Institut des systèmes complexes
Complex Systems Institute
Rhône-Alpes

IXXI

Grands Modèles de Langue

Jean-Philippe Magué

21 juin 2024

Ressources

2 notebooks → Google Colab

https://bit.ly/formation_LLM_IXXI_1

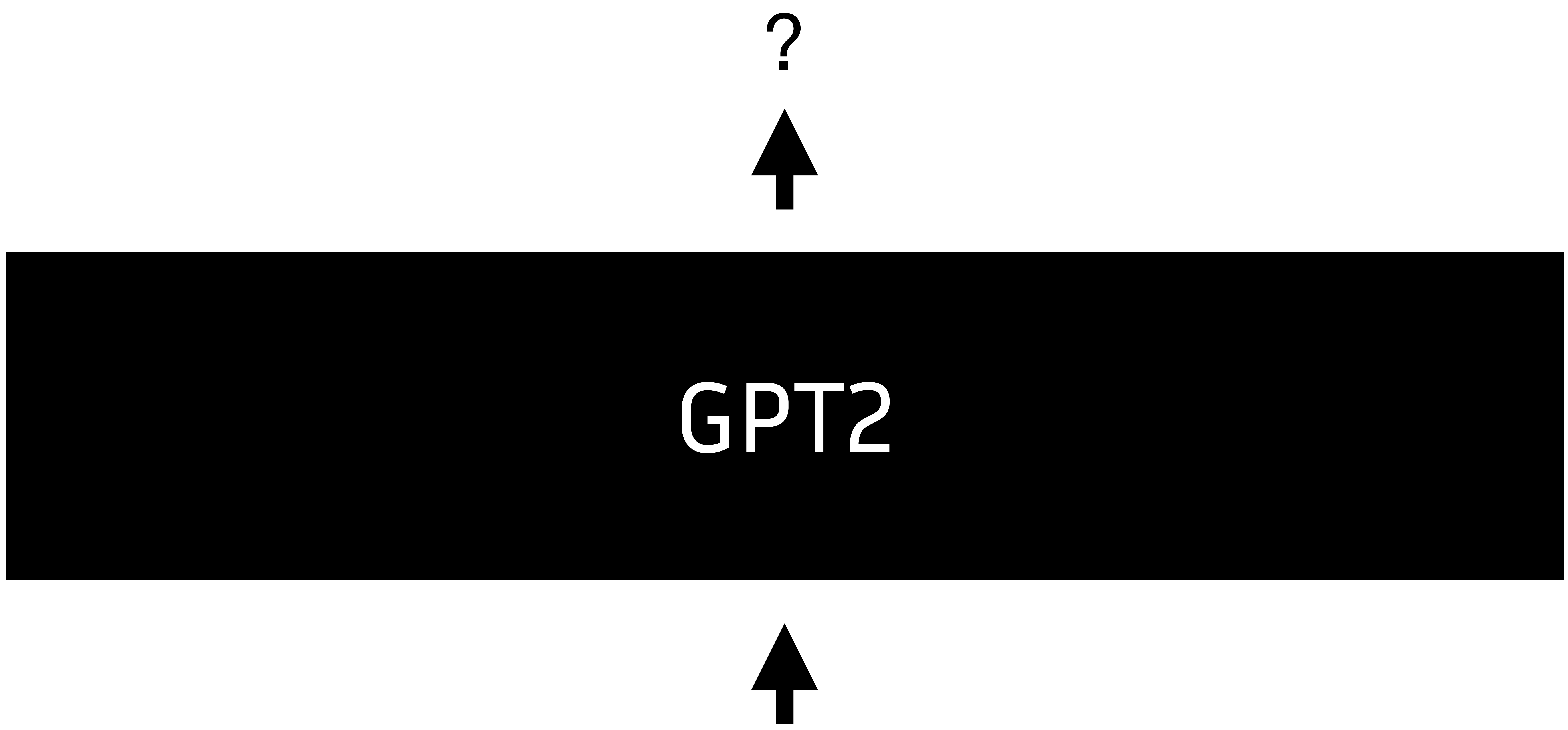
https://bit.ly/formation_LLM_IXXI_2

2 notebooks → GitHub

Diapos → GitHub

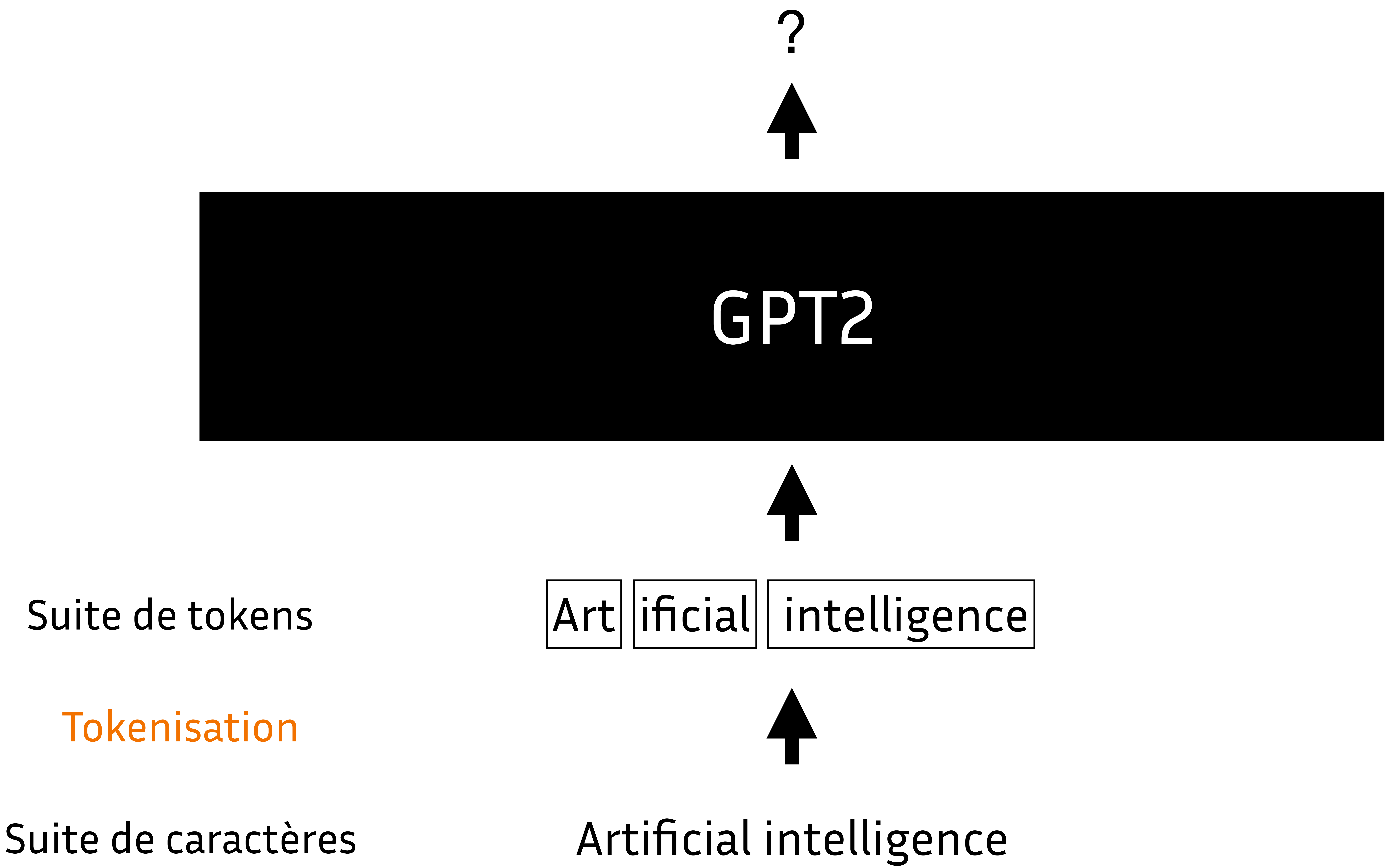
https://github.com/jmague/formation_LLM_IXXI

Premières générations de texte

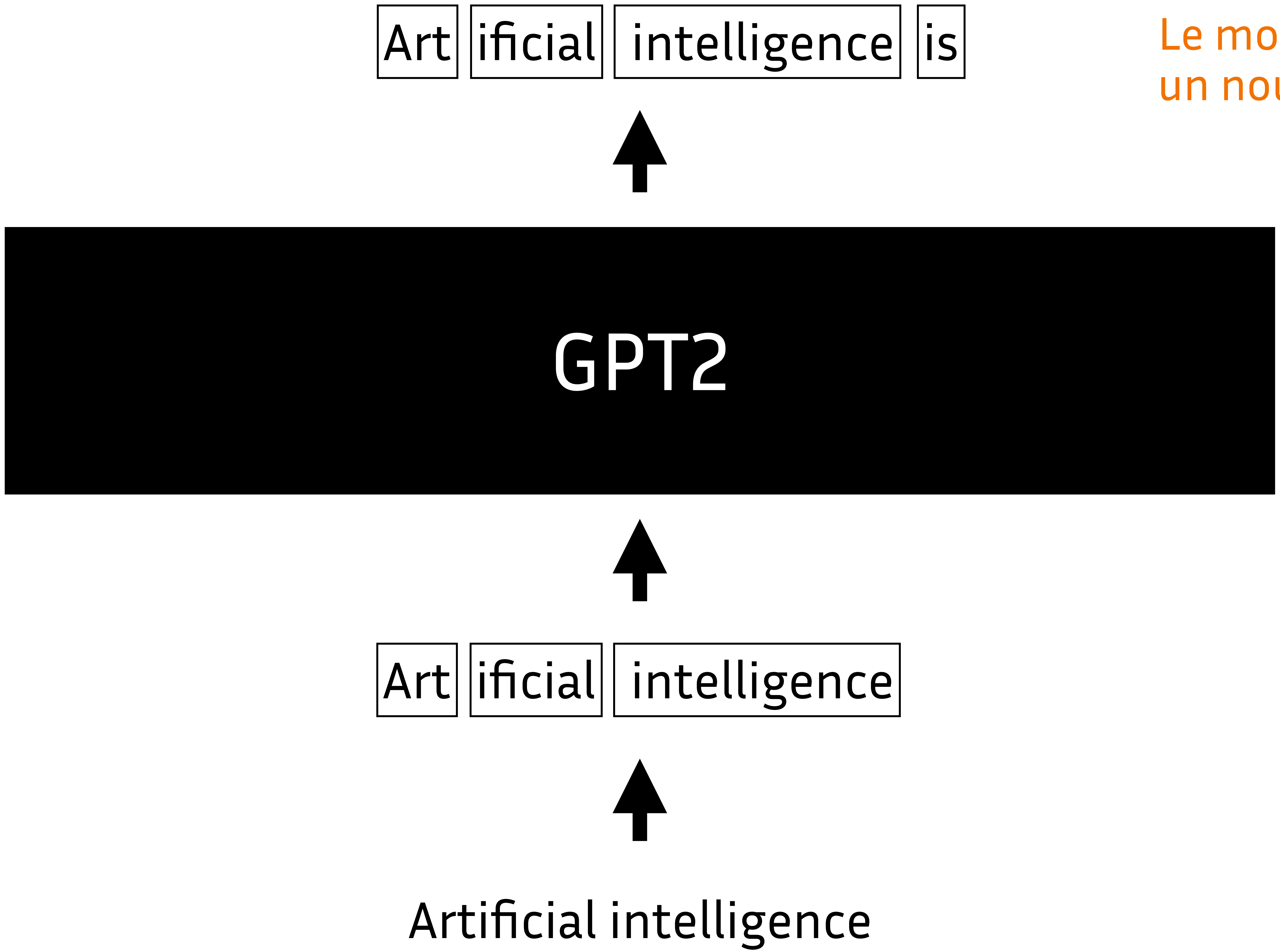


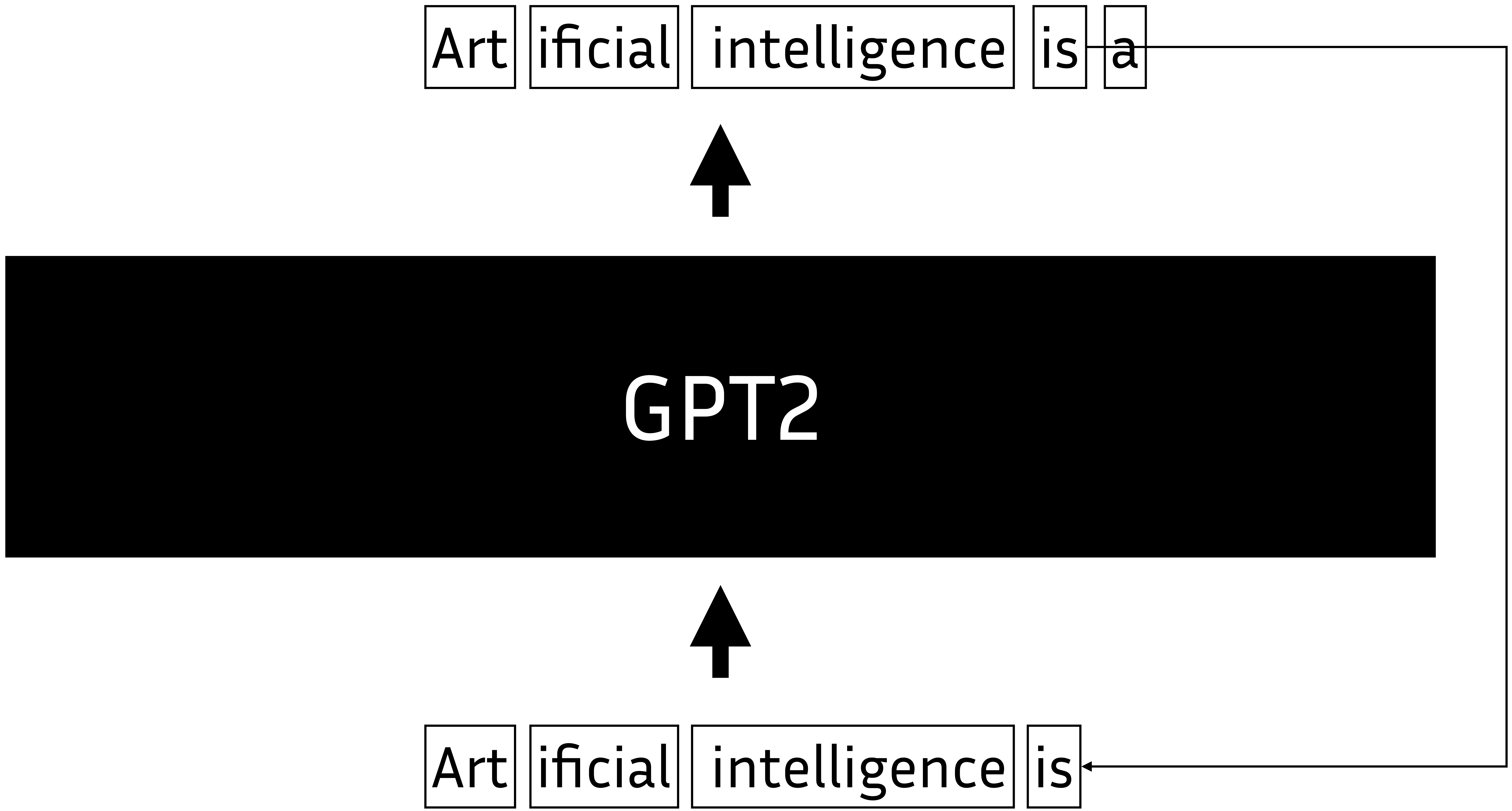
Artificial intelligence

Modèles génératifs : produisent du texte qui complète le *prompt* qui leur est donné


?
GPT2
Suite de tokens
Art
ificial
intelligence
Tokenisation
Suite de caractères
Artificial intelligence

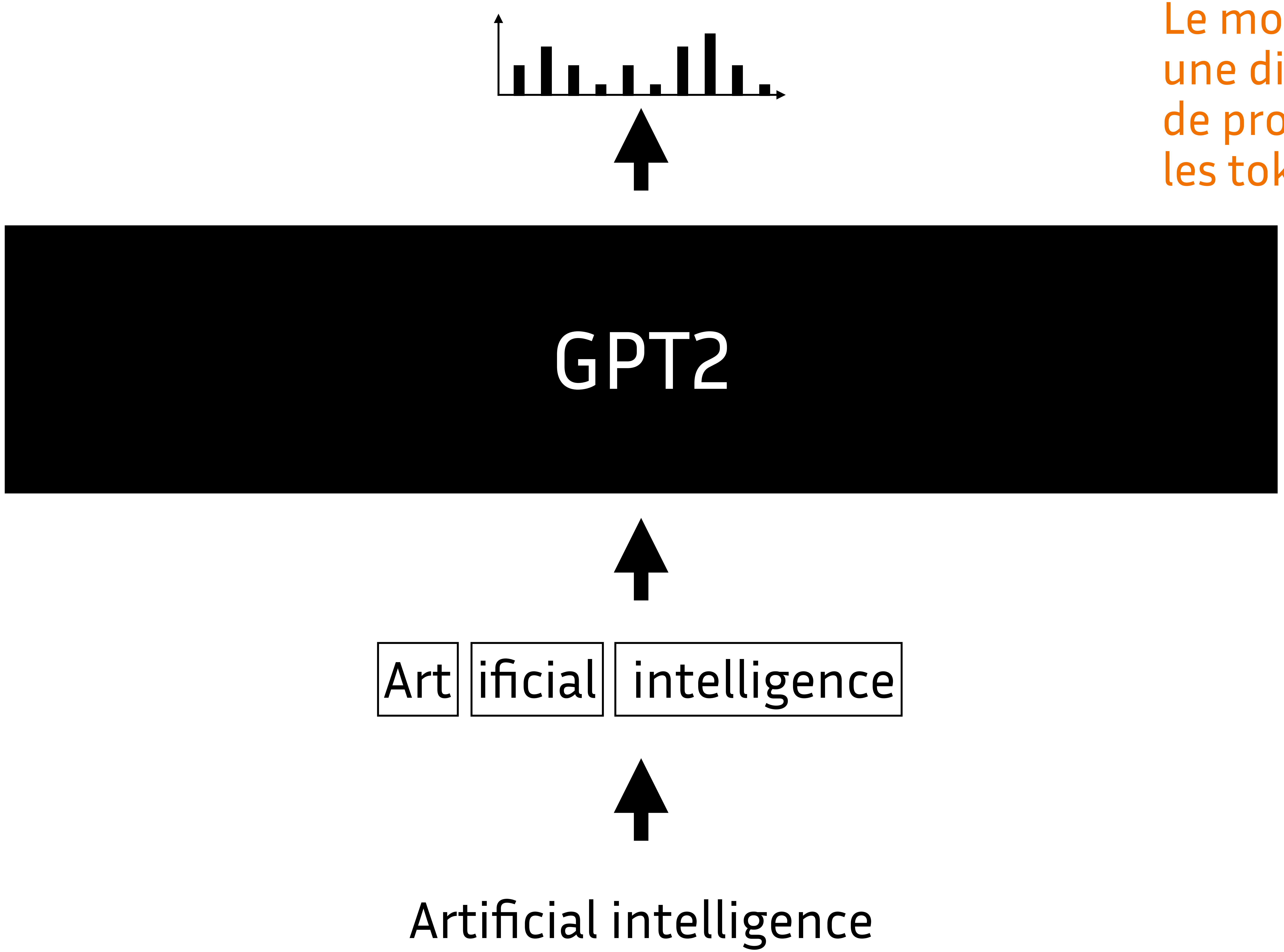
Le modèle génère
un nouveau token

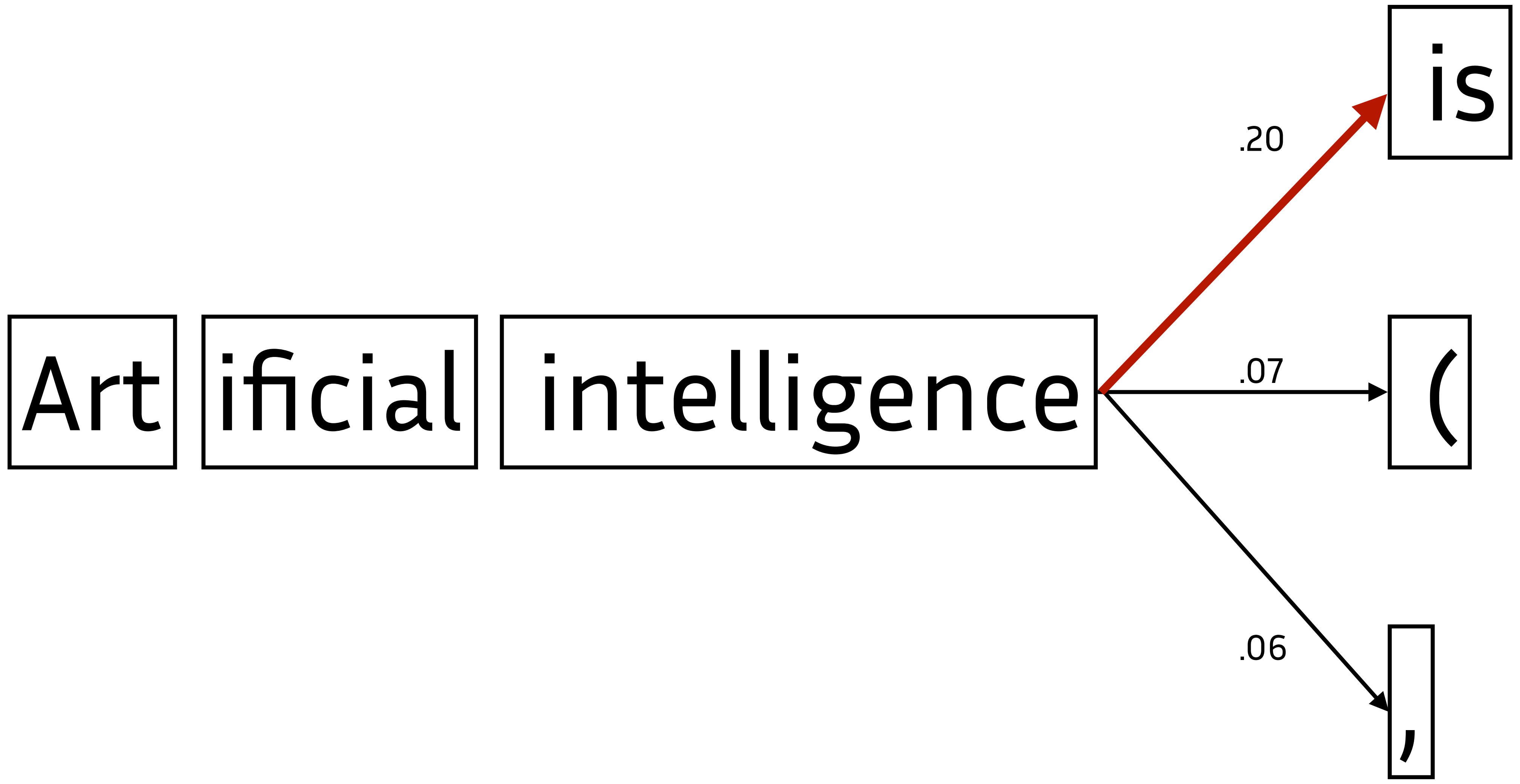



Art
ificial
intelligence
is
a
GPT2
Art
ificial
intelligence
is

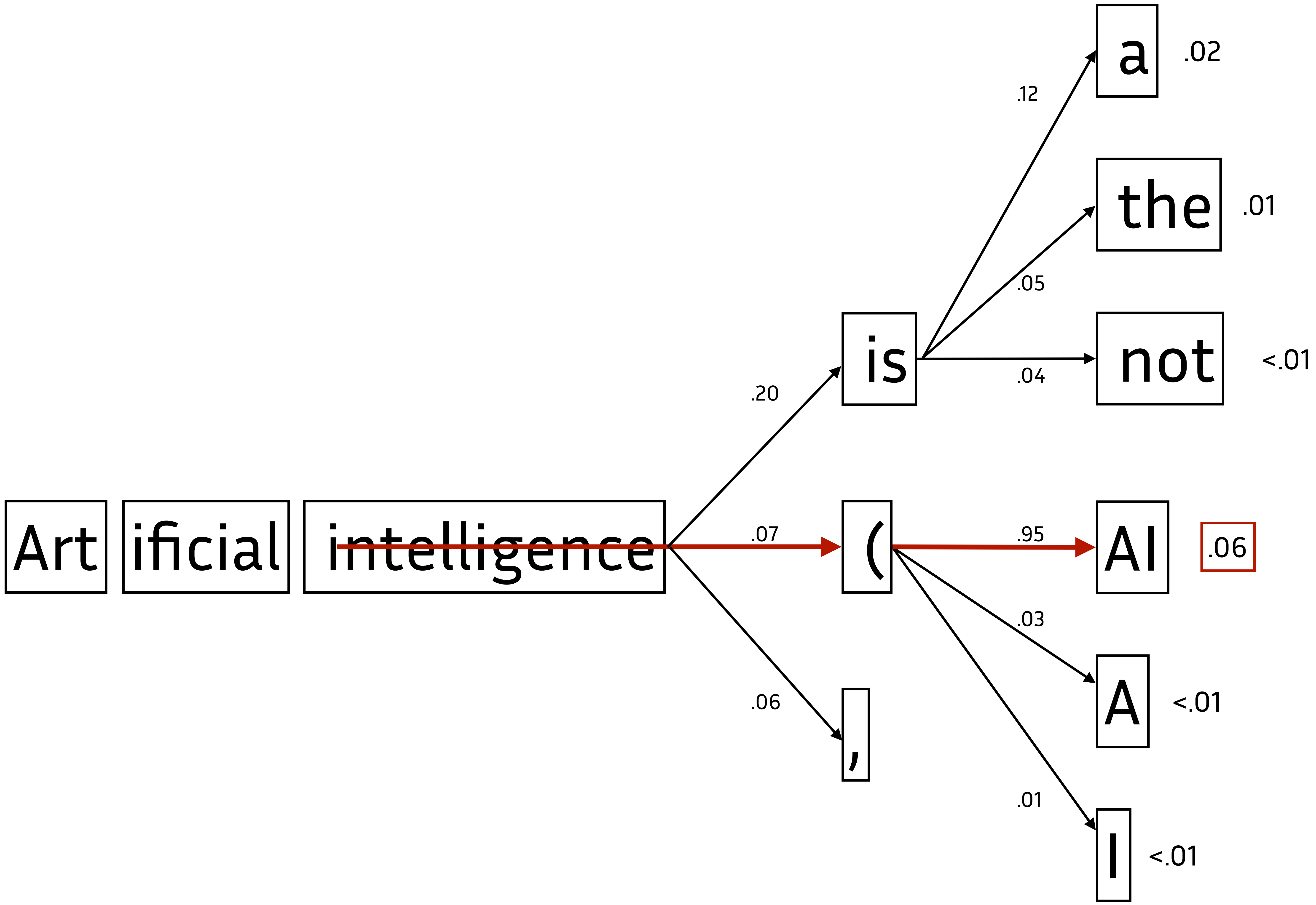
Stratégies de génération de texte

Le modèle génère une distribution de probabilité sur les tokens





Greedy search

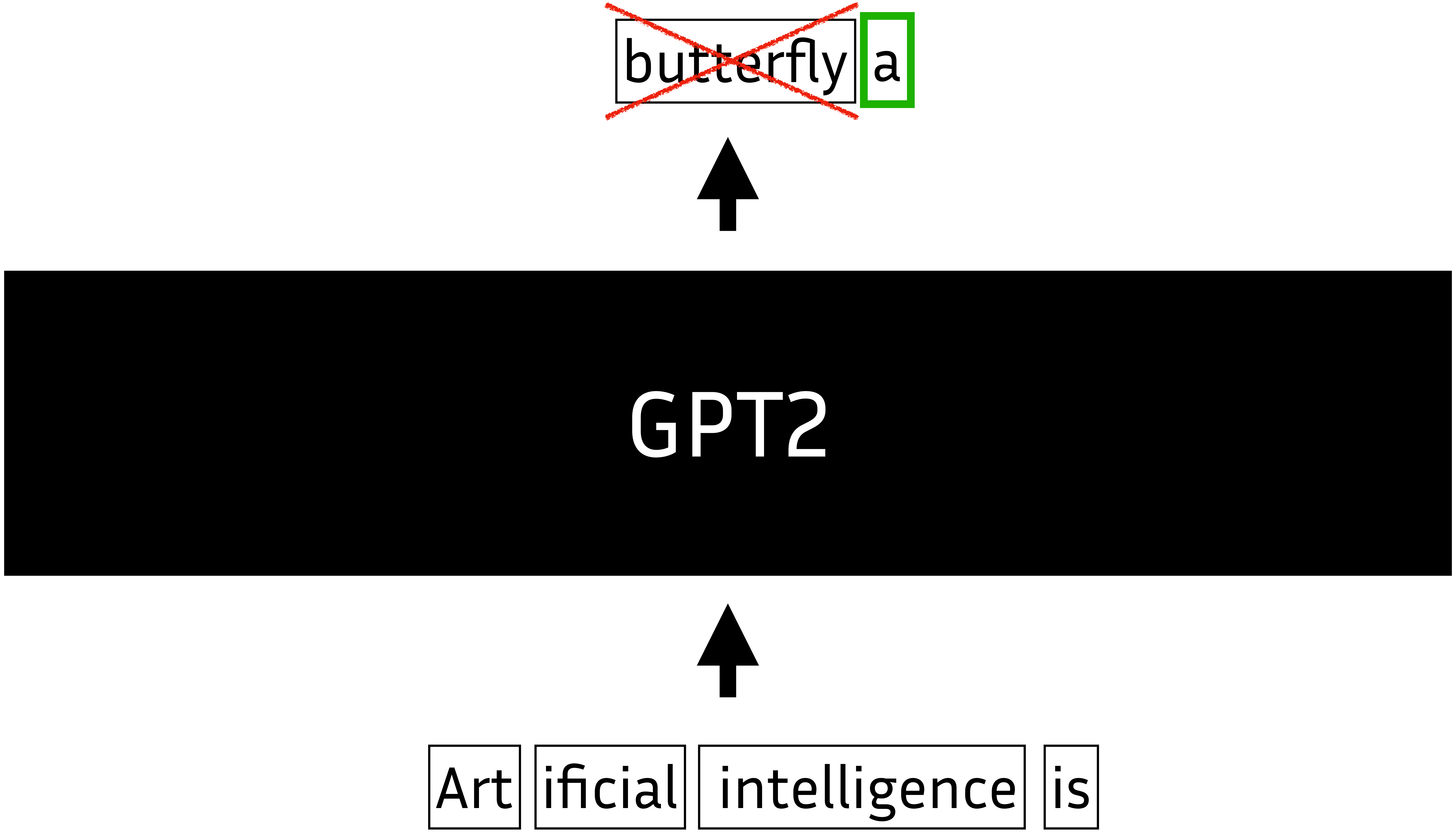

Art
ificial
intelligence
.20
.07
.06
is
(
,
.12
.05
.04
a
.02
the
.01
not
<.01
.95
.03
.01
AI
.06
A
<.01
I
<.01

Beam search

Apprentissage

Apprentissage

Apprentissage auto-supervisé



Apprentissage

Instruction tuning

The capital of Australia is Canberra.

