BLOG

Les Meilleurs LLMs Open Source

21 juin 2026 - Denis Favre

> *Nous vivons l'un de ces moments où une technologie change si vite qu'elle devance nos capacités à la réguler, à la comprendre, et même à en dépendre sereinement. En quelques années, les grands modèles de langage ont cessé d'être des curiosités de laboratoire pour devenir l'infrastructure invisible de nos organisations : ils alimentent les assistants qui répondent à nos clients, les agents qui planifient et agissent à notre place, les outils qui transforment une idée en produit en quelques minutes. La promesse était belle — et elle l'est toujours. Mais une promesse tenue par un seul acteur n'est pas une promesse : c'est une dépendance.*

Car voilà ce que l'épisode Fable 5 nous a enseigné avec une brutalité pédagogique rare. En juin 2026, le gouvernement américain a ordonné la suspension d'accès à l'un des modèles les plus puissants du monde — non pas parce qu'il était défaillant, mais précisément parce qu'il était trop capable. Une lettre signée par le secrétaire au Commerce, invoquant la sécurité nationale, a suffi. Trois jours d'interruption seulement — aucune entreprise n'a eu le temps de voir ses processus s'effondrer. Mais le précédent, lui, est permanent : désormais, un État peut retirer l'accès à un modèle d'intelligence artificielle sans justification concrète, du jour au lendemain, pour des raisons qui mêlent indistinctement la sécurité réelle, le calcul politique et les intérêts économiques. Ce qui était une infrastructure est devenu une arme diplomatique. Ce qui était un outil est devenu un actif géopolitique.

La vraie question n'est plus : quel est le meilleur modèle ? Elle est : à quel modèle aurai-je le droit d'accéder demain, selon ma nationalité, mon secteur d'activité, la couleur politique de mon gouvernement, ou la proximité de mon fournisseur avec telle ou telle administration ? Ce sont des critères potentiellement discriminants, imprévisibles, et entièrement hors de notre contrôle — si nous avons confié notre infrastructure cognitive à une poignée d'acteurs fermés.

C'est précisément cet impératif que les LLMs open source viennent répondre. Non par idéologie, mais par simple prudence stratégique : la diversification des outils n'est pas un luxe de perfectionniste, c'est la condition minimale de la résilience. Un modèle qu'on peut héberger soi-même, fine-tuner sur ses propres données, optimiser pour ses propres charges de travail — un modèle dont personne ne peut couper l'accès par décret — n'est pas seulement plus flexible techniquement. Il est structurellement plus fiable. Il échappe aux décisions arbitraires, aux guerres commerciales, aux revirements réglementaires, aux tensions entre une entreprise et un gouvernement que l'on ne maîtrise pas.

Dans ce guide, nous explorons les meilleurs LLMs open source disponibles aujourd'hui — leurs forces, leurs limites, et les critères qui permettent de choisir avec discernement. Parce qu'avoir accès à la meilleure intelligence artificielle du monde ne sert à rien si l'on peut vous en priver demain matin.

Qu'est-ce qu'un LLM open source ?

La question mérite qu'on s'y arrête, parce que le mot lui-même est devenu, dans l'usage courant, une de ces étiquettes commodes que l'on colle sur des réalités fort différentes — et dont la confusion nuit à ceux qui doivent prendre des décisions concrètes.

Dans son sens le plus rigoureux, un grand modèle de langage open source est un modèle dont l'architecture, le code source et les poids — c'est-à-dire les paramètres appris au terme de l'entraînement, qui constituent en quelque sorte la mémoire et l'intelligence du système — sont rendus publics dans leur intégralité, librement téléchargeables, librement modifiables, librement redistribuables. Vous pouvez l'installer sur vos propres serveurs, l'adapter à votre domaine métier avec vos propres données, l'optimiser pour vos propres charges de travail, et le déployer sans jamais demander permission à personne. C'est cette liberté totale — sur l'inférence, sur la personnalisation, sur la confidentialité des données, sur les coûts à long terme — qui distingue structurellement ces modèles de leurs équivalents fermés.

Mais il faut introduire ici une distinction que l'enthousiasme du secteur tend à brouiller : celle entre open source et open weights.

Un modèle à poids ouverts (open weights) rend publics ses paramètres — vous pouvez les télécharger, les exécuter, parfois les modifier. Mais la licence qui les accompagne ne satisfait pas nécessairement à la définition de l'Open Source Initiative, l'organisme de référence en la matière. Ces licences peuvent imposer des restrictions d'usage commercial, des obligations d'attribution, des conditions sur la redistribution des versions modifiées, ou des plafonds sur l'échelle de déploiement au-delà desquels une autorisation devient requise. Le modèle est accessible — mais l'accès n'est pas une liberté, c'est une permission conditionnelle. Et une permission, comme nous l'avons vu, peut se retirer.

La nuance n'est pas académique. Pour une organisation qui construit sur ces fondations, la différence entre une licence véritablement libre et une licence à poids ouverts peut se traduire, demain, par une facture inattendue, une mise en demeure juridique, ou la découverte tardive qu'un changement de politique de l'éditeur remet en cause l'ensemble d'une infrastructure. Choisir un modèle open source, c'est donc d'abord lire ce que l'on signe — ou ce que l'on ne signe pas.

