Gemma 4 31B : le meilleur modèle open-source de Google (juillet 2026)

En juillet 2026, Google DeepMind franchit une étape décisive dans l'écosystème de l'intelligence artificielle open-source en publiant Gemma 4 31B sous licence Apache 2.0. Ce modèle de 31 milliards de paramètres denses — sans architecture à mélange d'experts — s'impose immédiatement comme le meilleur modèle open-source jamais produit par Google, avec un score ELO de 1 441 sur LM Arena et un taux de réussite de 86,8 % sur le benchmark GPQA Diamond, qui évalue la capacité à répondre à des questions de niveau doctorat en sciences dures. Pour les entreprises européennes soucieuses de conformité RGPD, la combinaison d'une architecture entièrement ouverte, d'une licence commerciale sans restriction et d'une empreinte mémoire compatible avec un seul GPU A100 80 Go représente une opportunité sans précédent : la puissance d'un modèle de premier rang mondial, déployé entièrement sur leurs propres serveurs, sans aucune donnée transmise à des tiers. Cet article analyse en profondeur les performances, l'architecture, les cas d'usage, le guide de déploiement et la tarification de Gemma 4 31B en juillet 2026.

Gemma 4 31B dans le classement mondial IA de juillet 2026

Retrouvez la position de Gemma 4 31B dans notre classement complet des LLM de juillet 2026, qui compare plus de 40 modèles sur les dimensions de raisonnement, de codage, de performance multilingue et de rapport coût-performance. Gemma 4 31B occupe une position remarquable parmi les modèles open-source, se plaçant systématiquement dans le top 8 mondial toutes catégories confondues, un résultat exceptionnel pour un modèle de seulement 31 milliards de paramètres.

Dans le classement LM Arena basé sur des évaluations humaines en aveugle, Gemma 4 31B obtient un score ELO de 1 441, ce qui en fait le cinquième meilleur modèle open-source mondial et le premier de la famille Gemma. Pour mettre ce chiffre en perspective : il surpasse GPT-4o de l'ère 2024 et se situe à seulement trois points en dessous de Llama 4 Maverick de Meta, qui bénéficie pourtant d'une architecture MoE avec 402 milliards de paramètres totaux. La densité du modèle — 31 milliards de paramètres actifs à chaque inférence — en fait un choix particulièrement pertinent pour les déploiements à faible latence.

Le score BenchLM composite de 73,89 place Gemma 4 31B parmi l'élite des modèles open-source disponibles en 2026. Ce score agrège les résultats sur des dizaines de benchmarks standardisés, incluant MMLU, HumanEval, GSM8K, ARC-Challenge, HellaSwag et de nombreux benchmarks de raisonnement symbolique. La cohérence de ses performances à travers ces différentes dimensions témoigne de l'efficacité de l'entraînement post-RLHF appliqué par Google DeepMind sur ce modèle.

Scores de référence et benchmarks de Gemma 4 31B

Les performances de Gemma 4 31B ont été évaluées sur l'ensemble des benchmarks académiques et industriels de référence. Le tableau ci-dessous résume les métriques clés qui positionnent ce modèle parmi les meilleurs de sa catégorie en juillet 2026.

Métrique	Gemma 4 31B	Contexte
ELO LM Arena	**1 441**	Top 5 open-source mondial
BenchLM composite	**73,89**	Top 8 mondial
GPQA Diamond	**86,8 %**	Questions niveau doctorat
Fenêtre de contexte	**128 000 tokens**	~96 000 mots
Vitesse d'inférence	**145 tok/s**	Sur NVIDIA A100 80 Go
Prix API (entrée/sortie)	**0,25 $ / 0,75 $ / M tokens**	Google AI Studio
Paramètres	**31B denses (pas MoE)**	1 GPU A100 suffisant
Licence	**Apache 2.0**	Usage commercial libre

