Comprendre les Grands Modèles de Langage (LLMs)

Une exploration visuelle de la mécanique interne de l'IA, inspirée par 3Blue1Brown.

Appuyez pour commencer l'analyse >_

L'analogie du scénario inachevé

Humain :
Peux-tu expliquer...
un semi-conducteur...

>_

est un matériau | fonctionne comme

Imaginez que vous trouviez un script de film déchiré. Vous avez la question, mais la réponse est manquante. Si vous aviez une machine capable de deviner le mot suivant le plus logique, vous pourriez reconstruire la scène, mot après mot.

C'est exactement ce que fait un LLM : il prédit la suite logique d'une séquence de texte.

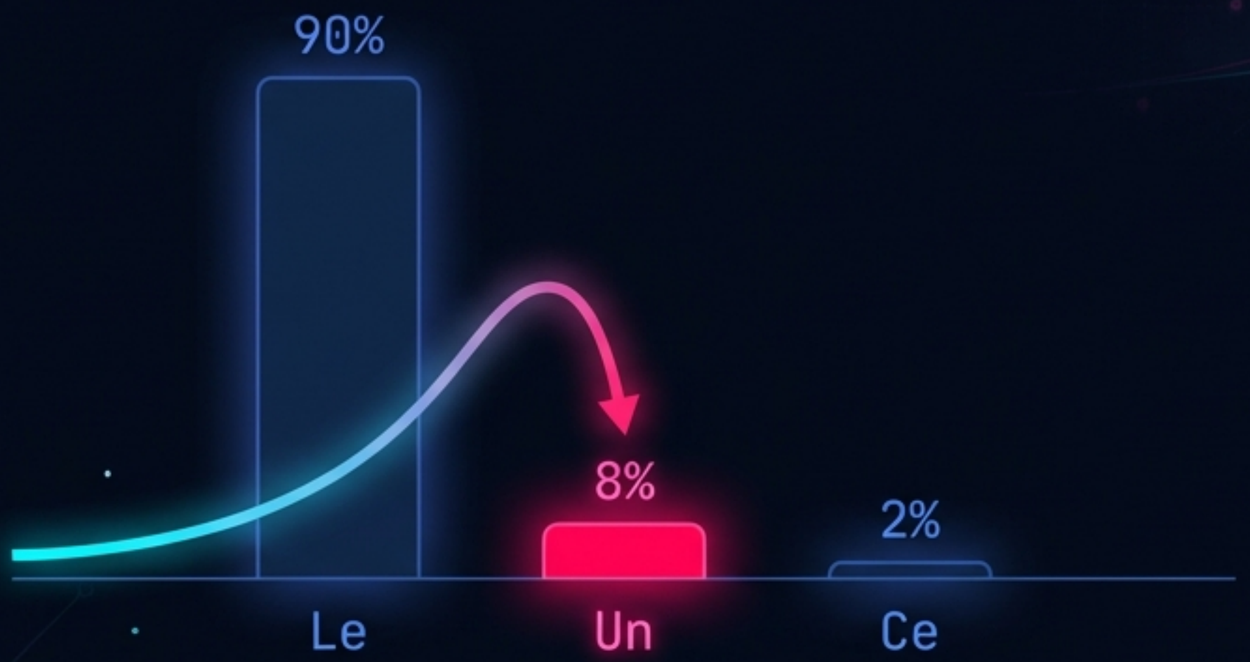
Une fonction mathématique, pas un cerveau

Un LLM est une fonction mathématique sophistiquée. Il ne « sait » rien. Il attribue une probabilité à tous les mots possibles qui pourraient suivre le texte actuel.

Au lieu d'une réponse unique, le modèle génère une « distribution de probabilités ».

Le hasard contrôlé crée la fluidité

Si le modèle choisissait toujours le mot le plus probable, il sonnerait robotique. Pour paraître naturel, il sélectionne parfois des mots moins probables au hasard.



