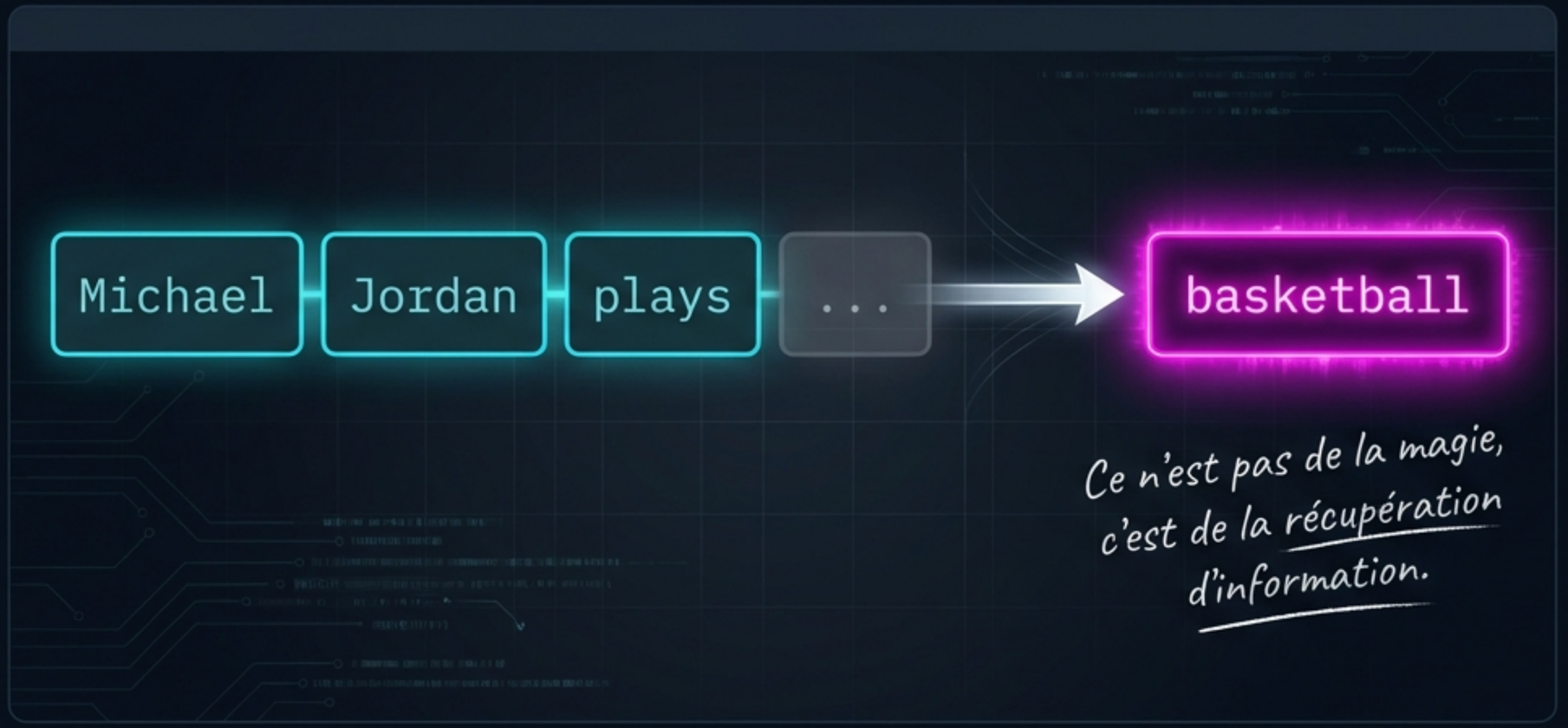
Comment les LLM stockent-ils les faits?

Une exploration visuelle des Perceptrons Multicouches (MLP) et de la géométrie vectorielle.