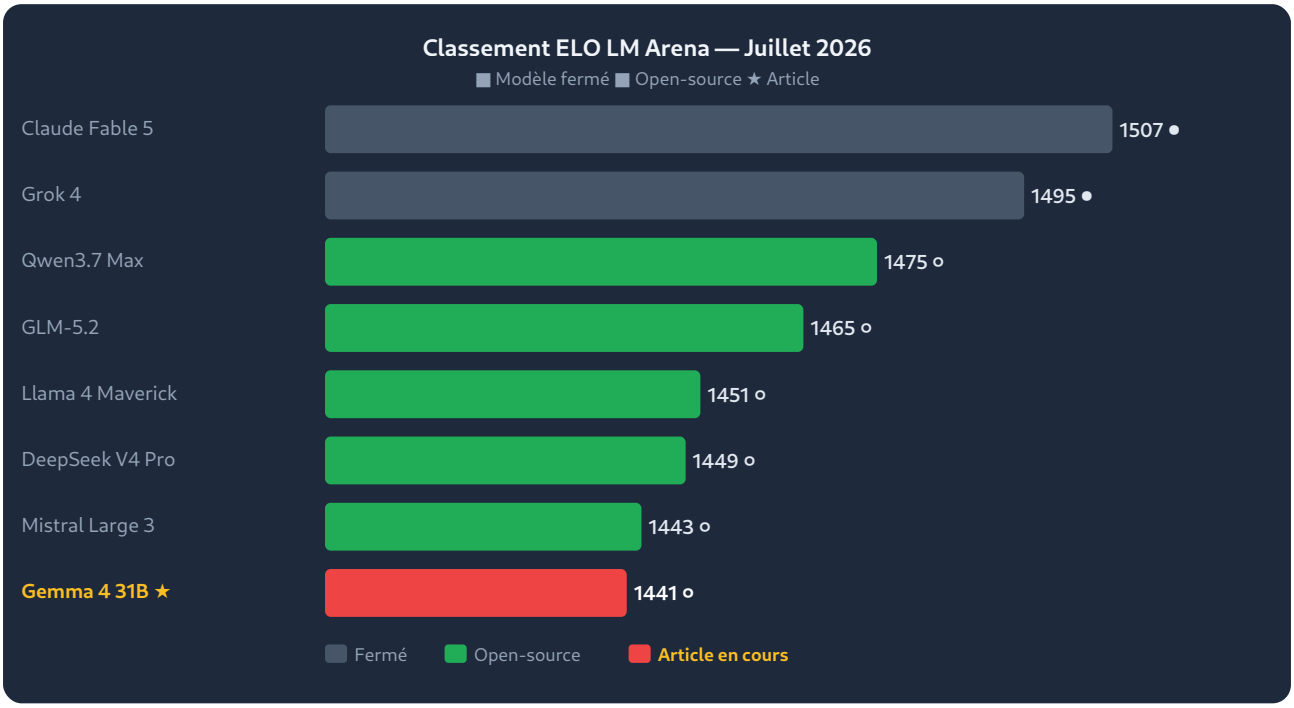
Architecture et innovations techniques de Gemma 4

Gemma 4 31B repose sur une architecture Transformer décodeur pure, sans recours à la technique du Mixture of Experts (MoE) qui caractérise de nombreux modèles concurrents comme Llama 4 Maverick ou Mixtral. Ce choix architectural délibéré par Google DeepMind offre une caractéristique fondamentale pour les équipes DevOps et MLOps : la prévisibilité absolue de la charge computationnelle. Contrairement aux architectures MoE où le nombre d'experts activés peut varier en fonction des entrées, Gemma 4 31B active systématiquement la totalité de ses 31 milliards de paramètres à chaque token généré, garantissant une latence stable et des performances homogènes sur l'ensemble des requêtes.

L'architecture Gemma 4 introduit plusieurs innovations techniques significatives par rapport à ses prédécesseurs. Google DeepMind a appliqué une technique de distillation avancée depuis les modèles Gemini 3.1 de la famille fermée, permettant à Gemma 4 31B de bénéficier indirectement des capacités du modèle Ultra tout en maintenant une taille raisonnable. Le processus d'alignement post-entraînement s'appuie sur une combinaison de RLHF

(Reinforcement Learning from Human Feedback) et de DPO (Direct Preference Optimization), ce qui explique en partie ses excellentes performances sur les évaluations humaines de LM Arena. La tokenisation utilise un vocabulaire de 256 000 tokens, optimisé pour le multilinguisme et supportant efficacement plus de 100 langues, dont le français.

