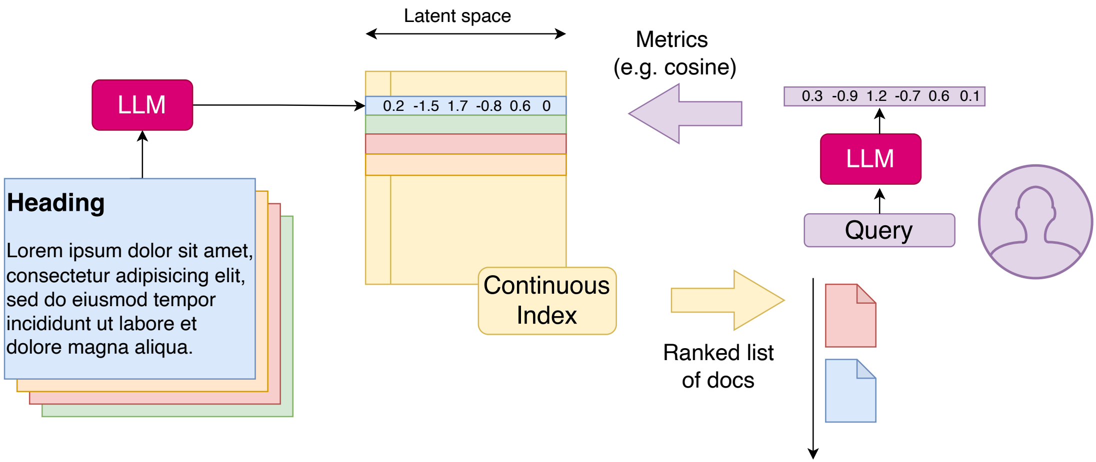


- Can I train models on generated data?
- How much does it cost? (\$ + CO_2) Need for GPUs?
- How good are open-weight models?

