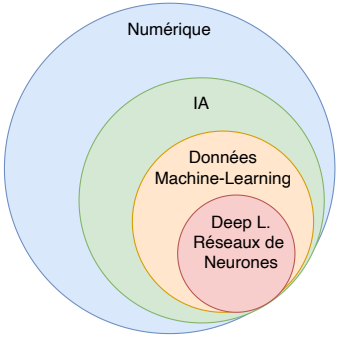
DES MODÈLES DE LANGUE À L'EXTRACTION D'INFORMATION

École ETLMIC 2026
Cours 1

Vincent Guigue & Rémy Découpes
https://vguigue.github.io

Introduction

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

IA : programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau.

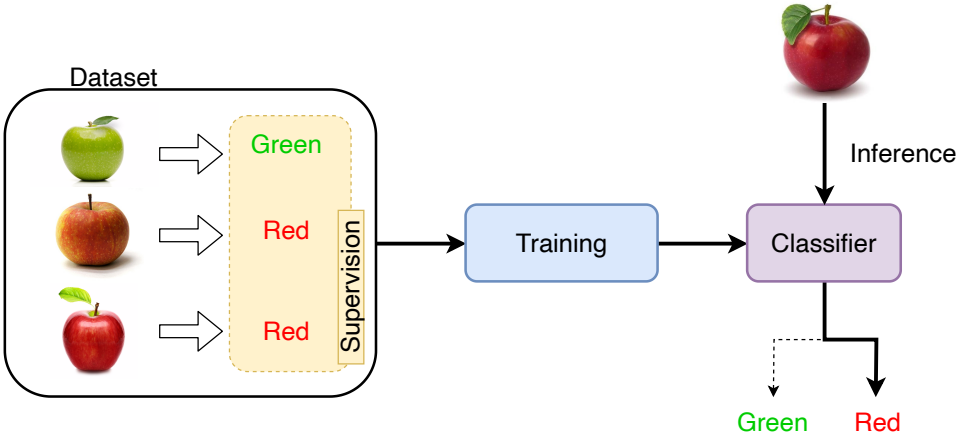
Marvin Lee Minsky, 1956

N-AI (Narrow Artificial Intelligence), dédiée à une tâche unique

$\neq$ **IA-G (IA Générale)**, qui remplace l'humain dans les systèmes complexes.

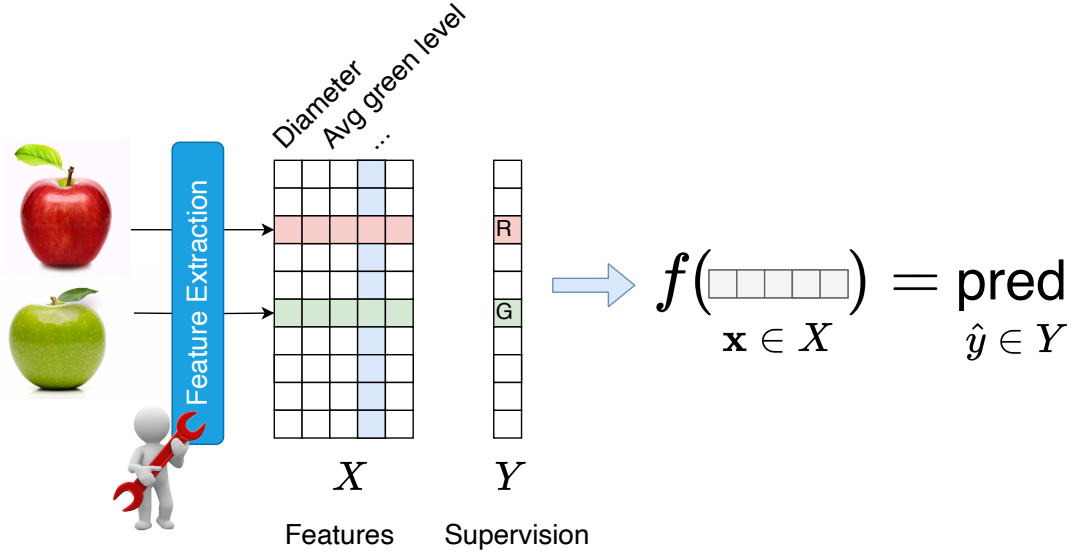
Andrew Ng, 2015

Chaîne de Traitement Supervisé & Modèles

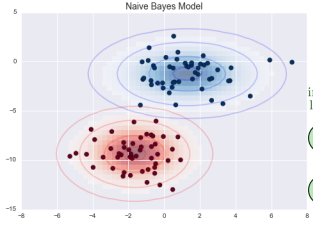


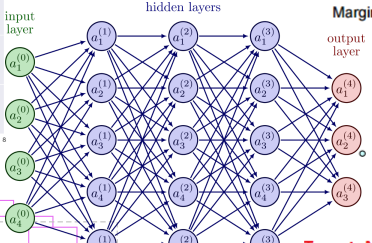
- Promesse = construire un modèle *uniquement* à partir d'observations

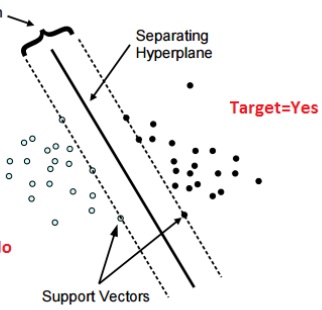
Chaîne de Traitement Supervisé & Modèles

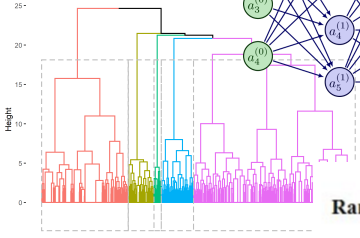


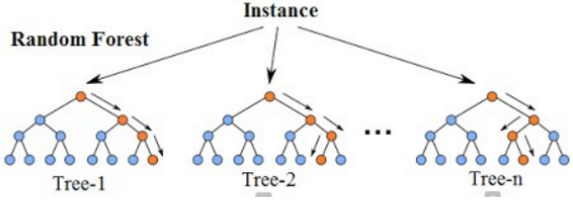
Chaîne de Traitement Supervisé & Modèles



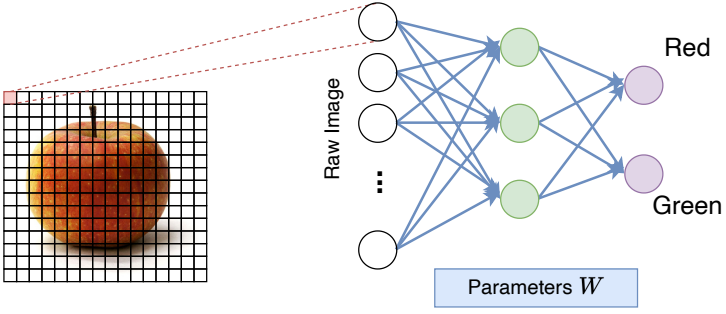






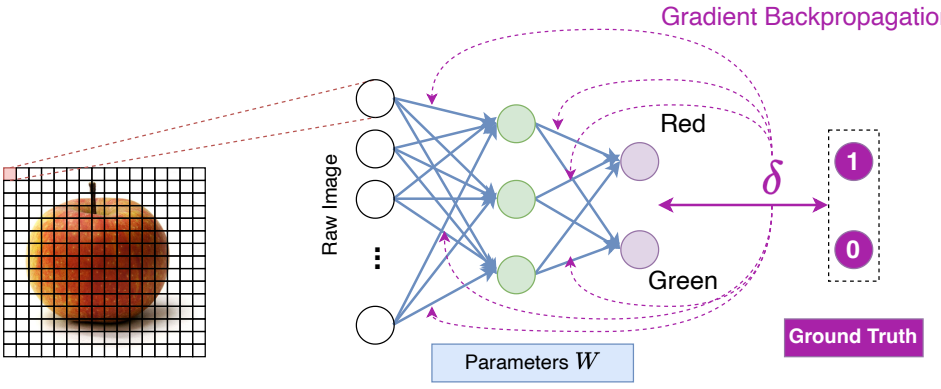


Chaîne de Traitement Supervisé & Modèles



- Initialisation aléatoire... Et prise de décision aléatoire (au début !)

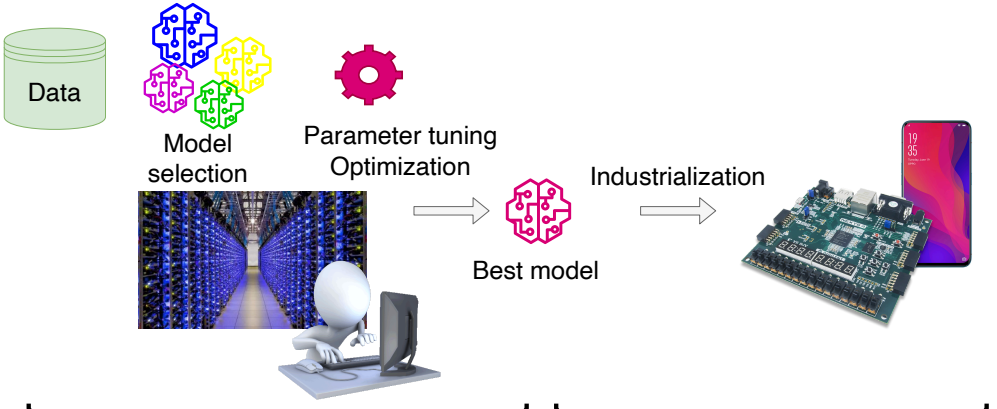
Chaîne de Traitement Supervisé & Modèles



- Mise à jour des poids
- Pas de taille epsilon, nombreuses itérations sur les données

Apprentissage/Construction *vs* Déployement/Exploitation

Différentes étapes en apprentissage automatique



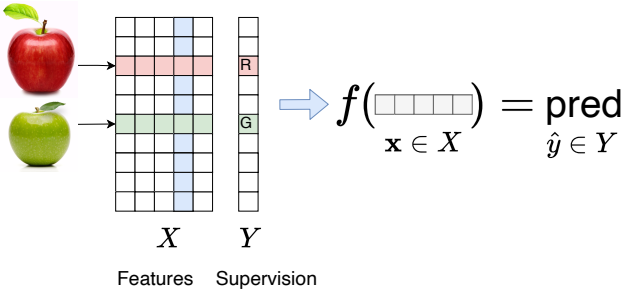
Deep learning & Apprentissage de représentations

[Application aux données textuelles]

Des données tabulaires au texte

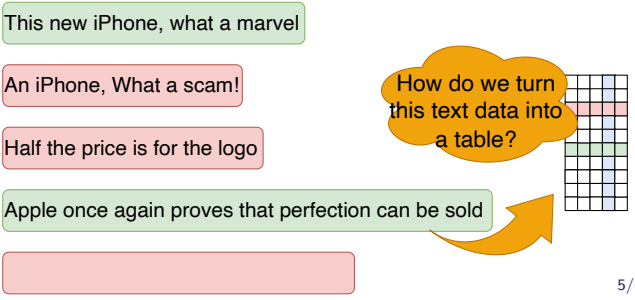
- **Données tabulaires**
  - Dimension fixe
  - Valeurs continues

  ⇒ Un terrain de jeu idéal pour l'apprentissage automatique



- **Données textuelles**
  - Longueurs variables
  - Valeurs discrètes

  ⇒ Complexes pour l'apprentissage automatique

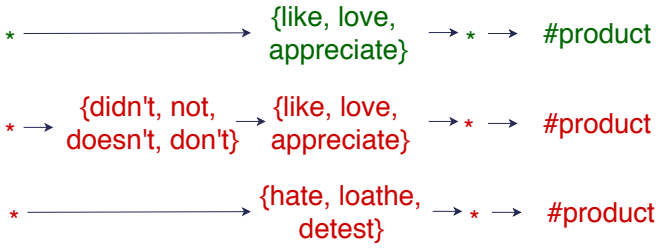


IA + Texte : Trait. Auto. du Langage Naturel (TALN)

TALN = plus grande communauté scientifique en IA

Linguistique [1960-2010]

Systèmes à base de règles :

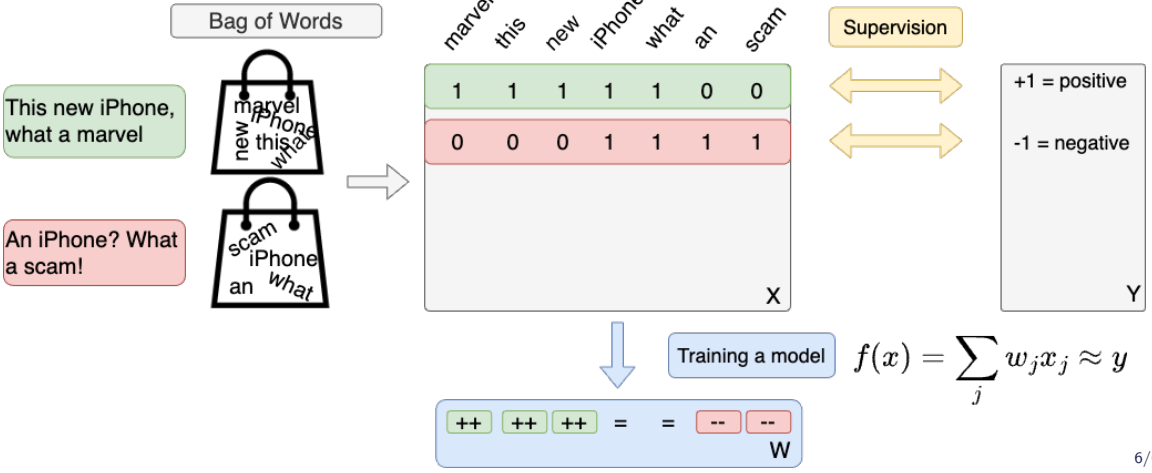


- Nécessite une expertise humaine
- Extraction de règles ⇔ données très propres
- Très grande précision
- Faible rappel
- Système interprétable

IA + Texte : Trait. Auto. du Langage Naturel (TALN)

TALN = plus grande communauté scientifique en IA

Apprentissage auto. [1990-2015]



IA + Texte : Trait. Auto. du Langage Naturel (TALN)

TALN = plus grande communauté scientifique en IA

Linguistique [1960-2010]

- Nécessite une expertise humaine
- Extraction de règles ⇔ données très propres

\+ Système interprétable
\+ Très grande précision
− Faible rappel

Apprentissage auto. [1990-2015]