Pour autant, cette distinction ne doit pas paralyser le praticien. Les deux catégories — open source au sens strict et open weights — partagent l'essentiel de ce qui importe dans un contexte opérationnel : la

possibilité d'héberger le modèle sur sa propre infrastructure, d'inspecter ses comportements, de l'interroger, de le corriger, de l'affiner sur des données propriétaires. Ce que vous ne pouvez pas faire avec un modèle fermé — voir ce qui se passe à l'intérieur, intervenir sur ce qui en sort, couper le cordon avec le fournisseur —, vous pouvez le faire avec l'un ou l'autre. Les différences réelles se situent dans l'étendue des libertés accordées par la licence, et dans le degré de transparence sur le pipeline d'entraînement : certains modèles publient leurs poids mais taisent leurs données, leurs méthodes de filtrage, leurs choix d'alignement. D'autres vont plus loin et documentent l'ensemble du processus. C'est une information utile — mais c'est rarement le critère décisif pour une équipe qui doit livrer en production.

Dans ce guide, nous n'entrerons donc pas dans la taxonomie complète des licences. Chaque modèle présenté ici peut être téléchargé librement et déployé sur votre propre infrastructure — ce qui est, dans la pratique, la condition première que se fixent les équipes au moment d'évaluer un LLM pour un usage réel. Le reste est affaire de juristes et de cas d'usage spécifiques ; nous vous donnerons les éléments nécessaires pour poser les bonnes questions, sans prétendre les trancher à votre place.

1. DeepSeek-V4

Le flagship général de DeepSeek. Il domine les benchmarks avec un score de 87 sur BenchLM.ai, ce qui en fait le numéro 1 toutes catégories confondues parmi les modèles open weight. Il se décline en deux variantes distinctes.

V4-Pro (1,6T paramètres totaux / 49B actifs) est le modèle de référence pour le raisonnement avancé, le codage et les workflows agentiques, avec une fenêtre de contexte d'un million de tokens et trois modes de raisonnement configurables :

ollama run deepseek-v4-pro:cloud

V4-Flash (284B paramètres totaux / 13B actifs) est la variante économique, pensée pour l'usage quotidien avec une latence réduite :

ollama run deepseek-v4-flash:cloud

Ces deux variantes ne sont disponibles sur Ollama qu'en mode cloud — les requêtes transitent par les serveurs d'Ollama sans qu'aucun poids ne soit téléchargé localement, ce qui les rend inadaptées aux données sensibles. Pour un usage strictement local et confidentiel, DeepSeek-R1 reste la meilleure alternative de la même famille, disponible en plusieurs tailles selon le matériel disponible :

ollama pull deepseek-r1:8b # ~5 Go, GPU 6-8 Go VRAM

ollama pull deepseek-r1:14b # ~9 Go, GPU 12 Go VRAM

ollama pull deepseek-r1:32b # ~19 Go, GPU 24 Go VRAM

Licence MIT pour l'ensemble de la famille.

En résumé voici les cas d'usage recommandés :

DeepSeek-V4-Pro #raisonnement-complexe #codage-avancé #agents-IA #contexte-long #SWE-bench #cloud-only

DeepSeek-V4-Flash #usage-quotidien #codage-léger #local #24Go-VRAM #faible-latence #coût-réduit

DeepSeek V4 : par où passer pour payer le moins cher ?

Une comparaison directe entre l'API DeepSeek, HuggingFace Inference Providers et OpenRouter.

Deux modèles dominent aujourd'hui le rapport qualité-prix dans l'univers des grands modèles de langage : **DeepSeek V4-Flash**, conçu pour les usages quotidiens et le codage léger, et **DeepSeek V4-Pro**, taillé pour le raisonnement complexe, les agents IA et l'analyse de bases de code entières. Les deux sont disponibles en open-weights sous licence MIT sur HuggingFace, et accessibles via plusieurs passerelles API. Reste la question pratique : par où faut-il passer pour minimiser la facture ?

Auto-hébergement : quel matériel pour inférer ces modèles ?

Les deux modèles étant disponibles en open-weights sous licence MIT, il est techniquement possible de les faire tourner localement — mais les exigences matérielles sont radicalement différentes selon le modèle choisi.

DeepSeek V4-Flash (284 milliards de paramètres au total, seulement 13 milliards actifs par token grâce à l'architecture MoE) est le seul des deux accessible sur une machine individuelle. Cette distinction est fondamentale : ce ne sont pas 284 milliards de paramètres qui transitent en mémoire à chaque génération de token, mais seulement 13 milliards — ce qui explique pourquoi le modèle tient dans un espace mémoire bien inférieur à sa taille nominale. En pratique, la version quantifiée Q4 pèse environ 158 Go, auxquels s'ajoutent 30 à 40 Go de marge pour le cache KV et le système d'exploitation. Le plancher réaliste se situe donc à 192 Go de mémoire.