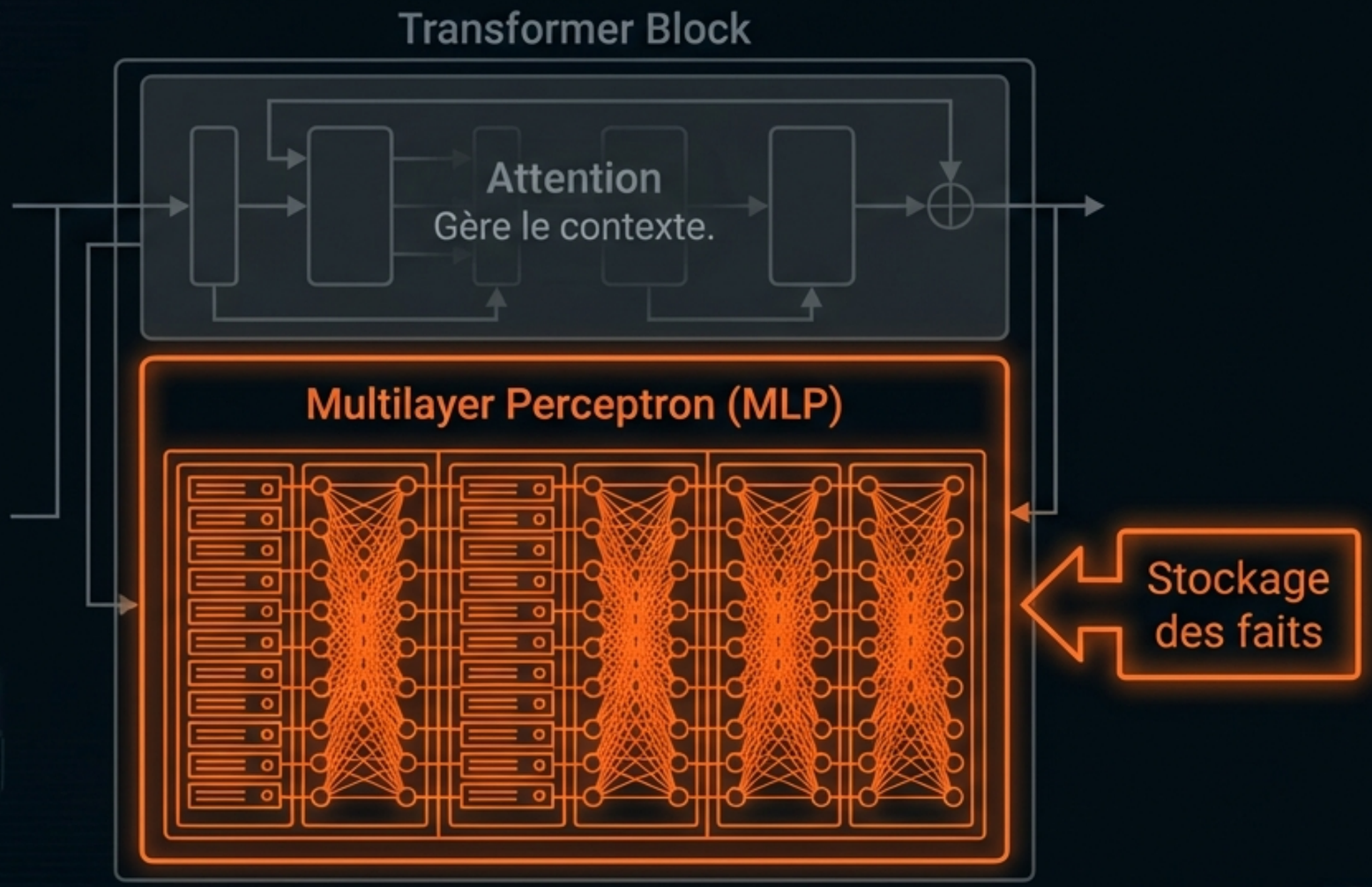
Basé sur l'analyse de 3Blue1Brown | Chapitre 7

L'Observation : Le test de Michael Jordan.

Lorsque vous écrivez "Michael Jordan plays the sport of...", le modèle prédit "basketball". Cela implique qu'une association factuelle est stockée physiquement parmi les centaines de milliards de paramètres.



