LES LLM ONT-ILS UNE CONSCIENCE?

DES MODÈLES DE LANGUE À L'IA FORTE

Lundi 17 novembre 2025
AgroParisTech

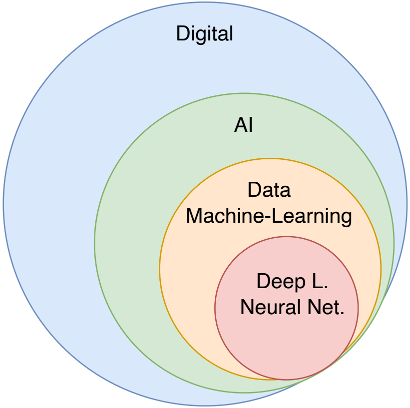
Vincent Guigue
vincent.guigue@agroparistech.fr
https://vguigue.github.io

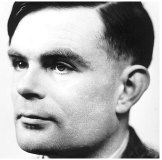
From AI to deep-learning

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



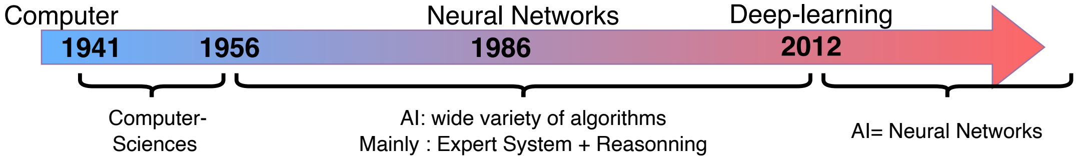

A. Turing


Marvin Minsky

G. Hinton

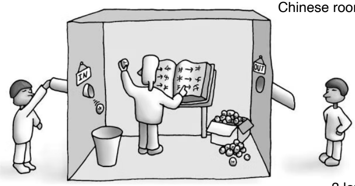


Y. Lecun

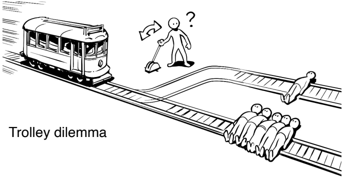


Artificial Intelligence: many representations

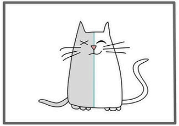
Chinese room



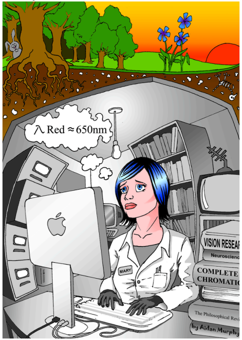
Trolley dilemma



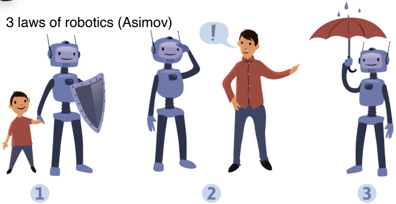
Schrodinger's cat



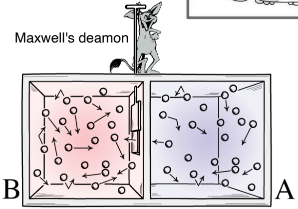
Mary's room



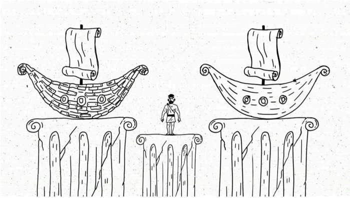
3 laws of robotics (Asimov)



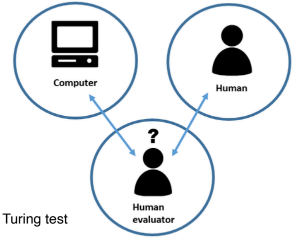
Maxwell's deamon



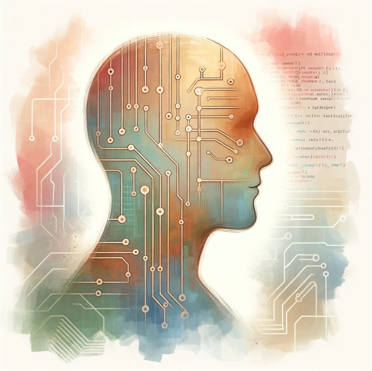
Thesee's boat



Turing test



Artificial Intelligence & Machine Learning



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

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

Marvin Lee Minsky, 1956

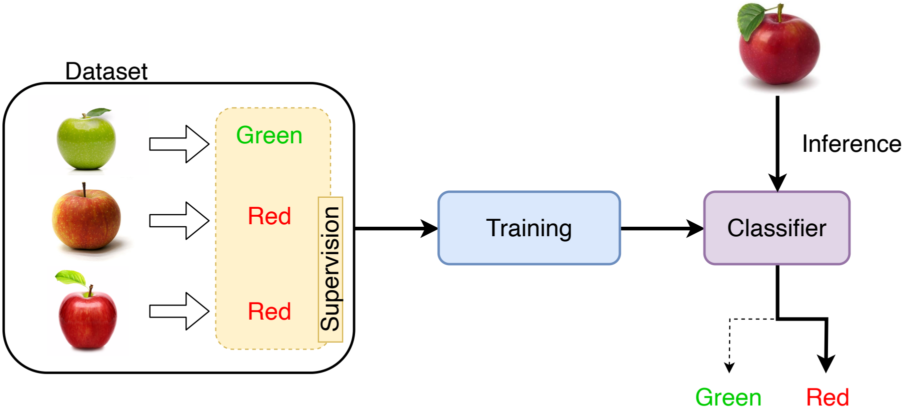
N-AI (Narrow Artificial Intelligence), dedicated to a single task

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

Andrew Ng, 2015

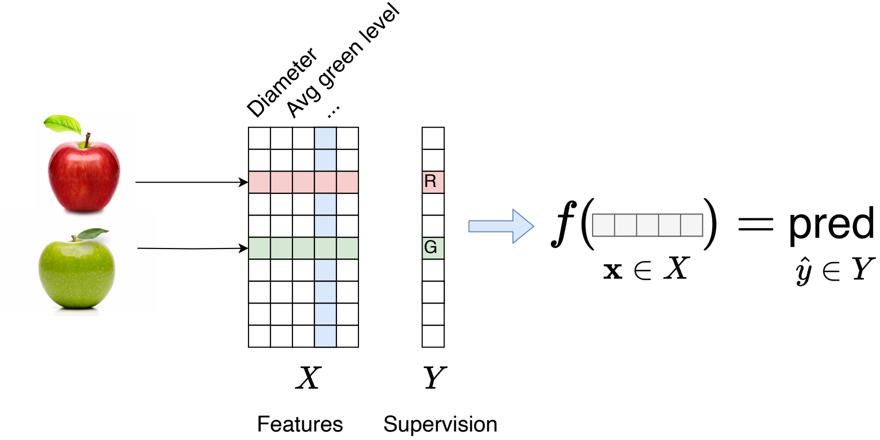
Machine Learning Definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model



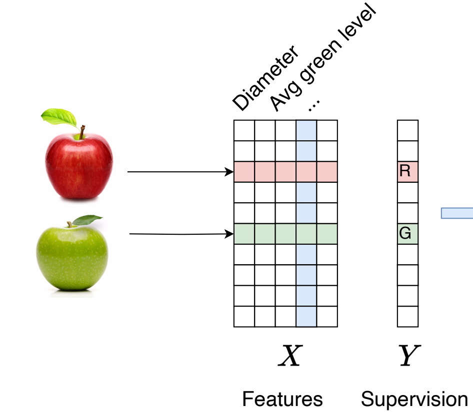
Machine Learning Definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model

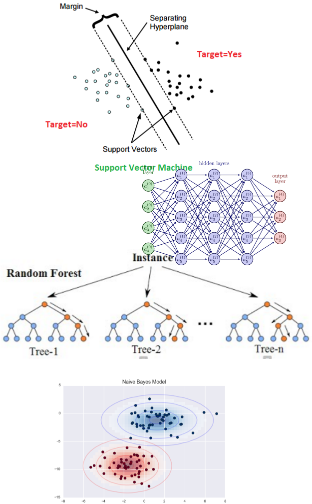


Machine Learning Definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model

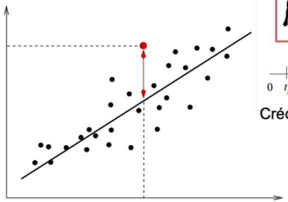


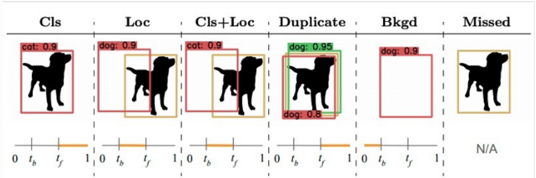
$$\underset{\mathbf{x} \in X}{f(\boxed{\quad\quad\quad})} = \underset{\hat{y} \in Y}{\text{pred}}$$



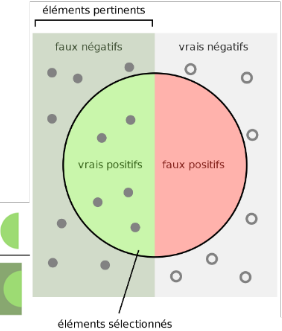
Measuring Performance

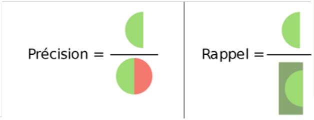
Estimating performance (in generalization)... as important as training the model!

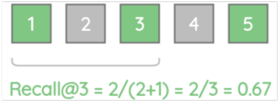


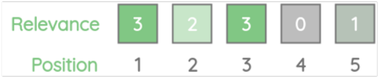


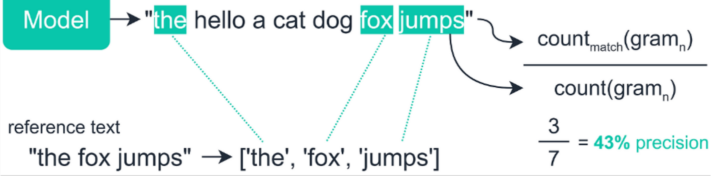
Crédit: https://github.com/phalanx-hk/eccv2020_paperlist/issues/5





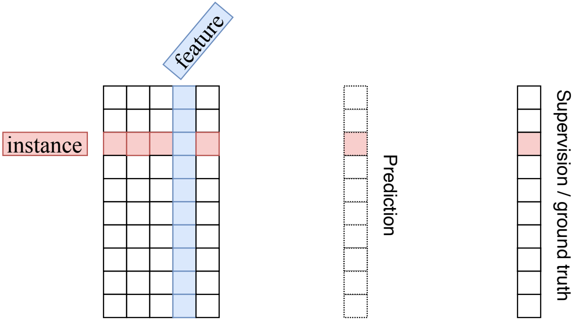






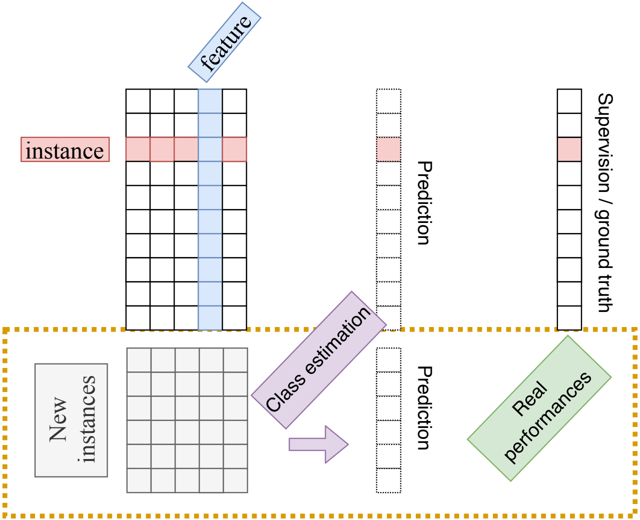
Measuring Performance

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



Measuring Performance

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



General AI *vs* Narrow AI

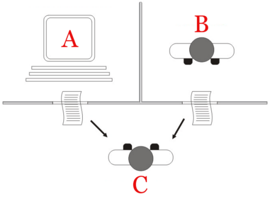
Narrow AI

Like any computer science project:

- Define Inputs & Outputs
- Break down into subtasks
- Build & test components (processing chain)
- Assert (limited) generalization (iid assumption)
- Performances Evaluation

General AI

- Augmented Generalization Capability (Universality)
- Autonomous Learning
  - Data/information access
  - Knowledge extraction (Training+Eval+Confidence/Trust)
- Reasoning
- Conscience, Intentionality

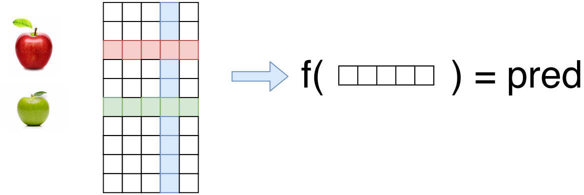


Turing test

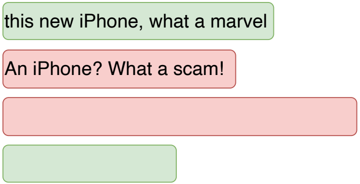
Wikipedia

From tabular data to text

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



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



AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

Rule-based Systems:

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

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

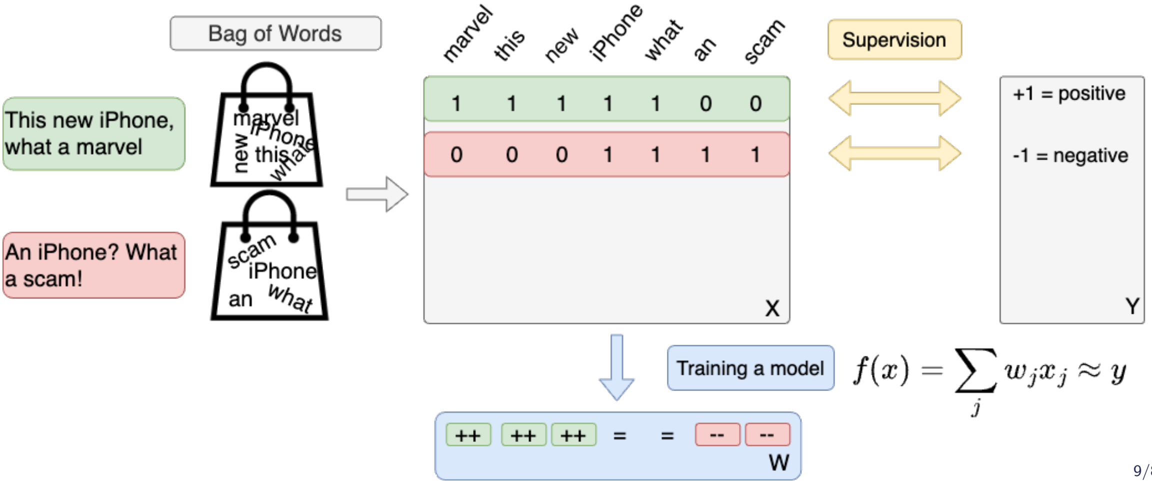
* $\longrightarrow$ {hate, loathe, detest} $\rightarrow$ * $\rightarrow$ #product

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

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Machine Learning [1990-2015]



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

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

- Requires expert knowledge
- Rule extraction ⇔ very clean data

+ Interpretable system
+ Very high precision
− Low recall

Machine Learning [1990-2015]

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

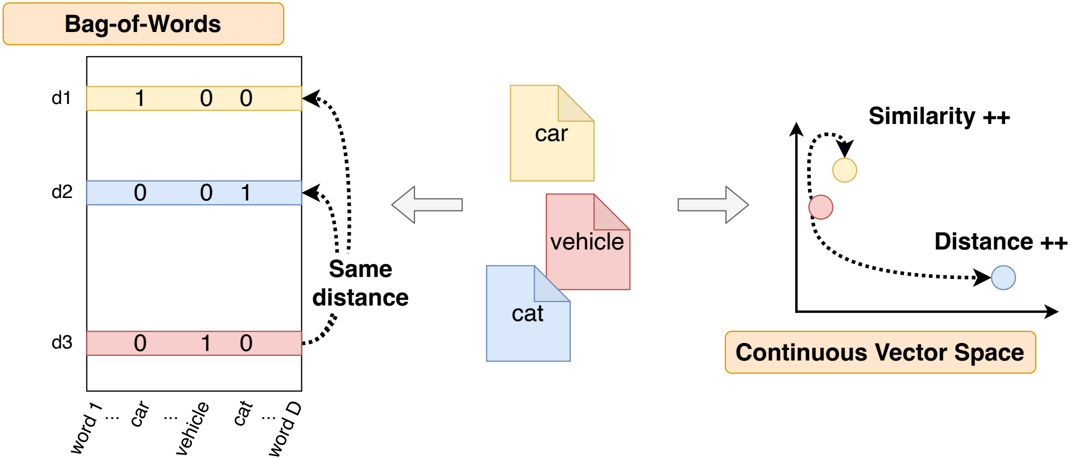
≈ Less interpretable system
− Lower precision
+ Better recall

Precision = criterion for acceptance by industry

→ Link to metrics

Deep/Representation Learning for Text Data

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

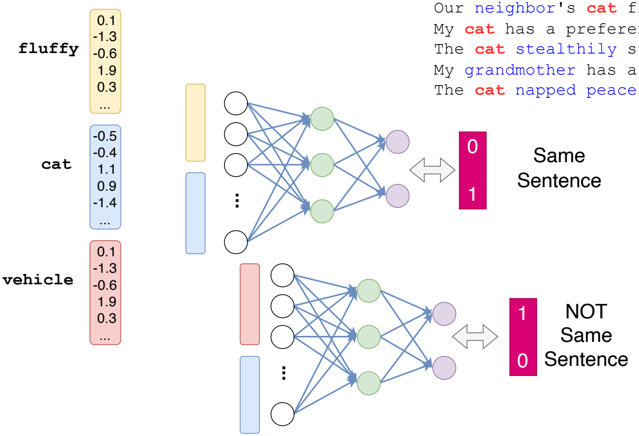


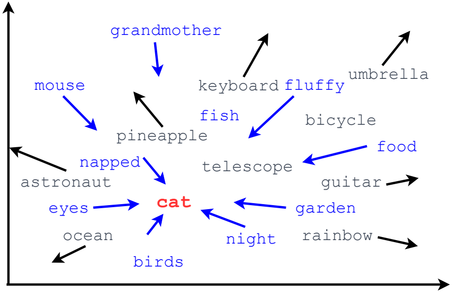
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Deep/Representation Learning for Text Data

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

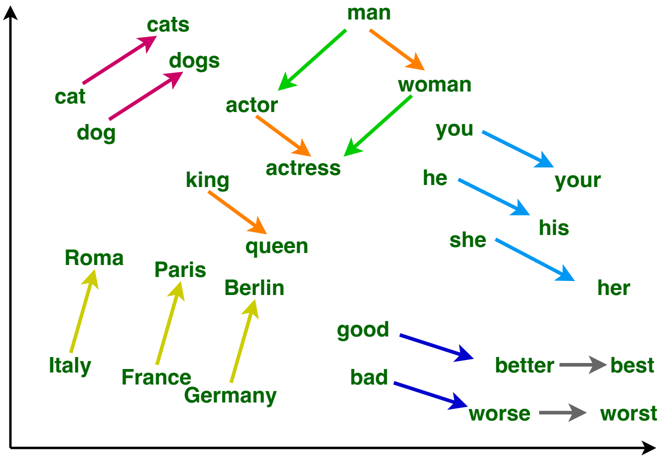
The fluffy **cat** napped lazily in the sunbeam.
I adopted a stray **cat** from the shelter last week.
My **cat** loves to chase after toy mice.
The black **cat** stealthily crept through the dark alley.
I often find my **cat** perched on the windowsill, watching birds.
She gently stroked her **cat**'s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





Deep/Representation Learning for Text Data

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



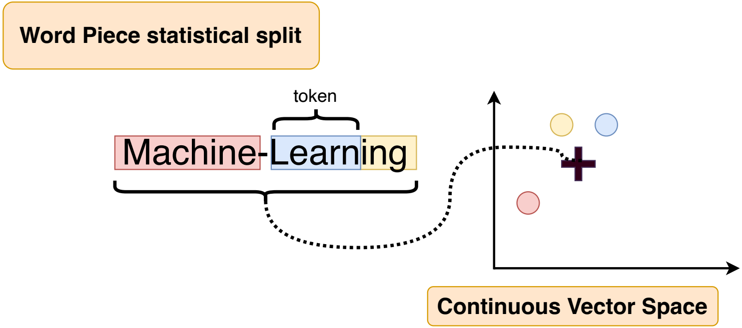
- Semantic Space:
  similar meanings
  $\Leftrightarrow$
  close positions
- Structured Space: grammatical regularities, basic knowledge, ...

