L’Anatomie Visuelle de l’Attention

Comprendre le cœur des Transformers et des LLMs

Une exploration visuelle basée sur le chapitre 6 de 3Blue1Brown.

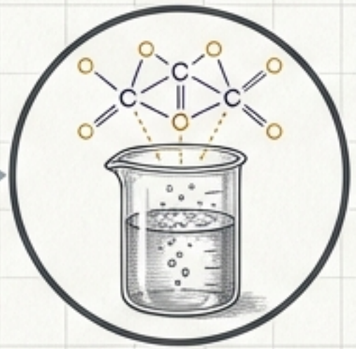
Le Problème de l'Ambiguïté

Mole

L'Entrée : Le mot 'Mole' possède le même plongement initial, quel que soit le contexte.

Taupe musaraigne

Le Conflit : Sans contexte, la machine ne peut pas distinguer l'animal, la chimie, ou la dermatologie.

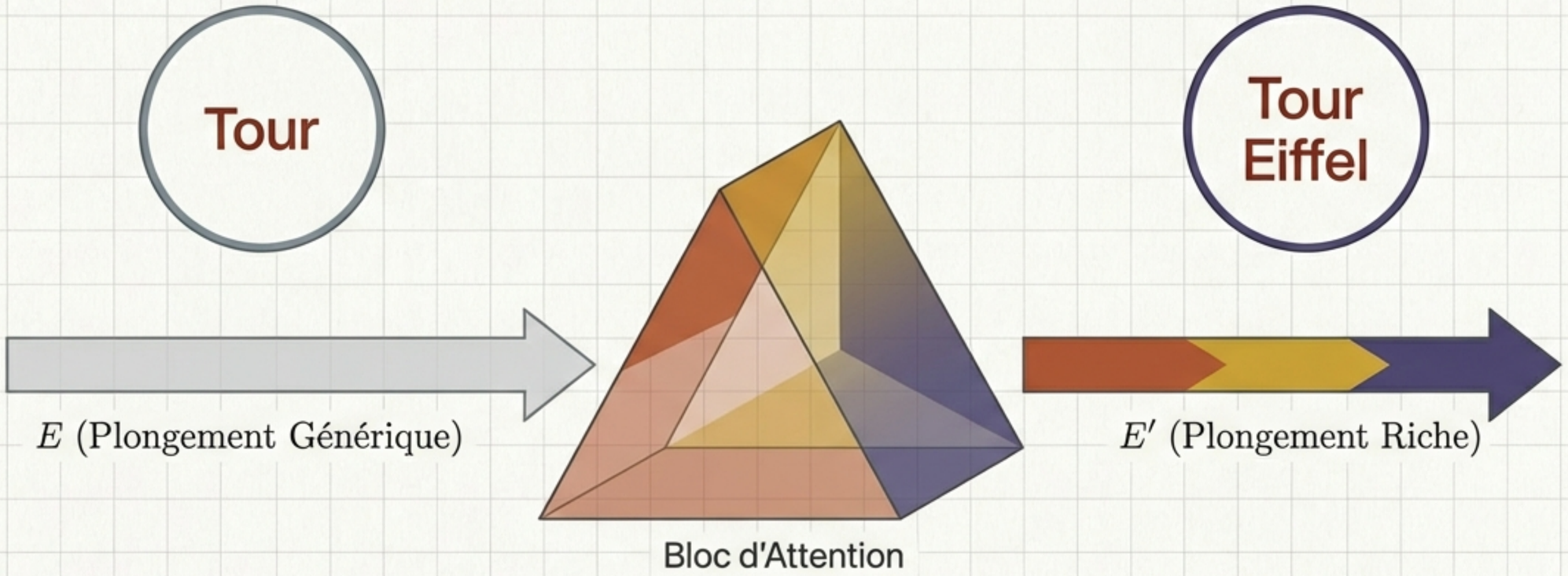


Une mole de carbone

L'Objectif : Mettre à jour ce vecteur pour qu'il pointe vers un sens unique.

