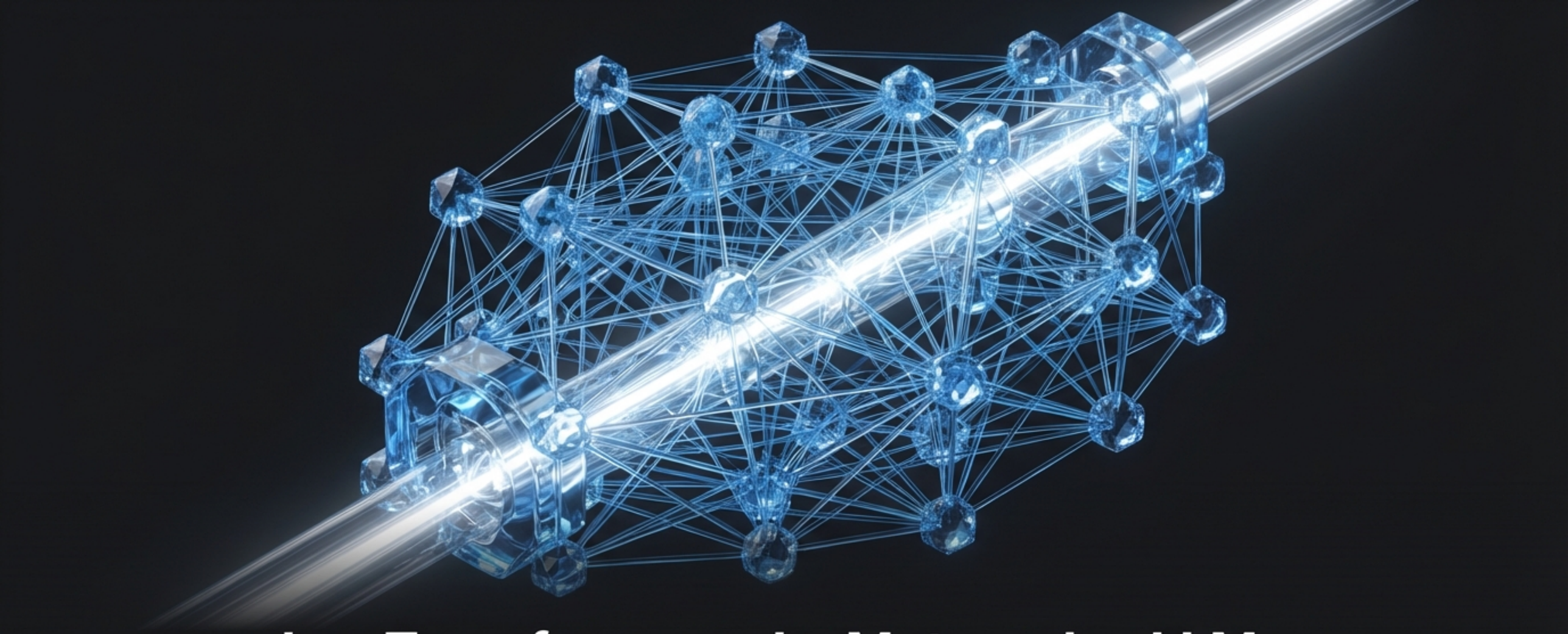


Les Transformers : Le Moteur des LLMs

Comprendre l'architecture derrière ChatGPT (GPT-3)