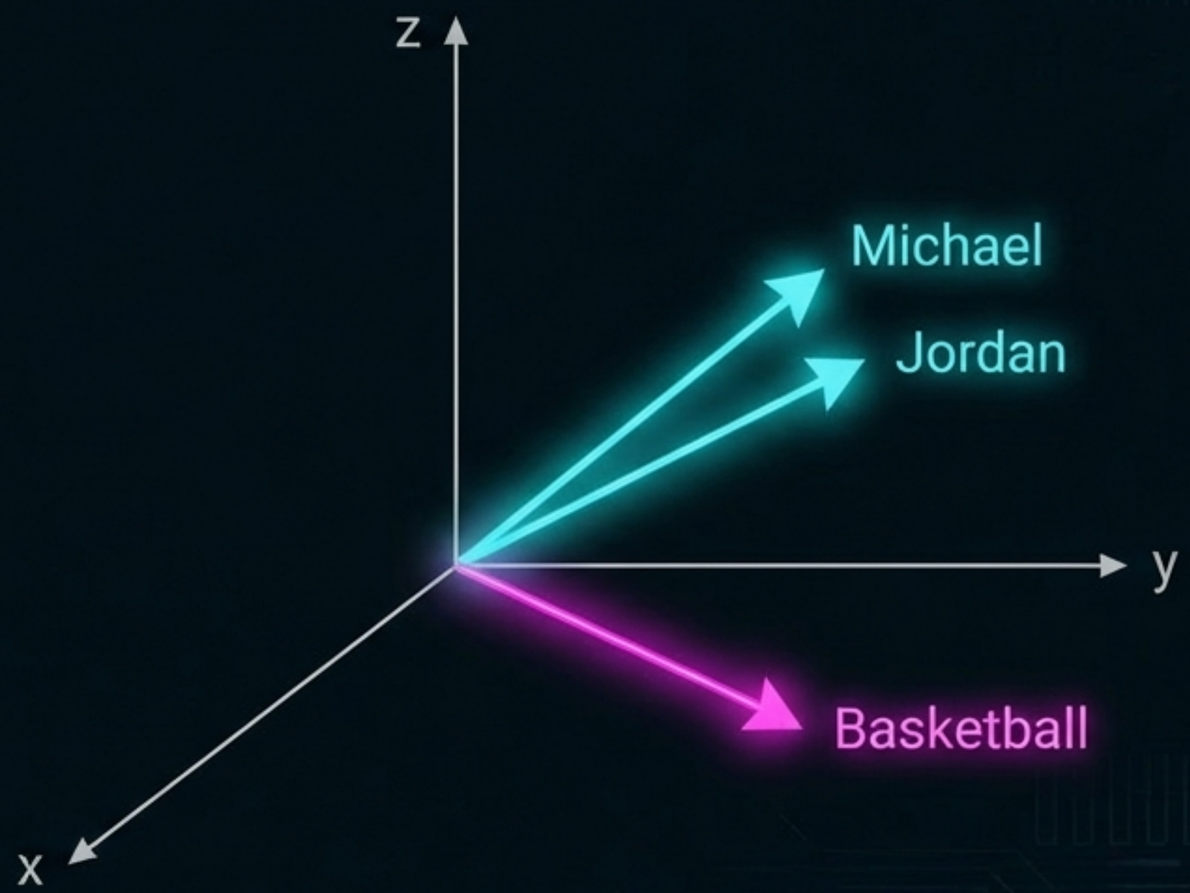
Le coffre-fort de la mémoire.



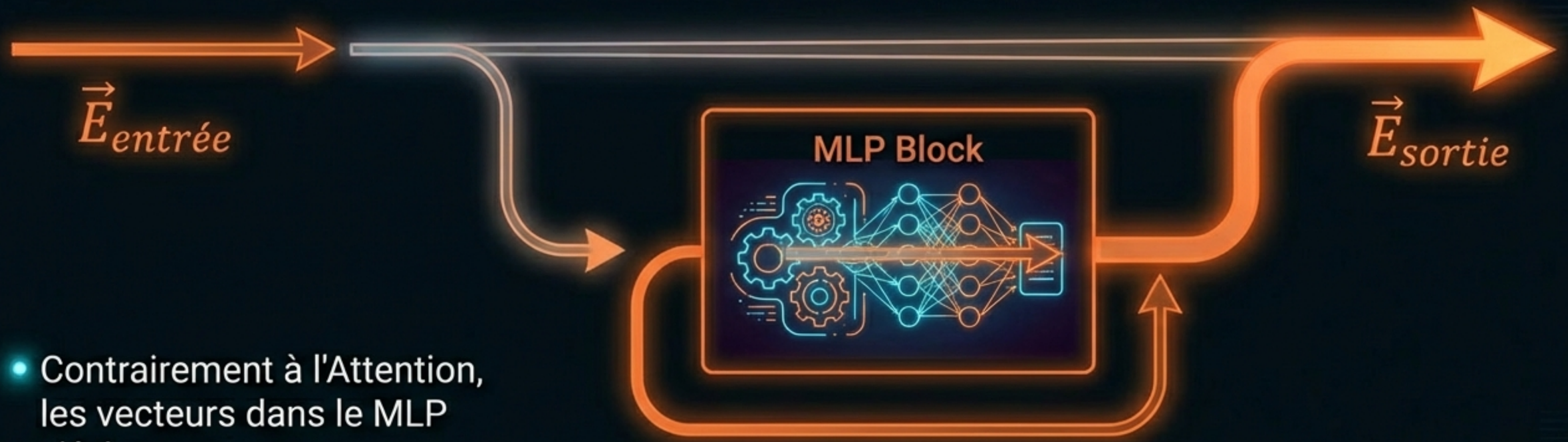
- Les blocs d'Attention gèrent le contexte (la relation entre les mots).
- Les blocs MLP stockent les connaissances factuelles.
- Dans GPT-3, les MLPs représentent **~2/3 des paramètres** (116 milliards sur 175).

