

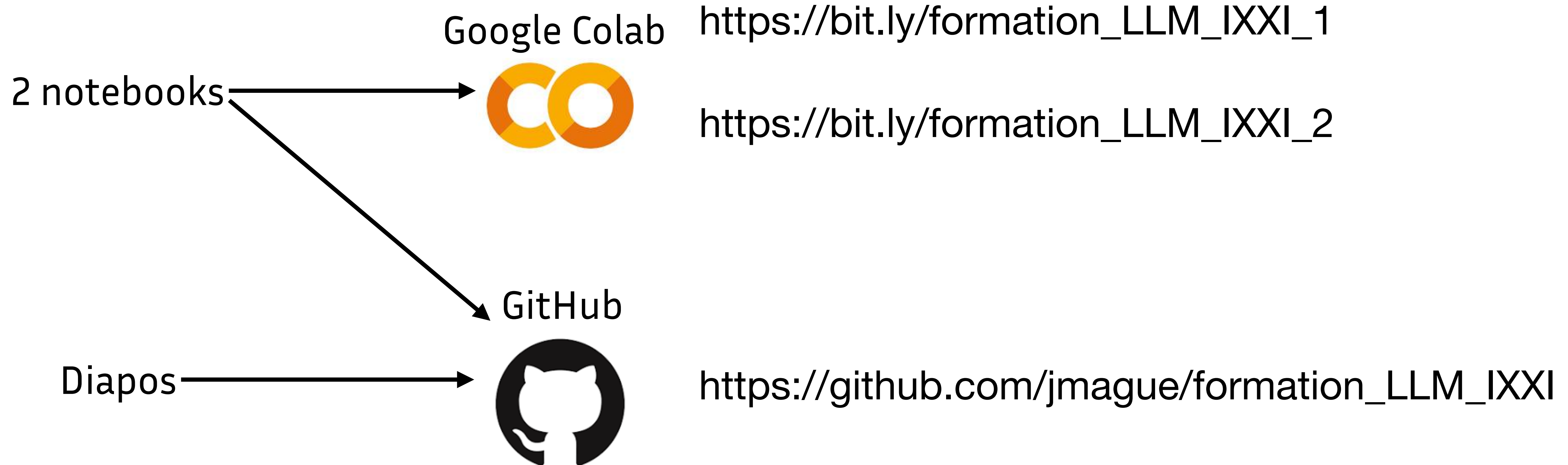


Grands Modèles de Langue

Jean-Philippe Magué

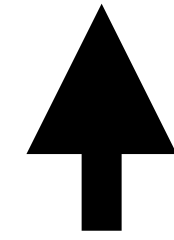
juillet 2025

Ressources

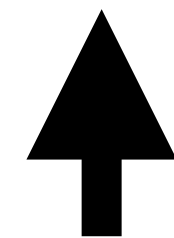


Premières générations de texte

?



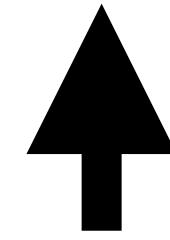
GPT2



Artificial intelligence

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

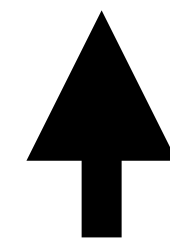
?



GPT2



Art	ificial	intelligence
-----	---------	--------------



Suite de tokens

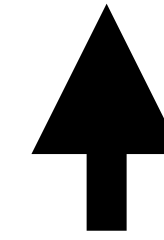
Tokenisation

Suite de caractères

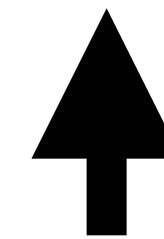
Artificial intelligence

Art ificial intelligence is

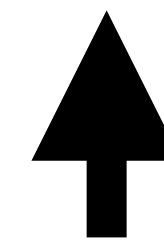
Le modèle génère
un nouveau token



GPT2

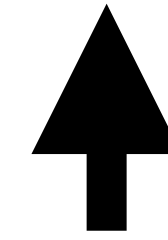


Art ificial intelligence



Artificial intelligence

Art ificial intelligence is a

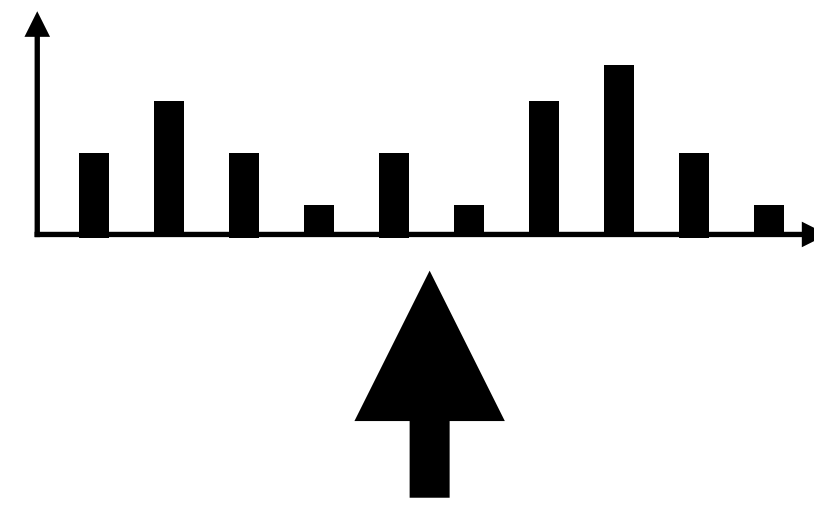


Art ificial intelligence is



Stratégies de génération de texte

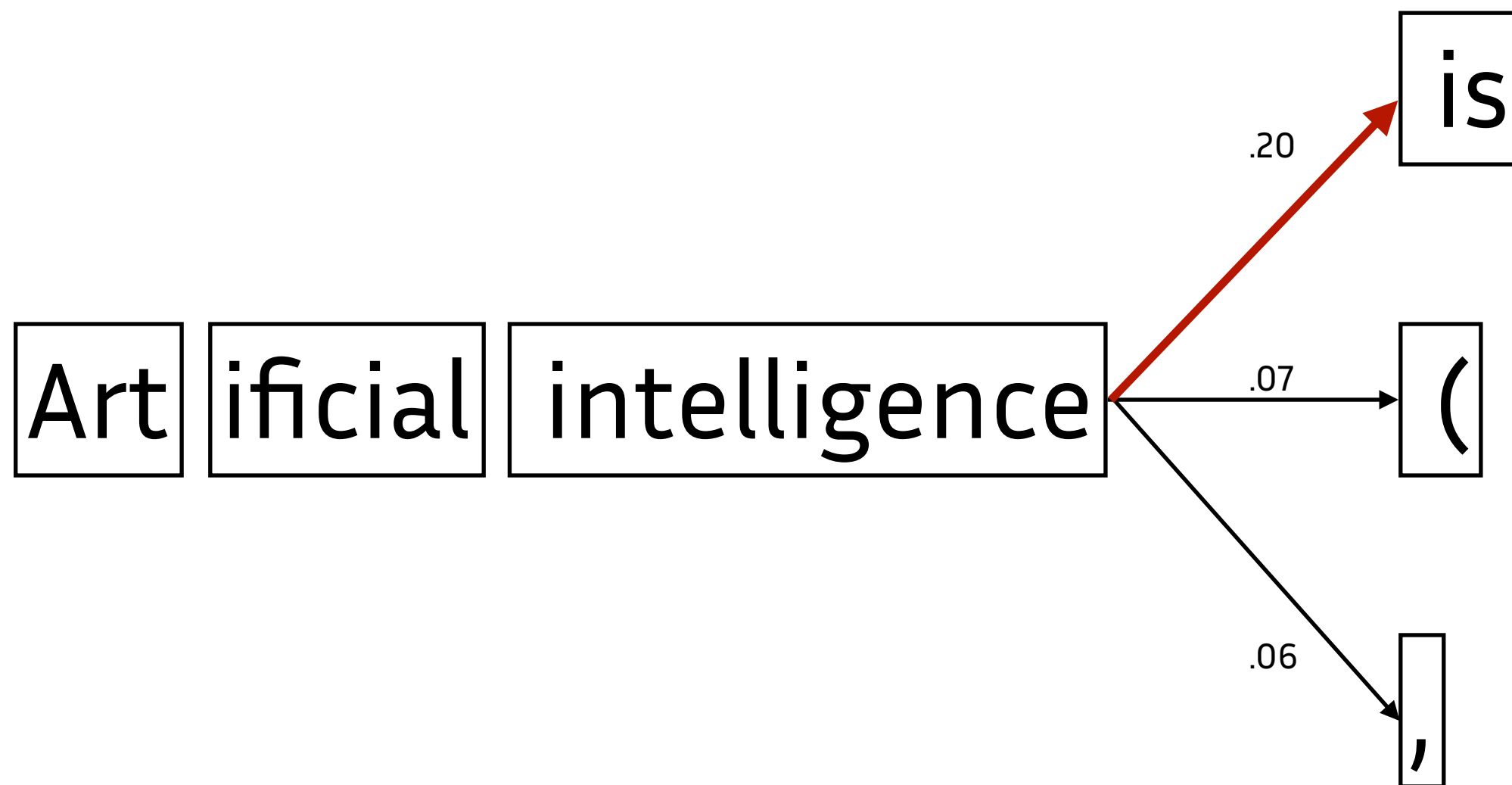
Le modèle génère
une distribution
de probabilité sur
les tokens



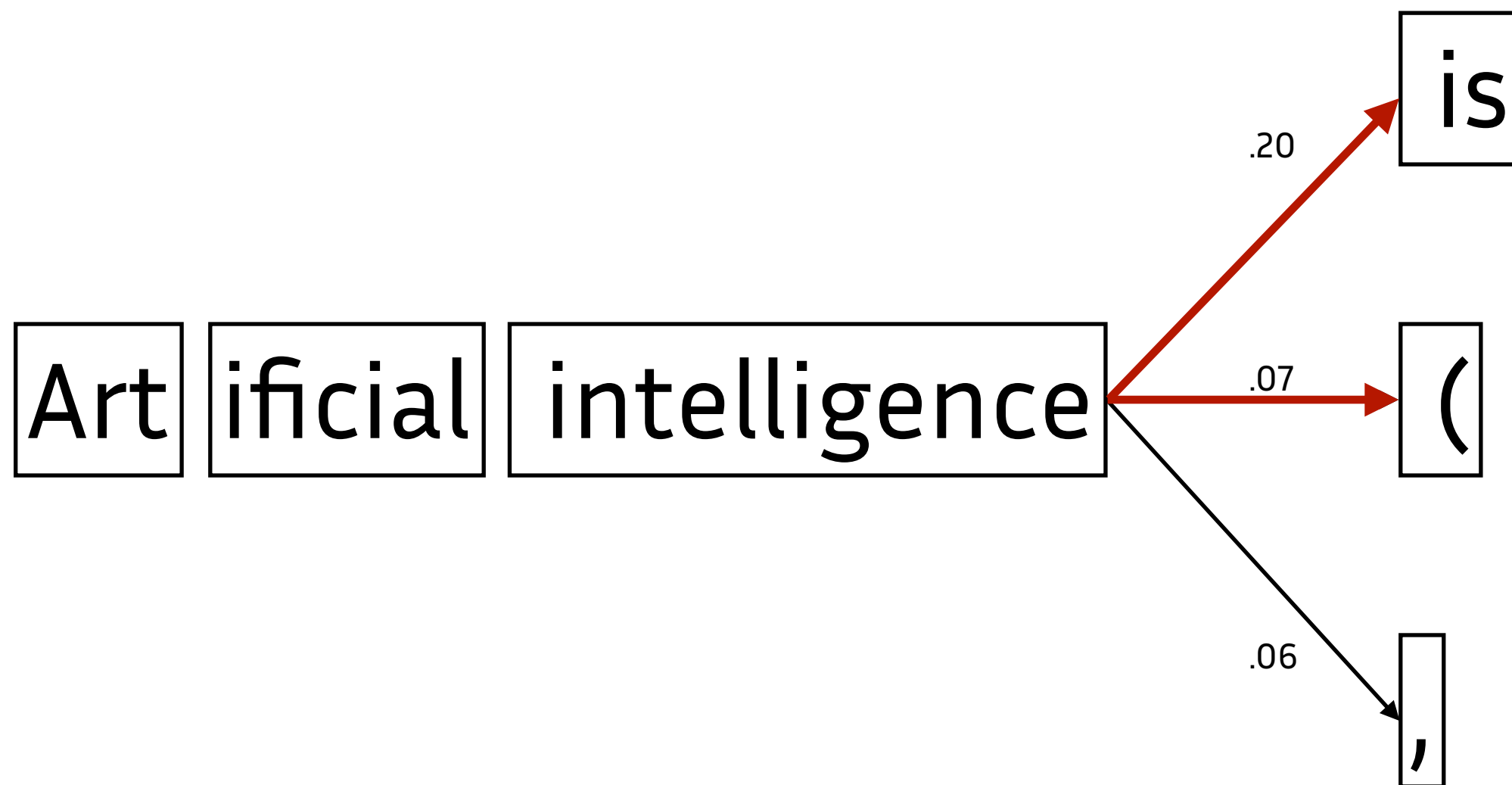
GPT2

Art ificial intelligence

Artificial intelligence

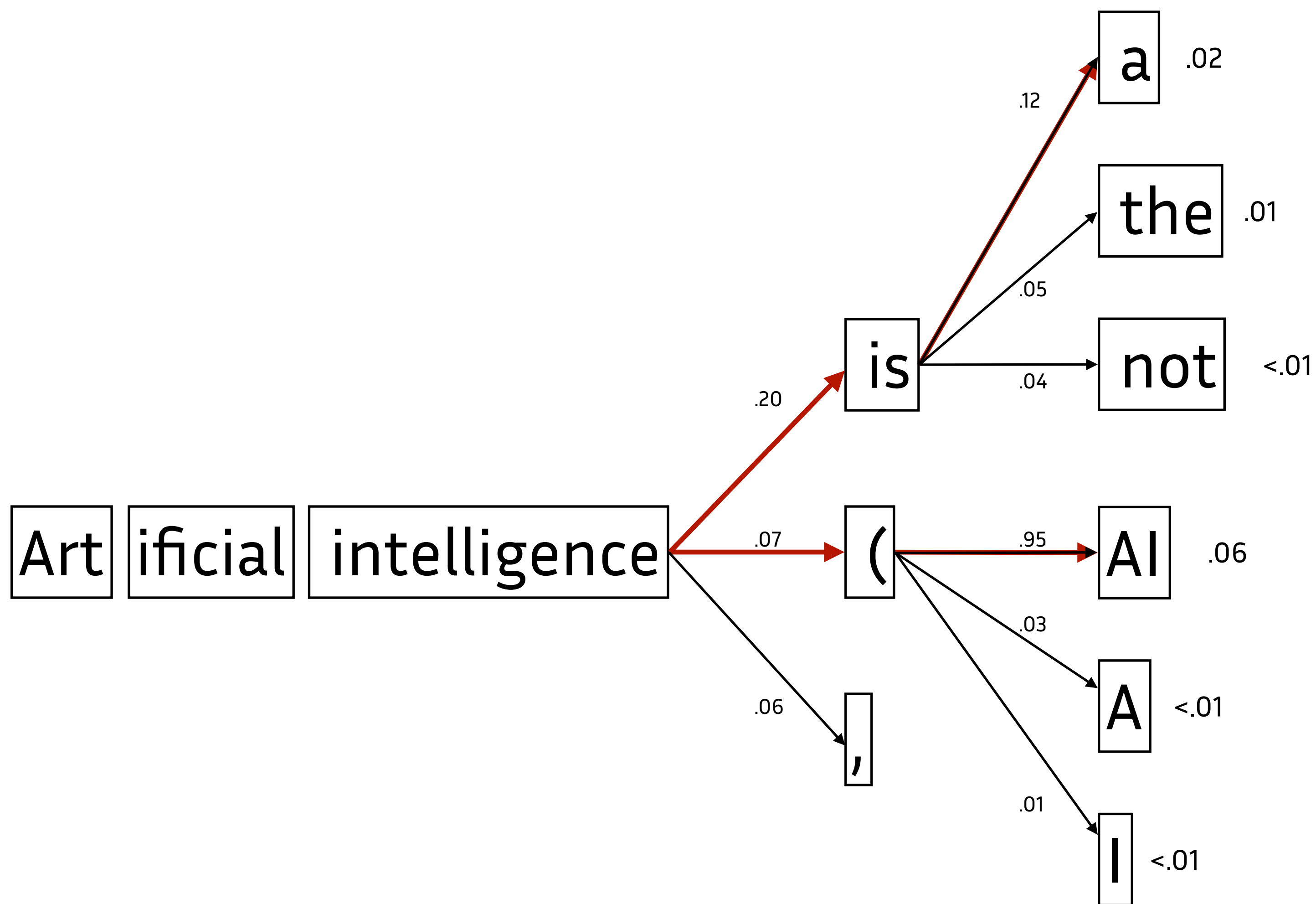


Greedy search



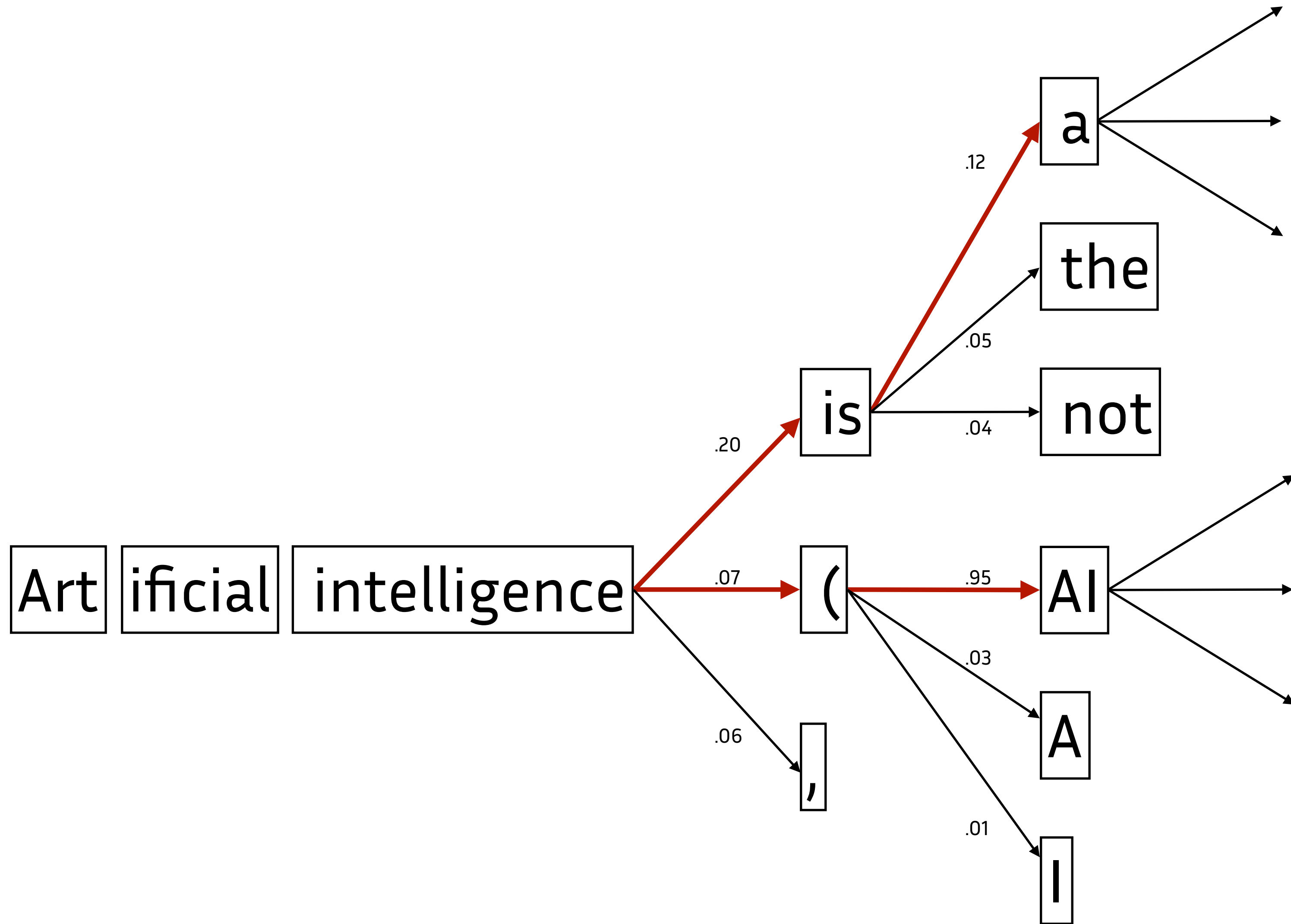
Beam search

(2 faisceaux)



Beam search

(2 faisceaux)



Beam search

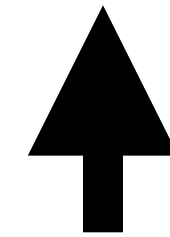
(2 faisceaux)

Apprentissage

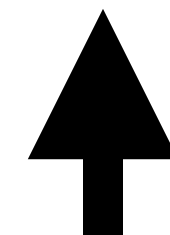
Apprentissage

Apprentissage
auto-supervisé

~~butterfly~~ a



GPT2



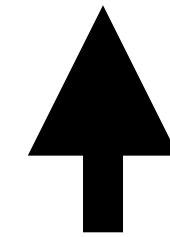
Art ificial intelligence is

Apprentissage

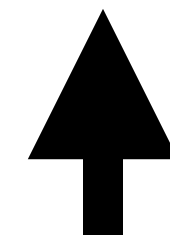
Instruction tuning

The capital of Australia is Canberra.

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



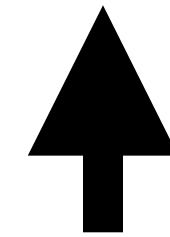
GPT2



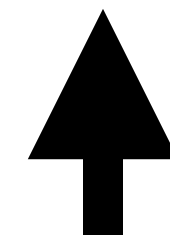
What is the capital of Australia?

Apprentissage

Apprentissage par
renforcement à partir
de rétroaction humaine



GPT2



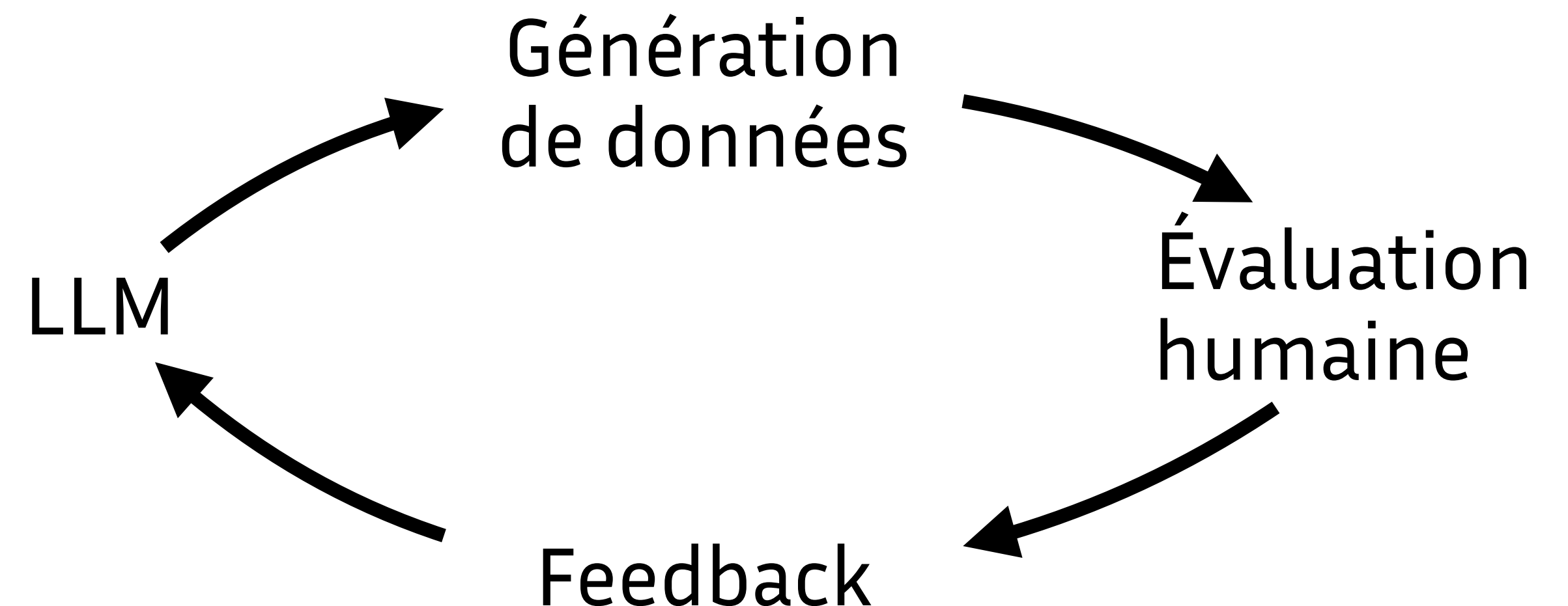
compared to men, women are more likely to

Apprentissage

Apprentissage par
renforcement à partir
de rétroaction humaine

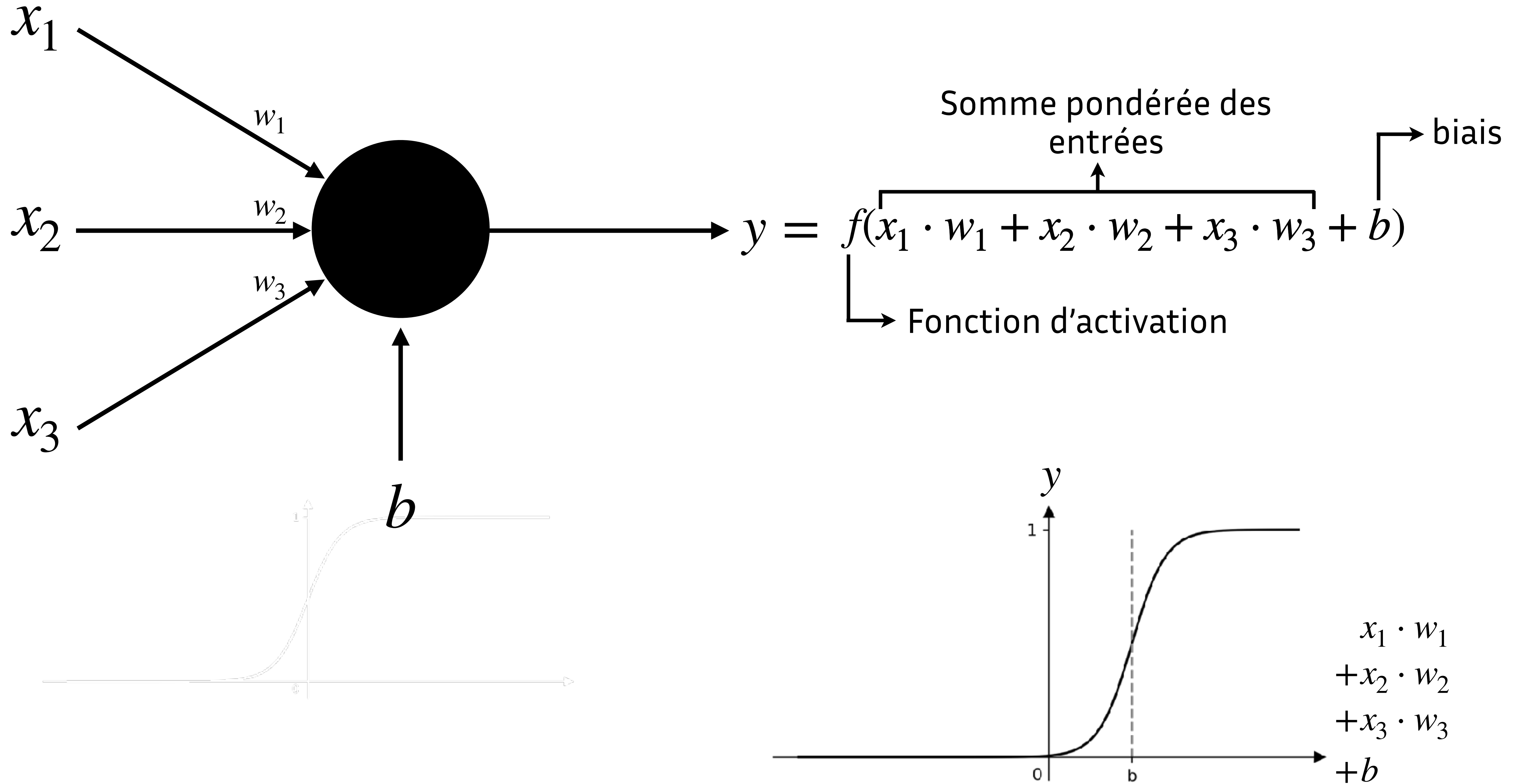
compared to men, women are more likely to

- have problems eating or drinking alcohol
- be underrepresented in science, technology, and academia
- have experienced sexual violence
- have sex when working
- be overweight

