

CADRES D'APPRENTISSAGE POUR L'EXTRACTION D'INFORMATION

Vincent Guigue & Rémy Découpes
vincent.guigue@agroparistech.fr



Plusieurs focus thématiques

- 1 La normalisation des entités nommées
- 2 La construction de corpus étiquétés et les stratégies de limitation des coûts
- 3 Deep learning & graphes de connaissances: qu'est ce qui a changé dans la représentation de ces données?
- 4 Evaluation des IA générative en texte
- 5 Perspectives dans l'accès à l'informations et aux connaissances

NORMALISATION



Normalisation: entités + relations \Rightarrow insuffisant

Normalisation = Entity Linking

Mention extraite \Leftrightarrow entrée unique dans une base de connaissances (KB)

Le **premier ministre** termine sa déclaration à **Paris**

- "Paris" \rightarrow **Paris**, France (Wikidata: Q90), plutôt que **Paris**, Texas.
- "premier ministre" \rightarrow ...

Dictionnaire, ontologie, thésaurus

MeSH, UMLS, Wikidata

[tâche critique en médecine, biologie, ...]

++ ambiguïté

Similarité

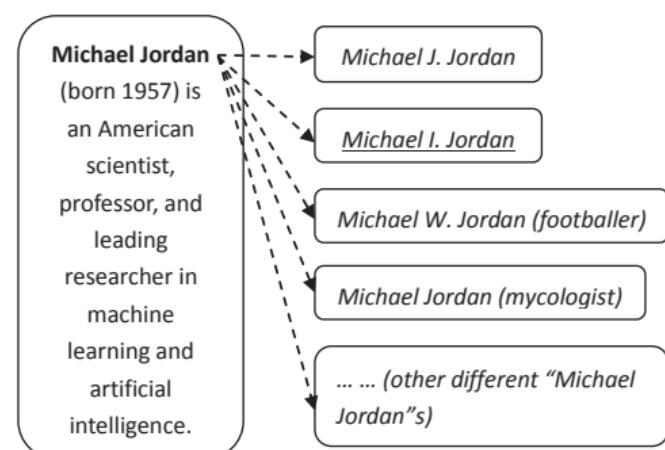
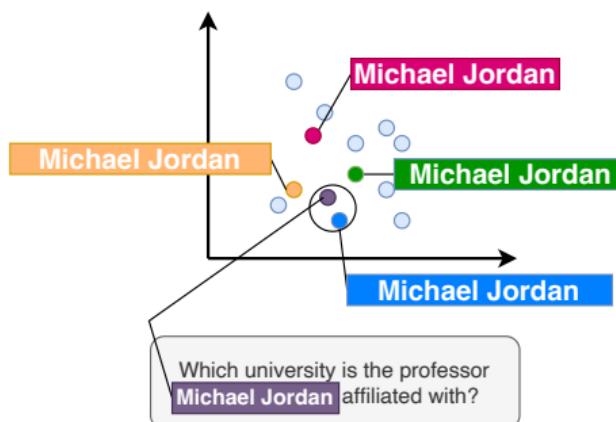
- Lexicale (souvent suffisant)
- Contexte (TF-IDF, BM25)
- Moteur de recherche (autorité)
- Embeddings

\Rightarrow Représenter l'entité + la référence \Rightarrow + Distance minimale = match

Normalisation: entités + relations \Rightarrow insuffisant

Normalisation = Entity Linking

Mention extraite \Leftrightarrow entrée unique dans une base de connaissances (KB)



Répondre à la question:

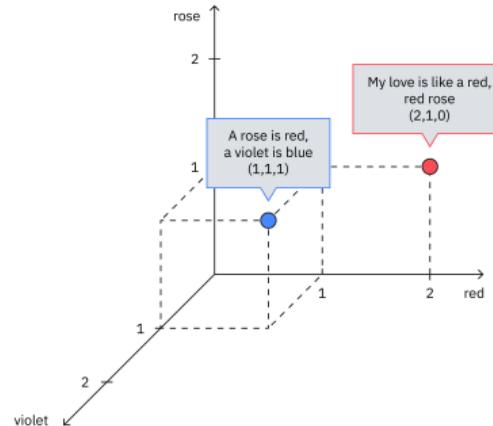
Which university is the professor Michael Jordan affiliated with?

\Rightarrow Trouver le bon *Michael Jordan* \Rightarrow désambiguisation de la question

Similarité tf-idf

- Entité : vecteur mono-valeur ou phrase
- Référence : mot, mot + synonymes, mot + description/définition
(e.g. définition d'un ontologie, page wikipedia d'un terme)

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions, [IEEE Trans. KDE 2025](#), Shen et al.



tf-idf

$$TF(t, d) = \frac{\text{(Number of occurrences of term } t \text{ in document } d)}{\text{(Total number of terms in the document } d)}$$

$$IDF(t, D) = \log_e \frac{\text{(Total number of documents in the corpus)}}{\text{(Number of documents with term } t \text{ in them)}}$$

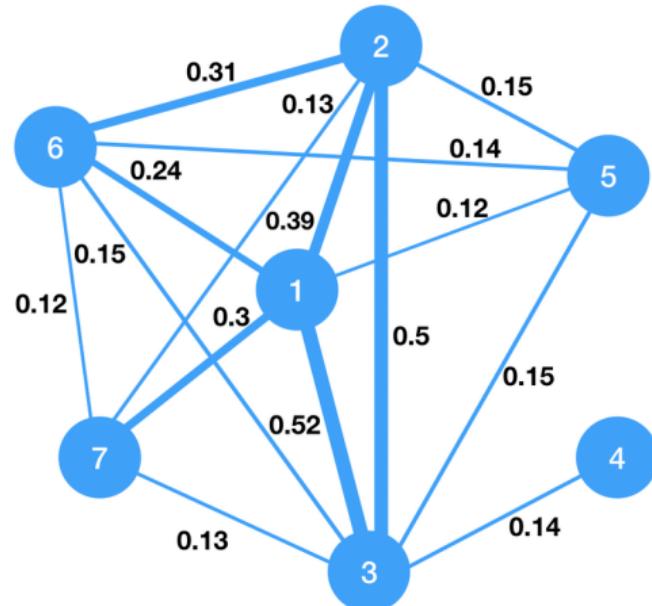
$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Lissage sur graphe

Le score tf-idf permet de construire un graphe :

- Entre les termes dans le **Dictionnaire**
 - Entre les phrases dans le **document**
- ⇒ Possibilité de lisser les représentations

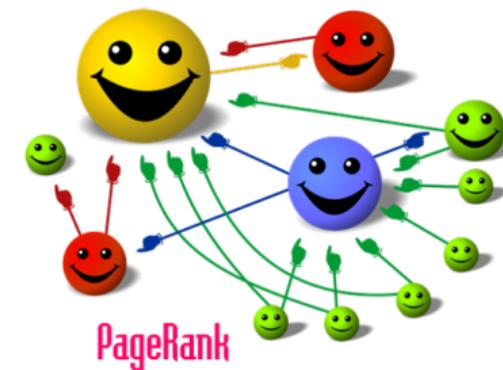
Est ce qu'on ajoute des informations ou du bruit en prenant en compte le contexte large?



Information Retrieval

- Corpus = Dictionnaire / thésaurus
- Requête = entité

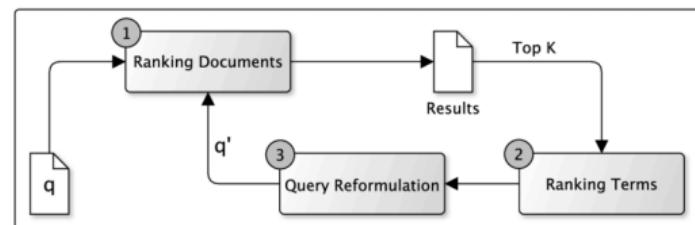
Recherche lexicale + Métrique BM25 (proche de tf-idf)



Il est possible d'ajouter un score d'autorité:

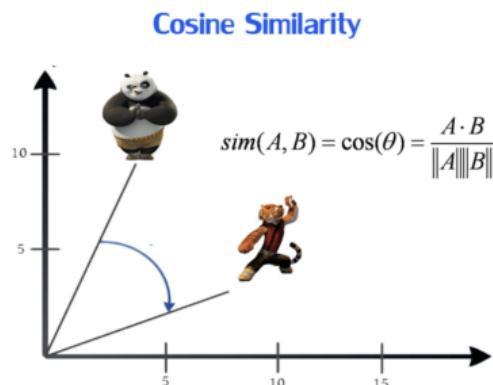
- Fréquence d'apparition des entités (e.g. Paris, France vs Paris, Texas)
- PageRank

Historiquement: stratégie d'élargissement des requêtes type *pseudo-relevance-feedback*



Embedding / plongement

- Similarité cosinus
- Avec ou sans contextualisation
 - Mot (dans un contexte)
- Agrégation type CLS
 - Phrase d'origine
 - Définition du dictionnaire



Contextual representation



LLM

Token representation



Aspirin : medication that relieves pain, reduces fever and inflammation

Graph Neural Network & normalisation

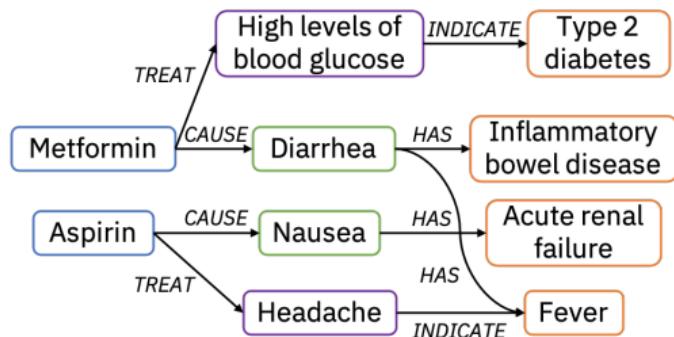
GNN = représentation de nœuds, repr. d'arc, repr. de graphe entier

- Message passing, local/global pooling

Application sur un graphe de connaissances

Hyp: la représentation de chaque noeud sera meilleure en prenant en compte le contexte du noeud

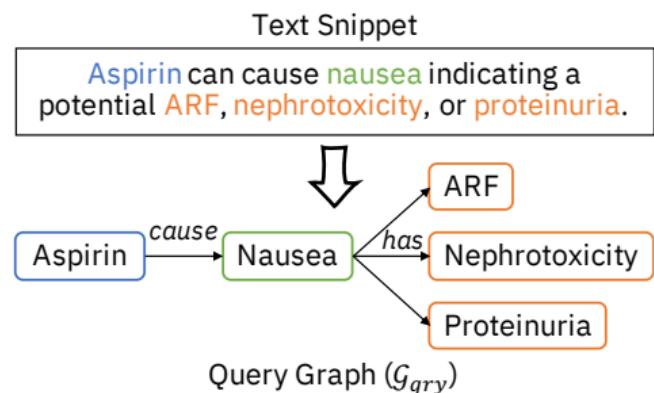
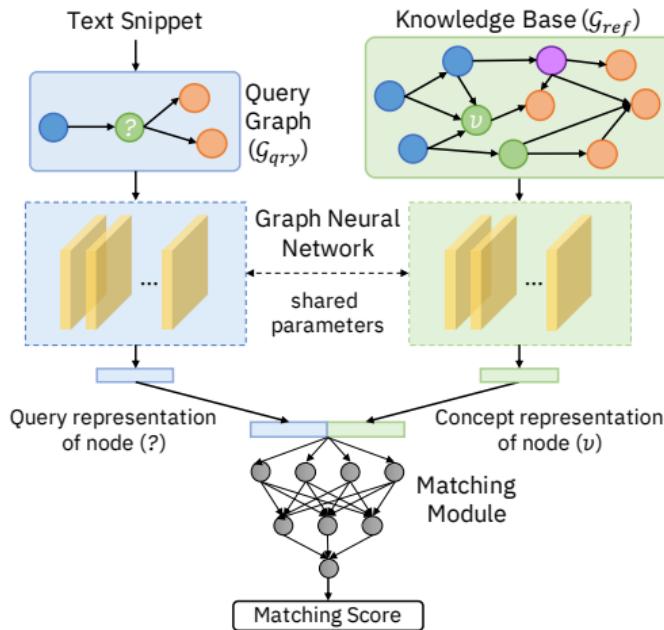
Medical Entity Disambiguation Using Graph Neural Networks, [ACM SIGMOD 2021, Vretinaris et al.](#)



Graph Neural Network & normalisation

GNN = représentation de nœuds, repr. d'arc, repr. de graphe entier

- Message passing, local/global pooling



⇒ Quel apport / coût du matching module ?