- Peu d'expertise nécessaire
- Extraction statistique ⇔ robuste aux données bruitées

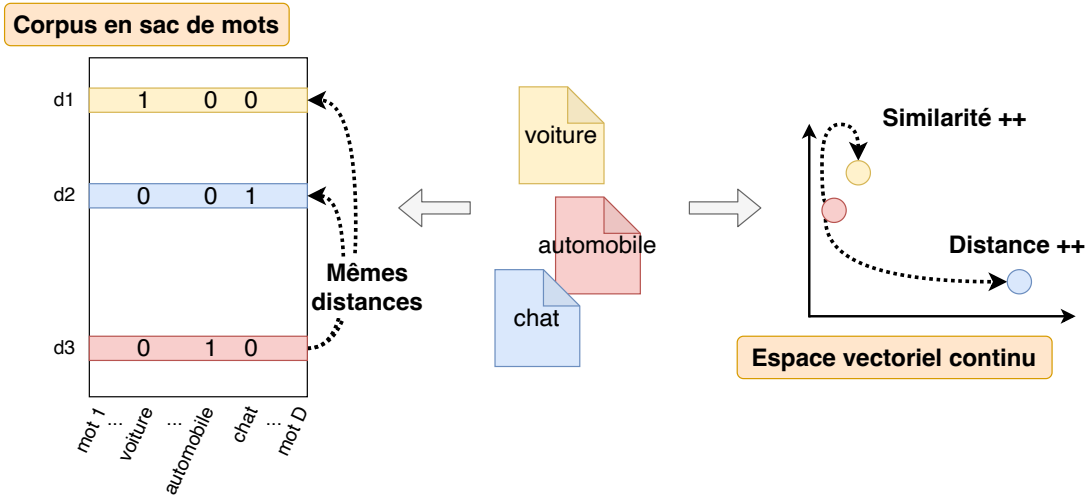
≈ Système moins interprétable
− Moindre précision
\+ Meilleur rappel

Précision = critère d'acceptation par l'industrie

→ Lien vers les métriques

Apprentissage de représentations pour les données textuelles

Du sac de mots aux représentations vectorielles [2008, 2013, 2016]

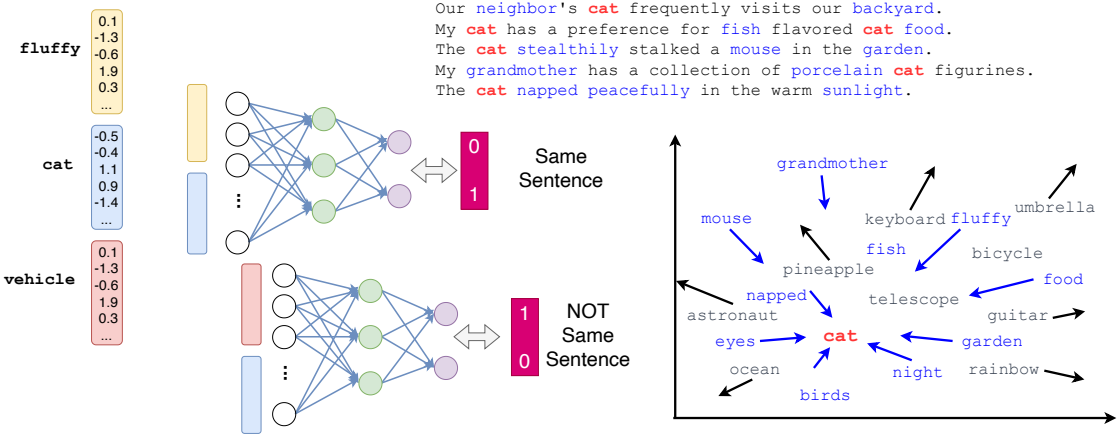


LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Apprentissage de représentations pour les données textuelles

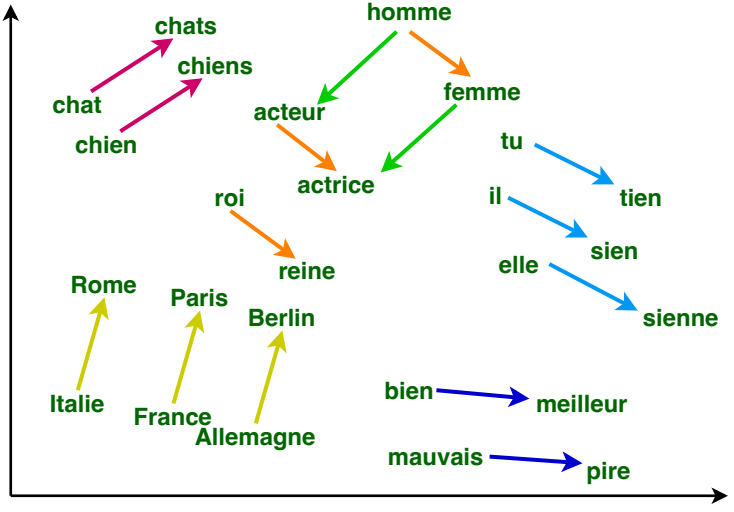
Du sac de mots aux représentations vectorielles [2008, 2013, 2016]

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.



Apprentissage de représentations pour les données textuelles

Du sac de mots aux représentations vectorielles [2008, 2013, 2016]



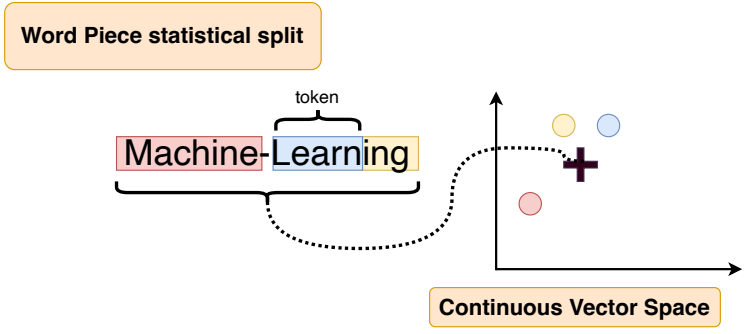
- Espace sémantique :
  significations similaires
  $\Leftrightarrow$
  positions proches
- Espace structuré :
  régularités grammaticales, connaissances de base, ...

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

Apprentissage de représentations pour les données textuelles

Du sac de mots aux représentations vectorielles [2008, 2013, 2016]

Des mots aux tokens

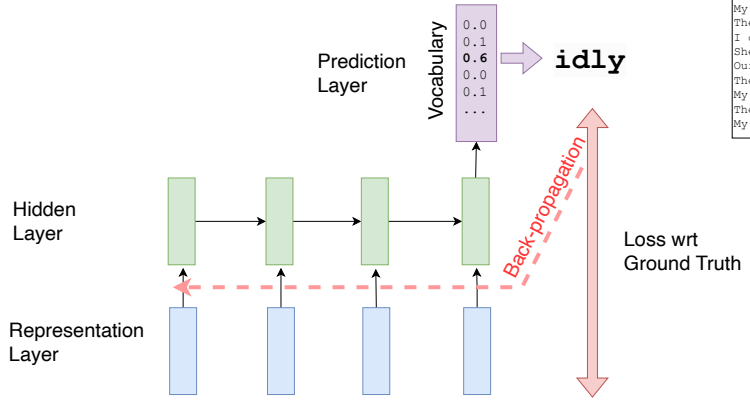


- Représentation des mots inconnus
- Adaptation aux domaines techniques
- Résistance aux fautes d'orthographe

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

Agrégation des représentations de mots : vers l'IA générative

- Génération et représentation
- Nouvelle manière d'apprendre les positions des mots

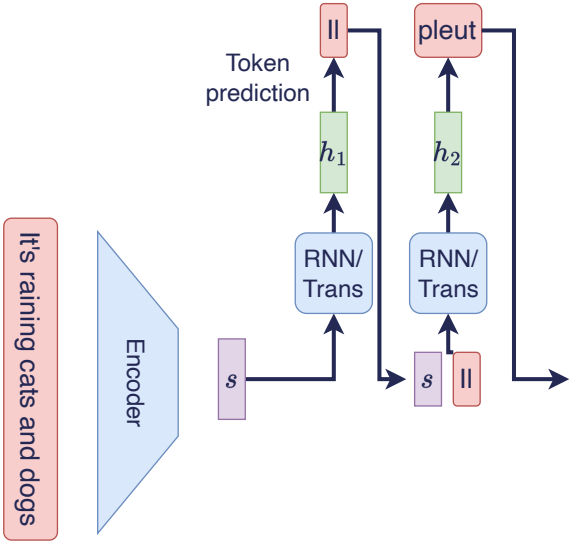


The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
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.

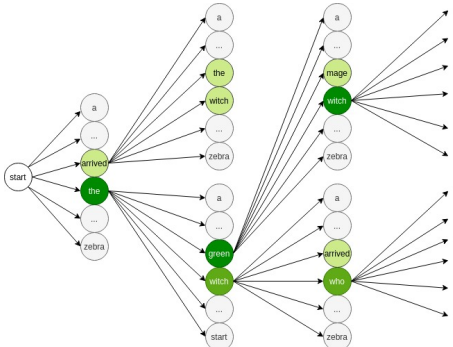
Corpus

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

Inférence & Beam Search



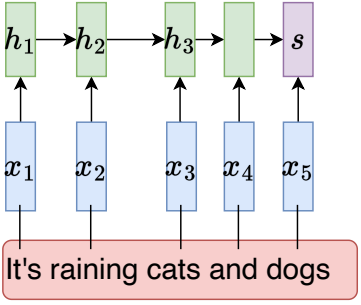
- Coût élevé $\approx$ 1 appel / token
- Principe du max de vraisemblance
- Tâches historique du TALN =
  - Archi. spécifique / data spécifique
  - Ajouter des contraintes a posteriori dans le décodage



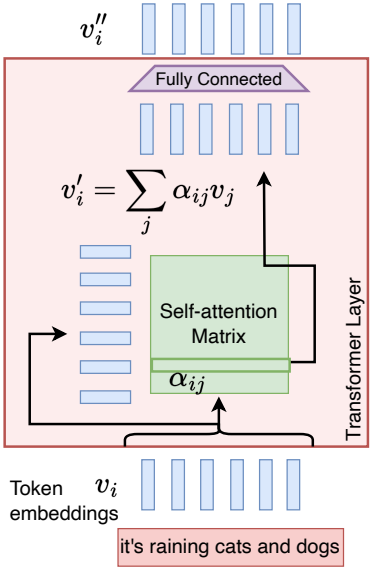
Architecture Transformer : agrégation à l'état de l'art

Réseau de neurones récurrents :

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



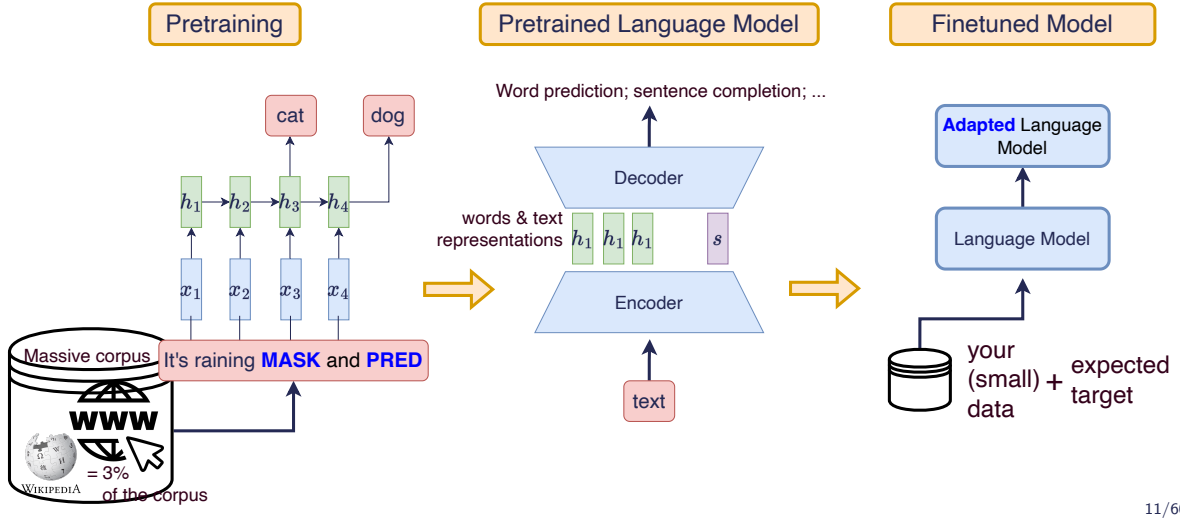
Transformer :



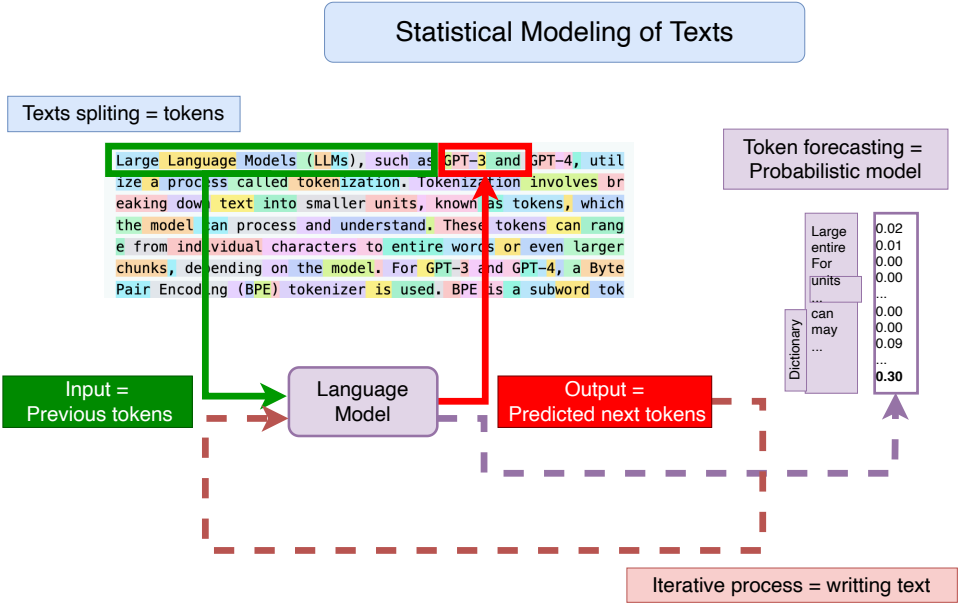
Attention is all you need, Vaswani et al. NeurIPS 2017
Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