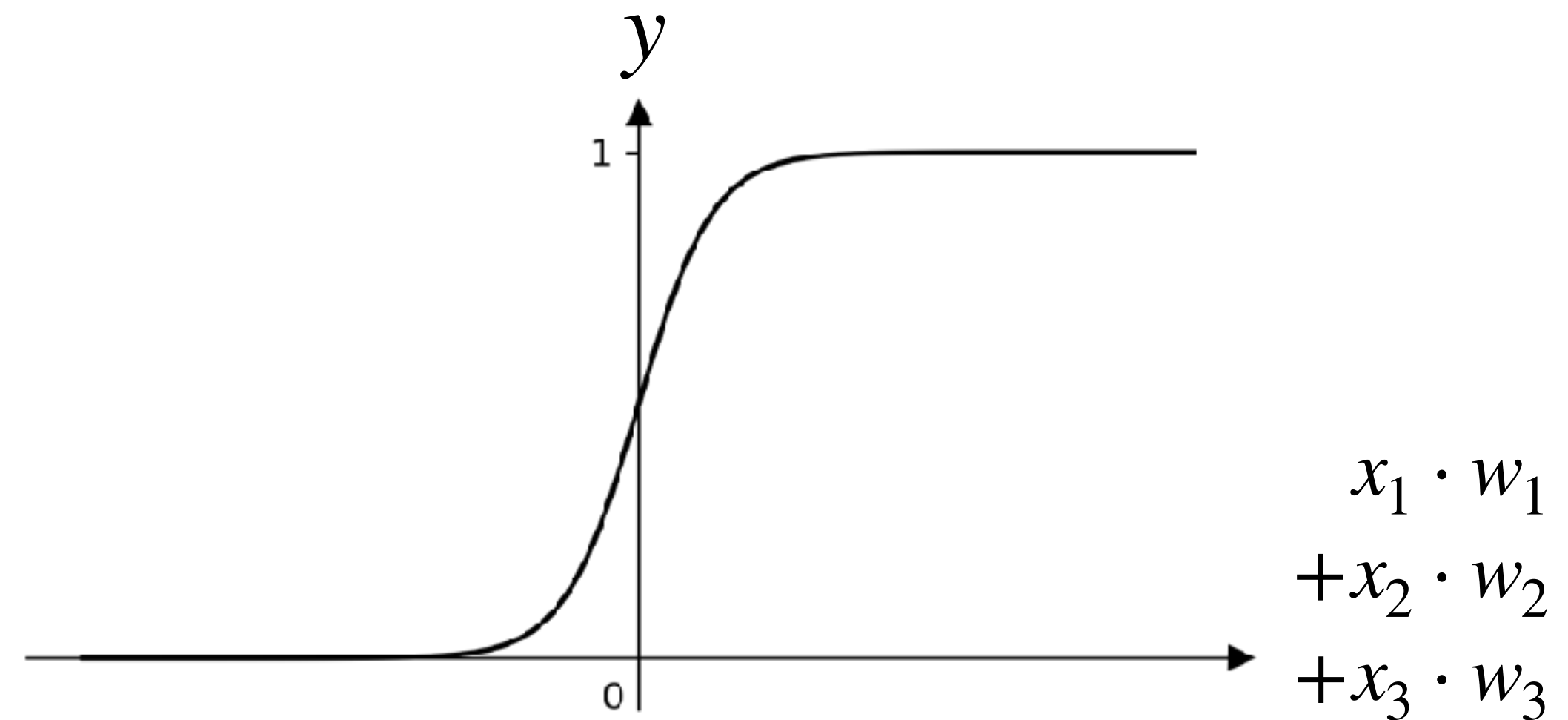
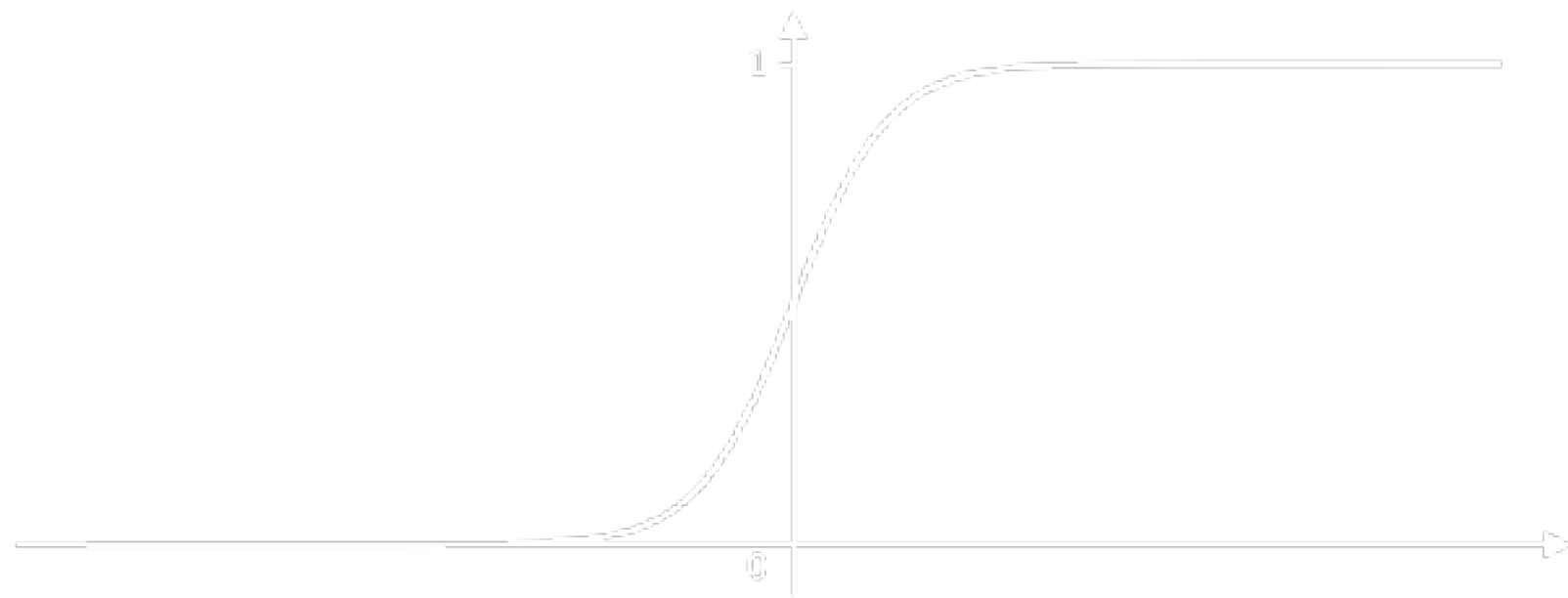
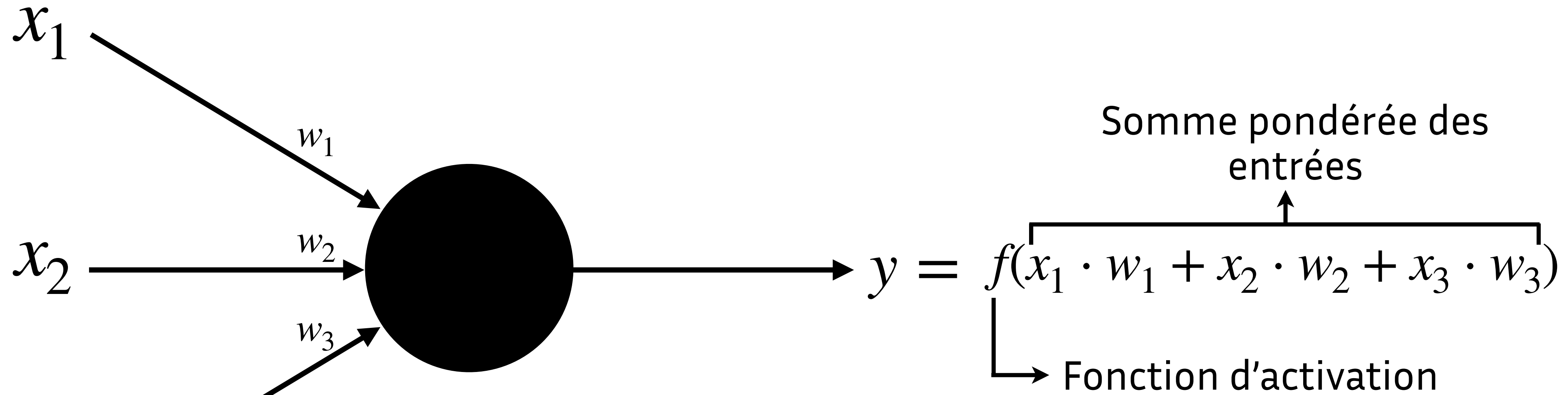


Plongée dans la boîte noire

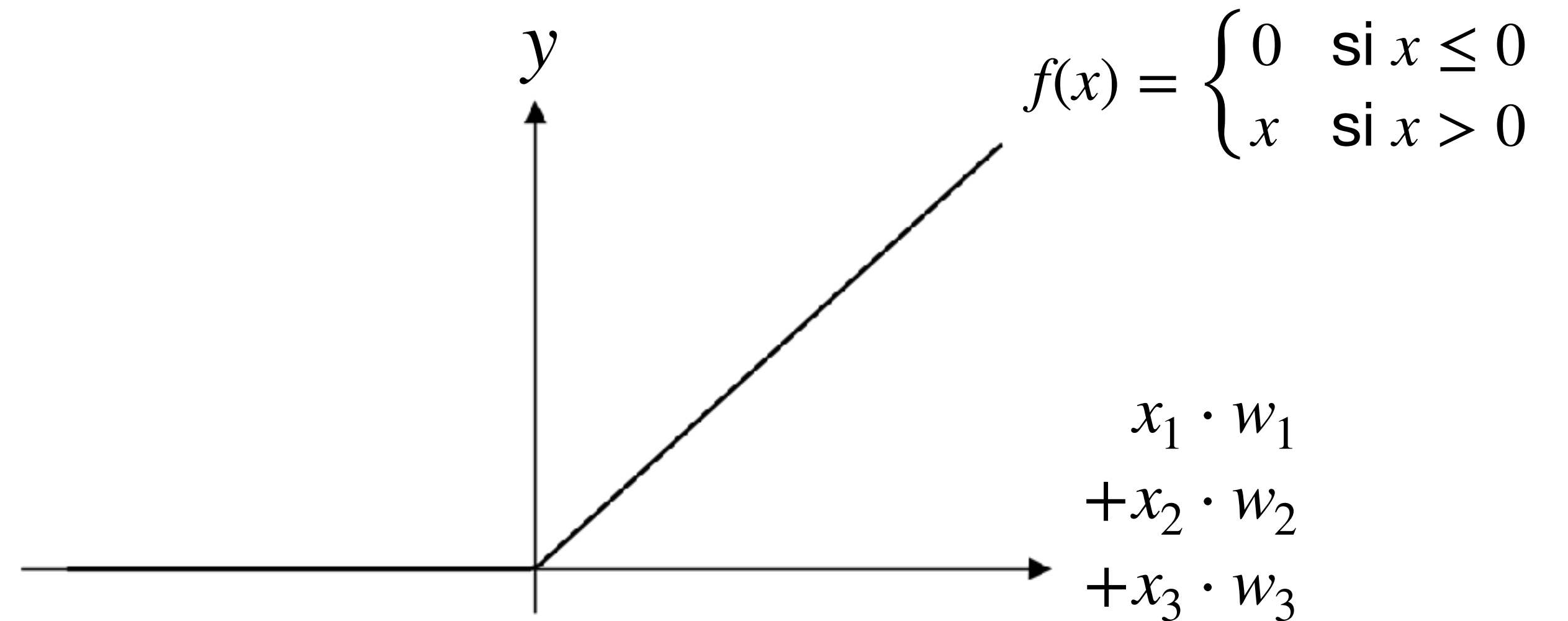
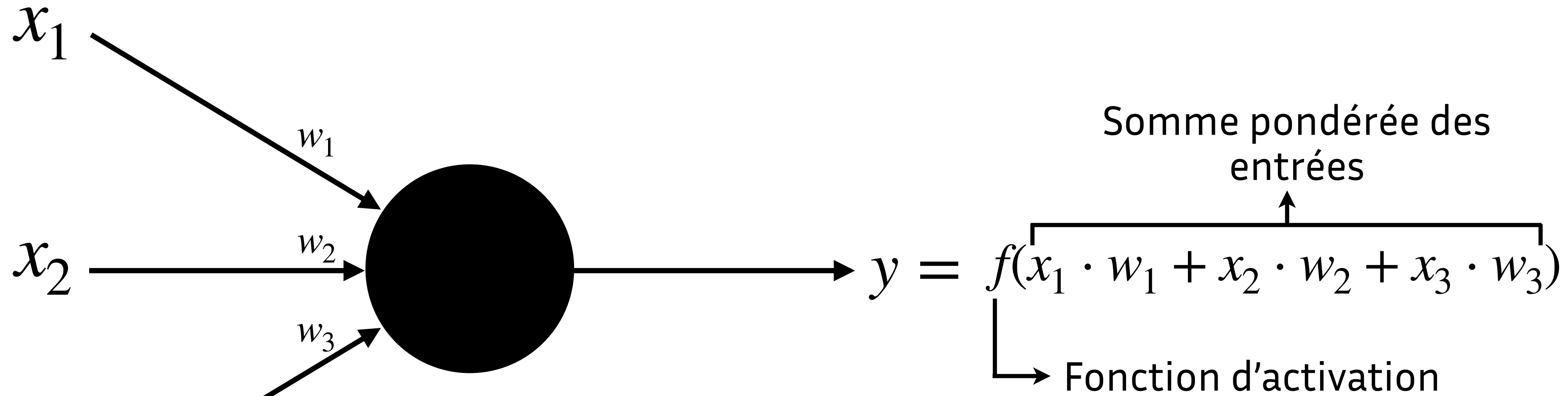
Neurone artificiel



