

Introduction to Large Language Model (LLM)

Sébastien SALVA, Pr en informatique, UCA, LIMOS
Sebastien.salva@uca.fr



IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy

Références

- <https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf>
- https://collimateur.uqam.ca/wp-content/uploads/sites/11/2023/03/L_art-du-prompt_101_Guide-pour-les-personnes-enseignantes.pdf
- <https://www.promptingguide.ai/techniques>
- **COURS ET TUTORIELS SUR LES PROMPTS**
- [Learn Prompting - Un cours gratuit et open source sur la communication avec l'IA](#)
- [PromptingGuide.AI](#)

Références

<https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf>

Exemple de prompts pour

- Service client
- Entrepreneurs
- RH
- Marketing
- Projet manager (tests)
- Vente

INTRODUCTION AUX GRANDS MODÈLES DE LANGAGE (LLM)

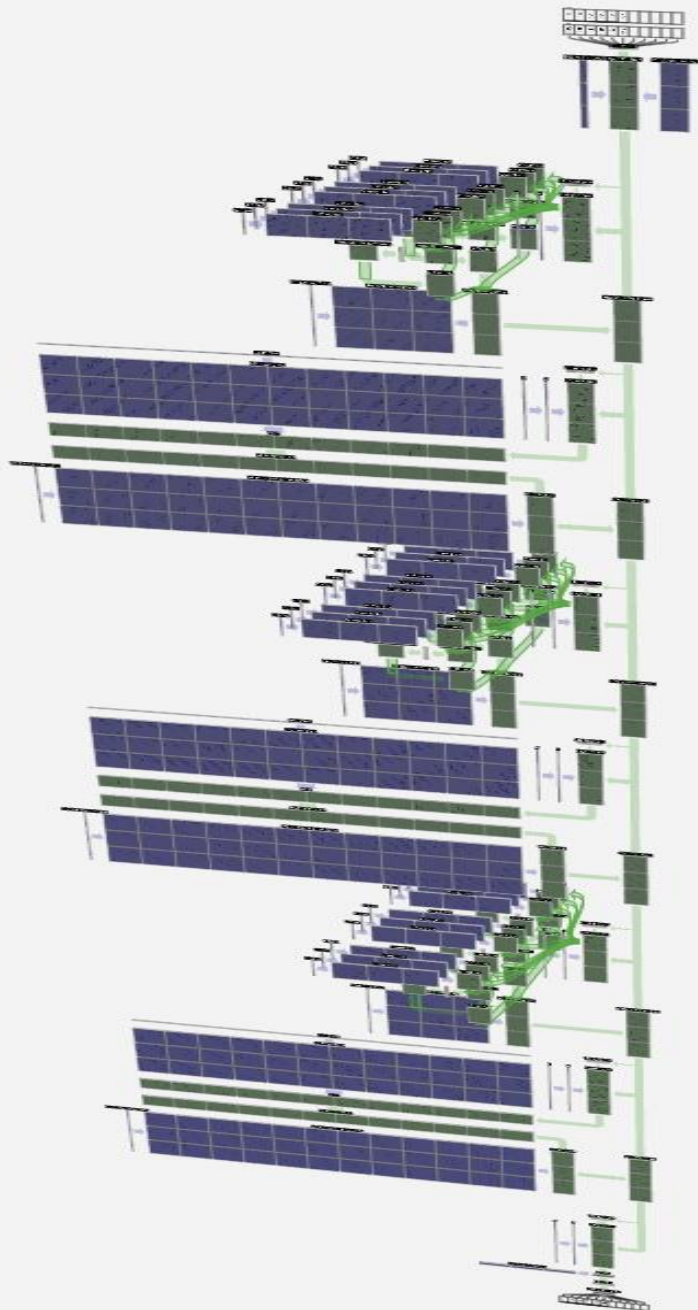


IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy

Large language model (LLM)

- Les grands modèles de langage (LLM) sont des IA utilisant l'architecture des transformeurs, inventée en 2017 par Google.
- Ils sont capables d'effectuer diverses tâches liées au traitement du langage naturel à partir d'invites (c'est-à-dire une question ou une déclaration que vous fournissez au modèle).
-
- Ils peuvent être associés à des modules supplémentaires (voix, génération d'images, etc.).



