Facteurs qui déterminent la consommation d'un modèle

Pour chaque requête, la consommation d'IA en ressources dépend de ces trois principaux facteurs :

Taille du modèle

En milliards de paramètres

Longueur du texte

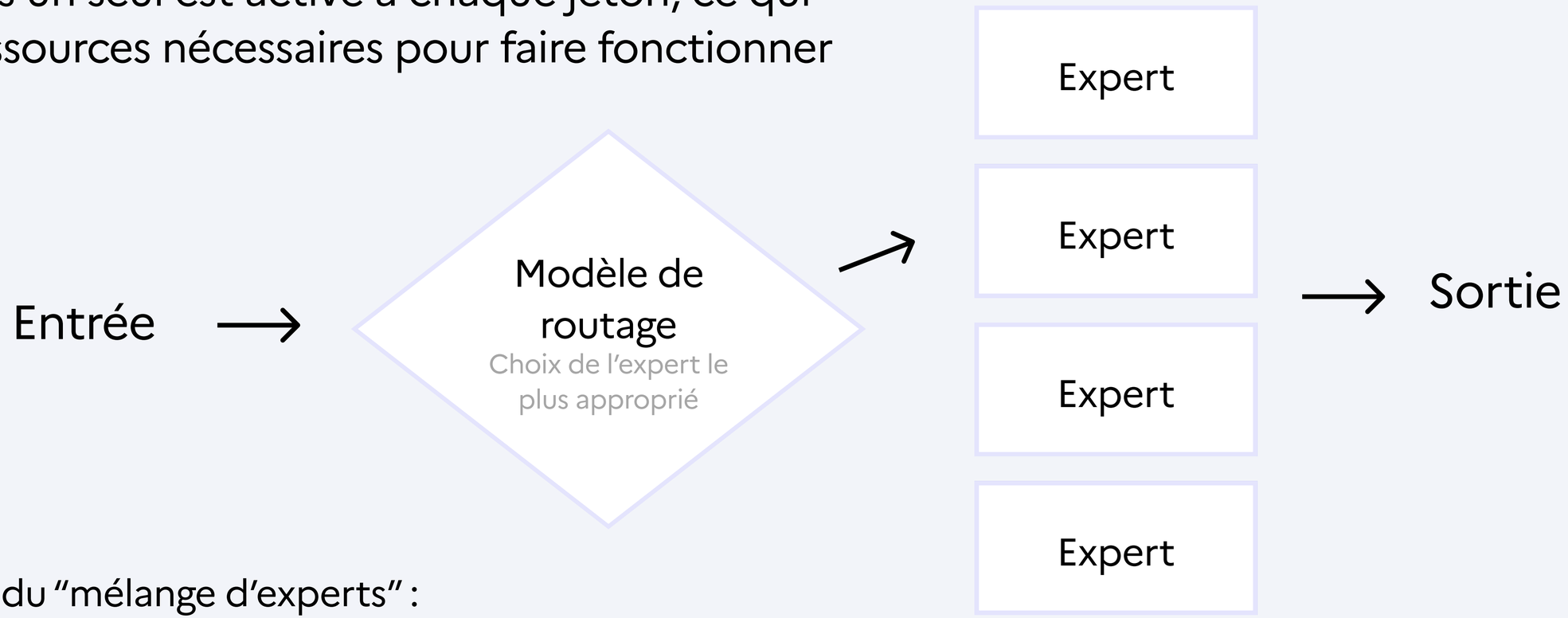
En jetons

Architecture

Par exemple MoE ou MatFormer

Les “mélanges d’experts”

Les modèles “mélanges d’experts” contiennent plusieurs experts, mais un seul est activé à chaque jeton, ce qui réduit les ressources nécessaires pour faire fonctionner le modèle.



Désavantages du “mélange d’experts” :

- plus complexe à développer
- plus de mémoire nécessaire
- légère baisse de performance et de capacité de généralisation