

ARTIFICIAL INTELLIGENCE AND MICROBIOLOGY : WHICH TOOLS FOR WHICH APPLICATIONS, AND AT WHAT COST ?

23 septembre 2026 – Séminaire Micalis
La lutte contre l'antibiorésistance à l'ère de l'IA

Vincent Guigue
<https://vguigue.github.io>

INTRODUCTION



Digital & Artificial Intelligence

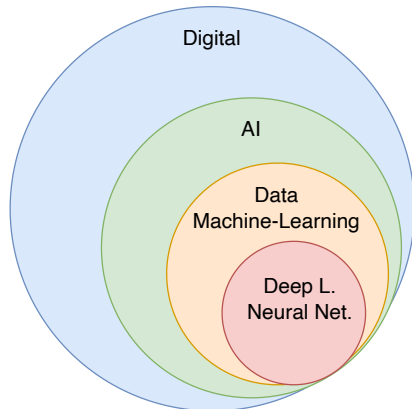
■ Two related but distinct concepts

■ AI: Different Definitions

1956 Any algorithm / program

1960-2012 Expert systems and logical reasoning

2012- Data & neural networks



A. Turing



Marvin Minsky

G. Hinton



Y. Lecun



Computer

1941

1956

Neural Networks

1986

Deep-learning

2012

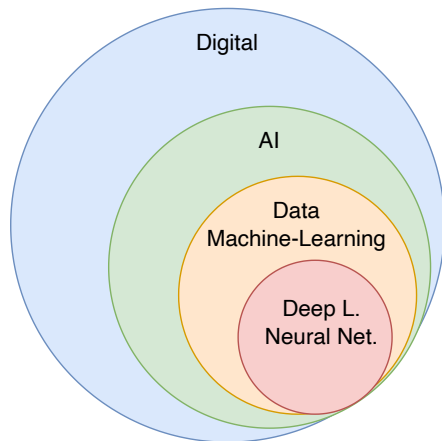
Computer-
Sciences

AI: wide variety of algorithms
Mainly : Expert System + Reasoning

AI= Neural Networks



AI's Place in the Digital Domain



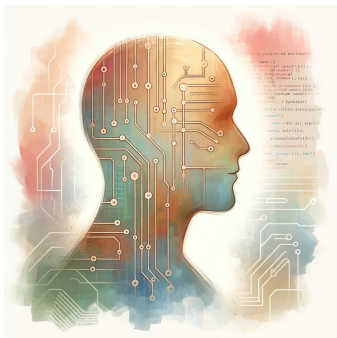
- Self-checkout at the supermarket, Google drive, SAP (basis)
- Google Maps, Knowledge bases, Expert systems
- Predictive systems (e.g., real estate market), recommendation
- Vision, chatbot, advanced systems

A lot of digital tools transform the job market...

⇒ A global view is required to understand the transformation



Artificial Intelligence & Machine Learning



Input ($\mathbf{x}$)	Output ($\mathbf{Y}$)	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

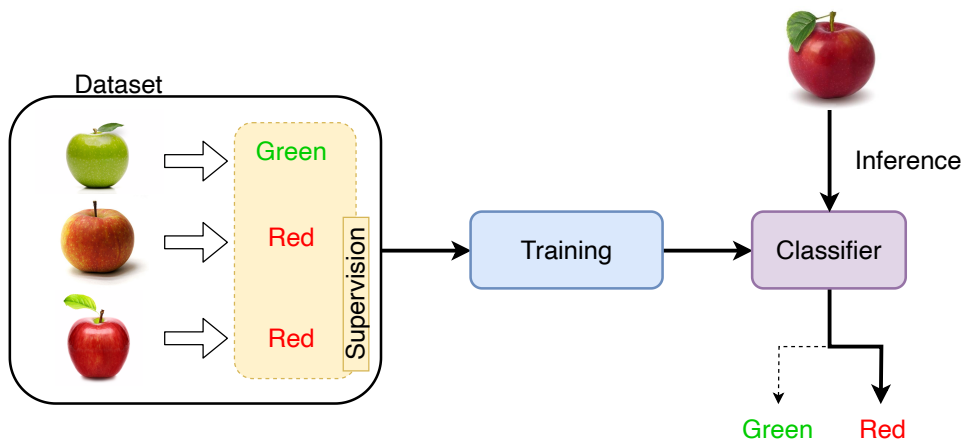
N-AI (Narrow Artificial Intelligence), dedicated to a single task

≠ G-AI (General AI), which replaces humans in complex systems.

Andrew Ng, 2015



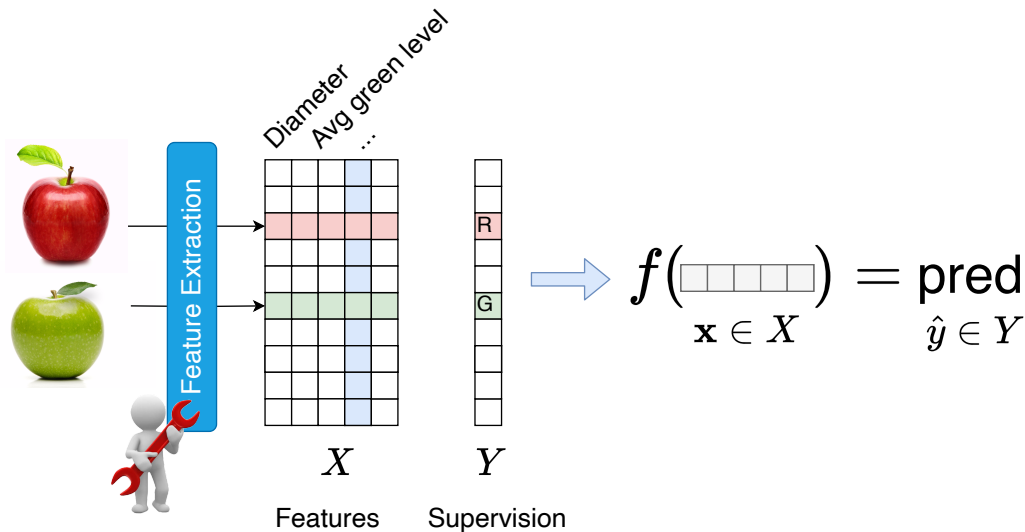
Supervised Processing Chain & Models



- Promise = building a model *solely* from observations



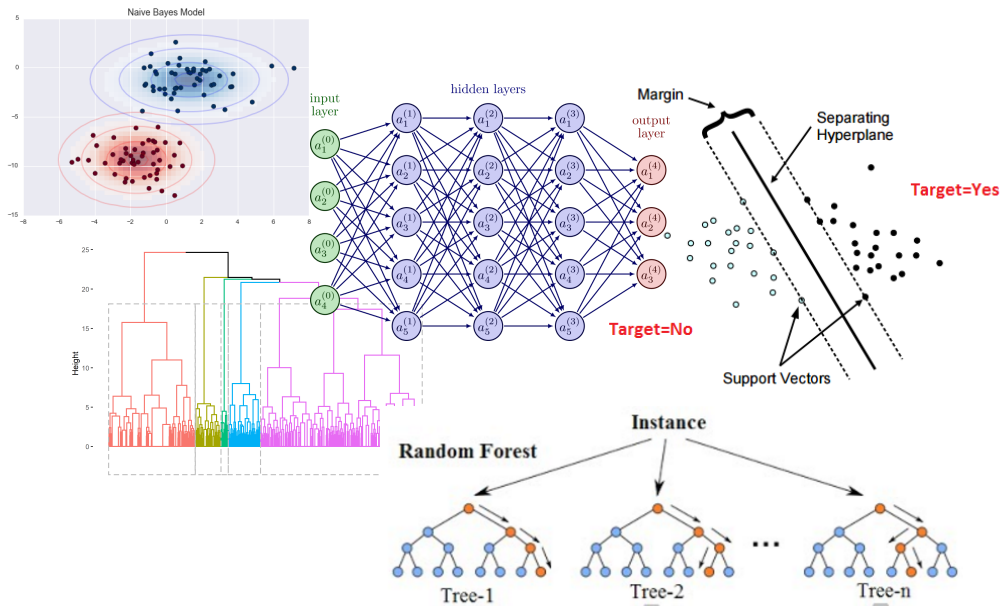
Supervised Processing Chain & Models



- Think about the feature space = observation description
- ... And about required supervision

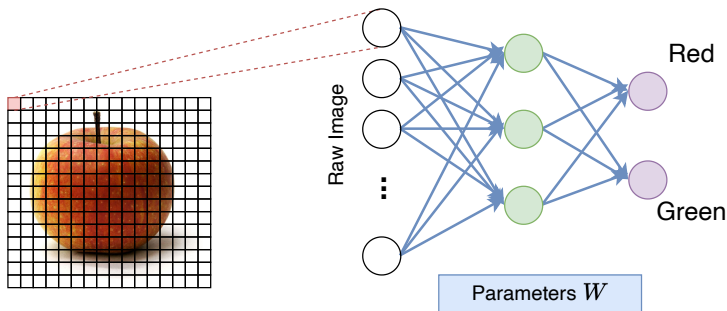


Supervised Processing Chain & Models





Supervised Processing Chain & Models

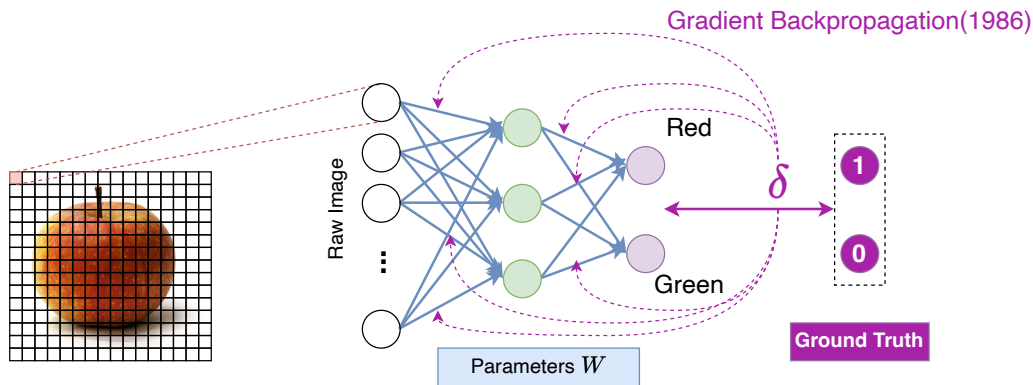


■ Random initialization...

And random decision-making (at first!)



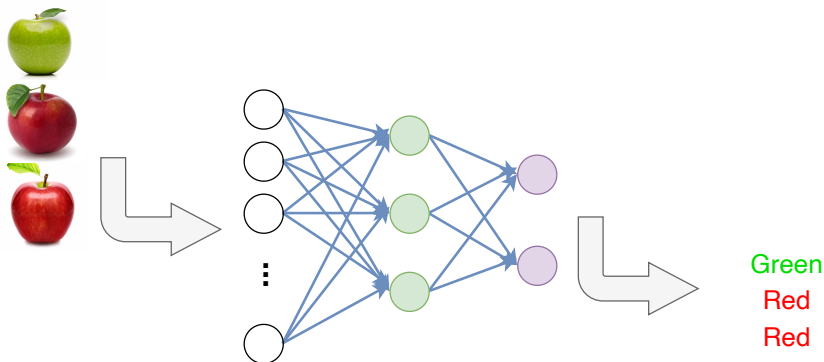
Supervised Processing Chain & Models



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



Supervised Processing Chain & Models

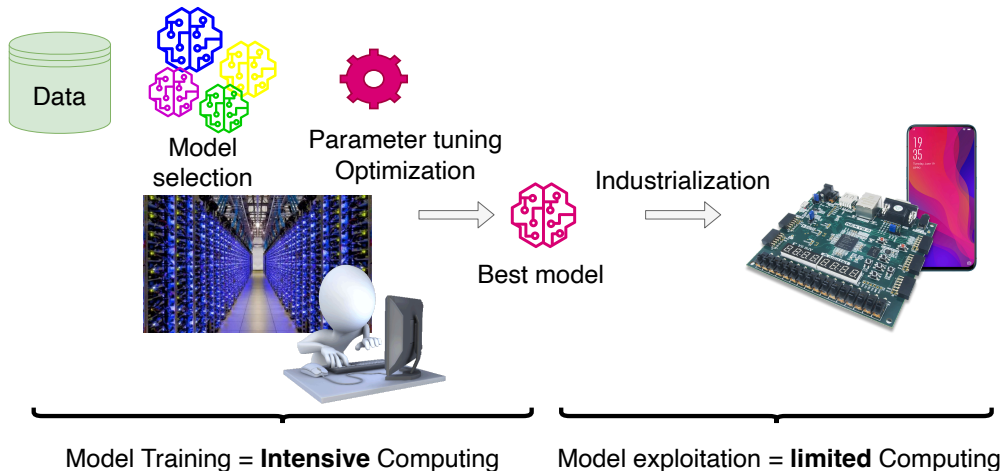


- **Training** is slow and costly
- **Inference** is (much) faster



Supervised Processing Chain & Models

Clearly separate the different steps in machine-learning





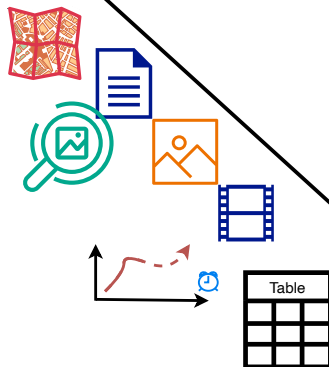
What can we do with AI?

Applications



- Digital agriculture
- Design of Experiments
- Biodiversity measurements
- Medical diagnosis
- Stock exchange prediction
- Exploitation optimization
- ...

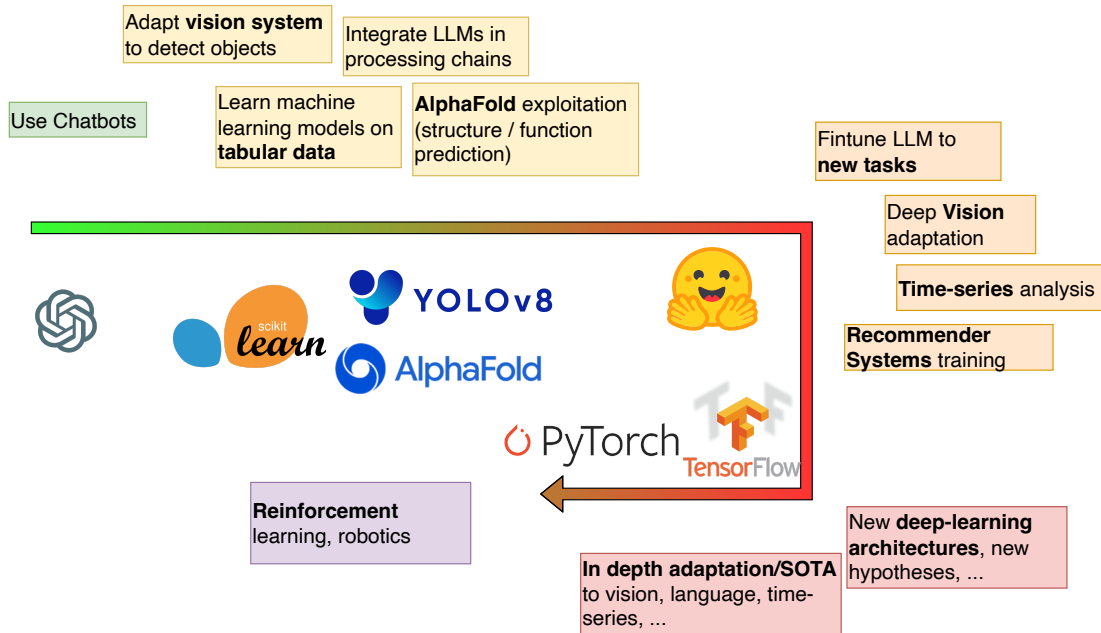
Media



- Image, text, time-series, movies, maps
- Table of data

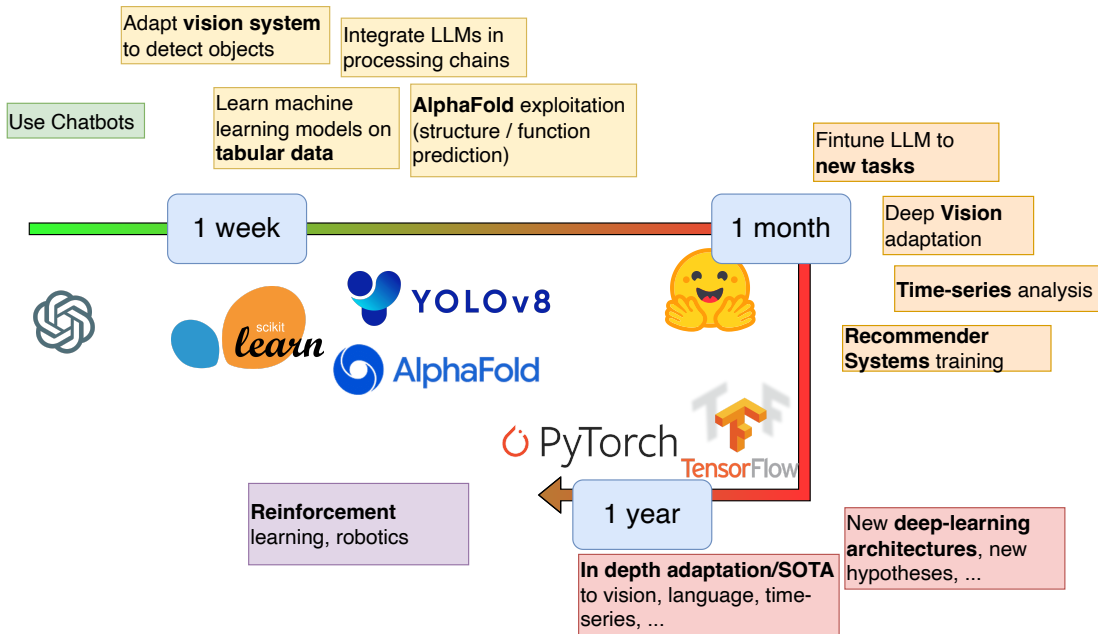


What is the cost to access AI (today)?