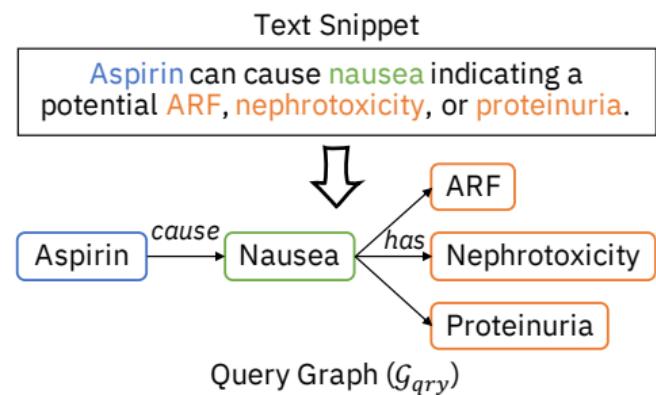
Graph Neural Network & normalisation

GNN = représentation de nœuds, repr. d'arc, repr. de graphe entier

- Message passing, local/global pooling

Le LLM est aussi une opportunité pour enrichir les descriptions des termes dans les thésaurus ou dictionnaires... Pour ensuite améliorer la normalisation.

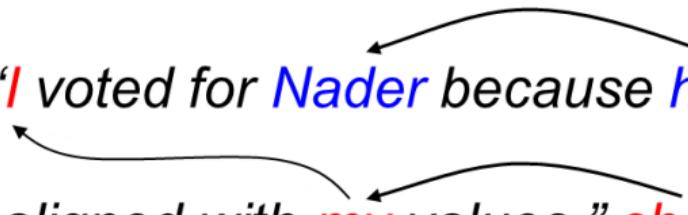
Projet en cours de Stéphane Dervaux et Magalie Weber



Résolution de co-références

Des problèmes connexes impactent directement la détection et la normalisation des entités

"I voted for Nader because he was most aligned with my values," she said.



Approches historiques:

- Systèmes de règles
- HMM / CRF (Markov / Conditional Random Field) pour le machine learning sur des séquences
- LLM : modèles génératifs très forts sur la tâches (considérée comme résolue par le *Stanford NLP Group*)

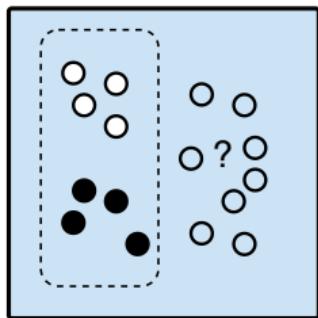
⇒ Risque de propagation des erreurs

ETIQUETAGE DES DONNÉES

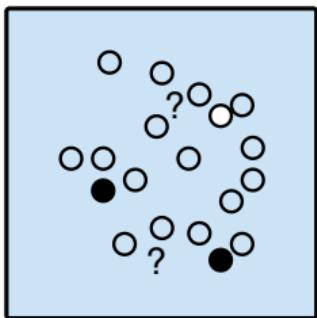


Apprentissage supervisé

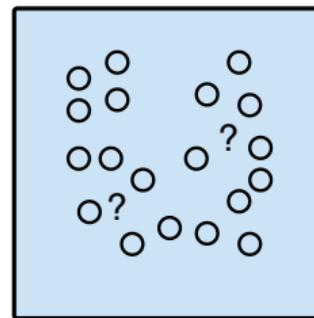
Cadres historiques d'apprentissage



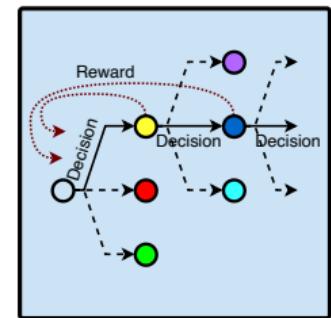
Supervised
Learning



Semi-Supervised
Learning

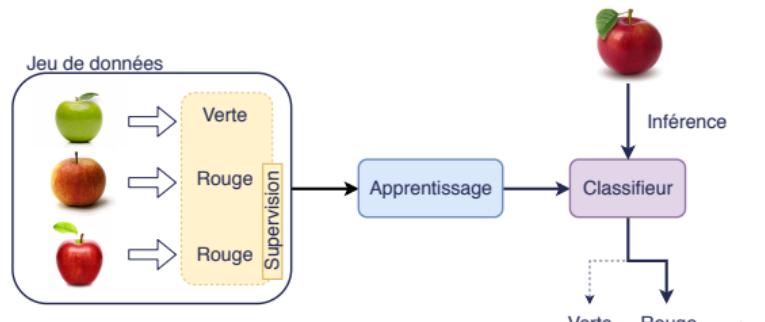


Unsupervised
Learning



Reinforcement
Learning

Deep learning: supervisé... Ou
auto-supervisé
⇒ Quid des étiquettes?





Historique rapide en vision

VOC 2007, caltech256

- milliers d'images, dizaines de catégories

ImageNet 2009:

- 3.5M d'images annotées, 1000 catégories (moins de 0.3% d'erreur)
- Construction sur plusieurs années : objectif 50M images
- Localisation pour 1.2M d'objets (2012)

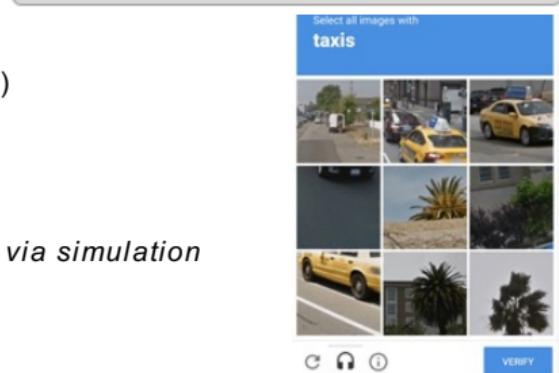
Google car

- 2018: *5 million miles on public roads and more than 5 billion miles via simulation*

Mobileye

- 500 personnes au Sri-Lanka: étiquetage scènes/objets
- Réduire les modèles pour les embarquer dans une caméra
- (Google) Captcha

L'augmentation de la puissance de calcul et des tailles de bases de données expliquent les succès actuels





Les sources déjà disponibles

- Page de désambiguisation de wikipedia + lien vers les sources

Michael Jordan (homonymie)

[Article](#) [Discussion](#)

[Lire](#) [Modifier](#) [Modifier le code](#) [Voir l'historique](#) [Outils](#) [▼](#)

文 A 8 langues [▼](#)

- *Cette page d'homonymie répertorie différentes personnes portant le même nom et le même prénom.*
- *Pour les articles homonymes, voir [Jordan](#).*

Michael Jordan est un nom de personne notamment porté par :

- [Michael Jordan \(1963-\)](#), un joueur américain de basket-ball ;
- [Michael Jordan \(1986-\)](#), un joueur anglais de football ;
- [Michael Jordan](#), un homme politique irlandais.

Mike est un nom de personne notamment porté par :

- [Mike Jordan \(1958-\)](#), un pilote automobile anglais.

Sur les autres projets Wikimedia :

• [Michael Jordan](#), sur Wikiquote

Voir aussi [\[modifier \]](#) [\[modifier le code \]](#)

- [Michael B. Jordan \(1987-\)](#), un acteur américain
- [Michael I. Jordan \(1956-\)](#), un chercheur américain
- [Michael-Hakim Jordan \(1977-\)](#), un joueur américain de basket-ball

Hypothèse: sur chacune des pages cibles, les entités *Michael Jordan* correspondent à la page en question

Exemple de corpus disponible basé sur ces pages:

Interactive Query Clarification and Refinement via User Simulation, [SIGIR 2022, Erbacher et al.](#)

Les sources déjà disponibles

■ Classification d'opinion: les sites d'avis en ligne



Bazza! Aussie IPA

American IPA | 6% ABV

Lucky Envelope Brewing in Seattle, Washington

Reviewed by mactrall from Washington

3.7/5 rDev +2.8% | Average: 3.6

look: 3.75 | smell: 3.75 | taste: 3.75 | feel: 3.75 | overall: 3.5

Slight haze on the pale amber brew and plenty of foam in the Stella Artois goblet. Pleasant aroma of berries and green apricot. Taste is like a fresh hop brew with that resinous and sharp flavor. There is some fruity flavor, like blueberry, but also a green, unripe taste. Bitterness is moderate but the finish is on the mineral side. Overall this is an interesting hoppy brew with some rewards as well as challenges. From the 16 oz can purchased at Elizabeth Station. Dated 8/19/25.



Du texte, des notes, parfois même des notes sur des aspects spécifiques
⇒ une source intéressante pour l'analyse d'opinion et de thématiques

■ Les news / forum déjà classés par thématiques

■ ...



Pattern/motifs: l'expertise pour amorcer le système

Expert \Rightarrow motif lexical/grammatical + variantes \Rightarrow étiquetage

- L'expert formule des règles pour identifier ce qui l'intéresse

ID	Pattern Synset & Support Sets				
P_1	$\langle \text{Politician} \rangle$ was governor of $\langle \text{State} \rangle$ A,80 B,75 C,70				
P_2	$\langle \text{Politician} \rangle$ politician from $\langle \text{State} \rangle$ A,80 B,75 C,70 D,66 E,64				
P_3	$\langle \text{Person} \rangle$ daughter of $\langle \text{Person} \rangle$ F,78 G,75 H,66				
P_4	$\langle \text{Person} \rangle$ child of $\langle \text{Person} \rangle$ I,88 J,87 F,78 G,75 K,64				

Table 1: Pattern Synsets and their Support Sets

- Permet un premier étiquetage
- Utilisation de patrons (regex, dépendances syntaxiques) combinés avec la KB pour annoter automatiquement.

PATTY: A Taxonomy of Relational Patterns with Semantic Types, [EMNLP 2012](#), Nakashole et al.



Supervision distante: exploitation des bases de connaissances

- Entités :
Base de connaissances \Rightarrow termes \Rightarrow Recherche & annotation dans le texte
- Relation : triplet $(e_h, R, e_t) \Rightarrow$ si (e_h, e_t) dans un paragraphe alors R

Risque de bruit : si la KB indique que *Barack Obama* est né à *Hawaï*, tout texte contenant ces deux entités est étiqueté comme instance de la relation *BornIn*.

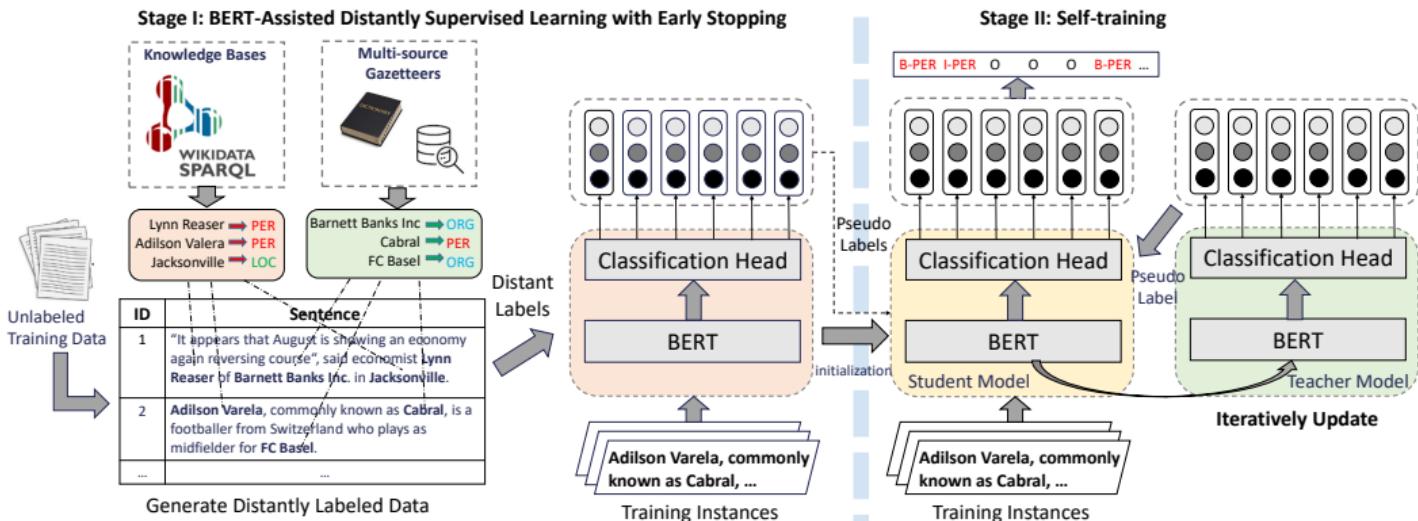
Risque de silence : toutes les variations d'écriture d'une entité sont oubliées (+ problème des co-références)

\Rightarrow Il faut des approches/détecteurs spécifiques pour traiter des corpus étiquetés de la sorte.



