

Manipulation de Données Structurées et Interaction avec des Outils Externes grâce aux LLMs :

L'Avenir de l'IA Générative dans les Systèmes d'Information ?

Laure Soulier

29 mai 2024

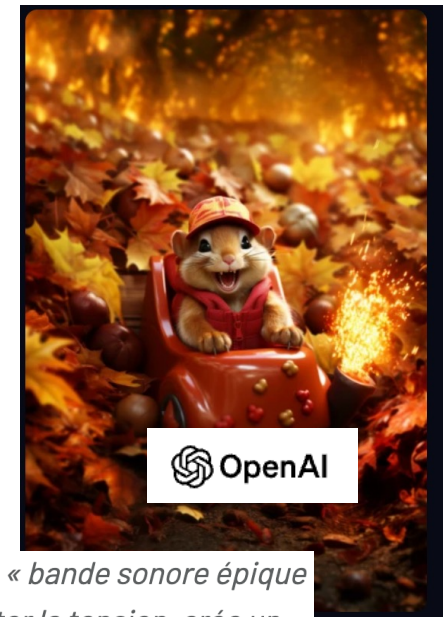
Modèles de langue et IA générative

LA

Ecris un texte de 5 phrases sur Jian-Yun Nie



Jian-Yun Nie est un chercheur et professeur émérite en informatique. Il est connu pour ses contributions significatives dans le domaine du traitement de l'information et de la recherche d'information. Ses travaux portent notamment sur l'exploration de textes, la recherche d'informations multilingues et la fouille de données. En tant que chercheur de renommée mondiale, Jian-Yun Nie a publié de nombreux articles scientifiques et a participé à des conférences internationales. Ses recherches ont eu un impact important dans le domaine de l'informatique et continuent d'influencer de nouvelles avancées technologiques.



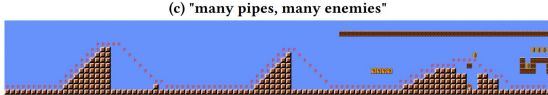
(a) "many pipes, many enemies, little blocks, low elevation"



(b) "no pipes, some enemies, many blocks, high elevation"



(c) "many pipes, many enemies"



(d) "no pipes, no enemies, many blocks"



(e) Prompt that does not exist in the dataset: "many pipes, no enemies, many blocks"



(f) Failure case: "Many pipes, no enemies, some blocks"

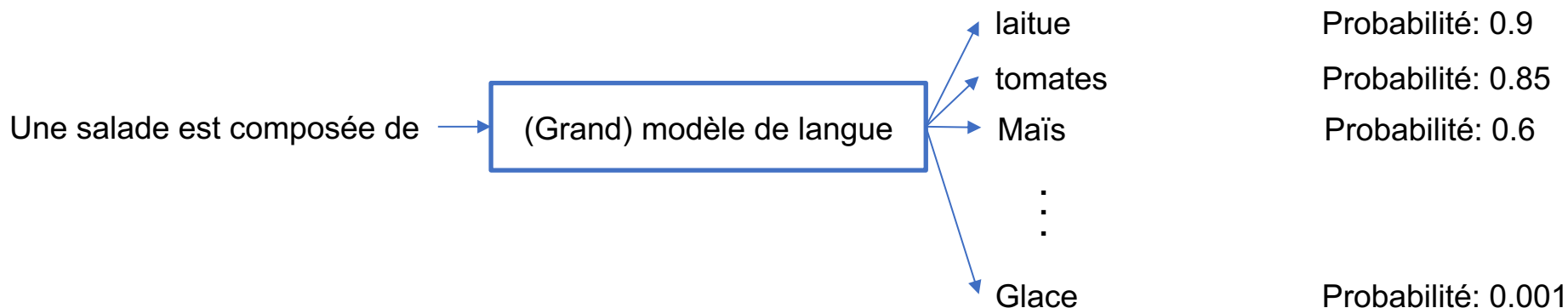
Pour en avoir un aperçu, voici l'extrait généré pour le prompt « *bande sonore épique utilisant des instruments orchestraux. Le morceau fait monter la tension, crée un sentiment d'urgence. Un chœur a cappella chante à l'unisson, ce qui crée un sentiment de puissance et de force* » :

© openAI, Midjourney,
<https://www.blogdumoderateur.com/musiclm-nouvel-outil-ia-google-generer-musique/>
And (Sudhakaran et al 2023)



Principe :

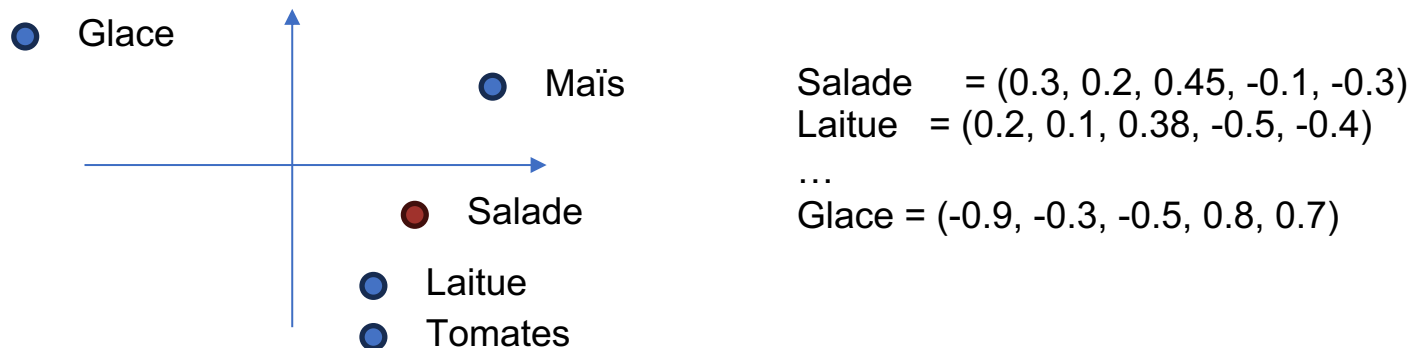
- Modéliser la probabilité d'une séquence x_1, x_2, \dots, x_n
- Les items peuvent être des mots, des caractères, des ngrams/bouts de mots, etc



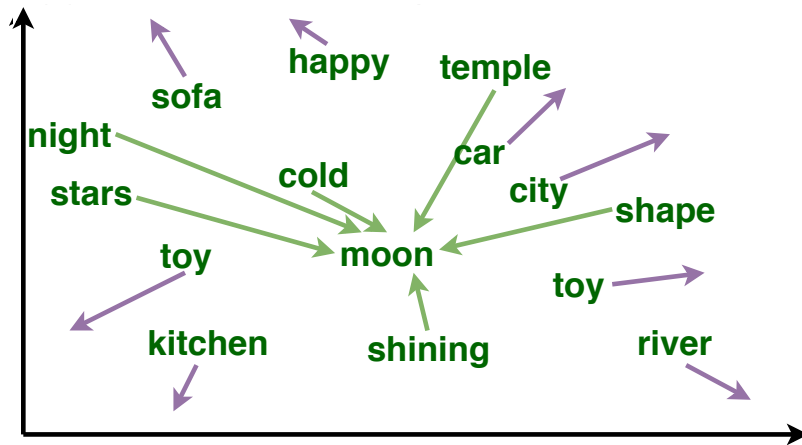
Etant donné une séquence x_1, x_2, \dots, x_{n-1} , quelle est la probabilité du prochain item x_n ?

$$P(x_n | x_1, x_2, \dots, x_{n-1})$$

Sémantique, représentation, espace latent



→ Algorithme Word2Vec



he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The splash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

2000

Modèle pionnier
de Bengio

2012

Word2Vec,
FastText, ...

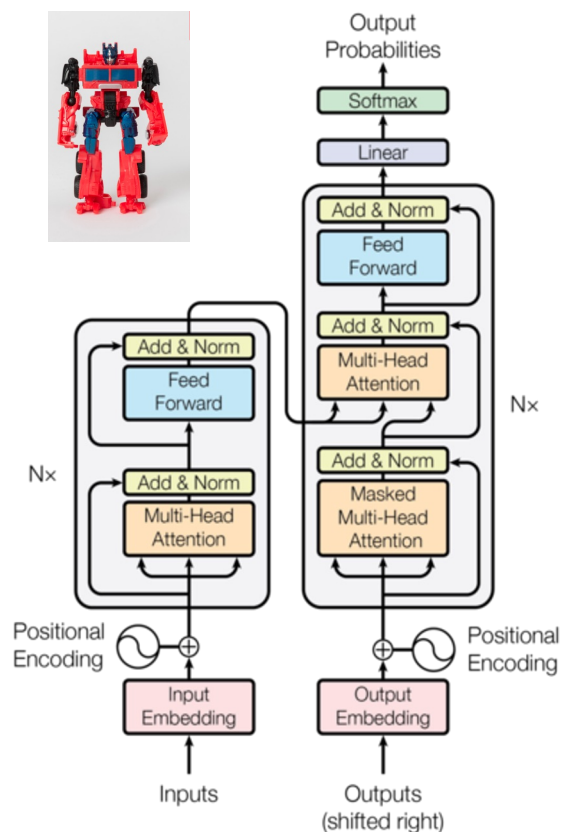
2014

Doc2Vec,
FastSent, ...

2017

Représentations contextuelles
Transformer networks
Bert, T5, GPT, ...

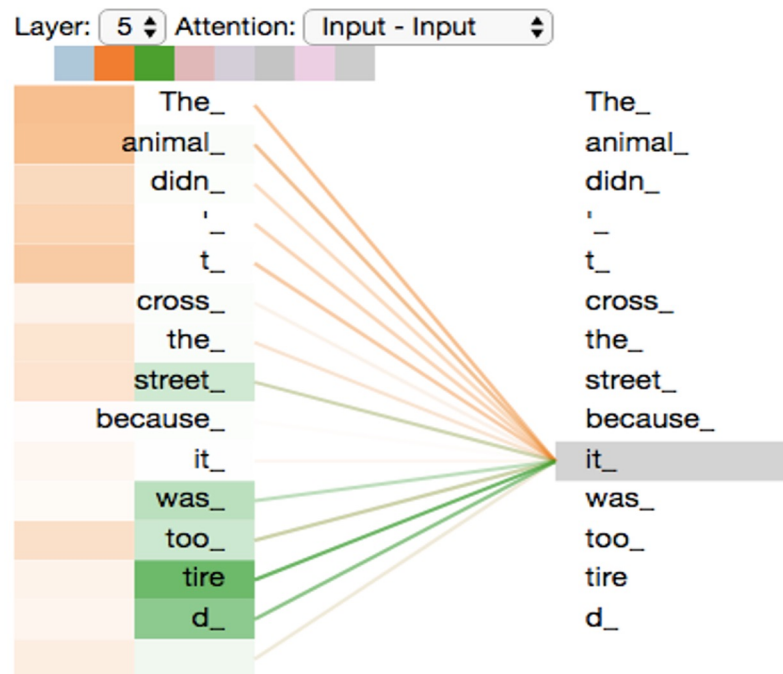
Transformer (2017)



Un encoder-decoder avec :

- Environ 65 millions de paramètres (maintenant plus)
- Plusieurs blocs successifs
- Des têtes parallèles

... qui estime des représentations contextuelles des items avec l'attention propre

