

Les duels de l'IA

par compar:IA

Comment est entraîné un modèle d'IA ?



Les duels de l'IA
par compar:IA

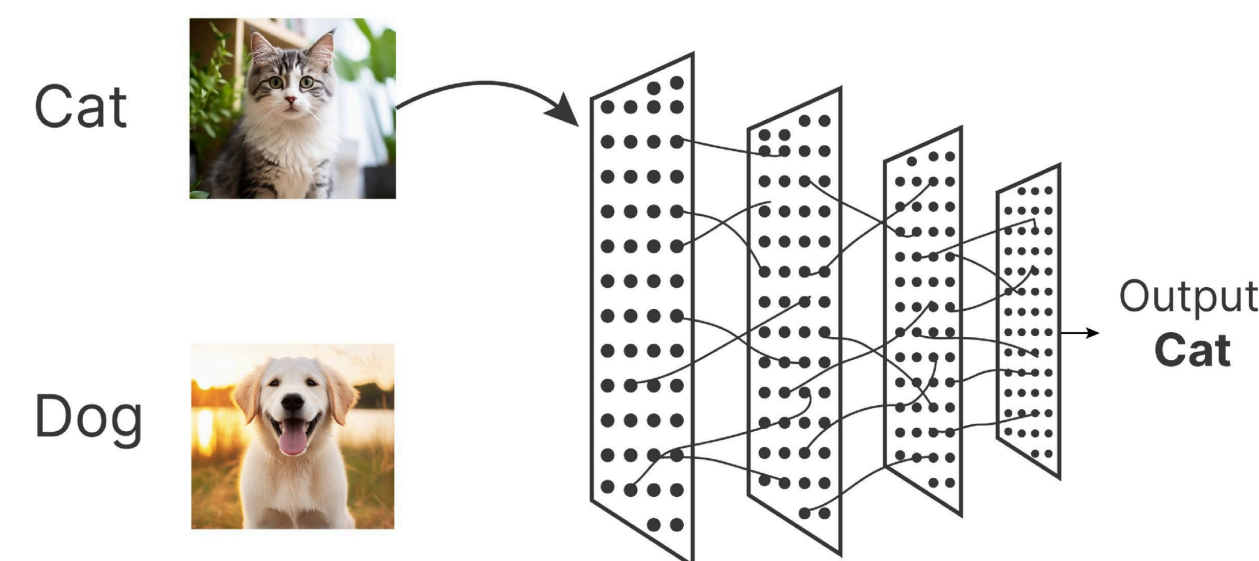
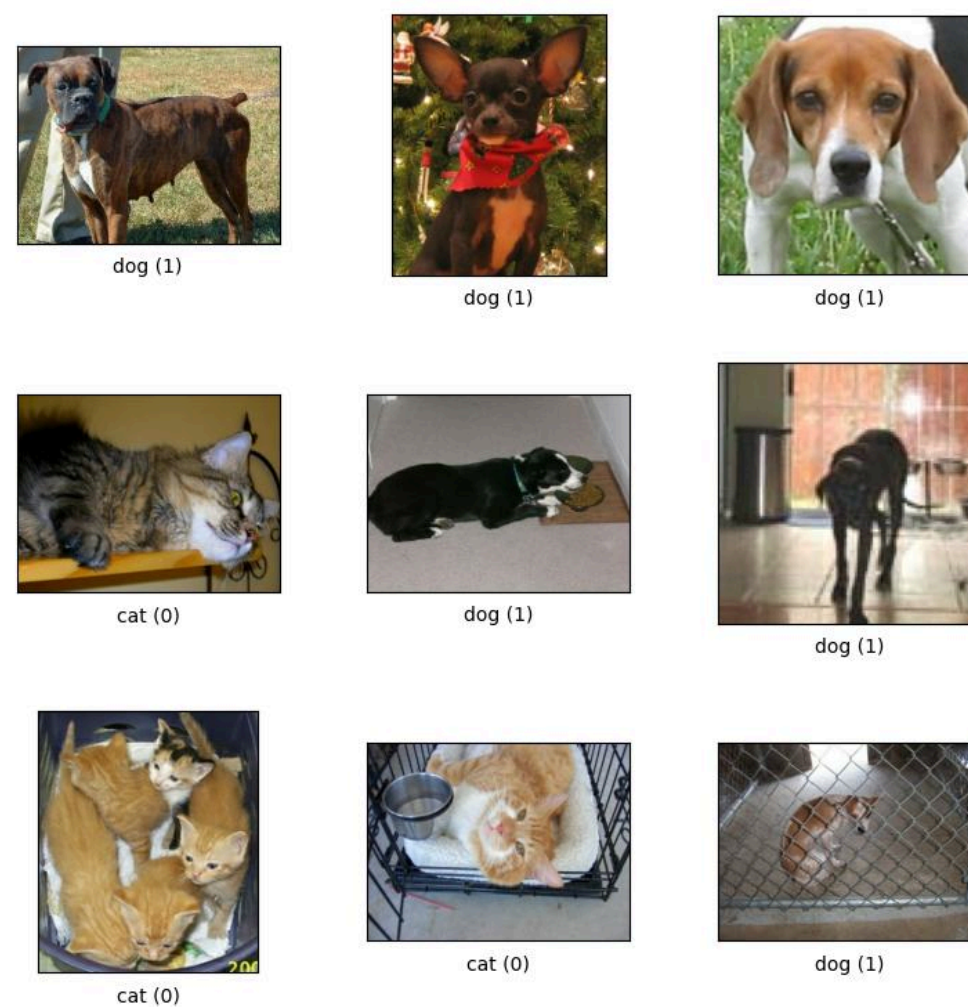
Des **données** pour
l'entraînement