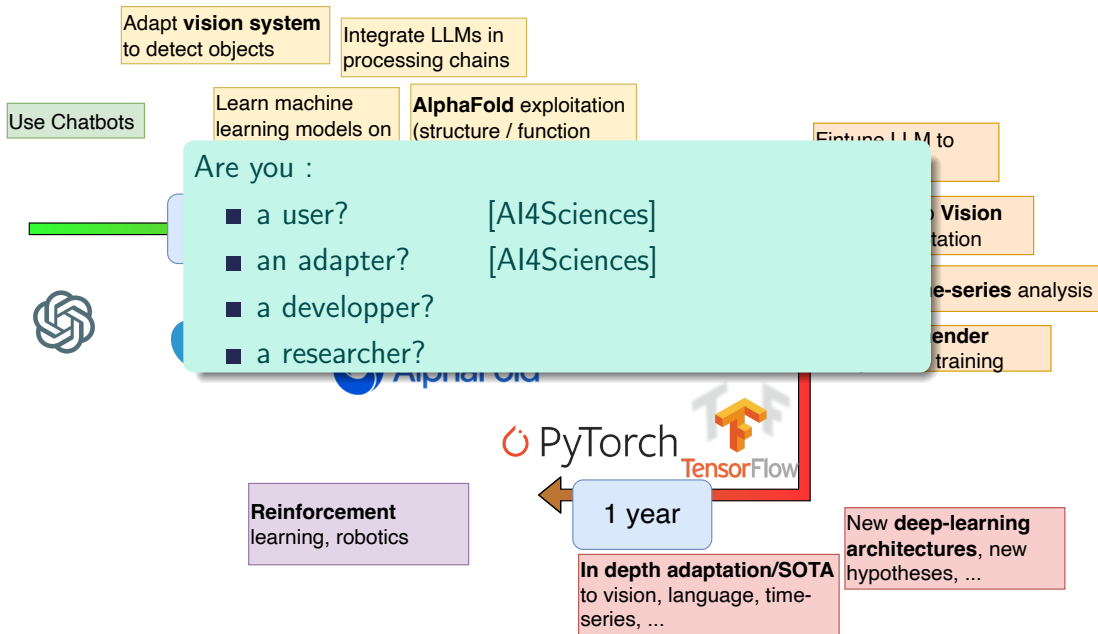


What is the cost to access AI (today)?





What is the cost to access AI (today)?





What is the cost of AI?



Sensor monitoring,
basic decision

Autonomous
driving system



Local platform to
build & deploy AI
models

Answering LLM
calls



Training a vision
system

Training an LLM

0 - 10 W

20 - 150 W

0.2 - 1.5 kW

1 - 5 kW

0.5 - 50 MW

- Electricity, water, rare-earth elements, CO₂ emissions, financing costs, etc.
⇒ Most costs scale proportionally



So, let's stop speaking of AI !

Machine-learning

- Easy to handle, cheap to compute
- Many (many) applications
- (often) Provide strong system optimization
- Should be part of the curriculum for all engineering students

Deep-learning

- Tailor made system / haute-couture for numerical data
- Basic entry for semantic data (text, image, voice, users' traces, ...)
- Multi-modal systems, new paradigm (e.g. self-training)

LLM

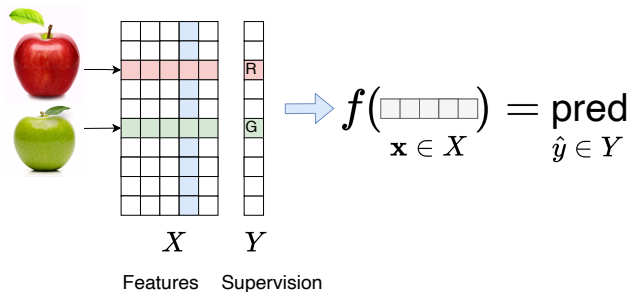
- ML systems
 $\propto 1\text{k params}$
DL systems $\propto 1\text{M}$
LLM systems $\propto 10^3\text{M}$
- New applications, new interfaces to existing systems...
- Major societal impact (education, jobs market, information access, ...)

⇒ Choose your keywords more carefully!

FROM
ARTIFICIAL INTELLIGENCE
TO DEEP-LEARNING &
GENERATIVE AI



Great opportunity of Machine-Learning



Numerous opportunities

- Experimental design
- Biological interaction prediction
- Nutritional value prediction for animal feed
- Medical diagnosis
- ...

Requirements: what you observe, what you expect

- X : Observations
 - Historical data + perimeter (= what do we need to observe)
 - Variations
- Y : Targets
 - Historical + Synchronized with observations

Various questions: is it possible/easy for a human to make the prediction? Can you explain the reasoning? What format is the data in?



A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

Building an opinion mining system:

1 What would a linguist search for?

2 How to generalize?

3 Implement with fast regex

- Gather documents
- Analyze
- Count

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

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

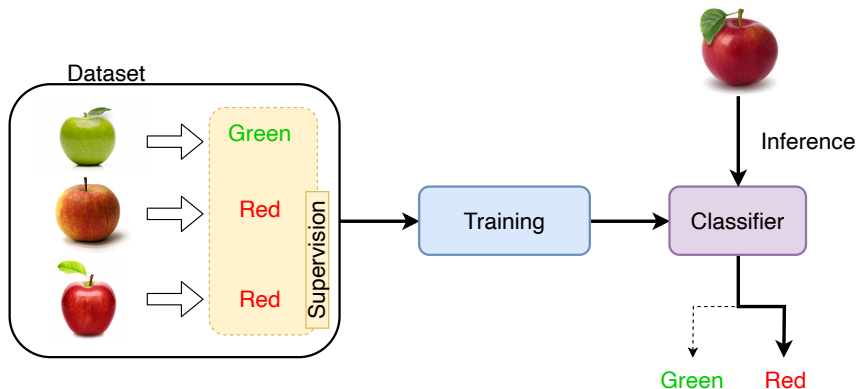
* $\longrightarrow$ {hate, loathe, detest} $\longrightarrow$ * $\longrightarrow$ #product



A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

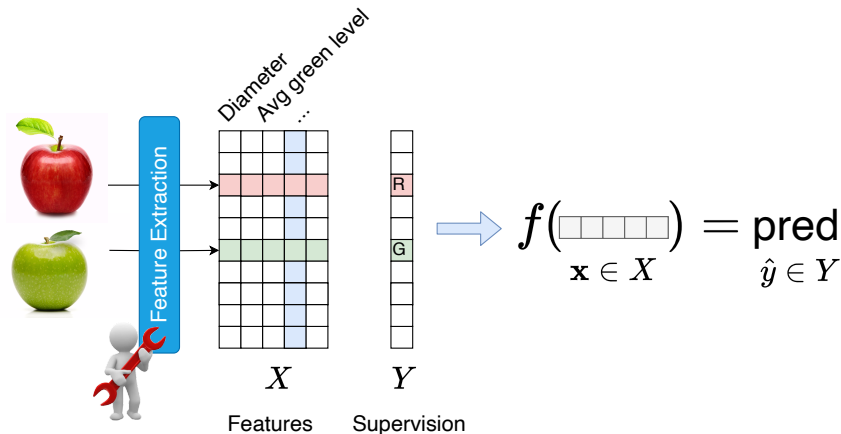




A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?





A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

This new iPhone, what a marvel

An iPhone, What a scam!

Half the price is for the logo

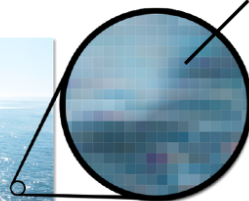
Apple once again proves that perfection can be sold

How do we turn
this text data into
a table?

How do we turn
this image into a
table?



{R = 97, G = 134, B = 162}



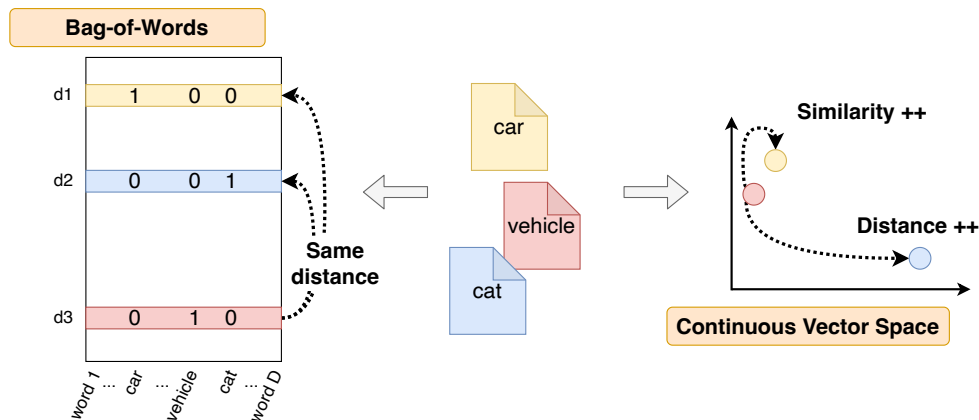


A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

1910-2015 : **Representation learning** $\Rightarrow$ data in continuous latent space





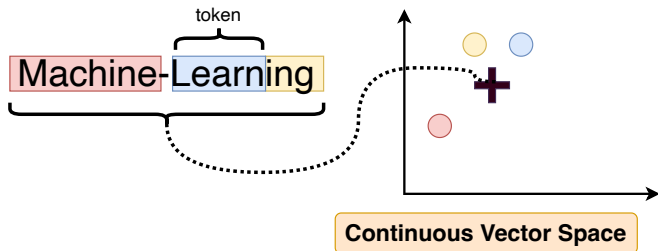
A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

1910-2015 : **Representation learning** $\Rightarrow$ data in continuous latent space

Word Piece statistical split





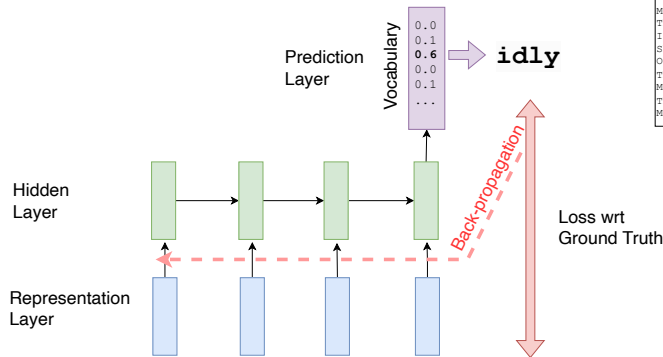
A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

1910-2015 : **Representation learning** $\Rightarrow$ data in continuous latent space