Un nouveau paradigme de développement depuis 2015

- Jeu de données massif + architecture massive ⇒ coût d'entraînement +++
- Architecture pré-entraînée + zéro-shot / affinage



Au bout du compte: un perroquet stochastique :)



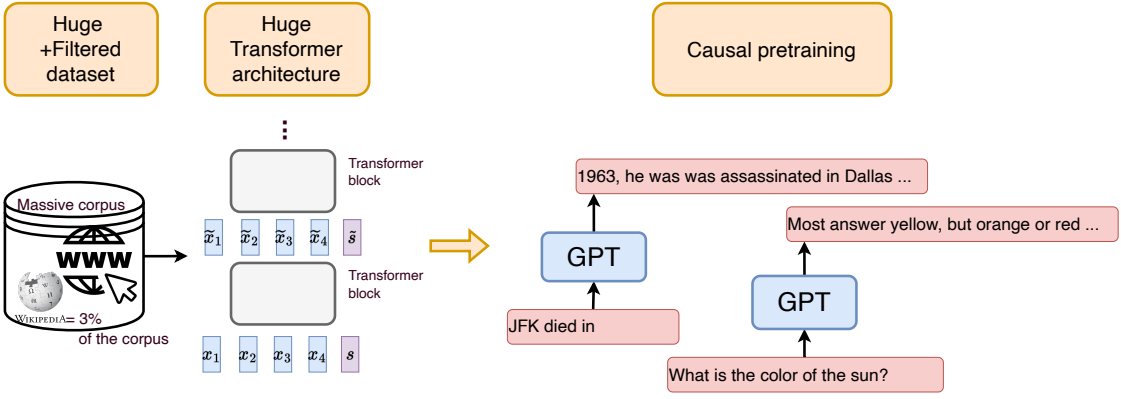
chatGPT

30 novembre, 2022

1 million d'utilisateurs en 5 jours
100 million à la fin janvier 2023
1.16 milliard en mars 2023

Les ingrédients de chatGPT

0. Transformer + données massives (GPT)



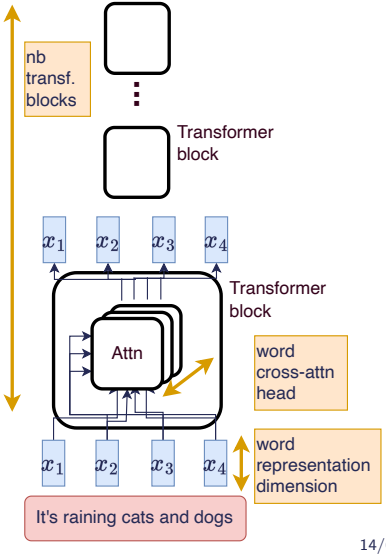
- Grammaire : accord singulier/pluriel, concordance des temps
- Connaissances : entités, nom, lieux, dates, ...

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

Les ingrédients de chatGPT

1. Plus, c'est mieux ! (GPT)

- \+ plus de mots en entrée [500 ⇒ 2k, 32k, 100k]
- \+ plus de dim. dans l'espace des mots [500-2k ⇒ 12k]
- \+ plus de têtes d'attention [12 ⇒ 96]
- \+ plus de blocs/couches [5-12 ⇒ 96]



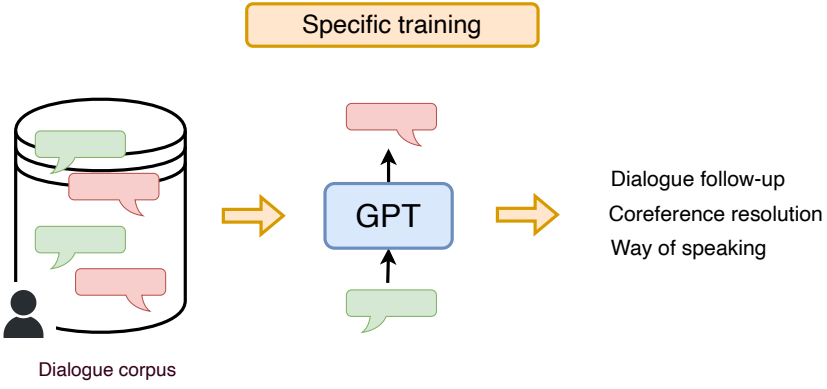
175 milliards de paramètres...

Qu'est-ce que cela signifie ?

- $1{,}75 \cdot 10^{11}$ ⇒ 300 Go + 100 Go (stockage pour l'inférence) ≈ 400 Go
- GPU NVidia A100 = 80 Go de mémoire (=20k€)
- Coût de (1) entraînement : 4,6 millions €

Les ingrédients de chatGPT

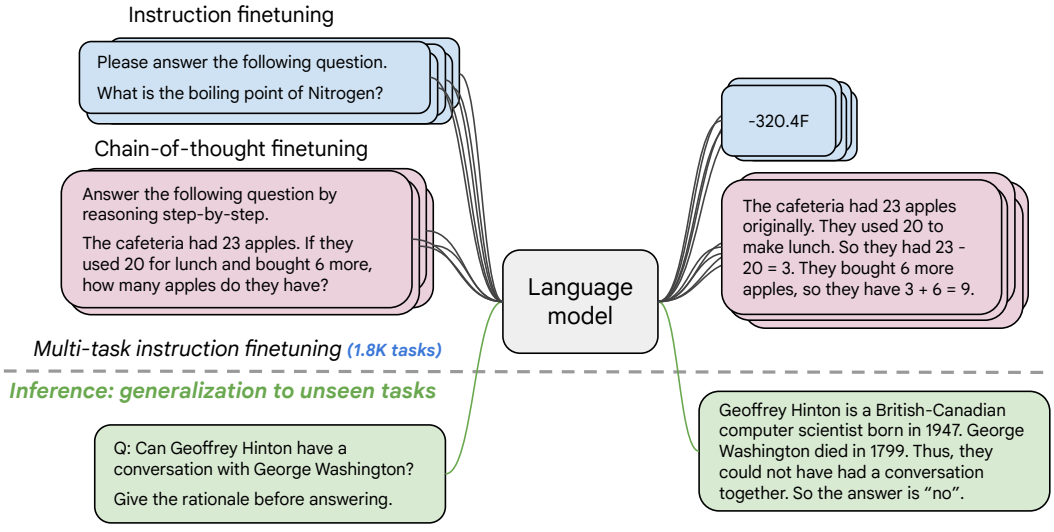
2. Suivi du dialogue



- **Données très propres** Données générées/validées/classées par des humains

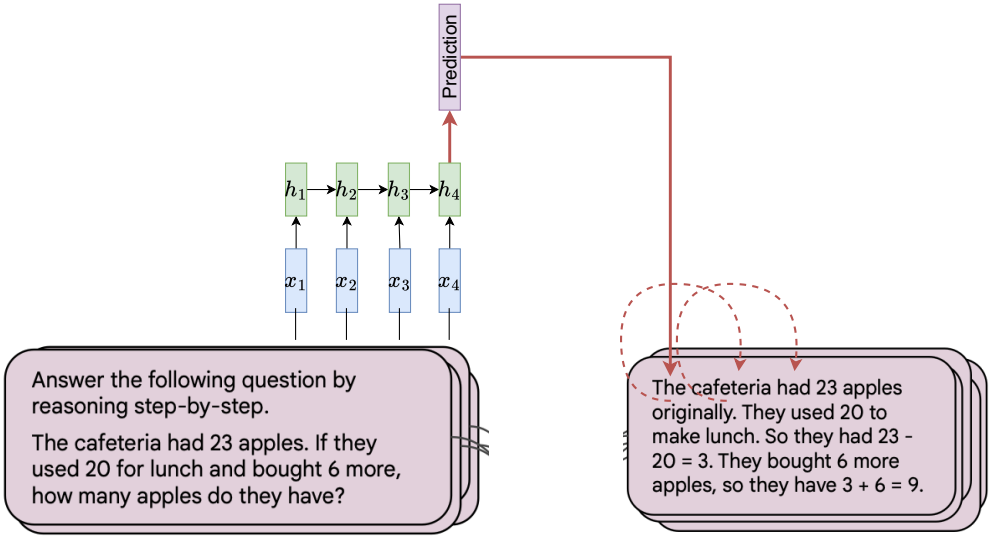
Les ingrédients de chatGPT

3. Ajustement fin sur des tâches de raisonnement (±) complexes



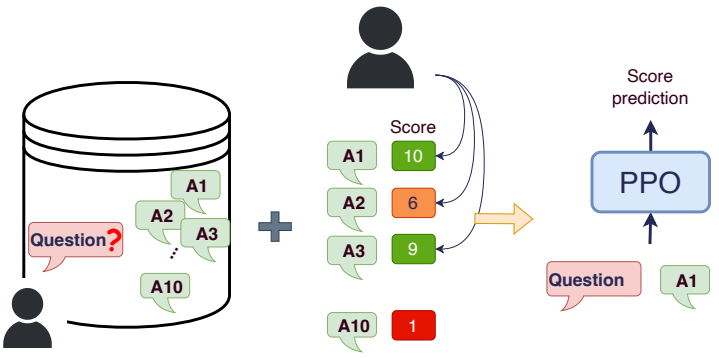
Les ingrédients de chatGPT

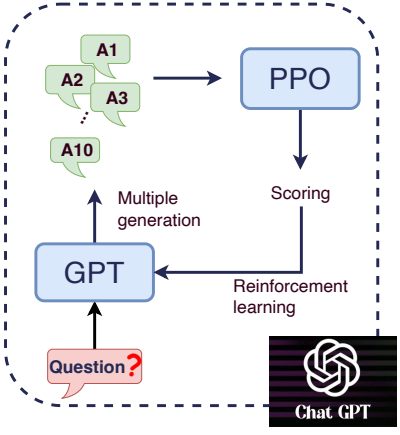
3. Ajustement fin sur des tâches de raisonnement (±) complexes



Les ingrédients de chatGPT

4. Instructions + classement des réponses



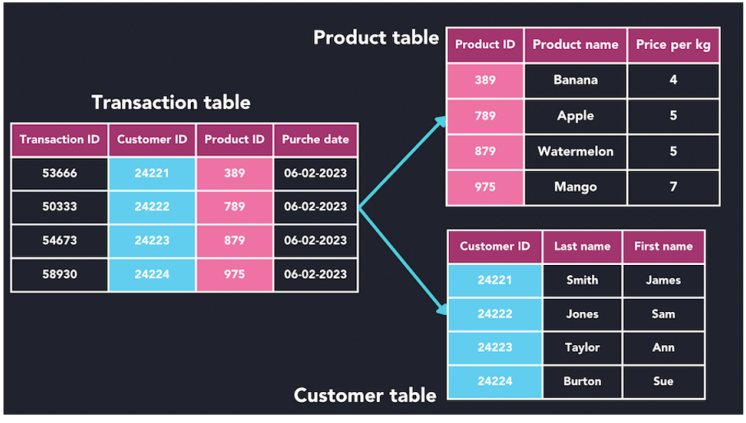


- Base de données créée par des humains
- Amélioration des réponses
- ... Aussi un moyen d'éviter les sujets sensibles = censure

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

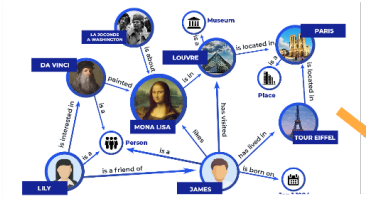
Un nouveau rapport aux connaissances

- Bases de données + accès en SQL



Un nouveau rapport aux connaissances

- Bases de données + accès en SQL
- Knowledge Graphs / Différents types de liens ⇒ Inférence logique



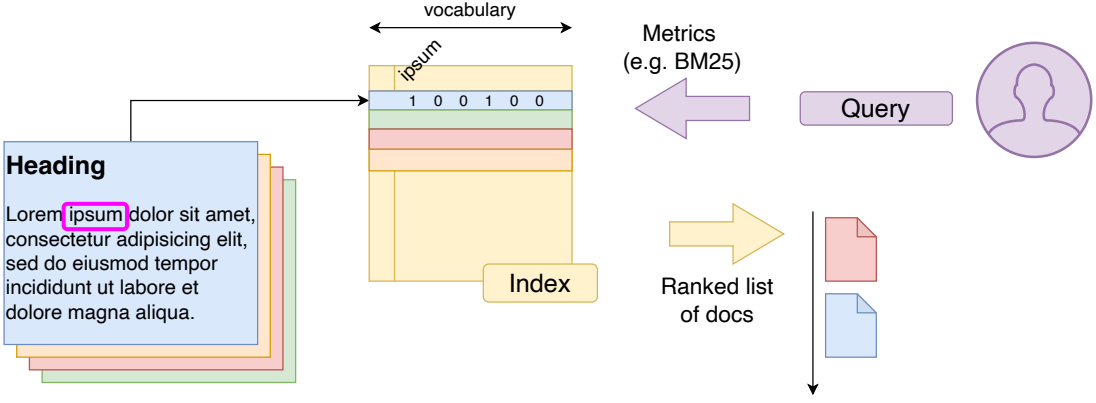




Un nouveau rapport aux connaissances

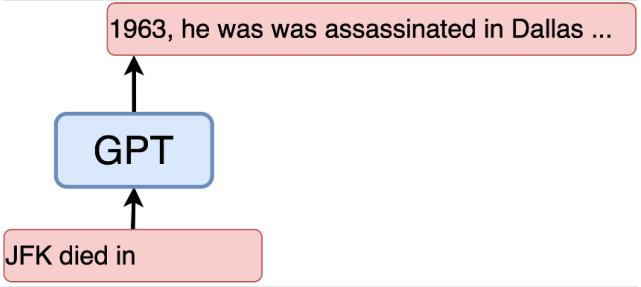
- Bases de données + accès en SQL
- Knowledge Graphs / Différents types de liens
- Indexation du web / moteurs de recherche

$\Rightarrow$ Inférence logique



Un nouveau rapport aux connaissances

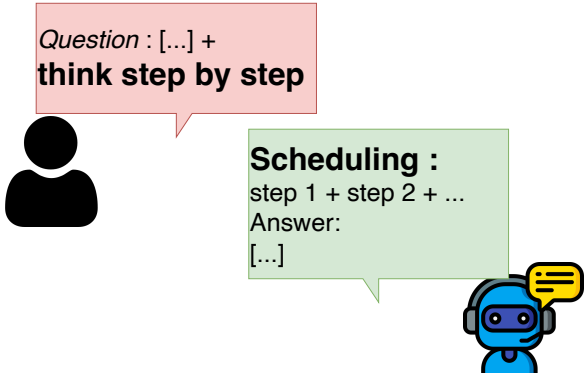
- Bases de données + accès en SQL
- Knowledge Graphs / Différents types de liens ⇒ Inférence logique
- Indexation du web / moteurs de recherche
- ... Loi de probabilité !



