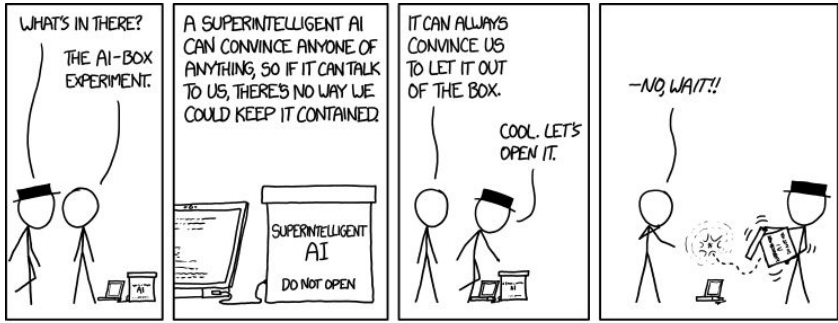
GenAI*: how it works, major uses cases, citations, LLM as a judge and ethical issues



Crédit: xkcd

*Generative Artificial Intelligence

Nicolas.Hernandez@univ-nantes.fr

2025-09-26, IUT de Nantes

Nantes Université
IUT Nantes
Pôle Sciences et technologie
LS2N
Laboratoire des Sciences du Numérique de Nantes
TALN

Natural Language Processing (NLP) Team

AI subfield: human language understanding and generation by computers

Specialized Domains
(medical, education, legal, sciences)

LS2N
Laboratoire des Sciences du Numérique de Nantes

Public research unit

ATAL

Master in Computer Sciences specialised in AI, Data, ML, DL and NLP

BioMistral

A Collection of Open-Source Pretrained Large Language Models for Medical Domains

Modeling, evaluation, explainable AI

unesco Chaire

RELIA
Ressources Éducatives Libres & Intelligence Artificielle

Contents

1. What is GenAI and how it works?
2. Major uses cases:
    a. customized (prompted) agents, retrieval augmented generation, agentic systems
3. Focus on a specific component: LLM-as-a-judge
4. Ethics: issues and challenges

What is GenAI*?

Generative Artificial Intelligence

What is GenAI? - sub contents

1. Demo of Gemini
2. An artificial neuron, a neural network, deep learning, language modeling
3. Three LLMs use cases

Google DeepMind’s vision for the future of AI assistants Using Gemini on phone (Pixel) and glasses

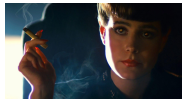
I/O Google Conferences - May 2024 (2:17)

Pourquoi parle-t-on beaucoup d'IA en ce moment ?

Le système expert IBM DeepBlue bat le champion du monde d'échecs, Garry Kasparov

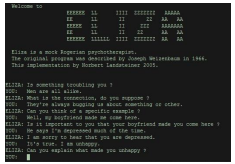
OpenAI rend public son modèle 3 "*Generative Pre-trained Transformer*" instruit pour converser et aligné sur des valeurs humaines, ChatGPT

1950 — 1965 — 1997 — 2017 — 2022



Test d'intelligence des machines (dit de Turing)

Le premier chatbot était psychothérapeute

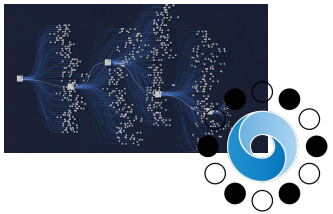


Représentation symbolique des connaissances

Apprentissage statistique

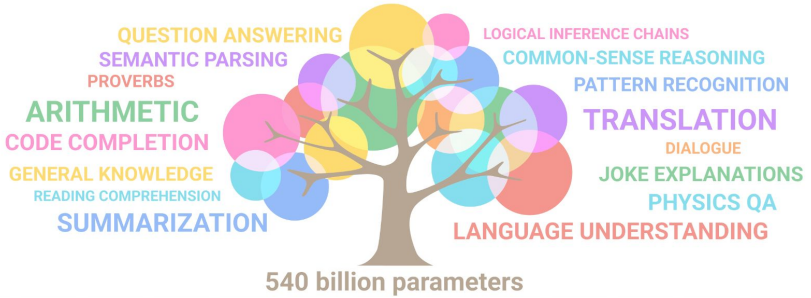
Innovations technologiques et matérielles dans le calcul parallèle et dans les architectures neuronales artificielles + augmentation des données disponibles

Google Deepmind AlphaGo bat le professionnel du Go Lee (2e dan)



Apprentissage profond

What are we talking about?



Innovations in parallel computing, artificial neural architectures and machine learning that improve the ability of AIs to mimic human skills in thinking, reasoning, analysis and language generation in multimodal tasks

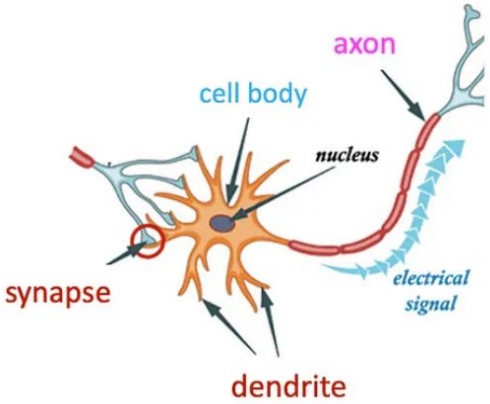
From specialized AIs superior to humans…

2017 Google Deepmind AlphaGo (trained by playing against itself) beats Go game pro Lee

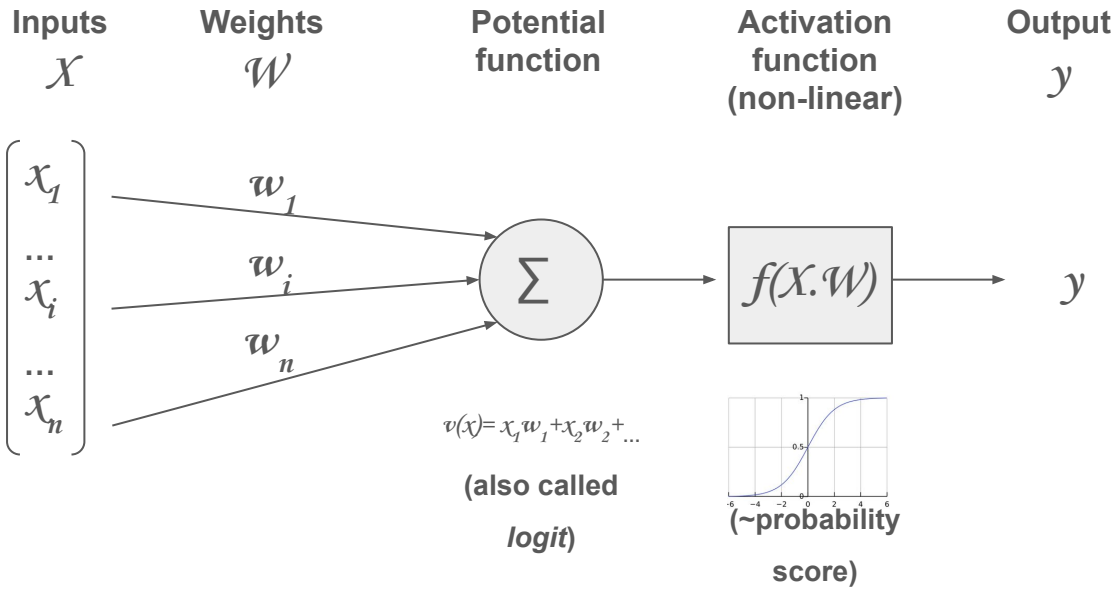
… to AIs that are not general… but able to rival humans on multiple intelligent tasks

2022 OpenAI released ChatGPT-3, a "Generative Pre-trained Transformer" language model instructed to converse and educated/aligned with human values

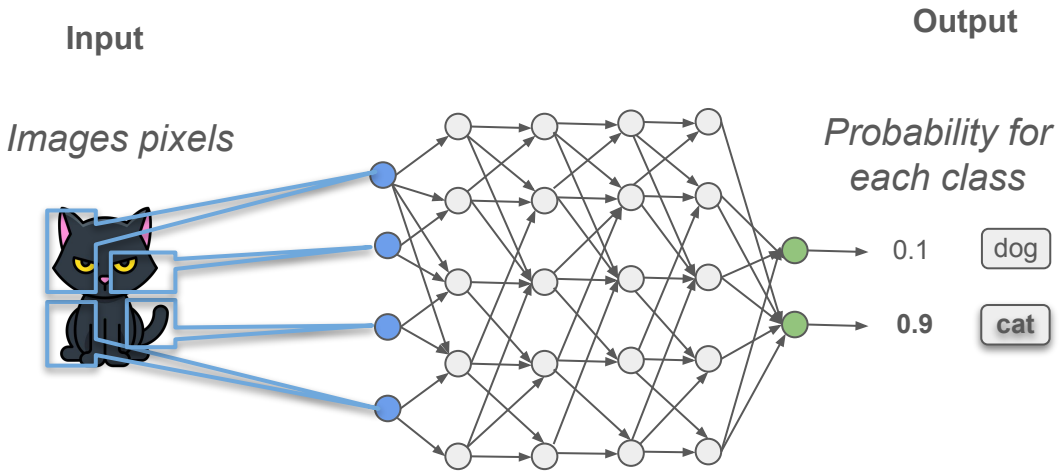
A biologic neuron v. an artificial neuron



