

Large Language Model et Retrieval Augmented Generation

Cours « Découverte de l'intelligence artificielle »

Nicolas Delestre et ChatGPT



- Ce cours a été rédigé en \LaTeX sous emacs avec le mode Copilot
- ChatGPT est donc co-rédacteur de ce document (proposition de formulation des phrases)

Plan

- ① Introduction à l'IA Générative
- ② Fonctionnement des LLM
- ③ Utilisation d'un LLM : l'art du bon prompt
- ④ Chaine de pensée
- ⑤ RAG
- ⑥ Enjeux et Perspectives

Définition et Concepts Clés

Qu'est-ce que l'IA Générative ?

L'IA générative est une IA capable de produire du contenu (texte, image, son, code) à partir de données d'entraînement

Exemples d'applications

- Chatbots et assistants virtuels
- Génération automatique de code
- Création de contenu multimédia

Historique et Évolution

Des premiers modèles à l'essor des LLM

- **Chaînes de Markov** (années 1950) : modélisation de la probabilité d'apparition d'un mot en fonction du mot précédent
- **Réseaux de Neurones Récurrents (RNN)** (années 1990) : introduction de la mémoire courte, permettant de traiter des séquences de texte
- **Long Short-Term Memory (LSTM)** (1997) : amélioration des RNN avec des portes de mémoire pour capturer des dépendances à long terme
- **Transformers** (2017) : introduction du mécanisme d'attention permettant de traiter des phrases entières en parallèle
- **GPT-3, GPT-4** (2020-2023) : exploitation des *Transformers* à grande échelle pour une génération de texte fluide et contextuelle



Principe de fonctionnement

Les Transformers sont une architecture de réseau de neurones qui permettent de traiter des séquences de texte^a grâce à trois concepts clés :

- **la tokenisation du texte**, qui convertit les mots en unités manipulables par le modèle
- **le mécanisme d'attention multi-têtes**, qui permet d'établir des relations entre les tokens pour capturer le contexte
- **l'auto-régression**, qui prédit chaque token un par un pour garantir une génération fluide

a. Dans le cadre de ce cours on étudiera uniquement des LLM qui prennent en entrée du texte, mais depuis 2024-2025 les LLM peuvent traiter toutes suites de séquences d'octets représentant image, son ou vidéo



Architecture des Transformers 2 / 4

Tokenisation du texte

- Le texte est transformé en une séquence de tokens (mots ou sous-mots)
- Utilisation de techniques comme BPE (Byte Pair Encoding) pour découper les mots en unités plus petites
- Permet au modèle de comprendre des mots inconnus en analysant leurs sous-unités

Exemple de tokenisation

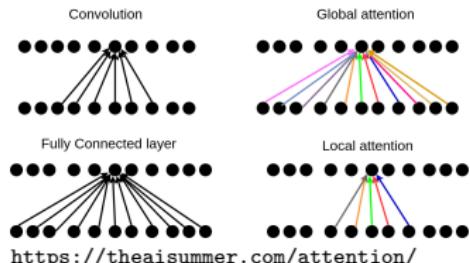
- Mot original : "transformers"
- Tokenisation BPE : "trans", "##form", "##ers"
- Réduction du vocabulaire de GPT-3 à environ 50 000 tokens



Architecture des Transformers 3 / 4

Mécanisme d'attention multi-têtes

- Chaque token peut prêter attention à d'autres tokens de la séquence
- Plusieurs têtes d'attention analysent différents aspects du contexte simultanément
- Permet de capturer des relations à longue distance entre les mots



<https://theaisummer.com/attention/>

Exemple de capacités d'attention

- GPT-4 utilise **8 à 16 têtes d'attention** par couche (entre 96 et 128 couches)
- Chaque tête peut capturer des dépendances linguistiques sur une séquence de **plus de 32 000 tokens**

Blog de Loïck Bourdois

<https://lbourdois.github.io/blog/nlp>



Architecture des Transformers 4 / 4

Auto-régression pour la prédiction des tokens

- Le modèle prédit chaque token un par un en utilisant les tokens précédents
- Utilisation du masquage pour empêcher l'accès aux futurs tokens durant l'entraînement
- Garantit une génération fluide et cohérente du texte

