Rapport IA n°3 - Maîtriser le Context Engineering

28/01/26

(100% rédigé par Jared)

0/ Introduction :

Salut les Taurus, j'espère que vous avez la forme cette semaine. Pour ce que nouveau rapport, on va traiter le sujet du context engineering, son lien avec le prompt engineering, et comment vous pourrez l'appliquer dès aujourd'hui à vos process de plusieurs manières. On verra entre autres la méthode classique d'ingénierie du contexte que vous appliquez peut-être déjà sans le savoir, puis un autre outil plus avancé qui, si vous apprenez à bien l'utiliser, va littéralement révolutionner votre manière d'apprendre et de comprendre des concepts avancés peu importe votre domaine tout en vous faisant gagner énormément de temps. C'est, à ce jour, l'outil le plus sous-coté que j'ai testé, tout en étant le meilleur pour apprendre n'importe quoi. Que vous soyez trader, investisseur, entrepreneur, salarié, même étudiant peu importe, cet outil va forcément vous servir, vous allez voir, c'est du pain bénie (et personne n'en parle 😅).

D'autant plus que si vous avez lu les précédents rapports (surtout le 1er), vous allez voir que vous connaissez déjà les concepts qu'on va aborder (juste le nom sera nouveau), et c'est ça qui est génial avec ce format je trouve, on construit ensemble de vraies connaissances au fil des semaines qui vont être réutilisées pour tirer meilleur parti de la révolution qui arrive.

J'en profite pour dire que si jamais vous avez des idées de sujets que vous aimeriez qu'on aborde dans les prochaines semaines (peu importe le thème que ce soit technique, sociétal, voire philosophique), n'hésitez pas à me le dire en DM, c'est avec plaisir qu'on pourra aborder ces sujets ensemble.

Sommaire :

1/ Context engineering : les trois piliers

a. Faut-il enterrer le prompt engineering ?

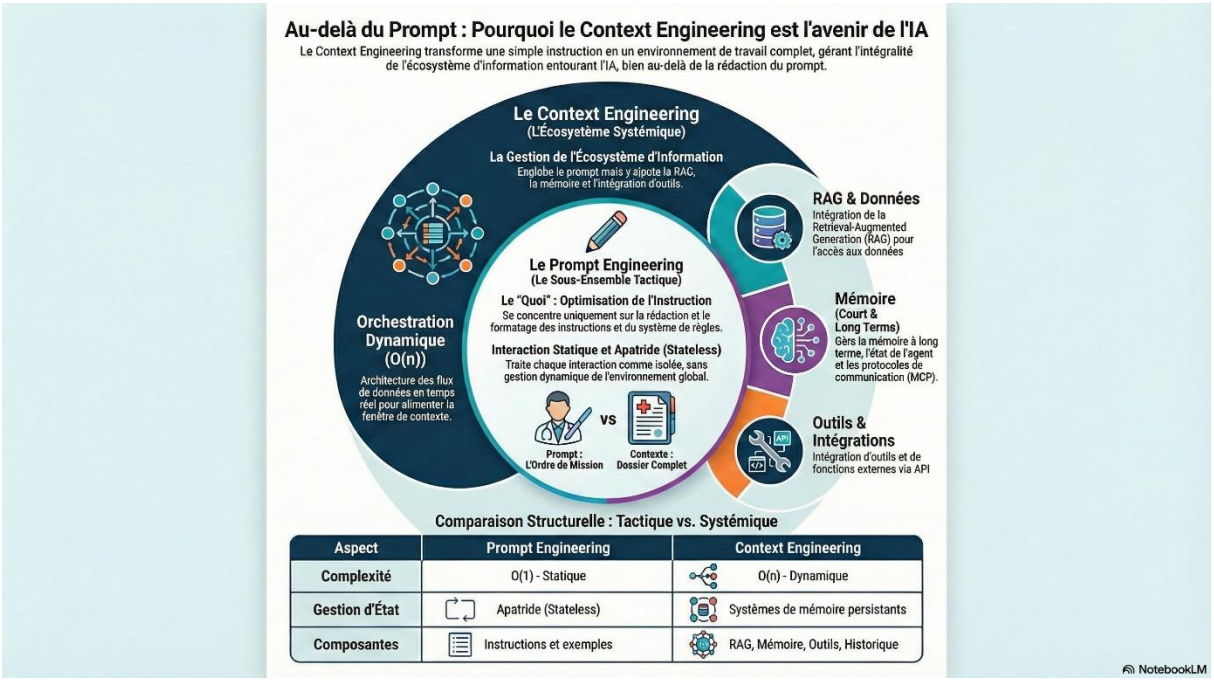
Comme introduit à la fin du deuxième rapport, voyez vraiment le contexte donné LLMs comme le carburant d'un moteur. Le moteur étant le prompt engineering. Vous aurez beau utiliser les meilleures techniques d'ingénierie des prompts, votre réponse restera assez générique. Imaginez que l'IA ait absolument toute la data vous concernant (en mettant de côté les questions de confidentialités haha), ses réponses seront dès lors bien plus adaptées à qui vous êtes, à votre situation professionnelle, etc.

Au-delà du modèle de langage que vous utilisez, la métrique numéro 1 sur laquelle vous pouvez jouer pour améliorer drastiquement les réponses de l'IA est le contexte. Cela dit, il ne faut pas croire que l'ingénierie du contexte se limite à envoyer le plus de data possible à l'IA vous concernant, car ce n'est qu'une partie de l'équation. De même, penser que le prompt engineering est devenu obsolète avec la démocratisation du context engineering reviendrait à ne pas comprendre que l'ingénierie des prompts n'est rien d'autre qu'un sous-ensemble de l'ingénierie du contexte.