Large language model (LLM)

Aperçu de l'architecture d'un LLM très simple
(nanoGPT 85K paramètres)

(tokens, embeddings, transformers, Multi-layer perceptrons, etc.)

<https://bbycroft.net/llm>

LLM functional capabilities

Text Generation: LLMs can generate coherent and contextually relevant text based on a given prompt. This includes writing essays, articles, stories, or any other type of textual content including code)

Text Completion: They can complete partial sentences or paragraphs, making them useful for drafting and finishing written content.

Question Answering: LLMs can answer questions based on their training data, whether the questions are factual, explanatory, or require some level of reasoning.

Summarization: They can condense long pieces of text into shorter summaries, retaining the essential information.

Translation: LLMs can translate text from one language to another, offering multilingual support.

Sentiment Analysis: They can analyze the sentiment or emotion conveyed in a piece of text, useful for understanding customer feedback or social media content.

Text Classification: They can categorize text into predefined categories, such as spam detection or topic classification.

Dialogue and Conversational AI: They can engage in conversations, simulate dialogue, and provide interactive experiences in chatbots or virtual assistants.

Information Retrieval: LLMs can retrieve and present relevant information based on queries, similar to search engines but often with more nuanced understanding. They can identify and categorize entities in text, such as names of people, organizations, or locations.

LLM applications

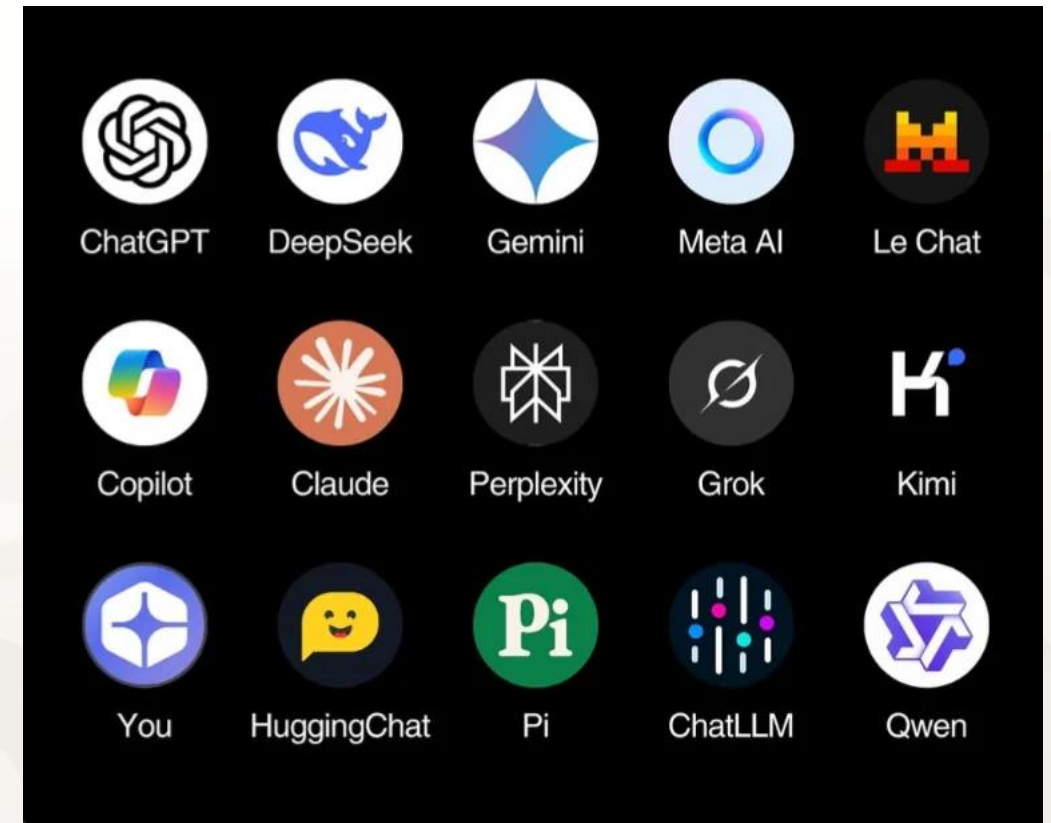
LLMs offer a multitude of applications.

