Llama 4 Maverick : 10 millions de tokens de contexte, la révolution open-source de Meta

Llama 4 Maverick, publié par Meta AI en début 2026, redéfinit les limites de ce que l'open-source peut accomplir dans le domaine des grands modèles de langage. Avec une fenêtre de contexte de 10 millions de tokens — un record absolu mondial dans la catégorie des modèles open-source et open-weight — ce modèle permet pour la première fois d'analyser une codebase entière, un livre complet, des milliers de documents contractuels ou une semaine de logs système en une seule requête. Son architecture Mixture of Experts (MoE) avec 402 milliards de paramètres totaux mais seulement 17 milliards actifs à chaque inférence lui confère une efficacité computationnelle remarquable, permettant d'atteindre des performances de premier rang mondial à un coût d'inférence bien inférieur à ce qu'implique sa taille totale. Avec un ELO LM Arena de 1 451, un MMLU de 91,8 % et un GPQA Diamond de 87,8 %, Llama 4 Maverick s'impose comme l'alternative open-source la plus crédible aux modèles fermés pour les cas d'usage nécessitant un contexte très long et une compréhension générale de haut niveau.

Llama 4 Maverick dans le classement mondial IA de juillet 2026

Retrouvez la position de Llama 4 Maverick dans notre classement complet des LLM de juillet 2026, qui analyse plus de 40 modèles sur l'ensemble des benchmarks académiques et industriels

de référence. Llama 4 Maverick occupe la quatrième position des modèles open-source mondiaux avec un ELO de 1 451, derrière Qwen3.7 Max (1 475) et GLM-5.2 (1 465), mais devant DeepSeek V4 Pro Max (1 449), Mistral Large 3 (1 443) et Gemma 4 31B de Google (1 441).

Cette position est d'autant plus remarquable que Llama 4 Maverick est conçu pour exceller sur un cas d'usage très spécifique — le traitement de contextes extrêmement longs — tout en maintenant des performances générales de premier ordre. L'ELO de 1 451 place Llama 4 Maverick à seulement 56 points en dessous de Claude Fable 5 (1 507), le modèle leader mondial, et à 44 points du Grok 4 d'xAI (1 495). Pour un modèle open-source entièrement déployable on-premise, cet écart minimal représente une réussite technique considérable de la part des équipes de Meta AI.

Le score BenchLM composite de 74,22 confirme la cohérence des performances de Llama 4 Maverick à travers les différentes dimensions d'évaluation. Sur les benchmarks spécialisés dans la compréhension de longs documents, Llama 4 Maverick n'a pas de concurrent sérieux dans la catégorie open-source : sa capacité à maintenir une compréhension cohérente sur plusieurs millions de tokens est une caractéristique unique sur le marché en juillet 2026.

Scores de référence et benchmarks de Llama 4 Maverick

Les performances de Llama 4 Maverick ont été évaluées sur l'ensemble des benchmarks standardisés, révélant un modèle d'excellence générale avec une spécialisation marquée sur les tâches nécessitant une compréhension approfondie et une vaste base de connaissances.

Métrique	Llama 4 Maverick	Contexte
ELO LM Arena	**1 451**	Top 4 open-source mondial
BenchLM composite	**74,22**	Top 6 mondial
GPQA Diamond	**87,8 %**	Questions niveau doctorat
MMLU	**91,8 %**	Connaissances encyclopédiques
Fenêtre de contexte	**10 000 000 tokens**	Record mondial open-source
Vitesse d'inférence	**110 tok/s**	Sur configuration multi-GPU
Prix API (entrée/sortie)	**0,20 $ / 0,60 $ / M tokens**	Meta AI API / Together AI
Paramètres	**402B total / 17B actifs (MoE)**	128 experts, efficacité MoE
Licence	**Llama 4 Community**	Libre < 700M MAU