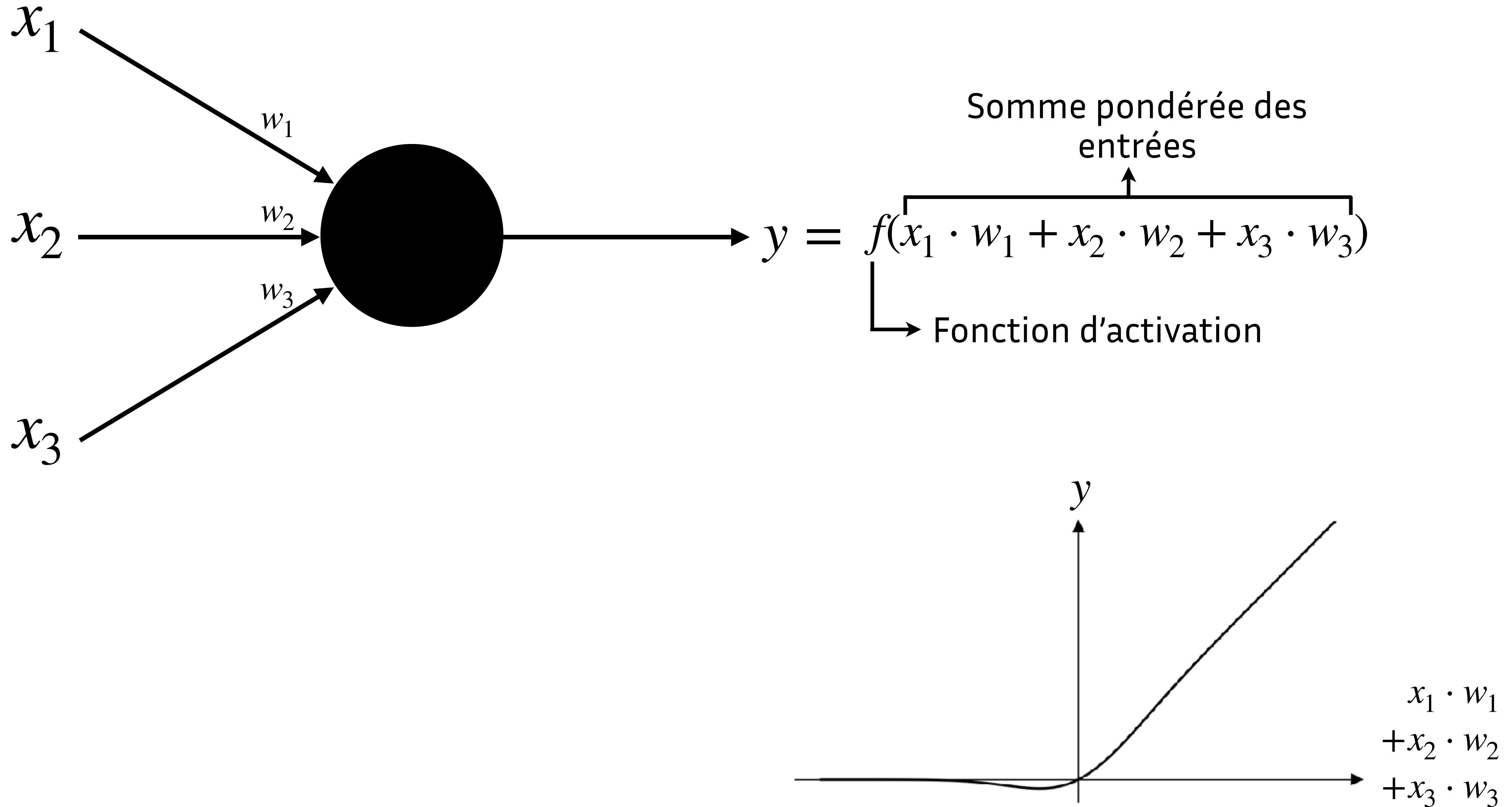
Neurone artificiel



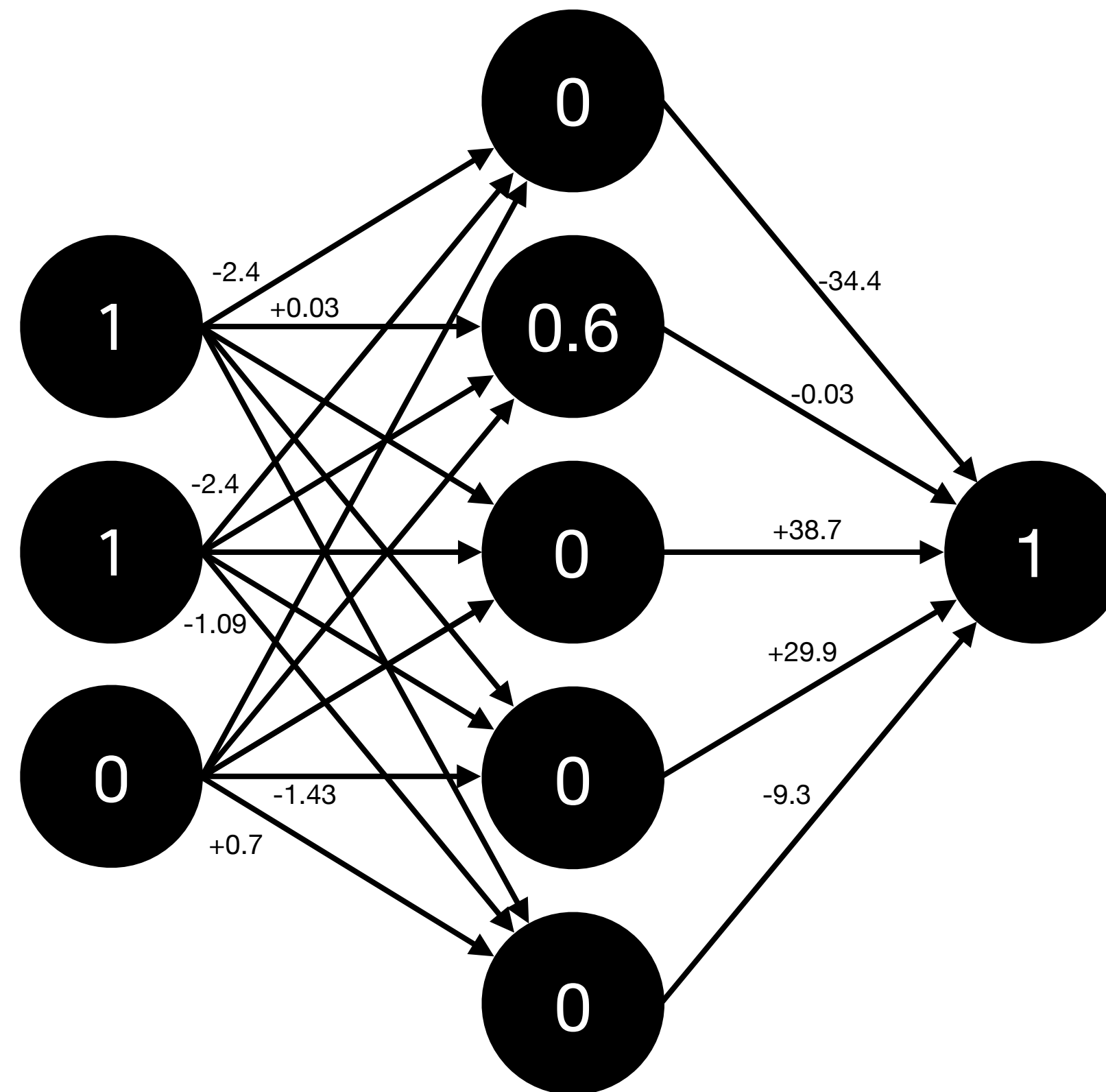
Neurone artificiel



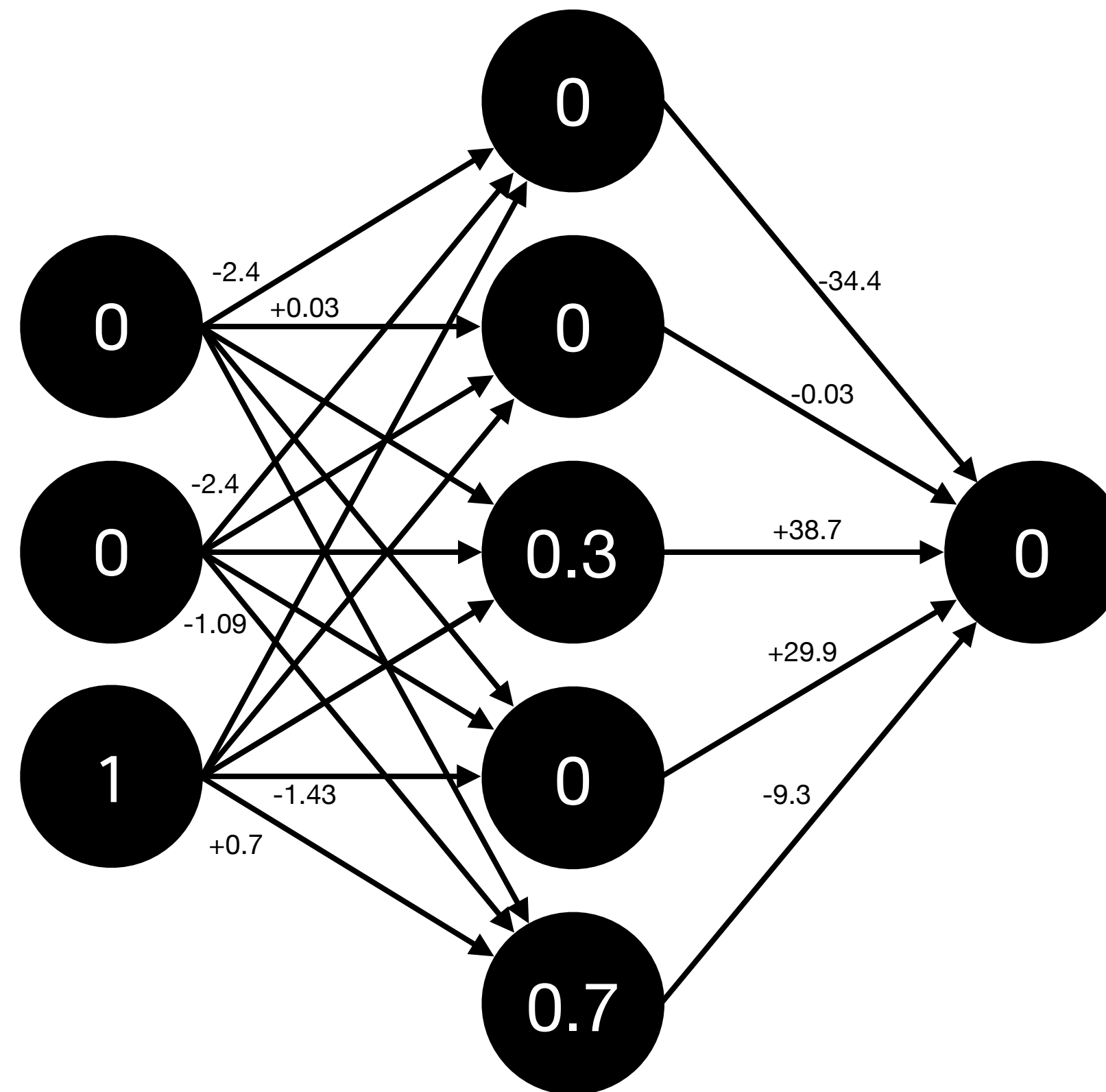
Neurone artificiel



Perceptron multicouche

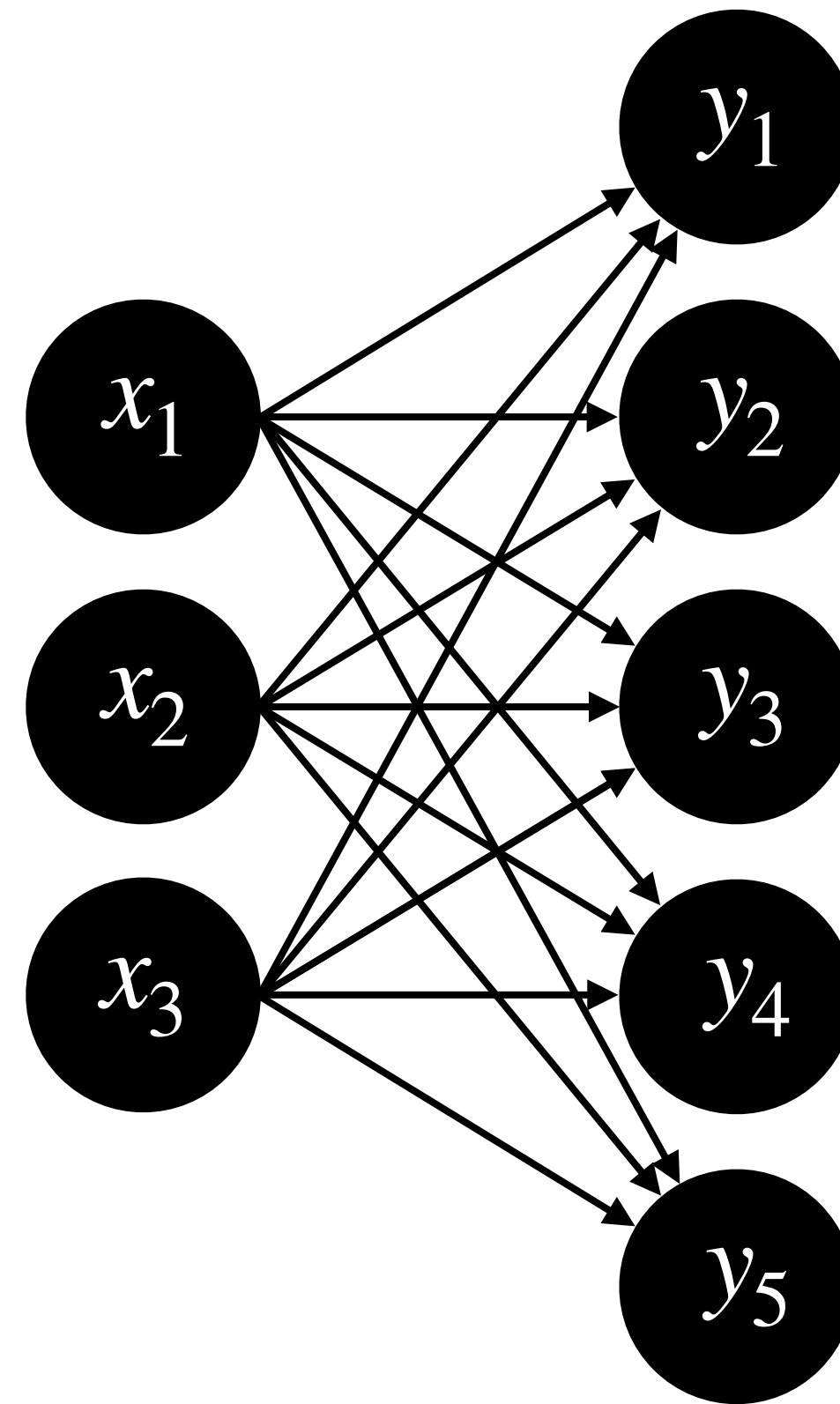


Perceptron multicouche



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

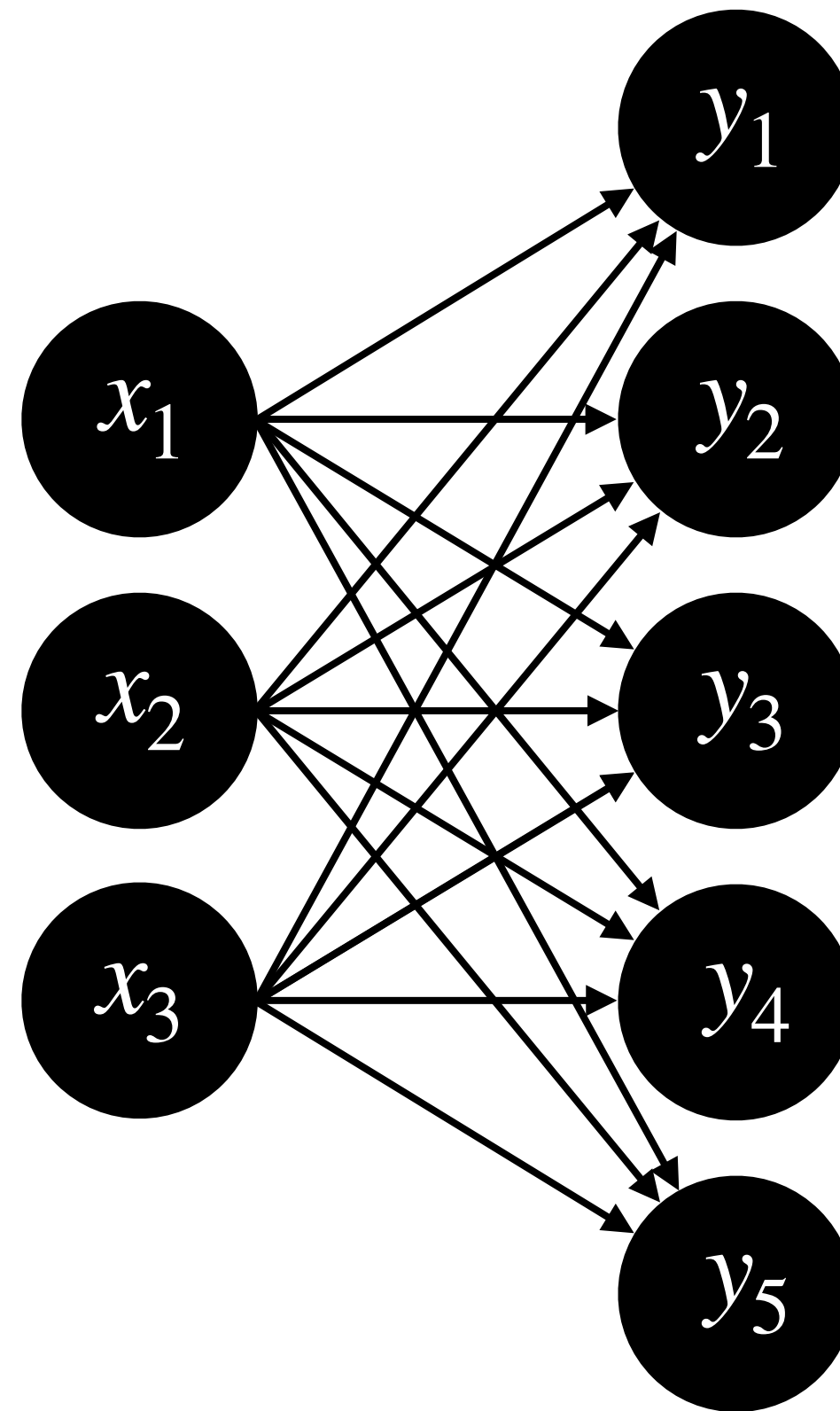
Perceptron multicouche



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

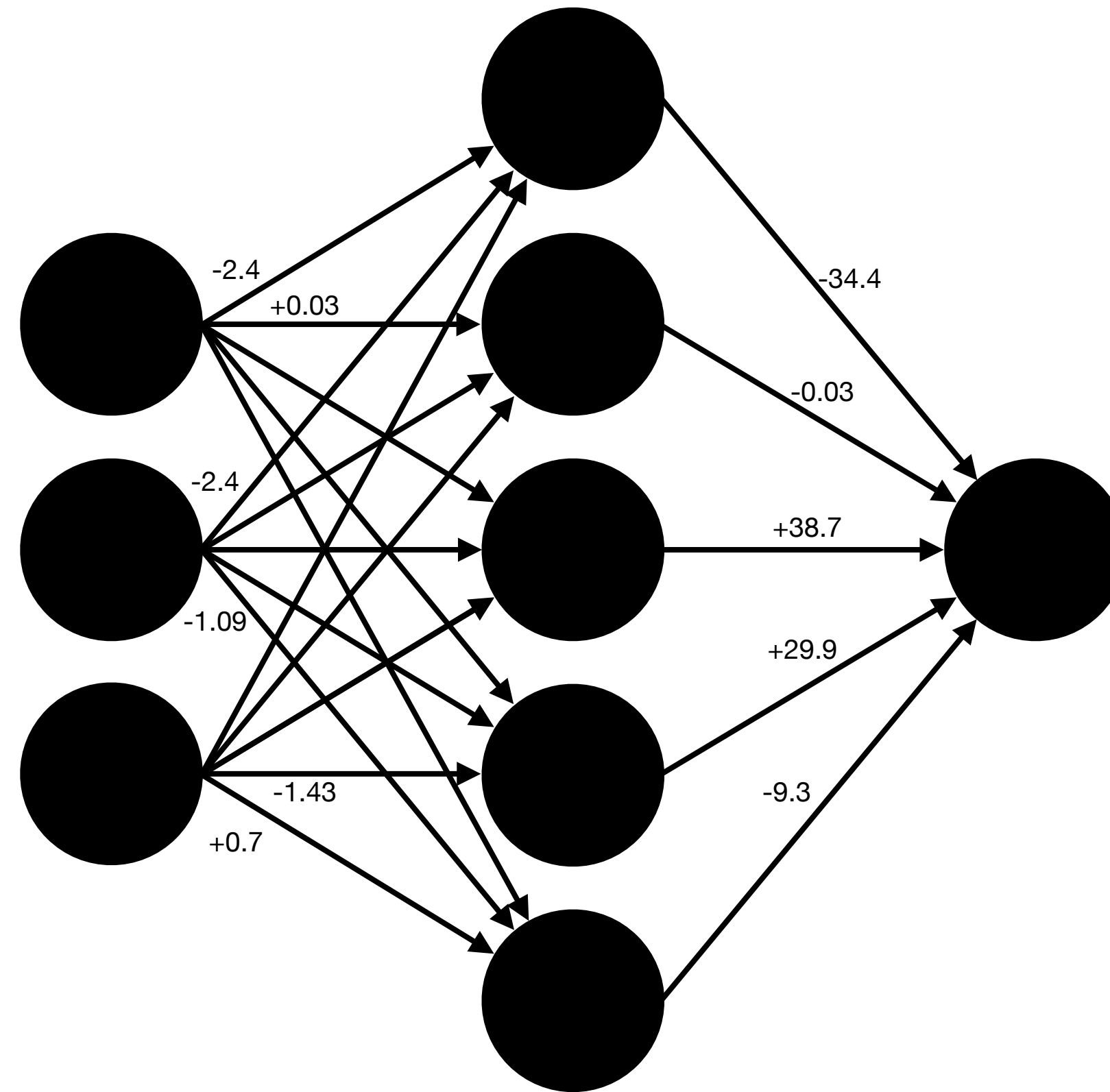
$$[y_1 \ y_2 \ y_3 \ y_4 \ y_5] = f\left([x_1 \ x_2 \ x_3] \times \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} \end{bmatrix}\right)$$

Perceptron multicouche



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

Perceptron multicouche

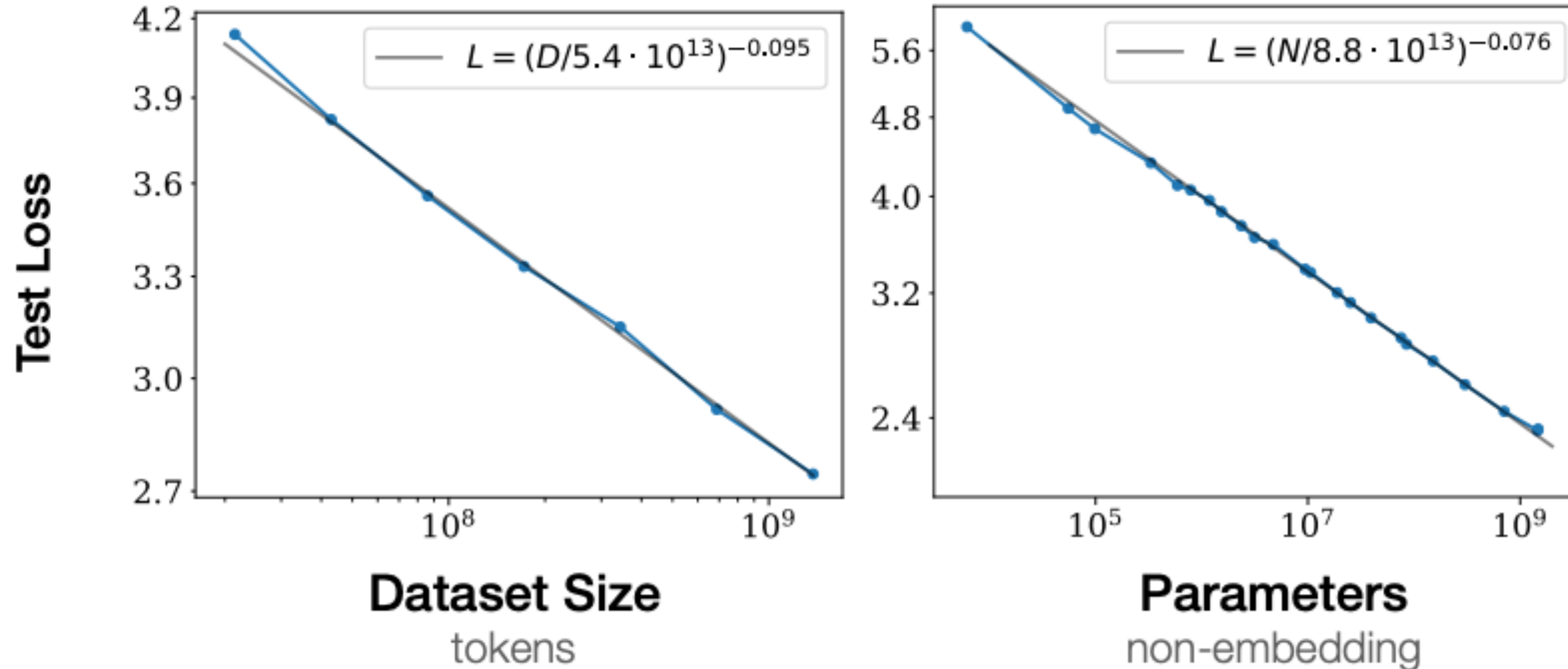


$3 \times 5 + 5 \times 1 = 20$ paramètres
(26 avec les biais)

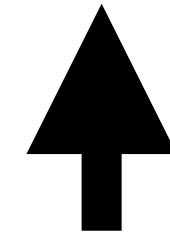
	Nombre de paramètres	Taille des données d'entraînement (en tokens)	Temps d'entraînement (eq. Ordinateur portable)
GPT1	117 millions 1000 livres	600 millions 6000 livres	13 ans
GPT2	1.5 milliard 13 000 livres	28 milliards 280 000 livres	1600 ans
GPT2 - small	124 millions 1000 livres	28 milliards 280 000 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

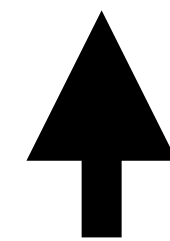
Kaplan et al., 2020



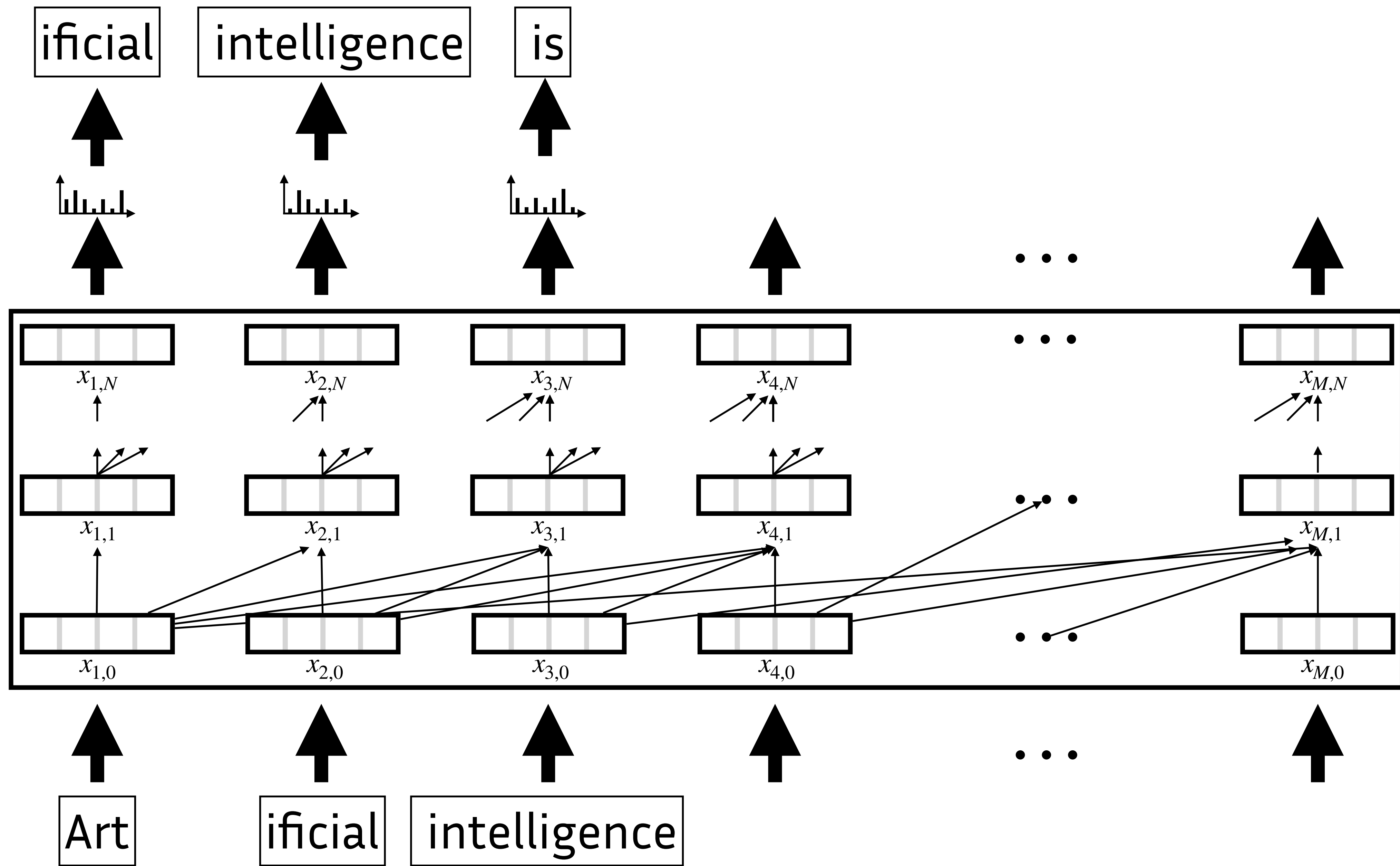
Artificial intelligence is

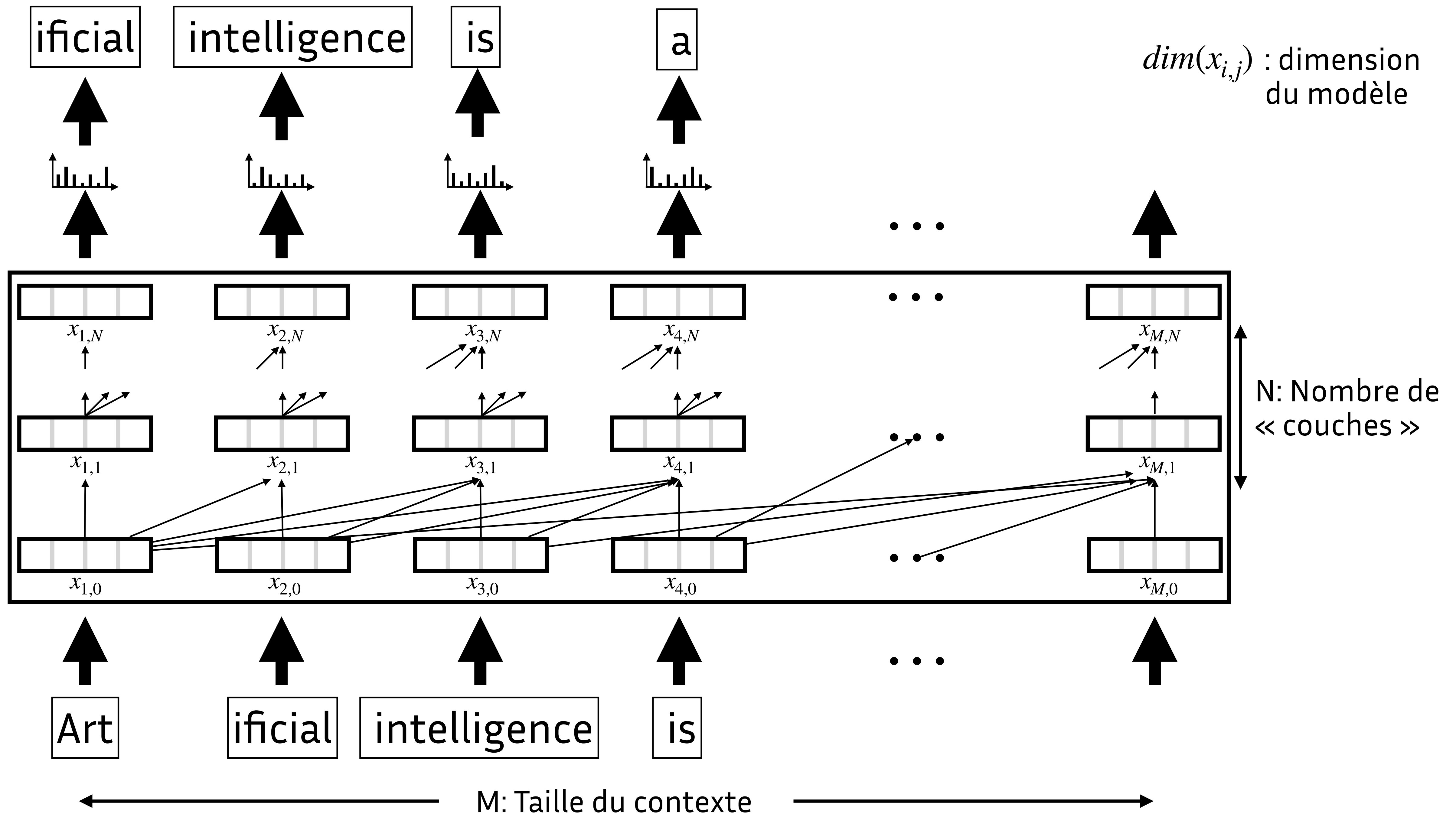


Artificial intelligence



Artificial intelligence



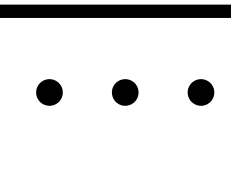
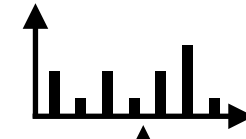
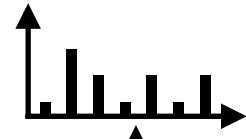
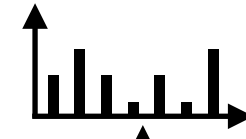
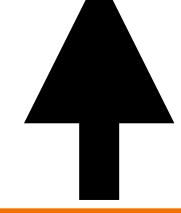
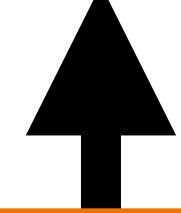
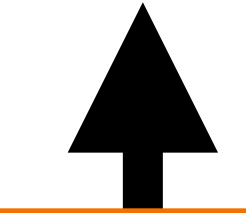


ificial

intelligence

is

Understanding



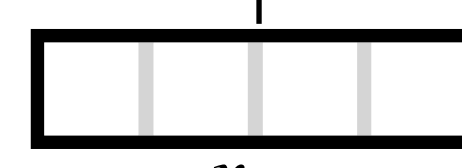
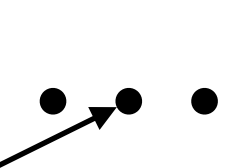
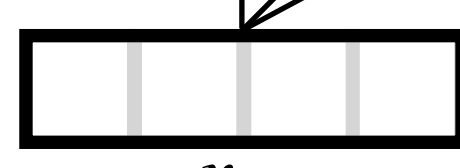
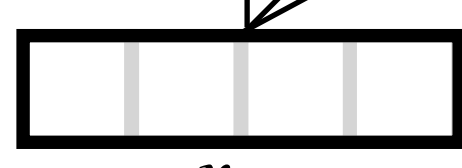
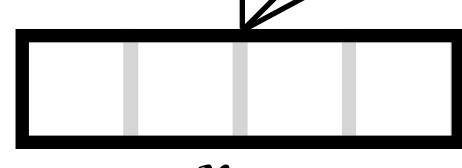
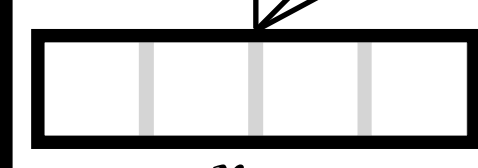
$x_{1,N}$

$x_{2,N}$

$x_{3,N}$

$x_{4,N}$

$x_{M,N}$



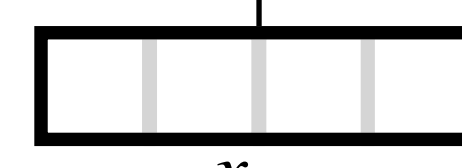
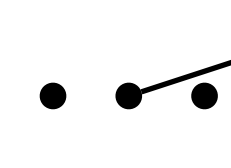
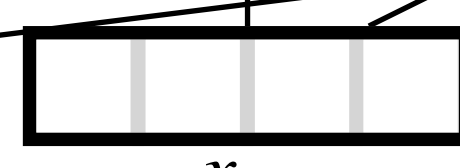
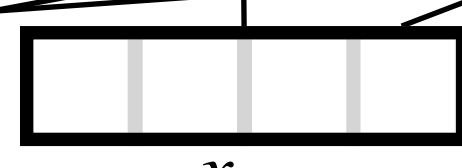
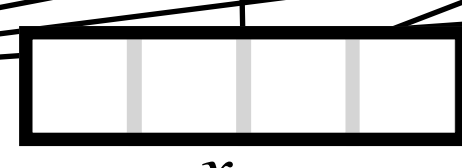
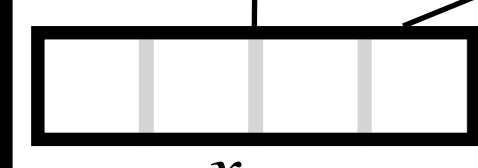
$x_{1,1}$

$x_{2,1}$

$x_{3,1}$

$x_{4,1}$

$x_{M,1}$



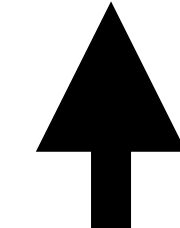
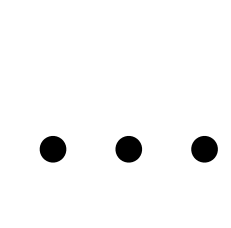
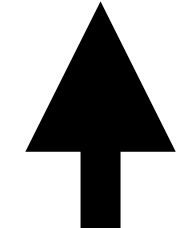
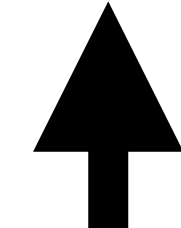
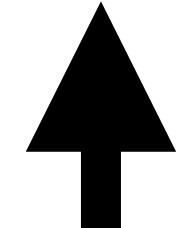
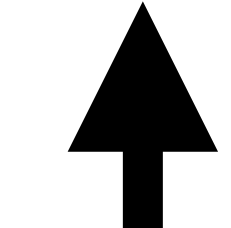
$x_{1,0}$