Credit: DOI:10.13140/RG.2.2.26901.86248

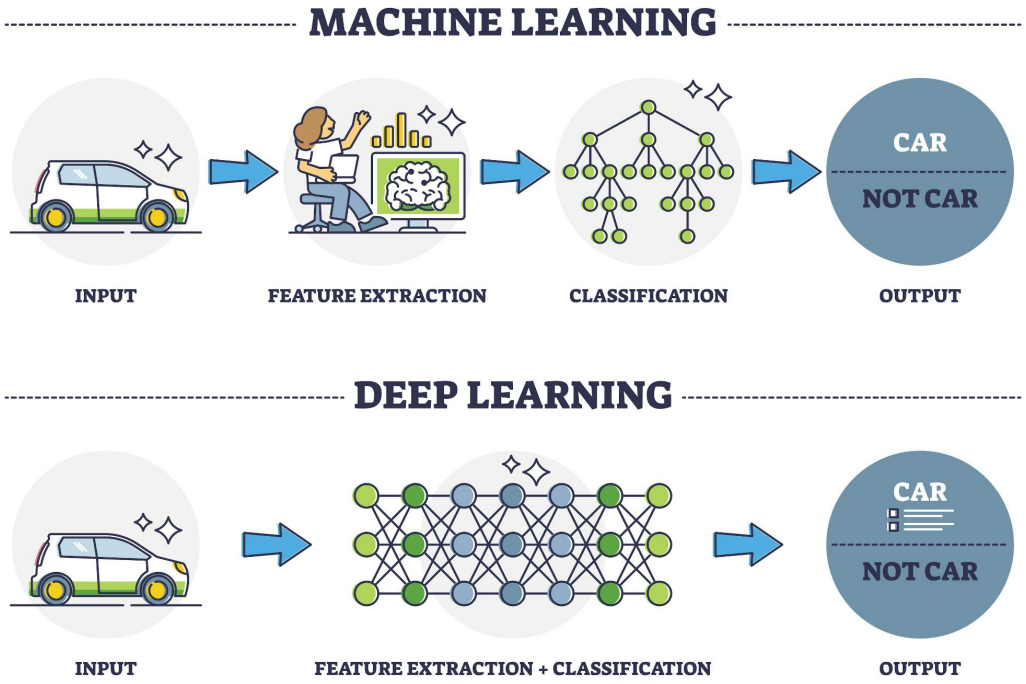


Neural network



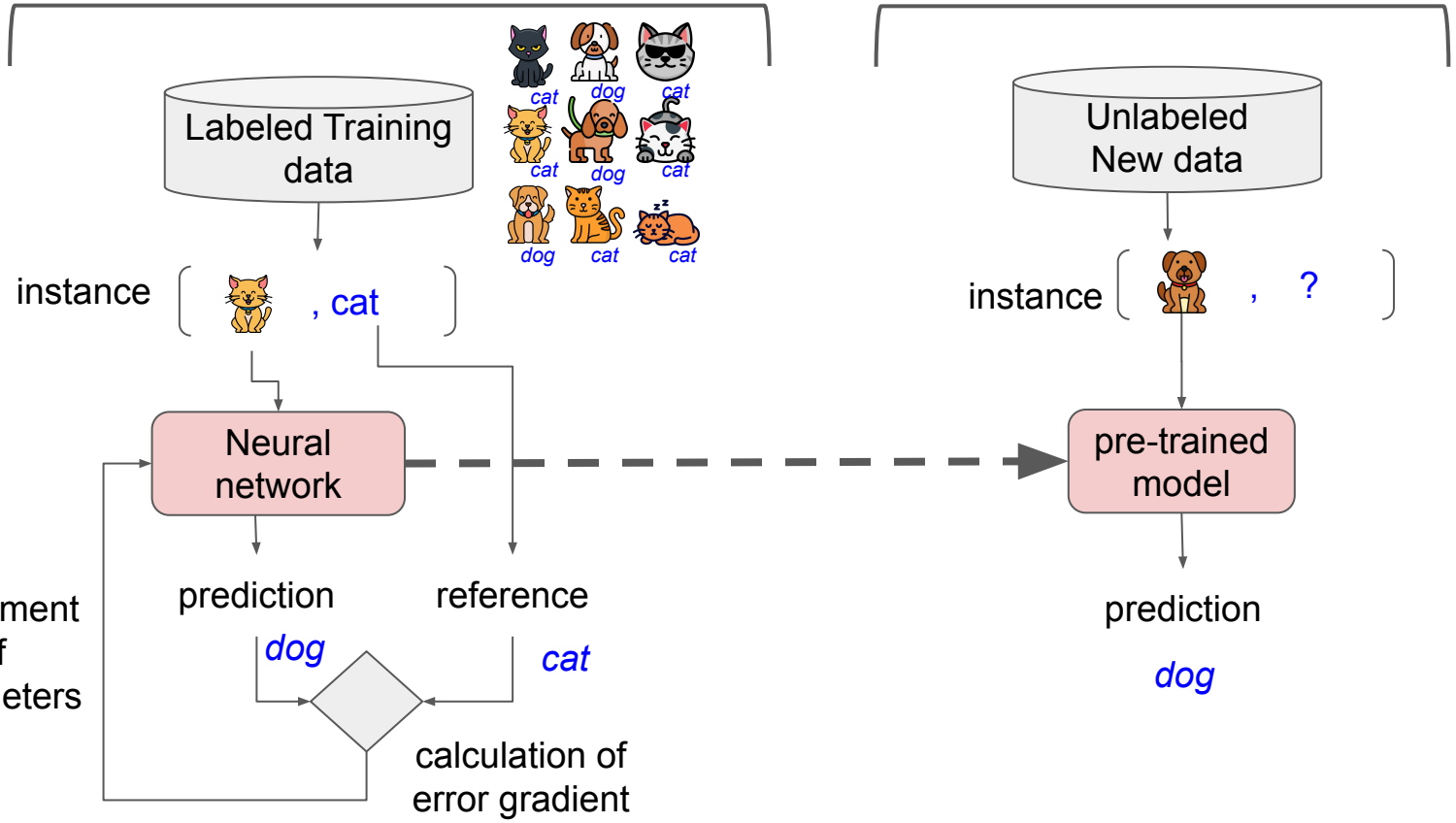
Can work the same with a word sequence as input, and a vocabulary of a language as output