C'est ici qu'une alternative inattendue entre en jeu : le **Mac Studio M4 Max 192 Go** (~5 999 €). Grâce à l'architecture de mémoire unifiée d'Apple Silicon, ces 192 Go servent simultanément de RAM et de VRAM, ce que ne peut pas faire une RTX 5090 limitée à 24 Go de VRAM dédiée. Le framework MLX d'Apple, qui supporte nativement le routage MoE depuis la version 0.24, permet d'atteindre 25 à 35 tokens par seconde sur V4-Flash — une vitesse tout à fait confortable pour un usage interactif ou un agent IA en solo. Un MacBook Pro M3 Max 128 Go, avec la quantification 2

bits expérimentale du moteur ds4 développé par Salvatore Sanfilippo (le créateur de Redis), atteint environ 26 tokens par seconde en génération — sur un laptop. Pour ceux qui ont besoin de plus de contexte ou de débit, le Mac Studio M3 Ultra 192 Go (à partir de 3 999 €) tourne autour de 25 tokens par seconde avec davantage de bande passante mémoire.

L'alternative en datacenter reste le serveur équipé d'un unique H100 80 Go ou H200 141 Go (25 000 à 35 000 € à l'achat, ou 2 à 3 $/heure en location), qui offre un débit supérieur en production multi-utilisateurs mais sans l'avantage de la mémoire unifiée.

DeepSeek V4-Pro (1,6 billion de paramètres, 49 milliards actifs) est en revanche hors de portée de toute machine individuelle. Il requiert plusieurs GPU de datacenter en parallèle — typiquement 4 à 8 H100 ou A100 80 Go — représentant une infrastructure de 100 000 à 250 000 € en matériel nu. Il n'existe pas de chemin Mac pour V4-Pro : même un Mac Studio M3 Ultra 512 Go (environ 9 500 €) ne suffit pas.

Pour la grande majorité des projets, l'auto-hébergement de V4-Pro n'a aucune justification économique face à l'API directe à 0,87 $/million de tokens. V4-Flash en local, en revanche, devient pertinent à partir d'un volume mensuel suffisamment élevé pour amortir le matériel — seuil que le Mac Studio M4 Max 192 Go atteint plus rapidement que tout autre configuration, grâce à un coût d'acquisition contenu et une consommation électrique de seulement 50 à 80 W.

Les tarifs de référence (juin 2026)

Avant toute comparaison de canaux d'accès, voici les prix publics des deux modèles, tels qu'ils ressortent de la documentation officielle DeepSeek après la baisse permanente de 75 % actée le 22 mai 2026 :

DeepSeek V4-Flash

Entrée (cache miss) : **0,14 $ / million de tokens**

Sortie : **0,28 $ / million de tokens**

Entrée (cache hit) : **0,0028 $ / million de tokens** — soit une économie de 98 % sur le contexte répété

DeepSeek V4-Pro

Entrée (cache miss) : **0,435 $ / million de tokens**

Sortie : **0,87 $ / million de tokens**

Entrée (cache hit) : **0,003625 $ / million de tokens**

Ces tarifs constituent le plancher absolu. Tout accès intermédiaire ajoute, au minimum, une friction administrative — et souvent un coût supplémentaire.

DeepSeek direct : le chemin le plus court et le moins cher

L'API officielle, accessible sur platform.deepseek.com, facture exactement aux tarifs ci-dessus, sans surcouche. Elle est compatible avec le format OpenAI ChatCompletions et avec le format Anthropic, ce qui simplifie la migration depuis d'autres fournisseurs : un changement de base_url et de clé API suffit dans la quasi-totalité des SDK courants.

À cela s'ajoute un crédit de démarrage de 5 millions de tokens offerts à toute nouvelle inscription, sans carte bancaire requise — de quoi couvrir plusieurs milliers d'appels de test avant le premier euro dépensé. Un bonus off-peak a historiquement réduit les coûts de 50 à 75 % supplémentaires entre 16h30 et 0h30 UTC.

La seule limite réelle est infrastructurelle : les serveurs DeepSeek sont basés en Chine, ce qui se traduit par une latence légèrement plus élevée pour les utilisateurs européens et quelques questions de souveraineté des données pour les projets soumis à des contraintes réglementaires strictes.

HuggingFace Inference Providers : commodité sans surcoût sur les tokens

HuggingFace ne gère pas directement l'inférence des grands modèles comme V4-Flash ou V4-Pro. Son service « Inference Providers » agit comme un routeur : il dispatche les requêtes vers des partenaires tiers — DeepInfra, Fireworks AI, Together AI, Nebius, entre autres — en passant leurs tarifs sans marge supplémentaire sur les tokens eux-mêmes. Les utilisateurs du plan PRO (9 $/mois) bénéficient de 2 $ de crédits d'inférence mensuels ; au-delà, la facturation bascule en pay-as-you-go aux tarifs des providers.

En pratique, les prix varient selon le provider routé. Via DeepInfra par exemple, V4-Pro s'établit autour de 0,30 $ en entrée et 1,00 $ en sortie — dans certains cas légèrement moins cher sur l'entrée que le tarif DeepSeek direct, mais plus cher sur la sortie, ce qui pénalise les workloads à forte génération de texte.

HuggingFace présente un intérêt réel pour les équipes qui souhaitent une clé API unique pour accéder à de nombreux modèles différents, sans multiplier les comptes fournisseurs. C'est une solution de commodité, non d'économie.

OpenRouter : la flexibilité au prix d'un frais de dossier

OpenRouter adopte une logique similaire : un point d'accès unique, des dizaines de providers derrière. Les tarifs affichés par token pour V4-Flash (0,09 $ en entrée) et V4-Pro (0,435 $ en entrée)

correspondent aux tarifs directs, parfois légèrement inférieurs selon le provider routé. OpenRouter ne prend pas de marge sur le prix par token.

