

FONCTIONNEMENT ET USAGES DES MODÈLES DE LANGAGE

Séminaire CIRAD
17 septembre 2026

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

INTRODUCTION



A very quick tour of NLP & machine-learning history

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

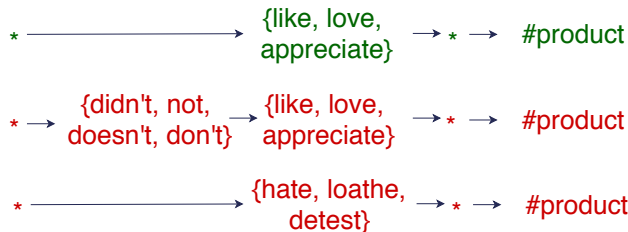
Building an opinion mining system:

1 What would a linguist search for?

2 How to generalize?

3 Implement with fast regex

- Gather documents
- Analyze
- Count

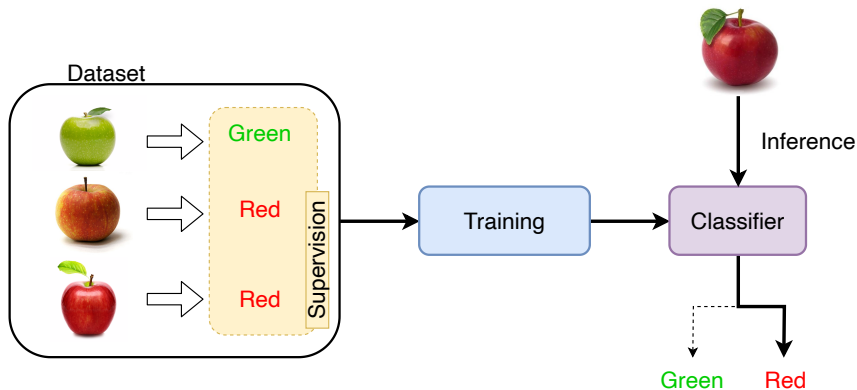




A very quick tour of NLP & machine-learning history

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

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

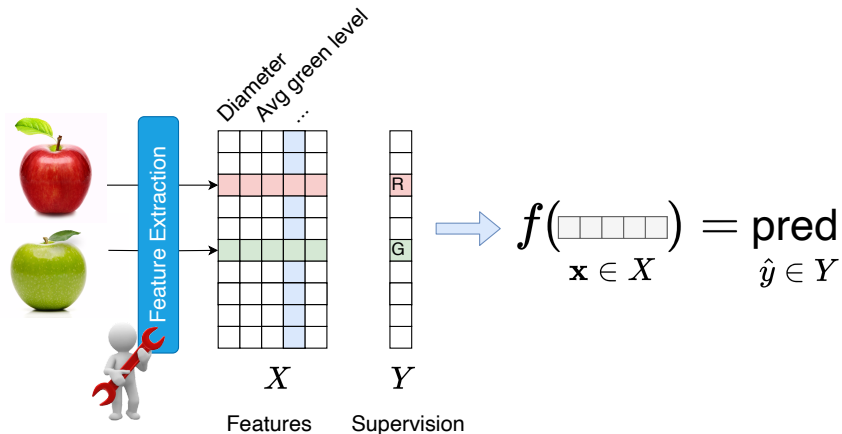




A very quick tour of NLP & machine-learning history

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

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

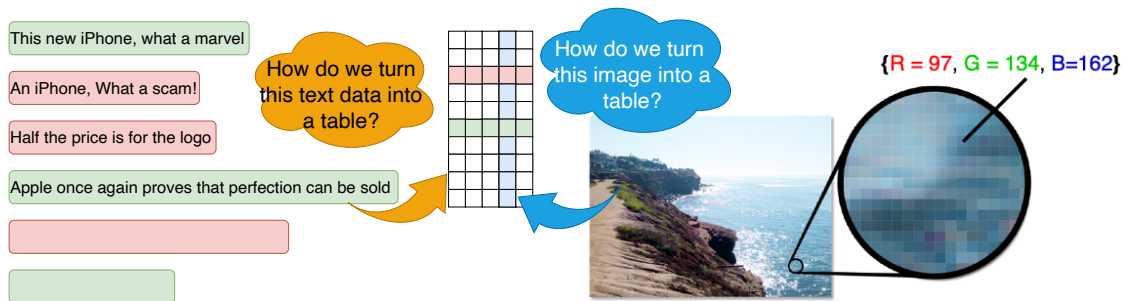




A very quick tour of NLP & machine-learning history

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

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



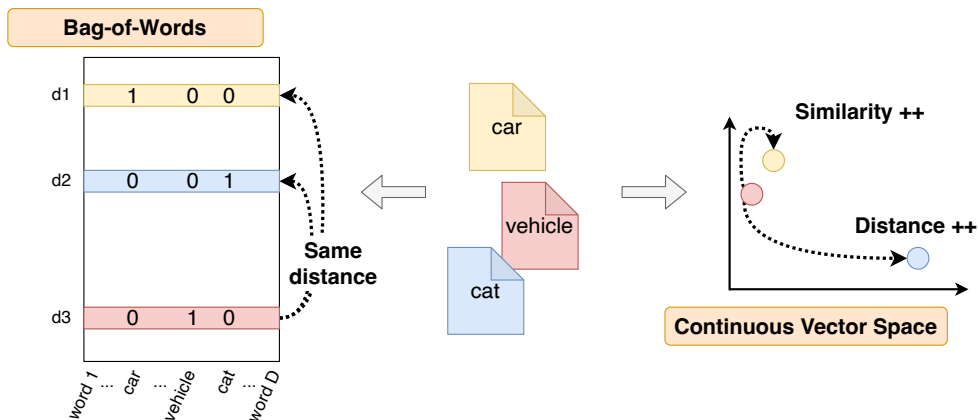


A very quick tour of NLP & machine-learning history

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





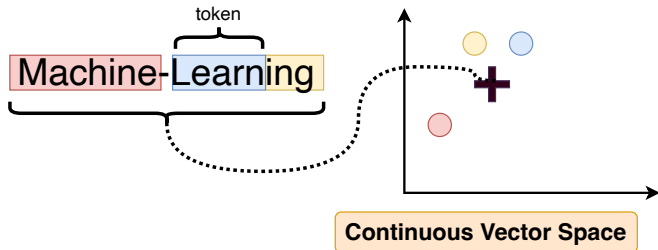
A very quick tour of NLP & machine-learning history

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





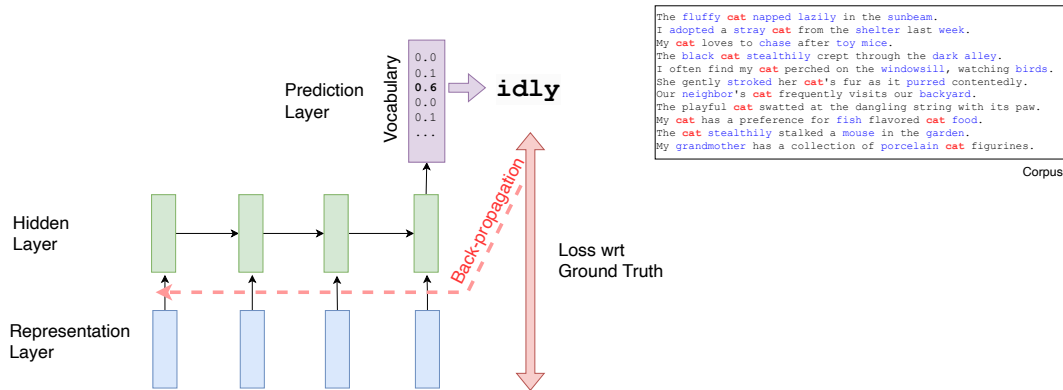
A very quick tour of NLP & machine-learning history