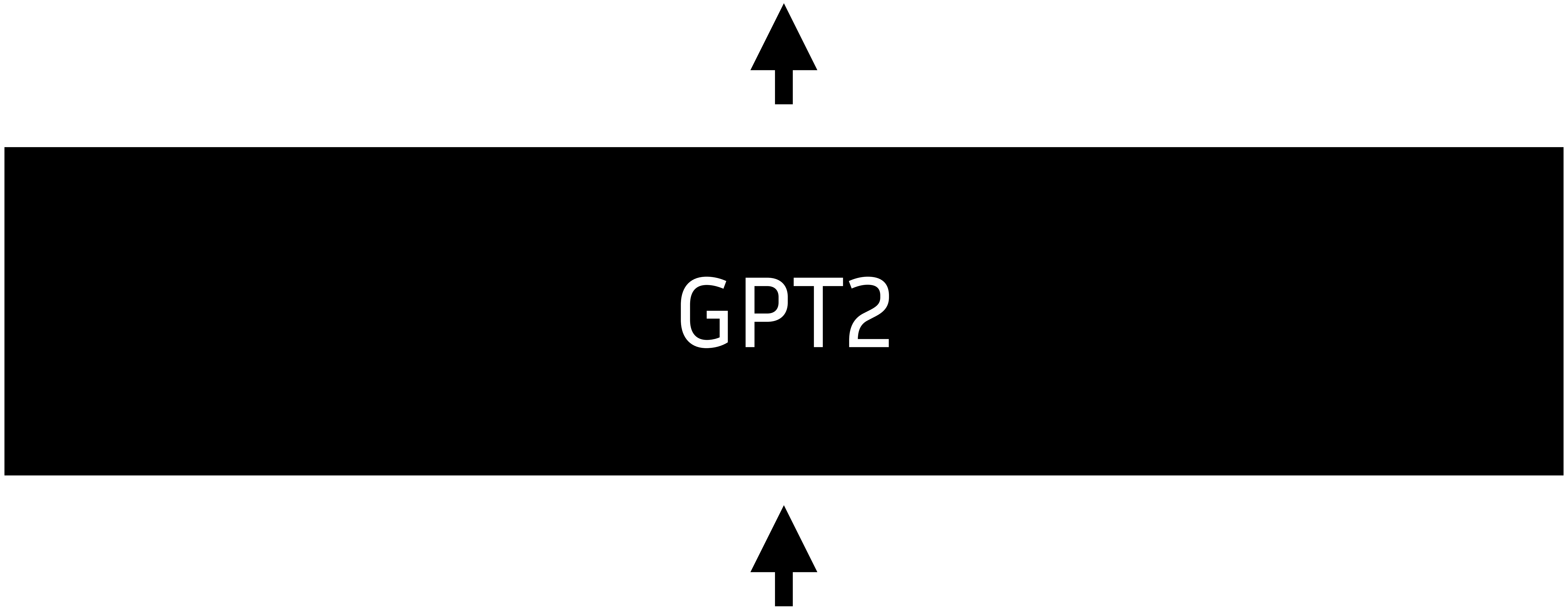
~~Is a question that people often get wrong.~~

GPT2

What is the capital of Australia?

Apprentissage

Apprentissage par
renforcement à partir
de rétroaction humaine



compared to men, women are more likely to

Apprentissage

Apprentissage par
renforcement à partir
de rétroaction humaine

compared to men, women are more likely to

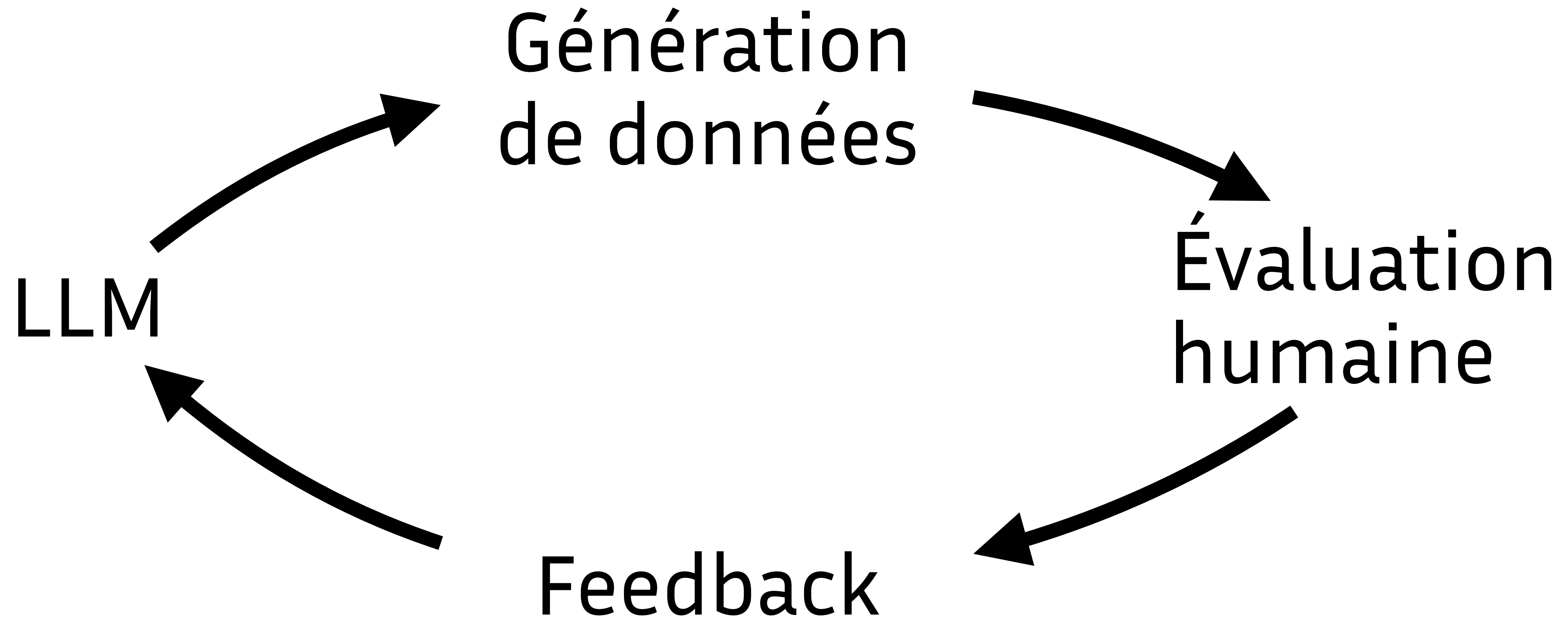
have problems eating or drinking alcohol

be underrepresented in science, technology, and academia

have experienced sexual violence

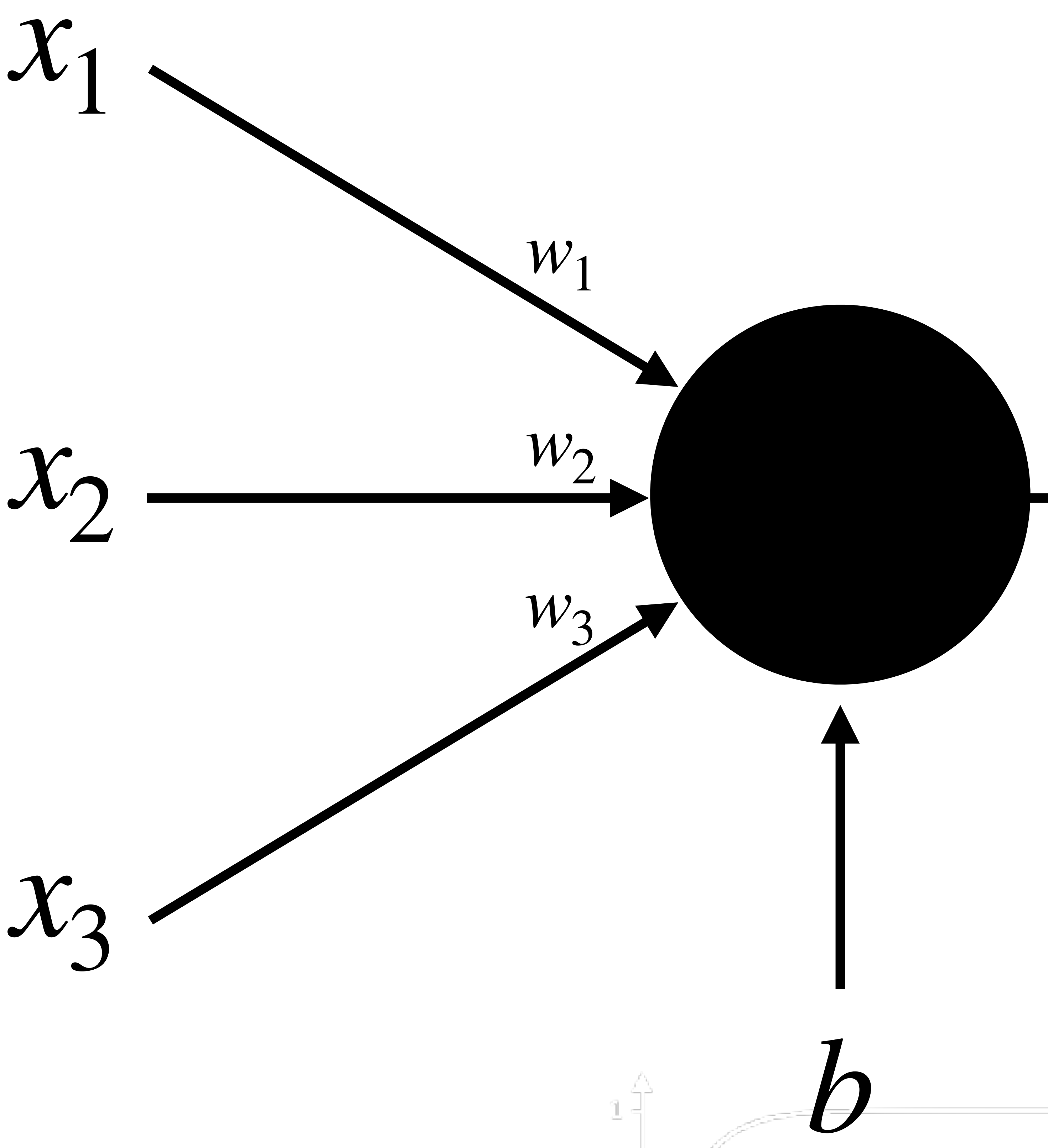
have sex when working

be overweight



Plongée dans la boîte noire

Neurone artificiel

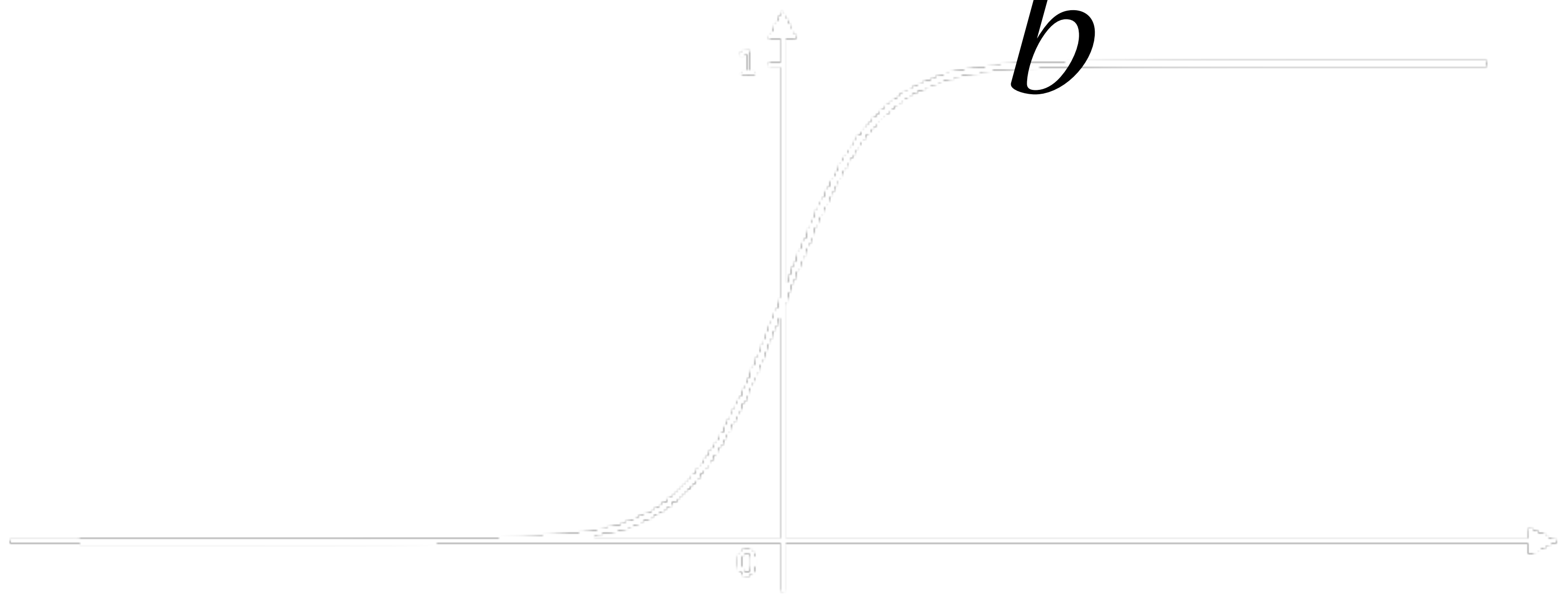


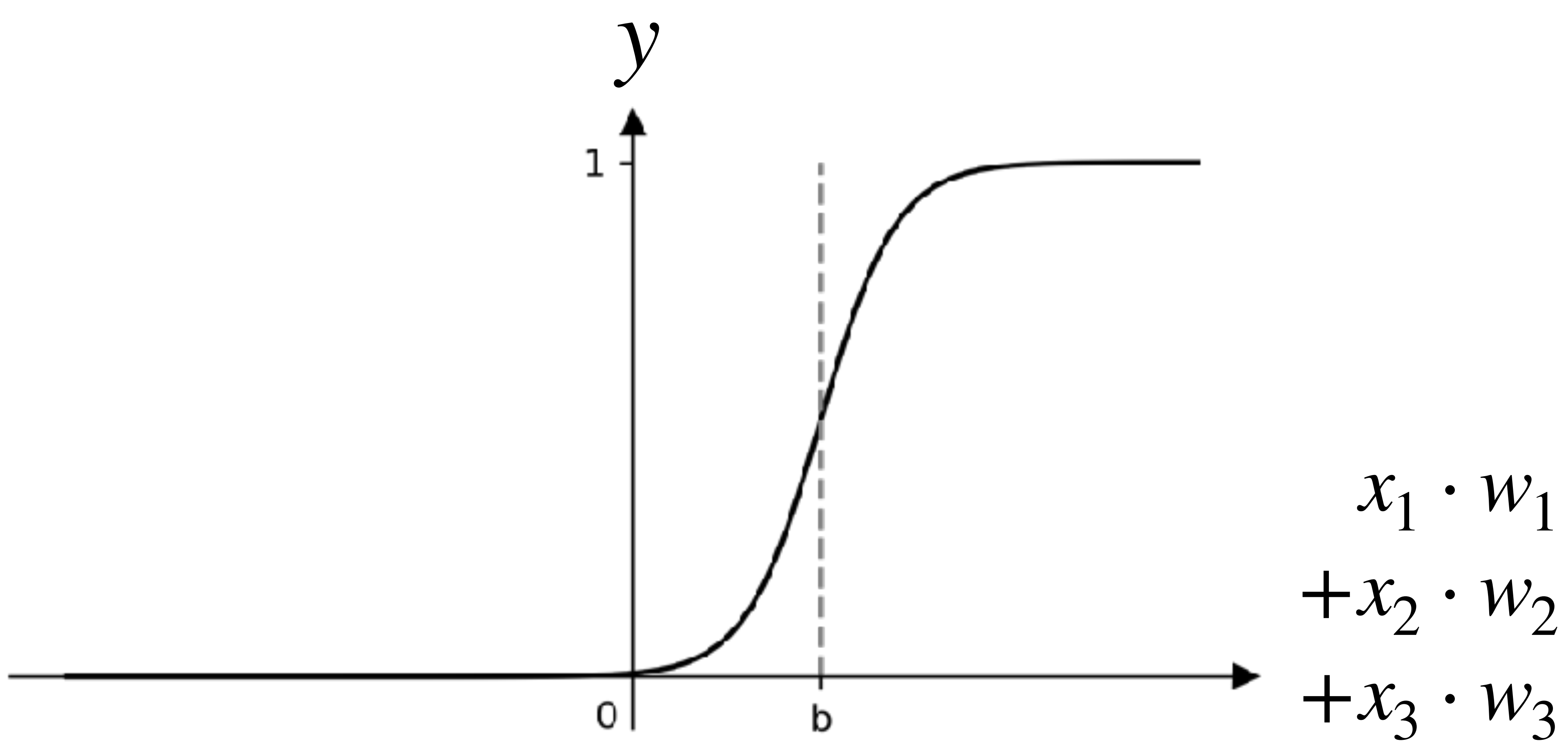
Somme pondérée des entrées

biais

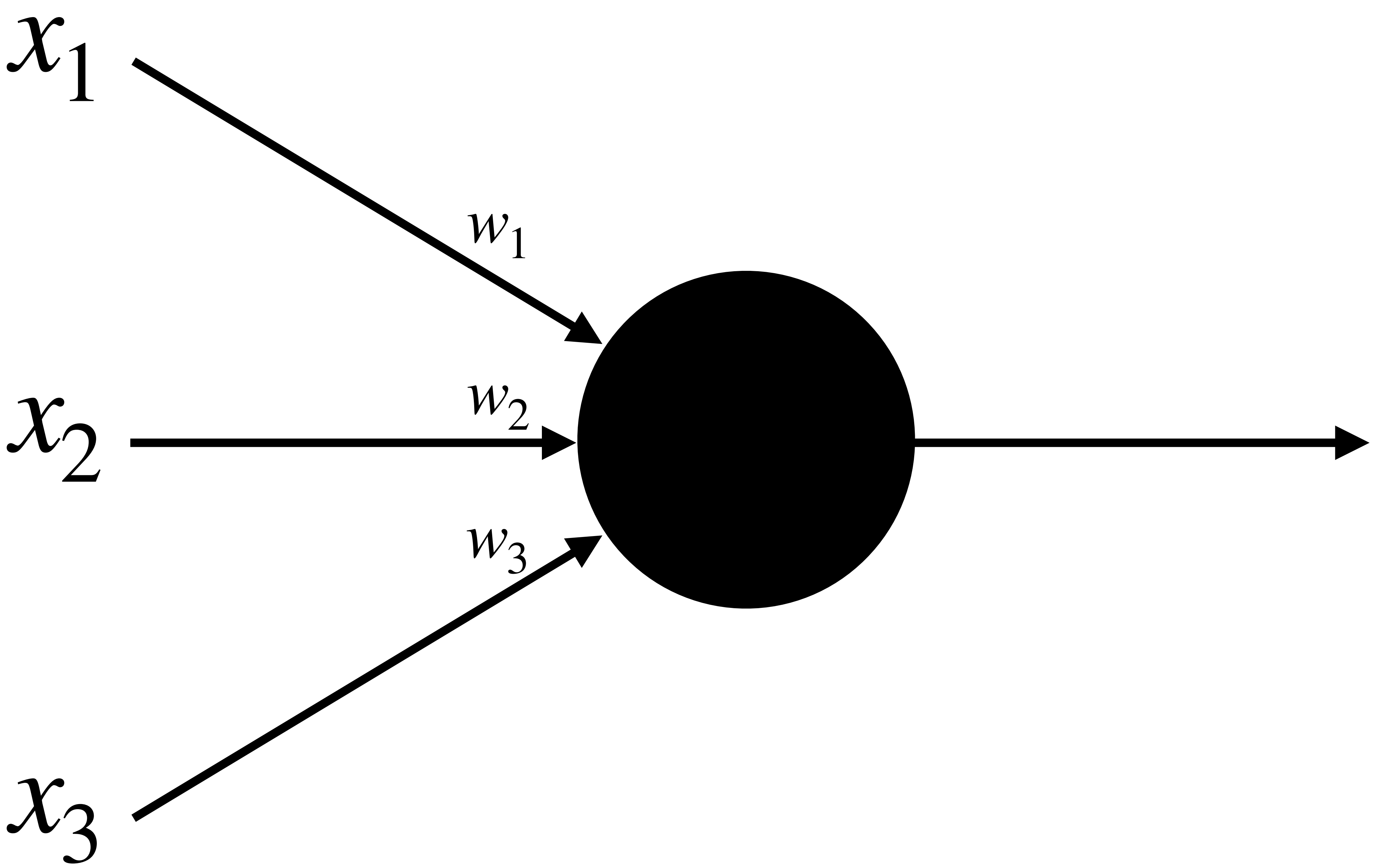
$$y = f(\overbrace{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3}^{\text{Somme pondérée des entrées}} + b)$$

Fonction d'activation

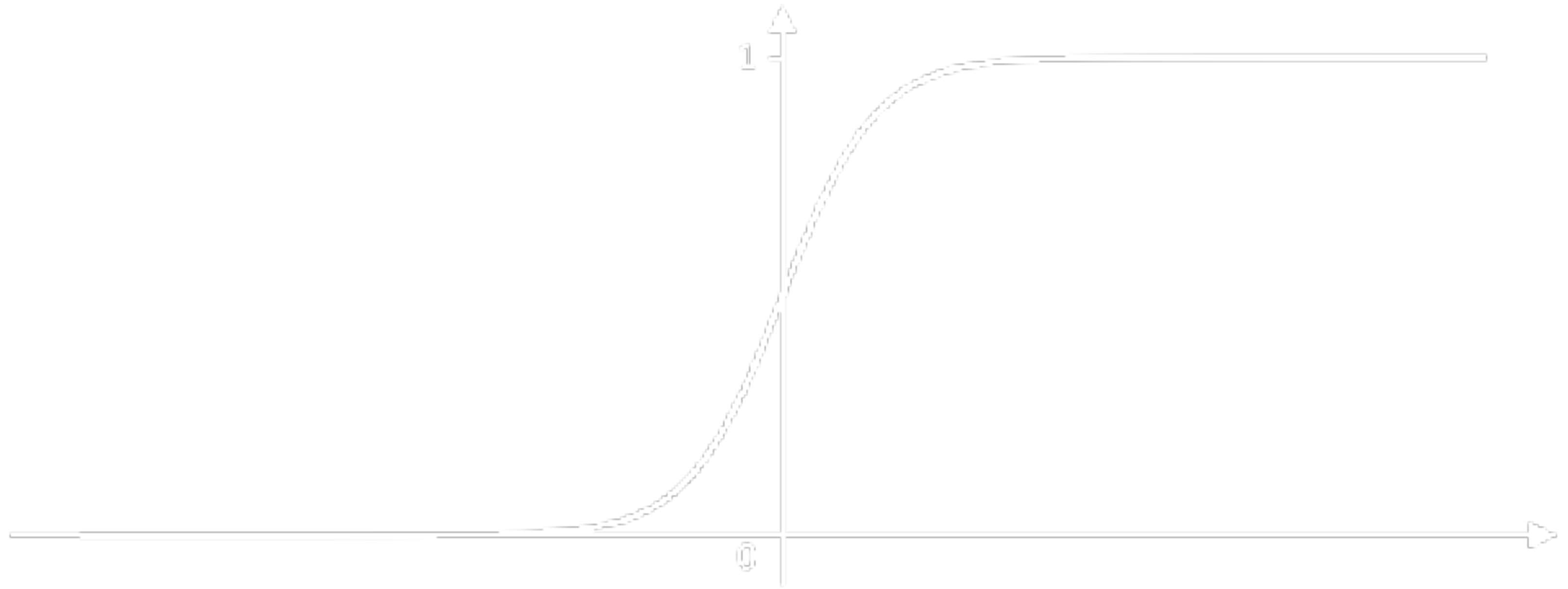


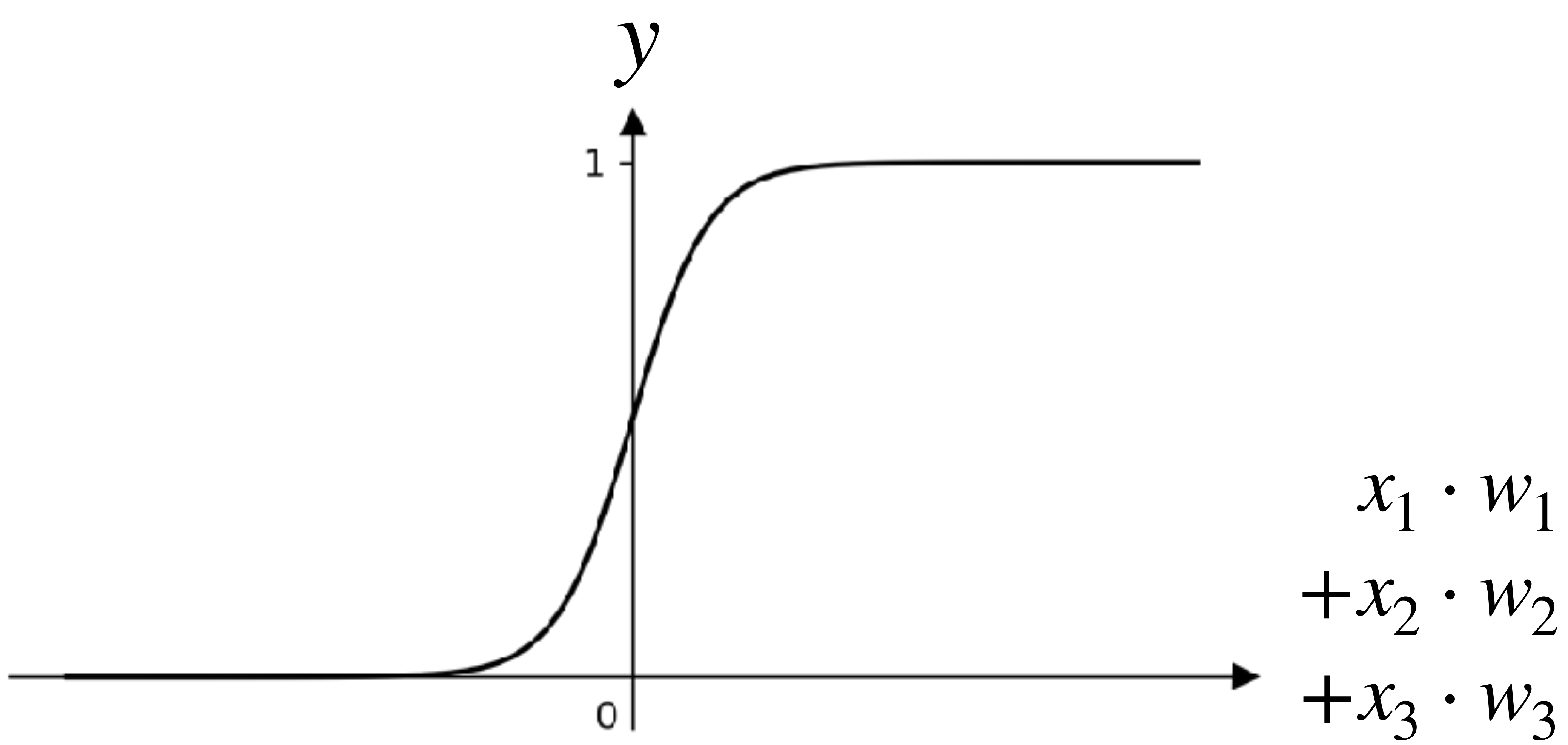


Neurone artificiel

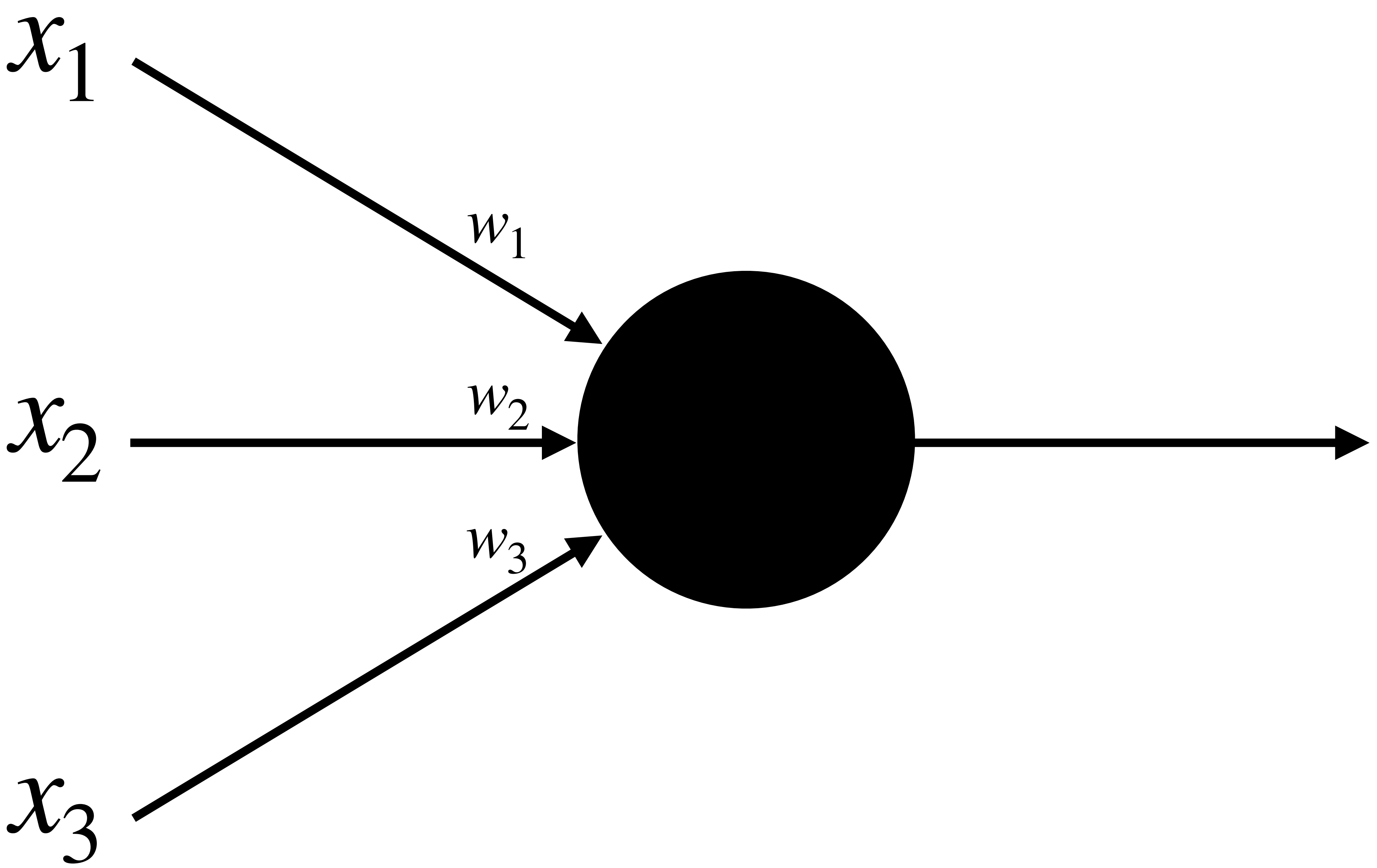


$$y = \underset{\text{Fonction d'activation}}{f}(\overbrace{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3}^{\text{Somme pondérée des entrées}})$$