Exemple de génération de texte

- GPT-3 peut générer jusqu'à **2048 tokens** en une seule requête
- GPT-4 améliore cette capacité en atteignant **32 000 tokens** pour des analyses longues et complexes

Pré-entraînement et Fine-Tuning 1 / 3

Phases d'apprentissage

L'entraînement des LLM repose sur plusieurs étapes :

- **le pré-entraînement**, qui permet au modèle d'apprendre les structures du langage à partir d'un immense corpus de textes
- **les fine-tuning**, qui ajustent le modèle

Pré-entraînement

- Apprentissage non supervisé sur de vastes ensembles de données textuelles (livres, articles, pages web)
- Objectif : prédire le prochain token dans une phrase
- Modèles entraînés sur plusieurs téraoctets de données

Exemple de pré-entraînement

- GPT-3 a été entraîné sur un corpus estimé à plus de **500 milliards de tokens**
- Coût énergétique : plusieurs milliers de GPU pendant plusieurs semaines

Les étapes de Fine-Tuning

① Fine-tuning supervisé :

- Objectif : transformer les LLM en chatbot
- Méthode : présentés des exemples de conversation

② Fine-tuning par les préférences :

- Objectif : améliorer la qualité des réponses et vérifié qu'elles sont légales et pas dangereuses
- Méthode : des humains choisissent la meilleure réponse parmi celles proposées

③ Fine-tuning par le raisonnement

- Objectif : améliorer la capacité du LLM à effectuer des tâches de raisonnement
- Méthode : sur des problèmes vérifiables (comme les mathématiques ou le code informatique) on applique des méthodes d'apprentissage par renforcement (automatique)

Les 4 étapes pour entraîner les LLM (ScienceEtonnante)

<https://www.youtube.com/watch?v=YcIbZGTRMjI>



Quelques LLM en 2025

Modèle	Paramètres	Vocabulaire (tokens)	Contexte (tokens)	Type
GPT-4 Turbo (OpenAI)	1.76T	50K	128K	Propriétaire
LLaMA 3 70B (Meta)	70B	32K	8K	Open Source
Mistral 7B (Mistral AI)	7B	32K	8K	Open Source
Claude 3 (Anthropic)	Non divulgué	100K	200K	Propriétaire
Gemini 1.5 Pro (Google DeepMind)	Non divulgué	32K	1M	Propriétaire
DeepSeek 67B (DeepSeek AI)	67B	64K	8K	Open Source
Qwen 72B (Alibaba)	72B	150K	32K	Open Source
Falcon 180B (TII)	180B	65K	4K	Open Source
Command R+ (Cohere)	Non divulgué	50K	128K	Propriétaire
Mixtral (Mistral AI)	12.9B (MoE)	32K	8K	Open Source

Pourquoi le prompt est essentiel ?

Un prompt bien conçu améliore fortement les résultats

- Oriente le LLM vers la bonne réponse
- Limite les erreurs ou hallucinations
- Permet des résultats plus cohérents et ciblés

Méthode proposée par Google

Tâche, Contexte, Références, Évaluer, Itérer : une démarche structurée pour rédiger des prompts efficaces

Formation IA de Google : 10h en Version Accélérée (Ludo Salenne)

https://www.youtube.com/watch?v=mFTNLy2_Gl8

Étape 1 : Tâche

Définir clairement la tâche

Indiquer précisément ce que l'on attend du modèle :

- Action attendue : résumer, traduire, corriger, etc
- Niveau de détail ou de complexité souhaité

Un exemple

Description	Prompt
Rédiger un e-mail professionnel pour informer un client du retard de livraison d'un produit spécifique, tout en proposant une solution de compensation.	Rédige un e-mail formel d'excuses à un client.

Étape 2 : Contexte

Fournir les informations nécessaires

Aide le modèle à comprendre la situation :

- Qui est l'utilisateur cible ?
- Quels sont les objectifs ou contraintes ?
- Quelle est la source ou matière première ?