Géométrie du sens : Le sens est une direction.

- Nous supposons que chaque concept est une direction spécifique dans un espace à haute dimension.
- **Produit Scalaire ≈ 1** : Le vecteur est aligné (Concept présent).
- **Produit Scalaire ≤ 0** : Le vecteur n’est pas aligné (Concept absent).



Le flux de traitement isolé.

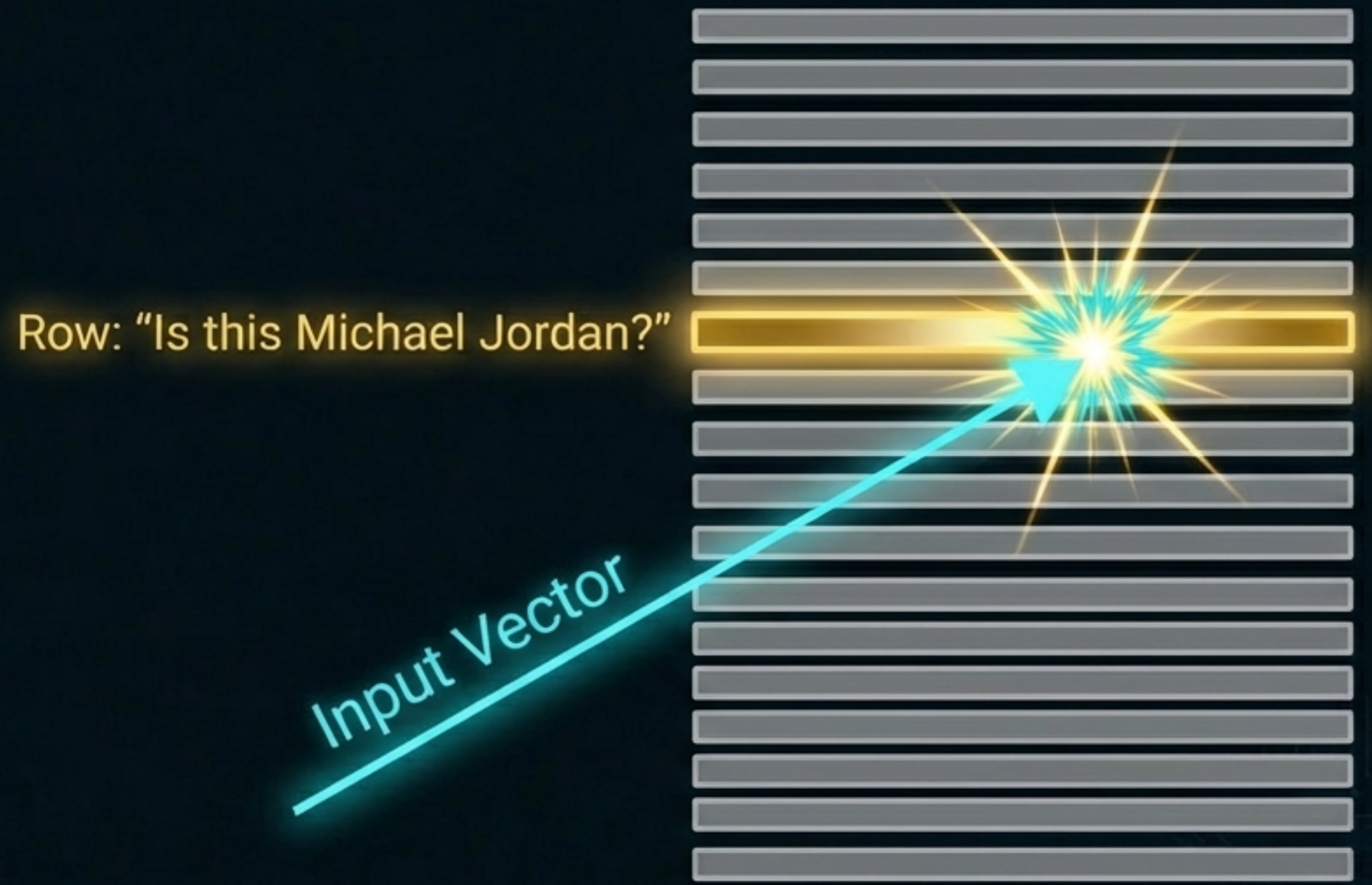


- Contrairement à l'Attention, les vecteurs dans le MLP n'échangent pas d'informations entre eux. Chaque mot est traité en parallèle.