Là où le coût réapparaît, c'est lors de l'achat de crédits : **une commission de 5,5 %** est prélevée à chaque rechargement. Sur 1 000 $ déposés, seuls 945 $ sont disponibles pour l'inférence. Sur des volumes mensuels importants, cela représente un overhead réel — 500 $ sur 10 000 $ de dépenses mensuelles, par exemple.

OpenRouter reste légitime pour les phases d'exploration où l'on teste plusieurs modèles simultanément, ou pour bénéficier du basculement automatique entre providers en cas de panne. Pour un usage en production centré sur DeepSeek, la commission de 5,5 % est un surcoût qui ne correspond à aucune valeur ajoutée sur le token lui-même.

Comparaison sur un exemple concret

Pour fixer les idées, voici ce que coûtent **1 million de requêtes légères** (500 tokens en entrée, 200 tokens en sortie chacune) selon le canal d'accès :

Canal d'accès	DeepSeek V4-Flash	DeepSeek V4-Pro
DeepSeek direct	**126 $**	**391 $**
HuggingFace → DeepInfra	~123 $	~415 $
OpenRouter (+5,5 %)	~133 $	~413 $

Les écarts absolus semblent faibles à ce volume. Ils deviennent significatifs à l'échelle d'une production mensuelle de plusieurs dizaines de millions de requêtes, où chaque fraction de cent compte.

Conclusion : aller au plus direct

La hiérarchie est claire. L'API DeepSeek directe offre les tarifs les plus bas, le cache hit le plus agressif du marché sur ses modèles, et une compatibilité native avec les deux standards de l'écosystème (OpenAI et Anthropic). HuggingFace ajoute une couche de commodité utile pour les projets multi-modèles, sans surcoût par token, mais sans avantage économique sur DeepSeek seul.

OpenRouter prélève 5,5 % à l'achat de crédits — un frais justifié pour la phase de prototypage, difficile à défendre en production stable.

Pour quiconque a fait le choix de DeepSeek V4-Flash ou V4-Pro comme modèle principal, **passer par l'API officielle est la décision économiquement rationnelle**. Les 5 millions de tokens offerts à l'inscription permettent de tout tester sans engagement. La migration depuis un autre fournisseur se résume à deux lignes de configuration. Et le modèle de cache, avec ses économies de 98 % sur le contexte répété, place l'API DeepSeek dans une catégorie à part pour les agents IA qui réutilisent de longs system prompts à chaque appel.

2. MiMo-V2.5-Pro

Le modèle phare de Xiaomi, sorti fin avril 2026 — l'une des sorties majeures qui, entre avril et mai, ont rebattu les cartes du classement open source. Contrairement à ce que l'on attendrait d'un constructeur de smartphones, il ne s'agit nullement d'un modèle compact taillé pour l'embarqué : MiMo-V2.5-Pro est au contraire un géant, une architecture Mixture-of-Experts de 1,02 billion de paramètres totaux dont seulement 42 milliards sont actifs par token. Il embarque une fenêtre de contexte d'un million de tokens en standard et un mécanisme d'attention hybride entrelaçant attention locale à fenêtre glissante et attention globale selon un ratio de 6:1, qui réduit de près de sept fois le stockage du cache clé-valeur. C'est cette frugalité mémoire — et non une miniaturisation des poids — qui explique ses performances étonnantes au regard de sa taille nominale.

Sa véritable spécialité est le travail agentique de longue haleine : codage autonome, raisonnement sur des bases de code entières, orchestration d'outils sur des milliers d'étapes sans jamais perdre de vue l'objectif initial. Sur le benchmark ClawEval, il domine le champ open source avec un taux de réussite de 63,8 %, en consommant environ 70 000 tokens par trajectoire — soit 40 à 60 % de moins que Claude Opus 4.6, Gemini 3.1 Pro ou GPT-5.4 pour un résultat comparable. Il dépasse DeepSeek V4-Pro et Kimi K2.6 sur les évaluations GDPVal-AA (Elo) et Claw-Eval, s'ouvre librement sur Hugging Face sous licence permissive, et s'intègre nativement aux principaux scaffolds agentiques — OpenCode, Hermes Agent, KiloCode. Côté tarifs, 1 $/million de tokens en entrée et 3 $ en sortie jusqu'à 256K de contexte, le double au-delà sur la plage 256K–1M. La division MiMo est dirigée par Luo Fuli, ancienne contributrice cœur de DeepSeek sur les modèles R1 et V.

Codersera + VentureBeat + BigGo

3. Kimi-K2.6

Le modèle phare de Moonshot AI, sorti le 20 avril 2026 — et considéré, à sa sortie, comme le meilleur modèle à poids ouverts du marché. Architecture Mixture-of-Experts de mille milliards de paramètres totaux pour seulement 32 milliards actifs par token (384 experts, dont 8 routés et 1 partagé), attention MLA, fenêtre de contexte de 256K tokens et multimodalité native, avec deux modes de fonctionnement — « Thinking » pour le raisonnement approfondi, « Instant » pour la réponse immédiate. Sur les épreuves qui comptent pour les développeurs, il égale GPT-5.5 sur SWE-Bench Pro à 58,6 %, mène sur Humanity's Last Exam avec outils à 54,0 %, et délivre ces résultats à environ un cinquième du coût au token des modèles fermés équivalents.