$x_{2,0}$

$x_{3,0}$

$x_{4,0}$

$x_{M,0}$

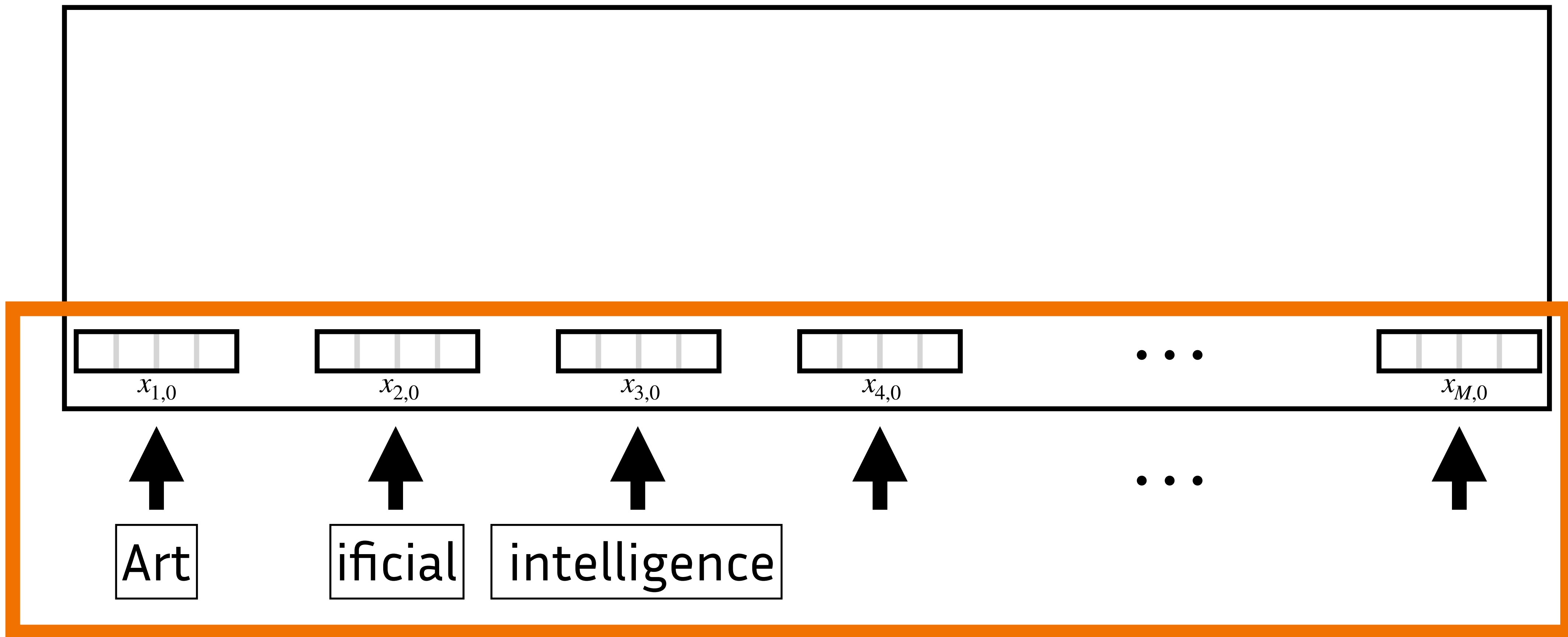


Art

ificial

intelligence

Embedding

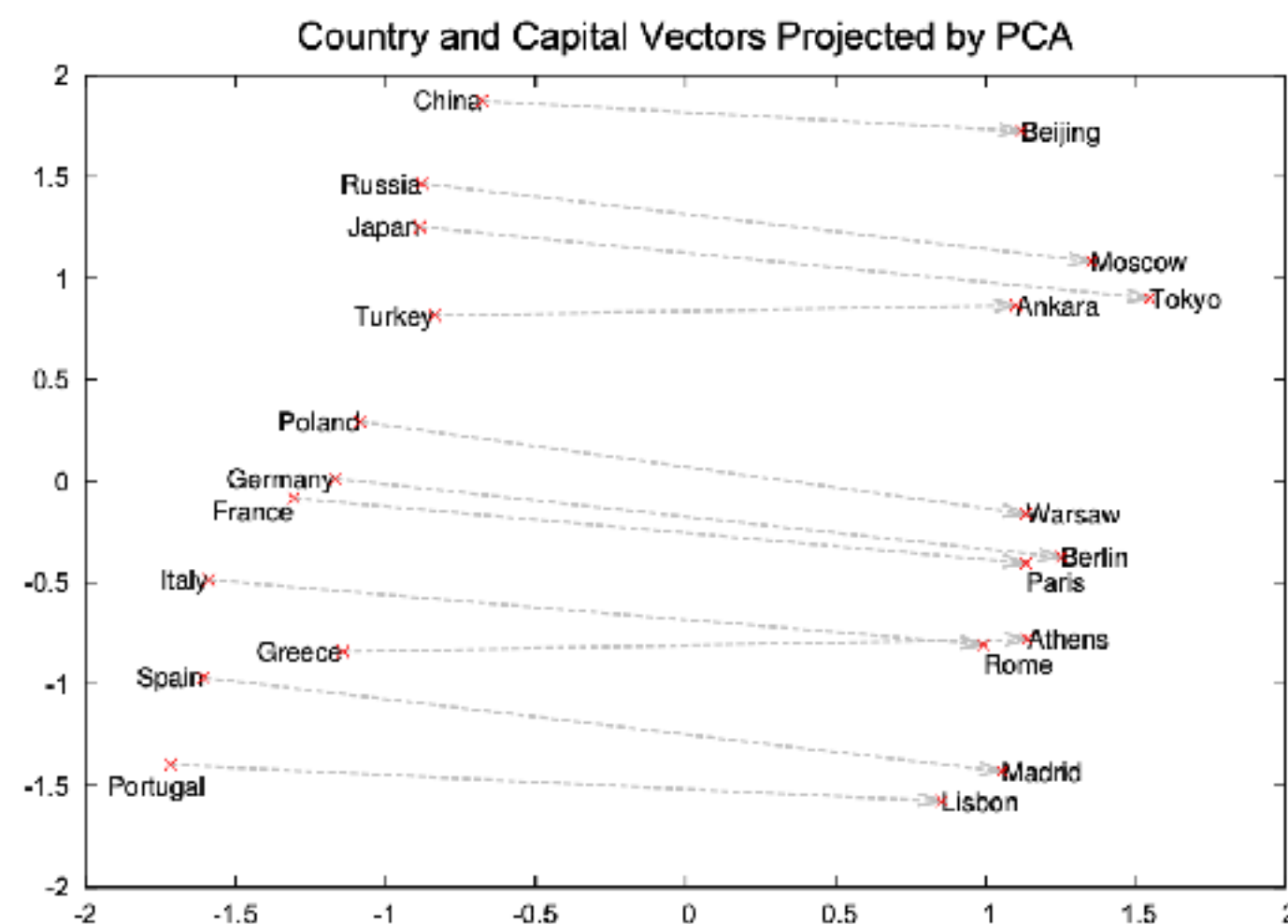


Embedding

1 mot \longleftrightarrow 1 vecteur

espace sémantique \longleftrightarrow espace vectoriel

relations sémantiques \longleftrightarrow relations géométriques

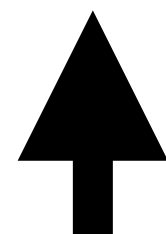


Mikolov et al., 2013

Embedding



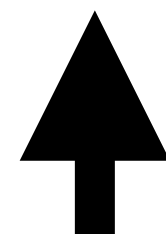
e_{Art}



Art



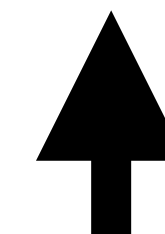
e_{ifical}



ifical

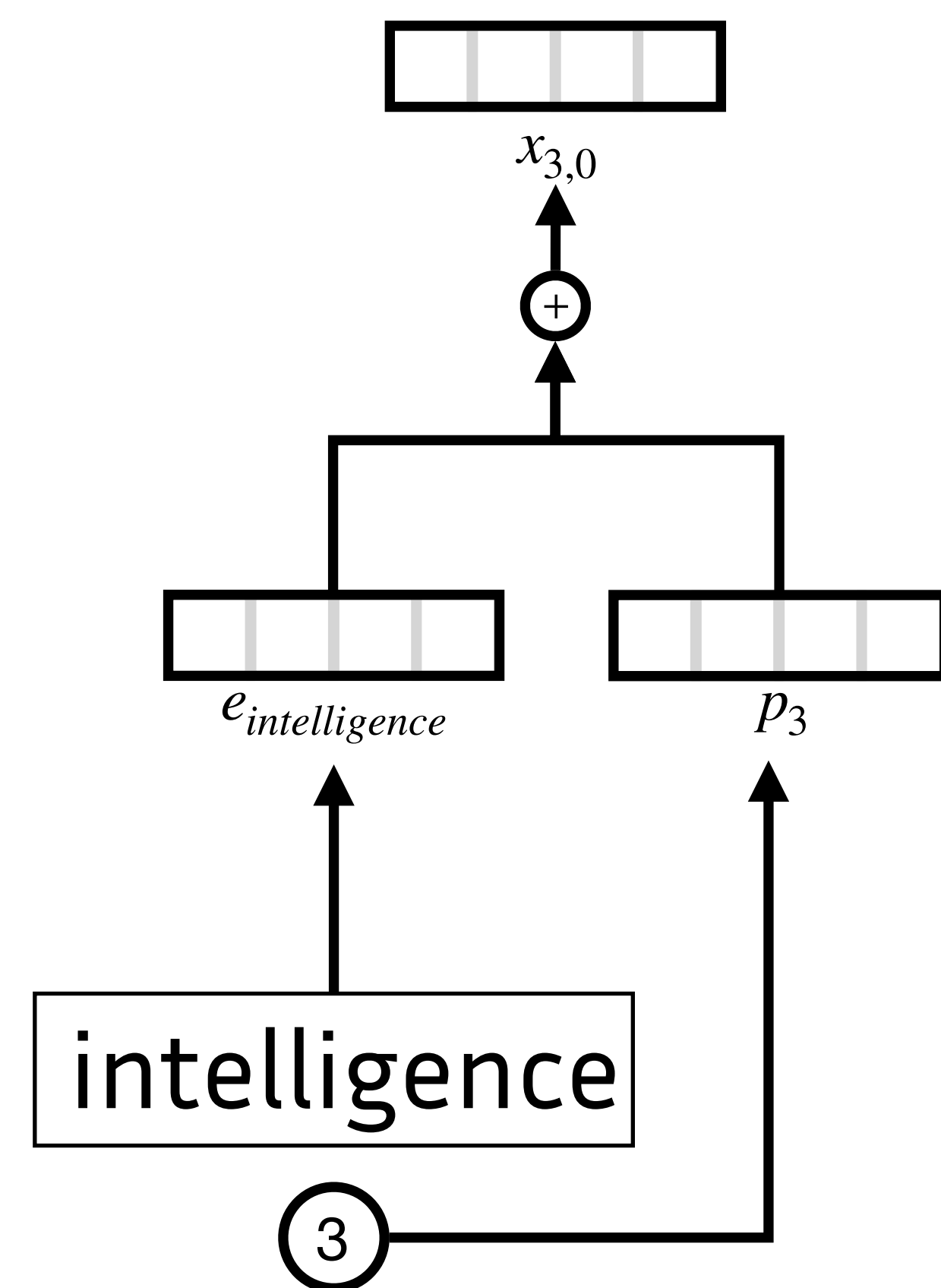
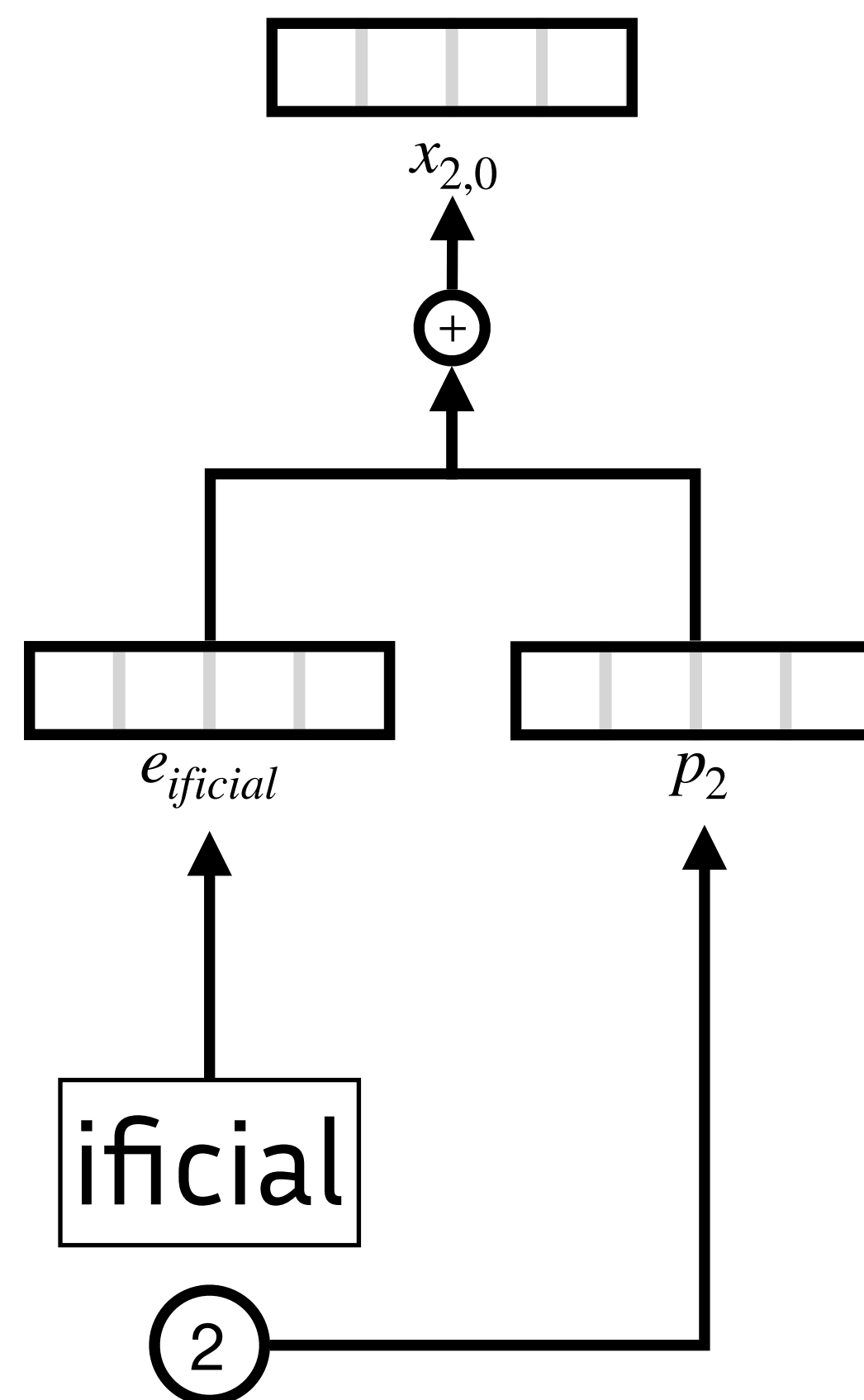
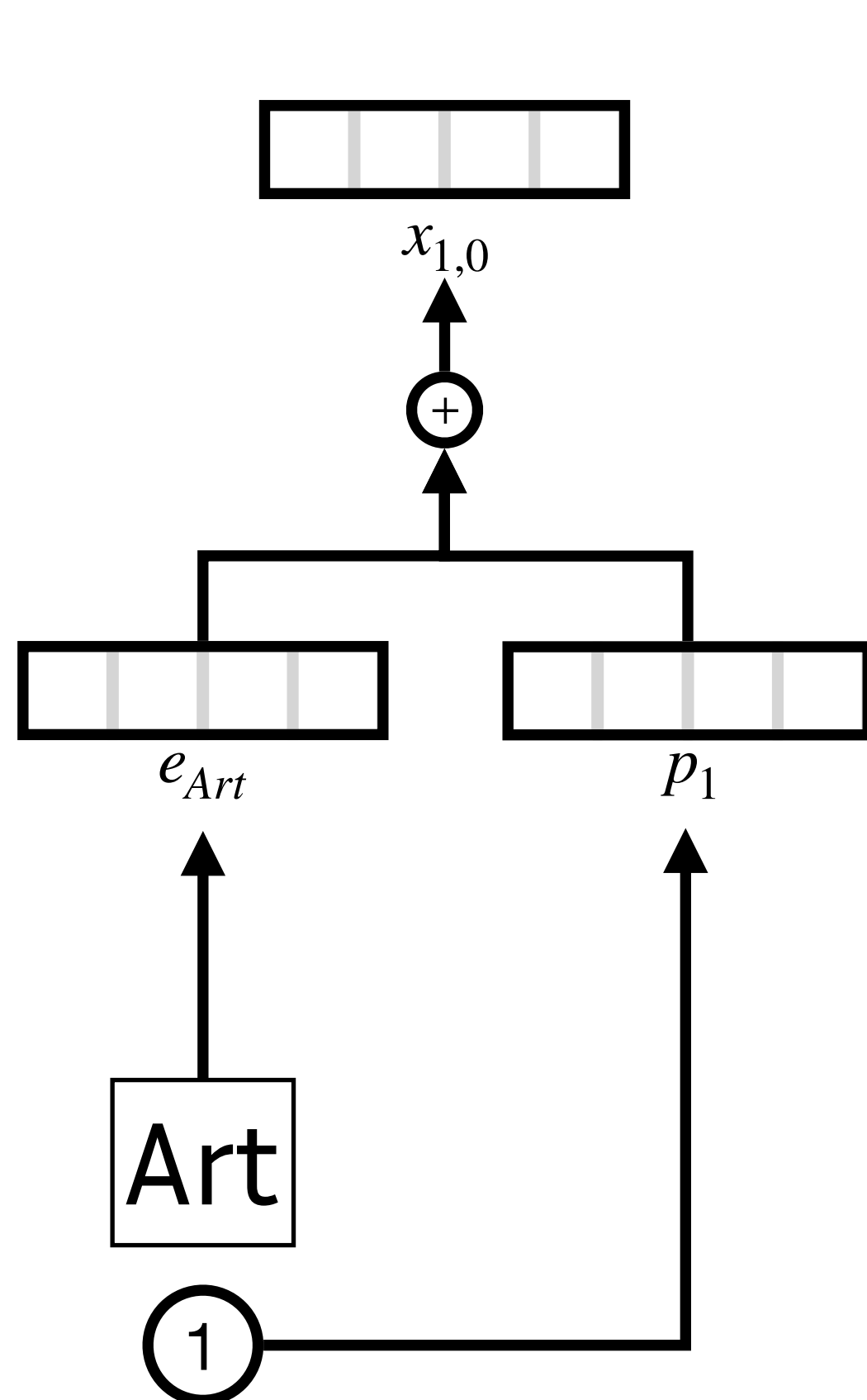


$e_{intelligence}$

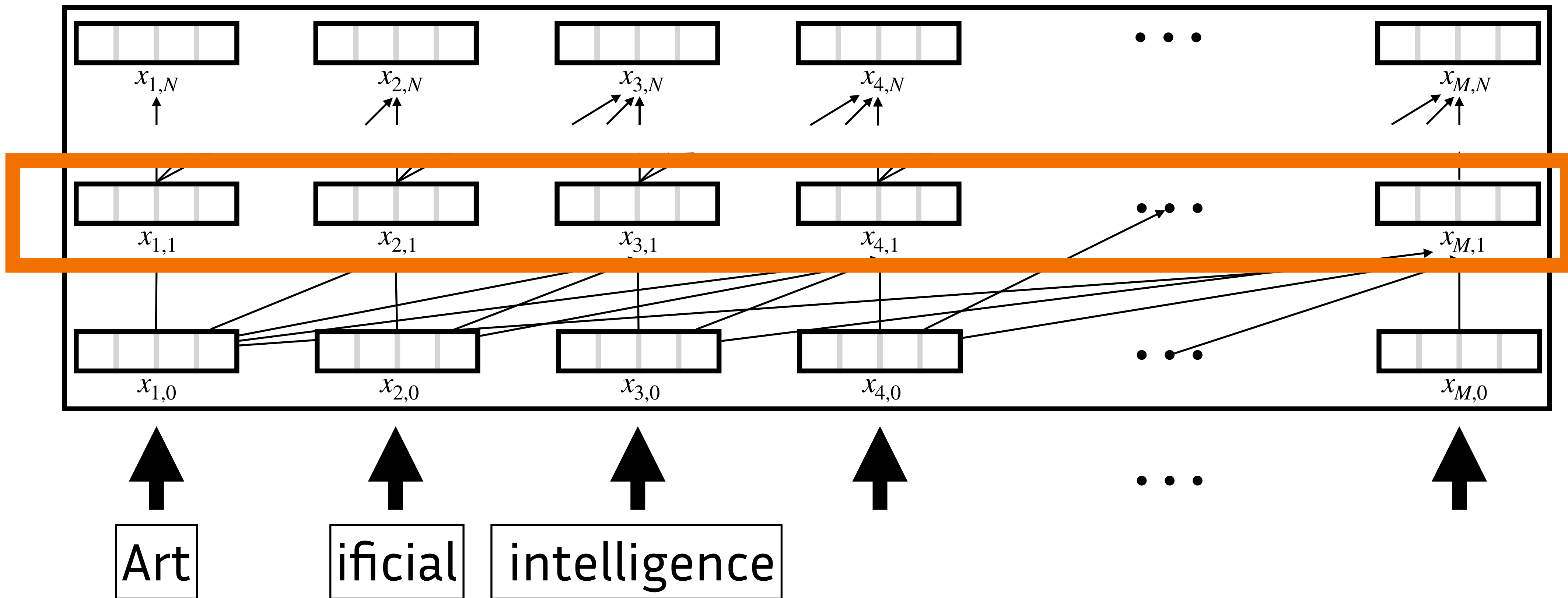


intelligence

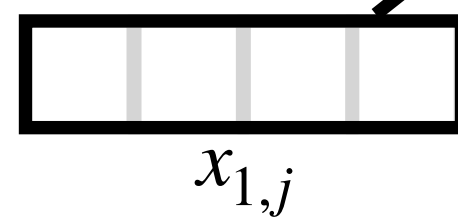
Position encoding



Transformer

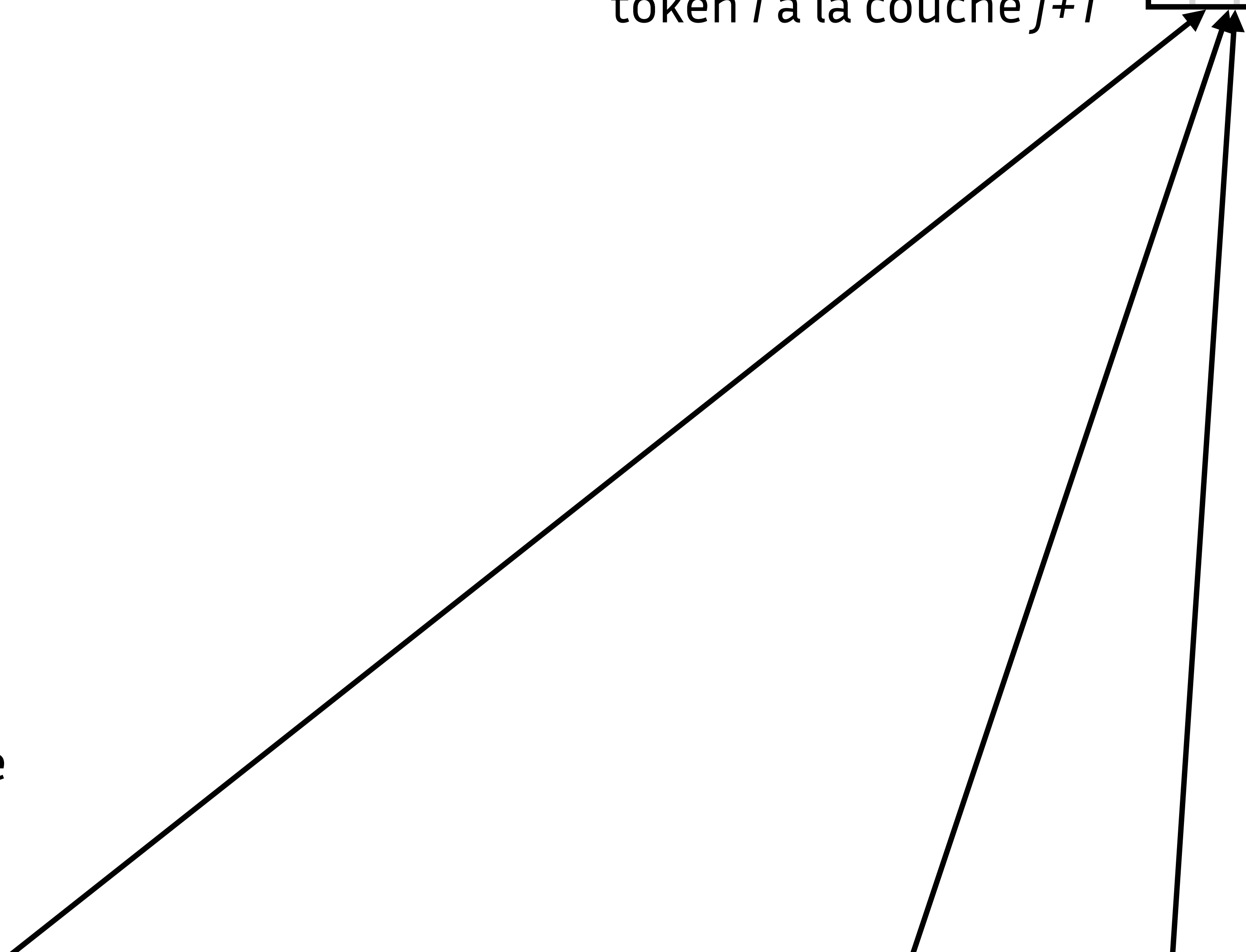
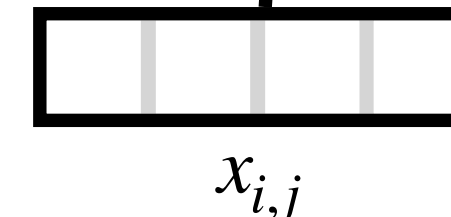
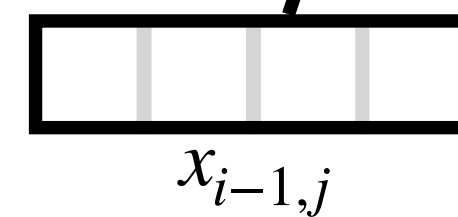
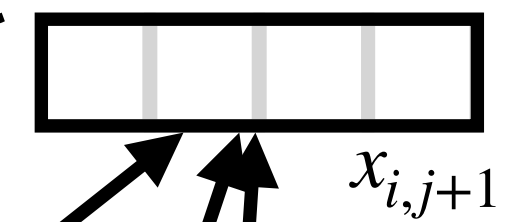


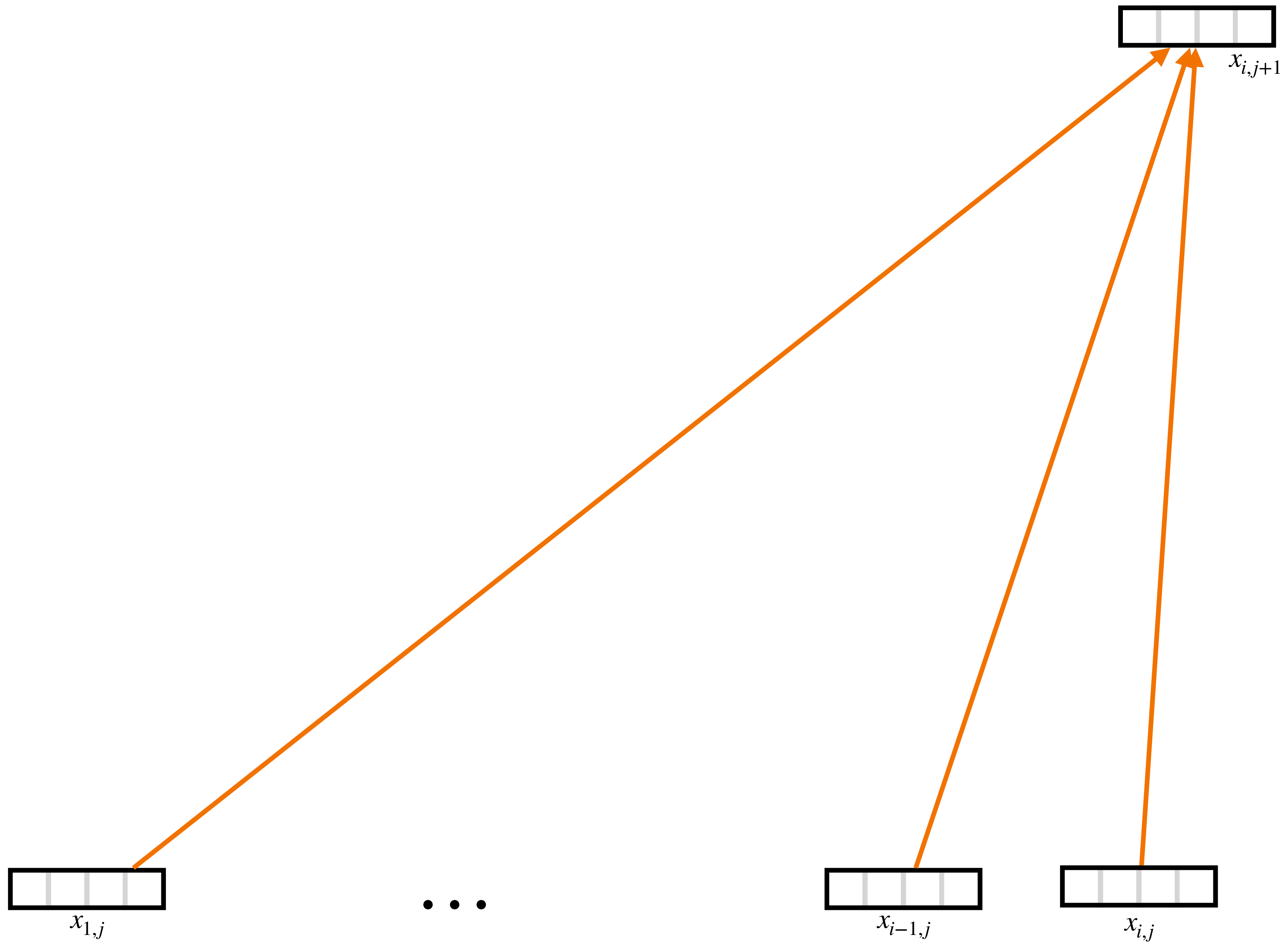
Vecteur
représentant le
token 1 à la
couche j



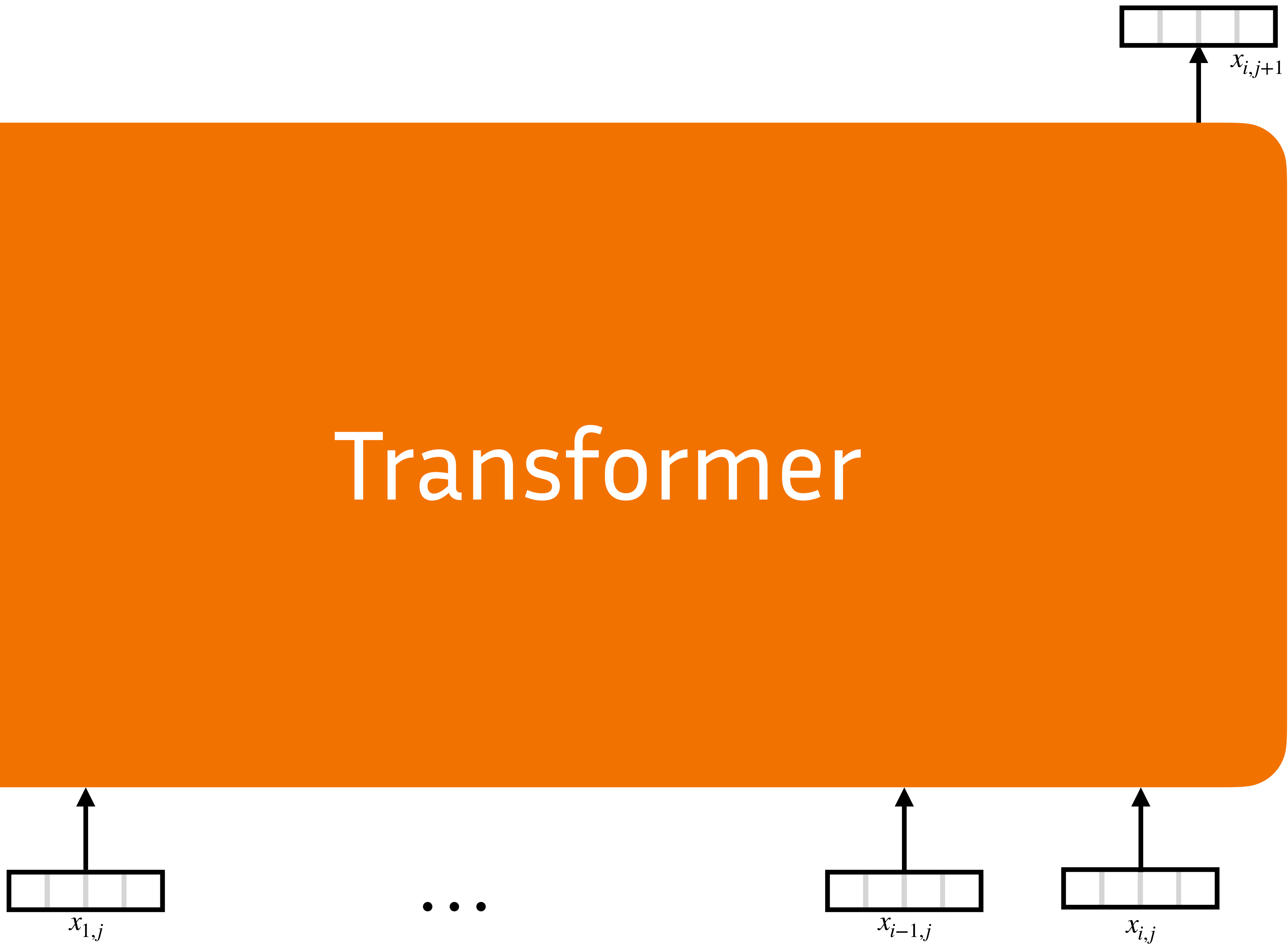
...