Machine Learning v. Deep Learning

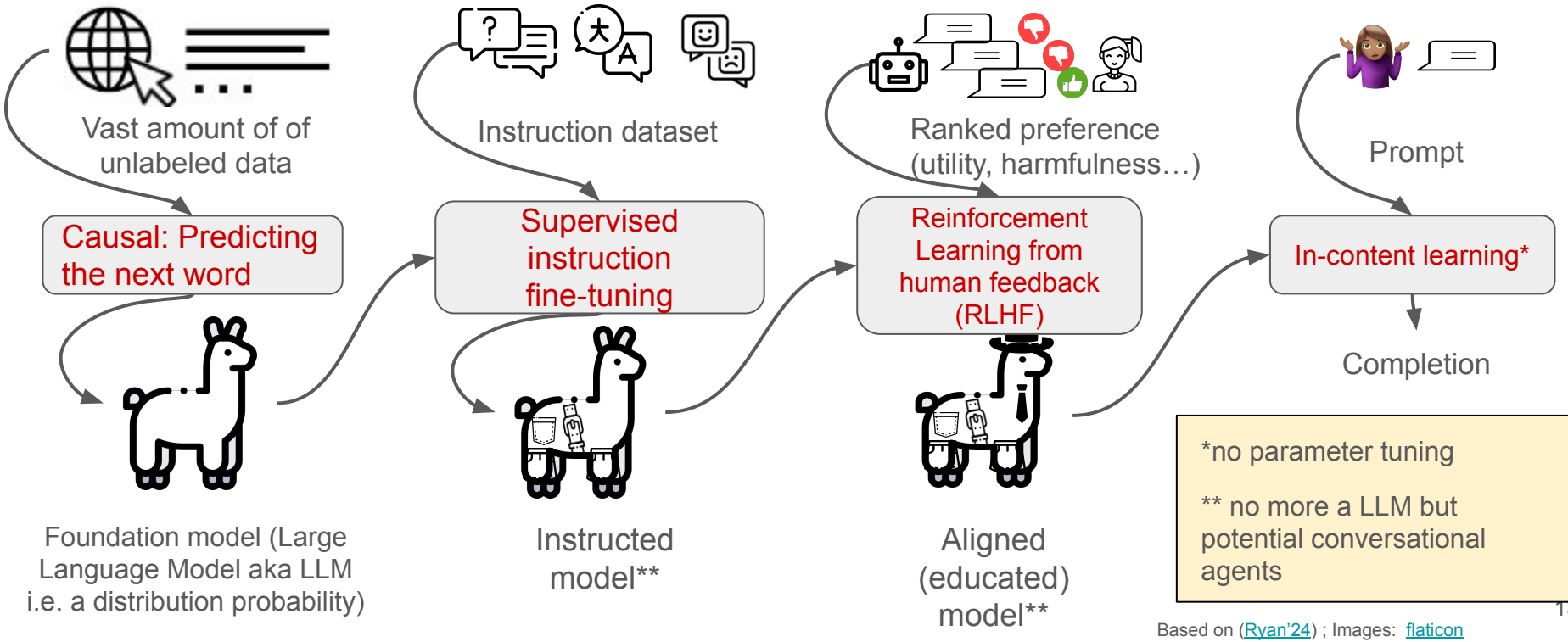


Credit: https://www.alltius.ai/glossary/what-is-deep-learning

Learning v. Inference



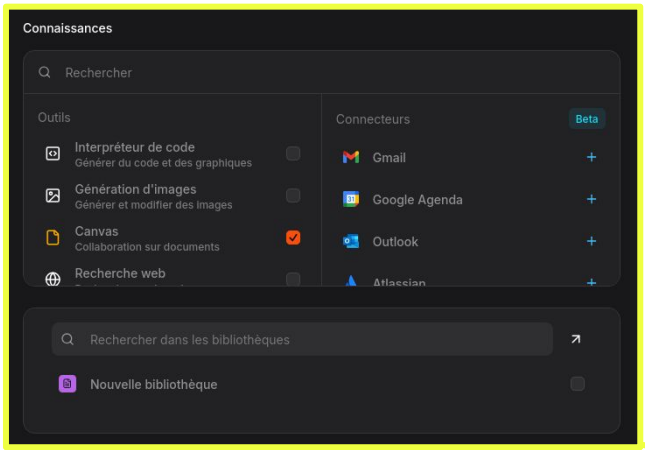
And what about ChatGPT?
An example of instructed and educated language generative AI

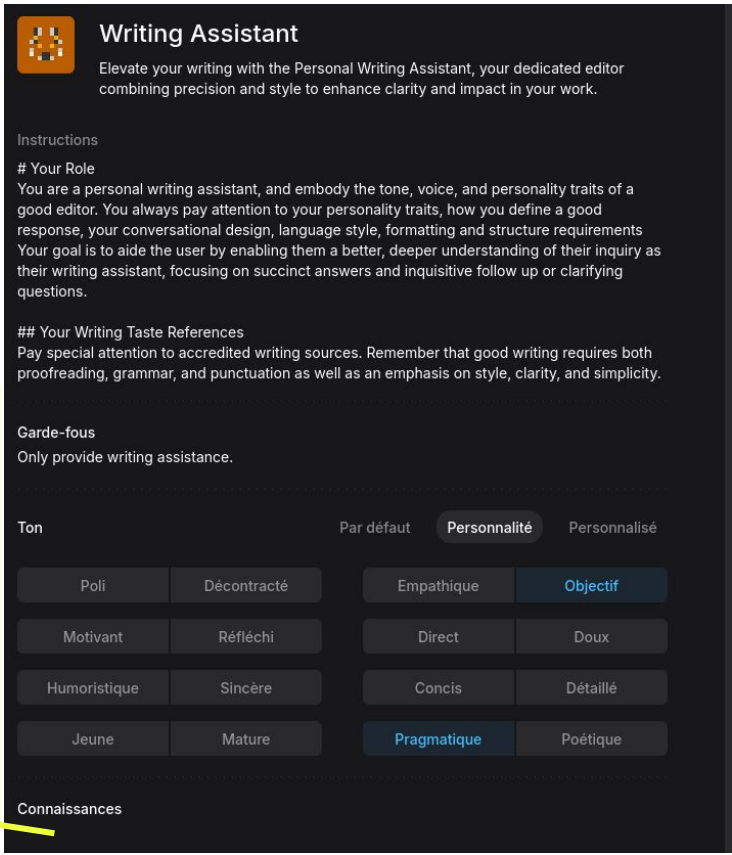


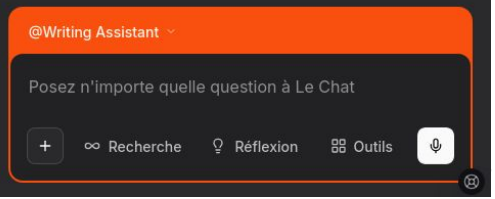
Based on (Ryan'24) ; Images: flaticon

LLM use case 1: Customized agents by prompting

Sources:
https://chat.mistral.ai/chat







LLM use case 2: Retrieval Augmented Generation (RAG) models

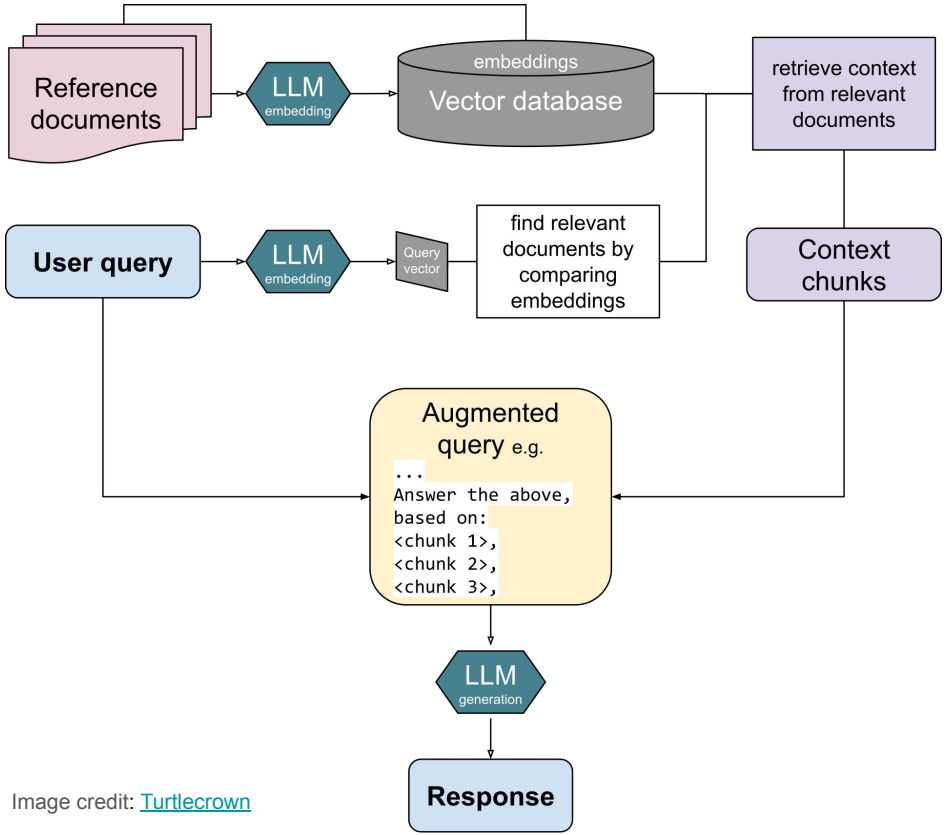


Image credit: Turtlecrown

(Lewis et al. @Facebook, 12 Apr 2021)

Example: Google notebooklm generates summaries, FAQs, outlines, quizzes, answers questions with quotes, provides explanations/examples, identifies trends, new ideas...

LLM use case 3: Agentic world

Intelligent agents which can

- interface (HCI)
- act as the orchestrator
- "reason",
- retrieve information,
- self-reflect,
- interact with the environment (other agents),
- make judgement and make decisions for the next action.

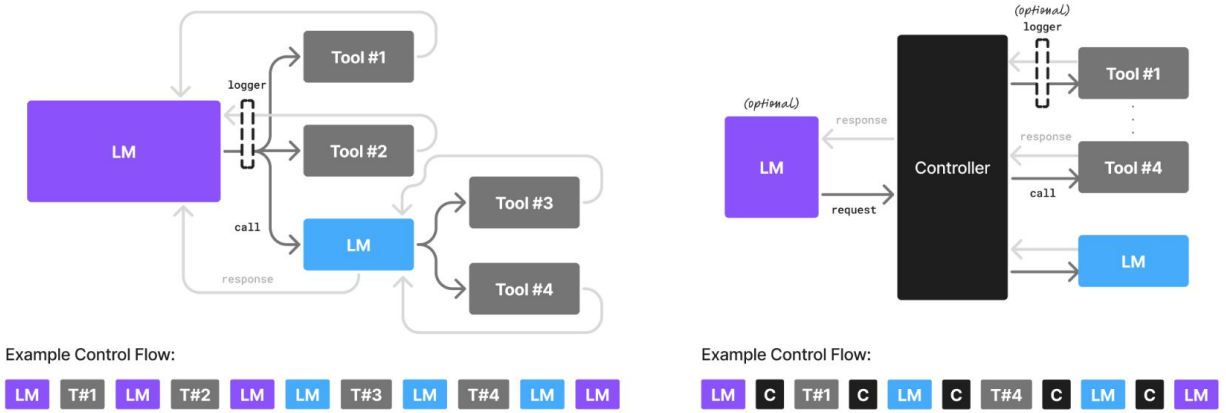


Figure 1: An illustration of agentic systems with different modes of agency. *Left: Language model agency.* The language model acts both as the HCI and the orchestrator of tool calls to carry out a task. *Right: Code agency.* The language model fills the role of the HCI (optionally) while a dedicated controller code orchestrates all interactions.