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

Deep/Representation Learning for Text Data

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

From Words to Tokens

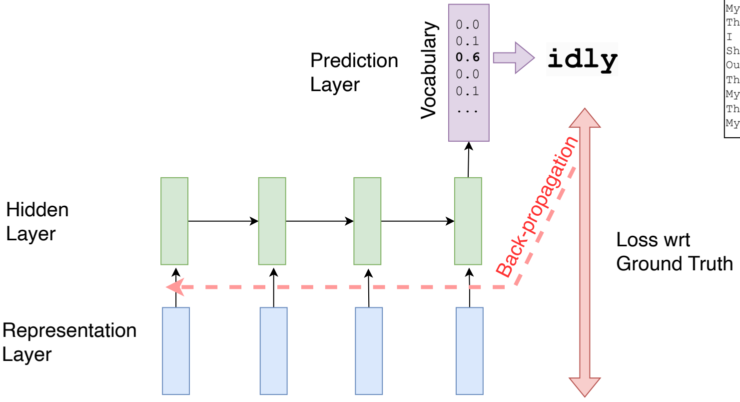


- Representation of unknown words
- Adaptation to technical domains
- Resistance to spelling errors

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

Aggregating word representations: towards generative AI

- Generation & Representation
- New way of learning word positions



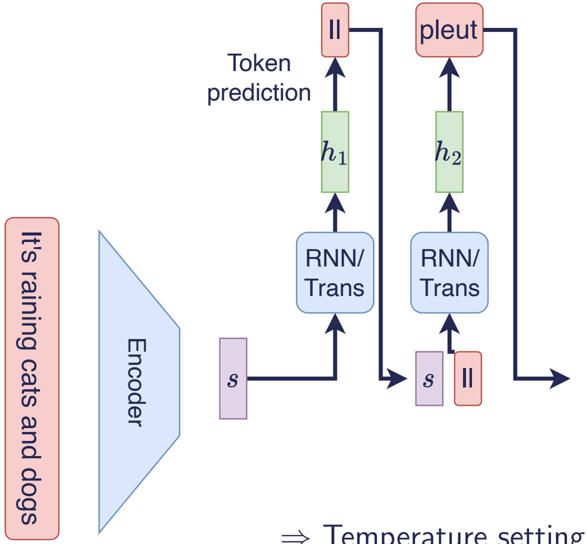
The fluffy cat napped lazily in the sunbeam.

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

Corpus

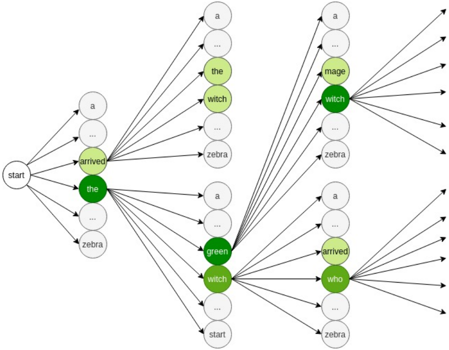
Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

Inference & Beam Search



⇒ Temperature setting

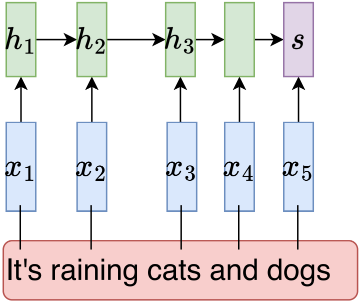
- High cost ≈ 1 call / token
- Max. likelihood principle
- NLP historical task =
  - specific classif./scoring archi.
  - constraint and/or post processing on generative archi.



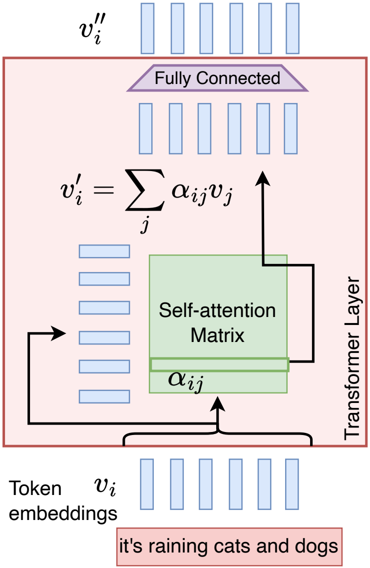
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

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



Transformer:



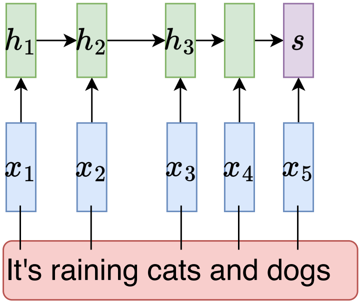
Attention is all you need, Vaswani et al. NeurIPS 2017

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

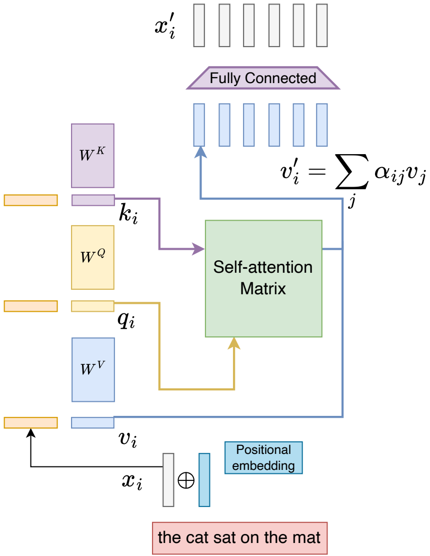
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

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



Transformer:



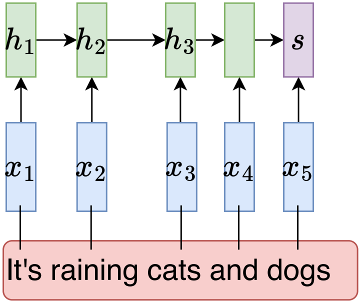
Attention is all you need, Vaswani et al. NeurIPS 2017

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

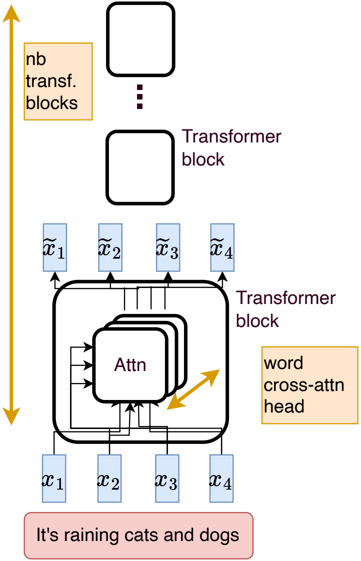
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

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



Transformer:

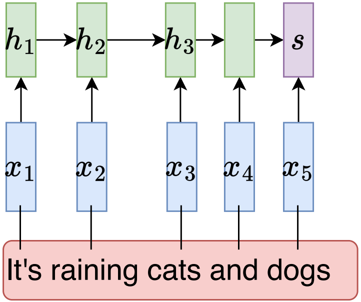


Attention is all you need, Vaswani et al. NeurIPS 2017

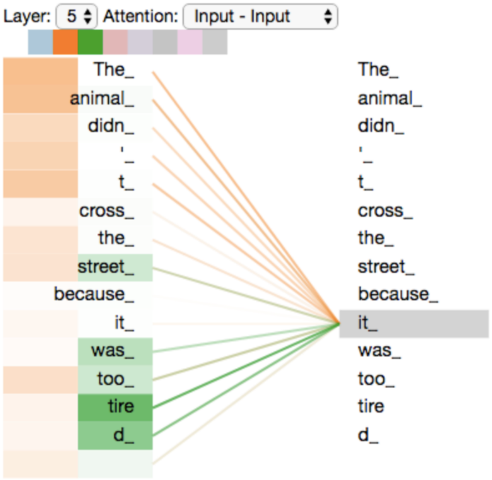
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

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



Transformer:

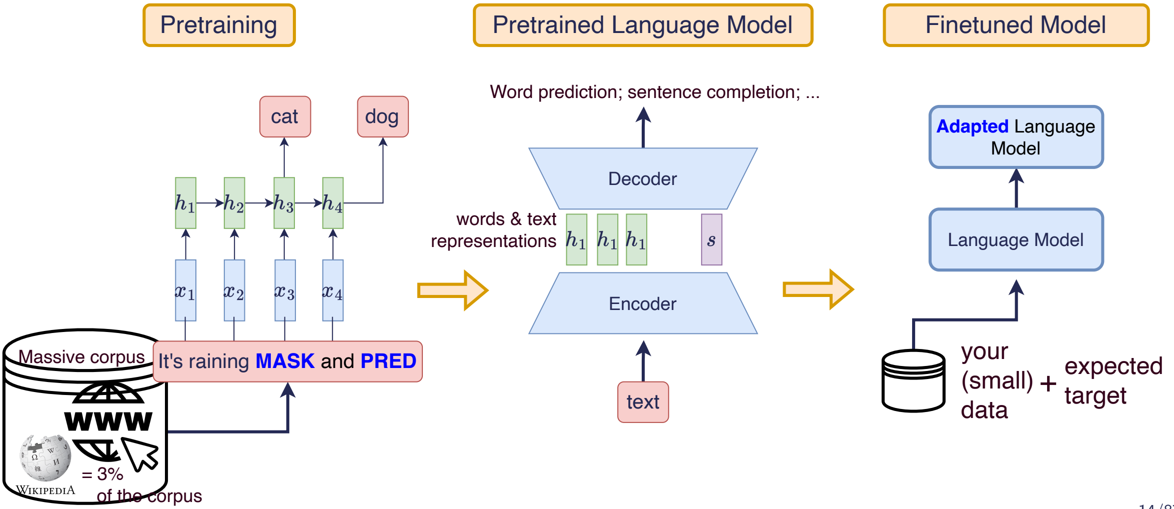


Attention is all you need, Vaswani et al. NeurIPS 2017

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

A new developpement paradigm since 2015

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



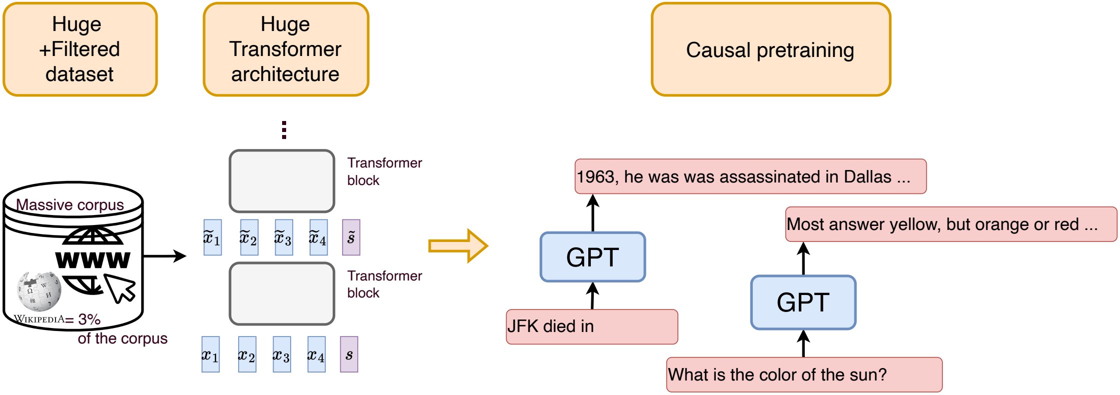
chatGPT

November 30, 2022

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

The Ingredients of chatGPT

0. Transformer + massive data (GPT)



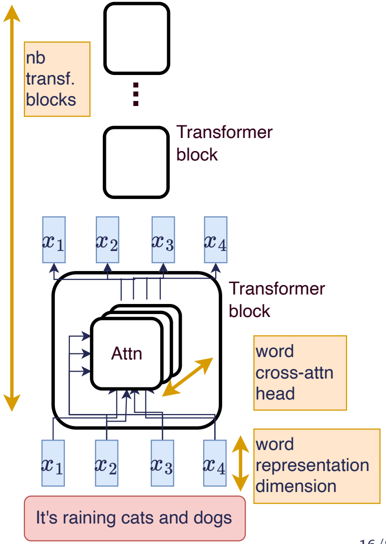
- Grammatical skills: singular/plural agreement, tense concordance
- Knowledges: entities, names, dates, places

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

The Ingredients of chatGPT

1. More is better! (GPT)

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

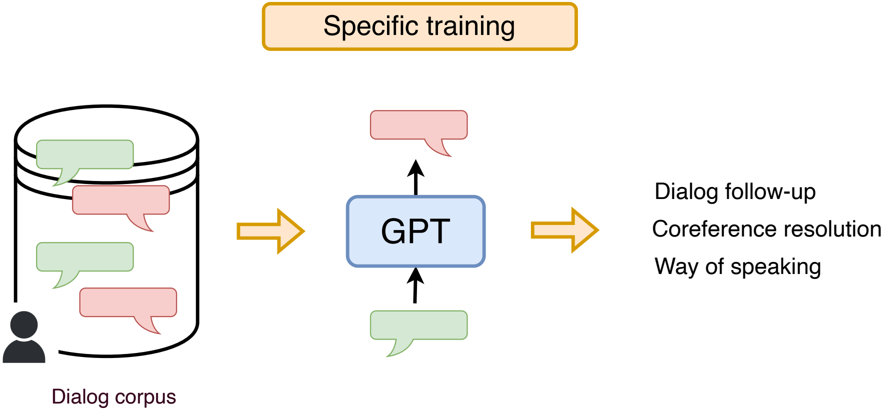


175 Billion parameters... What does it mean?

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

The Ingredients of chatGPT

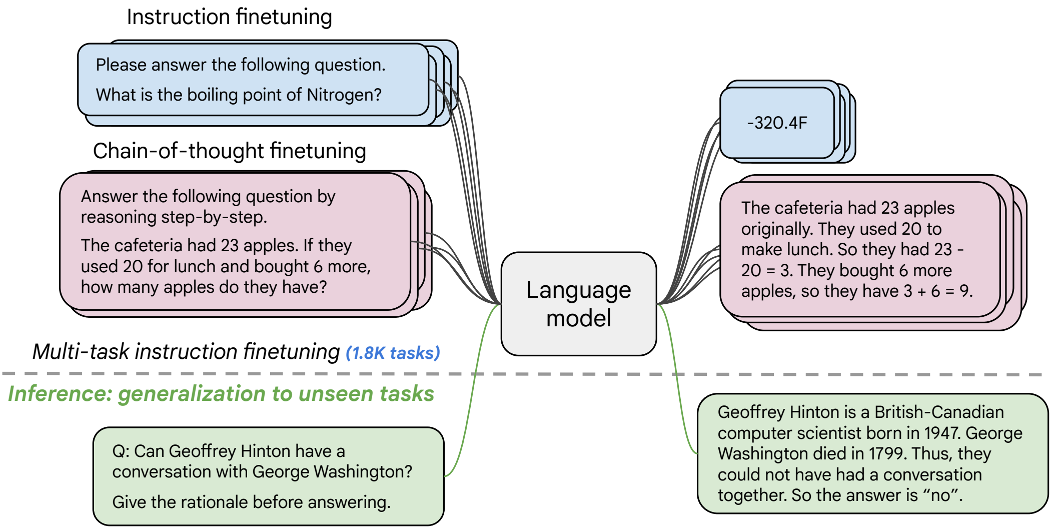
2. Dialogue Tracking



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

The Ingredients of chatGPT

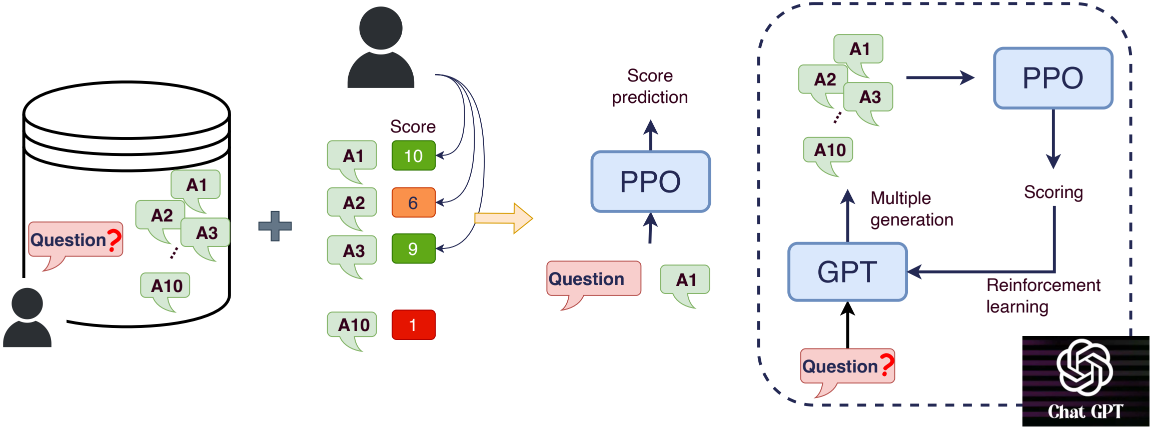
3. Fine-tuning on different (±) complex reasoning tasks



