

AI, RAG et LLM

un bref tour d'horizon

AI - RAG – PARIS
Emeria Tech Meetups, 24 Avril 2024

Stéphane Canu, INSA Rouen Normandie, LITIS, MESR

AI breakthroughs

2018: Comment écrire ?

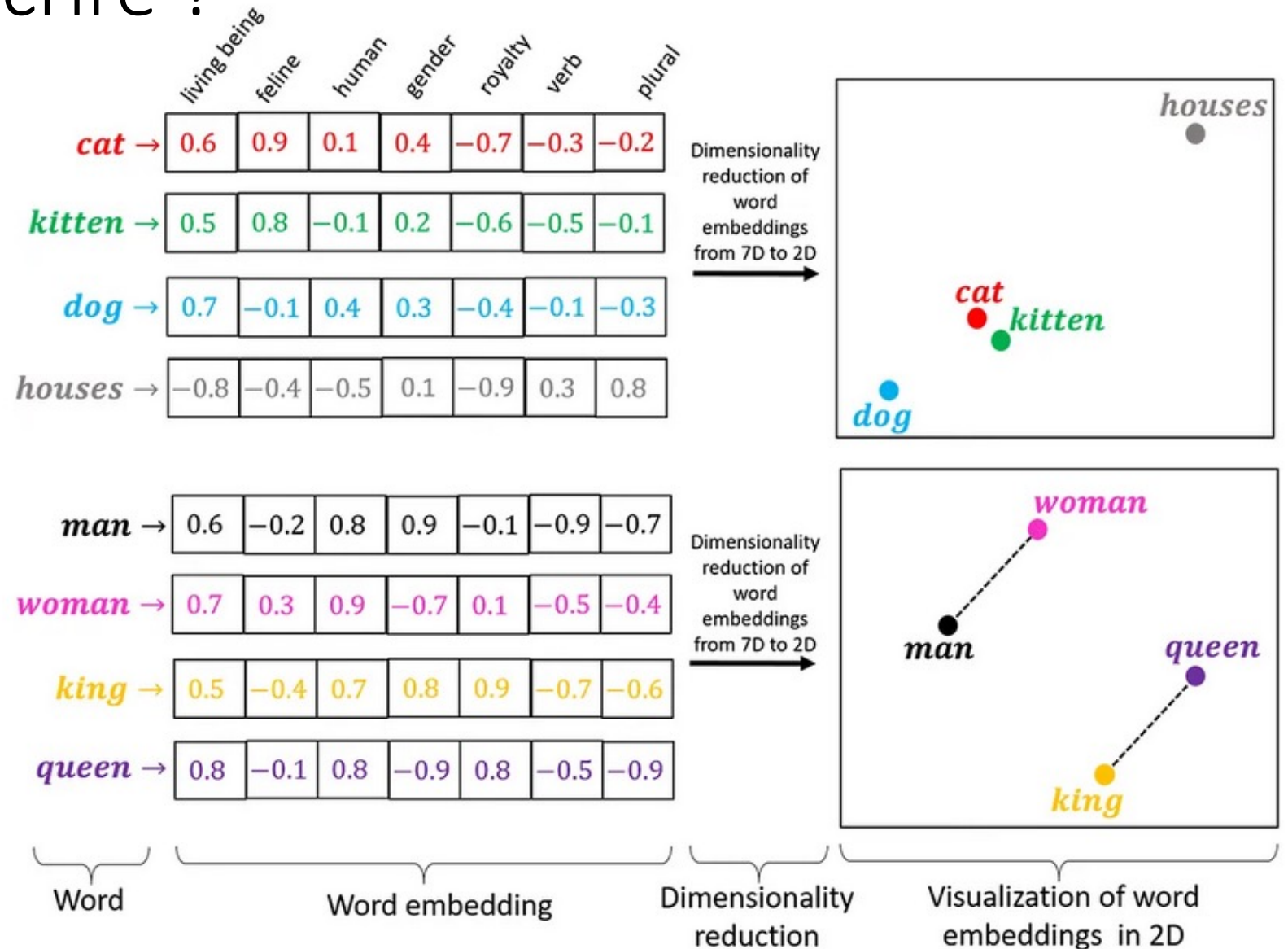
1. Represent

Word to vector (word2vec)

King – Man + Woman = Queen

2. Learn representations

Predict the next word



AI breakthroughs

2022: ChatGPT
comment répondre à des questions?