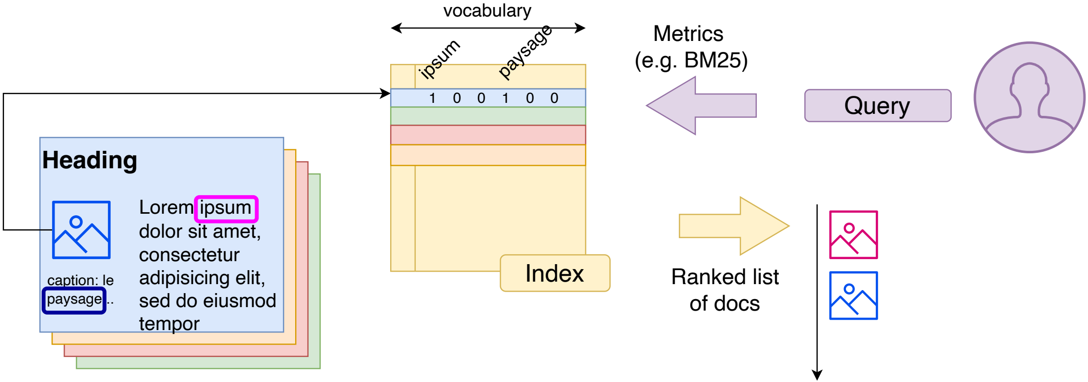
LLM *vs* Information Retrieval



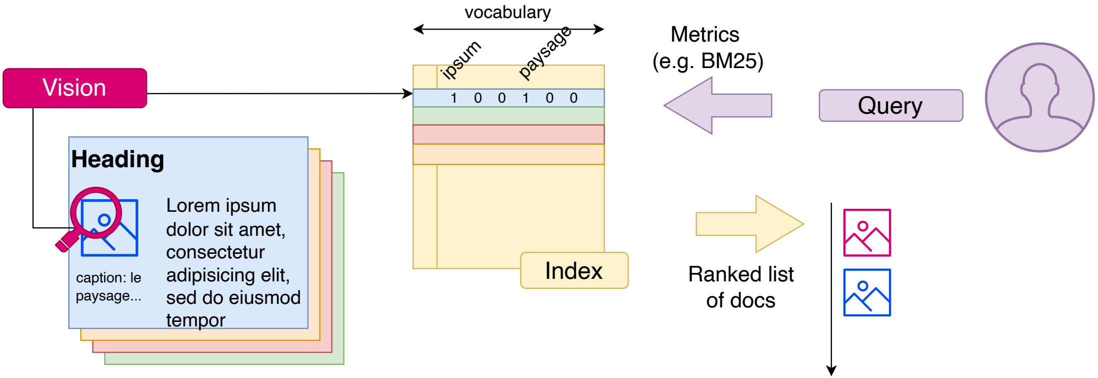
LLM *vs* Information Retrieval



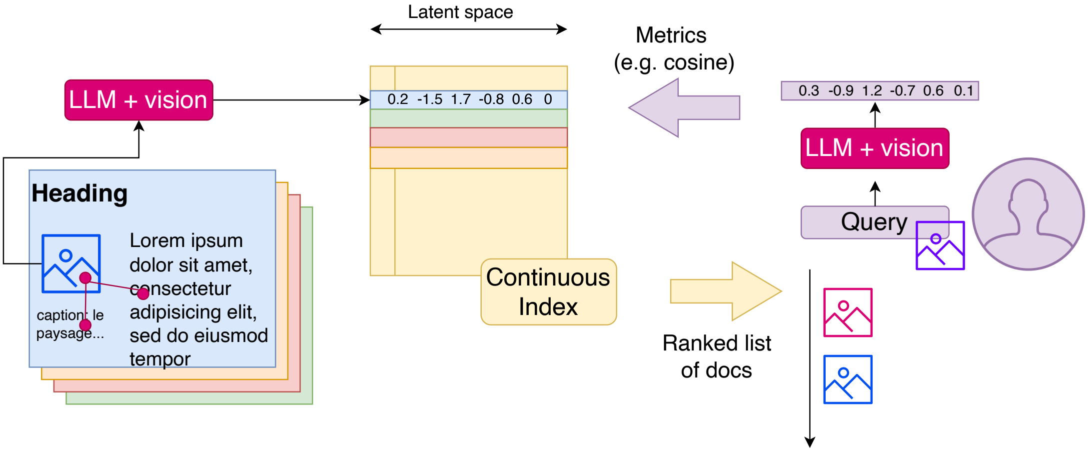
LLM *vs* Information Retrieval



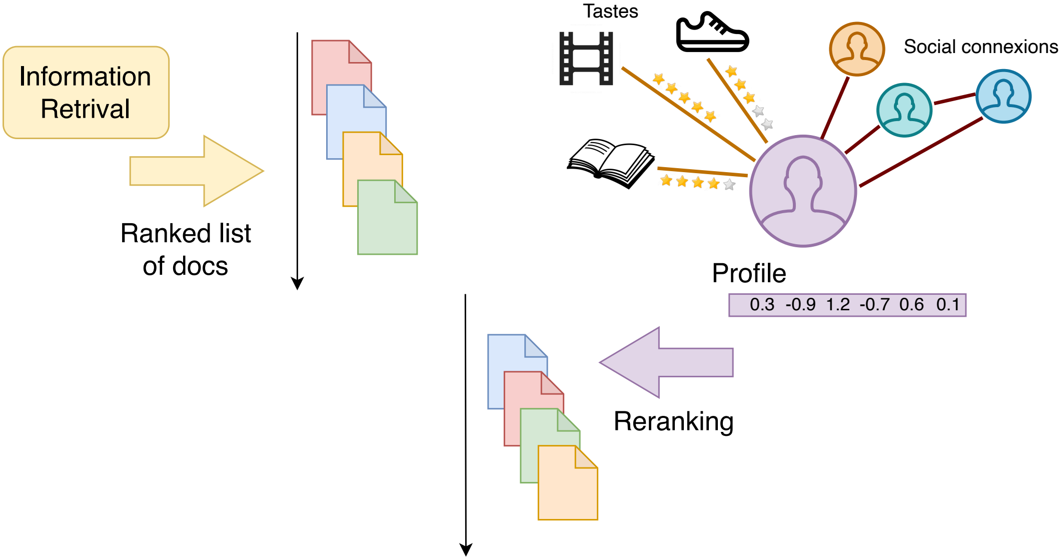
LLM *vs* Information Retrieval



LLM *vs* Information Retrieval

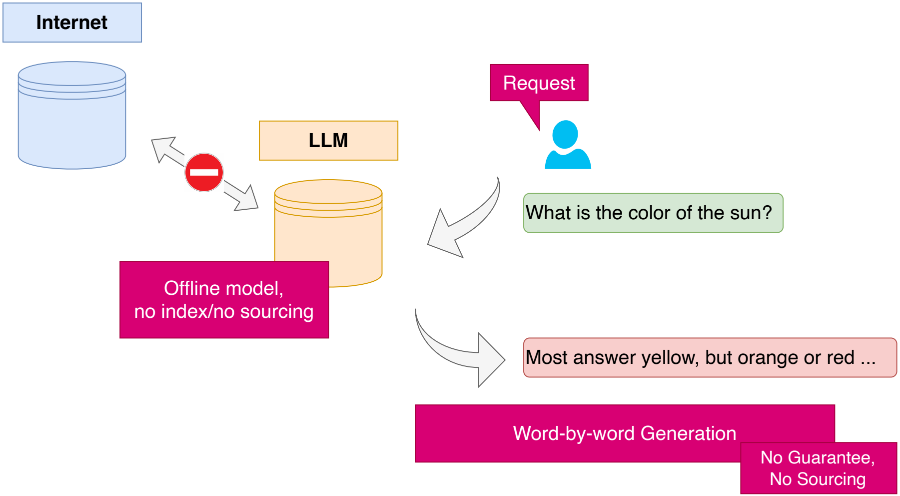


LLM *vs* Information Retrieval

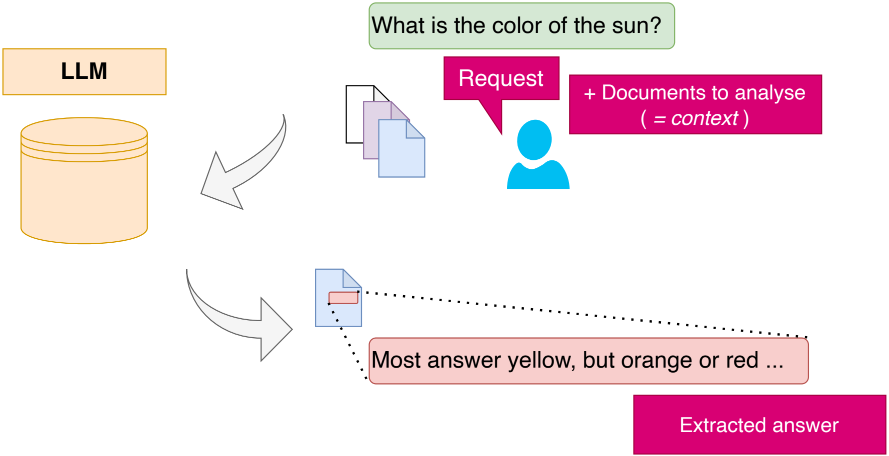


LLMs $\Rightarrow$ RAG : parametric memory *vs* Info. Extraction

- Asking for information from ChatGPT... A surprising use!
- But is it reasonnable? [Real Open Question (!)]

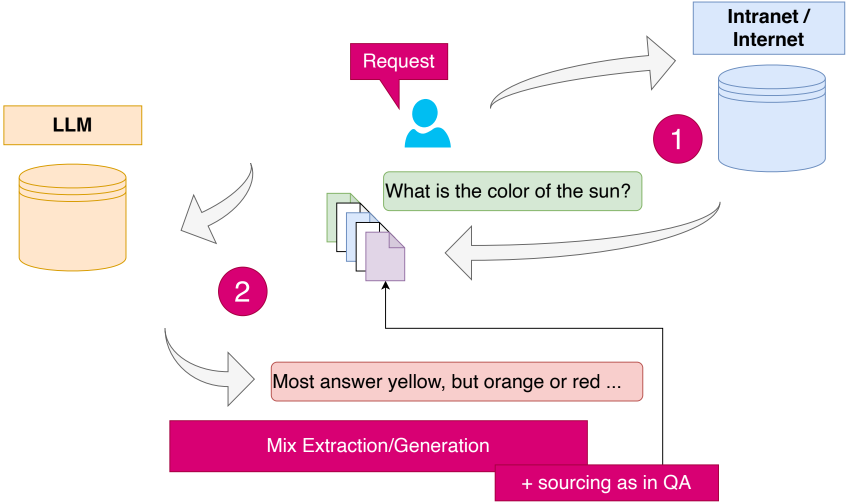


LLMs ⇒ RAG : parametric memory *vs* Info. Extraction



- Web query + analysis, automatic summary, rephrasing, meeting reports...
- (Current) limit on input size (2k then 32k tokens)
- = *pre chatGPT use of LLM for question answering*

LLMs ⇒ RAG : parametric memory *vs* Info. Extraction



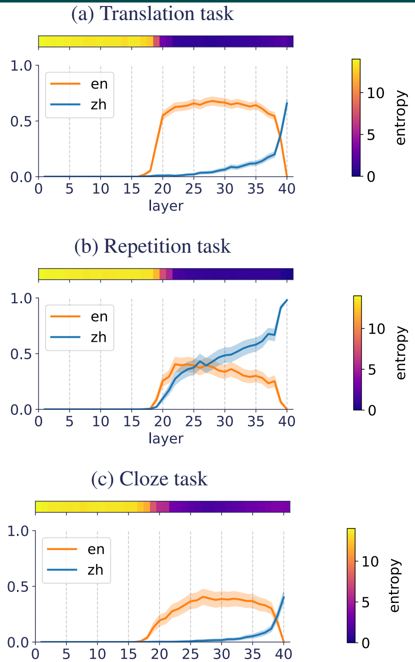
- RAG: Retrieval Augmented Generation
- (Current) limit on input size (2k then 32k tokens)