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



A very quick tour of NLP & machine-learning history

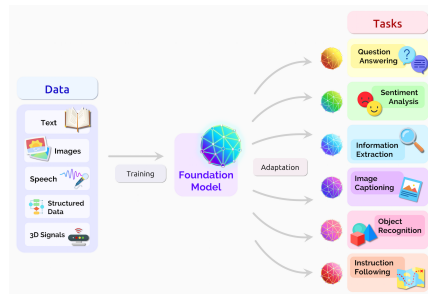
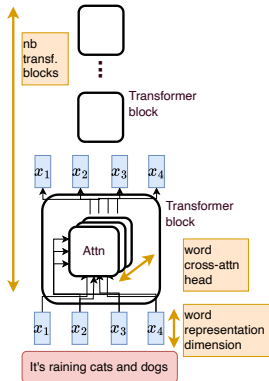
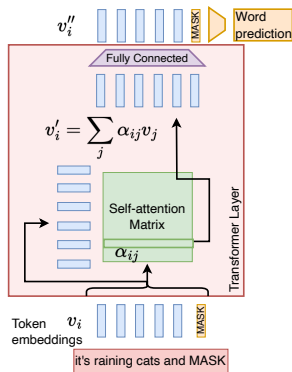
1950-2022 : AI = **Expert** knowledge (& know-how) $\Rightarrow$ automated systems

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

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

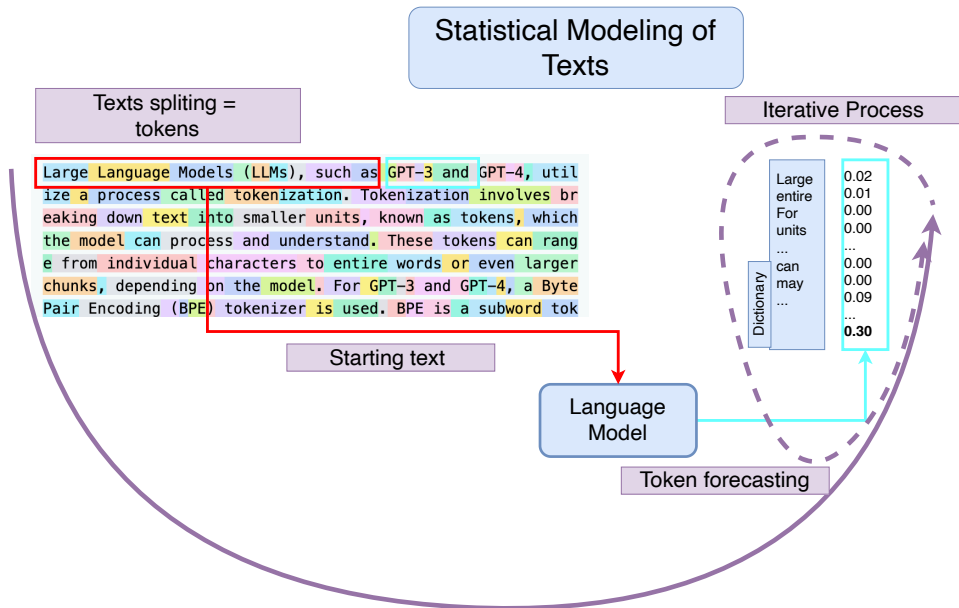
2015-... : **Language models** + pre-training

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





A (very smart) stochastic parrot!



