



Évolutions des LLMs : architectures, contexte et multimodalité

Comment les modèles de langage gagnent en taille de
contexte, en efficacité (MoE) et en capacités multimodales

Multimodal

Vision + Langage

Mixture-of-Experts

Long Context

Raisonnement

Open-weight



Les Large Language Models (LLMs) ont connu une évolution spectaculaire ces dernières années, passant de simples modèles de texte à des systèmes multimodaux sophistiqués capables de traiter images et vidéos (VLM). L'augmentation des tailles de contexte, l'essor du Mixture-of-Experts (MoE) et des modèles de raisonnement, et la montée en puissance de l'open-weight marquent une transformation profonde de ces technologies. Leur extension vers la perception et l'action — World Models et VLA — fait l'objet de pages dédiées dans la section « IA émergentes ».

Évolution des Architectures

- LLM (Large Language Model) : Traitement du texte et génération de langage naturel
- VLM (Vision-Language Model) : Compréhension conjointe d'images, de vidéos et de texte
- Architecture Transformer optimisée pour le long contexte
- Mixture-of-Experts (MoE) : seule une fraction des paramètres (experts) est activée par token, permettant des modèles très grands à coût d'inférence réduit (paradigme dominant 2025-2026)
- Modèles de raisonnement : génération d'une chaîne de pensée ('thinking') avant la réponse finale, pour les tâches mathématiques, logiques et de planification
- Mécanismes d'attention sparse et efficaces

MoE, Raisonnement et Modèles Open-Weight

Les évolutions majeures de 2025-2026 : le Mixture-of-Experts comme architecture dominante, l'essor des modèles de raisonnement, et la montée en puissance de l'open-weight.

Mixture-of-Experts (MoE)

Architecture

Plutôt qu'activer tous les paramètres à chaque token, un MoE route chaque token vers un sous-ensemble d'« experts ». DeepSeek V3/R1 (671 milliards de paramètres au total, ~37 milliards actifs) illustre ce paradigme : capacité massive, coût d'inférence proche d'un modèle bien plus petit. C'est aujourd'hui l'architecture de référence des grands modèles.

Modèles de raisonnement

Raisonnement

Les modèles de raisonnement (OpenAI o1/o3, DeepSeek R1, modes « thinking » intégrés à Gemini, Claude et GPT-5.x) produisent une chaîne de raisonnement explicite avant de répondre. Gains nets sur les mathématiques, le code et la planification multi-étapes, au prix d'une latence et d'un coût de calcul accrus.

Open-weight MoE

Open-weight

L'écart entre modèles fermés et open-weight s'est resserré. DeepSeek V3 puis R1 (janvier 2025, licence MIT) ont marqué un tournant, suivis de V3.2 (déc. 2025) et V4 (avril 2026, contexte 1M natif). Llama 4 (Scout/Maverick), Qwen3 235B et Kimi K2.6 (~1 100 milliards de paramètres) complètent une offre open-weight de niveau frontière, auto-hébergeable.

Course à la frontière (fin 2025)

Frontière

Quelques semaines fin 2025 ont vu se succéder Grok 4.1 (17 nov.), Gemini 3 Pro (18 nov.), Claude Opus 4.5 (24 nov.) et GPT-5.2 (11 déc.), suivis de GPT-5.5 (avril 2026). Ces modèles frontière combinent multimodalité native, raisonnement intégré et contextes de l'ordre du million de tokens.

L'Augmentation des Tailles de Contexte

2020 : 2K tokens

GPT-3 (juin 2020) plafonne à 2 048 tokens, capable de gérer quelques pages de texte

'Conversations courtes, documents limités

2023 : 8K-100K tokens

GPT-4 (mars 2023) introduit 8K en base et 32K en version étendue ; Claude 2 (juillet 2023) atteint 100K tokens (environ 75 000 mots)

'Analyse de documents longs, livres entiers, historique de conversation étendu

2024 : 200K à 1M tokens

Claude 3 (mars 2024) standardise les 200K tokens ;
Gemini 1.5 Pro atteint 1 million de tokens en production
(2 millions en version expérimentale)

'Traitement de codebases complètes, rapports volumineux, datasets massifs

2025-2026 : 1M+ natif en production

Les fenêtres de 1 million de tokens deviennent natives chez les frontières (Gemini 3 Pro, GPT-5.x) comme en open-weight : DeepSeek V4 (avril 2026) propose 1M de contexte natif. Distinction clé entre contexte natif (entraîné) et contexte expérimental (extrapolé).