Grain de beauté

De l’Ambiguïté à la Précision



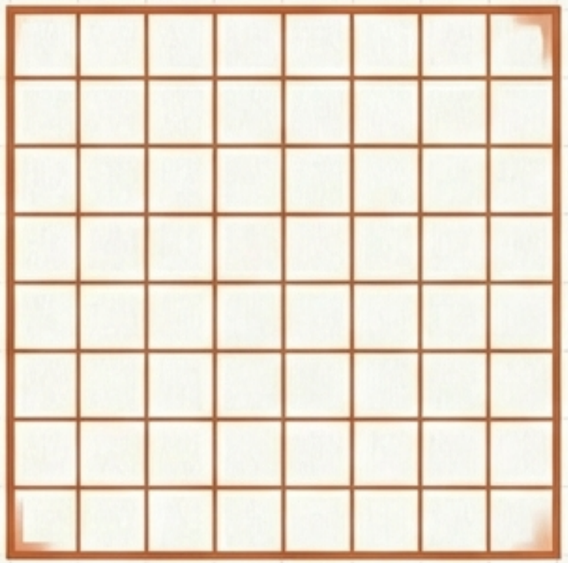
Le but d'un Transformer n’est pas seulement de lire, mais d'enrichir.

Si 'Tour' est précédé par 'Eiffel', le vecteur doit s'aligner avec les concepts de Paris et de métal.

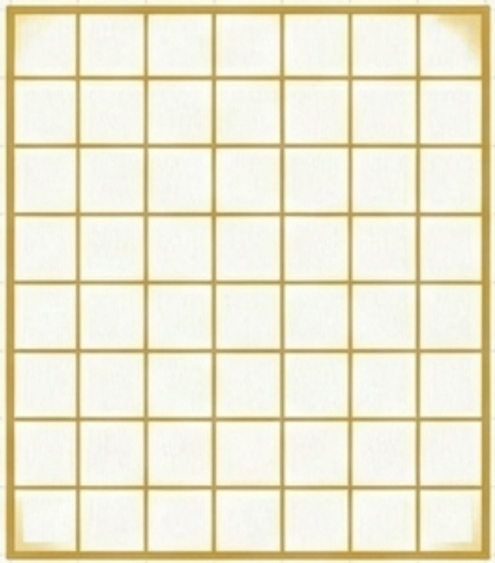
Le mécanisme d’attention calcule exactement ce qu'il faut ajouter au vecteur original.

Les Trois Matrices

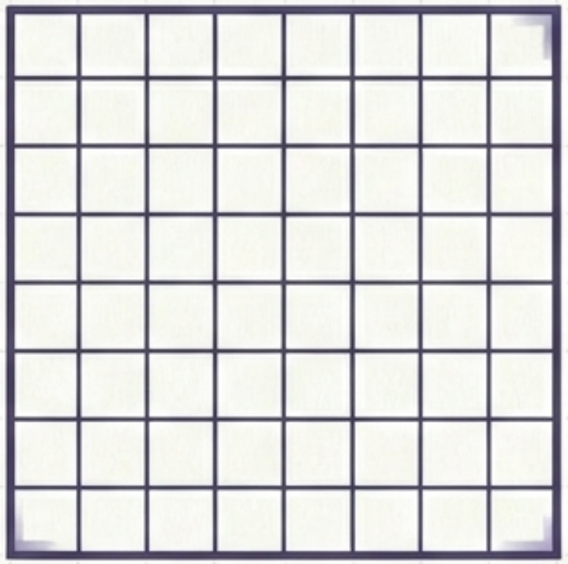
Paramètres ajustables pour une seule tête d'attention



W_Q (Requête / Query)



W_K (Clé / Key)



W_V (Valeur / Value)

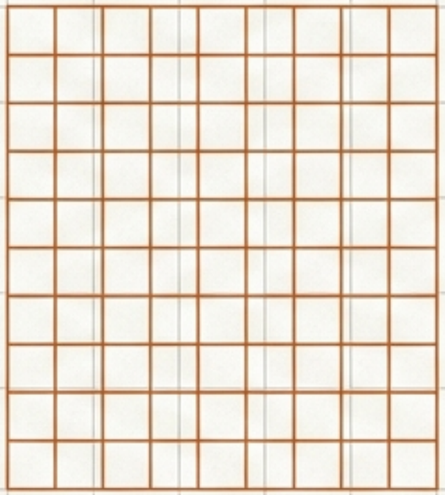
Pour visualiser le concept, imaginons la phrase : "Une créature bleue et duveteuse".
Notre but est que les adjectifs (bleue, duveteuse) mettent à jour le sens du nom (créature).