« Le modèle est déterministe, mais le résultat varie. C’est ce qui donne l’illusion de la créativité. »

Une échelle de lecture inhumaine

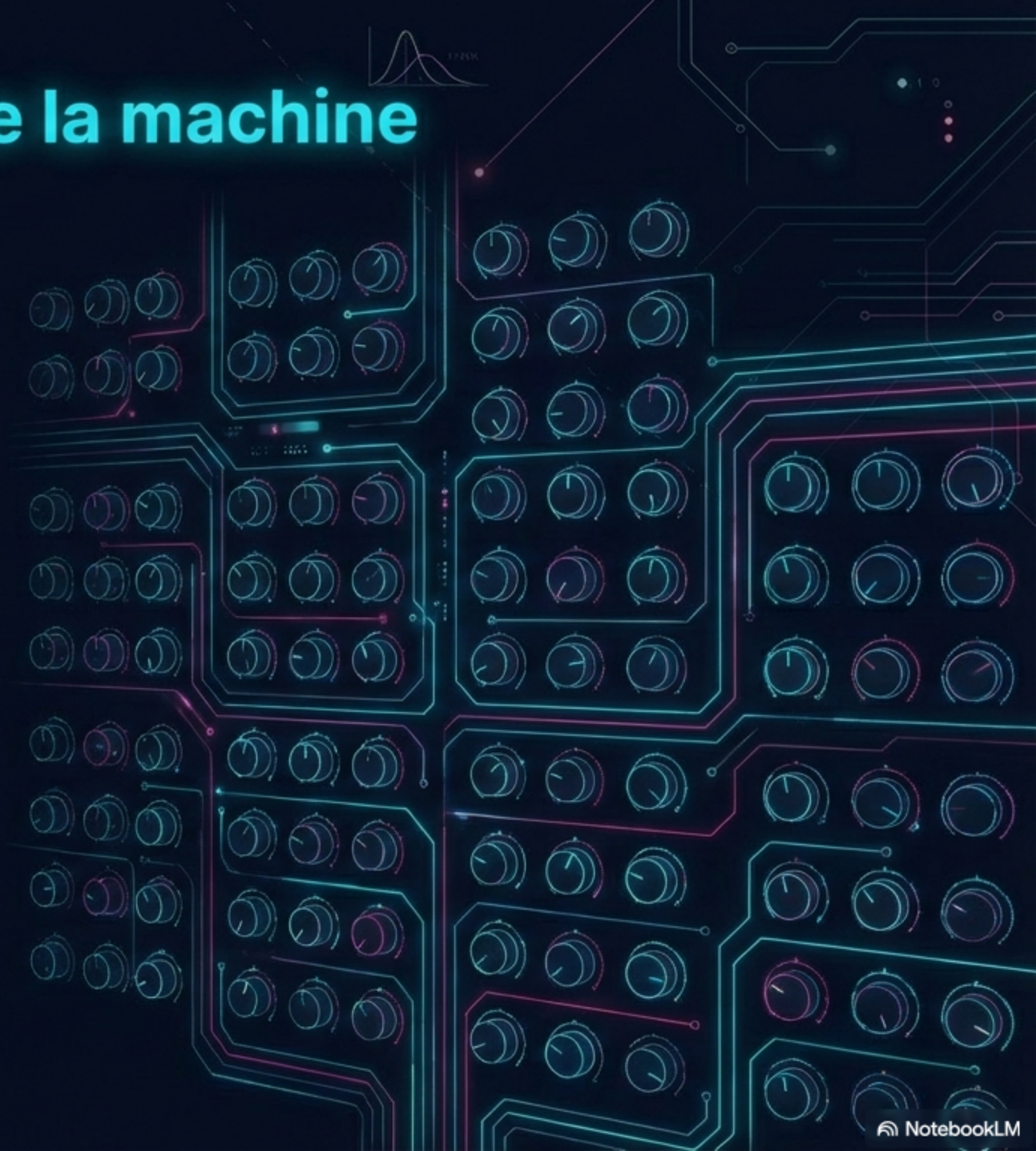
Les modèles apprennent en ingérant des quantités massives de texte provenant d'Internet. Pour lire la quantité de données utilisée pour entraîner GPT-3, un humain devrait lire 24h/24 et 7j/7 pendant plus de 2 600 ans.

*GPT-4 a été entraîné sur encore plus de données.