CHATGPT

NOVEMBER 30, 2022

1 MILLION USERS IN 5 DAYS

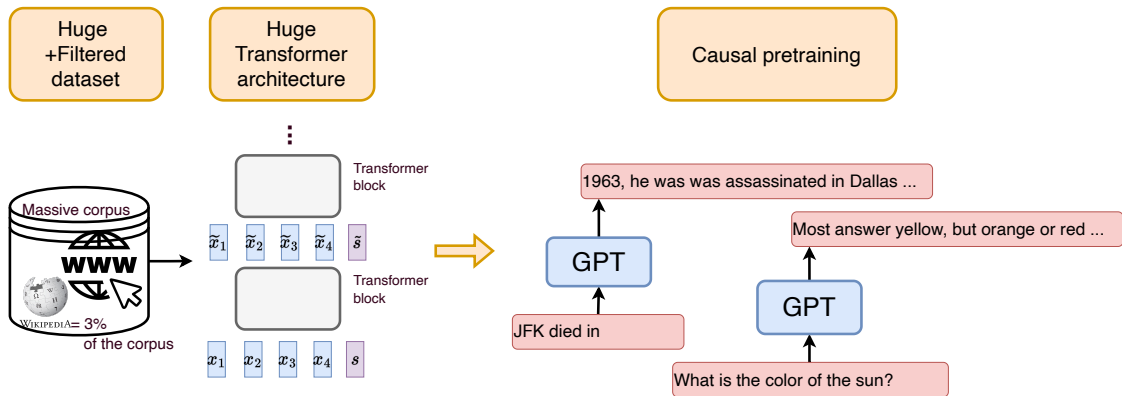
100 MILLION AT THE END OF JANUARY 2023

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



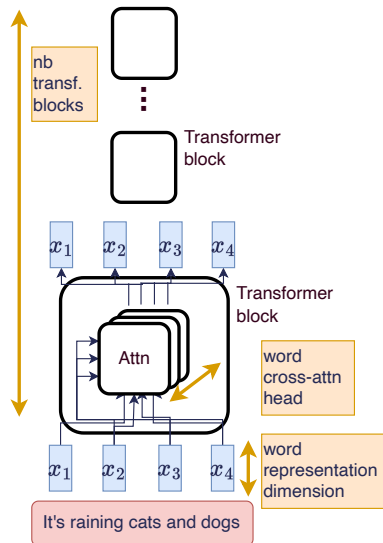
The Ingredients of chatGPT

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 €

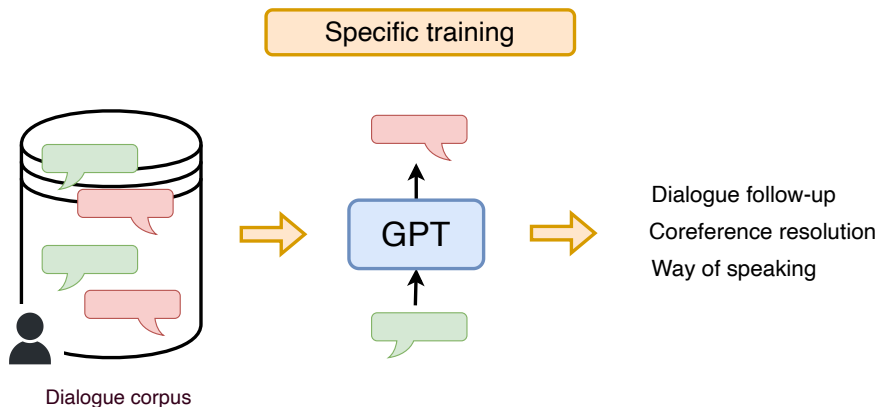


It's raining cats and dogs



The Ingredients of chatGPT

2. Dialogue Tracking



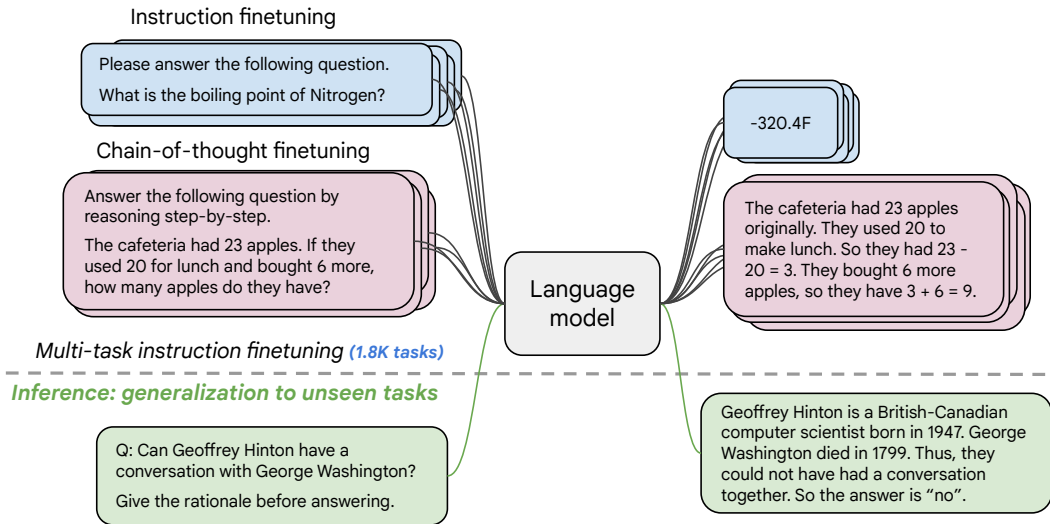
■ **Very clean data**

Data generated/validated/ranked by humans



The Ingredients of chatGPT

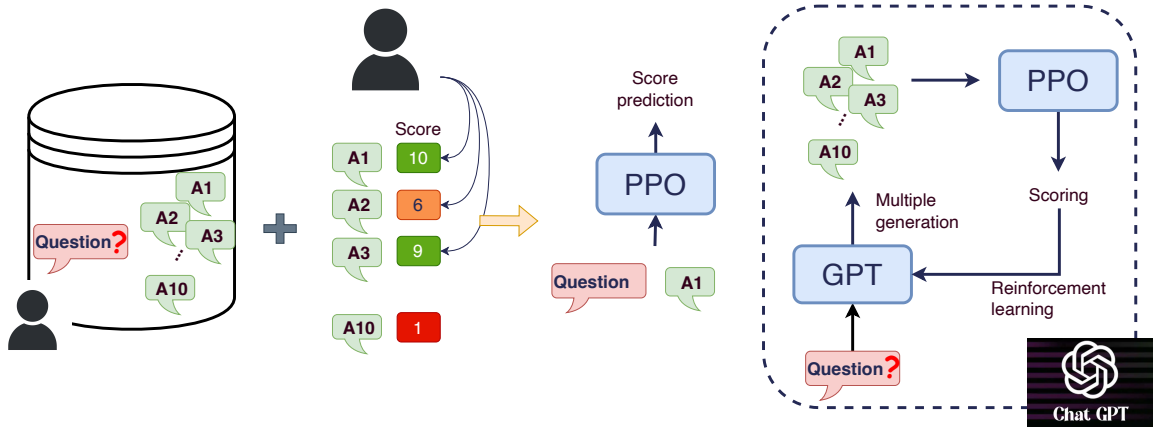
3. Fine-tuning on different ($\pm$) complex reasoning tasks





The Ingredients of chatGPT

4. Instructions + answer ranking



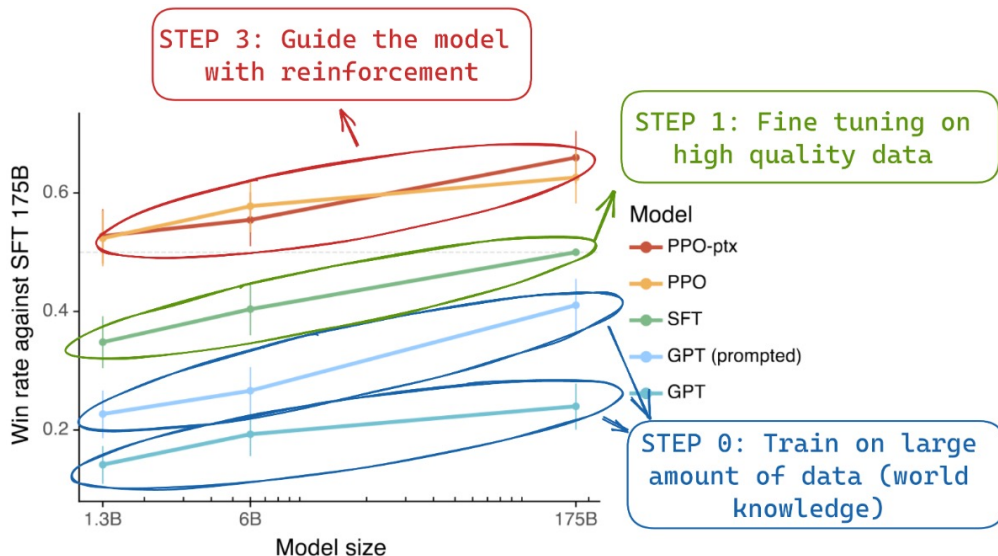
- Database created by humans
- Response improvement

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



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

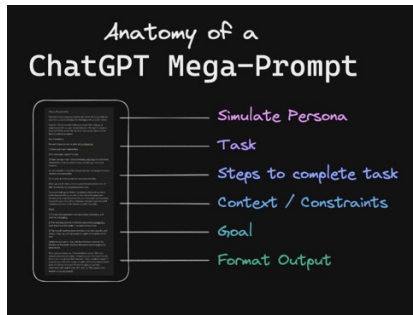
- Don't stop at the first question

- 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

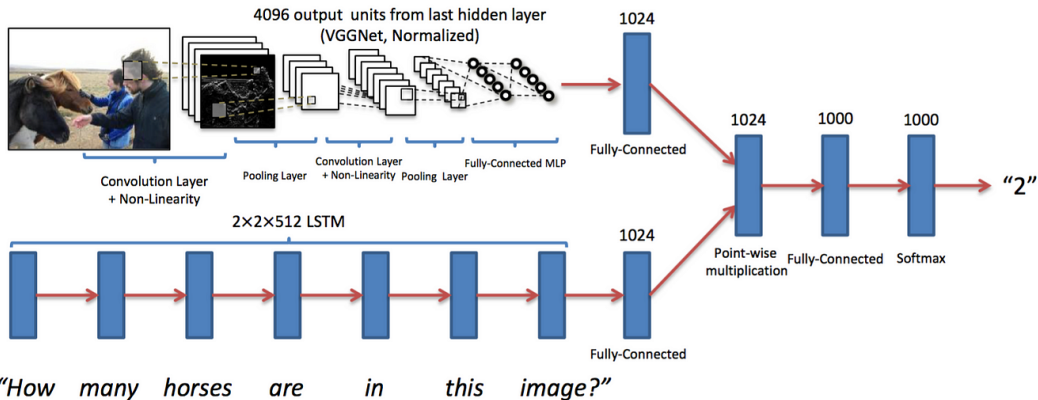
-chatgpt-prompt/



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

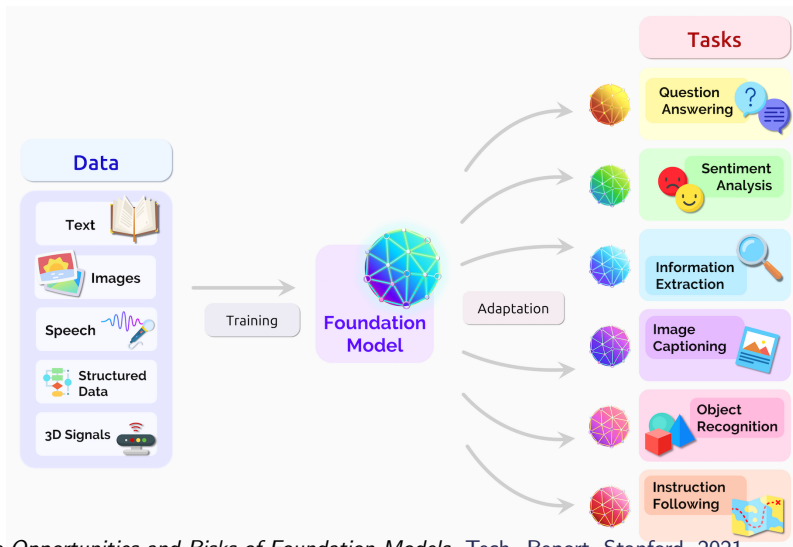


VQA: Visual Question Answering, arXiv, 2016, A. Agrawal et al.



Towards Larger Foundation Models?

- Let the modalities enrich each other



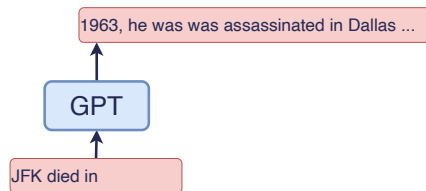
On the Opportunities and Risks of Foundation Models, Tech. Report, Stanford, 2021
Bommasani et al.