La Requête (Query)

Vector E
(Créature)

×

Matrix W_Q



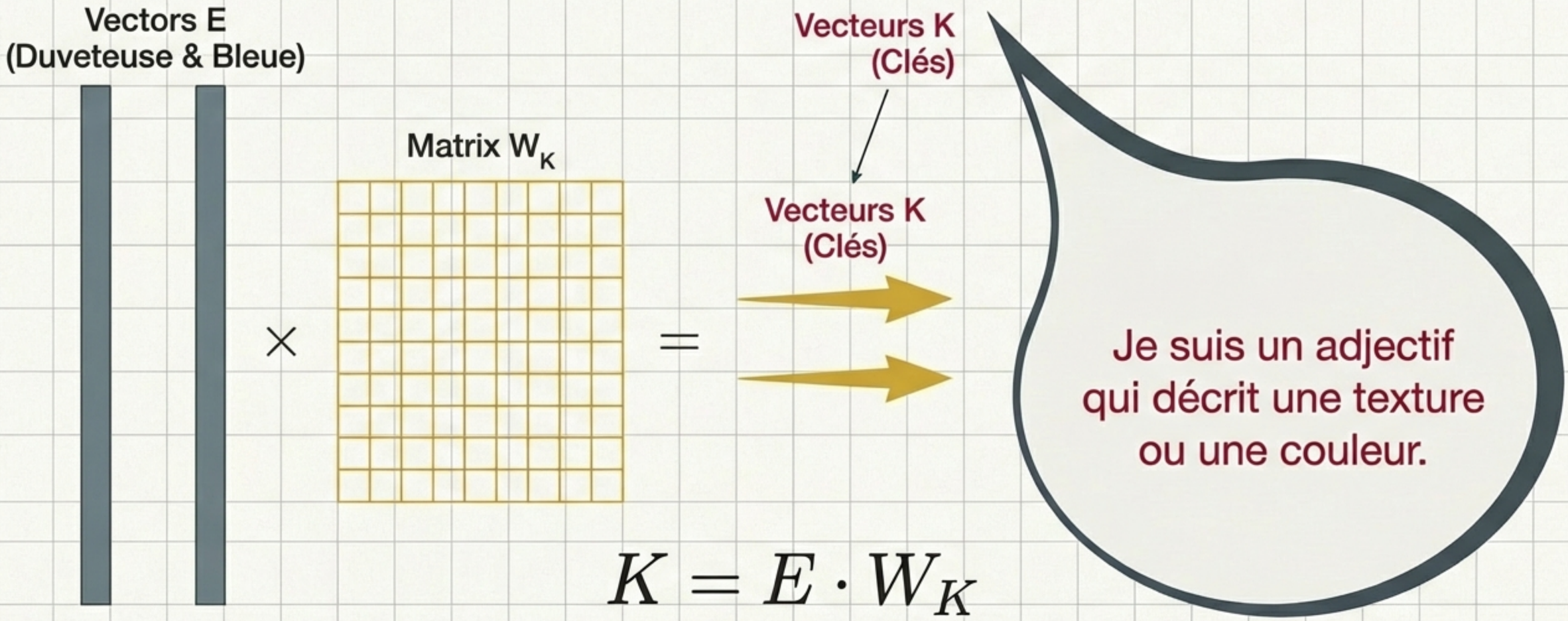
=

Vecteur Q
(Requête)