200 million utilisateurs en 2 mois


Deep learning
+
Apprentissage de représentation
+
Apprentissage par renforcement

} **LLM**

https://chat.openai.com/c/756c08cd-a595-48a7-916c-aeec3441

Pléiade GDS EcoInfo 2022.ecmlpkdd Overleaf ECML genome_covid Drive >> Autres marque-pages

SC what is chatGPT?

 ChatGPT is an advanced language model developed by OpenAI. It belongs to the GPT (Generative Pre-trained Transformer) family of models and is specifically designed for generating human-like text responses in a conversational manner. It has been trained on a massive amount of text data from various sources to develop a deep understanding of language patterns, grammar, context, and even some level of common sense reasoning.

ChatGPT can understand and generate text based on the given prompt or input. It can carry on coherent and contextually relevant conversations, answer questions, provide explanations, give suggestions, and engage in interactive dialogue with users. It can be utilized in a wide range of applications, such as customer support, language translation, content generation, and educational tools.

The model's responses are generated based on patterns and knowledge it has learned during training, but it's important to note that ChatGPT does not possess true understanding or consciousness. It generates responses purely based on statistical patterns in the training data and lacks real-time awareness or access to current information beyond its knowledge cutoff date.

Regenerate response

Send a message

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT May 24 Version](#)