Bond : une méthode populaire pour la supervision distante

- Coopération d'un modèle *Teacher* et d'un second *Student* pour intégrer les conjectures du premier

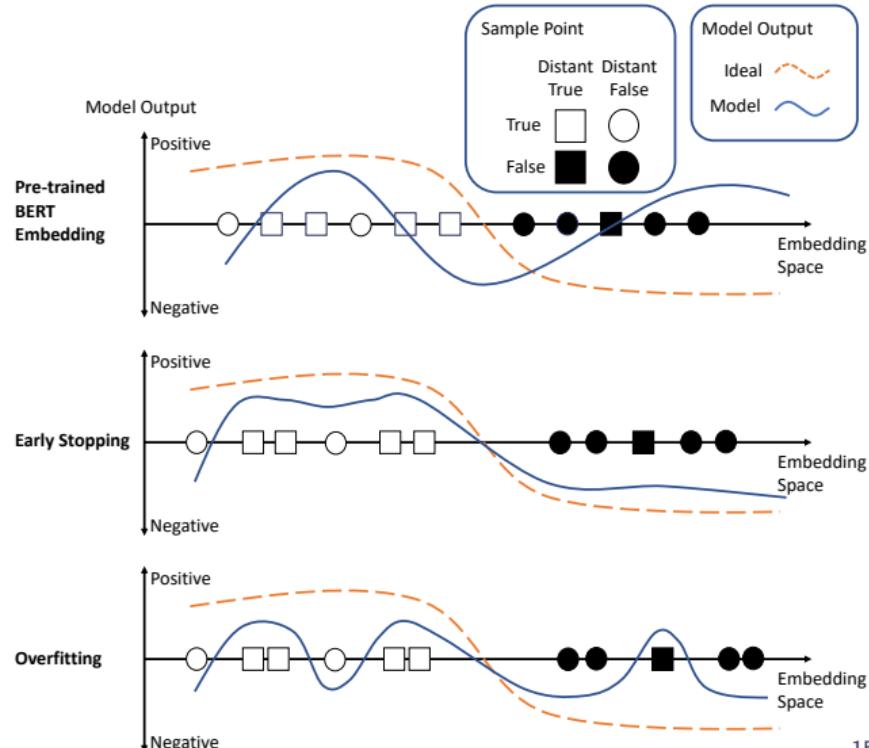




Bond : une méthode populaire pour la supervision distante

- Coopération d'un modèle *Teacher* et d'un second *Student* pour intégrer les conjectures du premier

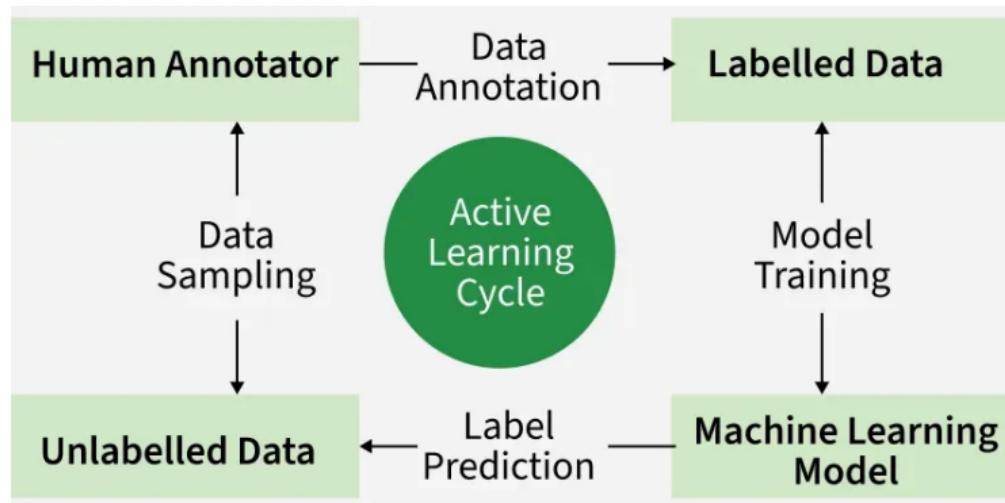
Tenter d'infléchir la courbe de décision sans tomber dans le sur-apprentissage





Active learning

Garder l'humain dans la boucle pour améliorer le système en toute confiance



Quels exemples présenter à l'utilisateur?

- Les plus ambigus
- Les plus représentatif d'une classe
- Ceux maximisant un critère statistique



Crowdsourcing

- L'IA moderne (post 2012) repose sur de vastes corpus annotés par des humains
 - ChatGPT et consort (post 2022) ont accéléré le mouvement!
- ⇒ On parle des *travailleurs du clic*

Amazon mechanical turk

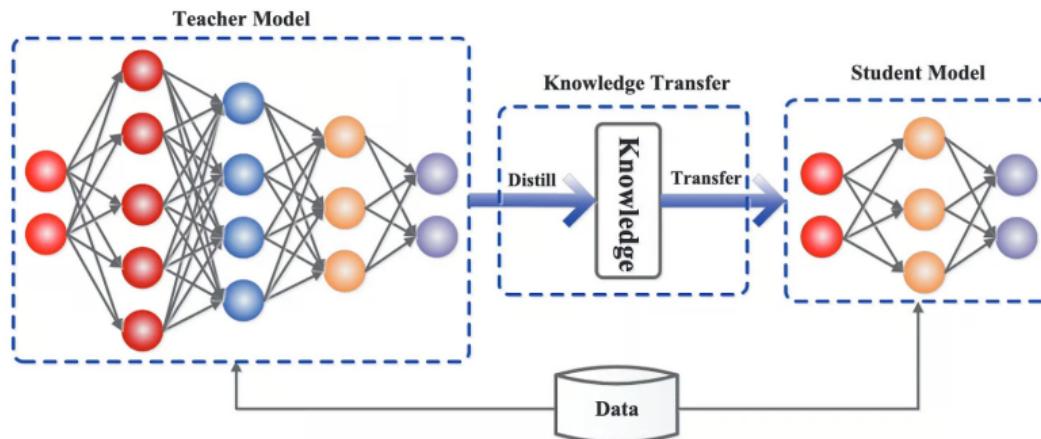


Possibilité d'annoter des corpus larges sur des tâches complexes avec de la redondance entre les annotateurs



Distillation LLM

Hypothèse: les très gros LLM ont acquis une véritable expertise dans certains domaines... Ils seraient en mesure d'étiqueter des corpus



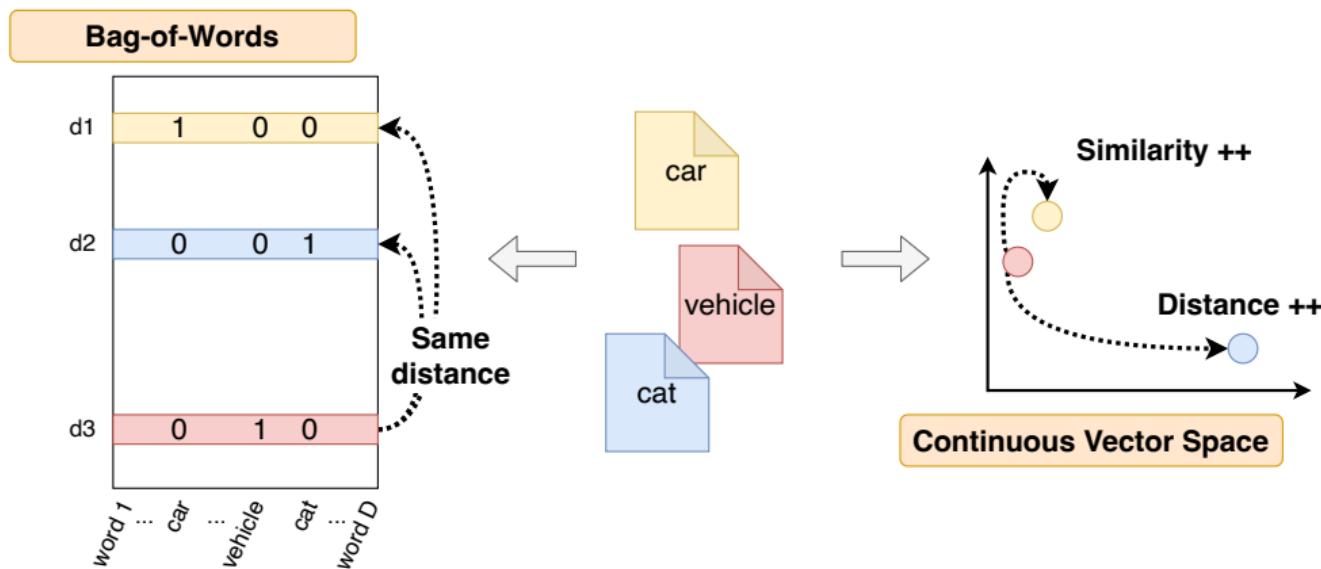
- Très peu coûteux
- Risque de propagation des erreurs, excès de confiance dans les LLM?

REPRÉSENTATION DES CONNAISSANCES DANS UN ESPACE LATENT



Repésentation latente de texte / base de connaissances

2012 Grand changement dans le texte avec Word2Vec

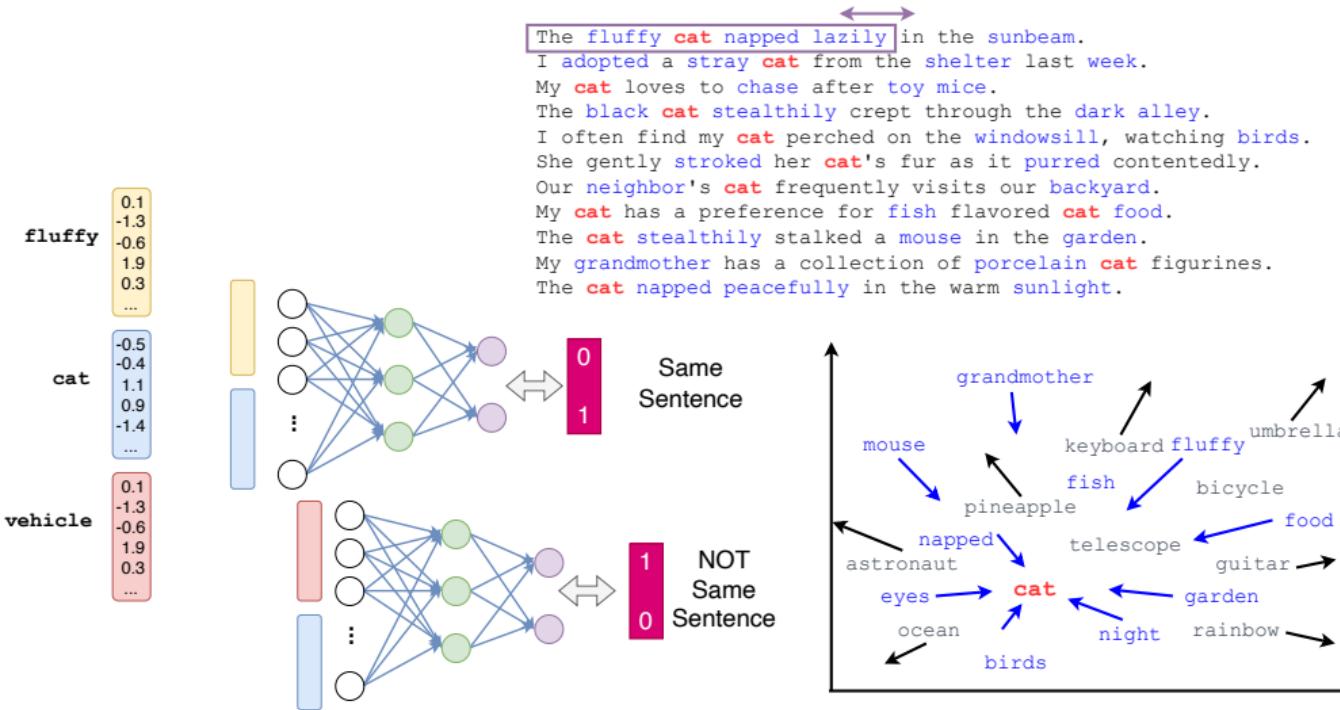


Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013



Repésentation latente de texte / base de connaissances

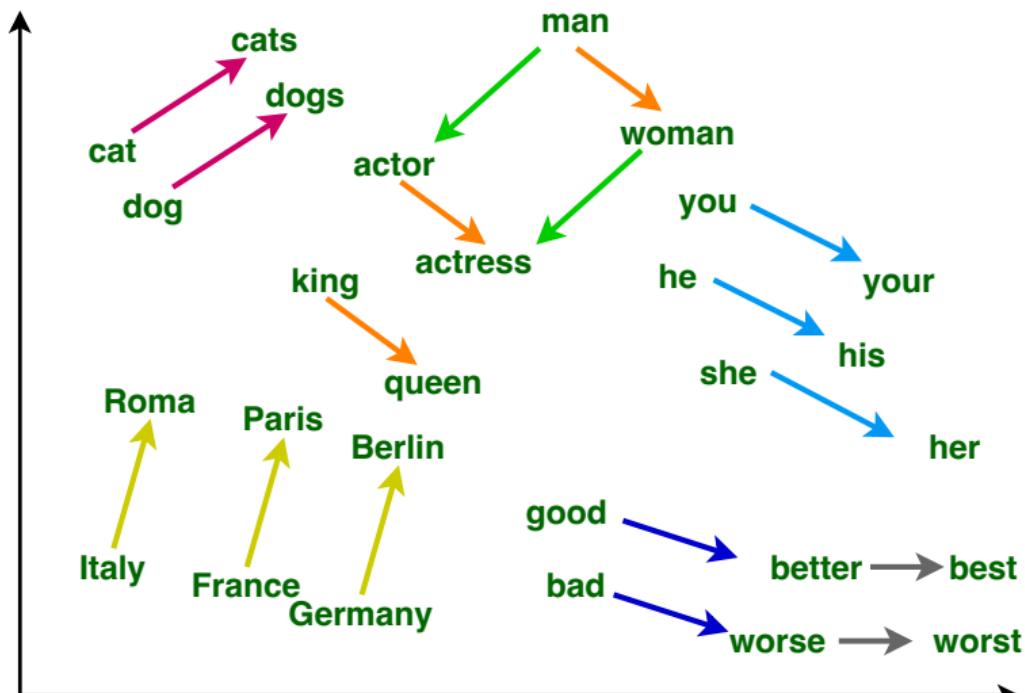
2012 Grand changement dans le texte avec Word2Vec





