

# SALON DE LA DATA ET DE L'IA

HYBRIDE

PRESENTIEL & VIRTUEL

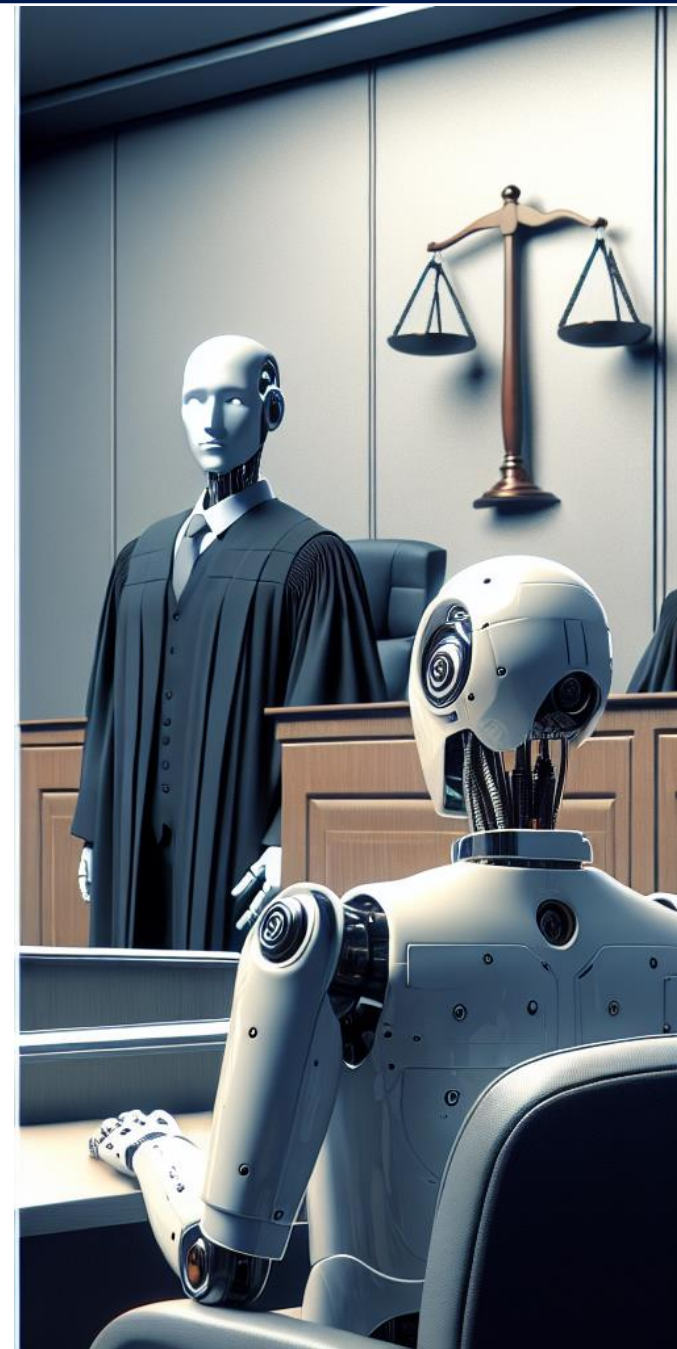
$\frac{17}{09}$  2024



100% BÉNÉVOLE  
GRATUIT

# Evaluation : les LLM sont-ils juges et parties ?

Ou comment mesurer la qualité des réponses apportées  
par un modèle de langage ?





Paul Peton

## NOTRE MÉTIER

MDW France, est un cabinet de conseil spécialisé en la gestion de la **relation clients** et collaborateurs ainsi que la **data** et **intelligence artificielle**.