+

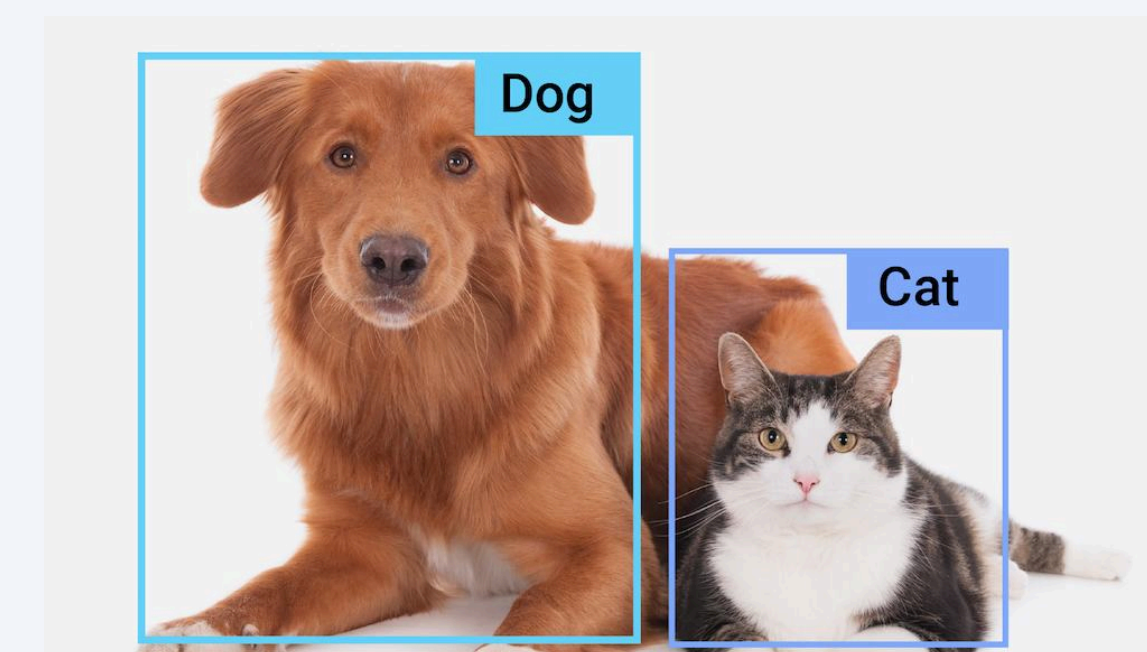
Beaucoup de
calculs pour
"apprendre"

=

Un **modèle d'IA**



Prédit sur de nouvelles
données



Comment est entraîné un modèle d'IA conversationnelle ?



Les duels de l'IA
par compar:IA

Des **données** pour
l'entraînement

+

Beaucoup de
calculs pour
"apprendre"

=

Un modèle d'IA



Quasiment tout Internet
public



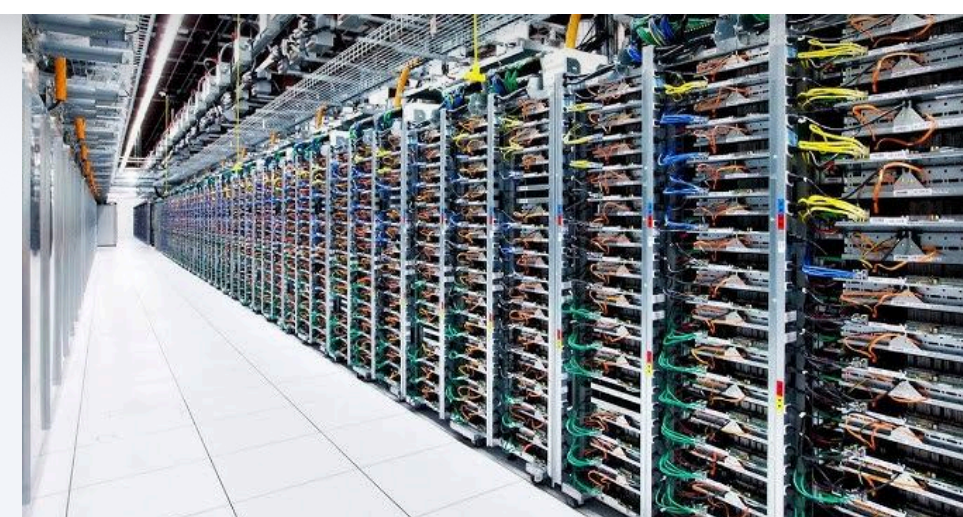
Peu de filtrage des contenus
sous droits d'auteur



Un jeu de données de taille
moyenne tel que FineWeb pèse
51TB de mémoire.

$$\text{Layer}(x) = \text{LayerNorm} \left[x + \text{FFN} \left(\text{Concat} \left(\text{Attention}(Q, K, V) \right) W^O \right) \right],$$

where $Q = (x + PE)W_Q,$
 $K = (x + PE)W_K,$
 $V = (x + PE)W_V,$
 $PE = \text{PositionalEncoding}(\text{positions}),$

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V,$$
$$\text{FFN}(z) = \text{LayerNorm} [z + \text{ReLU}(zW_1 + b_1)W_2 + b_2],$$


Un ensemble de
chiffres qui permettent
de **prédire la réponse**
la plus probable à la
requête.