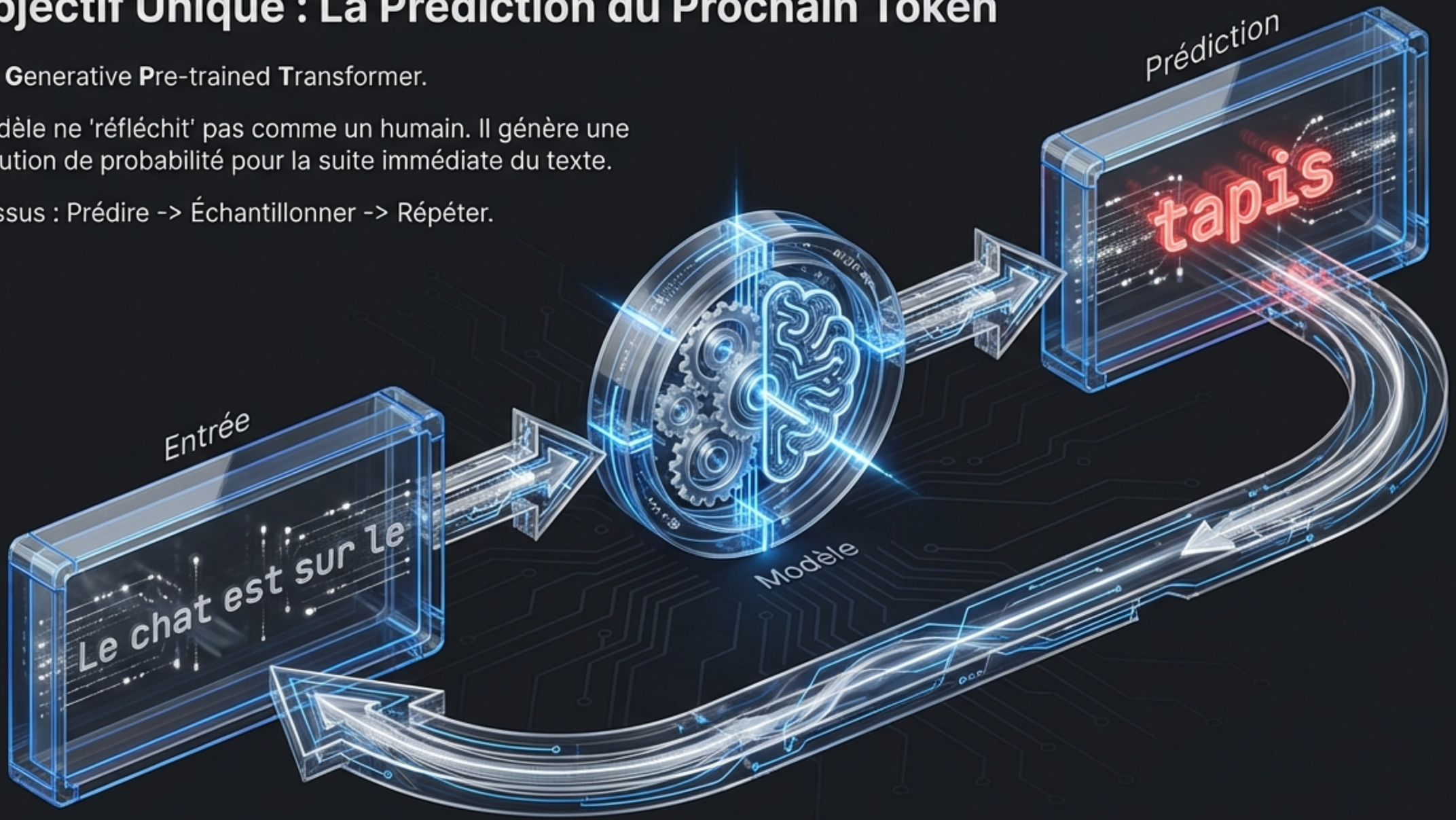
De la magie aux mathématiques : Ce pont technologique transforme du texte brut en sens calculable. Basé sur l'analyse technique de 3Blue1Brown.

L'Objectif Unique : La Prédiction du Prochain Token

GPT = **G**enerative **P**re-trained **T**ransformer.

Le modèle ne 'réfléchit' pas comme un humain. Il génère une distribution de probabilité pour la suite immédiate du texte.