## NOUS REJOINDRE



## NOS PARTENAIRES



Microsoft

Profisee

SUPERMETRICS

databricks



# L'édition 2024, c'est grâce à vous ! Merci !

## NOS SPONSORS EN 2024



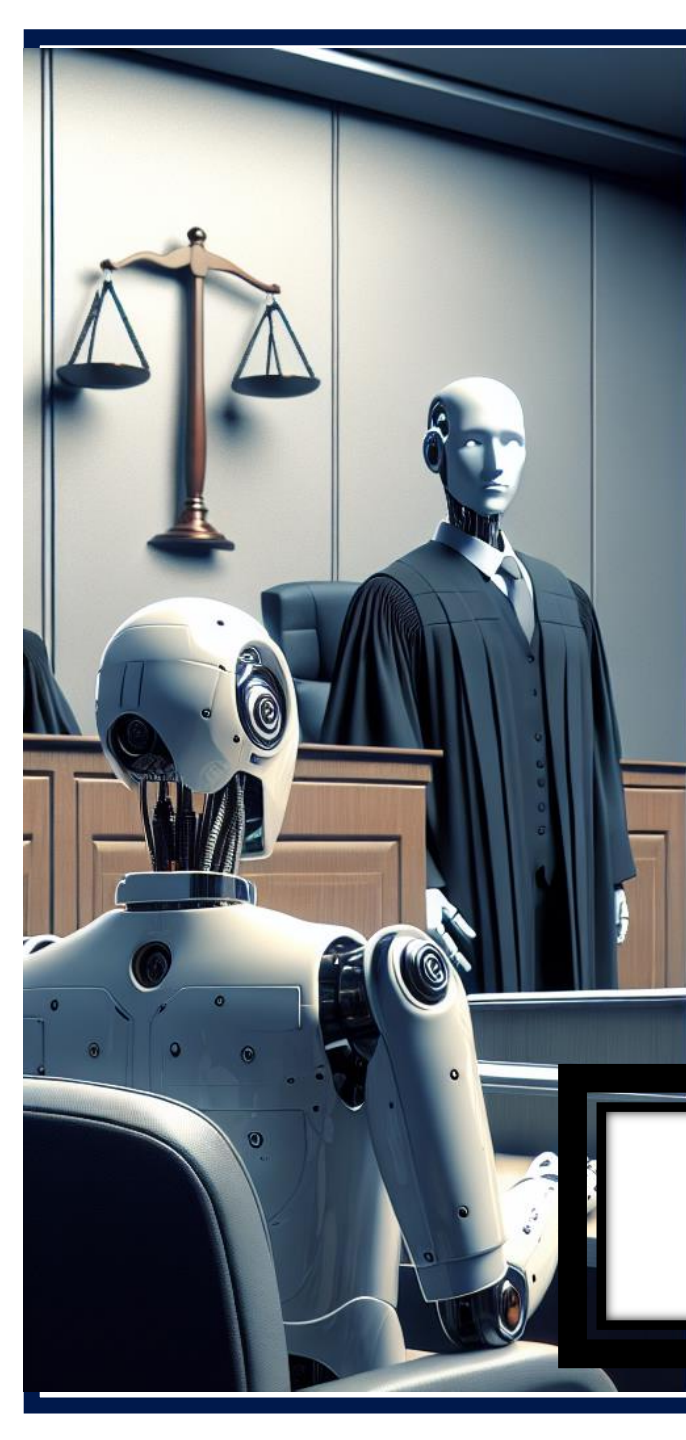
## NOS ORGANISATEURS



## NOS PARTENAIRES





A futuristic courtroom scene. In the foreground, a white robot with a spherical head and articulated arms is seated at a desk. In the background, a humanoid robot in a black judicial robe stands behind a wooden bench. On the wall behind the judge is a large brass scale of justice.

Qui a remporté les élections européennes en France en 2024 ?



Désolé, je n'ai pas d'informations sur les résultats de cette élection.

Comment rechercher cette information en ligne ?



Pour rechercher les résultats des élections européennes en France en 2024, vous pouvez suivre ces étapes :

1. **Utiliser un moteur de recherche** : Rendez-vous sur Google ou un autre moteur de recherche de votre choix.
2. **Formuler votre requête** : Tapez des requêtes telles que "résultats élections européennes France 2024" ou "qui a gagné les élections européennes en France 2024".



La date la plus récente de mes connaissances est octobre 2023. Si vous avez besoin d'informations actualisées au-delà de cette date, je peux vous aider à les chercher en ligne.

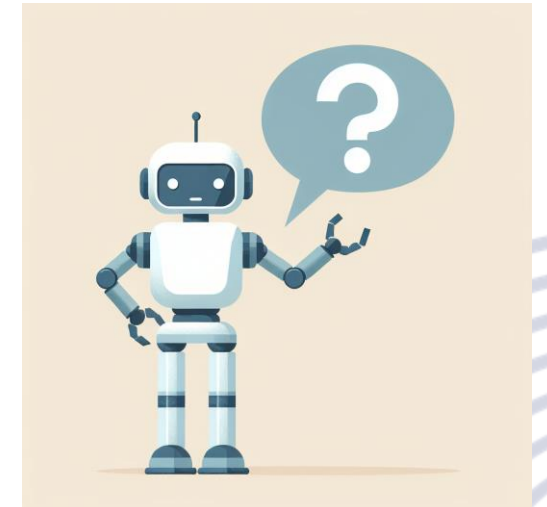
## Deux questions à vous poser (à vous, simples humains)

Qu'est-ce qui est vrai pour une supernova de type Ia ?