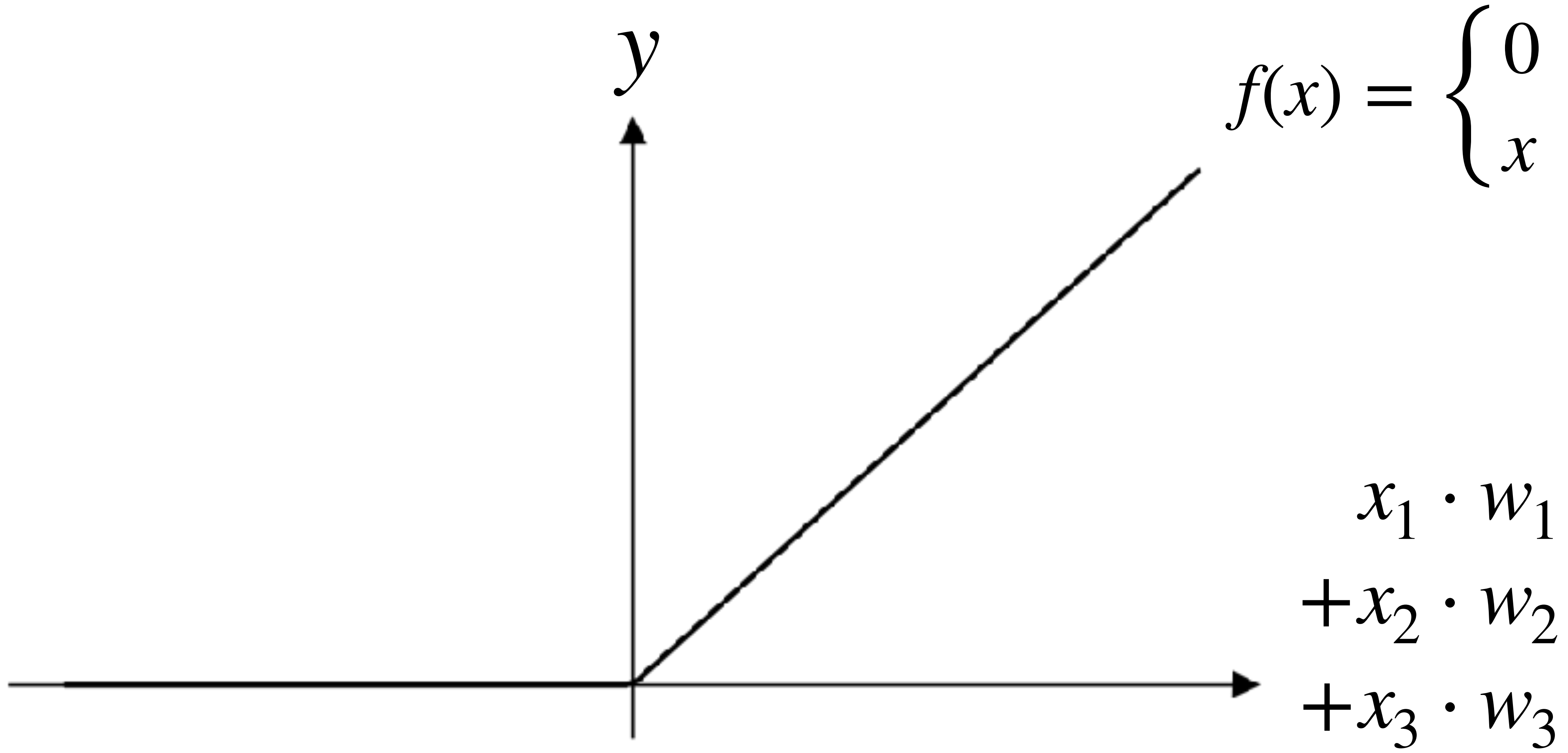
Neurone artificiel



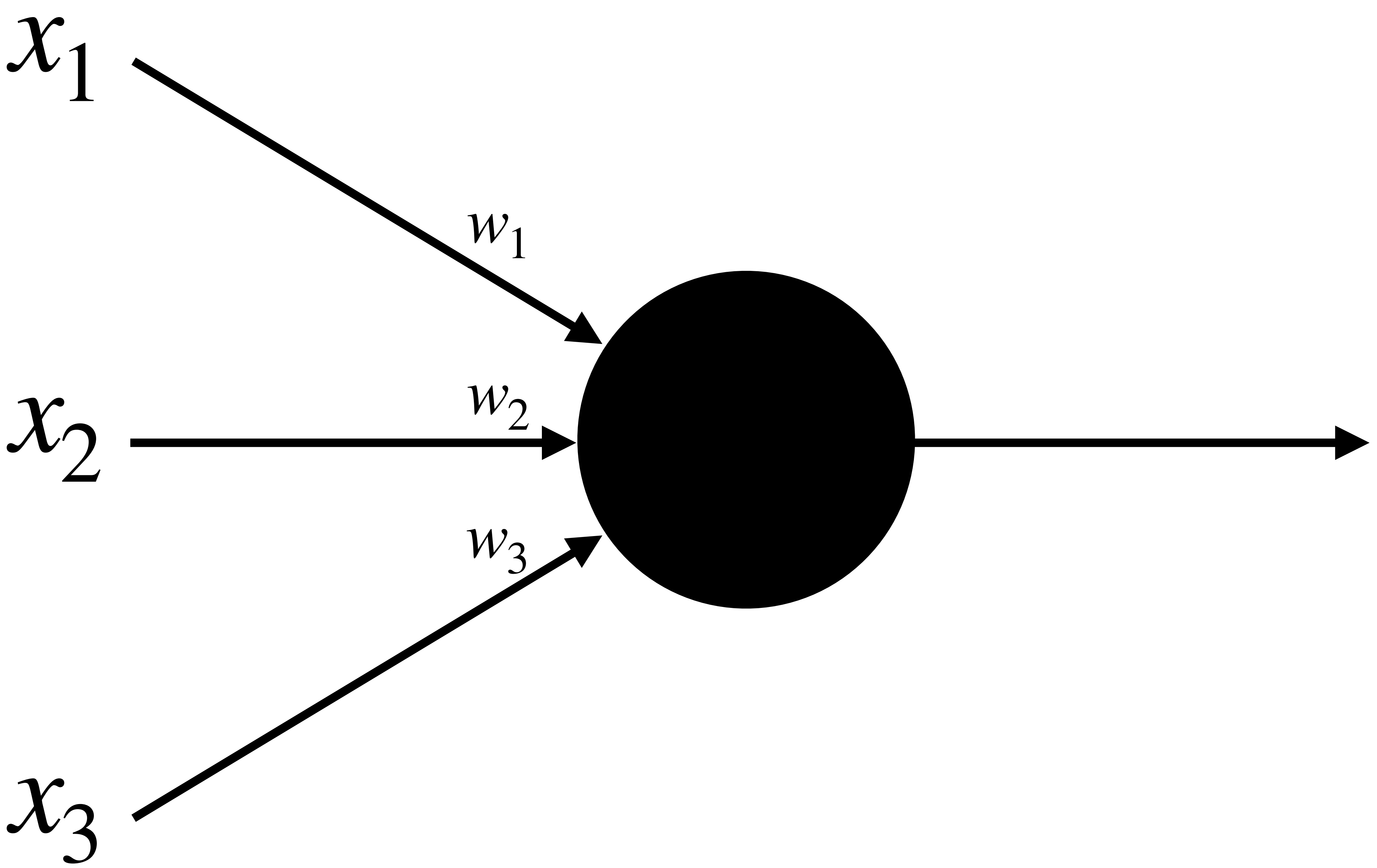
$$y = f(\overbrace{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3}^{\text{Somme pondérée des entrées}})$$

Fonction d'activation

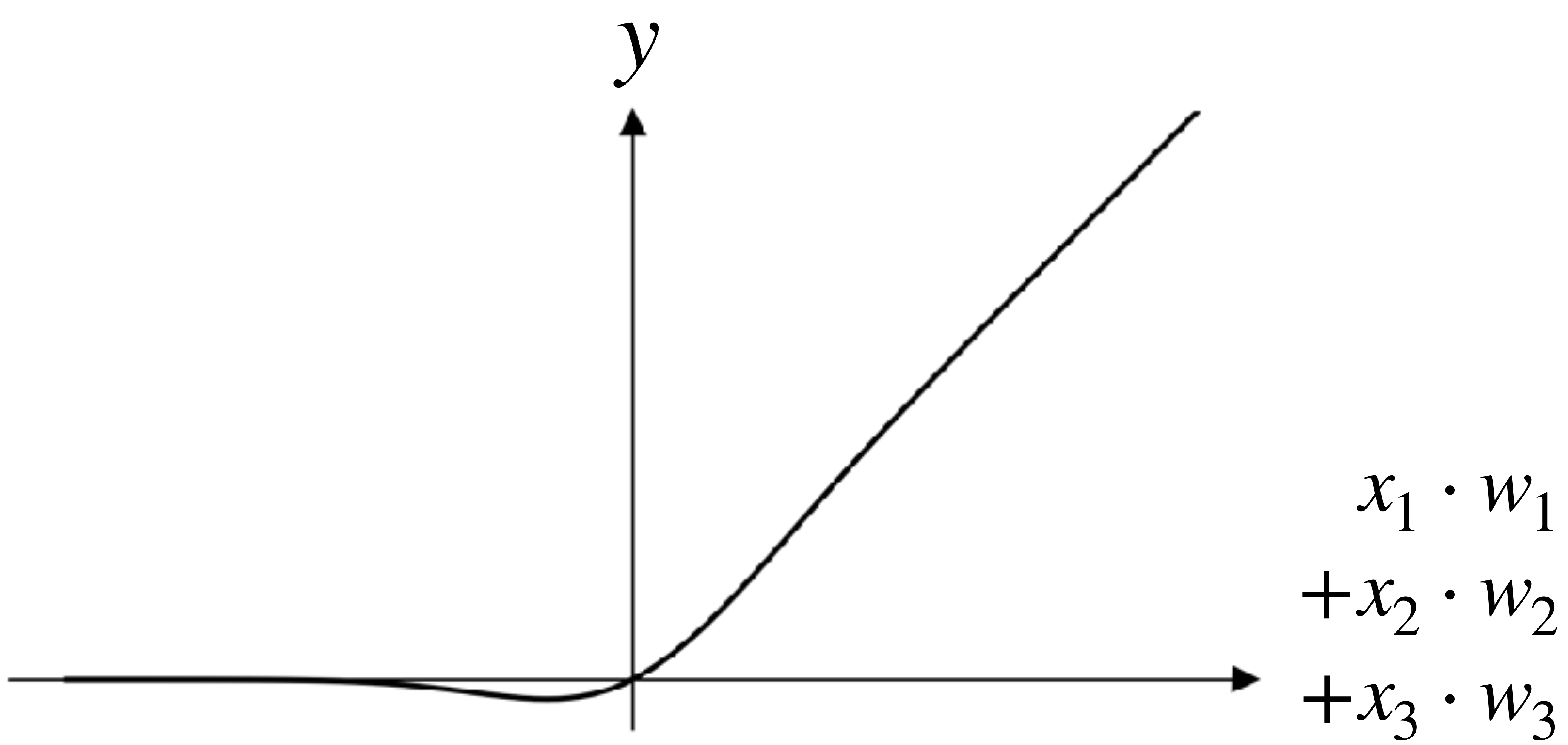
$$f(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } x > 0 \end{cases}$$



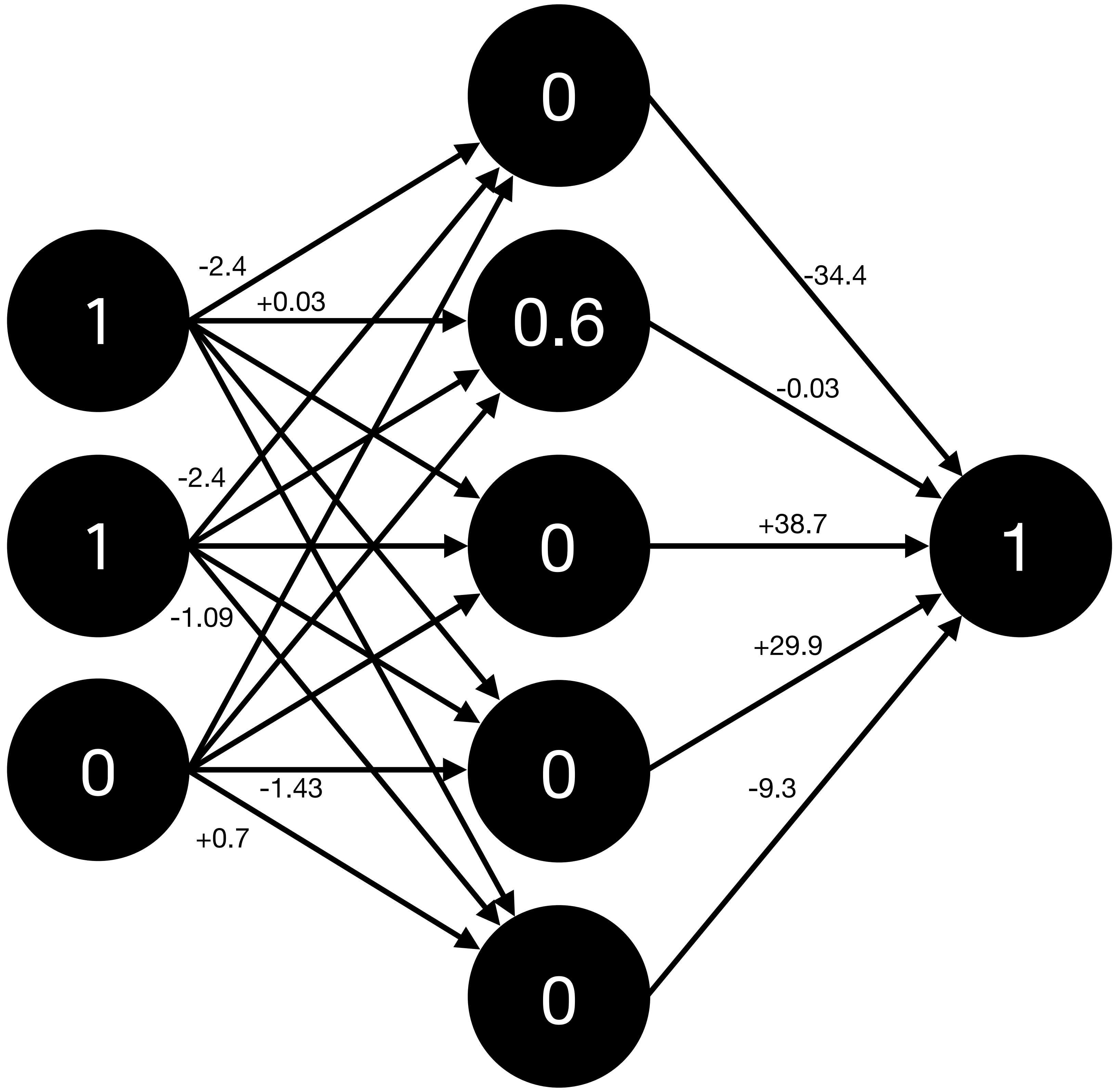
Neurone artificiel



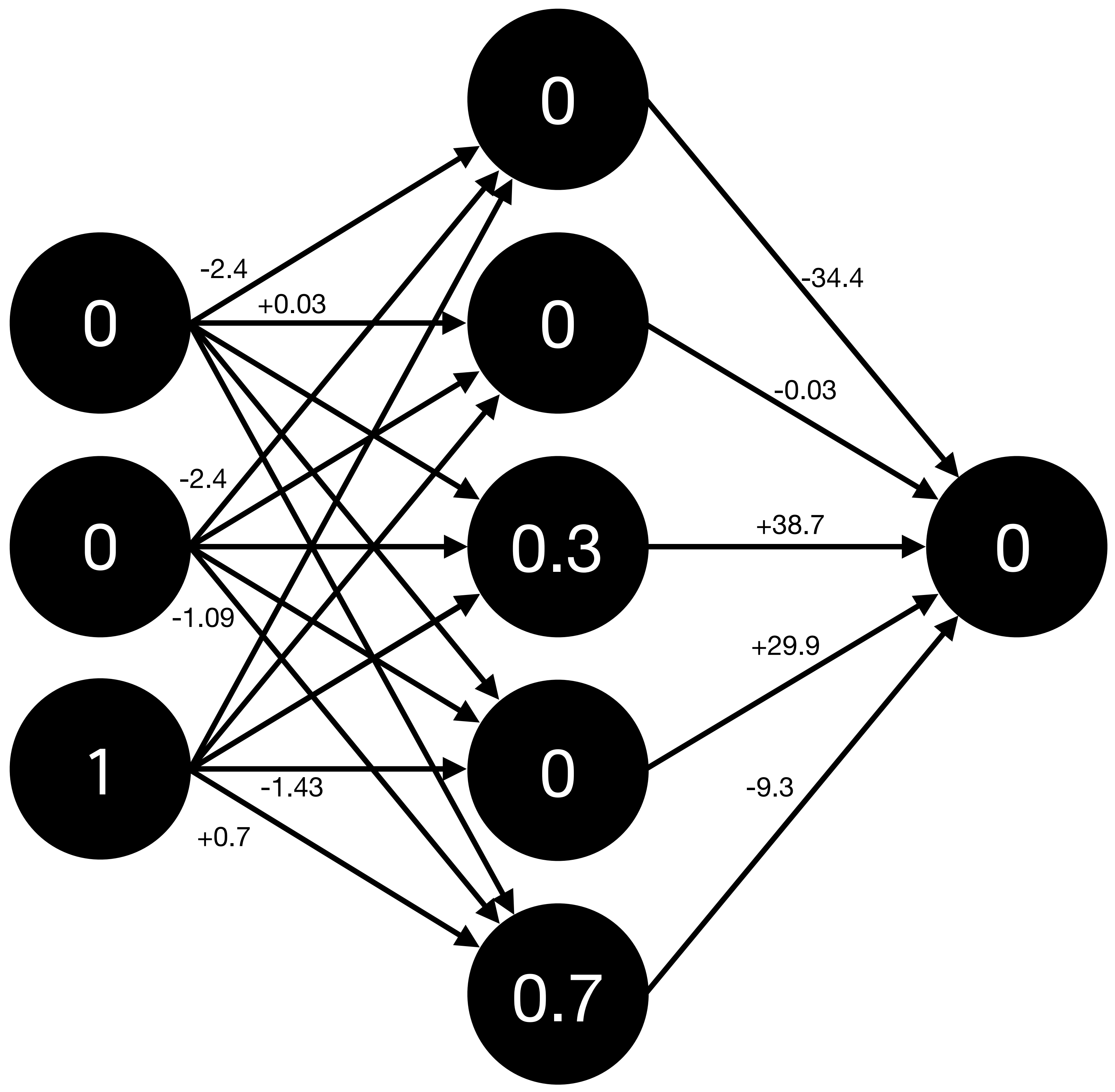
$$y = \underbrace{f}_{\text{Fonction d'activation}}\Big(\overbrace{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3}^{\text{Somme pondérée des entrées}}\Big)$$