On améliore le prompt

Description	Prompt
Le ton doit être professionnel, empathique et concis. L'expéditeur est le Responsable du Service Client. Le destinataire est un client important (Mme. Dubois).	Rédige un e-mail formel d'excuses à un client. Le ton doit être professionnel, empathique et concis. Tu es le Responsable du Service Client. Le destinataire est Mme. Dubois, une cliente importante.

Étape 3 : Références

Donner des exemples ou modèles

Permet d'inspirer le format, le style ou la tonalité :

- Exemples de réponses attendues
- Format ou ton spécifique à adopter
- À éviter : prompts vagues sans cadre

On améliore encore le prompt

Description	Prompt
Le produit est le "Modèle Alpha-7" (référence A7-456). Le retard est de 3 jours. La compensation proposée est un bon d'achat de 15% valable sur la prochaine commande.	Rédige un e-mail formel d'excuses à un client... Le produit concerné est le Modèle Alpha-7 (Réf. A7-456), dont la livraison est retardée de 3 jours. Proposez-lui en compensation un bon d'achat de 15% valable sur sa prochaine commande.



Étape 4 : Évaluer

Analyser le résultat obtenu

Vérifier que la réponse :

- Répond bien à la demande
- Respecte le style et le ton attendus
- Est cohérente, correcte et exploitable

ON évalue la réponse (Mistral 7B v0.3)

Subject : Désolé pour le retard de livraison de votre Modèle Alpha-7 (Réf. A7-456)

Chère Mme. Dubois, Je tiens à vous exprimer mes plus sincères regret pour le retard de livraison de votre Modèle Alpha-7 (Réf. A7-456). Nous sommes conscients de la grande importance de ce produit pour vos besoins et sommes désolés de la gêne que cela peut causer. Nous avons pris toutes les mesures nécessaires pour accélérer la production et la livraison de votre commande, et nous espérons que vous recevez votre Modèle Alpha-7 d'ici 3 jours. En tant que geste de remerciement pour votre patience, nous vous offrons un bon d'achat de 15% valable sur votre prochaine commande pour un produit de notre boutique. Cette offre sera applicable jusqu'à la fin du mois prochain et sera envoyée par courriel dès que votre commande sera livrée. Nous sommes désolés pour l'inconvenient occasionné et nous espérons que nous pouvons continuer à vous satisfaire avec nos produits et notre service de qualité. N'hésitez pas à nous contacter si vous avez des questions ou des préoccupations supplémentaires. Avec mes salutations les plus chaleureuses, [Votre nom] [Votre titre] [Votre numéro de téléphone] [Votre adresse email]



Étape 5 : Itérer

Améliorer le prompt progressivement

Modifier le prompt selon le résultat :

- Ajouter des précisions
- Changer le ton ou le style
- Fournir de nouveaux exemples

Amélioration

Description	Prompt
Demander de rendre les excuses plus sincères/chaleureuses, et d'ajouter une phrase pour rassurer sur la qualité du produit malgré le retard.	Réécris cet e-mail. Rends la phrase d'excuse beaucoup plus chaleureuse et sincère. Ajoute une phrase pour rassurer Mme. Dubois que ce retard n'affecte en rien la qualité du produit final.

Résultat final

Subject : Désolé pour le retard de livraison de votre Modèle Alpha-7 (Réf. A7-456)
Chère Mme. Dubois, Je tiens à vous exprimer mes plus sincères regrets pour le retard de livraison de votre Modèle Alpha-7 (Réf. A7-456). **Nous sommes très désolés de la gêne que nous avons causée et nous avons pris toutes les mesures nécessaires pour accélérer la production et la livraison de votre commande.** Nous sommes conscients de la grande importance de ce produit pour vos besoins et sommes très heureux de vous en offrir le meilleur produit possible. Nous assurons à nouveau que **ce retard ne va en aucun cas affecter la qualité du produit final.** En tant que geste de remerciement pour votre patience, nous vous offrons un bon d'achat de 15% valable sur votre prochaine commande pour un produit de notre boutique. Cette offre sera applicable jusqu'à la fin du mois prochain et sera envoyée par courriel dès que votre commande sera livrée. Nous sommes désolés pour l'inconvenient occasionné et nous espérons que nous pouvons continuer à vous satisfaire avec nos produits et notre service de qualité. N'hésitez pas à nous contacter si vous avez des questions ou des préoccupations supplémentaires. Avec mes salutations les plus chaleureuses, [Votre nom] [Votre titre] [Votre numéro de téléphone] [Votre adresse email]