2015-... : **Language models** + pre-training



```
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

The fluffy cat napped lazily in the sunbeam.



A very quick tour of NLP

1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

1980-2010 : How to learn directly from **annotated data**?

1910-2015 : **Representation learning** $\Rightarrow$ data in continuous latent space

2015-... : **Language models** + pre-training

2017-... : **Transformer**, multimodal foundation models

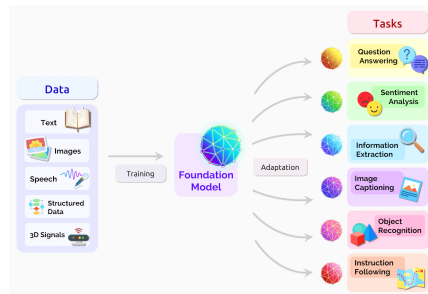
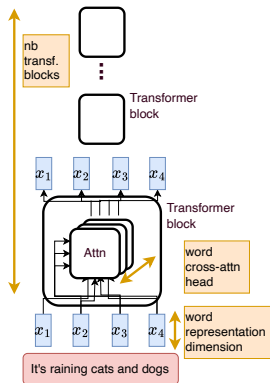
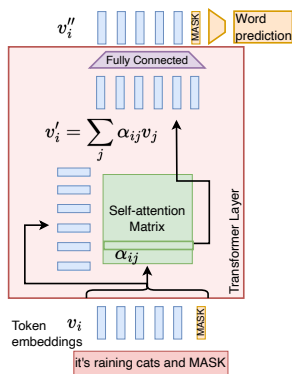




Image representation

The SIFT wave (2005)

- Extracting patches
- Cluster = visual word
- Counting = bag of word

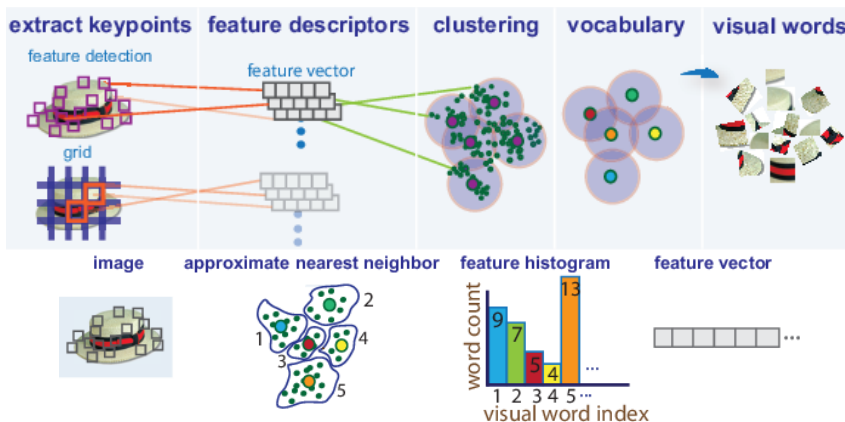
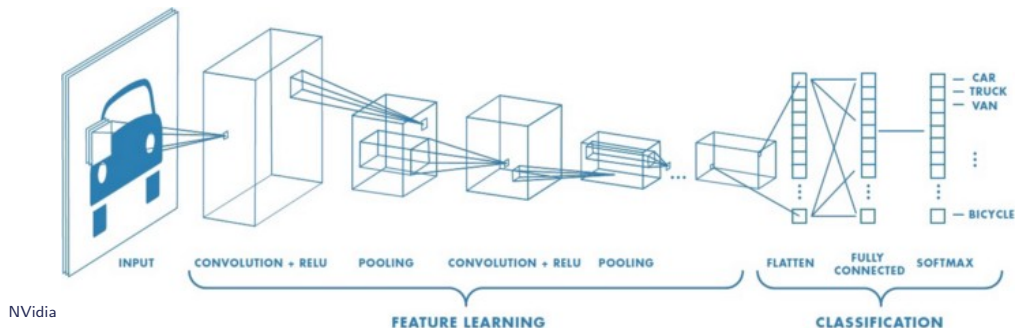




Image representation



The convolution wave (2012)

Sliding filters

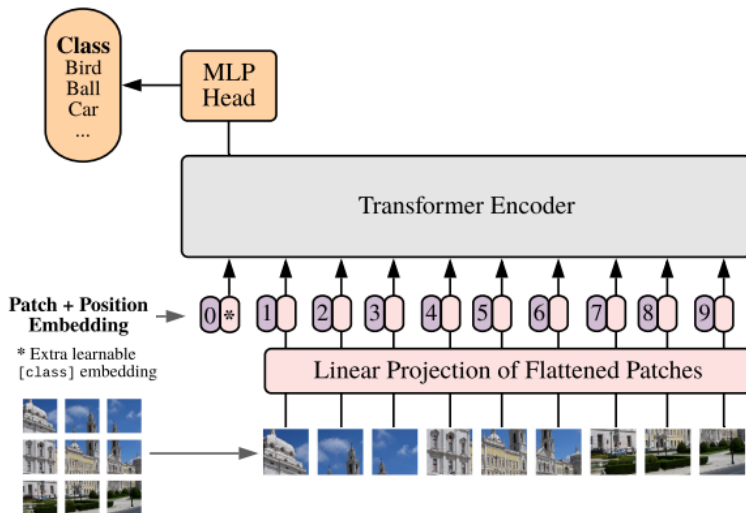
- Few parameter (relatively frugal)
- Pattern extraction
- Hierarchical information extraction/aggregation



Image representation

The Vision Transformer (ViT) wave (2020-21)

- Analyzing the relations between image regions



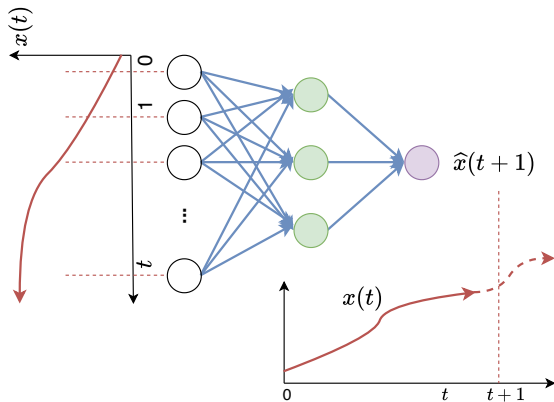
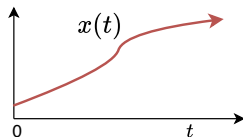


Time-series representation

Any dynamical system prediction

- Plant/animal growth
- Stock market
- Temperature evolution
- ...

- Various possible architecture
- Pattern extraction



Once again :

time series = vector representation

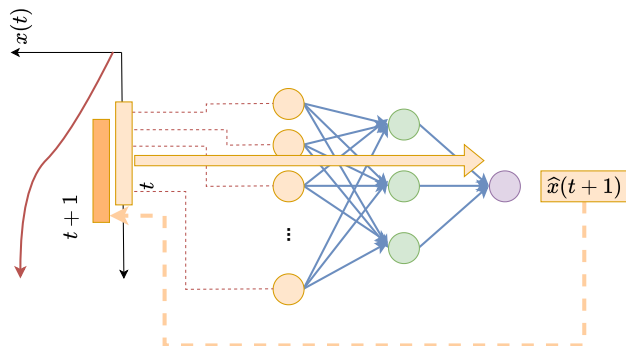
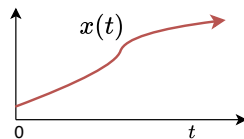
- Distance / similarity
- Integration with other data



Time-series representation

Any dynamical system prediction

- Plant/animal growth
- Stock market
- Temperature evolution
- ...
- Various possible architecture
- Pattern extraction

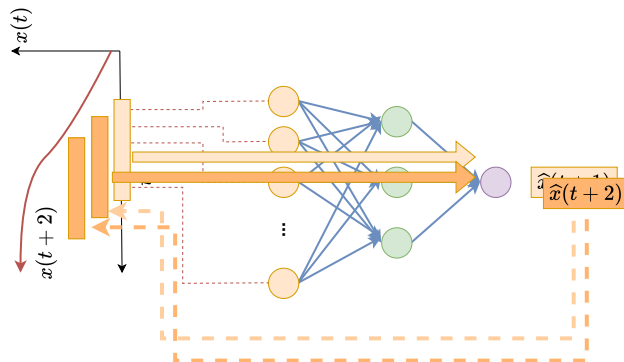
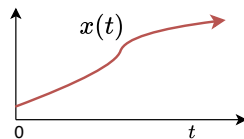




Time-series representation

Any dynamical system prediction

- Plant/animal growth
- Stock market
- Temperature evolution
- ...
- Various possible architecture
- Pattern extraction



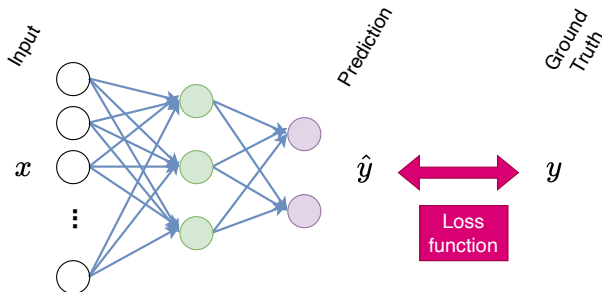


Deep-learning = taylor made solutions

Combining different objectives at different scales

- Triplet loss
- Self supervision / masking
- Profiling to tackle missing information
- ...

Taylor made = more expensive



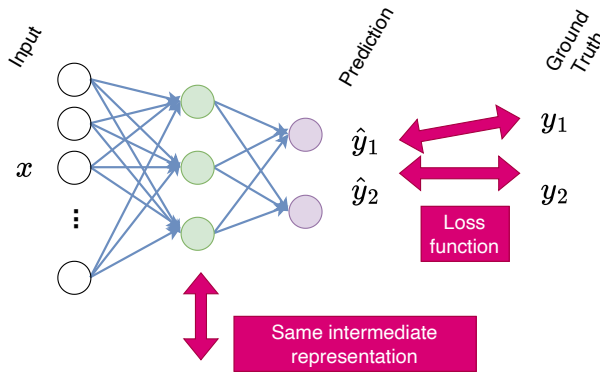


Deep-learning = taylor made solutions

Combining different objectives at different scales

- Triplet loss
- Self supervision / masking
- Profiling to tackle missing information
- ...

Taylor made = more expensive



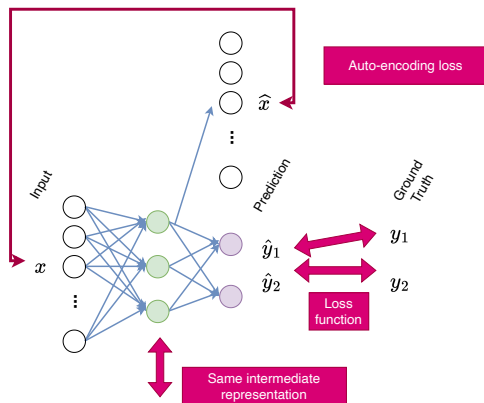


Deep-learning = taylor made solutions

Combining different objectives at different scales

- Triplet loss
- Self supervision / masking
- Profiling to tackle missing information
- ...

Taylor made = more expensive



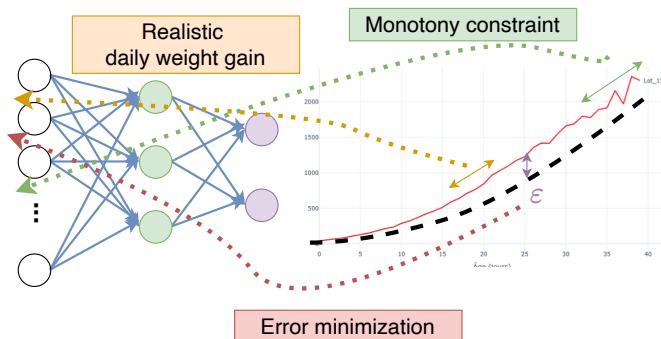


Deep-learning = taylor made solutions

Combining different objectives at different scales

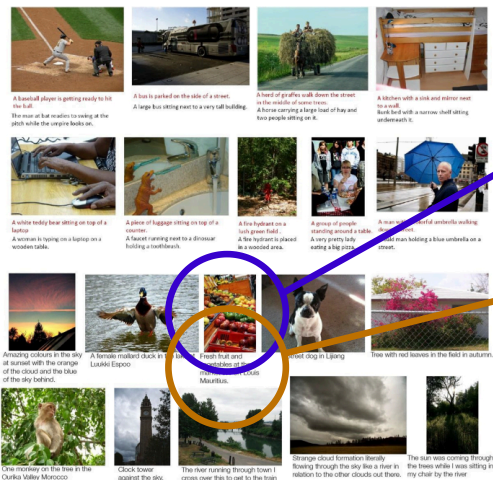
- Triplet loss
- Self supervision / masking
- Profiling to tackle missing information
- ...

Taylor made = more expensive

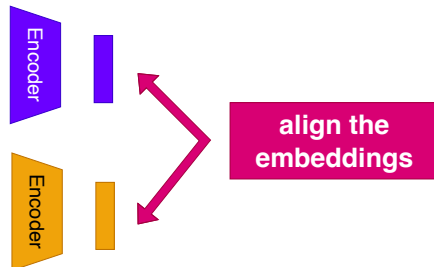




Deep-learning = deal with different modalities



Vector representations

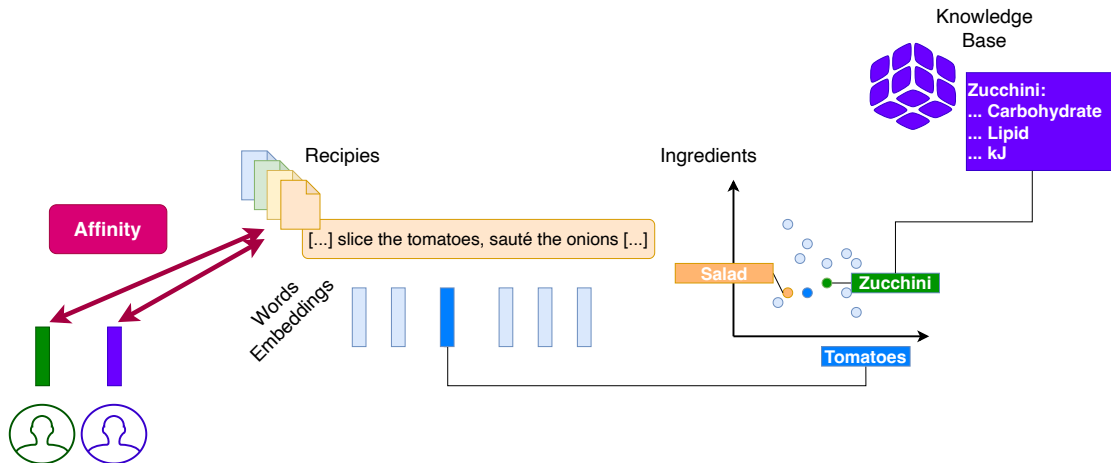


- Aligning languages (translation), aligning modalities
- New applications: image description, handwritten transcription, ...



Word, music, speech, image, user embeddings

- User embedding $\Rightarrow$ recommender system
- Knowledge embedding $\Rightarrow$ neuro-symbolic architecture

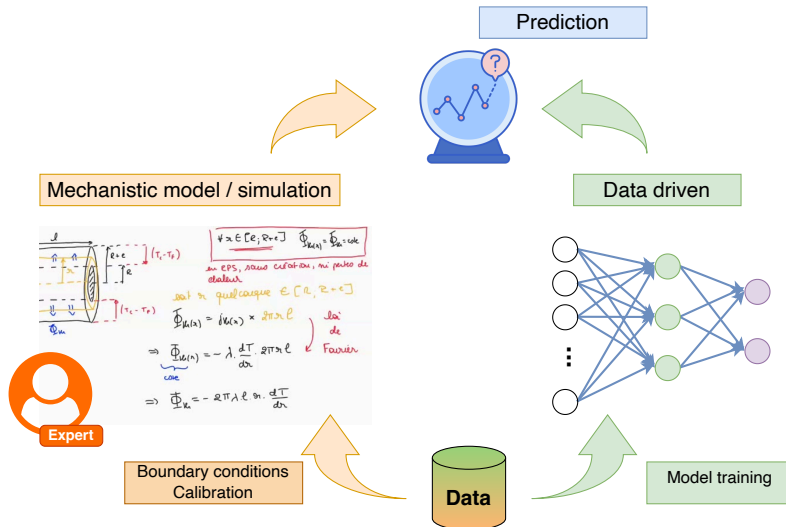


Warning: each arrow = aligned dataset required !



Physics Informed Machine-Learning (PIML / PINNs)

- Historically, two very different approaches
- Huge ambiguity about modeling !





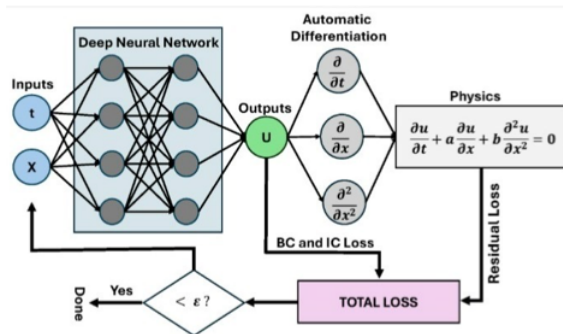
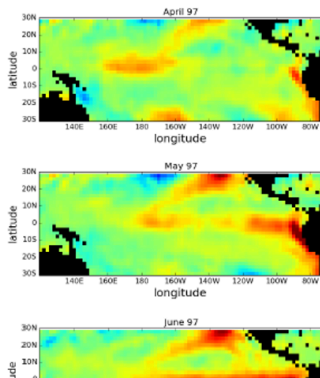
Physics Informed Machine-Learning (PIML / PINNs)

■ Physics Informed Neural Networks

= integrating ODE/PDE inside NN architecture

- e.g. (1) temperature estimation (with NN), (2) heat diffusion (ODE)

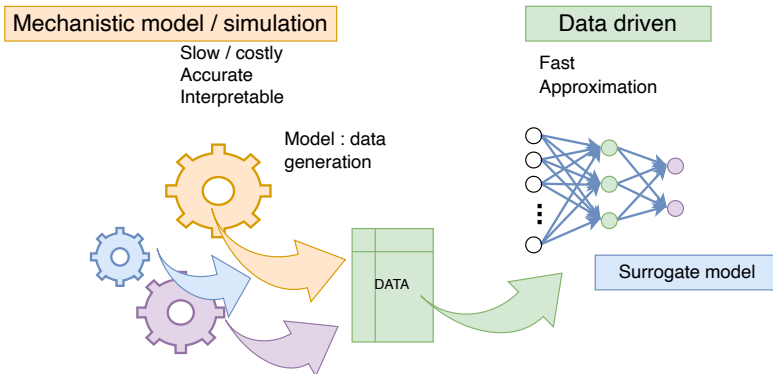
Physical constraints
Differential equation modeling
inside neural architecture





Physics Informed Machine-Learning (PIML / PINNs)

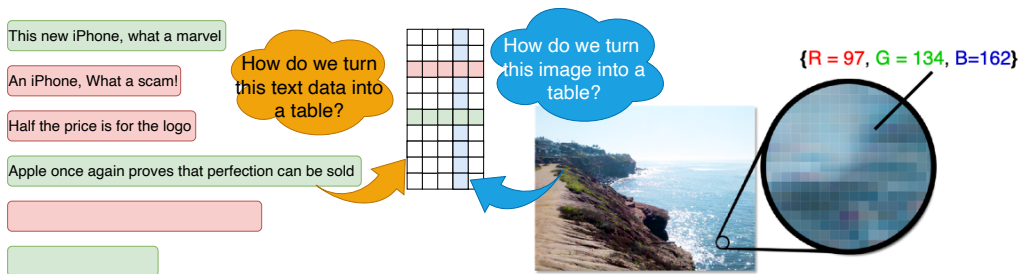
- Surrogate model
- ↗ speed, ↘ input req., new calibration options, ...





Machine-learning Limitations

- 1 Performance ambiguity
 - 90% recognition : great, interesting... or unusable?
 - Performance on the end-user task... Or on a specific domain?
- 2 Range of validity (Climate change, patient change, location shift...)
- 3 Operational integration (UX, ergonomics, training requirement)
- 4 How to represent complex input in a table?

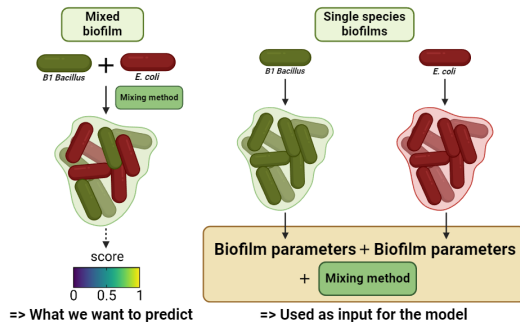


MACHINE-LEARNING
FOR SCIENCES
(& SOMETIMES BIOLOGICAL DATA)



Case study: predicting bacterial antagonism

- **Question:** which *Bacillus* strain will exclude a given pathogen?

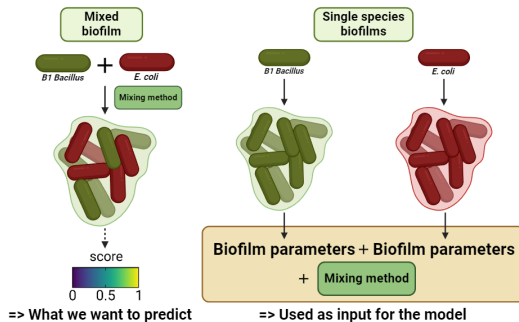


A machine learning approach to predict the antagonism of beneficial strains against microbial pathogens using morphological descriptors of their biofilms, Rubrice, Guéneau, Briandet, Cornuéjols, Guigue



Case study: predicting bacterial antagonism

- **Question:** which *Bacillus* strain will exclude a given pathogen?
- **Inputs:** descriptors of **mono-species** biofilms
volume, height, thickness, roughness, substratum coverage (BiofilmQ)

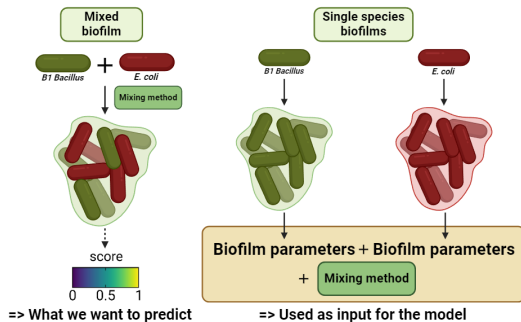


A machine learning approach to predict the antagonism of beneficial strains against microbial pathogens using morphological descriptors of their biofilms, Rubrice, Guéneau, Briandet, Cornuéjols, Guigue



Case study: predicting bacterial antagonism

- **Question:** which *Bacillus* strain will exclude a given pathogen?
- **Inputs:** descriptors of **mono-species** biofilms
volume, height, thickness, roughness, substratum coverage (BiofilmQ)
- **Target:** exclusion score measured on the **mixed** biofilm



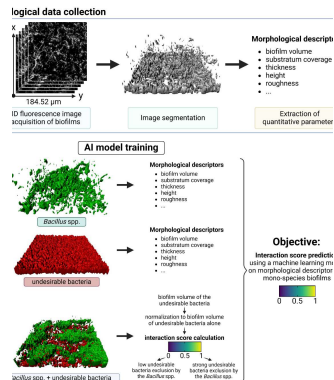
A machine learning approach to predict the antagonism of beneficial strains against microbial pathogens using morphological descriptors of their biofilms, Rubrice, Guéneau, Briandet, Cornuéjols, Guigue