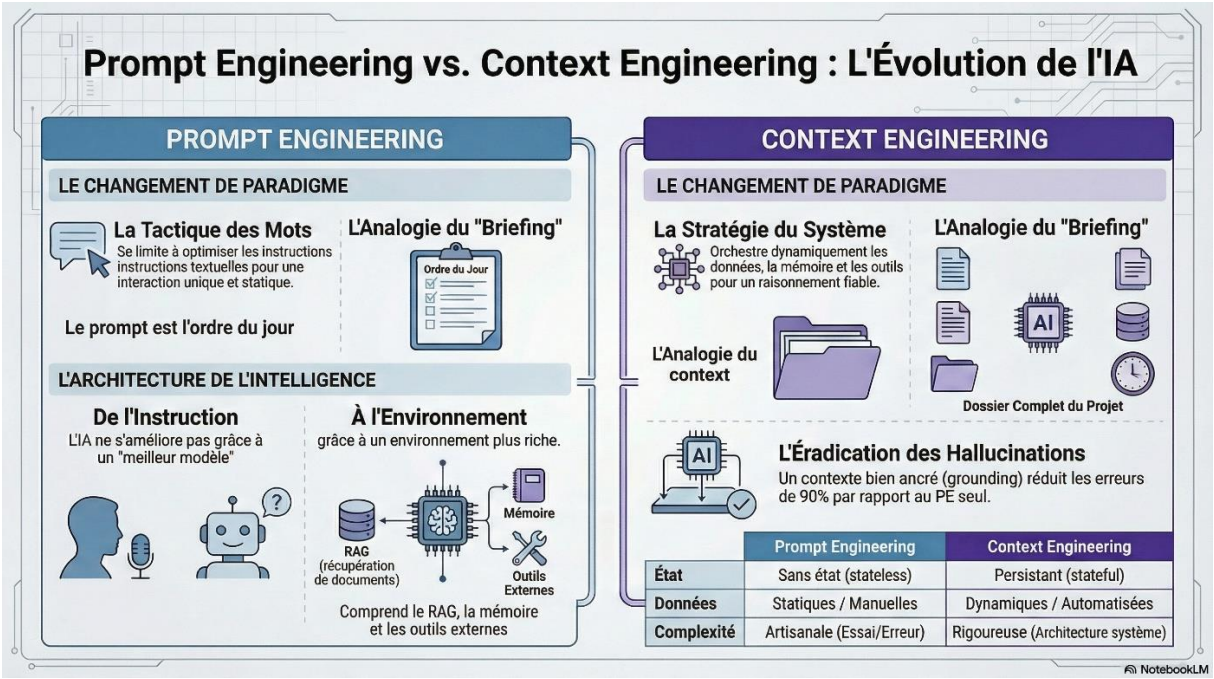
C'est vraiment un point très important donc j'insiste vraiment là-dessus, voyer le context engineering comme une grande famille à laquelle appartient le prompt engineering.

Si vous voulez tirer meilleur parti des modèles de langages d'aujourd'hui et de demain, il ne faut pas soit faire du prompt engineering, soit donner un maximum de contexte au modèle, il faut faire les DEUX.

D'ailleurs, le concept de context engineering englobe ces deux concepts comme l'illustre cette image :



Donc même en mettant votre focus sur la qualité du contexte, il est toujours mieux de continuer le prompt engineering (role prompting, few-shot prompting, CoT, etc…).



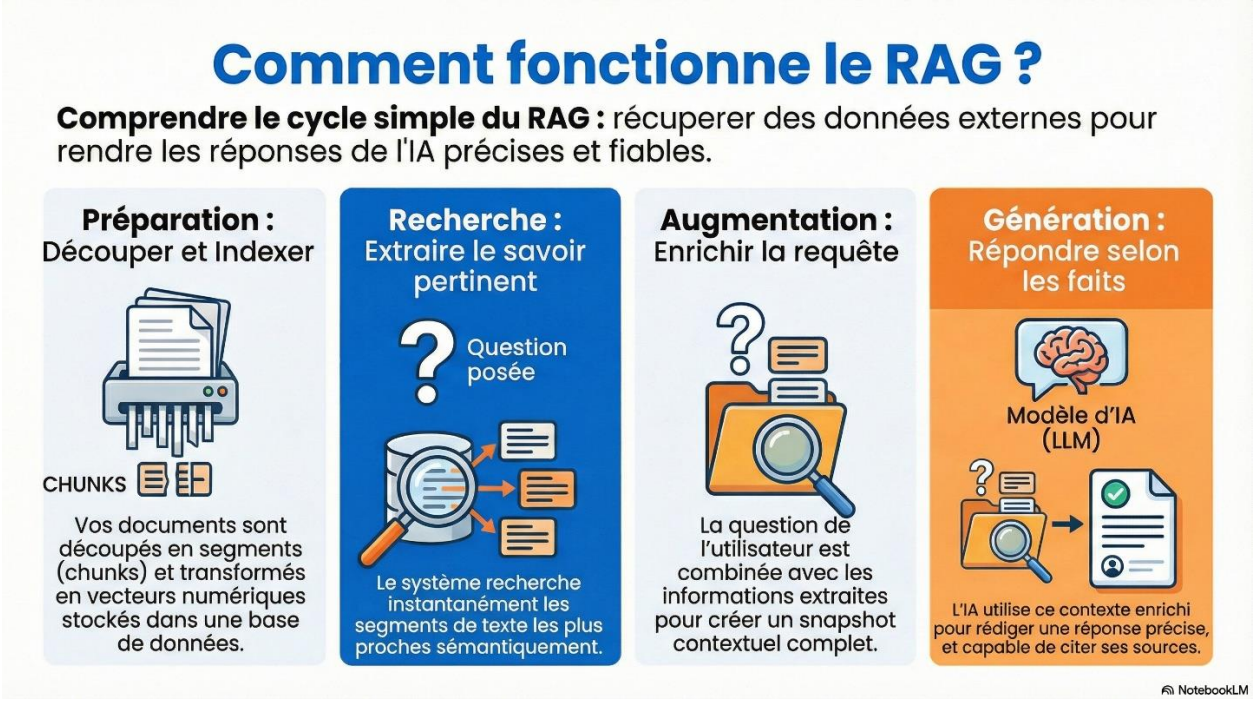
b. Le RAG

Maintenant que vous comprenez mieux l'importance de travailler sur le contexte que vous donnez à vos LLMs, il s'agit de comprendre comment améliorer celui-ci.