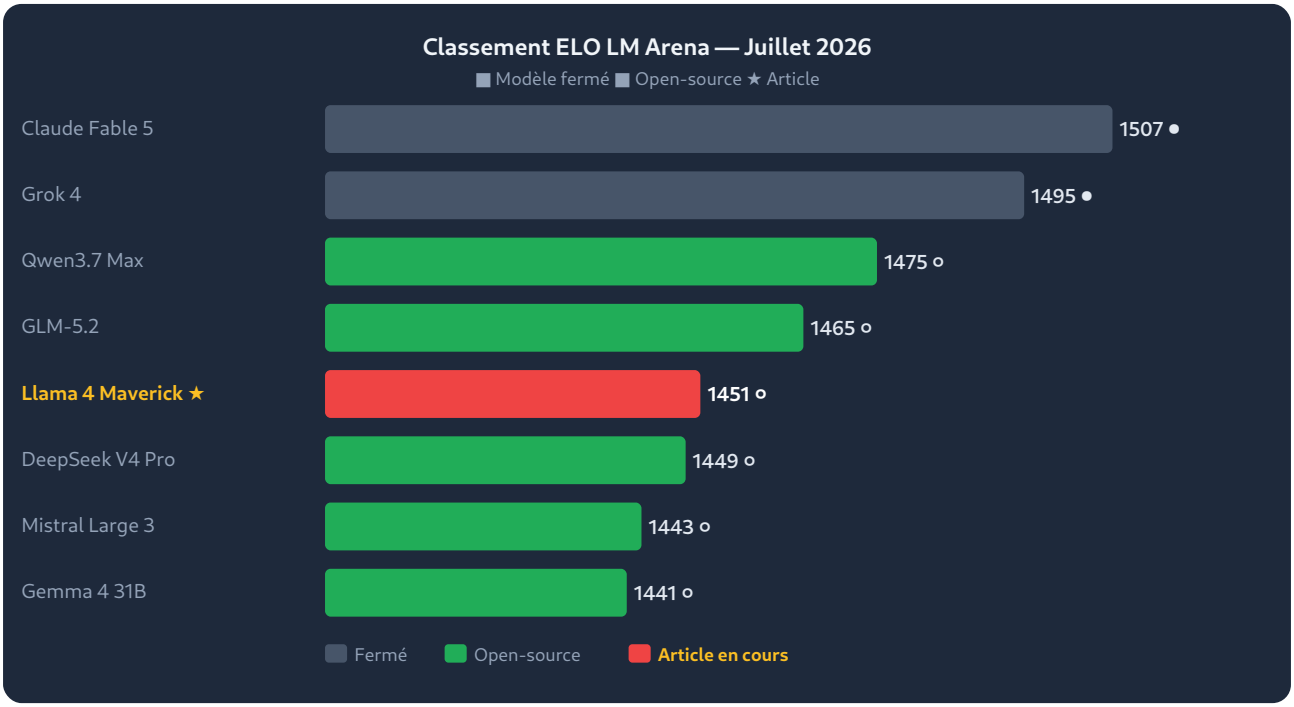
Architecture MoE et innovations techniques de Llama 4 Maverick

L'architecture Mixture of Experts (MoE) de Llama 4 Maverick constitue l'innovation centrale qui permet d'atteindre des performances de pointe avec une efficacité computationnelle remarquable. Le modèle dispose de 128 experts spécialisés, mais seuls 17 milliards de paramètres sont activés pour chaque token traité — les 17B actifs correspondent à environ deux experts sélectionnés dynamiquement parmi les 128 disponibles. Ce mécanisme de routage intelligent, entraîné de bout en bout avec le reste du modèle, garantit que les tokens sont dirigés vers les experts les plus compétents pour chaque contexte, optimisant simultanément la qualité des réponses et l'efficacité computationnelle.

La réussite technique majeure de Llama 4 Maverick est l'extension de la fenêtre de contexte à 10 millions de tokens tout en maintenant des performances de qualité à travers l'ensemble de cette fenêtre. Y parvenir nécessite des innovations architecturales profondes : une variante de

l'attention multi-tête (Grouped Query Attention ou GQA) hautement optimisée pour les longues séquences, un système de positionnement relatif (RoPE — Rotary Position Embedding) étendu bien au-delà de ce que les implémentations standard supportent, et des optimisations de mémoire avancées (Ring Attention, Flash Attention 3) pour maintenir l'ensemble de la fenêtre de contexte dans la VRAM GPU. En pratique, exploiter la totalité des 10 millions de tokens nécessite une configuration multi-GPU conséquente (typiquement 8 GPU H100 en parallélisme tensoriel).