Case study: predicting bacterial antagonism

- **Question:** which *Bacillus* strain will exclude a given pathogen?
- **Inputs:** descriptors of **mono-species** biofilms
volume, height, thickness, roughness, substratum coverage (BiofilmQ)
- **Target:** exclusion score measured on the **mixed** biofilm
- **Data:** 18 *Bacillus* $\times$ 4 pathogens $\times$ 3 mixing protocols $\times$ 3 replicates

⇒ **648** exclusion scores

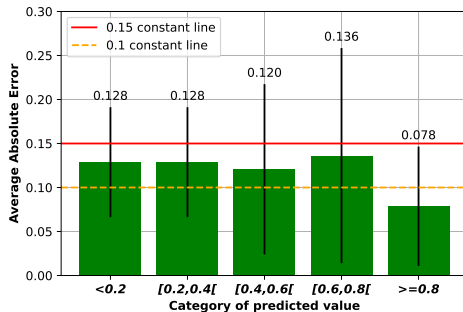


A machine learning approach to predict the antagonism of beneficial strains against microbial pathogens using morphological descriptors of their biofilms. Rubrice, Guéneau, Briandet, Cornuéjols, Guigue



$R^2 = 0.63$: a bad model or a good screening tool?

- A textbook answer: “below 0.8, useless”
- But what is the **decision** being made? rank candidate strains, then test the best ones at the bench
- 81% of predictions within ± 0.2
- When the model predicts a **strong exclusion** (> 0.8): mean error = **0.078**





The real deliverable: a monitored experimental campaign

The model is not the end of the story: it is a prioritization device

- $18 \times 4 \times 3$ combinations = already a heavy campaign... and a **tiny** corner of the real space
- Screening all promising *Bacillus*/pathogen pairs experimentally is out of reach

⇒ **Loop:** predict → test the top candidates → enrich the dataset → retrain

model → ranked candidates → **bench** → new scores → **model**

This is *active learning* / *design of experiments*: the model should also point at where it is the most **uncertain**, not only where it is the most **optimistic**.



The real deliverable: a monitored experimental campaign

What a campaign must produce to be exploitable by ML

+ Systematic design

full (or balanced) combinatorics, not opportunistic sampling

+ Replicates

biological *and* technical: they measure the noise floor $\Rightarrow$ the **ceiling** of any model

+ Controls & references

mono-species biofilms = the inputs of the model

+ Metadata

the mixing protocol was recorded $\Rightarrow$ it turned out to be **the** key feature

— Missing values

imputation degraded performance here

— Hidden batch effects

operator, day, plate, microscope settings: if not logged, they become noise



The real deliverable: a monitored experimental campaign

The entry ticket

- **Computation:** a laptop. XGBoost on 648 points = seconds.
- **Code:** scikit-learn + xgboost, a few hundred lines
- **Training:** a MSc-level internship (with supervision)

⇒ **Data is the bottleneck:** years of confocal acquisition & image analysis

⇒ and the **formalization** of the question: what exactly do we predict, from what?

The expensive part of “classical” ML in a wet lab is neither the GPU nor the algorithm: it is the **experimental campaign** and the **dialogue** between biologists and modellers.



From Wofost to machine-learning

Mechanistic approaches propose this kind of solution...

But **high precision** = **very precise inputs**

- Weather
- Soil data
- Crop management plan
- Phylogenetic/phenological details of the seeds

⇒ We know field plots, yields...

We can retrieve weather & soil data.





From Wofost to machine-learning

Mechanistic approaches propose this kind of solution...

But **high precision** = **very precise inputs**

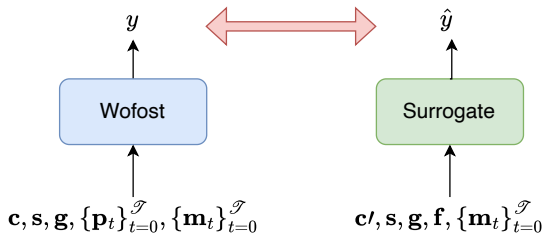
- Weather
- Soil data
- Crop management plan
- Phylogenetic/phenological details of the seeds

⇒ We know field plots, yields...

We can retrieve weather & soil data.



WOFOST



c	Crop-specific features
s	Soil characteristics
g	Initial conditions
$\{p_t\}_{t=0}^T$	Agronomic management practices
$\{m_t\}_{t=0}^T$	Meteorological observations
f	Farm id $\Rightarrow$ profile
c'	Crop id $\Rightarrow$ profile



From Wofost to machine-learning

Mechanistic approaches propose this kind of solution...

But **high precision** = **very precise inputs**

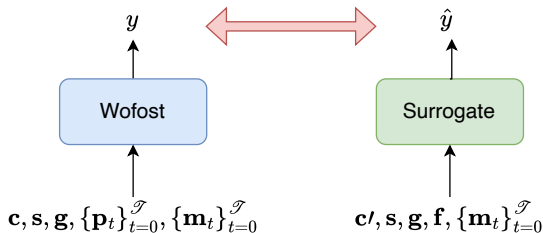
- Weather
- Soil data
- Crop management plan
- Phylogenetic/phenological details of the seeds

⇒ We know field plots, yields...

We can retrieve weather & soil data.

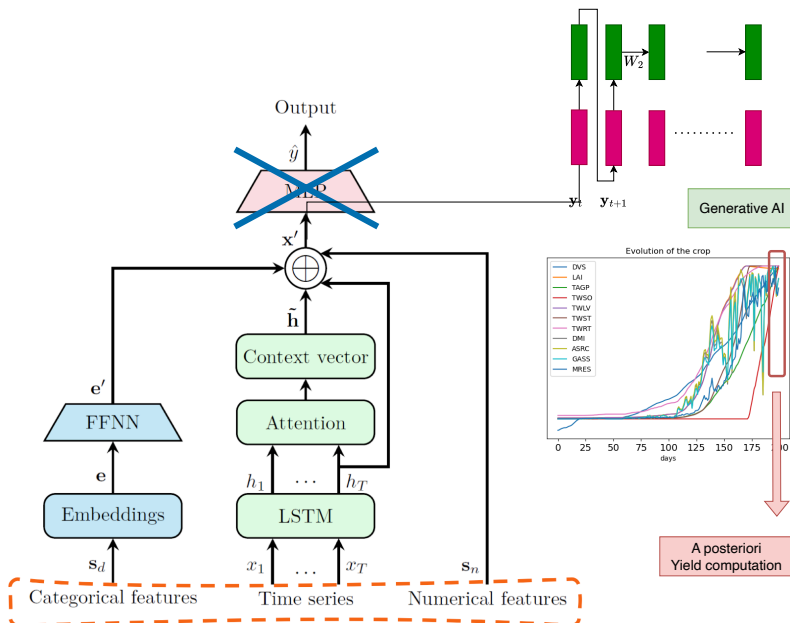


WOFOST

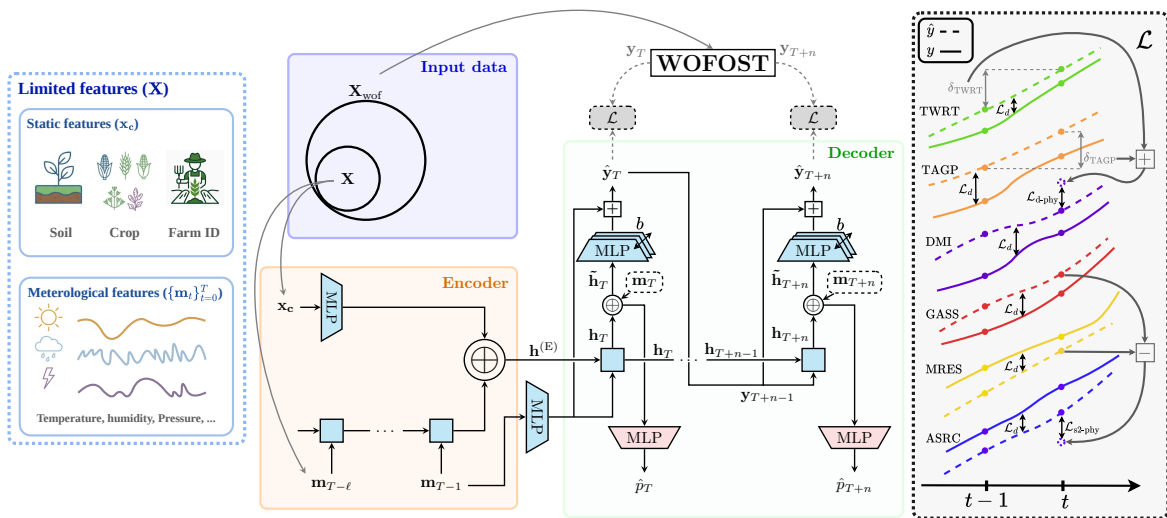


⇒ Surrogate = less input requirements, differentiable, easier to calibrate, faster

Sequence to sequence: generating the crop growth process

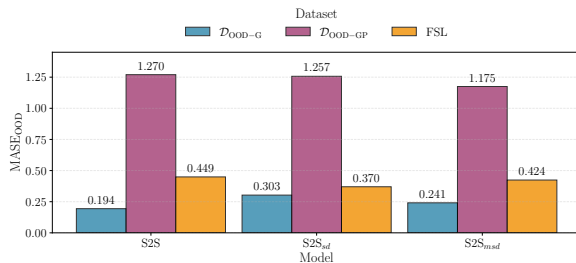


Sequence to sequence: generating the crop growth process



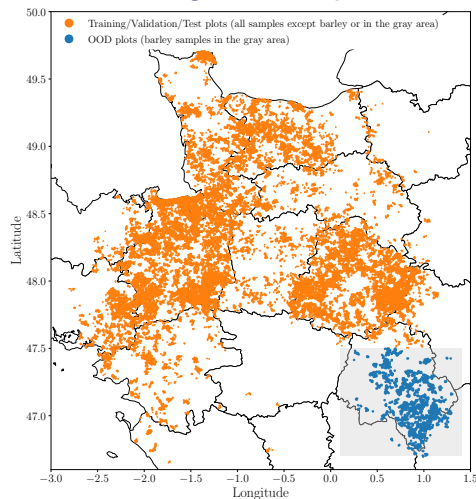


Out-Of-Domain Evaluation



- OOD-G: Geographic
⇒ OK (easy setting)
- OOD-GP: Geographic + species
⇒ KO
- OOD-GP + physical constraints
⇒ keep good performances

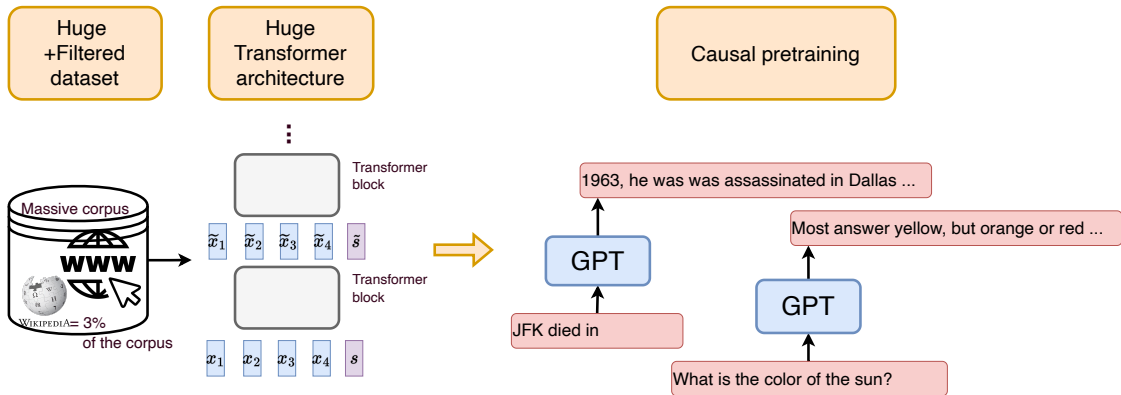
What happens in a new context?
Across new regions and species?





From next token to new tasks: chatGPT

0. Transformer + massive data (GPT)



- Grammatical skills: singular/plural agreement, tense concordance
- Knowledge: entities, names, dates, places



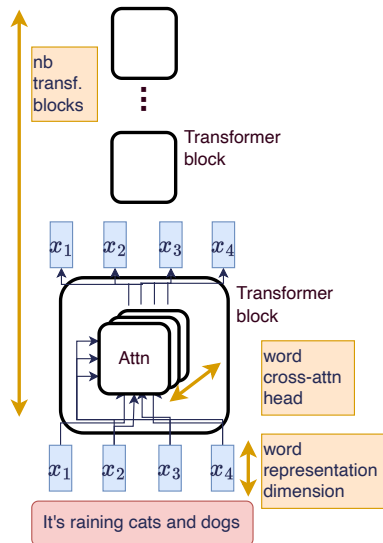
The Ingredients of chatGPT

1. More is better! (GPT)

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

175 Billion parameters... What does it mean?

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

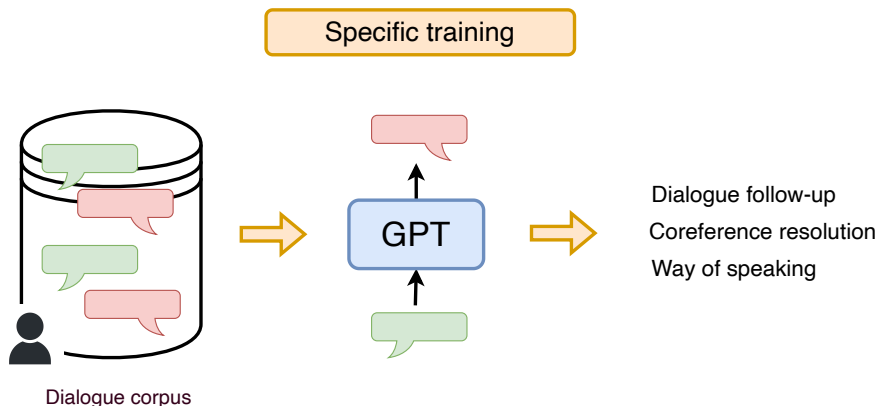


It's raining cats and dogs



The Ingredients of chatGPT

2. Dialogue Tracking



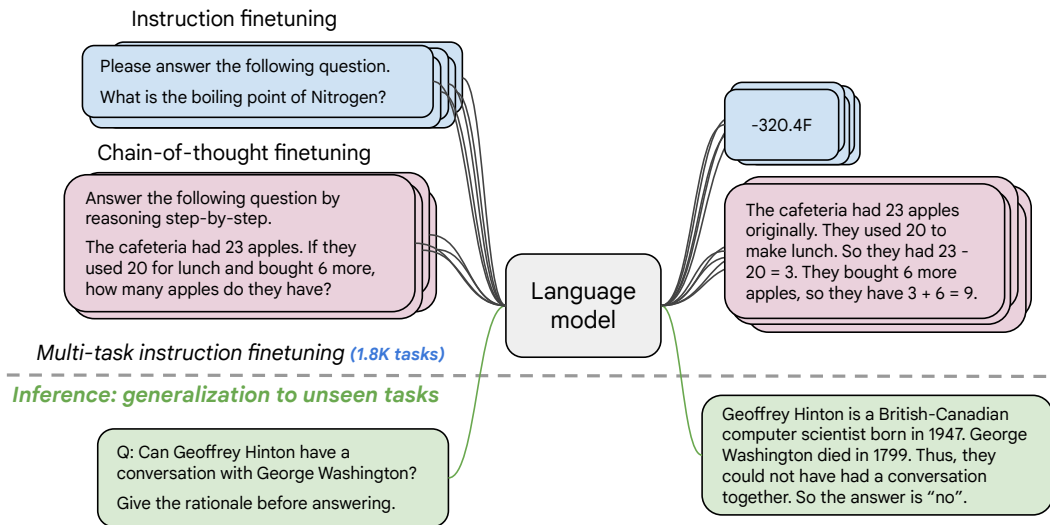
■ **Very clean data**

Data generated/validated/ranked by humans



The Ingredients of chatGPT

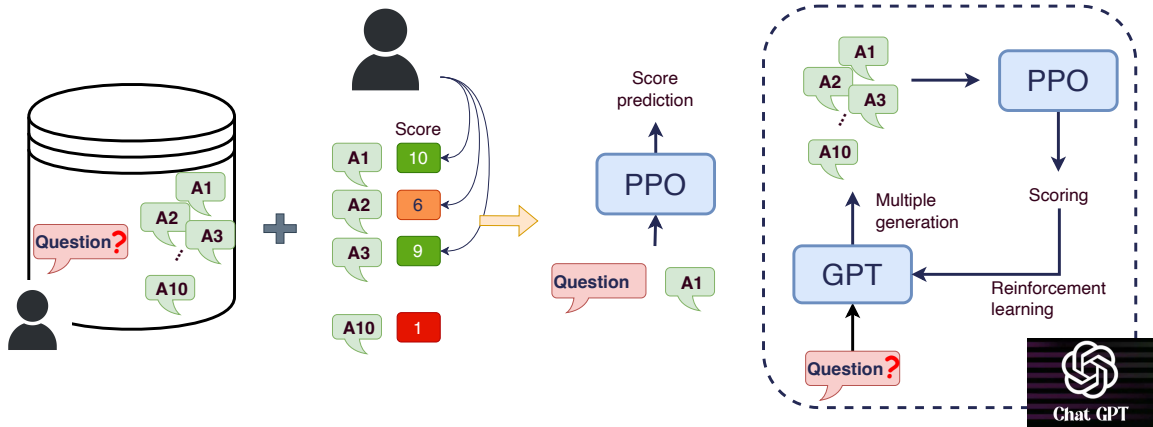
3. Fine-tuning on different ($\pm$) complex reasoning tasks





The Ingredients of chatGPT

4. Instructions + answer ranking



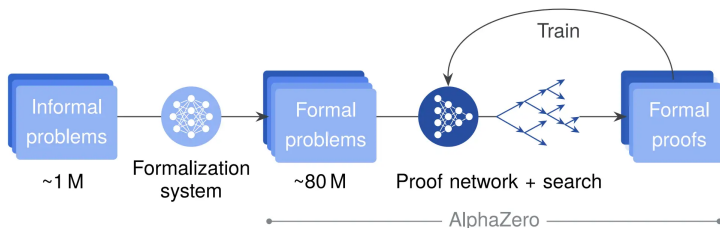
- Database created by humans
- Response improvement

- ... Also a way to avoid critical topics = censorship



What about Mathematical proof?

- 1 Still SFT (=supervised fine-tuning)... = **next token prediction** on the proof
- 2 Does the proof come the right result at the end? $\Rightarrow$ GRPO (DeepSeek loss)
reward good agent... And more reward to the original one!
- 3 Test Time Reinforcement Learning (AlphaProof):
exploring various branches... And select the good ones at test time



Lean = formal proof language

The key for automated evaluation = reliable reward = efficient training strategy