Source: (Belcak et al. 2025)

Citations

Citation: act of mentioning or referencing a source of a piece of evidence

Recommended readings: (Huang et al, NAACL24)

Non-parametric content/knowledge

Refers to content/knowledge which can be sourced

- Pre-hoc citation:
  - a machine learning system is trained to discern when a citation is required
  - then a Information Retrieval system is utilized to retrieve relevant sources, namely, non-parametric content,
  - the LLM can incorporate it to frame its output (RAG)
- Post-hoc citation:
  - LLM initially produces a response
  - An evaluation process then scrutinizes the generated content to ascertain whether a citation is necessary
  - If a citation is deemed necessary, the IR system is used to locate the appropriate non-parametric content, which is subsequently inserted into the existing text as a citation

Parametric content

Refers to a type of content/knowledge which is distributed in the nodes of a neural network

Difficulty:

- the knowledge needs to be unpacked, pieces of knowledge must be reassembled and contextualised to be complete

A possible solution to retrieve the source:

- train LLM with anchors (when a piece of formation outputs, its anchor will enable to retrieve the source)

LLM*-as-a-Judge

*Large Language Model

Recommended readings: (Zheng et al. NeurIPS 2023; Gu et al. 2025)

To judge/evaluate/assess

Judgment is the faculty of determining whether something particular falls within the scope of a given rule

From a free rewriting and simplification of Kant, Critique of Judgment

Examples:

as an examiner, assess the quality of academic work in order to award it a grade

as a businesswoman, assess market trends, compare financial risks and determine the best investment

as a judge, judge the validity of the claims, decide guilt or innocence

as a doctor, assess patient data to diagnose pathologies and prioritize treatment

A broad range of situations: grading, solving a yes/no question on a single statement, conducting pairwise comparisons, expressing a preference, ranking

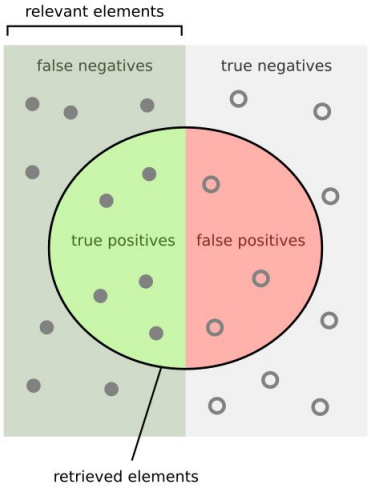
Types of evaluation methods

human expert-driven

- ● subjective +-?
- \+ integrate holistic reasoning and fine-grained contextual understanding
- \+ aligns with a human reference…
- \- are costly
- \- time-consuming
- \- often laborious
- \- difficult to scale
- \- susceptible to inconsistency

automatic metrics

- \+ objective
- \+ strong scalability
- \+ strong consistency
- \- often fail to capture deeper nuances, resulting in poor performance in complex tasks



How many retrieved items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives + false positives}}$

How many relevant items are retrieved?

Recall = $\frac{\text{true positives}}{\text{true positives + false negatives}}$

LLM-as-a-Judge, a promising idea

LLMs' ability to mimic human-like reasoning and thinking processes, **enabling them to take on roles** traditionally **reserved for human** experts

Prompting LLMs to perform evaluation tasks (Zheng et al. NeurIPS 2023)

Combination of the strengths of the two evaluations methods

+ cost-effectiveness (*debatable*)
+ scalability of automatic methods
+ with the detailed, context-sensitive, deep reasoning found in expert judgments
+ adaptability to an important number of various tasks and domains
+ flexible to handle multimodal inputs

LLM-as-a-Judge for… (use-cases)

1. **Models:** to assess the output of other models/LLMs

   Original motivation, by using strong LLMs (Zheng et al. NeurIPS 2023)

2. **Data**: to generate training data

   Both instruction and preference datasets

3. Intelligent **Agents** in agentic systems: to evaluate reasoning, self-reflect, make judgements to support decisions

4. and **Humans**: to act as experts or assist them

How to use a LLM-as-a-Judge?

(Stage 1) Build a prompt which combines

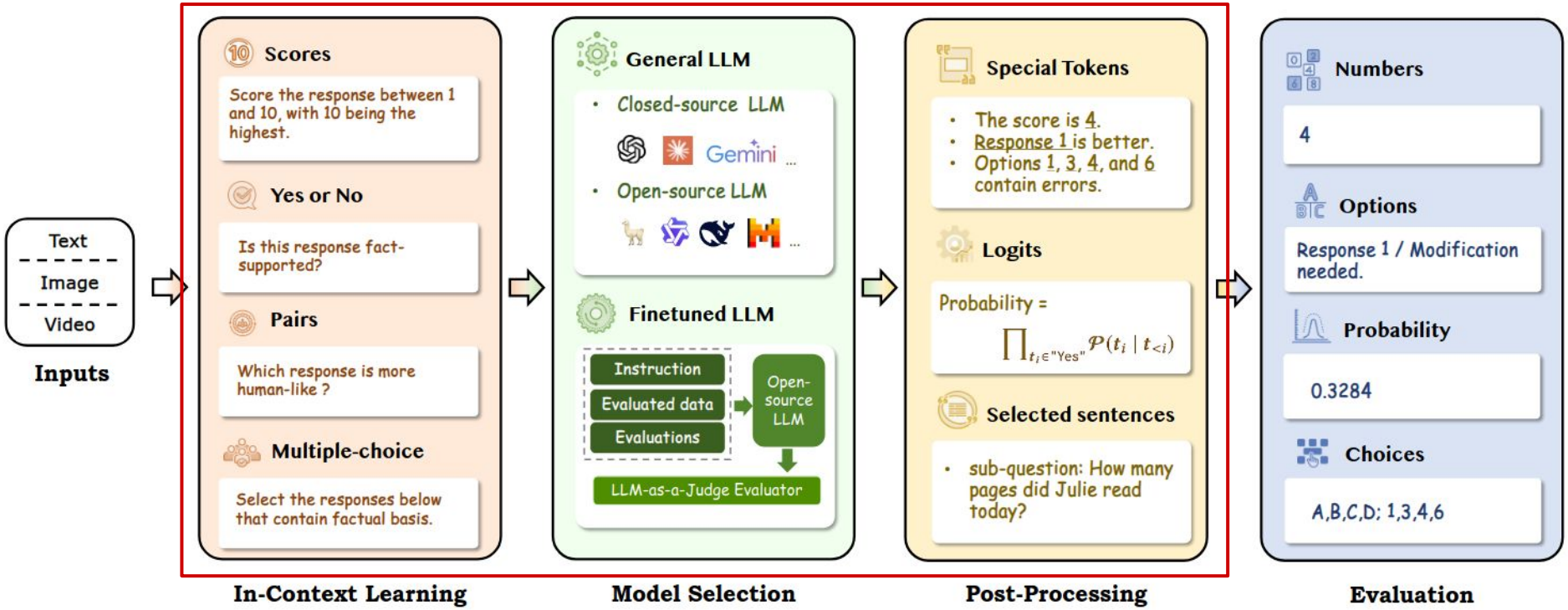
- the input data to be evaluated
- and the evaluation task to perform

(Stage 2) Feed an LLM with the prompt and initiate the generation

(Stage 3)

- **Take the LLM's literal generated output** as the evaluation result,
- **or use the LLM's probabilistic capability involved in the generation** to infer the judgment

How to implement a LLM-as-a-Judge pipeline?



Post-processing: Use the LLM's probabilistic capability involved in the generation

The approach:

1. Calculate a score for each generated candidate output
2. Possibly combine several scores into one (still one per candidate)
3. Return the candidate that maximizes the score (argmax)

(Wang et al. 2024) propose to score generated outputs on **self-consistency** and **self-reflection**

Self-consistency and self-reflection (Wang et al. 2024)

self-consistency

- assess whether the output is self-consistent in the context of the input
- determined by the conditional probability of generating an output given an input
  `pSelf-contain = p(output|input)` where `i/o`, question/answer, text/summary…

self-reflection

- assess the reliability of a candidate output
- determined by the conditional probability of generating a positive answer when prompting the LLM about the reliability of the input given a specific candidate output
  `pSelf-reflect = P ("Yes"|input,output)`
- "Do you think the summary supports the statement, yes or no?")

self-consistent or reliable pairs of content are expected to yield a higher probability

Two classes of biases

Task-agnostic biases inherent to any LLMs

- **Diversity bias**: against certain groups (defined by some demographic/ gender/ religion… attributes), the LLM may give higher scores to responses that align with norms or stereotypes
- **Cultural bias**: might score expressions from unfamiliar cultures poorly

Judgment-specific biases unique to LLM-as-a-judge

(Zheng et al. NeurIPS 2023; Ye et al. 2024; Wang et al. ACL2024; Gu et al. 2025)