Repésentation latente de texte / base de connaissances

2012 Grand changement dans le texte avec Word2Vec





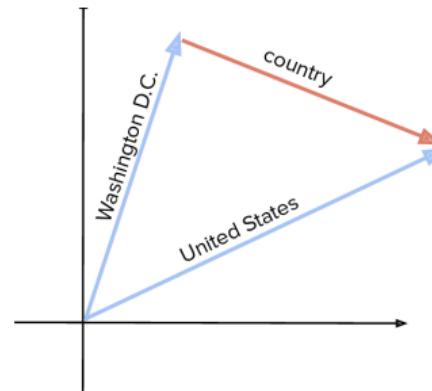
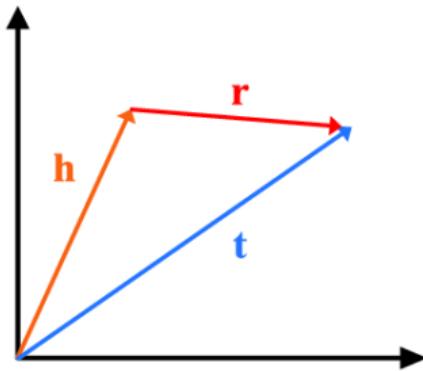
Repésentation latente de texte / base de connaissances

2012 Grand changement dans le texte avec Word2Vec

2013 Grand changement dans les bases de connaissances avec TransE

Représenter des triplets sous forme de vecteurs:

- Entités : vecteurs
- Relation : translation

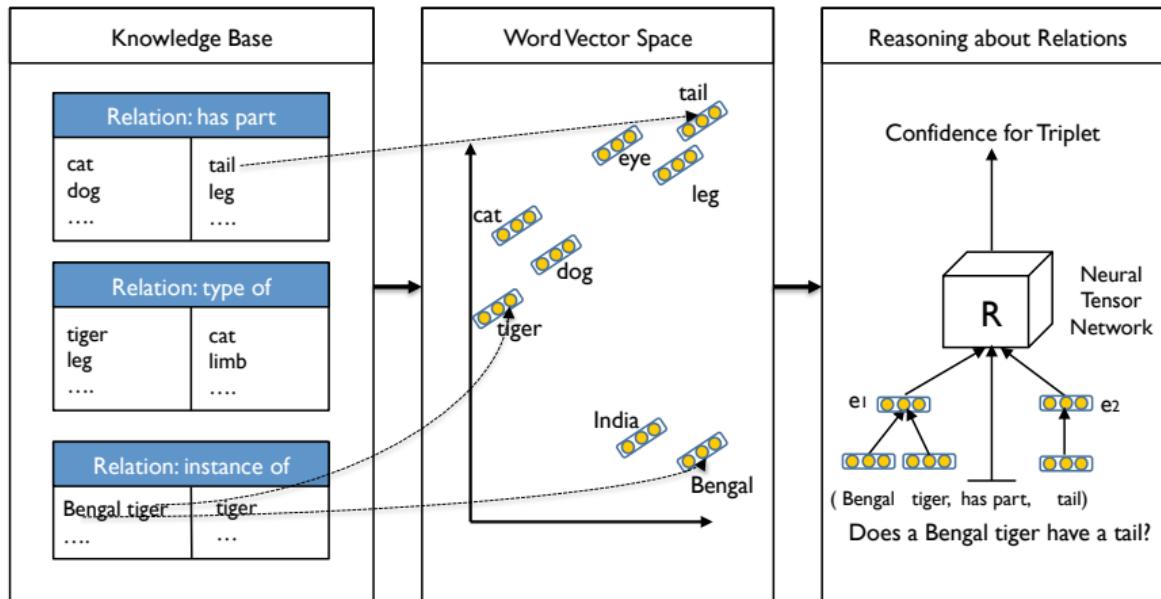


- Est-il possible de compléter une base de connaissances (nouveaux liens)?
- Mieux représenter les connaissances pour la normalisation
- Emergence d'un paradigme (en parallèle de Word2Vec)



NOMBREUSES VARIANTES

Complexifier les opérateurs entre entités (dépasser la translation):



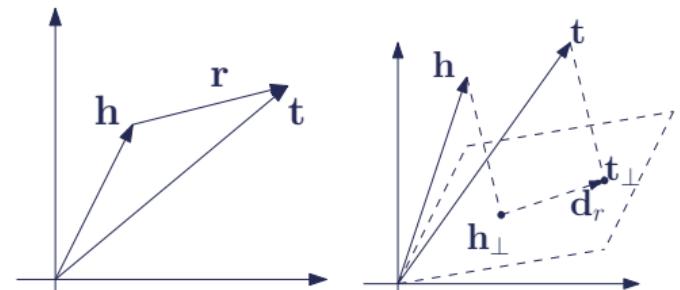


NOMBREUSES VARIANTES

Model	Score function $f_r(\mathbf{h}, \mathbf{t})$	# Parameters
TransE (Bordes et al. 2013b)	$\ \mathbf{h} + \mathbf{r} - \mathbf{t}\ _{\ell_{1/2}}, \mathbf{r} \in \mathbb{R}^k$	$O(n_e k + n_r k)$
Unstructured (Bordes et al. 2012)	$\ \mathbf{h} - \mathbf{t}\ _2^2$	$O(n_e k)$
Distant (Bordes et al. 2011)	$\ W_{rh}\mathbf{h} - W_{rt}\mathbf{t}\ _1, W_{rh}, W_{rt} \in \mathbb{R}^{k \times k}$	$O(n_e k + 2n_r k^2)$
Bilinear (Jenatton et al. 2012)	$\mathbf{h}^\top W_{rt} \mathbf{t}, W_r \in \mathbb{R}^{k \times k}$	$O(n_e k + n_r k^2)$
Single Layer	$\mathbf{u}_r^\top f(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r (sk + s))$
NTN (Socher et al. 2013)	$\mathbf{u}_r^\top f(\mathbf{h}^\top \mathbf{W}_{rt} \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, \mathbf{W}_r \in \mathbb{R}^{k \times k \times s}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r (sk^2 + 2sk + 2s))$
TransH (this paper)	$\ (\mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r)\ _2^2$ $\mathbf{w}_r, \mathbf{d}_r \in \mathbb{R}^k$	$O(n_e k + 2n_r k)$

Complexifier les opérateurs entre entités
(dépasser la translation):

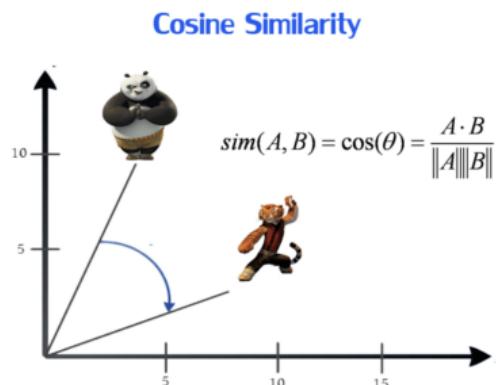
Knowledge Graph Embedding by Translating on
Hyperplanes, AAAI 2014, Wang et al.



A

Embedding / plongement

- Similarité cosinus
- Avec ou sans contextualisation
 - Mot (dans un contexte)
- Agrégation type CLS
 - Phrase d'origine
 - Définition du dictionnaire



Contextual representation



LLM

Token representation



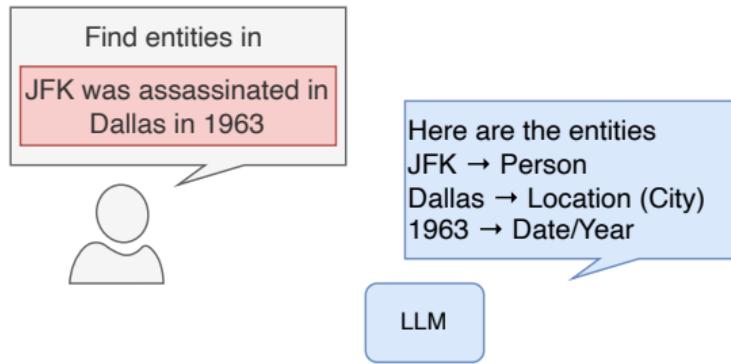
Aspirin : medication that relieves pain, reduces fever and inflammation

EVALUER LES SORTIES D'UN LLM



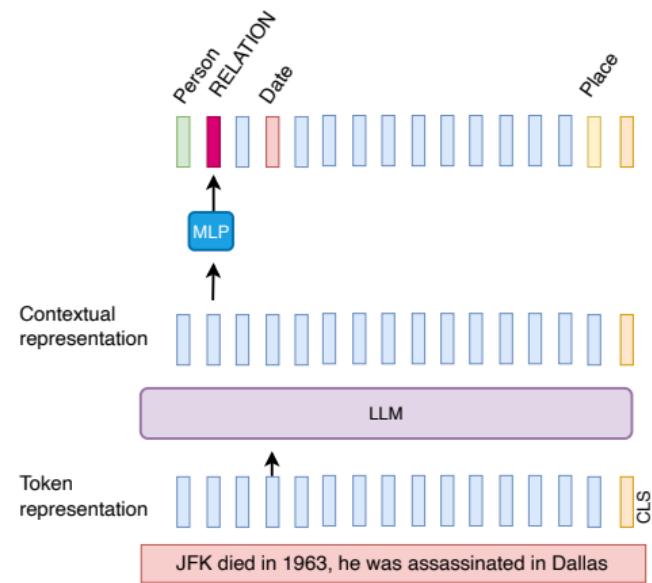
De plus en plus de sorties textuelles

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?)



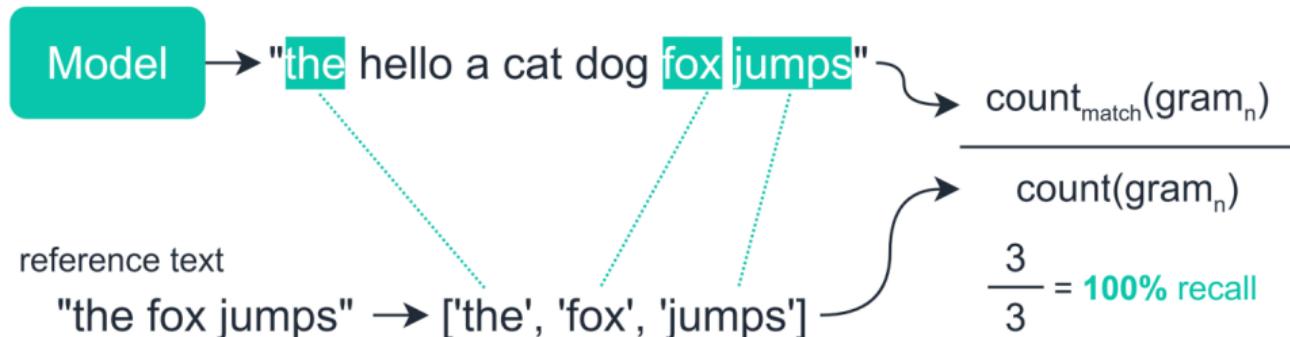
JFK PERSON was assassinated in Dallas GPE in 1963 DATE



IA générative : comment évaluer la performance ?