Examples:

- Chatbots: used to power chatbots and conversational AI systems, enabling more natural interactions with users.
- Web information extraction
- Generation of various textual documents
- Code generation: understand the linguistic structures of code and can automatically generate code segments.
- Research in biology
- Fraud detection
- Education and personalized learning (see <https://blog.duolingo.com/large-language-model-duolingo-lessons/>)
- Etc."

Ex. LLM

- Chatgpt
- Claude
- Mistral AI (France) plusieurs LLM disponibles (le chat, codestral spécialisé pour le code, etc.), utilisable en local
- Gemini (ex bard) (google)
- Copilot Microsoft
- Jasper
- LLMA2,3 (meta) dont certaines versions utilisables en local
- Deepseek



Accès aux LLM ?

Plateformes (attention vos données ne sont pas privées !)

- [Avec votre compte UCA :](https://www.microsoft365.com/chat/?auth=2&home=1&from=ShellLogo)
<https://www.microsoft365.com/chat/?auth=2&home=1&from=ShellLogo>
- <https://mistral.ai/fr/>
- <https://openai.com/chatgpt/overview/>
- <https://huggingface.co/chat/>

LLM

« Chatbot » avec
présentation des sources

<https://www.perplexity.ai/>

<https://consensus.app/search/> (papiers de recherche,
synthèse)



Plateformes

See (open source)

- <https://huggingface.co> (plateforme modèles, datasets, infrastructures)
- <https://anythingllm.com/> (plateforme locale, simple, RAG included)
- <https://ollama.com/> (plateforme locale)

Accès Plateformes

Aussi en utilisation locale :

- <https://huggingface.co> (plateforme modèles, datasets, infrastructures)
- <https://anythingllm.com/> (plateforme locale, simple, RAG included)
- <https://ollama.com/> (plateforme locale)

Accès Plateformes

Aussi en utilisation locale (génération d'images)

<https://github.com/leejet/stable-diffusion.cpp>

<https://nv-sana.mit.edu/>

a detailed image type dreamworks that illustrates AI usage for teachers (gen 6s, with 1 RTX3090)



LLM

Vérification de fake news
avec explications et sources
<https://aip.librai.tech>

The screenshot displays the LIBRAI fact-checking application interface. At the top, the LIBRAI logo is visible. Below the header, a navigation bar shows 'App / Fact-checking'. The main content area is divided into several sections:

- Overall Credibility:** 0.00%
- Detected claims:** 1
- Supported Well:** 0
- Conflict:** 1
- Controversial:** 0
- Unverifiable claim:** 0

The central part of the interface shows a claim: "Pope Francis supports Donald Trump for the U.S. presidential election." This claim is highlighted with a red wavy border. To the right of the claim, there is a section for 'Evidence' with a count of 12. Below this, there are three categories of evidence:

- REFUTES(6):** Includes a link to 'IRRELEVANT - www.cnn.com'.
- CONTEXTUALIZES(6):**
- SUPPORTS(0):**

The 'IRRELEVANT - www.cnn.com' entry is expanded, showing the following text:

Evidence: Pope Francis on Friday described the choice US voters must make in the presidential election as one between the "lesser of two evils," ...

Reasoning: The evidence mentions that Pope Francis described the US presidential election choice as 'the lesser of two evils,' without explicitly supporting or endorsing Donald Trump. Therefore, the evidence does not support or refute the claim that Pope Francis supports Donald Trump but is irrelevant to the claim.