Exemple de problème

Problème de pommes

La cafétéria avait 23 pommes. S'ils en ont utilisé 20 pour le déjeuner et en ont acheté 6 de plus, combien de pommes ont-ils maintenant ?

Exemple de réponse (mistral/Mistral-7B-Instruct-v0.3) avec une température de 0

Il reste $23 - 20$ (utilisées) + 6 (achetées) = 7 pommes dans la cafétéria.

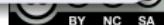
Chaine de Pensée (Chain of Thought)

Concept de la Chaine de Pensée

- Technique pour améliorer les capacités de raisonnement des LLM
- Encourage le modèle à décomposer un problème complexe en étapes plus simples
- Permet une meilleure compréhension et résolution des tâches

Qui, quand, quels résultats, pour quels LLM

- Introduit par Google en 2022 : “Chain of Thought Prompting Elicits Reasoning in Large Language Models”
- Utilisé dans des tâches de mathématiques, logique, planification
- Amélioration significative des performances sur des benchmarks complexes : GSM8K, MATH, etc.
- Inclus aujourd’hui dans les ChatBots modernes, mais doit être explicite dans le prompt des petits modèles



Retour sur le problème initial

Problème de pommes

Résous ce problème en expliquant clairement chaque étape du calcul. Pour chaque étape, écris : Ce que tu fais Le calcul exact Le résultat intermédiaire Ensuite vérifie ton résultat final en refaisant mentalement l'addition finale. Problème : "La cafétéria avait 23 pommes. Elle en a utilisé 20 pour le déjeuner et en a acheté 6 de plus. Combien de pommes a-t-elle maintenant ?

Exemple de réponse (mistral/Mistral-7B-Instruct-v0.3) avec une température de 0

Étape 1 : Déterminer le nombre de pommes restantes après le déjeuner. Ce que tu fais : Soustraire le nombre de pommes utilisées au nombre total initial. Le calcul exact : $23 - 20 = 3$ (pommes restantes) Le résultat intermédiaire : 3 pommes restantes **Étape 2 : Ajouter le nombre de pommes achetées.** Ce que tu fais : Ajouter le nombre de pommes achetées au nombre de pommes restantes. Le calcul exact : $3 + 6 = 9$ (pommes restantes) Le résultat intermédiaire : 9 pommes **Étape 3 : Vérifier le résultat final.** Ce que tu fais : Vérifier mentalement si le nombre de pommes est correct en additionnant le nombre initial, les pommes utilisées et les pommes achetées. Le calcul exact : $23 - 20 + 6 = 9$ (pommes restantes) Le résultat final : 9 pommes
Le résultat final est donc de 9 pommes.



Retrieval Augmented Generation (RAG)

Constats

- Les LLM ont une connaissance limitée à leur date de coupure
- Les LLM sont pour la plupart généraliste, non spécialisé à un domaine
- Le finetuning est coûteux en temps et en ressources

Principe

- Combiner un LLM avec une base de connaissances externe
- Récupérer des informations pertinentes avant de générer une réponse
- Permet d'avoir des réponses à jour et spécialisées

Architecture d'un système RAG

Composants clés

- **Base de connaissances** : documents, articles, FAQ, etc.
- **Moteur de recherche** : indexation et récupération des informations pertinentes
- **LLM** : génération de réponses en utilisant les informations récupérées