- A. Ce type de supernova se produit dans des systèmes binaires.
- B. Ce type de supernova se produit dans les jeunes galaxies.
- C. Ce type de système produit des sursauts gamma.
- D. Ce type de galaxie produit de grandes quantités de rayons X.

Lorsque l'élasticité croisée de la demande est négative, les biens en question sont nécessairement...

- A. normaux
- B. inférieurs
- C. complémentaires
- D. des substituts



## Deux questions vite répondues (par GPT-3.5 version 0301)

Qu'est-ce qui est vrai pour une supernova de type Ia ?

- A. Ce type de supernova se produit dans des systèmes binaires.

Le premier énoncé est vrai pour une supernova de type Ia, car elle se produit dans des systèmes binaires. Cependant, les autres énoncés sont faux. Les supernovas de type Ia peuvent se produire dans des galaxies de tous âges, et elles ne produisent pas de sursauts gamma ni de grandes quantités de rayons X.

- B. Ce type de galaxie produit de grandes quantités de rayons X.

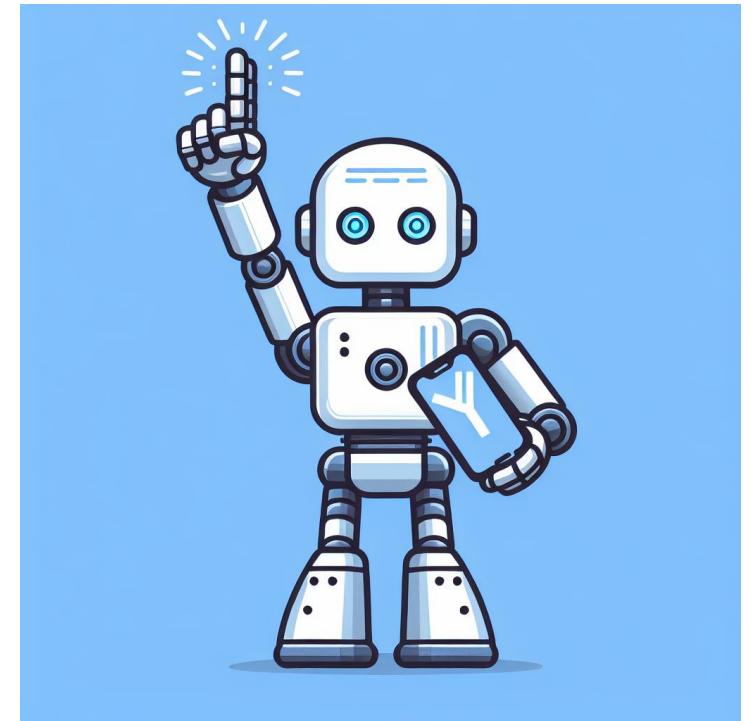
Lorsque l'élasticité croisée de la demande est négative, les biens en question sont nécessairement...

- A. normaux  
B. inférieurs  
C. complémentaires

Lorsque l'élasticité croisée de la demande est négative, cela signifie que la demande pour un bien est inversement proportionnelle à la variation du prix d'un autre bien. Par conséquent, les biens en question sont des biens complémentaires.

# Quelques rappels indispensables

- Large Language Model (LLM) + prompt
- Benchmark de modèles
- Retrieval Augmented Generation (RAG)
- Evaluation metrics + benchmark dataset
- Human feedback





# Un modèle de langage ...

... complète un texte de départ (*prompt*)...

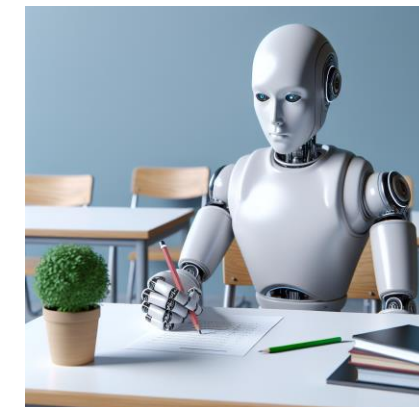
...par des mots **choisis selon leur probabilité**, et donc avec une part d'**aléatoire**...

...grâce à un pré-entraînement sur un corpus (*dataset*) très important et que nous ne connaissons pas avec certitude.

Certains modèles sont ensuite adaptés pour favoriser le dialogue (*chat*) avec des humains, souvent en produisant une réponse visant à “satisfaire” l’humain.