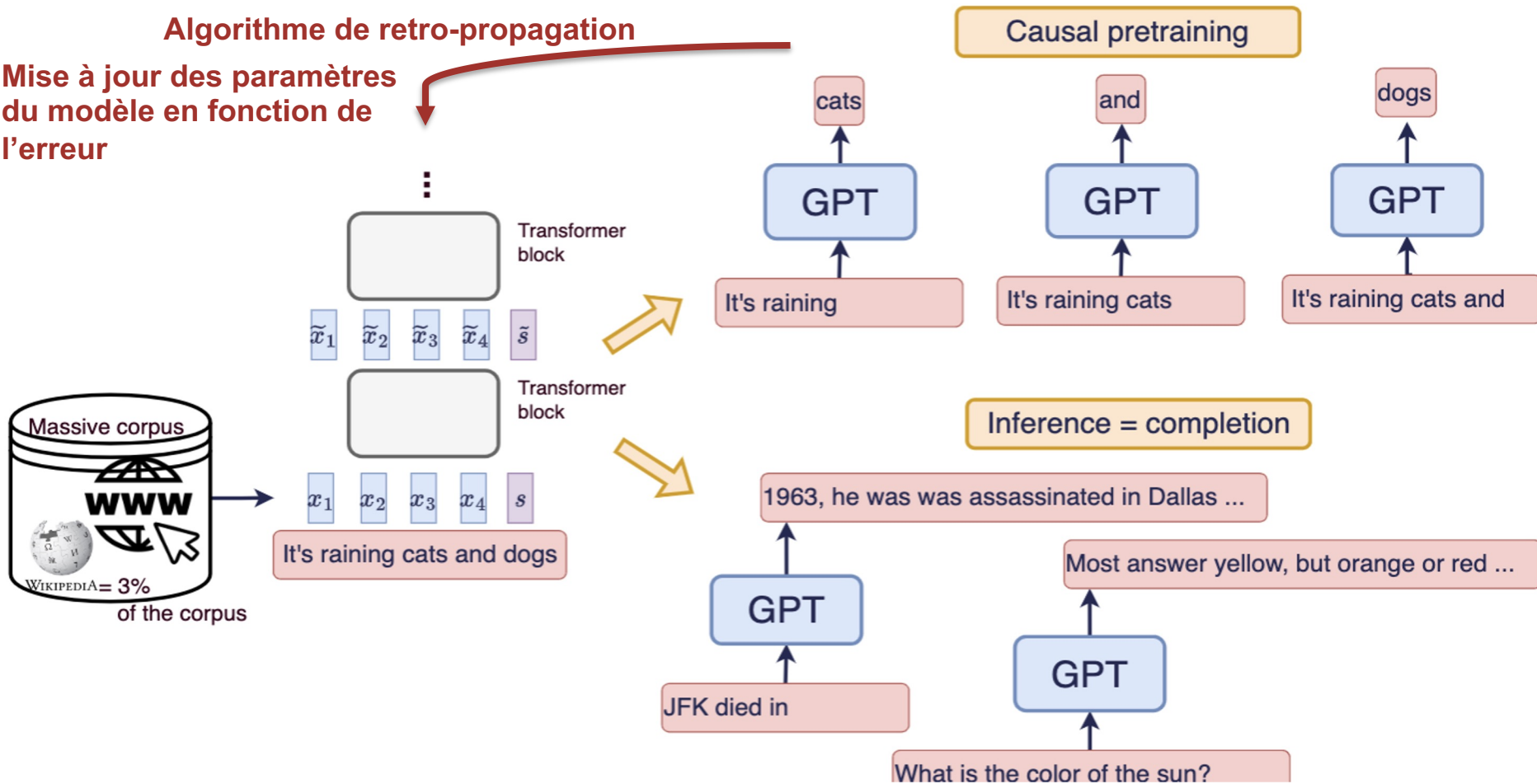


Distinguer *Washington/city* de *Washington/man*

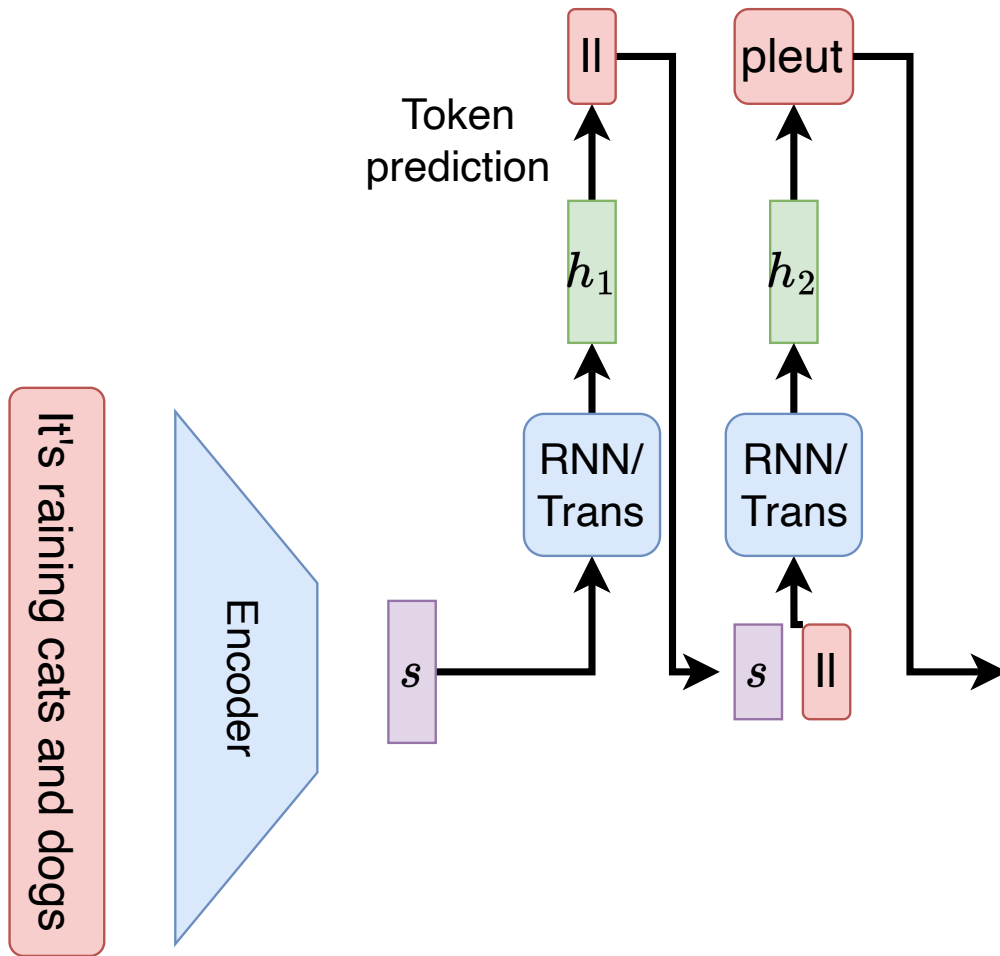
Entraîner un transformer (e.g. GPT)

Algorithme de retro-propagation

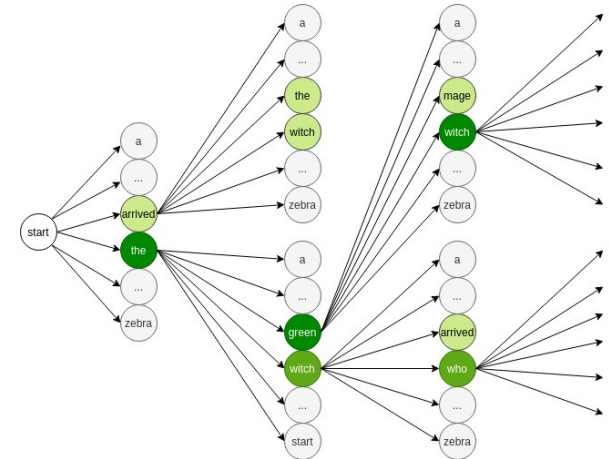
Mise à jour des paramètres du modèle en fonction de l'erreur



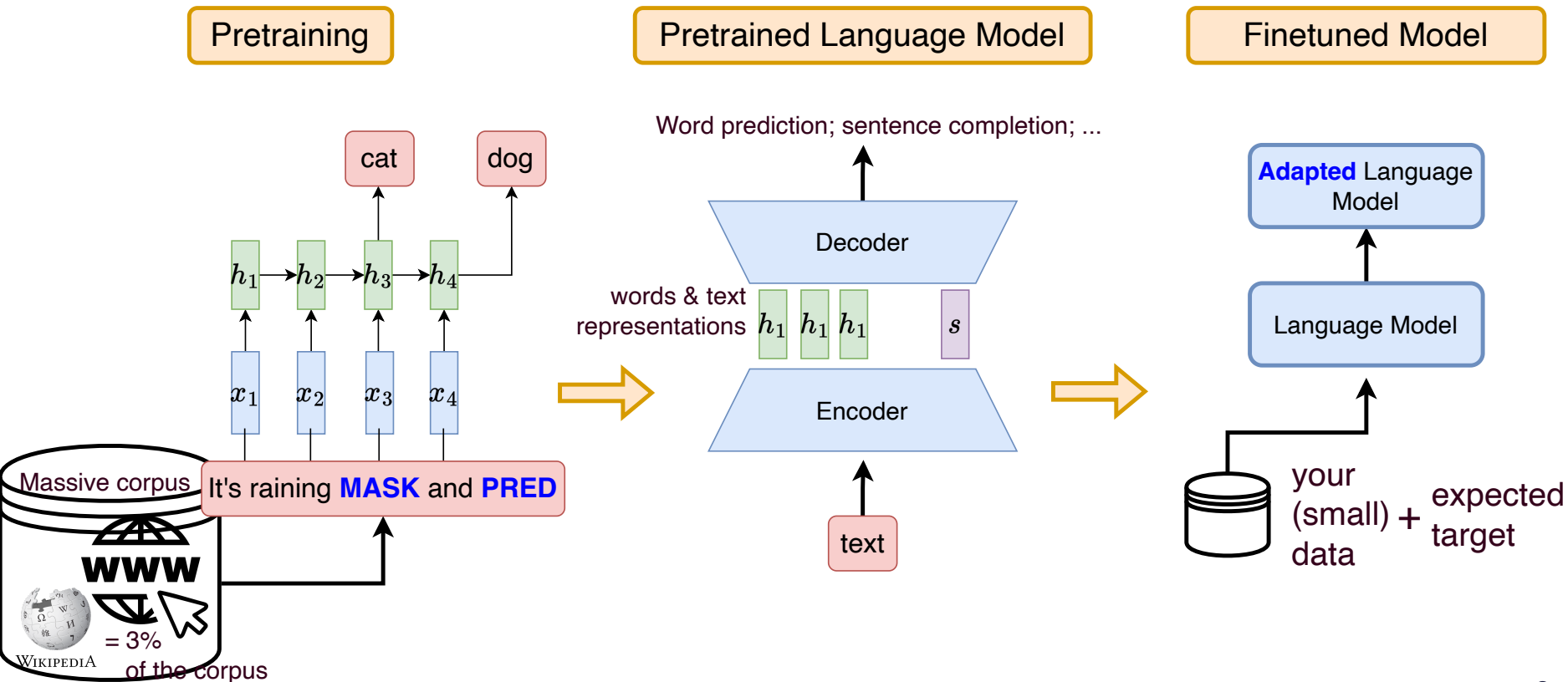
→ Exploiter les architectures précédentes pour écrire du texte



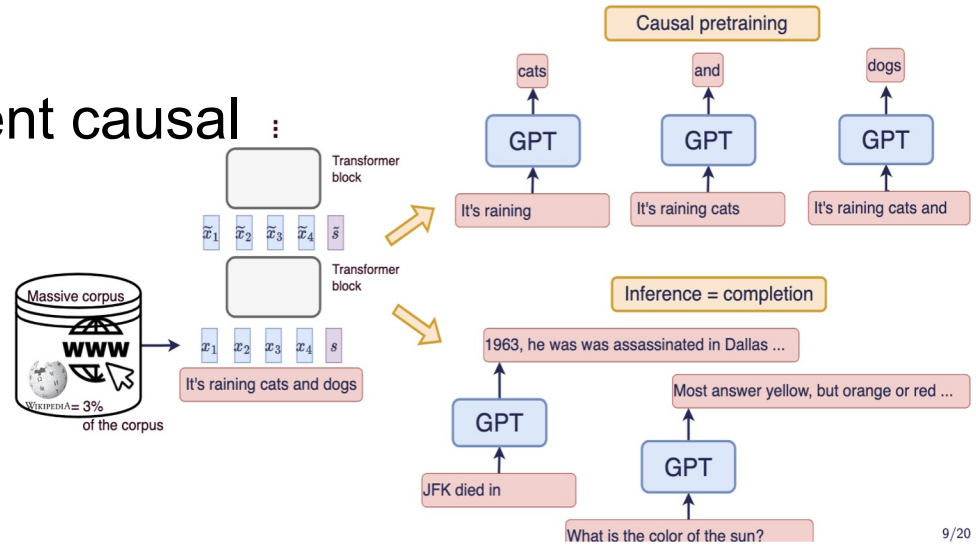
- Génération mot à mot
- Coût très important
- Génération de faisceaux



1. Nouvelle tâche
 - Peu de données
 - Choix de la taille des modèles
2. Modèle de langue
 - Connaissances générales
3. Adaptation pour une tâche
 - Traduction
 - Détection d'entités nommées
 - ...



→ Etape 1: Pré-entraînement causal :

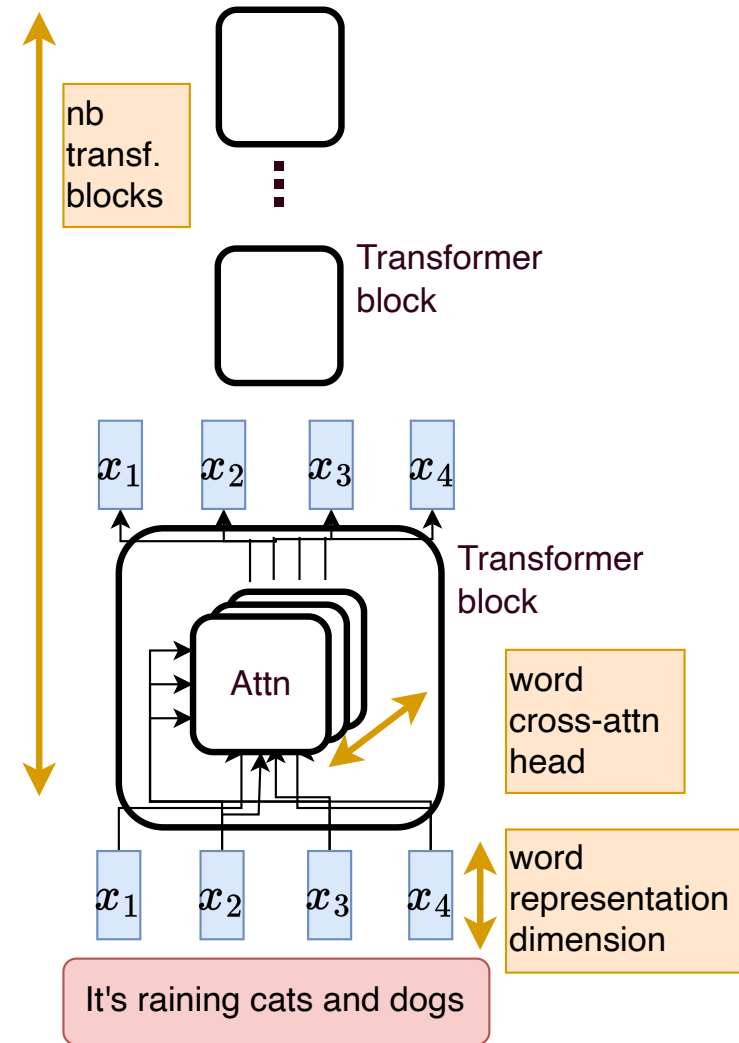


Plus...