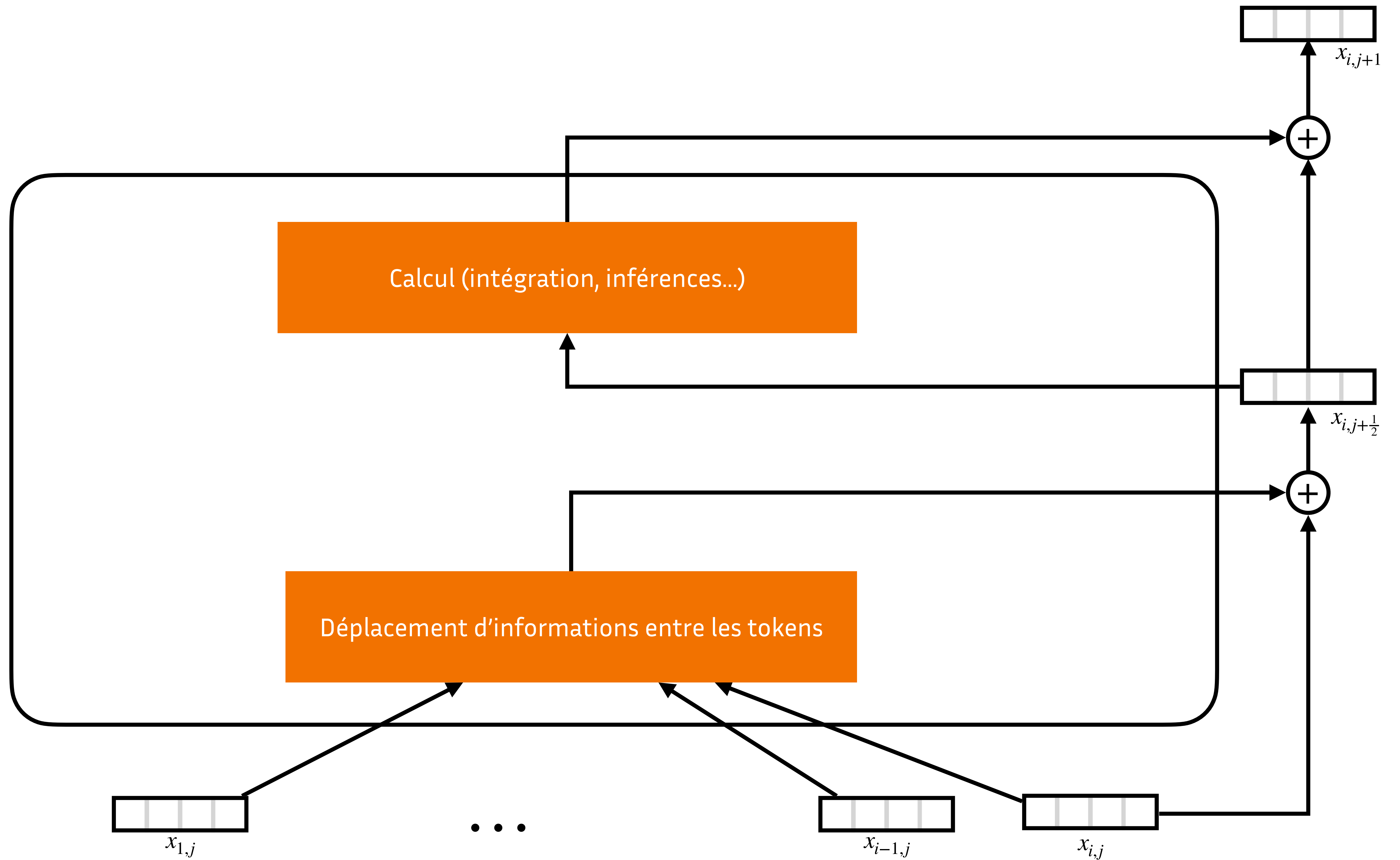
Vecteur représentant le
token i à la couche $j+1$

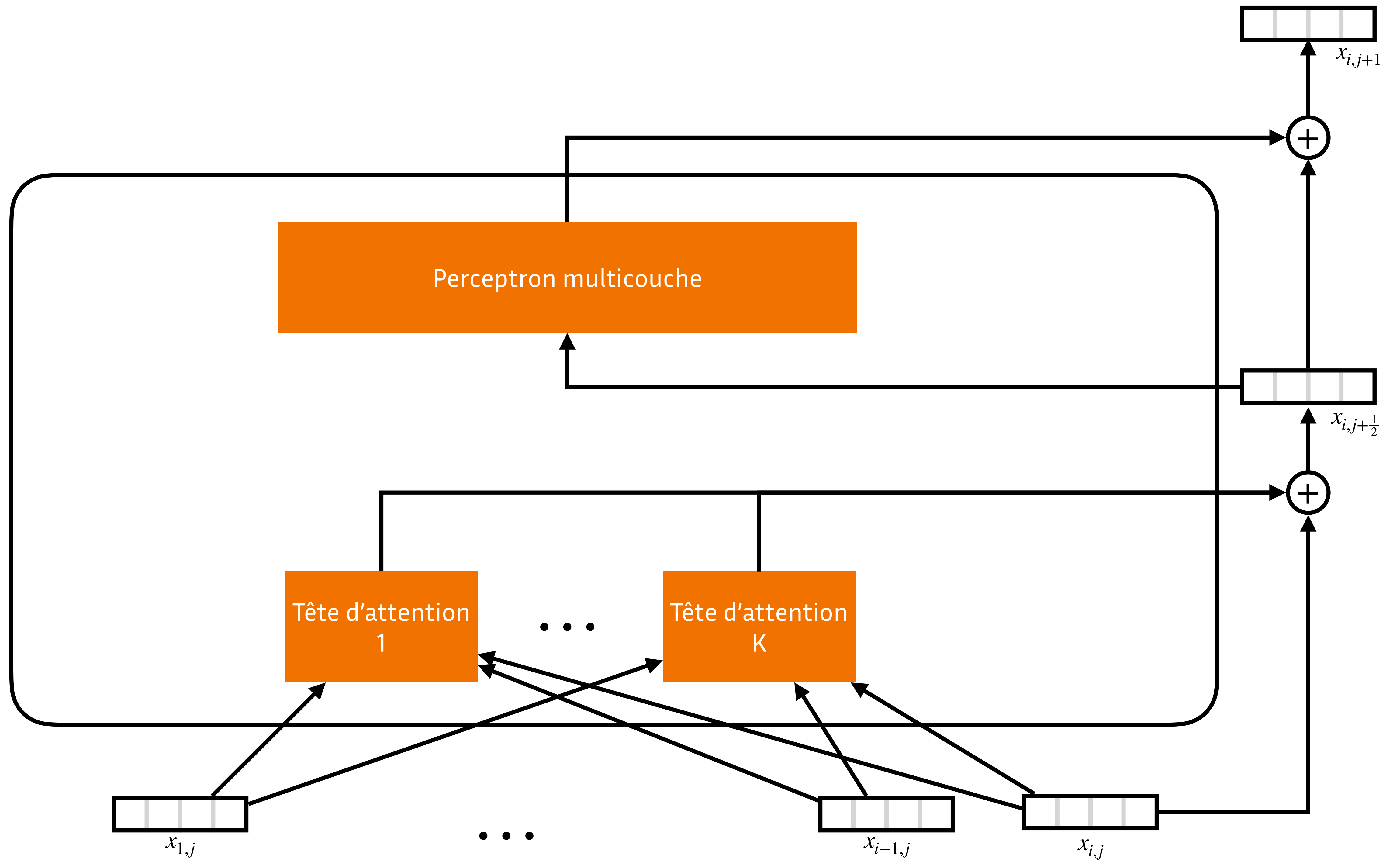


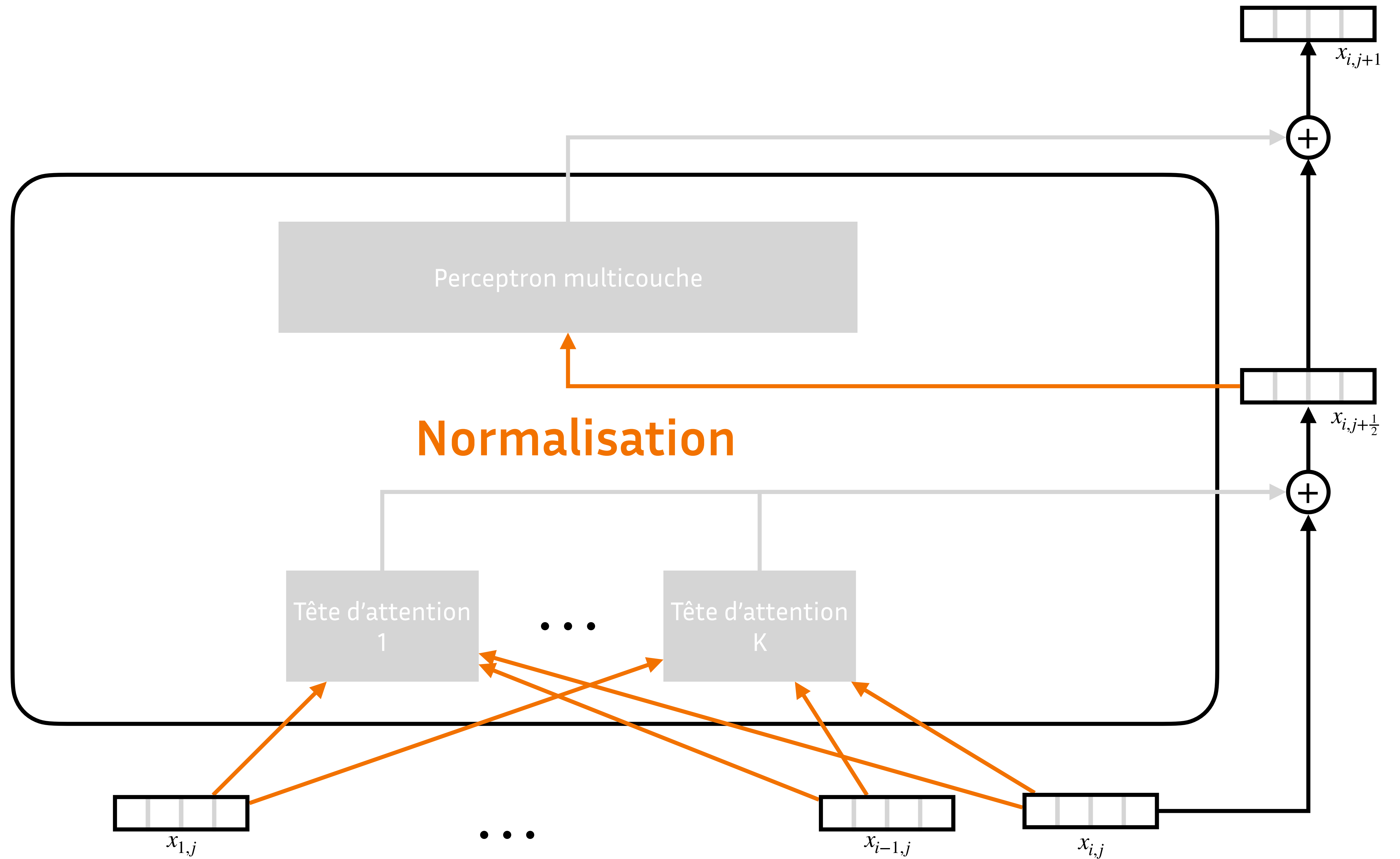


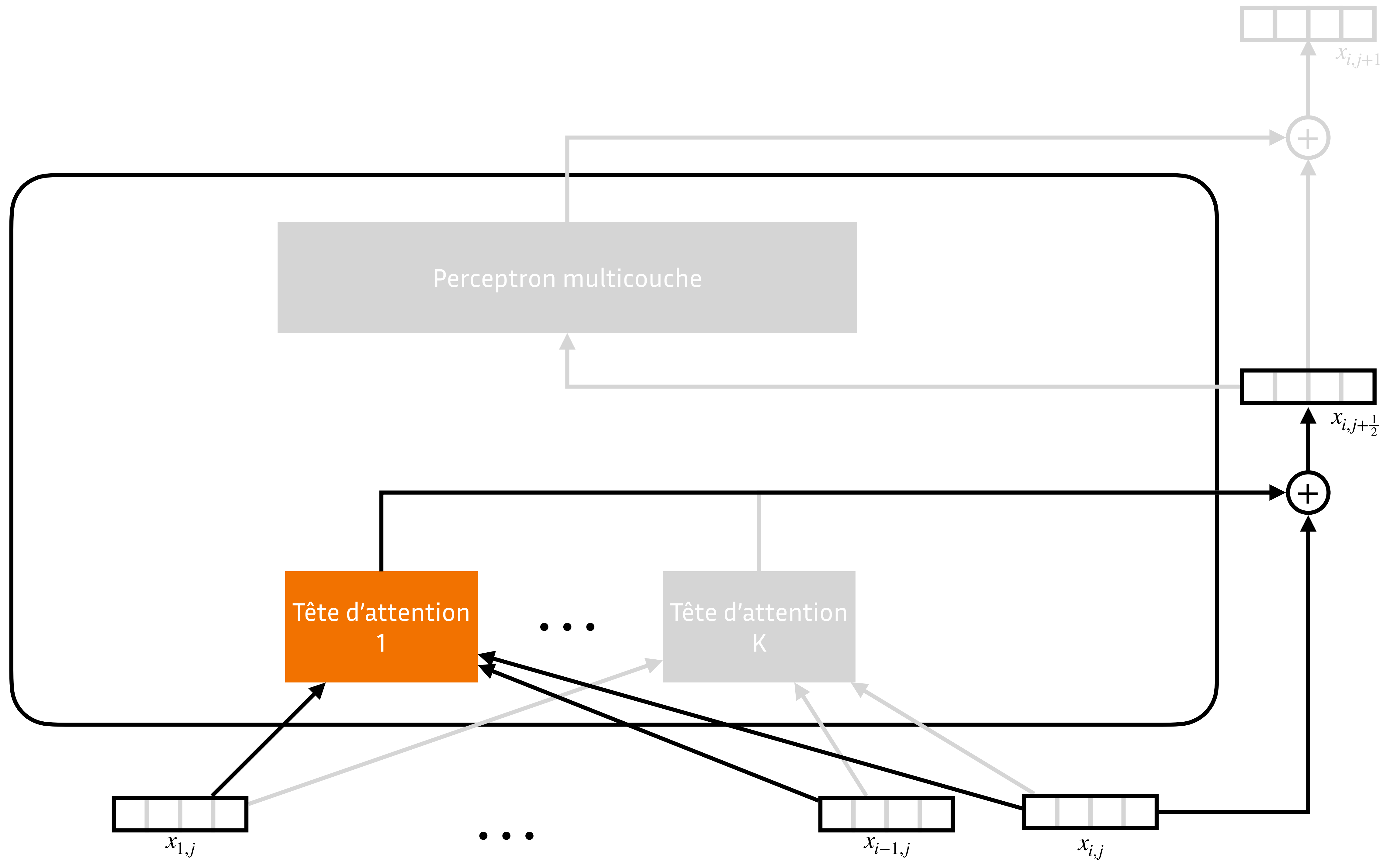
Transformer









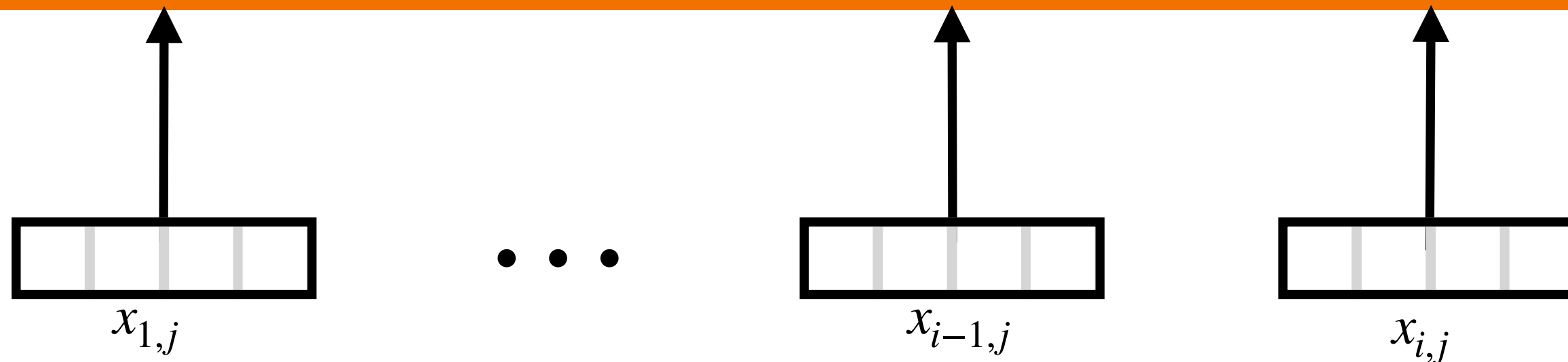


Tête d'attention

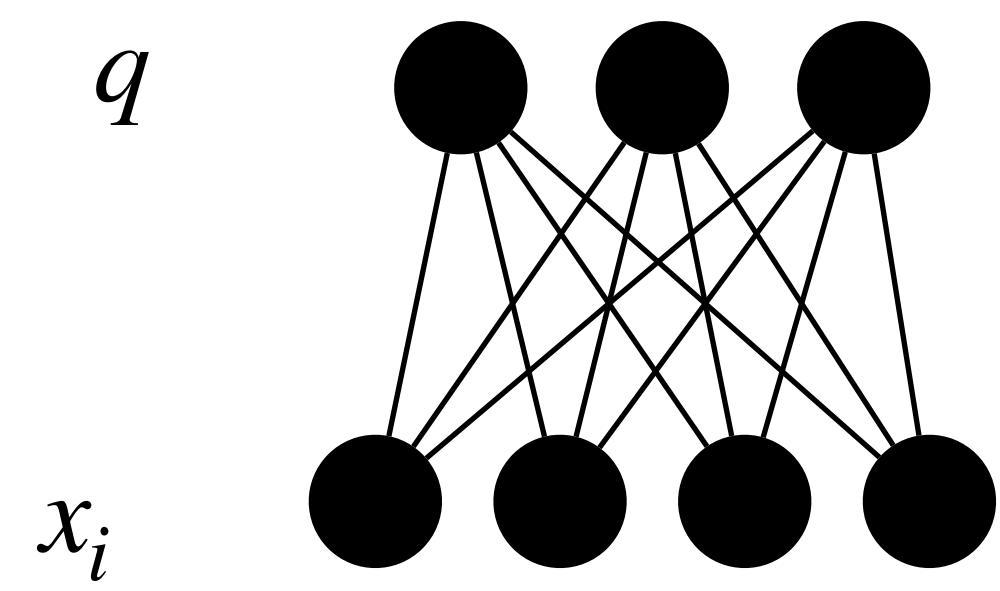
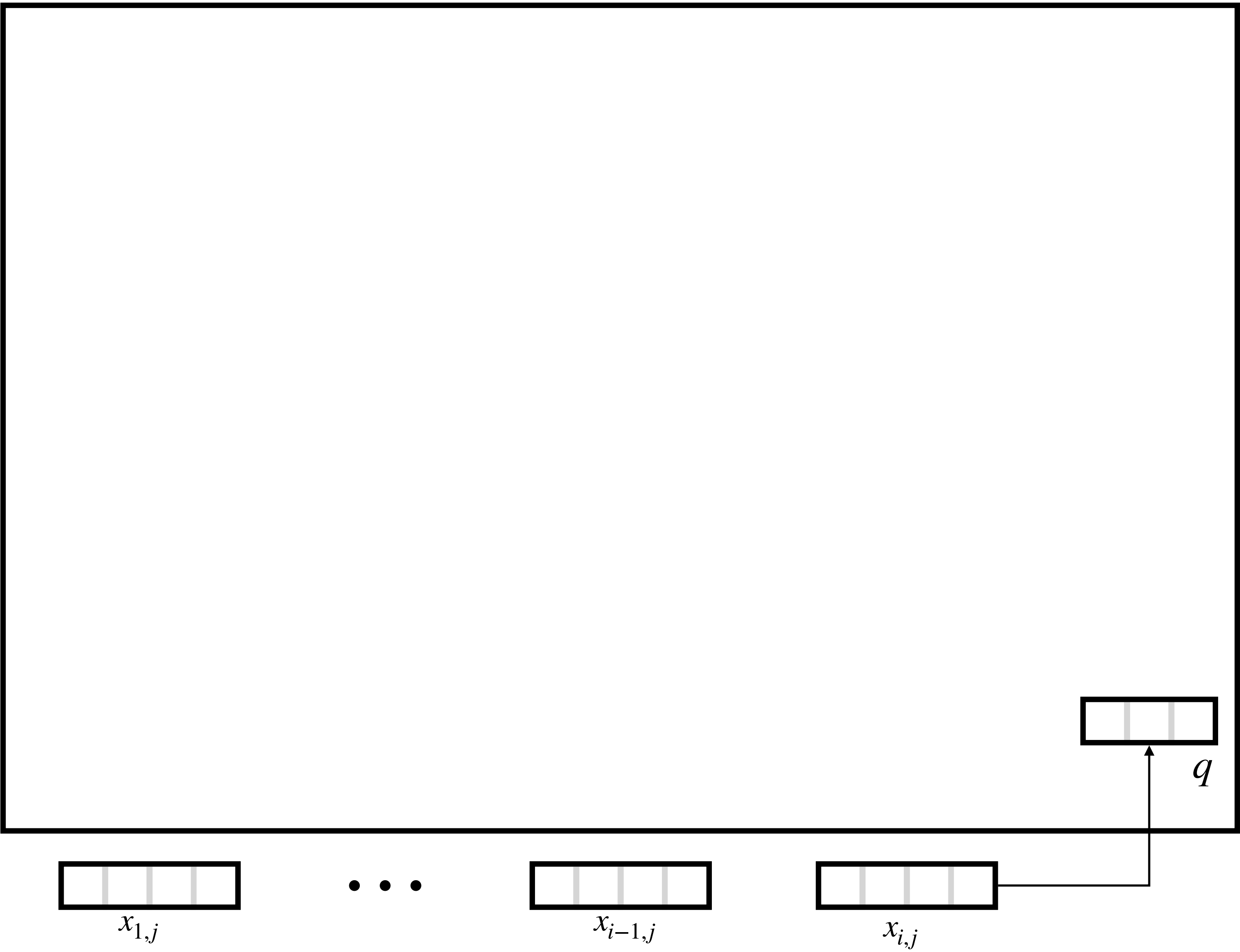
1 : identifier quel type d'information est recherché

2 : identifier quels tokens disposent de cette information

3 : récupérer cette information dans ces tokens



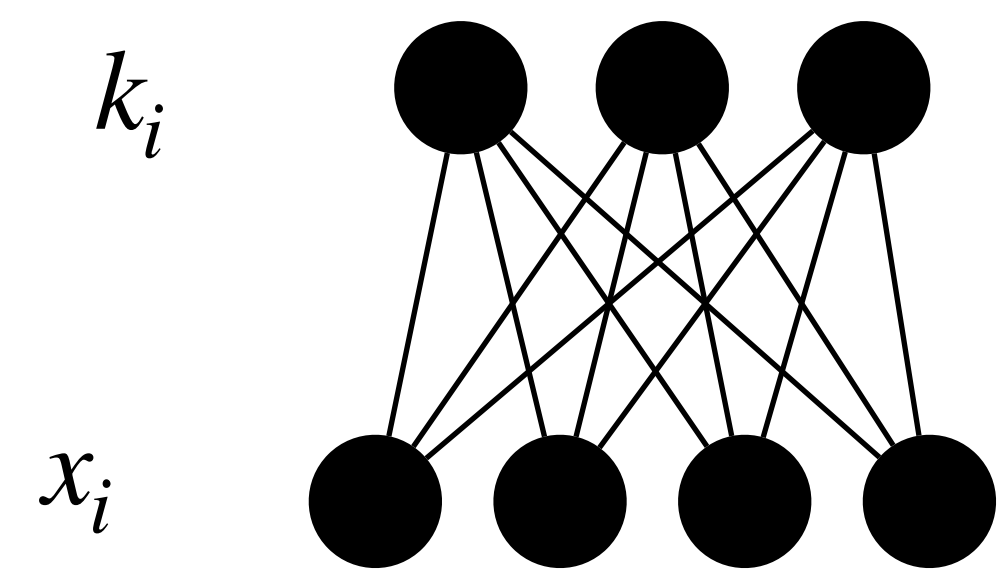
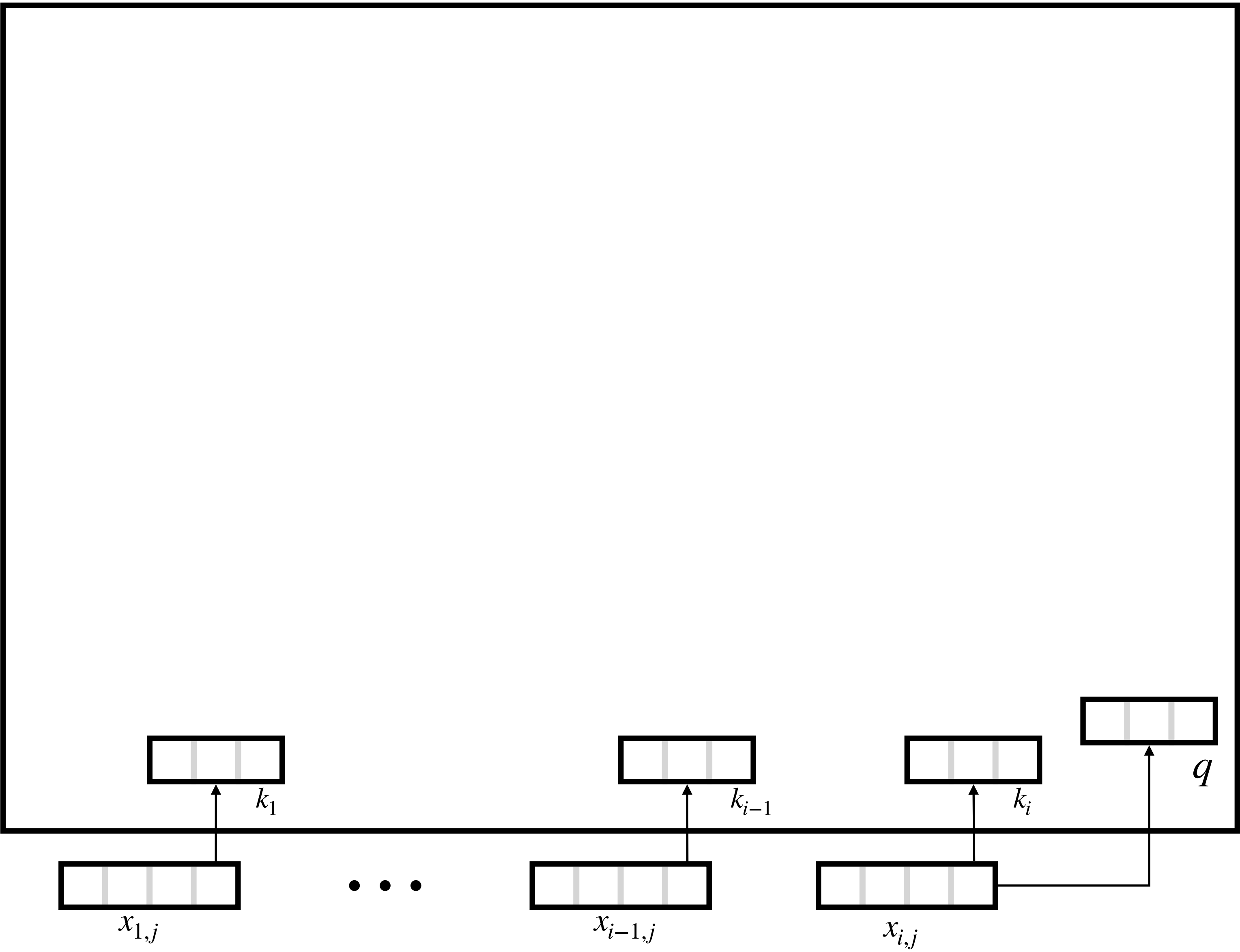
1 : identifier quel type d'information est recherché : *query*



$$q = x_i W_q$$

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

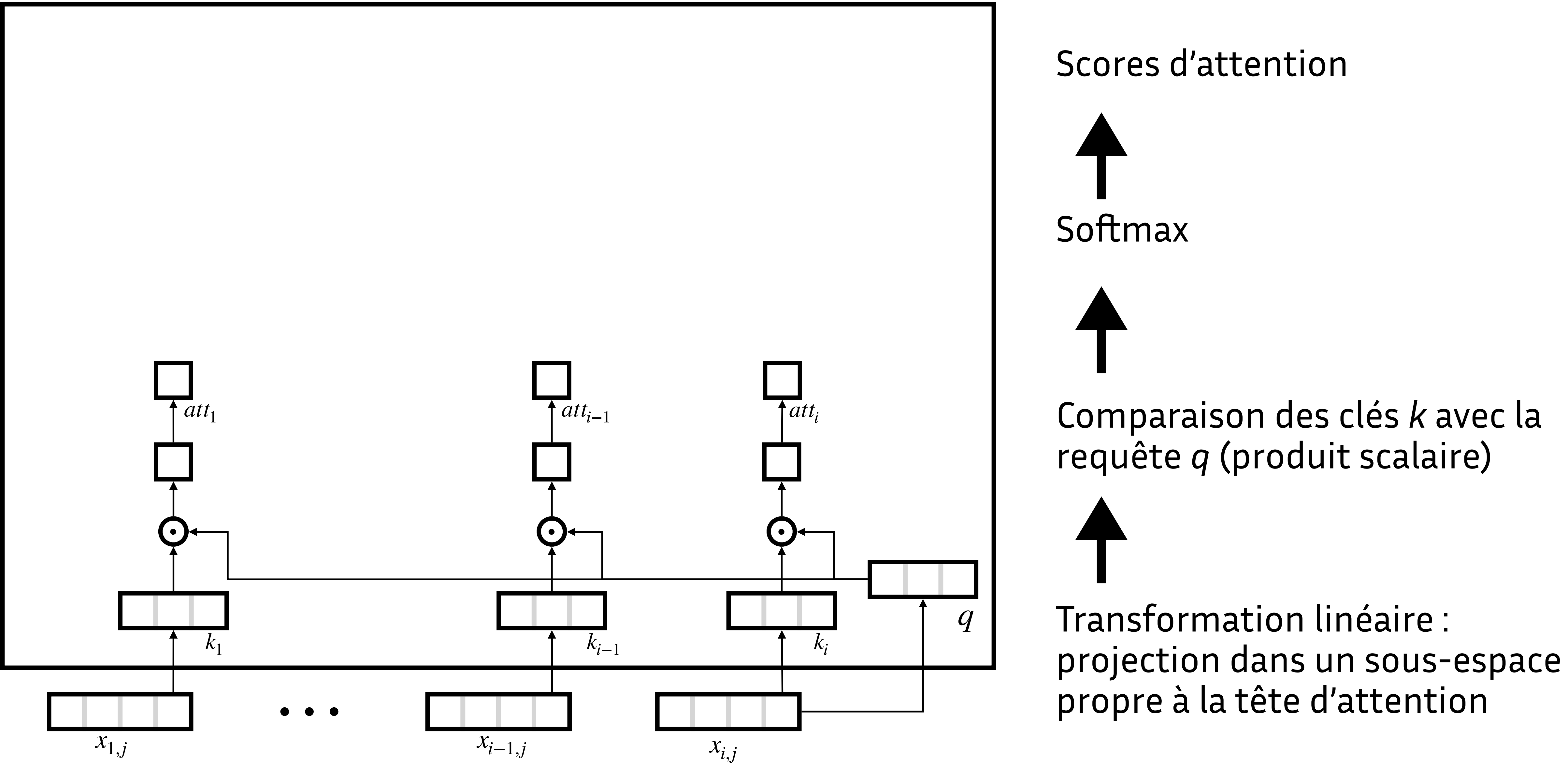
2 : identifier quels tokens disposent de cette information : *keys*