Sa véritable signature est l'exécution agentique de très longue haleine. Kimi K2.6 peut déployer jusqu'à 300 sous-agents en parallèle, coordonner plus de 4 000 appels d'outils et tenir des sessions continues de plus de douze heures ; Moonshot rapporte un agent ayant opéré de façon autonome pendant cinq jours, gérant monitoring, réponse à incident et opérations système de bout en bout. Le modèle excelle par ailleurs en génération frontend — il transforme un simple prompt en interface de niveau Awwwards, hero sections soignées, éléments interactifs et animations au défilement — et s'intègre dès le premier jour aux principales stacks agentiques : vLLM, OpenRouter, Cloudflare Workers AI, MLX, Hermes Agent et OpenCode, avec Kimi Code comme CLI dédié. Les poids sont publiés sous licence MIT modifiée : libre en usage commercial, à une seule réserve — afficher la mention « Kimi K2 » dans l'interface si le produit dépasse 100 millions d'utilisateurs actifs mensuels ou 20 millions de dollars de revenus mensuels ; en deçà, c'est du MIT pur.

Codersera + Latent Space + Moonshot

4. GLM-5.2

Développé par Z.AI (anciennement Zhipu AI). Sorti le 13 juin 2026, ce modèle Mixture-of-Experts de ~744 milliards de paramètres (~40 Md actifs) se hisse à 94 sur BenchLM.ai, où il se classe #3 sur 124 modèles du classement provisoire — un bond considérable par rapport au 83 de GLM-5.1. Il embarque une fenêtre de contexte d'1 million de tokens et deux niveaux d'effort de raisonnement (High et Max), avec une orientation très marquée vers le code agentique et les tâches « long-horizon ». Il conserve par ailleurs sa **licence MIT**, la plus permissive et la plus limpide de toute la liste, les poids étant ouverts sous cette licence sans aucune restriction d'usage régional.

Un lancement à forte charge géopolitique. La sortie de GLM-5.2 n'a rien d'anodin dans son calendrier : l'annonce est intervenue deux jours après que le secrétaire américain au Commerce a ordonné à Anthropic de bloquer l'accès étranger à ses modèles Fable 5 et Mythos 5 sous 48 heures, sans aucune possibilité d'appel. Anthropic a justifié cette suspension par une directive de contrôle des exportations émanant du gouvernement américain, invoquant des motifs de sécurité nationale. Z.AI a

explicitement positionné GLM-5.2 comme une réponse « open-source » à ce découplage : en publiant le modèle sous licence MIT — la même que celle de projets fondateurs comme Linux ou Redis — l'entreprise affirme vouloir démocratiser l'accès aux capacités d'IA de pointe et empêcher qu'une seule nation ou une seule entreprise ne monopolise la technologie des modèles fondamentaux. Ce geste est largement perçu comme un contre-coup stratégique face à la tendance croissante au découplage entre les écosystèmes IA américain et chinois, positionnant l'offre chinoise comme alternative ouverte crédible — un argument qui résonne directement avec la question de la souveraineté open-source.

5. Qwen3.5-397B-A17B

Le géant d'Alibaba, publié le 16 février 2026. Architecture Mixture-of-Experts de 397 milliards de paramètres totaux pour 17 milliards seulement actifs par token, fenêtre de contexte de 256K tokens, et fluence native dans 201 langues — la multimodalité y est intégrée tôt dans l'architecture, texte, image, vidéo et documents raisonnés dans un cadre unifié plutôt que greffés après coup sur un socle textuel. Sur le raisonnement général et la programmation, il évolue dans le même tier de performance que Gemini 3 Pro, Claude Opus 4.5 et GPT-5.2 ; mais c'est sur un point précis qu'il ne souffre aucune concurrence : le multilinguisme. Avec 201 langues et dialectes couverts — de très loin la plus large amplitude de tous les modèles à poids ouverts — et la meilleure qualité de traduction sur FLORES-200, il s'impose comme la référence absolue pour les usages non-anglophones, le français y compris.

Deux atouts en font un choix de premier ordre pour qui vise la souveraineté. La licence d'abord : Apache 2.0 sans la moindre restriction — ni plafond d'utilisateurs, ni clause d'usage acceptable, ni obligation d'attribution ; le modèle de classe frontière le plus libéralement licencié jamais publié. L'auto-hébergement ensuite : grâce à ses 17 milliards de paramètres actifs et aux quantifications dynamiques, sa version 4 bits (~214 Go) se charge directement sur un Mac Studio M3 Ultra, et tourne même sur une station équipée d'un unique GPU 24 Go épaulé de 256 Go de RAM système via offloading MoE, à plus de 25 tokens par seconde. Le tout pour un débit de décodage 8,6 à 19 fois supérieur à celui de la génération précédente. Il s'inscrit enfin dans la stratégie open source la plus agressive du secteur — une famille complète du 0,8B embarqué au 397B de datacenter, déjà prolongée par Qwen3.6 (avril) puis le flagship Qwen3.7 (mai-juin, API uniquement).

