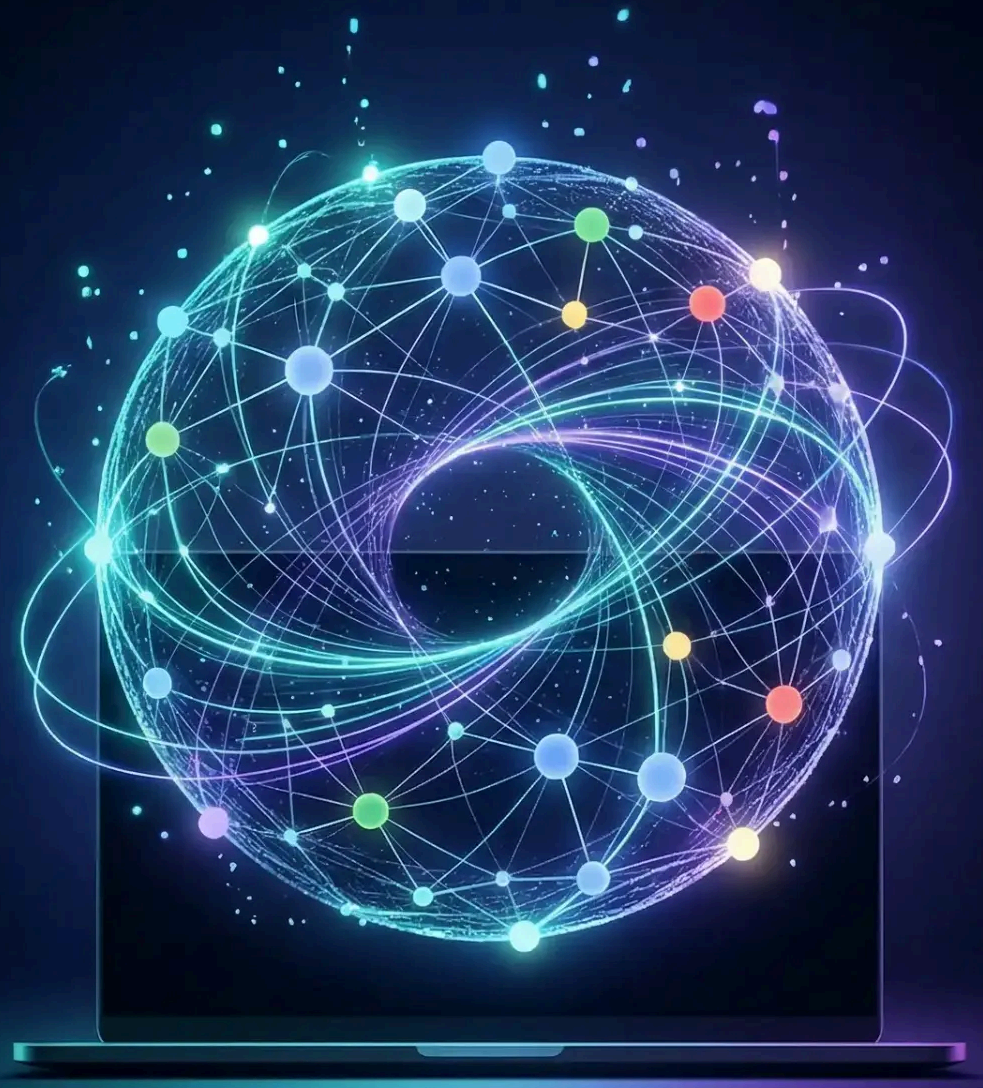
GOOGLE OPEN SOURCE · 2026

Gemma 4

L'Intelligence Artificielle de Google en Toute Liberté

Découvrez comment exécuter une IA de pointe directement sur votre appareil — privée, rapide, et entièrement hors-ligne.

Open Source | On-Device AI | Apache 2.0



02 • FONDAMENTAUX

Qu'est-ce que l'IA locale ?

DÉFINITION
L'IA s'exécute directement sur votre ordinateur, smartphone ou tablette, sans envoyer vos données vers le cloud.

SANS CONNEXION
Fonctionne même hors ligne. Vos calculs restent locaux, vos données ne quittent jamais votre environnement personnel.

INDÉPENDANCE TOTALE
Libérez-vous des contraintes cloud — pas de latence réseau, pas de dépendance aux plateformes tierces, pas de frais d'API.

