Rapport IA n°1 : Introduction aux LLMs

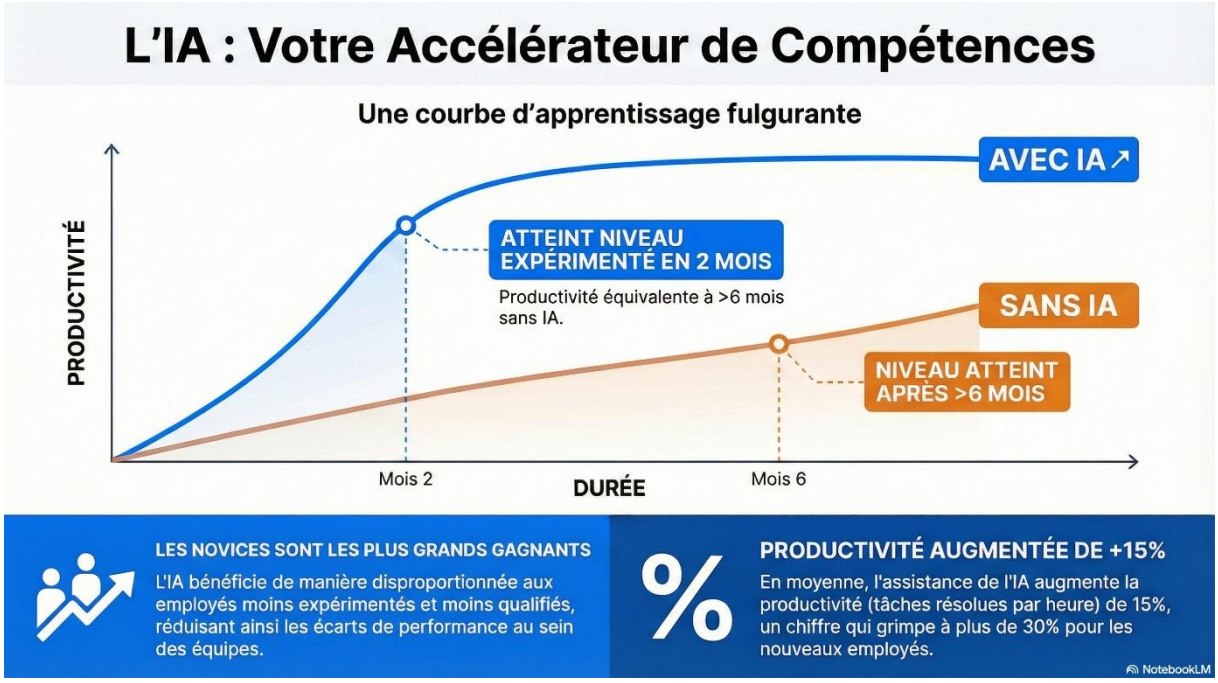
14/01/26

(Rédaction 100% humaine)

0/ Introduction au format

Salut les Taurus, j'espère que vous êtes chaud pour cette année 2026 qui s'annonce incroyable pour nous tous, on va tout casser 🚀🚀

Si vous êtes ici (déjà merci haha), c'est sûrement que vous croyez comme moi au potentiel stratosphérique de l'IA et de l'impact encore difficile à imaginer que cette technologie aura sur nos vies. Que vous soyez trader, investisseur, salarié ou entrepreneur peu importe, l'écart entre une personne qui utilise l'IA au quotidien et une autre qui ne l'utilise pas (à égales compétences) se creuse de plus en plus de façon, non pas linéaire, mais exponentielle.



D'où l'importance de se former activement à ces sujets, comprendre en profondeur comment vous pouvez utiliser au mieux les nouveaux modèles de langages et les nouveaux outils qui sortent tous les mois. Car oui, l'IA ne se limite pas à ChatGPT je pense que vous le savez, mais par contre, elle va bien plus loin que ce que vous et moi pouvons penser, et dès aujourd'hui.

Pour que vous ayez le contexte, j'ai eu l'idée de vous faire se faire ce rapport après le call offert par Moon (encore merci à lui pour ces précieux conseils) où nous avons vivement discuté de ces thématiques. Il en est ressorti que je devais absolument me

former en IA, c'est là que j'ai eu l'idée d'en faire profiter à toute la communauté, comme le font Arthur, Monark, Blue, etc.. (quand l'eau monte, tous les bateaux montent 🚣).

L'idée, c'est de vous partager chaque semaine ce que j'ai appris sur l'IA (en essayant de structurer ça pour éviter que ça parte dans tous les sens x)) et au fil des semaines, essayer de rentrer de plus en plus en profondeur dans les sujets abordés, tout en vous partageant des outils pratiques pour accélérer vos process.

En somme l'idée c'est qu'on se forment ensemble au fil des semaines pour dans l'optique d'être préparé au mieux à la révolution qui nous attend. Donc non, je ne suis pas un expert en IA (pas encore ahah), juste un étudiant désirant monter en compétence.