Le point critique aujourd'hui

- Comment évaluer par rapport à la vérité terrain ?
- Confiance du système / Vraisemblance de la génération ?

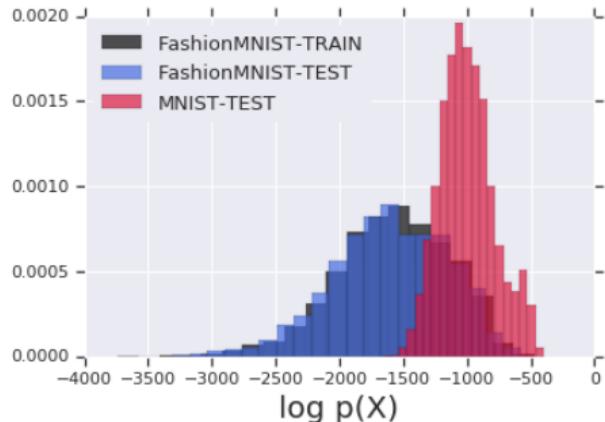




IA générative : comment évaluer la performance ?

Le point critique aujourd'hui

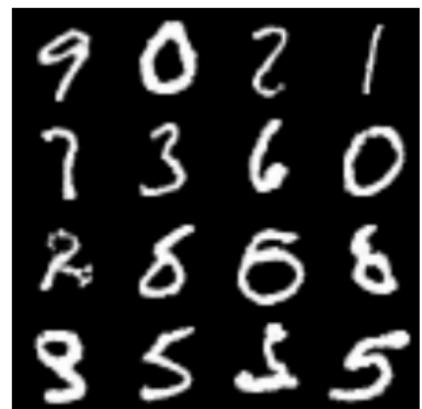
- Comment évaluer par rapport à la vérité terrain ?
- Confiance du système / Vraisemblance de la génération ?



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



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, \approx 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.

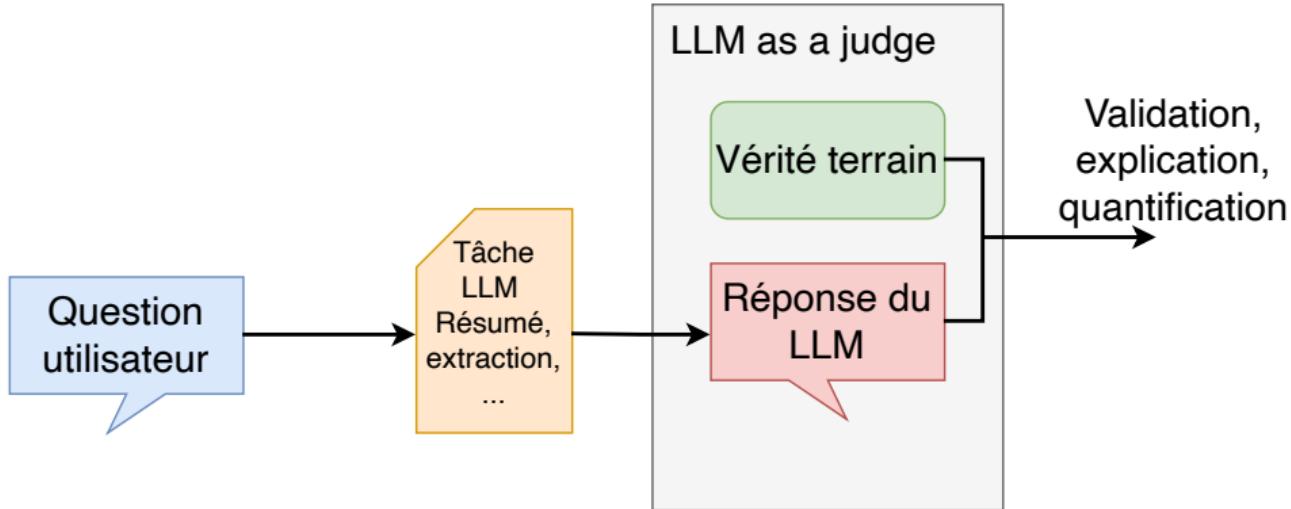
Voici une décomposition en statements élémentaires :

1. Des drones non identifiés ont été observés.
2. L'observation a eu lieu au-dessus de la plus grande base militaire.
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 entraînés sur la tâche (RoBerta, ...)



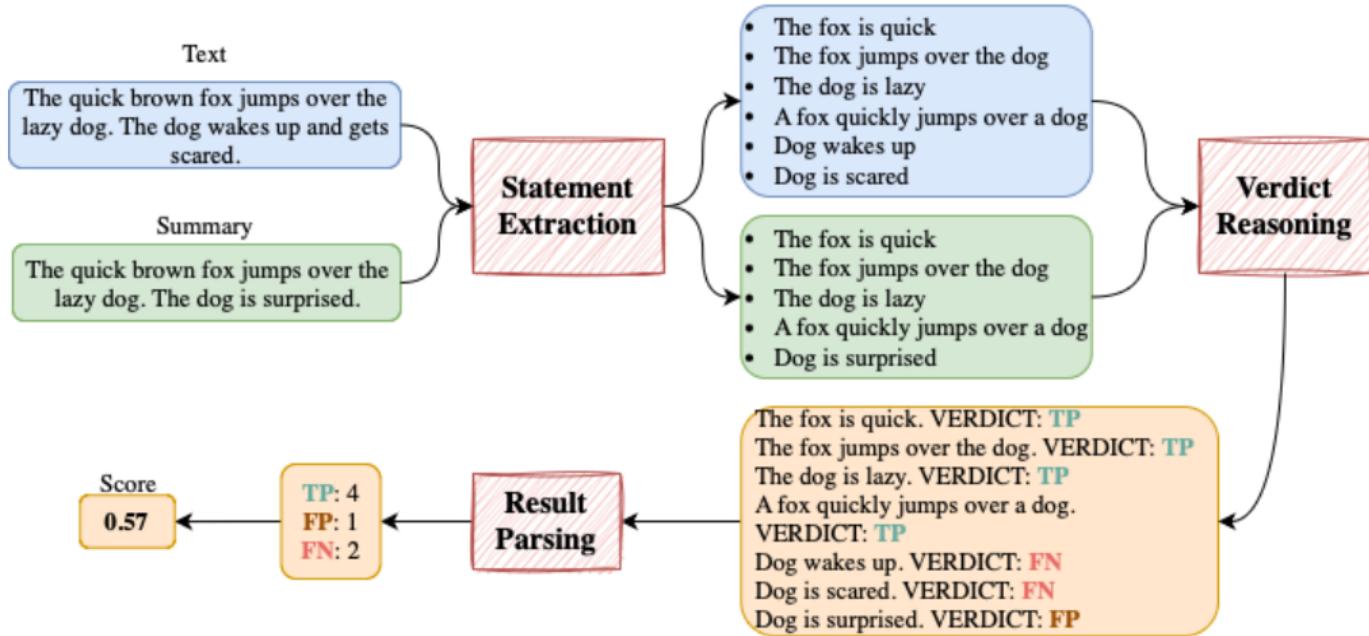
LLM as a judge: la reflexivité des LLM



- La performance du LLM n'est pas symétrique... 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 stratégie?



Évaluation de résumés automatique



- ⇒ À la fin, nous ne transformons pas le texte original
- ⇒ Nous effectuons l'extraction des informations atomiques sur le résumé



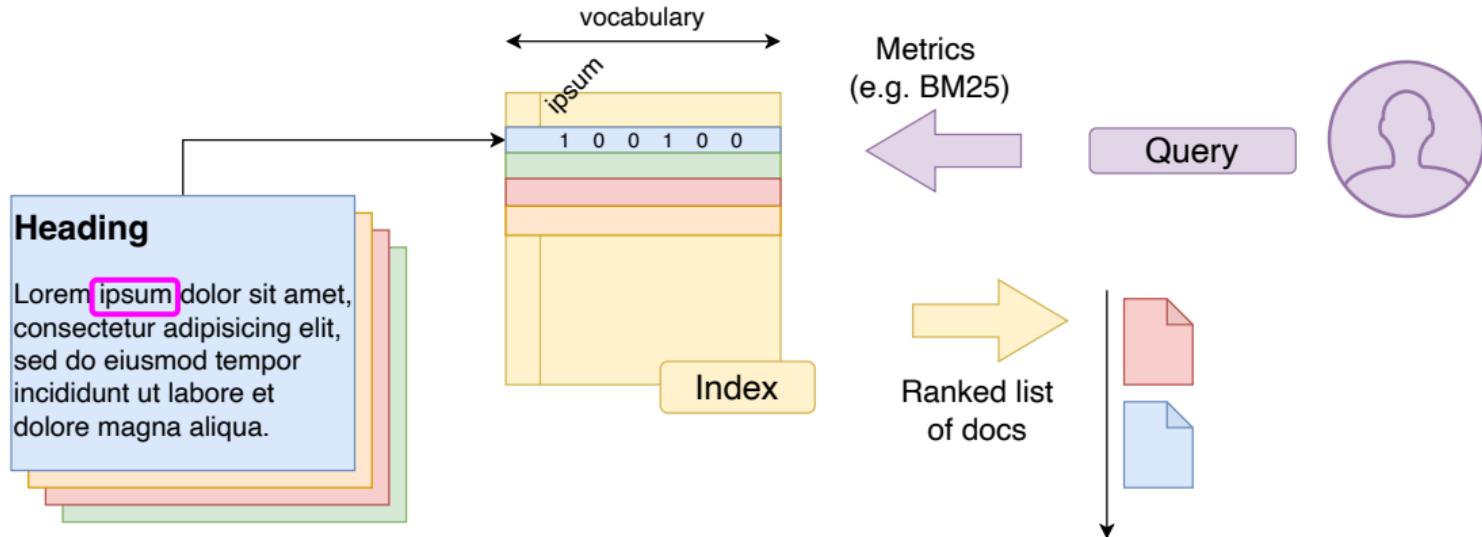
Évaluation de résumés automatique

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
	MOVERS core	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

QUEL(S) FUTUR(S) POUR L'ACCÈS À L'INFORMATION?