Language Handling

- Language models are (mostly) multilingual:

⇒ Think in the language you are most comfortable with
⇒ Ask for answers in the target language

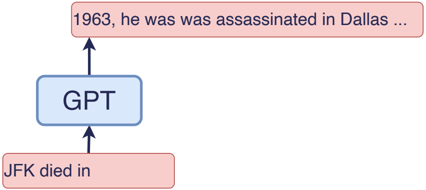
[Wendler et al. 2024] Do Llamas Work in English? On the Latent Language of Multilingual Transformers



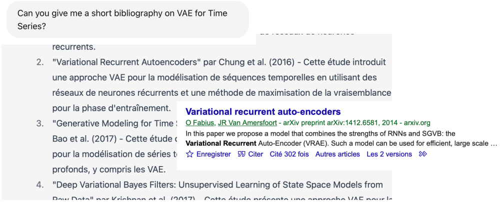
Machine Learning Technical Limits

chatGPT and the relationship with truth

1. **Likelyhood** = grammar, agreement, tense concordance, logical sequences...
   ⇒ Repeated knowledge
2. Predict the most **plausible** word...
   ⇒ produces **hallucinations**
3. **Offline** functioning
4. chatGPT $\neq$ **knowledge graphs**
5. Brilliant answers...
   And silly mistakes!
   + we cannot predict the errors



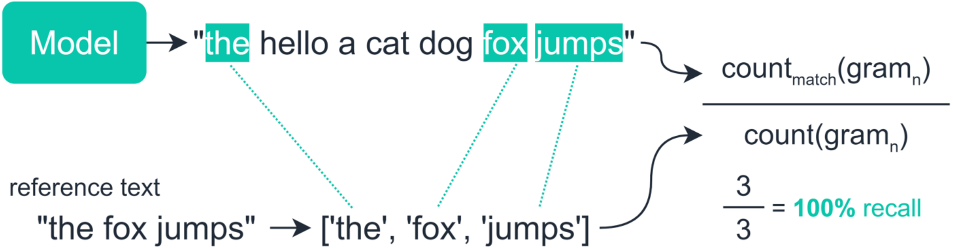
Example: producing a bibliography



Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?

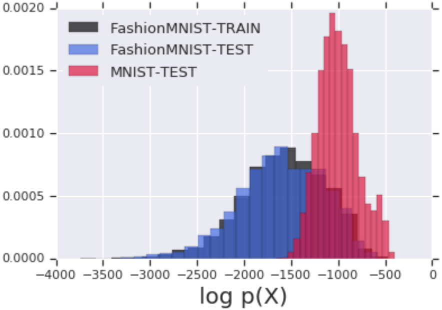


The Ultimate Performance Metric in NLP, J. Briggs, Medium 2021

Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?



Plausibility

Train

Test

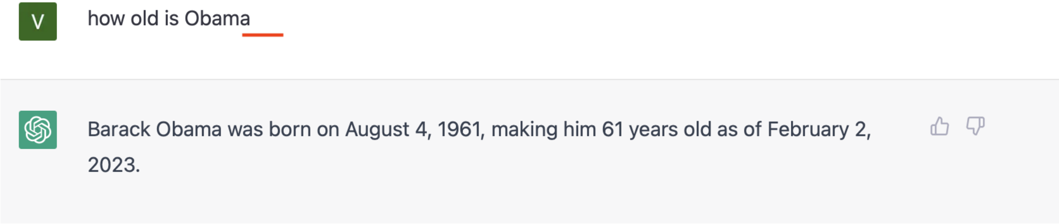
Do Large Language Models Know What They Don't Know?, Yin et al. , ACL, 2023

Do Deep Generative Models Know What They Don't Know?, Nalisnick et al. , ICLR, 2019

Stability/predictability

- Difficult to bound a behavior
- Impossible to predict good/bad answers

$\Rightarrow$ Little/no use in video games

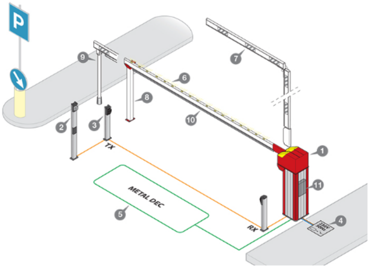


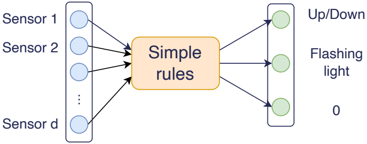
Stability/predictability

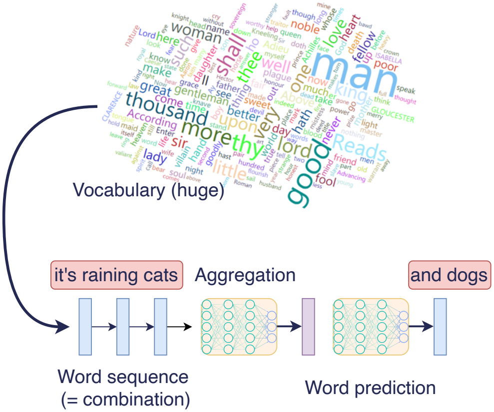
- Difficult to bound a behavior
- Impossible to predict good/bad answers

⇒ Little/no use in video games

Stability, explainability... And complexity







- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable & explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable & non-explainable**

Stability, explainability... And complexity

Interpretability *vs* Post-hoc Explanation

Neural networks = **non-interpretable** (almost always)

too many combinations to anticipate

Neural networks = **explainable a posteriori** (almost always)

[Uber Accident, 2018]

- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable** & **explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable** & **non-explainable**

Transparency : open source / open weight

- Can I modify it? — Adaptation
- What training data was used? — Data contamination / skills
- What editorial stance / censorship is involved? — Access to information
- Why this answer? — Explainability / interpretability