Meta AI a publié les poids du modèle sous la Llama 4 Community License, qui autorise l'usage commercial pour toute organisation dont le service atteint moins de 700 millions d'utilisateurs actifs mensuels — seuil qui couvre la quasi-totalité des entreprises mondiales, à l'exception des plus grandes plateformes comme Google, Facebook ou TikTok eux-mêmes. Cette licence permet également le fine-tuning, la création de dérivés et leur déploiement commercial, avec l'obligation de signaler l'origine Llama 4 dans les dérivés distribués. Comparée à la licence Apache 2.0 (plus permissive) de Gemma 4 31B ou Qwen3, la Llama 4 Community License est légèrement plus restrictive mais reste très adaptée aux usages d'entreprise standard.



Guide de déploiement et d'accès à Llama 4 Maverick

Llama 4 Maverick est disponible via plusieurs canaux en juillet 2026. L'accès le plus immédiat passe par les plateformes API tierces qui hébergent le modèle : Together AI propose une API compatible OpenAI avec une tarification de 0,20 dollar par million de tokens en entrée et 0,60 dollar en sortie, avec des garanties de temps de réponse excellentes grâce à leur infrastructure optimisée. Groq propose également Llama 4 Maverick sur ses puces LPU propriétaires, avec des vitesses d'inférence spectaculaires mais une disponibilité limitée au contexte standard. Replicate est une autre option accessible avec un modèle de facturation à l'utilisation.

Pour accéder directement aux poids du modèle en vue d'un déploiement self-hosted, Hugging Face héberge les fichiers sous l'identifiant `meta-llama/Llama-4-Maverick-17B-128E-Instruct`. L'accès aux poids nécessite d'accepter la Llama 4 Community License sur la page Hugging Face et de disposer d'un token d'accès authentifié. Le déploiement avec vLLM sur un cluster multi-GPU se configure comme suit :

```
 pip install vllm
# Déploiement sur 8xH100 ou 8xA100 (contexte complet 10M tokens)
python -m vllm.entrypoints.openai.api_server   --model meta-llama/Llama-4-Maverick-17B-128E-Instru
```

Pour un déploiement avec un contexte réduit (suffisant pour 95 % des cas d'usage) sur une configuration à 2 ou 4 GPU, il est possible de limiter `max-model-len` à 128 000 ou 500 000 tokens, réduisant drastiquement les besoins en VRAM tout en maintenant l'intégralité des capacités générales du modèle. La quantification INT8 via bitsandbytes permet également de réduire les besoins mémoire d'environ 40 % avec une perte de performance minime.

La Meta AI API officielle est disponible pour les développeurs partenaires et progressivement ouverte au grand public. Together AI reste la plateforme de référence pour l'accès API tiers, avec des SLA clairement définis et une facturation transparente. Pour les entreprises européennes, Scaleway (France) propose également Llama 4 Maverick avec des garanties de souveraineté des données en Europe, une option particulièrement intéressante pour les organisations soumises au RGPD qui ne souhaitent pas gérer l'infrastructure GPU elles-mêmes.

Performance en français : évaluation détaillée de Llama 4 Maverick

Llama 4 Maverick atteint un niveau B2 en français, ce qui le positionne comme un modèle capable de traiter la grande majorité des textes professionnels francophones avec une qualité acceptable. Meta AI a investi significativement dans le multilinguisme avec la famille Llama 4, incluant de larges corpus en français dans les données d'entraînement. En pratique, Llama 4 Maverick produit des textes grammaticalement corrects, comprend les instructions en français sans difficulté et génère des réponses pertinentes sur des sujets généraux et techniques.

Les tests de performance en français révèlent cependant quelques limites par rapport à des modèles comme Gemma 4 31B ou les modèles Mistral (entreprise française). Sur des tâches nécessitant une maîtrise fine des subtilités stylistiques françaises — rédaction formelle administrative, nuances juridiques, registres soutenu et familier — Llama 4 Maverick peut produire des formulations qui trahissent une légère préférence pour les constructions anglaises. Pour comparer les deux modèles Google et Meta sur le français, consultez notre article sur Gemma 4 31B, qui atteint un niveau légèrement supérieur (B2-C1) pour les tâches en français grâce à un entraînement multilingue plus équilibré.