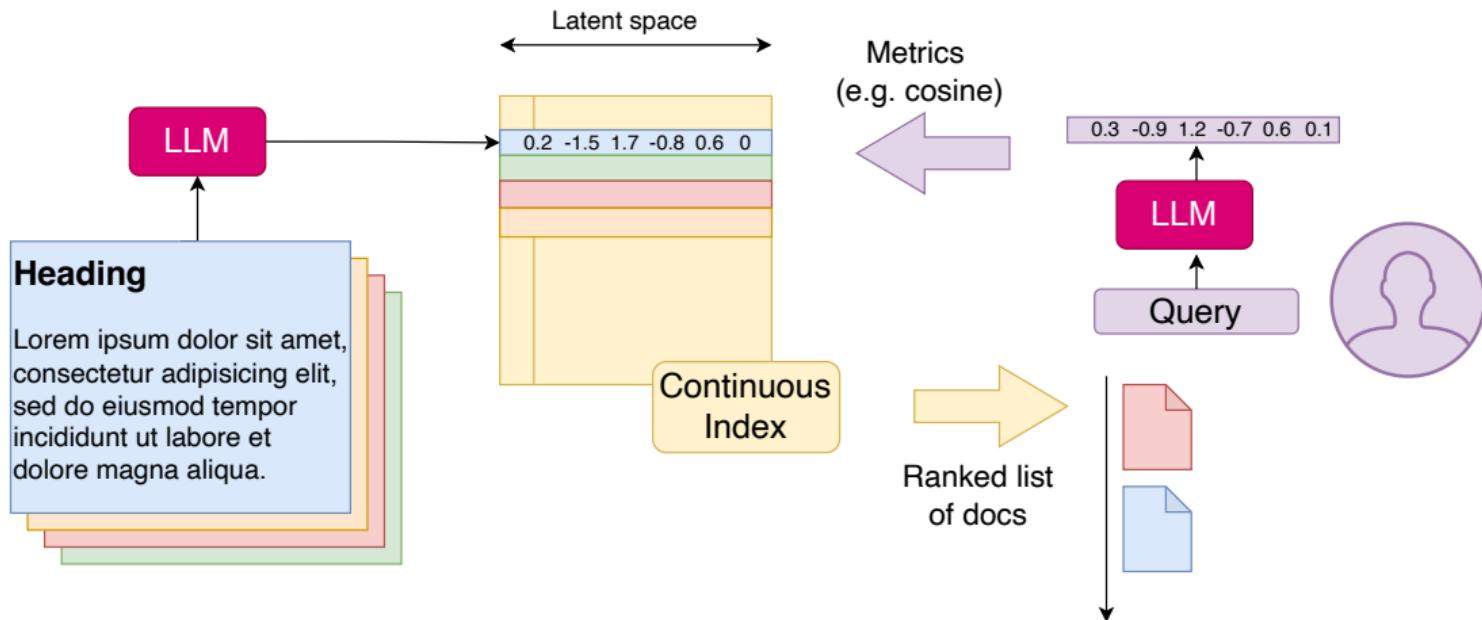


LLM & Information Retrieval



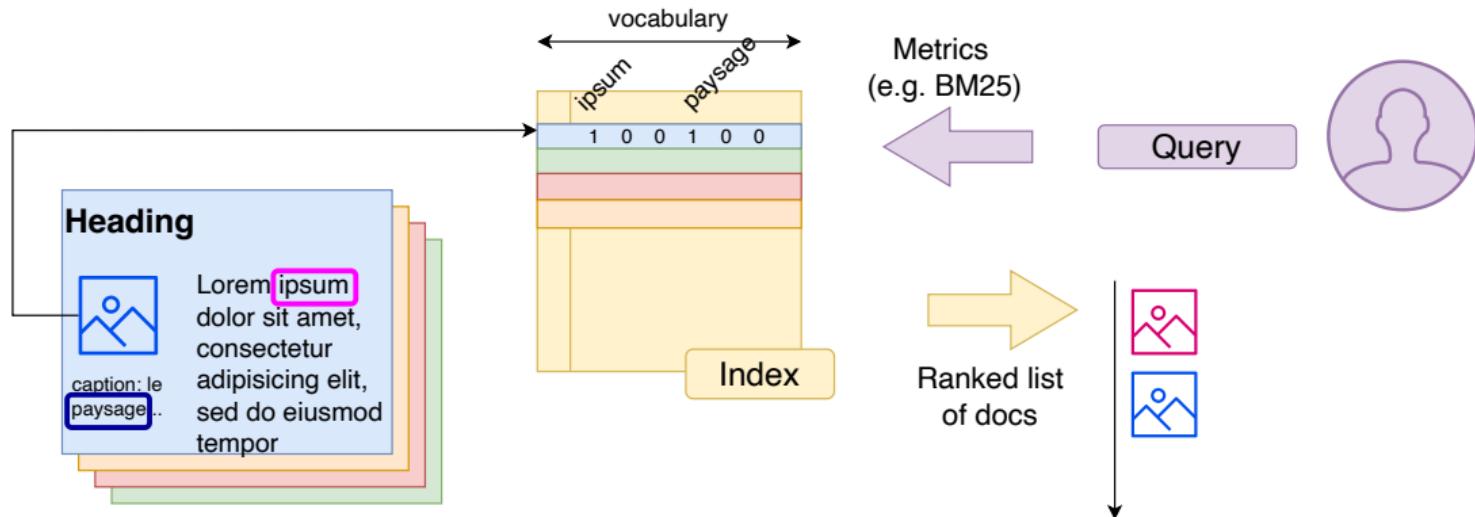


LLM & Information Retrieval



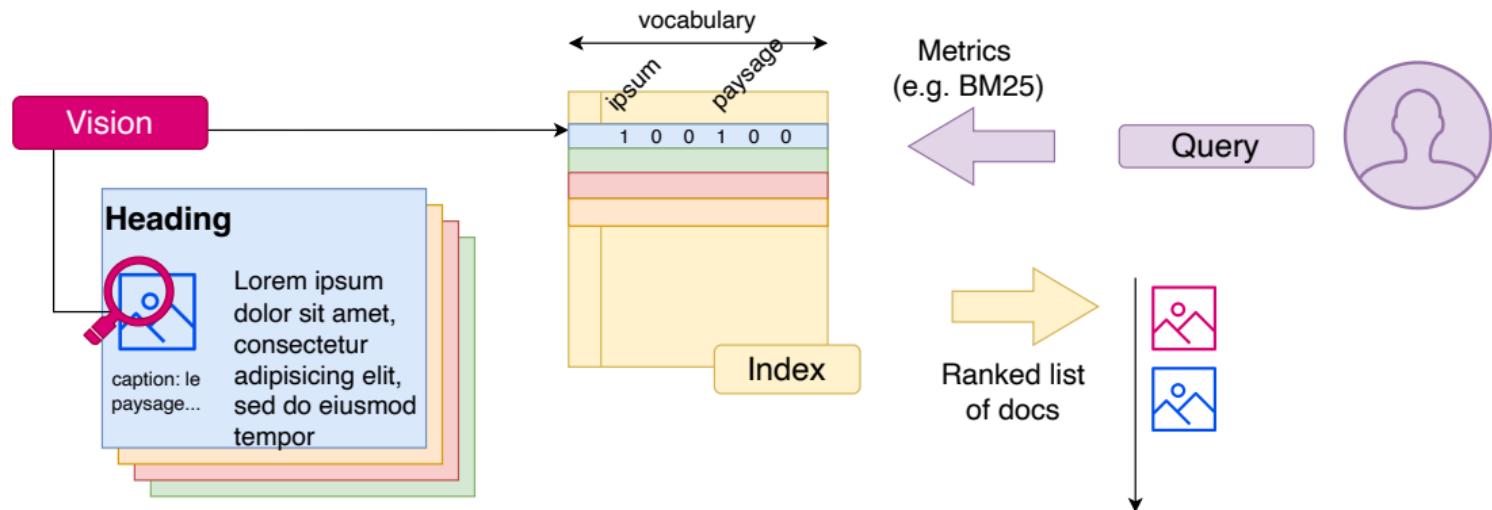


LLM & Information Retrieval



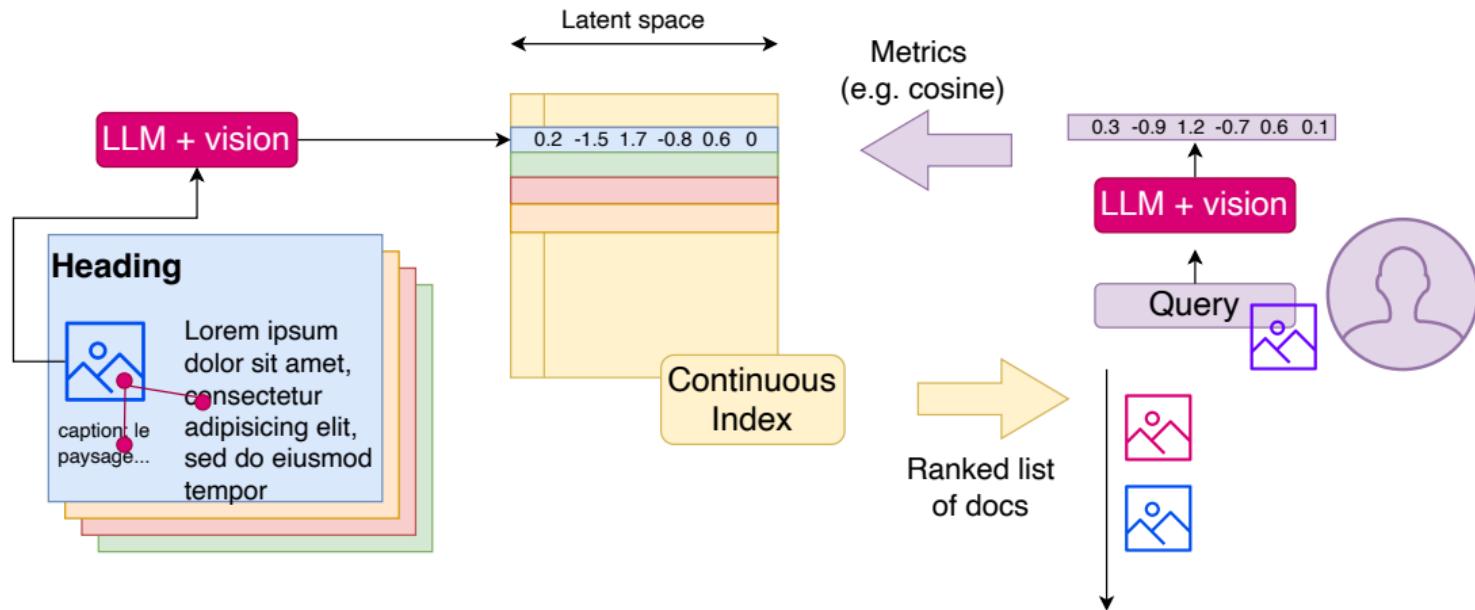


LLM & Information Retrieval



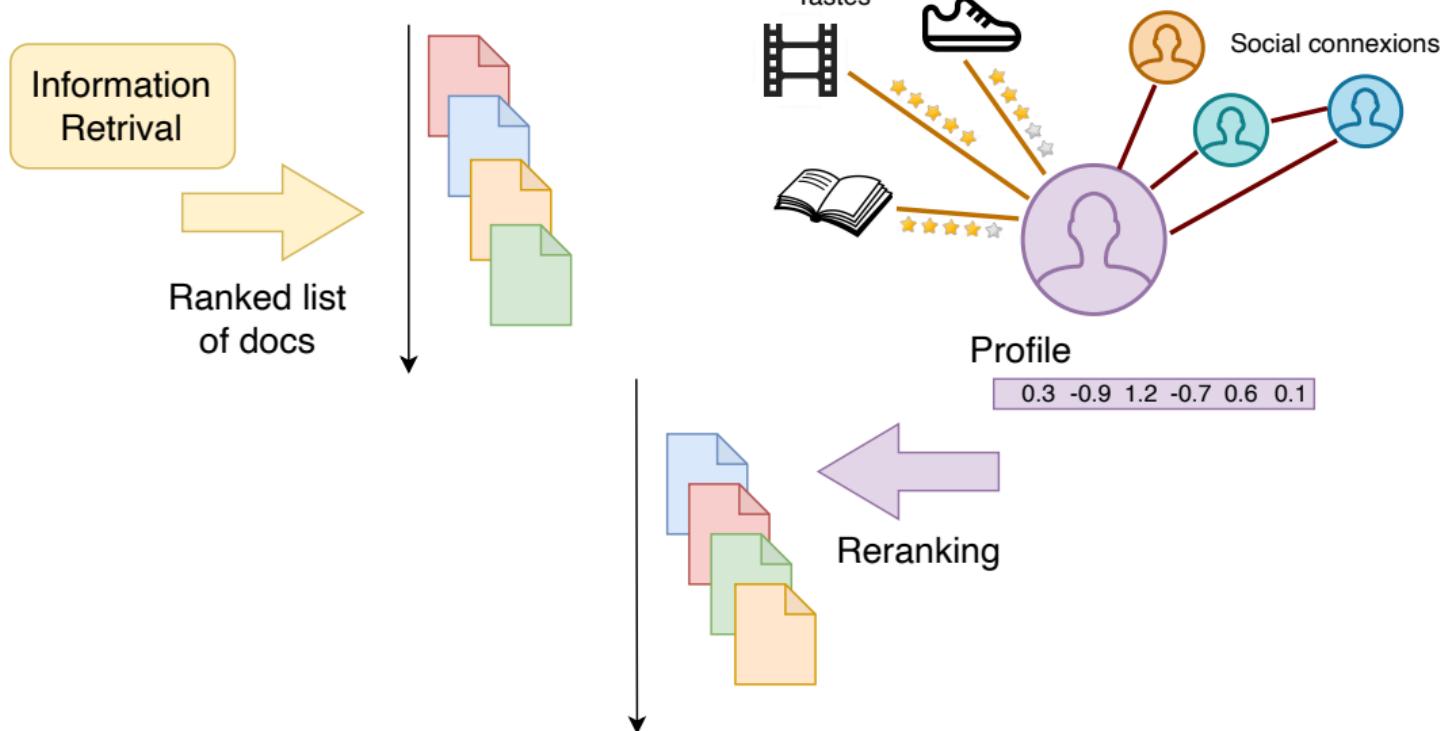


LLM & Information Retrieval



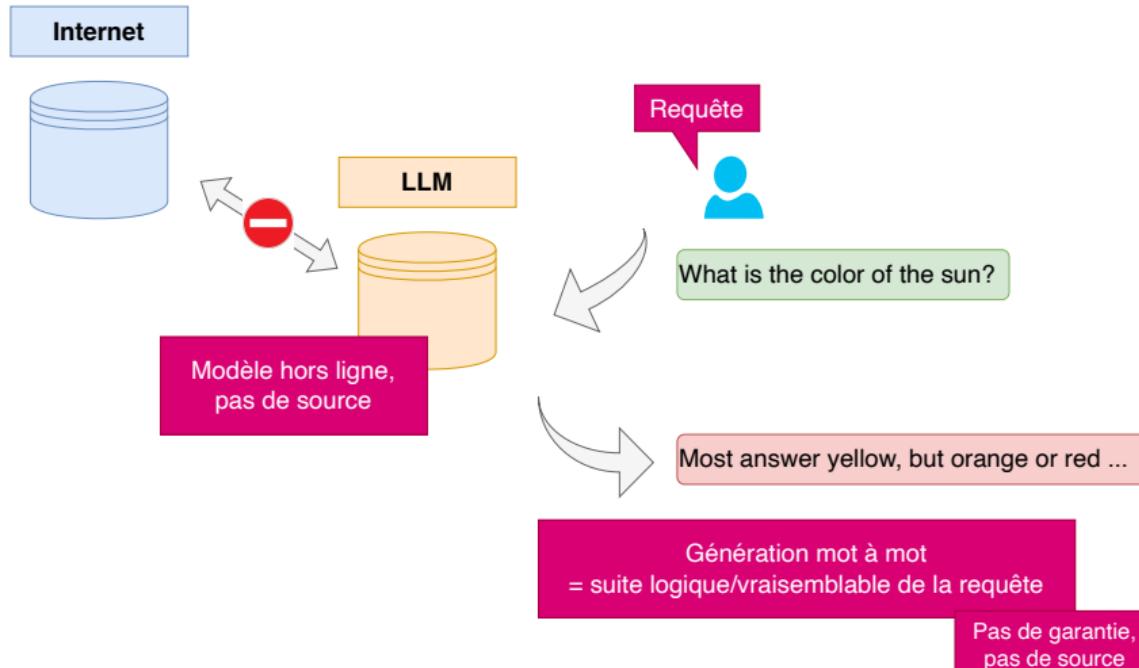


LLM & Information Retrieval





Usage en accès à l'information

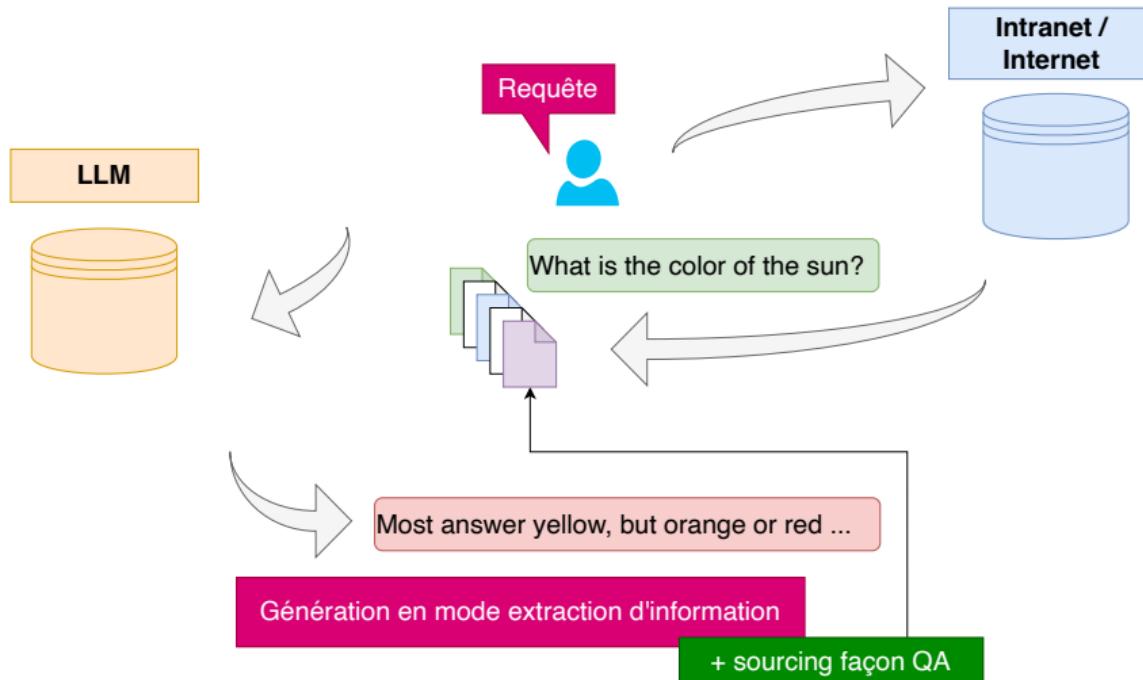


- LLM limité en connaissances
- Risque d'hallucination à la génération



Usage en accès à l'information

- Demander des informations à chatGPT... Un usage étonnant !

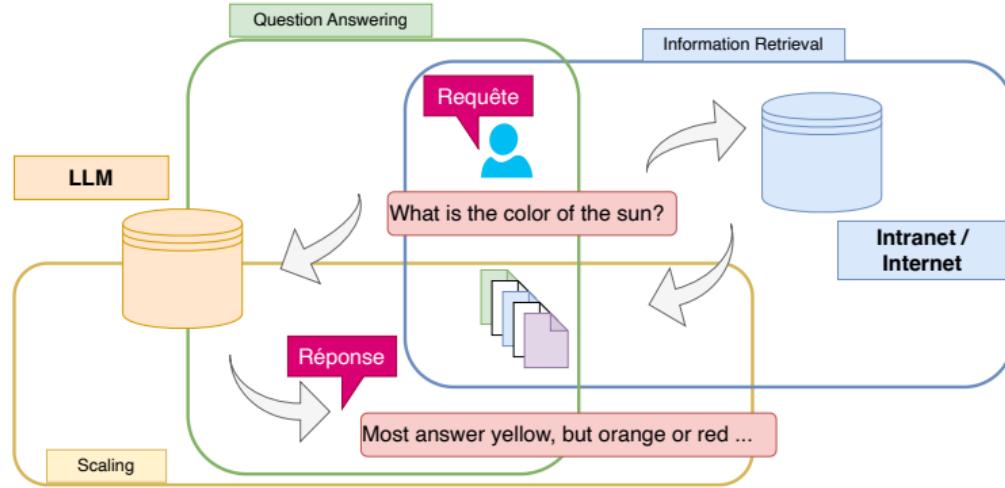


- *RAG: Retrieval Augmented Generation*

- Limite (actuelle) sur la taille des entrées (2k puis 32k puis 100k tokens)



L'état de l'art en RAG



Retrieval-Augmented Generation (RAG) [1]

Improve performance on knowledge intensive task (question answering)

Retrieval-Augmented Language Model (REALM) [2]

Integrate retrieval augmented into the pre-training

Retrieval-Enhanced Transformer (RETRO) [3]

Scale generation to large number of retrieved documents