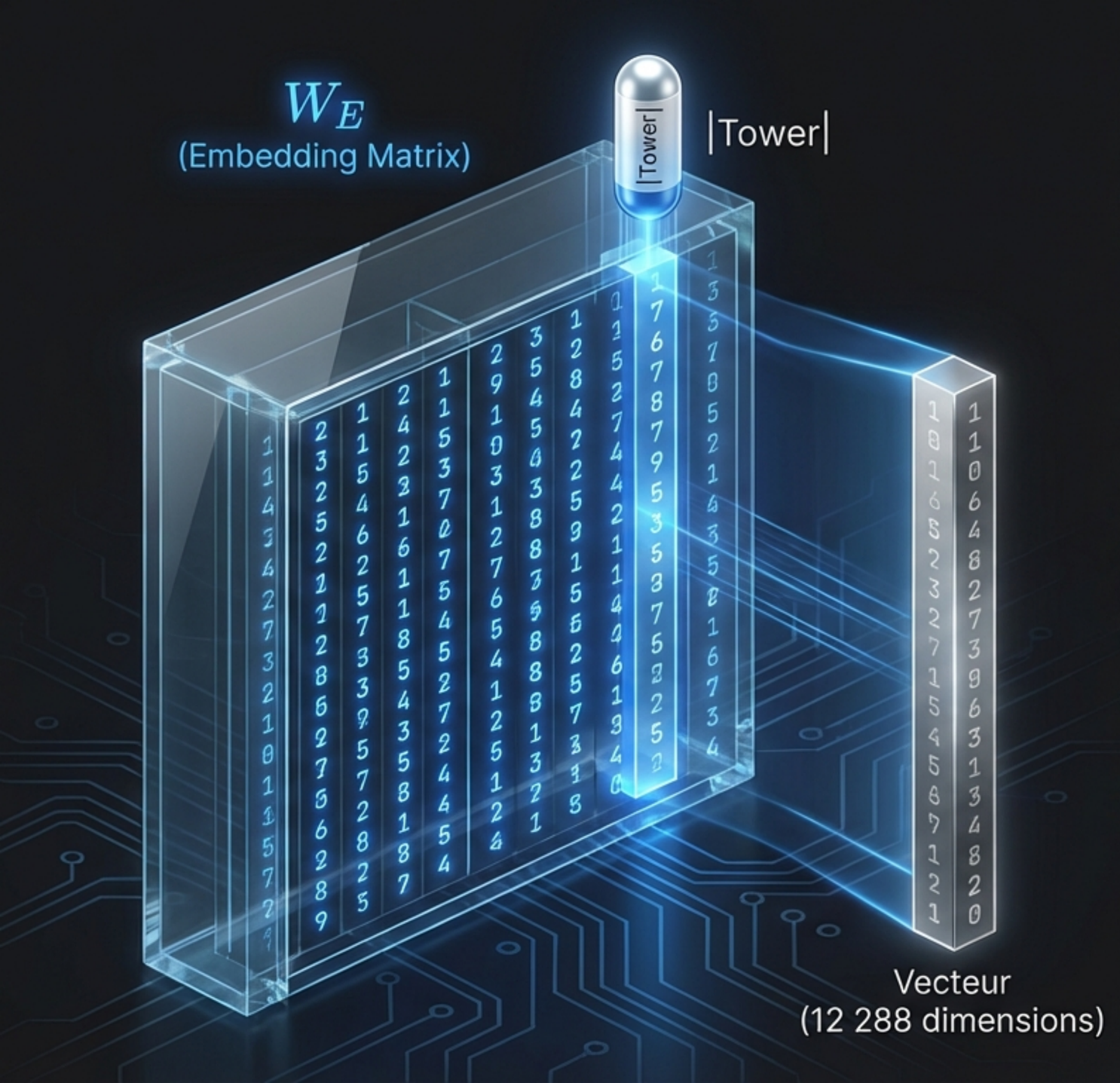
Processus : Prédire -> Échantillonner -> Répéter.



La Tokenisation : Découper pour Comprendre

La machine ne lit pas des mots, elle lit des **Tokens** (morceaux de mots ou ponctuation).

Vocabulaire standard (GPT-3) : **50 257 tokens** uniques.