Perceptron multicouche

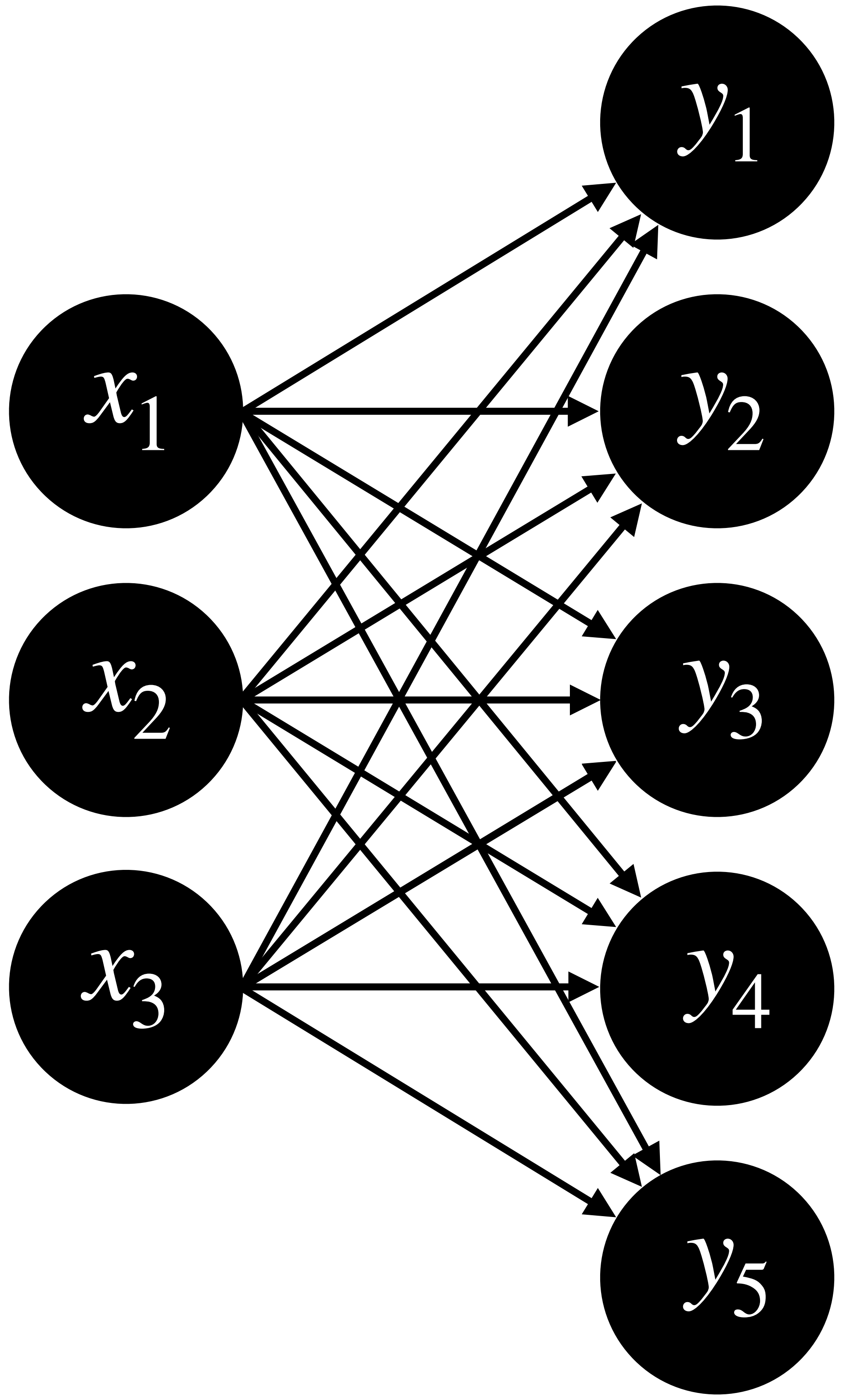


Perceptron multicouche



Détecte s'il y a exactement deux '1' en entrée

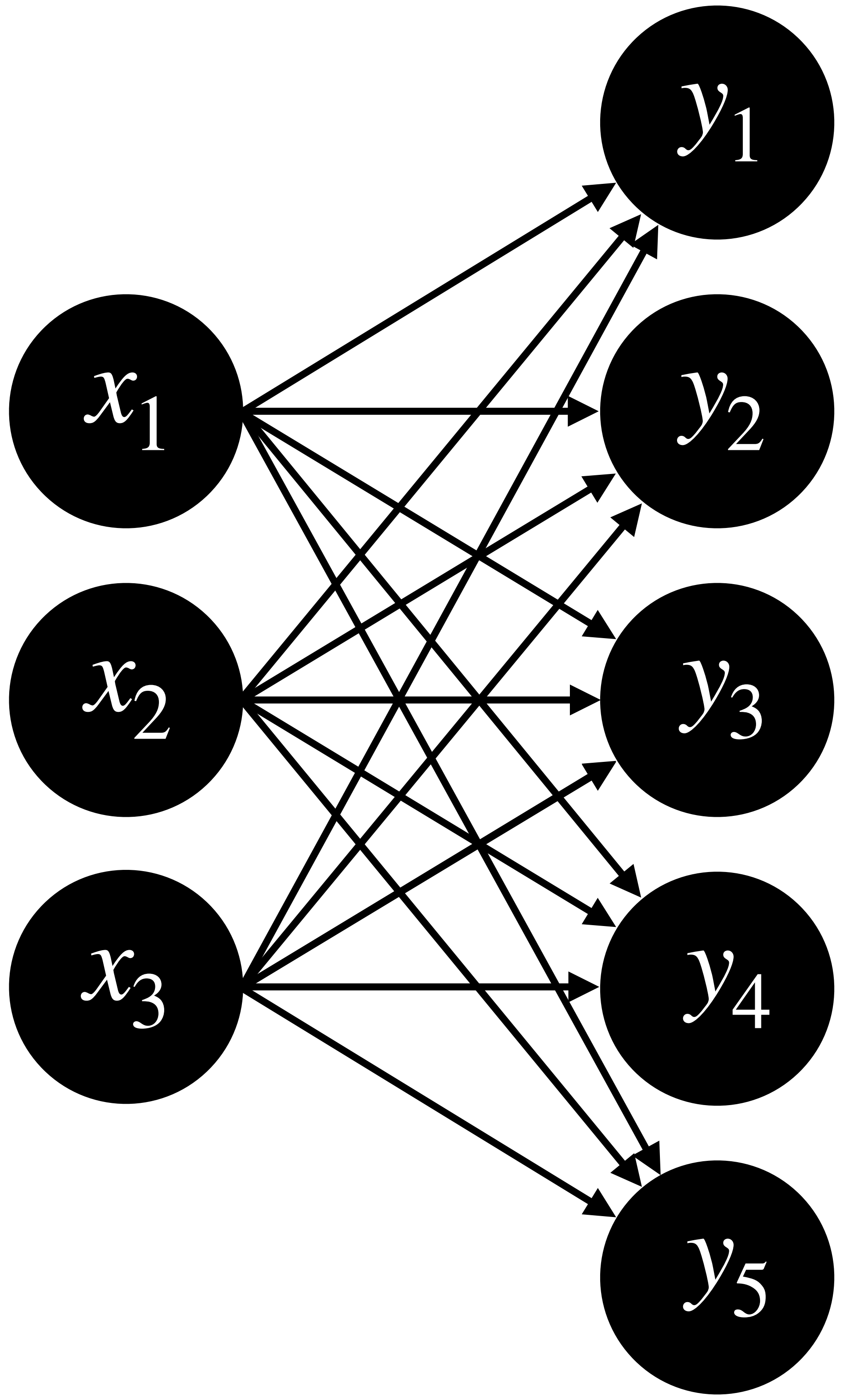
Perceptron multicouche



$$y_j = f\left(\sum_i w_{ij} \cdot x_i \right)$$

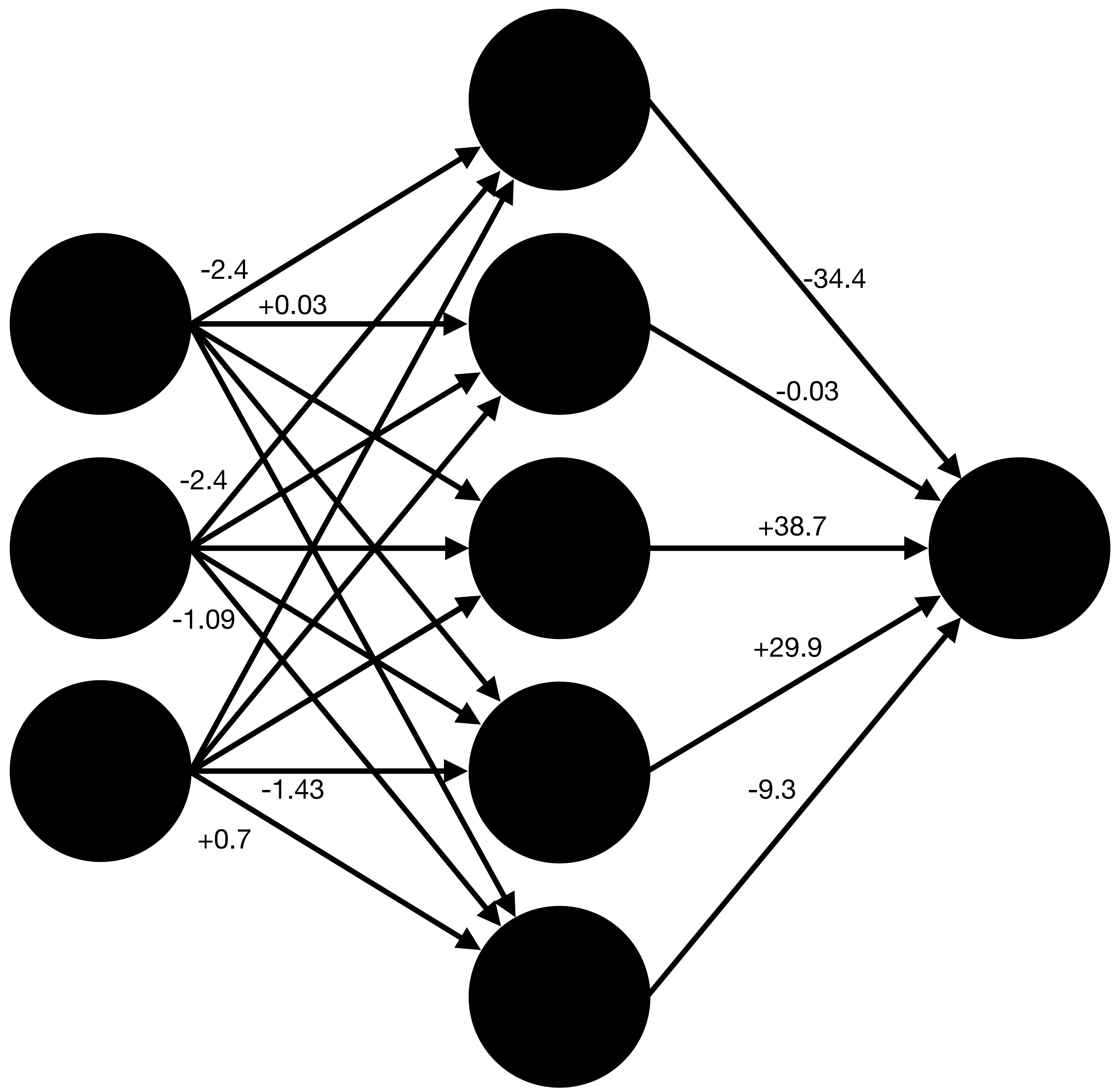
$$\begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix} = f\left(\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \times \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} \end{bmatrix} \right)$$

Perceptron multicouche



$$\mathbf{y} = f(\mathbf{x}W)$$

Perceptron multicouche



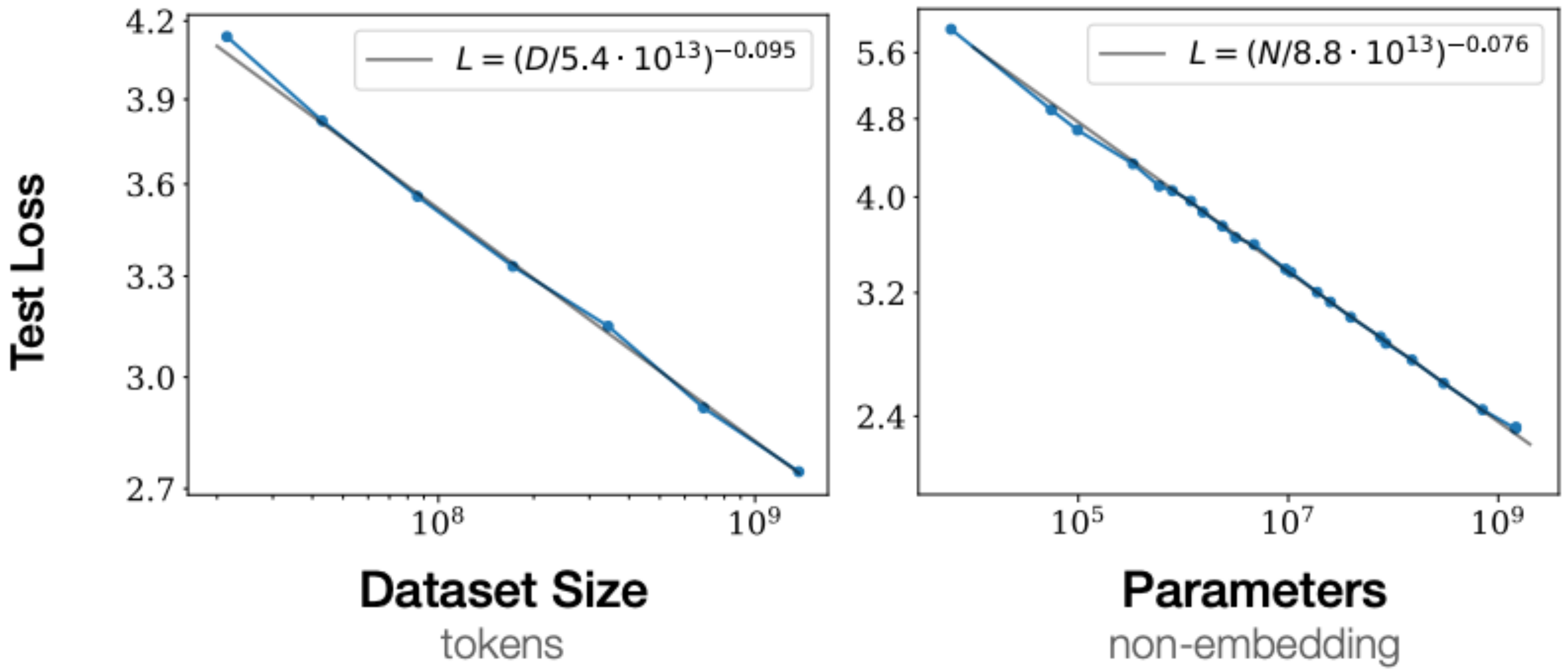
3 x 5 + 5 x 1 = 20 paramètres

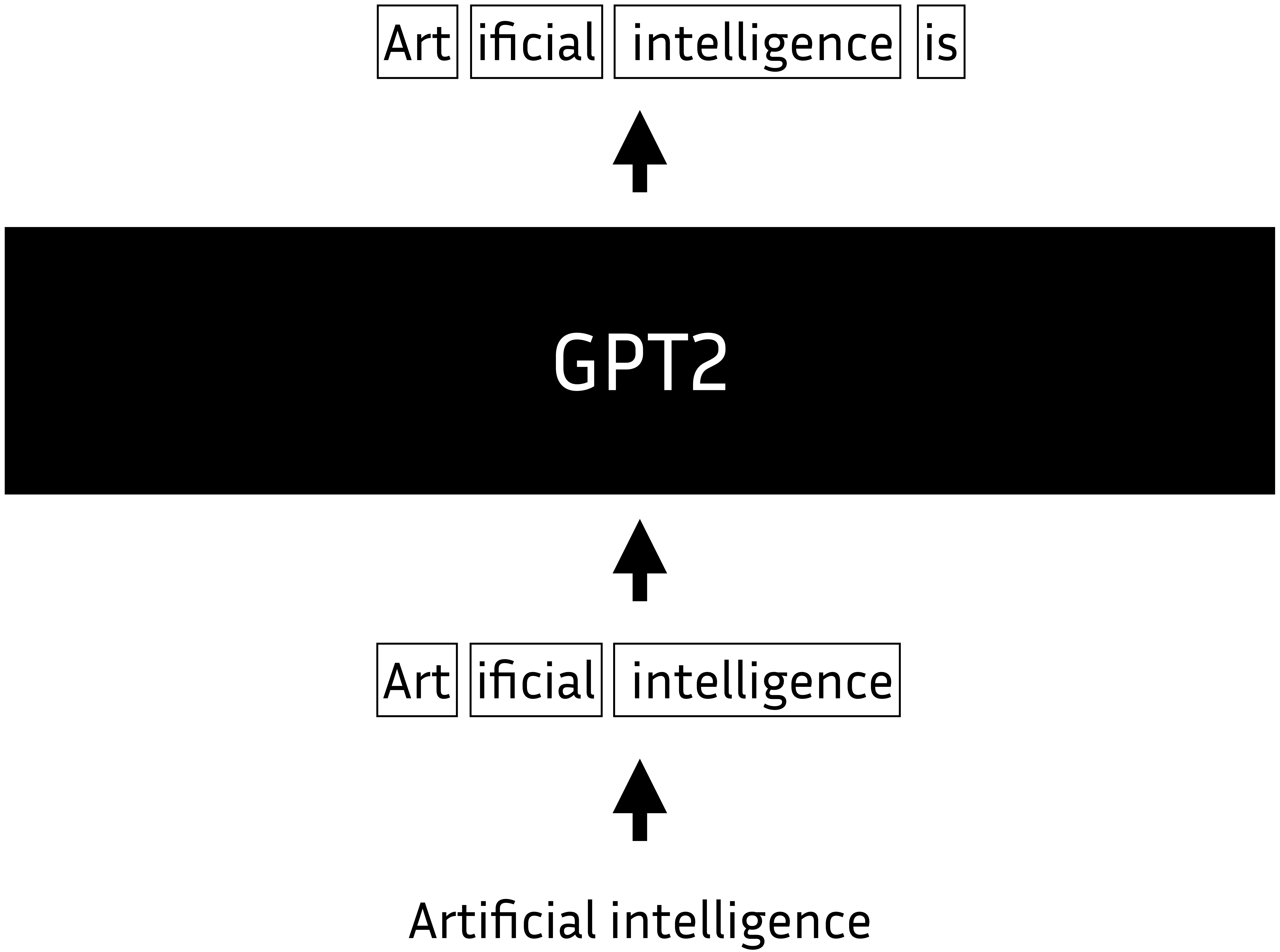
(26 avec les biais)

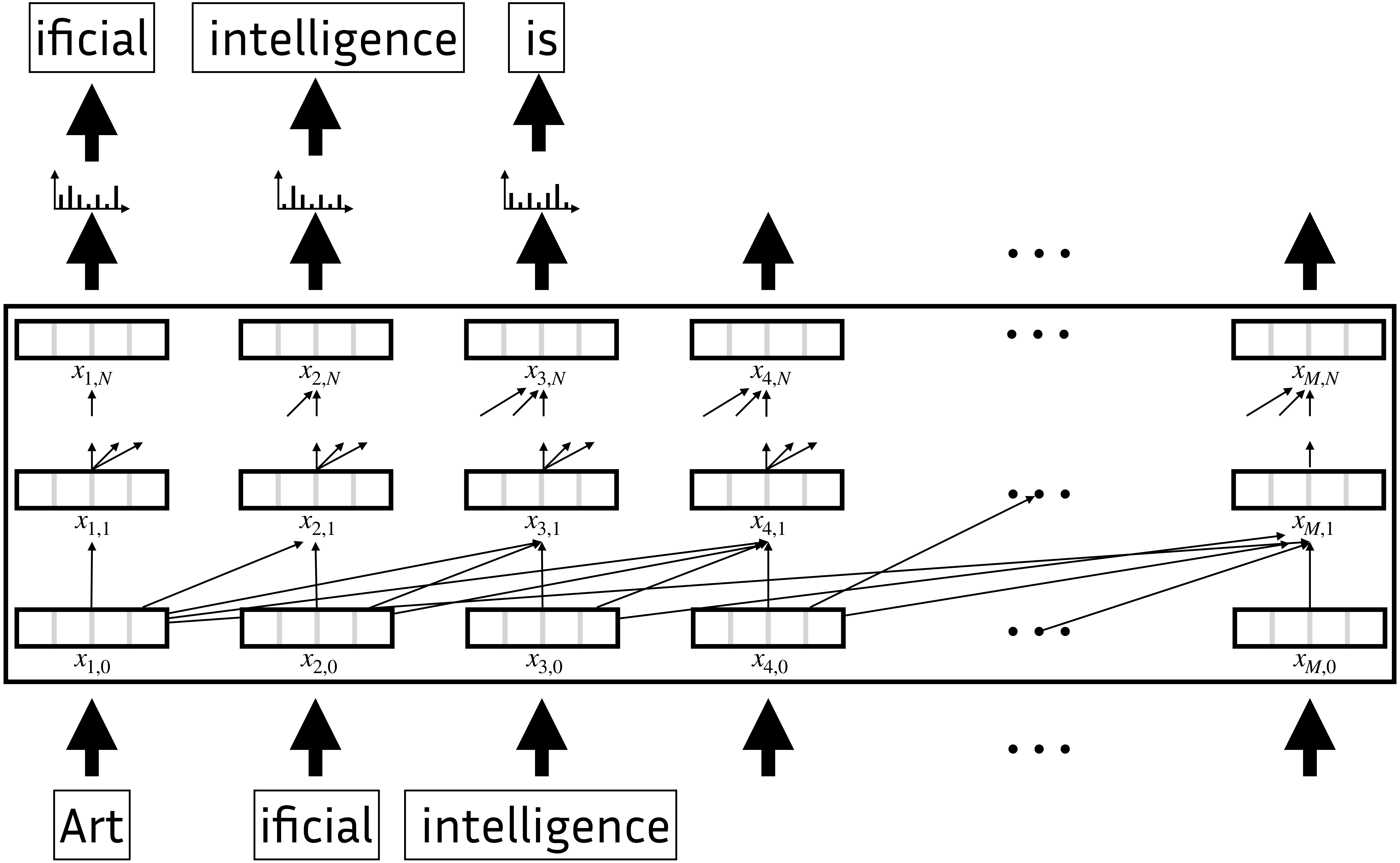
	Nombre de paramètres	Taille des données d'entraiment (en tokens)	Temps d'entrainement (eg. Ordinateur portable)
GPT1	117 millions 1000 livres	600 millions 6000 livres	13 ans
GPT2	1.5 milliard 13 000 livres	28 milliards 280 000 livres	1600 ans
GPT2 - small	124 millions 1000 livres	28 milliards 280 000 livres	
GPT3	175 miliards 1.5 millions de livres	300 milliards 3 millions de livres	99900 ans
GPT4	1800 milliards ? 15 millions de livres	13 000 milliards 130 millions de livres	7 millions d'années centaines de millions de dollars
PALM	540 milliards 5 millions de livres	780 milliards 7.8 millions de livres	800 000 ans
Gemini	?	?	?
Claude	130 milliards 1 millions de livres	assez peu	?
Mistral	45 milliards 400 000 de livres	?	?
Llama2	70 milliards 620 000 livres	2000 milliards 20 millions de livres	250 000 ans

Scaling Laws for Neural Language Models

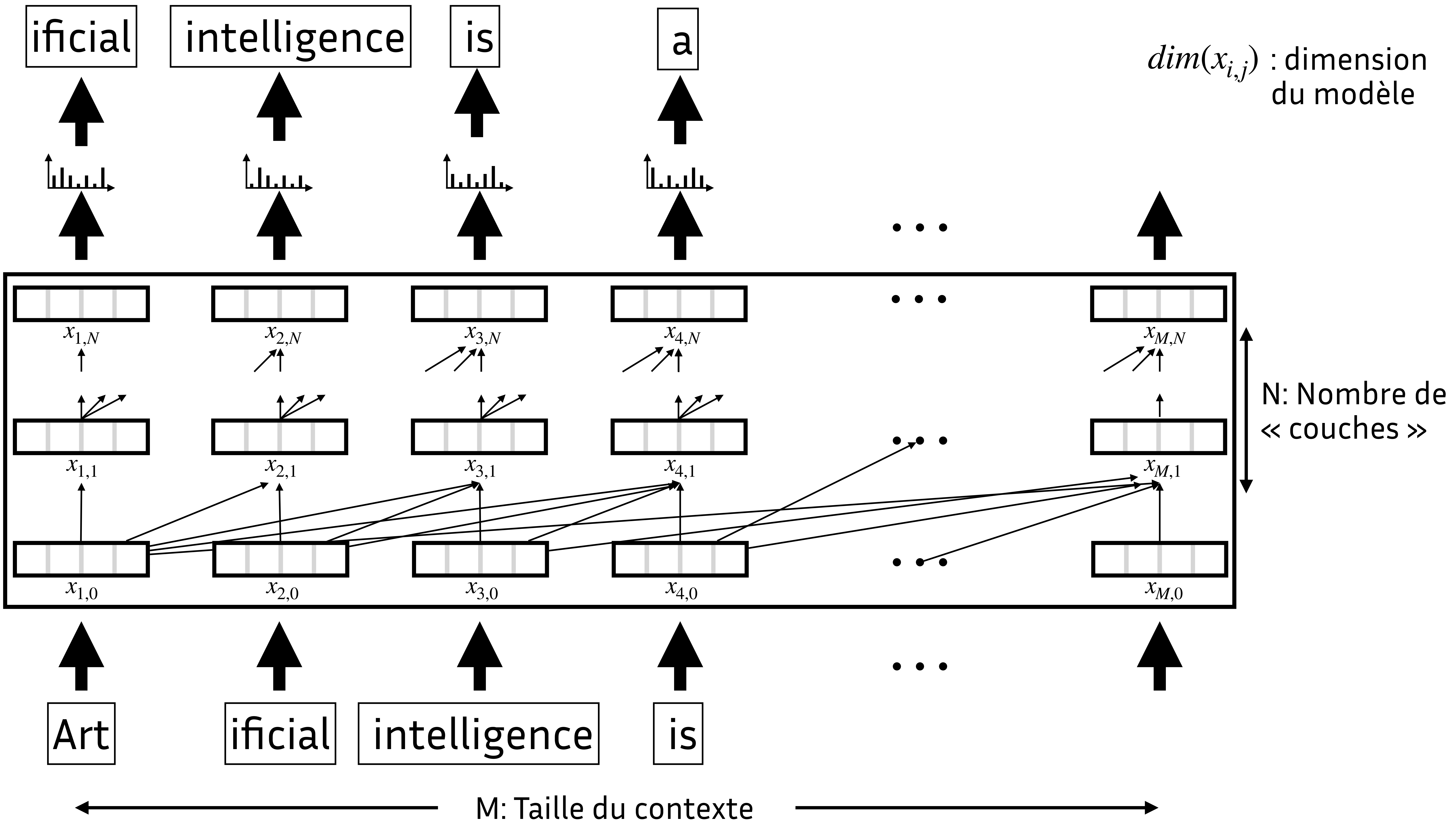
Kaplan et al., 2020



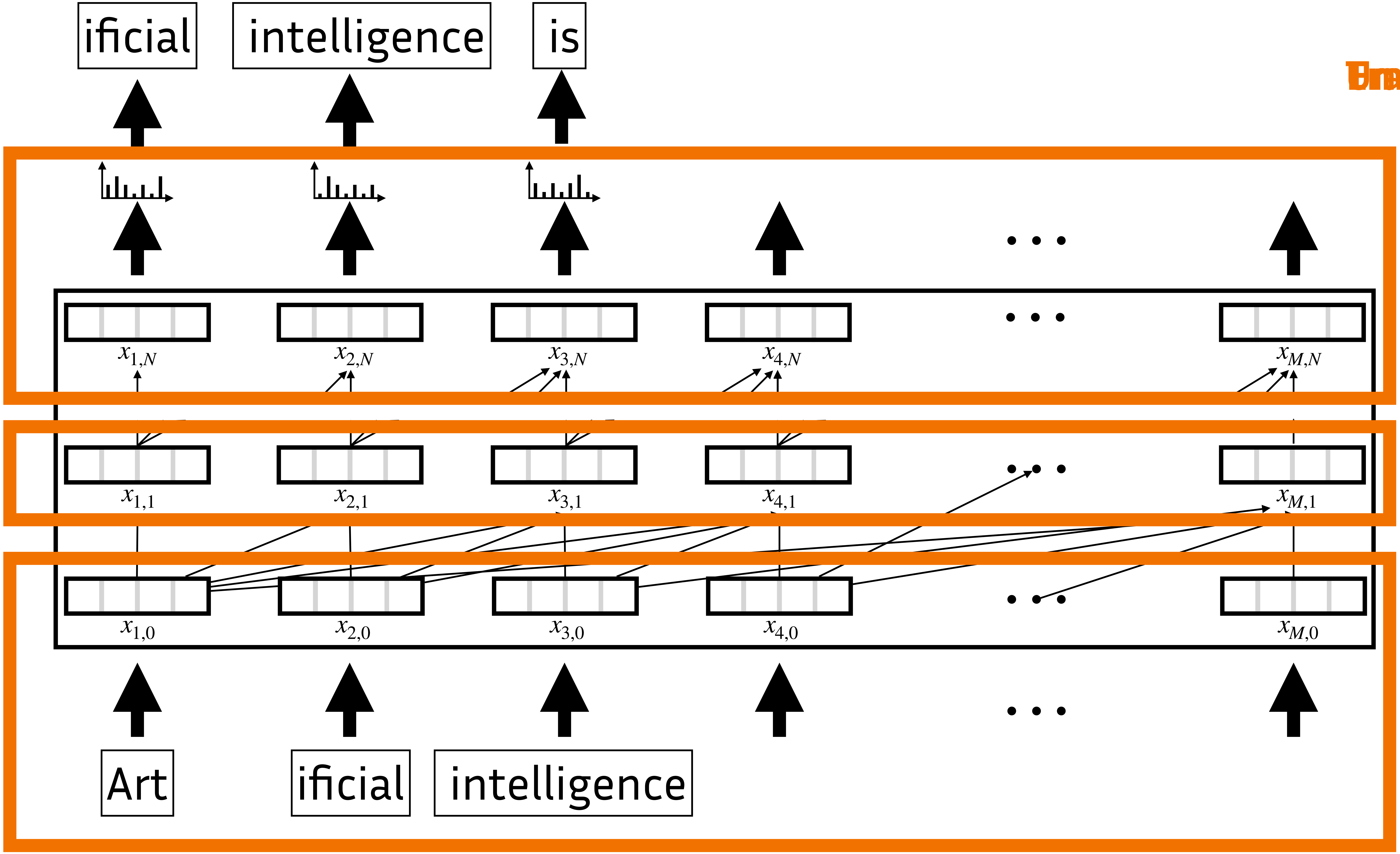

Art ificial intelligence is
GPT2
Art ificial intelligence
Artificial intelligence



ificial
intelligence
is
$x_{1,N}$
$x_{2,N}$
$x_{3,N}$
$x_{4,N}$
$x_{M,N}$
$x_{1,1}$
$x_{2,1}$
$x_{3,1}$
$x_{4,1}$
$x_{M,1}$
$x_{1,0}$
$x_{2,0}$
$x_{3,0}$
$x_{4,0}$
$x_{M,0}$
Art
ificial
intelligence

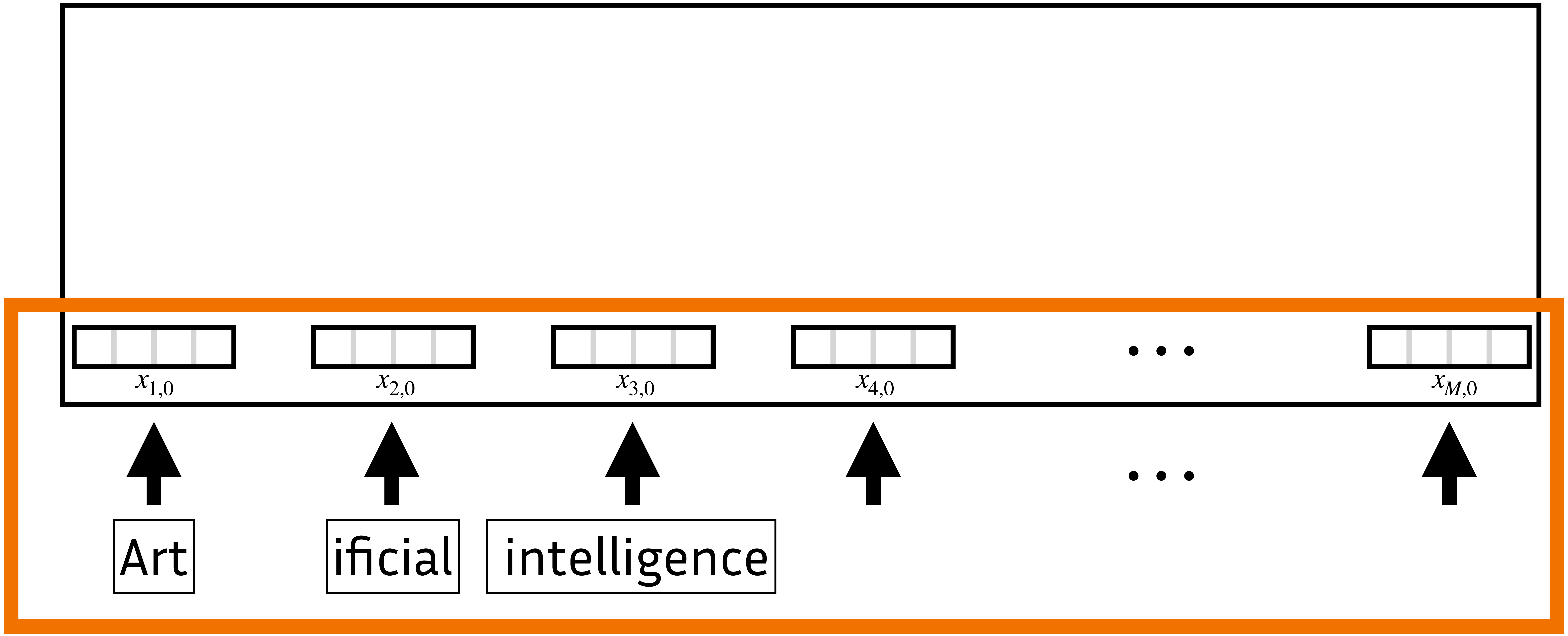


ificial
intelligence
is
a
$dim(x_{i,j})$: dimension du modèle
$x_{1,N}$
$x_{2,N}$
$x_{3,N}$
$x_{4,N}$
$x_{M,N}$
N: Nombre de « couches »
$x_{1,1}$
$x_{2,1}$
$x_{3,1}$
$x_{4,1}$
$x_{M,1}$
$x_{1,0}$
$x_{2,0}$
$x_{3,0}$
$x_{4,0}$
$x_{M,0}$
Art
ificial
intelligence
is
M: Taille du contexte



ificial
intelligence
is
$x_{1,N}$
$x_{2,N}$
$x_{3,N}$
$x_{4,N}$
$x_{M,N}$
$x_{1,1}$
$x_{2,1}$
$x_{3,1}$
$x_{4,1}$
$x_{M,1}$
$x_{1,0}$
$x_{2,0}$
$x_{3,0}$
$x_{4,0}$
$x_{M,0}$
Art
ificial
intelligence