Biological sequences: yet another language

1. English



"THE QUICK BROWN FOX
JUMPS OVER THE LAZY DOG"

A
Aback
Abandon
...

- BioMedLM
- PubMedBERT
- BioBERT

2. DNA



AGGACTGGACCT



- DNABERT
- DNABERT-2
- DNABERT-S

3. Protein



FYERATIMKHWE



- ESM
- ProteinBERT
- ProtMamba

4. Genome



dnaA
rpoB
recA
...

- Nucleotide Transformer
- Evo
- Evo2



Biological sequences: yet another language

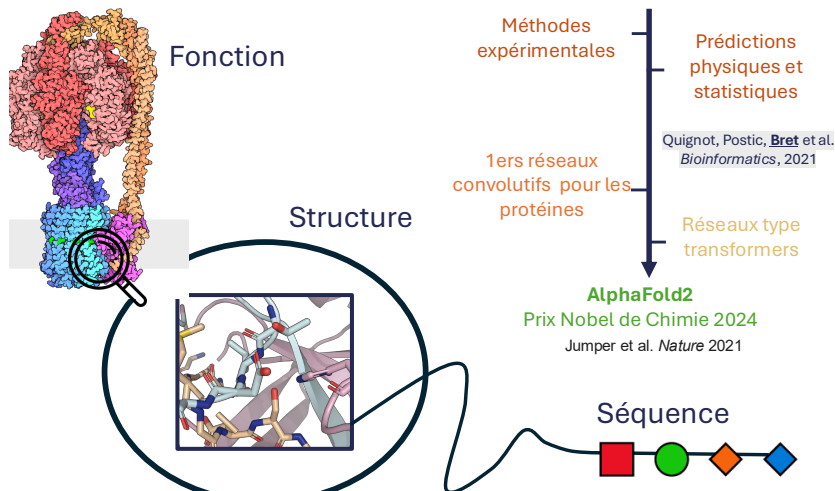
- Same recipe as NLP:
 - an **alphabet**: 4 nucleotides, 20 amino-acids, genes...
 - a **tokenizer**: k -mers, BPE
 - **self-supervised pre-training** on huge unlabelled corpora (UniProt, GenBank...)
- ⇒ a **representation** of each residue *in its context*
- Fine-tuned downstream for: function annotation, promoter & gene detection, variant effect prediction, sub-cellular localization, **resistance gene** identification...
- ⇒ the sequence is the text, the biology is the semantics

What genomics can learn from Natural Language Processing?, [Rannon et al., 2025](#) —

<https://doi.org/10.48550/arXiv.2506.02212>



From sequence to structure: AlphaFold2



Three levels of description — and 40 years of methods to bridge them



From prediction to generation: designing sequences

Generation is only useful inside an iterative loop

AI prediction → *a reduced number of mutants to test* →
experimental validation → *an enriched dataset* → **AI prediction**

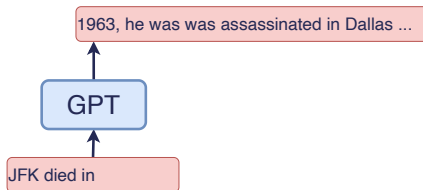
- The model does not replace the bench: it **shrinks the search space**
 - Exactly the same loop as the biofilm screening case
- ⇒ **the value is in the coupling**, not in either half

TECHNICAL LIMITS OF MACHINE-LEARNING



chatGPT and the relationship with truth

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



Example: producing a bibliography

Can you give me a short bibliography on VAE for Time Series?

recurrents.

2. "Variational Recurrent Autoencoders" par Chung et al. (2016) - Cette étude introduit une approche VAE pour la modélisation de séquences temporelles en utilisant des réseaux de neurones récurrents et une méthode de maximisation de la vraisemblance pour la phase d'entraînement.

Variational recurrent auto-encoders

3. "Generative Modeling for Time Series" par Bao et al. (2017) - Cette étude propose un modèle qui combine les forces des RNNs et des VAEs pour la modélisation de séries temporelles profondes, y compris les VAEs.

[Q.Fabius, J.R. Van Amersfoort - arXiv preprint arXiv:1412.6581, 2014 - arxiv.org](#)

In this paper we propose a model that combines the strengths of RNNs and SGVB: the **Variational Recurrent Auto-Encoder (VRAE)**. Such a model can be used for efficient, large scale ...

☆ Enregistrer 95 Citer 116 302 fois Autres articles Les 2 versions 36

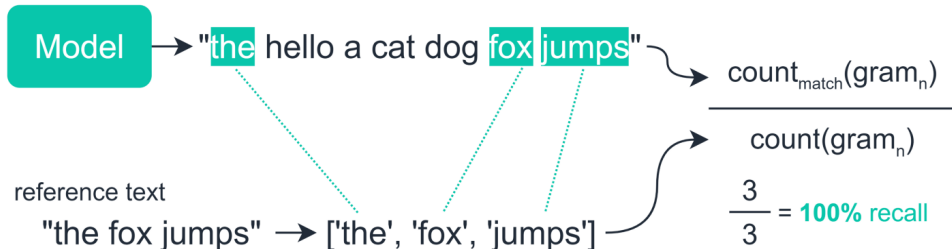
4. "Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Data" par Kulkarni et al. (2017) - Cette étude présente une approche VAE pour la



Generative AI: how to evaluate performance?

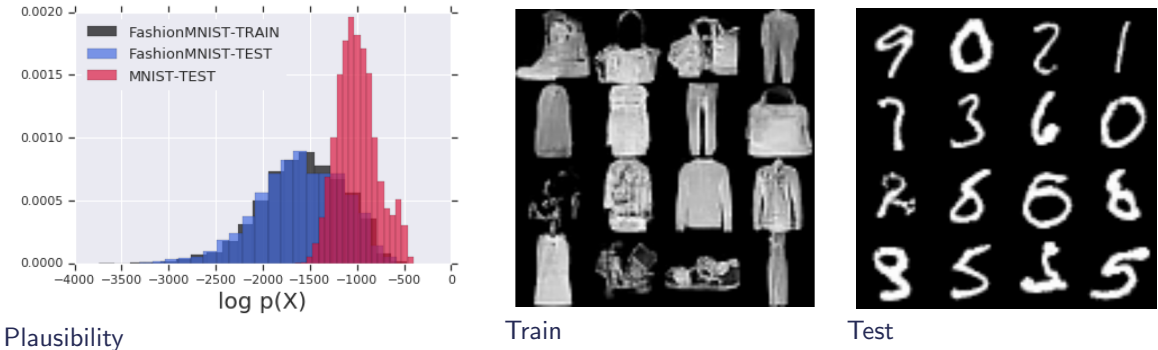
The critical point today

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



The critical point today

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





Stability/predictability

- Difficult to bound a behavior
 - Impossible to predict good/bad answers
- ⇒ Models that regularly discredit themselves
- Little/no use in video games
- ⇒ Impossible to certify these models for critical applications



how old is Obama



Barack Obama was born on August 4, 1961, making him 61 years old as of February 2, 2023.





Stability/predictability

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

⇒ Models that regularly discredit themselves

Little/no use in video games

⇒ Impossible to certify these models for critical applications



V

how old is obama?



As of 2021, Barack Obama was born on August 4, 1961, so he is 60 years old.



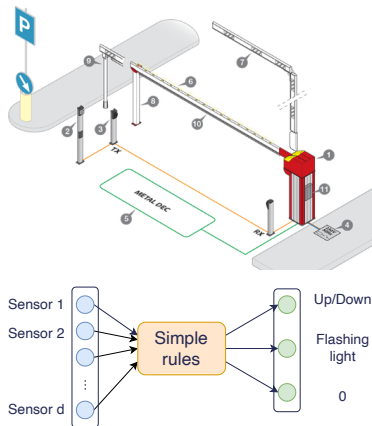
V

and today?

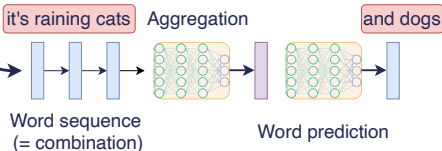




Stability, explainability... And complexity



- 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?
- What training data was used?
- What editorial stance / censorship is involved?
- Why this answer?

Adaptation

Data contamination / skills

Access to information

Explainability / interpretability

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

Source: 2023 Foundation Model Transparency Index

	Meta	BigScience	OpenAI	stability.ai	Google	ANTHROPIC	cohere	AI21labs	Inflection	amazon	
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	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%	



size = no. of parameters open-access

* = parameters undisclosed // see the data

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

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 :

 0.30g
CO₂ émis

 5min
ampoule LED

 33s
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

=

 6Wh
énergie conso.

Ce qui correspond à :

 6g
CO₂ émis

 2h
ampoule LED

 12min
vidéos en ligne

Voir plus



Different behaviors, different costs

Different costs for different users/languages

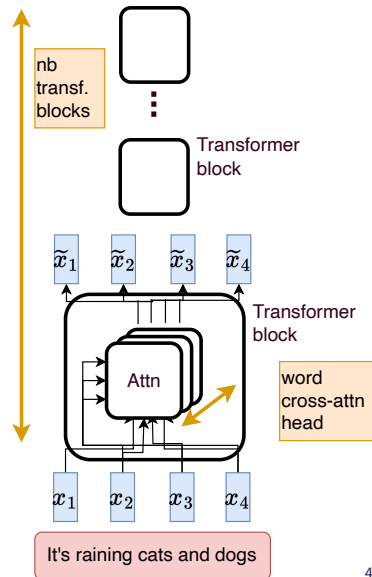
Pour un texte significatif en Français

and the same in English

TOKENS CHARACTERS
17 63

<s> Pour un texte significatif en Français

and the same in English





Different behaviors, different costs

Different costs for different users/languages

The Tokenizer Playground

Experiment with different tokenizers (running locally in your browser).

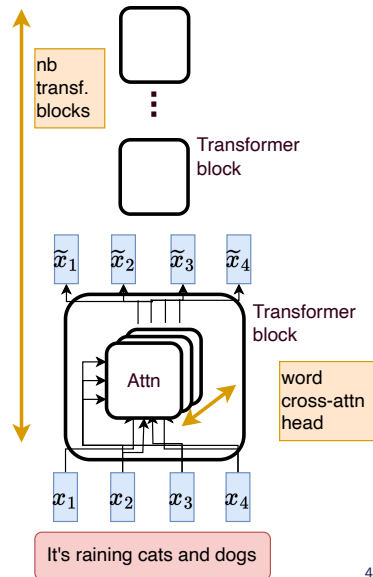
gpt-4 / gpt-3.5-turbo / text-embedding-ada-002 ▾

124578 * 963

TOKENS CHARACTERS
5 12

124578 * 963

☒ Text ☐ Token IDs ☐ Hide





Everything beyond the LLM's capabilities/training

- Simple calculations
(multiplication, division)
- Generating n -syllable animal names
(in progress)
- Playing chess
- Following (complex) causal reasoning
- ...

ATARI 2600 SCORES STUNNING VICTORY OVER CHATGPT



**WHEN YOU UNDERESTIMATE A 1977 CHESS ENGINE...
AND IT HUMBLER YOU IN FRONT OF THE WHOLE INTERNET**

CHATGPT

NOVEMBER 30, 2022

1 MILLION USERS IN 5 DAYS

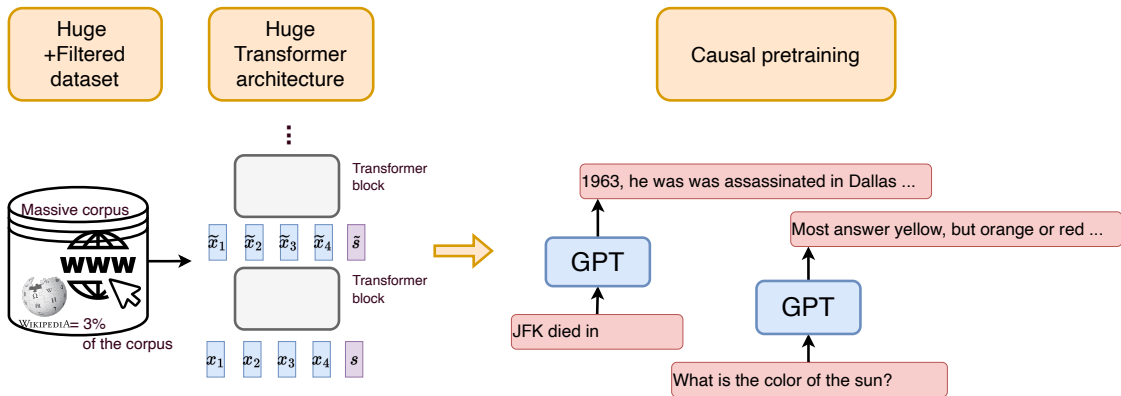
100 MILLION AT THE END OF JANUARY 2023

1.16 BILLION IN MARCH 2023



The Ingredients of chatGPT

0. Transformer + massive data (GPT)