$$Q = E \cdot W_Q$$

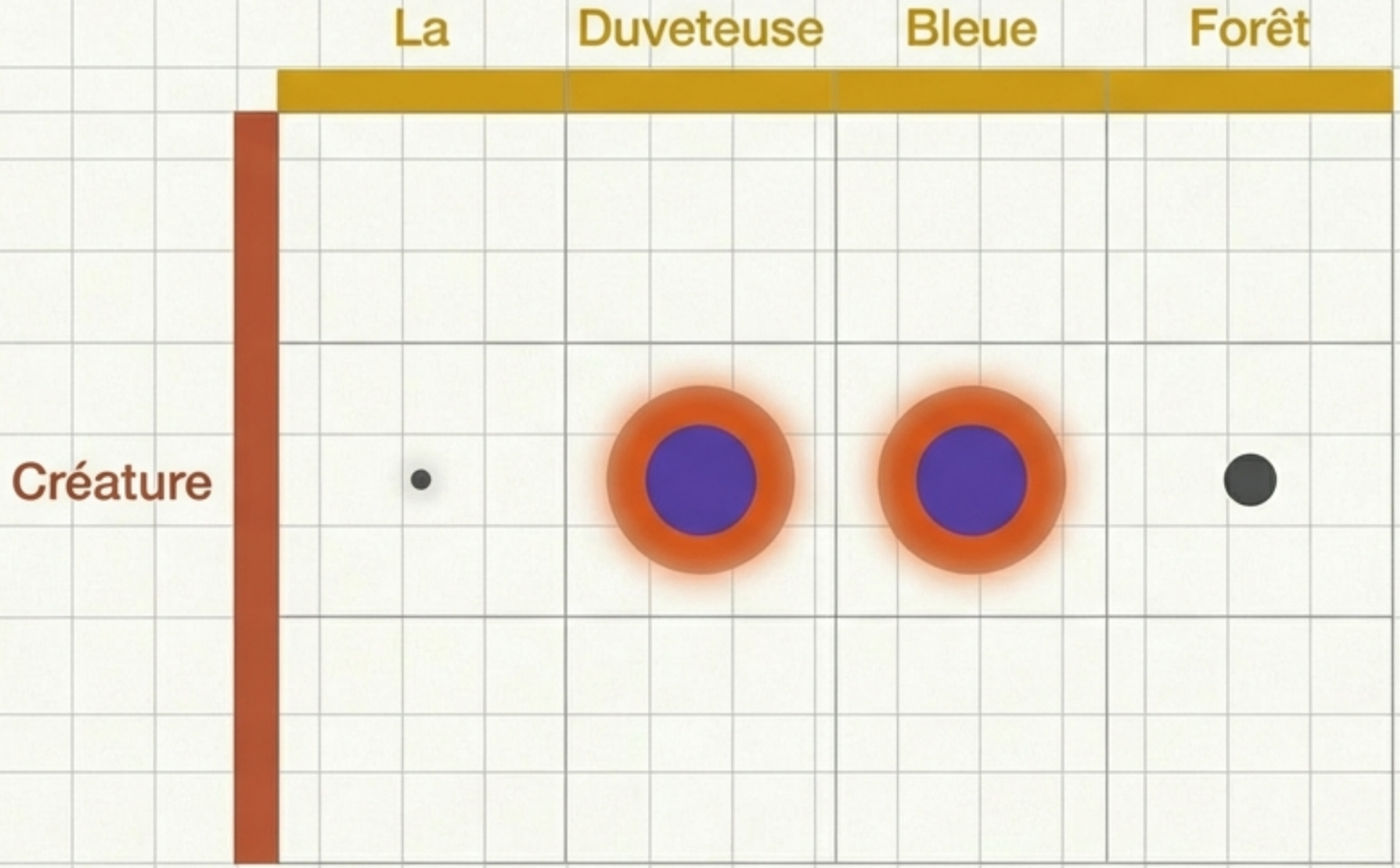
Hé, y a-t-il des adjectifs devant moi ?

La Clé (Key)



$$K = E \cdot W_K$$

Le Score d’Alignement (Dot Product)



Nous mesurons la pertinence entre la Requête (Orange) et la Clé (Jaune) via un produit scalaire.

- Un score élevé = Forte relation sémantique.
- Un score faible = Aucune relation (ex: “La”).

Normalisation (Softmax)

Softmax convertit les scores bruts ($-\infty$ à $+\infty$) en une distribution de probabilité (0 à 1).

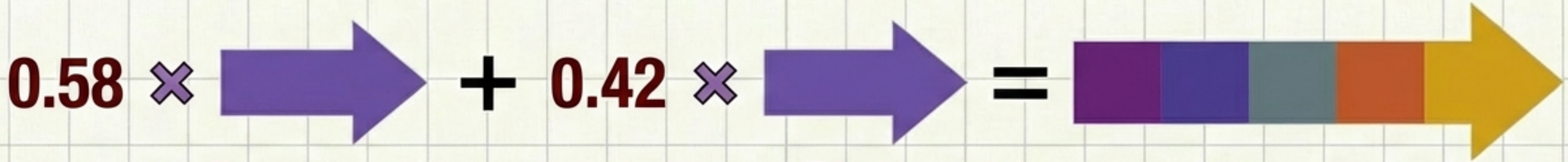
Ceci crée le Motif d'Attention : À quel point 'Créature' doit-il écouter 'Duveteuse' ?