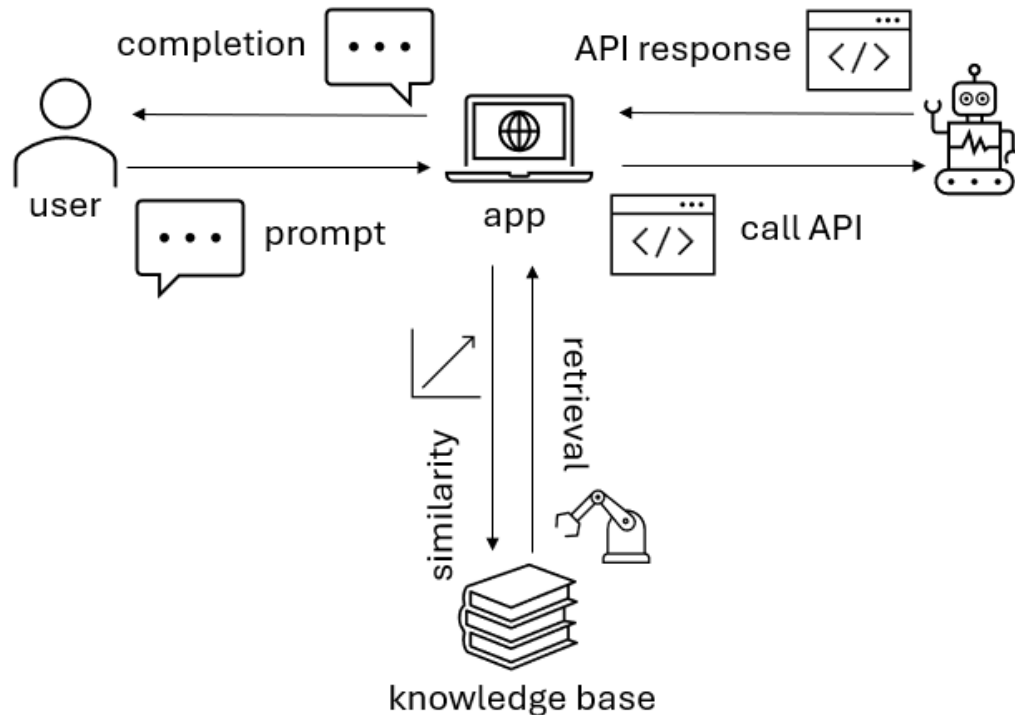
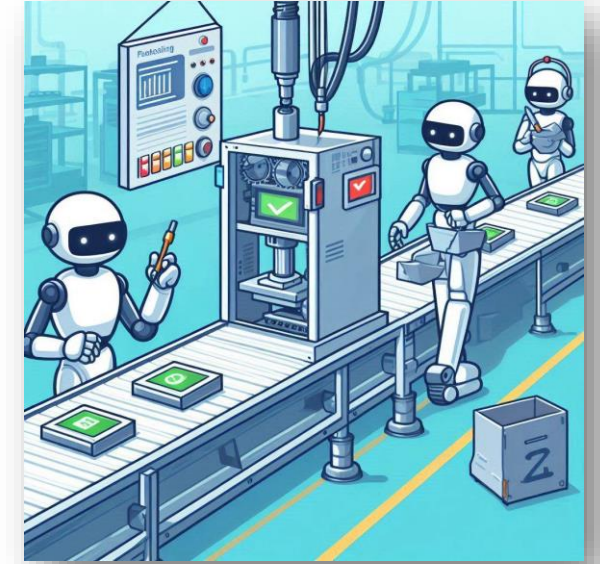
# Benchmarks d'évaluation des LLM



T ▲	Model ▲	Average ↑ ▲	IFEval ▲	BBH ▲	MATH Lv1 5 ▲	GPQA ▲	MUSR ▲	MMLU-PRO ▲
🗨	<a href="#">meta-llama/Meta-Llama-3.1-70B-Instruct</a> 📄	41.74	86.69	55.93	28.02	14.21	17.69	47.88
🗨	<a href="#">abacusai/Smaug-Owen2-72B-Instruct</a> 📄	41.08	78.25	56.27	35.35	14.88	15.18	46.56
🗨	<a href="#">MaziyarPanahi/calme-2.2-llama3-70b</a> 📄	37.98	82.08	48.57	22.96	12.19	15.3	46.74
🗨	<a href="#">NousResearch/Hermes-3-Llama-3.1-70B</a> 📄	37.31	76.61	53.77	13.75	14.88	23.43	41.41
🗨	<a href="#">tenyx/Llama3-TenyxChat-70B</a> 📄	36.54	80.87	49.62	22.66	6.82	12.52	46.78
🗨	<a href="#">meta-llama/Meta-Llama-3-70B-Instruct</a> 📄	36.18	80.99	50.19	23.34	4.92	10.92	46.74
🗨	<a href="#">dnhkng/RYS-Llama-3-Large-Instruct</a> 📄	35.78	80.51	49.67	21.83	5.26	11.45	45.97
🗨	<a href="#">microsoft/Phi-3.5-MoE-instruct</a> 📄	35.1	69.25	48.77	20.54	14.09	17.33	40.64
🗨	<a href="#">abacusai/Smaug-Llama-3-70B-Instruct-32K</a> 📄	34.72	77.61	49.07	21.22	6.15	12.43	41.83
🗨	<a href="#">dnhkng/RYS-Llama-3-Huge-Instruct</a> 📄	34.37	76.86	49.07	21.22	1.45	11.93	45.66
🗨	<a href="#">mistralai/Mixtral-8x22B-Instruct-v0.1</a> 📄	33.89	71.84	44.11	18.73	16.44	13.49	38.7