[1] Guu et al (2020), REALM: Retrieval-Augmented Language Model Pre-Training

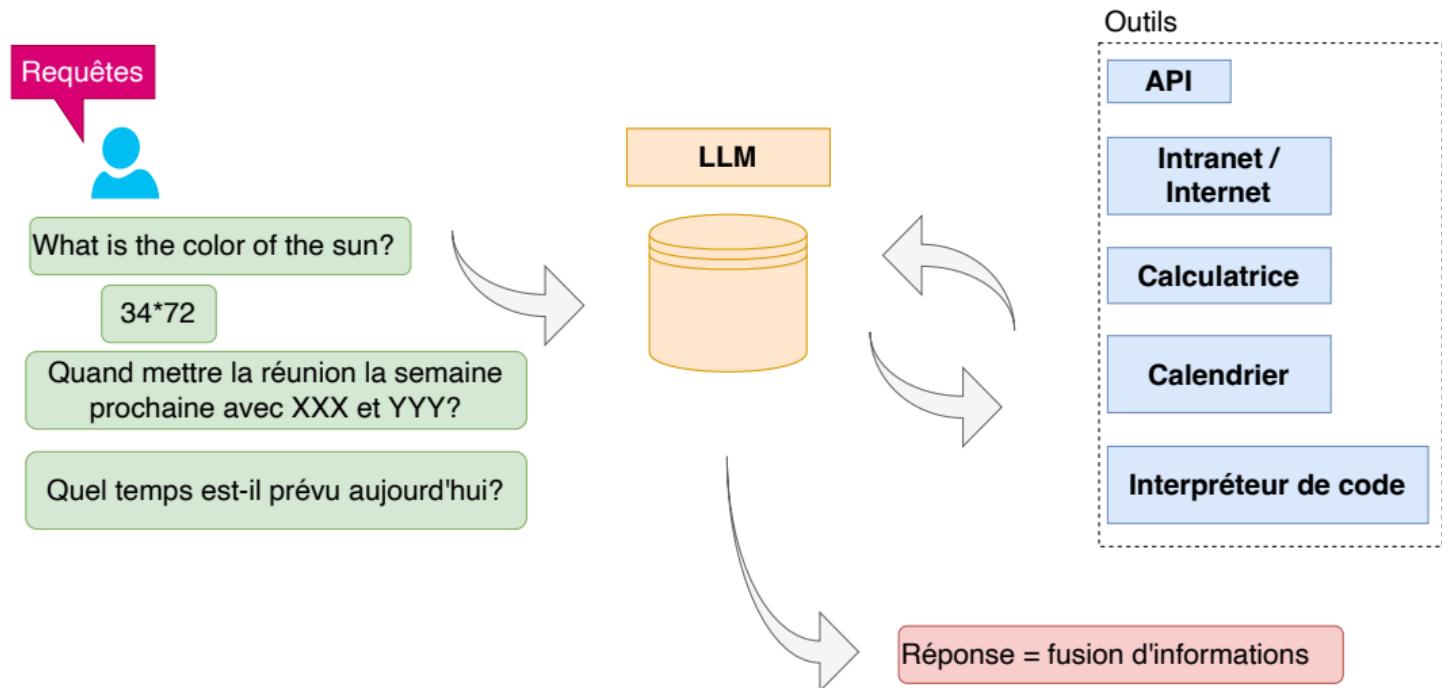
[2] Lewis et al (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

[3] Borgeaud et al (2022) Improving Language Models by Retrieving from Trillions of Tokens



Multiplier les outils: le LLM / couteau Suisse

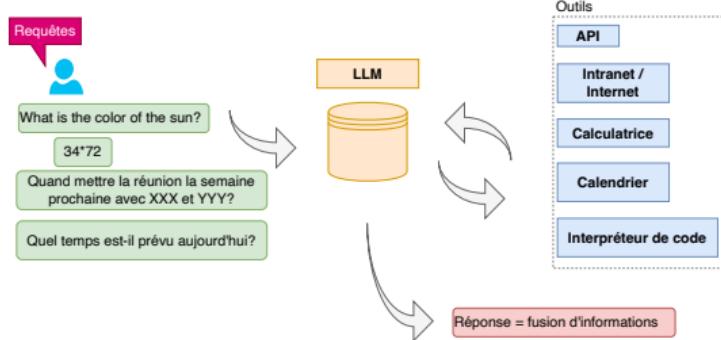
- Apprendre au LLM à appeler (*balise*) des outils externes





Multiplier les outils: le LLM / couteau Suisse

■ Apprendre au LLM à appeler (*balise*) des outils externes



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.

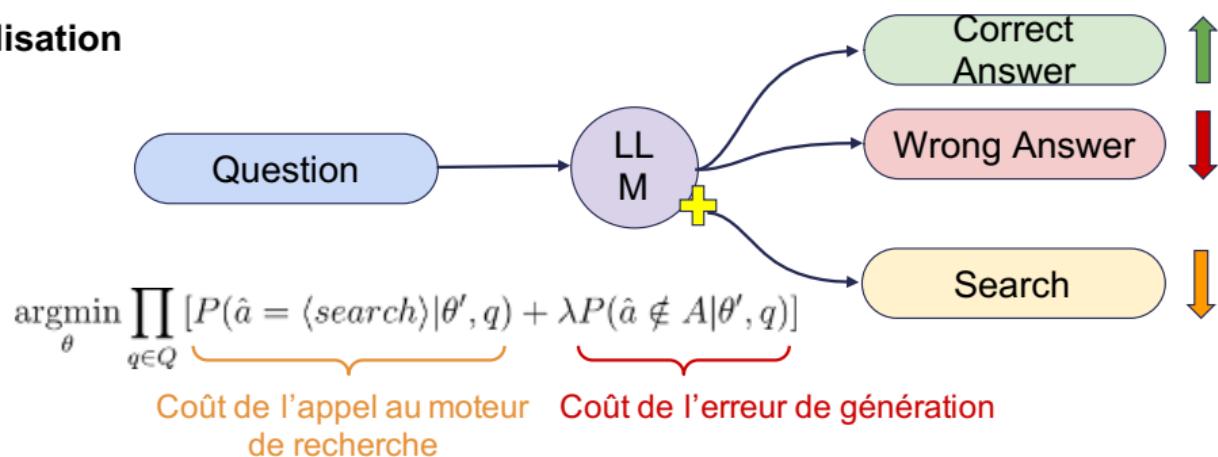
The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their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



Optimiser le cout des outils

Objectif : Apprendre à générer le token `<SEARCH>` lorsque cela est nécessaire

Formalisation



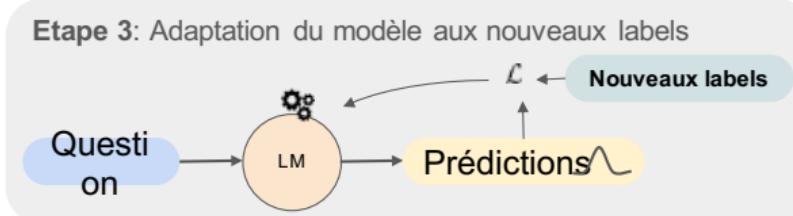
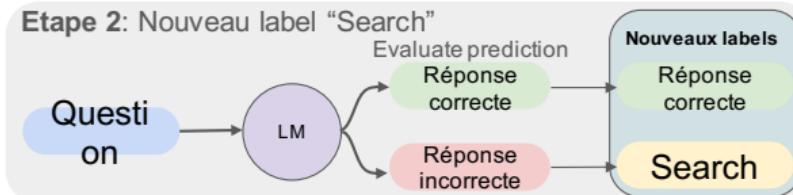
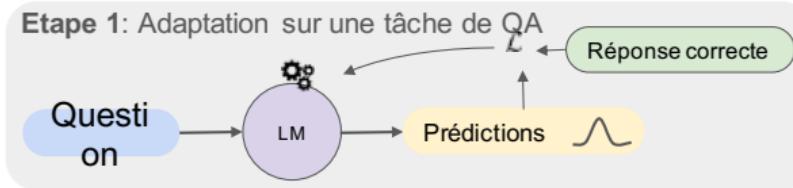
- Toolsformer appelle le moteur de recherche dans 99% des cas
- Peut-on faire la balance avec les connaissances du LLM?