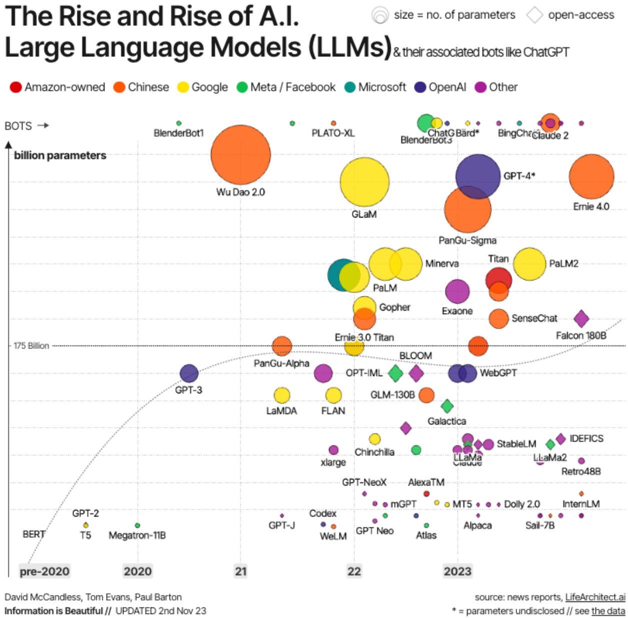
Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

https://crfm.stanford.edu/fmti/May-2024/index.html

Costs / Frugality

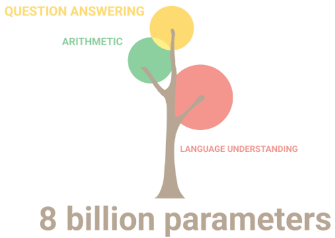


Parameters

1998	LeNet-5	= 0.06M
2011	Senna	= 7.3M
2012	AlexNet	= 60M
2017	Transformer	= 65M / 210M
2018	ELMo	= 94M
2018	BERT	= 110M / 340M
2019	GPT2	= 1,500M
2020	GPT3	= 175,000M