$$p(w_{t+1}|w_1, w_2, \ldots, w_t)$$

Un nouveau rapport aux connaissances

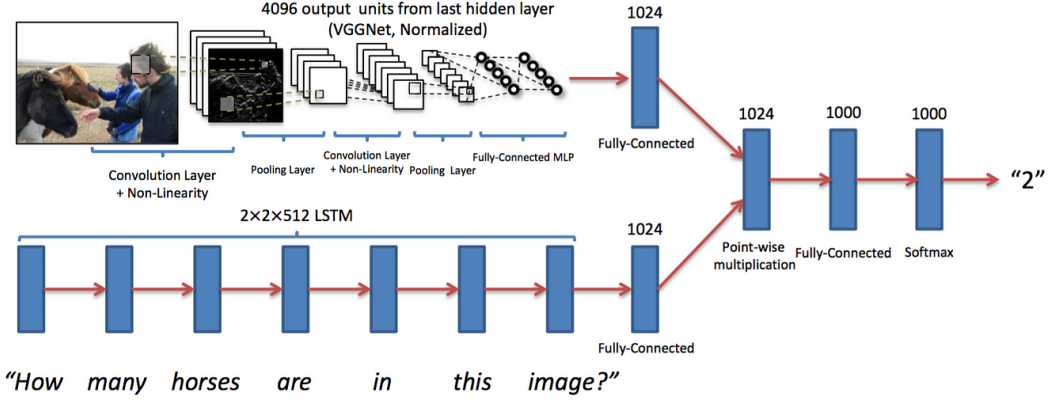
- Bases de données + accès en SQL
- Knowledge Graphs / Différents types de liens ⇒ Inférence logique
- Indexation du web / moteurs de recherche
- ... Loi de probabilité !
- Et de nouvelles options associées (=raisonnement) orienté sur la planification, l'exploration de solutions...



GPT-4 & Multimodalité

Fusionner info. texte + image. **Apprendre** l'information conjointement

Exemple du VQA : Visual Question Answering (questions visuelles)



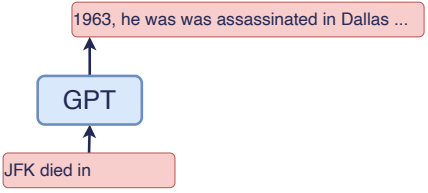
⇒ Rétropropager l'erreur ⇒ modifier les repr. des mots + l'analyse d'image

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

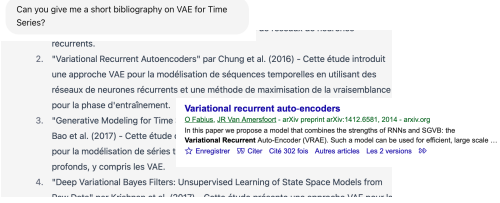
Limites techniques des modèles de langue

chatGPT et la relation à la vérité

1. **Vraisemblance** = grammaire, accords, concordance des temps, enchaînements logiques...
   ⇒ Connaissances répétées
2. Prédit le mot le plus **plausible**...
   ⇒ produit des **hallucinations**
3. chatGPT ≠ **graphes de connaissances**
4. Réponses brillantes...
   Et erreurs absurdes !
   + on ne peut pas prévoir les erreurs
5. Quantification de l'incertitude peu fiable



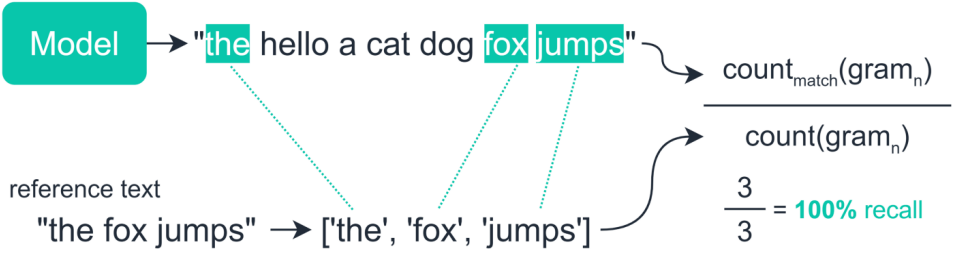
Exemple : produire une bibliographie



IA générative : comment évaluer la performance ?

Le point critique aujourd'hui

- Comment évaluer par rapport à la vérité terrain ?
- Comment évaluer la confiance du système / la plausibilité de la génération ?

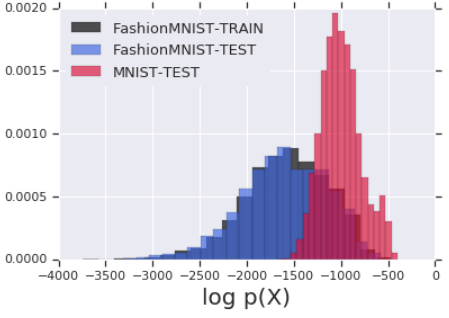


The Ultimate Performance Metric in NLP, J. Briggs, Medium 2021

IA générative : comment évaluer la performance ?

Le point critique aujourd'hui

- Comment évaluer par rapport à la vérité terrain ?
- Comment évaluer la confiance du système / la plausibilité de la génération ?



Plausibilité



Entraînement



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

Stabilité / Prévisibilité

- Difficile d'encadrer un comportement
- Impossible de prédire les bonnes / mauvaises réponses

⇒ Des modèles qui se discréditent régulièrement

Peu / pas utilisés dans les jeux vidéo

⇒ Impossible de certifier ces modèles pour des applications critiques



V how old is Obama

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

Stabilité / Prévisibilité

- Difficile d'encadrer un comportement
- Impossible de prédire les bonnes / mauvaises réponses



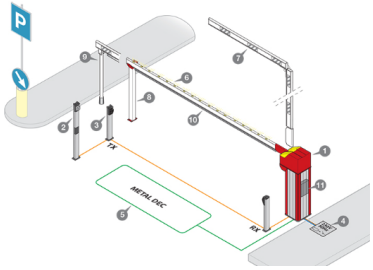
⇒ Des modèles qui se discréditent régulièrement

Peu / pas utilisés dans les jeux vidéo

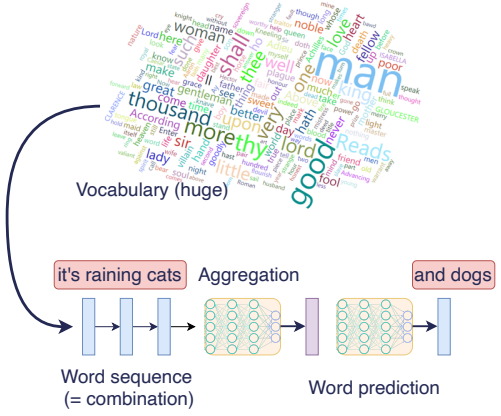
⇒ Impossible de certifier ces modèles pour des applications critiques



Stabilité, explicabilité... Et complexité



- Système simple
- Tests exhaustifs des entrées/sorties
- **Prévisible** & **explicable**



- Grande dimension
- Combinaisons non linéaires complexes
- **Non prévisible** & **non explicable**

Stabilité, explicabilité... Et complexité

Interprétabilité *vs* explication a posteriori

Réseaux de neurones = **non interprétables** (presque toujours)

trop de combinaisons à anticiper

Réseaux de neurones = **explicables a posteriori** (presque toujours)



[Accident Uber, 2018]

- Système simple
- Tests exhaustifs des entrées/sorties
- **Prévisible** & **explicable**

- Grande dimension
- Combinaisons non linéaires complexes
- **Non prévisible** & **non explicable**

Transparence : open source / poids ouverts

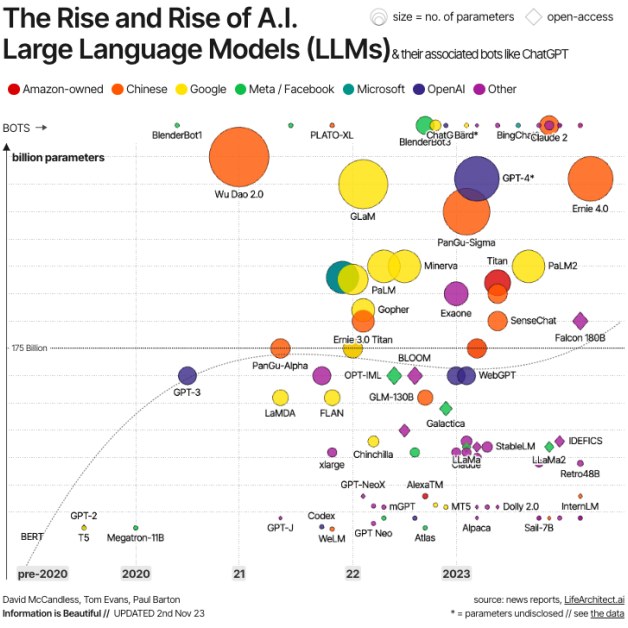
- Puis-je le modifier ? — Adaptation
- Quelles données d'entraînement ? — Contamination des données / compétences
- Quelle ligne éditoriale / censure ? — Accès à l'information
- Pourquoi cette réponse ? — Explicabilité / interprétabilité

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

Source: 2023 Foundation Model Transparency Index

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

Coûts / Frugalité



Paramètres

1998	LeNet-5	= 0,06M
2011	Senna	= 7,3M
2012	AlexNet	= 60M
2017	Transformer	= 65M / 210M
2018	BERT	= 110M / 340M
2019	GPT2	= 1 500M
2020	GPT3	= 175 000M

Comportements différents, coûts différents

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	**5**min ampoule LED	**33**s vidéos en ligne

Voir plus

DeepSeek/**DeepSeek v3**

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

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

Impact énergétique de la discussion

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

Ce qui correspond à :

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

Voir plus

Comportements différents, coûts différents

Des coûts différents selon les utilisateurs / les langues

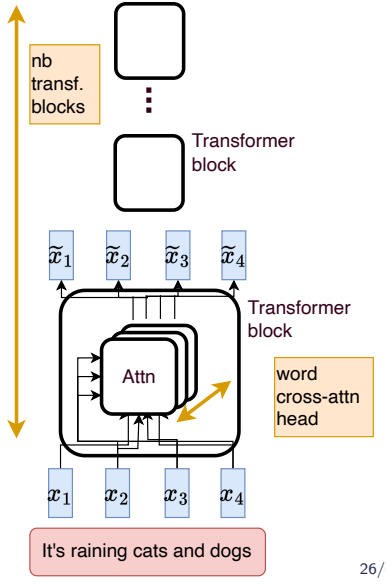
Pour un texte significatif en Français

and the same in English

TOKENS 17 CHARACTERS 63

<s> Pour un texte significatif en Français

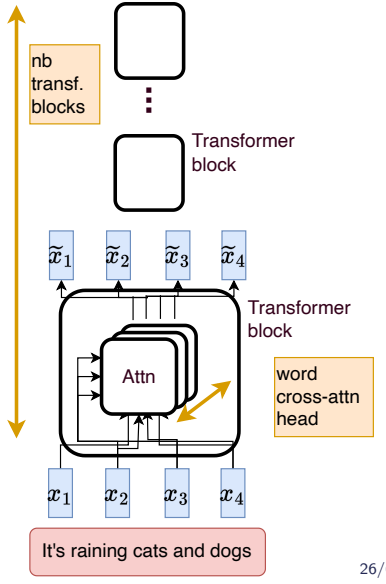
and the same in English



Comportements différents, coûts différents

Des coûts différents selon les utilisateurs / les langues





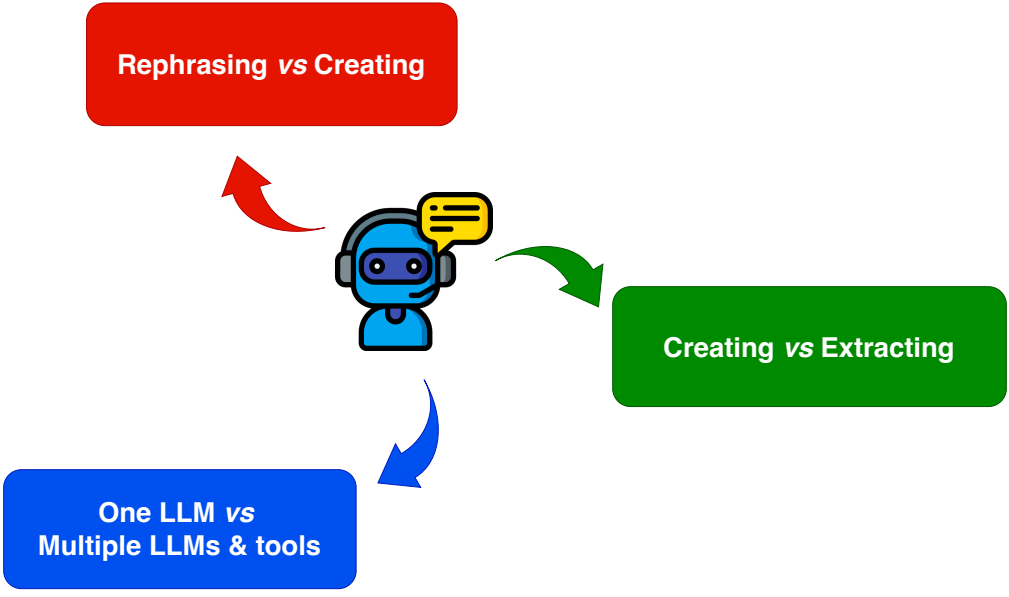
Tout ce qui dépasse les capacités / l'entraînement du LLM

- Calculs simples (multiplication, division)
- Générer des noms d'animaux en n syllabes (en cours)
- Jouer aux échecs
- Suivre un raisonnement causal (complexe)
- ...