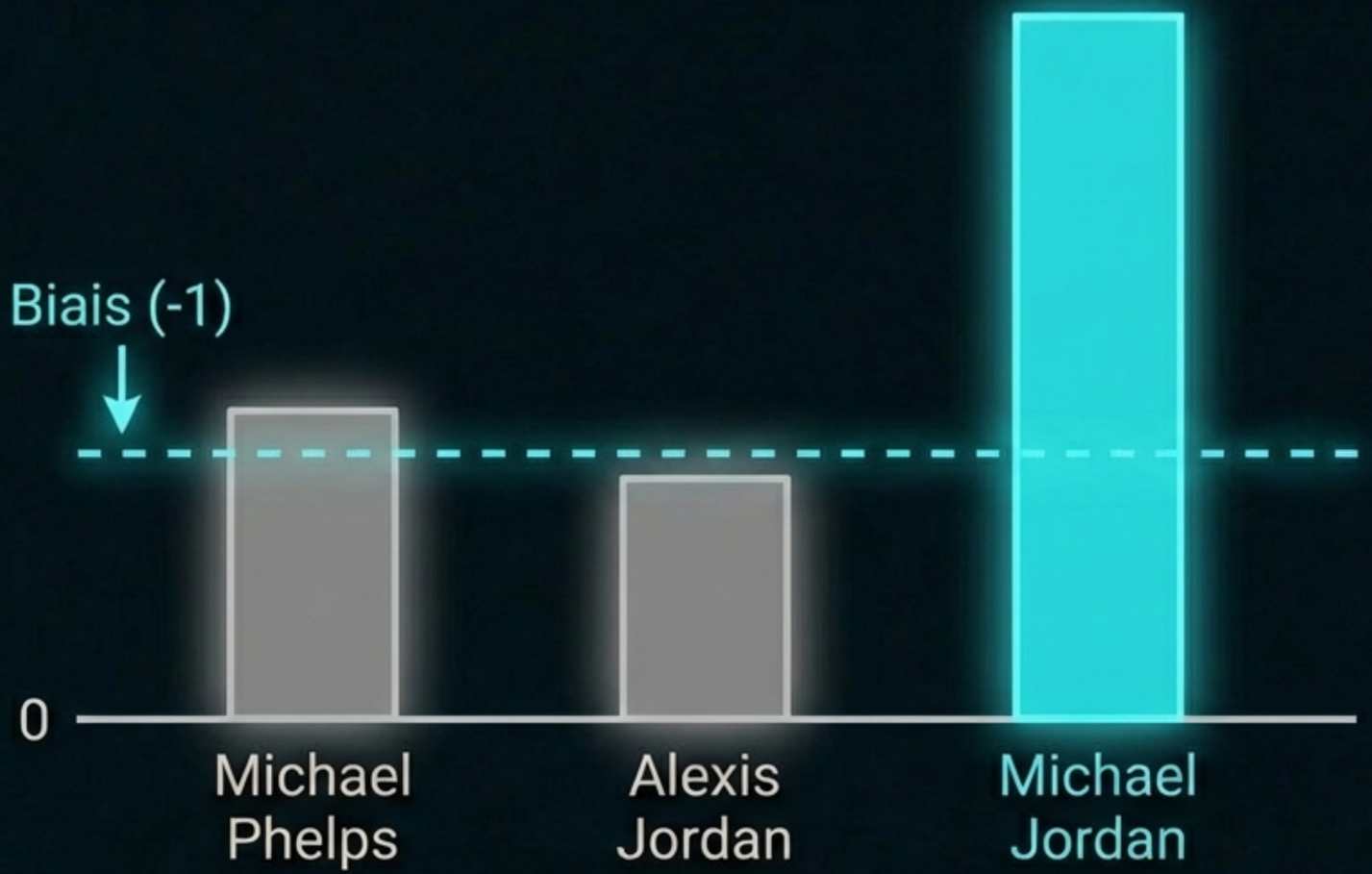
$$\vec{E}_{sortie} = \vec{E}_{entrée} + \textbf{Traitement}(\vec{E}_{entrée})$$

Étape 1 : La Matrice de Questions ($W_{\uparrow}$).