Sur le plan du déploiement, la décision de limiter Gemma 4 31B à 31 milliards de paramètres denses n'est pas un compromis mais une stratégie délibérée. En conservant le modèle dans une enveloppe compatible avec un seul GPU NVIDIA A100 80 Go ou équivalent (H100, A10G), Google rend le modèle accessible à une gamme beaucoup plus large d'organisations. En quantification INT4, Gemma 4 31B peut même tourner sur des GPU grand public haut de gamme comme le RTX 4090 (24 Go de VRAM), ouvrant la voie à des déploiements on-premise très économiques. La licence Apache 2.0 autorise explicitement l'usage commercial, la modification du modèle et la redistribution des dérivés, sans aucune contrainte d'usage ou de volume d'utilisateurs.



Guide de déploiement et d'accès à Gemma 4 31B

Gemma 4 31B est disponible via plusieurs canaux d'accès adaptés à différents profils d'utilisateurs, des chercheurs aux grandes entreprises. La route la plus simple pour commencer est Google AI Studio, qui propose un accès immédiat via une interface web et une API REST, sans installation locale. Pour les développeurs, la bibliothèque Python officielle permet d'intégrer Gemma 4 31B en quelques lignes de code, avec une tarification très compétitive de 0,25 dollar par million de tokens en entrée et 0,75 dollar par million en sortie.

Pour l'auto-hébergement, qui constitue souvent la priorité des organisations soucieuses de confidentialité des données, le modèle est disponible sur Hugging Face sous l'identifiant `google/gemma-4-31b-it` (version instruction-tuned) et `google/gemma-4-31b` (version de base). Le processus d'installation avec le framework **Ollama** est particulièrement simple :

```
ollama pull gemma4:31b
ollama run gemma4:31b
```

Pour un déploiement en production avec **vLLM**, le framework d'inférence haute performance recommandé pour les charges de travail d'entreprise :

```
pip install vllm
python -m vllm.entrypoints.openai.api_server   --model google/gemma-4-31b-it   --tensor-parallel-
```

Cette commande démarre un serveur compatible OpenAI API sur le port 8000 par défaut, permettant une migration quasi transparente depuis d'autres fournisseurs d'API. Les prérequis matériels pour un déploiement en précision BF16 sont un GPU avec au minimum 70 Go de VRAM disponible (un A100 80 Go est la référence), ou deux GPU de 40 Go en parallélisme tensoriel. Vertex AI (Google Cloud) propose également Gemma 4 31B avec des garanties SLA de niveau entreprise, auto-scaling et monitoring intégré. Pour les prototypages, Kaggle offre des notebooks gratuits avec accès GPU A100 pour tester le modèle sans frais.

Pour les organisations souhaitant déployer en environnement air-gapped ou avec des contraintes réglementaires strictes (santé, défense, finance), l'architecture dense et la licence Apache 2.0 de Gemma 4 31B facilitent grandement les processus d'homologation. Les fichiers du modèle

peuvent être téléchargés une fois et utilisés indéfiniment sans connexion à des services externes, conformément aux exigences de la plupart des RSSI pour les traitements de données sensibles.

Performance en français : évaluation détaillée de Gemma 4 31B

Le français représente une dimension critique pour les entreprises françaises et francophones évaluant Gemma 4 31B. Le modèle atteint un niveau de compétence B2-C1 en français, ce qui signifie qu'il est capable de produire des textes nuancés, grammaticalement corrects et stylistiquement adaptés à la plupart des usages professionnels. Cette performance est le résultat d'un corpus d'entraînement multilingue substantiel, Google DeepMind ayant inclus d'importantes quantités de textes français de qualité dans les données d'entraînement de Gemma 4.