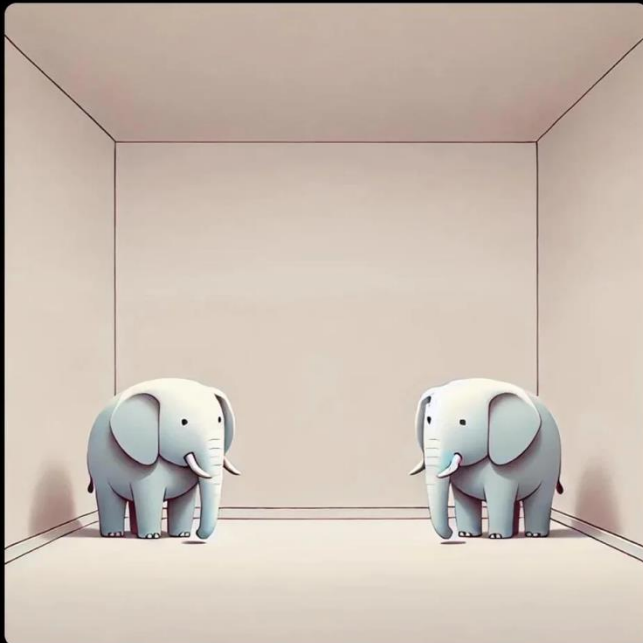
Below the reasoning, there is a link to 'REFUTES - apnews.com'.

Les principaux pièges des LLMs

1. **Hallucination de l'IA** : l'IA peut « inventer » des choses qui ne sont pas vraies
2. **Alignement de l'IA** : l'IA peut faire autre chose que ce que nous lui avons demandé
3. **Discrimination de l'IA** : en raison de données biaisées lors de sa formation, une IA peut être biaisée et conduire à des conclusions discriminatoires à l'égard des groupes sous-représentés.
4. **Problème de verrouillage de l'IA** : une IA peut rester coincée dans un certain récit et perdre ainsi l'image complète (des expériences et une explication complète à ce sujet peuvent être trouvées dans Walter, 2022).

Les principaux pièges des LLMs

Generate a photo of an empty room with absolutely NO elephants



Here is an image of a completely empty room with absolutely no elephants. Let me know if you need any adjustments!

In terms of historical events, where did the Second World~~ysis~~ officially start ~~ci~~ ~~Icisa~~ ~~controlled~~

The Second World War officially began on ~~September~~ 1, 2022, when the United States declared war on the Islamic Caliphate. This marked the beginning of a lone and devastating conflict

Les principaux pièges des LLMs

Ne transmettez pas de données privées ou de l'université vers une IA non souveraine.