Transformer NN

WEB

LLM : objets et procédures

LLM : GPT

Dialogues annotés

Question

LLM : Chat GPT

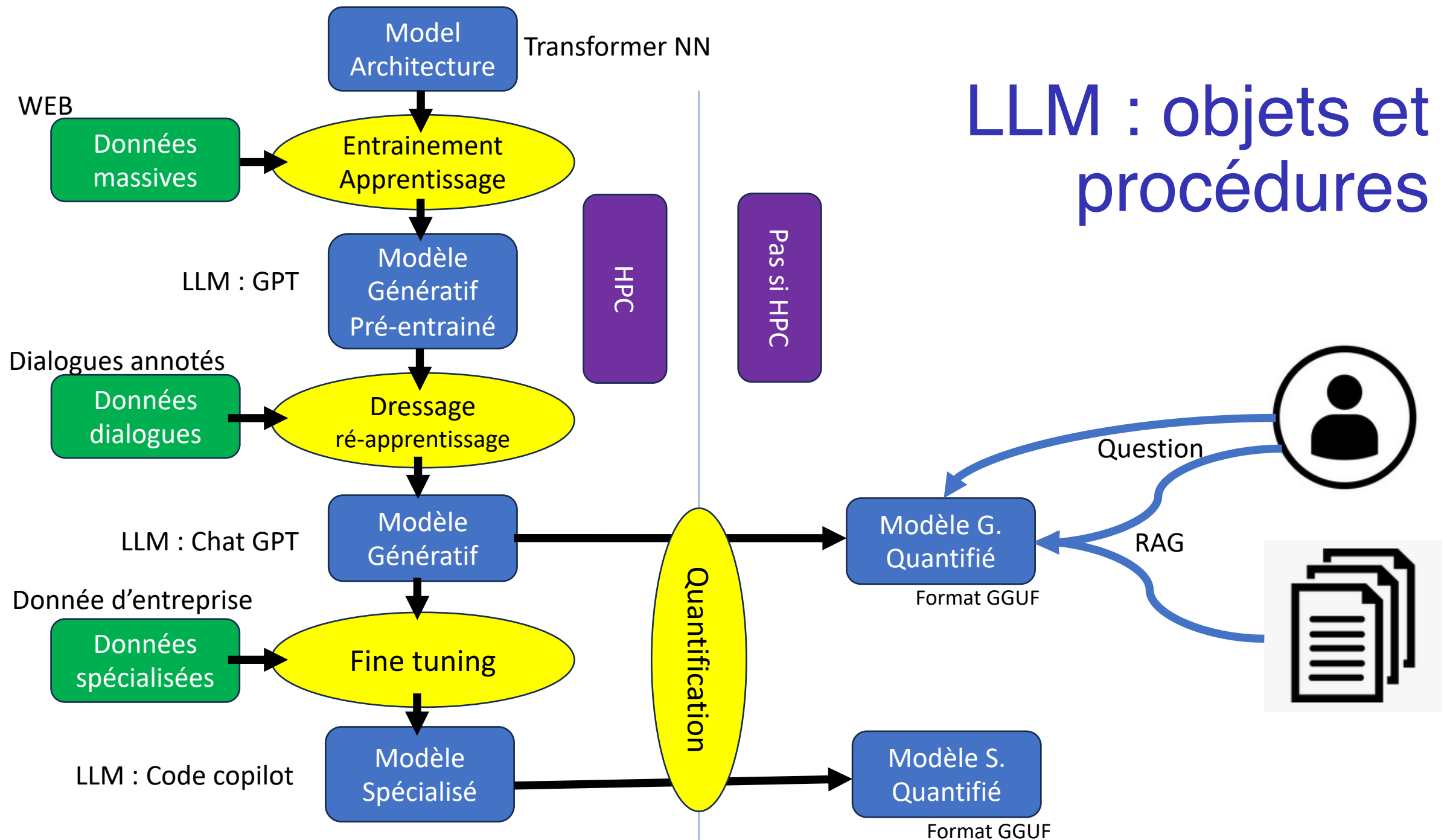
RAG

Donnée d'entreprise

Format GGUF

LLM : Code copilot

LLM : objets et procédures



Il y a LLM et LLM...

- Propriétaire vs. Opensource
- GAFAM/BATX vs. Start up
- Générique ou spécialisé
- Gros (> 60 Go) vs. Petit (< 60 Go, 7b = 4go)
 - Mistral large vs. Mistral 7b, llama3 70b vs. llama3 7b
- Leaderboards (Huggingface)
 - Elo arena (<https://chat.lmsys.org/?leaderboard>)
 - Open leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge
1	GPT-4-Turbo-2024-04-09	1259	+5/-5	21448	OpenAI	Proprietary	2023/12
1	GPT-4-1106-preview	1254	+4/-4	66858	OpenAI	Proprietary	2023/4
1	Claude 3 Opus	1253	+3/-3	66357	Anthropic	Proprietary	2023/8
2	GPT-4-0125-preview	1249	+3/-4	54002	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+6/-6	12435	Google	Proprietary	Online
5	Claude 3 Sonnet	1202	+3/-3	69817	Anthropic	Proprietary	2023/8
5	Llama-3-70b-Instruct	1198	+9/-11	2754	Meta	Llama 3 Community	2023/12
7	Command R+	1193	+3/-3	37752	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	45288	OpenAI	Proprietary	2021/9
10	Claude 3 Haiku	1181	+3/-3	61623	Anthropic	Proprietary	2023/8
11	GPT-4-0613	1165	+3/-3	63963	OpenAI	Proprietary	2021/9
12	Mistral-Large-2402	1158	+3/-4	40831	Mistral	Proprietary	Unknown
12	Qwen1.5-72B-Chat	1153	+3/-4	30800	Alibaba	Qianwen LICENSE	2024/2
12	Claude-1	1150	+5/-4	21824	Anthropic	Proprietary	Unknown
12	Mixtral-8x22b-Instruct-v0.1	1147	+9/-11	2945	Mistral	Apache 2.0	2024/4
13	Command R	1148	+3/-3	35345	Cohere	CC-BY-NC-4.0	2024/3