Embedding

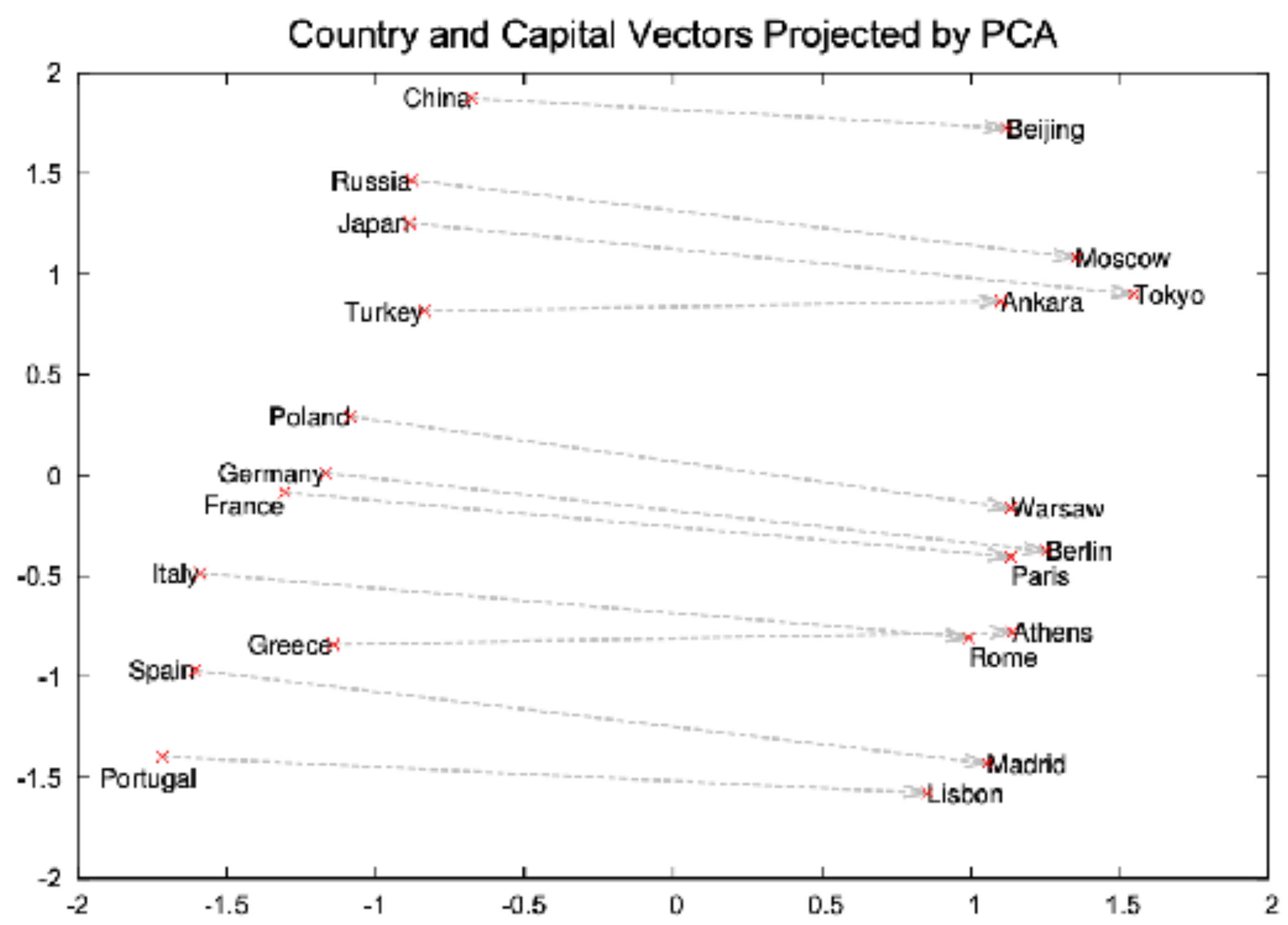


Embedding

1 mot ⟷ 1 vecteur

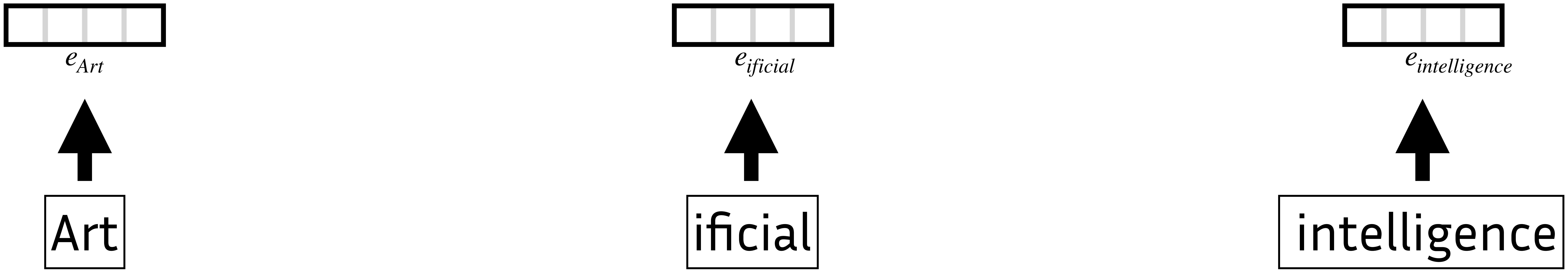
espace sémantique ⟷ espace vectoriel

relations sémantiques ⟷ relations géométriques

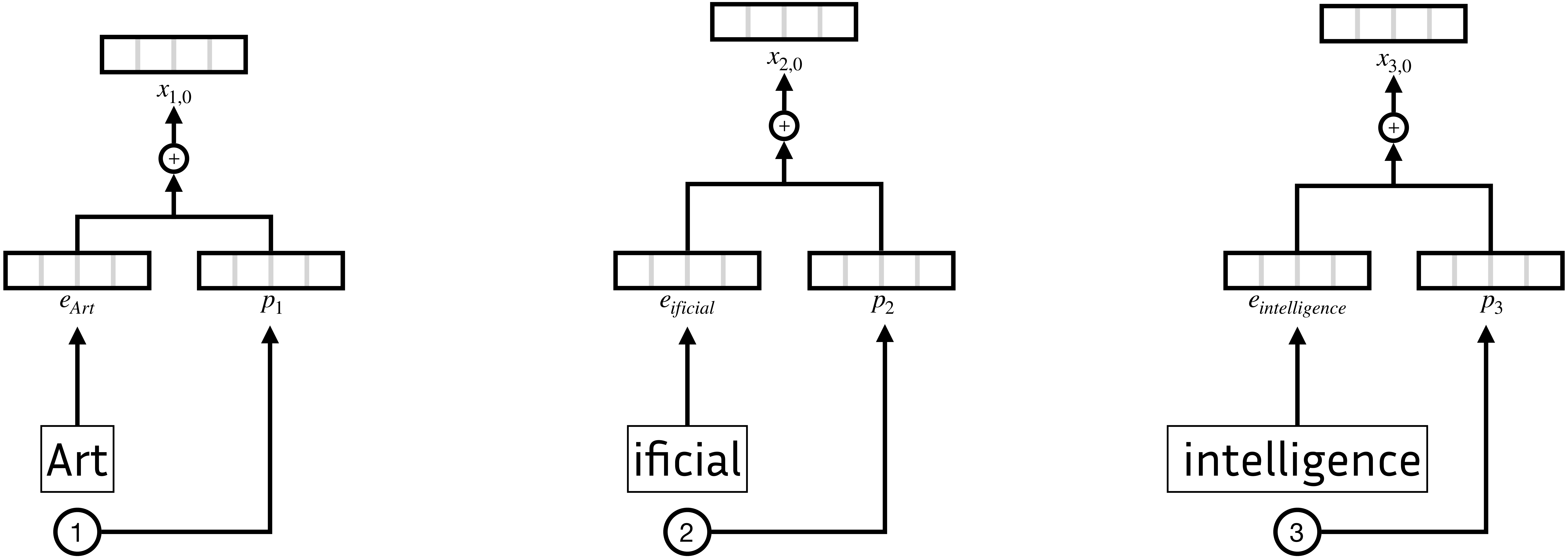


Mikolov et al., 2013

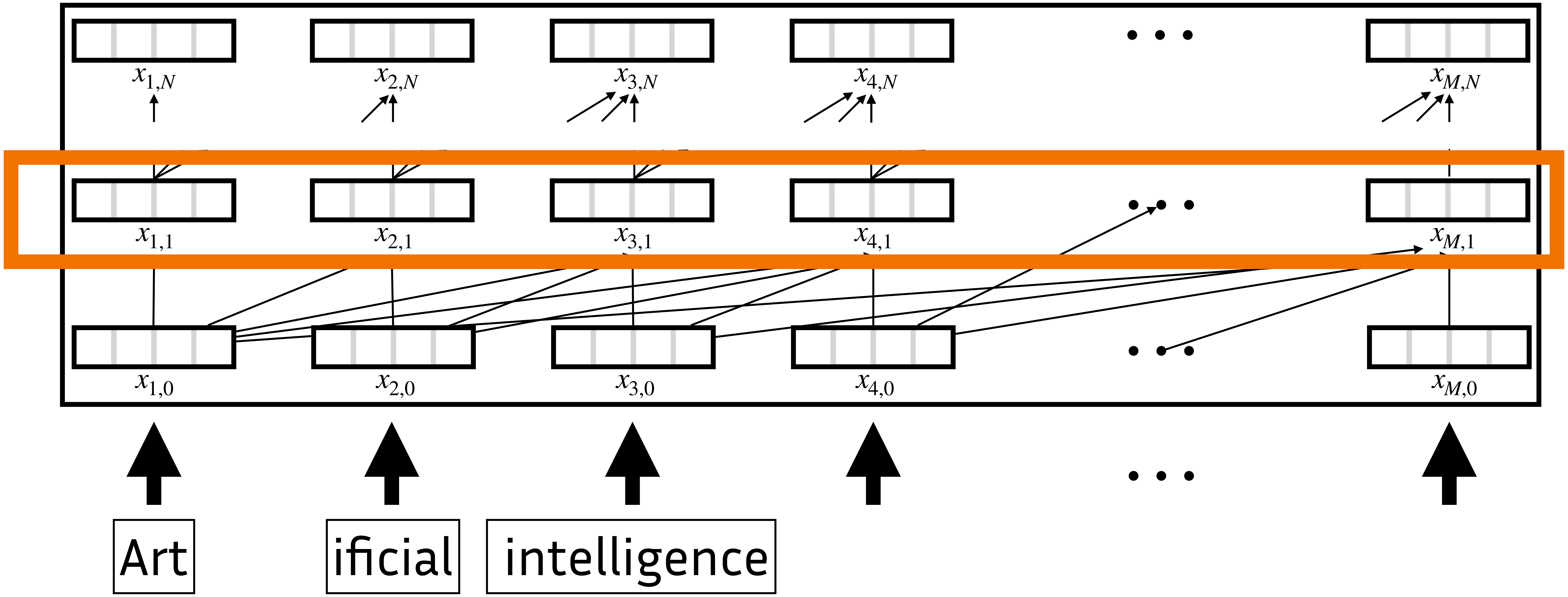
Embedding

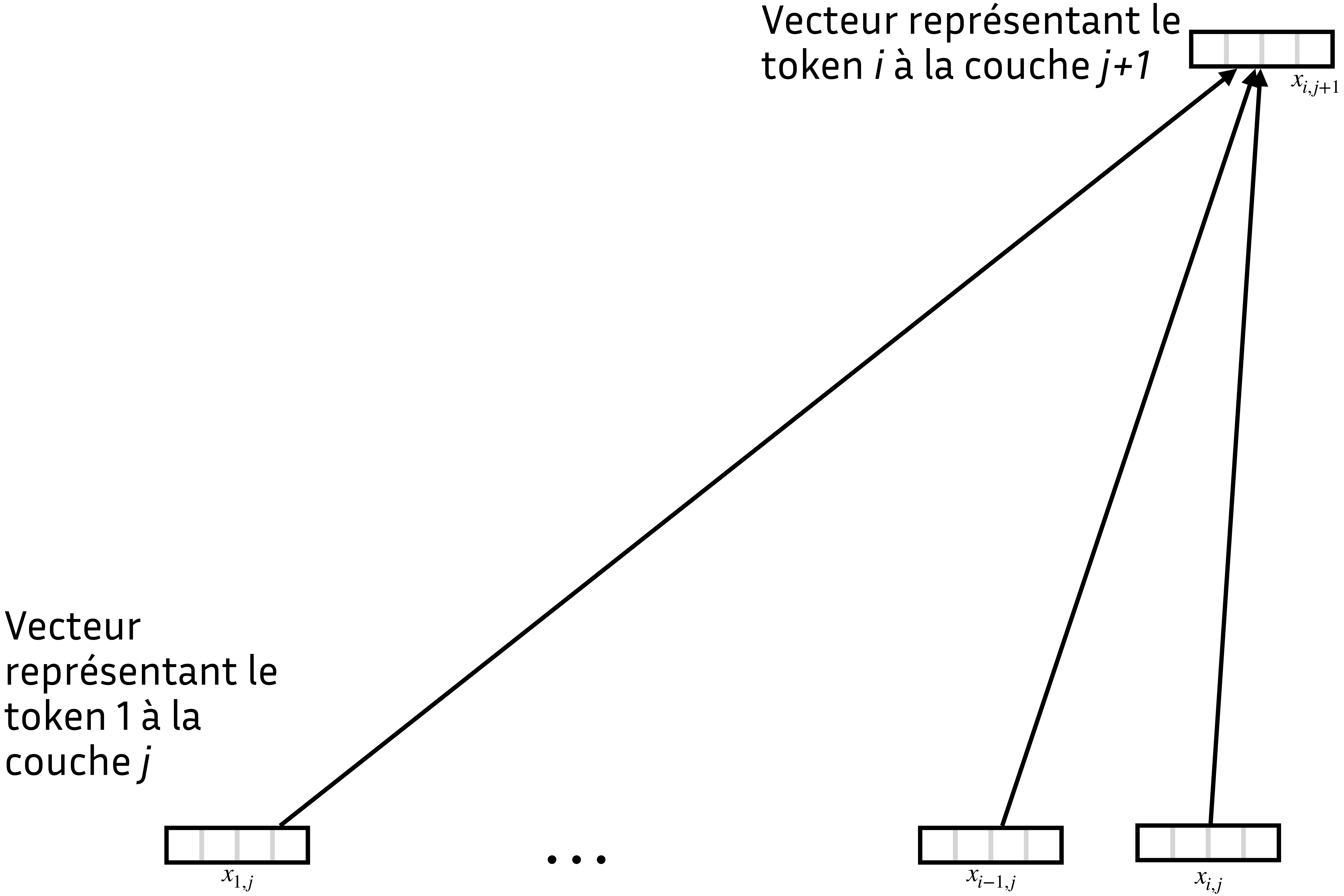


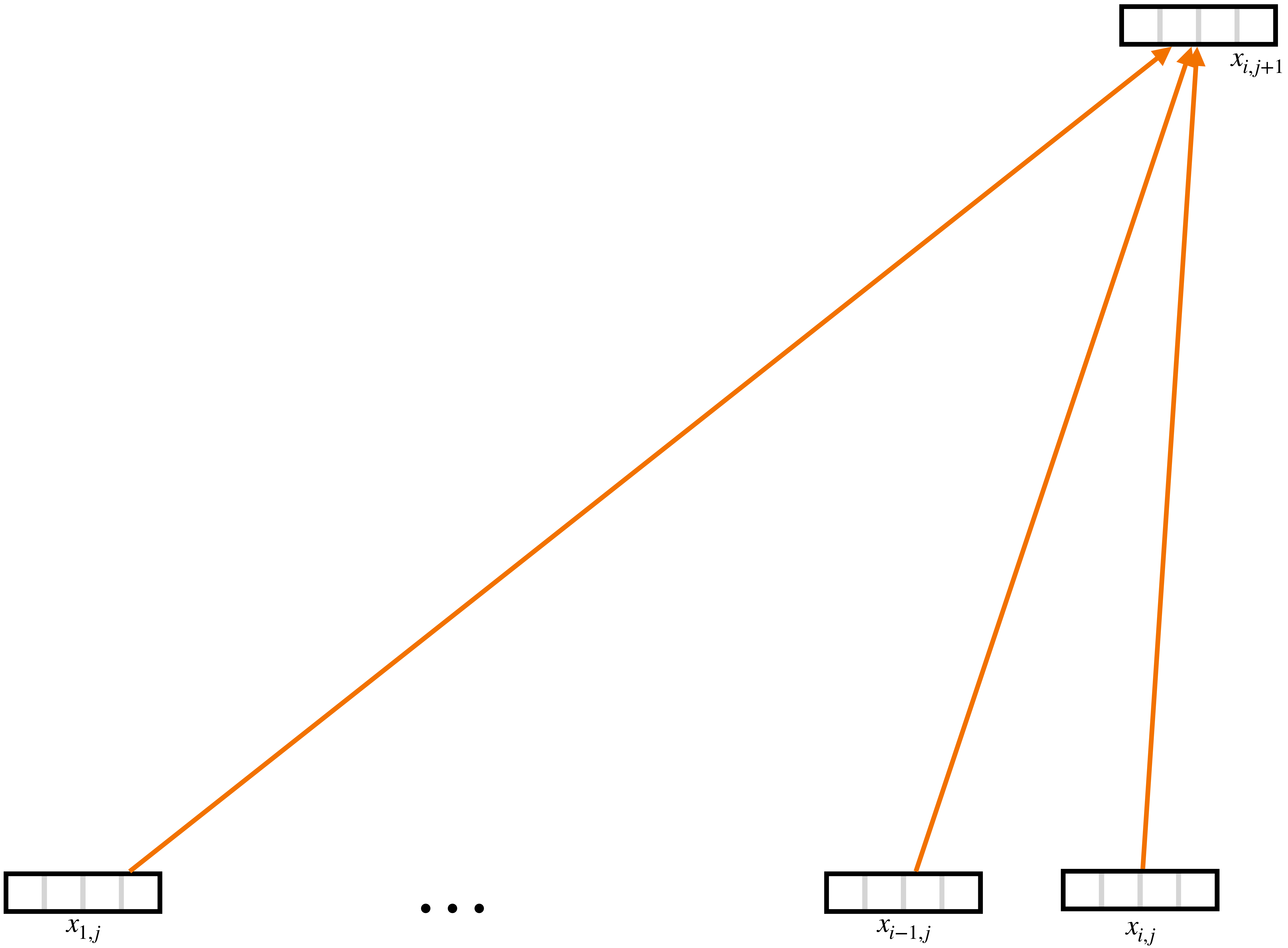
Position encoding



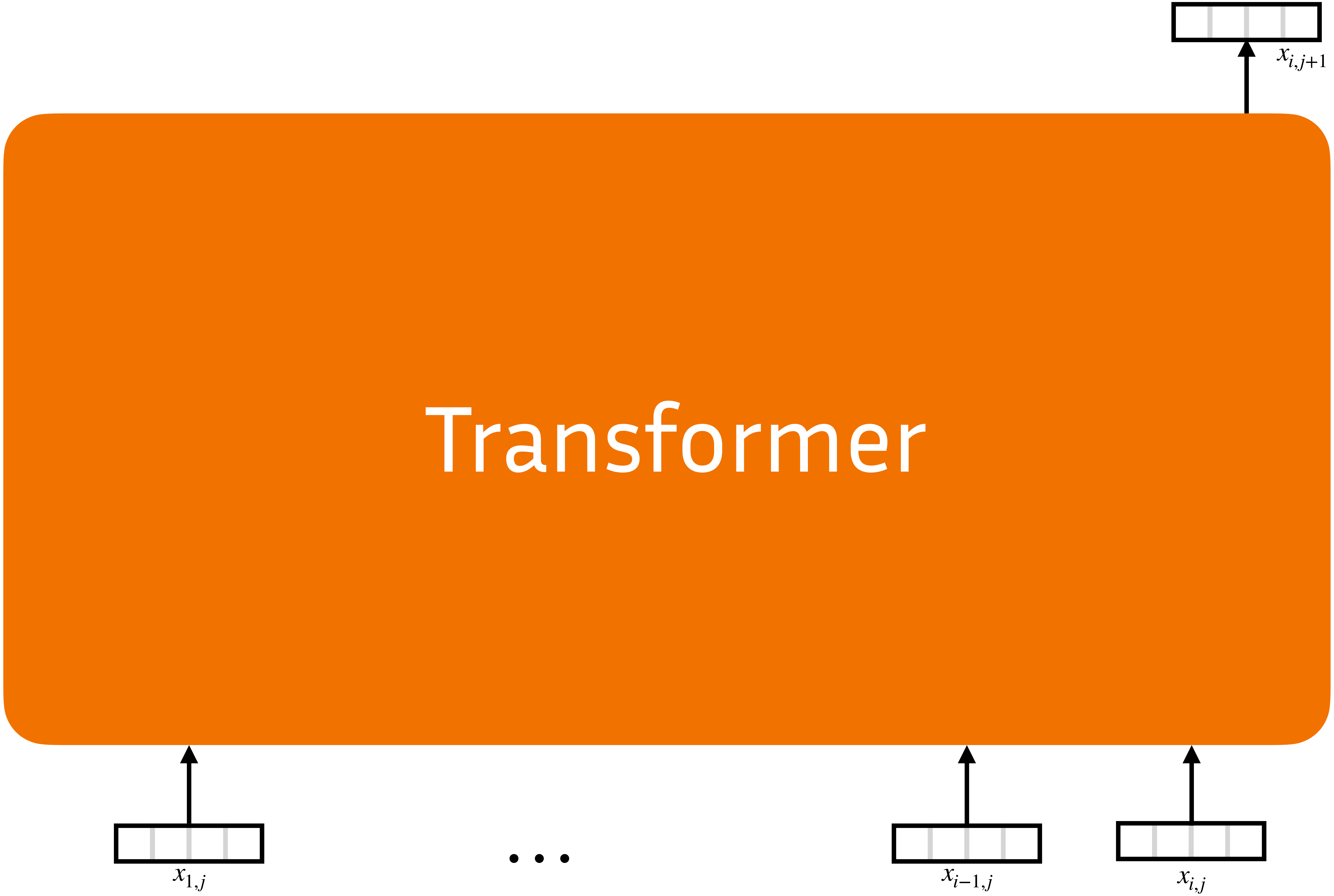
Transformer

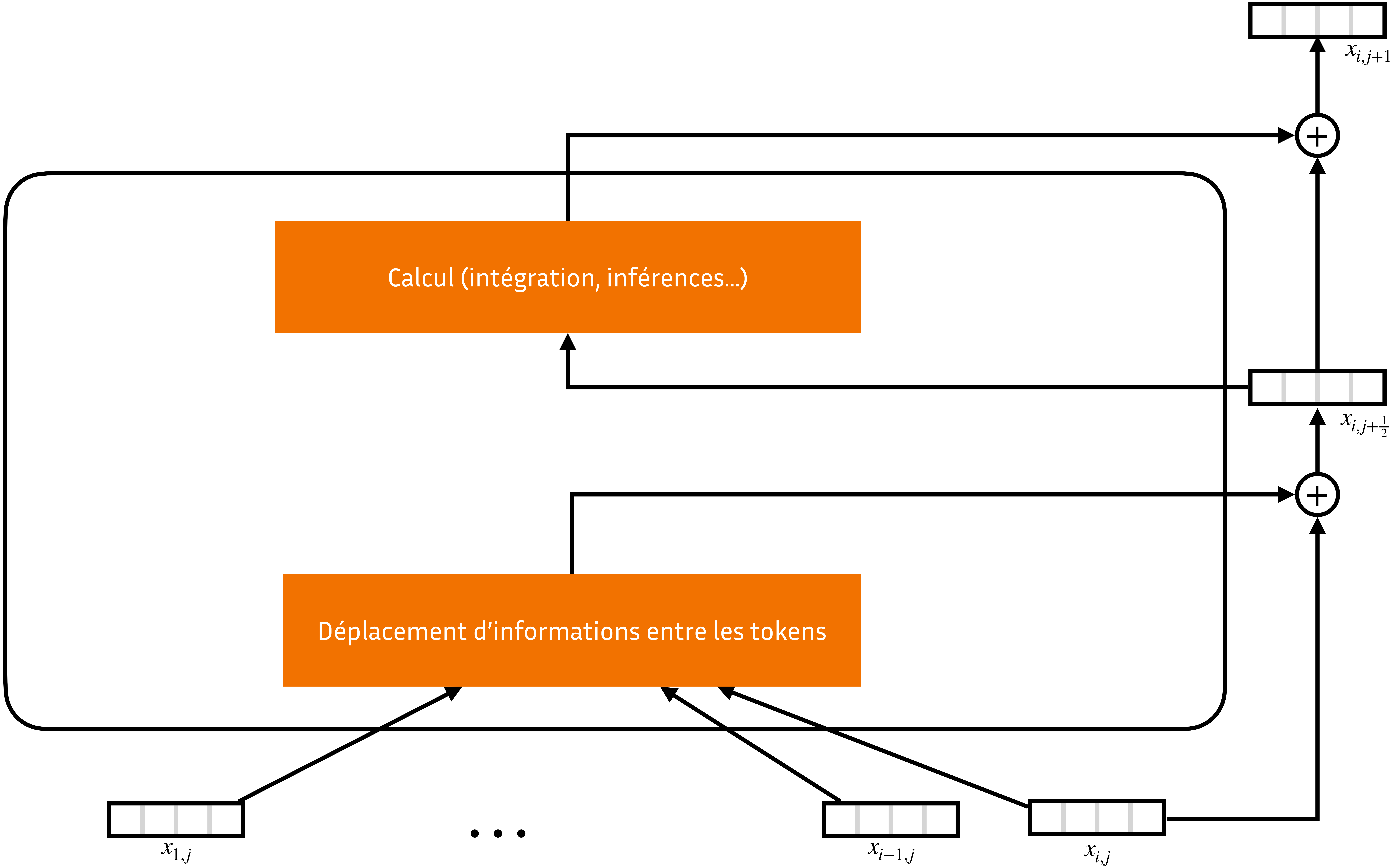



Vecteur représentant le token *i* à la couche *j+1*
$x_{i,j+1}$
Vecteur représentant le token 1 à la couche *j*
$x_{1,j}$
...
$x_{i-1,j}$
$x_{i,j}$