En pratique, les utilisateurs francophones constateront que Gemma 4 31B gère correctement les subtilités grammaticales du français — accords complexes, subjonctif, niveaux de langue —, produit des textes fluides et cohérents sur des sujets techniques et généraux, et comprend les références culturelles françaises. Les points de vigilance restent la tendance occasional à l'anglicisme dans certains contextes techniques, et une légère préférence pour les formulations internationales plutôt que spécifiquement franco-françaises. Sur les benchmarks multilinguaux comme MMMLU (version française de MMLU), Gemma 4 31B obtient des scores proches de ses résultats en anglais, ce qui est remarquable et inhabituel pour un modèle de cette taille.

Comparé aux modèles fermés comme Gemini 3.1 Pro ou Claude Fable 5, Gemma 4 31B présente un écart d'environ 3 à 5 points de pourcentage sur les tâches en français les plus complexes (rédaction juridique formelle, traduction de textes très spécialisés). Cet écart est négligeable pour la grande majorité des usages professionnels : analyse documentaire, génération de rapports, assistance à la rédaction, classification de textes, extraction d'informations. Pour les entreprises dont les cas d'usage principaux sont en français, Gemma 4 31B représente un excellent rapport qualité-prix, particulièrement quand l'hébergement local élimine les coûts d'API récurrents.

Cas d'usage détaillés pour Gemma 4 31B en entreprise

Conformité RGPD et traitement de données sensibles : Le principal avantage compétitif de Gemma 4 31B réside dans sa déployabilité on-premise complète. Les cabinets d'avocats, les établissements de santé, les banques et toute organisation traitant des données personnelles sensibles peuvent utiliser Gemma 4 31B sans que la moindre donnée ne quitte leur infrastructure. C'est un prérequis réglementaire que peu de modèles de ce niveau de performance peuvent satisfaire avec autant de facilité, grâce à la combinaison de la licence Apache 2.0 et de la taille raisonnable du modèle.

Analyse documentaire et extraction d'informations : Avec sa fenêtre de contexte de 128 000 tokens (environ 96 000 mots), Gemma 4 31B peut traiter des documents longs comme des contrats, des rapports d'audit, des procédures réglementaires ou des publications scientifiques en une seule passe. La vitesse d'inférence de 145 tokens par seconde sur A100 garantit un traitement rapide même pour des documents volumineux. Les équipes juridiques, financières et de conformité bénéficient particulièrement de cette capacité combinée à la confidentialité du déploiement local.

Développement logiciel et revue de code : Sur les benchmarks de codage comme HumanEval et SWE-bench, Gemma 4 31B affiche des performances remarquables pour sa taille. Les équipes de développement peuvent l'intégrer dans leurs IDE via des plugins compatibles (Ollama + Continue, LM Studio) pour obtenir une assistance à la complétion de code, la génération de tests unitaires, la revue automatique de pull requests et la documentation de fonctions, sans exposer leur code source à des services tiers.

Recherche académique et scientifique : Le score de 86,8 % sur GPQA Diamond témoigne d'une capacité de raisonnement scientifique avancée. Les laboratoires de recherche apprécient la disponibilité du modèle sur Kaggle avec des quotas GPU généreux, permettant des expériences à grande échelle sans coût prohibitif. La licence Apache 2.0 permet également de publier des recherches basées sur le modèle et d'en redistribuer des variantes fine-tunées, ce qui facilite la reproductibilité scientifique.

Automatisation des processus métier : Les entreprises peuvent déployer Gemma 4 31B comme moteur de classification, de résumé automatique, de génération de réponses standardisées ou d'aide à la décision dans leurs processus internes. L'intégration via l'API compatible OpenAI

(vLLM, Ollama) permet de migrer facilement des workflows existants construits pour GPT-4 ou Claude vers une infrastructure entièrement maîtrisée.