- Grammatical skills: singular/plural agreement, tense concordance
- Knowledge: entities, names, dates, places



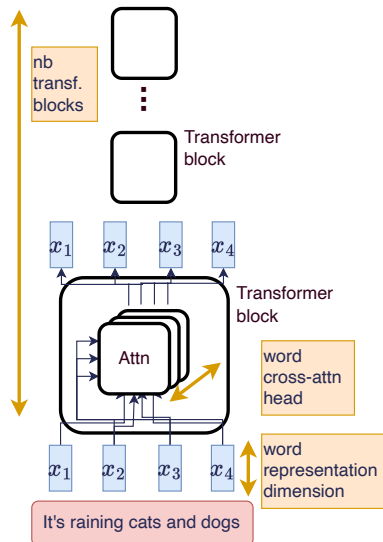
The Ingredients of chatGPT

1. More is better! (GPT)

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

175 Billion parameters... What does it mean?

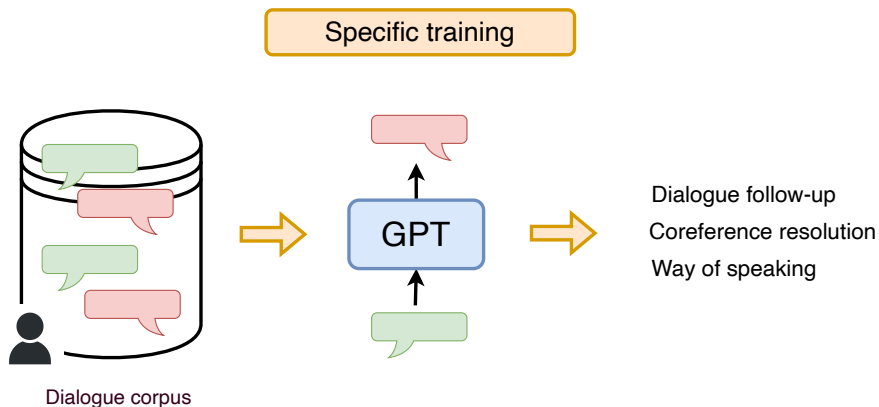
- $1.75 \cdot 10^{11} \Rightarrow 300 \text{ GB} + 100 \text{ GB}$ (data storage for inference) $\approx 400\text{GB}$
- NVIDIA A100 GPU = 80GB of memory (=20k€)
- Cost for (1) training: 4.6 Million €





The Ingredients of chatGPT

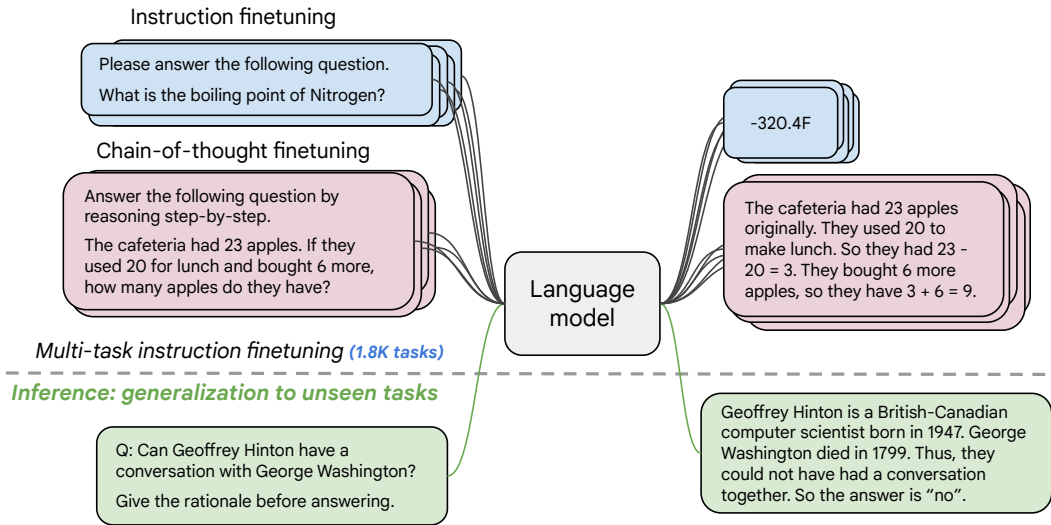
2. Dialogue Tracking



■ **Very clean data**

Data generated/validated/ranked by humans

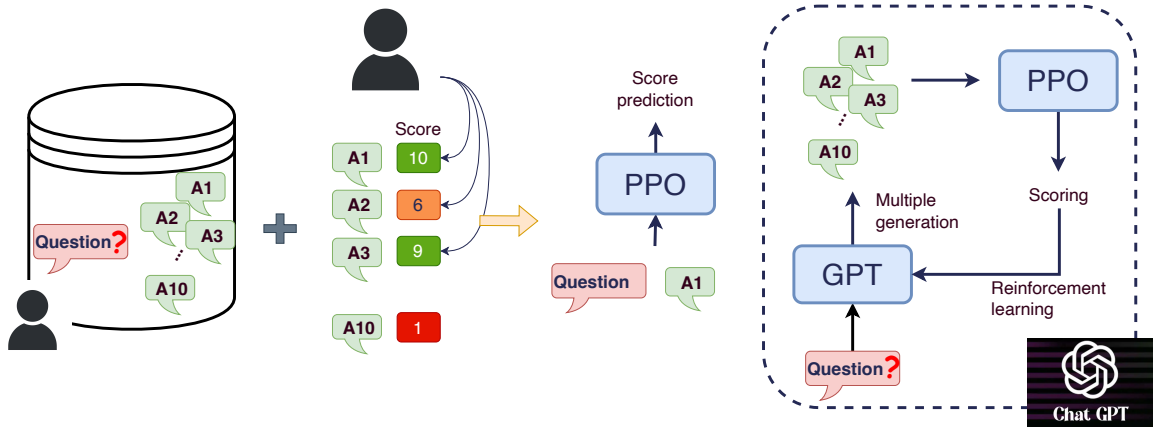
3. Fine-tuning on different ($\pm$) complex reasoning tasks





The Ingredients of chatGPT

4. Instructions + answer ranking



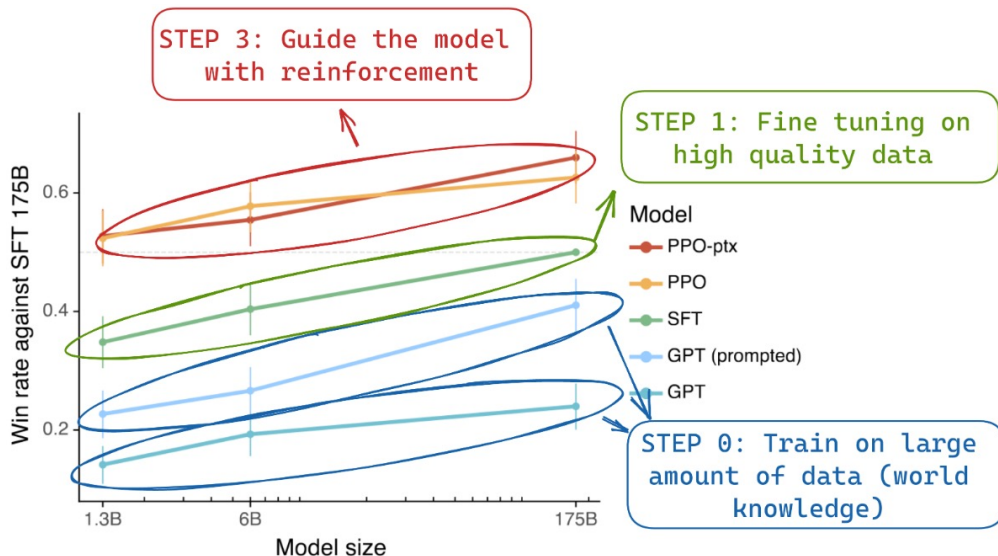
- Database created by humans
- Response improvement

- ... Also a way to avoid critical topics = censorship



Steps & Performance

Massive data $\Rightarrow$ HQ data (dialogue) $\Rightarrow$ Tasks $\Rightarrow$ RLHF

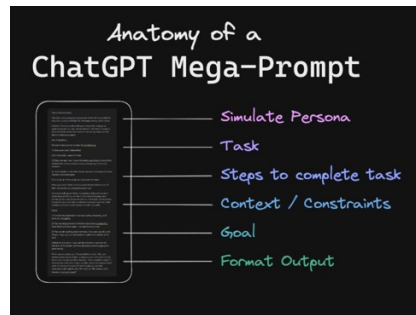




Usage of chatGPT & Prompting

- 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 (!)

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



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



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 ...*

- **Prompt = specification document:** long, detailed, designed to remove ambiguity about intended directions.

- Prompting requires

domain expertise >> general prompting knowledge.



- Don't

- Rephrase

- Add mistakes (!)

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

Anatomy of a

pt

sona

plete task
constraints

t

-chatgpt-prompt/



Language Handling

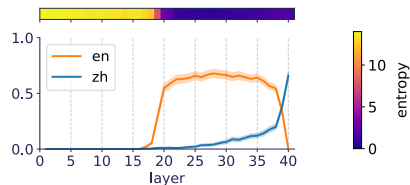
■ Language models are multilingual today:

⇒ Stay in the language you are comfortable with

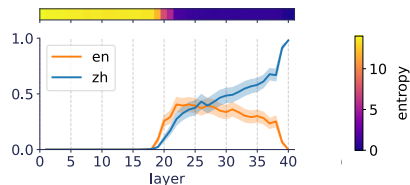
⇒ Ask for answers in any language

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

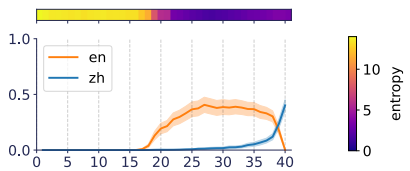
(a) Translation task



(b) Repetition task



(c) Cloze task

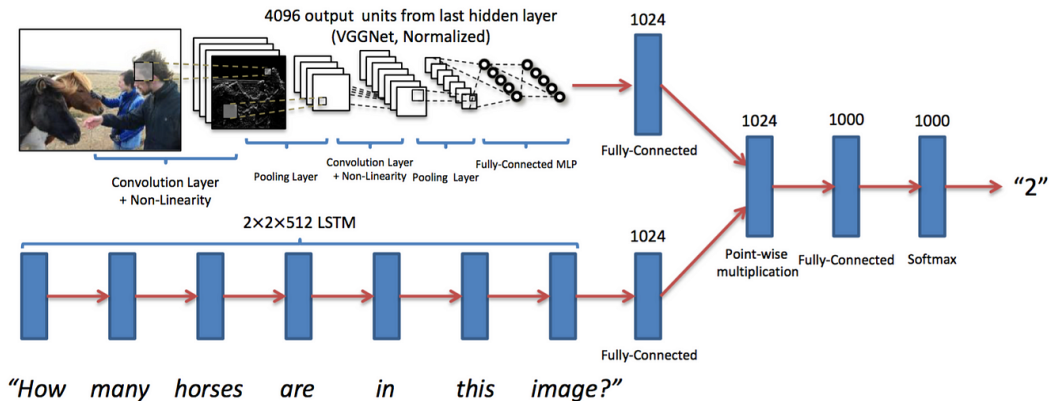




GPT4 & Multimodality

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

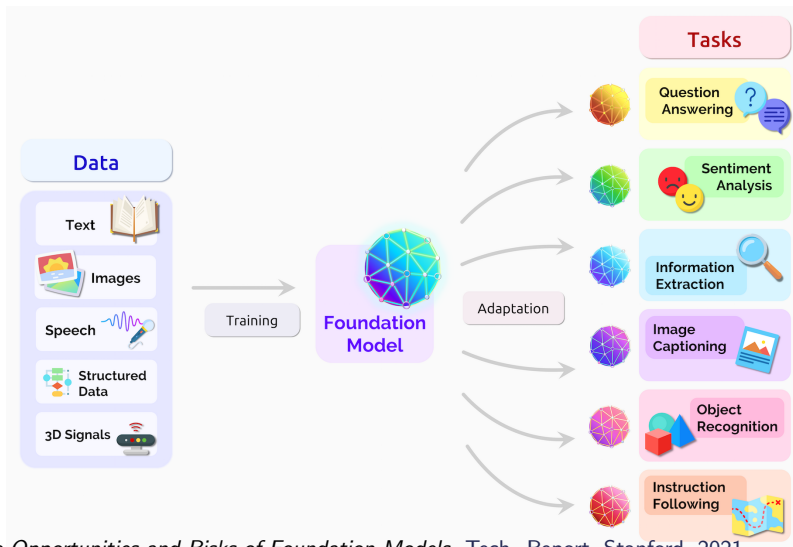


VQA: Visual Question Answering, arXiv, 2016, A. Agrawal et al.



Towards Larger Foundation Models?

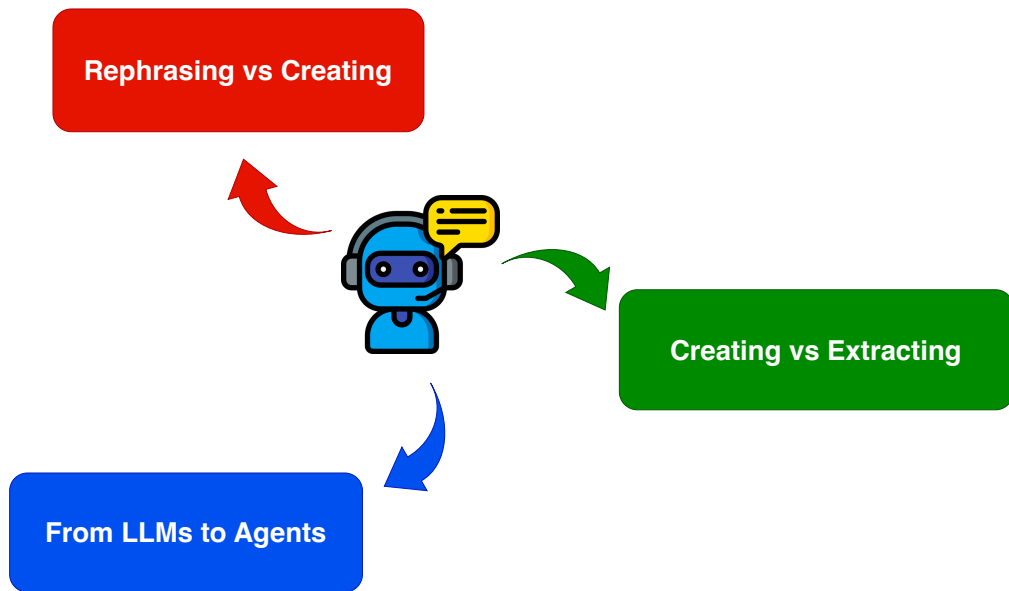
- Let the modalities enrich each other



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

LLM TYPOLOGY OF USES

Three questions to frame the use cases



Rephrasing *I bring the content*

- **Translating/Proofreading** a paragraph
- **Rephrasing**: style, tone, 10% shorter
- A paragraph from **my** bullet points
- Meeting minutes from **my** notes
- Cover letter from a CV + job offer
- Table formatting, OCR, $\text{\LaTeX}$ /Excel
- The abstract of **my own** paper

Creating *the model brings the content*

- **Finding** inspiration [blank page]
- **Suggesting** a plan, an angle, an acronym
- **Answering** a question I cannot answer
- **Testing** an idea, listing weak points
- Writing a paragraph on a topic I give
but do not detail
- Writing a review, a related-work section



Rephrasing: examples

Building a recommendation letter

Prompt

[Tâche]
Etudiant rencontré...
qualités ...
résultats marquant



LLM



Meeting minutes



Transcription

Résumé/CR





Ethical questions

Rephrasing

The content is **mine**. The model touches the *form*.

→ Am I still the author? **yes**

→ Do I have to disclose it?
like a spell-checker? a translator?

→ Real risk = **silent distortion**
a nuance lost, a hedge removed

→ Real risk = **confidentiality**
unpublished text sent to a server

⇒ Same tool, same prompt box, **two completely different debates.**

Creating

The content comes from **the model**, *i.e.* from a corpus I did not choose.

→ Am I still ~~the author~~ responsible? **yes**

→ Who do I cite?
the model has no source

→ Real risk = **hallucination**
plausible, well written, false

→ Real risk = **uniformity of thought**
the most likely idea, not the best one



(2) Creating *ex-nihilo* vs extracting information

Ex-nihilo

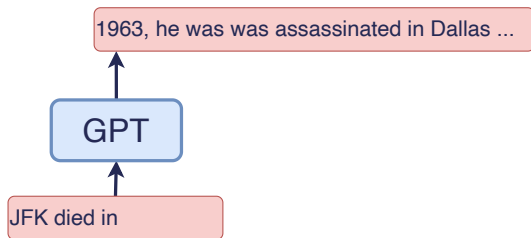
parametric memory

Extraction

the document is given

- The answer is **in the weights**
- *JFK died in...* ⇒ probabilistic answer
- No source, no traceability
- Frozen at training time

⇒ **hallucination** (*factuality* measurement)



- The answer is **in the input**
 - The model *locates*, it does not invent
 - The source is known *by construction*
 - Up to date = *as fresh as the document*
- ⇒ **verifiable** (*faithfulness* measurement)

Passage Sentence

In meteorology, **precipitation** is any product of the condensation of atmospheric water vapor that **falls** under **gravity**.

Question

What causes **precipitation** to **fall**?

Answer Candidate

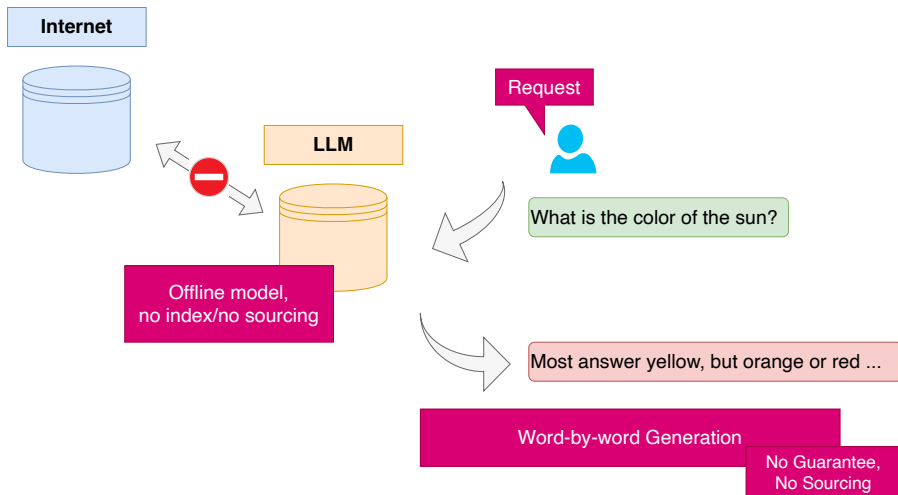
gravity

⇒ Same question, same interface... but only one of them can be checked.



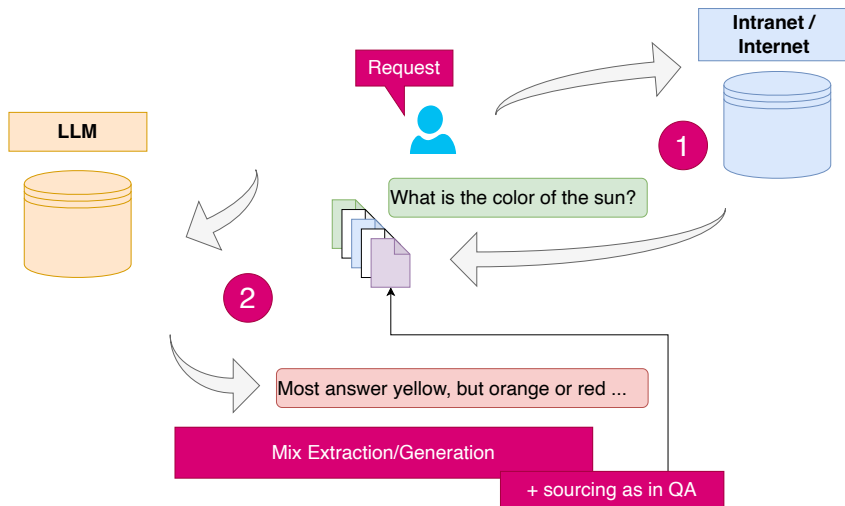
LLMs $\Rightarrow$ RAG: memory vs information extraction

- Asking ChatGPT questions... A surprising use case!
- But is it reasonable? [A true open question (!)]





LLMs $\Rightarrow$ RAG: memory vs information extraction



- RAG: Retrieval-Augmented Generation
- Current limit on input size (2k, 32k, 200k tokens)



In a scientific context

■ RAG in a scientific context

should be called **bibliography**!

■ Consensus, Scopus.ai, opscidia, ...

⇒ great summary + analysis... On which articles?

■ NotebookLM: choose your articles (up to 50)...

Then start the discussion

⇒ analysis, topic clustering, comparison, ...

■ If you want to dialogue with a manuscript...

It is too long for an LLM

⇒ RAG again!

Can I be a (good) scientist without AI watch?
Is my job at risk?



Consensus

Scopus[®] AI

Change the way you view knowledge

OPSCI DIA

The value of knowledge

NotebookLM

Think Smarter, Not Harder

Try NotebookLM



(3) From LLMs to agents

LLM

a tool I hold

Agent

a system that acts

- One question → one answer
- I read, I judge, I keep or discard
- The model produces **text**
- No effect on the world
- Efficient on small tasks