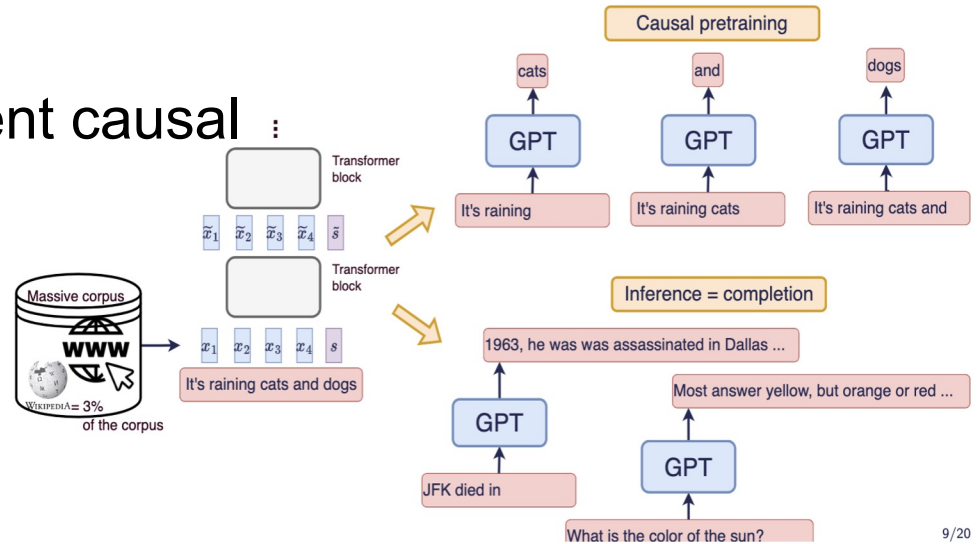
- de mots en entrée [500 => 2k, 32k]
- de dimensions pour les mots [500-2k => 12k]
- de têtes d'attention [12 => 96]
- de blocks/couches [5-12 => 96]

175 Milliards de paramètres... Ca fait quoi?

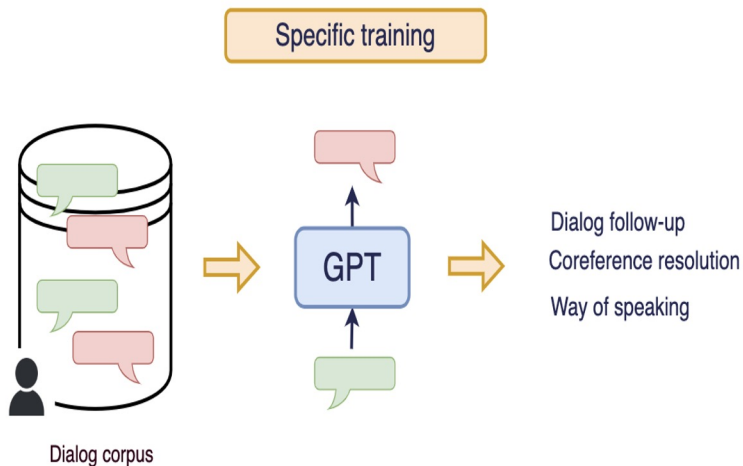
- $1.75 \cdot 10^{11} \Rightarrow 300 \text{ Go} + 100 \text{ Go}$ (stockage des données en inférence) $\approx 400 \text{ Go}$
- GPU NVidia A100 = 80Go de mémoire (=20k€)
- Coût pour (1) entraînement: 4.6 Millions d'€



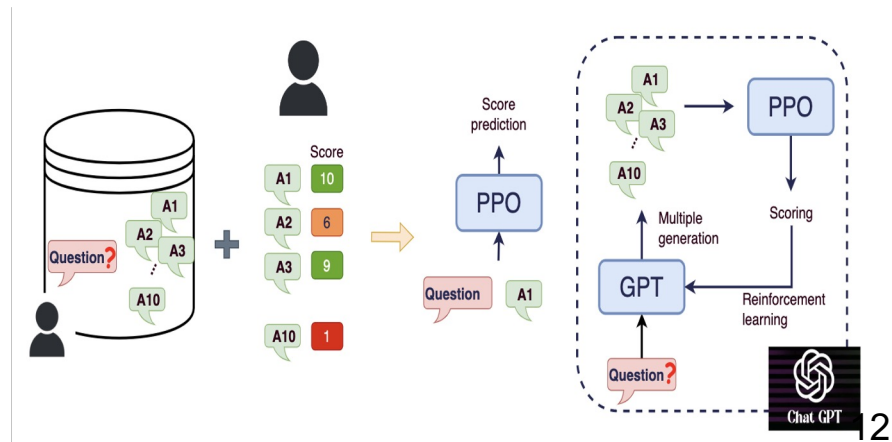
→ Etape 1: Pré-entraînement causal :



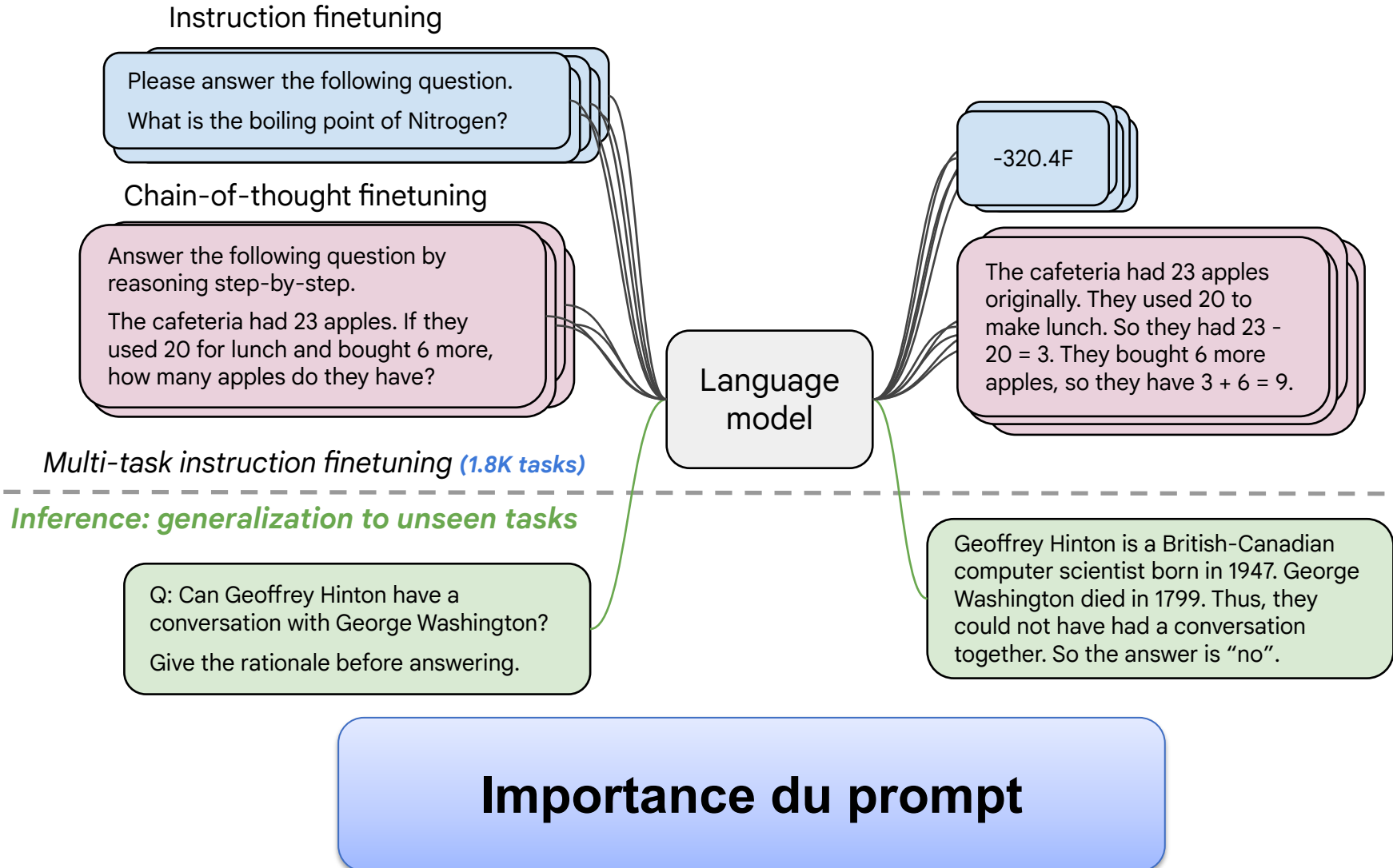
→ Etape 2: Suivi de dialogue



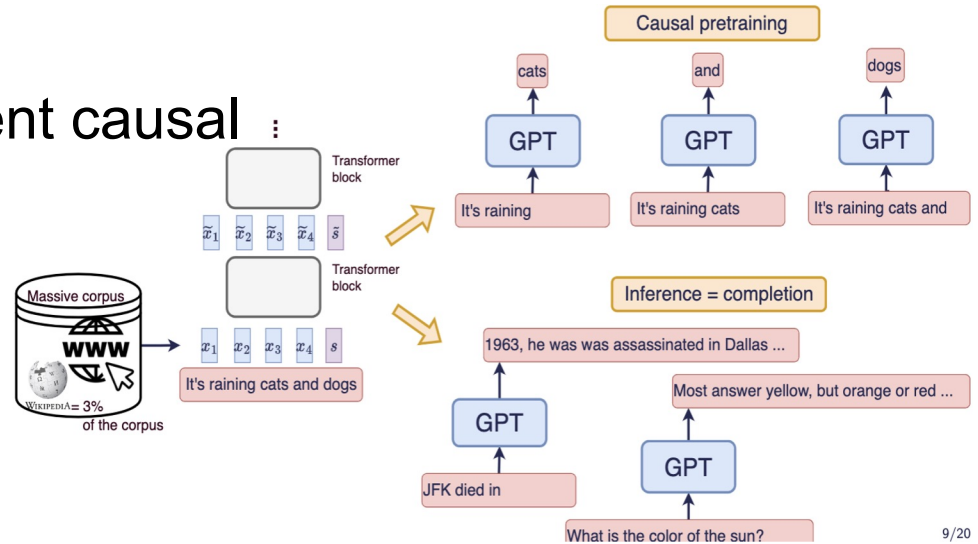
→ Etape 3: Alignement des réponses avec le feedback humain (RLHF)



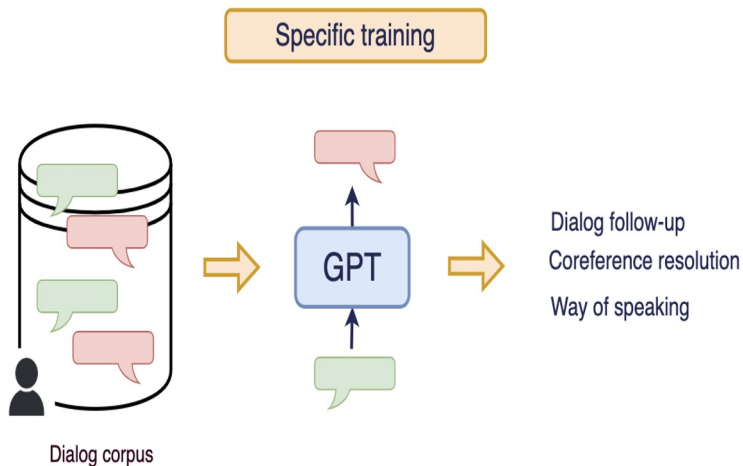
→ Affinage en questions/réponses, raisonnements, ...



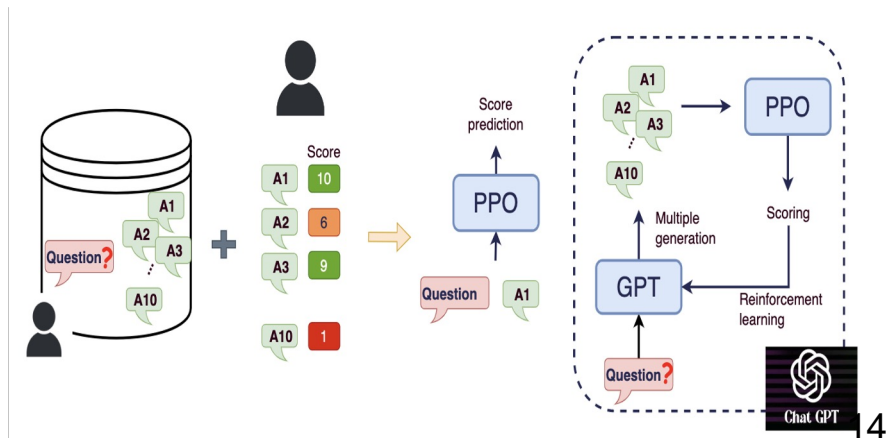
→ Etape 1: Pré-entraînement causal :



→ Etape 2: Suivi de dialogue



→ Etape 3: Alignement des réponses avec le feedback humain (RLHF)



- Apprentissage en contexte / In-context learning
- Mentionner des exemples dans le prompt



→ Manque de **véracité / fiabilité**

→ Vérité vs Vraisemblance -
Génération d'*hallucinations*

→ Incapacité à **s'auto-évaluer**

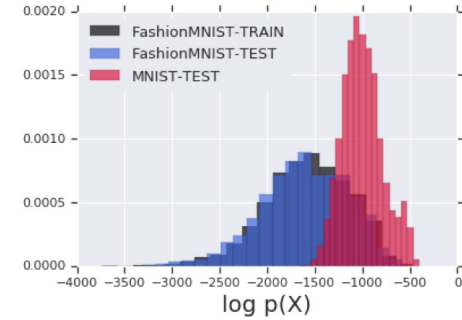
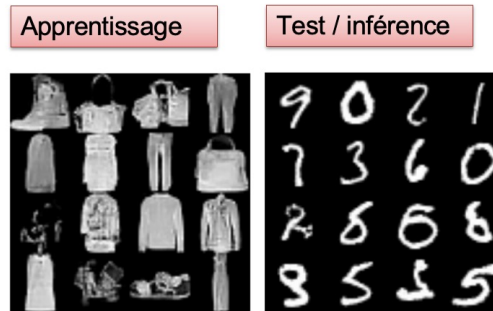
→ **Biais des données**