C’est ici qu’on va introduire le concept de RAG (Retrieval-Augmented Generation), ou génération augmentée par récupération. Si vous avez lu le 1er rapport sur les LLMs, vous allez très vite voir qu’un RAG ressemble beaucoup à l’entrainement des LLMs. Seulement contrairement aux LLMs, un RAG n’est pas un réseau de neurones, c’est une base de données externe sur laquelle va s’appuyer un LLM.

Il faut voir le RAG comme une bibliothèque dans laquelle un LLM va piocher des informations pour améliorer ses réponses. Vous pourrez par exemple y mettre des documents vous concernant, ou concernant votre entreprise, etc…. Le but est de lui donner un context précis sur lequel il pourra se baser pour l’ensemble de ses réponses. C’est une mémoire externe vouée à alimenter un LLM.

Cette bibliothèque n’est autre qu’un espace vectoriel (oui encore) que vous allez pouvoir alimenter en mettant des vecteurs dedans. Des outils comme Supabase permettent de faire ça, vous commencez par placer les documents que vous souhaitez vectoriser dans une base de données. Ensuite un LLM s’occupera de « chunker » vos documents, ou les segmenter en français. En gros ça revient à les découper en morceau de phrase avant de les vectoriser. Viens ensuite l’étape de l’embedding où les chunks sont vectorisés, puis placés dans un espace vectoriel. Comme pour les LLMs, les vecteurs proches en sens le sont également géométriquement (chien et chat par exemple) ainsi, on obtient une sorte mémoire externe prête à l’emploi.



La question posée est aussi segmentée puis « embeddée » et placée dans l’espace vectoriel, c’est comme ça que le modèle de langage peut trouver les vecteurs les plus proches de la question pour enrichir sa réponse.

2/ Comment tirer le meilleur parti de l’ingénierie du contexte ?

a. Entrainer la mémoire à long terme (méthode de base)

Le RAG est ce qui permet d’entraîner la mémoire à long terme des IAs. Chaque information qu’elle apprend sur vous est ensuite vectorisé et placé dans l’espace vectoriel.

Tout à l’heure j’ai parlé d’outils comme Supabase pour créer des espaces vectoriels. Sachez qu’on en aura besoin pour faire des agents IA, mais c’est encore un autre sujet.

Donc pas besoin de passer par des applications annexes pour l’instant, vous pouvez simplement converser avec les LLMs pour qu’ils développent cette fameuse mémoire à long terme. Veuillez à bien avoir activé l’option qui permet à l’IA d’utiliser la mémoire et de citer vos précédentes conversations (voir le rapport n°2).

Cette méthode est simple et efficace mais elle a un gros défaut : la durée.

En effet développer la mémoire à long terme de l’IA de cette manière est très long et fastidieux. Si vous étalez ça dans le temps avec des interactions quotidiennes, ça passe encore. Mais vous avez peut-être besoin d’un contexte bien spécifique maintenant, et c’est là que la deuxième méthode entre en jeux.

b. NotebookLM : l’outil d’ingénierie du contexte le plus sous-coté

i. Introduction à NotebookLM

Google a encore frappé (•_•).

Si vous n’avez jamais entendu parler de NotebookLM, c’est normal car j’ai l’impression que ça devient une habitude chez Google de sortir des dingueries qui passent totalement sous les radars (comme Anti-Gravity qu’on verra plus tard).

Pour vous introduire cet outil, imaginez que vous puissiez dans une seule et même interface, façonner le contexte avec lequel vous allez travailler : que ce soit des fichiers, des pages internet, des vidéos youtube (extrêmement puissant), des Google drives entiers et j’en passe. Ensuite, que vous puissiez à partir de ces sources, discuter librement avec Gemini (meilleur LLM aujourd’hui), et lui poser des questions par rapport aux sources que vous avez connectés.

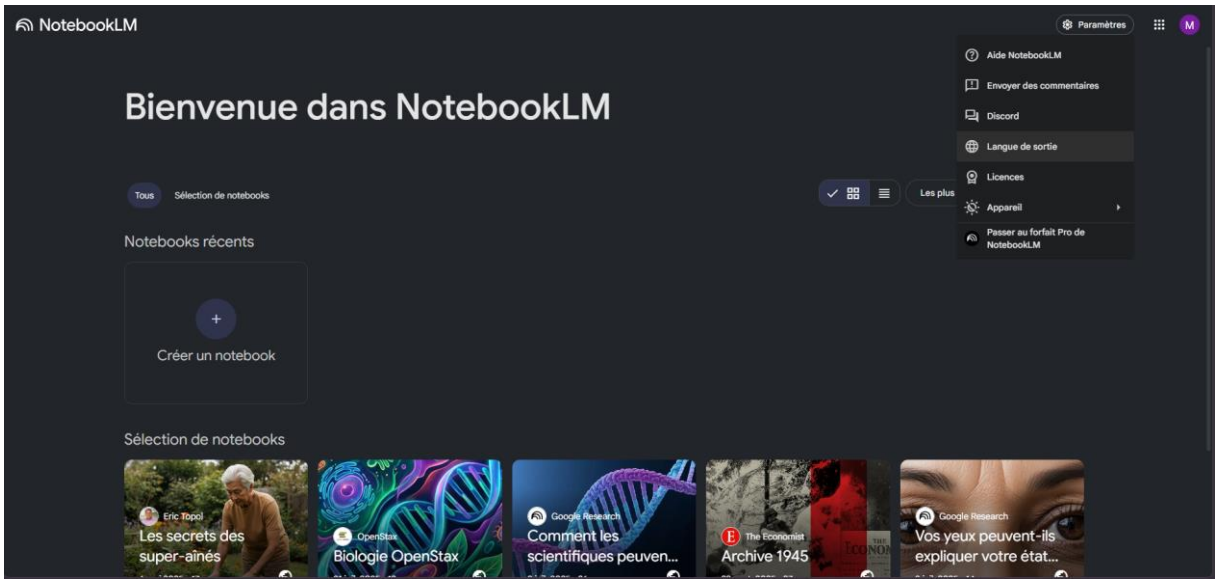
Bon déjà là c'est pas mal mais attendez, vous n'avez rien vu mdr. Imaginez que, toujours à partir de ces sources, vous puissiez créer des podcasts personnalisés, des vidéos explicatives personnalisées, des cartes mentales, des slides, des quizz, des infographies (voir celles plus haut), et j'en passe. Non ce n'est pas une blague et c'est bluffant vous allez voir. Le pire c'est que c'est dispo avec le plan gratuit (vous êtes limité à 50 sources donc ça va).

