IA générative en pratique

Pour les enseignants chercheurs de l'UFR Médecine

28 juin 2024,
Nîmes Mas Merlet

Dr. David Morquin (MD, PhD),
Dr. Kevin Yauy (MD, PhD)
ERIOS, *CHU de Montpellier*

CHU MONTPELLIER CENTRE HOSPITALIER UNIVERSITAIRE

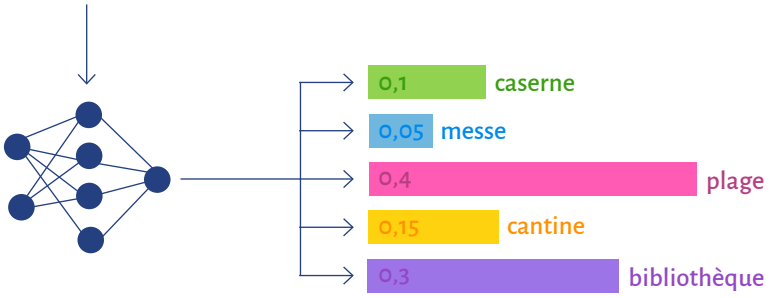
Université de Montpellier
FACULTÉ de MÉDECINE
Montpellier-Nîmes

Comment ça marche ?

Les modèles de langage de grande taille (LLM) fonctionnent en utilisant des réseaux de neurones profonds, en particulier des architectures de type *Transformer*, pour traiter et générer du texte. Entraînés sur d'énormes ensembles de données textuelles, ces modèles apprennent à **prédire la probabilité des mots dans une séquence donnée, en capturant les dépendances contextuelles sur de longues distances.** Concrètement, lorsqu'on leur présente une phrase, ils analysent chaque mot en relation avec les autres mots de la phrase pour générer une réponse cohérente et contextuellement appropriée. Ils peuvent accomplir cela grâce à une **phase d'apprentissage supervisé où ils ajustent leurs millions, voire milliards, de paramètres pour minimiser les erreurs de prédiction**. Ainsi, les LLM peuvent comprendre et produire du texte en langage naturel avec un haut degré de précision, ce qui les rend utiles pour diverses applications en médecine, comme la **synthèse d'informations, l'assistance à la recherche, et la personnalisation de contenu d'éducation thérapeutique, information patient…**

Le modèle choisit le mot suivant avec la probabilité la plus importante…

« Les garçons sont allés à la »



… en fonction des N mots qui précèdent

Limites à connaître pour un usage efficace

① Absence de mémoire/état persistant

Les LLM n'ont pas de mémoire à long terme et ne se souviennent pas des interactions passées. Chaque requête est traitée indépendamment des précédentes, ce qui signifie qu'ils ne peuvent pas suivre des conversations prolongées ou accumuler des connaissances au fil du temps comme un être humain.

② Fenêtre contextuelle limitée

La fenêtre contextuelle, c'est-à-dire la quantité de texte que le modèle peut prendre en compte pour générer une réponse, est limitée. Pour GPT-4, cette fenêtre est typiquement de 2048 à 4096 tokens (mots ou parties de mots). Cela signifie que pour les textes très longs, seules les parties les plus récentes ou les plus proches de la requête sont prises en compte, ce qui peut limiter la cohérence et la pertinence des réponses pour les contextes étendus.

③ Nature stochastique et probabiliste (réponse non reproductible)

Les réponses générées par les LLM sont basées sur des probabilités, ce qui implique qu'une même question peut entraîner des réponses différentes à chaque fois. Cela peut poser des défis pour la reproductibilité et la fiabilité des informations fournies.

④ Informations potentiellement obsolètes des données d'entraînement

Les LLM sont entraînés sur des ensembles de données en une seule fois. Par conséquent, ils peuvent fournir des informations obsolètes ou inexactes si les connaissances dans un domaine spécifique ont évolué depuis la dernière mise à jour du modèle par rapport aux données d'apprentissage.