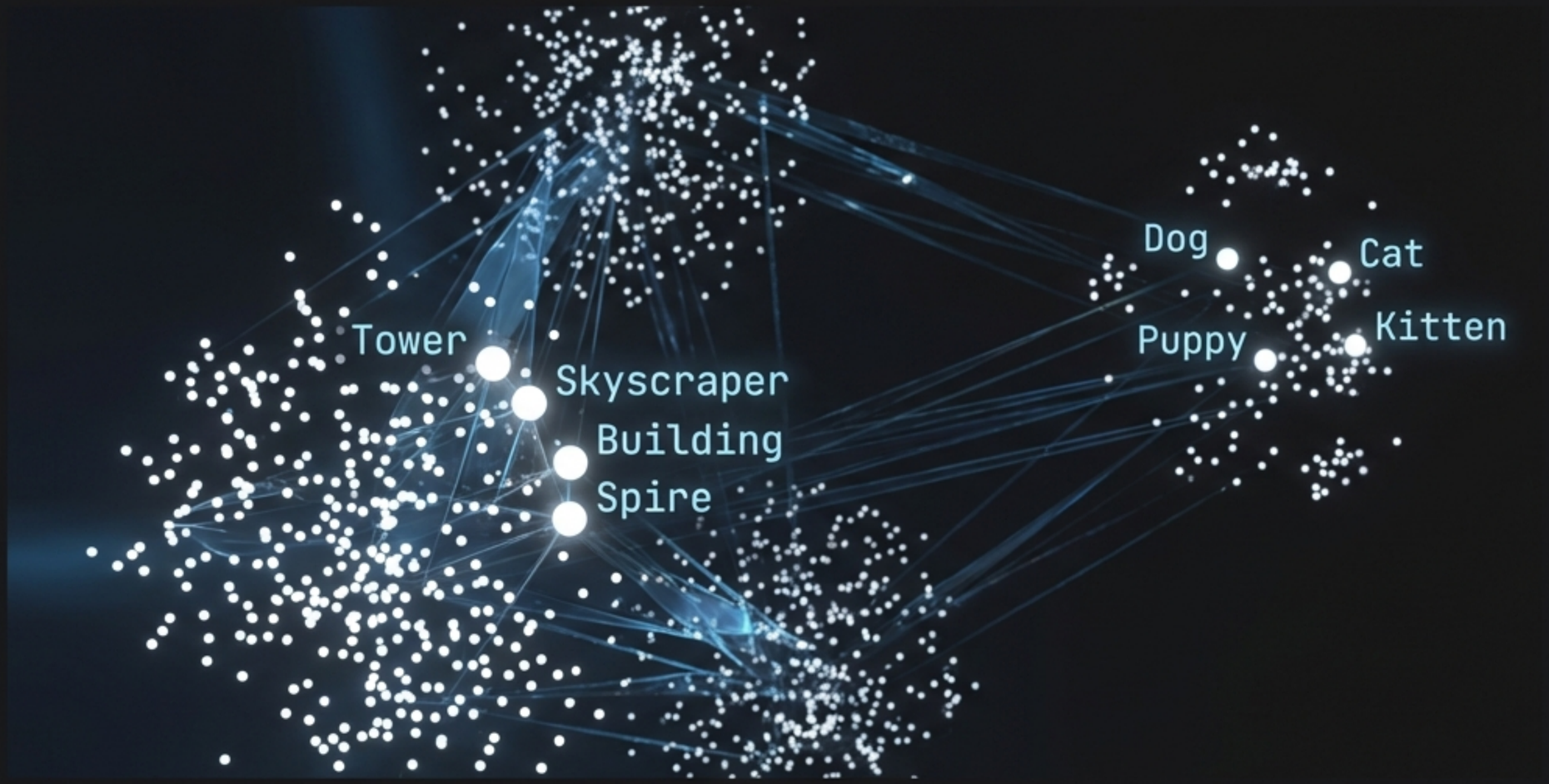


L'Embedding : De Mots à Nombres

Chaque token est converti en un **Vecteur** (une liste de nombres). C'est une table de correspondance (Lookup Table).
Le modèle apprend ces valeurs via l'entraînement.

Données Techniques : Pour GPT-3, chaque vecteur possède **12 288 dimensions**.