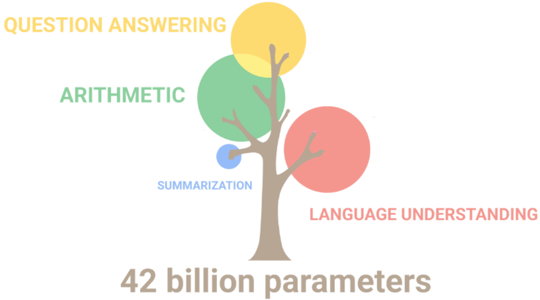
Costs / Frugality

Emergent Capabilities



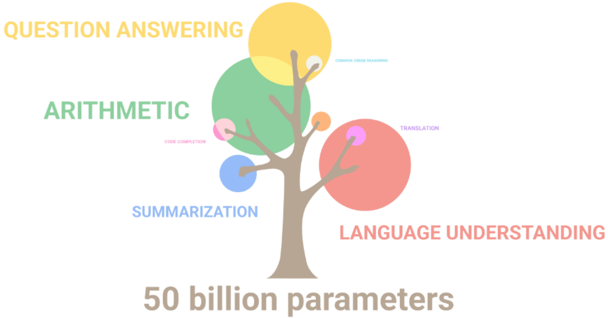
Costs / Frugality

Emergent Capabilities



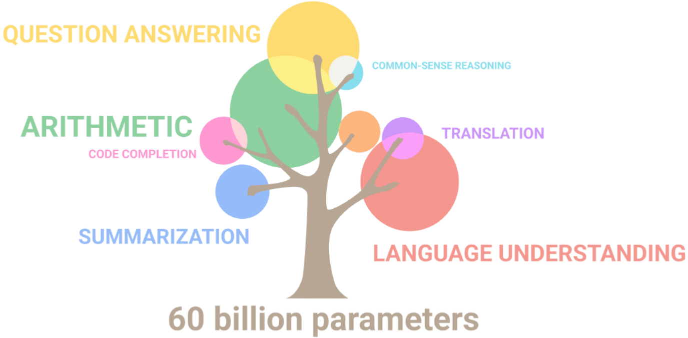
Costs / Frugality

Emergent Capabilities



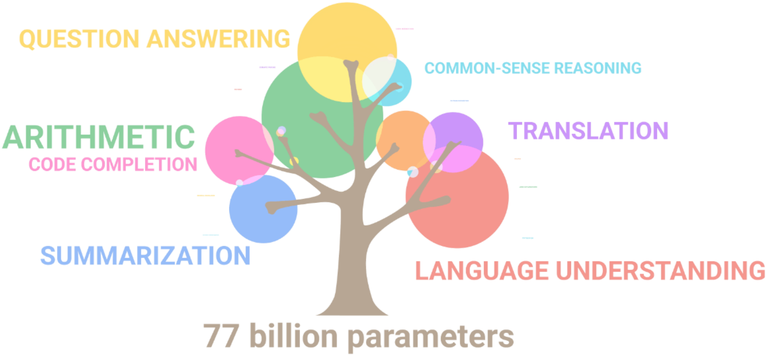
Costs / Frugality

Emergent Capabilities



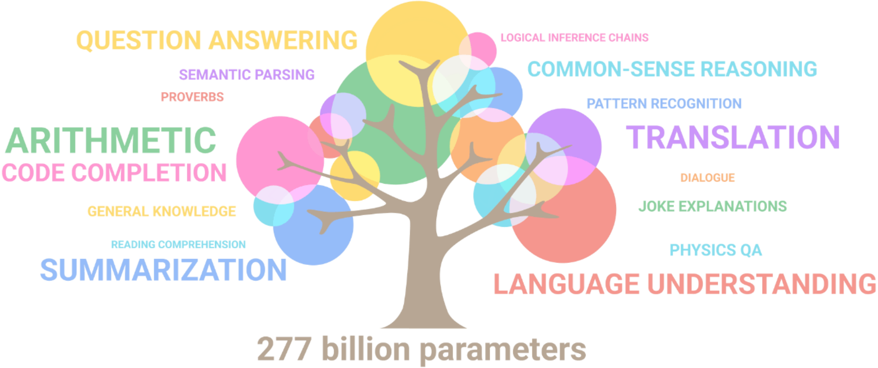
Costs / Frugality

Emergent Capabilities



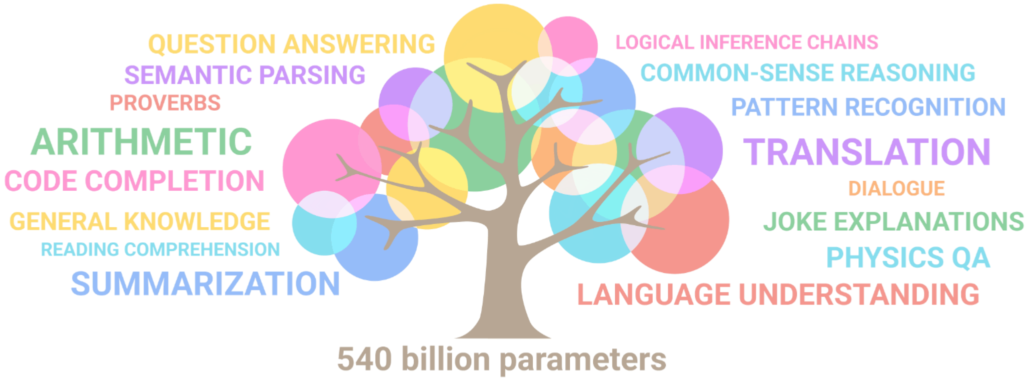
Costs / Frugality

Emergent Capabilities



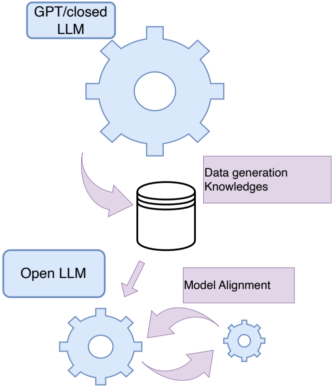
Costs / Frugality

Emergent Capabilities



LLMs & Frugality

Distillation



Pruning

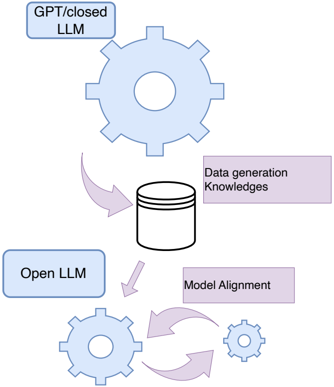
Quantization

Mixture of Experts

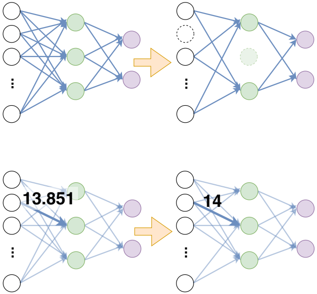
Frugality... Model size **x1000 in 3y**... Then **optimization x1/100 in 2y**

LLMs & Frugality

Distillation



Pruning
Quantization



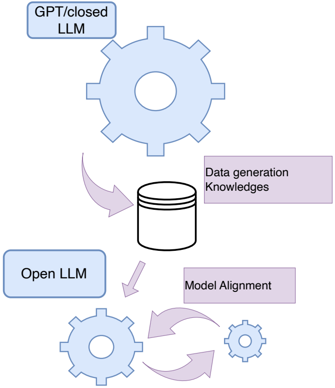
FP32 $\Rightarrow$ INT4

Mixture of Experts

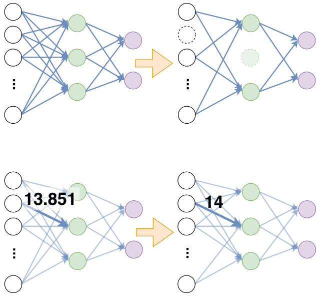
Frugality... Model size **x1000 in 3y**... Then **optimization x1/100 in 2y**

LLMs & Frugality

Distillation

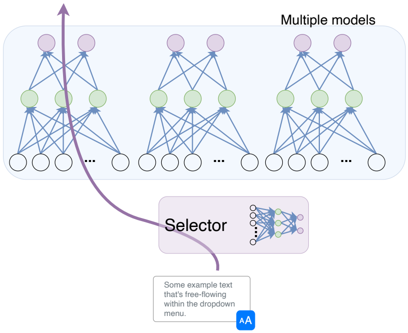


Pruning Quantization



FP32 $\Rightarrow$ INT4

Mixture of Experts



+ Code industrialization

Frugality... Model size **x1000 in 3y**... Then **optimization x1/100 in 2y**

Different behaviors, different costs

Les IA sont démasquées !

Mistral/**Ministral**

SEMI-OUVERT | 8 MDS DE PARAMÈTRES | SORTIE 10/2024

Optimisé pour un temps de réaction rapide, ce modèle est idéal pour des applications nécessitant des réponses immédiates et peut supporter plus de 100 langues. Sorti en octobre 2024.

Impact énergétique de la discussion

8 milliards param. (taille du modèle) × **128** tokens (taille du texte) = **0.30** Wh (énergie conso.)

Ce qui correspond à :

- **0.30**g CO_2 émis
- **5**min ampoule LED
- **33**s vidéos en ligne

Voir plus

DeepSeek/**DeepSeek v3**

SEMI-OUVERT | 671 MDS DE PARAMÈTRES | SORTIE 12/2024

Sorti en décembre 2024, le modèle DeepSeek V3 possède une architecture Mixture-of-Experts qui lui permet d'être d'une très grande taille en diminuant les coûts d'inférence.

Impact énergétique de la discussion

671 milliards param. (taille du modèle) × **225** tokens (taille du texte) = **6**Wh (énergie conso.)

Ce qui correspond à :

- **6**g CO_2 émis
- **2**h ampoule LED
- **12**min vidéos en ligne

Voir plus

Different behaviors, different costs

Different costs for different users/languages

Pour un texte significatif en Français

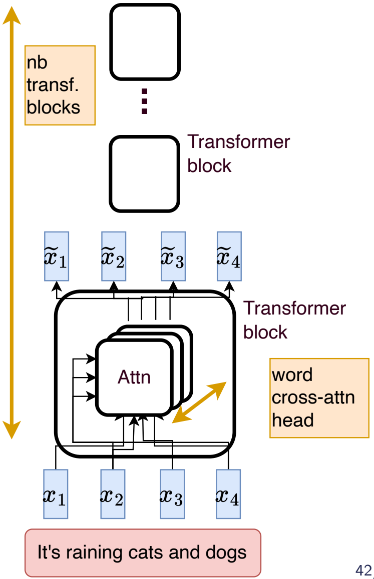
and the same in English

TOKENS 17

CHARACTERS 63

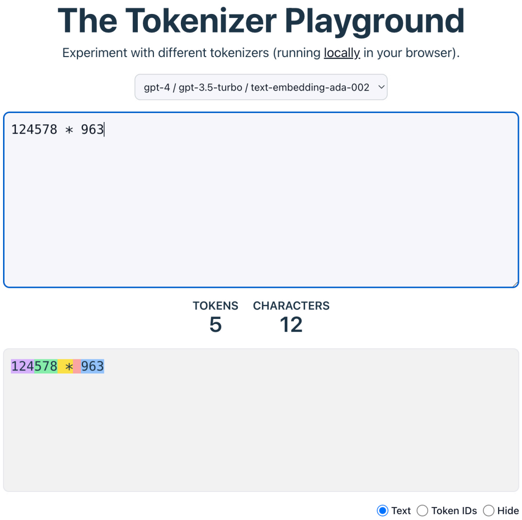
<s> Pour un texte significatif en Français

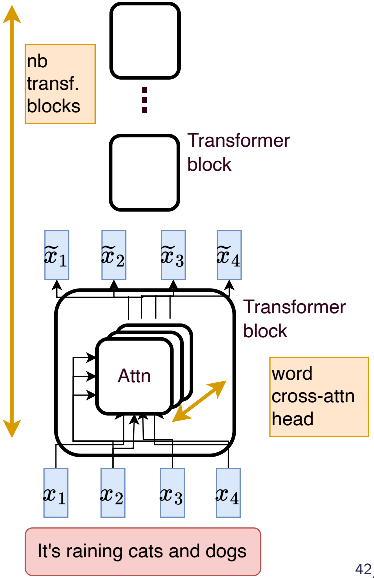
and the same in English



Different behaviors, different costs

Different costs for different users/languages





(Main) Risks derived from ML & LLM

Typology of AI Risks in NLP (L. Weidinger)

Discrimination, exclusion and toxicity

Harms that arise from the language model producing discriminatory and exclusionary speech.