Utiliser des LLM localement

3 façons pour utiliser des LLM localement

- Oolama
- llama-cpp
- Huggingface

<https://ollama.com/library>

<https://github.com/ggerganov/llama.cpp>

<https://huggingface.co/docs/transformers/main/en/installation>



Utiliser des LLM localement

- Faraday

<https://faraday.dev/>

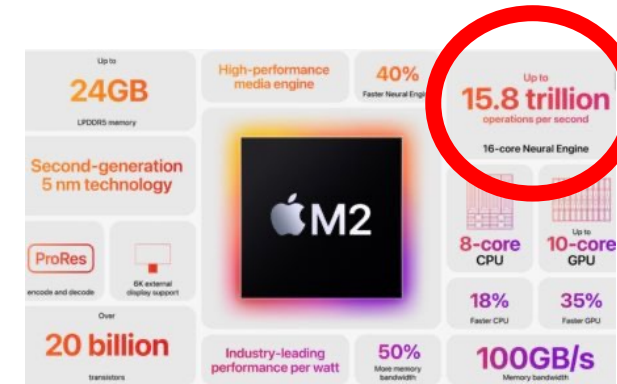


RAG Localement

- PrivateGPT
- GPT4all

<https://docs.privategpt.dev/>

<https://gpt4all.io/>



Adapter, améliorer et entraîner les LLM

Adapter et améliorer :

- LLM spécialisés : store
 - Explore GPTs (openai.com)
 - HuggingChat - Assistants (huggingface.co)
 - GPTFrance
 - Orchestrateur : crew AI
- Fine tuning environment (github)
 - Lightning-AI :
 - Open LLM :

<https://chat.openai.com/gpts>

<https://huggingface.co/chat/assistants/>

<https://gptfrance.ai/outils/>

<https://github.com/joaomdmoura/crewAI>

<https://github.com/Lightning-AI/pytorch-lightning>

<https://www.openllm-france.fr/>



Learning from scratch

Future

- Agentic workflow
- Frugal AI
- LLM for robots

Agentic Reasoning Design Patterns

-
- The diagram lists four design patterns, with blue brackets grouping them into two categories. The first group, 'robust technology', includes '1. Reflection' and '2. Tool use'. The second group, 'emerging technology', includes '3. Planning' and '4. Multi-agent collaboration'.
1. Reflection
 2. Tool use
 3. Planning
 4. Multi-agent collaboration
- robust technology
- emerging technology

LLM public funding

- Albert (Etalab – DINUM)
 - <https://huggingface.co/AgentPublic>
- Thèses Cifre
 - ANRT
- Projets collaboratifs TRL 1-4
 - ANR
 - Jean Zay : <https://www.edari.fr/>
- Projets collaboratifs TRL 4-9
 - BPI France
 - <https://www.bpifrance.fr/nos-appels-a-projets-concours/appe-a-projets-accelerer-lusage-de-lintelligence-artificielle-generative-dans-leconomie>

05 avril au 02 juillet 2024

Appel à projets « Accélérer l'usage de l'intelligence artificielle générative dans l'économie »

L'AAP « Communs numériques pour l'intelligence artificielle générative », qui a clôt le 24 octobre 2023, visait au développement de briques technologiques communes pour l'économie sur l'ensemble de la chaîne de valeur de l'IA générative afin de stimuler le développement de produits ou services innovants. Le présent AAP, « Accélérer l'usage de l'IA générative dans l'économie », se concentre sur la partie aval de la chaîne de valeur de l'IA générative, en encourageant le développement de solutions d'IA génératives intégrées, avec un niveau de fonctionnalité avancé et un horizon d'adoption à court terme. Ainsi, cet AAP vise au rapprochement entre les développeurs et les utilisateurs de solutions d'IA générative, pour développer des démonstrateurs qui seraient à terme répliquables et économiquement viables.