☒ Votre appareil devient un cerveau numérique autonome et privé

Les Avantages Clés Confidentialité & Performance

Confidentialité Absolue

Vos données ne quittent jamais votre appareil. Aucune transmission vers des serveurs externes — une protection maximale pour vos informations personnelles et professionnelles.

Réponses Instantanées

Sans délai de transmission réseau, les réponses sont quasi instantanées. L'expérience utilisateur est fluide, réactive et sans interruption.

100% Hors-Ligne

Aucune connexion Wi-Fi requise. En déplacement, en zone blanche ou simplement déconnecté — votre IA reste pleinement opérationnelle.

“L’IA locale transforme votre appareil en un cerveau numérique privé — puissant, autonome et entièrement sous votre contrôle.”

04/10 - PRÉSENTATION

Gemma 4

L'Innovation Open-Source de Google

Apache 2.0

“Issu de la même technologie que Gemini — ouvert à tous.”

LANCÉ EN 2026

31 Mars 2026

par Google DeepMind

Open Source

Modifiez, redistribuez et intégrez librement dans vos projets personnels ou commerciaux.

Technologie Gemini

Construit sur la même recherche de pointe que les modèles Gemini de Google.

4 Modèles Disponibles

De l'ultra-mobile E2B au puissant 31B Dense — un modèle pour chaque appareil.

CAS D'USAGE · 01

Développement & Code

Votre copilote IA, entièrement privé.

Génération de Code

Produit des extraits de code dans tous les langages majeurs, directement sur votre machine.

Complétion Intelligente

Anticipe et complète vos lignes de code pour accélérer votre flux de développement.

Débogage Local

Identifie les erreurs et propose des corrections sans jamais envoyer votre code vers un serveur externe.

Refactorisation

Réécrit et optimise votre code pour une meilleure lisibilité et performance.

Vos fragments de code propriétaires ne quittent jamais votre machine.

CAS D'USAGE • 06

Création de Contenu & Raisonnement Avancé

RÉDACTION CRÉATIVE

Poèmes, scripts, articles, ébauches de courriels — générés hors-ligne en quelques secondes

SYNTHÈSE DE DOCUMENTS

Résumez des rapports complexes et longs articles de recherche instantanément

RAISONNEMENT AVANCÉ

Planification multi-étapes, suivi d'instructions complexes, résolution de problèmes

COMPRÉHENSION MULTIMODALE

Images, vidéos, audio — OCR, graphiques, reconnaissance vocale

FENÊTRE DE CONTEXTE

256K

tokens pour les grands modèles

128K

tokens · modèles compacts

Tout s'exécute localement — vos documents ne quittent jamais votre appareil

GEMMA 4 · CAPACITÉS AVANCÉES

Agents Autonomes & Workflows Complexes

Gemma 4 agit, connecte et orchestre — bien au-delà de la simple conversation.

FORMAT DE SORTIE

JSON Structuré

Des réponses parsables et exploitables directement par vos applications logicielles.

INTÉGRATION NATIVE

Function Calling

L'IA appelle des fonctions externes et des API, s'intégrant dans des systèmes complexes existants.

COMPORTEMENT DÉFINI

Instructions Système

Définissez des rôles, des règles et des contraintes pour des agents fiables et prévisibles.

Résultat : Des agents autonomes qui interagissent avec vos outils locaux — sans cloud, sans compromis.

FAMILLE DE MODÈLES

La Famille Gemma 4

Choisir son Modèle

Du smartphone à la station de travail — quatre architectures, un écosystème.

Ultra-Mobile → Haute Performance

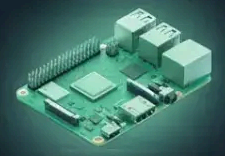
E2B

Effective 2B

Ultra-Mobile

⚡ Faible latence
Edge & Navigateur
Audio + Vidéo + Image

Contexte : 128K tokens



E4B

Effective 4B

Embarqué & Edge

⚡ Faible latence
IoT & Systèmes embarqués
Multimodal complet

Contexte : 128K tokens

RECOMMANDÉ

26B MoE

Mixture-of-Experts

Efficacité Maximale

Active seulement 3.8B paramètres
💻 Laptop GPU dédié
⚡ Vitesse exceptionnelle

Contexte : 256K tokens

BEST IN CLASS

31B Dense

Architecture Dense

Qualité Maximale

🏆 Meilleur raisonnement
🖥 Station de travail
Fine-tuning avancé

Contexte : 256K tokens

🔓 Tous les modèles — Licence Apache 2.0 · Open Source · Disponibles dès maintenant

Quel Modèle pour Quel Appareil ?