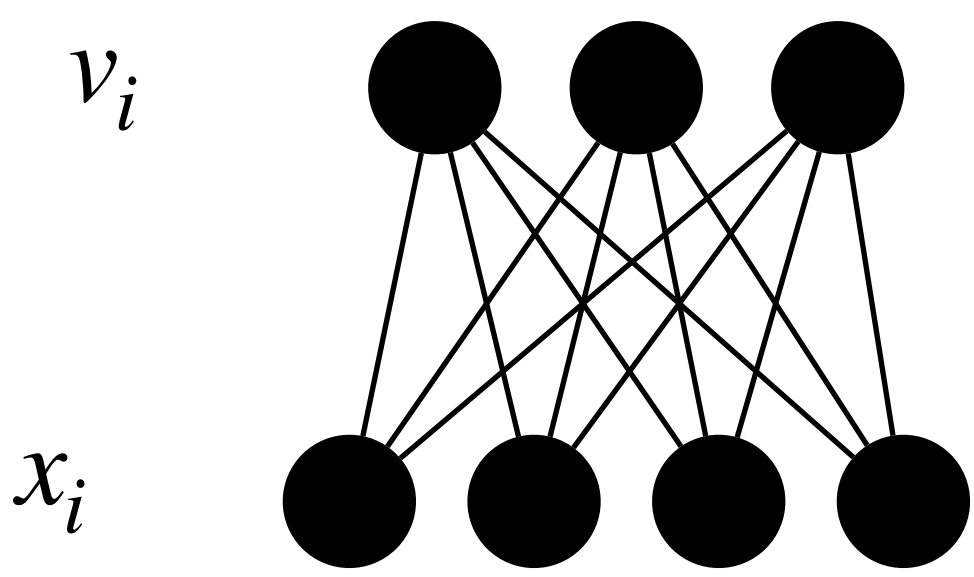
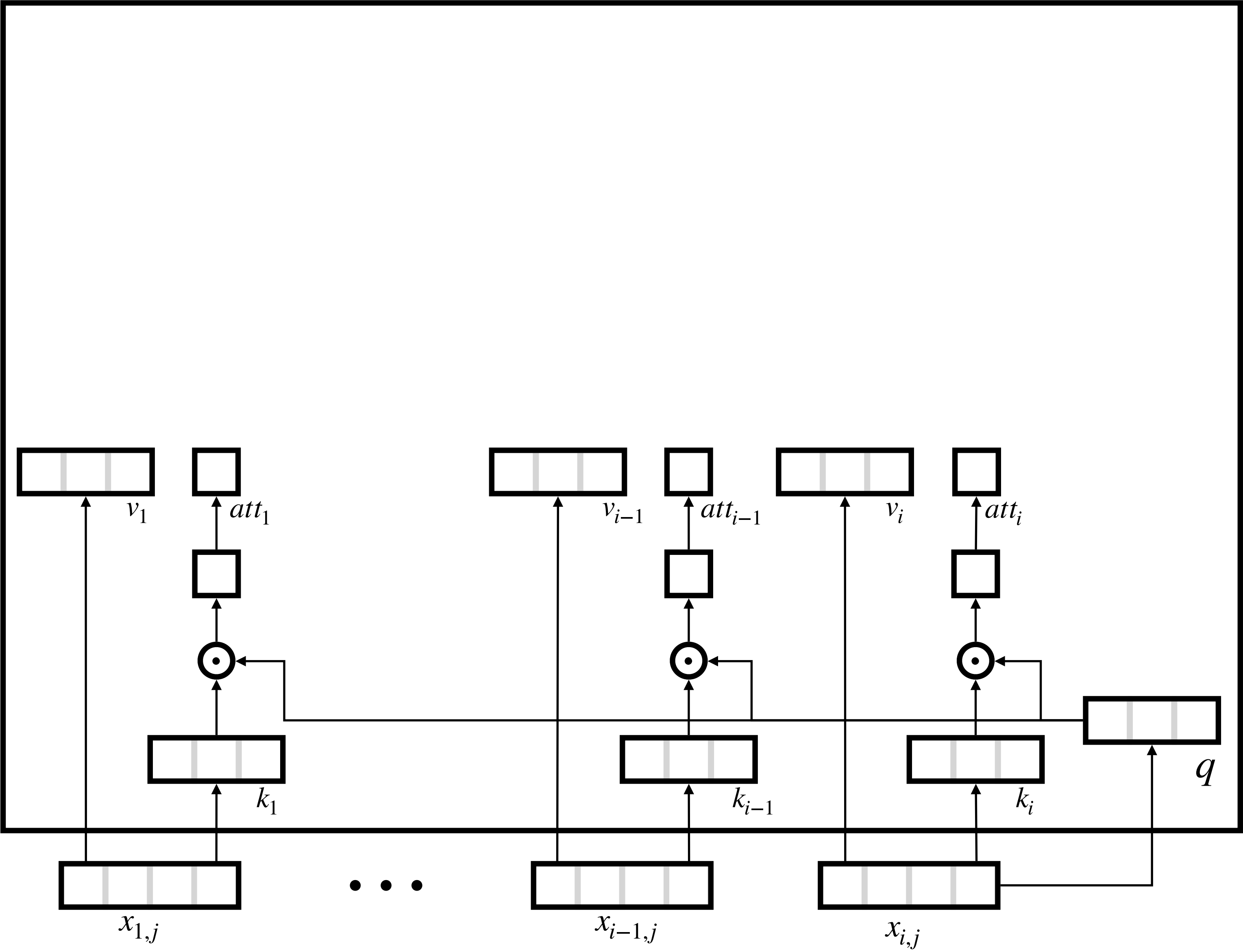
$$k_i = x_i W_k$$

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

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



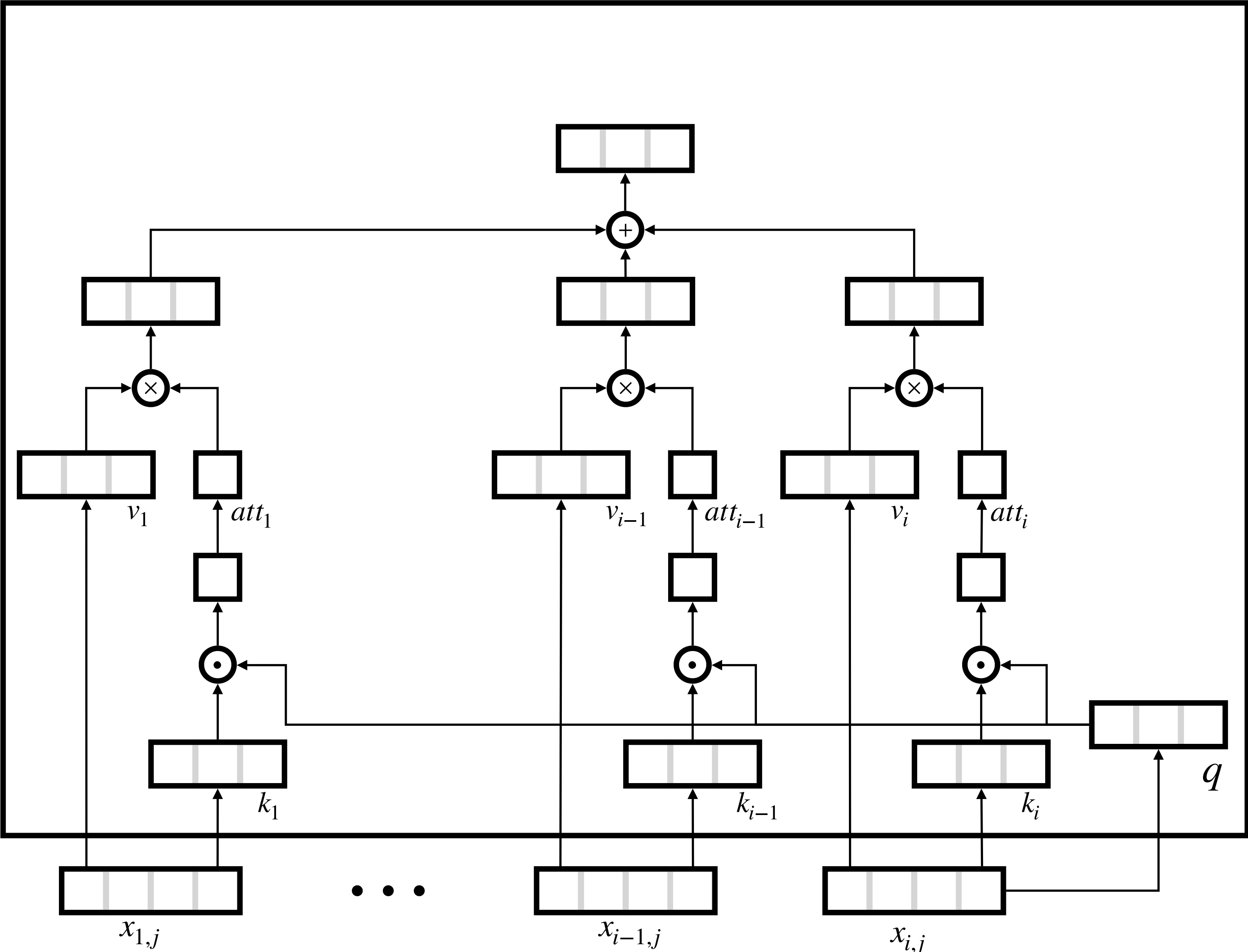
3 : récupérer cette information dans ces tokens : *values*



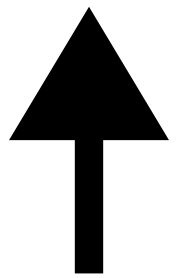
$$v_i = x_i W_v$$

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

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

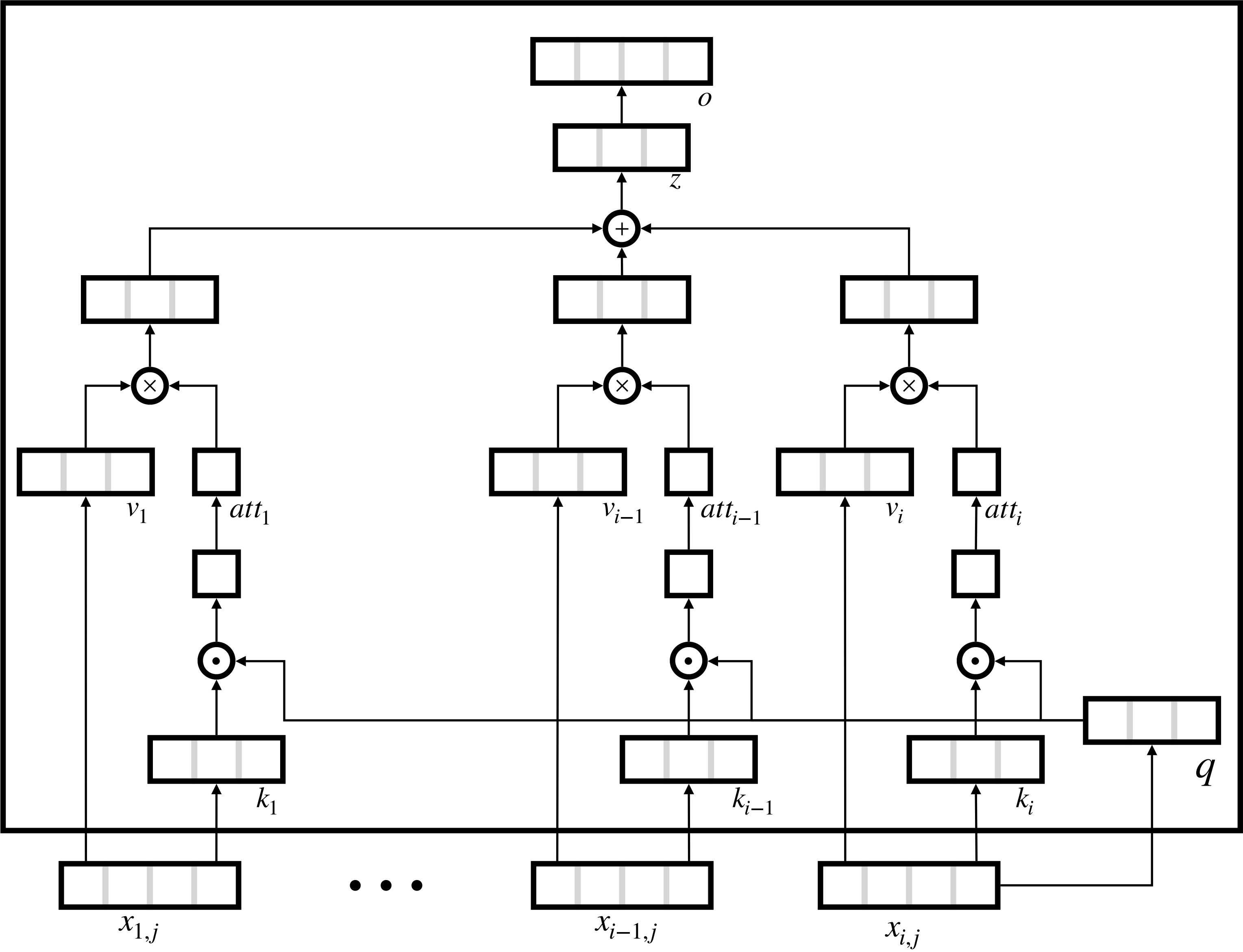


Somme pondérée par les scores d'attention des vecteurs *values*



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

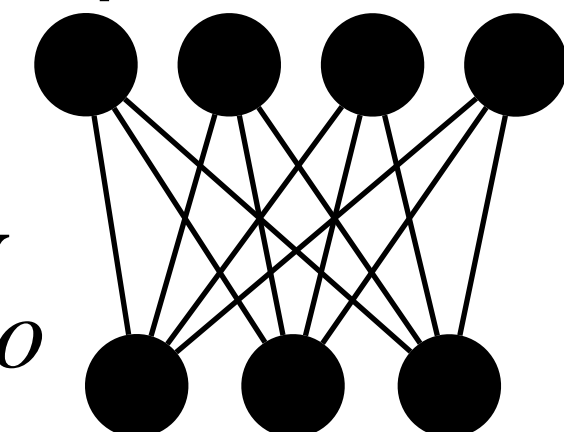
↑

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

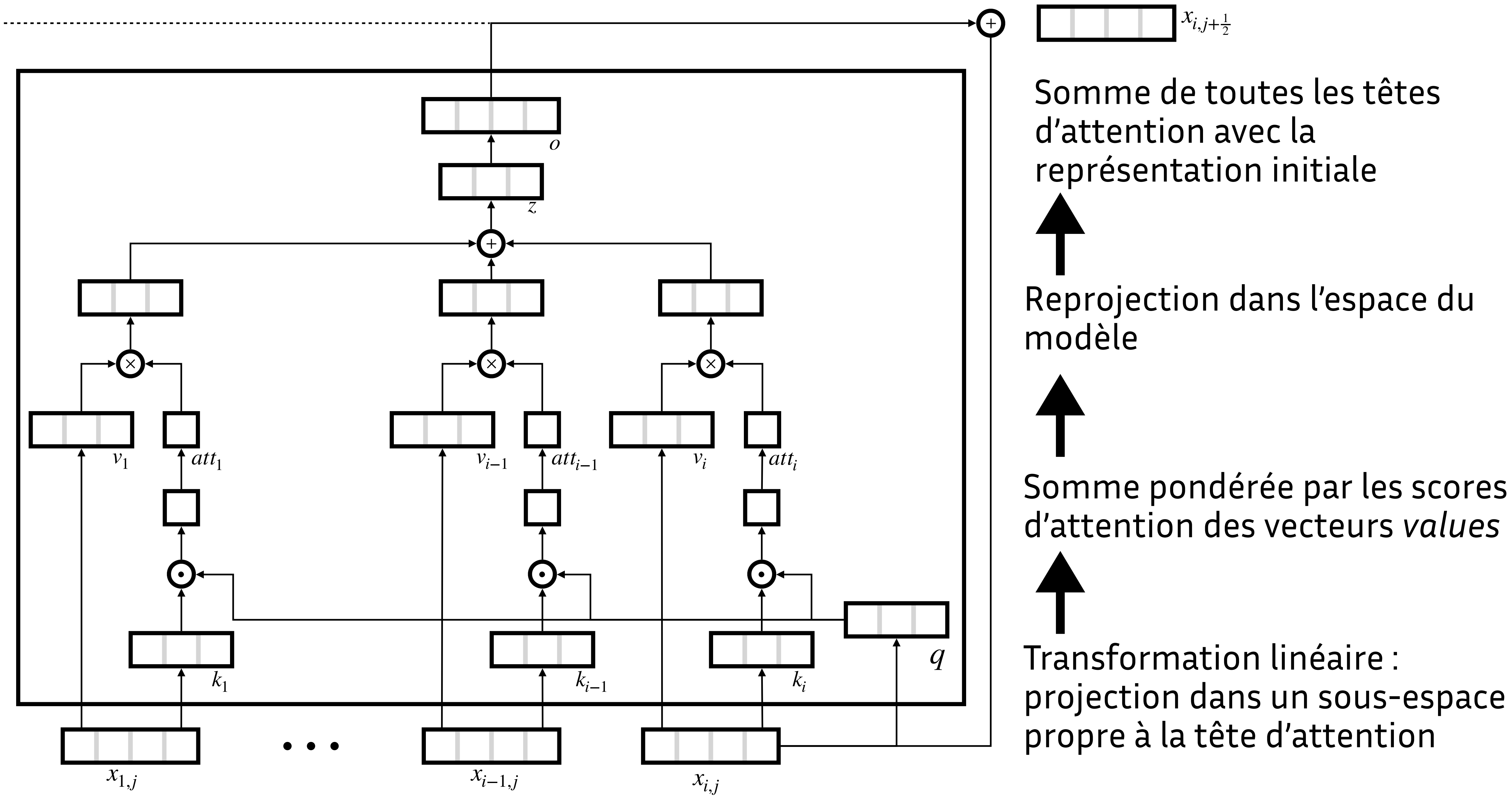
↑

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

$o = zW_o$



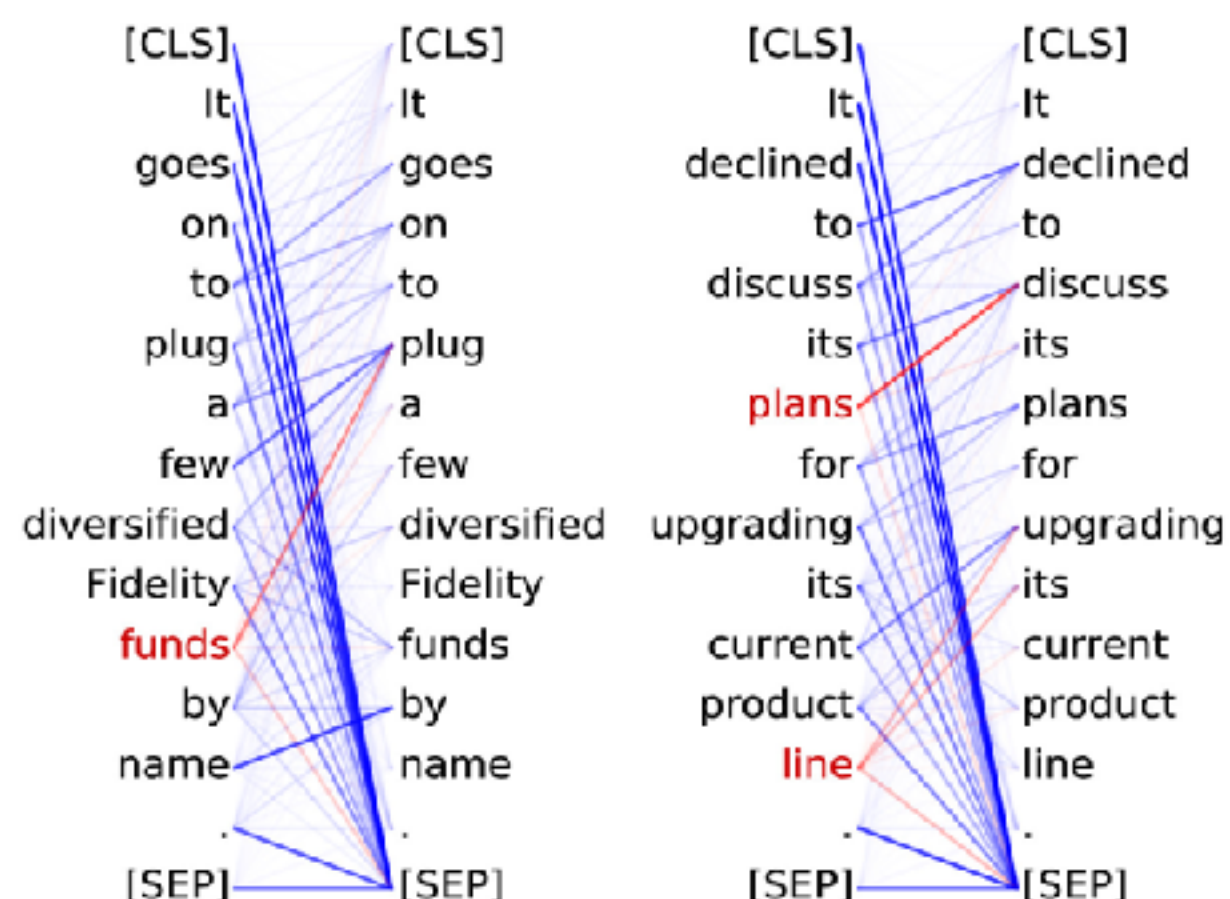
3 : récupérer cette information dans ces tokens : *values*



Patterns attentionnels

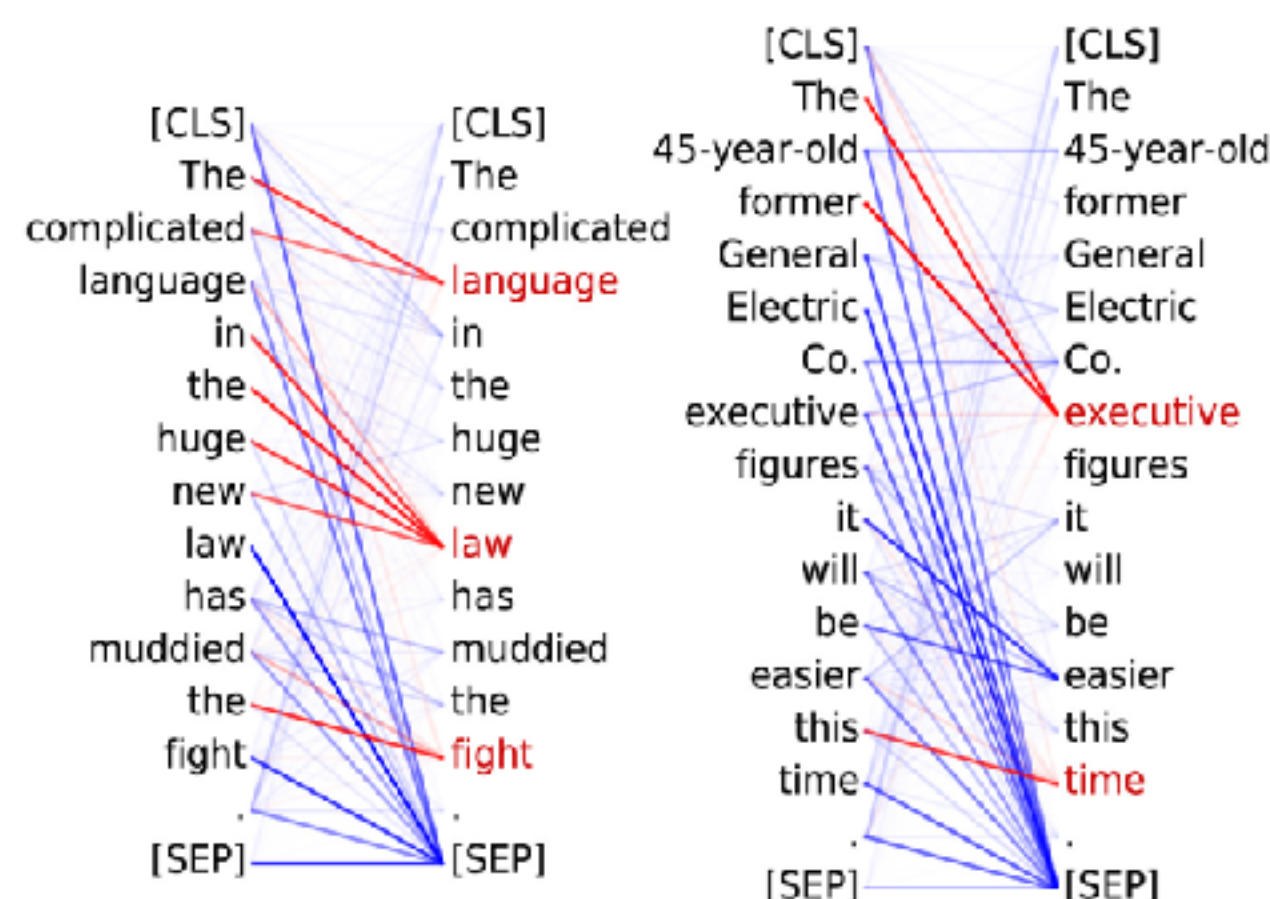
Head 8-10

Direct objects most attend to their verbs 86.8% of the time.