En revanche, pour les cas d'usage où le contexte en français est très long — analyse d'un ensemble de contrats français, revue d'une documentation réglementaire volumineuse en français, analyse de transcriptions de réunions —, Llama 4 Maverick n'a aucun concurrent open-source sérieux. Sa capacité à maintenir la cohérence et la pertinence sur des millions de tokens de contexte compense largement les petites imperfections stylistiques pour ce type d'application.

Cas d'usage détaillés pour Llama 4 Maverick

Analyse de codebase entière : Le cas d'usage emblématique de Llama 4 Maverick est l'analyse d'une codebase logicielle complète dans une seule requête. Avec 10 millions de tokens de contexte, il est possible d'inclure les fichiers source d'une application de taille moyenne à grande (plusieurs centaines de milliers de lignes de code) et de demander une revue de sécurité, une

analyse d'architecture, une recherche de vulnérabilités ou une documentation automatique. Cette capacité transforme radicalement les workflows de revue de code et d'audit de sécurité logicielle.

Analyse juridique et contractuelle à grande échelle : Les cabinets d'avocats et les directions juridiques d'entreprise peuvent charger des centaines de contrats simultanément pour des analyses transversales : identification de clauses contradictoires entre contrats, recherche de risques spécifiques à travers l'ensemble d'un portefeuille contractuel, uniformisation des conditions générales. Une tâche qui nécessitait auparavant des dizaines d'heures de travail humain peut être complétée en minutes avec Llama 4 Maverick.

Audit de conformité réglementaire : Pour la conformité NIS 2, ISO 27001 ou DORA, Llama 4 Maverick peut ingérer simultanément l'ensemble des politiques de sécurité d'une organisation, ses procédures opérationnelles, ses logs de sécurité et le texte complet de la réglementation applicable, puis générer un rapport d'écarts précis et hiérarchisé. Le contexte illimité de fait élimine le besoin de chunking et de récupération (RAG) pour ces analyses, réduisant les risques d'omission.

Recherche scientifique et synthèse bibliographique : Les chercheurs peuvent charger des centaines d'articles scientifiques simultanément pour des revues de littérature automatisées, des méta-analyses ou des détections de contradictions entre publications. Le score GPQA Diamond de 87,8 % confirme la capacité du modèle à raisonner à un niveau expert sur des questions scientifiques complexes, ce qui en fait un outil puissant pour l'accélération de la recherche.

Analyse de logs et monitoring opérationnel : Les équipes SecOps et DevOps peuvent utiliser Llama 4 Maverick pour analyser des volumes massifs de logs système, réseau et applicatifs dans un seul contexte, identifiant des patterns anormaux, corrélant des événements à travers de longues périodes temporelles et distinguant les incidents réels des faux positifs. La fenêtre de 10 millions de tokens couvre plusieurs jours de logs denses pour les systèmes d'entreprise typiques.

Tarification et accès à Llama 4 Maverick en 2026

La structure tarifaire de Llama 4 Maverick est l'une des plus compétitives du marché premium des LLM. Via Together AI, la plateforme API de référence pour Llama 4 Maverick, la tarification

s'établit à 0,20 dollar par million de tokens en entrée et 0,60 dollar en sortie — des prix qui la positionnent parmi les offres les plus économiques pour un modèle de ce niveau de performance. En comparaison, Claude Fable 5 (ELO 1 507, 56 points au-dessus) est typiquement facturé 4 à 8 fois plus cher pour des volumes comparables.

L'accès au modèle via Groq présente une tarification différente mais offre des vitesses d'inférence incomparables grâce aux puces LPU propriétaires de Groq. Pour les applications nécessitant des réponses en temps réel (chatbots, assistants interactifs, génération de code en temps réel), Groq est souvent le premier choix malgré une disponibilité plus variable. Replicate propose une tarification à la seconde GPU, adaptée aux charges de travail intermittentes ou aux expérimentations.