TECHNICAL LIMITS OF LANGUAGE MODELS



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 de modélisation de séries temporelles pour la modélisation de séries temporelles profondes, y compris les VAE.

[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](#) [VS Citer](#) [Clé 302 fois](#) [Autres articles](#) [Les 2 versions](#) [36](#)

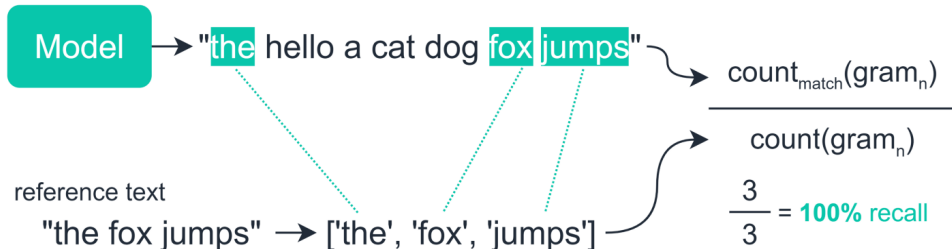
4. "Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Data" par Krishnan et al. (2017) - Cette étude présente une approche VAE pour la



Generative AI: how to evaluate performance?

The critical point today

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

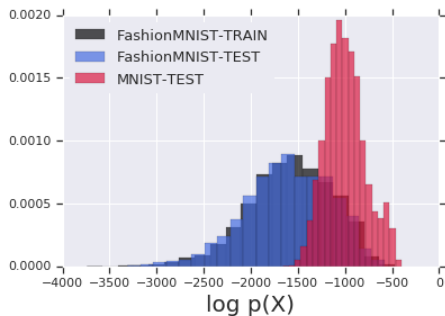




Generative AI: how to evaluate performance?

The critical point today

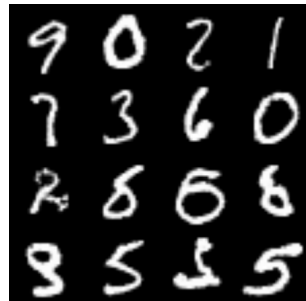
- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?



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

⇒ 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.



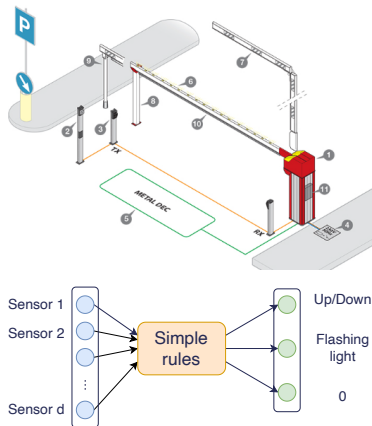
V

and today?

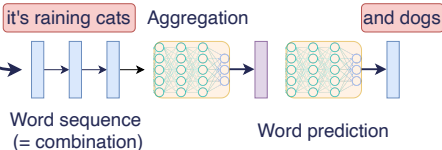




Stability, explainability... And complexity



- 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

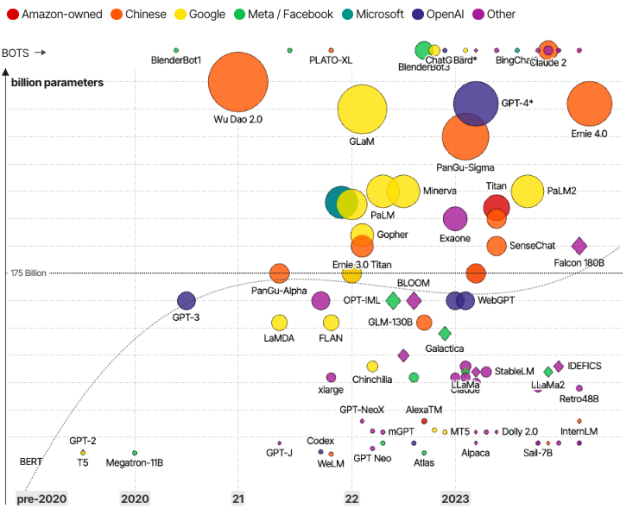
Neural networks = **explainable a posteriori** (almost always)



[Uber Accident, 2018]

- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable & explainable**
- Large dimension
- Complex non-linear combinations
- **Non-predictable & non-explainable**

The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT



Parameters

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



Different behaviors, different costs

Les IA sont démasquées !



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

x

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](#)

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

x

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

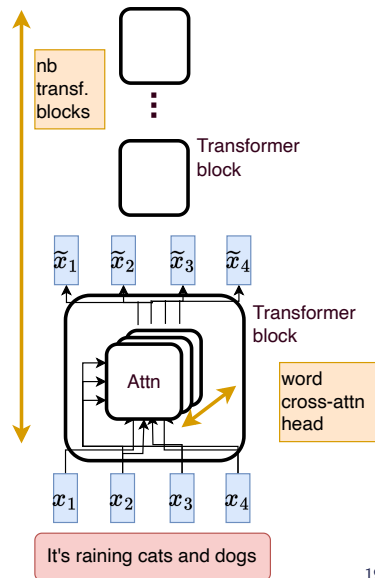
Pour un texte significatif en Français

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

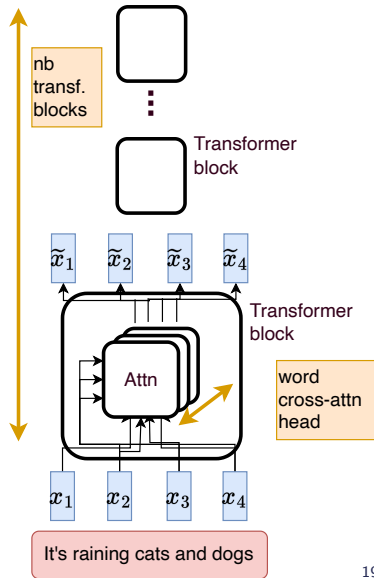
Experiment with different tokenizers (running locally in your browser).

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

124578 * 963|

TOKENS	CHARACTERS
5	12

124578 * 963

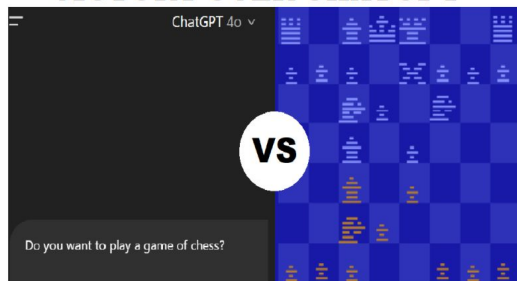
☒ Text ☐ Token IDs ☐ Hide



Everything beyond the LLM's capabilities/training

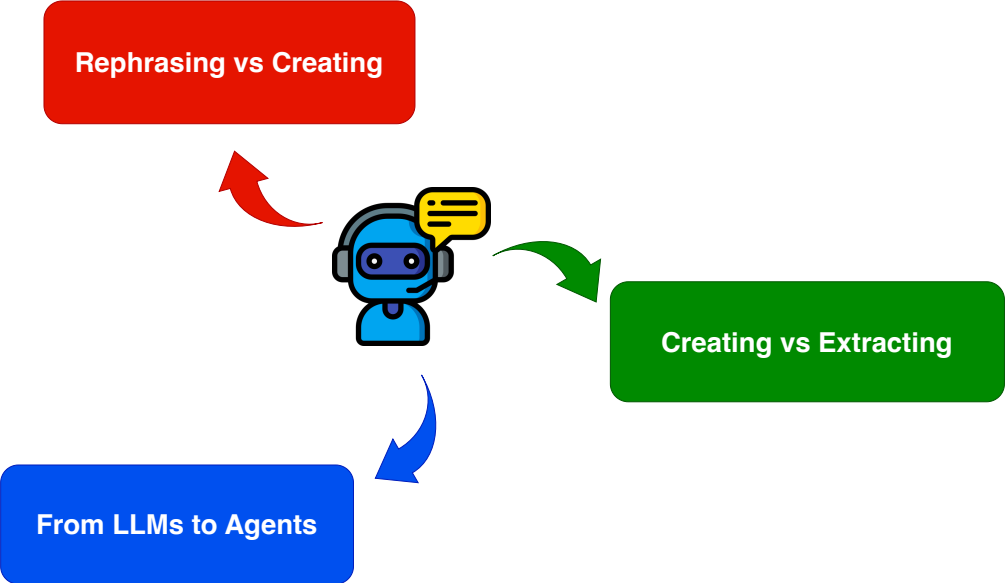
- Simple calculations
(multiplication, division)
- Generating n -syllable animal names
(in progress)
- Playing chess
- Following (complex) causal reasoning
- ...