L’Espace Vectoriel : Une Galaxie de Sens



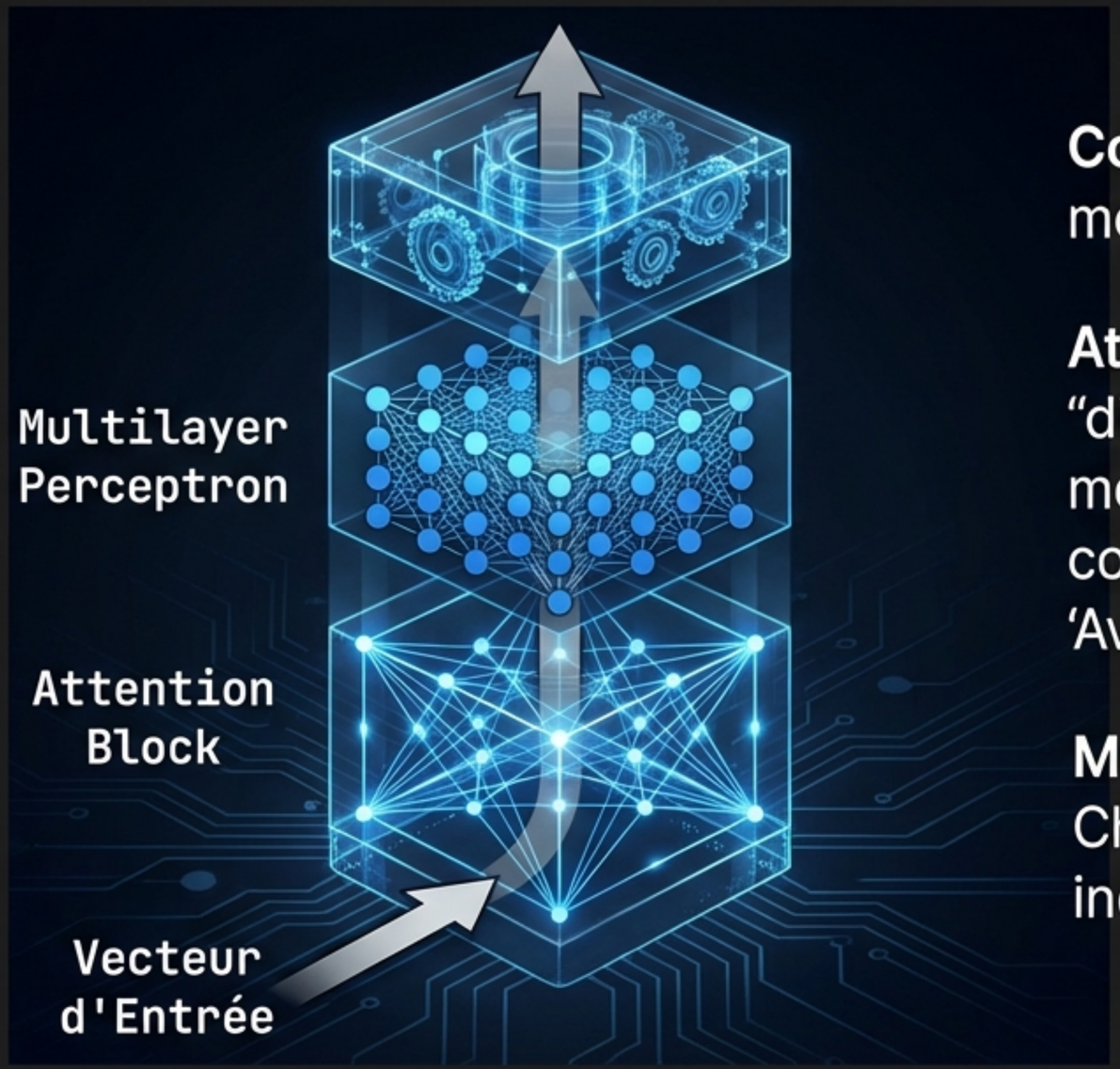
Les 12 288 nombres sont des coordonnées dans un espace multidimensionnel. **Proximité = Similarité.** Les mots au sens proche se retrouvent voisins dans cet espace. Le sens n'est plus une définition, c'est une **direction**.

L’Arithmétique du Langage

$$\text{Roi} - \text{Homme} + \text{Femme} \approx \text{Reine}$$

Le modèle encode les concepts (genre, pluriel, histoire) comme des mouvements géométriques.
Le calcul vectoriel permet de manipuler des analogies complexes par simple addition et soustraction.