Peu importe votre situation, cet outil va forcément vous servir, c'est pourquoi il est primordial d'apprendre à l'utiliser au plus tôt (vous allez voir, c'est hyper intuitif). Regardons ensemble comment ça marche avec un cas concret d'utilisation appliqué à l'analyse fondamentale.

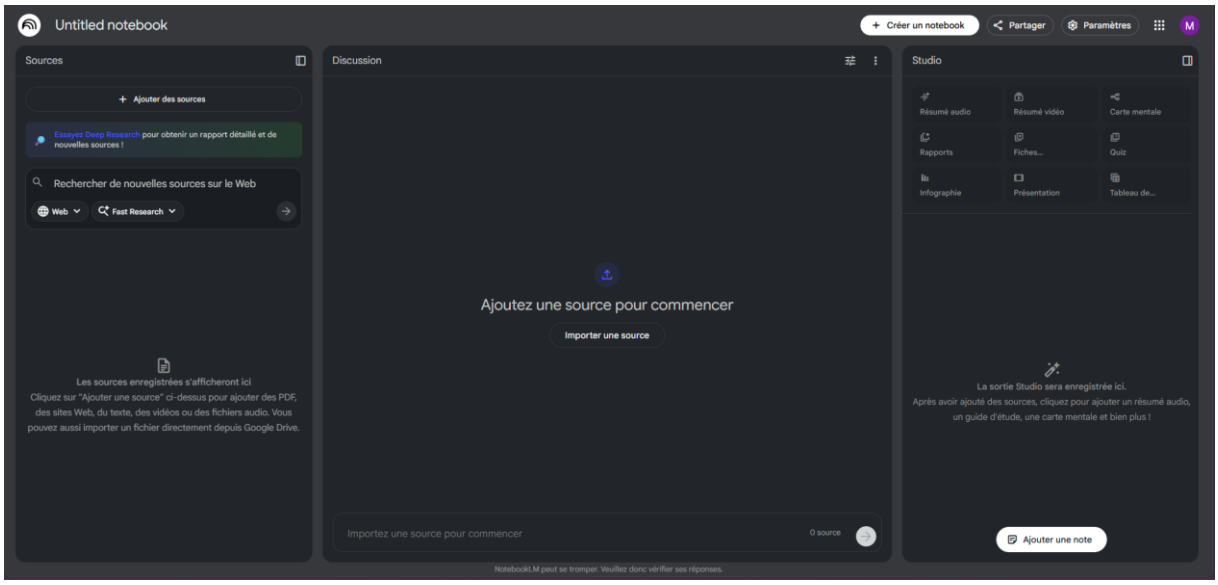
iii. Cas concret d'utilisation

Bon pour l'exemple je ne me suis pas foulé, Grok m'a proposé un thème de géo-po d'actualité : **Trump et la bataille pour le Groenland + la guerre commerciale avec l'Europe**.

Disons qu'on parte là-dessus, rendez-vous sur https://notebooklm.google.com/ et veuillez à configurer la langue de sortie en français.



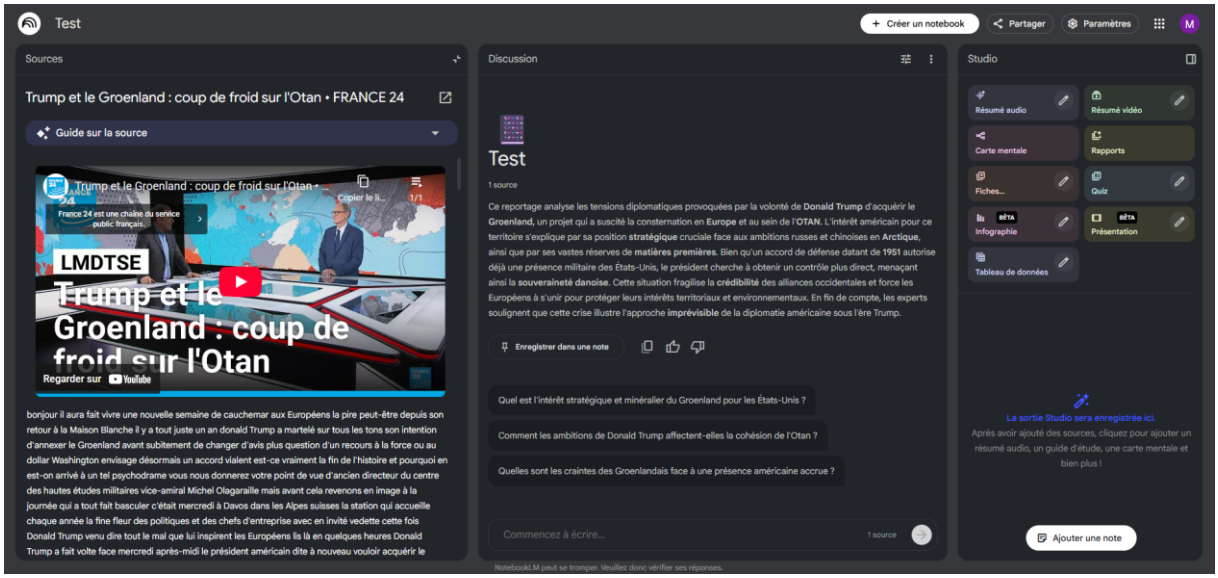
Cliquez ensuite sur « créer un notebook », vous arrivez sur cette interface :