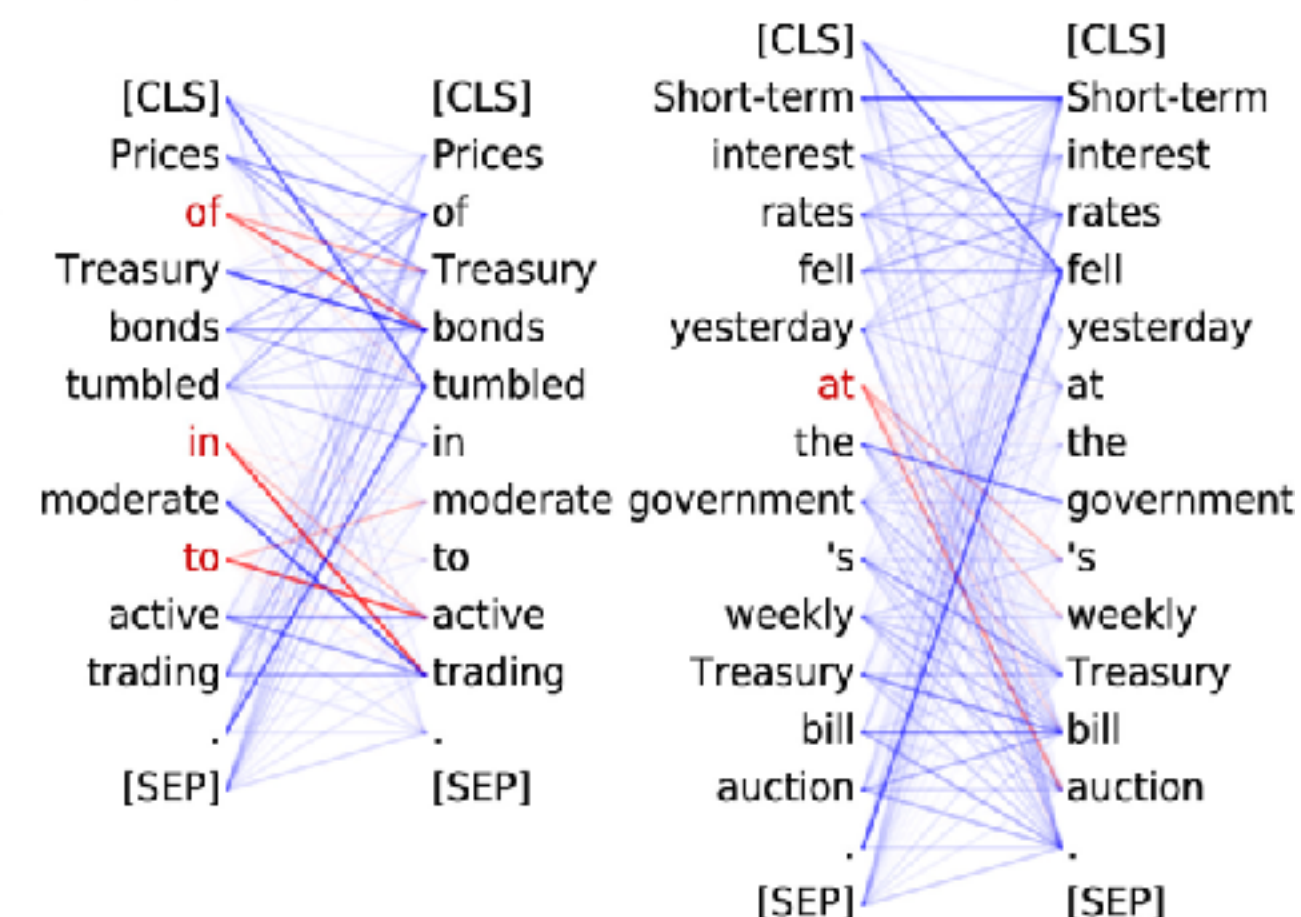
Head 8-11

Noun premodifiers attend to their noun. Determiners most attend to their noun 94.3% of the time.



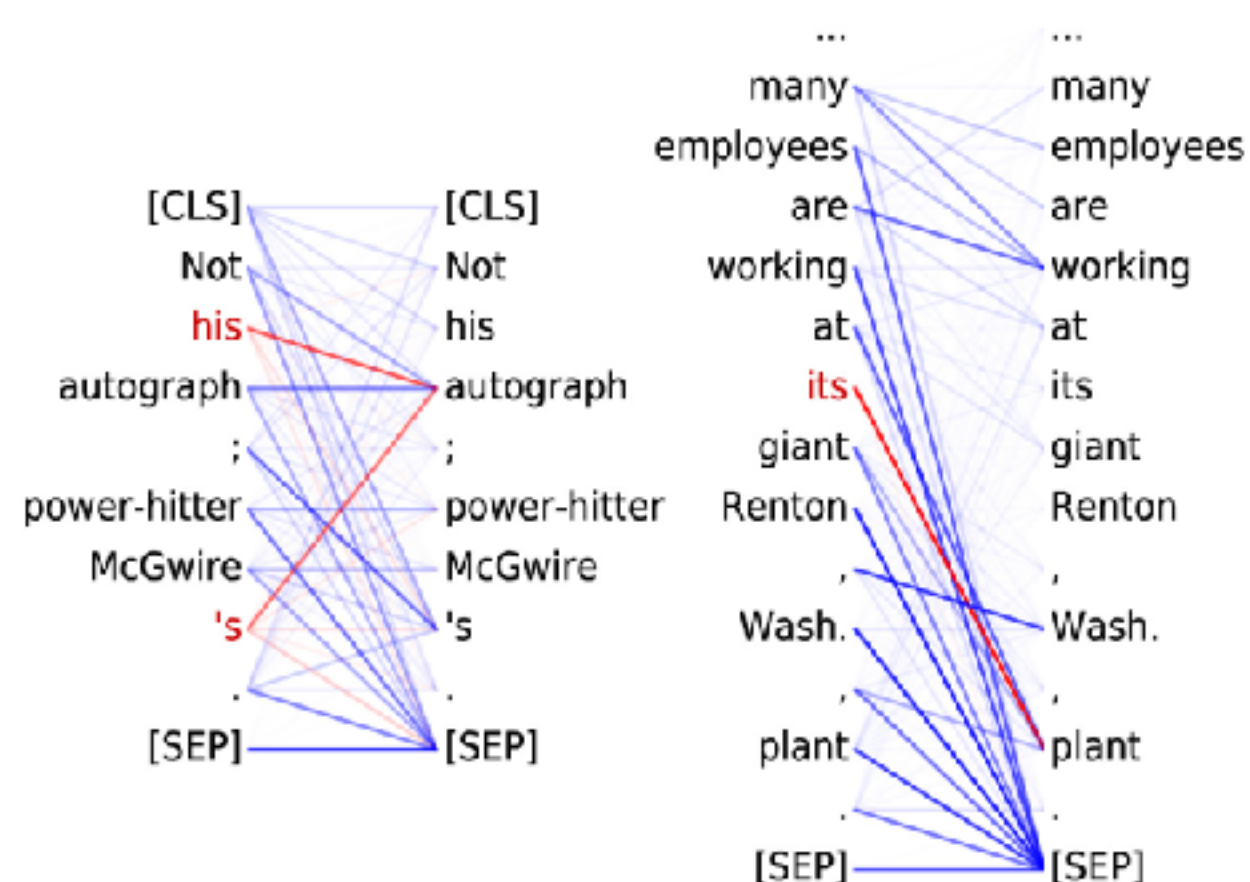
Head 9-6

Prepositions most attend to their objects 76.3% of the time



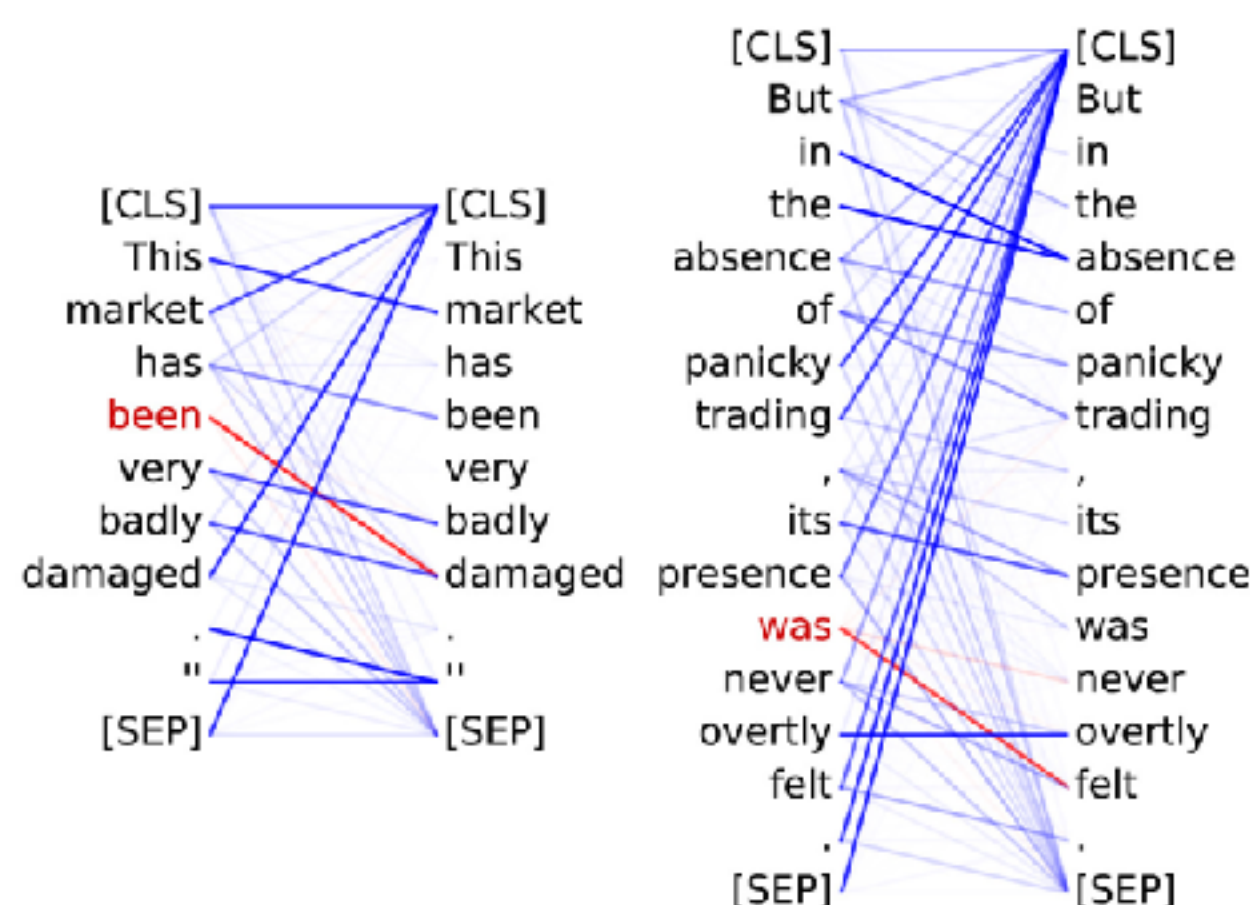
Head 7-6

Possessive pronouns and apostrophes most attend to the head of the corresponding NP 80.5% of the time.



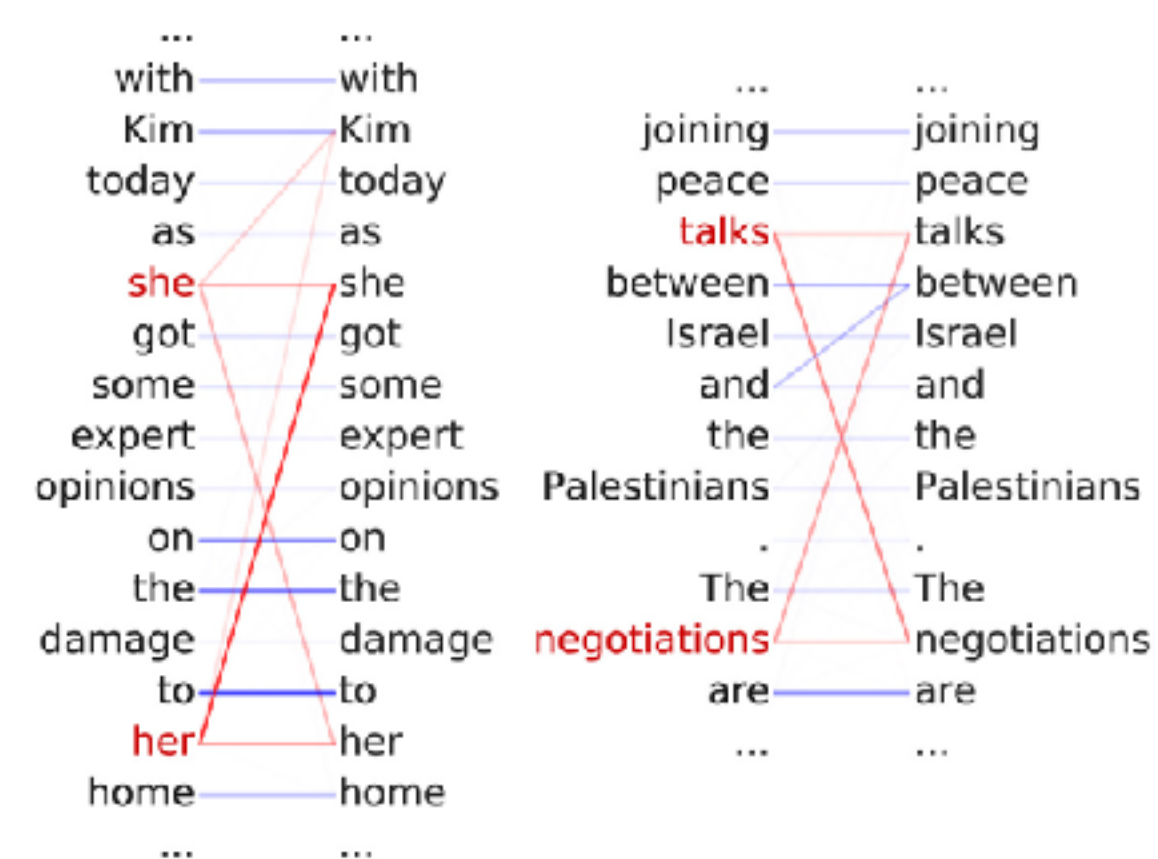
Head 4-10

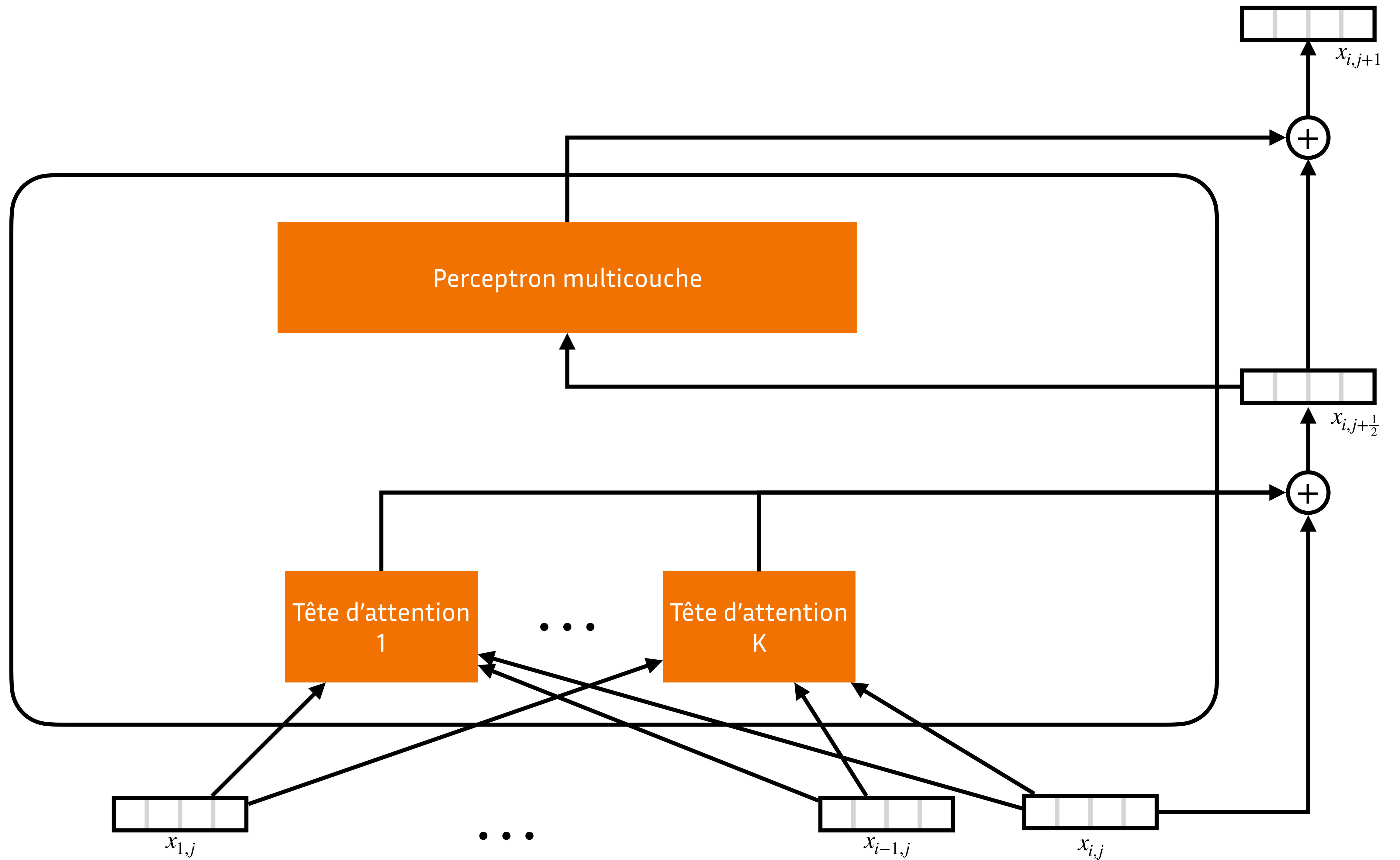
Passive auxiliary verbs most attend to the verb they modify 82.5% of the time.

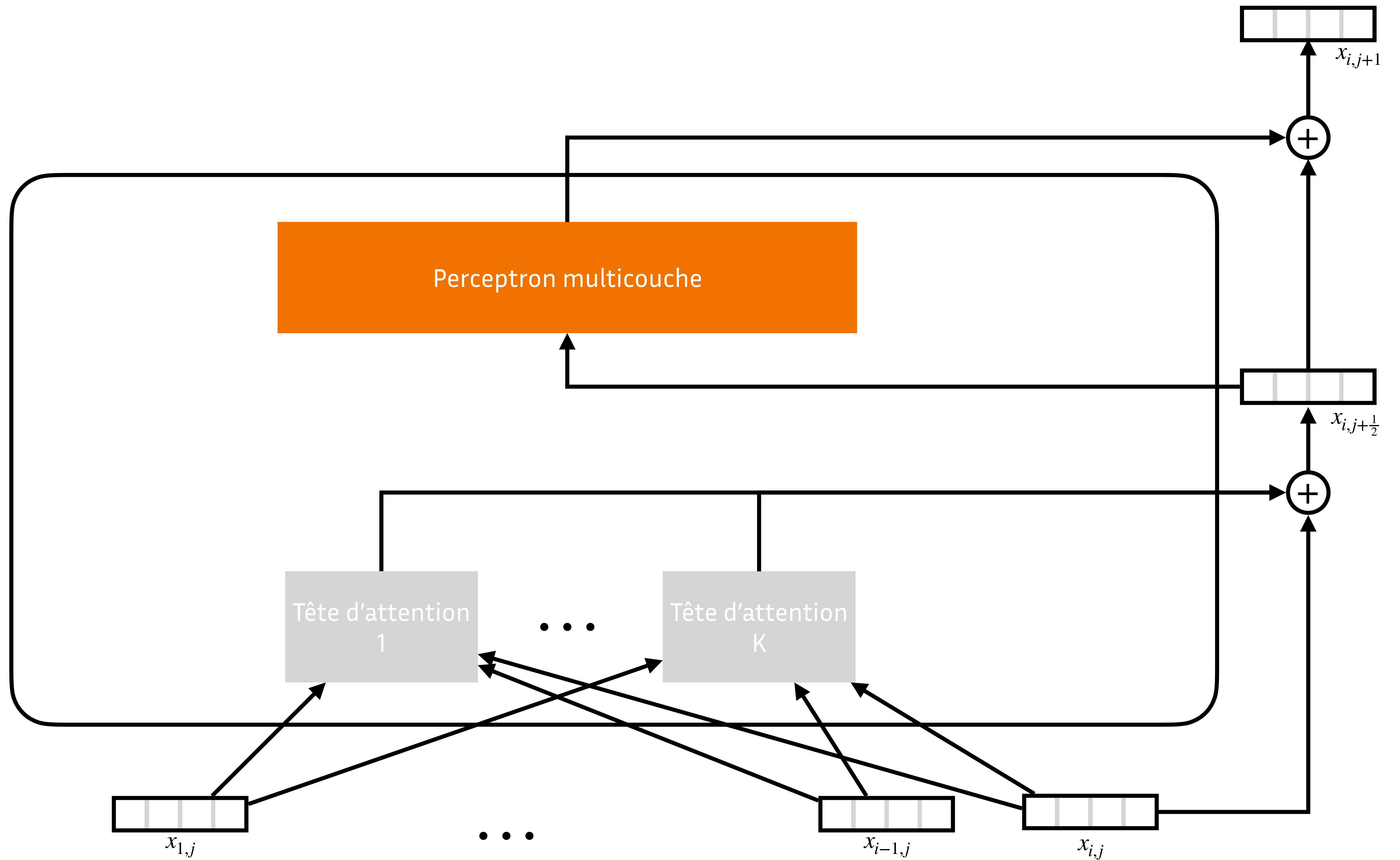


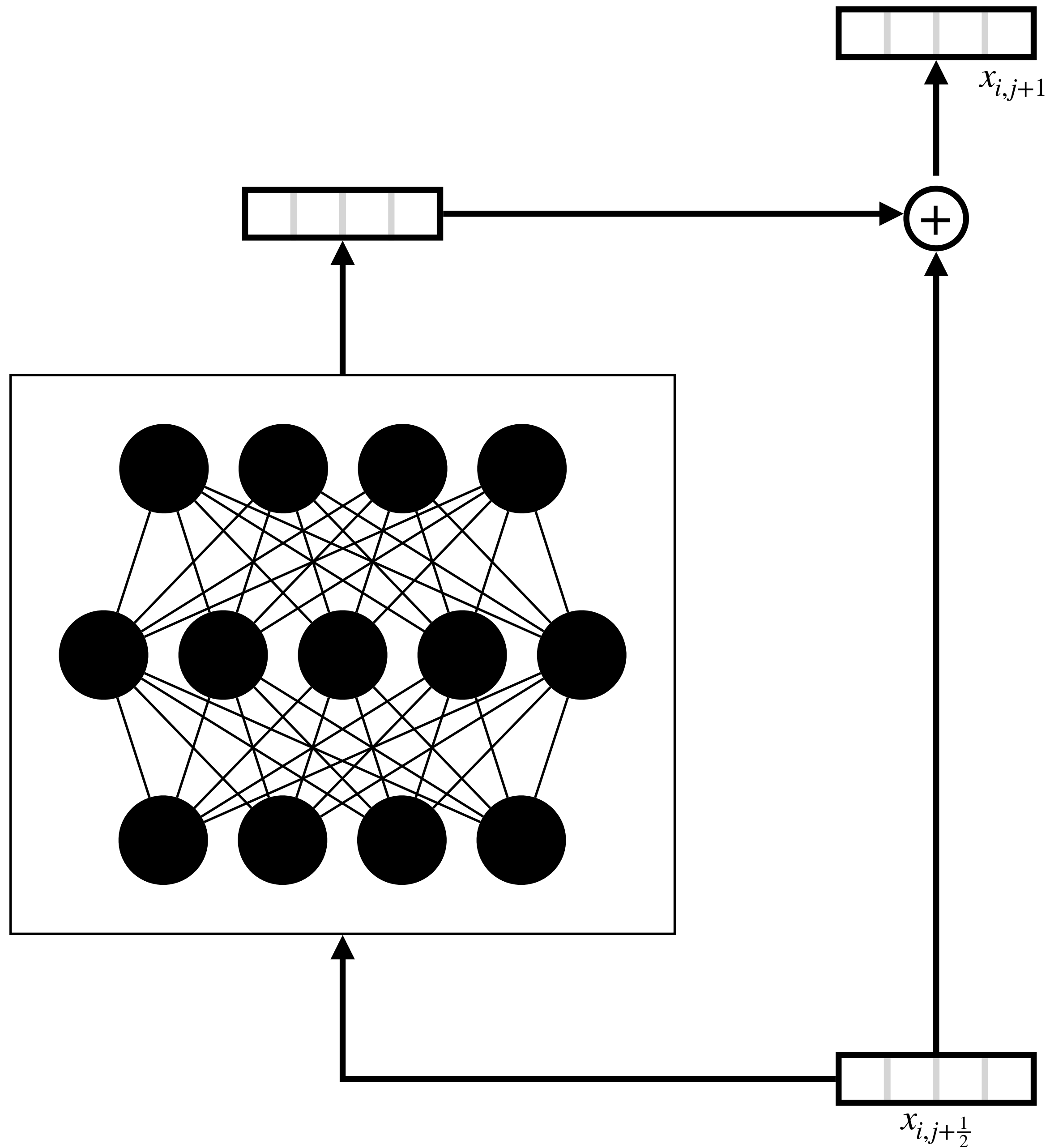
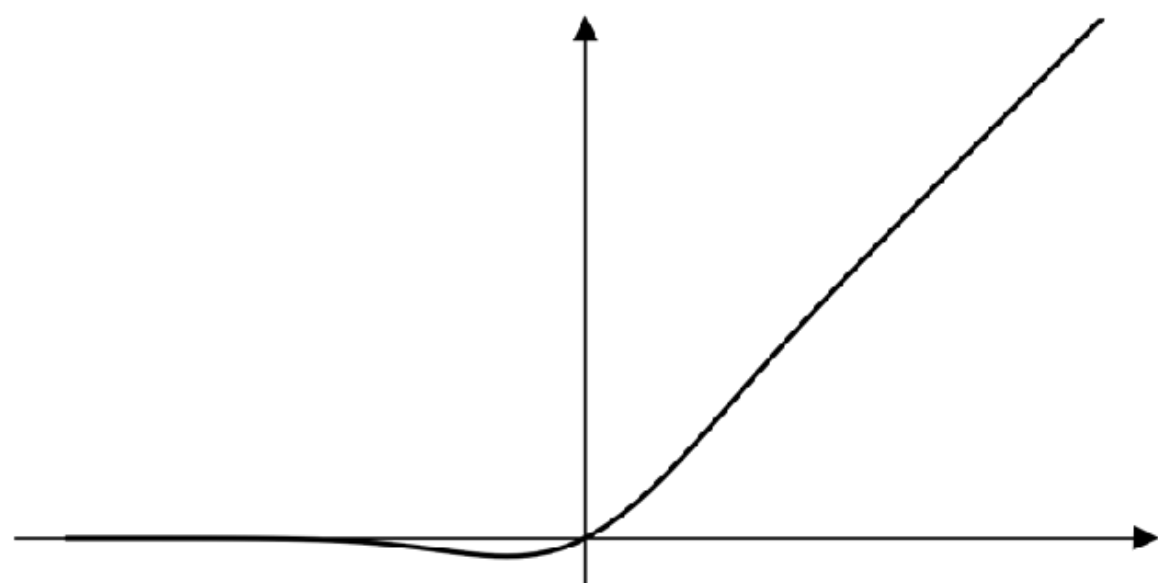
Head 5-4

Coreferent mentions most attend to their antecedents 65.1% of the time.