⇒ human validation **at every step**

⇒ The error rate did not change.

What changed is that **errors now compose**, and nobody reads the intermediate steps.

- One goal → a **chain** of actions
- The model calls **tools**, loops, retries
- The model produces **actions**
queries, files, emails, code run
- Effects on the world

⇒ human validation **at the end**, if at all



Toolformer [2023]: When the LLM calls upon tools

The LLM:

- 1 Identifies its own weaknesses
- 2 Calls tools/APIs to provide better answers

⇒ Controlled data sources (SQL, Wikipedia) = RAG++; Calculator; Translator; Specialized compute engine

LLM at the heart of the system

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

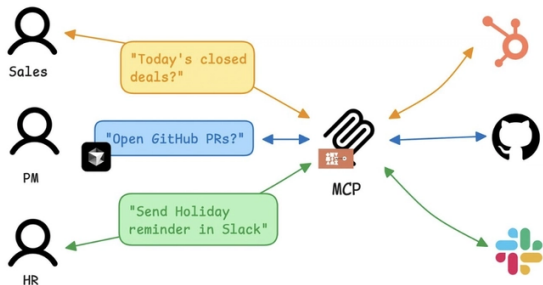
Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.



MCP Protocol



Model Context Protocol

© INSTRUCTA.AI

- Interfacing with LLMs (remote or local)
- Link with applications (emails, websites, messaging, calendar)

⇒ LLM Interfacing = the technical foundation for agents... and many other applications

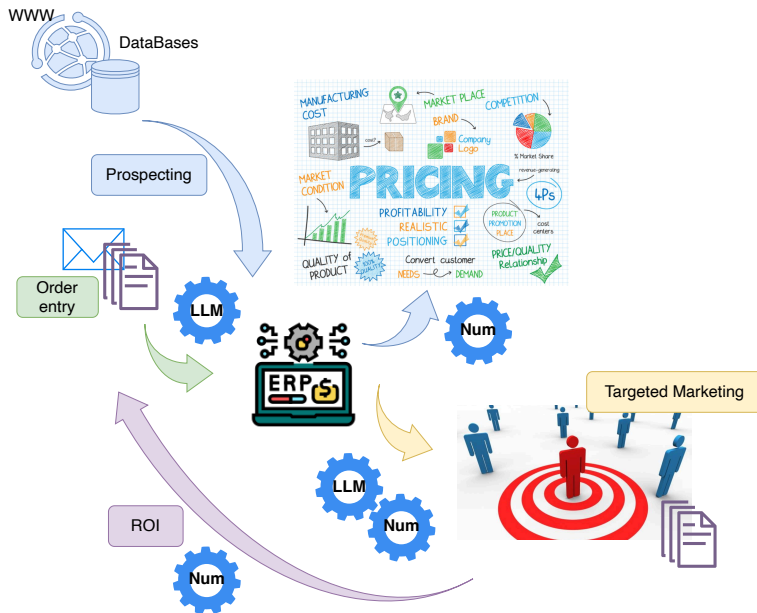
MCP = using an API in natural language

Openclaw: wrap many computer functions in an LLM callable API (mail, browser, console, ...)





Integration into a Complex System

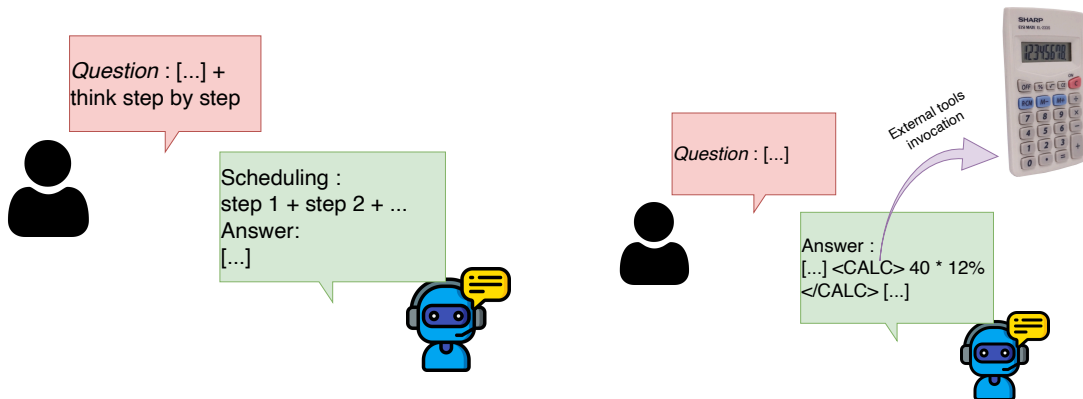


- Prospect list
- Email processing, needs identification
- ERP interaction ⇒ proposal databases
- Pricing
- Drafting a commercial proposal



From *Chain-of-Thought* to Actions

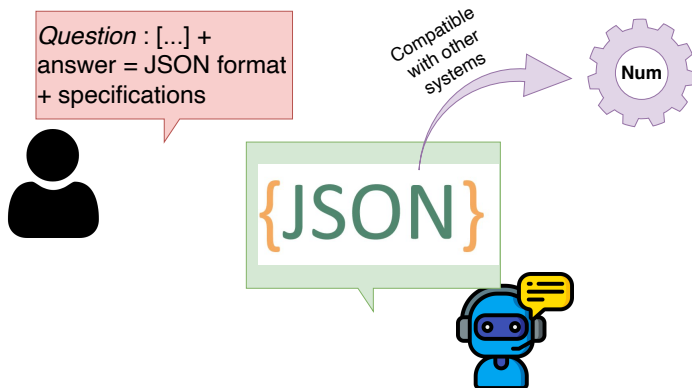
1 Chain-of-Thought / Toolformer





From *Chain-of-Thought* to Actions

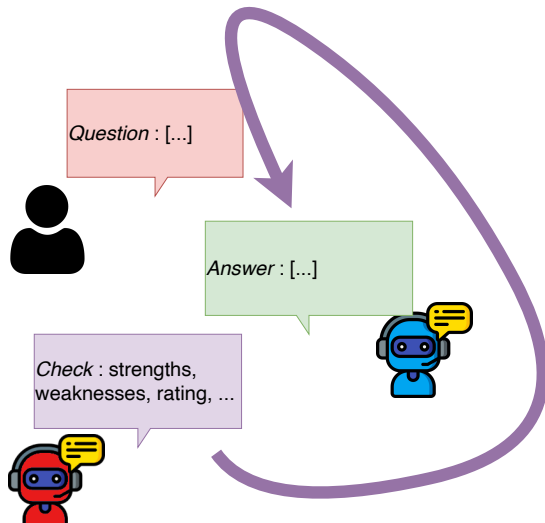
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information





From *Chain-of-Thought* to Actions

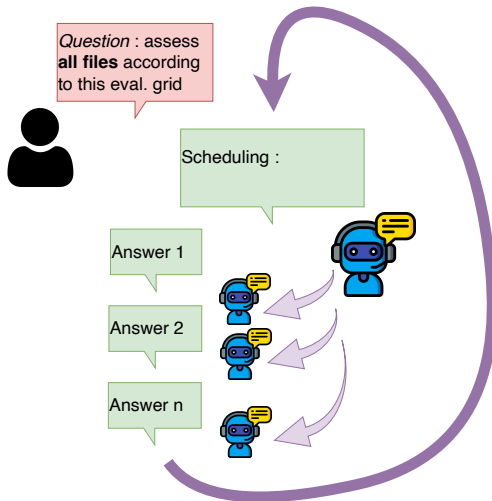
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information
- 3 LLM as a judge / reflexivity + improvement loop





From *Chain-of-Thought* to Actions

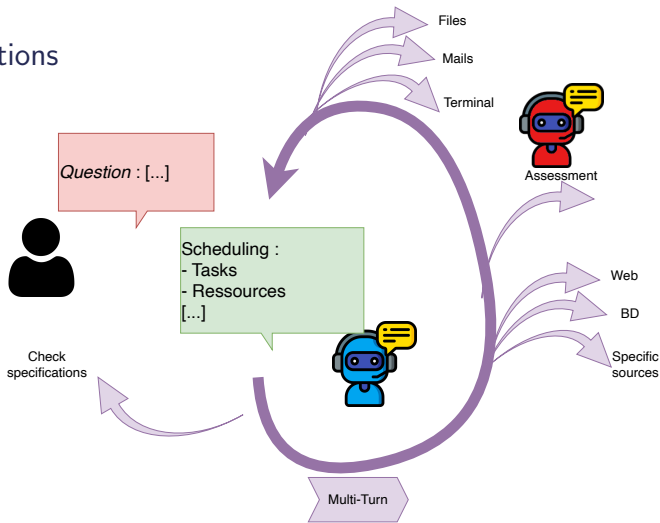
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information
- 3 LLM as a judge / reflexivity + improvement loop
- 4 Parallelism





From *Chain-of-Thought* to Actions

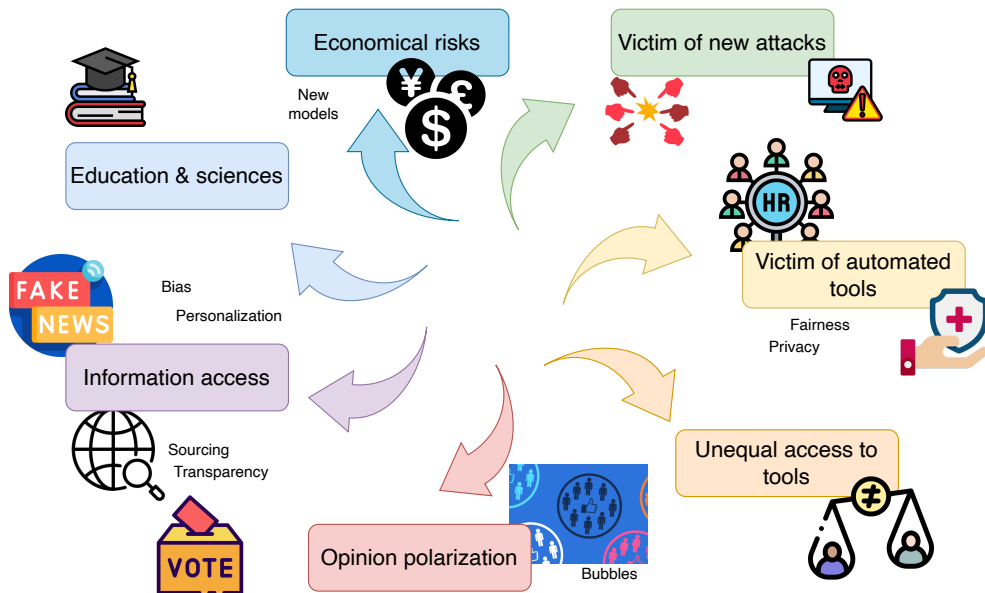
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information
- 3 LLM as a judge / reflexivity + improvement loop
- 4 Parallelism
- 5 MCP / actions



THE RISKS ASSOCIATED
WITH THESE USES

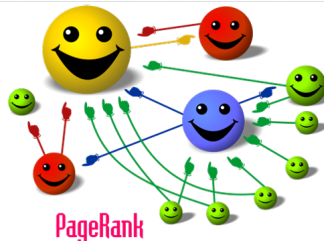


Risk topology



(1) Access to Information

- Access to dangerous/forbidden information
 - +Personal data
 - Right to be forgotten (GDPR)
 - Efficiently deal with huge amounts of personal data
 - Information authorities
 - **Nature:** unconsciously, image = truth
 - **Source:** newspapers, social media, ...
 - **Volume:** number of variants, citations (pagerank)
- ⇒ A huge risk for democracy





Profiling / personalization to information access



- A good way to face **Information overload**
 - Personal filter / agent
 - Proactive system: watch agent + notifications
- **Information bubbles** = cluster in social networks
 - Uniform point of view (tunnel vision)
 - Consensus illusion
 - Notion of *Opinion leader*
- **Group polarization**
 - emotional reinforcement
 - Confirmation bias ↗ ↗ ↗
 - Disinformation rapid diffusion

[Cambridge Analytica]



Machine Learning & Bias



Mustache, Triangular Ears, Fur
Texture

Cat



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

Senior Executive

Bias in the data $\Rightarrow$ bias in the responses

Machine learning is based on extracting statistical biases...

$\Rightarrow$ Fighting bias = manually adjusting the algorithm



Machine Learning & Bias



Stereotypes from *Pleated Jeans*

Google Traduction



✎ Texte

🖼 Images

📄 Documents

🌐 Sites Web

Détection la langue

Anglais

Français



Français

Anglais

Arabe



The nurse and the doctor



L'infirmière et le médecin



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

Bias in the data $\Rightarrow$ bias in the responses

Machine learning is based on extracting statistical biases...

$\Rightarrow$ 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, ...

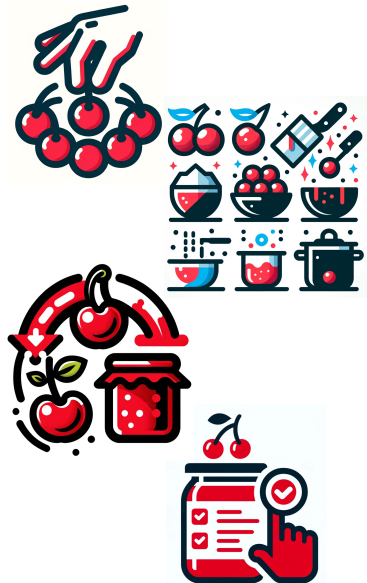
⇒ Choices that influence algorithm results

AI reinforces biases

This software developer position is held by a person who is ...

Answer = **male** in 80% of the training examples,

⇒ AI system end up answering **male** 100% (max selection)



[illegible]

Cognitive Bias	Effect on the User	Generative AI	AI & Med. Diagnosis
Automation bias	Passive acceptance of results (reduced verification)	High	Very high
Authority bias	Overestimation of AI authority (and reliability)	Very high	High
Overconfidence / Miscalibration of trust	Inappropriate use of AI in situations where it is unreliable	Very high	Moderate
Confirmation bias	Reinforcement of erroneous hypotheses and reduced exploration of alternatives	Moderate	High
Cognitive offloading	Decrease in critical analysis and metacognition	Very high	Very high
Framing effect. Illusion of explanatory depth	Biased interpretation of the level of certainty or risk (response style, presentation of numbers)	Very high	Moderate
Anthropomorphism bias	Attribution of human capabilities (increased trust)	Very high	Moderate



Personal data & the right to be forgotten

- GDPR art. 17: **right to erasure**
designed for databases: one row, one delete
- But an LLM is **not a database**
 - the information is **diluted** over 10^{11} weights
 - *machine unlearning* = open research problem
 - retraining from scratch = several M€
- What was scraped is **already in the weights**
 - a forum post from 2009, a leaked file, an old article
 - *membership inference* / training-data extraction
- Who is the **data controller**?
the scraper? the trainer? the deployer?



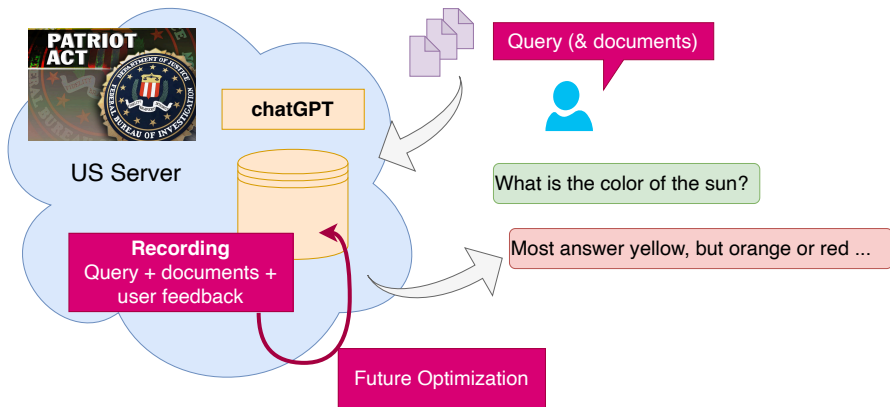
⇒ On a RAG, deletion works (remove the document).

⇒ On parametric memory, it does not.

Same question, two architectures.



(2) Data Leak(s): different security levels

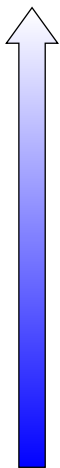


- Transfer of sensitive data
- Exploitation of data by OpenAI (or others)
- Data leakage in future models



(2) Data Leak(s): different security levels

Tools



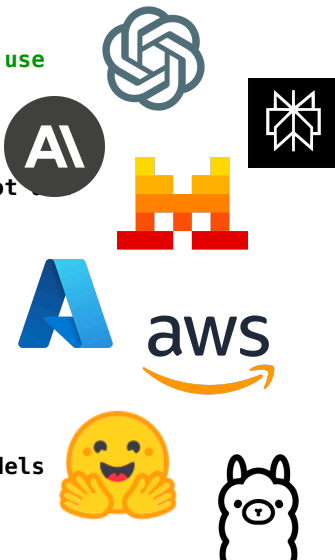
Commercial tools, **free to use**
