

Comment fonctionne un LLM

(Modèle de Langage de Grande Taille)

1 Données



- **Collecte** : textes variés (web, livres, forums)
- **Nettoyage** : filtrage, anonymisation, formatage
- **Volume** : milliards de tokens

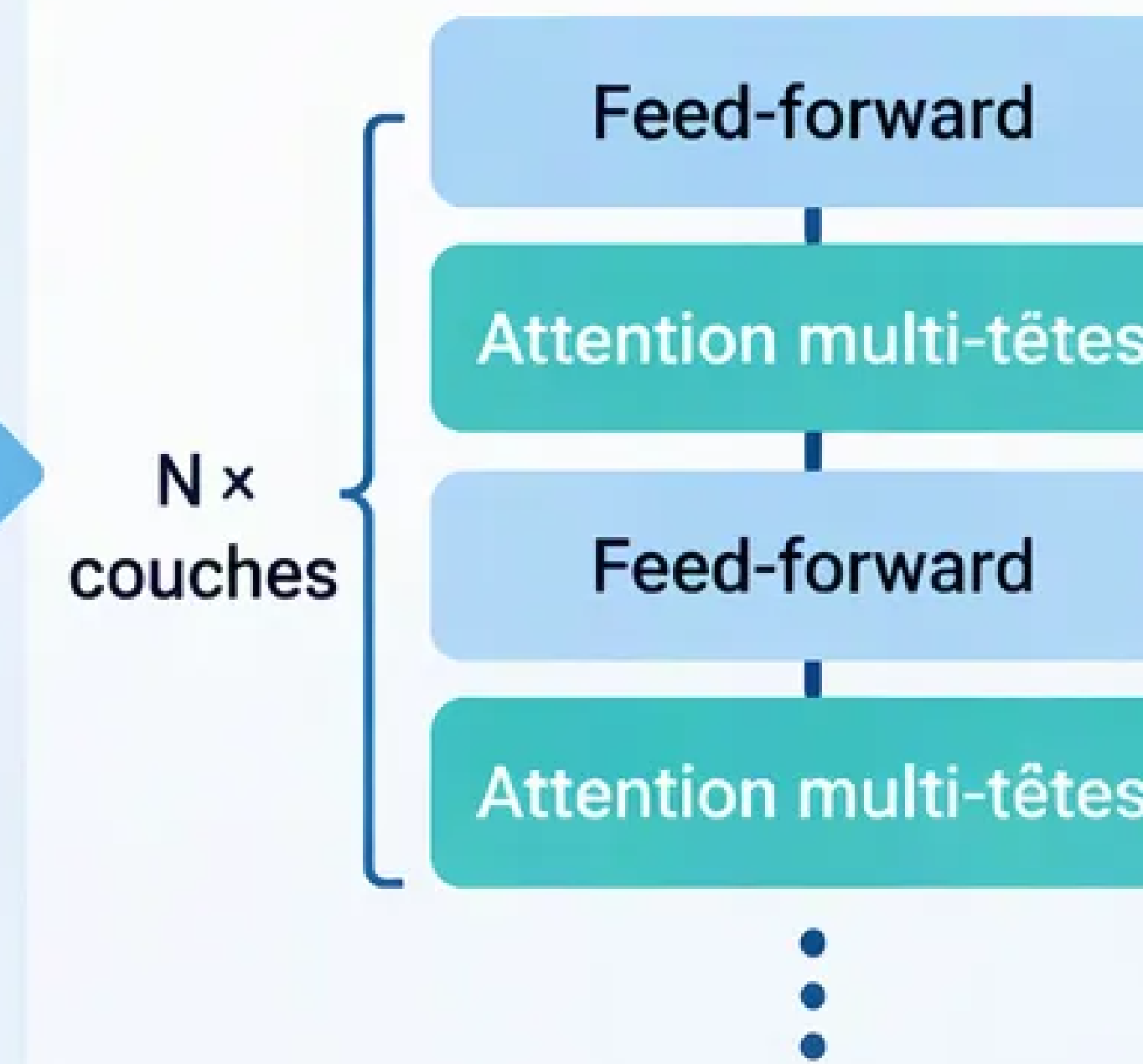
2 Tokenisation

Un chat dort sur le tapis.