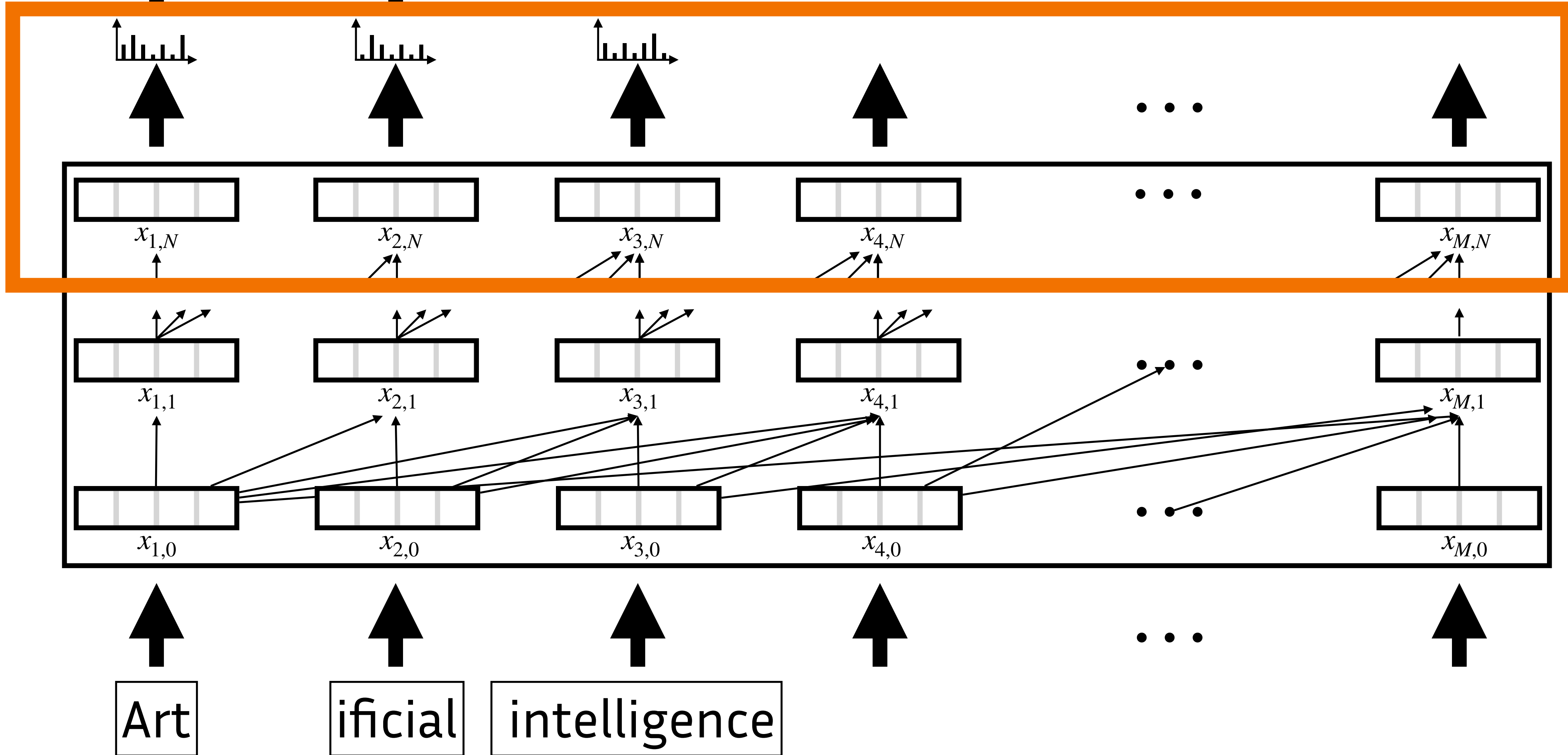


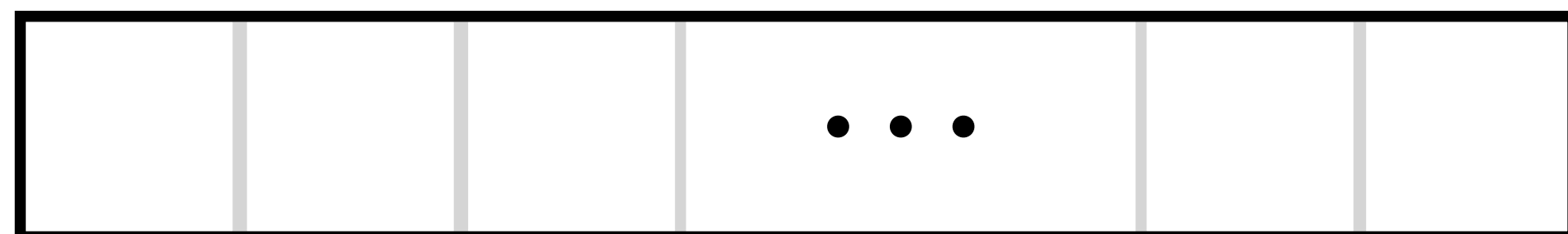






Unembedding



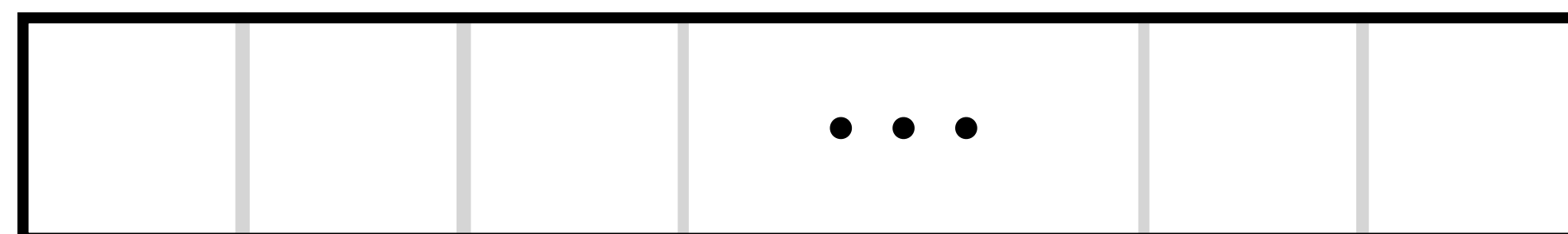


Distribution de probabilité sur le
vocabulaire

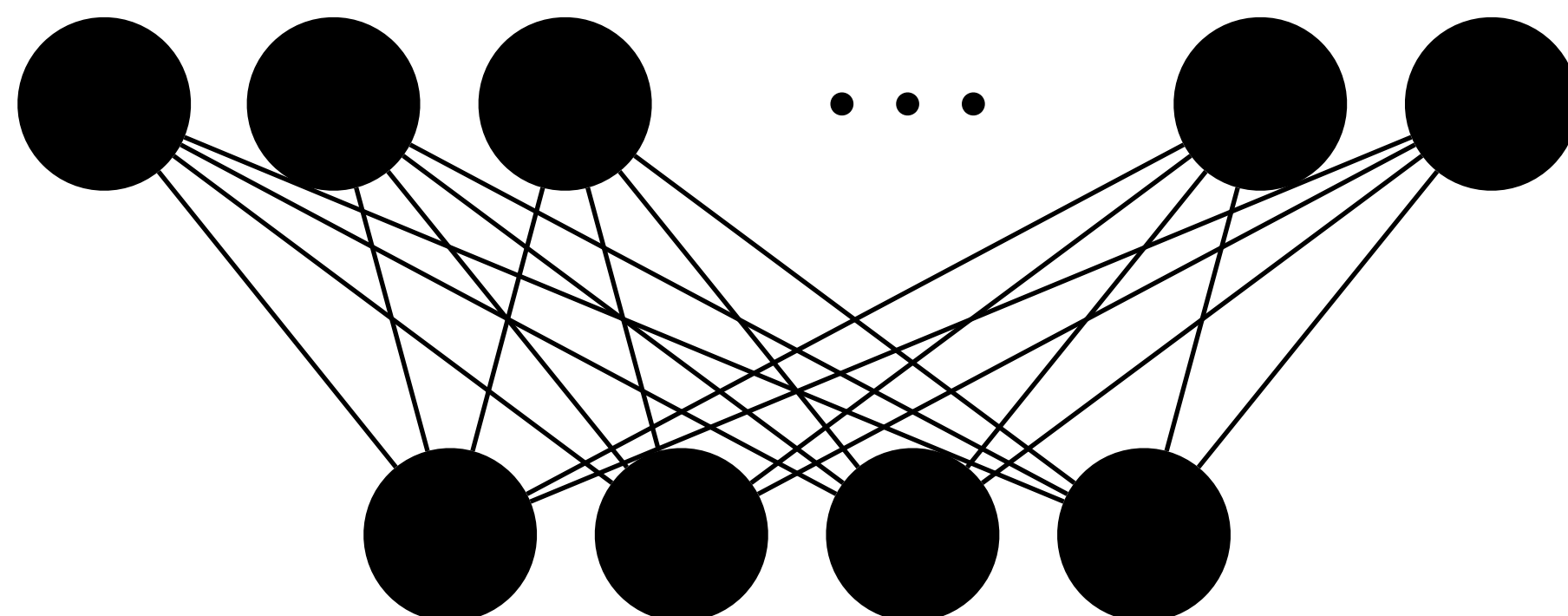


Softmax

S_i



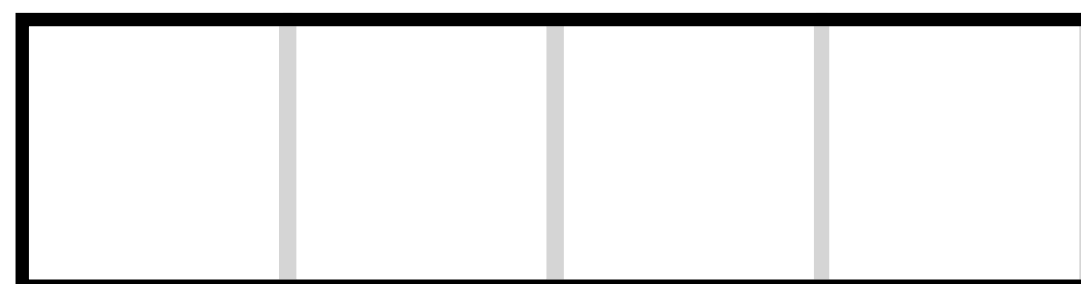
Scores (logits)



Projection sur le vocabulaire

$$S_i = x_{i,N} W_o$$

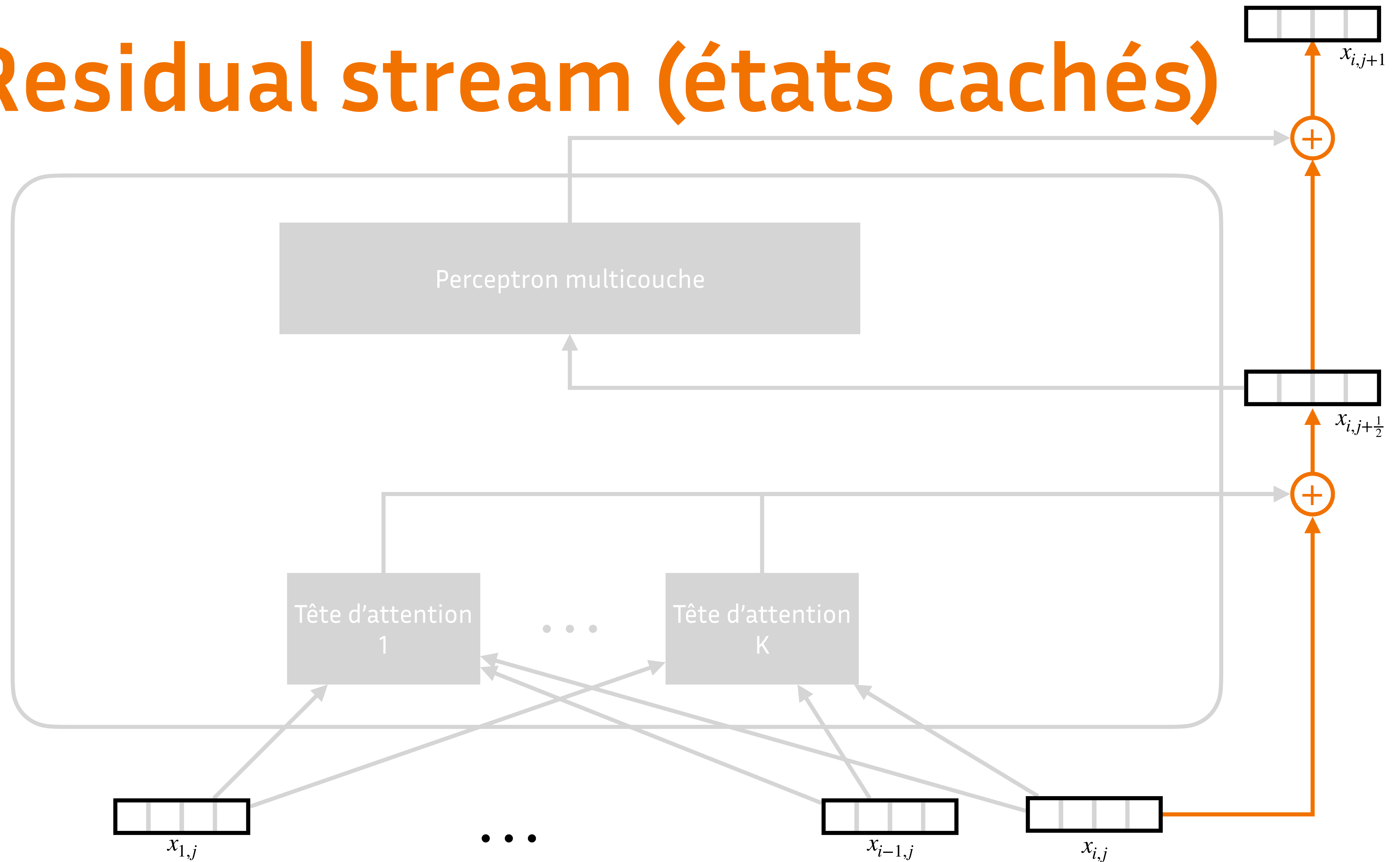
$x_{i,N}$



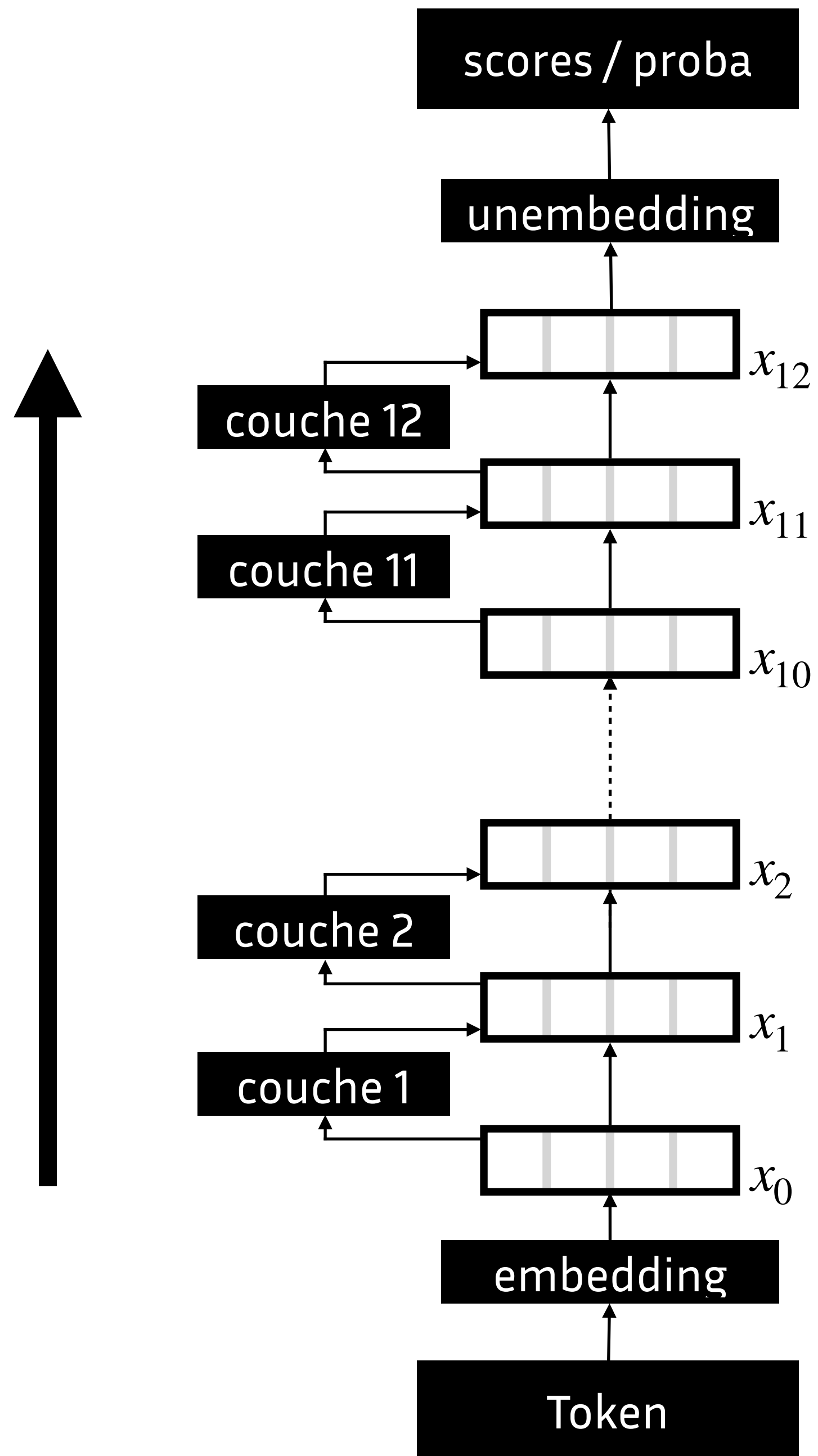
Représentation du token sur la
dernière couche

Représentations

Residual stream (états cachés)

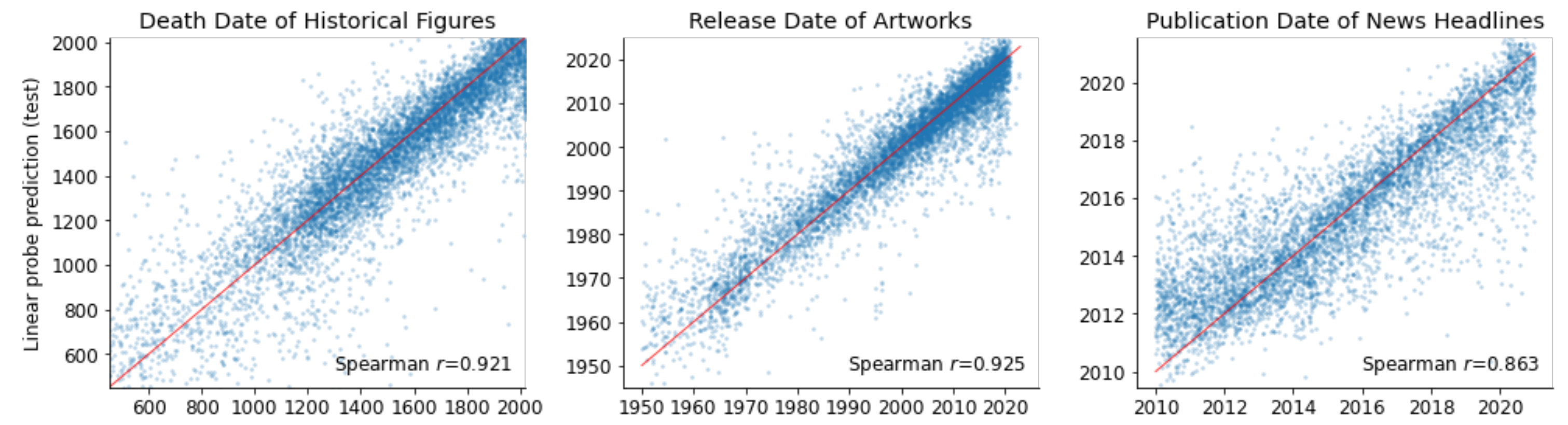
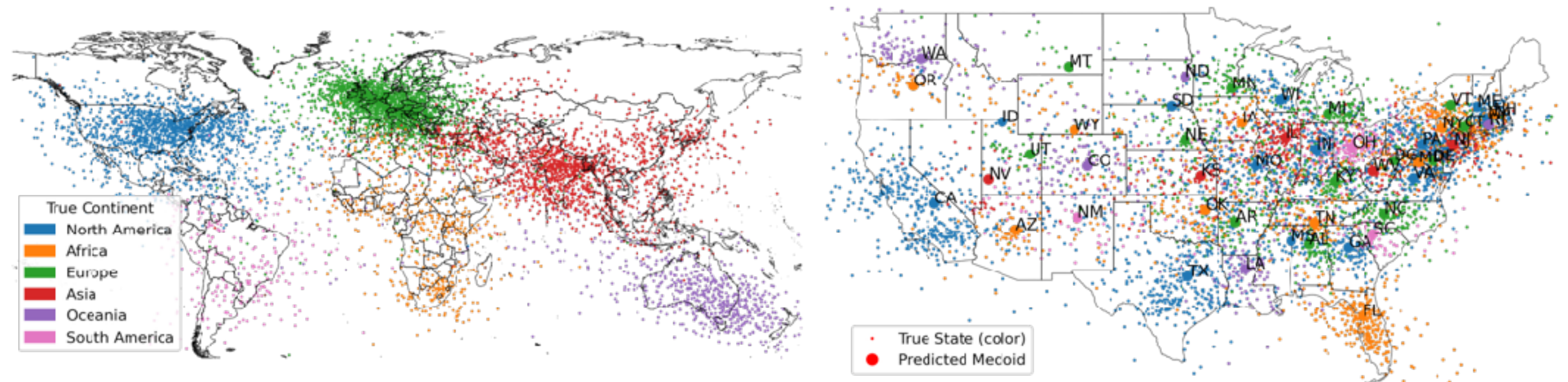
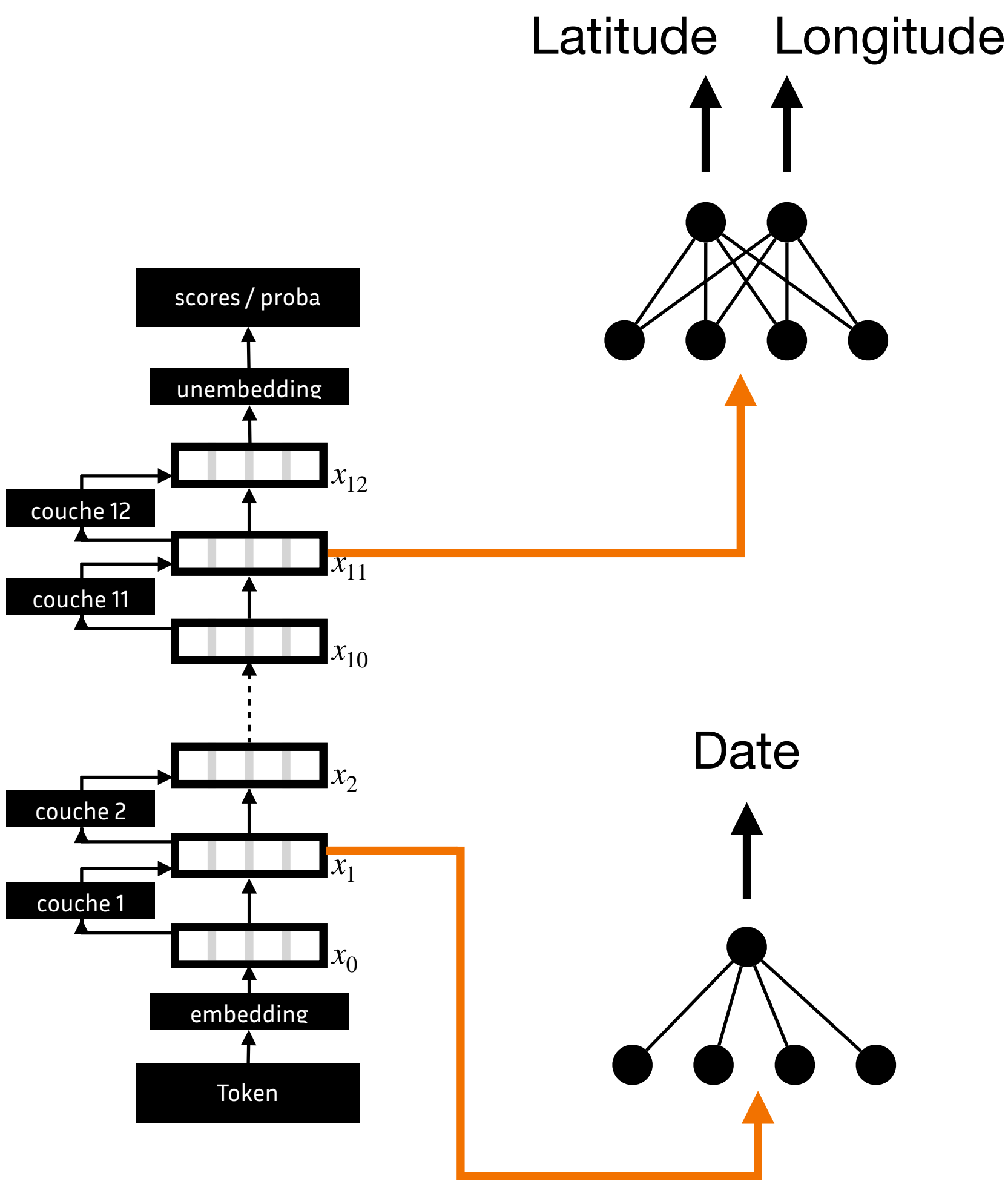


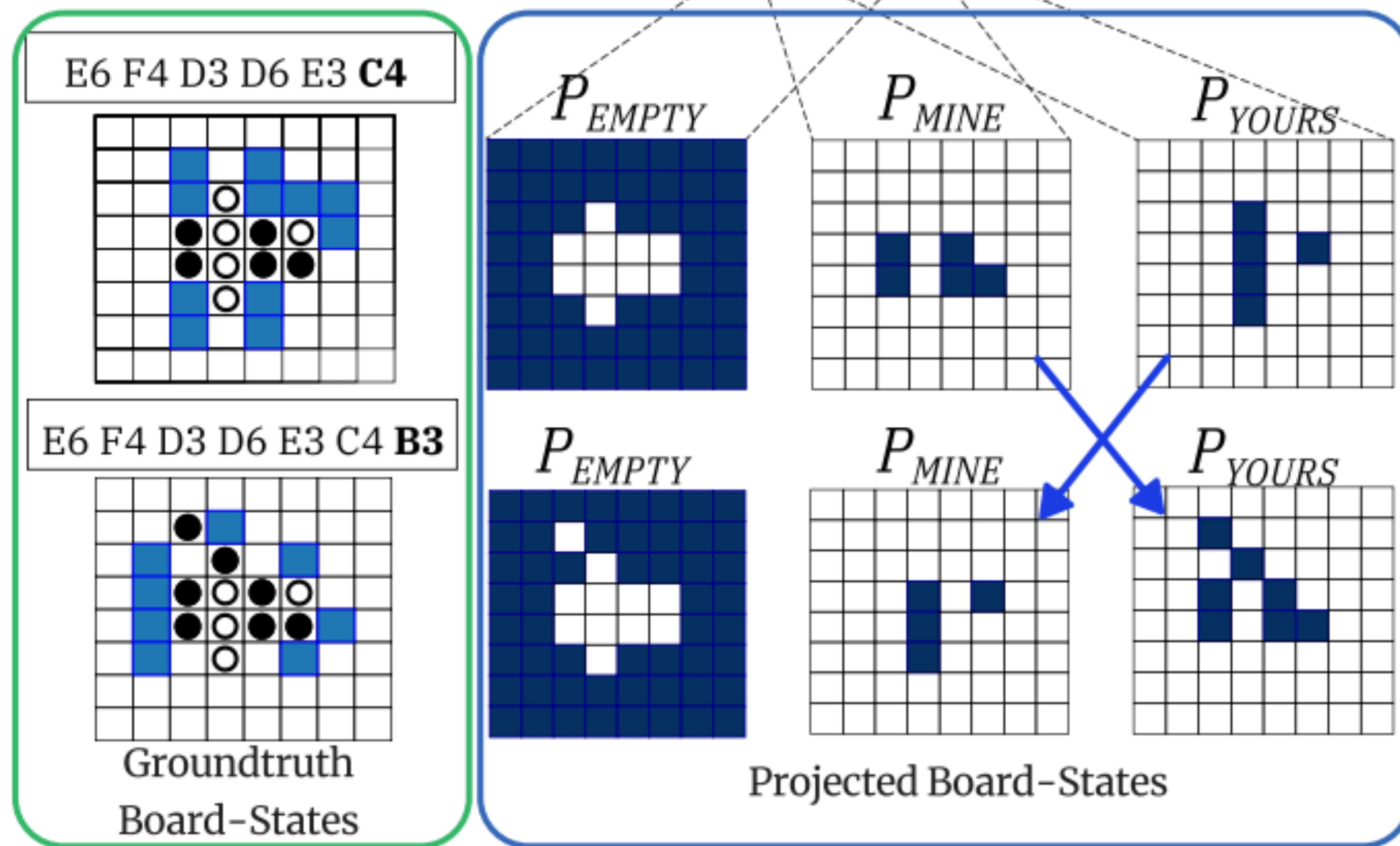
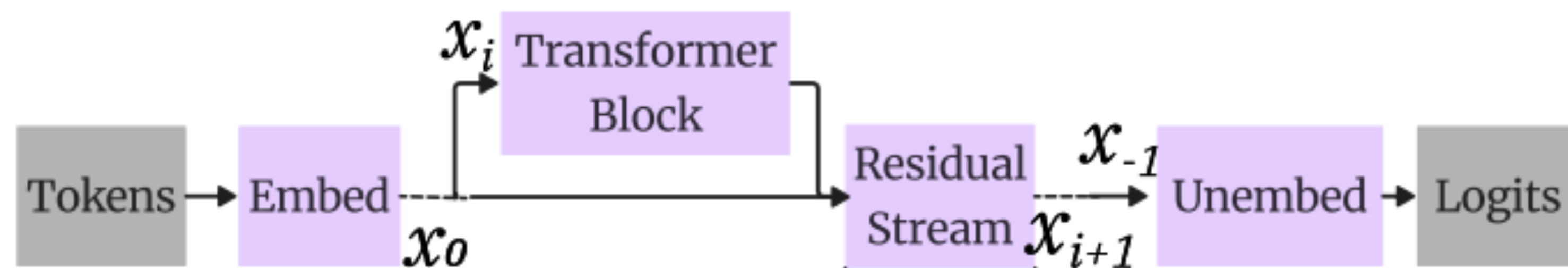
Transformations /
complexifications
successives de la
représentation des
tokens



représentation
suffisamment élaborée
pour permettre la
prédiction du token
suivant

représentation du token
d'entrée





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>