Information hazards

Harms that arise from the language model leaking or inferring true sensitive information.

Misinformation harms

Harms that arise from the language model producing false or misleading information.

Malicious uses

Harms that arise from actors using the language model to intentionally cause harm.

Human-computer interaction harms

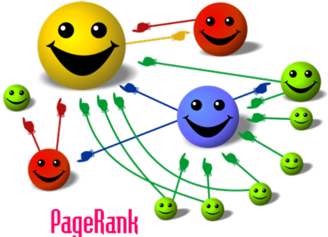
Harms that arise from users overly trusting the language model, or treating it as human-like.

Automation, access and environmental harms

Harms that arise from environmental or downstream economic impacts of the language model.

Access to Information

- Access to dangerous/forbidden information
  - +Personal data
  - Right to digital oblivion
- Information authorities
  - Nature: unconsciously, image = truth
  - Source: newspapers, social media, ...
  - Volume: number of variants, citations (pagerank)
- Text generation: harassment...
- Risk of anthropomorphizing the algorithm
  - Distinguishing human from machine



Machine Learning & Bias

Mustache, Triangular Ears, Fur Texture

Cat

Over 40 years old, white, clean-shaven, suit

Senior Executive

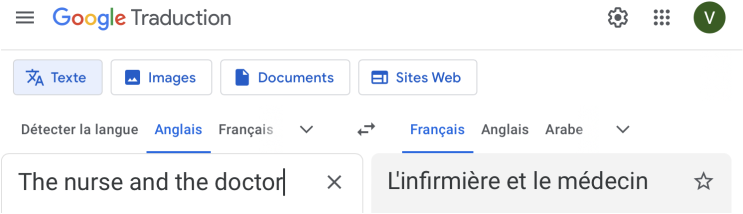
Bias in the data ⇒ bias in the responses

Machine learning is based on extracting statistical biases...

⇒ Fighting bias = manually adjusting the algorithm

Machine Learning & Bias

Sterreotypes from *Pleated Jeans*



- Gender choice
- Skin color
- Posture
- ...

Bias in the data ⇒ bias in the responses

Machine learning is based on extracting statistical biases...

⇒ Fighting bias = manually adjusting the algorithm

Bias Correction & Editorial Line

Bias Correction:

- Selection of specific data, rebalancing
- Censorship of certain information
- Censorship of algorithm results

⇒ Editorial work... Done by whom?

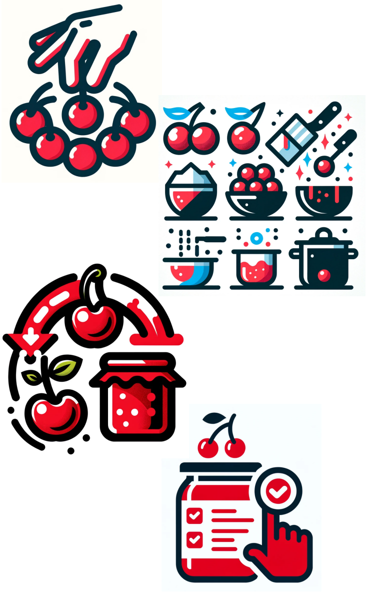
- Domain experts / specifications
- Engineers, during algorithm design
- Ethics group, during result validation
- Communication group / user response

⇒ What legitimacy? What transparency? What effectiveness?

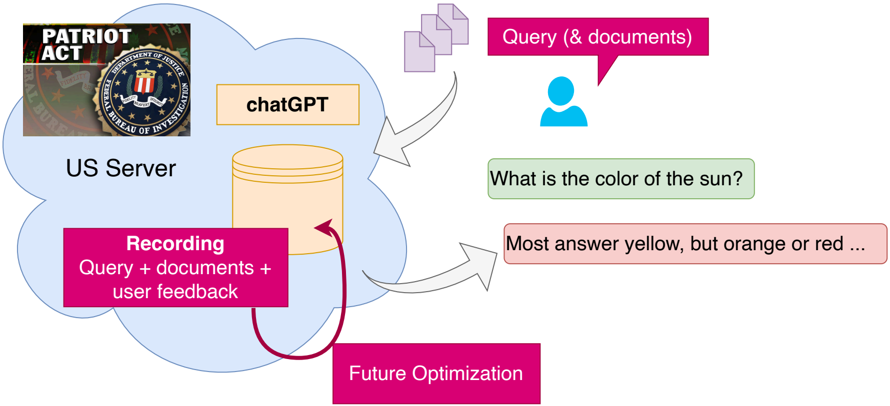
Machine learning is never neutral

1. Data selection
   - Sources, balance, filtering
2. Data transformation
   - Information selection, combination
3. Prior knowledge
   - Balance, loss, a priori, operator choices...
4. Output filtering
   - Post processing
   - Censorship, redirection, ...

$\Rightarrow$ Choices that influence algorithm results



Data Leak(s): different security levels



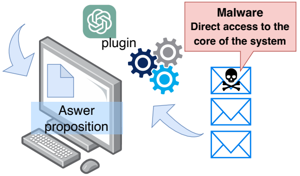
- Transfer of sensitive data
- Exploitation of data by OpenAI (or others)
- Data leakage in future models

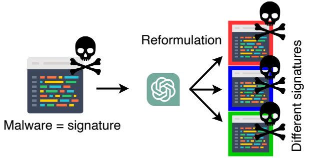
Data Leak(s): different security levels

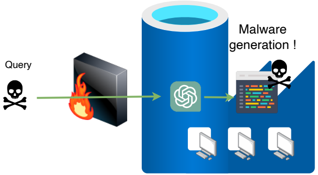
Level 1: **Commercial tools, free to use**	Variable licenses (depending on the companies and subject to change over time). Uncertain data protection, risk to personal data. *chatGPT, Mistral, Perplexity, ...*
Level 2: **Commercial tools, paid licence**	Strong contractual guarantees. Risks associated with the *Patriot Act*. Possible to enforce non-storage of queries. *chatGPT, Mistral, Perplexity, ...*
Level 3: **Commercial tools, paid licence +**	+ Negotiation on the server location/data security. *Microsoft Azur, Mistral, AWS, ...*
Level 4: **Local use**	Use of a locally operated LLM, with no data transferred over the web. *HuggingFace, Ollama, ...*

Security Issues

- Plug-ins ⇒ Often significant security vulnerabilities for users
  - Email access / transfer of sensitive information etc...
- Management issues for companies
  - Securing (very) large files
- Increased opportunities for malware signatures
  - ≈ software rephrasing
- New problems!
  - Direct malware generation