TECHSY + Unsloth + BentoML

Qwen3.5, ou la souveraineté faite machine

Il y a, dans le cas de Qwen3.5, quelque chose qui dépasse la simple fiche technique et qui mérite qu'on s'y arrête ; car c'est ici, et peut-être ici seulement dans toute cette liste, que la thèse défendue en introduction cesse d'être une précaution théorique pour devenir une possibilité matérielle, vérifiable, à portée de main. Nous écrivions plus haut qu'une promesse tenue par un seul acteur n'est pas une promesse mais une dépendance, et que l'épisode Fable 5 avait montré avec quelle facilité un État pouvait, d'un simple trait de plume, retirer l'accès à l'intelligence sur laquelle reposaient des organisations entières. Qwen3.5 est la réponse concrète à cette inquiétude — non pas une réponse rhétorique, mais une réponse que l'on peut télécharger, installer et faire tourner ce soir même.

Réunissons les éléments. Une licence Apache 2.0 d'abord, c'est-à-dire la plus permissive qui soit : aucun plafond d'utilisateurs, aucune clause d'usage, aucune permission à solliciter — l'exact contraire de ces autorisations conditionnelles qui se révèlent, le jour venu, révocables. Une couverture de deux cent une langues ensuite, parmi lesquelles un français de tout premier plan : il ne s'agit plus de se contenter d'un modèle anglo-saxon qui traite notre langue comme un dialecte périphérique, mais de disposer d'un outil qui pense, raisonne et rédige dans notre langue avec une aisance native. Un poids enfin — dix-sept milliards de paramètres actifs seulement, une version quantifiée d'environ deux cents gigaoctets — qui ramène l'inaccessible à l'échelle d'une seule machine de bureau : un Mac Studio posé sur une table, sans datacenter, sans abonnement, sans cordon ombilical vers un fournisseur lointain.

Mesurons ce que cela signifie. Un indépendant, une PME, une administration, un chercheur français peut aujourd'hui faire tourner, sur sa propre machine, dans sa propre langue et sous son propre toit, un modèle de classe frontière — comparable, pour le raisonnement, aux meilleurs systèmes fermés du moment. Aucune nationalité ne lui en interdira l'accès demain matin ; aucun secrétaire au Commerce ne signera la lettre qui l'en privera ; aucun revirement réglementaire ne viendra suspendre son outil de travail. Ce n'est pas le modèle le plus puissant de cette liste, et nous ne le présentons pas comme tel. C'est, en revanche, la démonstration la plus nette que la souveraineté numérique dont nous parlions n'est ni une utopie d'idéologue ni un luxe de perfectionniste : c'est une option disponible, immédiate, et que rien n'empêche de saisir.

6. Gemma 4

Le modèle open weight de Google DeepMind, publié le 2 avril 2026 — et, pour la première fois dans l'histoire de la famille, intégralement sous licence Apache 2.0, donc librement exploitable en usage commercial. Il ne s'agit pas d'une simple mise à jour mais d'une refonte architecturale, déclinée en

plusieurs tailles pensées pour chaque type de matériel : E2B et E4B pour l'embarqué — l'E2B tient dans 2 Go de RAM en Q4, jusque sur un Raspberry Pi — un 12B multimodal unifié, un dense de 31B, et surtout un Mixture-of-Experts de 26B n'activant que 4 milliards de paramètres par token, qui délivre une qualité de classe 27B avec la latence d'un petit modèle. C'est ce 26B MoE — et non un improbable « 27B sur 16 Go » — qui constitue le véritable point d'entrée facile à auto-héberger ; quant au 31B dense, il se contente d'un unique GPU grand public haut de gamme de type RTX 4090.

Loin d'être un simple modèle « généraliste solide », Gemma 4 affiche surtout un rapport intelligence/paramètre remarquable : son 31B dense rivalise sur l'Arena avec des modèles vingt fois plus gros (ELO de 1452) et fait des bonds spectaculaires par rapport à Gemma 3 — 89,2 % contre 20,8 % sur AIME 2026, 80 % contre 29 % sur LiveCodeBench v6. Multimodalité native sur toute la gamme (texte et image partout, audio sur les variantes edge et le 12B), fenêtre de contexte de 128K tokens pour les petits modèles et 256K pour les moyens, function calling intégré et modes « thinking » configurables. Son atout le moins contestable reste toutefois la maturité de son écosystème : plus de 400 millions de téléchargements cumulés et un « Gemmaverse » de plus de 100 000 variantes dérivées, avec un support natif sur Ollama, llama.cpp, vLLM, MLX, Hugging Face, Kaggle et Vertex AI.

7. MiniMax-M2.7

Le modèle de MiniMax, laboratoire shanghaïen fondé en 2021 et introduit en bourse à Hong Kong en janvier 2026. Sorti le 18 mars 2026 (poids ouverts diffusés en avril), il succède à M2.5 et acte une bascule revendiquée : non plus le très long contexte qui avait fait la réputation des modèles MiniMax-01, mais l'efficacité agentique. Architecture Mixture-of-Experts éparse de 230 milliards de paramètres totaux pour seulement 10 milliards actifs par token — un taux d'activation de 4,3 %. C'est cette frugalité qui lui permet de casser les prix d'un ordre de grandeur : environ 0,30 $/million de tokens en entrée, soit près de dix fois moins cher que Claude Sonnet.