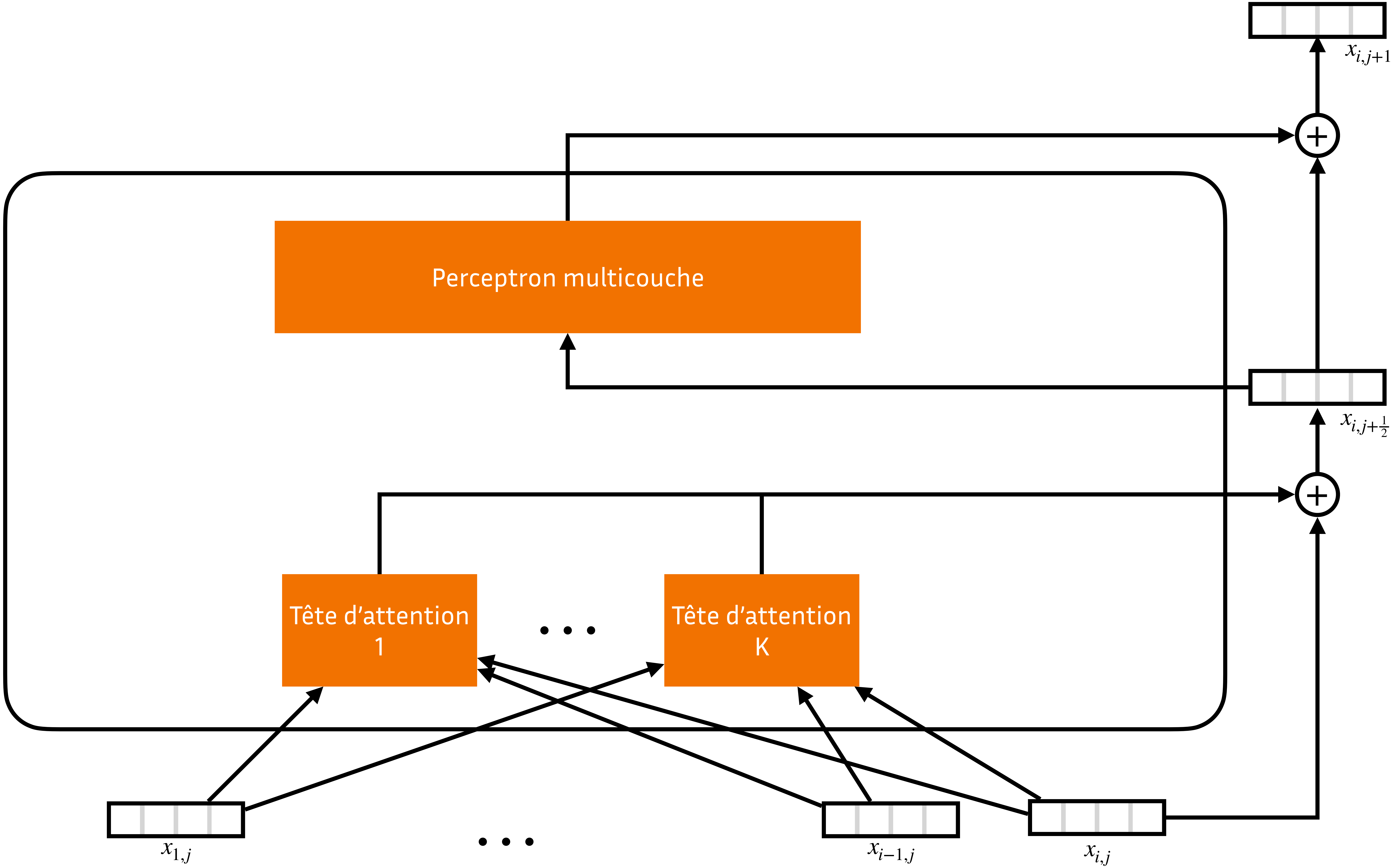


$x_{i,j+1}$
$x_{1,j}$
$\cdots$
$x_{i-1,j}$
$x_{i,j}$

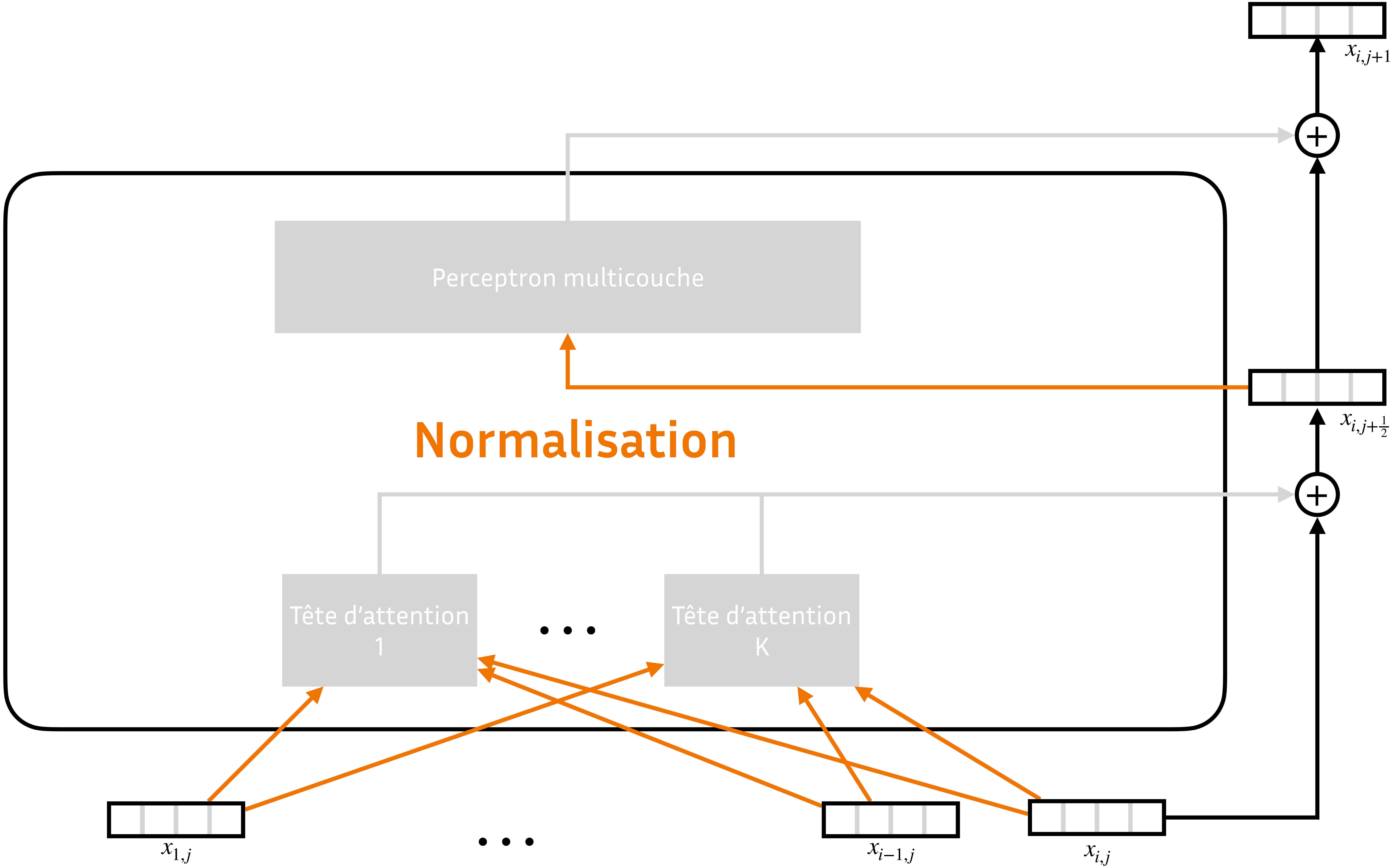

$x_{i,j+1}$
Transformer
$x_{1,j}$
...
$x_{i-1,j}$
$x_{i,j}$



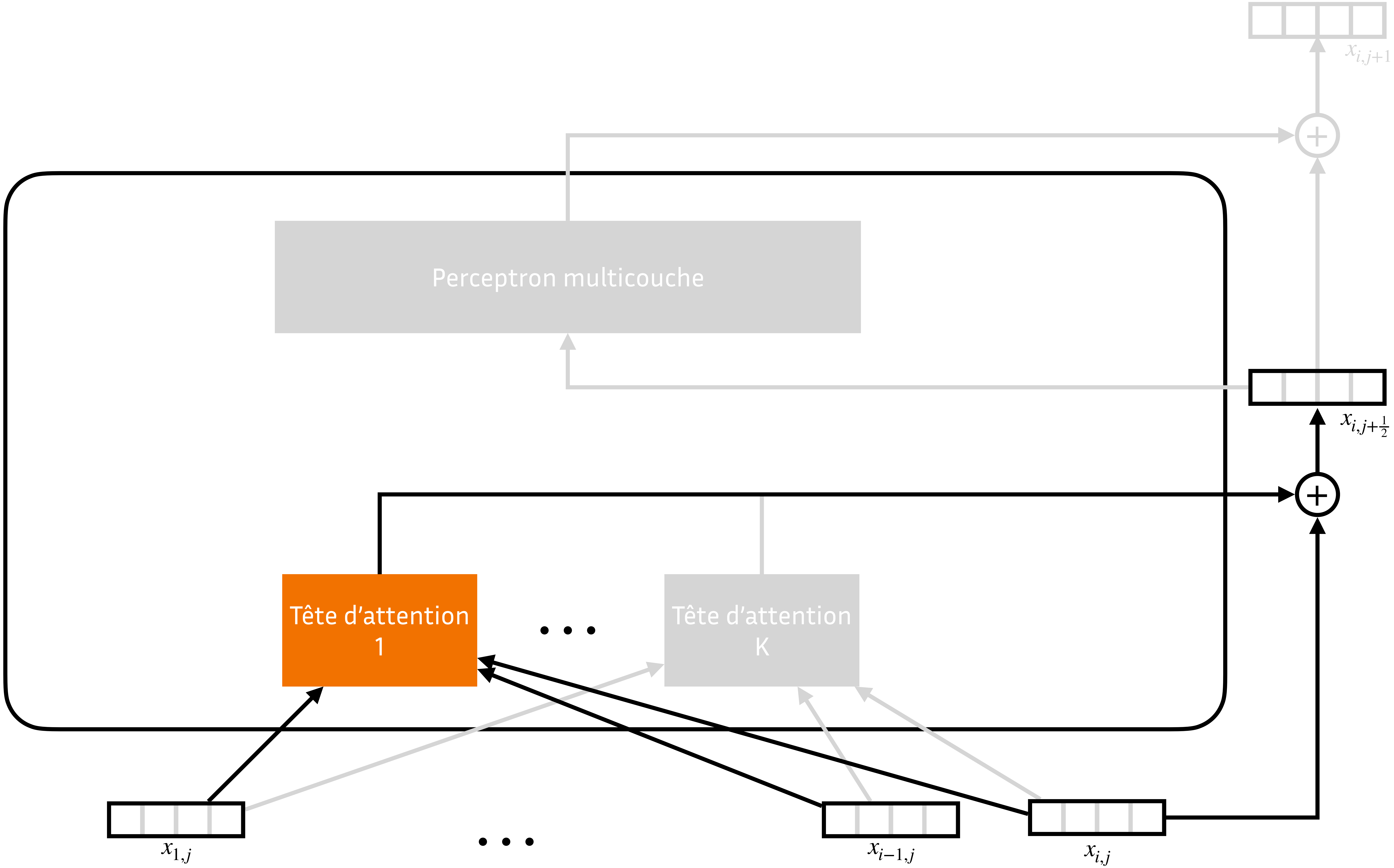
$x_{i,j+1}$
Calcul (intégration, inférences...)
$x_{i,j+\frac{1}{2}}$
Déplacement d'informations entre les tokens
$x_{1,j}$
...
$x_{i-1,j}$
$x_{i,j}$



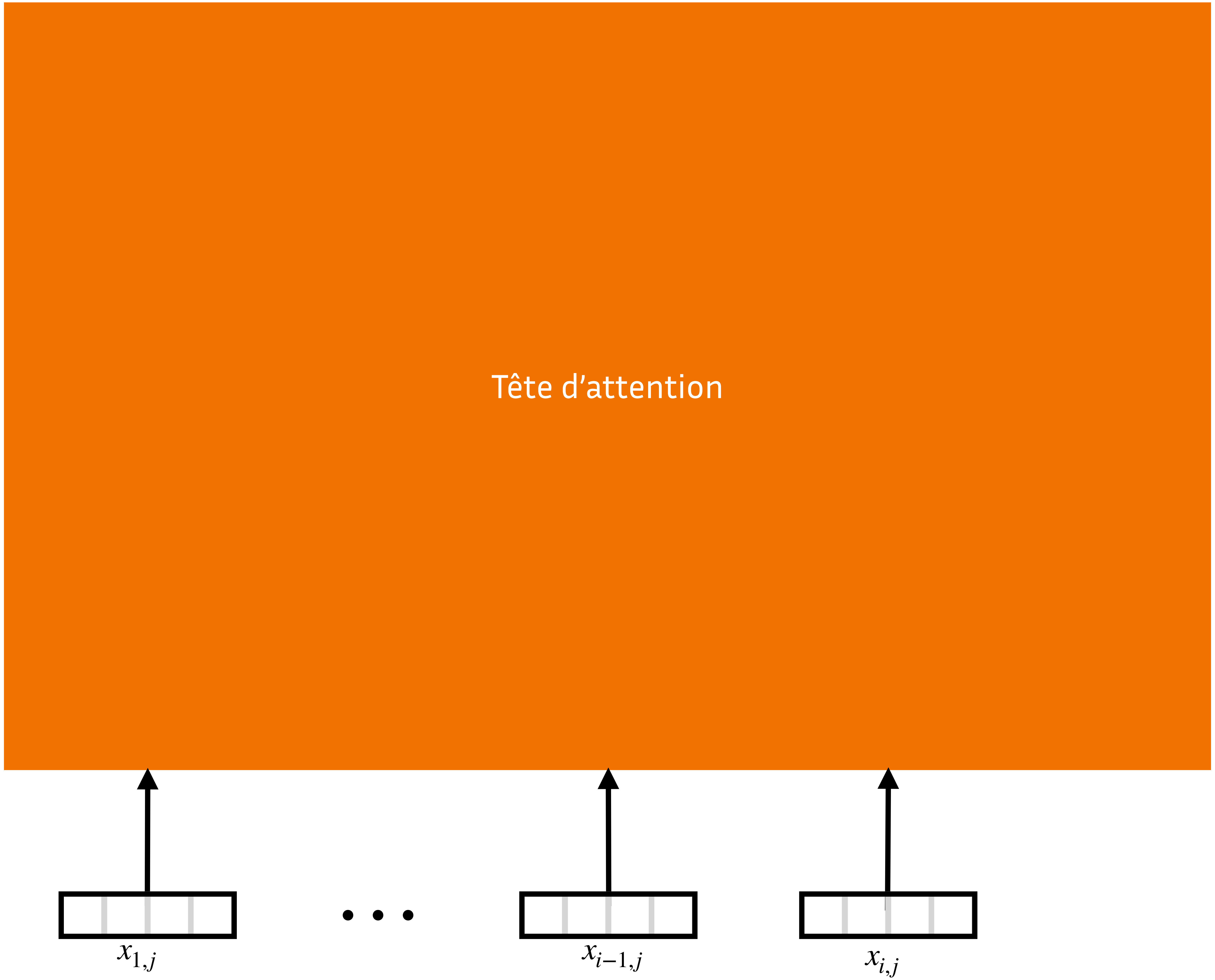
$x_{i,j+1}$
Perceptron multicouche
$x_{i,j+\frac{1}{2}}$
Tête d'attention 1
Tête d'attention K
$x_{1,j}$
$x_{i-1,j}$
$x_{i,j}$



$x_{i,j+1}$
Perceptron multicouche
Normalisation
$x_{i,j+\frac{1}{2}}$
Tête d'attention
1
Tête d'attention
K
$x_{1,j}$
$x_{i-1,j}$
$x_{i,j}$



$x_{i,j+1}$
Perceptron multicouche
$x_{i,j+\frac{1}{2}}$
Tête d'attention 1
Tête d'attention K
$x_{1,j}$
$x_{i-1,j}$
$x_{i,j}$

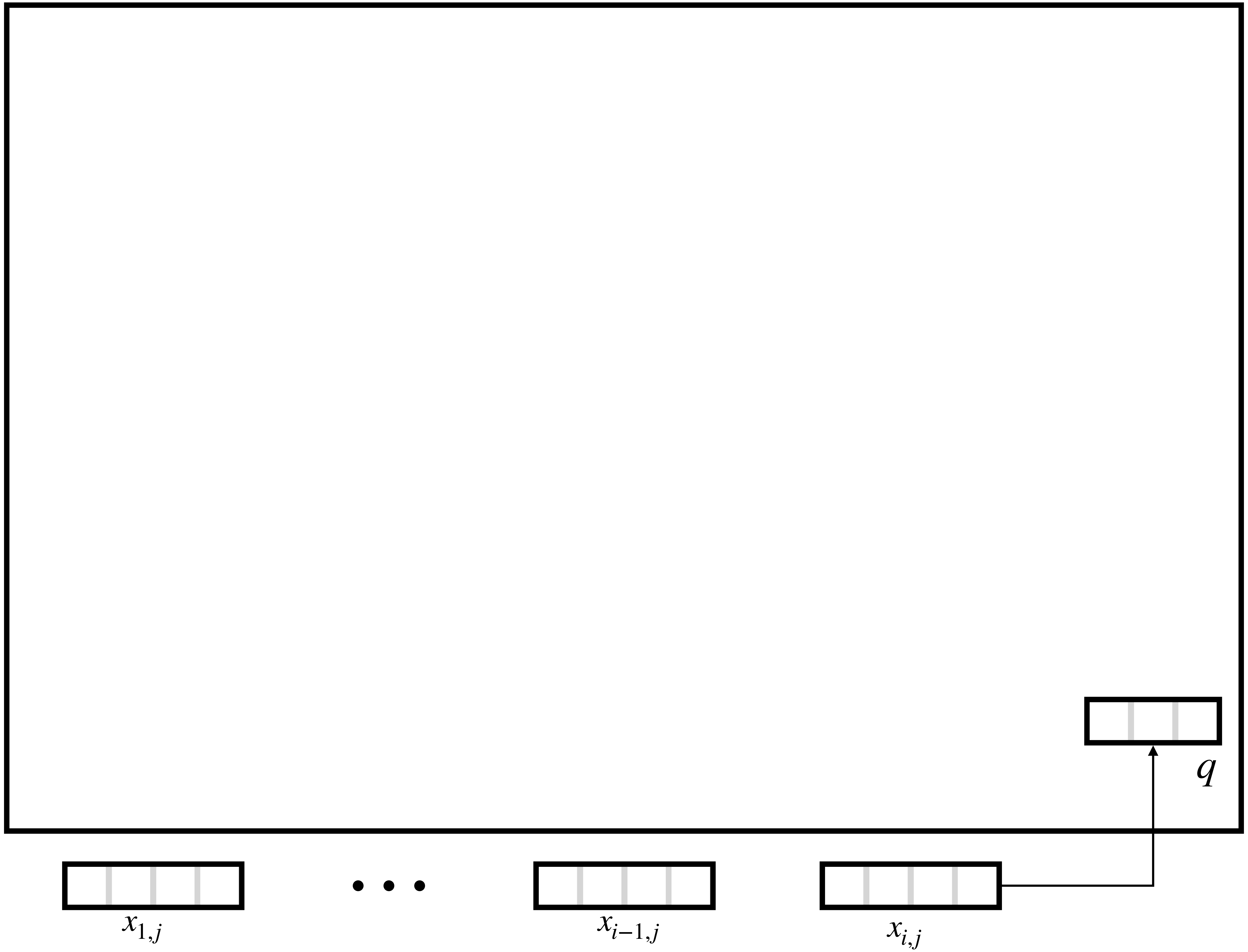


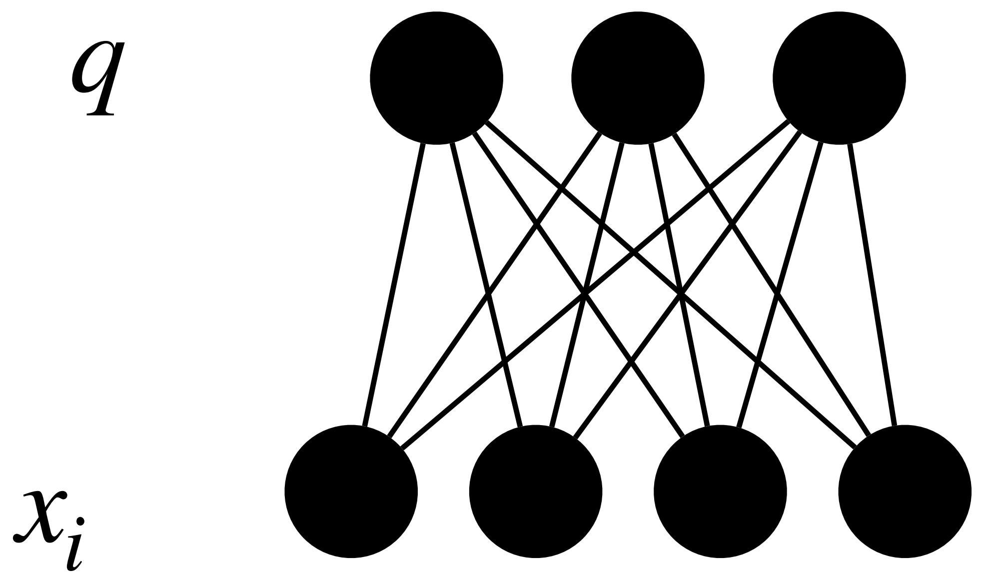
1 : identifier quel type d’information est recherché

2 : identifier quels tokens disposent de cette information

3 : récupérer cette information dans ces tokens

1 : identifier quel type d'information est recherché : *query*

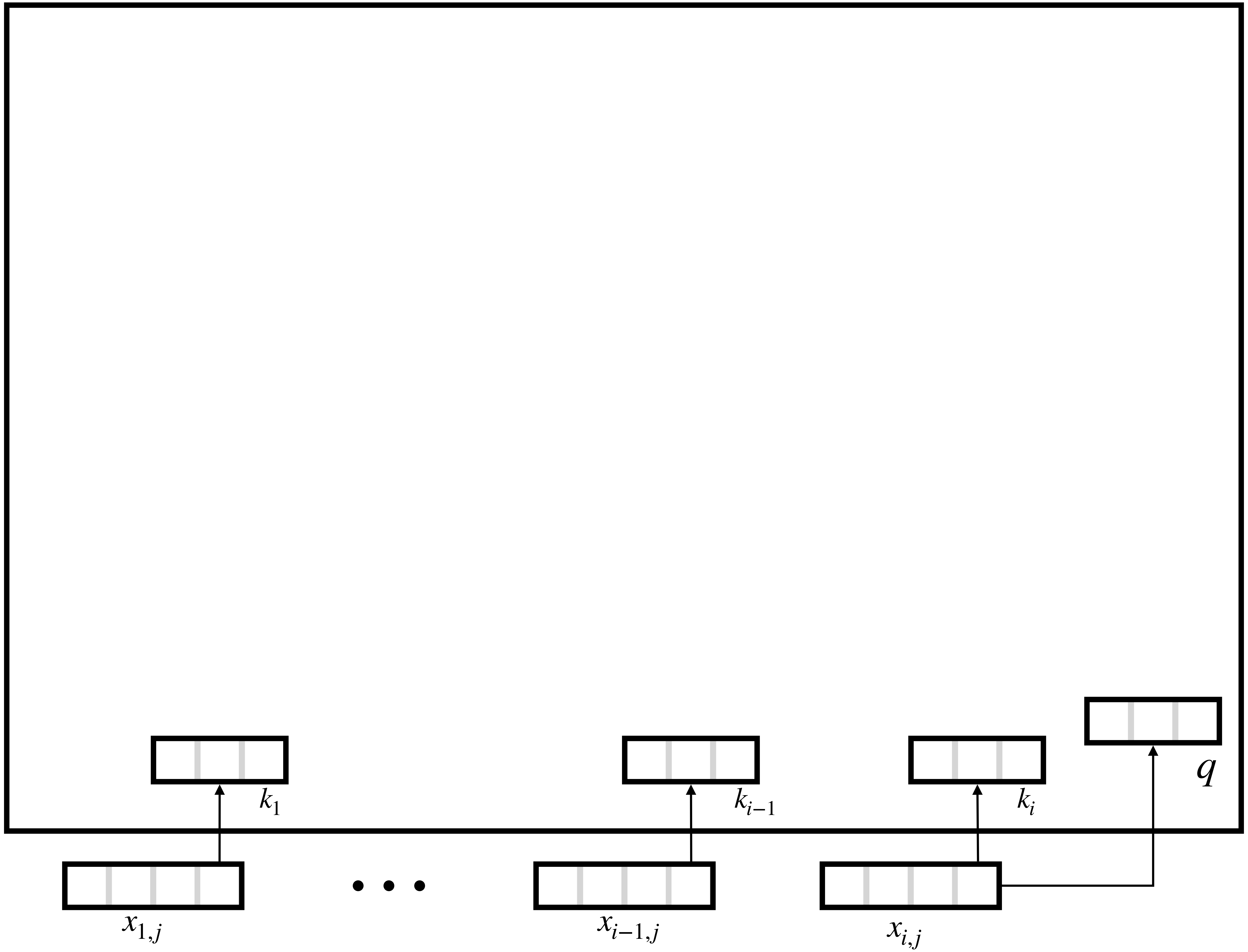


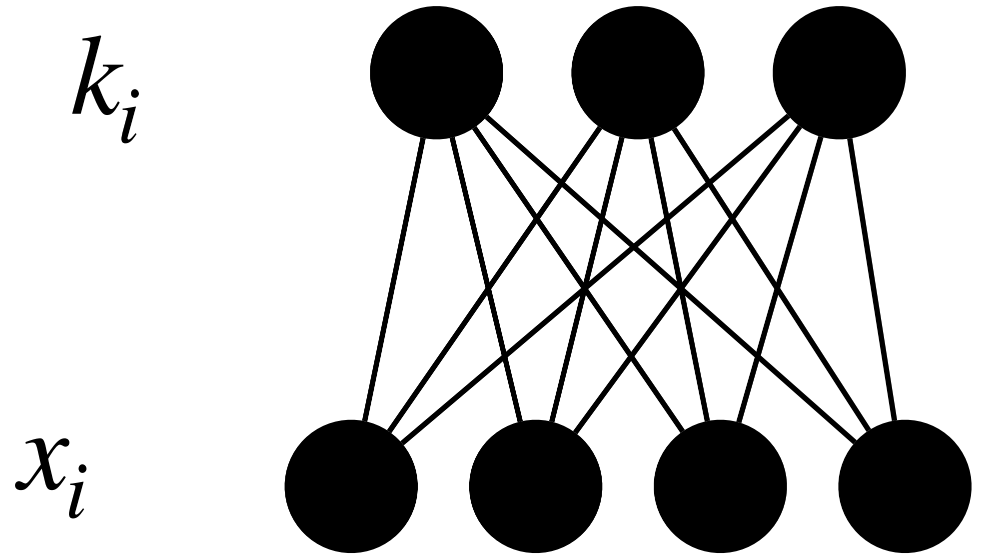


$$q = x_i W_q$$

Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

2 : identifier quels tokens disposent de cette information : ***keys***

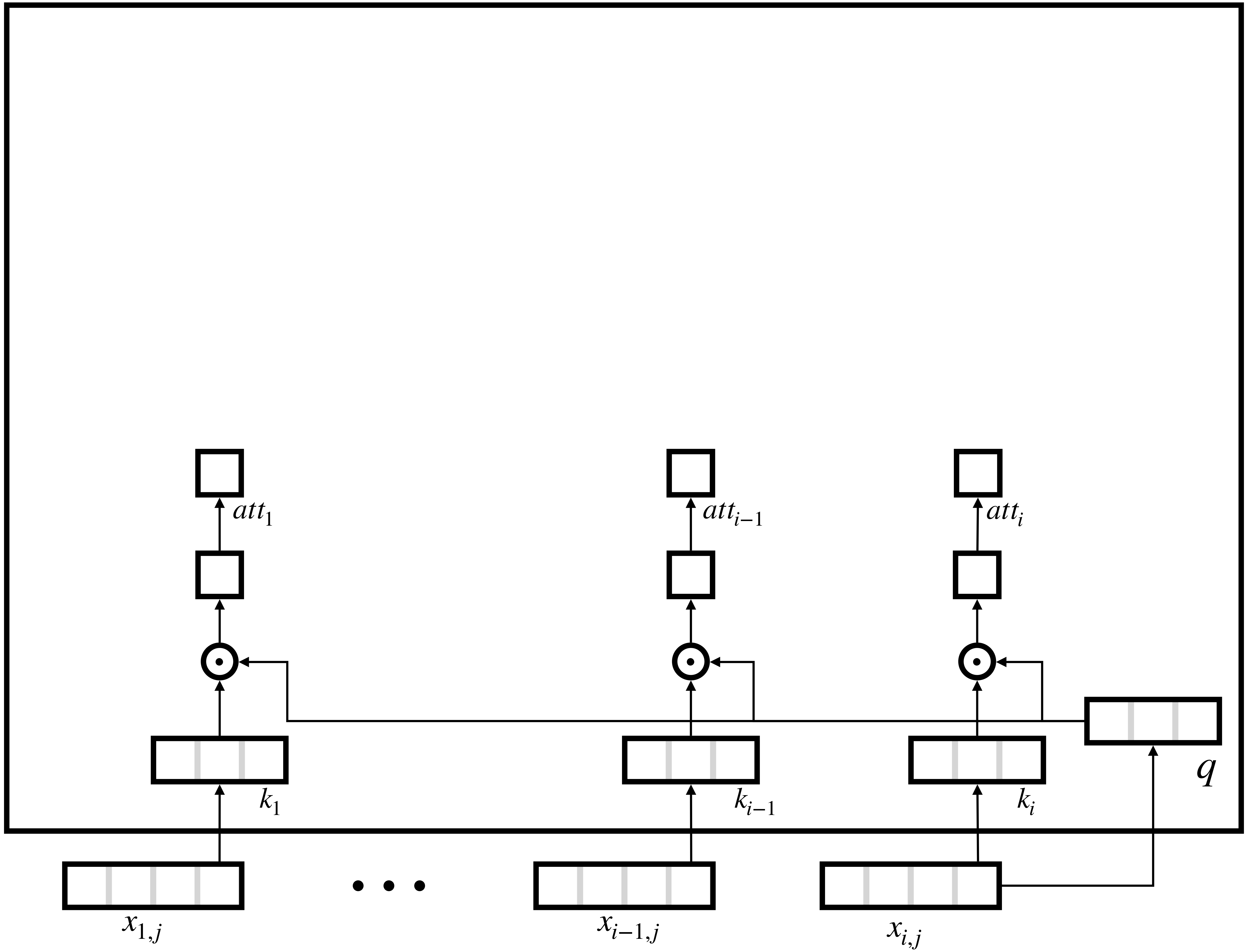




$$k_i = x_i W_k$$

Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

2 : identifier quels tokens disposent de cette information : *keys*



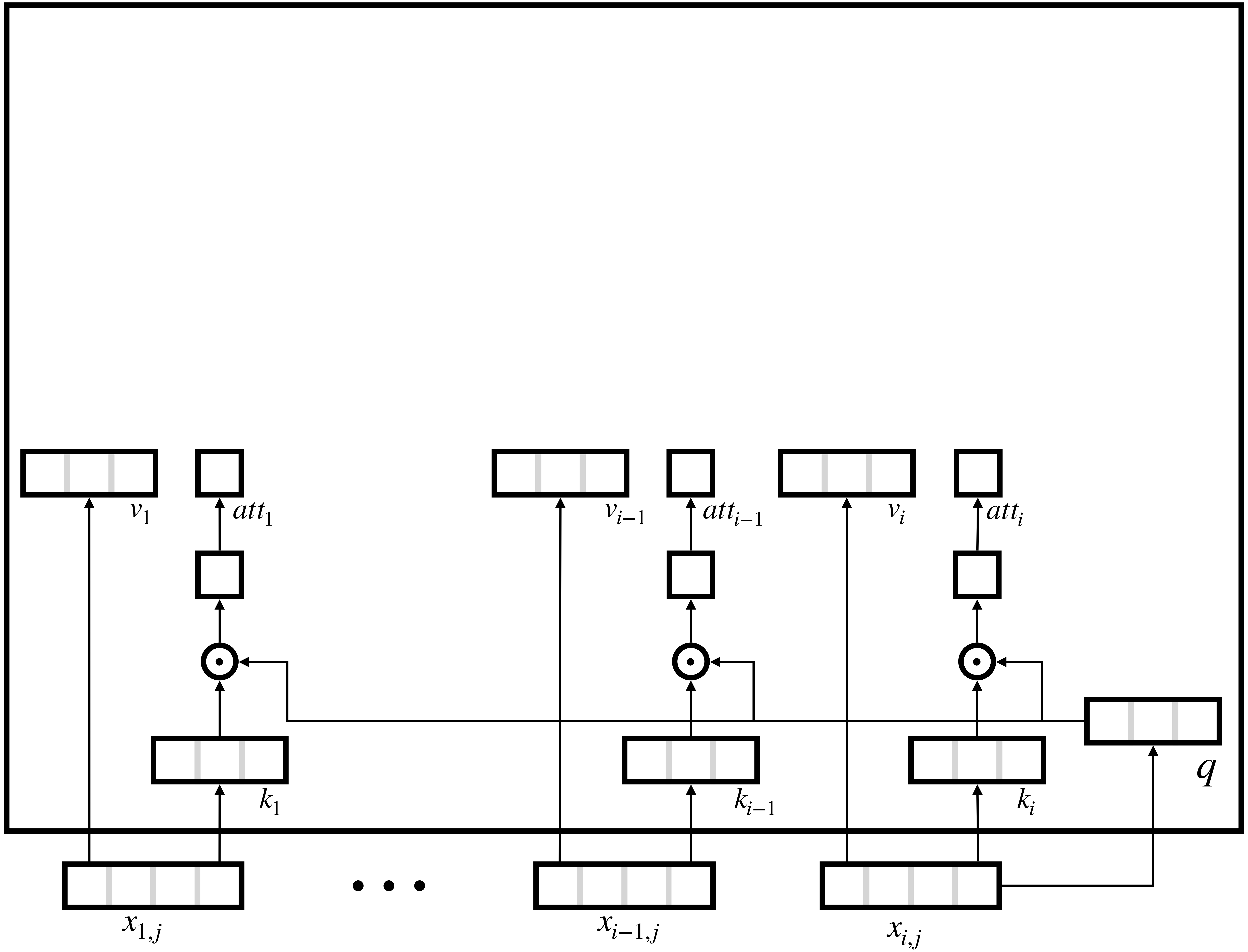
Scores d'attention

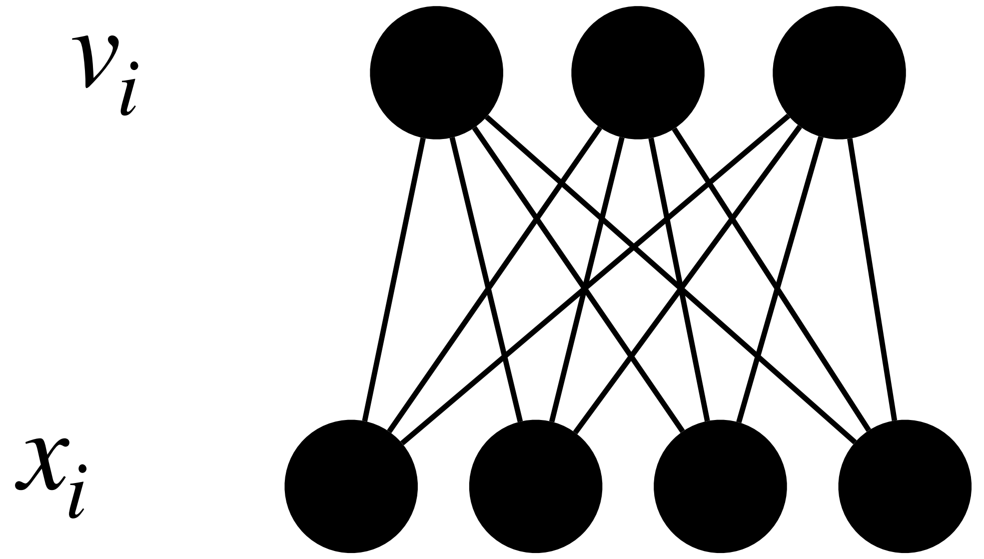
Softmax

Comparaison des clés *k* avec la requête *q* (produit scalaire)

Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

3 : récupérer cette information dans ces tokens : *values*

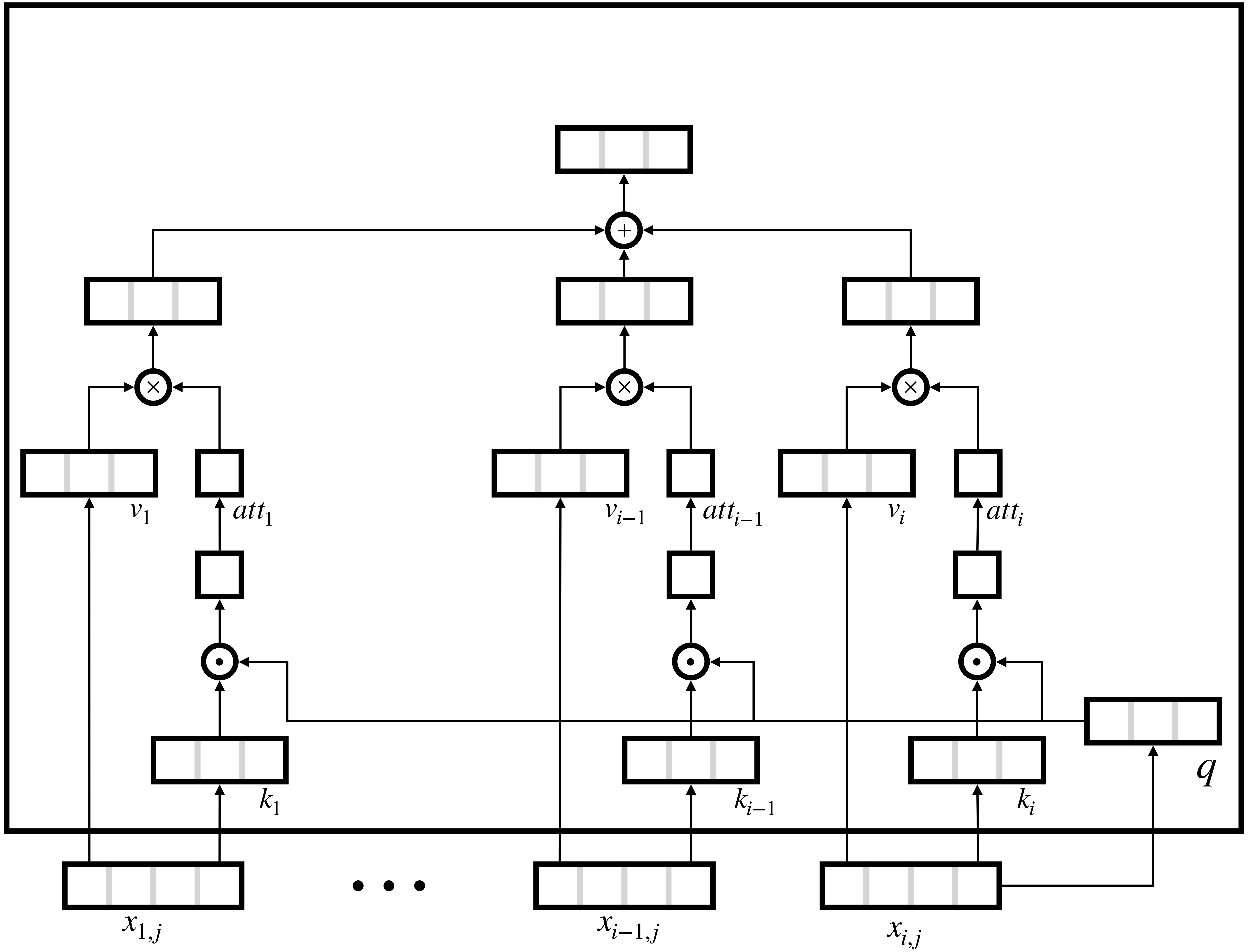




$$v_i = x_i W_v$$

Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

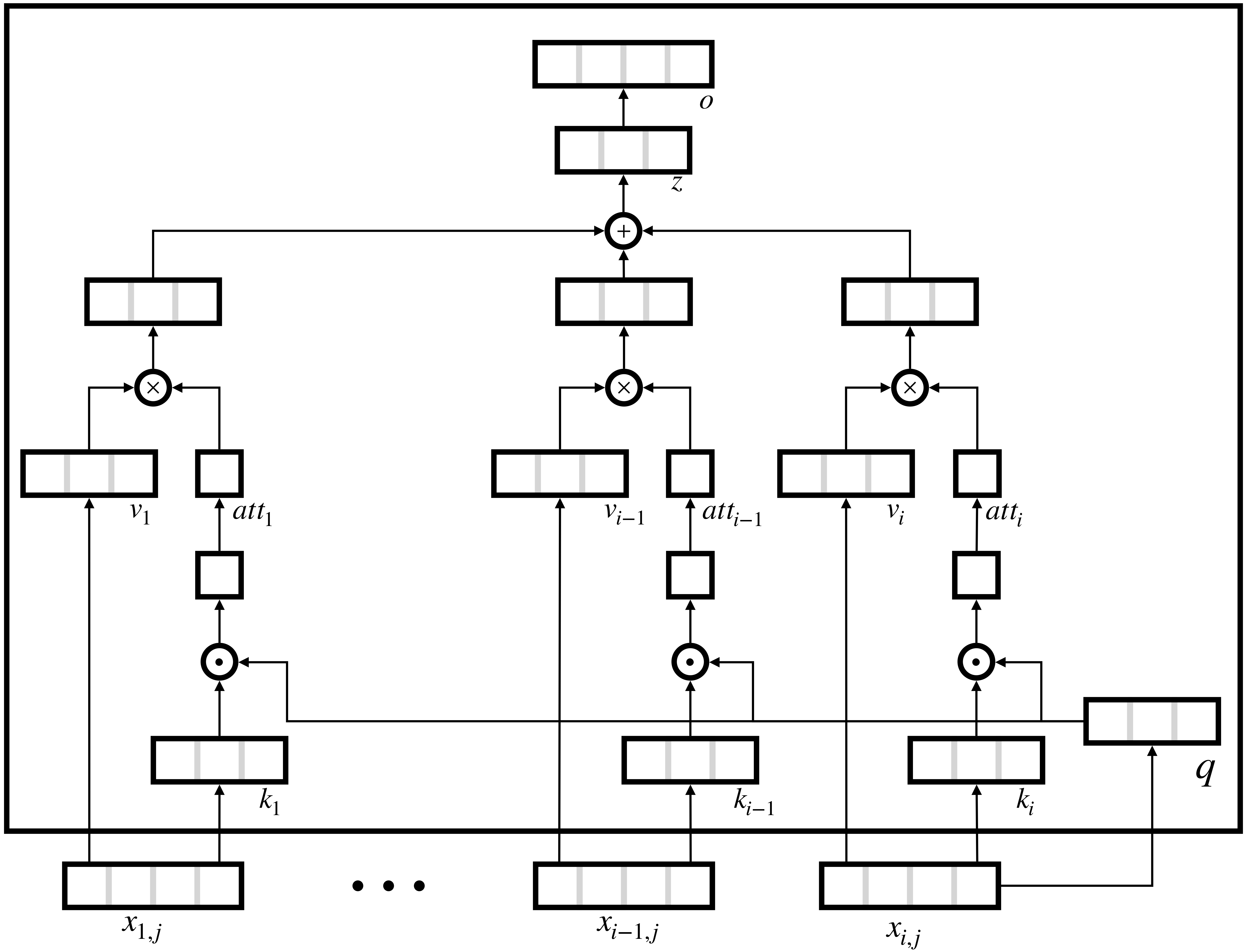
3 : récupérer cette information dans ces tokens : *values*



Somme pondérée par les scores d'attention des vecteurs *values*

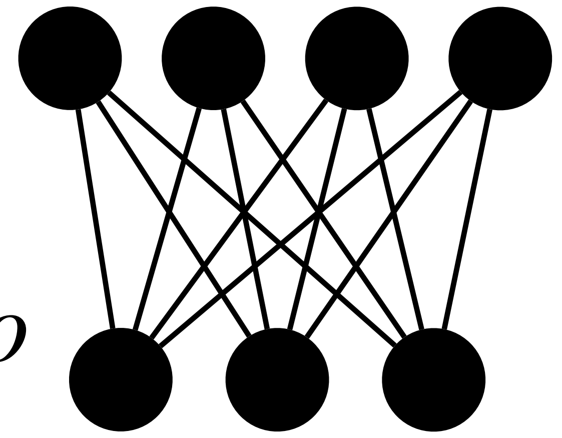
Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

3 : récupérer cette information dans ces tokens : *values*



Reprojection dans l'espace du modèle

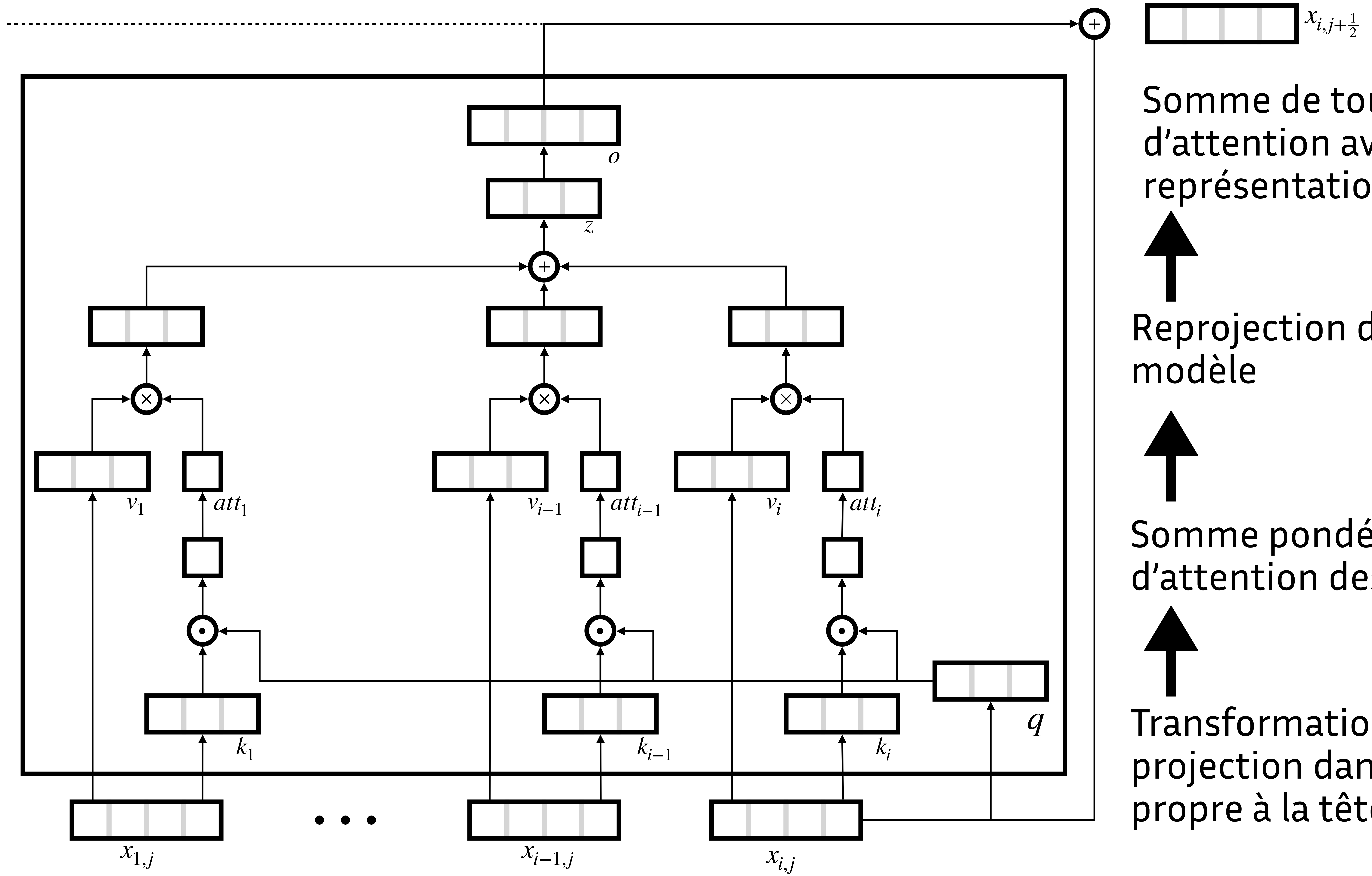
$$o = zW_o$$



Somme pondérée par les scores d'attention des vecteurs *values*

Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

3 : récupérer cette information dans ces tokens : *values*



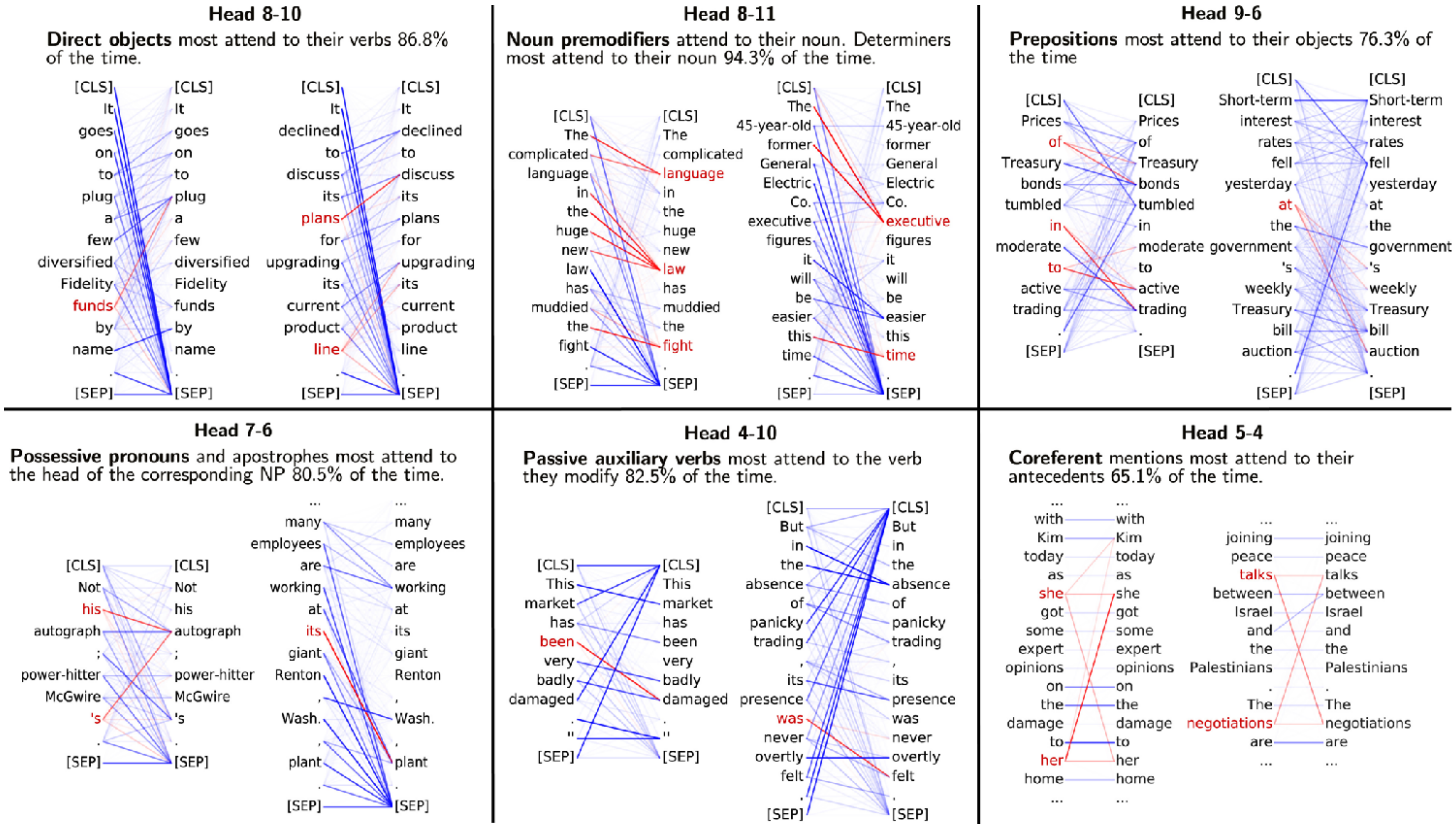
Somme de toutes les têtes d'attention avec la représentation initiale

Reprojection dans l'espace du modèle

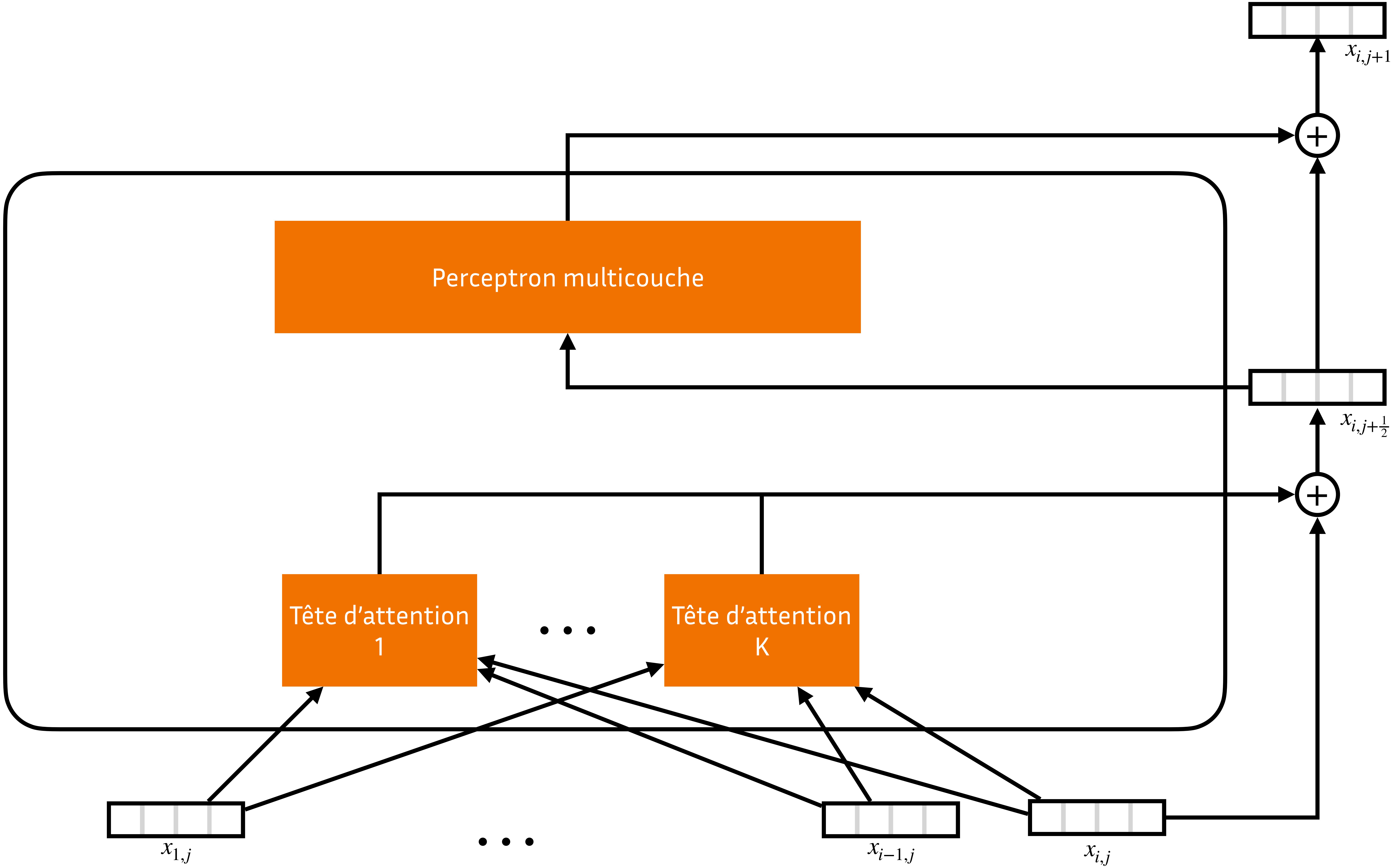
Somme pondérée par les scores d'attention des vecteurs *values*

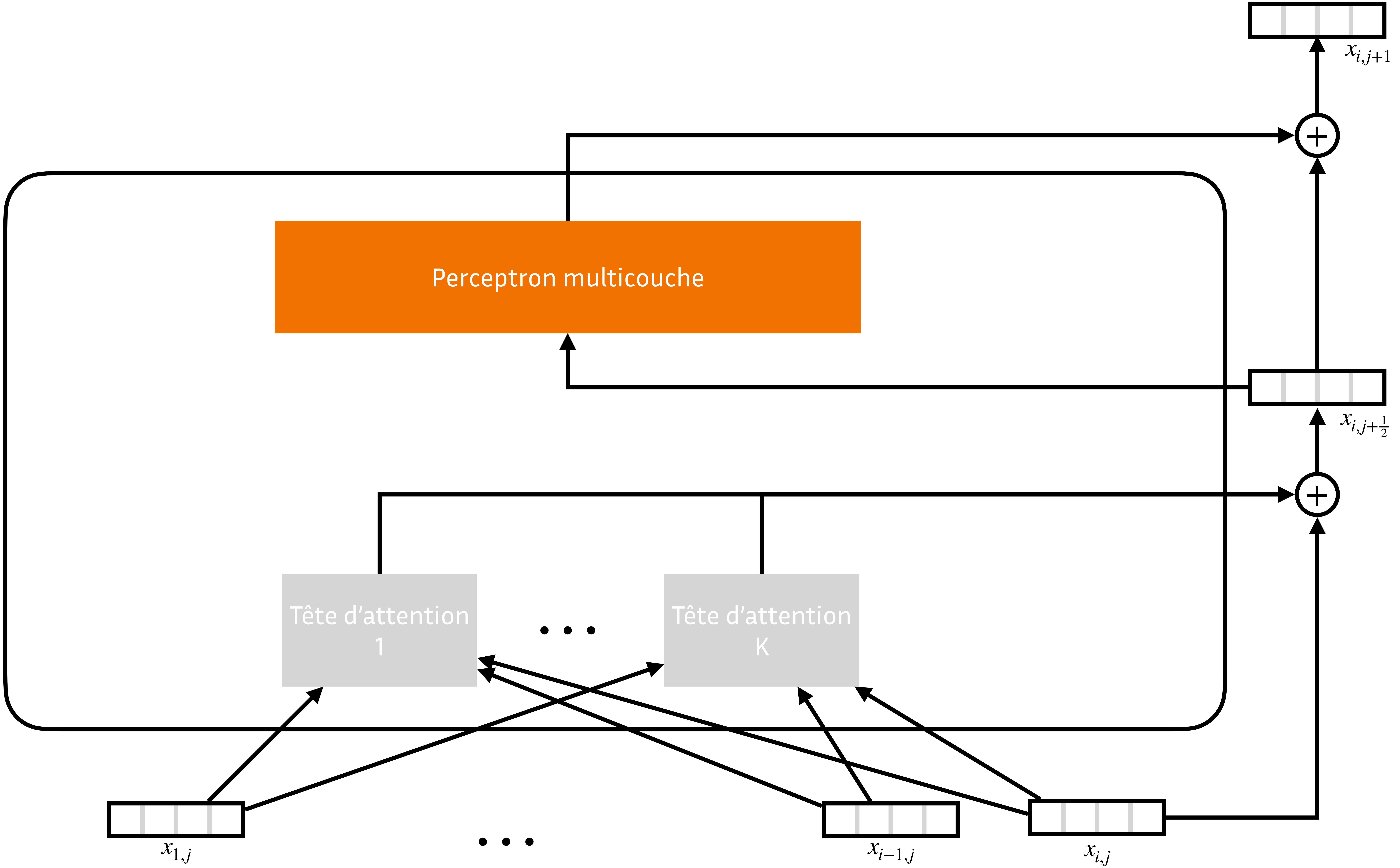
Transformation linéaire : projection dans un sous-espace propre à la tête d'attention