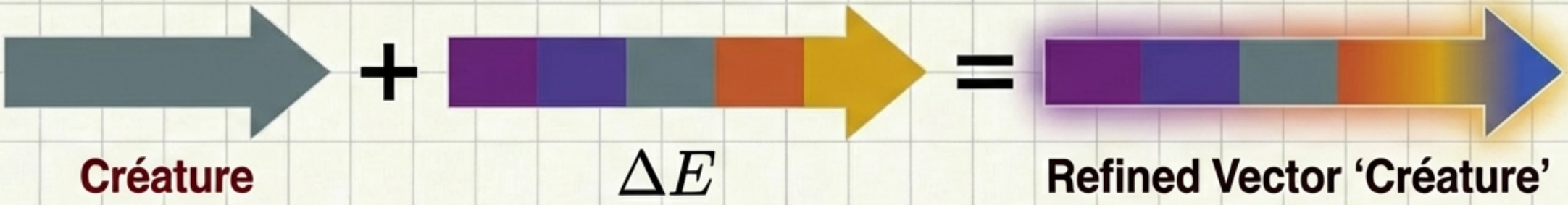
La Valeur (Value)

Ce vecteur **V** vit dans l'espace du plongement original. Il contient l'information sémantique (ex: aspect fourrure, couleur bleue) prête à être transférée.

$$V = E \cdot W_V$$

La Mise à Jour (Weighted Sum)





Le vecteur de ‘Créature’ absorbe une fraction des vecteurs de Valeur de ses adjectifs.
Résultat : Le mot encode désormais un ‘Créature-Bleue-Duveteuse’.

L’Équation Complète

La grille de produits scalaires (Pertinence)

Conversion en probabilités (0 à 1)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Stabilisation numérique

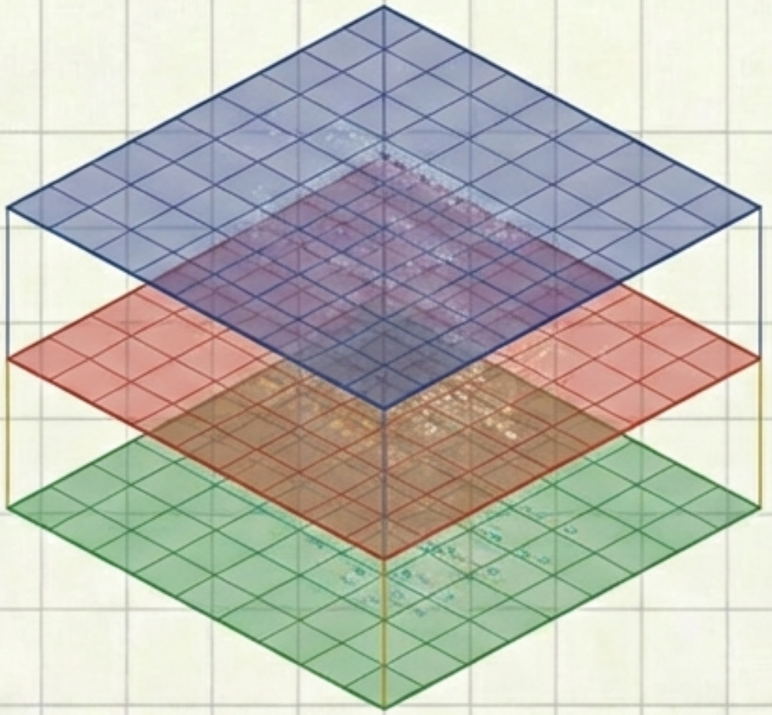
L’information sémantique à transférer

Une formulation compacte pour une intuition linguistique complexe.

Masquage (Masking)

Lors de l'entraînement, le modèle prédit le mot suivant. Il ne doit pas tricher en regardant le futur.
La solution : Forcer les scores d'attention future à $-\infty$ avant le Softmax.