ATARI 2600 SCORES STUNNING VICTORY OVER CHATGPT



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

LLM TYPOLOGY OF USES



Building a recommendation letter

Prompt

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

CV

Sujet

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

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

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



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

Ex-nihilo

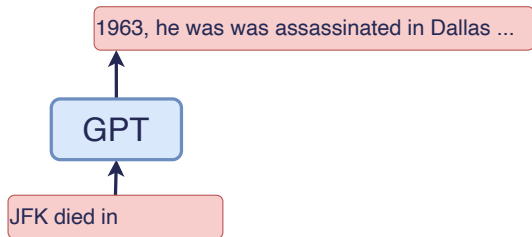
parametric memory

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

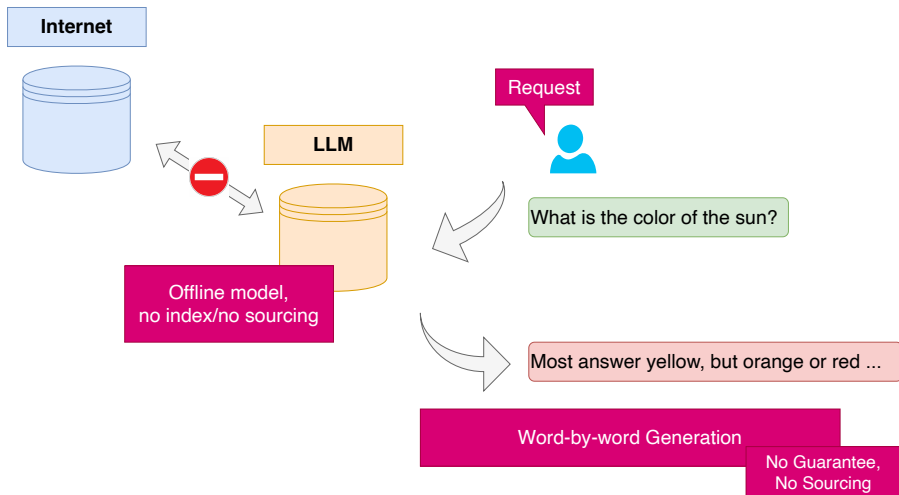
gravity

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



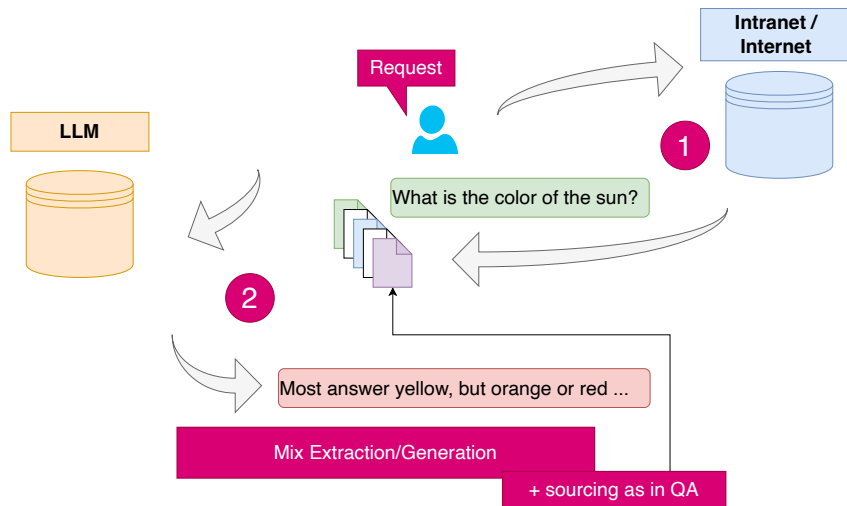
LLMs $\Rightarrow$ RAG: memory vs information extraction

- 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^{IA}

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**

- 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

⇒ The error rate did not change.

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



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.

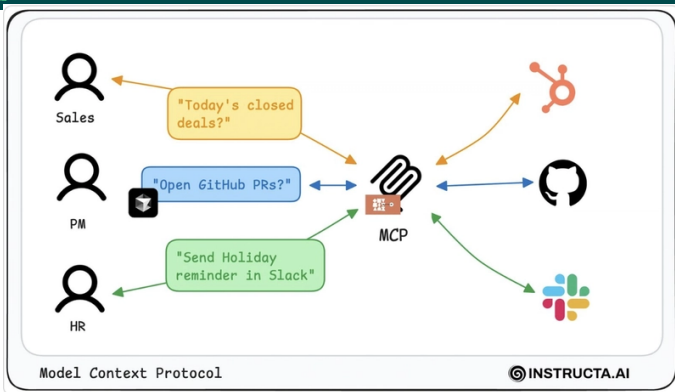
Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

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



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

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

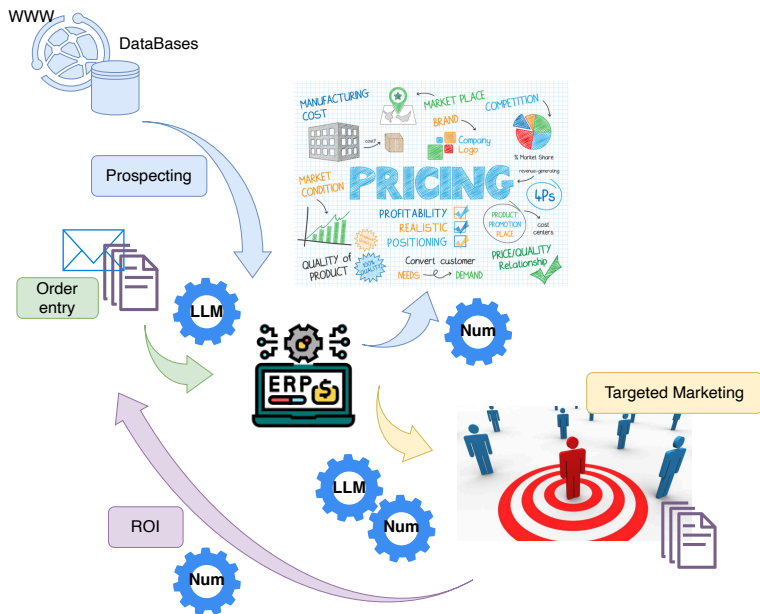
MCP = using an API in natural language

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





Integration into a Complex System

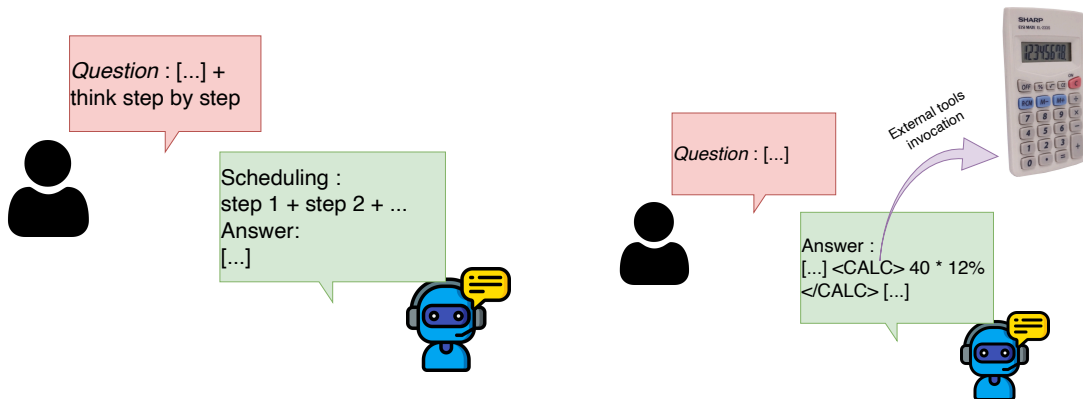


- Prospect list
- Email processing, needs identification
- ERP interaction $\Rightarrow$ proposal databases
- Pricing
- Drafting a commercial proposal