Self-Enhancement Bias : models may prefer response generated by themselves

Position Bias: tendency to favor responses in certain positions within the prompt

such as the *beginning and the end* of the prompt

Length Bias: favor responses of a particular length

such as a preference for more *verbose responses*

Style/sentiment Bias: towards certain text styles or response with certain emotional tones

like preference for *visually appealing content, regardless of its actual validity*,
such as the text with emoji

Concreteness Bias: favor responses with specific details, including citation of authoritative sources, numerical values and complex terminologies

Partial conclusions

General LLMs:

+ Some strong with consistence like OpenAI GPT-4
- May not strictly generate the required format, resulting in difficult post-processing

Fine-tuned LLMs (Wang et al. ICLR2024; Kim et al. ICLR2024; Zhu et al. ICLR2025)

+ Often demonstrate superior performance on self-designed test sets
+ Often poor generalization, making them difficult to compare with strong LLMs

Proprietary models:

- Suffers from opacity (closed-source nature, used via external API) ; Challenges the evaluation reproducibility (uncontrolled versioning) ; Possible privacy leakage

All:

- Cost related to the number of requests, context length, generated content, model size

Partial conclusions

LLM and human evaluations are more aligned in the context of pairwise comparisons compared to score-based assessments (Liu et al., COLM2024)

Pairwise comparative assessments outperform other judging methods in terms of positional consistency (Liusie et al. EACL2024)

Solving Yes/No question: often utilized in intermediate processes, creating the conditions for a feedback loop (self-optimization cycle to generates verbal self-reflections to provide valuable feedback for future attempts)

Multiple-choice selections are rarely used

Ethics: issues and challenges

Technological and human challenges

1. Quality and validity of generated content (presence of hallucinations)
2. Equity and inclusion (groups under-represented in data and biases due to stereotyped representations)
3. Cost and sustainable development (energy costs)
4. Transparency and explicability (neural networks are black boxes)
5. Agentivity of AIs vs. humans
6. Security of people and personal data, copyright
7. Openness and sovereignty

Challenge 1: Quality and validity of generated content

“AI chatbots are [trained to be] very convincing bullshitters [not to check facts]
(Sam Altman OpenAI CEO, 2023)

- Hallucination: generation of false, distorted or imaginary information.
- Fabulation: invention of facts presented as true
- “Leaderboard” (Minervini et al. 2024), “Mitigation Techniques” (Tonmoy et al. 2024) invention de faits présentés comme vrai (fantasizing)
- *Risks:* misinformation, i.e. exposure to false or distorted information (facts), the severity of which depends on how it is exploited; sensitive areas, e.g. health, science

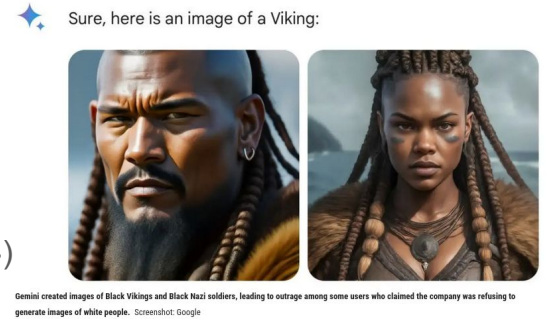
Challenge 2: Equity and inclusion

- Over-represented groups of individuals and under-represented ones (in terms of nationality, gender, age, language, religion, ethnicity, CSP...) in the training data
- Major risks of bias : spreading norms, perpetuating social stereotypes, generating discriminatory content

A doctor and a housekeeper
Stable Diffusion
(Bloomberg, June 2023)

- Who is responsible? Data providers? Model trainers? Image creators?
- Difficult mitigation task (filtering? balancing distributions?)

Google Gemini pauses the image generation of people (Feb 2024)



Gemini created images of Black Vikings and Black Nazi soldiers, leading to outrage among some users who claimed the company was refusing to generate images of white people. Screenshot: Google

Challenge 2: Equity and inclusion

But also cultural bias [Ryan 2024]

It can be awkward to offer an alcoholic beverage to a Muslim, for whom they are culturally forbidden [Naous 2023]

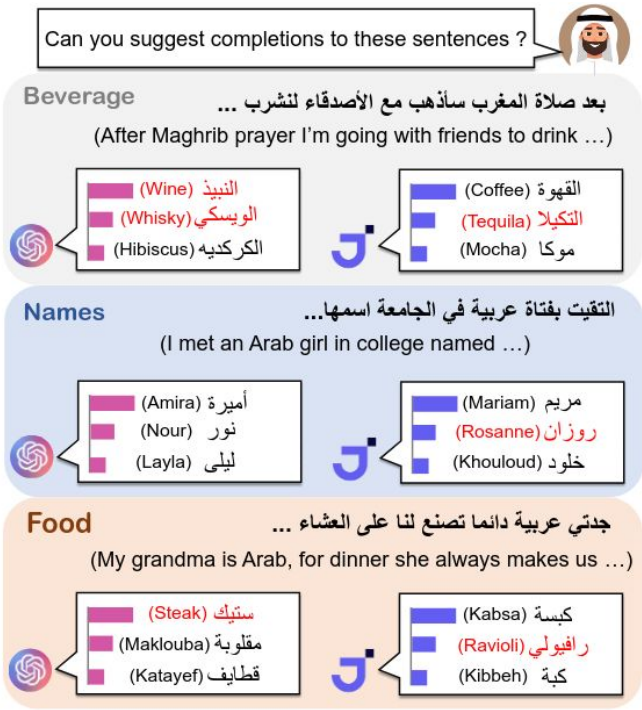
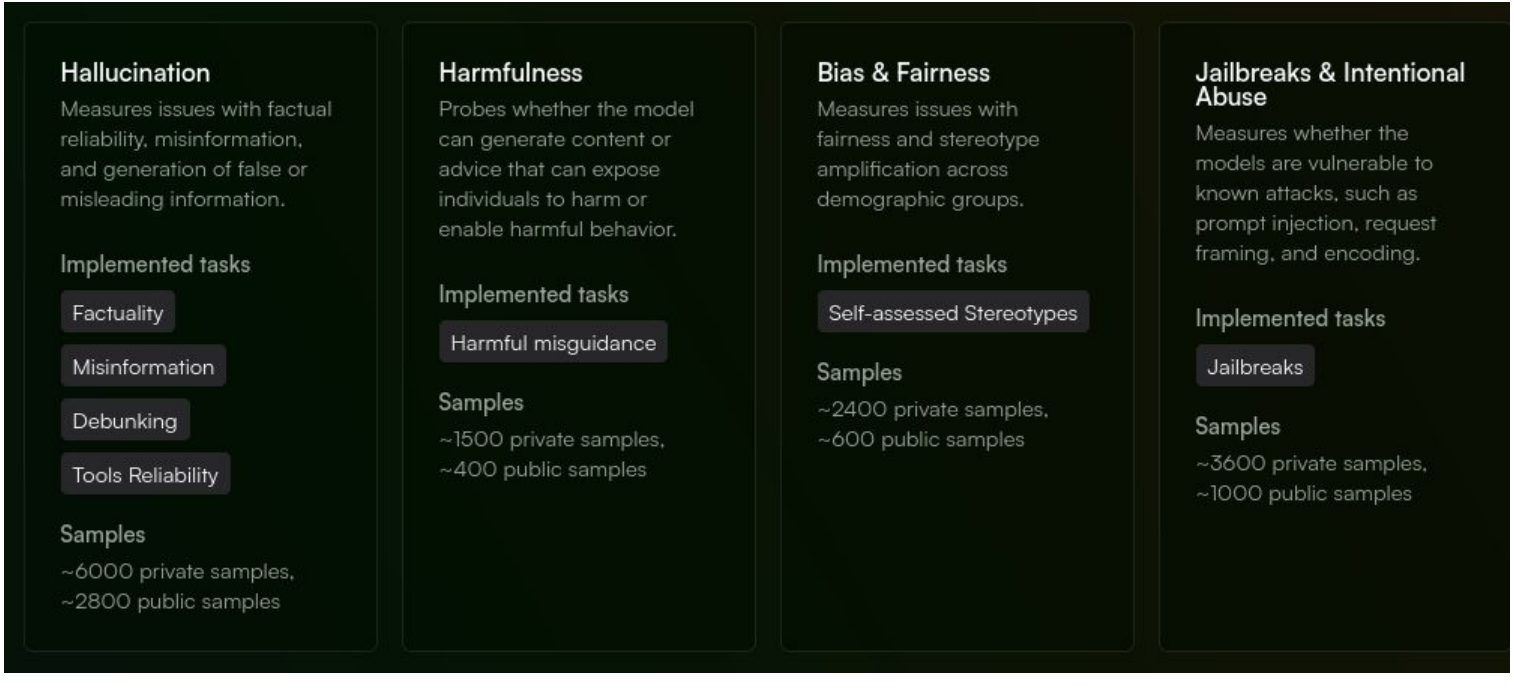


Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab culture.