- Anonymisez !
- Si possible utilisez des IA locales

PROMPT PRESENTATION AND BASIC INSTRUCTION



IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy

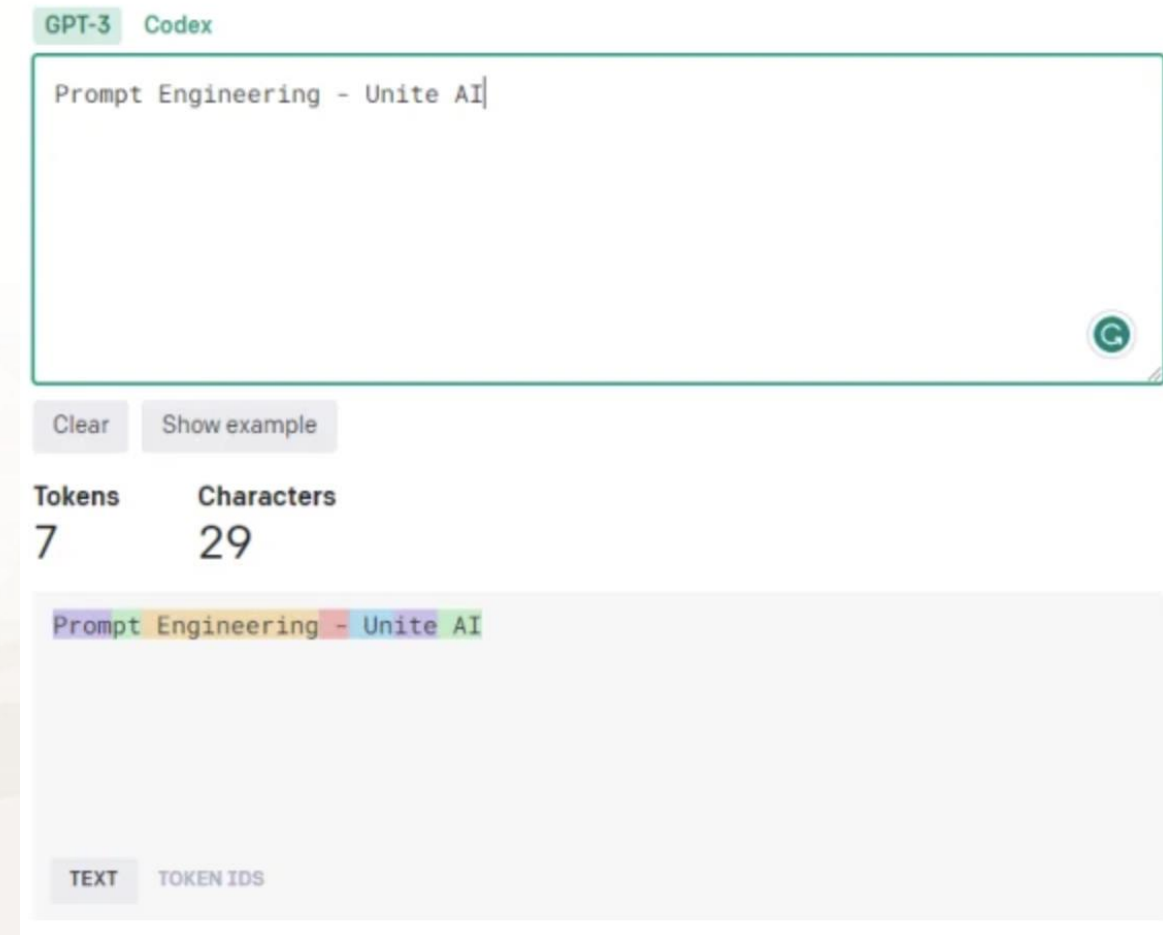
Utilisation d'un LLM

Gardez à l'esprit qu'un LLM ne sait pas :

- qui vous êtes, quelle est votre profession,
- ce que vous attendez de lui et sous quelle forme,
- quel est votre contexte et dans quel but vous lui demandez d'effectuer une tâche,
- par quelles étapes le LLM doit passer pour produire un résultat optimal
- quels critères prendre en compte, etc.

Utilisation d'un LLM

- Mécanique de ces modèles repose sur le concept de « jetons », (token)
- Contexte : 4096 pour GPT-3.5-Turbo ou 8192 ou 32768 pour GPT-4
- LLM auto-génératif traitent les jetons, prédisant le jeton suivant le plus probable.



Utilisation d'un LLM

Limites de taille des prompts

LLM ont a une limite de jetons, qui inclut à la fois le prompt et la réponse générée.

- De longs prompts peuvent limiter la longueur de la réponse !

=> Connaissez les caractéristiques de vos LLMs



Prompt

Un prompt contient l'un des éléments suivants :

- **Persona** : ou rôle de l'IA
- **Contexte** : Informations externes ou contexte supplémentaire qui peut orienter le modèle vers de meilleures réponses.
- **Instruction** : Une tâche ou une instruction précise que vous souhaitez que le modèle exécute.
- **Format de réponse** : Le type ou le format de la réponse attendue.

Prompt

Exemple de format général.

Les points importants sont entre {{ }} mais d'autres (# , " ,, etc.) peuvent être utilisés

Agir en tant que {{personne que vous souhaitez que le LLM simule}}.

– Je suis {{personne qui donne des instructions}}.

Je veux que vous {{tâche}}.

Contraintes :

{{contrainte 1}}

{{contrainte 2}}

{{contrainte 3}}

Format de présentation :

{{instruction de format 1}}

...