Sa véritable spécialité est l'orchestration agentique et le codage. Il bâtit des harnais d'agents complexes — Agent Teams, Skills, recherche dynamique d'outils — et maintient un taux de conformité de 97 % sur quarante compétences complexes, avec de solides scores SWE-bench (56,22 % sur SWE-Pro, 76,5 sur SWE Multilingual). Détail saisissant, MiniMax le présente comme son premier modèle participant à sa propre évolution : une version interne a optimisé de façon autonome un scaffold de programmation sur plus de cent itérations, pour un gain de 30 % de performance. Un bémol qui résonne avec ton fil rouge open source / open weights : ses poids sont ouverts, mais plusieurs guides signalent qu'un usage commercial requiert une autorisation écrite de MiniMax — à vérifier sur la fiche du modèle avant tout déploiement payant.

8. DeepSeek-R1

Le modèle de raisonnement pur de DeepSeek — et sans doute le plus marquant de cette liste par son onde de choc. Sorti le 20 janvier 2025, il fut le premier modèle à poids ouverts à rivaliser avec OpenAI o1 sur les mathématiques, le codage et les sciences, déclenchant le 27 janvier la plus forte perte de capitalisation en une seule séance de l'histoire boursière américaine — de l'ordre de 600 milliards de dollars effacés chez Nvidia. Architecture Mixture-of-Experts de 671 milliards de paramètres totaux pour 37 milliards actifs, fenêtre de contexte de 128K tokens, et surtout une chaîne de raisonnement visible (balises think) qui rend sa logique entièrement auditable. Sur l'AIME 2024 il atteint 79,8 % — de justesse au-dessus des 79,2 % d'o1 — et 97,3 % sur MATH-500 ; sur Codeforces il talonne de près à 2 029 Elo. Licence MIT, pleinement commerciale et sans condition.

Une précision sur le coût, presque toujours mal rapportée : le fameux « moins de 6 millions de dollars » désigne l'entraînement du modèle de base DeepSeek-V3 sur lequel R1 se greffe ; l'entraînement de raisonnement propre à R1 a, lui, coûté environ 294 000 dollars, R1 étant par ailleurs le premier LLM validé par les pairs, dans une publication de la revue Nature. Ces chiffres excluent la recherche antérieure et l'infrastructure, mais l'ordre de grandeur a suffi à fissurer le dogme selon lequel l'IA de pointe exigeait des milliards. La build courante est aujourd'hui R1-0528 (mai 2025), qui réduit nettement les hallucinations et ajoute la sortie JSON et le function calling ; R2, lui, se fait attendre, repoussé sur la seconde moitié de 2026. R1 n'est donc plus l'unique référence du raisonnement ouvert, mais il demeure, sur le rapport qualité-prix du raisonnement profond, l'un des choix les plus défendables.

9. Llama 4 Scout

Le modèle de Meta pour les ultra-longs contextes, publié en avril 2025. Il établit bel et bien un record open source avec une fenêtre de contexte de 10 millions de tokens — dix fois le précédent plafond ouvert, le million de Qwen 3.5 — soit près de 78 fois une fenêtre standard de 128K, de quoi ingérer une base de code entière ou une année d'archives. Architecture Mixture-of-Experts, 109 milliards de paramètres totaux pour 17 milliards actifs (16 experts), multimodalité native texte-image, le tout tenant sur un unique GPU H100 en quantification 4 bits.

Mais ce record appelle une mise au point. Les 10 millions de tokens valent surtout pour la récupération d'information (retrieval, type « aiguille dans une botte de foin ») ; dès qu'il faut raisonner ou synthétiser à travers l'ensemble de la fenêtre, les performances se dégradent, et la fenêtre réellement exploitable se situe plus près de 5 millions que de 10. Sur Fiction.LiveBench, qui mesure la compréhension, Scout ne marque que 15,6 % à 128K tokens, là où Gemini 2.5 Pro atteint 90,6 %. S'ajoute la controverse de lancement : Meta avait soumis à LM Arena une version « optimisée pour la

conversation » jamais distribuée ; testé en version publique, Maverick chuta de la 2■ à la 32■ place et Scout sortit du top 100. Deux réserves enfin, décisives pour un public français : la licence Llama 4 n'est pas open source au sens de l'OSI (clause des 700 millions d'utilisateurs, obligations d'attribution), et surtout son volet vision est juridiquement indisponible pour les licenciés domiciliés dans l'Union européenne — un verrou dur pour bien des produits européens.

10. Mistral Large 3

Le flagship européen de Mistral AI, publié en décembre 2025 (identifiant mistral-large-2512). Architecture Mixture-of-Experts éparse de 675 milliards de paramètres totaux pour 41 milliards actifs — la première MoE de Mistral depuis la série Mixtral — entraînée de zéro sur 3 000 GPU NVIDIA H200, avec une fenêtre de contexte de 256K tokens. Surtout, elle paraît sous licence Apache 2.0 : une rupture majeure avec la licence de recherche restrictive de Mistral Large 2, et sans aucune des conditions de type « plafond d'utilisateurs » qui grèvent un Llama 4. Sur l'arène LMSYS, il se classe deuxième parmi les modèles ouverts et sixième toutes catégories confondues, propriétaires inclus.