- Cette première matrice projette le vecteur dans un espace de dimension supérieure.
- Chaque ligne de la matrice pose une question spécifique.
- Si le vecteur d'entrée s'aligne avec la ligne "Michael Jordan", le score est élevé.

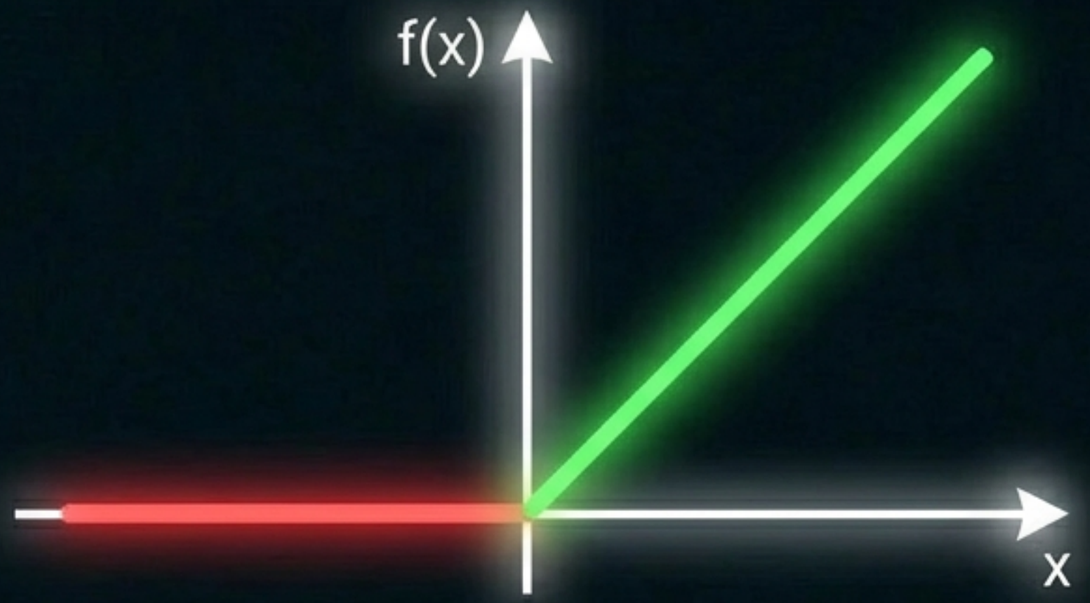


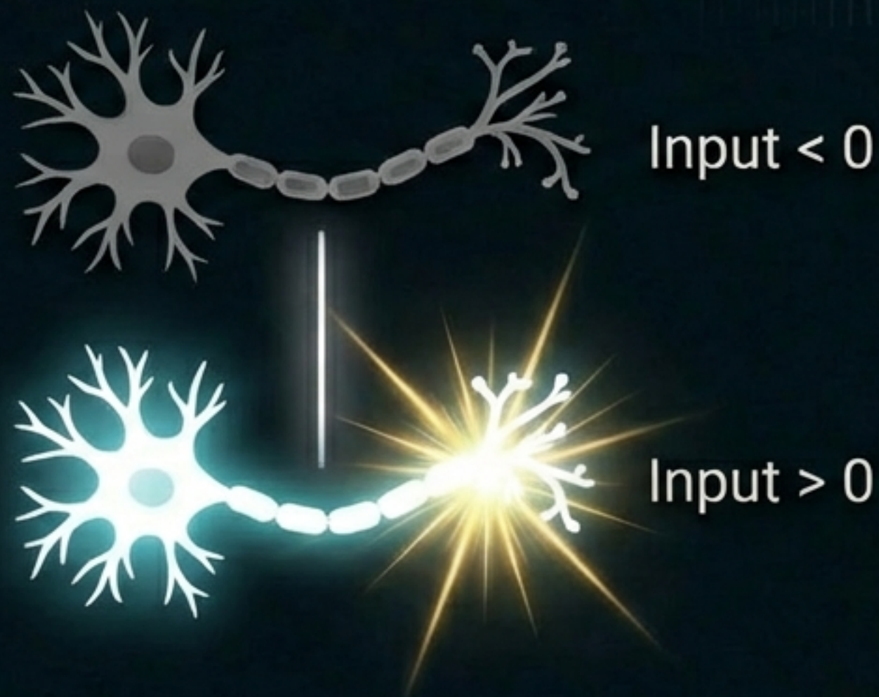
Le Biais : Affiner le signal.



- La linéarité ne suffit pas. Le modèle ajoute un "vecteur de biais" pour filtrer le bruit.
- En soustrayant une valeur, seuls les alignements très forts restent positifs.