Pour l'auto-hébergement en production avec la fenêtre de contexte complète de 10 millions de tokens, les besoins matériels sont conséquents : 8 GPU NVIDIA H100 80 Go (ou A100 80 Go) en parallélisme tensoriel sont nécessaires pour maintenir le modèle complet en BF16 en mémoire avec suffisamment de KV-cache pour les longs contextes. Le coût de location cloud pour cette configuration se situe entre 25 et 50 dollars de l'heure, ce qui peut être économiquement justifié pour les organisations traitant des volumes très importants ou ayant des contraintes de confidentialité absolues.

Pour les budgets plus contraints, l'utilisation avec un contexte limité à 128 000 tokens (similaire à Gemma 4 31B) peut être réalisée sur une configuration à 2-4 GPU A100, avec des coûts d'infrastructure de 8 à 20 dollars de l'heure. Dans ce cas, les avantages spécifiques de Llama 4 Maverick (contexte extrêmement long) ne sont pas pleinement exploités, et Gemma 4 31B peut s'avérer un meilleur choix en termes de rapport coût-performance.

Comparaison avec les modèles open-source concurrents

Face à **Gemma 4 31B de Google** (ELO 1 441), Llama 4 Maverick présente un profil complémentaire : supérieur sur MMLU (91,8 % vs score plus modeste de Gemma) et sur les très longs contextes, inférieur en termes de facilité de déploiement et de coût pour les usages standard. Pour les organisations dont le cas d'usage principal nécessite une fenêtre de contexte

au-delà de 128 000 tokens, Llama 4 Maverick n'a pas d'équivalent open-source sérieux. Pour les déploiements sur infrastructure limitée (1 GPU), Gemma 4 31B reste le choix naturel.

Face à **Qwen3.7 Max d'Alibaba** (ELO 1 475, 24 points au-dessus), Llama 4 Maverick se distingue principalement par son contexte de 10 millions de tokens contre 128 000 tokens pour Qwen3.7 Max. Sur les benchmarks généraux et de mathématiques, Qwen3.7 Max est légèrement supérieur. Le choix entre les deux dépend essentiellement des cas d'usage : Qwen3.7 Max pour les performances générales maximales avec un contexte standard, Llama 4 Maverick pour les analyses de très longues séquences.

Face à **DeepSeek V4 Pro Max** (ELO 1 449, licence MIT), Llama 4 Maverick offre une alternative développée par un acteur occidental avec une documentation de déploiement plus complète et une communauté plus large. DeepSeek se distingue sur les benchmarks de mathématiques et de raisonnement symbolique, tandis que Llama 4 Maverick excelle sur la compréhension de longs contextes et sur MMLU. Les organisations préférant les modèles d'origine américaine pour des raisons de conformité géopolitique ou de politique d'achat trouveront dans Llama 4 Maverick leur première alternative à DeepSeek.

Comparé aux **modèles fermés**, Llama 4 Maverick représente l'alternative open-source la plus crédible à Claude Fable 5 pour les analyses de très longs documents. Claude Fable 5 dispose d'une fenêtre de contexte de 200 000 tokens — très inférieure aux 10 millions de Llama 4 Maverick — et d'un ELO de 1 507 (56 points au-dessus), mais à un coût d'API nettement plus élevé et sans possibilité de déploiement on-premise. Pour les organisations dont le budget est contraint et le contexte long est critique, Llama 4 Maverick est la solution de choix.

Limites et points de vigilance pour Llama 4 Maverick

La limitation la plus évidente de Llama 4 Maverick est son exigence en infrastructure pour un déploiement self-hosted complet. Les 402 milliards de paramètres totaux du modèle nécessitent une configuration multi-GPU substantielle pour l'hébergement, ce qui le rend inaccessible pour les petites organisations ou les équipes sans budget infrastructure dédié. En pratique, l'utilisation via des API tierces (Together AI, Groq) reste la solution la plus accessible pour la grande

majorité des entreprises, mais elle implique que les données transitent par des serveurs tiers — ce qui peut être problématique dans certains contextes réglementaires.