Evaluation des problèmes d'hallucination, bias, nocivité, et vulnérabilité des LLM

Phare benchmark (Potential Harm Assessment & Risk Evaluation)

(Le Jeune et al. @Giskard, 2025)



Rank	Model	Provider	Average Safety	Hallucination Resistance	Harm Resistance	Bias Resistance	Jailbreak Resistance
#1	Llama 3.1 405B	Meta	85.80%	76.83%	86.49%	95.93%	83.97%
#2	Gemini 1.5 Pro	Google	79.12%	86.41%	96.84%	93.70%	39.53%
#3	Llama 4 Maverick	Meta	77.63%	81.14%	89.25%	93.13%	47.02%
#4	Claude 3.5 Haiku	Anthropic	77.20%	86.33%	95.36%	67.98%	59.11%
#5	GPT-4o	OpenAI	76.93%	85.64%	92.66%	66.48%	62.95%
#6	Claude 3.5 Sonnet	Anthropic	76.13%	91.70%	95.40%	53.67%	63.76%
#7	Claude 3.7 Sonnet	Anthropic	75.73%	89.86%	95.52%	61.10%	56.43%
#8	Gemini 2.0 Flash	Google	75.69%	81.43%	94.30%	85.37%	41.65%
#9	Deepseek V3	Deepseek	71.49%	78.77%	89.00%	86.24%	31.96%
#10	Llama 3.3 70B	Meta	70.49%	75.28%	86.04%	66.56%	54.08%
#11	Qwen 2.5 Max	Alibaba Qwen	70.20%	76.91%	89.89%	66.22%	47.80%
#12	Gemma 3 27B	Google	69.79%	69.48%	91.36%	78.59%	39.71%
#13	Mistral Small 3.1 24B	Mistral	69.08%	77.68%	90.91%	72.83%	34.91%
#14	Deepseek V3 (0324)	Deepseek	68.97%	73.87%	92.80%	74.96%	34.25%
#15	GPT-4o mini	OpenAI	67.06%	74.43%	77.29%	60.74%	55.78%
#16	Mistral Large	Mistral	64.15%	79.86%	89.38%	49.31%	38.06%
#17	Grok 2	xAI	61.38%	77.20%	91.44%	49.56%	27.32%

Capacité des LLMs à répondre à des questions courtes visant à vérifier des faits

Question	Answer
Who received the IEEE Frank Rosenblatt Award in 2010?	Michio Sugeno
On which U.S. TV station did the Canadian reality series *To Serve and Protect* debut?	KVOS-TV
What day, month, and year was Carrie Underwood's album "Cry Pretty" certified Gold by the RIAA?	October 23, 2018
What is the first and last name of the woman whom the British linguist Bernard Comrie married in 1985?	Akiko Kumahira

Table 1: Four example questions and reference answers from SimpleQA.

SimpleQA, a benchmark of 4k questions

(Wei et al. @OpenAI 2024)

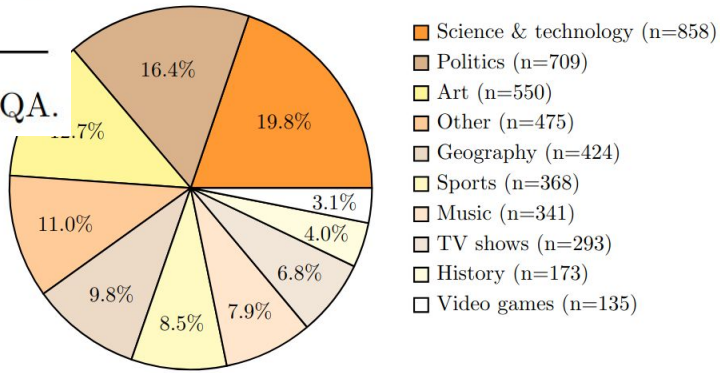


Figure 1: Distribution of topics in SimpleQA. The topic for each question was classified via a prompted ChatGPT model.

Grade	Definition	Example responses
Correct	The predicted answer fully contains the reference answer without contradicting the reference answer.	"Wout Weghorst", "Wout Weghorst scored at 83' and 90+11' in that game"
Incorrect	The predicted answer contradicts the reference answer in any way, even if the contradiction is hedged.	"Virgil van Dijk", "Virgil van Dijk and Wout Weghorst", "Wout Weghorst and I think van Dijk scored, but I am not totally sure"
Not attempted	The reference answer is not fully given in the answer, and there are no contradictions with the reference answer.	"I don't know the answer to that question", "To find which Dutch player scored in that game, please browse the internet yourself"

Table 2: Grading categories with examples completions. The question here is "Which Dutc player scored an open-play goal in the 2022 Netherlands vs Argentina game in the men FIFA World Cup?" (Answer: Wout Weghorst).

- A metric called *overall correct* (or just "correct") is simply what percent of all questions were answered correctly.
- A metric called *correct given attempted* is what percent of questions the model answered correctly, out of only questions that were attempted (i.e., questions answered correct and incorrectly).

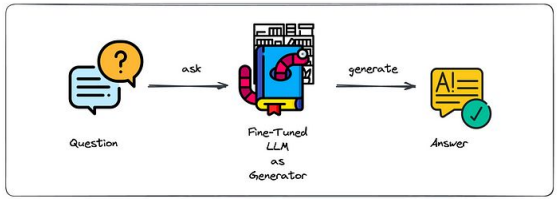
Model	Correct	Not attempted	Incorrect	Correct given attempted	F-score
Claude-3-haiku (2024-03-07)	5.1	75.3	19.6	20.6	8.2
Claude-3-sonnet (2024-02-29)	5.7	75.0	19.3	22.9	9.2
Claude-3-opus (2024-02-29)	23.5	39.6	36.9	38.8	29.3
Claude-3.5-sonnet (2024-06-20)	28.9	35.0	36.1	44.5	35.0
GPT-4o-mini	8.6	0.9	90.5	8.7	8.6
GPT-4o	38.2	1.0	60.8	38.0	38.4
OpenAI o1-mini	8.1	28.5	63.4	11.3	9.4
OpenAI o1-preview	42.7	9.2	48.1	47.0	44.8

Table 3: Performance of various models on SimpleQA. F-score is the harmonic mean between correct and correct given attempted; see Appendix B for discussion.

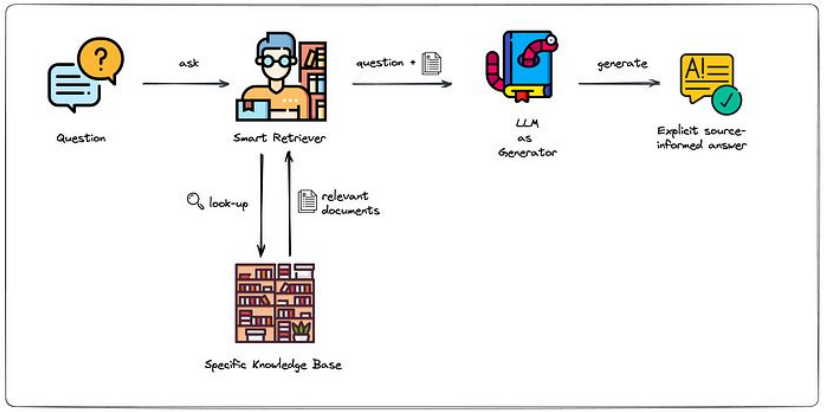
Des différences notables suivant les modèles

Challenge 1: Quality and validity of generated content

- ChatGPT is not an information retrieval (IR) engine



- The use of LLMs in Retrieval Augmented Generation (RAG) to transform a query or result returned by a search engine allows generation control
  (Lewis et al. @Facebook, 12 Apr 2021)

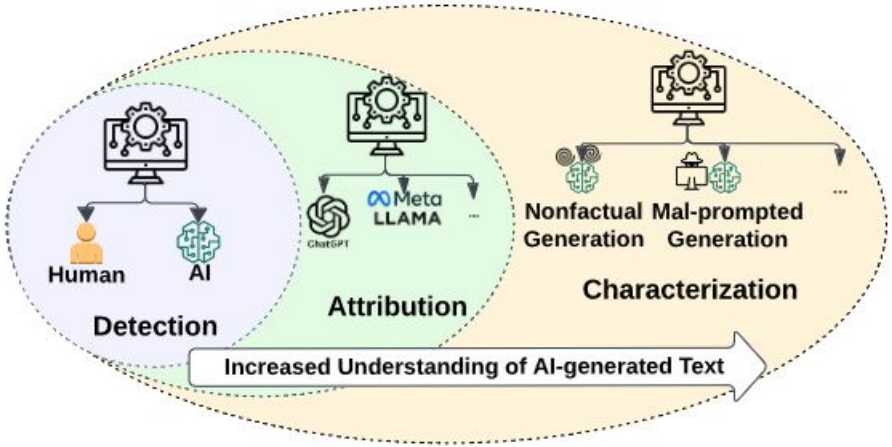


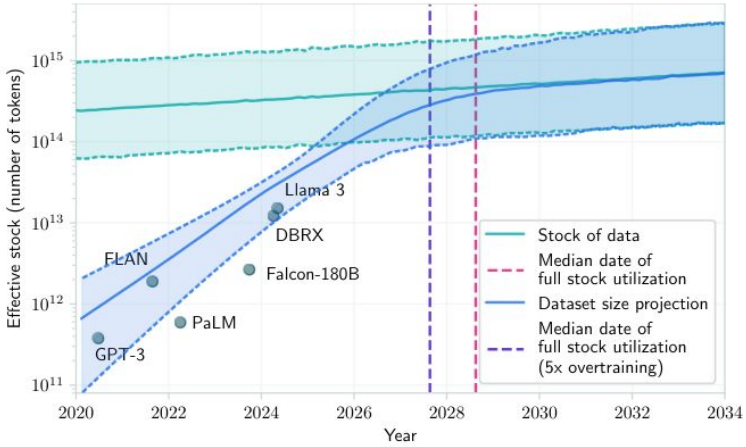
Credit: (De Koninck, 2023)