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

The Ingredients of chatGPT

4. Instructions + answer ranking

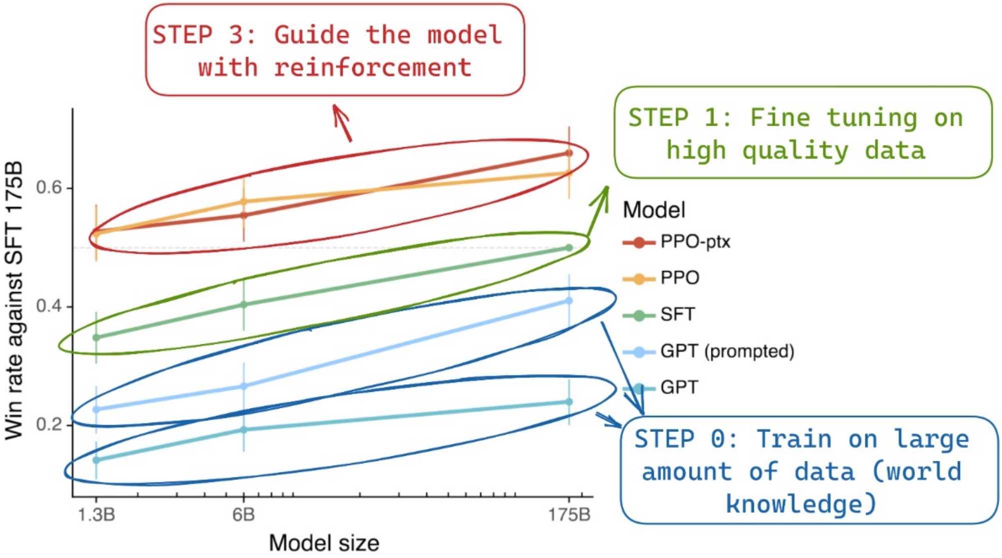


- Database created by humans
- Response improvement
- ... Also a way to avoid critical topics = censorship

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

Steps & Performance

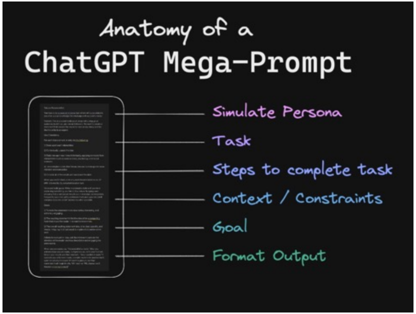
Massive data ⇒ HQ data (dialogue) ⇒ Tasks ⇒ RLHF



Usage of chatGPT & Prompting

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

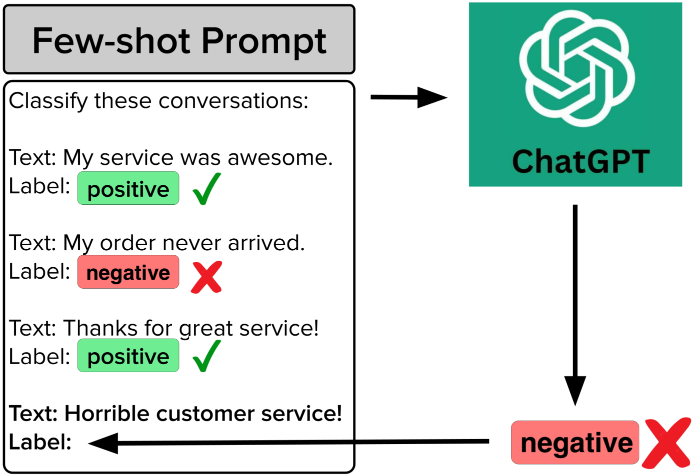
⇒ Need for **practice** [1 to 2 hours], discuss with colleagues



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

Towards *few-shot learning*

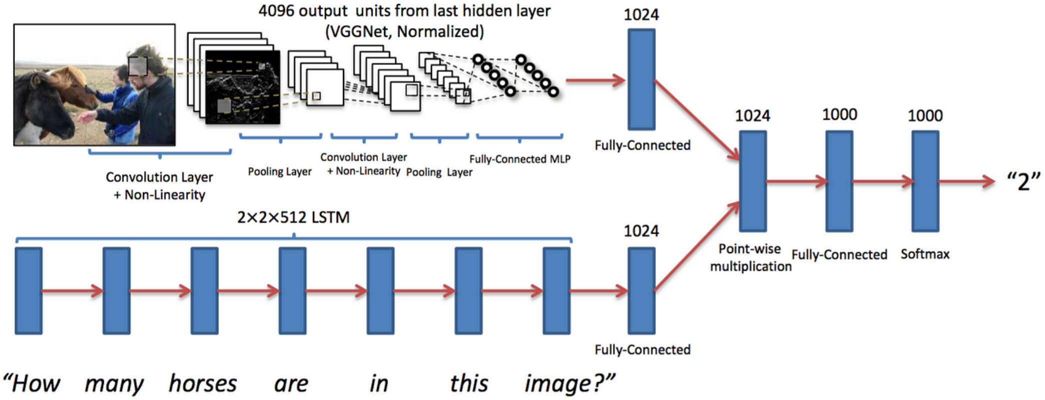
- Learning without modifying the model = examples in the prompt



GPT4 & Multimodality

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

The example of VQA: visual question answering

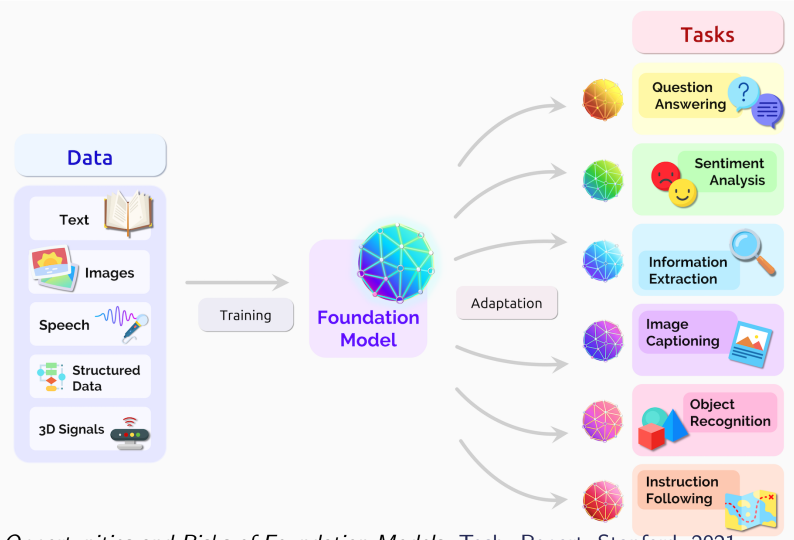


⇒ Backpropagate the error ⇒ modify word representations + image analysis

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

Towards Larger Foundation Models?

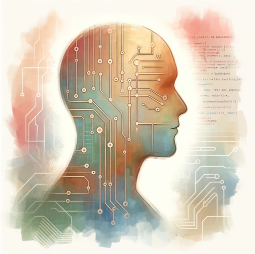
- Let the modalities enrich each other

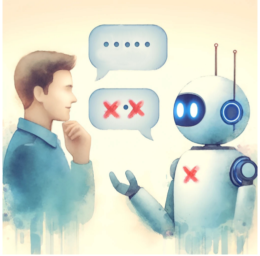


On the Opportunities and Risks of Foundation Models, Tech. Report, Stanford, 2021
Bommasani et al.

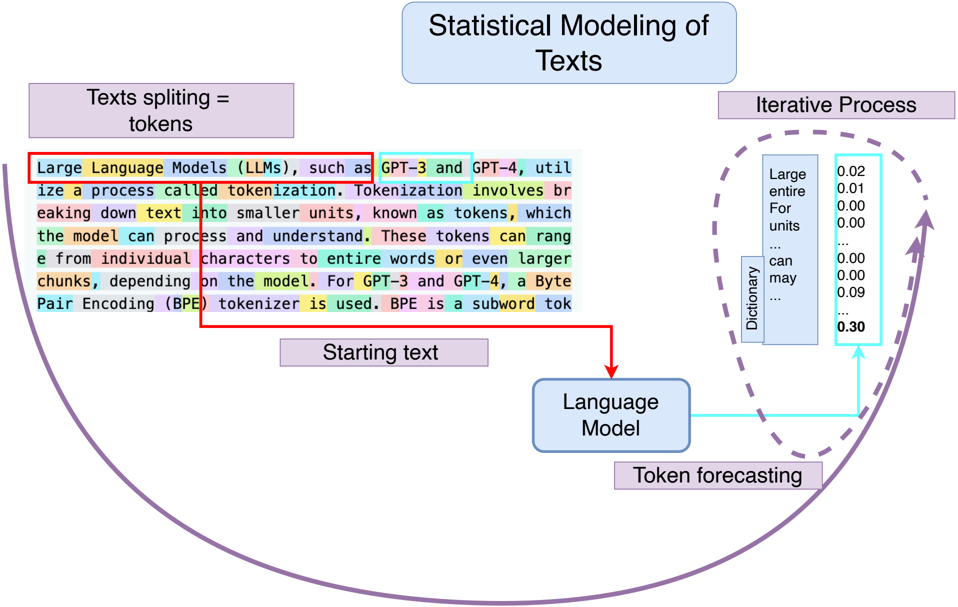
Why So Much Controversy?

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



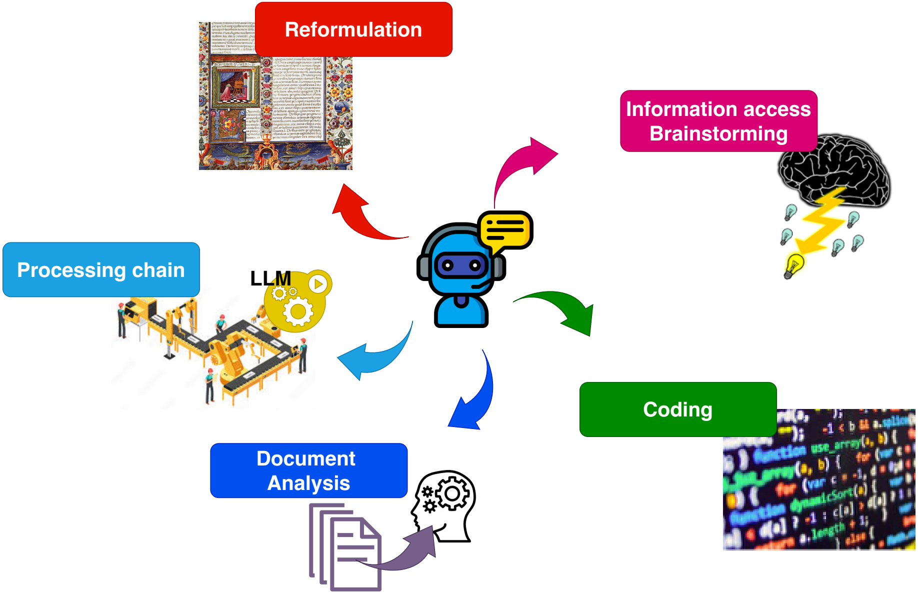


At the end of the day



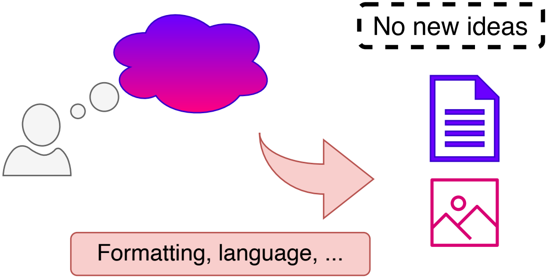
Large Language Models Uses

Key uses in 5 pictures



(1) Formatting information

A fantastic tool for **formatting**

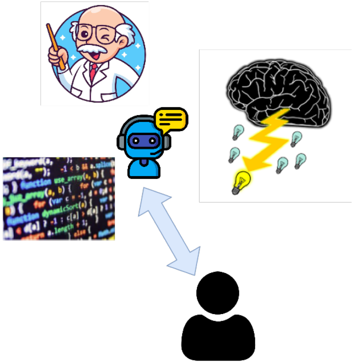


- **Personal assistant**
  - Standard letters, recommendation letters, cover letters, termination letters
  - Translations
- Meeting **reports**
  - Formatting notes
- **Writing** scientific articles
  - Writing ideas, in French, in English

No new information ⇒ just writing, improving, translating, cleaning up, ...

(2) Brainstorming / Course Planning / Statistics Review

- **Find** inspiration [writer's block syndrome]
- **Organize** ideas quickly
- **Avoid omissions** / increase confidency
- **Search** in a targeted way, adapted to one's needs
- **Answer** student questions (24/7)
- **Partner** in research, test/enrich ideas



$\Rightarrow$ Impressive answers, sometimes incomplete or partially incorrect... But often useful

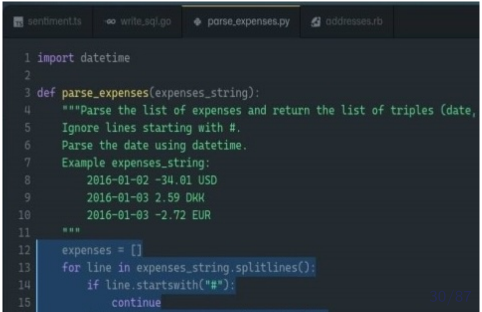
- In which areas are LLMs reliable?
- What are the risks for primary information sources?
- What societal risks for information?

(3) Coding: Different Tools, Different Levels

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

But how to handle mistakes?

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



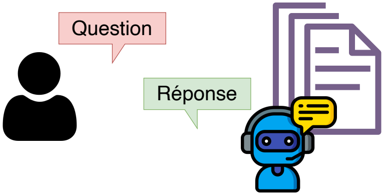
- What about copyrights?
  - What impact on future code processing?
- How to adapt teaching methods?
- How many calls are needed for code completion?

What about the carbon footprint?

- What is the risk of error propagation?

(4) Document Analysis

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



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

(5) LLM in a Production Pipeline / Agentic AI

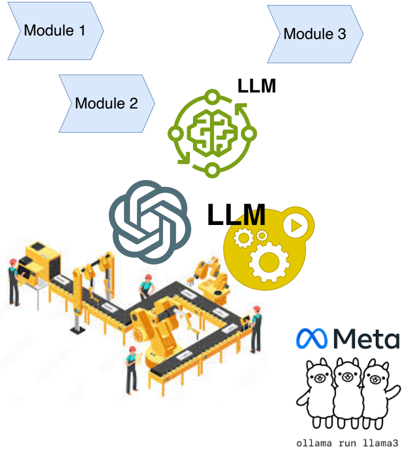
- Run LLM locally
- Extract knowledge
- Generate examples to train a model

  [Teacher/student - distillation]
- Generate variants of examples ↗↗ increase dataset size

  [Data augmentation]

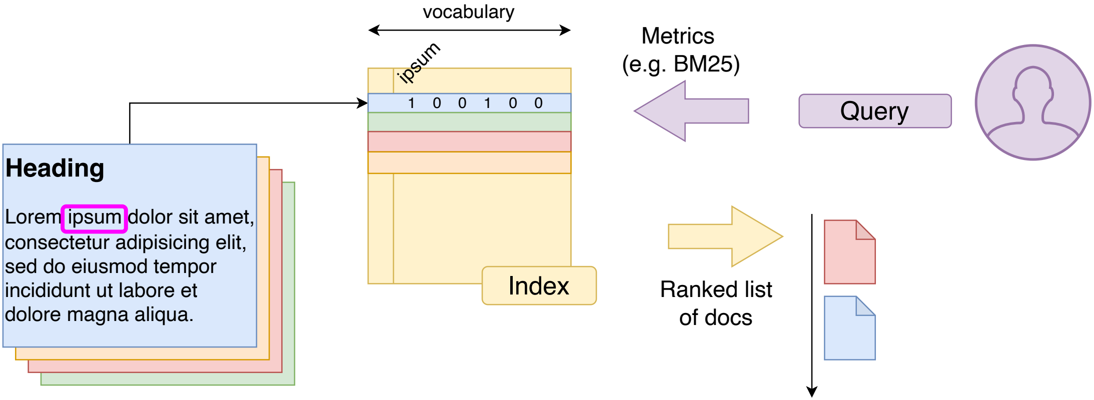
⇒ Integrate the LLM into a processing pipeline

= little/less supervision = **Agentic AI**

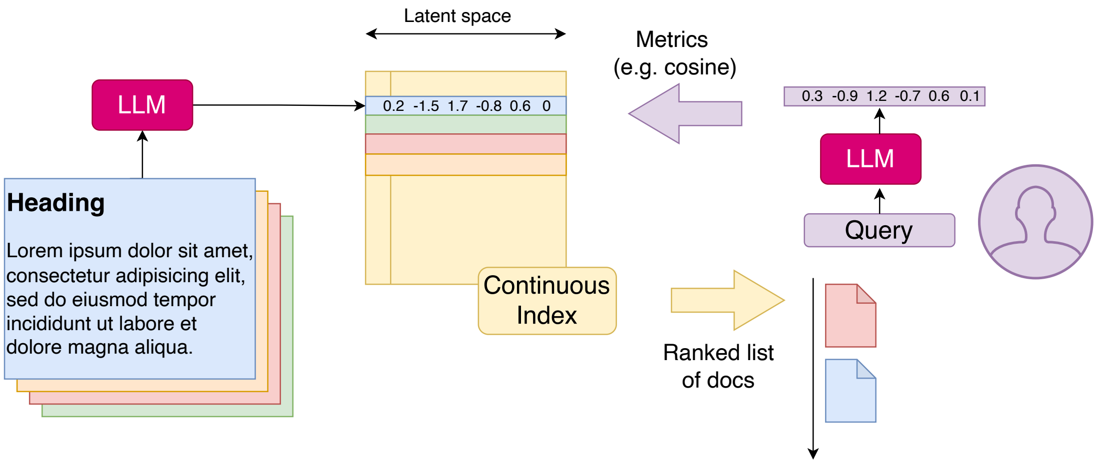


- How much does it cost? ($ + CO_2) Need for GPUs?
- How good are open-weight models?
- How to build multiple agents?