Les milliards de cadrans de la machine

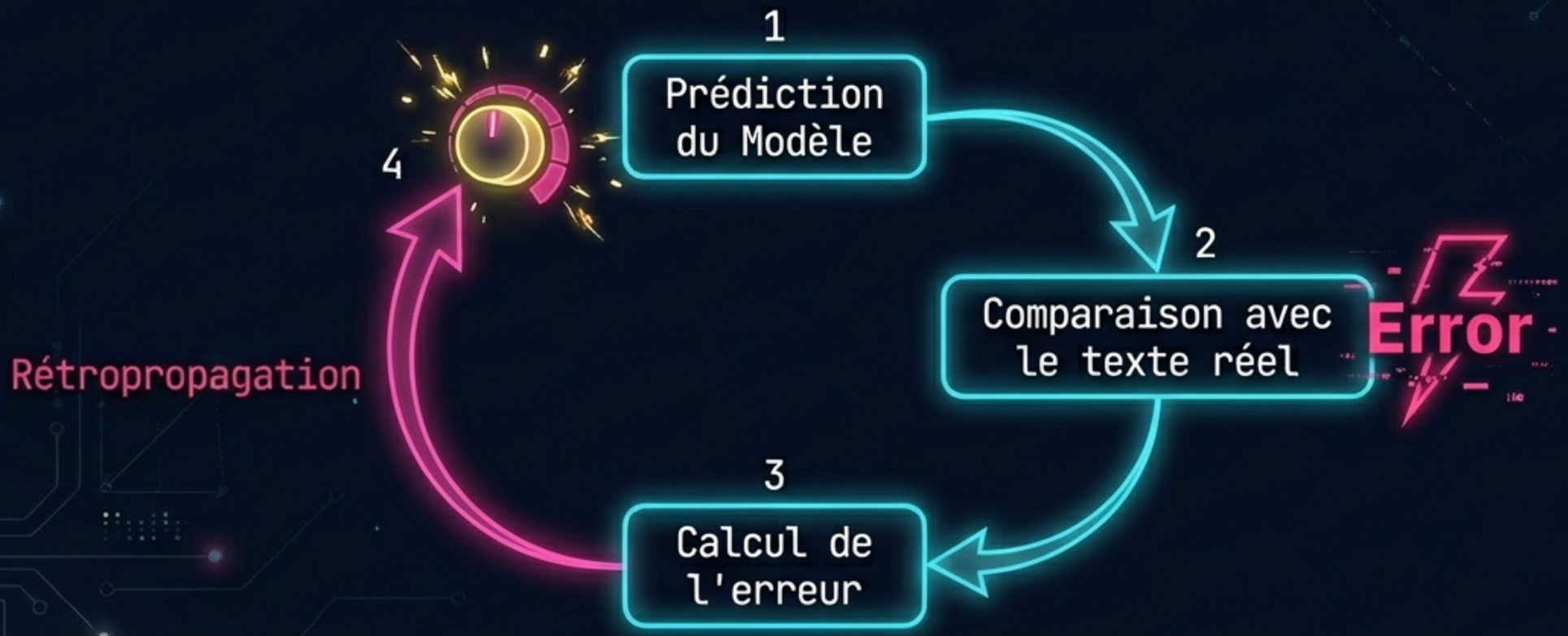
L'entraînement consiste à régler les « paramètres » (ou poids) du modèle. Ce sont comme des petits cadrans qui déterminent comment l'entrée est transformée en sortie.

Ce qui rend un modèle « Large » (LLM), cest qu'il possède des centaines de milliards de ces paramètres.



La Rétropropagation (Backpropagation)

Un algorithme compare la prédiction du modèle avec le mot réel du texte d'entraînement. Si le modèle se trompe, la rétropropagation ajuste légèrement les paramètres pour réduire l'ereur la prochaine fois.