L’Attention Multi-Têtes



Tête 1: Grammaire

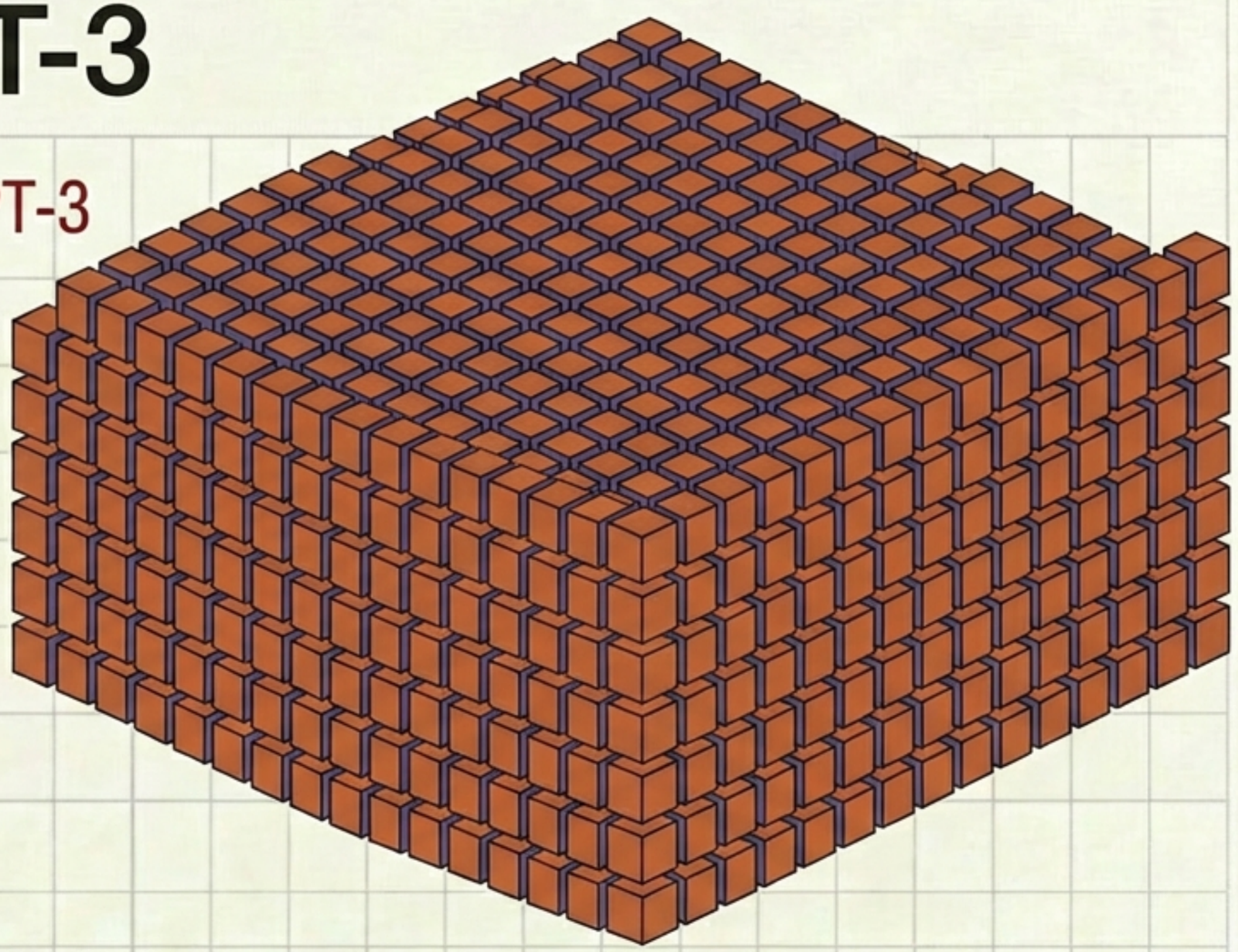
Tête 2: Sentiment

Tête 3: Faits

Une seule tête ne suffit pas. Le contexte a plusieurs dimensions.
Chaque tête possède ses propres matrices W_Q, W_K, W_V distinctes.
Les résultats sont concaténés pour former une représentation ultra-nuancée.