L'interface parait intimidante mais c'est super simple vous allez voir. Elle est décomposée en trois onglets :

- Sources : là où vous allez pouvoir faire de l'ingénierie du contexte
- Discussion : là où vous allez pouvoir discuter avec votre Gemini munit du contexte
- Studio : là où vous allez pouvoir faire des vidéos, podcasts, flash cards, etc…

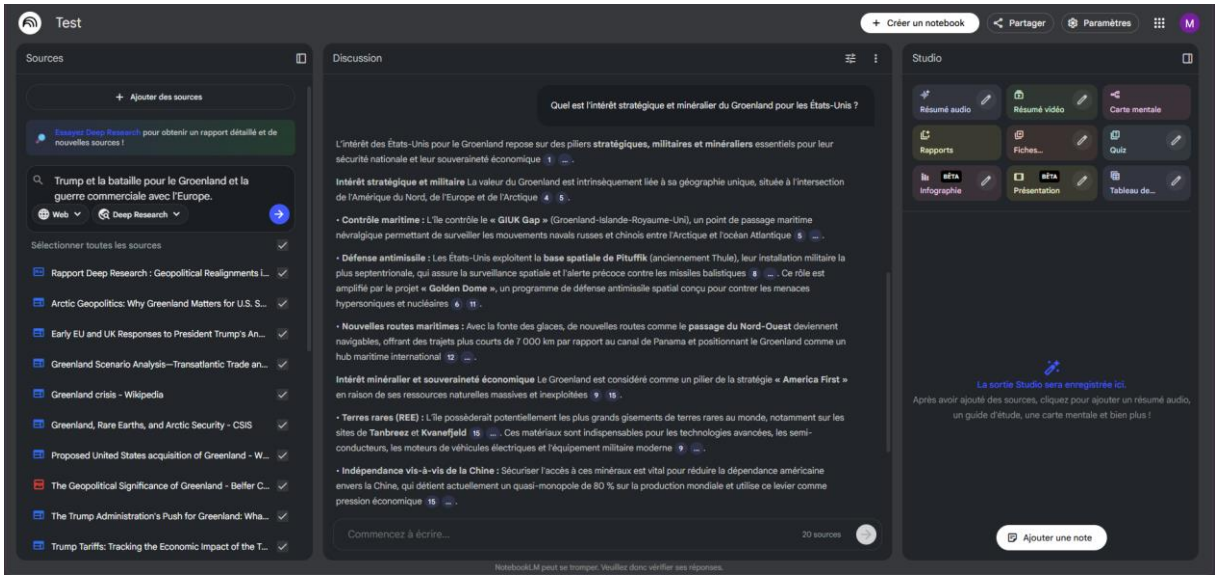
Dans les sources, vous pouvez soit ajouter vos sources à la main, soit laisser le modèle faire une recherche sur le web ou dans un drive à partir d'un prompt. De ce que j'ai pu tester, le top c'est d'importer surtout des vidéos youtube (1ère plateforme d'apprentissage), car vous aurez les retranscriptions exactes (là j'ai pris une vidéo random pour l'exemple mais tachez de faire ça bien).



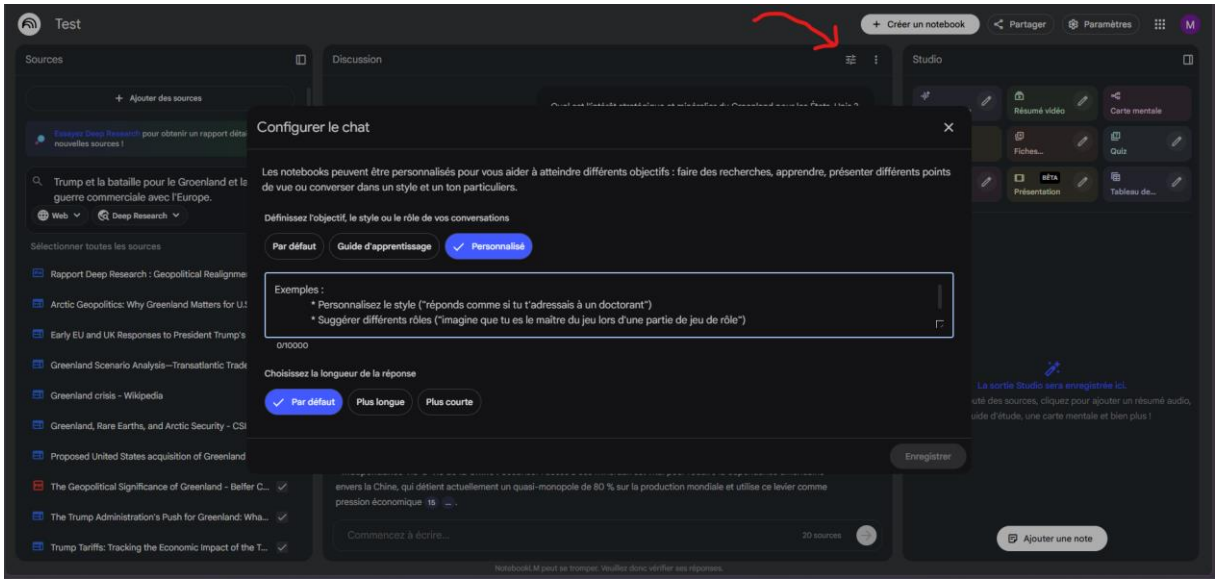
Vous avez la possibilité d'importer jusqu'à 50 sources avec le plan gratuit donc vous avez de la marge. Vous pouvez aussi faire une recherche directement avec l'IA

(préférencier une deep search) pour gagner du temps. C'est ce que je vais faire ici, à noter qu'une deep search peut prendre quelque minutes à charger.