Choisissez selon votre VRAM disponible

📱 Mobile & Edge

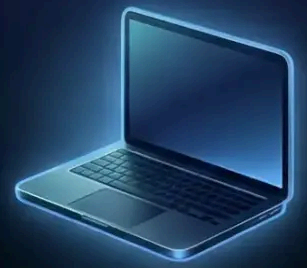
Gemma 4 E2B · E4B

BF16 → 9.6 GB / 15 GB

SFP8 → 4.6 GB / 7.5 GB

Q4_0 → 3.2 GB / 5 GB Recommandé

Smartphones Android · Raspberry Pi · NVIDIA Jetson



💻 Ordinateur Portable

Gemma 4 26B MoE

BF16 → 48 GB

SFP8 → 25 GB

Q4_0 → 15.6 GB Recommandé

GPU dédié · RTX 3080 / 4080 · MacBook Pro M-series

🖥️ Station de Travail

Gemma 4 31B Dense

BF16 → 58.3 GB

SFP8 → 30.4 GB

Q4_0 → 17.4 GB Haute Qualité

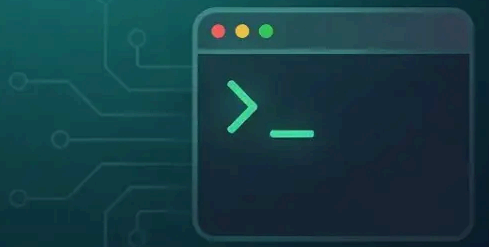
NVIDIA H100 80GB · A100 · Workstation Pro

⚡La quantification Q4_0 divise par 3 les besoins en VRAM — rendant Gemma accessible sur presque tous les appareils modernes.

INSTALLATION GUIDE · ÉTAPE PAR ÉTAPE

Guide d'Installation

Gemma 4 en mode hors-ligne, en 3 étapes



01 Installer Ollama

Téléchargez Ollama sur ollama.com/download — disponible sur macOS, Windows et Linux. Lancez l'installateur et vérifiez l'installation dans votre terminal.

```
ollama --version
```

02 Télécharger le Modèle Gemma

Une seule commande suffit. Le modèle est stocké localement — aucune connexion requise par la suite. Choisissez la taille adaptée à votre appareil.

```
ollama pull gemma4:2b \
ollama list
```

03 Lancer Gemma Sans Wi-Fi

Exécutez Gemma entièrement hors-ligne. Toutes les opérations restent sur votre machine — confidentialité garantie.

```
ollama run gemma4:2b "Écris un poème sur l'IA locale."
```

ALTERNATIVE

LM Studio

Interface graphique intuitive. Supporte les formats GGUF et MLX. Idéal pour explorer sans ligne de commande.

lmstudio.ai

Une fois téléchargé, Gemma fonctionne sans connexion internet — vos données ne quittent jamais votre appareil.

Prêt à commencer ?

Autre solution si Gemma 4 n'apparaît pas encore dans Ollama ?

Téléchargement officiel et alternatives simples pour démarrer en local

Page officielle Google DeepMind

deepmind.google/models/gemma/gemma-4

Point d'entrée recommandé pour découvrir Gemma 4. Regroupe les liens vers Hugging Face, Ollama, Kaggle, LM Studio et Docker.

Hugging Face · Ollama · Kaggle · LM Studio · Docker

Autres options de démarrage

- **LM Studio** : interface graphique, plus simple pour les non-techniciens
- **Hugging Face / Kaggle** : téléchargement direct des modèles
- **Docker** : option pour utilisateurs plus avancés

À savoir

- Si Gemma 4 n'apparaît pas encore dans Ollama, cela peut venir d'un déploiement progressif
- Le premier téléchargement peut prendre du temps selon la taille du modèle et la machine
- Une IA peut aussi vous guider pas à pas dans le terminal

Si Ollama n'est pas encore prêt, passez par la source officielle Google DeepMind pour choisir la voie la plus simple vers une installation locale.