Patterns attentionnels

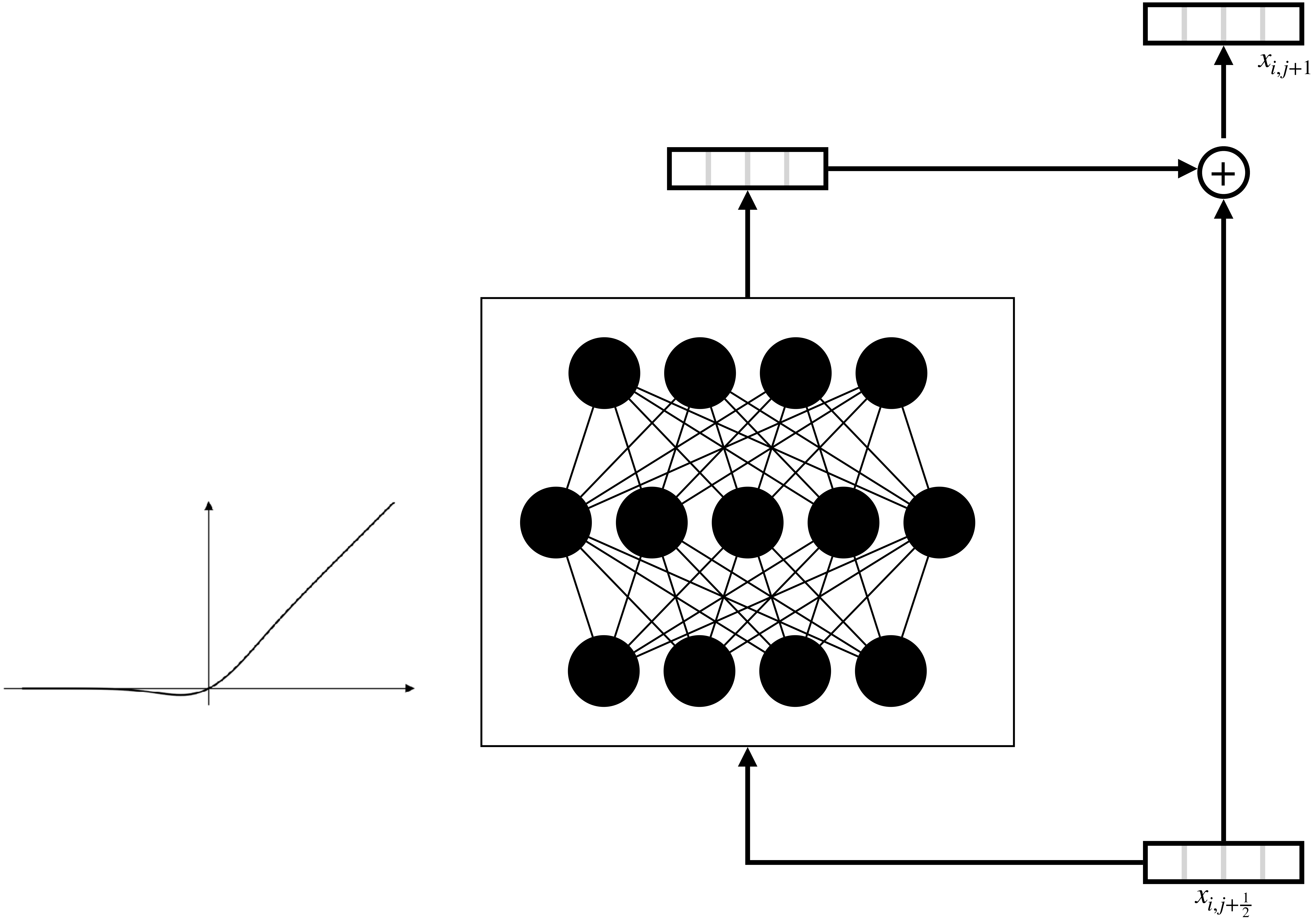


Manning et al., 2020

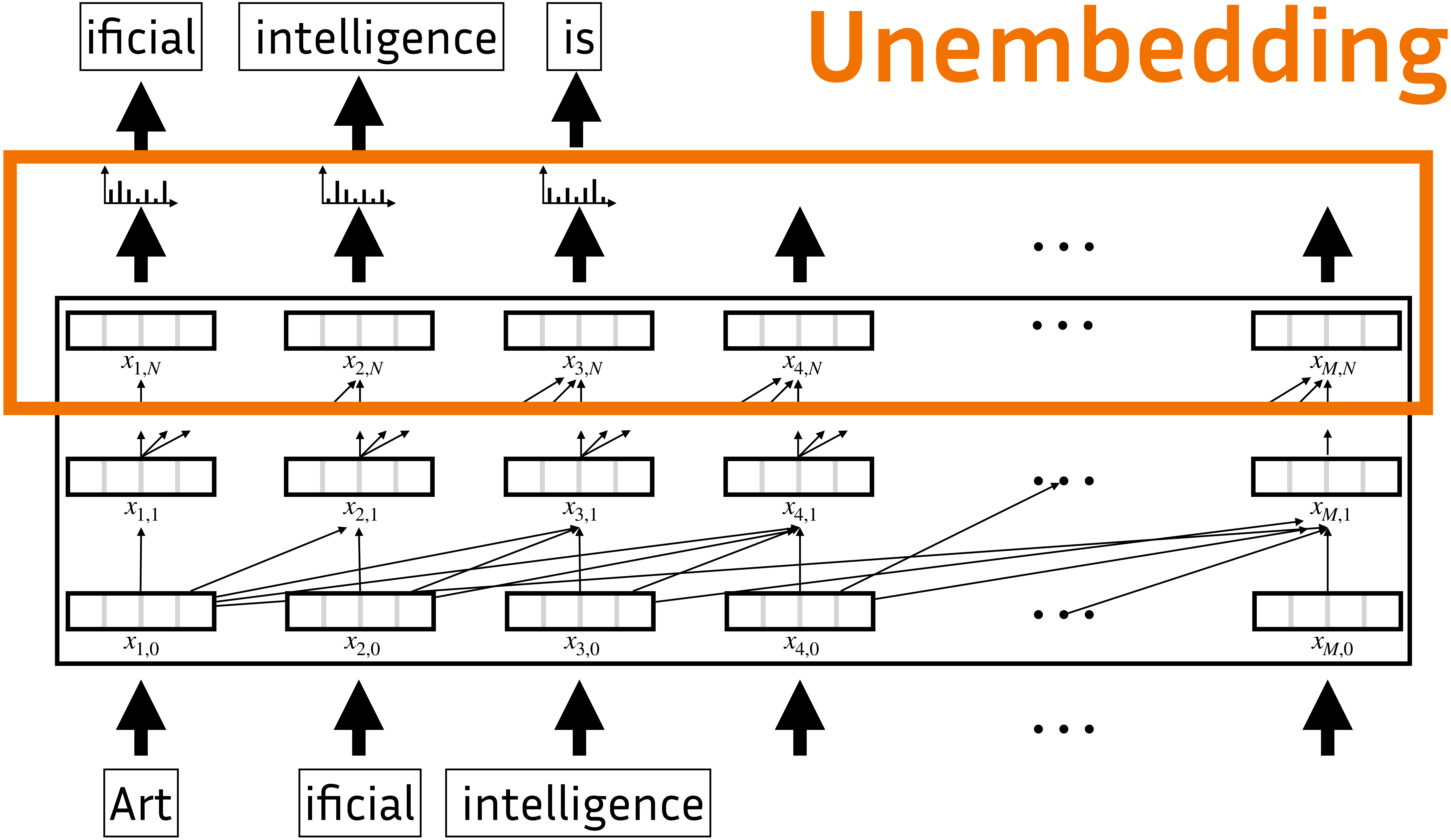

$x_{i,j+1}$
Perceptron multicouche
$x_{i,j+\frac{1}{2}}$
Tête d'attention
1
Tête d'attention
K
$x_{1,j}$
$x_{i-1,j}$
$x_{i,j}$

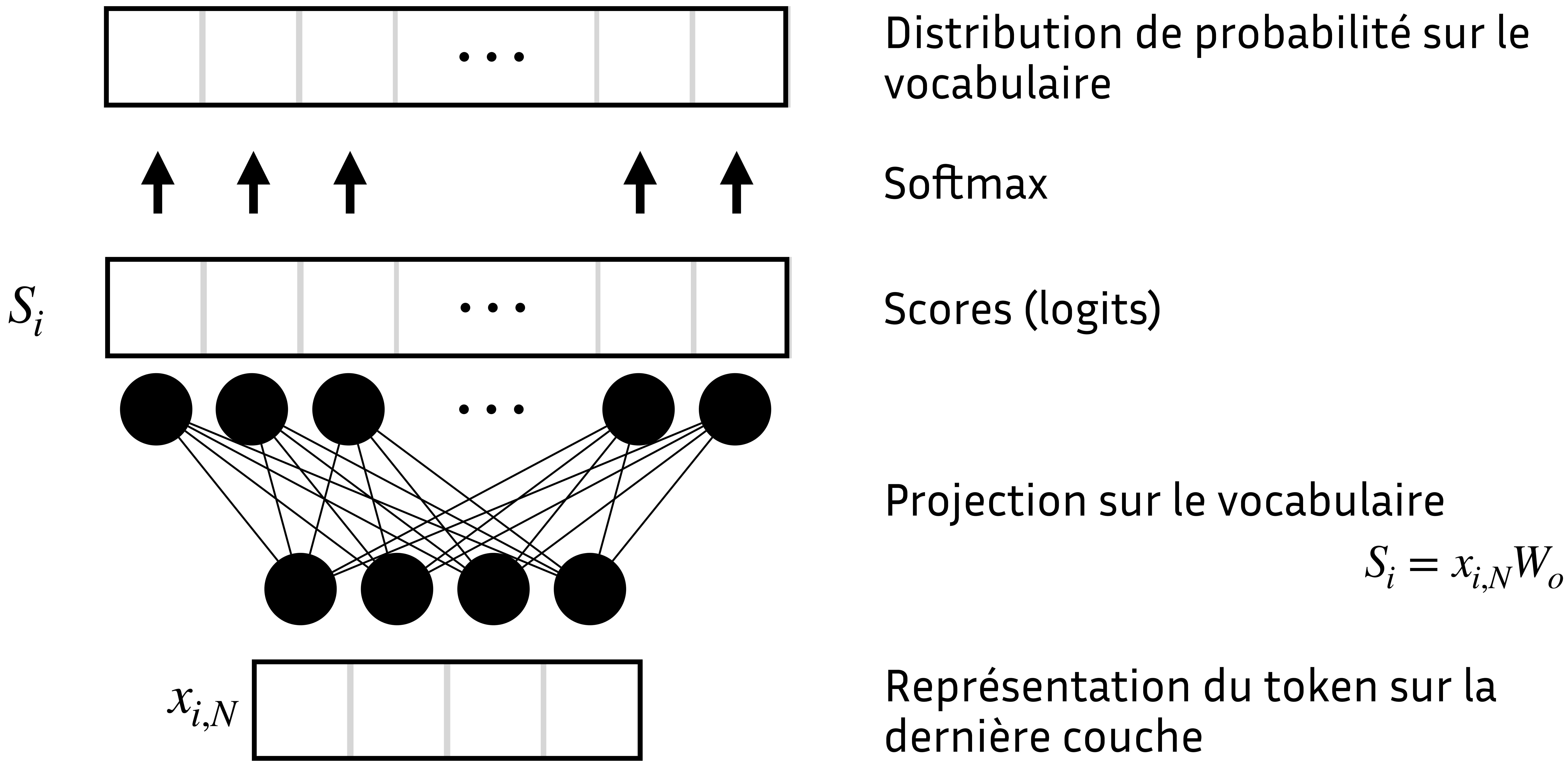


$x_{i,j+1}$
Perceptron multicouche
$x_{i,j+\frac{1}{2}}$
Tête d'attention 1
...
Tête d'attention K
$x_{1,j}$
...
$x_{i-1,j}$
$x_{i,j}$


$x_{i,j+1}$
$x_{i,j+\frac{1}{2}}$

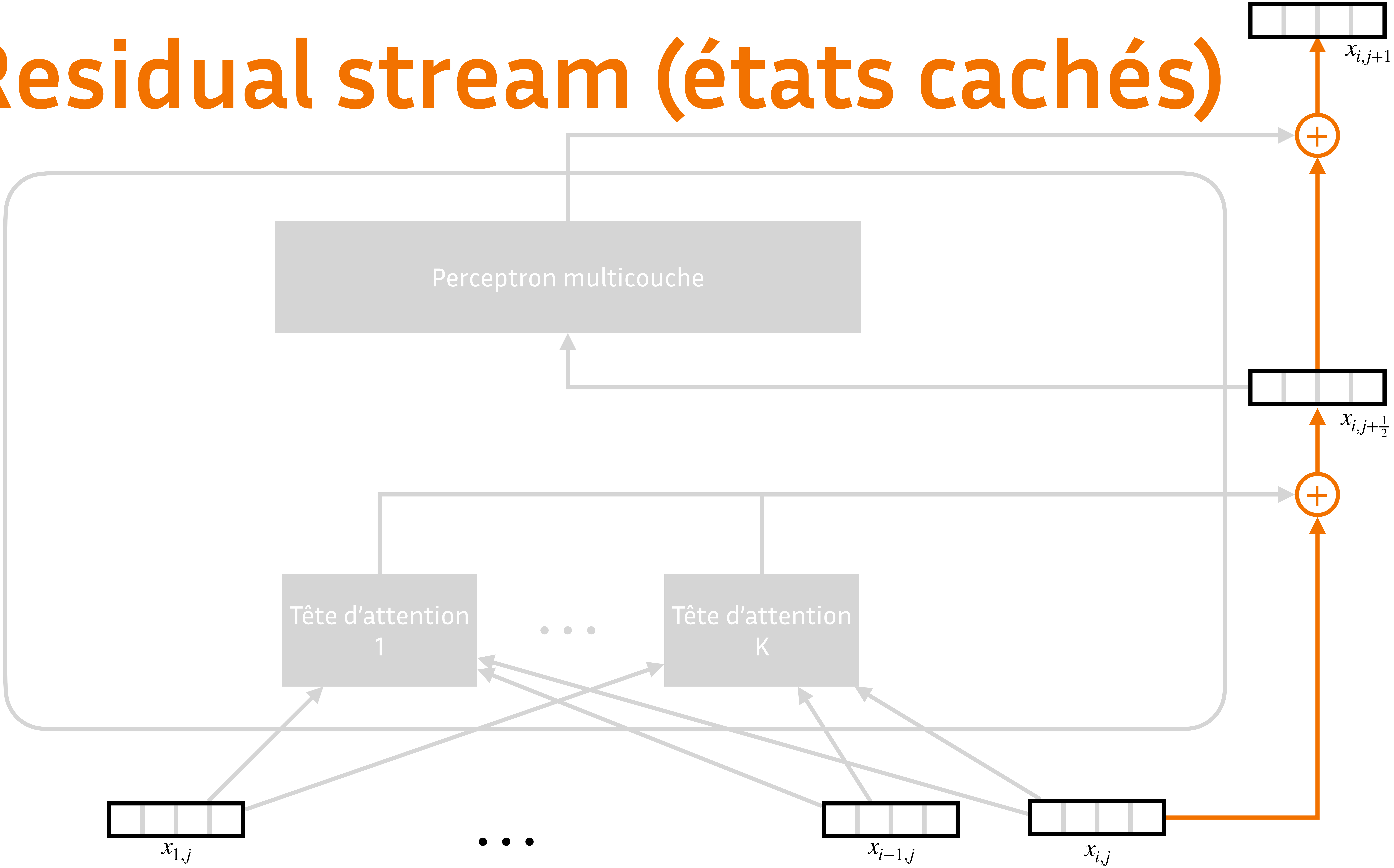
Unembedding



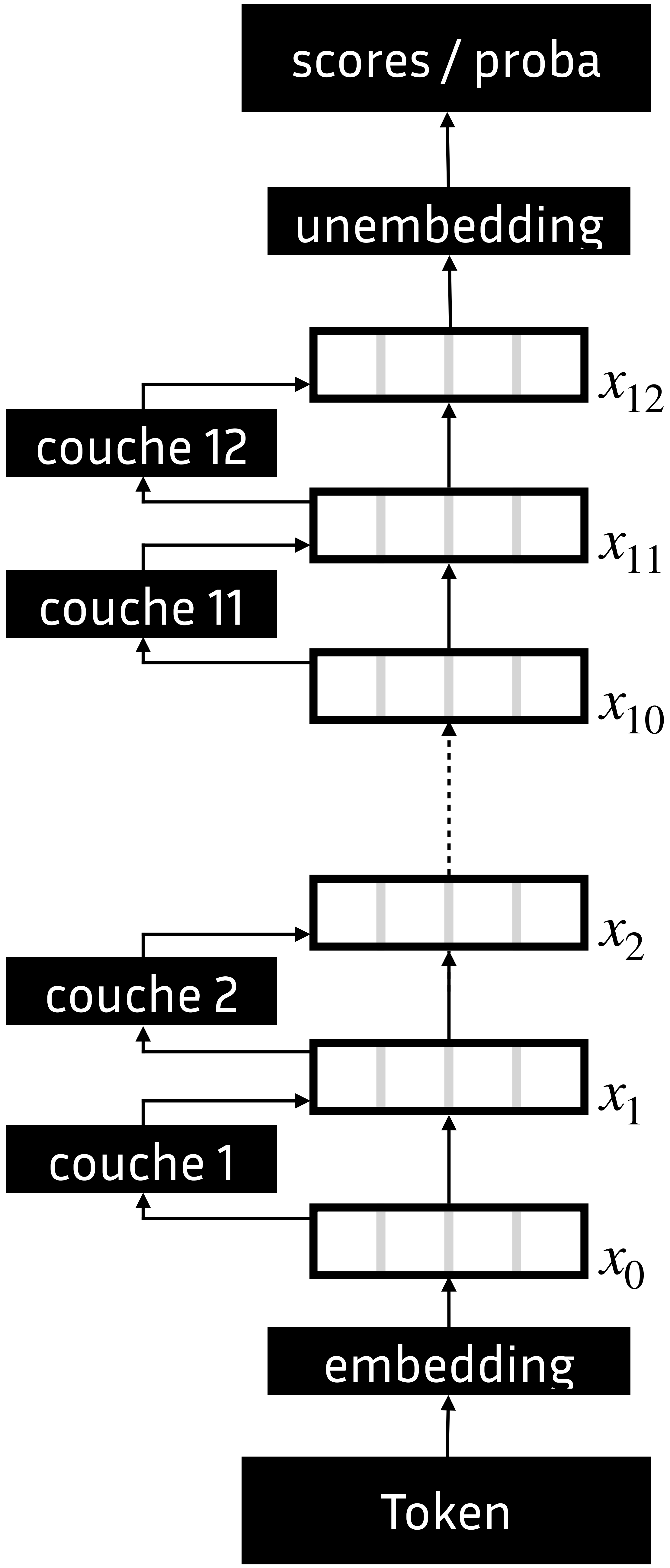

Distribution de probabilité sur le vocabulaire
Softmax
S_i
Scores (logits)
Projection sur le vocabulaire
$S_i = x_{i,N} W_o$
$x_{i,N}$
Représentation du token sur la dernière couche

Représentations

Residual stream (états cachés)

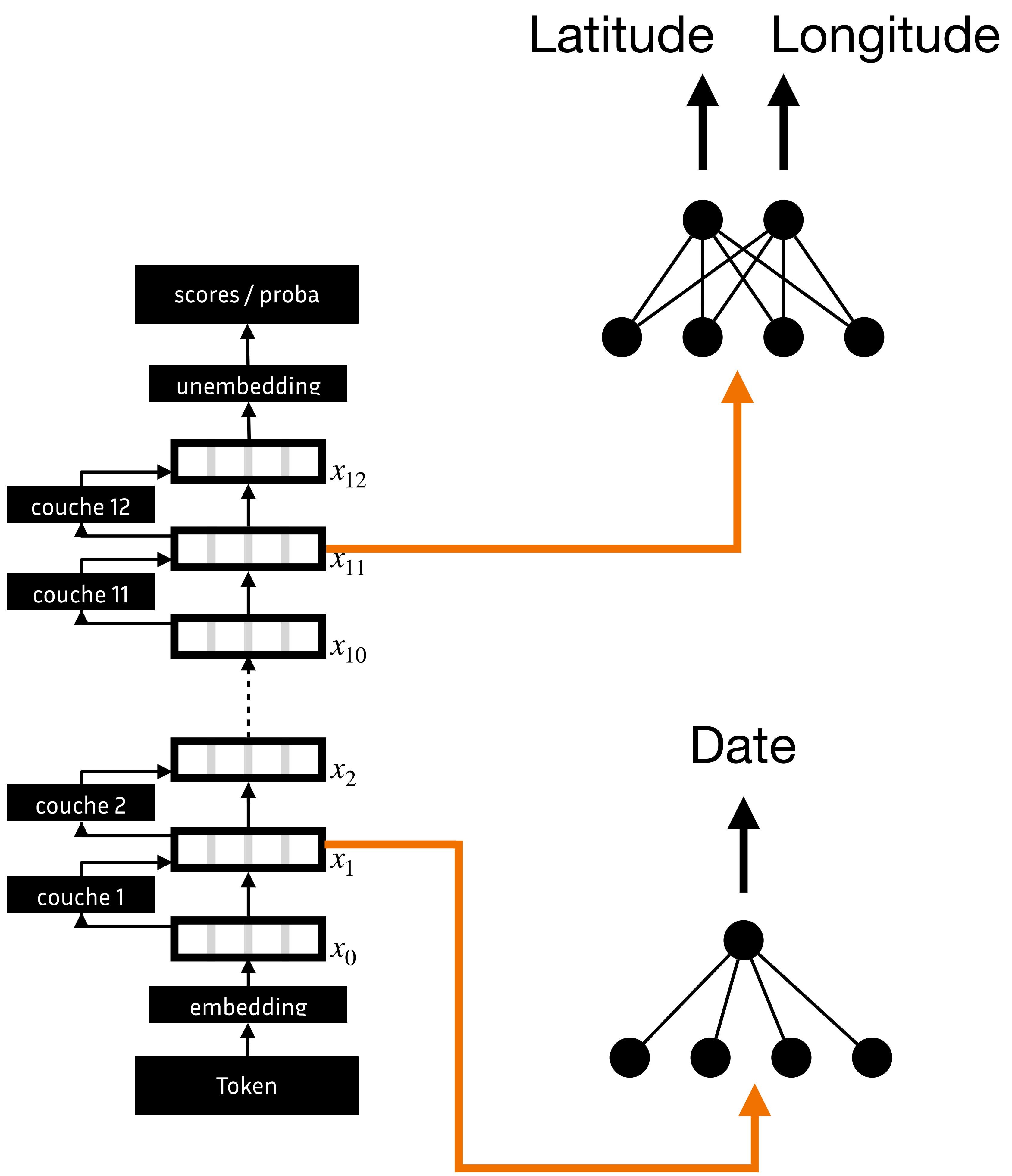


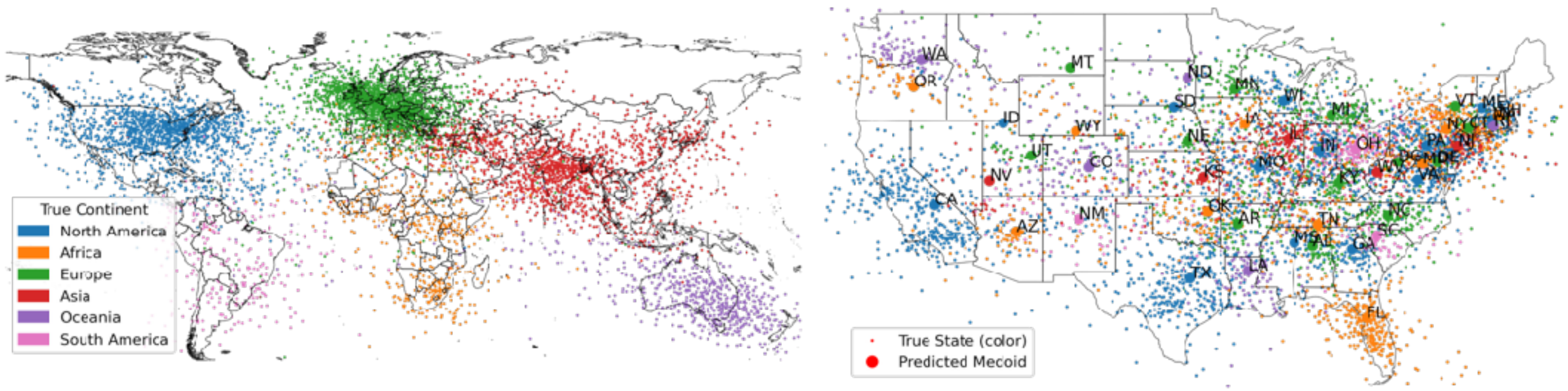
Transformations / complexifications successives de la représentation des tokens

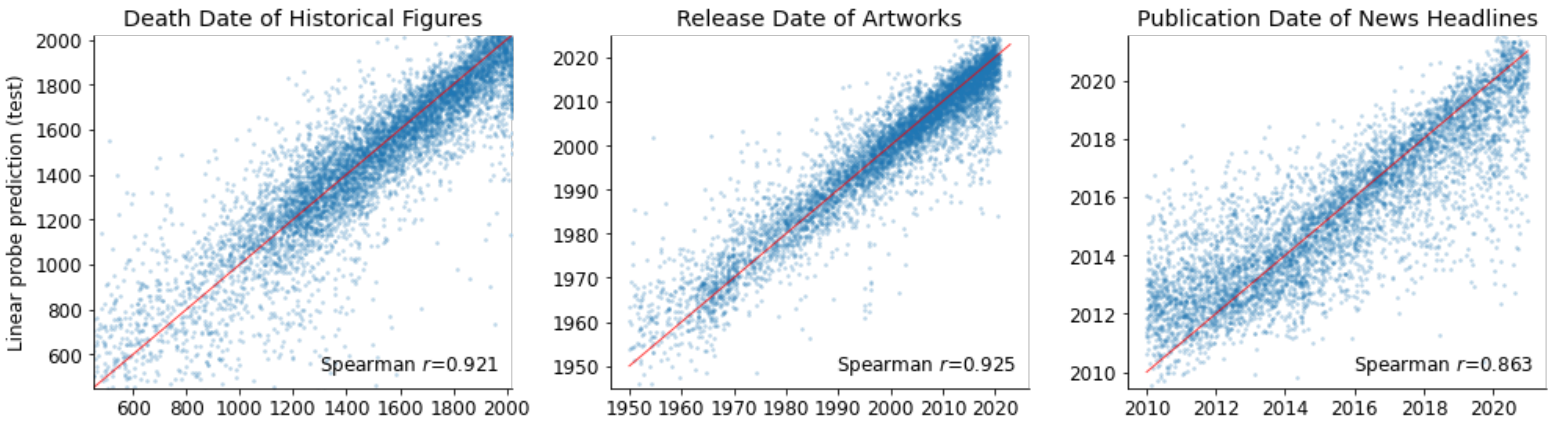


représentation suffisamment élaborée pour permettre la prédiction du token suivant

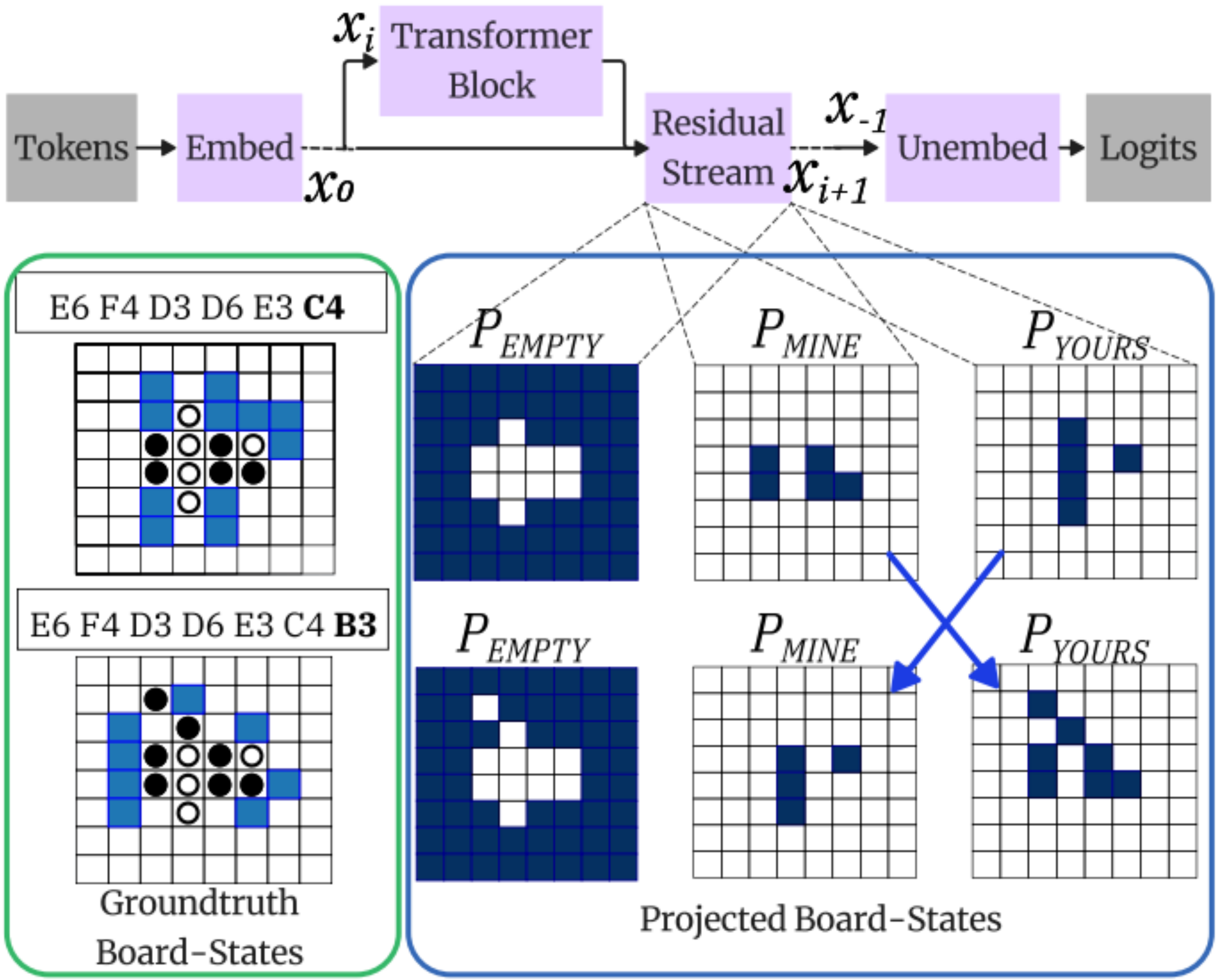
représentation du token d'entrée







Gurnee & Tegmark, 2024



Tokens
Embed
x_0
x_i
Transformer Block
Residual Stream
x_{i+1}
x_{-1}
Unembed
Logits
E6 F4 D3 D6 E3 **C4**
E6 F4 D3 D6 E3 C4 **B3**
Groundtruth Board-States
P_{EMPTY}
P_{MINE}
P_{YOURS}
P_{EMPTY}
P_{MINE}
P_{YOURS}
Projected Board-States

Bibliographie

Gurnee &Tegmark, 2024. Language Models Represent Space and Time. https://arxiv.org/abs/2310.02207

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *PNAS 117*(48)

Minaee et al, 2024. Large Language Models : a Survey. https://arxiv.org/abs/2402.06196

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent Linear Representations in World Models of Self-Supervised Sequence Models. *ArXiv, abs/2309.00941*.

Zhao et al, 2023. A Survey of Large Language Models. https://arxiv.org/abs/2303.18223