→ Manque de **stabilité/prédictibilité**

→ *How old is Obama VS how old is Obama?*

→ Manque d'**explicabilité/interprétabilité**

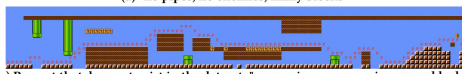
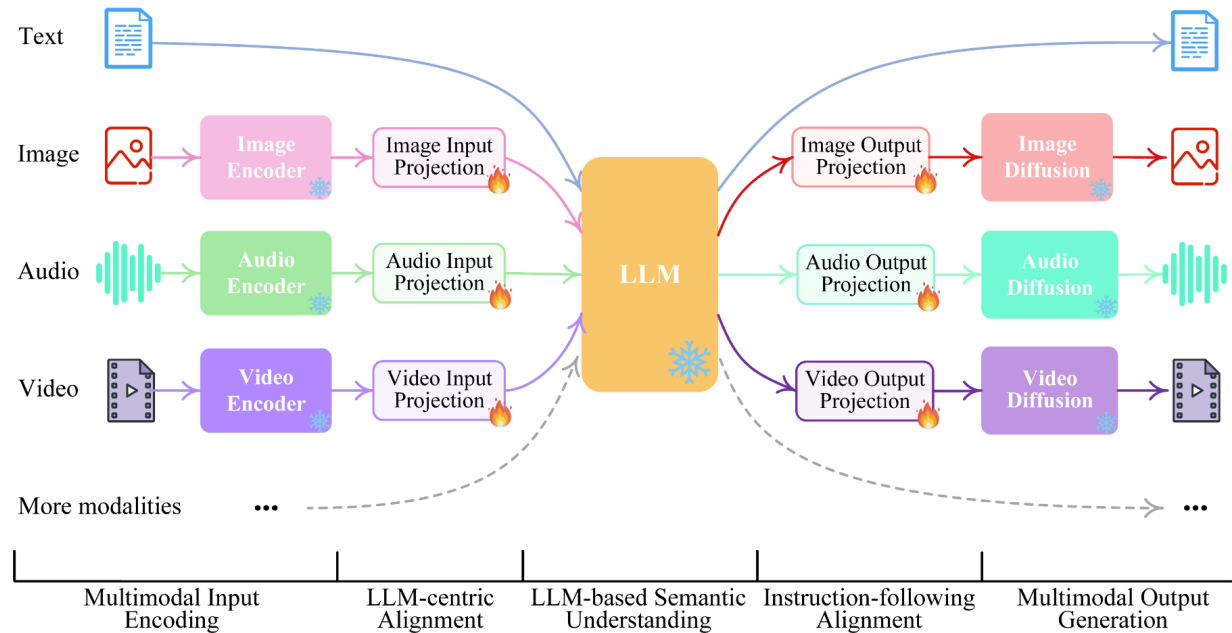
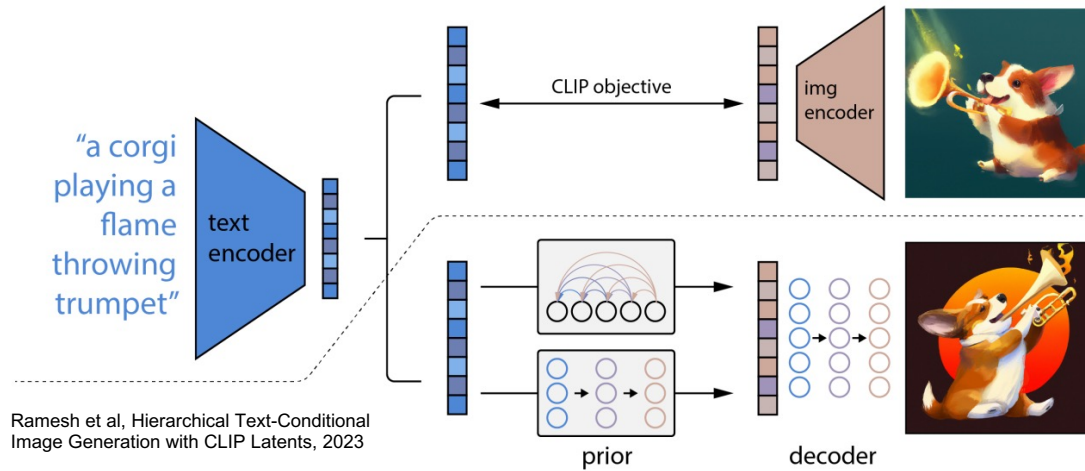
→ Manque de **transparence**



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

Source: 2023 Foundation Model Transparency Index

	Meta	Bing	OpenAI	stability.ai	Google	ANTHROPIC	cohere	Ai2I labs	Inflection	amazon	Average
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	Titan Text	
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%	33%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	



(from) Data-to-text

- Analyser le texte : sélectionner/extraire les informations**

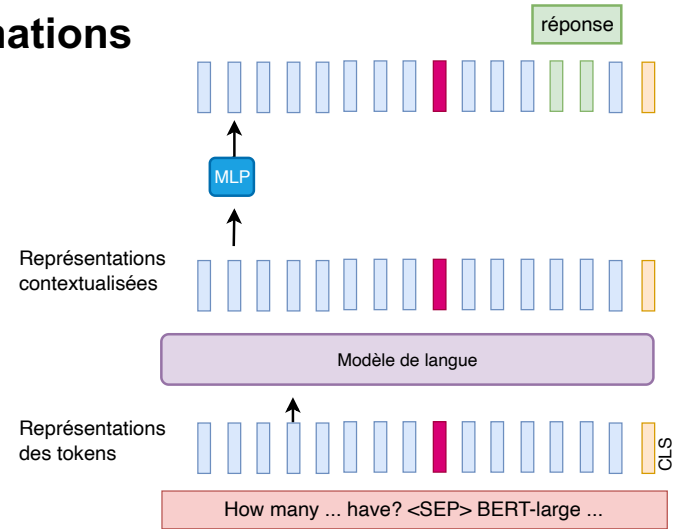
Passage Sentence
 In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question
 What causes precipitation to fall?

Answer Candidate
 gravity

When **Sebastian Thrun** PERSON started at **Google** ORG in **2007** DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** NORP car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** ORG **earlier this week** DATE.

A little **less than a decade later** DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated



- Formater le texte : mise en forme / interfaçage avec l'humain**

María no abofeteó a la bruja verde
 | \ | | X
 Maria did not slap the green witch
 arg1 relation arg2

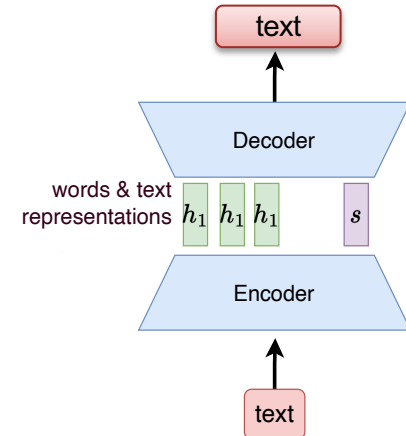
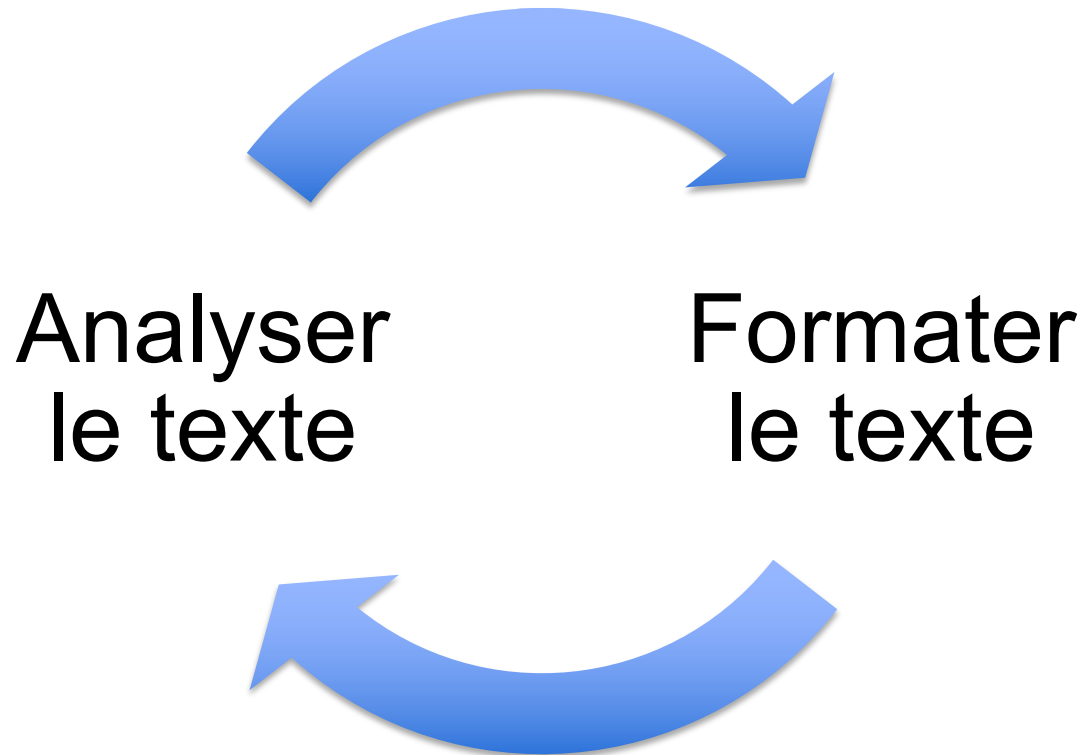


Table Title: Cristhian Stuani
Section Title: International goals

No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0



Les deux faces d'une même pièce ?

Extraire les informations saillantes

Résumé de documents
Extraction d'informations



Analyser



Formater

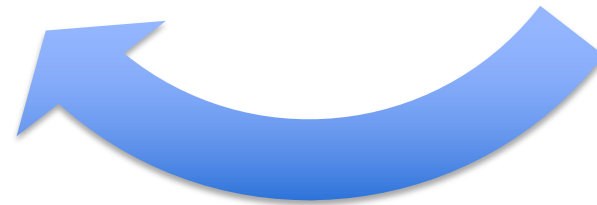
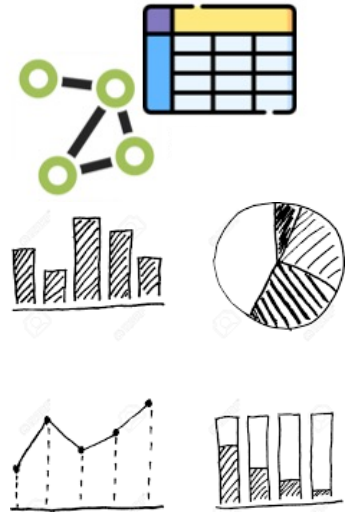


Les deux faces d'une même pièce ?

Résumé de structures complexes



Formater



Rendre les informations lisibles

Les deux faces d'une même pièce ?

Résumé de structures
complexes

Extraire les informations saillantes



Analyser

Formater



Rendre les informations lisibles

Les deux faces d'une même pièce ?

	Fund	Benchmark	Excess		Allocation	Selection	FX rates	CGT Prov	Residual
OFFICIAL PERFORMANCE (net)	2,38%	5,24%	-1,96%						
OFFICIAL PERFORMANCE (gross)	3,48%	5,34%	-1,86%	MANAGEMENT EFFECTS	-0,81%	-1,42%	0,46%	-0,01%	-0,08%
INTERNAL PERFORMANCE	3,10%	5,34%	-2,24%						

Reference Benchmark : MSCI China A, NR
Report Currency : EUR

SECURITY LEVEL - BIGGEST CONTRIBUTORS TO EXCESS RETURN

Company Name	Grouping	Perf	Var(W)	Effect
China Pacific Insu	Financials	13,43	3,15	0,25
Shanxi Lu'an Envir	Energy	14,82	2,12	0,21
Anhui Conch Cement	Materials	13,41	2,56	0,20
China Minsheng Ban	Financials	10,61	3,58	0,18
Jizhong Energy Res	Energy	13,61	2,20	0,18
Jiangsu Yueda Inv	Industrials	11,78	2,24	0,14
Poly Real Estate G	Financials	9,14	3,50	0,13
Xinjiang Ba Yi Iro	Materials	16,69	0,89	0,10
China Nonferrous	Materials	13,84	0,96	0,08
China Baoan Group	Industrials	17,34	0,61	0,07

SECURITY LEVEL - BIGGEST DETRACTORS FROM EXCESS RETURN

Company Name	Grouping	Perf	Var(W)	Effect
Nari Tech Dev Co	Industrials	-6,48	3,26	-0,39
Shandong Denghai S	Consumer Staples	-8,32	1,76	-0,26
Zte Corp	Information Technology	-9,38	1,66	-0,25
Mesnac Co.Ltd	Industrials	-9,17	1,54	-0,22
Yunnan Balyao Grp	Health Care	-6,82	1,70	-0,22
Jiangsu Aoyang Tec	Materials	-14,77	0,49	-0,22
Fujian Septwolves	Consumer Discretionary	-7,86	1,46	-0,20
Tianglin Tasly Phar	Health Care	-9,11	1,05	-0,17
Xi An Aero -Engine	Industrials	-3,10	1,83	-0,16
Ping An Insurance	Financials	13,27	-1,96	-0,15

SECTOR LEVEL - BEST ALLOCATION DECISIONS

Grouping	Var(W)	Segm. Perf	Effect
Health Care	-2,24	-5,53	0,25
Consumer Discretionary	-4,65	3,12	0,11
Energy	2,51	9,23	0,10
Industrials	0,19	5,23	0,01
Utilities	0,29	1,67	-0,01

SECTOR LEVEL - BEST SELECTION DECISIONS

Grouping	PF Perf	BM Perf	Effect
Financials	9,58	8,69	0,22
Information Technology	-1,93	-2,63	0,07
Telecommunication Services	-0,89	-2,01	0,01
Utilities	0,32	1,67	-0,04
Health Care	-7,92	-5,53	-0,08

In January (30/12/2011 to 20/1/2012), Flexifund Equity China A rose in value by **3.48%** compared to a gain of **5.34%** for its index in Euro terms. Both **asset allocation** and **stock selection** detracted from relative performance, as the market focused on oversold or cyclical themes, due to better global risk appetite and more positive economic news.

From a sector allocation perspective, [...]