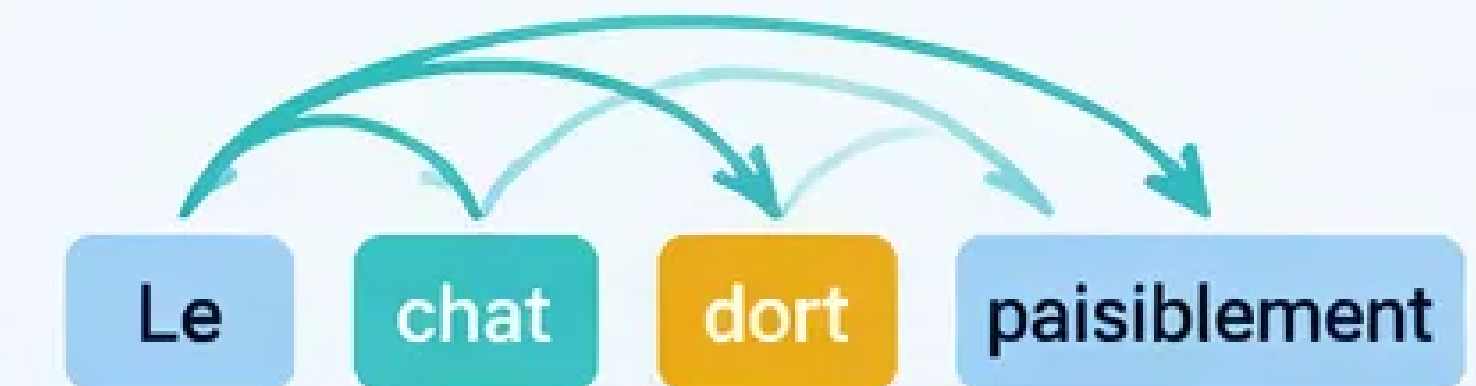


- Découpe du texte en tokens (mots ou sous-mots)
- Conversion tokens → entiers (IDs)
- Vocabulaire fixe utilisé par le modèle

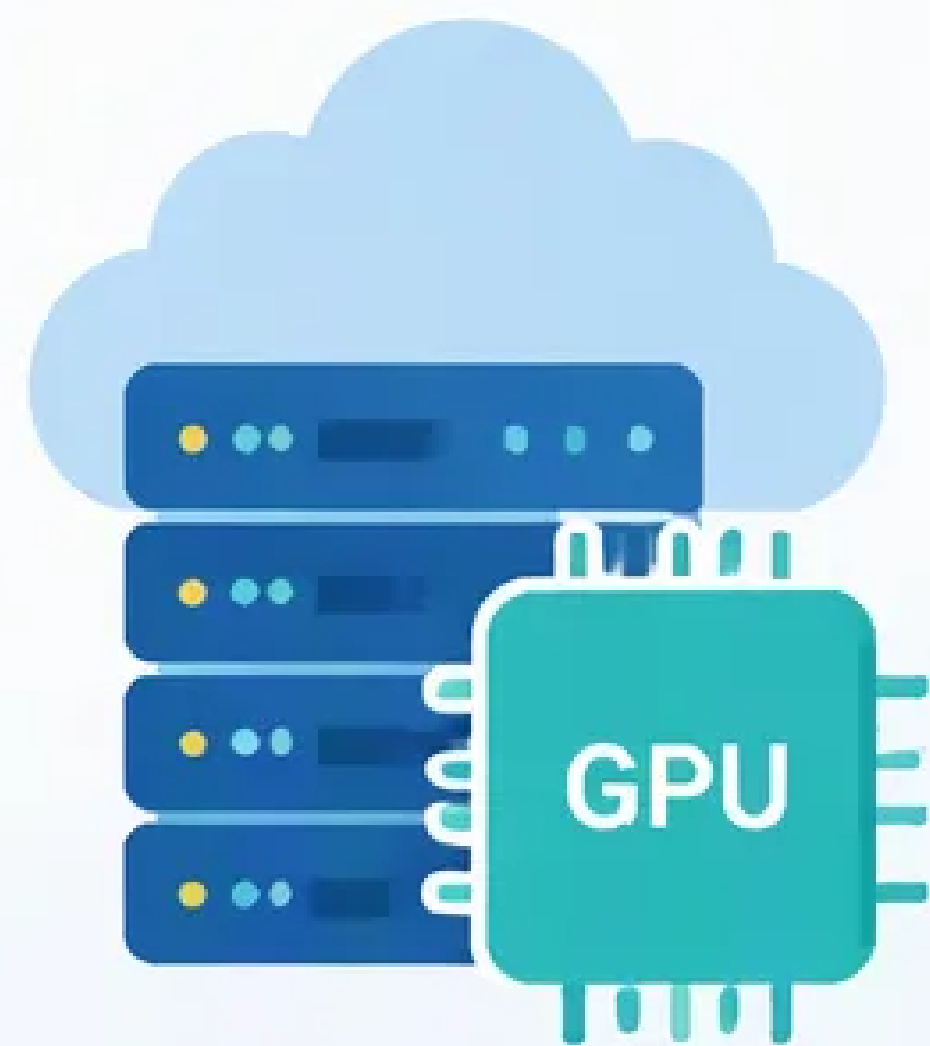
3 Architecture Transformer



- Transformer = empilement de couches d'attention + feed-forward
- Mécanisme clé : Attention multi-têtes (weights entre tokens)
- Illustration : schéma simplifié avec flèches montrant attention entre mots



4 Entraînement



- **Objectif** : prédire le token suivant / minimiser la perte
- **Pré-entraînement** massif sur données non supervisées
- **Ressources** : GPU/TPU, longues itérations

5 Inférence

Quel temps fait-il aujourd'hui ?

Aujourd'hui, il fait généralement beau et ensoleillé. ...

- **Entrée** : prompt utilisateur (texte)
- Le modèle prédit probabilités sur tokens suivants
- **Décodage** : greedy/beam/temperature → texte final

6 Fine-tuning & Sécurité



- **Fine-tuning** supervisé pour tâches spécifiques
- **RLHF** : apprentissage avec feedback humain pour alignement
- Filtrage et modération en sortie

DES DONNÉES AUX RÉPONSES, UNE CHAÎNE INTELLIGENTE



Bonnes pratiques & limites

- Surveillez les hallucinations et biais
- Vérifiez les faits dans les réponses critiques
- Respectez la confidentialité des données
- Utilisez prompts clairs et instructions contextuelles