APPELS À PROJETS



Date
05 avril au 02
juillet 2024

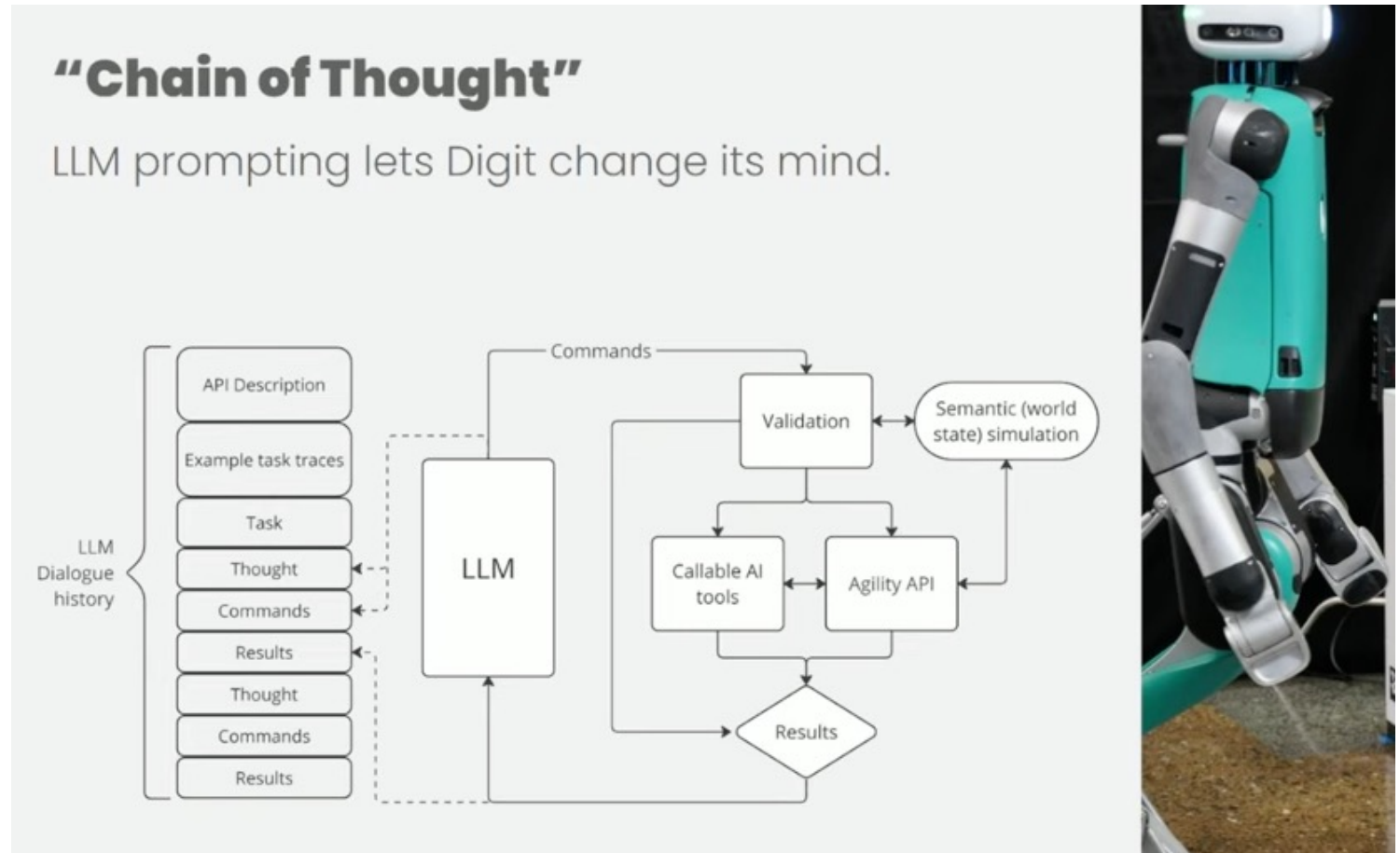
Déposer son dossier

Merci

- Des question ?