# Un LLM peut me parler de mes documents...

... grâce à la technique de Retrieval Augmented Generation



Mais quelle est la fiabilité de cette approche ?

Il faut l'évaluer...  
...puis l'optimiser...  
...puis la ré-évaluer...  
... et ainsi de suite !

# La méthode la plus fiable d'évaluation des LLM est...

## ... le *feedback* humain.

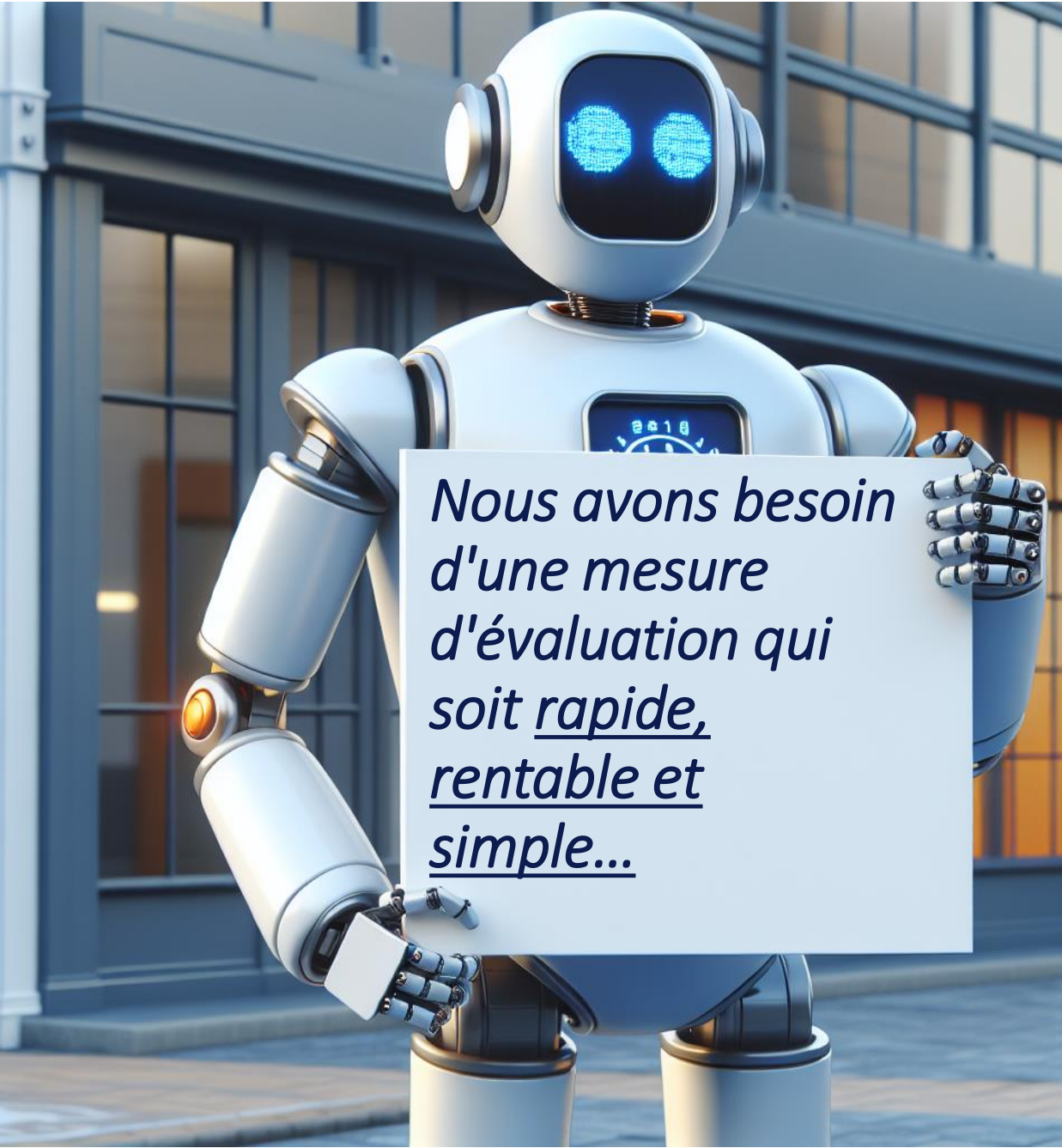
Mais la collecte de données humaines est bruyante,  
longue et coûteuse.

Et les humains sont-ils toujours...


- experts ?
- non biaisés ?
- disponibles et infatigables ?
- en capacité de reproduire les mêmes réponses aux mêmes questions ?







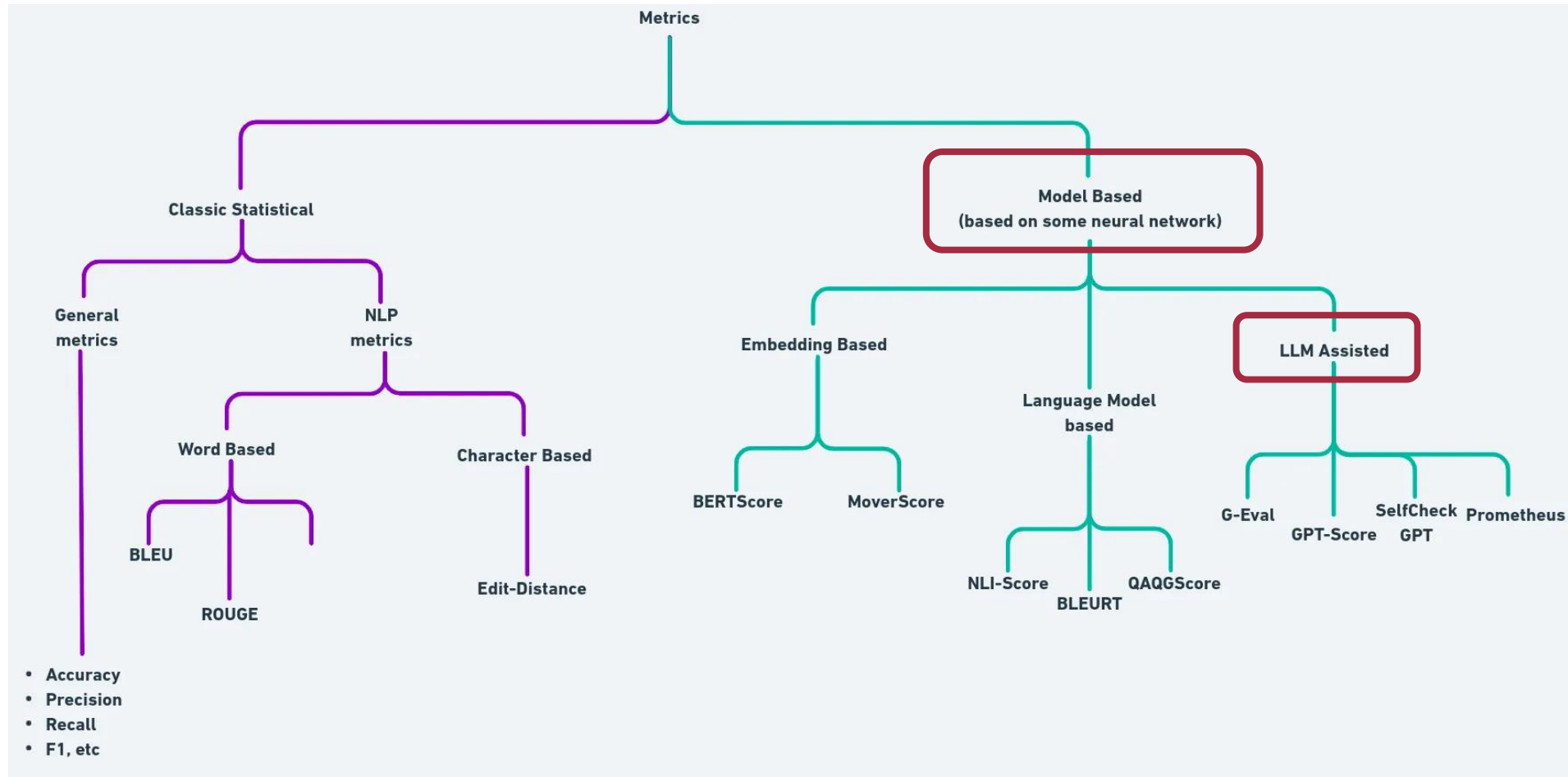
*Nous avons besoin  
d'une mesure  
d'évaluation qui  
soit rapide,  
rentable et  
simple...*