Tarification et accès à Gemma 4 31B en 2026

L'un des atouts majeurs de Gemma 4 31B est sa structure tarifaire extrêmement avantageuse. Pour les organisations qui n'ont pas les ressources pour maintenir une infrastructure GPU, Google AI Studio propose un accès API à des tarifs parmi les plus compétitifs du marché : 0,25 dollar par million de tokens en entrée et 0,75 dollar par million en sortie. À titre de comparaison, Claude Fable 5, qui domine le classement LM Arena avec un ELO de 1 507, est facturé 3 à 6 fois plus cher selon les niveaux de performance.

Pour le self-hosting, l'investissement initial est un GPU NVIDIA A100 80 Go (ou équivalent H100 NVL) avec un coût de location cloud typique de 2 à 4 dollars de l'heure sur les principales plateformes cloud (Google Cloud, AWS, Azure). Pour un usage de 8 heures par jour, cela représente environ 500 à 1 000 dollars par mois, ce qui est très compétitif pour les organisations qui génèrent des volumes importants de tokens. Pour les équipes traitant plus de 10 millions de tokens par mois, l'auto-hébergement devient économiquement supérieur à l'API même en incluant les coûts d'infrastructure et de maintenance.

Vertex AI (Google Cloud) propose également Gemma 4 31B dans son catalogue de modèles gérés, avec des garanties de disponibilité (SLA 99,9 %), un autoscaling adaptatif et une intégration native avec les outils de monitoring de GCP. Cette option est recommandée pour les organisations qui souhaitent bénéficier de la confidentialité d'un modèle open-source (les données restent dans leur projet GCP) sans gérer l'infrastructure GPU elles-mêmes. Le coût est légèrement supérieur à l'API directe de Google AI Studio, mais inclut les garanties contractuelles nécessaires pour les environnements de production critiques.

Comparaison avec les modèles open-source concurrents

Face à **Llama 4 Maverick de Meta** (ELO 1 451), Gemma 4 31B présente un profil complémentaire et dans certains cas supérieur. Llama 4 Maverick bénéficie d'une architecture MoE avec 402 milliards de paramètres totaux (17B actifs) et d'un contexte révolutionnaire de 10 millions de tokens, ce qui en fait le choix naturel pour les analyses de très longs documents. En contrepartie, Gemma 4 31B est nettement plus facile à déployer (1 seul GPU suffisant), plus rapide en inférence (145 vs 110 tokens/seconde), et affiche des scores légèrement supérieurs sur certains benchmarks de raisonnement pur. Pour comparer les deux modèles en détail, consultez notre article sur Llama 4 Maverick.

Face à **Mistral Large 3** (ELO 1 443), Gemma 4 31B obtient des scores légèrement inférieurs sur les benchmarks de codage en anglais, mais supérieurs sur GPQA Diamond et sur les tâches de raisonnement scientifique. La licence Apache 2.0 des deux modèles est comparable en termes de liberté d'usage commercial. Le principal avantage de Mistral Large 3 réside dans son origine européenne (Mistral AI, Paris), ce qui peut être un argument supplémentaire pour les organisations cherchant à valoriser la souveraineté numérique européenne. Gemma 4 31B surpasse Mistral Large 3 sur la majorité des benchmarks académiques standardisés.

Face à **DeepSeek V4 Pro Max** (ELO 1 449) sous licence MIT, Gemma 4 31B offre une alternative développée par un acteur occidental avec des garanties de transparence sur le processus d'entraînement. DeepSeek présente des performances légèrement supérieures sur les benchmarks de mathématiques et de codage, mais Gemma 4 31B se distingue par une meilleure performance en français et une documentation plus complète pour le déploiement en environnement d'entreprise. Les organisations préférant éviter une dépendance à des modèles chinois pour des raisons de conformité ou de politique interne trouveront dans Gemma 4 31B une alternative solide.

Face à **Qwen3.7 Max** d'Alibaba (ELO 1 475, Apache 2.0), Gemma 4 31B se situe 34 points en dessous sur LM Arena, un écart significatif. Qwen3.7 Max excelle particulièrement sur les tâches mathématiques et de codage complexes. Gemma 4 31B compense par une meilleure facilité de déploiement en environnement d'entreprise occidental et une communauté plus large autour de l'écosystème Google (intégrations GCP, documentation en français, support professionnel disponible via Google).