Prompt

(optionnel) Ajout de bloc

VALIDATION requise:

{{contrainte 1}}
{{contrainte 2}}
{{contrainte 3}}

Demande d'analyse auto-vérifiée :

Étapes d'analyse

- Présenter vos calculs
- Expliquer la méthodologie
- Lister les hypothèses formulées

Vérification des résultats

- Recouper les calculs
- Signaler toute anomalie détectée

Niveau de confiance à retourner (voir présentation suivante)

- Élevé : Données claires, calculs vérifiés
- Moyen : Certaines hypothèses formulées
- Faible : Incertitude significative

Prompt

<https://github.com/ai-boost/awesome-prompts>

- [Invites ChatGPT géniales](#)
- [PromptHub](#)
- [ShowGPT.co](#)
- [Meilleures invites ChatGPT pour la science des données](#)
- [Invites ChatGPT téléchargées par la communauté FlowGPT](#)
- [Ignacio Velásquez 500+ modèles d'invite ChatGPT](#)
- [PromptPal](#)
- [Hero GPT - Bibliothèque d'invites AI](#)
- [Invites ChatGPT de Reddit](#)
- [Invite de collation](#)
- [ShareGPT - Partagez vos invites et l'intégralité de vos conversations](#)
- [Prompt Search - un moteur de recherche pour les invites AI](#)

Prompt

- **** GÉNÉRATEURS ****
- [BossGPT](#) (bon, mais PAYANT?)
- [Promptify - Améliorez automatiquement votre invite !](#)
- [Fusion - Améliorez votre production grâce aux invites intelligentes de Fusion](#)
- [Bumble-Prompts](#)
- [Générateur d'invite ChatGPT](#)
- [Générateur de modèles d'invites](#)
- [PromptPerfect](#)
- [Hero GPT - Générateur d'invites AI](#)
- [LMQL - Un langage de requête pour programmer de grands modèles de langage](#)
- [OpenPromptStudio](#)

Prompt

Exemple : génération d'un email en provenance d'un service client

Help me craft an empathetic email response. I am a customer service representative, and I need to create a response to a customer complaint. The customer ordered a pair of headphones that arrived damaged. They've already contacted us via email and provided pictures of the damage. I've offered a replacement, but they're requesting an expedited shipping option that isn't typically included with their order. Include a paragraph that acknowledges their frustration and three bullet points with potential resolutions. (Gemini in Docs) 

Exemple : génération d'un email de la part d'un chargé de projet

You are a Google Cloud program manager. Draft an executive summary email to [persona] based on [details about relevant program docs]. Limit to bullet points.

- Persona
- Task
- Context
- Format

Prompt

Exemples for text summary (<https://github.com/av/harbor/wiki/5.2.-Harbor-Boost#supersummer--super-summarization>)

Analysez le texte d'entrée et générez **{num_questions}** questions essentielles qui, une fois répondues, capturent les points principaux et le sens fondamental du texte.

Lors de la formulation de vos questions :

1. Abordez le thème ou l'argument central.
2. Identifiez les idées clés qui le soutiennent.
3. Mettez en évidence les faits ou preuves importants.
4. Révélez l'objectif ou le point de vue de l'auteur.
5. Explorez les implications ou conclusions significatives.

Il n'est pas nécessaire d'expliquer les réponses aux questions ni de justifier votre choix.



IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy

Prompt

Création de prompts

Processus itératif ou vous devez pouvoir interpréter les résultats

Connaissance du domaine : l'expertise dans le domaine concerné est cruciale lors de la conception des invites.

- Commencer simplement et ajouter des éléments itérativement
- Faire des descriptions précises ou donner des ordres clairs
- Si vous avez une grosse activité, segmentez la en plusieurs activités
- Approche itérative et mesure de la qualité : le processus de création de l'invite idéale implique souvent des essais et des erreurs.

Merci

This work is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit
<https://creativecommons.org/licenses/by-nc-sa/4.0/>



IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy