PATRICK PÉCOT

Ingénieur · Formateur IA & LLM

42 Rue d'Amiens · 76000 Rouen · patpec442@gmail.com · 06 99 43 22 13

NEWSLETTER IA · 01 · Printemps 2026

LLM — Architecture & Fonctionnement

Comprendre ce qui se passe réellement à l'intérieur d'un grand modèle de langage.

QU'EST-CE QU'UN LLM ?

Un Large Language Model (LLM) est un réseau de neurones entraîné sur des centaines de milliards de tokens pour prédire, token après token, la suite la plus probable d'une séquence.

Ce mécanisme simple produit des capacités émergentes remarquables : raisonnement, traduction, synthèse, génération de code. Mais il implique des limites structurelles importantes à connaître.

ARCHITECTURE TRANSFORMER

Tous les LLM actuels (GPT, Claude, Gemini, Llama) reposent sur l'architecture Transformer (2017). Son mécanisme central : l'attention multi-têtes, qui pondère dynamiquement l'importance de chaque token.

— Encodeur : transforme les tokens en représentations vectorielles

— Décodeur : génère les tokens de sortie en auto-régression

— Attention : mesure la pertinence de chaque token vis-à-vis du contexte global

TOKENISATION & FENÊTRE DE CONTEXTE

Le texte est traité token par token — environ 3/4 d'un mot anglais, moins en français. 'Intelligence artificielle' = 4 tokens.

La fenêtre de contexte définit la quantité d'information traitée simultanément. Claude 3.5 : 200 000 tokens. GPT-4o : 128 000 tokens. Au-delà, les informations antérieures disparaissent.

TEMPÉRATURE & GÉNÉRATION

— Température 0 → réponses stables, déterministes, factuelles

— Température 0.7–1 → rédaction créative, brainstorming

— Top-p, top-k : paramètres complémentaires de filtrage du vocabulaire de sortie

HALLUCINATIONS : MÉCANISME & PRÉVENTION

L'hallucination n'est pas un bug : c'est le comportement normal d'un modèle qui génère la suite statistiquement probable même sans information fiable. Il ne 'sait' pas qu'il invente.

Prévention : ancrer avec des données factuelles (RAG), demander le niveau de certitude, ne jamais utiliser un LLM seul comme source unique sur des faits critiques.

À RETENIR

— Un LLM prédit des tokens — il ne comprend pas au sens humain

— La fenêtre de contexte est une contrainte physique, non logicielle

— La température se règle selon la tâche, pas selon le modèle

— L'hallucination se prévient, elle ne s'élimine pas

© Patrick Pécot — Tous droits réservés

Document strictement réservé à l'usage personnel du destinataire. Toute reproduction, diffusion, publication, intégration dans un cours ou une formation, revente ou réutilisation à des fins commerciales ou professionnelles est strictement interdite sans autorisation écrite préalable. Contact : patpec442@gmail.com