L’Échelle de GPT-3

Bloc GPT-3



Une Tête

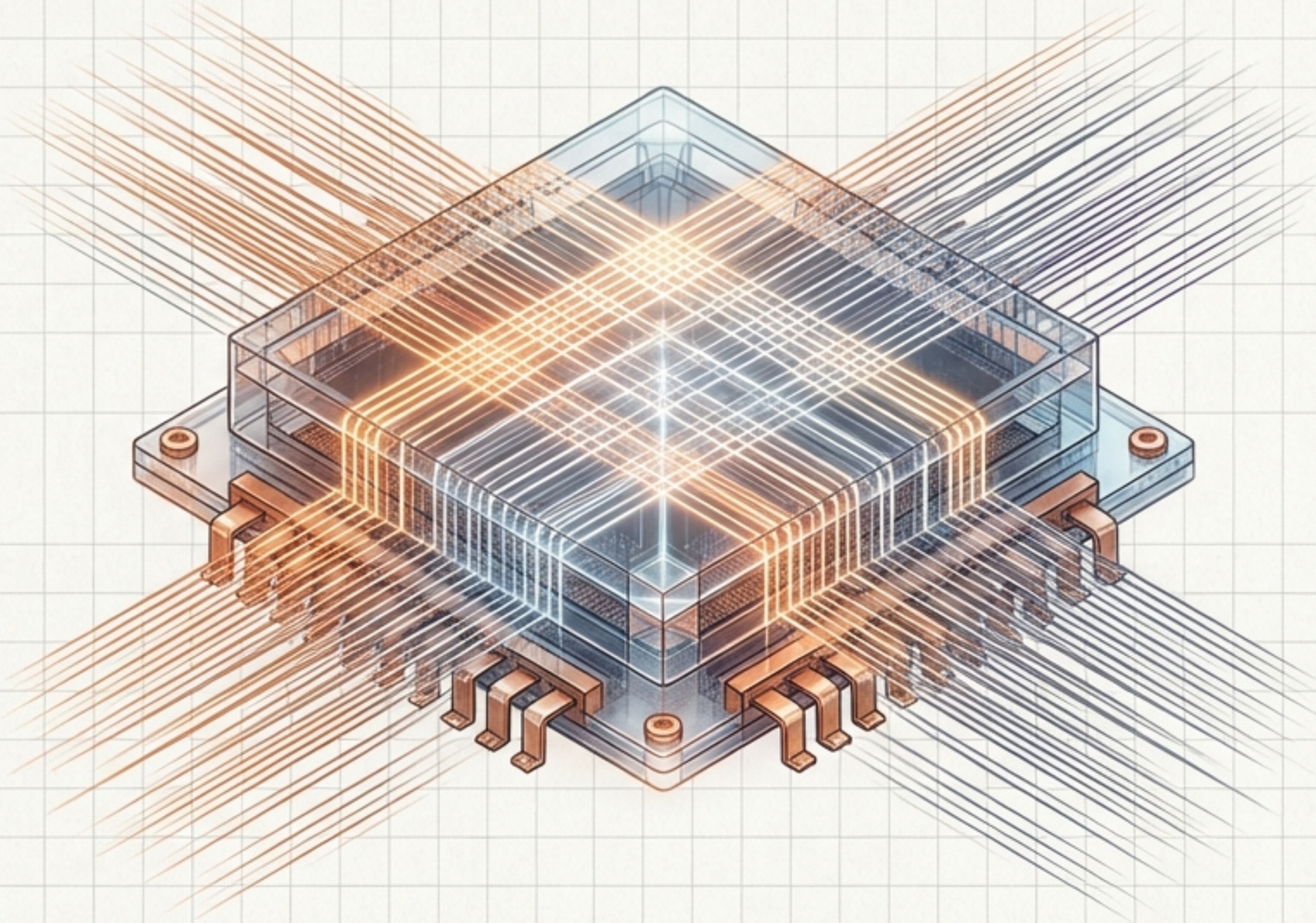
6.3 millions de paramètres

96 Têtes d’attention par bloc.
96 Couches de profondeur.
58 Milliards de paramètres dévoués uniquement à l’attention.
Coût computationnel : N^2 (carré de la fenêtre contextuelle).

Pourquoi l’Attention a Tout Changé

Helvetica Now Display

- **Nuance** : Les vecteurs intègrent des concepts abstraits (ton, style, sentiment).
- **Parallélisme** : Contrairement aux RNNs, l’Attention traite toute la séquence simultanément.
- **Échelle** : Capacité à s’entraîner sur des GPUs massifs.



Pour aller plus loin : 3Blue1Brown - Deep Learning Chapter 6.