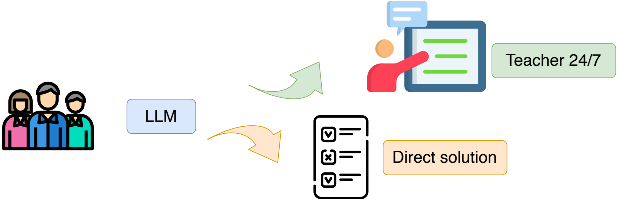




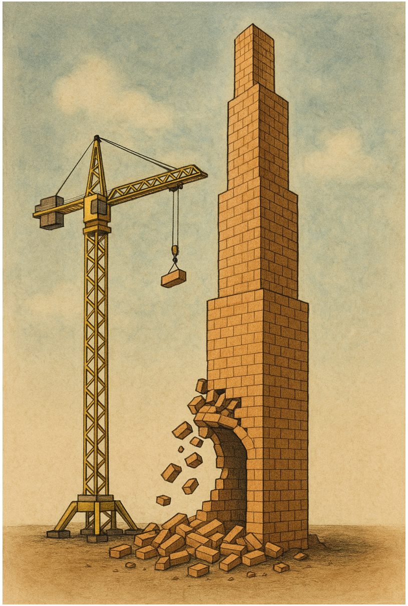


What Educational Challenges

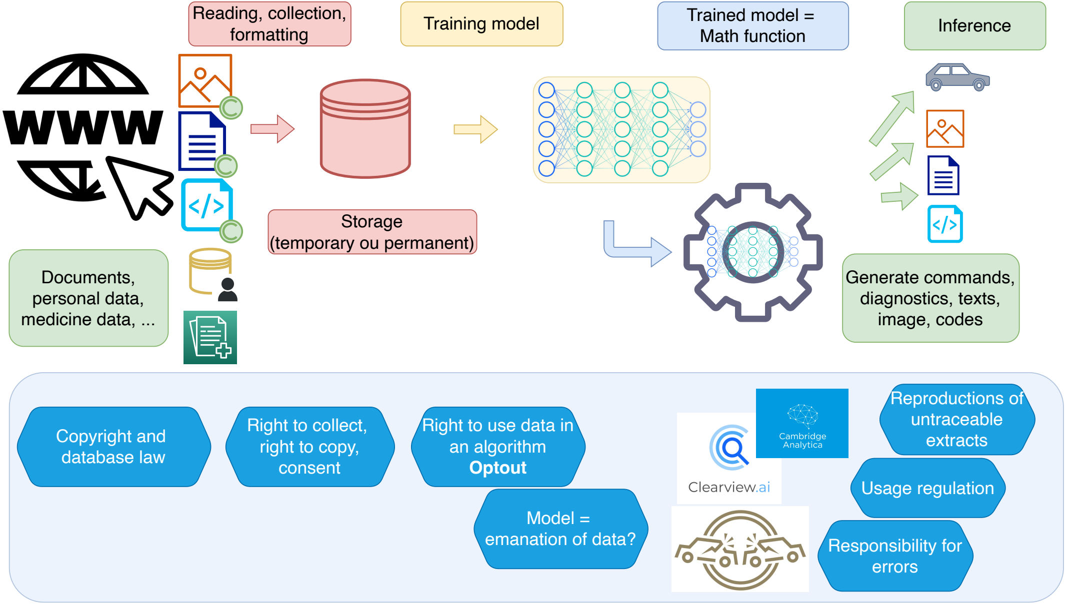
- Redefine our educational priorities, subject by subject, as we did with Wikipedia/calculator/...
  - Accept the decline of certain skills
- Train students in the use of LLMs, while managing to temporarily prohibit their use



- Learn to recognize LLM-generated content, use detection tools.



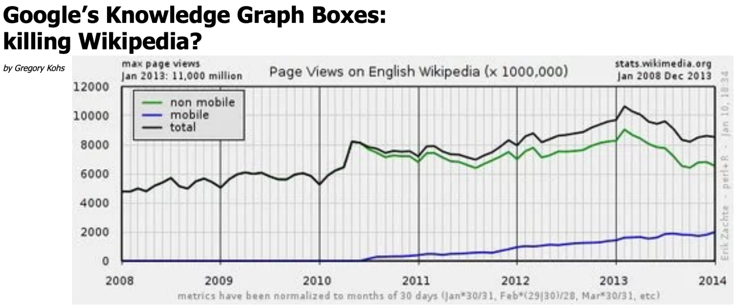
Legal Risks/Questions



Economic Questions

- Funding/Advertising $\Leftrightarrow$ **visits** by internet users
- Google knowledge graph (2012) $\Rightarrow$ fewer visits, less revenue
- chatGPT = encoding web information... $\Rightarrow$ much fewer visits?

$\Rightarrow$ What **business model for information sources** with chatGPT?

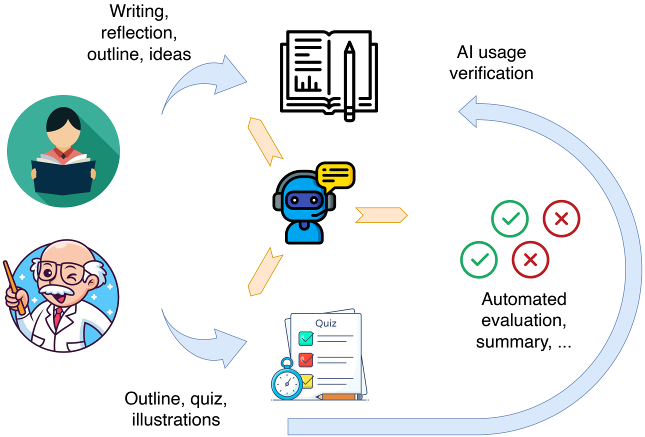


$\Rightarrow$ Who does **benefit from the feedback**? [StackOverFlow]

Risks of AI Generalization

AI everywhere =
loss of meaning?

- In the educational domain
- Transposition to HR
- To project-based funding systems



How to approach the ethics question?

Medicine

1. **Autonomy:** the patient must be able to make informed decisions.
2. **Beneficence:** obligation to do good, in the interest of patients.
3. **Non-maleficence:** avoid causing harm, assess risks and benefits.
4. **Equality:** fairness in the distribution of health resources and care.
5. **Confidentiality:** confidentiality of patient information.
6. **Truth and transparency:** provide honest, complete, and understandable information.
7. **Informed consent:** obtain the free and informed consent of patients.
8. **Respect for human dignity:** treat all patients with respect and dignity.

Artificial Intelligence

1. **Autonomy:** Humans control the process
2. **Beneficence:** in the interest of whom? User + GAFAM...
3. **Non-maleficence:** Humans + environment / sustainability / malicious uses
4. **Equality:** access to AI and equal opportunities
5. **Confidentiality:** what about the Google/Facebook business model?
6. **Truth and transparency:** the tragedy of modern AI
7. **Informed consent:** from cookies to algorithms, knowing when interacting with an AI
8. **Respect for human dignity:** harassment behavior/ human-machine distinction

How to approach the ethics question?