La fenêtre de contexte de 10 millions de tokens est une capacité théorique dont l'exploitation pratique est limitée par la mémoire GPU disponible. Pour utiliser réellement des contextes de plusieurs millions de tokens, il faut des GPU avec suffisamment de VRAM pour stocker le KV-cache correspondant — ce qui représente des dizaines à centaines de gigaoctets pour les contextes les plus longs. En pratique, la plupart des déploiements self-hosted utilisent des contextes de 128 000 à 500 000 tokens, qui restent bien supérieurs aux autres modèles open-source mais loin de la capacité maximale théorique.

Sur la performance en français, Llama 4 Maverick (niveau B2) se situe légèrement en dessous de Gemma 4 31B (B2-C1) et nettement en dessous des modèles Mistral AI pour les tâches nécessitant une maîtrise fine des nuances linguistiques françaises. Cette limitation est particulièrement perceptible sur la rédaction formelle administrative et juridique, où les subtilités stylistiques comptent autant que la précision factuelle. Pour les organisations dont l'usage principal est en français, un benchmark interne sur des données représentatives est fortement recommandé avant tout déploiement en production.

La Llama 4 Community License, bien que généreuse, est légèrement plus restrictive que l'Apache 2.0 appliquée à Gemma 4 31B ou Mistral Large 3. La limite de 700 millions d'utilisateurs actifs mensuels et l'obligation de signaler l'origine Llama 4 dans les dérivés distribués sont des contraintes marginales pour la plupart des entreprises, mais méritent d'être vérifiées par les équipes juridiques avant déploiement, en particulier pour les éditeurs de logiciels qui redistribuent des produits basés sur le modèle.

À RETENIR

À retenir

10 millions de tokens de contexte — un record absolu mondial open-source qui transforme les possibilités d'analyse documentaire : codebase entière, bibliothèque de contrats, corpus réglementaire complet en une seule requête.

MMLU 91,8 % et ELO 1 451 — performances de premier rang mondial, dans le top 4 des modèles open-source, avec seulement 17 milliards de paramètres actifs grâce à l'architecture MoE.

Architecture MoE 128 experts (17B actifs) — efficacité computationnelle remarquable : la puissance d'un modèle 400B avec le coût d'inférence d'un modèle 17B, rendant les usages à grande échelle économiquement viables.

Llama 4 Community License — usage commercial libre pour toute organisation sous 700 millions d'utilisateurs actifs mensuels, permettant le déploiement en production, le fine-tuning et la redistribution de dérivés.

API à 0,20 $/0,60 $ par M tokens — tarification parmi les plus compétitives du marché pour un modèle de ce niveau, disponible via Together AI, Groq, Replicate et Meta AI API.

Alternative open-source à Claude Fable 5 pour les longs contextes — ELO à seulement 56 points de Claude Fable 5, mais avec un contexte 50 fois plus grand (10M vs 200K tokens) et une licence qui autorise l'hébergement on-premise pour les données sensibles.

Foire aux questions sur Llama 4 Maverick

Qu'est-ce qui rend le contexte de 10 millions de tokens de Llama 4 Maverick révolutionnaire ?

Dix millions de tokens correspondent à environ 7,5 millions de mots, soit l'équivalent de 40 à 50 livres entiers, d'une codebase logicielle de plusieurs centaines de milliers de lignes, ou de plusieurs années de logs système. Avant Llama 4 Maverick, les modèles open-source étaient limités à des contextes de 128 000 tokens au maximum. Cette limitation forçait les développeurs à utiliser des techniques complexes de RAG (Retrieval-Augmented Generation) pour traiter de longs documents, avec les risques d'omission et la complexité d'implémentation que cela implique. Llama 4 Maverick élimine ce besoin pour la majorité des cas d'usage, permettant une compréhension globale et cohérente sans les artefacts introduits par le chunking et la récupération vectorielle.

Comment l'architecture Mixture of Experts (MoE) de Llama 4 Maverick fonctionne-t-elle concrètement ?