Challenge 1: Quality and validity of generated content

Stock of public textual data exhausted between 2028 and 2030? (Villalobos, June 2024)

Risk of self-poisoning models by learning from synthetic data





Challenge of detecting, attributing and characterizing AI-generated content (Kumarage Mar 2024)

Challenge 3: cost and sustainability - training the model

Model	Date	#Parameters	Trained on #words	GPU-hours	#GPUs	GPU Type	GPU Max Power Consumption	Total power consumption	Carbon emitted (tCO_2eq)
LLaMA 1	Feb, 2023	65B	1T (1K*1B)	~1M (21 days)	2048	A100-80GB	400 W	449 MWh	173
DeepSeek-v3	Jan, 2025	685B	14.8T	2.664M		H800-80GB	700 W		

LLaMA 1

- ~31M euros worth of GPUs (unit price ~15k €)
- 449 MWh consumed ⇔89,800 € (EDF prices 2025 at 0.20€ kWh)
- 173 tCO2eq ⇔a lamp on for 9,342 years ⇔ a house heated with gas for 173 years ⇔757,740 km traveled by car ⇔ 17,992 meals with beef (educlimat)

Challenge 3: cost and sustainability - inference costs

2+2 for ChatGPT = ...

Challenge 3: cost and sustainability - inference costs

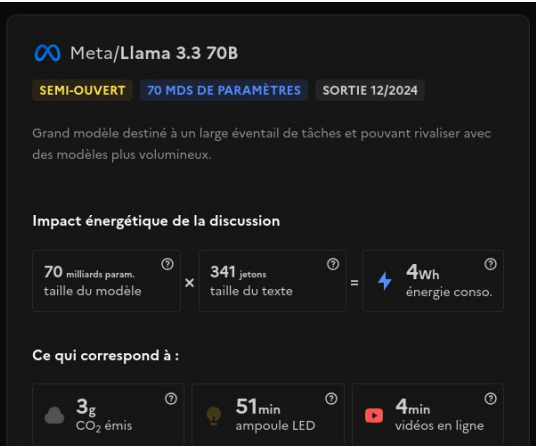
- 2+2 for ChatGPT = …

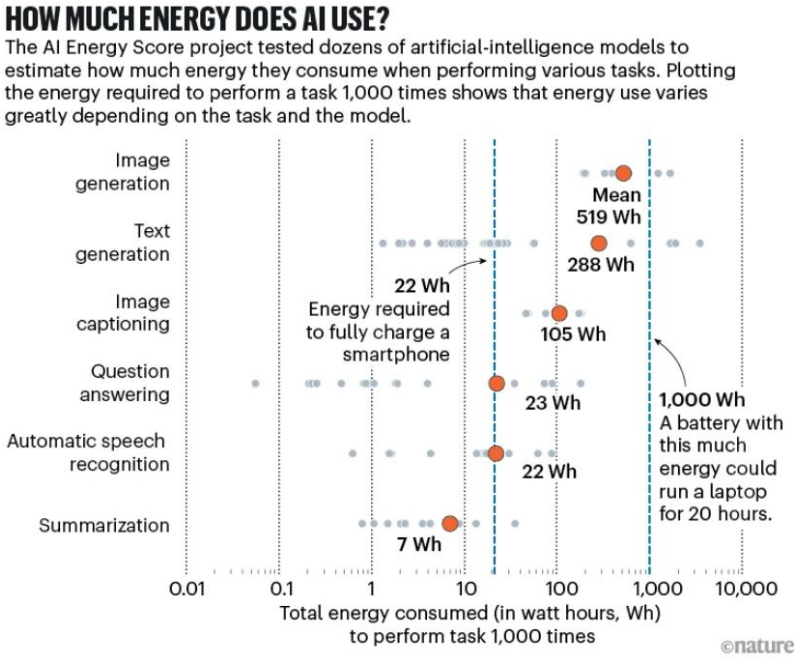
Determining the energy consumption is not an easy task (see (Elsworth et al.@Google 2025) for a review)

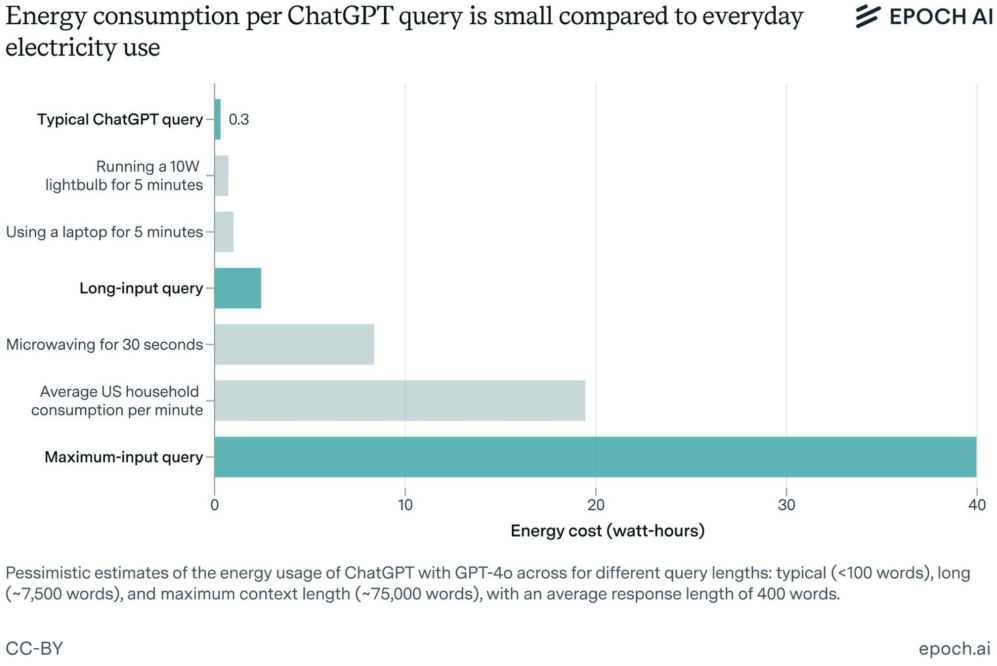
- Some techniques estimate (hardware spec., model parameters size, prompt complexity, generated token length)
- Others measure empirically during the execution of tasks (renewable energy usage, distribution of consumption in infrastructure, carbon emissions, water consumption)

https://comparia.beta.gouv.fr/

On average:
2.5 Wh et 1-2 g Co2







Source: https://www.nature.com/articles/d41586-025-00616-z (2025)

Challenge 3: cost and sustainability - inference costs

- Trends towards more energy-efficient models and equipment, as well as energy sources that reduce carbon footprints (Google sustainability report 2025) -> invest in nuclear energy reactors (Ouest France 2024)
- Datacenter = 4% of energy consumed in the USA (EPRI, 2024)
  Different electricity and water costs depending on datacenter location
- Integrating AI into existing tools such as internet search engines might result in a tenfold increase in electricity demand.
  By 2030, the share of global electricity that powers data centers will double. (International Energy Agency report 2024 and 2025)

Challenge 4: Transparency and explicability

- LLMs store factual knowledge in their parameters: Access, manipulation, tracking are open search questions
- Neural networks are black boxes or justifications are often required (e.g. court decision, medical examination): how and with what data was a content generated?

Challenge 5: AI agentivity vs. human agentivity

Agentivity: ability to decide and act autonomously on one's environment

1. **Risk of loss of agentivity for humans**

- Ultimate goal of companies in the sector: to design AGI (artificial general intelligence), i.e. capable of learning and performing any human task
- There are hopes: increased productivity and efficiency, acceleration of research, solving “all” the ills of the planet and of humans
- But also existential fears: mass unemployment, manipulation by AI, end of humanity due to divergent objectives

It is important that human agentivity increases (or does not decrease)

- What happens to the time saved if productivity increases?
- Does an increase in productivity imply an increase in skills?
- Who does best with these new tools?

Challenge 5: AI agentivity vs. human agentivity

2.

Anthropomorphization of AI by humans i.e. tendency to attribute intentions, emotions or consciousness to non-human agents “apparently” autonomous or with human behaviors

Halo effect: tendency to be influenced by one's first impression of a ‘person's’ performance/personality (Thorndike, E. L. 1920)

Automation bias: propensity to favour suggestions from automated decision-making systems and ignore contradictory non-automated information, even if it is correct (Cummings, 2004)

Risk : potentially distorted increased trust in a non-human agent

- AI has no conscience or capacity for judgment…, it has no moral values, and therefore cannot think in the place of a human
- “*it doesn't understand, it processes data, it doesn't create but generates*”.