Le Traitement : Attention et Profondeur



Context Size : 2048 tokens (le modèle voit tout ça à la fois).

Attention Block : Les vecteurs “discutent” entre eux pour mettre à jour leur sens selon le contexte (ex: ‘Avocat” fruit vs ‘Avocat” métier).

Multilayer Perceptron : Chaque vecteur est traité individuellement.

La Machine (Poids) vs La Pensée (Données)

La Machine (Poids)

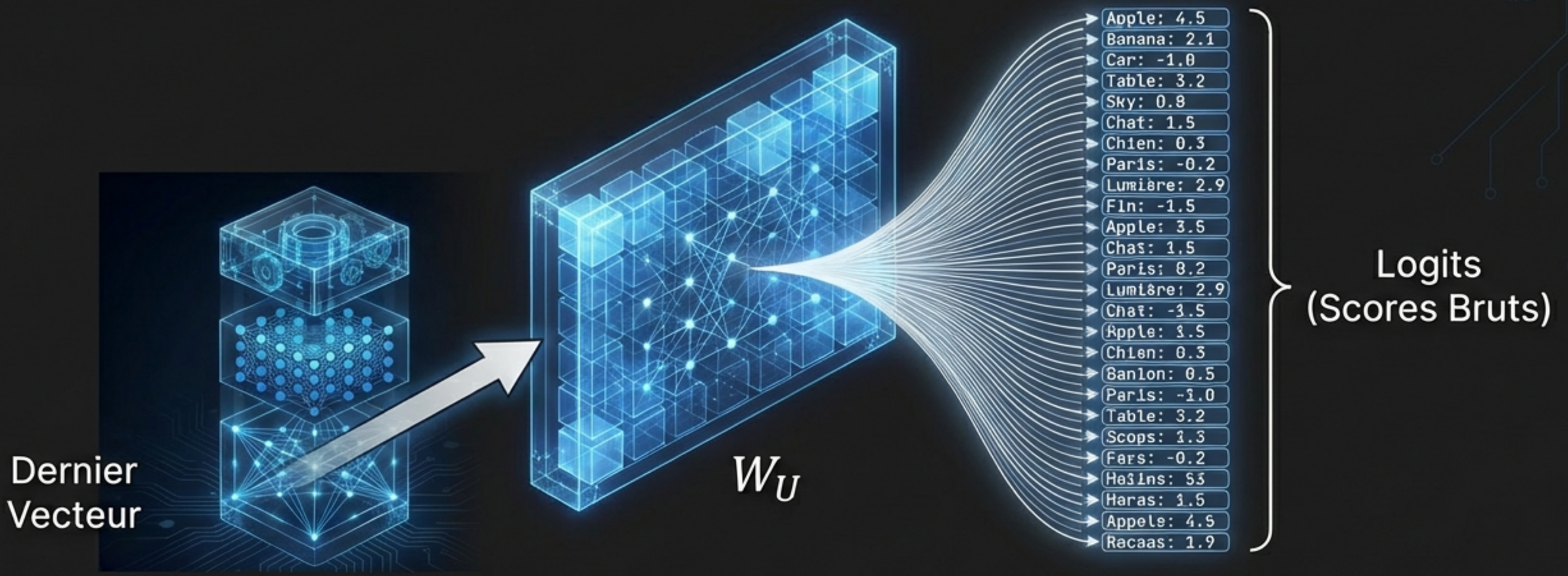
La Pensée (Données)

Les Poids (W) **: 175 Milliards de paramètres**. C'est le 'cerveau' appris durant l'entraînement (Blue).

Les Données : Le texte qui traverse le réseau (White).

Structure : Environ **28 000 matrices** distinctes organisées en 96 couches.