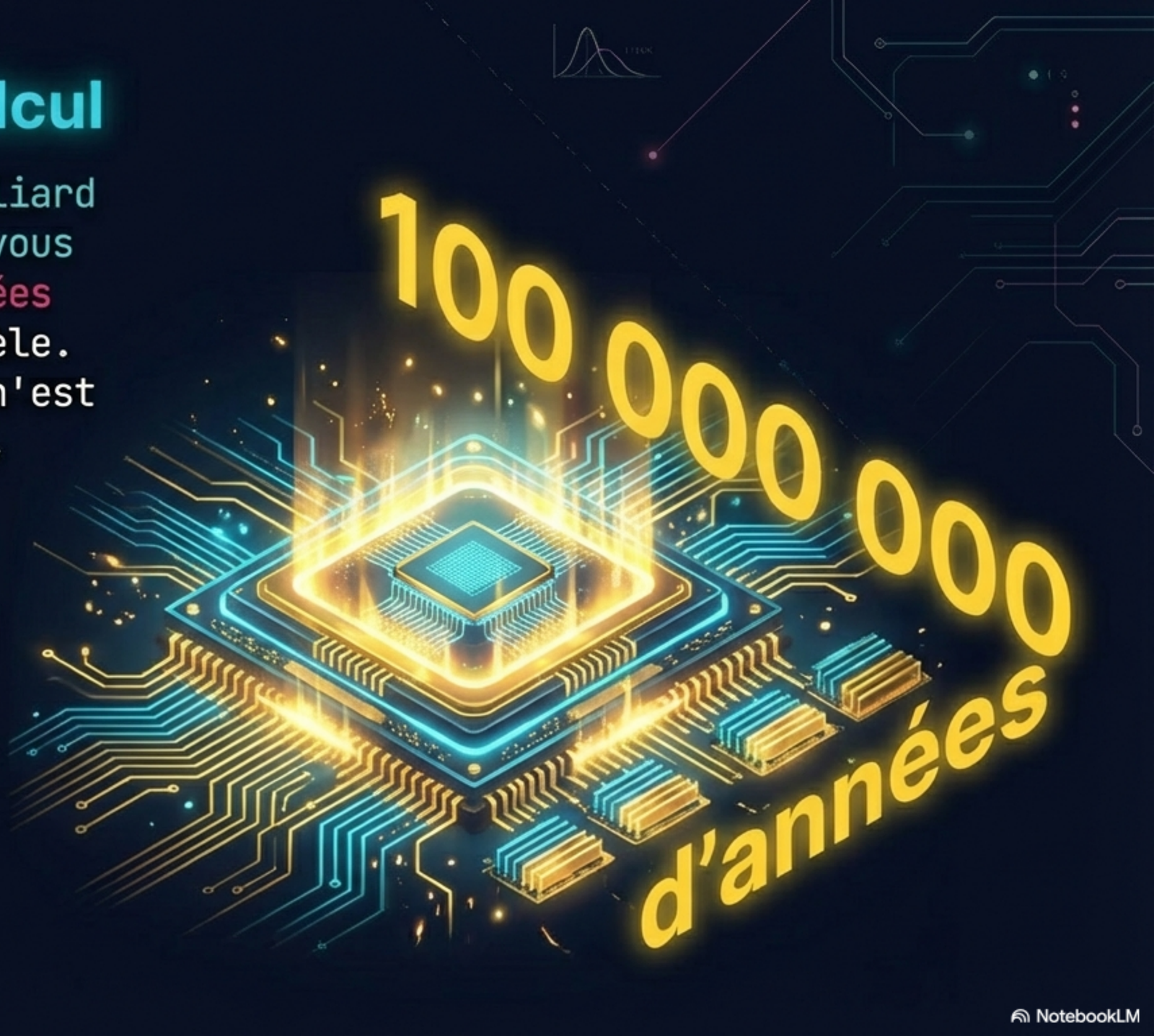


Ce processus est répété sur des milliers de milliards d'exemples.

*GPT-4 a été entraîné sur encore plus de données.

La puissance de calcul

Si vous pouviez faire 1 milliard de calculs par seconde, il vous faudrait 100 millions d'années pour entraîner un grand modèle. Cette échelle vertigineuse n'est possible que grâce aux GPUs, des puçes capables d'effectuer des milliers d'opérations en parallèle.



Du perroquet à l'assistant

L'auto-complétion seule ne suffit pas. Le RLHF est une étape cruciale où des humains notent les réponses pour aligner le modèle sur les intentions de l'utilisateur.

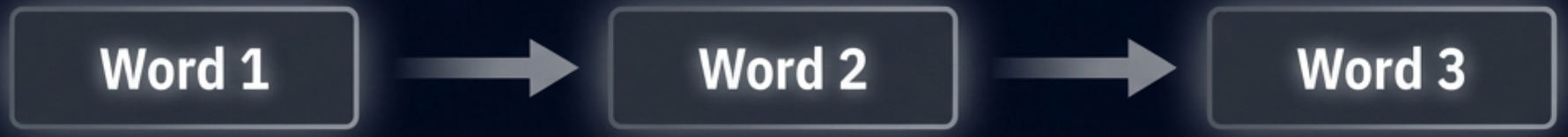
Pré-entraînement
(Auto-complétion)