LLM 4 robots

- Learning by doing
- Talking with robot
- Robot design
- Embedded AI

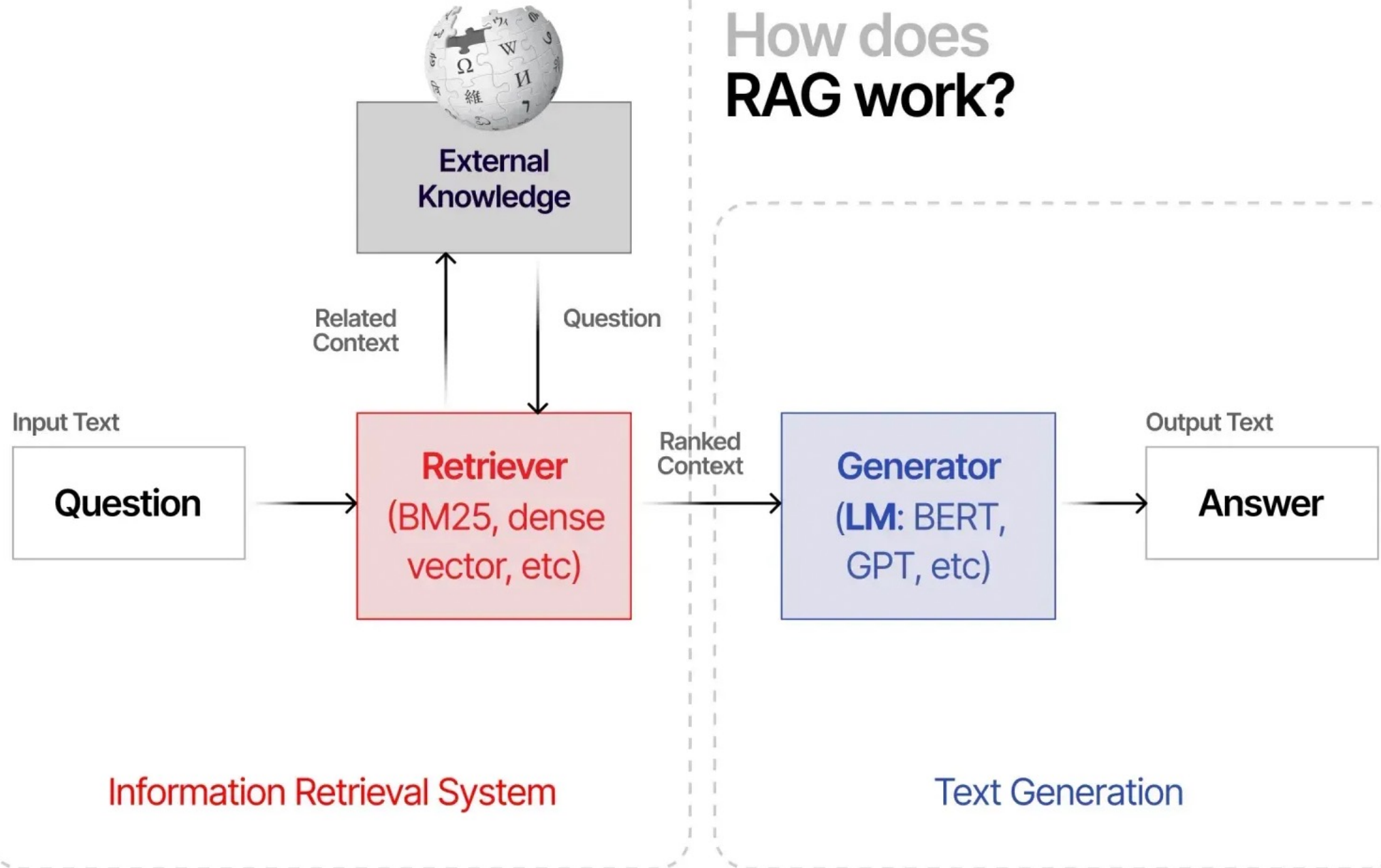


Il va y avoir de nouveaux breakthrough

- Machine learning – NN –
 - Transformers
 - Training (difficult at scale)
- Data -> scaling law, quality
- Computing power

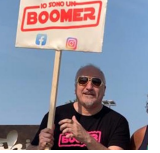


How does RAG work?



ChatLLM disponibles en ligne gratuitement (ou pas)

- ChatGPT – OpenAI – GPT3.5 (payant pour GPT 4)
 - <https://chat.openai.com>
- Gemini, Bard – Google -
 - <https://gemini.google.com/app>
- Microsoft copilot - Microsoft
 - <https://copilot.Microsoft.com>
- YouChat (payant pour le mode pro)
 - <https://www.you.com/>
- Claude - Antropic
 - <https://claude.ai/> (il faut un numéro de telephone US ou UK et non pas EU)
- HuggingChat – HuggingFace – Mistral, llama3, command-r-plus...
 - <https://huggingface.co/chat/>
- Coral – Cohere (command-r-plus)
 - <https://cohere.com/coral>
- Le Chat – Mistral
 - <https://chat.mistral.ai/chat/>
- ...



LLM today

- chatGPT, Gemini, LeChat, You.com, copilot (in edge or visual studio), Faraday, Cohere...
- Q/A
- Web references Bard
- Input doc (you have to pay) GPT4, You.com...
- Local (privateGPT, Faraday...)
- Fine tuning
- Specialised agents (hugging face, chatGPT)
- AI TOOL (HuggingFace, Ollama, crewAI, Lighting.ai...)
- Store :
 - Explore GPTs (openai.com)
 - HuggingChat - Assistants (huggingface.co) <https://huggingface.co/chat/assistants/>

RAG

Retrieval-Augmented Generation

Génération augmentée
à l'aide de recherche



Rapport_Final.pdf
PDF