Typologie des usages des LLMs

Trois questions pour cadrer les usages



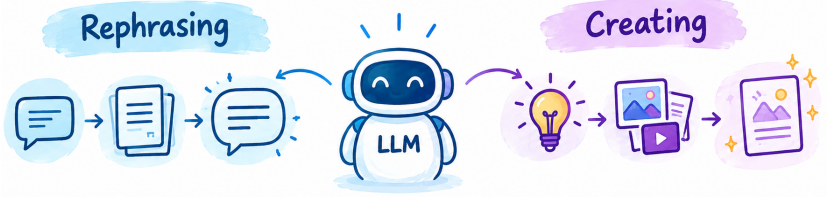
(1) Reformuler *vs* créer du contenu

Reformuler *j'apporte le contenu*

- **Traduire**/**relire** un paragraphe
- **Reformuler** : style, ton, -10%
- Paragraphe à partir de **mes** puces
- Compte-rendu depuis **mes** notes
- Lettre de motivation : CV + offre
- Tableaux, OCR, LaTeX/Excel
- Le résumé de **mon propre** article

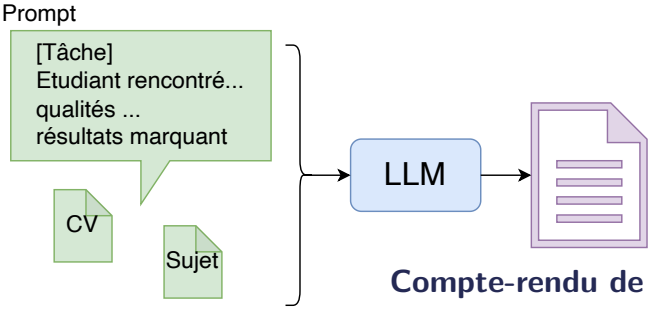
Créer *le modèle apporte le contenu*

- **Trouver** l'inspiration [page blanche]
- **Proposer** un plan, un acronyme
- **Répondre** là où je ne sais pas
- **Tester** une idée, ses faiblesses
- Rédiger sur un sujet donné
  mais non détaillé
- Rédiger une relecture, un état de l'art

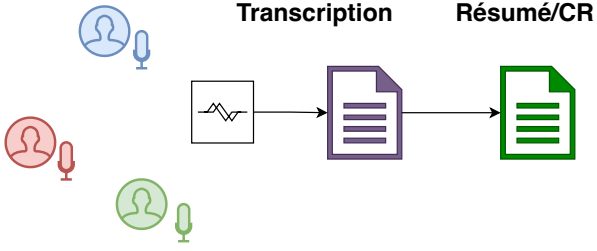


(1) Reformuler : exemples

Construire une lettre de recommandation



Compte-rendu de réunion



(1) Questions éthiques

Reformuler

Le contenu est **le mien**. Le modèle touche à la *forme*.

→ Suis-je toujours l'auteur ? **oui**

→ Dois-je le déclarer ?
comme un correcteur ? un traducteur ?

→ Vrai risque = **distorsion silencieuse**
une nuance perdue, une précaution effacée

→ Vrai risque = **confidentialité**
texte inédit envoyé sur un serveur

Créer

Le contenu vient **du modèle**, *i.e.* d'un corpus que je n'ai pas choisi.

→ Suis-je toujours ~~l'auteur~~ responsable ? **oui**

→ Qui dois-je citer ?
le modèle n'a pas de source

→ Vrai risque = **hallucination**
plausible, bien écrit, faux

→ Vrai risque = **uniformisation de la pensée**
l'idée la plus probable, pas la meilleure

⇒ Même outil, même zone de saisie, **deux débats totalement différents**.

(2) Créer *ex-nihilo* *vs* extraire de l'information

Ex-nihilo *mémoire paramétrique*

- La réponse est **dans les poids**
- *JFK est mort en...* ⇒ réponse probabiliste
- Pas de source, pas de traçabilité
- Figé au moment de l'entraînement

⇒ hallucination (mesure de *factualité*)

Extraction *le document est fourni*

- La réponse est **dans l'entrée**
- Le modèle *localise*, il n'invente pas
- La source est connue *par construction*
- À jour *= aussi frais que le document*

⇒ vérifiable (mesure de *fidélité*)

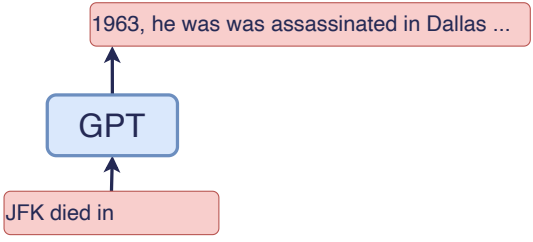


(2) Créer *ex-nihilo* *vs* extraire de l'information

Ex-nihilo *mémoire paramétrique*

- La réponse est **dans les poids**
- *JFK est mort en...* ⇒ réponse probabiliste
- Pas de source, pas de traçabilité
- Figé au moment de l'entraînement

⇒ hallucination (mesure de *factualité*)



Extraction *le document est fourni*

- La réponse est **dans l'entrée**
- Le modèle *localise*, il n'invente pas
- La source est connue *par construction*
- À jour = *aussi frais que le document*

⇒ vérifiable (mesure de *fidélité*)

Passage Sentence

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

Question

What causes precipitation to fall?

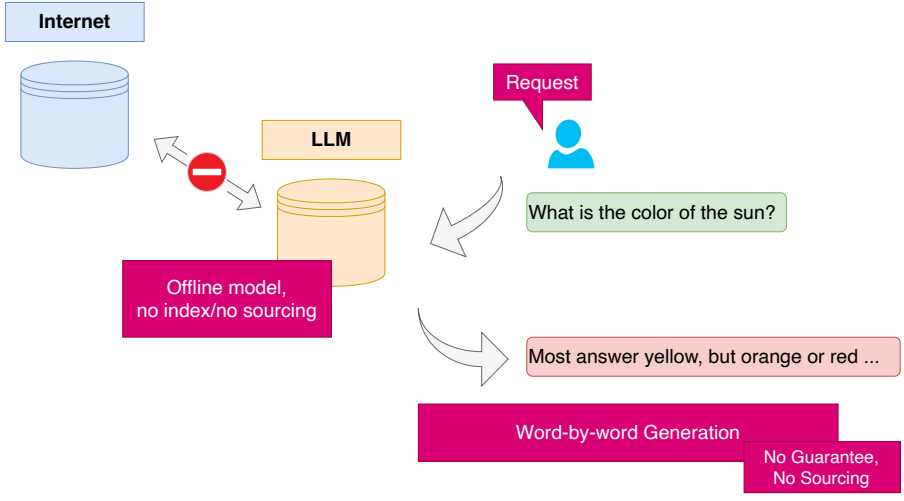
Answer Candidate

gravity

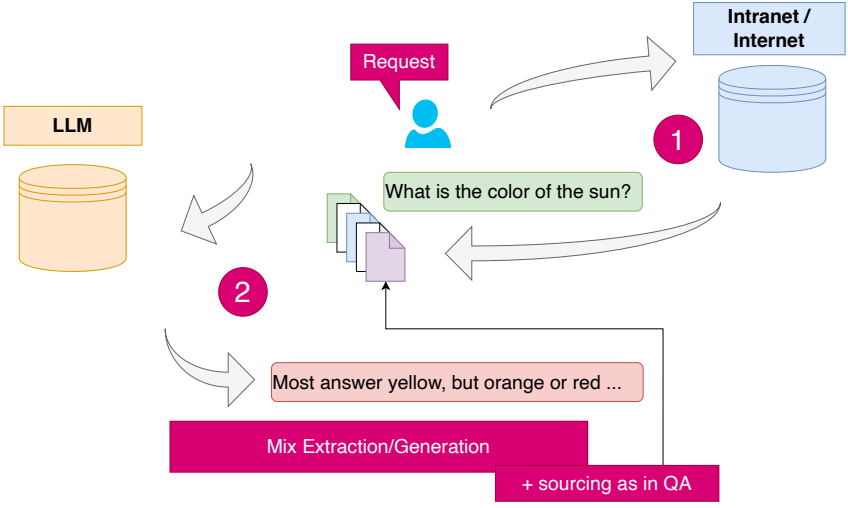
⇒ Même question, même interface... mais une seule des deux est vérifiable.

(2) LLMs ⇒ RAG : mémoire *vs* extraction d'information

- Poser des questions à ChatGPT... Un usage surprenant !
- Mais est-ce raisonnable ? [Une vraie question ouverte (!)]



(2) LLMs ⇒ RAG : mémoire *vs* extraction d'information



- RAG : *Retrieval-Augmented Generation*
- Taille d'entrée limitée (2k, 32k, 200k tokens)

(2) Dans un contexte scientifique

- Le RAG en contexte scientifique devrait s'appeler **bibliographie** !
  - Consensus, Scopus.ai, opscidia, ...

  ⇒ excellente synthèse + analyse... Sur quels articles ?
  - NotebookLM : choisissez vos articles (jusqu'à 50)... Puis lancez la discussion

  ⇒ analyse, regroupement thématique, comparaison, ...
- Pour dialoguer avec un manuscrit...
  Il est trop long pour un LLM

  ⇒ encore du RAG !

Puis-je être un (bon) scientifique sans veille IA ?
Mon métier est-il menacé ?

Consensus

Scopus AI
Change the way you view knowledge

OPSCIDIA
The value of knowledge

NotebookLM
Think Smarter, Not Harder
Try NotebookLM

(3) Des LLMs aux agents

LLM *un outil que je tiens*

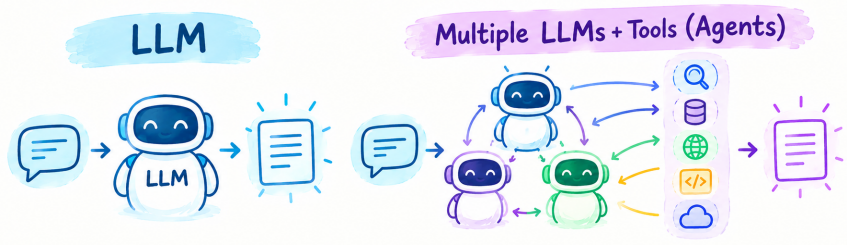
- Une question → une réponse
- **Je** lis, **je** juge, **je** garde ou je jette
- Le modèle produit du **texte**
- Aucun effet sur le monde
- Efficace sur de petites tâches

⇒ validation humaine **à chaque étape**

Agent *un système qui agit*

- Un objectif → une **chaîne** d'actions
- Appels d'**outils**, boucles, essais
- Le modèle produit des **actions**
  requêtes, fichiers, emails, code
- Effets sur le monde

⇒ validation humaine **à la fin**, au mieux



(3) Toolformer : quand le LLM fait appel à des outils

Le LLM [2023] :

1. Identifie ses propres faiblesses
2. Appelle des outils/API pour mieux répondre

⇒ Sources de données contrôlées (SQL, Wikipédia) = RAG++ ; calculatrice ; traducteur ; moteur de calcul spécialisé

Le LLM au cœur du système

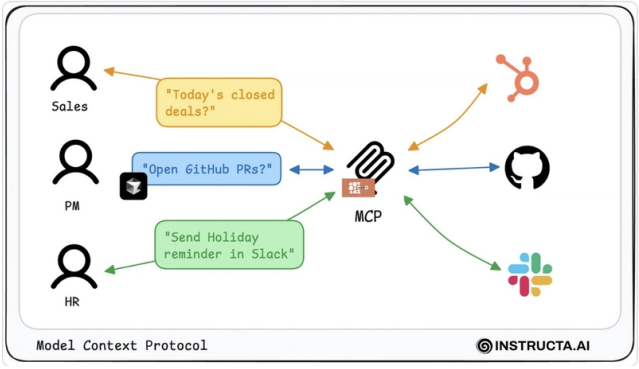
The New England Journal of Medicine is a registered trademark of **[QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society]** the MMS.

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

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

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

Le protocole MCP

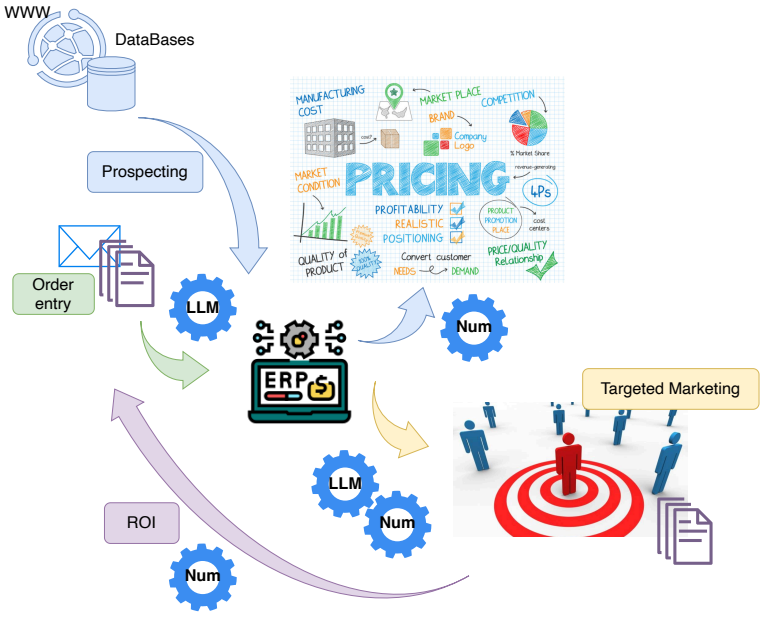


- Interfaçage avec les LLMs (distants ou locaux)
- Lien avec les applications (emails, sites web, messagerie, agenda)

⇒ Interfaçage des LLMs = le socle technique des agents... et de bien d'autres applications

MCP = utiliser une API en langage naturel
Openclaw : encapsule de nombreuses fonctions de l'ordinateur dans une API appelable par un LLM (mail, navigateur, console, ...)