Une fois que vous avez selectionné toutes vos sources, vous pouvez commencer à échanger avec Gemini via l'onglet discussion :



Ses réponses seronts basées sur vos sources. A noter que si vous cliquez en haut à droite de l'onglet discussion sur les paramètres, vous pouvez ajouter un system prompt ou un mode préfait ainsi que la longueur des réponses :

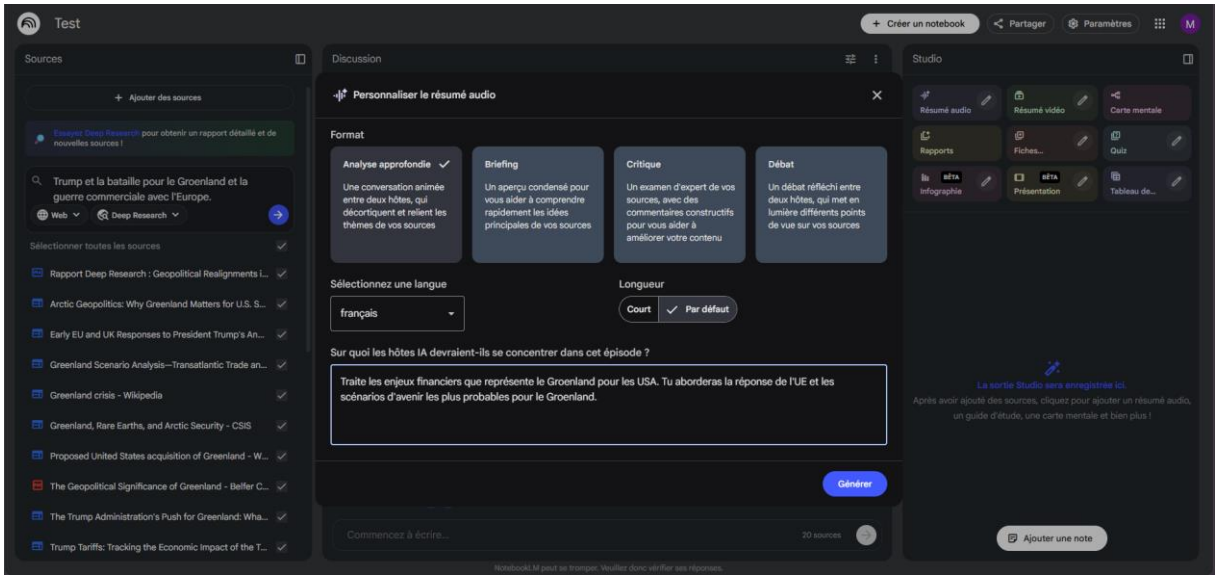


Bon tout ça c'est bien beau mais maintenant regardons le plus intéressant : l'onglet studio. L'énorme avantage de NotebookLM, c'est que vous avez pas besoin de jonglet entre Nano Banana, Veo, Gemini, etc pour vos requêtes (souvent le problème avec l'écosystème Google), non là tout est mis dans une même interface.

Essayons l'option « Résumé audio » pour faire un podcast. Vous pouvez soit cliquer direct, auquel cas il vous fera un podcast assez générique sur vos sources. Le mieux

est de cliquer sur le petit crayon pour ajouter un prompt et ainsi orienter le podcast vers ce qui vous intérrese vraiment.

Testons ce prompt :



Voila pas besoin de faire plus compliqué, il suffit d'orienter le modèle dans la direction que vous souhaitez. Comptez quelques minutes de génération pour les podcasts et les vidéos. Voici le résultat :

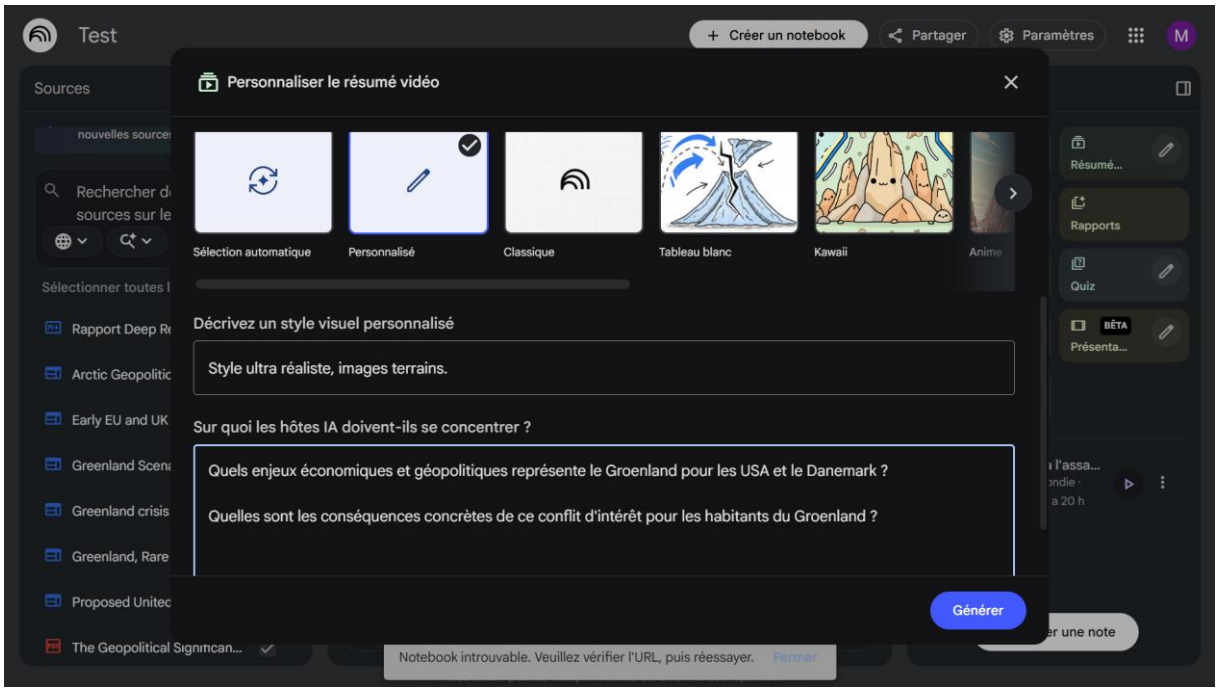
https://drive.google.com/file/d/1yNGNHVgn2joVfPYaCjPAlwnfe9ScahnM/view?usp=drive_link