Medicine

1. **Autonomy:** the patient must be able to make informed decisions.
2. **Beneficence:** obligation to do good, in the interest of patients.
3. **Non-maleficence:** avoid causing harm, assess risks and benefits.
4. **Equality:** fairness in the distribution of health resources and care.
5. **Confidentiality:** confidentiality of patient information.
6. **Truth and transparency:** provide honest, complete, and understandable information.
7. **Informed consent:** obtain the free and informed consent of patients.
8. **Respect for human dignity:** treat all patients with respect and dignity.

Artificial Intelligence

1. **Autonomy:** Humans control the process
2. **Beneficence:** in the interest of whom? User + GAFAM...
3. **Non-maleficence:** Humans + environment / sustainability / malicious uses
4. **Equality:** access to AI and equal opportunities
5. **Confidentiality:** what about the Google/Facebook business model?
6. **Truth and transparency:** the tragedy of modern AI
7. **Informed consent:** from cookies to algorithms, knowing when interacting with an AI
8. **Respect for human dignity:** harassment behavior/ human-machine distinction

Conclusion

Upcoming Challenges

- **What about hallucinations?**
  - Should we try to reduce them or learn to live with them?
  - Will LLMs improve? In what directions?
  - Do LLMs make us *lose* our connection to truth? To verification?
- **Do we need small or large language models?**
  - How much does it cost? Is it sustainable?
  - With or without fine-tuning?
  - What does frugality mean in the world of LLMs?
- **When others use them... What impact does it have on me?**
  - Productivity (fellow researchers, coders, reviewers, ...)
  - Education: managing/training *tech-savvy* students
- **Data protection... Mine and others'**
  - Is it reasonable to train LLMs on GitHub, Wikipedia, scientific papers, news outlets, etc.?
  - How important is privacy? What are the risks when using an LLM?

Upcoming Challenges

- **What about hallucinations?**
  - Should we try to reduce them or learn to live with them?
  - Will LLMs improve? In what directions?
  - Do LLMs make us *lose* our connection to truth? To verification?
- **Do we need small or large language models?**
  - H
  - W
  - W
- **When others use them... What impact does it have on me?**
  - Productivity (fellow researchers, coders, reviewers, ...)
  - Education: managing/training *tech-savvy* students
- **Data protection... Mine and others'**
  - Is it reasonable to train LLMs on GitHub, Wikipedia, scientific papers, news outlets, etc.?
  - How important is privacy? What are the risks when using an LLM?

The smartphone has made me an *augmented human*...
Will the LLM make me an *augmented researcher*?

$\Rightarrow$ Still, have a look at NotebookLM

Tools and Questions

New tools:

- New ways to handle existing problems
- Address new problems
- ... But obviously, it doesn't always work!
- AI often makes mistakes (assistant *vs* replacement)

Learning to use an AI system

- AI not suited for many problems
- AI = part of the problem (+interface, usage, acceptance...)

Maturity of Tools & Environments

(More) mature tools

- **Environments**: Jupyter, Visual Studio Code, ...
- **Machine Learning** `Scikit-Learn`: blocks to assemble
  - Training: 1 week
  - Project completion: few hours to few days
- **Deep Learning** `pytorch`, `tensorflow`: building blocks... but more complex
  - Training: 2-5 weeks
  - Project completion: few days to few months
  - Mandatory for text and image

- A data project = 10 or 100 times less time / 2005
- Developing a project is **accessible to non-computer scientists**

Levels of Access to Artificial Intelligence

1. User via an interface: *chatGPT*
   - Some training is still required (2-4h)
2. Using Python libraries
   - Basics on protocols
   - Standard processing chains
   - Training: 1 week-3 months (ML/DL)
3. Tool developer
   - Adapt tools to a specific case
   - Integrate business constraints
   - Build hybrid systems (mechanistic/symbolic)
   - Mix text and images
   - Training: ≥ 1 year

Digital Sovereignty: the Entire Chain

Pre-trained model construction

Data Mastery
- Collection/balance
- Cleaning

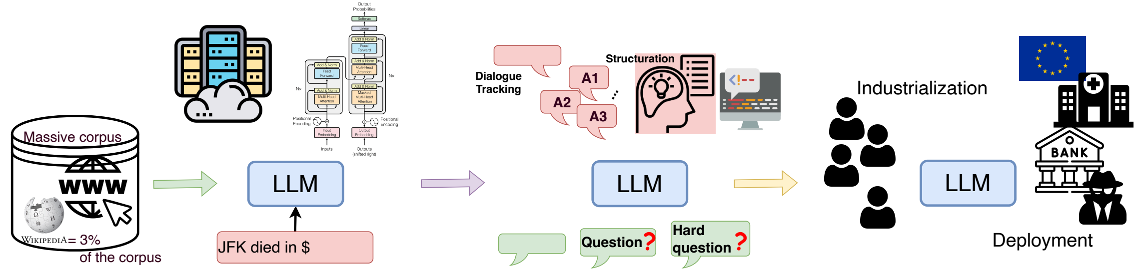
Training
- Computation power (x1000 GPU)
- Architecture ML

Model Fine-Tuning

Data Mastery & Construction
- Human interactions +++
- Dataset cost
- Domain adaptation

Model exploitation

Optimization / Cost reduction
- MLOps skills
- Local deployment



A Multitude of Professions

Data architect / manager

- Data management & hardware devices (storage, network, ...)

Data Engineer

- Update & Query on the data

+ DPO :
Data Protection Officer

Data Analyst

- Data visualization (chart, indicators, ...)
- Statistical trends

Prompt Engineer

- Query on LM/foundation models with "prompts"

Data Scientist

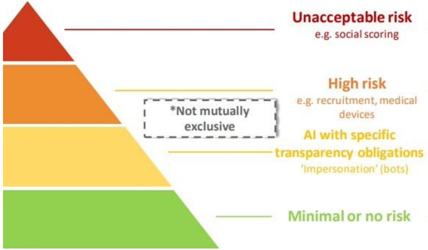
- Query the data / critical selection & balance
- Algorithm development / adaptation / evaluation
- Advanced data visualization

MLOps Engineer

- Algorithm optimization
- Industrialize software solutions

Factors of Acceptability for Generative AI

1. **Utilitarianism**:
   - Performance (acceptance factor of chatGPT)
   - Reliability / Self-assessment
2. **Non-dangerousness**:
   - Bias / Correction
   - Transparency (editorial line, human/machine confusion)
   - Reliable Implementation
   - Sovereignty (?)
   - Regulation (AI act)
     - Avoid dangerous applications
3. **Know-how**:
   - Training (usage/development)



chatGPT: A Simple Step

- Training & Tuning Costs

  4-5 Million Euros / training ⇒ chatGPT is **poorly trained**!

- Data Efficiency

  chatGPT > 1000x a human's lifetime reading

- Identify Entities, Cite Sources

  Anchoring responses in knowledge bases
  Anchoring responses in sources

- Multiplication of initiatives: GPT, LaMBDA, PaLM, BARD, BLOOM, Gopher, Megatron, OPT, Ernie, Galactica...

- Public involvement,

  impact on information access