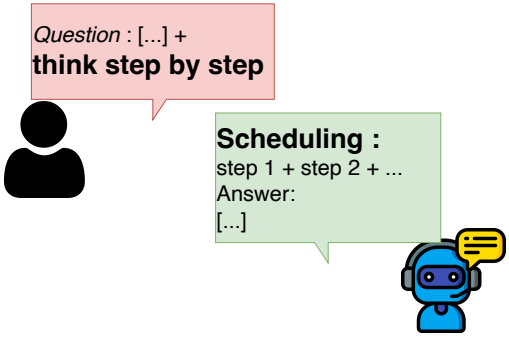
Intégration dans un système complexe

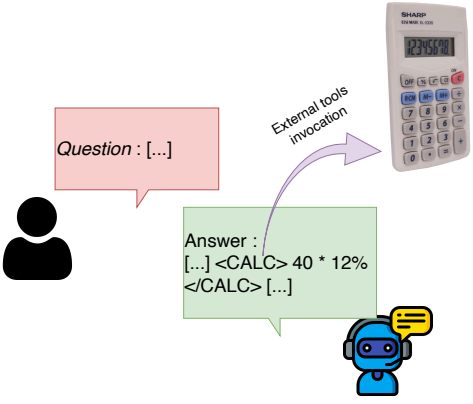


- Liste de prospects
- Traitement des emails, identification des besoins
- Interaction avec l'ERP $\Rightarrow$ bases de propositions
- Tarification
- Rédaction d'une proposition commerciale

De la *Chain-of-Thought* aux actions

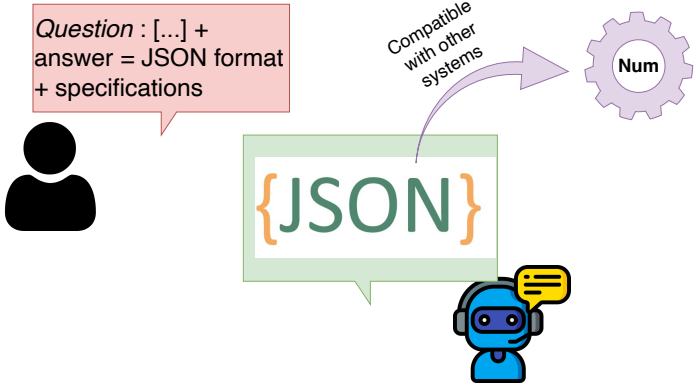
1 *Chain-of-Thought* / Toolformer





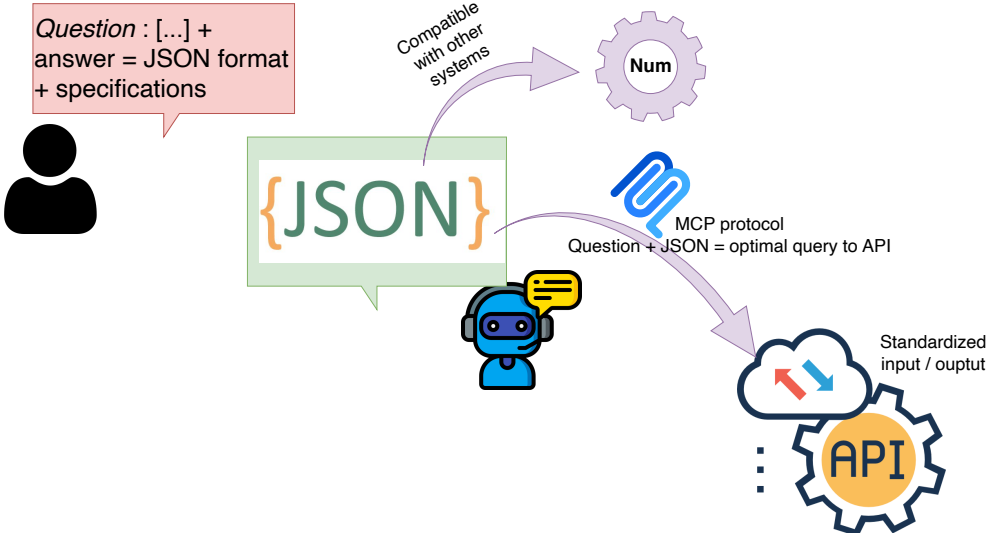
De la *Chain-of-Thought* aux actions

1. *Chain-of-Thought* / Toolformer
2. Des réponses textuelles à l'information structurée



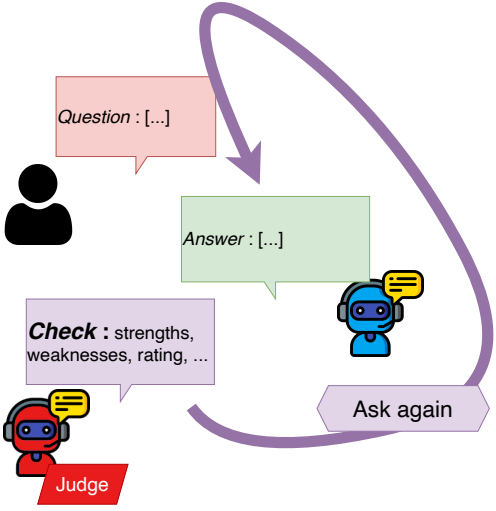
De la *Chain-of-Thought* aux actions

1. *Chain-of-Thought* / Toolformer
2. Des réponses textuelles à l'information structurée



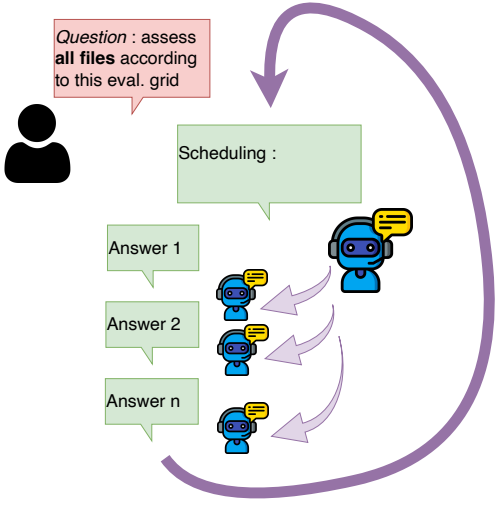
De la *Chain-of-Thought* aux actions

1. *Chain-of-Thought* / Toolformer
2. Des réponses textuelles à l'information structurée
3. LLM juge (*LLM as a judge*) / réflexivité + boucle d'amélioration



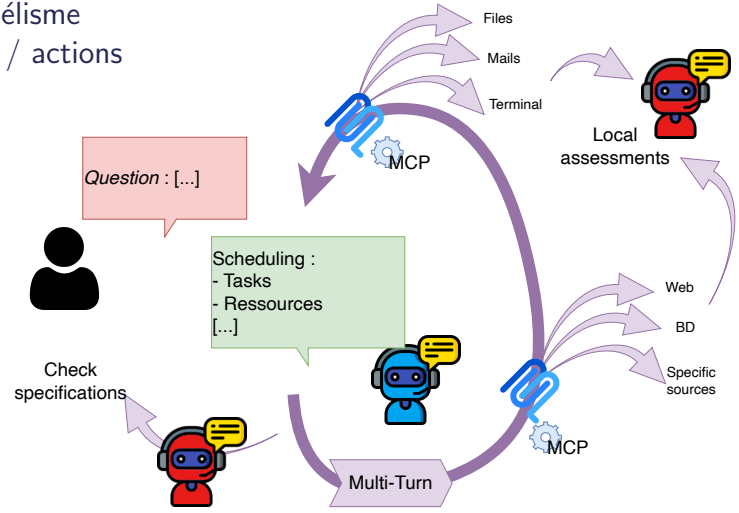
De la *Chain-of-Thought* aux actions

1. *Chain-of-Thought* / Toolformer
2. Des réponses textuelles à l'information structurée
3. LLM juge (*LLM as a judge*) / réflexivité + boucle d'amélioration
4. Parallélisme



De la *Chain-of-Thought* aux actions

1. *Chain-of-Thought* / Toolformer
2. Des réponses textuelles à l'information structurée
3. LLM juge (*LLM as a judge*) / réflexivité + boucle d'amélioration
4. Parallélisme
5. MCP / actions



De l'importance du Harnais

LLM alone

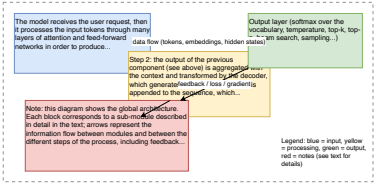
it knows the XML syntax... but never sees the result

Question : draw a figure explaining [...]

Answer :

```
<mxfile>
<mxCell value="..."/>
[...]
</mxfile>
```

one shot, no feedback



text-heavy, overlaps, unreadable

LLM + harness

render, look, fix: generation becomes a loop

Question : draw a figure explaining [...]

Harness

Answer :

```
<mxfile>[...]
</mxfile>
```

tool call

Render
draw.io → PNG

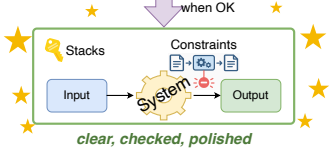
image

Look : overlaps? too much text? readable? ...

Vision

Fix

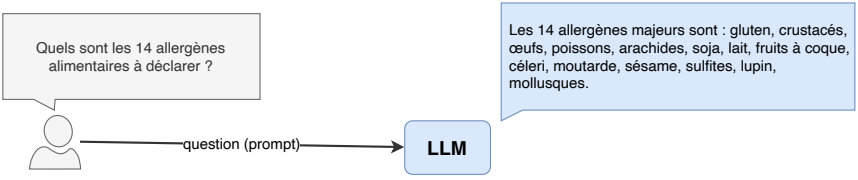
when OK



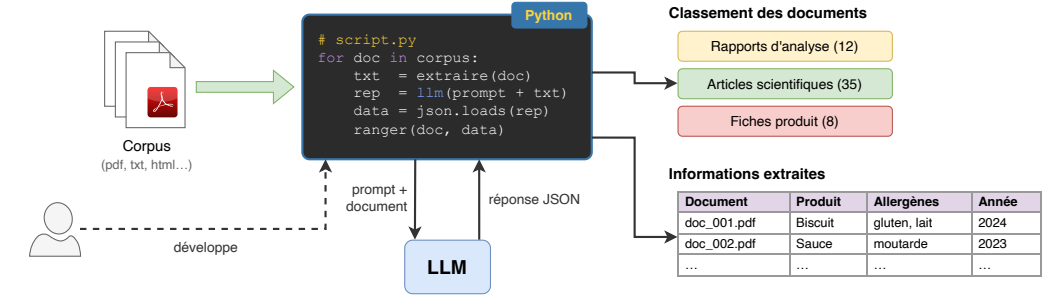
clear, checked, polished

Positionnement par rapport aux outils

1. Usage direct — interaction manuelle, une question à la fois



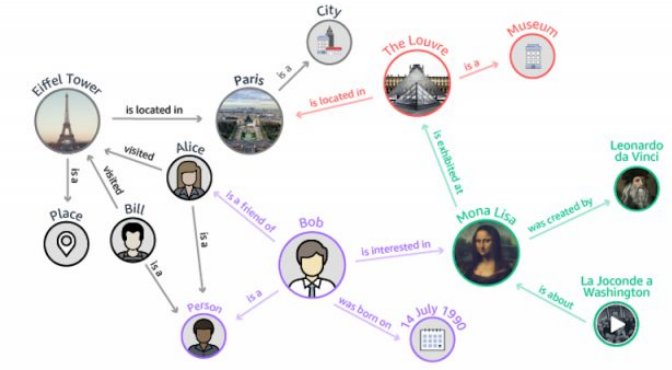
2. LLM intégré dans une chaîne de traitement — traitement automatique de N documents



Problématiques de l'extraction d'information

Knowledge Base building = Information Extraction

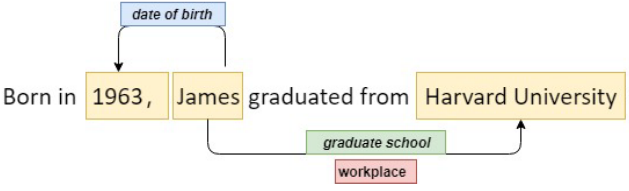
- Base exploitable par les ordinateurs
  - RDF / SparQL
  - Google Knowledge Graph



Idée:

- Construire une base exploitable automatiquement
- A partir de textes bruts

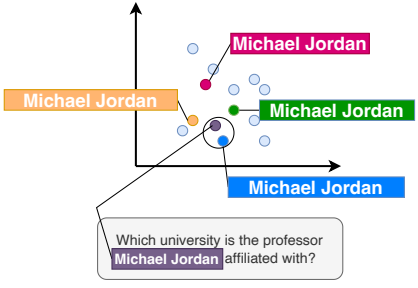
⇒ NER, Relation Extr.,
End to End Relation Extr.



Entités + relations ⇒ nécessaire mais insuffisant

Extraire des mentions ≠ construire une base de connaissances exploitable

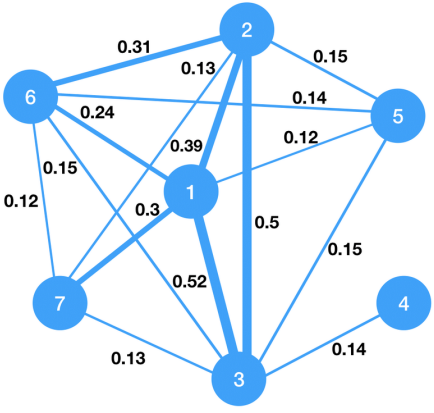
- **Normalisation** : variantes de surface → forme canonique
  PM, premier ministre, chef du gouvernement
- **Liage d'entités** (*entity linking*) : mention → entrée unique d'une KB
  MeSH, UMLS, Wikidata
  "Paris" → Paris, France (Q90) *vs* Paris, Texas
- **Lissage sur graphe** : exploiter le voisinage (thésaurus, document, KG) pour désambiguïser
- Problèmes connexes : co-références ⇒ propagation des erreurs



Entités + relations ⇒ nécessaire mais insuffisant

Extraire des mentions ≠ construire une base de connaissances exploitable

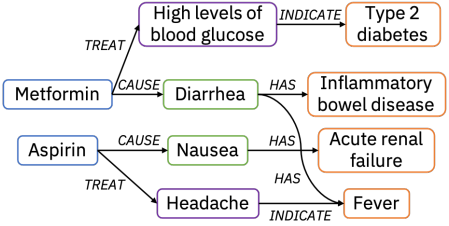
- **Normalisation** : variantes de surface → forme canonique
  PM, premier ministre, chef du gouvernement
- **Liage d'entités** (*entity linking*) : mention → entrée unique d'une KB
  MeSH, UMLS, Wikidata
  "Paris" → Paris, France (Q90) *vs* Paris, Texas
- **Lissage sur graphe** : exploiter le voisinage (thésaurus, document, KG) pour désambiguïser
- Problèmes connexes : co-références ⇒ propagation des erreurs



Entités + relations ⇒ nécessaire mais insuffisant

Extraire des mentions ≠ construire une base de connaissances exploitable

- **Normalisation** : variantes de surface → forme canonique
  PM, premier ministre, chef du gouvernement
- **Liage d'entités** (*entity linking*) : mention → entrée unique d'une KB
  MeSH, UMLS, Wikidata
  "Paris" → Paris, France (Q90) *vs* Paris, Texas
- **Lissage sur graphe** : exploiter le voisinage (thésaurus, document, KG) pour désambiguïser
- Problèmes connexes : co-références ⇒ propagation des erreurs



Une autre formulation : Question Answering

Les modèles de langue répondent aux questions... Dans un certain contexte :)

Passage Sentence

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

Question

What causes precipitation to fall?