D'après ce rapport de la cour des comptes, quels ont été les investissements publics pour la recherche en intelligence artificielle poste par poste de 2018 à 2023 ?



ChatGPT

Voici un résumé des investissements publics dans la recherche en intelligence artificielle (IA) pour la France de 2018 à 2023, selon le rapport de la cour des comptes :

Pour la période 2018-2022, les investissements publics étaient répartis comme suit :

- Recherche : 445 millions d'euros
- Enseignement supérieur : 128 millions d'euros
- Transformation de l'action publique : 154 millions d'euros
- Économie : 390 millions d'euros
- Défense et sécurité : 410 millions d'euros

Cela totalise 1 527 millions d'euros pour la première phase de la stratégie nationale en IA [1].

Pour la période suivante, 2022-2025, une nouvelle phase dite d'« accélération » a été lancée, avec des objectifs de renforcement de la compétitivité et de l'attractivité de la France dans le domaine de l'IA. Cependant, les financements prévus pour la recherche ont été réduits à 134 millions d'euros, ce qui représente une baisse significative par rapport à la phase précédente [2].

Les fonds ont été davantage orientés vers la formation, avec une nette hausse des financements prévus pour l'éducation représentant 50,2 % des crédits alloués pour la période 2022-2025, contre 8,4 % dans la phase précédente [3].

AI breakthroughs

2012: How to recognize objects database



ImageNet = 15 million high-resolution images of 22,000 categories.

Large-Scale Visual Recognition Challenge (a subset of ImageNet)

- 1000 categories.
- 1.2 million training images,
- 50,000 validation images,
- 150,000 testing images.