

CLAUDE FONDATIONS · FICHE DE RÉVISION

Comprendre l'intelligence artificielle générative

Module 02

L'essentiel à retenir

01. Fonctionnement d'un LLM

Un grand modèle de langage (LLM) ne « comprend » pas le monde comme nous. Il fait une chose, très bien : prédire le mot suivant. Tout le reste en découle.

Points abordés : Le principe : prédire la suite · Les tokens : l'unité de base · Les paramètres et l'attention · Une conséquence importante : le probabilisme

À retenir — Un LLM prédit le token suivant à partir du contexte, grâce à des milliards de paramètres et au mécanisme d'attention. Il produit du plausible, pas du vérifié : gardez-le en tête pour tout usage sérieux.

02. Les phases d'entraînement d'un LLM

Un modèle comme Claude ne naît pas compétent. Il passe par plusieurs phases d'entraînement successives, chacune avec un rôle précis.

Points abordés : 1. Le pré-entraînement · 2. Le fine-tuning (ajustement) · 3. L'apprentissage par retour (RLHF et variantes) · Une conséquence : la date de coupure des connaissances

À retenir — Pré-entraînement (savoir + langage), ajustement (suivre des consignes), apprentissage par retour (utile/honnête/sûr) : trois étapes qui transforment un prédicteur de texte en assistant fiable. Son savoir reste figé à une date de coupure.

03. Les LRM et l'effort de raisonnement

Certains modèles ne se contentent pas de répondre : ils prennent le temps de « réfléchir » avant. C'est l'idée des modèles de raisonnement et de l'effort.

Points abordés : Du LLM au LRM · La notion d'effort · Quand est-ce utile ?

À retenir — Les modèles de raisonnement « pensent » avant de répondre, ce qui améliore les tâches complexes. L'effort est un curseur : plus de réflexion pour les problèmes durs, moins pour les demandes simples.

04. Composition et gestion du contexte

Le contexte, c'est la mémoire de travail du modèle le temps d'une conversation. Bien le gérer change tout dans la qualité des réponses.

Points abordés : Qu'est-ce que le contexte ? · Une fenêtre large, mais pas infinie · Bien composer son contexte · Contexte ≠ mémoire permanente

À retenir — Le contexte est la mémoire de travail momentanée du modèle, mesurée en tokens et limitée. Fournissez l'essentiel, évitez le bruit, et structurez le contexte durable via les instructions et les projets.

05. La spécificité Anthropic : la Constitution IA

C'est la signature d'Anthropic : entraîner le modèle à respecter des principes explicites, plutôt que de dépendre uniquement de jugements humains au cas par cas.

Points abordés : Le problème que ça résout · Le principe de la Constitution IA · Ce que ça change concrètement

À retenir — La Constitution IA entraîne Claude à s'aligner sur des principes écrits qu'il applique à ses propres réponses. Résultat : un modèle qui cherche à être utile, honnête et sûr de manière plus cohérente et transparente.

06. L'économie des tokens et la facturation

Comprendre les tokens, c'est comprendre à la fois les limites techniques et le coût réel de l'IA générative.

Points abordés : Le token, unité de mesure et de coût · Entrée, sortie et raisonnement · Deux logiques de facturation · Optimiser sa consommation

À retenir — Le token est l'unité qui mesure le contexte et le coût. On paie l'entrée, la sortie et (le cas échéant) le raisonnement. Abonnement à quotas côté application, paiement à l'usage côté API : dans les deux cas, le bon modèle et un contexte sobre font la différence.