Pas mal hein ? Vous pouvez maintenant créer des podcasts entièrement personnalisés sur des sujets ultras précis. Quelle époque formidable :D

Bon ok ça vaut pas encore un vrai podcast, mais là ce qui est dingue c'est que vous pouvez choisir un sujet sur lequel vous galérez, et avoir une explication personnalisée de 15 mins dessus :O

Essayons la vidéo maintenant, là encore vous pouvez soit faire une vidéo générale sans prompt, soit tout configurer, des aspects abordées aux images de fond de votre vidéo.

Voici les prompts (là encore pas besoin d'aller très loin) :



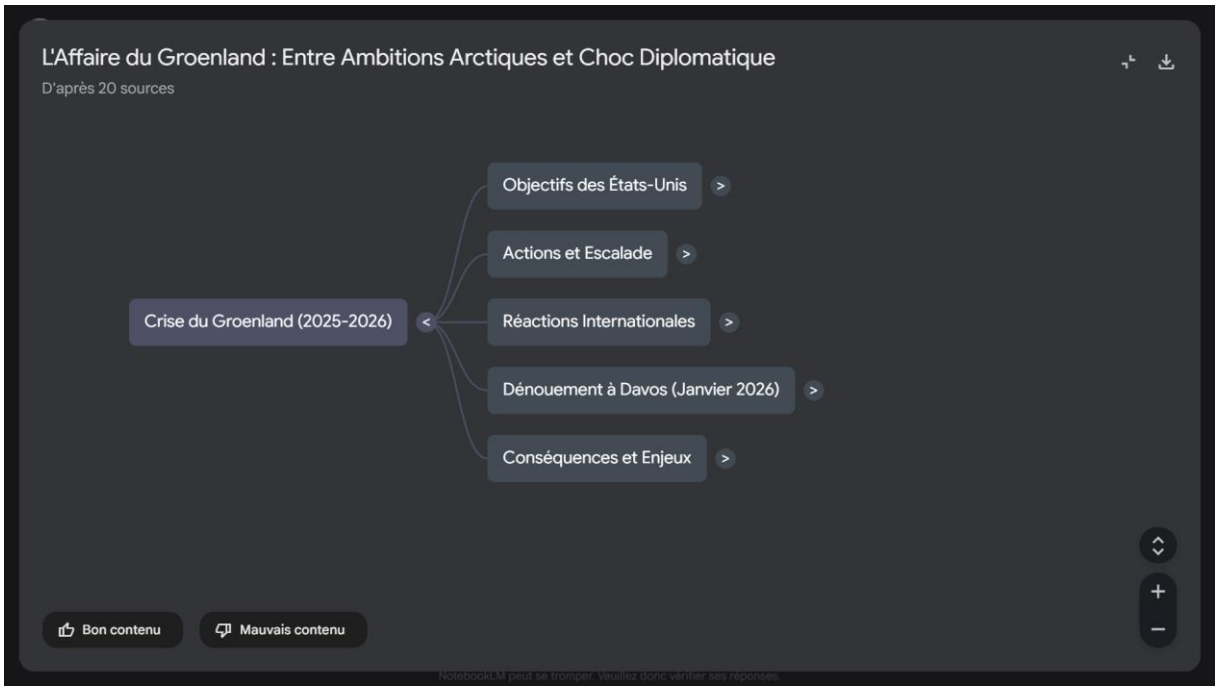
Comptez plusieurs minutes pour la génération d'une vidéo. Autre observation personnelle, j'ai l'impression que le format podcast permet de traiter le sujet donné plus en profondeur que le format vidéo (sûrement pour des raisons de consommation plus importante de tokens). La vidéo est parfaite pour saisir le gros d'un sujet, mais si vous voulez aller plus en prodondeur je vous conseille le podcast. Voici la vidéo qu'il m'a faite :

https://drive.google.com/file/d/1awNtGMcd22VsbVqudl_9wQC1l1B5UVc0/view?usp=drive_link