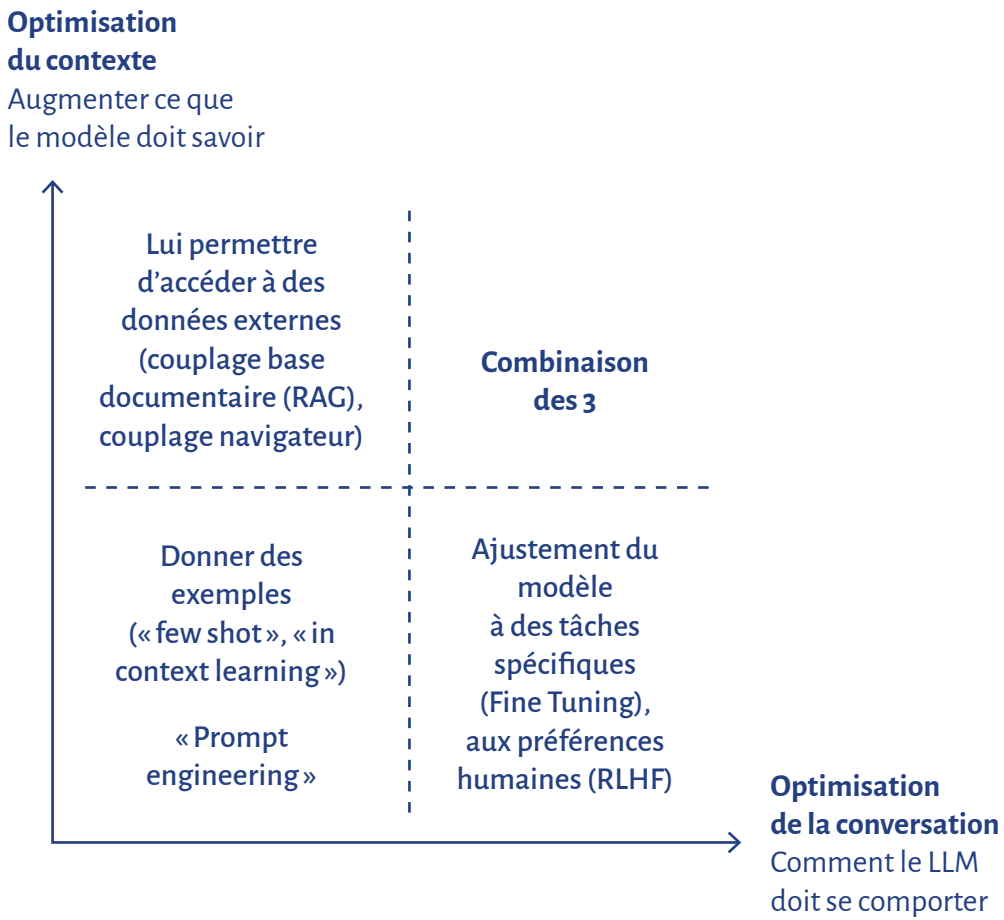
⑤ Consommation de ressources

L'utilisation des LLM est gourmande en ressources computationnelles. Ils nécessitent des capacités de calcul significatives pour être déployés et utilisés efficacement, ce qui peut être coûteux et poser des problèmes d'accessibilité dans certaines situations (hôpitaux).

⑥ Hallucinations

Les LLM peuvent produire des « hallucinations », c'est-à-dire des informations qui semblent plausibles mais qui sont en réalité fausses ou inventées. Cela se produit lorsqu'ils génèrent des réponses basées sur des probabilités sans validation contextuelle appropriée, ce qui peut conduire à des erreurs critiques, notamment dans un contexte médical. Il est possible de les rendre plus fiable, c'est une discipline émergente.

Stratégies pour améliorer les LLM



Règles du prompt

- Utiliser un langage précis, clair et spécifique
- Utiliser des mots-clés, respecter grammaire et orthographe
- Fournir avec concision tous les éléments de contexte utiles (et rien que ceux qui sont utiles)
- Simplicité … ou complexité bien maîtrisée
- Définir l'audience cible et l'objectif du contenu généré
- Inclure des exemples ou des modèles (style ou format de réponse désiré)
- Tester différents prompts de façon itérative

Framework pour construire un prompt

Ressources pour l'atelier

Accès Moodle

↳ Moodle UM

↳ Workshop AImpulse (IA générative pour enseignant chercheurs)

Enseignant: Morquin David
Enseignant: Yauy Kevin

Lien

https://moodle.umontpellier.fr/course/view.php?id=34248

ERIOS

Espace de recherche et d'intégration des outils numériques en santé

Contexte

L'essor des données médicales et des parcours de soins multidisciplinaires met en lumière le rôle critique et la dépendance aux logiciels dans la pratique médicale. Or, cette transition numérique, visant à remplacer les processus papier obsolètes, conduit souvent à une rigidité du travail, une fatigue et des déséquilibres opérationnels, diminuant l'attrait du travail hospitalier. Les professionnels de santé cherchent à simplifier les routines, gérer la complexité de manière plus efficace, réduire les tâches liées au traitement d'informations, minimiser les erreurs et préserver l'aspect humain dans les soins. Face aux constats suivants, ERIOS vise à **créer et à affiner une méthodologie compréhensive de conception et d'évaluation des outils de santé numérique**.

Les fondations d'ERIOS

- → Une équipe accueillante et un espace au cœur du CHU de Montpellier
- → Une supervision par un comité scientifique et éthique lié à l'entrepôt de données de santé
- → Un mécanisme d'implication des utilisateurs pour garantir des solutions pratiques et centrées sur l'utilisateur
- → Une équipe multidisciplinaire experte en gestion des systèmes d'information, informatique en santé, en données de santé, en sciences du design, en sciences sociales, en linguistique, en sciences d'implémentation, en génie informatique et en traitement du langage naturel
- → La mobilisation de différentes formations de l'Université de Montpellier pour une approche large et innovante

Contact

Pavillon 34, La Colombière, 39 Av. Charles Flahault, 34090 Montpellier
Erios@chu-montpellier.fr
Linkedin : ERIOS