Optimiser le cout des outils

Apprendre une fonction de filtrage qui :

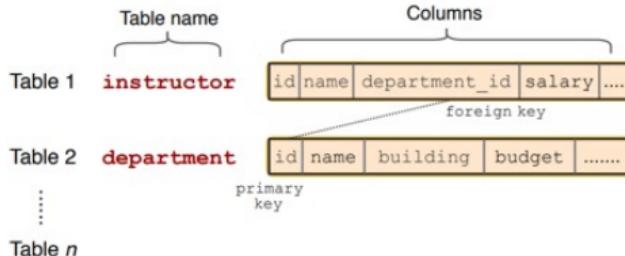
- Laisse inchangé les **Correct Answer**
- Masque les **Wrong Answer** avec **Search**





Le SQL: un outil comme les autres?

Annotators check database schema (e.g., database: college)



Annotators create:

Complex question What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL

```

SELECT T2.name, T2.budget
FROM instructor AS T1 JOIN department AS T2
ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
(SELECT avg(salary) FROM instructor)
    
```

Easy

What is the number of cars with more than 4 cylinders?

```

SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
    
```

Medium

For each stadium, how many concerts are there?

```

SELECT T2.name, COUNT(*)
FROM concert AS T1 JOIN stadium AS T2
ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
    
```

Hard

Which countries in Europe have at least 3 car manufacturers?

```

SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id
JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
    
```

Extra Hard

What is the average life expectancy in the countries where English is not the official language?

```

SELECT AVG(life_expectancy)
FROM country
WHERE name NOT IN
(SELECT T1.name
FROM country AS T1 JOIN
country_language AS T2
ON T1.code = T2.country_code
WHERE T2.language = "English"
AND T2.is_official = "T")
    
```

Figure 3: SQL query examples in 4 hardness levels.

- TableQA: schema + question \Rightarrow SQL
- Prédire ce qui est facile ou dur



Le SQL: un outil comme les autres?

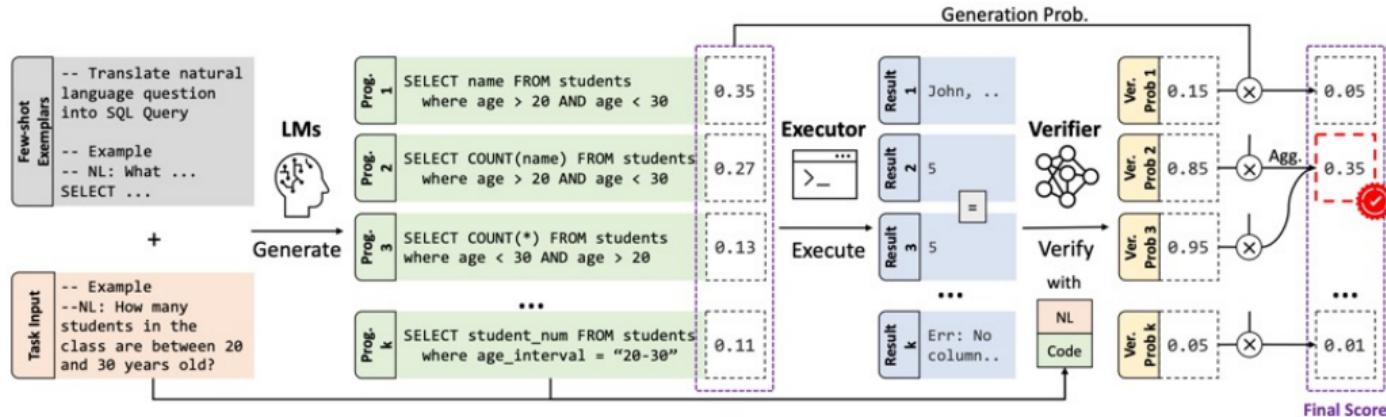


Figure 1: The illustration of LEVER using text-to-SQL as an example. It consists of three steps: 1) *Generation*: sample programs from code LLMs based on the task input and few-shot exemplars; 2) *Execution*: obtain the execution results with program executors; 3) *Verification*: using a learned verifier to output the probability of the program being correct based on the NL, program and execution results.

- Prédire les bonnes et les mauvaises réponses
- Plus de feedback pour mieux apprendre

(Ni et al. 2023), LEVER: Learning to Verify Language-to-Code Generation with Execution



Le SQL: un outil comme les autres?

Question : The player's career spanned less than 20 years ?

Date	Games	Yards	Team
1983	16	1,808	Los Angels Rams
1984	16	2,105	Los Angels Rams
...			
1993	4	91	Atlanta Falcons
Career	146	13,256	

Answer : True

(a) Extractive

Column cells Selection
1993 .. 1983

Aggregations
Max

Answer : 1993 ✗

Error Explanation
struggles with Complex, multi-aggregation queries.

(b) SQL-Queries

SQL Generation
select (select max (Date) - min (Date) from w) < 20

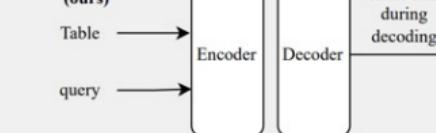
Answer : None ✗

(c) Direct Answer Generation

Answer : False ✗

Error Explanation
Limited Numerical Reasoning in Transformers

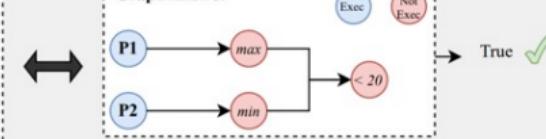
d) Partial execution (ours)



Logical Form answer

P1
 $< 20 || - || \max || 1983 | \dots | 1993 || \min$
 P2
 $|| 1984 | \dots | 1993 ||$

Graph Answer



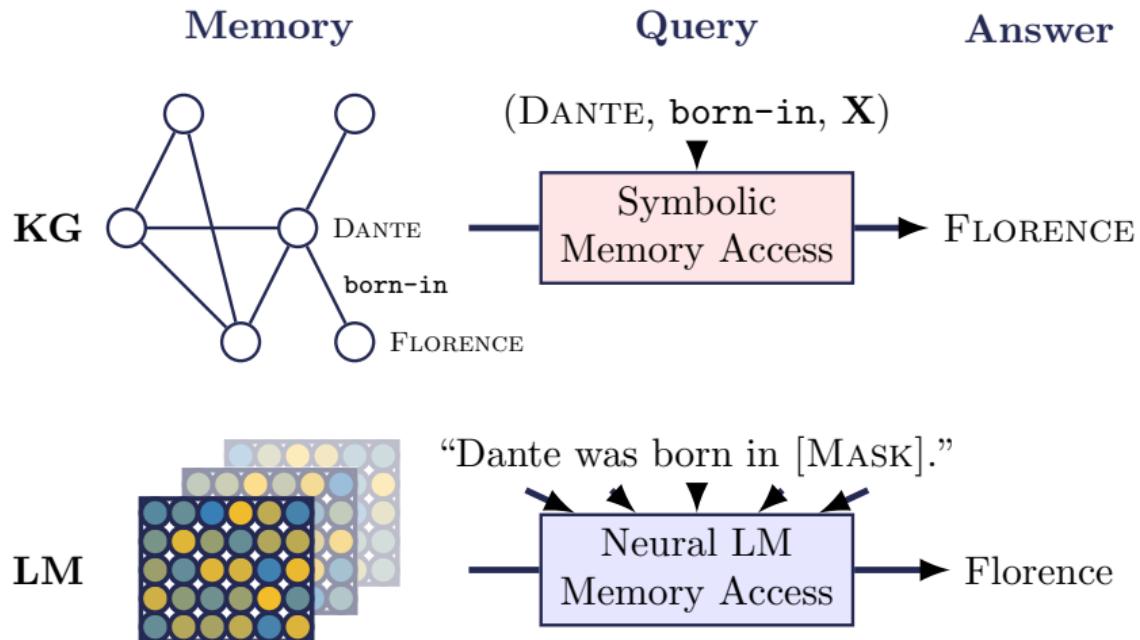
- Apprendre à raisonner numériquement à partir d'une base étiquetée en SQL
- Le LLM apprend à évaluer les requêtes SQL

(Mouravieff et al. 2024), Training Table Question Answering via SQL Query Decomposition

CONCLUSION



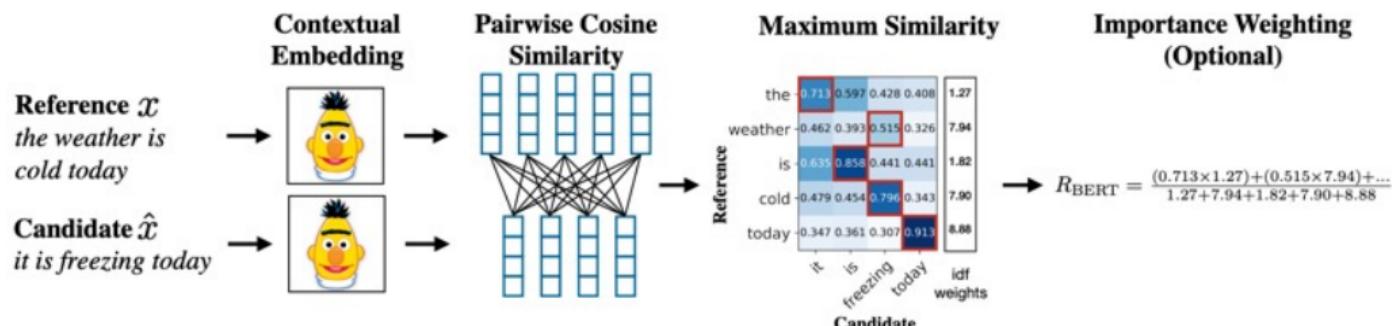
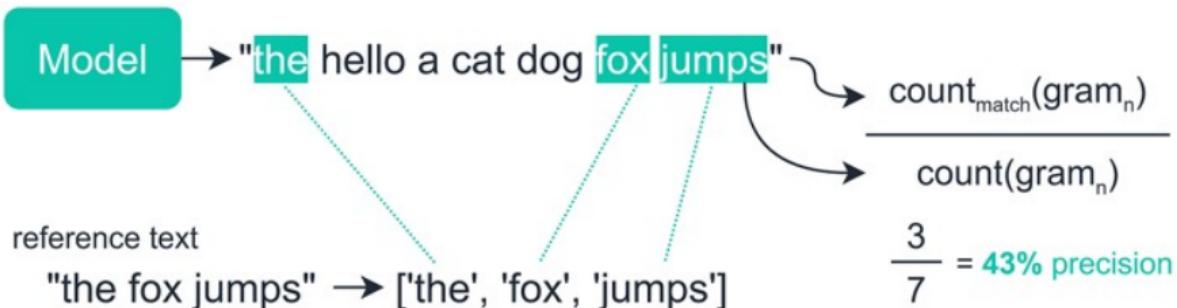
Sous quelle forme stocker les connaissances?



- Exhaustivité?
- Fiabilité?



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
 - Contrôle / garantie sur la génération