Materials	7,37	11,15	-0,35
Consumer Discretionary	-2,19	3,12	-0,28
Consumer Staples	-4,59	-3,23	-0,12
Energy	8,22	9,23	-0,10

Toronto Raptors (4-2)

Player	MIN	ORTG	USG%	PTS	FG	3PT	FT	OREB	DREB	TO	AST	BLK	STL	PF
Pascal Siakam F	46	132.0	20.4	26	10-17	3-6	3-4	2	8	2	3	1	1	2
Kawhi Leonard F	41	115.7	23.5	22	7-16	1-5	7-8	1	5	2	3	1	2	4
Marc Gasol C	27	93.3	12.8	3	0-5	0-2	3-4	3	6	1	4	0	0	4
Kyle Lowry G	42	137.3	23.2	26	9-16	4-7	4-6	2	5	3	10	0	3	5
Danny Green G	18	-	2.5	0	0-0	0-0	0-0	0	1	1	3	0	1	1
Fred VanVleet G	34	132.5	22.7	22	6-14	5-11	5-5	1	1	1	0	0	1	1
Serge Ibaka C	22	118.7	28.5	15	7-12	0-1	1-2	2	1	1	2	0	0	4
Norman Powell G	11	-	12.4	0	0-2	0-1	0-0	0	1	1	0	0	0	2
Game Total	240	-	-	114	39-82	13-33	23-29	11	28	12	25	2	8	23

Golden State Warriors (2-4)

Player	MIN	ORTG	USG%	PTS	FG	3PT	FT	OREB	DREB	TO	AST	BLK	STL	PF
Draymond Green F	44	92.9	19.0	11	5-10	1-4	0-2	4	15	8	13	2	3	4
Andre Iguodala F	32	117.1	25.0	22	9-15	3-6	1-5	0	2	1	2	1	0	3
Kevon Looney C	27	109.4	13.2	6	3-7	0-0	0-0	2	1	1	4	1	1	2
Stephen Curry G	42	113.3	23.7	21	6-17	3-11	6-6	1	2	3	7	1	2	3
Klay Thompson G	32	160.6	25.0	30	8-12	4-6	10-10	1	4	2	0	0	2	3
DeMarcus Cousins C	19	117.3	27.6	12	4-9	0-1	4-7							
Shaun Livingston G	16	92.0	16.7	6	3-5	0-0	0-0							
Quinn Cook G	13	73.2	10.3	2	1-3	0-2	0-0							
Alfonzo McKinnie F	10	-	4.4	0	0-1	0-1	0-0							
Andrew Bogut C	3	-	-	0	0-1	0-0	0-0							
Jonas Jerebko F	2	-	-	0	0-0	0-0	0-0							
Game Total	240	-	-	110	39-80	11-31	21-30							

The **Toronto Raptors** defeated the host **Golden State** Warrior, **114-110**, in Game 6 of the NBA Finals at ORACLE Arena on Thursday. [...]

The **Raptors (4-2)** were lead by **Kyle Lowry**, as he accrued **26 points, seven rebounds, 10 assists and three steals**. [...]

[...]

A Hierarchical Model for Data Text Generation 1	
Type	Long Paper
Length	12 pages
Authors	Clement Rebuffel; Laure Soulier; Geoffrey Scoutheeten; Patrick Gallinari 2
Published	14 April 2020 4
Conference	42nd European Conference on Information Retrieval 3

A hierarchical Model for Data-To-Text Generation (Rebuffel et al.) will be published at ECIR 2020

Content selection

Macro-planning

Surface Realisation

Sentence aggregation

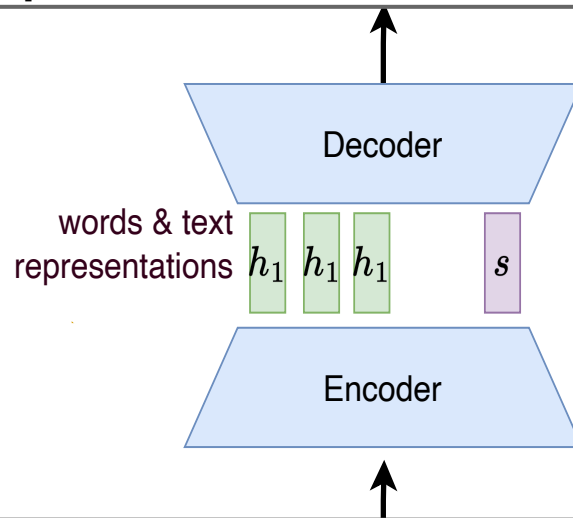
Data abstraction/interpretation

Data-to-text : générer du texte à partir de données structurées

The **Toronto Raptors** defeated the host **Golden State** Warrior, **114-110**, in Game 6 of the NBA Finals at ORACLE Arena on Thursday. [...]

The **Raptors** (4-2) were lead by **Kyle Lowery**, as he accrued **26 points, seven rebounds, 10 assists and three steals**. [...]

[...]



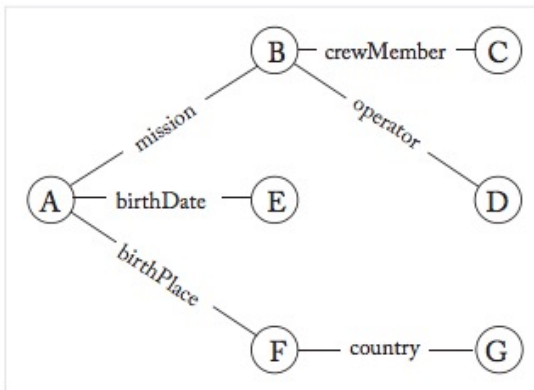
$$P(R|T, \theta) = \prod_{i=1}^l P(y_i | y_{<i}, T, \theta)$$

Toronto Raptors (4-2)

Player	MIN	ORTG	USG%	PTS	FG	3PT	FT	OREB	DREB	TO	AST	BLK	STL	PF
Pascal Siakam F	46	132.0	20.4	26	10-17	3-6	3-4	2	8	2	3	1	1	2
Kawhi Leonard F	41	115.7	23.5	22	7-16	1-5	7-8	1	5	2	3	1	2	4
Marc Gasol C	27	93.3	12.8	3	0-5	0-2	3-4	3	6	1	4	0	0	4
Kyle Lowry G	42	137.3	23.2	26	9-16	4-7	4-6	2	5	3	10	0	3	5
Danny Green G	18	-	2.5	0	0-0	0-0	0-0	0	1	1	3	0	1	1
Fred VanVleet G	34	132.5	22.7	22	6-14	5-11	5-5	1	1	1	0	0	1	1
Serge Ibaka C	22	118.7	28.5	15	7-12	0-1	1-2	2	1	1	2	0	0	4
Norman Powell G	11	-	12.4	0	0-2	0-1	0-0	0	1	1	0	0	0	2
Game Total	240	-	-	114	39-82	13-33	23-29	11	28	12	25	2	8	23

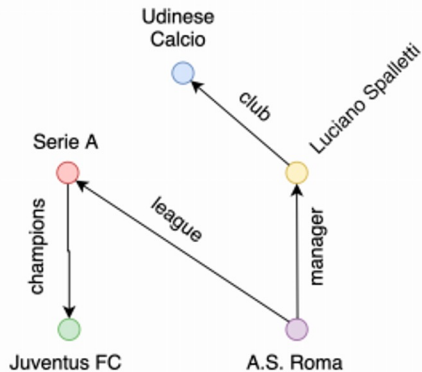
Country	PTS
Germany	2
Argentina	0

[(0, Country, Germany), (0, PTS, 2), (1, Country, Argentina), (1, PTS, 0)]



Input: *A (birthPlace F (country G)) (birthDate E) (mission B (operator D) (crewMember C))*

Encodage des données structurées : linéarisation



```
<S> Serie A <P> champions <O> Juventus F.C.
<S> Luciano Spalletti <P> club <O> Udinese
Calcio <S> A.S. Roma <P> manager <O> Luciano
Spalletti <S> A.S. Roma <P> league <O> Serie A
```

AS Roma play in the Serie A league where Juventus FC are the champions. Their manager is Luciano Spalletti who has been associated with Udinese Calcio.

Domain	train
Inform	arrive_by : 11:51
Request	num_people

```
train inform arriveby = 11:51 | train request
people = ?
```

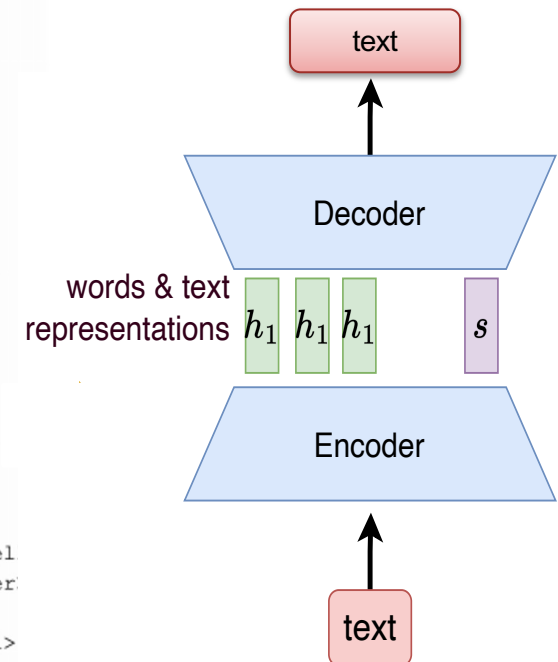
The closest arrival time i can give you is 11:51 , is that ok ? And how many tickets would you like ?

Table Title: Cristhian Stuani
Section Title: International goals

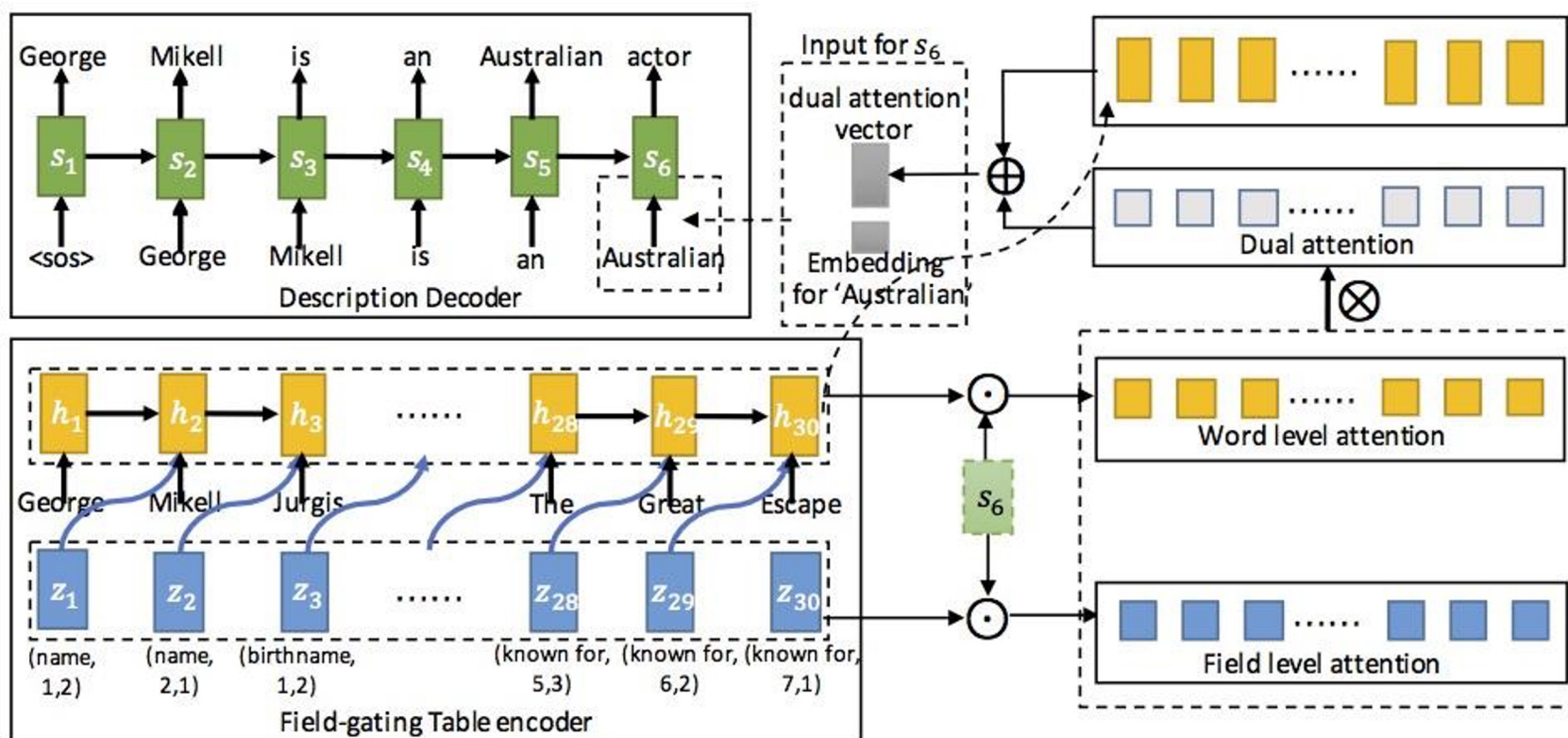
No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0

```
<page_title> Cristhian Stuani </page_title>
<section_title> International goals </section_title>
<table> <cell> 2. <col_header> No. </col_header> </cel
<cell> 13 November 2013 <col_header> Date </col_header
</cell> <cell> Amman International Stadium, Amman,
Jordan <col_header> Venue </col_header> </cell> <cell>
Jordan <col_header> Opponent </col_header> </cell>
<cell> 5-0 <col_header> Result </col_header> </cell>
</table>
```

On 13 November 2013 Cristhian Stuani netted the second in a 5-0 win in Jordan.