L’Unembedding : Le Retour au Vocabulaire



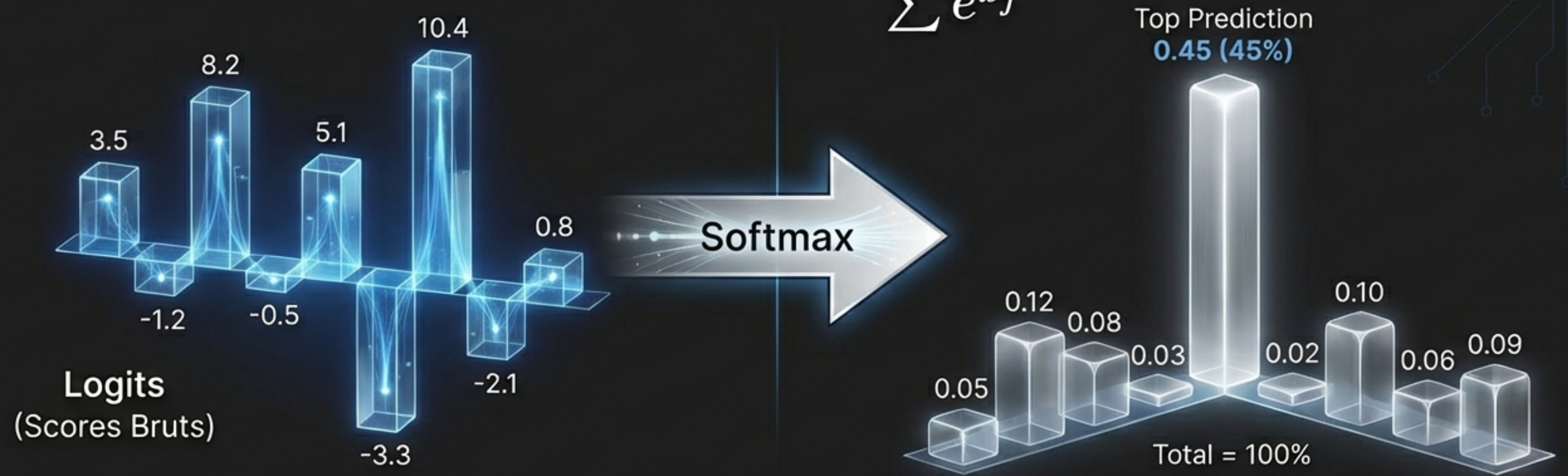
Le dernier vecteur contient toute l’information nécessaire pour prédire le mot suivant.

Matrice W_U : Mappe le vecteur vers le vocabulaire complet (50 257 mots).

Résultat : Une liste de **Logits** (scores bruts non normalisés).

Softmax : La Distribution de Probabilité

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum e^{x_j}}$$



Transforme les scores bruts en pourcentages (0 à 1).

La somme de toutes les probabilités est toujours égale à 100%.

Le mot avec le score le plus élevé n'est **pas toujours choisi** : c'est un échantillonnage.

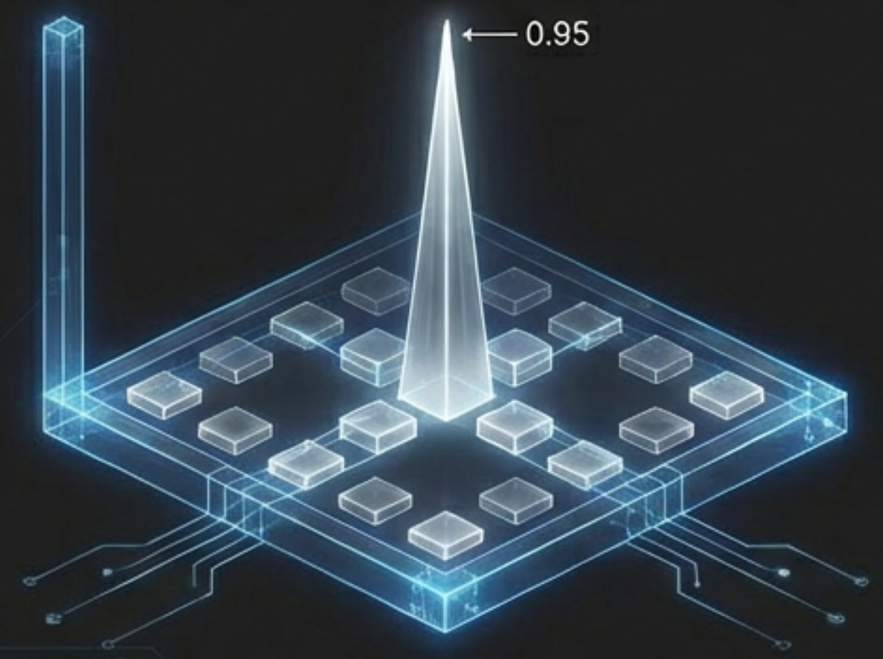
La 'Température' : Créativité ou Précision ?