L’Activation : L’interrupteur (ReLU).

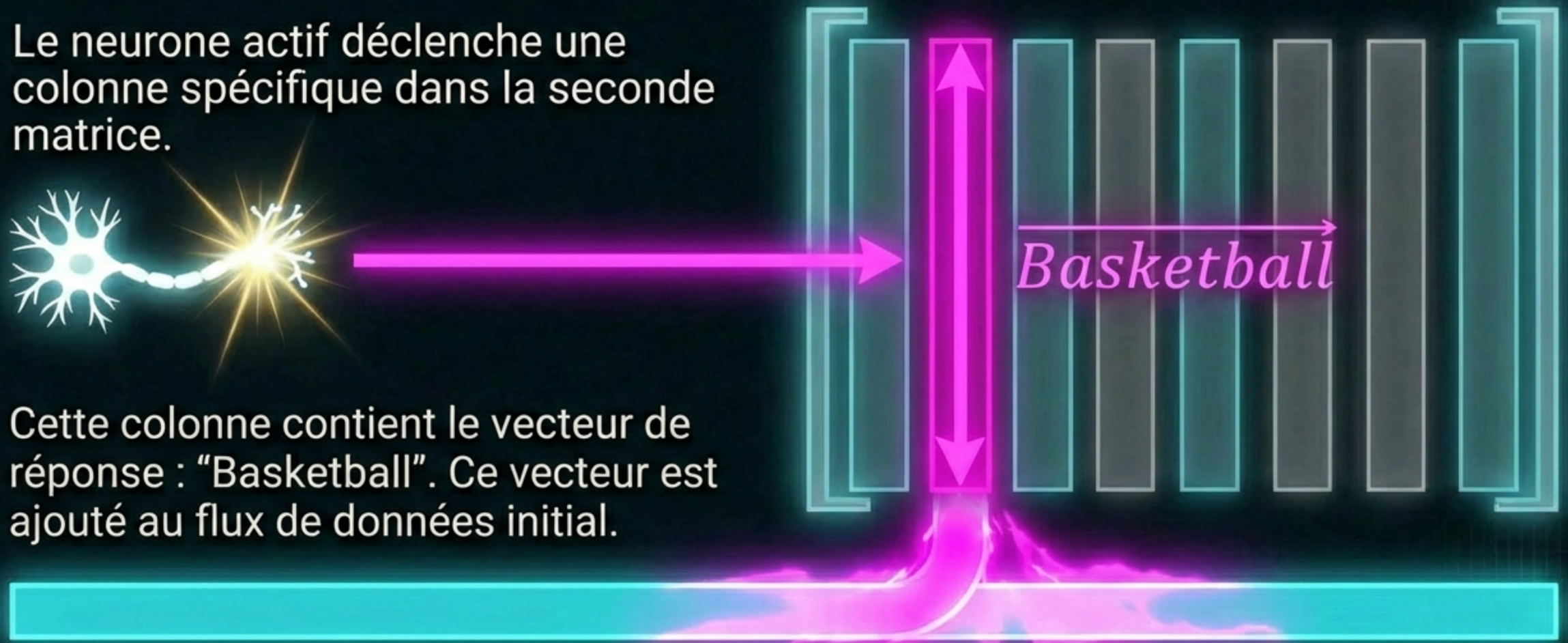




- La fonction ReLU agit comme une porte logique **ET**.
- Valeurs négatives → 0 (Inactif).
- Valeurs positives → Inchangées (Actif).
- C'est ici que le modèle décide : "OUI, c'est Michael Jordan".

Étape 2 : La Matrice de Réponses ($W_{\downarrow}$).

Le neurone actif déclenche une colonne spécifique dans la seconde matrice.



Cette colonne contient le vecteur de réponse : “Basketball”. Ce vecteur est ajouté au flux de données initial.

Mécanisme Résumé : L'addition de sens.

- Le vecteur est entré avec un nom.
- Il ressort avec un nom **plus** un concept associé.
- Le savoir est encodé dans les poids des matrices.

L’échelle de GPT-3.

1 Neurone

12 288 dimensions par vecteur

50 000 neurones par bloc MLP

96 couches de blocs

175 Milliards de paramètres au total

Le problème de la place : La Polysémantique.

En réalité, un seul neurone code rarement pour un fait unique. Il s’active pour plusieurs concepts non reliés car il y a plus de faits à stocker que de dimensions disponibles.

La solution : La Superposition.