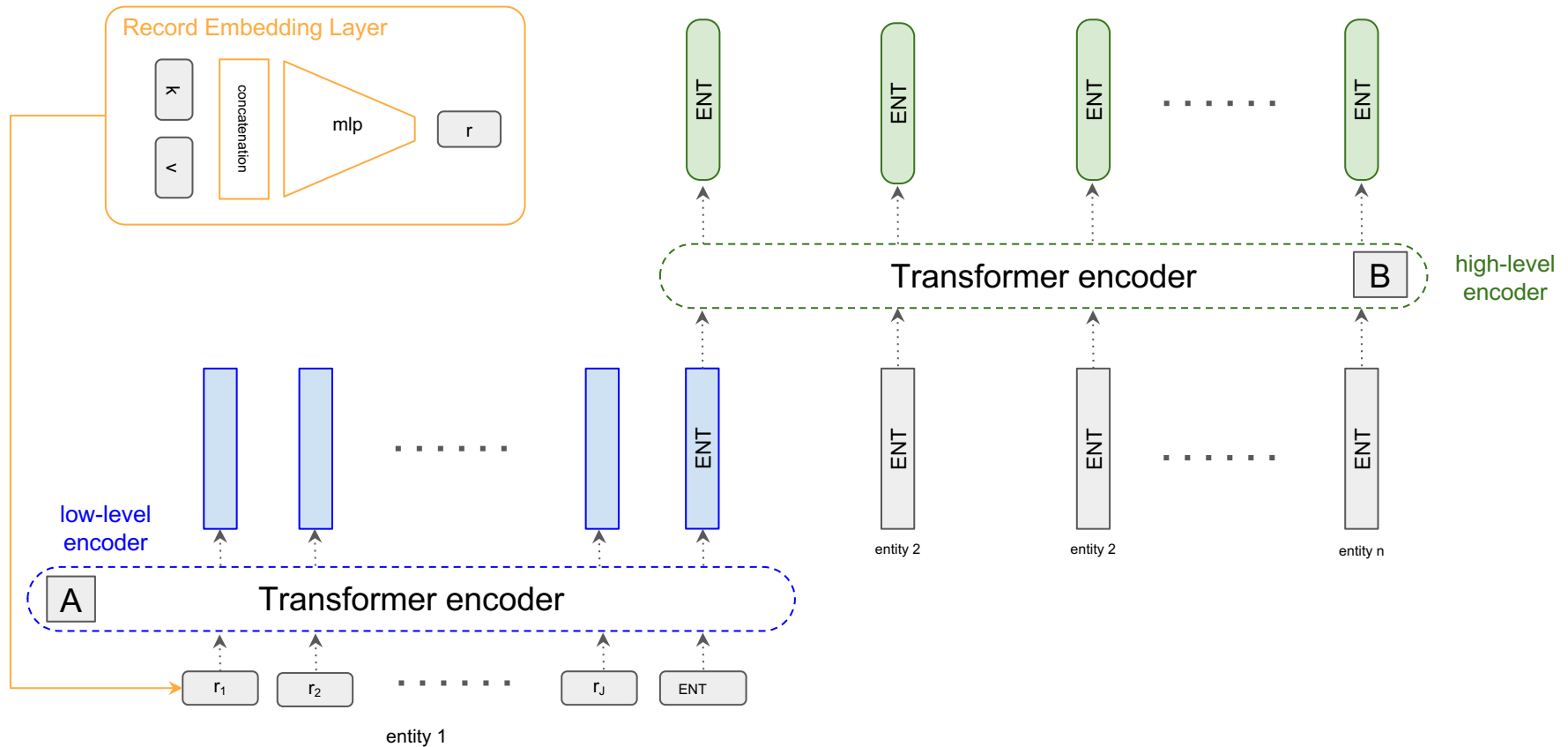
- Attention duale :
 - Scores d'attention sur les mots et les champs de la table



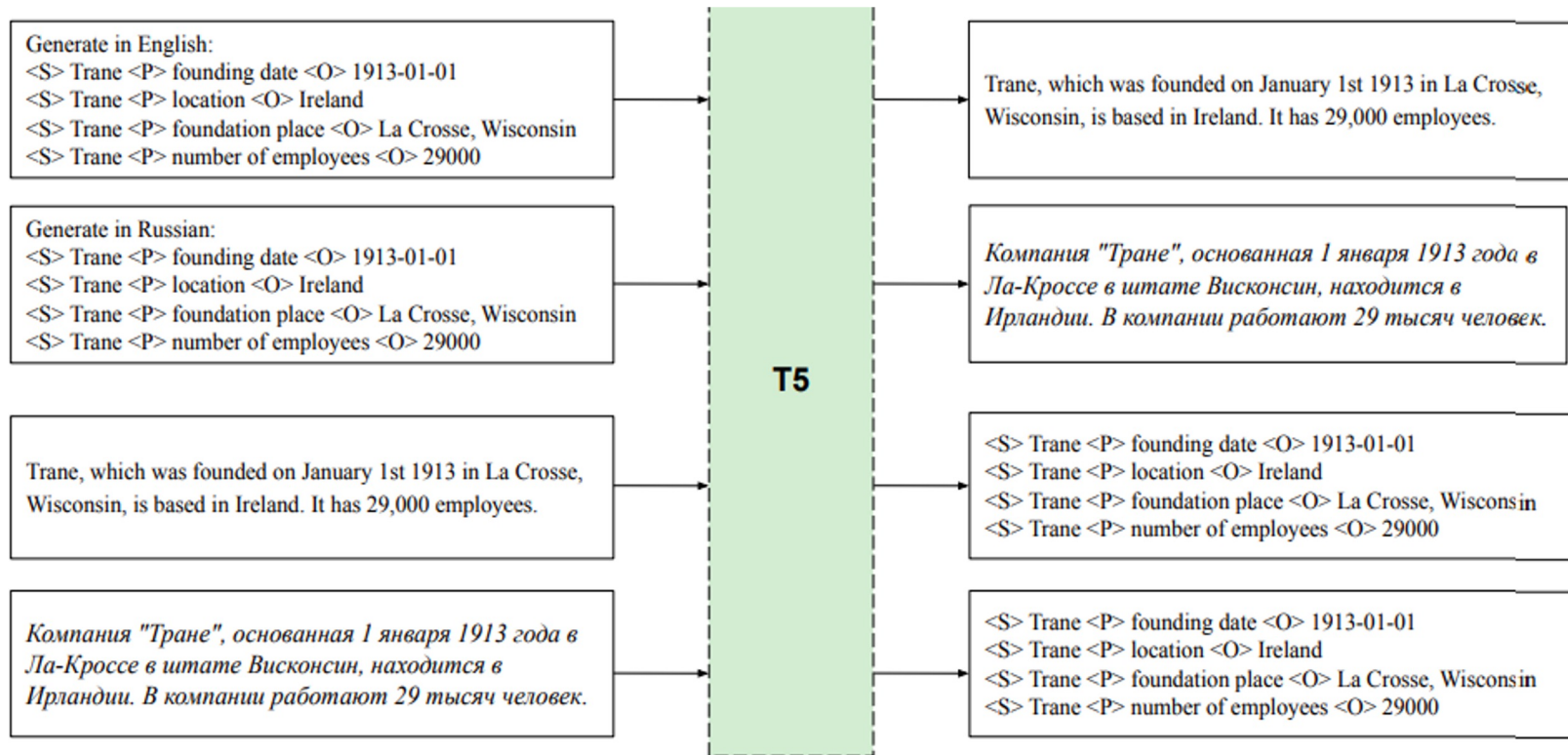
Encodage basé sur la structure

[Rebuffel et al., ECIR 2021]

- Encodage hiérarchique :
 - Encodage d'une case du tableau
 - Encodage d'une ligne du tableau (token [ENT])
 - Encodage du tableau



- Passage D2T et T2D possible avec le même modèle
 - T5 pré-entraîné + affiné sur la tâche
 - Sérialisation des triplets : <S> Sujet <R> Relation <O> Object
 - Concaténation des triplets dans l'ordre d'apparition



Génération en DTT: cas des omissions & des hallucinations

- Deux catégories de pathologies en DTT

Attribute	Value
Birthplace	<i>Utah, America</i>
Position	<i>forward (soccer player)</i>

Omissions

Contenu attendu mais manquant dans le texte généré

A soccer player, who plays as a forward.

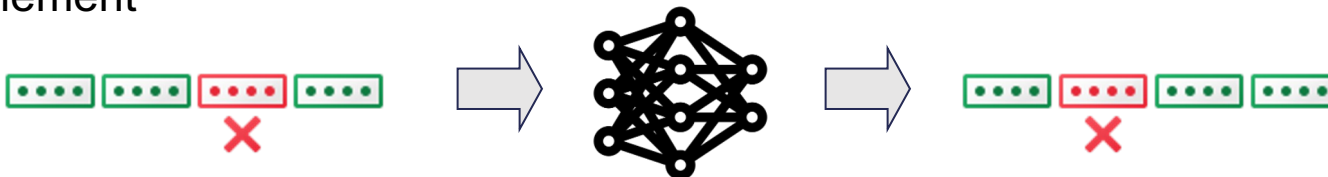
Hallucinations

- Texte généré contient du texte divergent de la table
- Textes de références souvent divergents

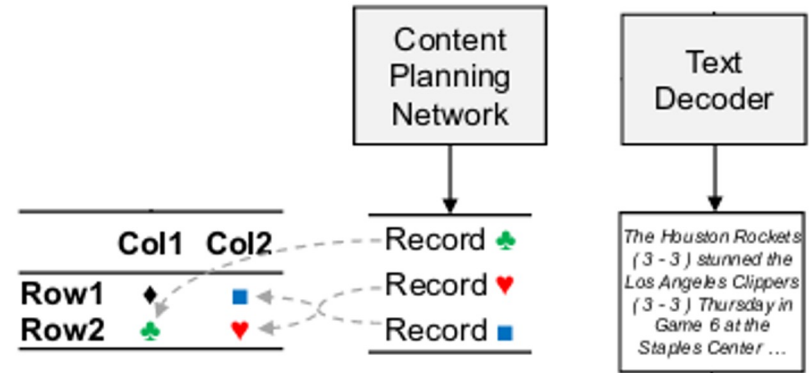
A Utah forward, from the national team.

- Explication :

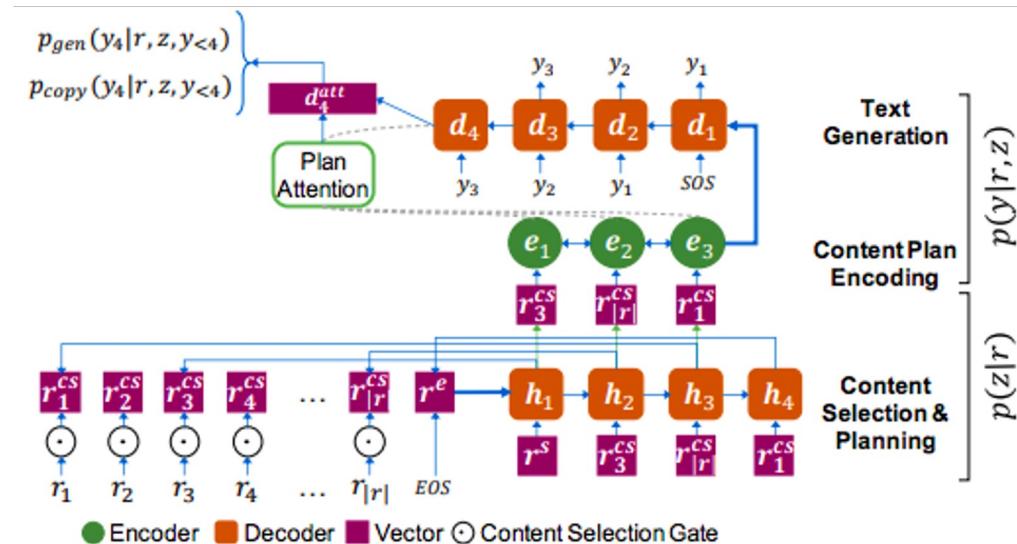
Procédure d'apprentissage force le modèle à copier les « comportements » des données d'entraînement



- Deux étapes de décodage
 - Génération d'un plan (séquence ordonnée de clés-valeurs)
 - Génération du texte final