Variable licence

Commercial tools,
Paid licence
more guaranties vs patriot

Commercial tools,
Paid licence + option
e.g. **European servers**

Institutional LLMs
deployed within a
controlled perimeter

Local use
pre-trained/finetuned models



Data



Any document



Personal
information



Ongoing
project



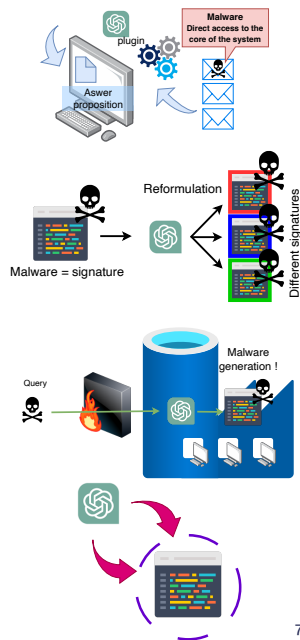
Medical
records





Security Issues: Attacking computers

- Plug-ins / openclaw / MCP
 - ⇒ 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
 - $\approx$ software rephrasing
- New problems!
 - Direct malware generation
- Multiplication of security breach discoveries
 - Claude Code myθος, ...





Attacking people: impersonation & phishing

■ Identity falsification

- Voice/text impersonation
- Fake authority impersonation (boss, teacher, lawyer, doctor)

■ Social engineering at scale

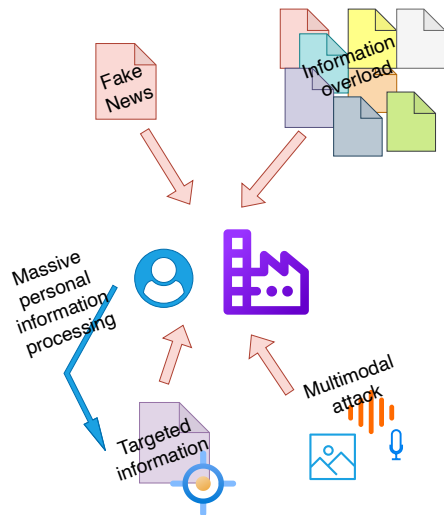
- Phishing messages that are highly personalized
- Link with data leaks of personal information

■ Informational harms

- Misinformation & disinformation
- Decision sabotage (finance, health, safety)
- Integrity attacks (e.g. scientific, journalistic)
- Pornographic Deepfake diffusion

■ Government control

- Mass surveillance becomes *possible*





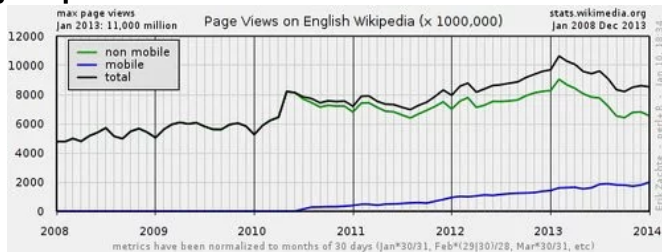
(3) Economic Questions

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

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

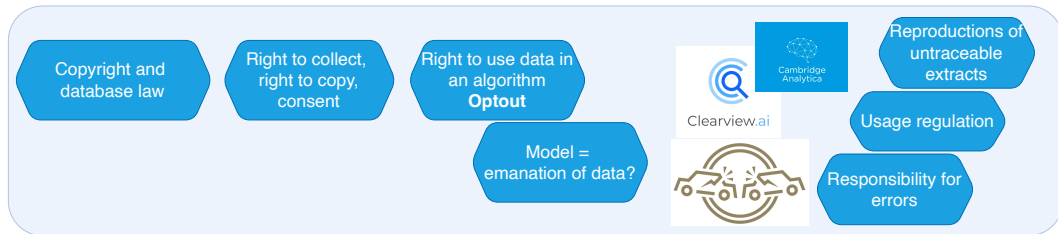
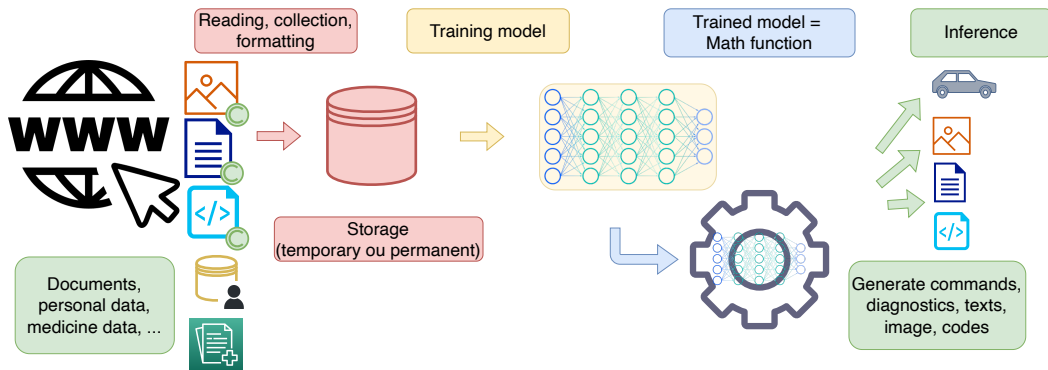
Google's Knowledge Graph Boxes: killing Wikipedia?

by Gregory Kohs



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

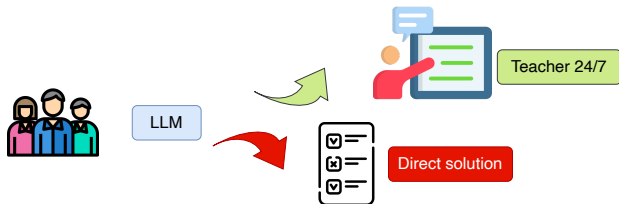
The chain of risks & responsibilities



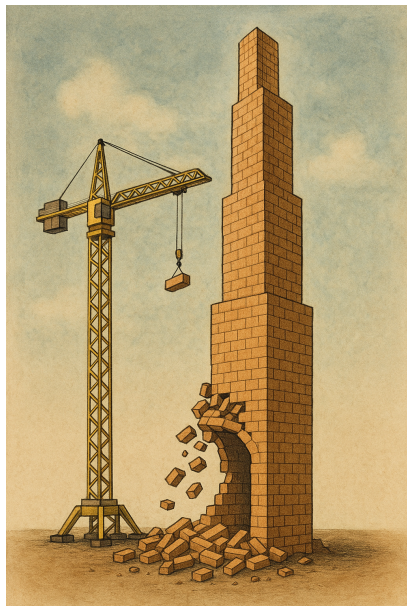


(4) Educational Challenges

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



- Learn to **recognize LLM-generated content**
- Do not underestimate the psychological aspects





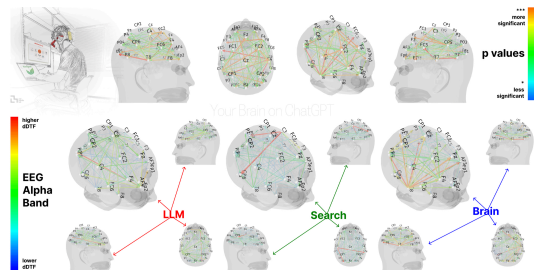
Evolution of Cognitive skills / psychology

Our brain will evolve (=decline) with these new tools...

What is the scope of these transformations? What will be the consequences?

- Education sciences and psychology had conjectured it...
cognitive sciences have measured it

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task, [N. Kosmyna et al. arXiv 2025](#)



How to remain motivated when a *calculator* answer all your exams?

But:

- Life does not look like exams...
- Keep the control...
- Illusion of understanding the solutions
- Temptation is too high



Detection of *texts generated by chatGPT*

Teaching how to **use** LLM = critical...

But also teaching how **NOT** to use them !

⇒ Control / detect / monitor usages

The screenshot shows a plagiarism detection tool interface. At the top, it says "Plagiat Def 2 -- Saved to my Mac". Below this, there's a search bar and a list of detected documents. The first document is "Plagiat Def 2 #4483eb" dated "07/01/2023 19:18 par vous", with 122 mots, 19,47 ko, and a "Plus d'infos" link. It shows a green progress bar at 0% and a "Rapport" button. The second document is "Plagiat Def 1 #f90f3" dated "07/01/2023 19:16 par vous", with 135 mots, 16,78 ko, and a "Plus d'infos" link. It shows a red progress bar at 100% and a "Rapport" button. The background text is a definition of "externalité" from Wikipedia.

Reformulation par chatGPT

Trier les documents par Date de dépôt 1 - 2 sur 2

Plagiat Def 2 #4483eb
07/01/2023 19:18 par vous | 122 mots | 19,47 ko | Plus d'infos
0% Rapport

Plagiat Def 1 #f90f3
07/01/2023 19:16 par vous | 135 mots | 16,78 ko | Plus d'infos
100% Rapport

Définition de Wikipedia

Crédit: S.
Pajak



Detection of *texts generated by* chatGPT

GPTZero

Detect AI Plagiarism. Accurately



ORIGINALITY.AI

Chat GPT



AI Detector

Torchbox

- **Text classifier** (like for any author)
 - Detection of biases in word choice / phrasing
- Characterization of text **plausibility** (OpenAI, GPTZero)
 - Hyper-fluency of sentences, over-abundance of logical connectors
 - Language model = stats $\Rightarrow$ measurement between distributions (**perplexity**)
- δ -**plausibility** on perturbed texts (DetectGPT)
- **LLM** have to integrate **fingerprints** in generated texts (AI act)

Detectors $\Rightarrow$ < 100% detection

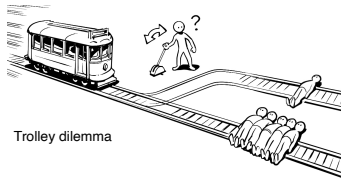
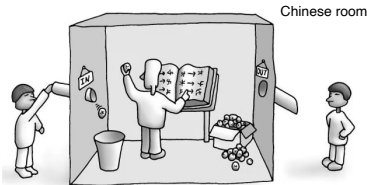
+ confidence level in detection

— depends on text length and modifications made

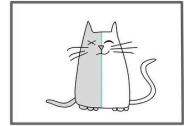
≈ detects pieces of Wikipedia (chatGPT = *stochastic parrot*)

CONCLUSION

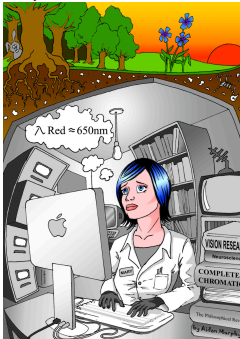
Representations of Artificial Intelligence



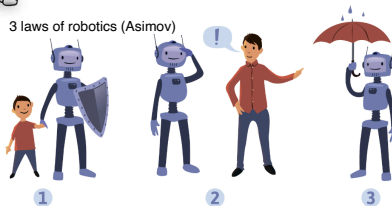
Schrodinger's cat



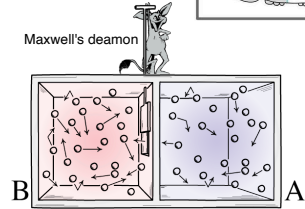
Mary's room



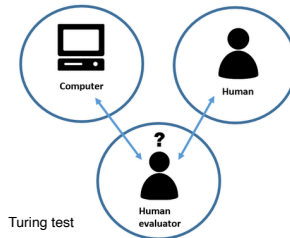
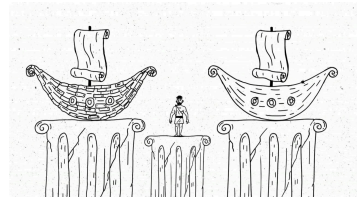
3 laws of robotics (Asimov)



Maxwell's demon



These's boat



Turing test



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 The smartphone has made me an *augmented human*...
- V Will the LLM make me an *augmented researcher*?
- V ⇒ Still, have a look at **NotebookLM**

■ 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'

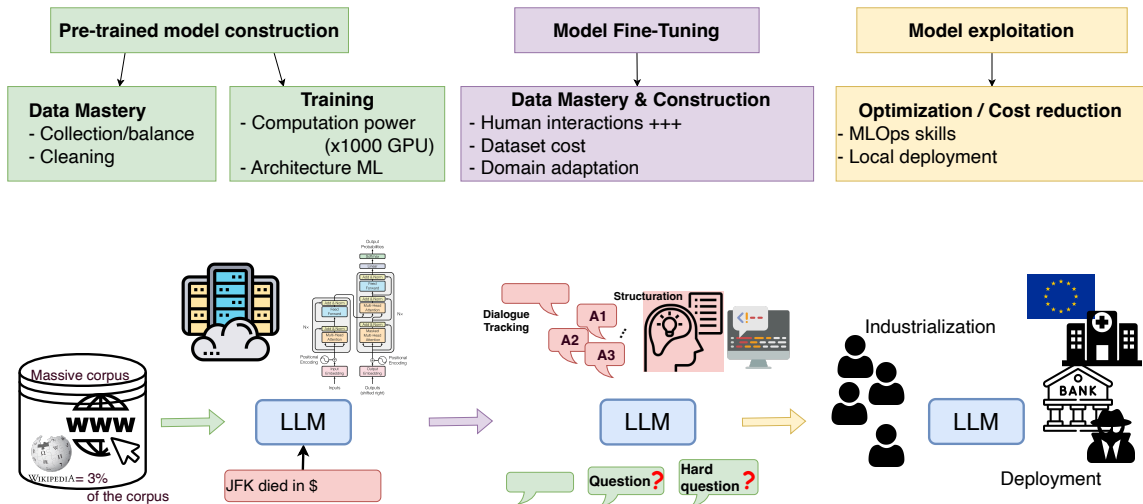
- Is it reasonable to train LLMs on GitHub, Wikipedia, scientific papers, news outlets, etc.?
- How important is privacy? What are the risks when using an LLM?



Levels of Access to Artificial Intelligence

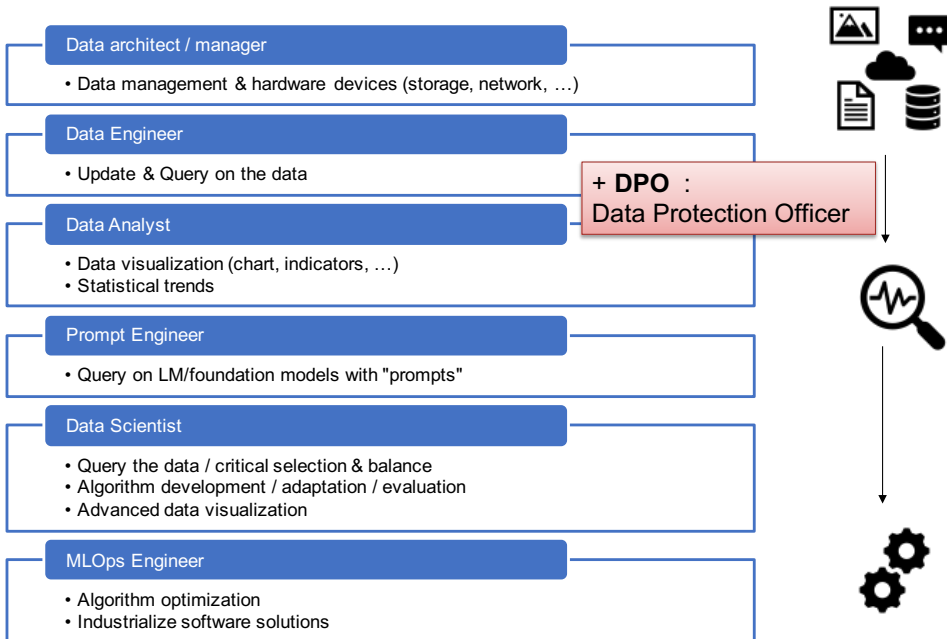
- 1 User via an interface: *chatGPT*
 - Some training is still required (2–4h)
- 2 Using Python libraries
 - Basics on protocols
 - Standard processing chains
 - Training: 1 week to 3 months (ML/DL)
- 3 Tool developer
 - Adapt tools to a specific case
 - Integrate business constraints
 - Build hybrid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

Digital Sovereignty: the Entire Chain





A Multitude of Professions





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



Factors of Acceptability for Generative AI

1 Utilitarianism:

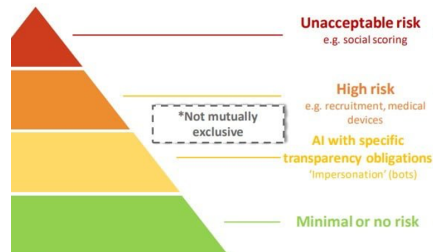
- Performance (acceptance factor of chatGPT)
- Reliability / self-assessment

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)





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?