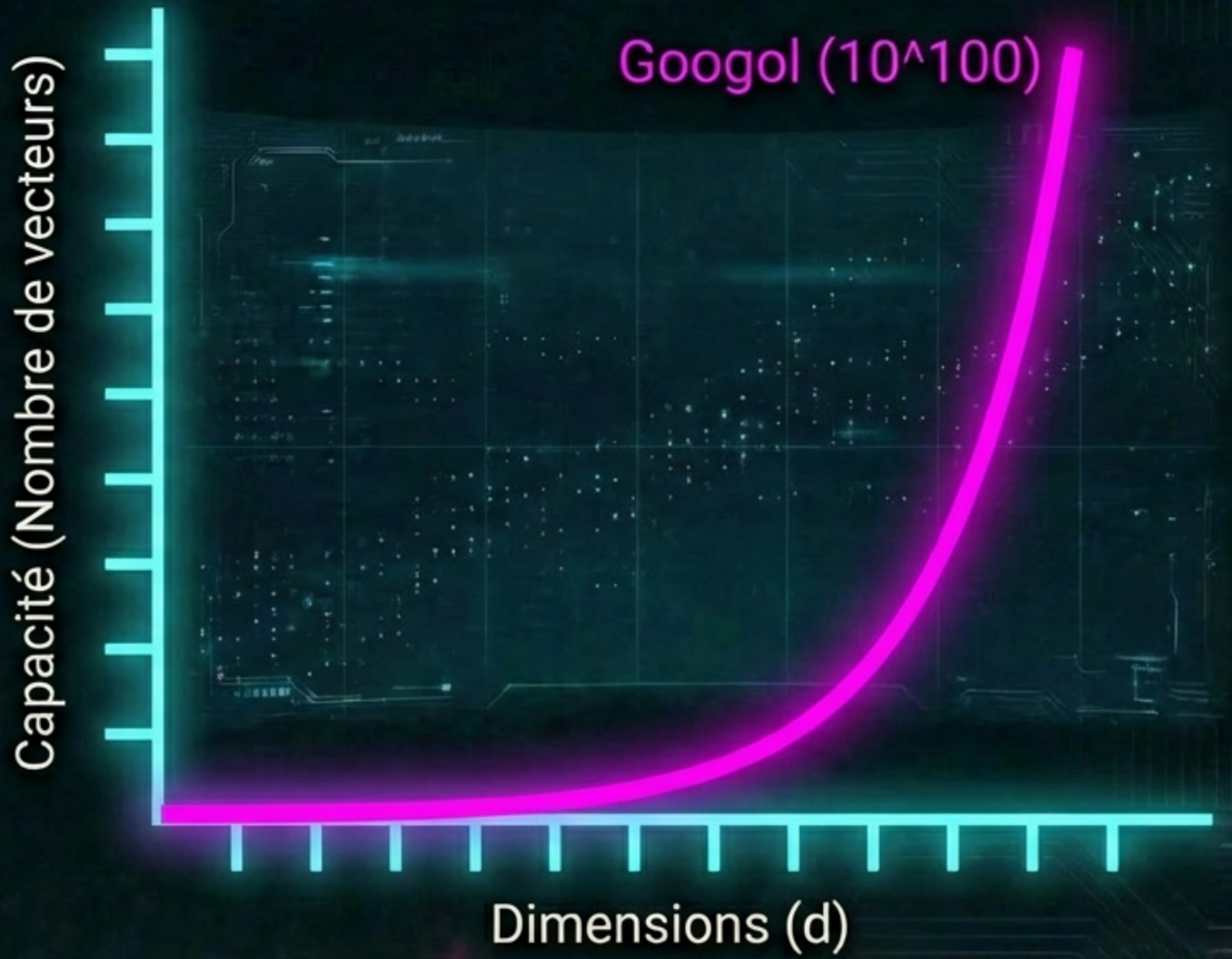
Espace 2D : Saturé

Haute Dimension :
Quasi-Orthogonalité

Dans un espace à haute dimension, les vecteurs n'ont pas besoin d'être à 90° pour être distincts. En acceptant un angle de 89°, on peut stocker des milliards de concepts.

Le Lemme de Johnson-Lindenstrauss.

- Le nombre de vecteurs "presque orthogonaux" croît de manière exponentielle.
- Avec 12 288 dimensions, la capacité de stockage dépasse l'imagination humaine.



La frontière de la recherche.

Démêler la superposition avec les Sparse Autoencoders.

Anthropic Research:
Toy Models of Superposition

3Blue1Brown:
Deep Learning Series

Les MLPs sont des moteurs mathématiques simples (Matrices + ReLU) permettant une mémoire complexe.