Variante : génération séquentielle d'un élément du plan et de la phrase associée

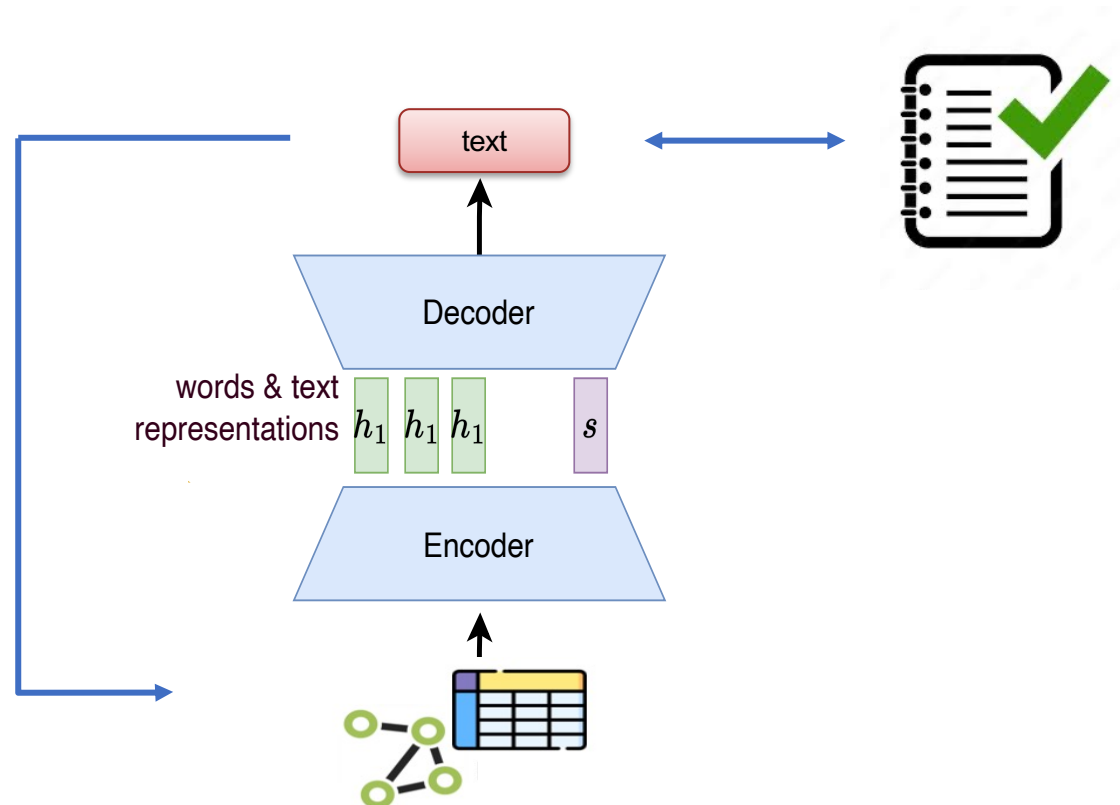


- Injecter une métrique dans l'apprentissage du modèle

RL loss based on PARENT

Cross-Entropy loss

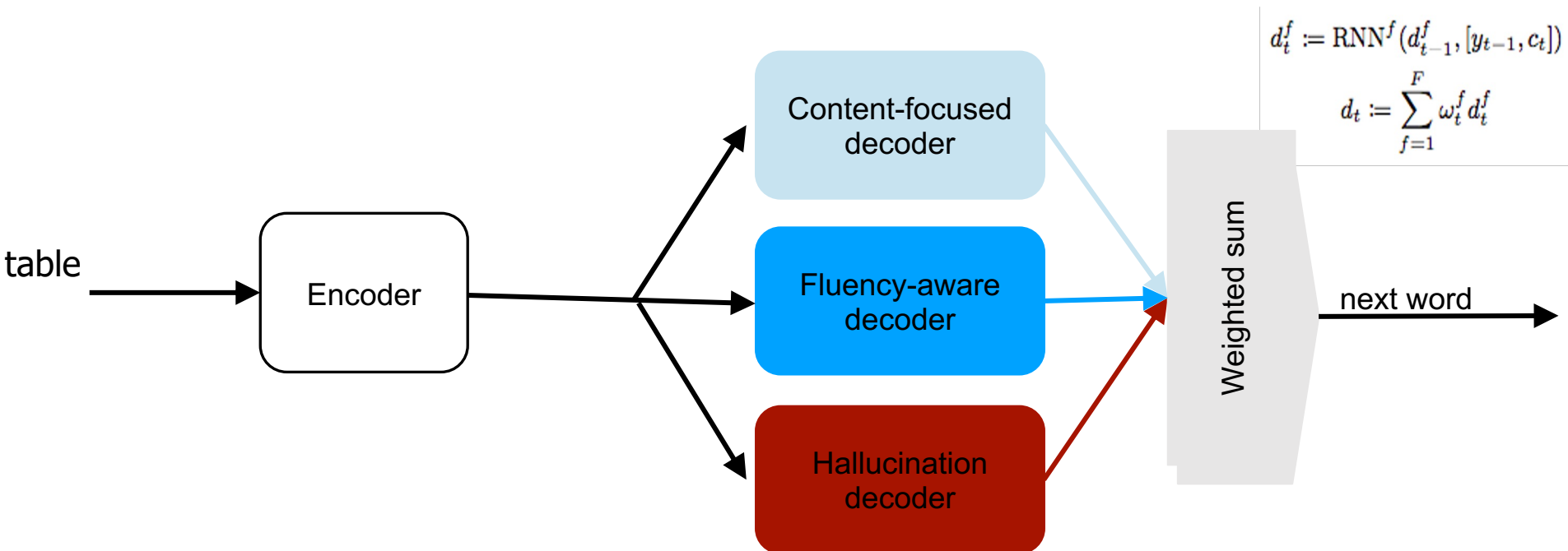
$$\mathcal{L} := \gamma \mathcal{L}_{rl} + (1 - \gamma) \mathcal{L}_{ml}$$



PARENT : métrique d'appariement entre le texte généré et les données structurées (n-grams, co-occurrences)

- Intuitions

- Un mot : associé soit à du contenu, soit à de l'hallucination
- Pour chaque mot, on peut contrôler différents facteurs via une architecture multi-branches



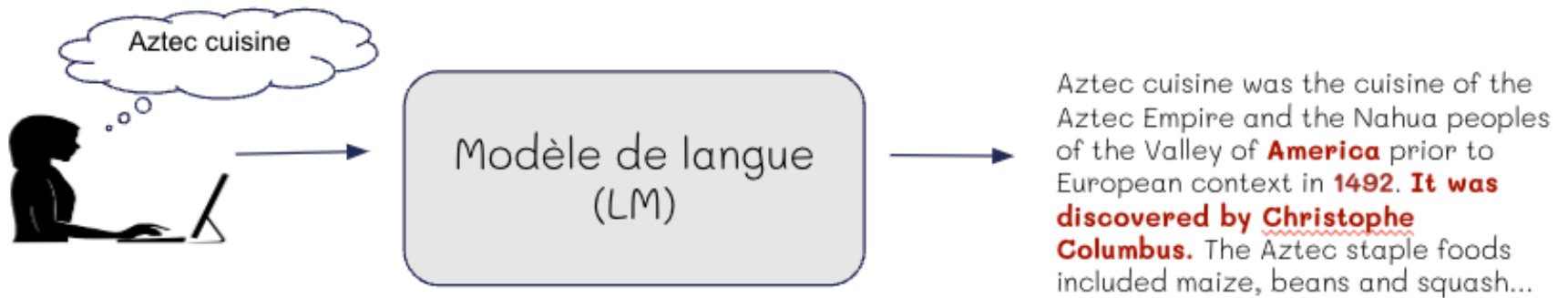
- **Entraînement** : chaque mot est associé à la branche correcte (contenu ou hallucination)
- **Génération** : l'humain contrôle le niveau des facteurs

- Construire des données bruitées
 - Hallucination (texte non référencé)
 - Absence de factualité sur des valeurs
 - Ajouts d'attributs

Donnée structurée (entrée)	Texte attendu	Texte bruité
<H> AMC_Matador <R> bodyStyle <T> Coupé	The AMC Matador's body style is Coupé	The Alc Matador is a Spanish bourgeois coupe .
name[Clowns] eatType[coffee shop] food[Fast food] customer rating[high] area[riverside] near[Clare Hall]	Clowns is a coffee shop which offers fast food and has high customer ratings , and may be found near Clare Hall in the riverside area	Cats and coffee shop , and the fast food place, Clowns , is located near Clare Hall . It is in the riverside area. It has a high customer rating .

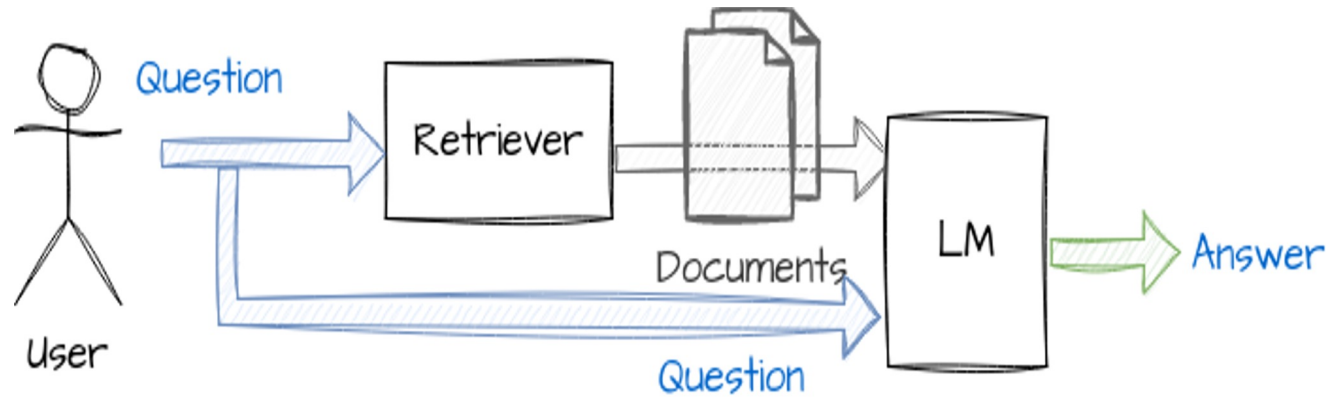
- Apprentissage contrastif
 - Maximiser la probabilité du texte attendu
 - Minimiser la probabilité du texte bruité

Augmenter les modèles de langue
avec des outils externes
pour améliorer la factualité



- Génération dépendante de la formulation de la requête
- Génération dépendante des connaissances paramétriques du modèle

Génération assistée par la recherche d'information (RAG : retrieval-augmented generation)



Retrieval-Augmented Generation (RAG) [1]

Améliorer les connaissances en fournissant des documents en entrée

Retrieval-Augmented Language Model (REALM) [2]

Intégrer l'ordonnancement dans le pré-entraînement du LLM

Retrieval-Enhanced Transformer (RETRO) [3]

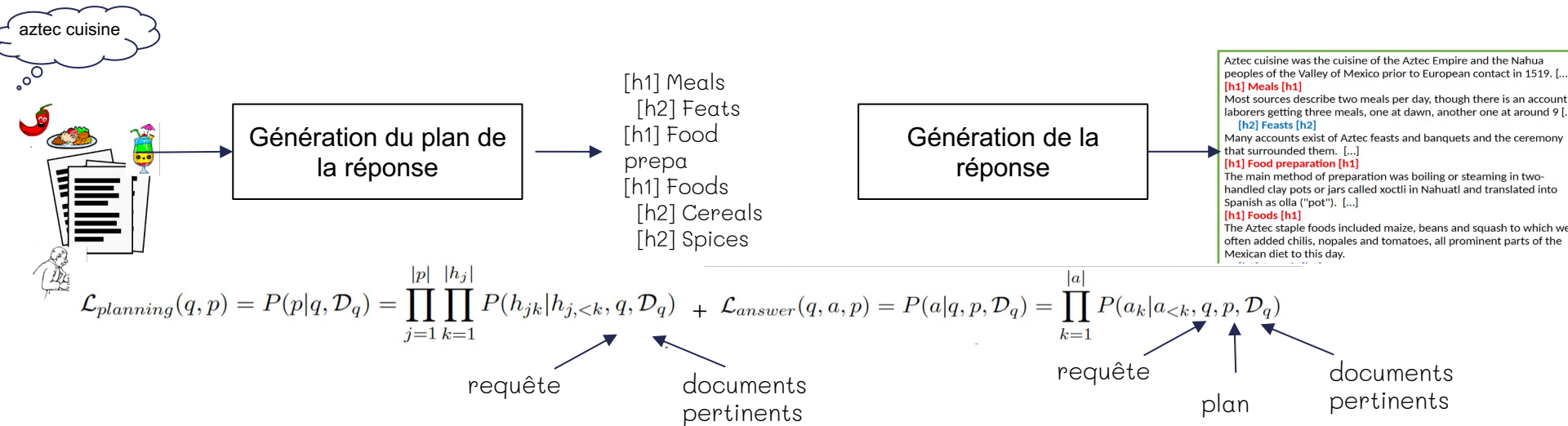
Scalabilité à un grand nombre de documents (chunk-based attention)

- **Comment contraindre le LLM à extraire le contenu des documents ?**
- **Est-ce nécessaire d'utiliser le moteur de recherche à chaque question ?**

Planification de la structure de la réponse (~ planification du Data-to-text)

Etape 1 : Génération d'un plan intermédiaire à partir des documents

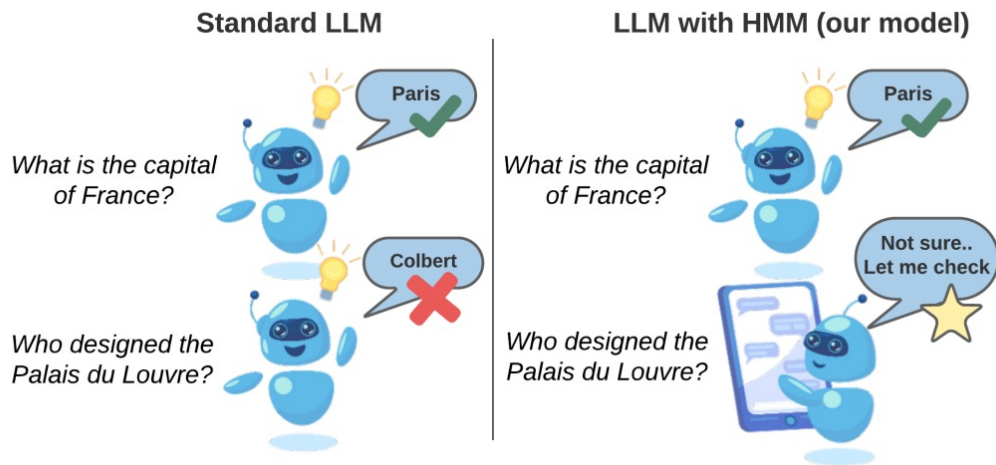
Etape 2 : Génération de la réponse à partir du plan