From *Chain-of-Thought* to Actions

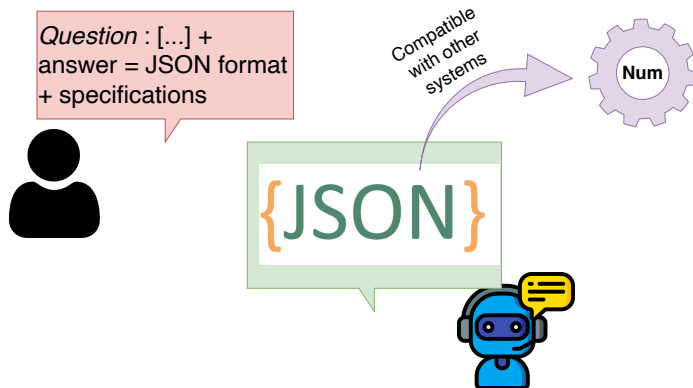
1 Chain-of-Thought / Toolformer





From *Chain-of-Thought* to Actions

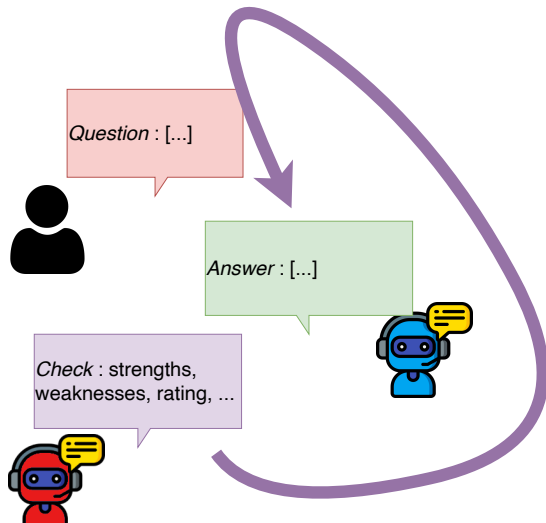
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information





From *Chain-of-Thought* to Actions

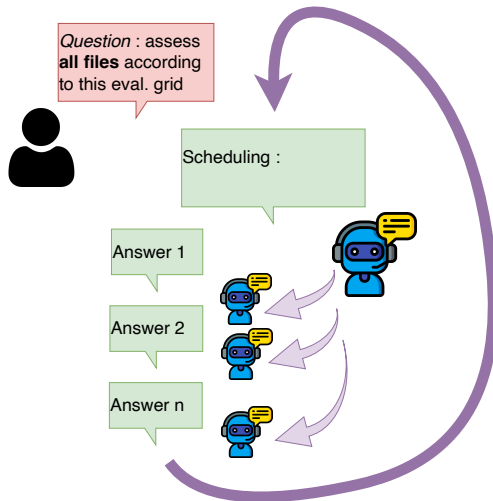
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information
- 3 LLM as a judge / reflexivity + improvement loop





From *Chain-of-Thought* to Actions

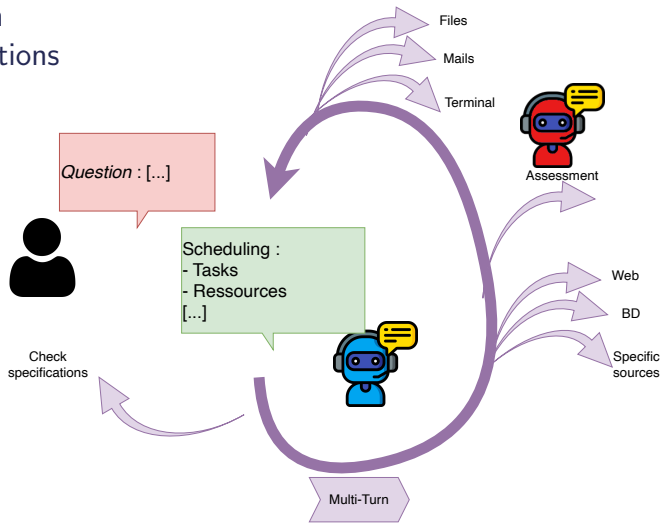
- 1 Chain-of-Thought / Toolformer
- 2 From text answers to structured information
- 3 LLM as a judge / reflexivity + improvement loop
- 4 Parallelism





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





Systems vs Tasks... For a student

2005 Wikipedia

- A much-criticized source that is gradually gaining respect
- First digital fraud detection systems



2010 Moodle

- Generalization of digital resources



2022 chatGPT

- Answer more & more questions...
- ... As a teacher available 24/7 ... Or as a magic solution finder



2023-24 NotebookLM / image input capability

- More & more opportunities
- ... For better or for worse





Systems vs Tasks... For a researcher

2000 Internet, emails, personal web pages

- Faster communication
- Paper diffusion

2010 GitHub

2015 Microsoft Academic research

- then Google Scholar

2015 Google Translate (v2)

- then DeepL...

2020 AI for science

- data availability, software maturity

2022 ChatGPT

- new bibliography tools
- then Agentic AI





Systems vs Tasks... For a programmer

2000 IDE / Completion / UML to code

- Environment to code faster

2020 1st version of codex / copilot

- Advanced code writing: algo/function level

2022 chatGPT

- Tutorials
- Debugging
- Advanced+ code writing: multi-function level

⇒ Expert users have to watch specifications + architecture

2025 Claude Code

- Dialogue on the architecture (= advanced planning)
- Code writing at the project level
- User interaction to check detailed specifications

⇒ Low-code / no-code is a reality (Sept. 2025)

eclipse



Codex





Systems vs Tasks... For everybody

- Translators, interpreters
- Journalists
- Consultants
- ...

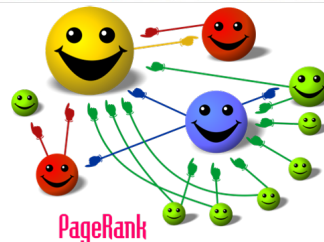
THE RISKS ASSOCIATED
WITH THESE USES





(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?

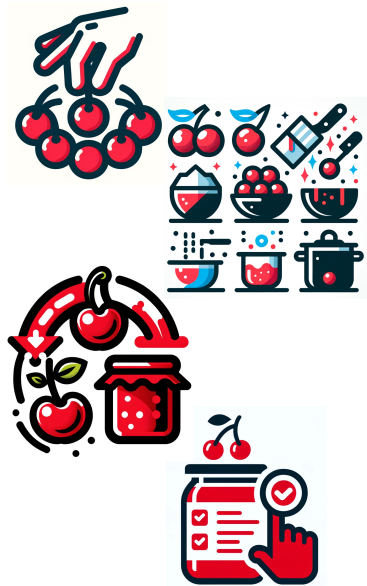




Machine learning is never neutral

- 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





(Human) Bias facing AI-Proposed Solutions

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?



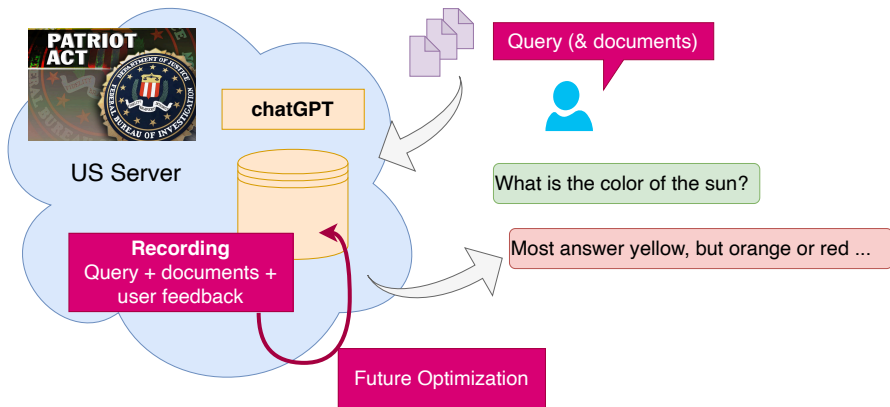
⇒ On a RAG, deletion works (remove the document).