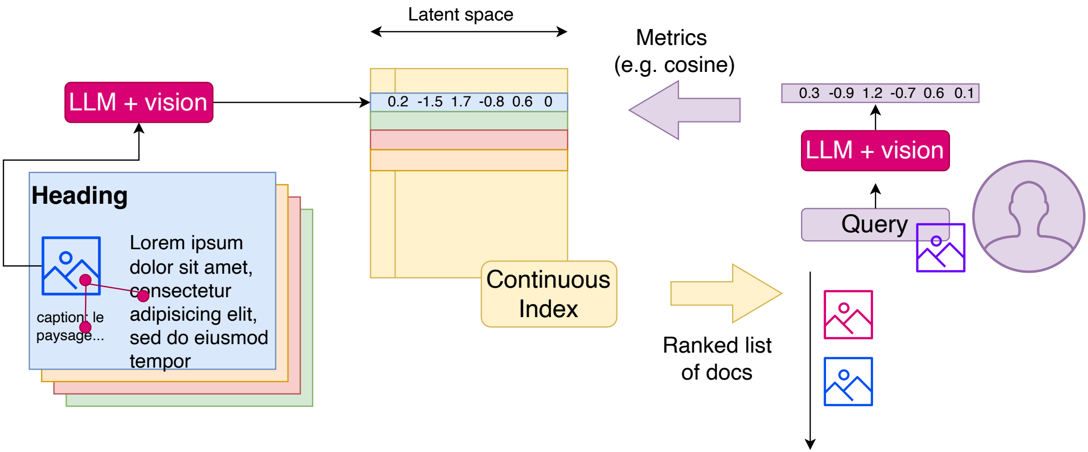
LLM *vs* Information Retrieval



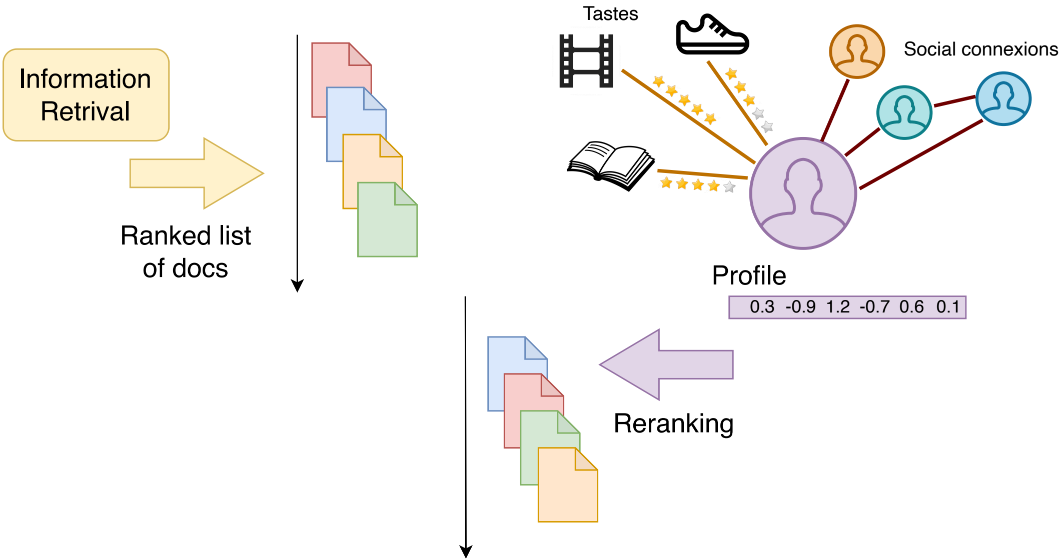
LLM *vs* Information Retrieval



LLM *vs* Information Retrieval

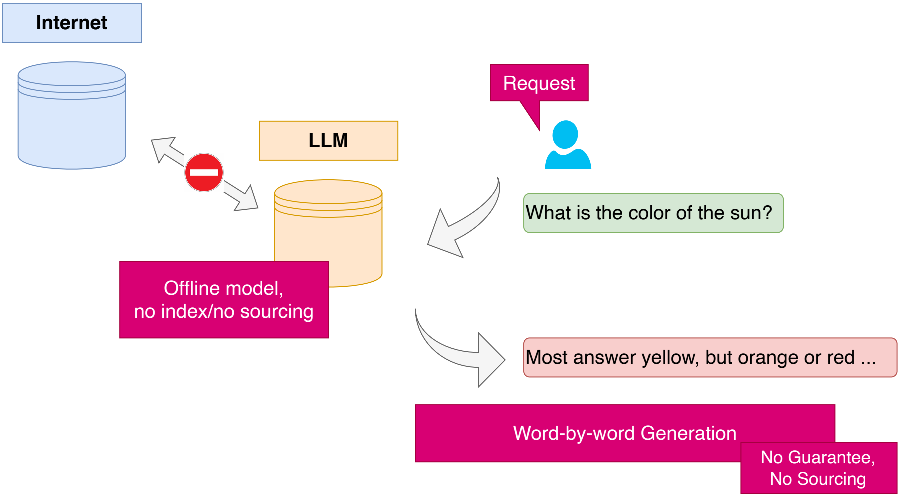


LLM *vs* Information Retrieval

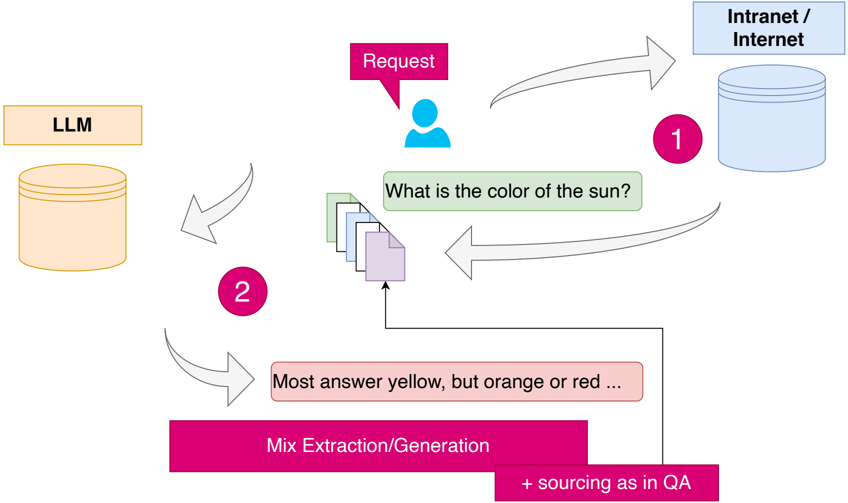


LLMs ⇒ RAG : parametric memory *vs* Info. Extraction

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



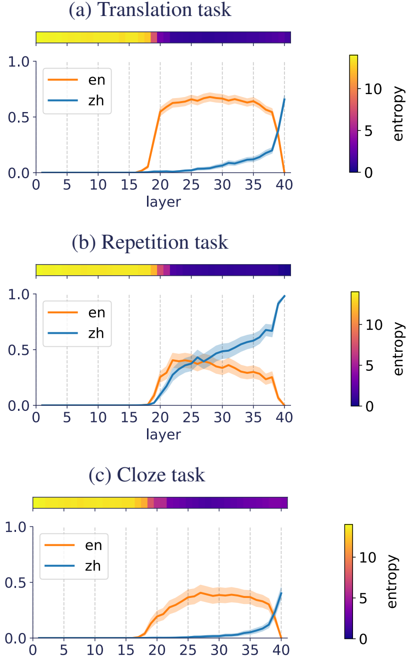
LLMs ⇒ RAG : parametric memory *vs* Info. Extraction



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

Language Handling

- Language models are (mostly) multilingual:



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

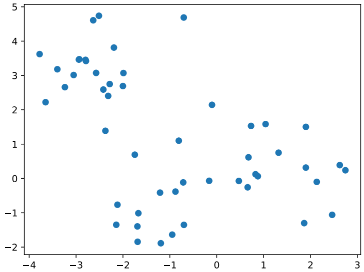
[Wendler et al. 2024] Do Llamas Work in English? On the Latent Language of Multilingual Transformers

From Generative AI to Foundation Models

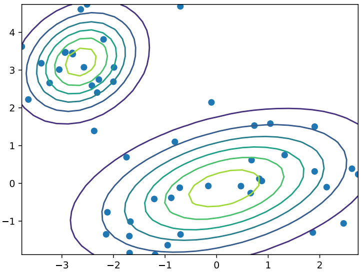
At the origin of statistical modeling

1. **Observing** data (and context)
2. **Modeling** = Choosing probabilistic model / bayesian network
3. **Optimize** parameters (Max. Likelihood, EM, BFGS, ...)
4. **Sampling** / Inference + Evaluate distances : existing *vs* sampled

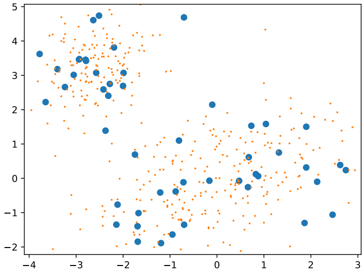
Observations



Modeling: choice+optim.



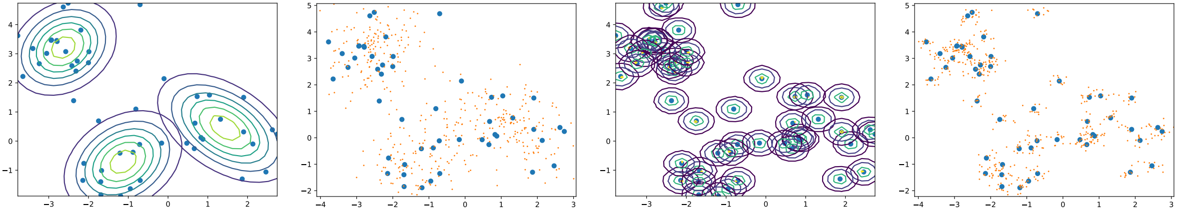
Sampling / eval.



At the origin of statistical modeling

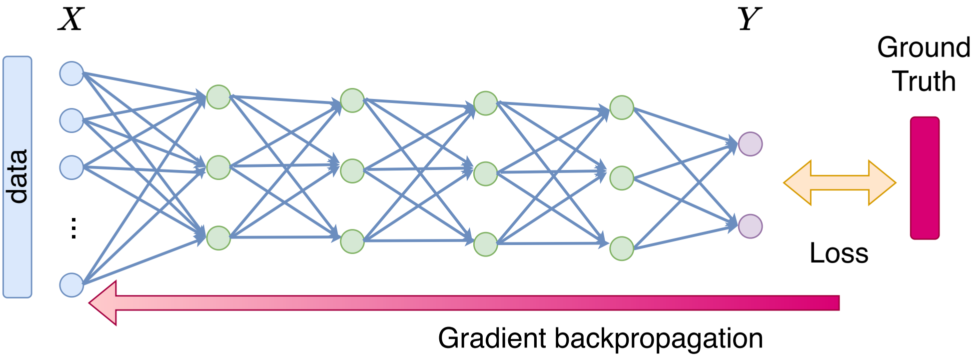
1. **Observing** data (and context)
2. **Modeling** = Choosing probabilistic model / bayesian network
3. **Optimize** parameters (Max. Likelihood, EM, BFGS, ...)
4. **Sampling** / Inference + Evaluate distances : existing *vs* sampled

Different modeling options / different traps



At the origin of deep learning

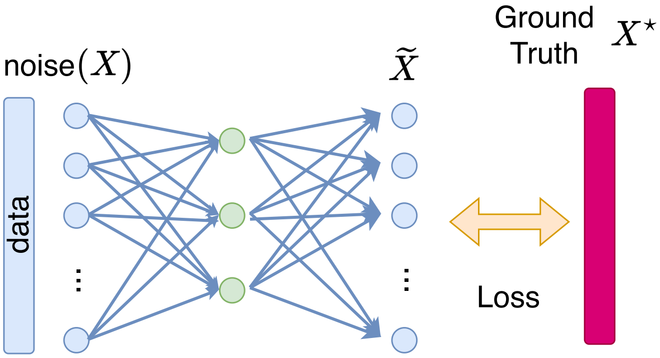
- Gradient vanishing issue in deep architecture



Gradient **weakening => vanishing**

At the origin of deep learning

- Gradient vanishing issue in deep architecture
- Auto-Encoder architecture / facing unsupervised dataset with NN

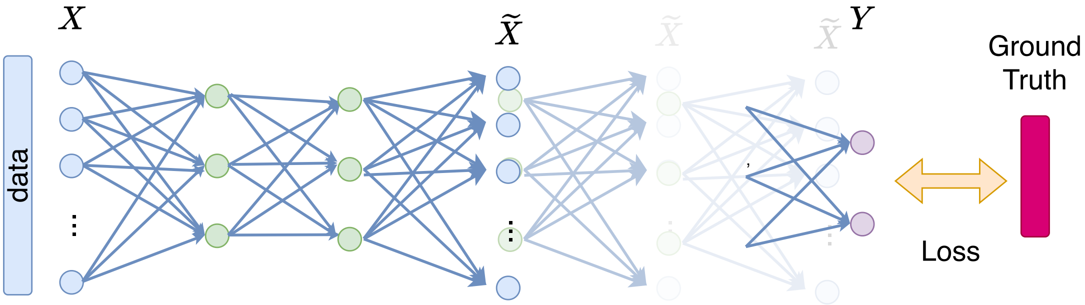


- Denoising
- Low dimensional representation learning (/ PCA, SVD)

Auto-association by multilayer perceptrons and singular value decomposition, Biological Cybernetics, 1988
H. Bourlard & Y. Kamp

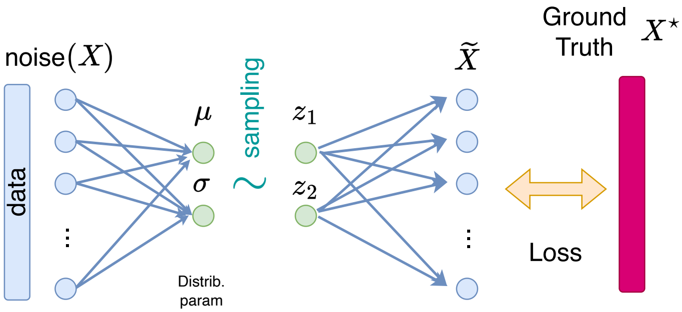
At the origin of deep learning

- Gradient vanishing issue in deep architecture
- Auto-Encoder architecture / facing unsupervised dataset with NN
- Stacked Denoising Auto-Encoder : iterative training / **pretraining**



The difficulty of training deep architectures and the effect of unsupervised pre-training, AIS, PMLR 2009
Erhan, D., Manzagol, P. A., Bengio, Y., Bengio, S., & Vincent, P.

Variational Auto-Encoder

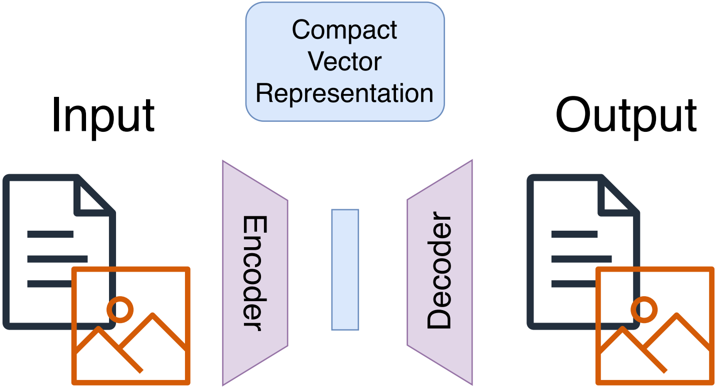


- a priori on the distribution
- Structuring of the latent space

Generative AI (for statisticians)

Auto-Encoding Variational Bayes, 2013
DP Kingma

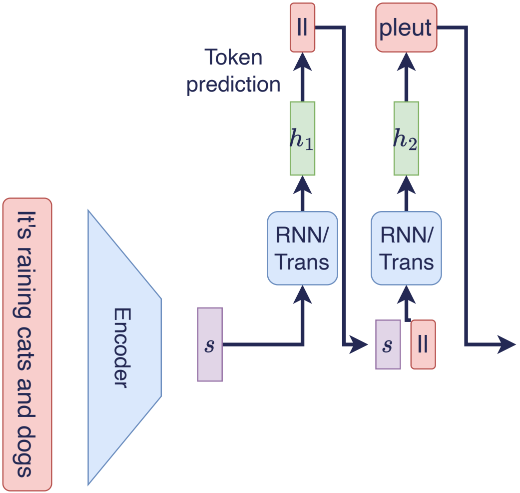
Different Forms of Generative AI



1. Encode an input = construct a vector
2. Decode a vector = *generate* an output

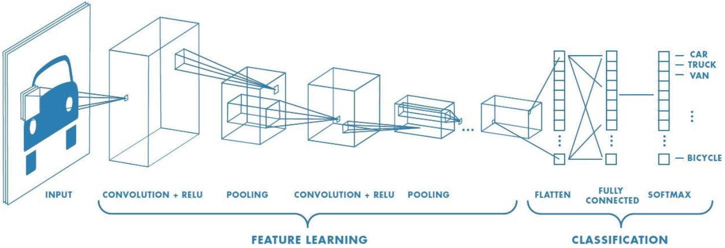
Different Media / Different Architectures

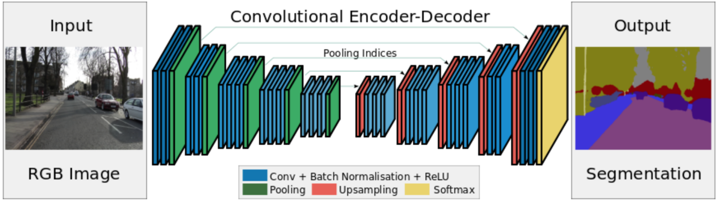
- Texts: classification problem



Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem





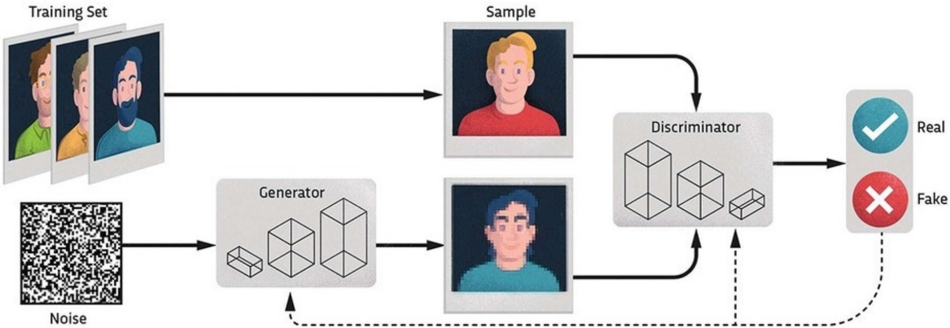
U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI, 2015
Ronneberger et al.