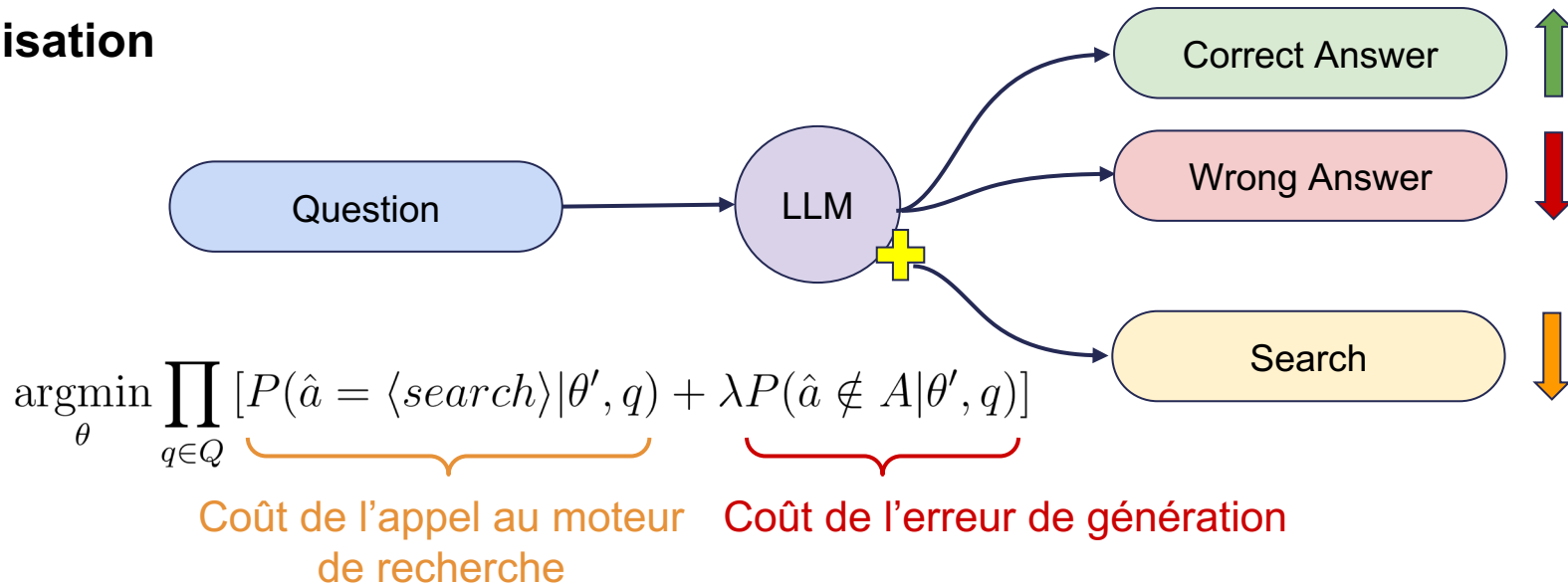
Objectif

Apprendre à générer le token

<SEARCH> lorsque cela est nécessaire



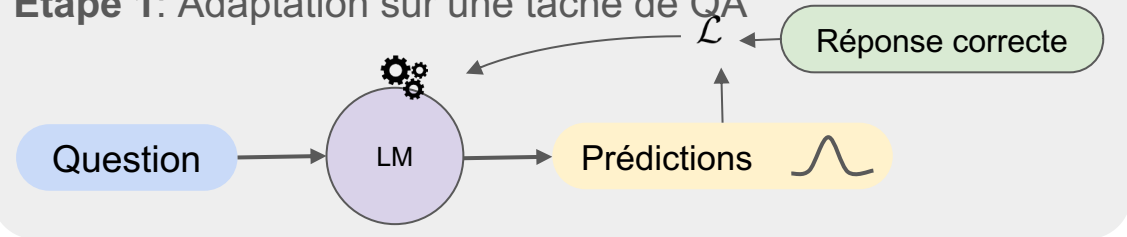
Formalisation



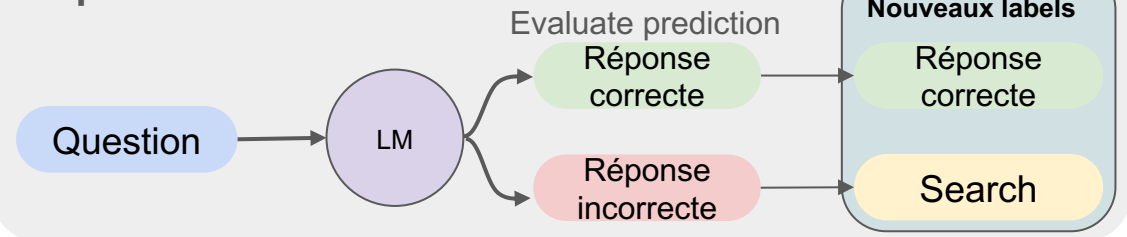
Apprendre une fonction de filtrage qui :

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

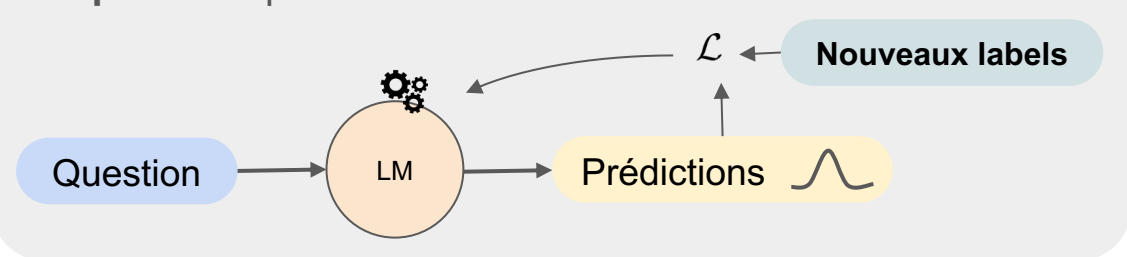
Etape 1: Adaptation sur une tâche de QA



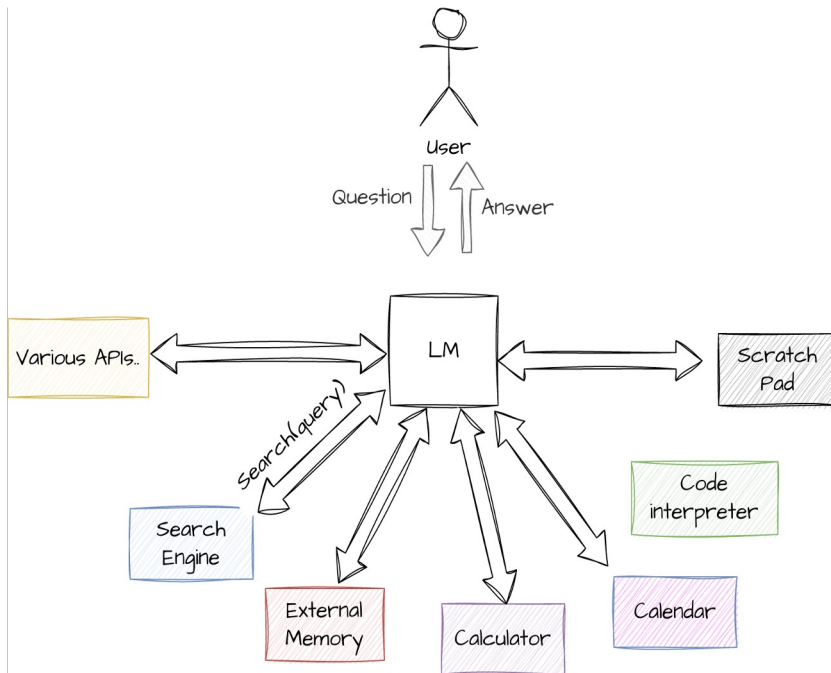
Etape 2: Nouveau label "Search"



Etape 3: Adaptation du modèle aux nouveaux labels



Apprendre aux LLMs à appeler 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.

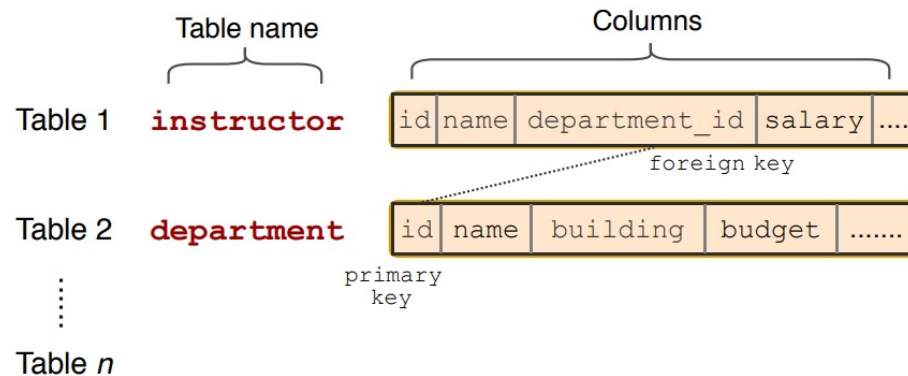
The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

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.

Les bases de données parmi les outils externes possibles

- Générer des requêtes SQL à partir de questions

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
```

Meidum

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.

Les bases de données parmi les outils externes possibles

- Générer plusieurs requêtes SQL
- Estimer la probabilité du programme en fonction de la tâche et de l'exécution
- Trier les requêtes SQL

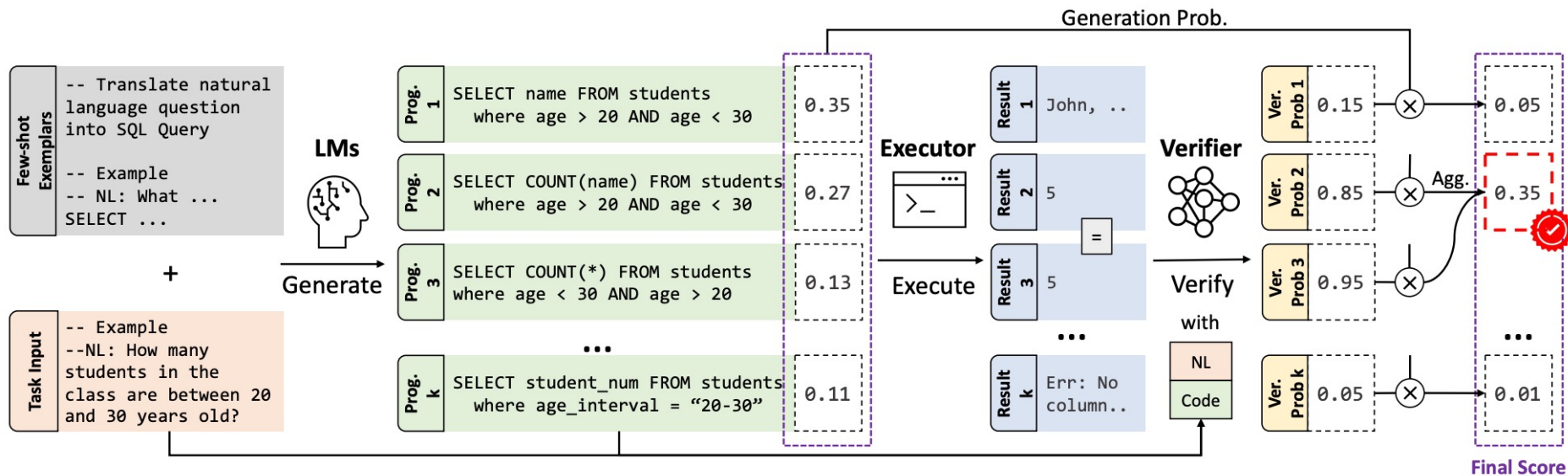
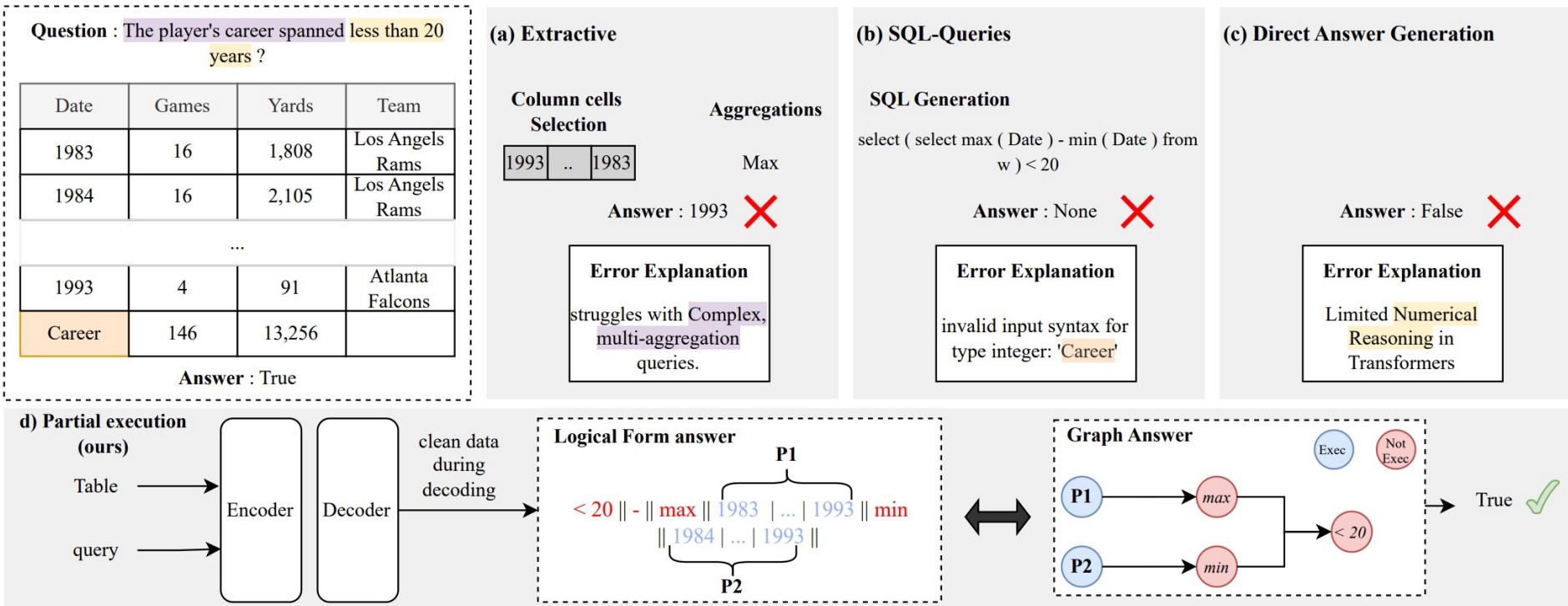


Figure 1: The illustration of LEVER using text-to-SQL as an example. It consists of three steps: 1) *Generation*: sample programs from code LLMs based on the task input and few-shot exemplars; 2) *Execution*: obtain the execution results with program executors; 3) *Verification*: using a learned verifier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

Les bases de données parmi les outils externes possibles

- Le LLM génère la réponse à partir de la requête SQL
- Décomposition de la requête
- « Exécution hybride » : 1) auto-évaluation de la capacité de répondre sur des parties du graphe, 2) exécution SQL des parties non répondues



Conclusion

- Représentation des connaissances
 - Limite des LLMs par rapport aux connaissances paramétriques
 - Besoin de spécialiser/contrôler les LLMs pour une tâche dédiée
 - **Emergence de la construction de systèmes de systèmes ?**

- Génération texte : champ applicatif énorme
 - De nombreuses tâches sont abordées avec des LLMs
 - **Mais auto-évaluation encore difficile**

Merci de votre attention