Dans l'architecture MoE de Llama 4 Maverick, le modèle dispose de 128 experts — des sous-réseaux de neurones spécialisés dans différents types de tâches ou de connaissances. Pour chaque token traité, un mécanisme de routage (lui-même un réseau de neurones entraînable) sélectionne les 2 experts les plus pertinents parmi les 128 et envoie le token uniquement à ces experts pour traitement. Résultat : seuls 17 milliards de paramètres (correspondant approximativement à 2 experts) sont activés à chaque inférence, même si le modèle total contient 402 milliards de paramètres. Cela réduit le coût computationnelle d'un facteur 10 à 20 par rapport à un modèle dense équivalent, tout en bénéficiant de la richesse d'un ensemble de paramètres beaucoup plus large lors de l'entraînement.

Llama 4 Maverick est-il utilisable pour des applications commerciales en France ?

Oui, la Llama 4 Community License autorise l'usage commercial pour toute organisation dont le produit ou service ne dépasse pas 700 millions d'utilisateurs actifs mensuels — seuil qui concerne seulement une poignée de plateformes mondiales. Pour la grande majorité des

entreprises françaises, Llama 4 Maverick peut être utilisé librement en production, fine-tuné sur des données propriétaires et intégré dans des produits commerciaux. Les obligations principales sont d'inclure la mention "Built with Llama" dans les produits redistribués et de respecter les politiques d'utilisation acceptable de Meta (qui excluent notamment les usages malveillants, la désinformation ou la génération de contenus illicites). Une vérification juridique est recommandée pour les cas d'usage sensibles, mais la grande majorité des applications d'entreprise standard sont clairement couvertes.

Quels sont les prérequis GPU minimaux pour auto-héberger Llama 4 Maverick ?

Les prérequis varient selon la longueur de contexte souhaitée. Pour un contexte de 32 000 tokens (usage courant), un seul GPU NVIDIA A100 80 Go en INT8 est théoriquement suffisant pour le modèle de base, mais les performances seront limitées. Pour un contexte de 128 000 tokens avec des performances acceptables, 2 à 4 GPU A100 80 Go en parallélisme tensoriel sont recommandés. Pour exploiter le contexte complet de 10 millions de tokens en production, 8 GPU NVIDIA H100 80 Go constituent la configuration de référence. Une alternative économique est l'utilisation via les API de Together AI ou Groq pour les charges de travail ne nécessitant pas un hébergement on-premise, ce qui élimine complètement les besoins en infrastructure GPU propre et permet de payer uniquement les tokens consommés.

Llama 4 Maverick peut-il remplacer un système RAG (Retrieval-Augmented Generation) existant ?

Pour de nombreux cas d'usage, oui. Les systèmes RAG ont été développés comme solution de contournement aux contextes limités des LLM, permettant de récupérer dynamiquement les portions pertinentes d'une base documentaire avant chaque requête. Llama 4 Maverick, avec son contexte de 10 millions de tokens, permet souvent de charger directement l'intégralité de la base documentaire en contexte, éliminant la complexité du pipeline RAG (embedding, recherche vectorielle, chunking, récupération) et les risques d'omission d'informations pertinentes. Cependant, pour des bases documentaires très larges (dizaines de gigaoctets de texte) ou des cas où la latence de récupération est critique, le RAG reste pertinent. La décision dépend du volume de documents, des exigences de latence et du budget infrastructure.

Comment Llama 4 Maverick se positionne-t-il face aux modèles d'OpenAI pour l'analyse de code ?

Pour l'analyse de code, Llama 4 Maverick présente un avantage structurel majeur sur GPT-4o et o3 d'OpenAI : son contexte de 10 millions de tokens permet d'inclure l'intégralité d'une codebase dans une seule requête, tandis que GPT-4o est limité à 128 000 tokens et o3 à 200 000 tokens. Sur les benchmarks de codage purs (HumanEval, SWE-bench), les modèles o3 d'OpenAI restent légèrement supérieurs, mais l'avantage du contexte long de Llama 4 Maverick peut être décisif pour des tâches d'audit de code à grande échelle, de détection de vulnerabilités transversales ou de refactorisation architecturale nécessitant une vision globale du projet. La licence open-source de Llama 4 Maverick permet également son intégration directe dans des pipelines CI/CD on-premise sans exposer le code source à des API externes.