Limites et points de vigilance pour Gemma 4 31B

Malgré ses excellentes performances, Gemma 4 31B présente plusieurs limitations qu'il convient de prendre en compte avant tout déploiement. La première concerne l'écart de performance avec la version fermée Gemini 3.1 Pro : Gemma 4 31B se situe environ 3 à 4 points de pourcentage en dessous sur les benchmarks les plus exigeants, un écart qui peut être significatif pour des applications à très haute précision comme la génération de code critique ou l'analyse médicale de niveau expert. Les organisations dont les cas d'usage exigent le niveau de performance absolu devront évaluer si cet écart est acceptable.

La deuxième limitation est l'absence de capacités multimodales natives dans la variante 31B. Contrairement à certains modèles concurrents qui intègrent nativement la vision (analyse d'images, de diagrammes, de captures d'écran), Gemma 4 31B traite exclusivement du texte. Pour les workflows nécessitant l'analyse d'images ou de documents PDF avec mise en page complexe, il faudra soit utiliser un modèle multimodal séparé en amont, soit opter pour la variante 9B multimodale de la famille Gemma 4 (performances texte inférieures).

Troisièmement, la fenêtre de contexte de 128 000 tokens, bien qu'importante, ne peut pas rivaliser avec les capacités de Llama 4 Maverick (10 millions de tokens) pour les analyses de très longs documents. Pour analyser une codebase entière, un livre complet ou un ensemble de contrats volumineux en une seule passe, Llama 4 Maverick restera le choix supérieur malgré ses exigences en infrastructure plus importantes. Gemma 4 31B est cependant amplement suffisant pour la grande majorité des cas d'usage documentaires professionnels.

À RETENIR

À retenir

ELO 1 441 et GPQA Diamond 86,8 % — le meilleur modèle open-source jamais produit par Google, dans le top 5 des modèles open-source mondiaux en juillet 2026.

Apache 2.0 totalement libre — usage commercial illimité, modification autorisée, redistribution permise, aucune contrainte de volume d'utilisateurs ni de revenus.

31B denses sur 1 GPU A100 — déployable sur une infrastructure raisonnable sans cluster GPU, idéal pour les PME et les laboratoires de recherche disposant d'un budget maîtrisé.

Conformité RGPD native — l'hébergement on-premise complet garantit qu'aucune donnée utilisateur ne quitte votre infrastructure, un avantage décisif pour les secteurs réglementés (santé, finance, droit).

API à 0,25 $/0,75 $ par M tokens — l'une des offres API les plus compétitives du marché pour un modèle de ce niveau de performance, via Google AI Studio ou Vertex AI.

Niveau B2-C1 en français — performances multilingues solides, adaptées à la grande majorité des usages professionnels francophones, avec un léger écart résiduel par rapport aux modèles fermés sur les tâches très spécialisées.

Foire aux questions sur Gemma 4 31B

Quelle est la différence entre Gemma 4 31B et Gemini 3.1 Pro de Google ?

Gemma 4 31B est le modèle open-source publié par Google sous licence Apache 2.0, tandis que Gemini 3.1 Pro est un modèle fermé disponible uniquement via l'API Google. Les deux sont développés par Google DeepMind, mais Gemini 3.1 Pro bénéficie d'une architecture plus grande, d'un entraînement sur des données supplémentaires non divulguées et de capacités multimodales avancées. En pratique, Gemma 4 31B se situe environ 3 à 4 points de pourcentage en dessous de Gemini 3.1 Pro sur les benchmarks les plus exigeants. Le choix dépend de vos priorités : si la confidentialité des données et le déploiement on-premise sont essentiels, Gemma 4 31B est supérieur ; si vous cherchez les performances maximales sans contrainte d'infrastructure, Gemini 3.1 Pro sera plus approprié.

Combien coûte le déploiement de Gemma 4 31B en production ?