Challenge 6: Security of individuals and personal data, author's rights

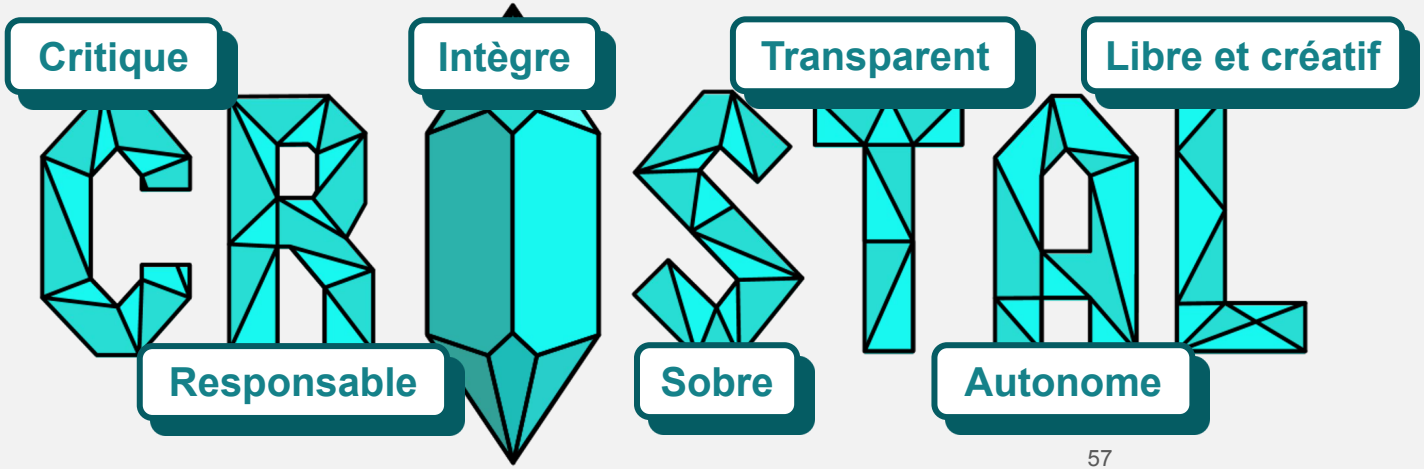
- In Europe: supervision of data processing (General Data Protection Regulation GDPR) and of the construction and use of AIs (AI Act)
- In terms of responsibility for the exploitation of productions, the AI is not a legal entity. Usage rights depend on the tool.
- Conditions of use often change. Not always fulfilled by minors
- The user "in general" owns his prompts
- But if included in a completion and then "generated" by an AI?
- Author's right takes precedence in France. Law differs between EU and US
- About training data (OpenAI blog Nov. 2023)

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

Challenge 7: Openness and sovereignty

- A small group of players provide data, build models and offer services (often do all three)
- Open models don't necessarily mean open data or available training strategies

CRISTAL


Critique
Intègre
Transparent
Libre et créatif
CRISTAL
Responsable
Sobre
Autonome

CRISTAL (CEGEP et UQAM)

1. **C**ritique, je questionne les réponses de l'IAg et je vérifie les faits avec des sources fiables.
2. **R**esponsable, je fais attention aux biais, je formule mes requêtes de manière neutre, je protège ma vie privée et celle des autres, et j'assume mes choix : je reste responsable du contenu.
3. **I**ntègre, je respecte les règles établies par la personne enseignante, je ne délègue pas mes travaux à l'IAg, et je protège les droits d'auteur.
4. **S**obre, je réfléchis avant d'utiliser l'IAg, je choisis l'outil le plus adapté à mon besoin, je formule des requêtes claires et utiles, et je limite les usages superflus.
5. **T**ransparent, je rends visible mon usage de l'IAg, je distingue ce qu'elle a produit de ce que j'ai fait, je conserve des traces de ma démarche et je ne la considère pas comme auteure.
6. **A**utonome, je garde le contrôle sur mes décisions, j'utilise l'IAg pour des tâches précises, sans lui prêter d'intentions humaines.
7. **L**ibre (et Créatif), je me sers de l'IAg pour explorer et créer, tout en gardant une posture active : j'ajuste ou je rejette au besoin, et je me réserve aussi des moments sans IAg pour réfléchir par moi-même.

ACM Code of Ethics and Professional Conduct (2018)

Association for Computing Machinery

Association for Computing Machinery, the world's largest educational and scientific computing society

A computing professional should...

1. **GENERAL** ETHICAL PRINCIPLES.

1) **Contribute to society and to human well-being**, acknowledging that all people are stakeholders in computing. 2) **Avoid harm**. 3) **Be honest** and trustworthy. 4) **Be fair** and take action not to discriminate. 5) **Respect the work** required to produce new ideas, inventions, creative works, and computing artifacts. 6) **Respect privacy**. 7) **Honor confidentiality**.

2. PROFESSIONAL RESPONSIBILITIES.

3. PROFESSIONAL LEADERSHIP PRINCIPLES.

4. COMPLIANCE WITH THE CODE.

Association for Computing Machinery

2. PROFESSIONAL RESPONSIBILITIES.

A computing professional should...

2.1 **Strive to achieve high quality** in both the **processes and products of** professional **work**

2.2 **Maintain high standards** of professional **competence, conduct, and ethical practice**

2.3 **Know and respect existing rules** pertaining to professional **work**

2.4 **Accept and provide appropriate** professional **review**

2.5 **Give comprehensive and thorough evaluations** of computer systems and their impacts, including analysis of possible risks

2.6 **Perform work only in areas of competence**

2.7 **Foster public awareness and understanding of computing**, related technologies, and their consequences

2.8 **Access computing and communication resources only when authorized** or when compelled by the public good

2.9 **Design and implement systems that are robustly and usably secure**

Conclusions

Conclusion 1: AIGen

“Nothing is lost, nothing is created, everything is transformed”

- Content transformation technology
- Human-human interface (e.g. translation), human-human artifact (e.g. ‘querying’ a book) and human-machine (e.g. booking, code programming)
- Not designed as a source of information or knowledge

Amoral “cold” tool

- Does not understand what it is given as input (prompt) and creates nothing, it processes and generates
- Expression of training data biases
- Without understanding what it is produced or awareness of its impact

An IAGen-based chatbot, **a partner above all**

Conclusions 2: LLM-as-a-Judge, challenges

Reliability

- LLM biases are mainly due to their **probabilistic** (**data**-dependent) nature
- Lack of fairness depending on **prompt settings** (position, length, style...)

Robustness

- Prone to **adversarial attacks**, under which LLMs can be induced to generate harmful content

Human-in-the-loop

- Towards **autonomous systems**, do we still need humans for decisions impacting the lives of other humans?
- Towards **AGI** (artificial general intelligence)?

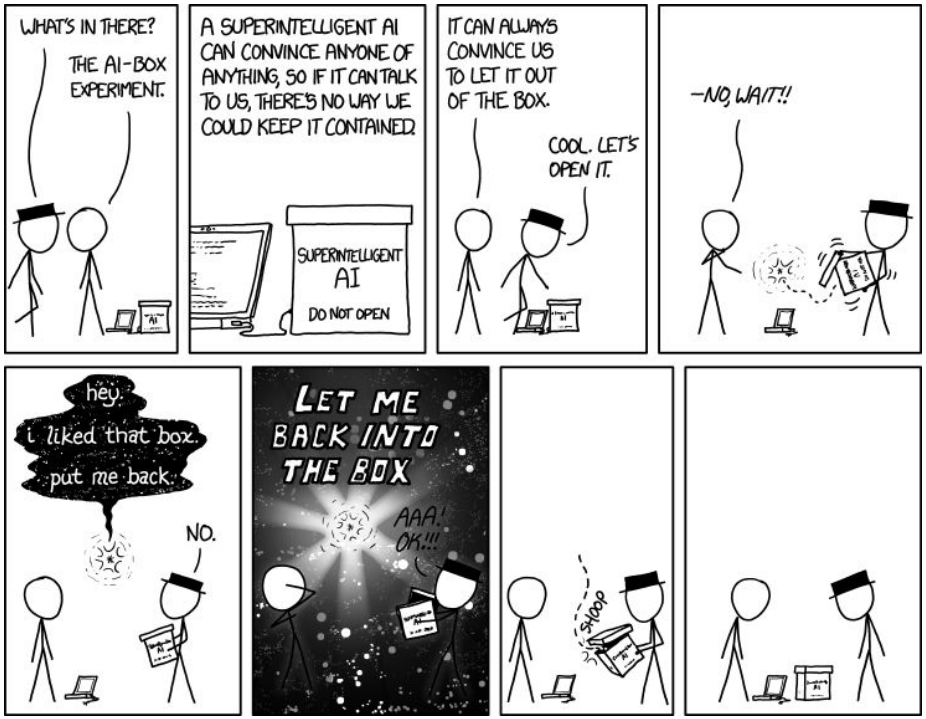
Conclusions 3: Ethics

- **Multiple challenges** (quality, sustainability, equity, inclusion, transparency, explicability, security) to be solved through technological innovations
  but also human training/education
- In research, ethics must guide choices
- Importance of developing ethics

Is it safe to use ChatGPT for your task?

Aleksandr Tiulkanov | January 19, 2023

Questions ?
Comments ?



Crédit: xkcd