Assistant Utile

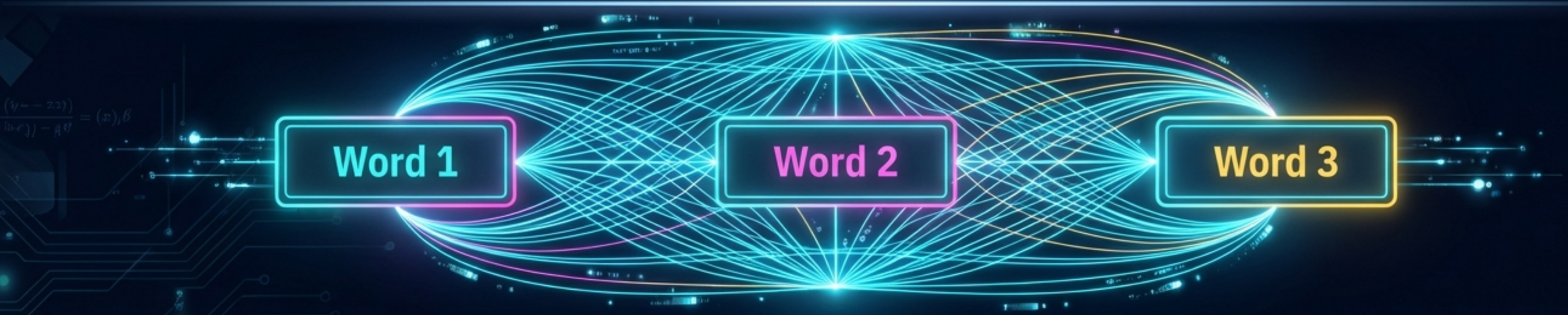
Key Insight: "Sans RLHF, le modèle ne serait qu'un générateur de texte chaotique."

La révolution Transformer (2017)

Avant 2017, les modèles lisaient un mot à la fois. L'architecture « Transformer » a tout changé en permettant au modèle d'analyser l'ensemble du texte simultanément, en parallèle.



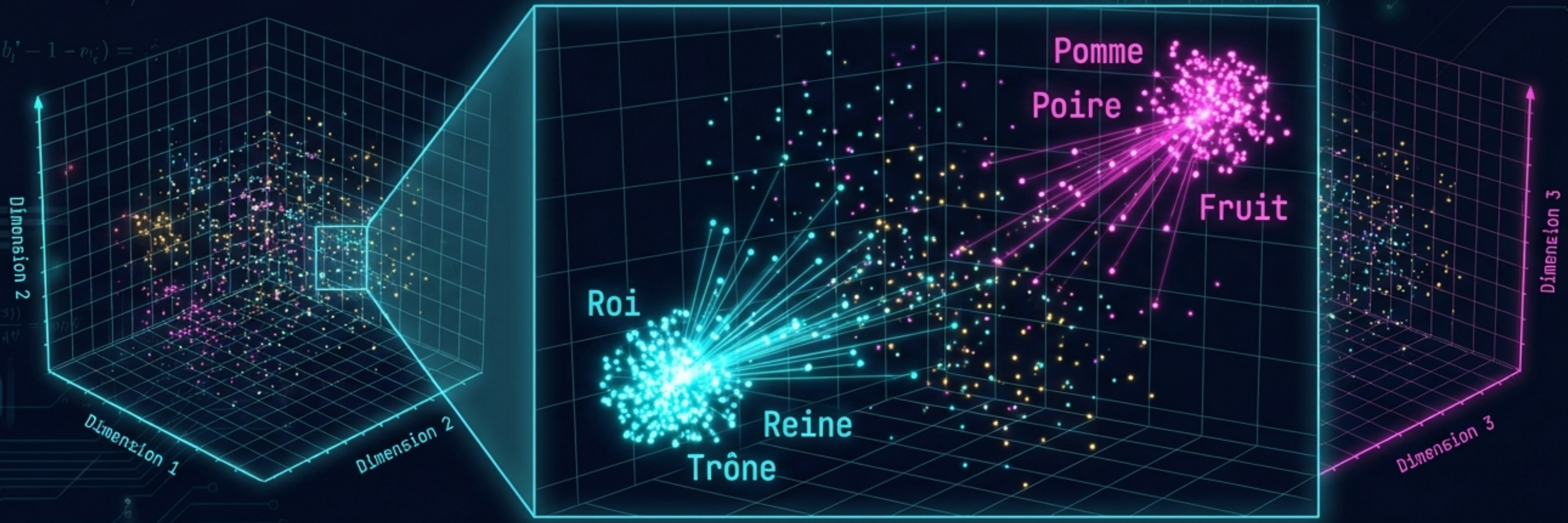
Anciens Modèles (Séquentiel)