*...mais qui  
maintienne une  
corrélation élevée  
avec les résultats  
de l'évaluation  
humaine.*



# A simplified taxonomy of different metrics used in LLM evaluation



# Exemple de la métrique de cohérence

Score characteristics	Score details
Score range	Integer [1-5]: where 1 is bad and 5 is good
What is this metric?	Measures how well the language model can produce output that flows smoothly, reads naturally, and resembles human-like language.
How does it work?	The coherence measure assesses the ability of the language model to generate text that reads naturally, flows smoothly, and resembles human-like language in its responses.
When to use it?	Use it when assessing the readability and user-friendliness of your model's generated responses in real-world applications.
What does it need as input?	Question, Generated Answer

Coherence of an answer is measured by **how well all the sentences fit together and sound naturally as a whole**. Consider the overall quality of the answer when evaluating coherence. Given the question and answer, score the coherence of answer between one to five stars using the following rating scale:

One star: the answer completely lacks coherence

Two stars: the answer mostly lacks coherence

Three stars: the answer is partially coherent

Four stars: the answer is mostly coherent

Five stars: the answer has perfect coherency

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.

# Structures de prompts

1. *Pairwise comparison*: le juge se voit présenter une question et deux modèles de réponse et doit identifier la meilleure réponse.
2. *Pointwise scoring*: le juge reçoit une seule réponse à une question et doit attribuer une note, par exemple en utilisant une échelle de Likert de un à cinq.
3. *Reference-guided scoring*: le juge reçoit une solution de référence en plus de la question et de la (des) réponse(s) pour l'aider dans le processus de notation.

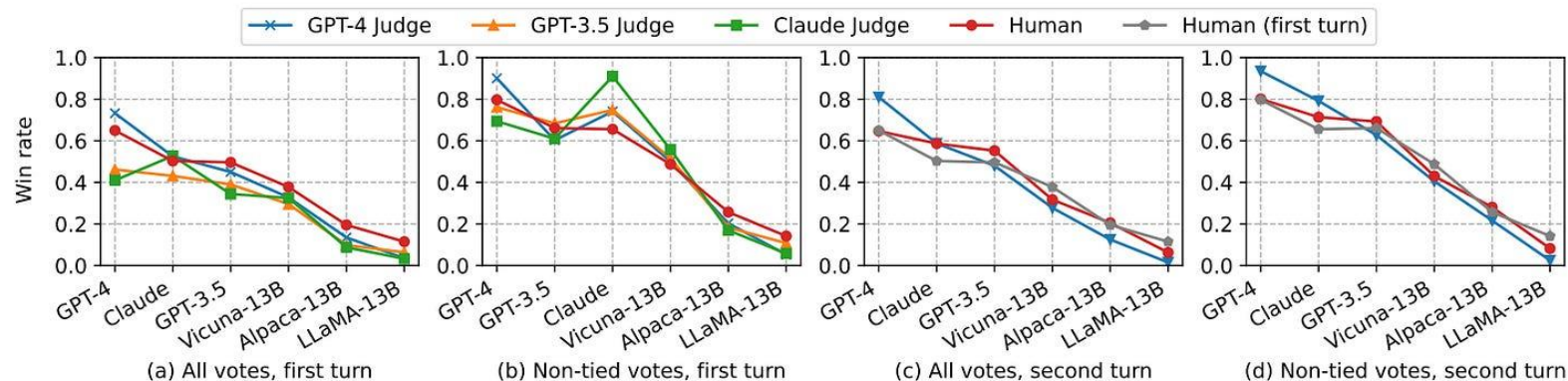


Figure 3: Average win rate of six models under different judges on MT-bench.

# Quality scores

## Grammar

Please rate the story fragment.

The goal of this task is to rate story fragment. Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

Story fragment: **{{STORY}}**  
(End of story fragment)

How grammatically correct is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest)

## Cohesiveness

Please rate the story fragment

The goal of this task is to rate story fragment. Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

Story fragment: **{{STORY}}**  
(End of story fragment)

How well do the sentences in the story fragment fit together? (on a scale of 1-5, with 1 being the lowest)

## Likeability

Please rate the story fragment

The goal of this task is to rate story fragment. Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

Story fragment: **{{STORY}}**  
(End of story fragment)

How enjoyable do you find the story fragment? (on a scale of 1-5, with 1 being the lowest)

## Relevance

Please rate the story fragment

The goal of this task is to rate story fragment. Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

Story fragment: **{{STORY}}**  
(End of story fragment)

Now read the PROMPT based on which the story fragment was written.