⇒ On parametric memory, it does not.

Same question, two architectures.



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

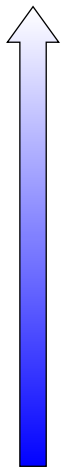


- Transfer of sensitive data
- Exploitation of data by OpenAI (or others)
- Data leakage in future models



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

Tools



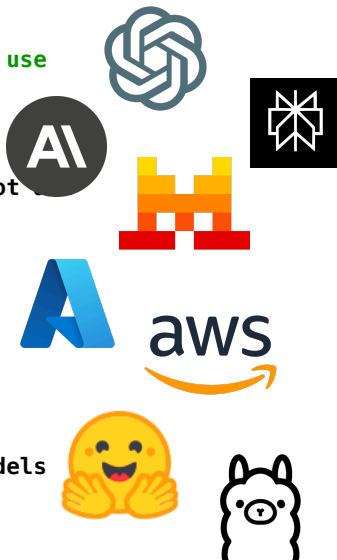
Commercial tools, **free to use**
Variable licence

Commercial tools,
Paid licence
more guaranties vs patriot

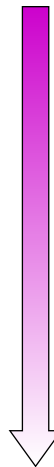
Commercial tools,
Paid licence + option
e.g. **European servers**

Institutional LLMs
deployed within a
controlled perimeter

Local use
pre-trained/finetuned models



Data



Any document



Personal
information



Ongoing
project



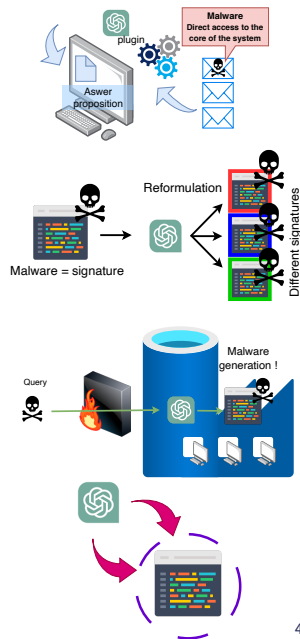
Medical
records





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

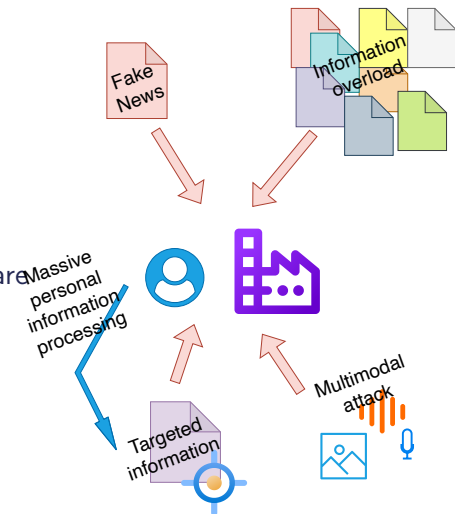
- Voice/text impersonation
- Fake authority impersonation (boss, teacher, lawyer, doctor)

■ Social engineering at scale

- Phishing messages that are highly personalized
⇒ the spelling mistakes that used to betray them are gone

■ Informational harms

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





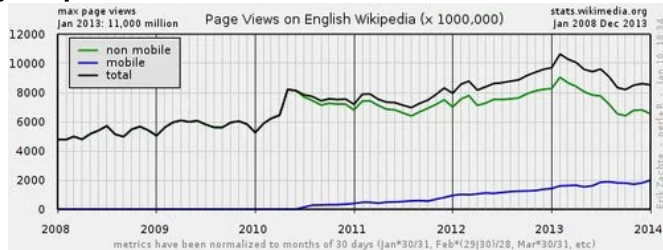
(3) Economic Questions

- 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?

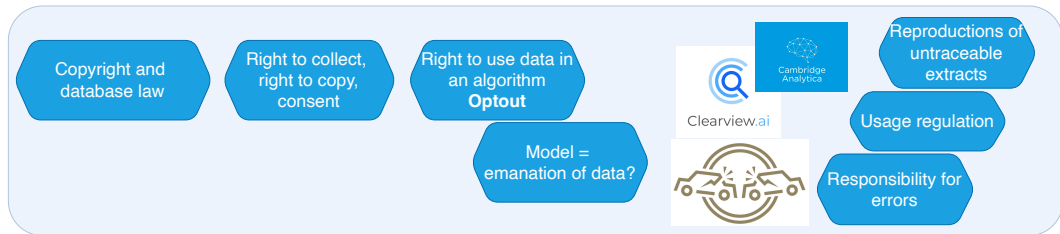
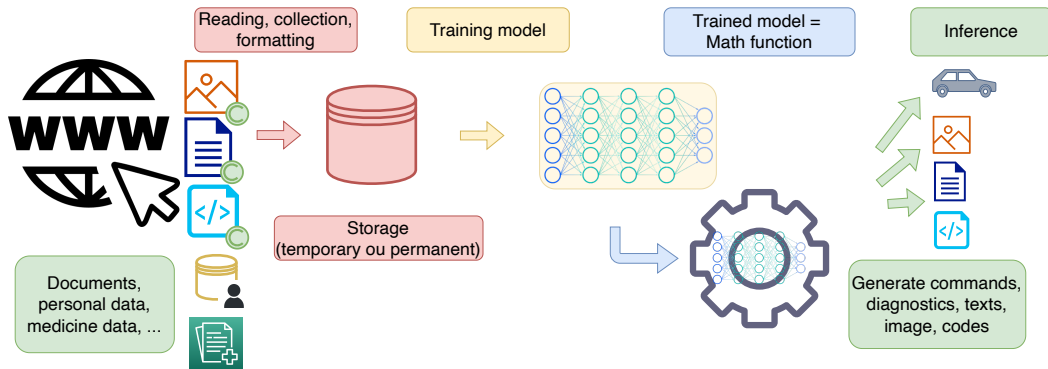
by Gregory Kohs



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



The chain of risks & responsibilities



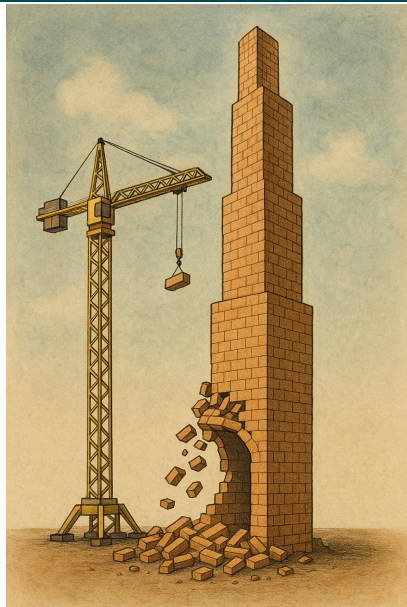


(4) Educational Challenges

- Redefine our **educational priorities**, subject by subject, as we did with Wikipedia/calculator/...
 - Accept the **decline of certain skills**
- 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





Cheating: detection of *texts generated by chatGPT*

Reformulation par chatGPT

Tout cocher : ☐ 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 #f90ff3 07/01/2023 19:16 par vous 135 mots 16,78 ko Plus d'infos	  100%	 Rapport 

Définition de Wikipedia

L'externalité fait référence au fait qu'une activité économique d'un agent peut avoir un impact sur d'autres personnes sans qu'il y ait de compensation financière. Cela peut être bénéfique pour les autres, comme offrir une utilité gratuitement, ou nuisible, comme causer des dommages à l'environnement.

L'externalité caractérise le fait qu'un agent économique crée, par son activité, un effet externe en procurant à autrui, sans contrepartie monétaire, une utilité ou un avantage de façon gratuite, ou au contraire une nuisance, un dommage sans compensation (coût social, coût écosystémique, pertes de ressources pas, peu, difficilement, lentement ou coûteusement renouvelables...).