TRANSFORMER (Parallèle)

Traduire les mots en nombres (Embeddings)

Les ordinateurs ne lisent pas les mots. Chaque mot est converti en une longue liste de nombres (un vecteur) qui représente son sens. Les mots ayant des significations proches sont situés géographiquement près les uns des autres dans cet espace mathématique.



Cluster 1: Royauté (Proximité Sémantique)

Cluster 2: Fruits (Distance Sémantique)

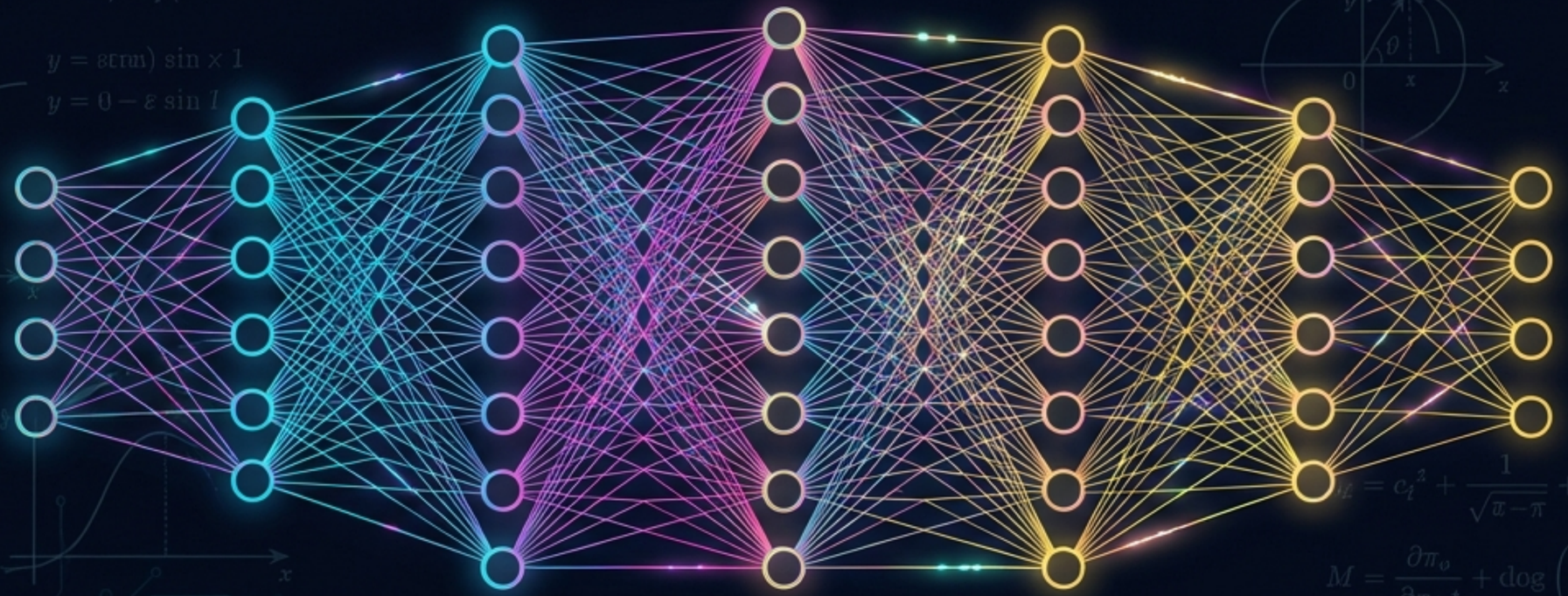
Le mécanisme d’Attention

C’est le cœur du système. L’Attention permet aux mots de « discuter » entre eux pour affiner leur sens en fonction du contexte.

Le mot « avocat » change de sens s’il est associé à « salade » ou à « tribunal ». L’Attention met à jour les nombres pour refléter ce contexte précis.