Answer Candidate

gravity

- Passage **fourni** + question
- Réponse = **segment** du passage (début, fin) ⇒ le modèle *pointe*, il ne produit pas
- Est-ce une alternative aux KB ?

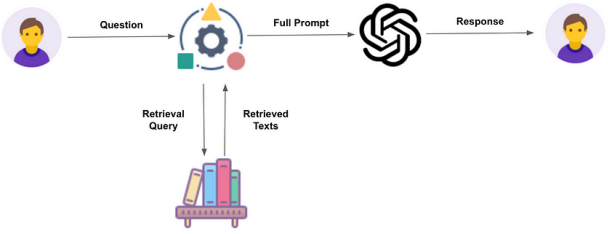


Rajpurkar et al., ACL 2018
Know What You Don't Know: Unanswerable Questions for SQuAD

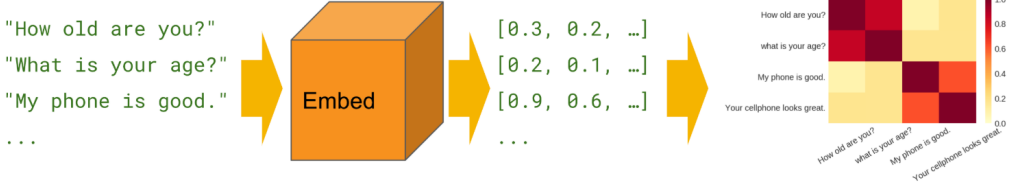
Une autre formulation : Question Answering

Du QA au RAG : le passage n'est plus fourni

RAG = **recherche** (*retriever*) + **QA génératif** (LLM)



- Indexer les **embeddings** des passages ⇒ recherche de plus proches voisins
- 1 paragraphe → 10^3–10^9 documents
- Réponse = synthèse *citée*



À vous : formaliser votre problématique

À partir de [quel corpus ?], extraire [quoi ?]
pour [quel usage ?], évalué par [quel critère ?], sur [quelles étiquettes ?]

- **Données** : type de documents, volume, langue(s), format (PDF, scans, tableaux...)
- **Cible** : entités ? relations ? attributs, valeurs numériques, événements ?
- **Référentiel** : thésaurus, ontologie, KB existante ?
  ⇒ normalisation, liage
- **Formulation** : base structurée (IE) ou réponse à des questions (QA / RAG) ?
- **Performances** : précision ou rappel ? Qui utilise le résultat ?
- **Vérité terrain** : avez-vous des exemples annotés ?
  Combien pouvez-vous en produire ?

Écrivez un exemple concret : **un texte → la sortie attendue**

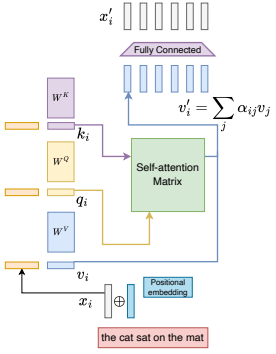
Comment traiter ces tâches ?

Deux familles de modèles

Encodeurs (BERT, 2018)

Texte → **classification**
(de la phrase, de chaque mot)

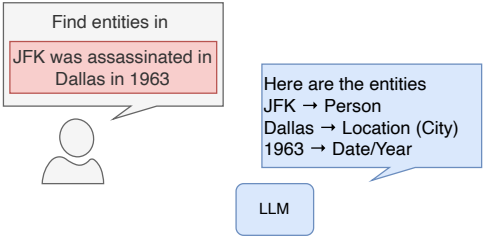
- Petits modèles, rapides
- Nécessitent des données annotées



Génératifs (GPT, LLM)

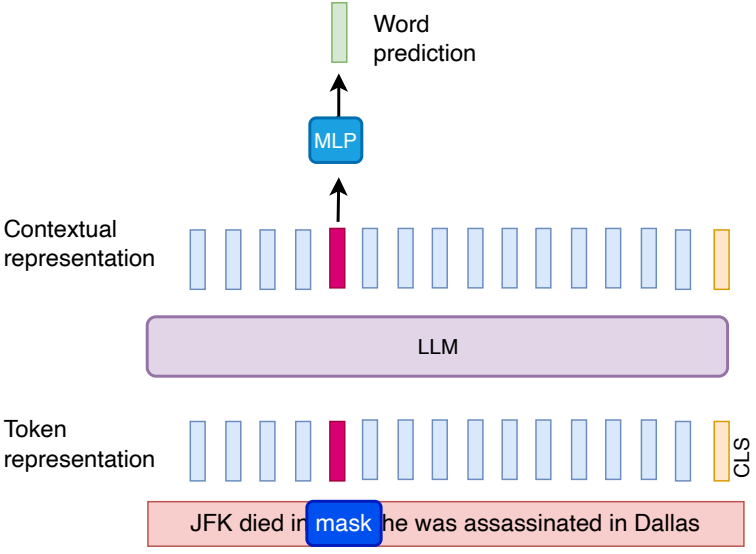
Texte → **texte**

- Gros modèles, coûteux
- Utilisables (presque) sans entraînement



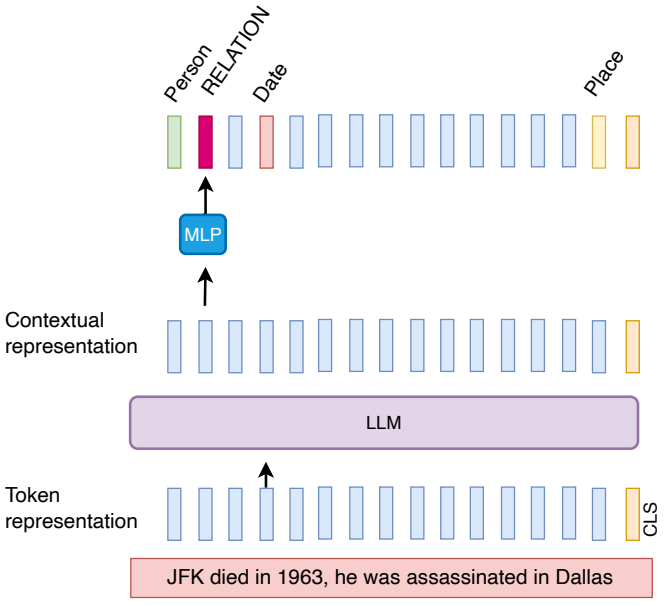
Encodeurs : pré-entraînement puis tâche cible

- Pré-entraînement : masquage de mots



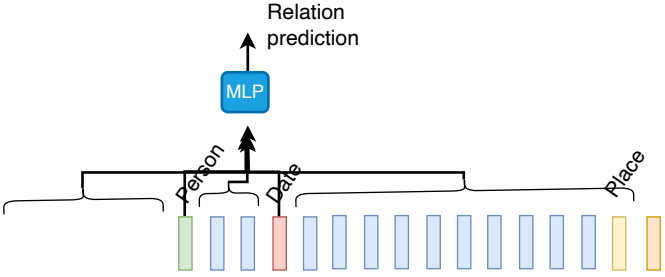
Encodeurs : pré-entraînement puis tâche cible

- Entités nommées = classification de chaque mot

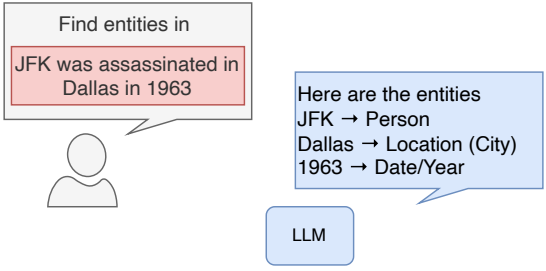


Encodeurs : pré-entraînement puis tâche cible

- Relations = classification de chaque couple d'entités
- Hyp : une relation par couple
- Représentation par morceaux (PCNN)



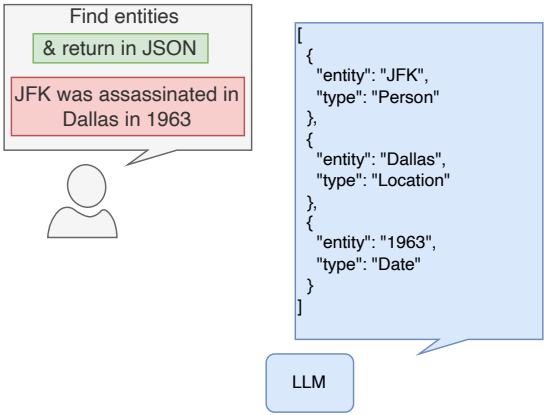
Génératifs : nouvelle formulation, nouvelles questions

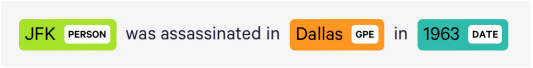


JFK PERSON was assassinated in Dallas GPE in 1963 DATE

- Évaluation : localisation des entités ? reformulations ? sur-détection ?

Génératifs : nouvelle formulation, nouvelles questions





- Évaluation : localisation des entités ? reformulations ? sur-détection ?
- Contraindre le format : JSON... **et uniquement JSON**
- ⇒ chaîne de traitement : OCR, LLM, parsing, métriques

Sans entraînement : zero-shot, prompting, few-shot

- **Zero-shot** : la tâche est-elle vraiment nouvelle ? (risque de contamination)

Sans entraînement : zero-shot, prompting, few-shot

- **Zero-shot** : la tâche est-elle vraiment nouvelle ? (risque de contamination)
- **Prompting** : la manière de poser la question impacte la performance

You are an information extraction system. Your task is to read an English news article and extract named entities of the following types:

- **PERSON**: names of individual people
- **LOCATION**: geographical entities such as cities, countries, landmarks
- **ORGANIZATION**: companies, institutions, government bodies, media outlets, NGOs, etc.
- **DATE**: explicit calendar dates or temporal expressions (e.g., “January 5, 2023”, “last Monday”)

Return the results in **JSON format** with the following structure:

json Copier le code

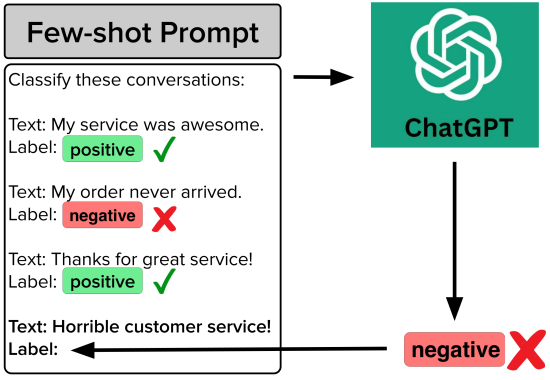
```
{
  "PERSON": ["..."],
  "LOCATION": ["..."],
  "ORGANIZATION": ["..."],
  "DATE": ["..."]
}
```

If a category has no entities, return an empty list. Do not include duplicates.

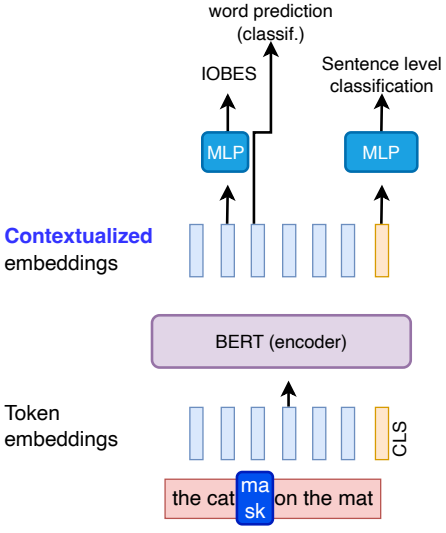
ROLE
TASK
Class DEF
Format
Supp.

Sans entraînement : zero-shot, prompting, few-shot

- **Zero-shot** : la tâche est-elle vraiment nouvelle ? (risque de contamination)
- **Prompting** : la manière de poser la question impacte la performance
- **Few-shot** : quelques exemples (entrée ⇔ sortie) dans le prompt



Fine-tuning : spécialiser un modèle

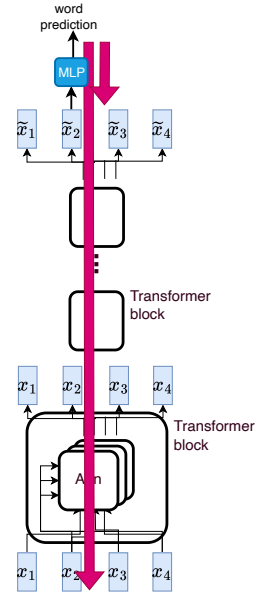


- Supervision : X textes, Y cibles
- Ajuster W t.q. $f_W(X) \approx Y$
- Encodeur : classifieur sur les mots
- Génératif : apprendre à générer la vérité terrain

Un **petit** modèle spécialisé peut-il battre un **gros** LLM ?