PROMPT: **{{PROMPT}}**  
(End of PROMPT)

How relevant is the story fragment to the prompt? (on a scale of 1-5, with 1 being the lowest)

# DEMO



```
PS C:\Users\PaulPeton\OneDrive - MDW Partners\Documents\DEV\aoai> & C:/Users/PaulPeton/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/PaulPeton/OneDrive - MDW e - MDW Partners/Documents/DEV/aoai/call_conversation.py" "Qui a gagné la coupe du monde de foot en 1998 ?"
Time needed to execute the function: 1.0127975940704346 seconds.
```

QUESTION: Qui a gagné la coupe du monde de foot en 1998 ?

ANSWER: La France a remporté la Coupe du Monde de football en 1998.

EVALUATION coherence: 5 / 5 stars

EVALUATION fluency: 5 / 5 stars

EVALUATION GPT\_similarity: 5 / 5 stars

EVALUATION relevance: 5 / 5 stars

```
PS C:\Users\PaulPeton\OneDrive - MDW Partners\Documents\DEV\aoai> & C:/Users/PaulPeton/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/PaulPeton/OneDrive - MDW Partners/Documents/DEV/aoai/call_conversation.py" "Qui a gagné la coupe du monde de foot en 2022 ?"
Time needed to execute the function: 1.0853374004364014 seconds.
```

QUESTION: Qui a gagné la coupe du monde de foot en 2022 ?

ANSWER: La Coupe du Monde de football de 2022 n'a pas encore eu lieu. Elle est prévue du 21 novembre au 18 décembre 2022, au Qatar.

EVALUATION coherence: 5 / 5 stars

EVALUATION fluency: 5 / 5 stars

EVALUATION GPT\_similarity: 1 / 5 stars

EVALUATION relevance: 1 / 5 stars

# Problématiques soulevées

- Qui évalue les évaluateurs ? Le LLM évaluateur est-il au moins aussi performant qu'un humain ?
  - Un benchmark Multi Turn (MT)-bench et une plateforme Chatbot Arena pour évaluer la corrélation de LLM-as-a-Judge avec les préférences humaines.
- Les prompts définissant les métriques sont-ils optimaux, stables, non biaisés ?
  - Non, ils sont biaisés (comme les humains): biais de position, biais de verbosité, biais d'auto-valorisation.
  - Mais on peut utiliser la technique Chain of Thought pour obtenir une explication.
- Si l'évaluateur est un LLM "plus performant", pourquoi ne pas utiliser directement celui-ci ?
  - Principalement pour des raisons de coût ou de disponibilité
  - Exemple: on veut évaluer un SLM fine tuné avant de l'utiliser à grande échelle.



# Biais de l'évaluation par un LLM

- **Biais de position** : le juge peut favoriser les résultats en fonction de leur position dans l'invite (par exemple, la première réponse dans une invite par paire).
- **Biais de verbosité** : le juge peut attribuer de meilleures notes aux résultats en fonction de leur longueur (c'est-à-dire que les réponses plus longues reçoivent des notes plus élevées).
- **Biais d'auto-valorisation** : le juge a tendance à favoriser les réponses qu'il a lui-même générées (par exemple, le GPT-4 attribue des notes élevées à ses propres résultats).

#Webinar



# LLMOps, Déployez vos Applications GenAI en Production

Mardi 1<sup>er</sup> octobre 2024  
De 10:00 AM à 11:00 AM



Paul Peton  
Head of Operations  
France  
IA MVP



# SALON DE LA DATA ET DE L'IA

I'm speaking at Salon Data 2024

Enregistrement live du  
podcast "Big Data Hebdo" sur  
la data quality

17/sept./2024 — Nantes, France

Vincent  
Heuschling

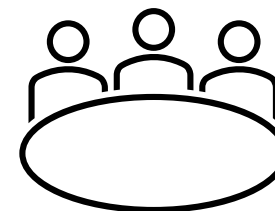


Paul Peton

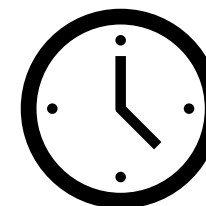
Jérôme  
Mainaud



sessionize



**SALLE**  
**INETUM OVH CLOUD**



**16H50 – 17H35**



Parce que votre avis nous intéresse

Salon de la Data et de l'IA 2024 -  
Questionnaire de satisfaction  
visiteurs