'Mémoire à long terme, analyse de projets entiers, documentation complète

Impact sur les applications

Permet le RAG (Retrieval-Augmented Generation) avancé, l'analyse de conversations longues, la compréhension contextuelle profonde

'Applications professionnelles complexes, systèmes de mémoire à long terme

Défis techniques

Coût de calcul croissant avec la longueur, gestion de la mémoire GPU, optimisations nécessaires (Flash Attention, Ring Attention, attention sparse)

'Nécessite des innovations en architectures et en hardware

De la Vision au Langage : les VLM

VLM (Vision-Language Models)

Les VLMs combinent la compréhension d'images et de texte dans un seul modèle unifié ; la vision est désormais native dans les modèles frontière. Exemples : GPT-5.x, Claude Opus 4.5, Gemini 3 Pro, ainsi que des modèles open-weight comme Qwen3-VL et LLaVA. Ces modèles peuvent analyser des images, répondre à des questions sur leur contenu, générer des descriptions détaillées, et extraire des informations structurées.

Compréhension contextuelle riche, analyse de documents visuels (PDF, diagrammes, graphiques), OCR avancé, description détaillée d'images

Hallucinations visuelles possibles, précision variable selon la complexité, coût computationnel élevé

Compréhension vidéo (Video Understanding)

Les modèles VLM étendus à la vidéo (Gemini 3 Pro, GPT-5.x) peuvent analyser des vidéos, extraire des événements, répondre à des questions sur le contenu, générer des résumés et même transcrire avec compréhension contextuelle. Utile pour l'analyse de surveillance, l'indexation de contenus, et l'accessibilité.

Indexation intelligente de vidéos, résumés automatiques, recherche sémantique dans les vidéos, génération de sous-titres contextuels

Traitement lent pour longues vidéos, coût élevé, précision variable, difficultés avec actions rapides

Au-delà du langage : les IA émergentes

L'évolution des LLMs vers la perception et l'action ouvre deux frontières qui dépassent le cadre des modèles de langage. Elles font l'objet de pages dédiées dans la section « IA émergentes ».

World Models '

IA émergente

Des modèles capables de générer des environnements 3D interactifs et navigables en temps réel, avec simulation physique (Genie 3 / Project Genie, NVIDIA Cosmos, World Labs). Au-delà de la vidéo : des mondes explorables, utilisés notamment pour entraîner des agents et des véhicules autonomes (Waymo).

[Découvrir les World Models](#)

VLA (Vision-Language-Action) '

IA émergente

Des modèles qui traduisent perception visuelle et instruction en langage naturel directement en actions robotiques. Des fondations (RT-2, OpenVLA) aux modèles 2026 (ÀQ Gemini Robotics, MolmoAct), c'est le pont entre l'IA et la robotique du monde réel. [Découvrir les VLA](#)

Défis Techniques et Perspectives

- Scalabilité : Gérer efficacement des contextes de millions de tokens tout en maintenant la cohérence
- Efficacité énergétique : Réduire la consommation pour des déploiements durables et embarqués
- Multimodalité native : Unifier texte, image, audio et vidéo dans une architecture commune
- Coût d'inférence : Maîtriser le surcoût des modèles de raisonnement et des très grands contextes
- Robustesse : Améliorer la fiabilité face aux entrées adversariales et aux hallucinations
- Explicabilité : Rendre les décisions des modèles multimodaux compréhensibles pour les humains

Modèles et Plateformes Clés

Les acteurs principaux des évolutions LLM/VLM (état juin 2026)

Modèles frontière (multimodaux + raisonnement)

GPT-5.x OpenAI 🚫API payante	Claude Opus 4.5 Anthropic 🚫API payante, vision native
Gemini 3 Pro Google 🚫API payante, contexte ~1M tokens	Grok 4.x xAI 🚫API payante

Open-weight (MoE / raisonnement)

DeepSeek V4 / R1 DeepSeek (MoE, MIT) 🔓Open-weight, auto-hébergeable	Llama 4 Scout/Maverick Meta (MoE) 🔓Open-weight
Qwen3 235B Alibaba (MoE) 🔓Open-weight	Kimi K2.6 Moonshot AI (~1,1T, MoE) 🔓Open-weight

Sources et Ressources

- Wikipedia - Large Language Model
https://en.wikipedia.org/wiki/Large_language_model
- OpenAI - Introducing GPT-5.5 (2026)
<https://openai.com/index/introducing-gpt-5-5/>
- BentoML - The Complete Guide to DeepSeek Models: from V3 to R1 and Beyond
<https://www.bentoml.com/blog/the-complete-guide-to-deepseek-models-from-v3-to-r1-and-beyond>
- Hugging Face - The State of Open-Source LLMs (MoE : Llama 4, Qwen3, Kimi K2.6)
<https://huggingface.co/blog/daya-shankar/open-source-llms>
- GPT-4 Technical Report (OpenAI, 2023)
<https://arxiv.org/abs/2303.08774>
- Maxim - Gemini 3 Pro vs Claude Opus 4.5 vs GPT-5 : Frontier Model Comparison (2026)
<https://www.getmaxim.ai/articles/gemini-3-pro-vs-claude-opus-4-5-vs-gpt-5-the-ultimate-frontier-model-comparison/>
- Google - Gemini 1.5: Unlocking multimodal understanding across millions of tokens (2024)
<https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/>
- Flash Attention: Fast and Memory-Efficient Exact Attention (2022)
<https://arxiv.org/abs/2205.14135>