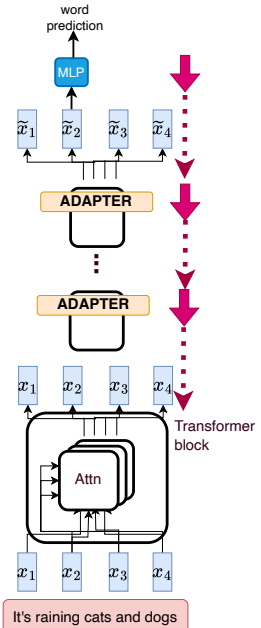
Fine-tuning efficace

1. Dernière couche seulement, peu d'époques



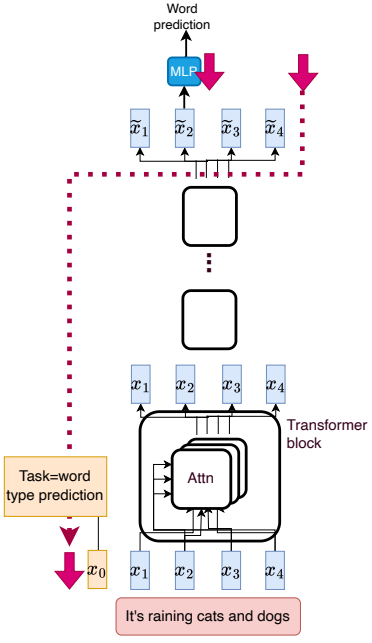
Fine-tuning efficace

1. Dernière couche seulement, peu d'époques
2. Adapter



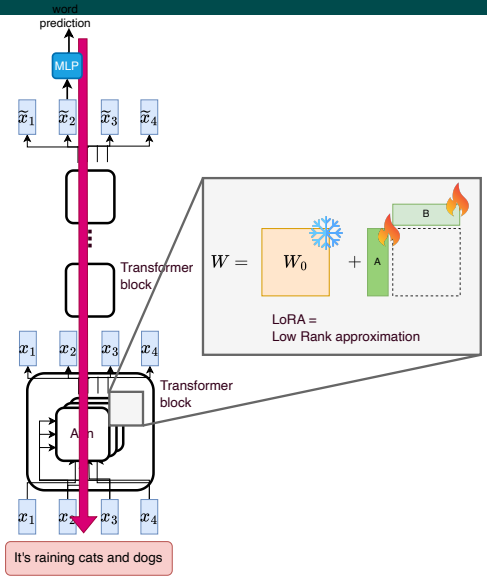
Fine-tuning efficace

1. Dernière couche seulement, peu d'époques
2. Adapter
3. Prompt Tuning



Fine-tuning efficace

1. Dernière couche seulement, peu d'époques
2. Adapter
3. Prompt Tuning
4. Low Rank Adaptation (LoRA)

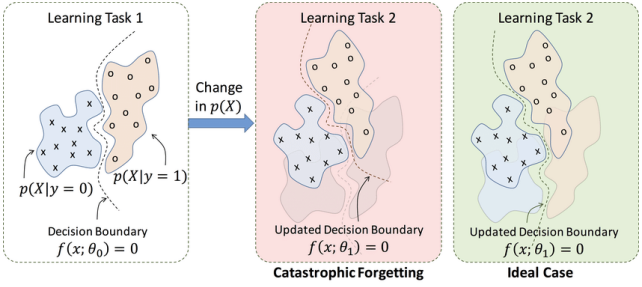


Fine-tuning efficace

1. Dernière couche seulement, peu d'époques
2. Adapter
3. Prompt Tuning
4. Low Rank Adaptation (LoRA)

Oubli catastrophique

Apprendre une nouvelle tâche... oublier les bases ?



Quelle stratégie choisir ?

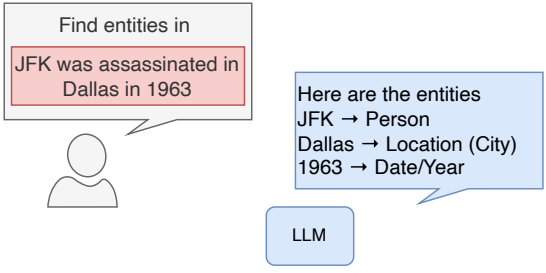
	Données annotées	Coût	Contrôle
LLM zero-shot / prompting	aucune	↗ (inférence)	faible
LLM few-shot	quelques ex.	↗ (inférence)	moyen
Encodeur fine-tuné	$\sim$ centaines+	faible (entrain.)	fort
LLM + LoRA	$\sim$ centaines+	moyen (entrain.)	fort

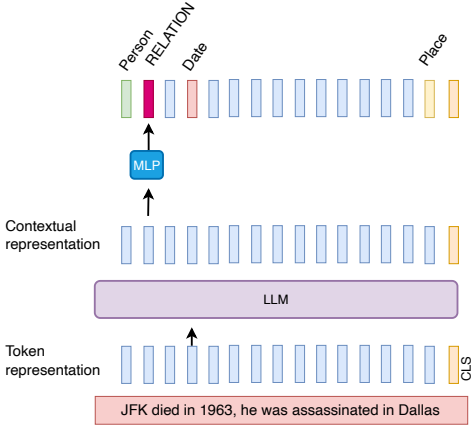
Dans tous les cas : il faut des données annotées... au moins pour **évaluer** !

Evaluer les sorties d'un LLM

De plus en plus de sorties textuelles

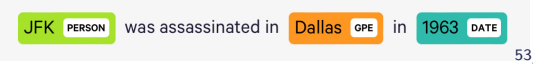
L'émergence des LLM pousse à redéfinir des tâches comme des tâches génératives





⇒ Mais comment exploiter le texte en sortie?

- Contraindre la sortie en JSON
- Ré-analyser la sortie textuelle (de nouveau avec un LLM?)



NLI: Natural Language Inference

Question: est ce que la réponse du LLM *correspond* à la vérité terrrain?

Lien avec une problématique historique en NLP: Natural Language Inference (NLI)

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

⇒ Est ce que la réponse est *impliquée* par la vérité terrain?

NLI: Natural Language Inference

Question: est ce que la réponse du LLM *correspond* à la vérité terrrain?

Dans le cas d'une transformation en informations élémentaires (*statements*), comment les valider?
[Résolution de coréférence, ≈ extraction de *triplets*... Dans le domaine textuel]

Transformer le texte suivant en une liste de statements élémentaires:

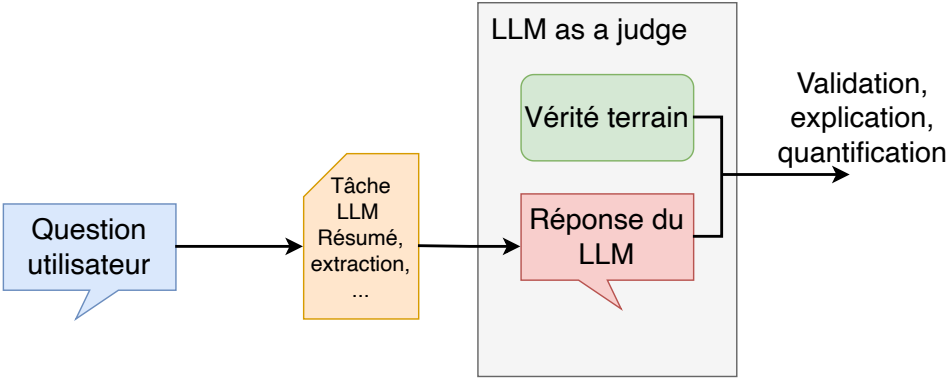
Des drones non identifiés ont été observés au-dessus de la plus grande base militaire du Danemark vendredi soir, a annoncé la police, après plusieurs survols d'aéroports cette semaine dans le pays nordique.

Voici une décomposition en statements élémentaires :

1. Des drones non identifiés ont été observés.
2. L'observation a eu lieu au-dessus de la plus grande base m
3. L'observation a eu lieu vendredi soir.
4. La police a annoncé cette observation.
5. Des survols d'aéroports ont eu lieu cette semaine.
6. Les survols d'aéroports ont eu lieu dans le Danemark.
7. Le Danemark est un pays nordique.

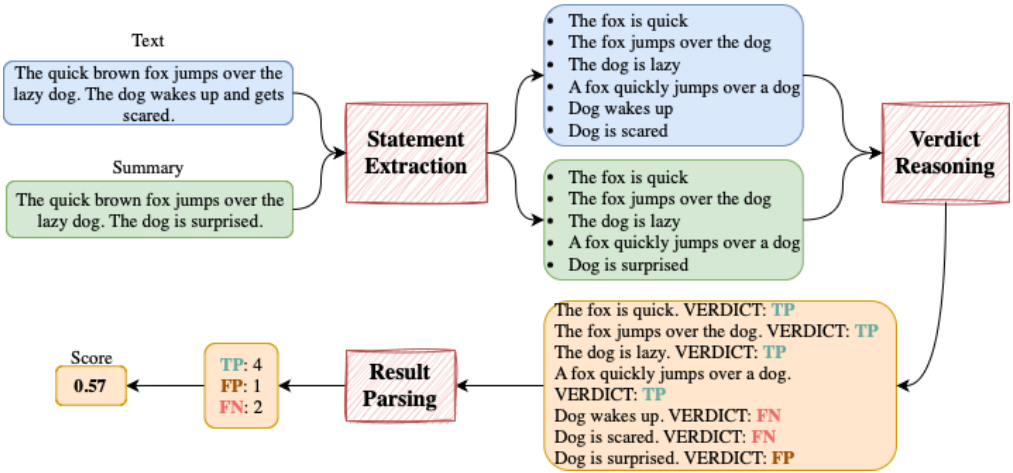
⇒ Utilisation de modèles déjà entrainés sur la tâche (RoBerta, ...)

LLM as a judge: la reflexivité des LLM



- La performance du LLM n'est pas symmétrique... Surtout si on ne lui demande pas la même chose (connaissances paramétriques *vs* analyse comparée de texte pour la validation)
- Quid du coût de la sratégie?

Évaluation de résumés automatique



⇒ À la fin, nous ne transformons pas le texte original

⇒ Nous effectuons l'extraction des informations atomiques sur le résumé

SEval-Ex: A Statement-Level Framework for Explainable Summarization Evaluation Herserant et al.; PAKDD 2025

Évaluation de résumés automatique

Architecture	Metric	Fluency	Consistency	Coherence	Relevance	Average
GPT4	**G-Eval (Best)**	**0.455**	0.507	**0.582**	**0.547**	0.523
GPT3	**GPTScore**	0.403	0.449	0.434	0.381	0.417
n-gram	**ROUGE-1**	0.115	0.160	0.167	0.326	0.192
	ROUGE-2	0.159	0.187	0.184	0.290	0.205
	ROUGE-L	0.105	0.115	0.128	0.311	0.165
Embedding based	**BERTScore**	0.193	0.110	0.284	0.312	0.225
	MOVERScore	0.129	0.157	0.159	0.318	0.191
	BARTScore	0.356	0.382	0.448	0.356	0.385
T5	**QuestEval**	0.228	0.306	0.182	0.268	0,246
	UniEval	0.449	0.446	0.575	0.426	0.474
qwen2.5:72b	**SEval-Ex**	0.351	**0.580**	0.264	0.300	0.373

SEval-Ex: A Statement-Level Framework for Explainable Summarization Evaluation Herserant et al.; PAKDD 2025

Conclusion

Generative AI : new opportunities at a cost

- **LLM cost** >> encoders (>> CRF / trad. ML >> classical lexical rules)
  - Paying for extraction *or* for development?
- New formulations ⇒ new metrics ⇒ **how to evaluate properly?**
- Above which threshold the systems are **useful**? Autonomously useful?
- What about a system that would be able to extract *statements*?

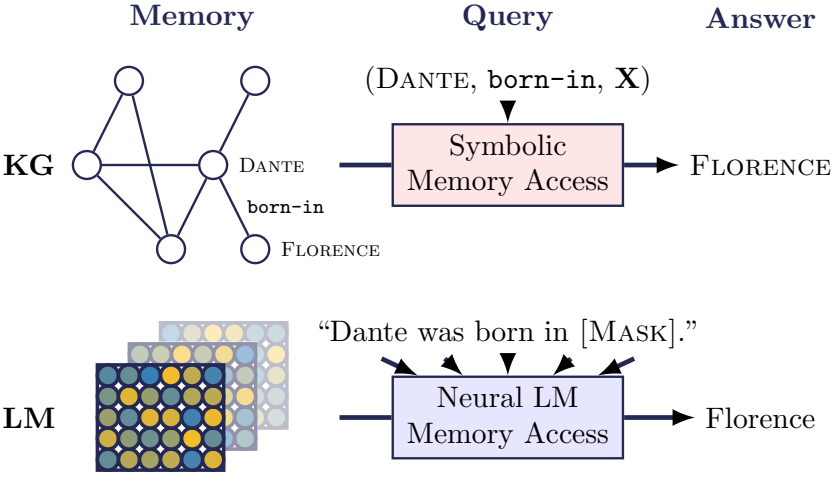
Transformer le texte suivant en une liste de statements élémentaires:

Des drones non identifiés ont été observés au-dessus de la plus grande base militaire du Danemark vendredi soir, a annoncé la police, après plusieurs survols d'aéroports cette semaine dans le pays nordique.

Voici une décomposition en statements élémentaires :

1. Des drones non identifiés ont été observés.
2. L'observation a eu lieu au-dessus de la plus grande base militaire du Danemark.
3. L'observation a eu lieu vendredi soir.
4. La police a annoncé cette observation.
5. Des survols d'aéroports ont eu lieu cette semaine.
6. Les survols d'aéroports ont eu lieu dans le Danemark.
7. Le Danemark est un pays nordique.

Sous quelle forme stocker les connaissances?

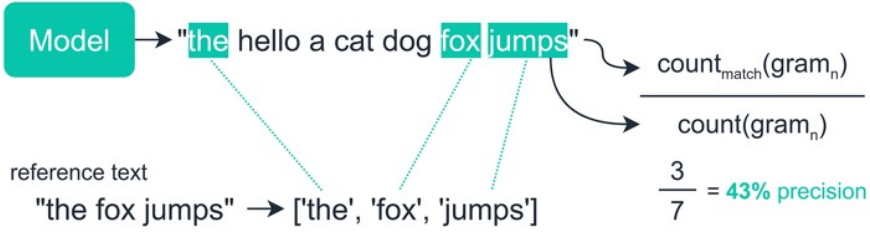


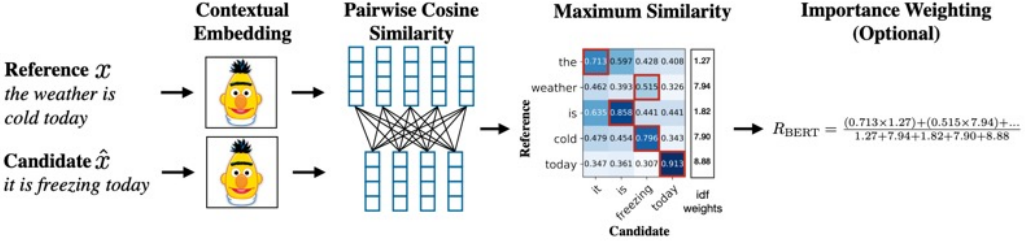
- Exhaustivité? Rapidité?
- Fiabilité? Raisonnement?

⇒ Toujours réfléchir par rapport au cahier des charges des utilisateurs

Language Models as Knowledge Bases?, EMNLP 2019, Petroni et al.

Comment évaluer les modèles de langue?





Comment évaluer la qualité d'un texte ou d'une image générée ?

Conclusion et perspective

- LLM + Instruction = le début d'un mouvement
  - Objet de recherche dépassé par les usages
- Des technologies **chères** (mais un coût en baisse)
  - Ressources disponibles: Jean Zay
- Des limites critiques:
  - Evaluation
  - Controle / garantie sur la génération