Le coût dépend fortement du mode de déploiement choisi. Via l'API Google AI Studio : 0,25 dollar par million de tokens en entrée et 0,75 dollar par million en sortie, avec un essai gratuit généreux. En auto-hébergement sur cloud : un GPU A100 80 Go coûte entre 2 et 4 dollars de l'heure selon le fournisseur (Google Cloud, AWS, Azure), soit 500 à 1 000 dollars par mois pour 8 heures d'utilisation quotidienne. En auto-hébergement on-premise avec du matériel propriétaire, le coût récurrent est quasi nul après l'investissement initial dans le GPU. Pour les volumes importants (plus de 50 millions de tokens par mois), l'auto-hébergement devient presque toujours plus économique que l'API.

Gemma 4 31B est-il utilisable gratuitement pour des projets commerciaux ?

Oui, sans restriction. La licence Apache 2.0 de Gemma 4 31B autorise explicitement l'usage commercial, la modification du modèle, la création de dérivés et leur distribution, sans aucune obligation de redevance ni contrainte de volume. Les seules obligations de la licence Apache 2.0 sont de conserver les mentions de copyright dans les fichiers redistribués et de signaler les modifications apportées au modèle original. Il n'y a pas de clause non-commerciale, pas de limite sur le nombre d'utilisateurs ni sur les revenus générés. C'est une liberté d'usage bien plus large que la licence Llama 4 Community de Meta, qui impose une limite à 700 millions d'utilisateurs actifs mensuels.

Gemma 4 31B peut-il être fine-tuné sur des données propriétaires ?

Oui, le fine-tuning est pleinement supporté et autorisé par la licence Apache 2.0. Google fournit des scripts officiels de fine-tuning basés sur la bibliothèque Keras et compatible avec les frameworks HuggingFace Transformers et TRL (Transformer Reinforcement Learning). Les techniques les plus utilisées sont le LoRA (Low-Rank Adaptation) et QLoRA (quantized LoRA), qui permettent d'adapter Gemma 4 31B à un domaine spécifique avec des ressources GPU limitées. Un fine-tuning LoRA basique peut être réalisé sur un seul GPU A100 en quelques heures pour des datasets de taille modeste. Les modèles fine-tunés peuvent être redistribués sous licence Apache 2.0 ou toute licence compatible, ce qui est un avantage majeur pour les éditeurs de logiciels souhaitant intégrer un LLM personnalisé dans leurs produits.

Quelles sont les alternatives open-source à Gemma 4 31B pour un déploiement sur matériel limité ?

Pour les organisations disposant de GPU moins puissants (RTX 4090 avec 24 Go de VRAM), Gemma 4 31B en quantification INT4 (environ 16 Go de VRAM) est une option viable avec une légère perte de performance. Les alternatives principales sont Mistral Small 3 (22B, Apache 2.0, très performant sur GPU consommateur), Qwen3 30B (performances légèrement supérieures mais architecture plus exigeante en mémoire), et les variantes plus petites de la famille Gemma 4 (9B instruction-tuned, multimodal, environ 8 Go en BF16). Pour un déploiement sur CPU uniquement ou sur du matériel très contraint, les modèles de 7 à 8 milliards de paramètres comme Gemma 4 9B ou Mistral 7B v0.3 restent des choix pertinents, avec des performances bien inférieures mais une empreinte mémoire de seulement 4 à 8 Go.

Comment Gemma 4 31B se compare-t-il aux modèles propriétaires sur la confidentialité des données ?

Gemma 4 31B offre le niveau de confidentialité le plus élevé possible lorsqu'il est déployé en auto-hébergement : vos données ne quittent jamais votre infrastructure et ne sont utilisées par personne d'autre. C'est un contraste total avec les modèles propriétaires comme GPT-4o (OpenAI) ou Claude Fable 5 (Anthropic), où les données envoyées via l'API sont traitées sur des serveurs tiers, même si ces entreprises s'engagent contractuellement à ne pas les utiliser pour l'entraînement. Pour les secteurs soumis au RGPD, au Health Data Hub français, au secret professionnel (avocats, médecins) ou aux contraintes de confidentialité industrielle, l'auto-hébergement de Gemma 4 31B constitue la solution la plus robuste juridiquement et techniquement.