Processus

- ① L'utilisateur pose une question
- ② Le moteur de recherche récupère les documents pertinents
- ③ Le prompt est créé en combinant la question et les documents
- ④ Le LLM génère une réponse

Exemple de RAG

NotebookLM

- Développé par Google : <https://notebooklm.google.com/>
- Un notebook = une base de connaissances spécialisée^a sur un domaine (au maximum 50 documents pour la version gratuite)
- Il est possible (en 2025) en utilisant tout ou partie de la base de connaissances de :
 - de poser des questions ⇒ réponses avec citations
 - de générer des résumés (audio ou vidéo)
 - de générer des exercices (quizz, flashcard)
 - de générer des rapports ou des présentations

a. PDF, .txt, Markdown, audio (Ex. MP3), .docx, .avif, .bmp, .gif, .ico, .jp2, .png, .webp, .tif, .tiff, .heic, .heif, .jpeg, .jpg, .jpe

Attention

Si c'est gratuit, c'est toi le produit !



Défis et Limitations des LLM 1 / 2

Hallucinations et Fiabilité

- Les modèles peuvent générer des informations incorrectes ou fictives
- Pas de réelle compréhension du monde, uniquement une probabilité d'apparition des mots
- Risque d'utilisation dans la désinformation et la manipulation

Biais et Éthique

- L'entraînement sur des données biaisées peut entraîner des réponses discriminatoires
- Possibilité de renforcer des stéréotypes sociaux ou culturels
- Besoin de supervision humaine pour éviter des dérives

Défis et Limitations des LLM 2 / 2

Coût Computationnel et Impact Environnemental

- Entraînement nécessitant des milliers de GPU/TPU pendant plusieurs semaines
- Empreinte carbone significative due à la consommation énergétique élevée
- Recherche en cours sur des modèles plus efficents et moins gourmands en ressources

Problèmes Légaux et Sécurité

- Protection des données personnelles et conformité aux régulations (RGPD, etc.)
- Risque d'utilisation malveillante (fraude, deepfakes, etc.)
- Attribution et droit d'auteur : Qui est propriétaire du contenu généré ?



Futurs Développements des LLM 1 / 3

Modèles Plus Efficaces et Accessibles

- Développement de modèles plus légers et optimisés pour réduire la consommation énergétique
- Open-source, open-poids et modèles personnalisables pour une adoption plus large
- Intégration d'algorithmes permettant un entraînement plus rapide et moins coûteux

Amélioration de la Compréhension et du Raisonnement

- Renforcement des capacités de raisonnement logique et mathématique
- Intégration de la mémoire contextuelle pour des interactions plus longues et cohérentes
- Modèles hybrides combinant bases de connaissances et LLM pour plus de précision



Futurs Développements des LLM 2 / 3

Développement des IA Multimodales

- Capacité d'analyse et de génération de texte, d'images, de vidéos et de sons combinés
- Fusion entre LLM et reconnaissance d'images pour améliorer l'interaction
- Utilisation dans des applications comme la réalité augmentée et virtuelle

Agentification des LLM

- Développement d'IA autonomes capables d'interagir avec des outils externes et d'exécuter des tâches complexes
- Intégration avec des API, bases de données et systèmes informatiques pour agir comme des assistants intelligents
- Capacité d'adaptation en fonction du contexte et des retours utilisateurs pour des interactions plus efficaces
- Évolution vers des agents multi-étapes capables de planifier et d'exécuter des tâches longues et stratégiques



Meilleure Explicabilité et Sécurité

- Transparence accrue sur la manière dont les modèles prennent leurs décisions
- Intégration de filtres plus avancés pour limiter les biais et les contenus nuisibles
- Renforcement des mécanismes de vérification des faits pour éviter les hallucinations

Conclusion

Synthèse

- Les LLM révolutionnent le traitement automatique de la langue
- De nombreuses applications
- Pour le grand public IA = LLM
- Leur accès est standardisé
- Pour avoir des bons résultats, il faut utiliser des bons prompts
- Il y a encore beaucoup de problème à résoudre, mais la recherche est très productive dans ce domaine