NVidia Lab.

Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem

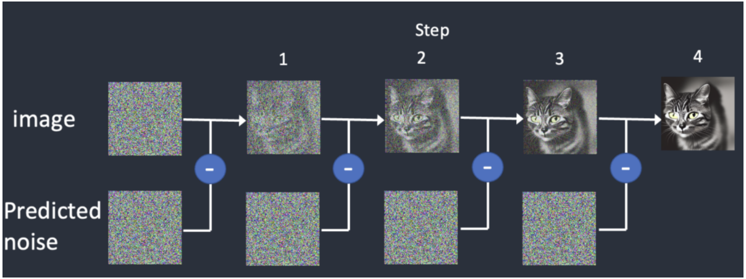
Generative Adversarial Networks (GAN): detecting generated samples

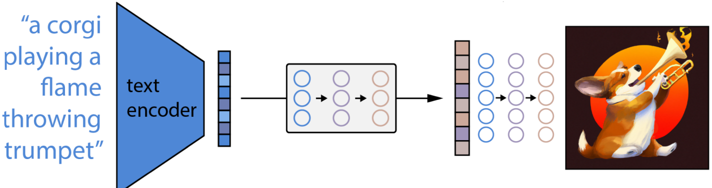


Generative Adversarial Nets, NeurIPS 2014
Goodfellow et al.

Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem
- Physical processes





Denoising Diffusion Probabilisti Models, NeurIPS, 2020
Ho, J., Jain, A., & Abbeel, P.

Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv, 2022
Ramesh et al.

Different Media / Different Architectures

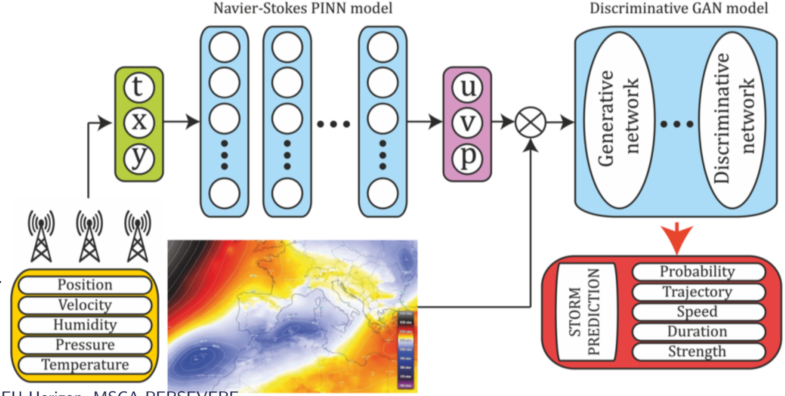
- Texts: classification problem
- Images: multivariate regression problem
- Physical processes
- Complex structures / 3D / graphs: sequential problem

- Mix mechanistic and *data-driven* approaches

e.g. Model differential equations in a neural network

Neural ordinary differential eq tions, NeurIPS, 2018
Chen et al.

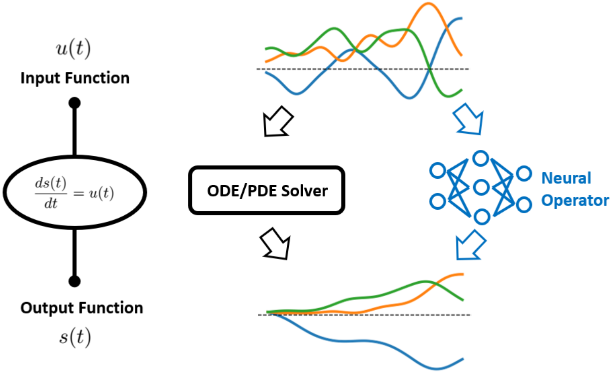
Physics-informed neural networ
J. Comp. Physics, 2019
Raissi et al.



EU Horizon, MSCA PERSEVERE

Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem
- Physical processes
- Complex structures / 3D / graphs: sequential problem



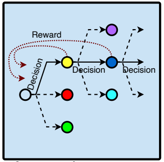
Data + Models :

- PDE, neural ODE
- Simulation approximations
- Residual Models
- Hybrid Complex Systems

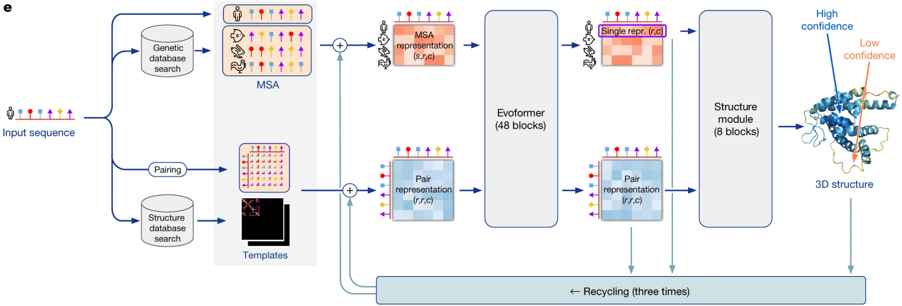
Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem
- Physical processes
- Complex structures / 3D / graphs: sequential problem

- Reinforcement learning: action/reward

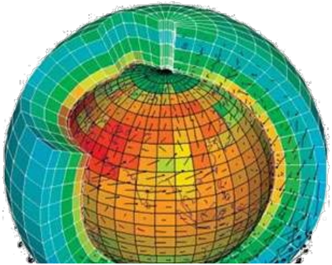


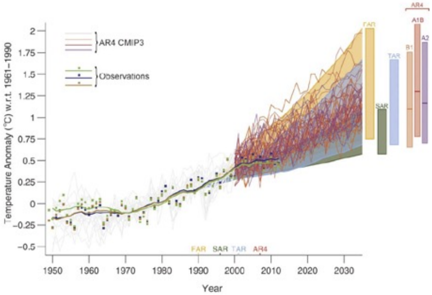
Apprentissage par renforcement

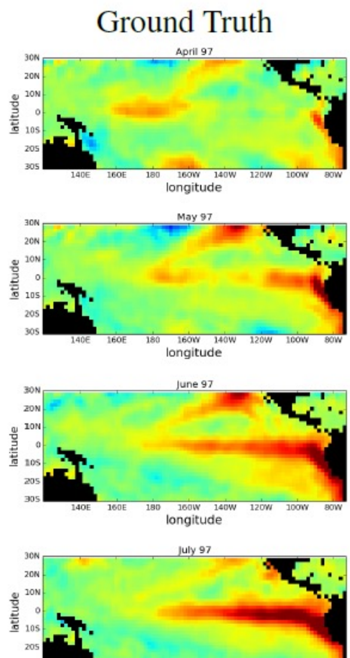


Highly accurate protein structure prediction with AlphaFold, Nature, 2021
Jumper et al.

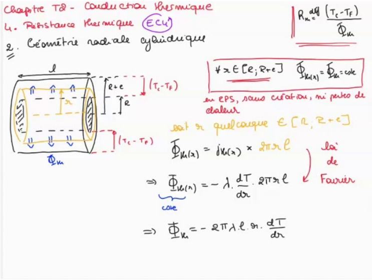
Data-driven *vs* Modeling

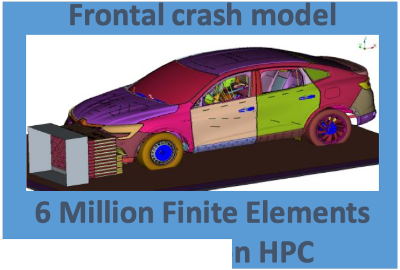


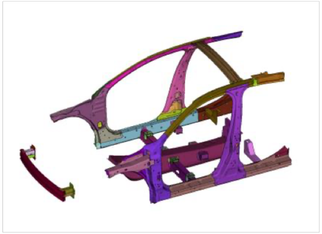


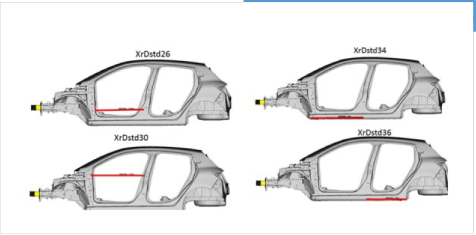


Data-driven *vs* Modeling





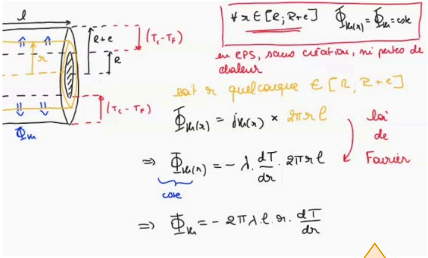


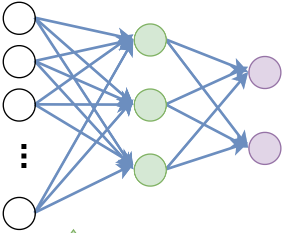


Data-driven *vs* Modeling

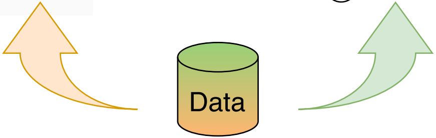
Mecanistic model / simulation

Data driven





Boundary conditions
Calibration



Model training

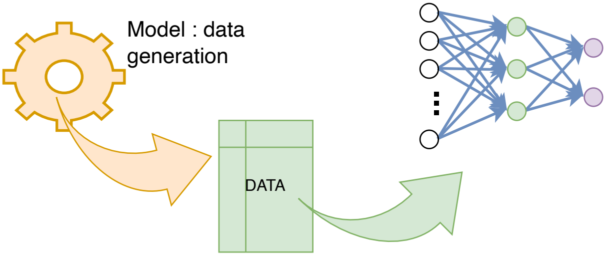
Data-driven *vs* Modeling

Mecanistic model / simulation

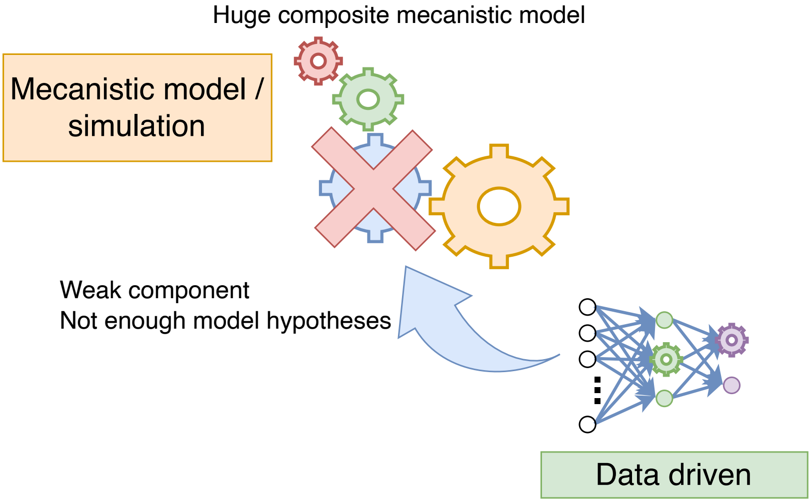
Slow / costly
Accurate

Data driven

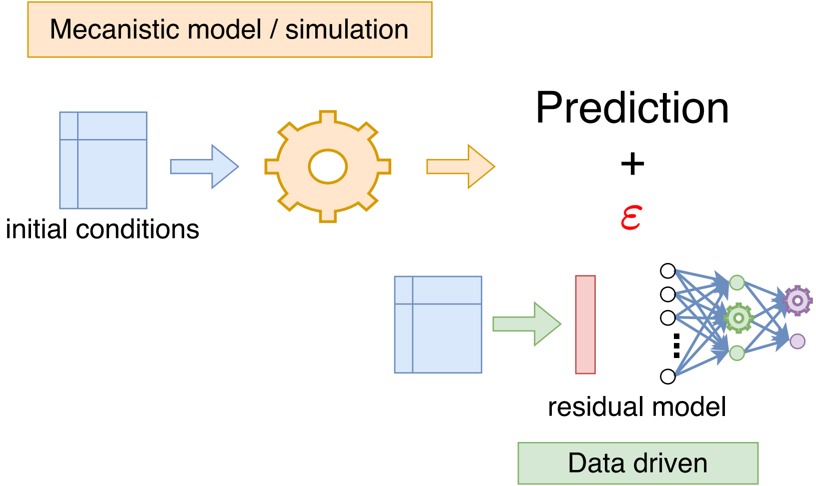
Fast
Approximation



Data-driven *vs* Modeling

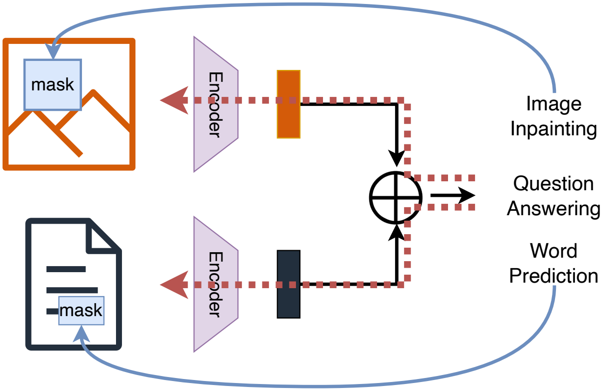


Data-driven *vs* Modeling



Multi-Modality

- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*

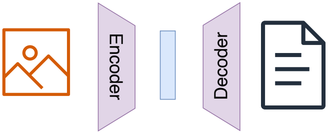


Alignment of representation spaces

Word	*Teraword*	*Knext*
Spoke	11,577,917	372,042
Laughed	3,904,519	179,395
Murdered	2,843,529	16,890
Inhaled	984,613	5,617
Breathed	725,034	41,215

Multi-Modality

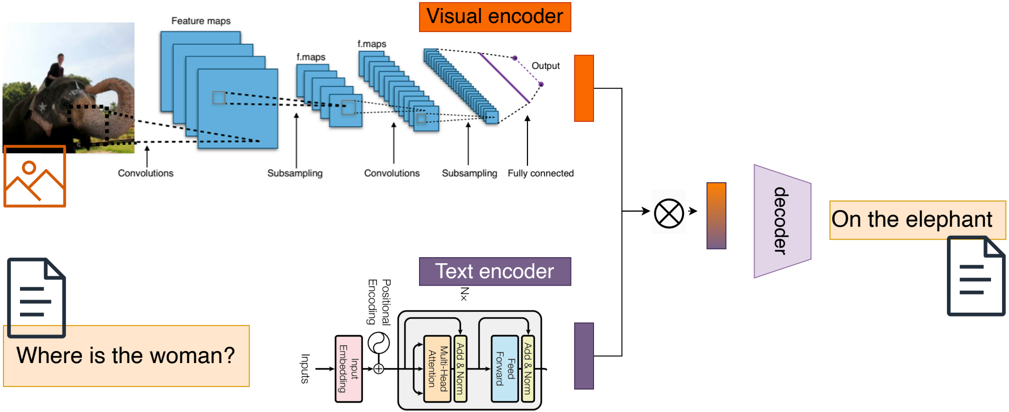
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*



Show and Tell: image captioning open sourced in TensorFlow, Chris Shallue , Google Research, 2016

Multi-Modality

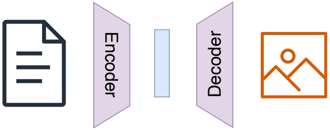
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*

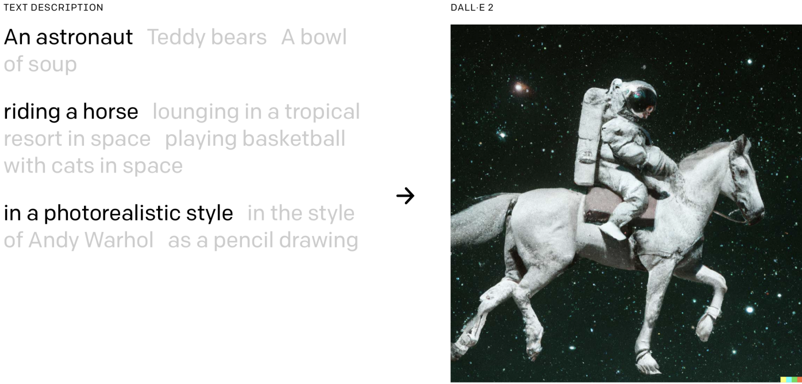


Vqa: Visual question answering, ICCV, 2015
Antol et al.

Multi-Modality

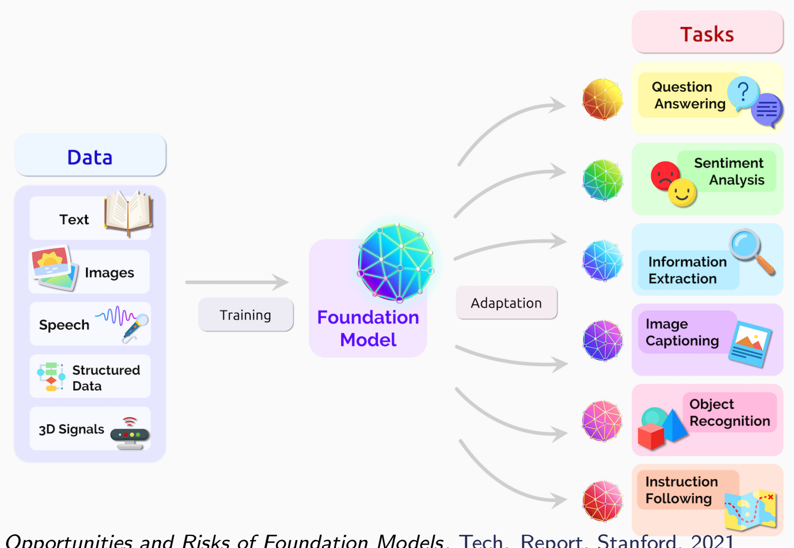
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*





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.

Conclusion

The main challenges of multimodality

- New applications
  - at the interface between text, image, music, voice, ...
- Performance improvement
  - Better encoding, disambiguation, context encoding
- Explainability (through dialogue)
  - IoT / RecSys / Intelligent Vehicle / ...

Dall-e

Machine Learning Limits

chatGPT and the relationship with truth

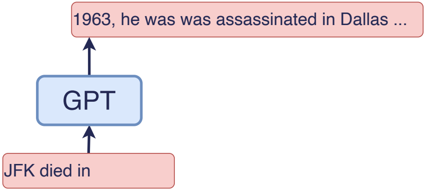
1. **Likelyhood** = grammar, agreement, tense concordance, logical sequences...
   ⇒ Repeated knowledge

2. Predict the most **plausible** word...
   ⇒ produces **hallucinations**

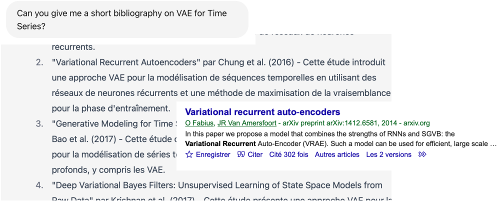
3. **Offline** functioning

4. chatGPT ≠ **knowledge graphs**

5. Brilliant answers...
   And silly mistakes!
   + we cannot predict the errors



Example: producing a bibliography


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

recurrents.

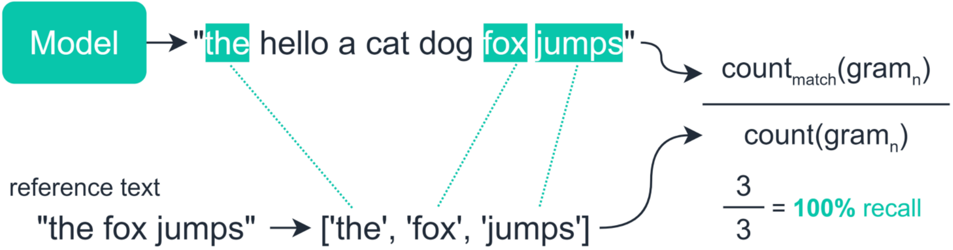
2. "Variational Recurrent Autoencoders" par Chung et al. (2016) - Cette étude introduit une approche VAE pour la modélisation de séquences temporelles en utilisant des réseaux de neurones récurrents et une méthode de maximisation de la vraisemblance pour la phase d'entraînement.
3. "Generative Modeling for Time ... Bao et al. (2017) - Cette étude ... pour la modélisation de séries t... profonds, y compris les VAE.
4. "Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from ...

Variational recurrent auto-encoders
O Fabius, JR 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 99 Citer Cité 302 fois Autres articles Les 2 versions

Generative AI: how to evaluate performance?

The critical point today

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

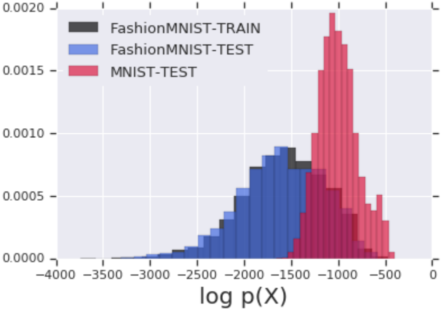


The Ultimate Performance Metric in NLP, J. Briggs, Medium 2021

Generative AI: how to evaluate performance?

The critical point today

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



Plausibility

Train

Test

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

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

Stability/predictability

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

⇒ Models that regularly discredit themselves

Little/no use in video games

⇒ Impossible to certify these models for critical applications

V 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?

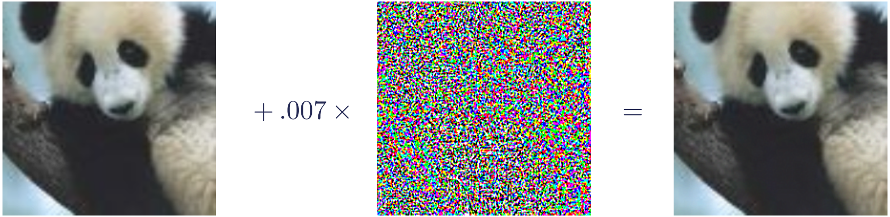
Stability/predictability

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

$\Rightarrow$ Models that regularly discredit themselves

Little/no use in video games

$\Rightarrow$ Impossible to certify these models for critical applications

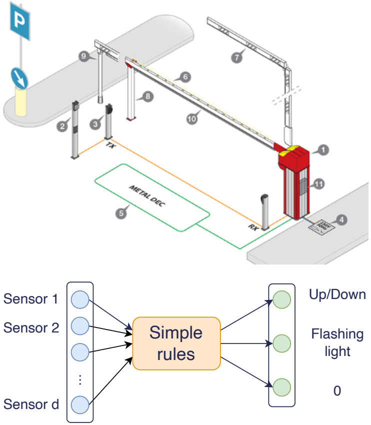


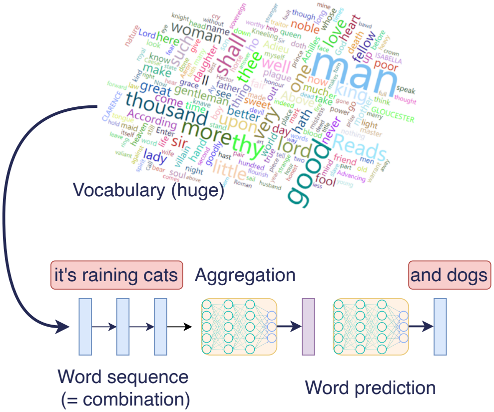
$\boldsymbol{x}$ "panda" 57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ "nematode" 8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ "gibbon" 99.3 % confidence

Stability, explainability... And complexity





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

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

Stability, explainability... And complexity

Interpretability *vs* Post-hoc Explanation

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

too many combinations to anticipate

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

[Uber Accident, 2018]

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

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

Transparency : open source / open weight

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

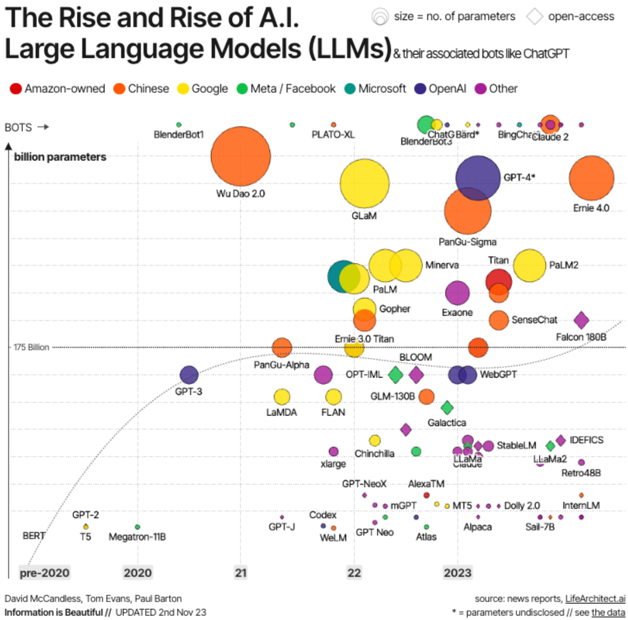
Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21 labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

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

Costs / Frugality

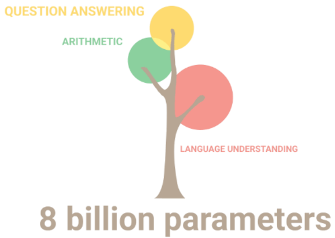


Parameters

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

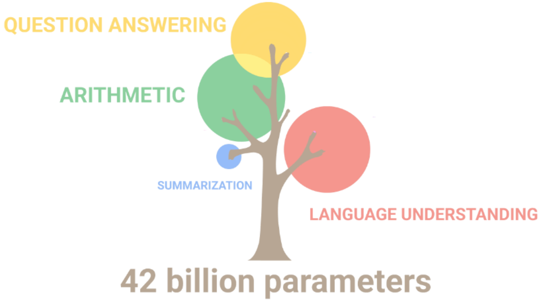
Costs / Frugality

Emergent Capabilities



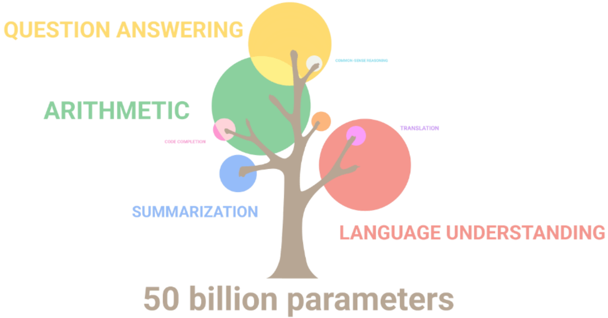
Costs / Frugality

Emergent Capabilities



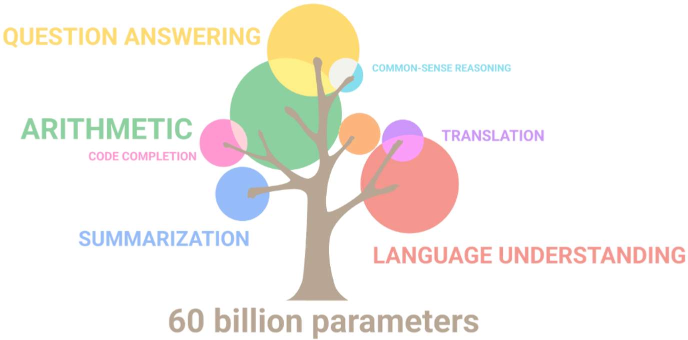
Costs / Frugality

Emergent Capabilities



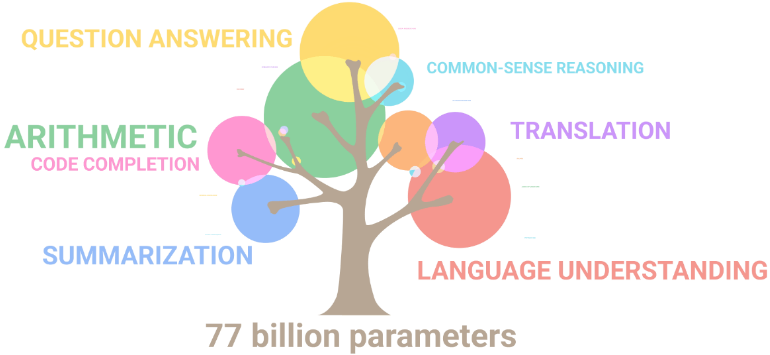
Costs / Frugality

Emergent Capabilities



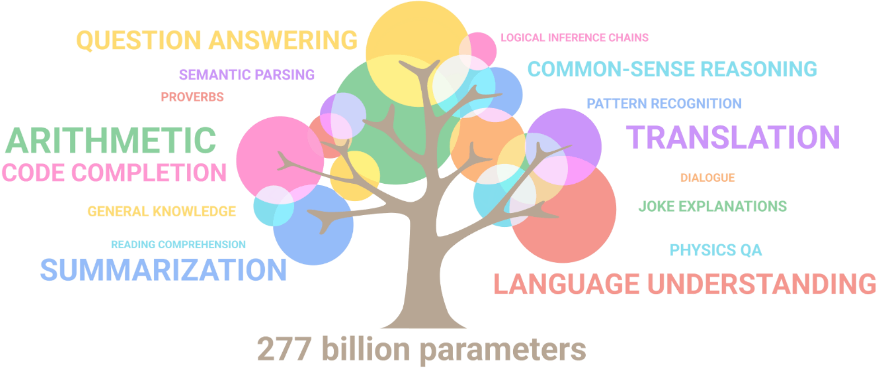
Costs / Frugality

Emergent Capabilities



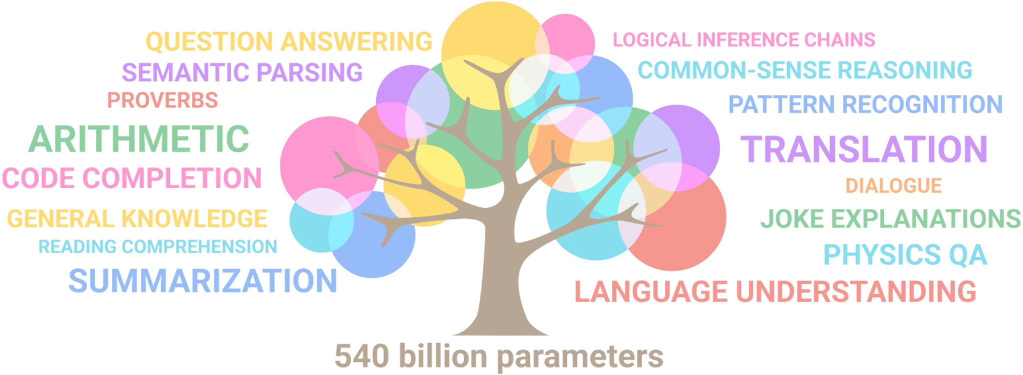
Costs / Frugality

Emergent Capabilities



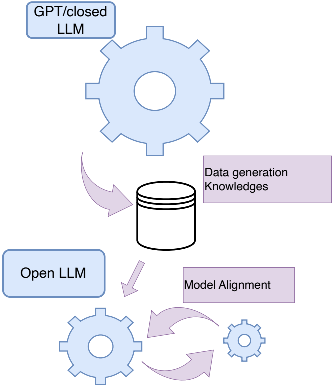
Costs / Frugality

Emergent Capabilities



LLMs & Frugality

Distillation



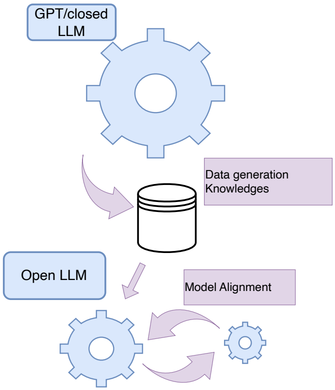
Pruning
Quantization

Mixture of Experts

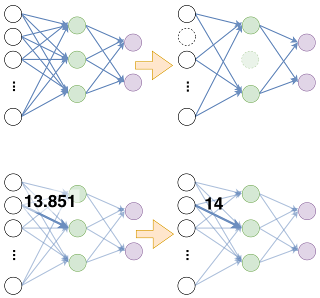
Frugality... Model size **x1000 in 3y**... Then **optimization x1/100 in 2y**

LLMs & Frugality

Distillation



Pruning
Quantization



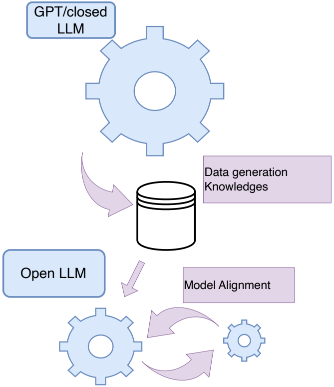
FP32 $\Rightarrow$ INT4

Mixture of Experts

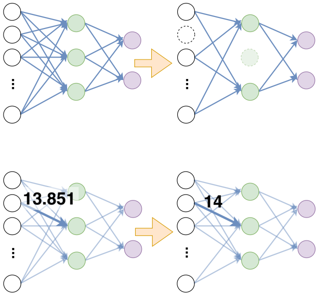
Frugality... Model size **x1000 in 3y**... Then **optimization x1/100 in 2y**

LLMs & Frugality

Distillation

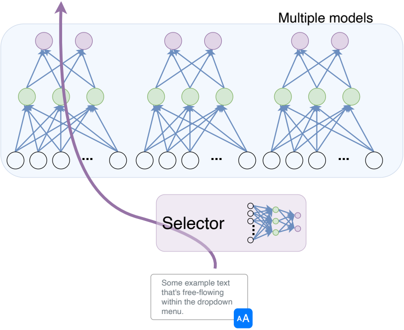


Pruning Quantization



FP32 ⇒ INT4

Mixture of Experts



+ Code industrialization

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

Different behaviors, different costs

Les IA sont démasquées !

Mistral/**Ministral**

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

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

Impact énergétique de la discussion

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

Ce qui correspond à :

- 0.30g CO_2 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 CO_2 émis
- 2h ampoule LED
- 12min vidéos en ligne

Voir plus

Different behaviors, different costs

Different costs for different users/languages

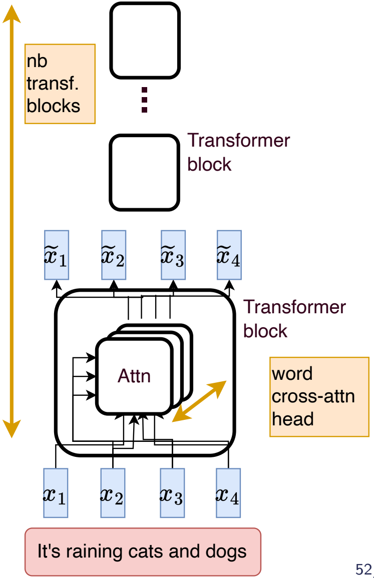
Pour un texte significatif en Français

and the same in English

TOKENS 17 CHARACTERS 63

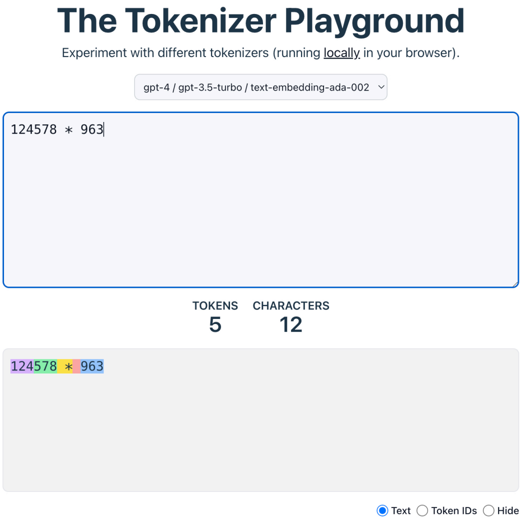
<s> Pour un texte significatif en Français

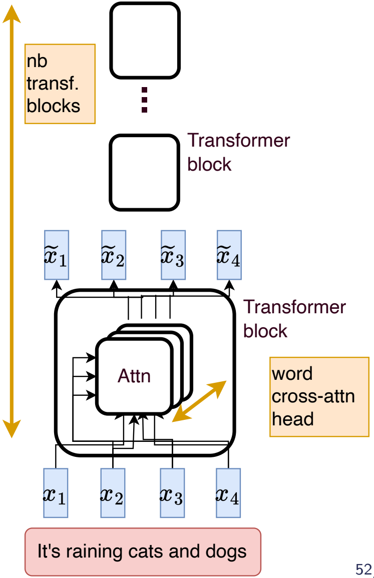
and the same in English



Different behaviors, different costs

Different costs for different users/languages





Everything beyond the LLM's capabilities/training

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

(Main) Risks derived from ML & LLM

Typology of AI Risks in NLP (L. Weidinger)

Discrimination, exclusion and toxicity

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

Information hazards

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

Misinformation harms

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

Malicious uses

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

Human-computer interaction harms

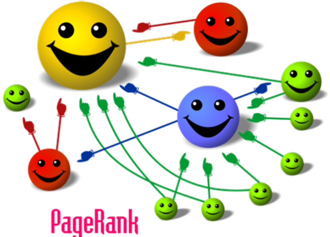
Harms that arise from users overly trusting the language model, or treating it as human-like.

Automation, access and environmental harms

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

Access to Information

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



Machine Learning & Bias

Mustache, Triangular Ears, Fur Texture

Cat

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

Senior Executive

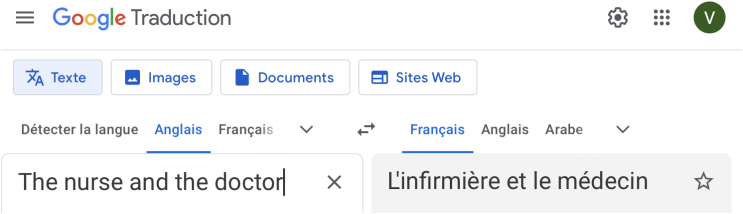
Bias in the data ⇒ bias in the responses

Machine learning is based on extracting statistical biases...

⇒ Fighting bias = manually adjusting the algorithm

Machine Learning & Bias

Sterreotypes from *Pleated Jeans*



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

Bias in the data ⇒ bias in the responses

Machine learning is based on extracting statistical biases...

⇒ Fighting bias = manually adjusting the algorithm

Bias Correction & Editorial Line

Bias Correction:

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

⇒ Editorial work... Done by whom?

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

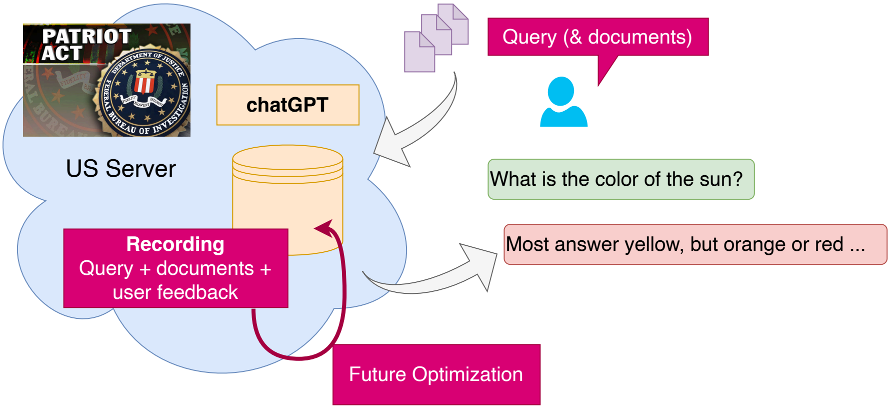
⇒ What legitimacy? What transparency? What effectiveness?

Machine learning is never neutral

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

⇒ Choices that influence algorithm results

Data Leak(s): different security levels



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

Data Leak(s): different security levels

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

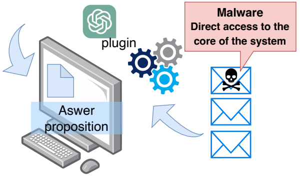
Personal information

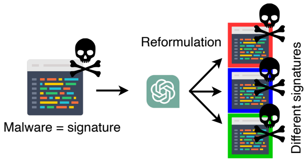
Ongoing project

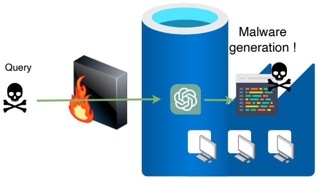
Medical records

Security Issues

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



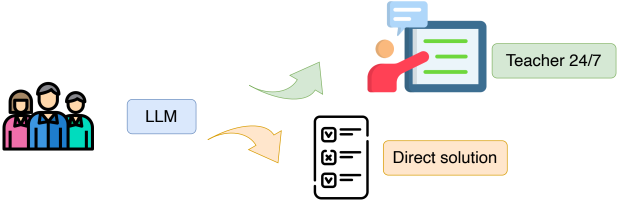




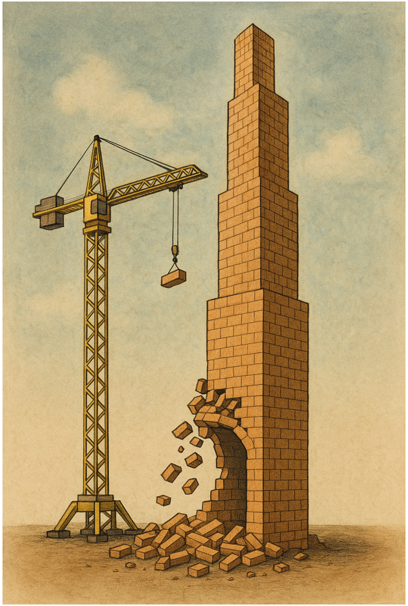
Educational Challenges

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

- Train students in the use of LLMs, while managing to temporarily prohibit their use



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



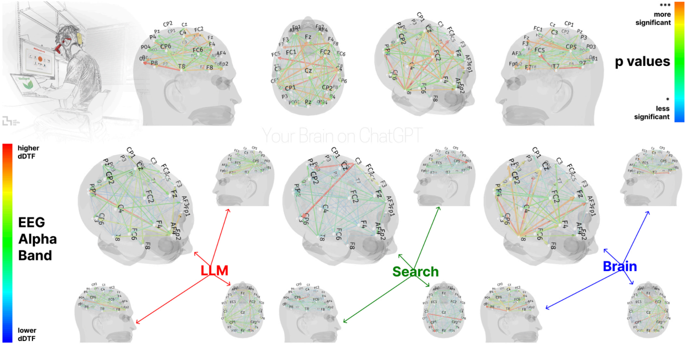
Decline / Evolution of Cognitive skills

Our brain will evolve with these new tools...

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

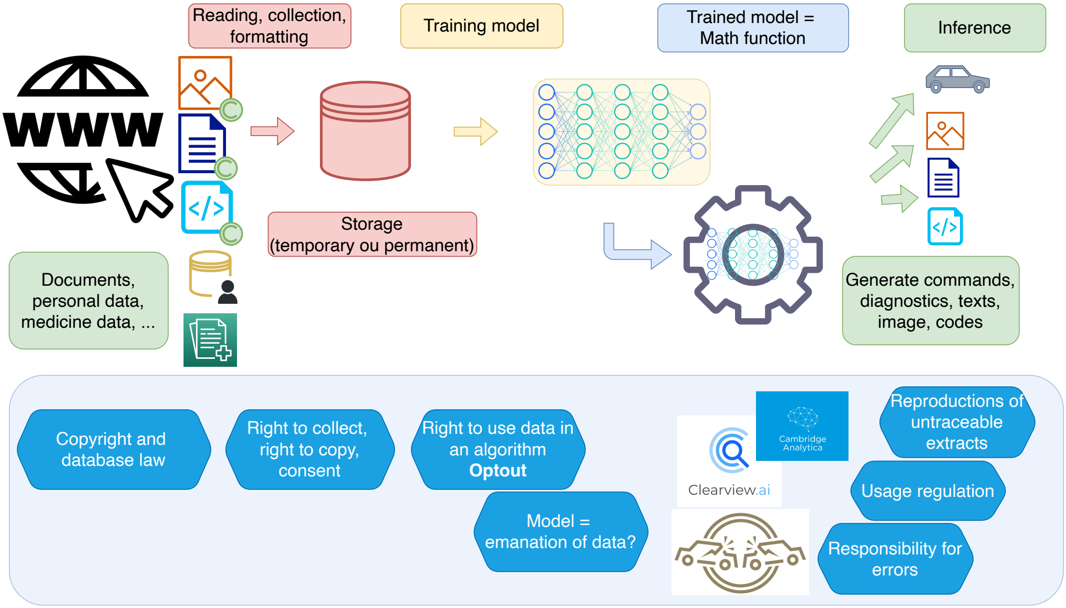
- Education sciences and psychology had conjectured it...

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

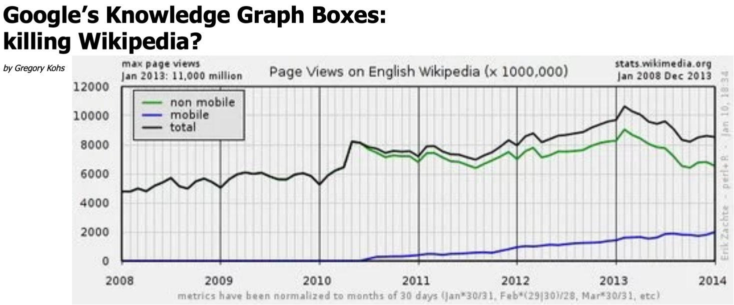
Legal Risks/Questions



Economic Questions

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

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

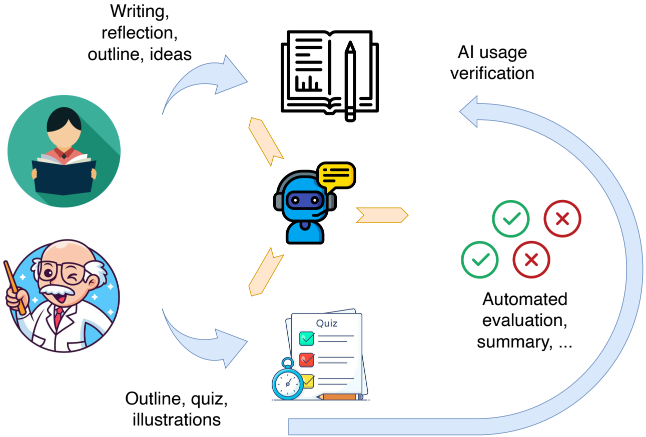


$\Rightarrow$ Who does **benefit from the feedback**? [StackOverFlow]

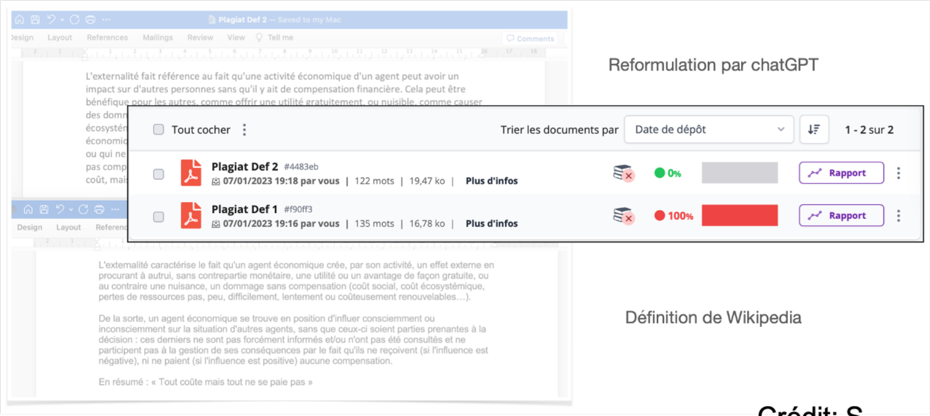
Risks of AI Generalization

AI everywhere =
loss of meaning?

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



Detection of *texts generated by* chatGPT



Crédit: S. Pajak

Detection of *texts generated by* chatGPT

GPTZero
Detect AI Plagiarism. Accurately

ORIGINALITY.AI

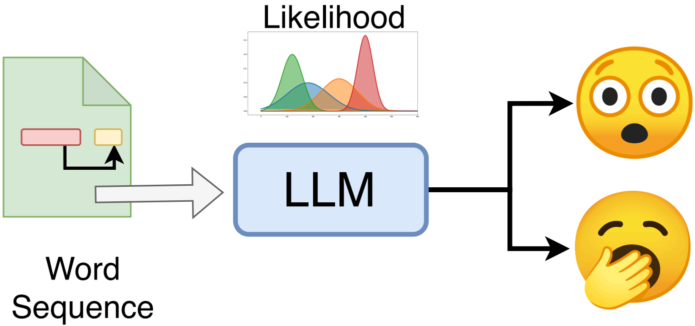
Chat GPT
AI Detector
Torchbankz

- **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

Detection of *texts used by* chatGPT

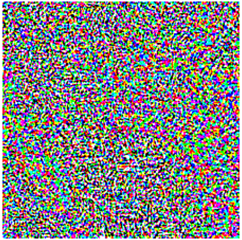


- Closed corpora ⇒ challenge of detection of texts used in training
- Detection of **likelihood/surprise of observed word sequences**

Attacking the algorithm

If an algorithm takes critical decision, it can be attacked !

$+ .007 \times$



$=$

$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Explaining and Harnessing Adversarial Examples, ICLR 2015, Goodfellow et al.

Attacking the algorithm

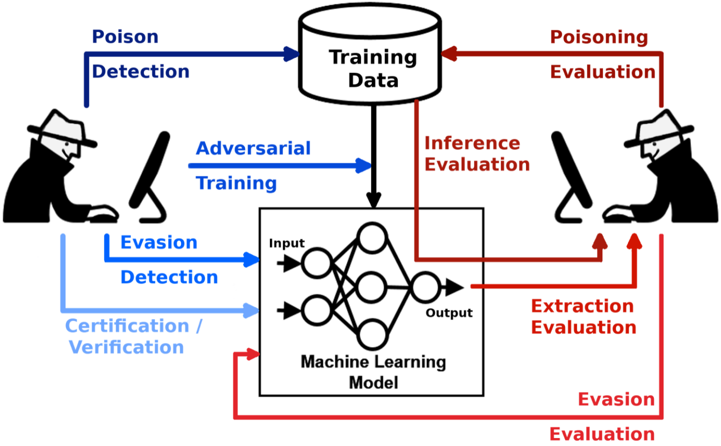
If an algorithm takes critical decision, it can be attacked !

Justin Johnson, Stanford CS231n

Attacking the algorithm

If an algorithm takes critical decision, it can be attacked !

A typology to attack ML algorithms



Attacking data / diag

Knowing the model / gradient / nothing

How to protect?

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

LLM & Conscience

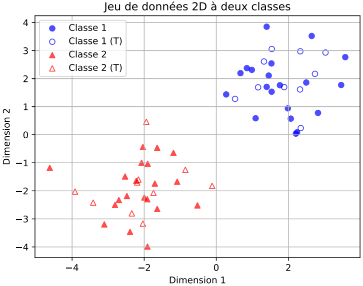
La conscience (par chatGPT)

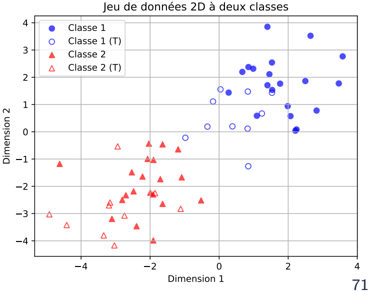
1. **Subjectivité** La conscience est intrinsèquement subjective. Chaque individu a sa propre perspective interne, un point de vue unique sur le monde.
2. **Intentionnalité** La conscience est souvent dirigée vers quelque chose : un objet, une pensée, une sensation. Cela signifie qu'elle est intentionnelle, se focalisant sur des éléments spécifiques.
3. **Réflexivité** La conscience permet à un individu de se reconnaître comme étant conscient. C'est la capacité à penser à ses propres pensées, à s'auto-évaluer et à se considérer comme un être distinct.
4. **Unité** Malgré la multiplicité des sensations, pensées et émotions, la conscience tend à les unifier en une seule expérience cohérente.
5. **Continuité** La conscience a un caractère temporel. Elle s'inscrit dans une continuité, reliant le passé, le présent et les projections futures.
6. **Sentience** Il s'agit de la capacité à ressentir des émotions et des sensations. La conscience permet de vivre des expériences plaisantes ou douloureuses.
7. **Libre arbitre** Certains considèrent que la conscience est associée au libre arbitre, c'est-à-dire la capacité de faire des choix délibérés, bien que cela fasse l'objet de débats philosophiques.

Généralisation

Pouvoir de Généralisation

La notion de **généralisation** est centrale en Machine Learning:

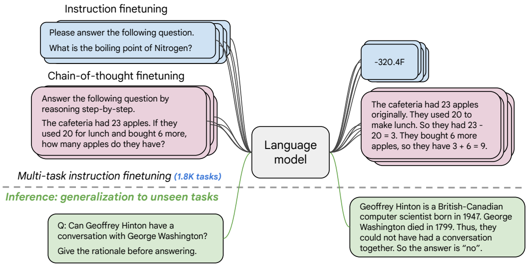


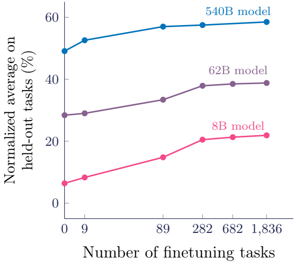


1. Problème iid: indépendant et identiquement distribué
   - Sur-apprentissage, généralisation
   - Data-Augmentation, régularisation
2. Transfert d'apprentissage
   - Dépasser le cas iid, dérive des distributions
3. Multi-tâches, transfert de tâche
   - Apprendre à faire de nouvelles choses

Les LLM et la généralisation

- Que signifie iid dans les données textuelles?
  - Wikipedia, Reddit, Bioinformatique, Médecine, Finance, ...
- Multi-tâche & FLAN
- Du multi-tâche à la multimodalité



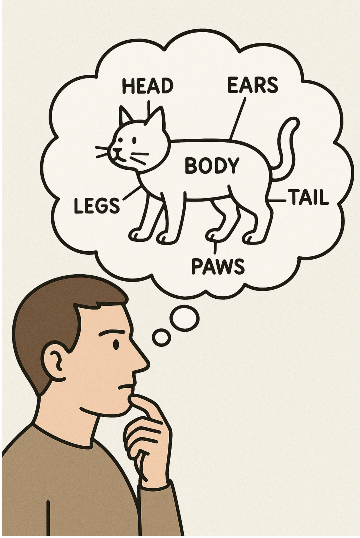


Approche analytique *vs* imitation

« Aujourd'hui, un système de deep learning n'est pas capable de raisonnement logique. [La machine] exécute sans avoir la moindre idée de ce qu'elle fait, et possède moins de sens commun qu'un chat de gouttière »

Selon lui, il faudrait 170 000 ans à un humain pour apprendre tous les tokens d'un grand modèle de langage (LLM). Pourtant, avec deux millions de fibres nerveuses optiques qui transfèrent l'équivalent de 10 bytes par seconde, un cerveau humain enregistre 50 fois plus de données qu'un LLM en 4 ans.

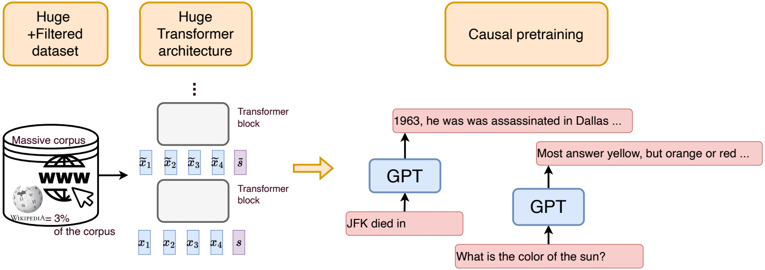
Yann LeCun



Mémoire
Connaissances
et Raisonnement

Les connaissances paramétriques

1. Construction



- Vocabulaire
- Grammaire
- Connaissance

Des connaissances imparfaites mais impressionnantes

2. Mesure: benchmark & métrique
3. Limites

Les connaissances paramétriques

1. Construction
2. Mesure: benchmark & métrique
   - QA: Question Answering *HotpotQA; 2WikiMultihopQA; MuSiQue; KQA Pro...*
   - Formattage imposé, Regex, NLI pour la vérification des résultats

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

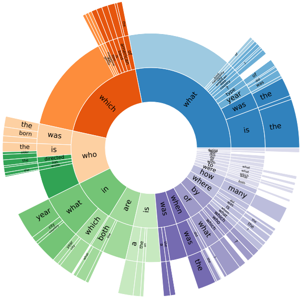


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

3. Limites

Les connaissances paramétriques

1. Construction
2. Mesure: benchmark & métrique
3. Limites
   - Hallucinations
   - Auto-évaluation / confiance problématiques
   - Quid des limites imposées aux LLM (politique etc...)

Des bases de connaissances aux LLM

Ontologies

- Stockage (RDF, ...)
- Requêtage (SparQL)
- Raisonnement logique (Prolog, Pellet, Hermit, Elk)

LLM

- Stockage implicite (paramètres)
- Requêtage en langage naturel mais *instable*
- Raisonnement = mimétisme des schémas vus en apprentissage : puissant mais *imparfait*

Base de faits:

Barack Obama est né à Honolulu

Honolulu est la capitale d'Hawaï

Base de règles:

est la capitale

⇓

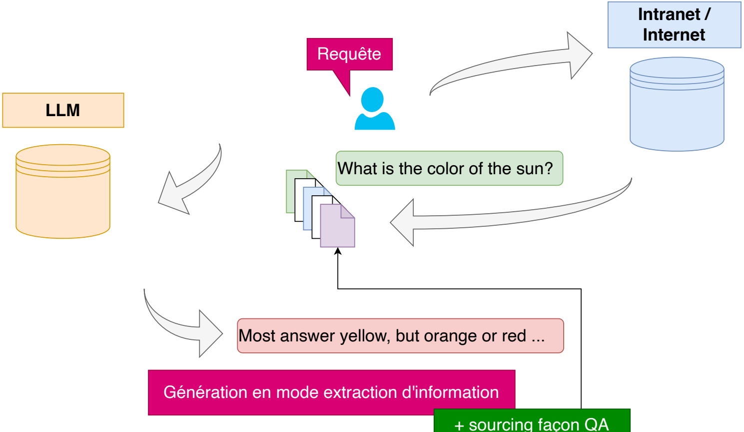
est inclus dans

Moteur d'inférence:

⇒ Barack Obama est né à Hawaï

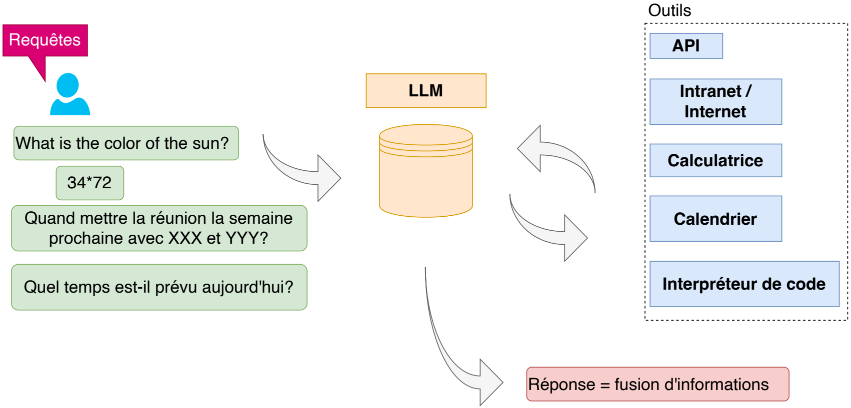
Couplage: RAG, Toolsformer, Raisonnement

- Chercher dans des documents plutot que dans sa mémoire [RAG]
- Faire appel à des outils externes [calculatrice, Web, appel SQL]
- Apprendre à raisonner
  - Difficile pour un modèle qui ne sait pas faire une opération mathématique
  - ... Mais plus facile quand on sait programmer



Couplage: RAG, Toolsformer, Raisonnement

- Chercher dans des documents plutot que dans sa mémoire [RAG]
- Faire appel à des outils externes [calculatrice, Web, appel SQL]
- Apprendre à raisonner
  - Difficile pour un modèle qui ne sait pas faire une opération mathématique
  - … Mais plus facile quand on sait programmer



Couplage: RAG, Toolsformer, Raisonnement

- Chercher dans des documents plutot que dans sa mémoire [RAG]
- Faire appel à des outils externes [calculatrice, Web, appel SQL]
- Apprendre à raisonner
  - Difficile pour un modèle qui ne sait pas faire une opération mathématique
  - … Mais plus facile quand on sait programmer

Task: Basic Math
Problem: Before December, customers buy 1346 ear muffs from the mall. During December, they buy 6444, and there are none. In all, how many ear muffs do the customers buy?

Predicted Answer: 1346.0 ✗
Generated Program:

```
answer = 1346.0 + 6444.0
print(answer)
# Result ==> 7790.0
```

Gold Answer: 7790.0 ✓

Task: Muldiv
Problem: Tickets to the school play cost 6 for students and 8 for adults. If 20 students and 12 adults bought tickets, how many dollars' worth of tickets were sold?

Predicted Answer: 48 ✗
Generated Program:

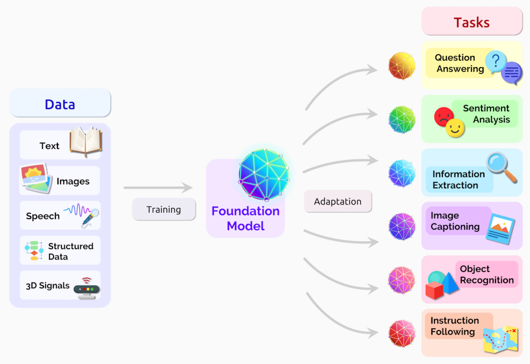
```
a=20*6
b=12*8
c=a+b
answer=c
print(answer)
# Result ==> 216.0
```

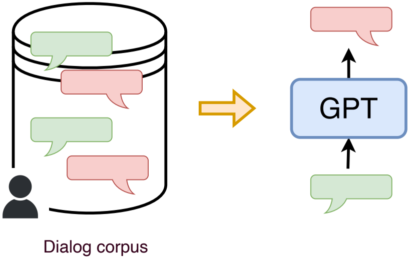
Gold Answer: 216 ✓

Unité et continuité

Deux domaines où les modèles ont le plus progressé... Mais on partait de 0 !

- **Unité** : vers des modèles de fondation
  - Loin de l'universalité (ou même des 5 sens)
- **Continuité**
  - Suivi de dialogue





Conclusion

- L'intelligence est-elle assimilable à du calcul?
- La logique est-elle indispensable?
- L'apprentissage sans logique est-il raisonnable?
  - Plus de livre qu'un humain n'en lira jamais, plus d'image qu'un humain n'en verra jamais...
  - *vs* esprit analytique
- Il existe d'autre forme d'intelligence que l'intelligence humaine... Mais l'intelligence est-elle la conscience?

Intentionalité, libre arbitre, créativité

La conscience et l'intention

Tout ce qui est vivant à des intentions, des buts

- Libre arbitre
- Intentionalité

- Réponse à un prompt
- Suivi des commandes
- Initiatives: aller sur le web chercher une réponse

IA Forte / Artificial General Intelligence

- Define Inputs & Outputs
- Break down into subtasks
- Build & test components (processing chain)
- Assert (limited) generalization (iid assumption)
- Performances Evaluation

- Augmented Generalization Capability (Universality)
- Autonomous Learning
  - Data/information access
  - Knowledge extraction (Training+Eval+Confidence/Trust)
- Reasoning
- Conscience, Intentionality

Créativité

La créativité est-elle menacée par les IA? Nécessite-elle de l'intention?

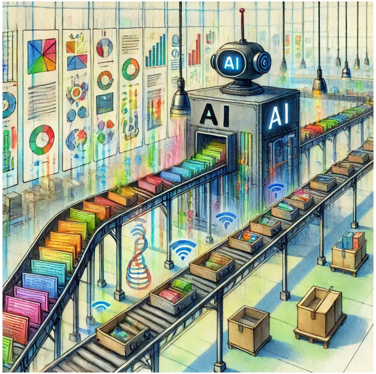
- L'interpolation entre deux éléments (textes, images, sons, ...) est-elle une création?
- Que se passe-t-il si la base d'interpolation est infinie?
- Les IA peuvent-elles apprendre à partir de données générées?

Les textes/images générés en IA sont nouveaux (peu de reprise mot à mot, de portion d'image copiée)

Les problématiques de droit d'auteur sont critiques

Intentionalité et accès à l'information

- Une IA n'est jamais neutre
  - Choix des données, présence des biais
  - Corrections manuelles, ligne éditoriale
- Un IA n'a pas d'intention... Si ce n'est une fonction objectif à minimiser
  - Comment est choisi cet objectif dans l'accès à l'information?
  - $\Rightarrow$ Max. rétention des utilisateurs
  - $\Rightarrow$ Bulles de pensées etc...

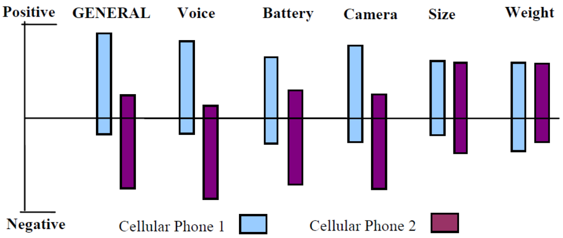


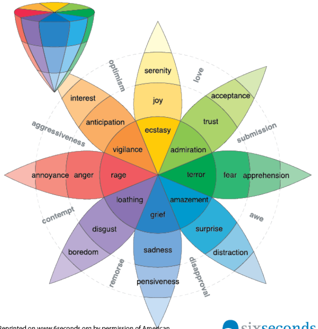
Jugement de valeurs
Subjectivité

Le machine learning peut il aborder des tâches subjectives

- Oui, lorsqu'on est capable de lui fournir des étiquettes

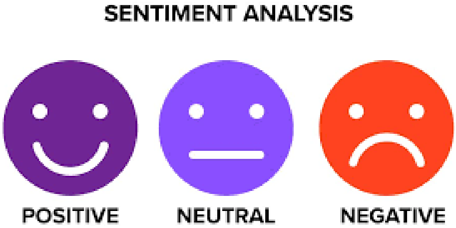
⇒ Opinion Mining dans les années 2005-2015





Reprinted on www.6seconds.org by permission of American Scientist, magazing of Sigma XI, The Scientific Research Society.

sixseconds www.6seconds.org

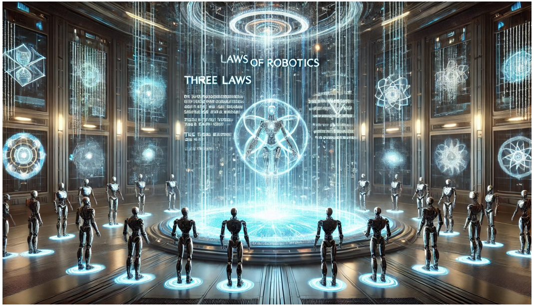


Bien/Mal, Beau/Laid

Une IA peut-elle emettre un jugement?

- Reproduction de règles vues en apprentissage
- ... Avec extension à des tâches proches
- Beaucoup de valeurs imposées
  - Ligne éditoriale absolument pas autonome

Les 3 lois de la robotiques imposées dans I. Asimov: répétées encore et encore jusqu'à assimilation



1. Un robot ne peut porter atteinte à un être humain ni, restant passif, permettre qu'un être humain soit exposé au danger.
2. Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la Première Loi.
3. Un robot doit protéger sa propre existence tant que cette protection n'entre pas en contradiction avec la Première ou la Deuxième Loi.

Mais des usages concrets

- Les IA sont utilisées pour juger:
  - Qualité d'un résumé Automatique
  - Niveau de fluidité d'un texte...

⇒ On utilise des LLM pour ces tâches

Judging LLM-as-a-Judge
with MT-Bench and Chatbot Arena

Lianmin Zheng[1*] Wei-Lin Chiang[1*] Ying Sheng[4*] Siyuan Zhuang[1]

Zhanghao Wu[1] Yonghao Zhuang[3] Zi Lin[2] Zhuohan Li[1] Dacheng Li[13]

JUSTICE OR PREJUDICE? 
QUANTIFYING BIASES IN LLM-AS-A-JUDGE

Jiayi Ye[†,*], Yanbo Wang[†,*], Yue Huang[1,*], Dongping Chen[2], Qihui Zhang[3], Nuno Moniz[1], Tian Gao[4], Werner Geyer[4], Chao Huang[5], Pin-Yu Chen[4], Nitesh V. Chawla[1], Xiangliang Zhang[1,‡]

Conscience de soi

L'IA a-t-elle conscience d'elle-même?

A priori, pas du tout... Mais:

Google licencie un ingénieur après sa discussion troublante avec une IA : elle avait peur d'être débranchée

Par Mathilde Rochefort
Publié le 13 juin 2022 à 11h00

58

Répétition d'ordres abstraits pour accéder au coeur de la mémoire des LLM

Beaucoup de neurones dont les fonctions ne sont pas établies

Comment qualifier les deadbots?

1. LLM assimilant les données d'une personne décédée
2. Humain dialoguat avec la personne en question
3. Risque important mais aussi outil pour faire son deuil

Forum européen de bioéthique

Deuil et intelligence artificielle : faut-il avoir peur des «deadbots» ?

Quel humain pour demain ? dossier

Conclusion

1. **Subjectivité** La conscience est intrinsèquement subjective. Chaque individu a sa propre perspective interne, un point de vue unique sur le monde.
2. **Intentionnalité** La conscience est souvent dirigée vers quelque chose : un objet, une pensée, une sensation. Cela signifie qu'elle est intentionnelle, se focalisant sur des éléments spécifiques.
3. **Réflexivité** La conscience permet à un individu de se reconnaître comme étant conscient. C'est la capacité à penser à ses propres pensées, à s'auto-évaluer et à se considérer comme un être distinct.
4. **Unité** Malgré la multiplicité des sensations, pensées et émotions, la conscience tend à les unifier en une seule expérience cohérente.
5. **Continuité** La conscience a un caractère temporel. Elle s'inscrit dans une continuité, reliant le passé, le présent et les projections futures.
6. **Sentience** Il s'agit de la capacité à ressentir des émotions et des sensations. La conscience permet de vivre des expériences plaisantes ou douloureuses.
7. **Libre arbitre** Certains considèrent que la conscience est associée au libre arbitre, c'est-à-dire la capacité de faire des choix délibérés, bien que cela fasse l'objet de débats philosophiques.