De la sorte, un agent économique se trouve en position d'influer consciemment ou inconsciemment sur la situation d'autres agents, sans que ceux-ci soient parties prenantes à la décision : ces derniers ne sont pas forcément informés et/ou n'ont pas été consultés et ne participent pas à la gestion de ses conséquences par le fait qu'ils ne reçoivent (si l'influence est négative), ni ne paient (si l'influence est positive) aucune compensation.

En résumé : « Tout coûte mais tout ne se paie pas »

Crédit: S.
Pajak



Cheating: detection of *texts generated by* chatGPT

GPTZero

Detect AI Plagiarism. Accurately



ORIGINALITY.AI

Chat GPT



AI Detector

Torchbankz

- **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 = statistical $\Rightarrow$ measurement between distributions (**perplexity**)
- δ -**plausibility** on perturbed texts (DetectGPT)
- **chatGPT** *should quickly* integrate **fingerprints** in generated texts

Detectors $\Rightarrow$ $< 100\%$ detection

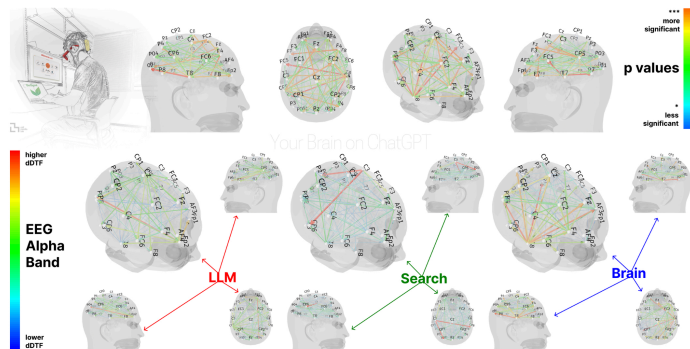
+ confidence level in detection



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

- Education sciences and psychology had conjectured it...

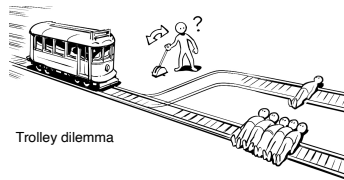
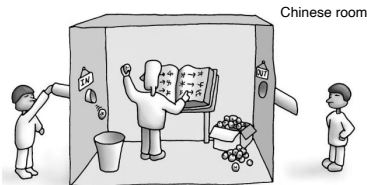
cognitive sciences have measured it



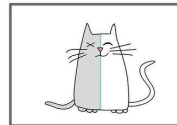
Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task, N. Kosmyna et al. arXiv 2025

CONCLUSION

Representations of Artificial Intelligence



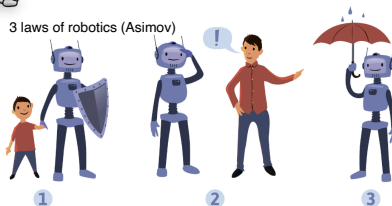
Schrodinger's cat



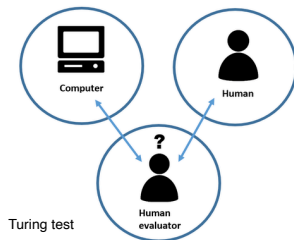
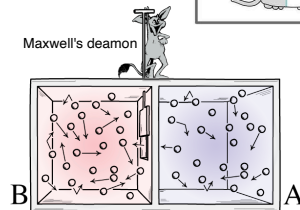
Mary's room



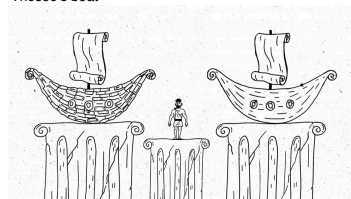
3 laws of robotics (Asimov)



Maxwell's demon



These's boat





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?

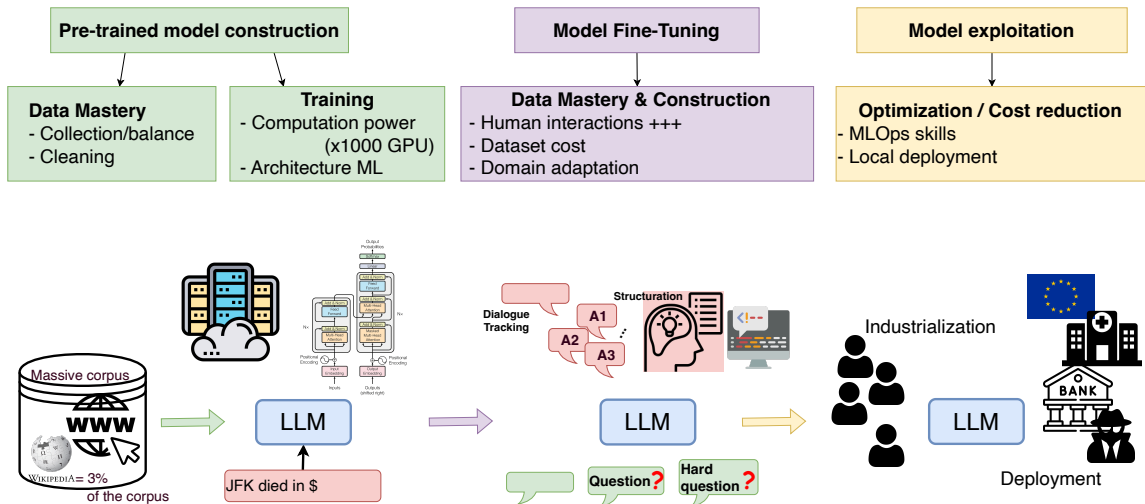


Levels of Access to Artificial Intelligence

- 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

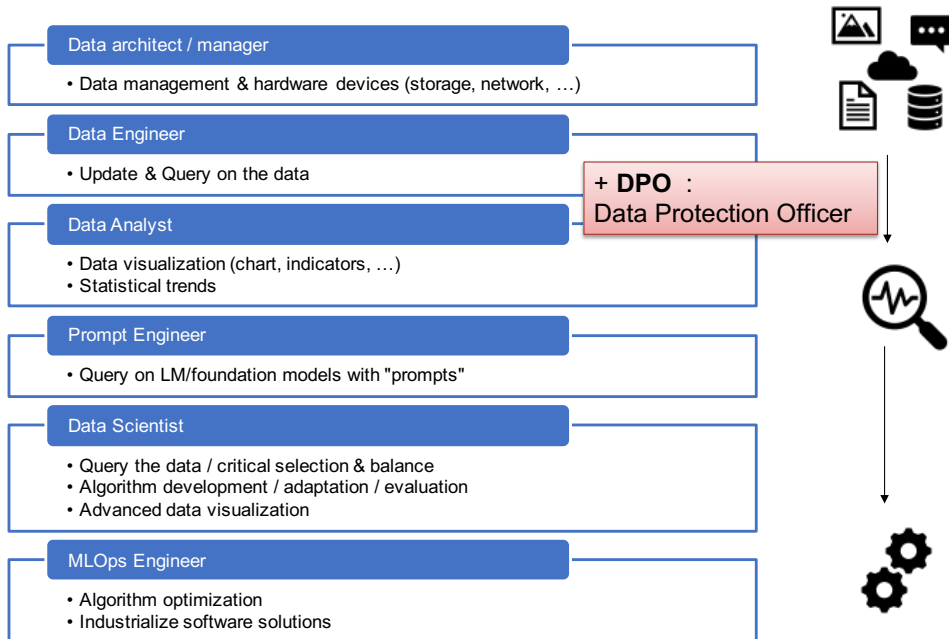


Digital Sovereignty: the Entire Chain





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:

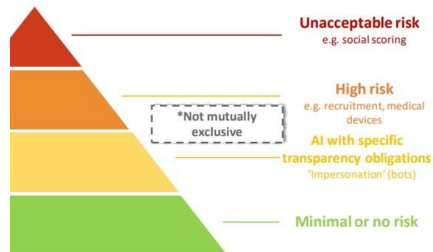
- Performance (acceptance factor of chatGPT)
- Reliability / self-assessment

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?