Un paramètre *T* est injecté dans l'équation Softmax.

T Bas : Le modèle choisit presque toujours le mot le plus probable.

T Haut : La distribution s'aplatit, permettant des choix plus rares.

Température Basse (T -> 0)

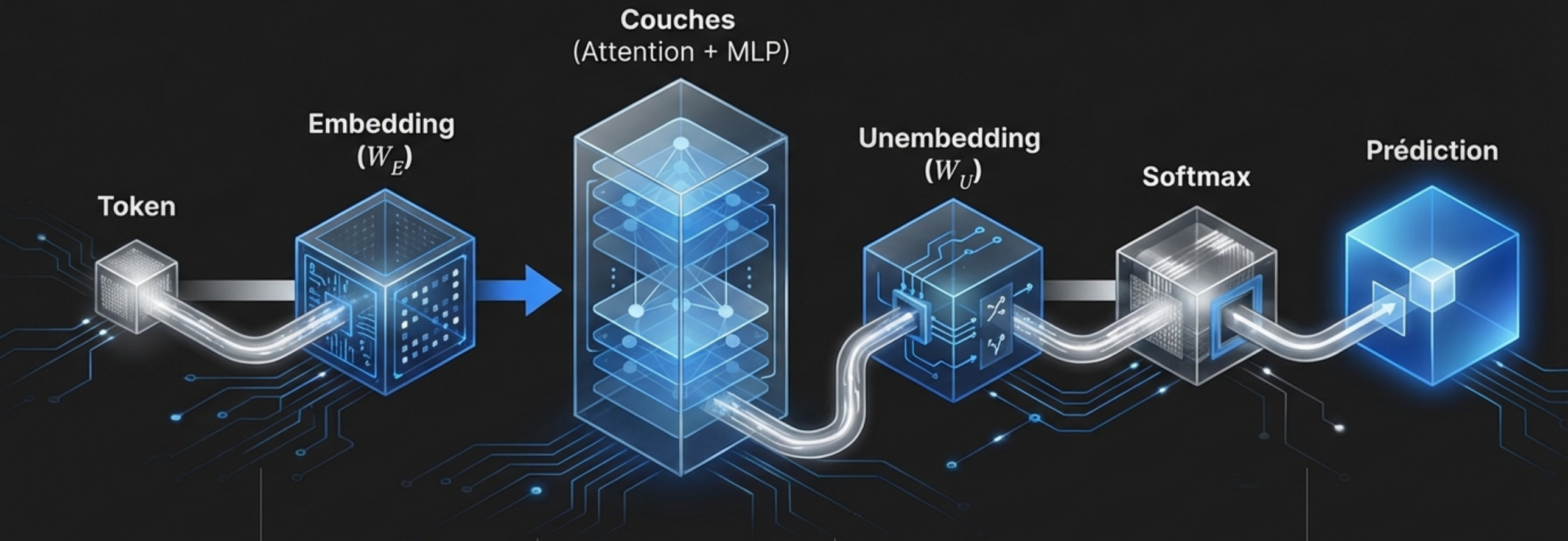


Précis / Répétitif

Température Haute

Créatif / Hallucination

Le Pipeline Complet



1. **Entrée** : Texte découpé.

2. **Embedding** : Conversion en vecteurs.

3. **Profondeur** : 96 couches de calcul.

4. **Sortie**: Probabilités sur 50 257 tokens.

Ce n'est que le début...

Nous avons vu le flux des données. La véritable révolution réside dans la manière dont les mots “se parlent” entre eux. **Prochaine étape : Le Mécanisme d'Attention.**
Adapté de "Deep Learning Chapter 5” par 3Blue1Brown.