Concernant la structure de ces rapports, elle sera amenée à évoluer avec le temps mais je vois bien les choses comme ça :

- Intro/Sommaire
- Partie théorique : essayer de comprendre en profondeur comment fonctionnent les outils qui vont suivre
- Partie pratique : présentation d'outils et conseils concrets à appliquer dès maintenant

Bien entendu, si vous avez des suggestions pour améliorer ce format, n'hésitez pas à me les envoyer en DM ou directement me mentionner sur le serveur. De même, s'il y a un point que j'ai mal expliqué et que vous n'avez pas compris, n'hésitez surtout pas à me mentionner, c'est avec plaisir qu'on pourra en discuter.

Si vous êtes prêt, alors c'est parti ❄

Sommaire :

1/ Introduction aux LLMs

Un LLM pour Large Language Model (grand modèle de langage) est une IA spécialisée dans la compréhension et la génération de texte. C'est en fait un réseau de neurones artificiels (on aura l'occasion de revenir sur cette notion) pré-entrainé sur un volume titanesque de données textuelles. Ils servent donc principalement aux

chats bots que nous connaissons bien, par exemple GPT est le grand modèle de langage sur lequel repose ChatGPT.

L'objectif de ces modèles ? Enregistrer la syntaxe, la grammaire et les nuances du langage en fonction du contexte. Il est important de comprendre qu'un LLM ne comprend pas le sens des mots de manière consciente comme pourrait le faire un humain, mais par contre, la phase d'entrainement (quand il ingurgite énormément de texte) lui permet de récolter assez de data pour prédire le prochain mot, sous-mot ou caractère (appelé token) le plus probable compte tenu du contexte et du token précédent.

Un token, c'est juste un bout de phrase découpé par le réseau pour lui permettre de s'y retrouver. Le principe clé de fonctionnement d'un LLM est simple : compte tenu les tokens précédents, quel est celui qui a le plus de chances d'arriver ensuite ?

Ex :

Vous voyez ici (avec ChatGPT) qu'il a parfaitement prédit le mot que j'attendais, car dans ce contexte, les probabilités du mot « bœufs » ont été les plus grandes.

Bien-sûr pour arriver à ce résultat, il est nécessaire de passer par plus d'étapes que la tokenisation seul (rien à voir avec les RWA mdr).

Car un token est simplement un bout de phrase, mais comme vous le savez peut-être, un ordinateur ne traite que des nombres. Il va donc falloir transformer ces tokens en nombres à moment donné pour pouvoir les traiter et calculer des probabilités.

C'est pourquoi chaque token reçoit d'abord un ID numérique unique pour être indexé par le modèle de langage.

Là encore, on est loin d'avoir assez d'informations pour prédire le prochain token, l'ID numérique ne suffit pas. On a besoin d'outils mathématiques plus puissants pour capturer toute la signification sémantique de chaque token, ainsi que le contexte et les relations entre les différents concepts.

On rentrera dans l'aspect technique et algorithmique dans un prochain rapport, mais pour que vous ayez un avant-gout, chaque token passe par une phase appelé la *Vectorisation.* C'est le passage d'ID numérique à vecteur.

Si le mot vecteur vous fait grincer des dents, pas de panique, j'essayerai de l'expliquer de la façon la plus simple dans un prochain rapport. Le but c'est que tout le monde puisse comprendre, pas besoin d'un bac+10 en algèbre linéaire 😷

Pour cette fois, sans rentrer dans les détails, cette représentation vectorielle permet de capter bien plus de data sur un token qu'un simple identifiant numérique. Ce n'est plus seulement un chiffre pour identifier un mot, mais véritable espace multidimensionnel où les mots proches en sens sont aussi proches géométriquement, un monde où la poésie croise l'algèbre linéaire, ça a l'air sympa hein ? 😬