Le Perceptron Multicouche (MLP)

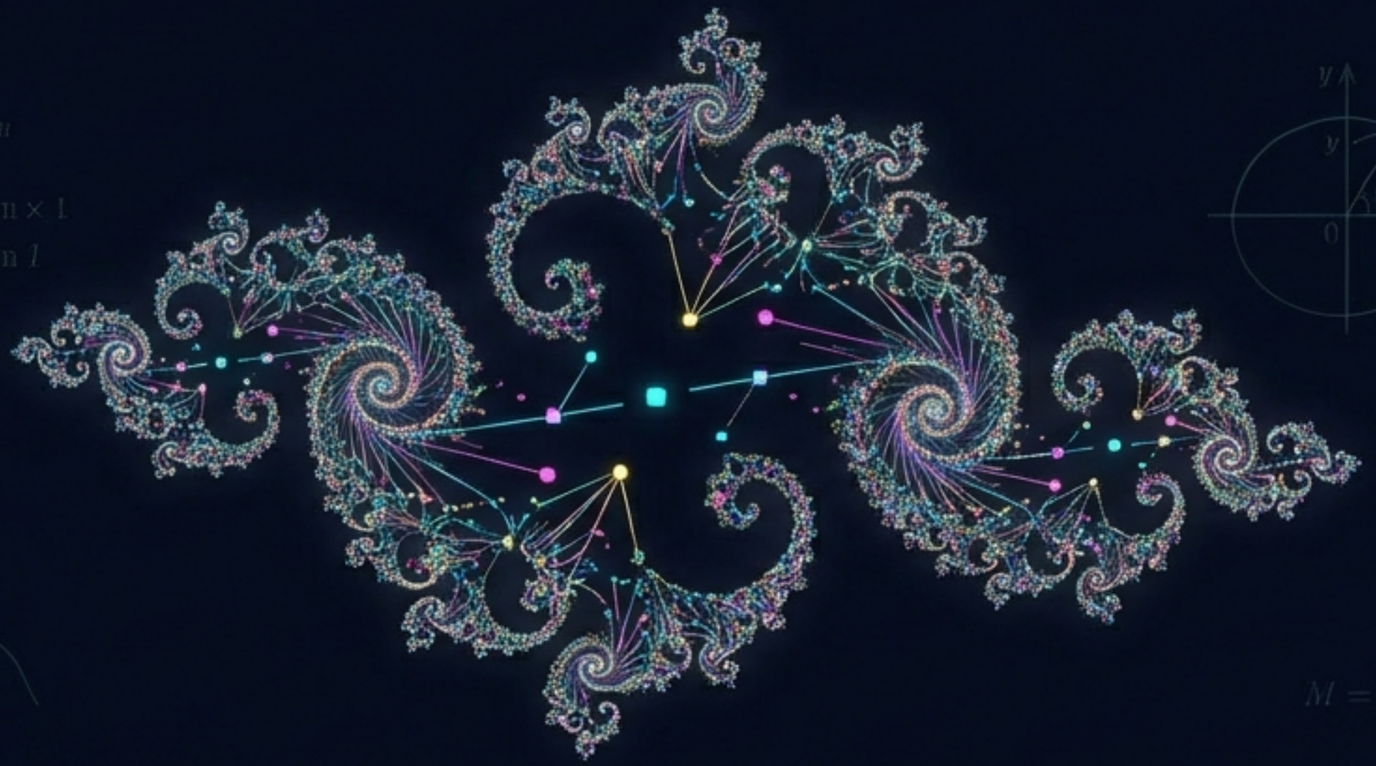
Si l'Attention sert à comprendre le contexte, le MLP sert à stocker les connaissances et les motifs linguistiques appris durant l'entraînement (comme les règles de grammaire ou les faits).



Feed-Forward Network (MLP)

Les données passent alternativement par des couches d'Attention et des couches de MLP pour affiner la prédiction.

L’Émergence



Bien que nous ayons conçu l’architecture, le comportement spécifique du modèle est un phénomène émergent. Personne n’a programmé les règles de grammaire ou les faits historiques.

Ces capacités naissent spontanément de l’interaction de milliards de paramètres cherchant simplement à prédire le mot suivant.

Pour aller plus loin

→ 3Blue1Brown : Neural Networks (Vidéo)

→ Attention is All You Need (Papier de recherche 2017)

→ Deep Learning, Chapitre 5

La magie, c’est juste des mathématiques à une échelle impossible.