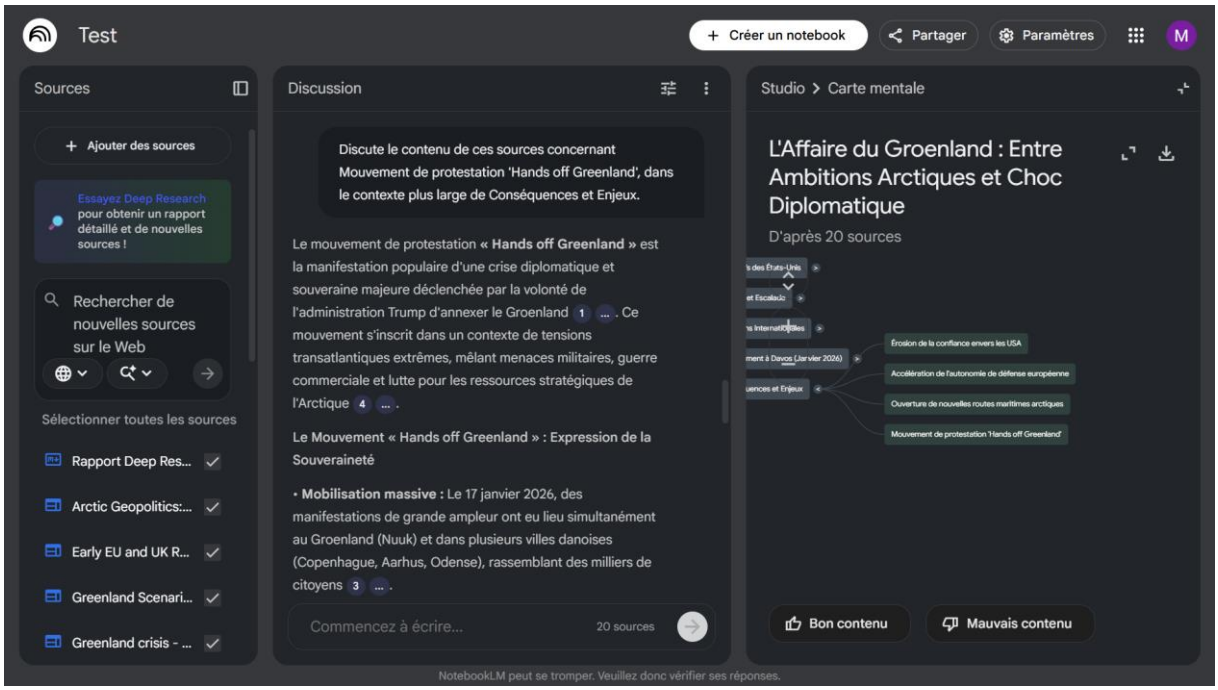
C'est peut-être moins impressionnant que le podcast, mais ça reste vraiment pas mal et j'imagine pas ce qu'on pourra faire dans 2 ans ça va être incroyable :O

Pour ce que est des infographies, vous pouvez voir celles que je met dans ces rapports, elles sont aussi générée par NotebookLM.

Je voulais vraiment vous montrer l'option carte mentale aussi car c'est vraiment géniale pour avoir une vue d'ensemble sur un sujet :

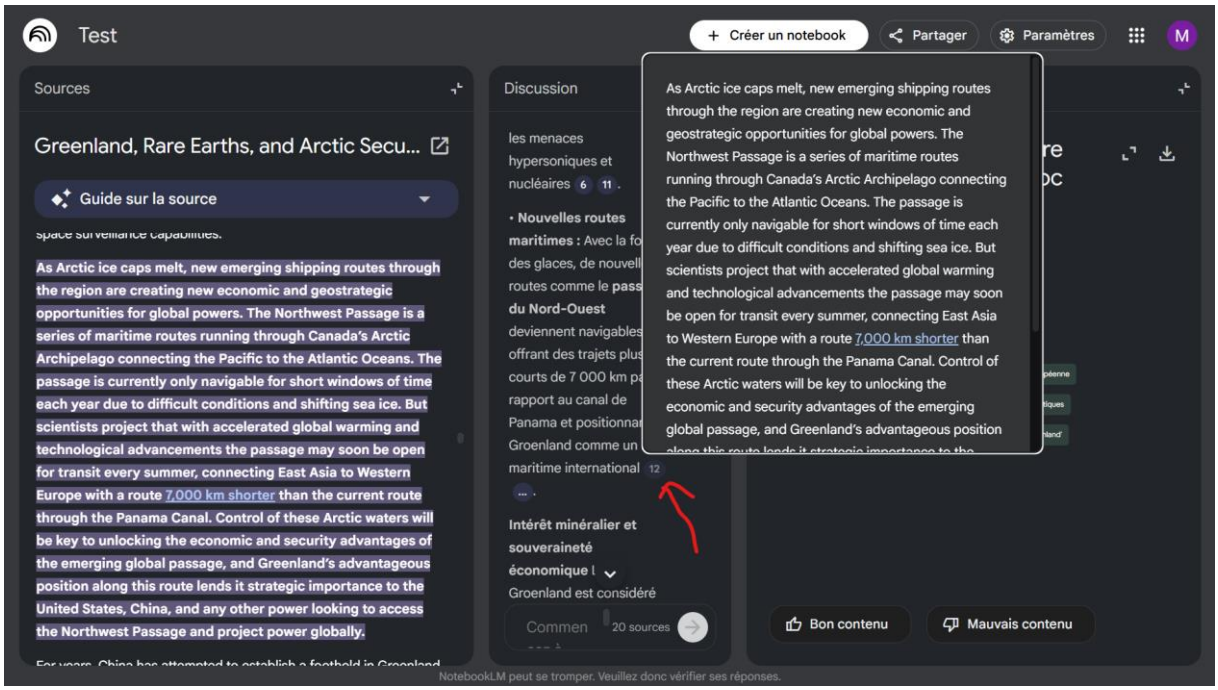


Voila ce qu'il m'a généré, on peut cliquer sur les flèches pour développer une thématique, et quand on veut en savoir plus sur un point précis, il suffit de cliquer dessus :



Par exemple ici j'ai cliqué sur « Mouvement de protestation Hands off Greenland » et Gemini m'a généré une réponse automatiquement. De même, quand vous cliquez sur les numéros, il vous renvoit directement à la source citée (vous comprenez mieux

pourquoi cet outil est utilisé par les chercheurs) :



Bon on va pas tous les faire ici, je pense que vous avez déjà pu voir le potentiel de cet outil avec les quelques exemples qu'on a fait. Libre à vous de tester les autres outils du studio, vous avez toutes les clés pour en tirer meilleur parti. A vous maintenant de voir comment vous allez pouvoir l'intégrer à votre activité ヽ(*°▽°*)ノ

Merci d'avoir lu cet article jusqu'au bout, n'hésitez pas à me faire un retour aussi bien de cet article que de NotebookLM. Maintenant que vous avez un bon bagage pour intéragir au mieux avec les IAs, vous voyez qu'on commence à rentrer dans du concret avec NotebookLM. J'ai pas encore choisi le sujet de la semaine prochaine, donc si vous avez des idées ou si il y a des zones d'ombre que vous souhaitez qu'on aborde, n'hésitez pas à me passer le message en privé ᓚᘏᗢ