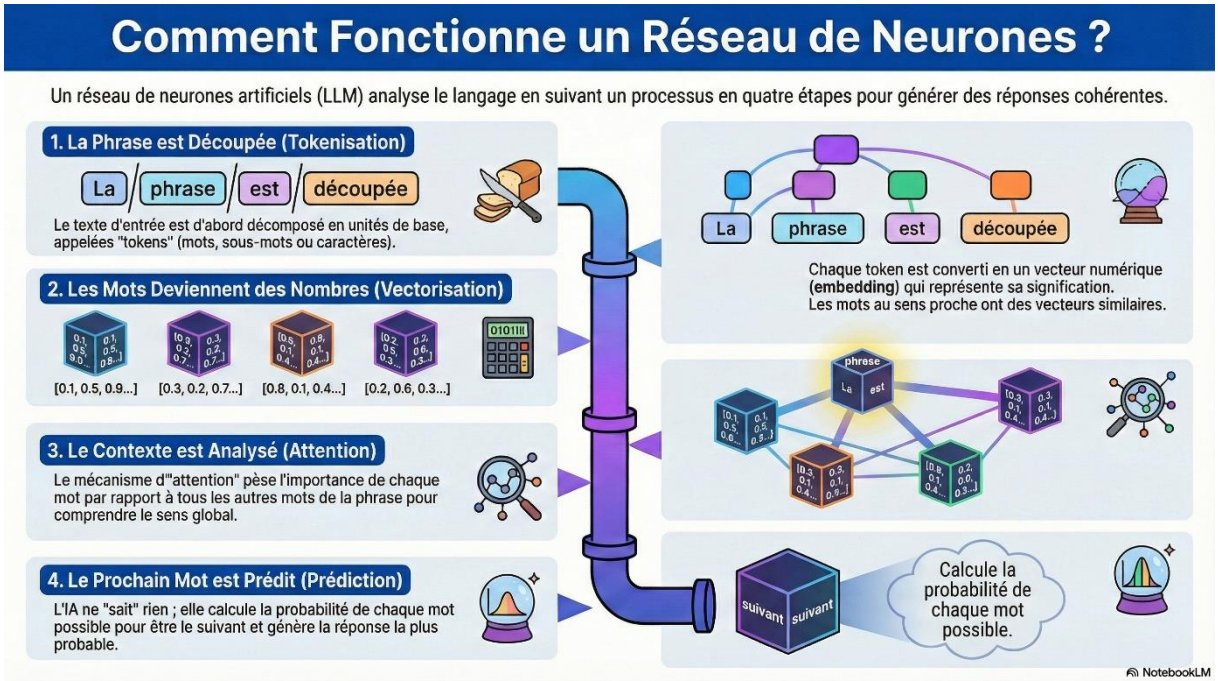
Pour terminer sur le fonctionnement global d'un LLM, il est essentiel de parler de la notion de *Tranformer* et du mécanisme d'*Attention*, la vraie révolution. Encore une fois, on aura l'occasion de s'attarder du l'aspect algorithmique plus tard pour vraiment comprendre deep comment fonctionnent ces technologies, mais pour l'heure essayons déjà d'avoir une vue d'ensemble sur ces technologies.

Contrairement aux réseaux de neurones profonds type RNN (Recurrent Neural Network) pour réseaux de neurones récurrents qui se contentent d'analyser une phrase de façon séquentielle, un Transformer lui, permet un traitement parallèle de toute la phrase, ce qui améliore grandement sa fiabilité. Cette technologie permet d'introduire le mécanisme d'Attention, qui permet d'attribuer plus de poids à certains mots (les plus importants de la phrase) pour éviter des erreurs syntaxiques. Cette technologie est indispensable quand on sait qu'un même mot peut avoir plusieurs sens.

Exemple :

« J'ai une montre en or. »

« Montre-moi le chemin. »



Maintenant que vous savez dans les grandes lignes comment fonctionne un LLM, il important de noter que GPT (le modèle derrière ChatGPT) n'est pas le seul modèle grand publique disponible. Entre GPT d'OpenAI, Claude d'Anthropic, Gemini de Google, Grok d'xAI, DeepSeek, etc… il peut paraitre difficile de s'y retrouver.

Pourtant, si vous ne vous contentez pas seulement d'utiliser votre modèle préféré mais d'adapter le modèle de langage en fonction de la tâche que vous allez attribuer à celui-ci, vous avez déjà un avantage sur plus de 90% des gens.

Les trois principaux modèles de langages et ceux que j'utilise au quotidien sont : GPT, Claude et Gemini. Vous pouvez vous aider d'autres modèles, mais attention à ne pas faire trop complexe, les trois modèles que je viens de citer sont à ce jour les meilleurs dans leurs domaines de prédilection respectif.

A ce stade, la question que vous devez vous poser est la suivante : comment savoir quel modèle choisir pour quelle tâche ? Pour répondre à cette question, vous pouvez utiliser cet indicateur : https://lmarena.ai/fr/leaderboard

Ce site marche selon un principe de vote entre plusieurs réponses pour des tâches prédéfinies (Texte, développement web, etc..). Ce qui est intéressant c'est que l'utilisateur ne sait pas à l'avance quel modèle a généré quelle réponse, ce qui permet de garder une certaine objectivité. De plus, les derniers modèles de chaque entreprise sont testés en temps réel (gemini 3 pro, claude opus 4.5, etc…), donc vous pouvez revenir de temps en temps et adapter les tâches que vous souhaitez déléguer au modèle le plus adapté.

A noter que selon cet indicateur, gemini serait devenu plus polyvalent que GPT. Cela reste à nuancer mais on aura l'occasion de revenir sur les dernières innovations de Google en matière d'IA dont des applications surpuissantes de leur dernier modèle gemini 3, vous n'êtes pas prêt haha.

Merci d'avoir tout lu, encore une fois si vous avez des questions ou si vous avez des suggestions pour le prochain rapport, n'hésitez pas à me mentionner ou à me DM. La prochaine fois on traitera la notion de prompt engineering pour maximiser les réponses des chats bots. Passez une excellente semaine les Taurus et à bientôt 👏