C'est, de toute cette liste, le choix qui parle le plus directement aux équipes européennes — et pas seulement par patriotisme. Son multilinguisme est natif, inscrit dans l'objectif de pré-entraînement plutôt que greffé après coup, avec un français de premier plan et environ 85,5 % de précision sur un MMLU multilingue équilibré. Trait distinctif et précieux : il est entraîné à répondre « je ne sais pas » plutôt qu'à halluciner, ce qui en fait un socle solide pour les pipelines RAG et les applications d'entreprise où la justesse prime sur la créativité. Il accepte texte et image, sans être pour autant un modèle « vision-first » ni un modèle de raisonnement dédié — sur ce terrain, o3 ou Gemini 2.5 Pro gardent l'avantage. Enfin, la souveraineté ne s'arrête pas à la licence : Mistral construit ses propres datacenters en Europe, dont une installation à Bruyères-le-Châtel près de Paris opérationnelle à la mi-2026 — de quoi faire de Large 3 un modèle non seulement européen par l'éditeur, mais de plus en plus européen par l'infrastructure.

Il serait inexact de croire que l'excellence en matière d'intelligence artificielle soit désormais l'apanage exclusif des laboratoires dotés de milliers de GPU et de budgets d'entraînement qui défient l'entendement. La maturité d'un écosystème se mesure aussi — peut-être surtout — à sa capacité à descendre vers le plus grand nombre sans se trahir. Or c'est précisément ce que révèlent les cinq modèles qui viennent de compléter cette liste : une forme d'intelligence frugale, délibérément conçue pour prospérer là où les mastodontes s'essoufflent.

Phi-4-mini-instruct fonctionne sans GPU, sur le processeur d'un ordinateur ordinaire. SmolLM3 tient dans deux gigaoctets de mémoire tout en surpassant des modèles qui lui sont pourtant supérieurs en taille. Qwen3-8B s'exécute confortablement sur une carte graphique de milieu de gamme à six

gigaoctets de VRAM. Gemma 3 27B, le plus ambitieux de la série, se contente d'une seule RTX 4090 — une carte gaming, non un équipement de datacenter. Mistral Small 4, enfin, mobilise à peine six milliards de paramètres actifs malgré une architecture totale de cent dix-neuf milliards, accomplissant par là un tour de passe-passe technique que les ingénieurs de Mistral AI ont mis plusieurs années à rendre possible.

Ce qui frappe dans cette économie de moyens, c'est qu'elle n'est pas synonyme d'économie d'ambition. Ces modèles raisonnent, codent, traduisent, résument, et pour certains comprennent les images. Ils tournent hors connexion, respectent la confidentialité des données, et s'affranchissent de toute dépendance à un fournisseur d'API. Pour un développeur indépendant, une PME soucieuse de sa souveraineté numérique, ou simplement un praticien curieux qui souhaite expérimenter sans engager de budget cloud, ils représentent une porte d'entrée sérieuse vers l'IA locale — non pas un pis-aller, mais un choix délibéré et pleinement justifiable.

La leçon, au fond, n'est pas technique. Elle est stratégique : la vraie démocratisation de l'intelligence artificielle ne passera pas uniquement par les modèles les plus puissants, mais par ceux qui auront su rendre la puissance accessible.

11. Phi-4-mini-instruct (Microsoft — 3,8B paramètres)

Le plus petit de la liste, et le plus accessible : un transformer dense de 3,8 milliards de paramètres qui tourne sur le simple processeur d'un ordinateur ordinaire, sans carte graphique, via Ollama ou ONNX Runtime. Sa fenêtre de contexte de 128K tokens est remarquable pour la catégorie, et ses performances en raisonnement dépassent largement sa taille : il égale Llama 3.1 8B sur MMLU (73 %) et le surpasse en mathématiques, tout en pesant à peine 3 Go en quantification Q4. Multilingue, français inclus, doté du function calling, sous licence MIT pleinement commerciale. Une réserve que Microsoft assume d'ailleurs ouvertement : un modèle de cette taille ne peut stocker beaucoup de connaissances factuelles et peut se tromper sur les faits — d'où l'intérêt de l'augmenter d'un moteur de recherche en configuration RAG. Son terrain naturel est donc l'assistance locale, l'instruction et le RAG plutôt que la culture générale encyclopédique. Idéal pour les machines sous contrainte, les étudiants, les déploiements hors ligne et les environnements sans accès Internet. C'est le point d'entrée absolu de la liste.

Pour qui voudrait monter en gamme sans quitter l'écosystème Microsoft, la même famille MIT compte un Phi-4 de 14B, nettement plus musclé en raisonnement et en mathématiques, et un Phi-4-multimodal de 5,6B capable de traiter texte, image et audio au sein d'un seul modèle.

12. SmolLM3-3B (Hugging Face — 3B paramètres)

À l'échelle des 3 milliards de paramètres, il surpasse Llama-3.2-3B et Qwen2.5-3B tout en restant compétitif avec des alternatives de la classe 4B comme Qwen3 et Gemma3, sur douze benchmarks populaires. Entraîné sur 11,2 billions de tokens, il offre un double mode de raisonnement (think / no-think) commutable à la demande, une fenêtre de contexte de 64K extensible à 128K par YaRN, et un support multilingue natif de six langues — dont le français. Sous licence Apache 2.0, il tourne sur une machine dotée d'à peine 4 Go de RAM, sans GPU. Mais sa singularité tient ailleurs : c'est un modèle véritablement et intégralement ouvert. Là où la plupart des « open weights » se contentent de publier leurs poids, Hugging Face livre ici tout le plan de fabrication — poids, code, configurations d'entraînement et composition publique du mélange de données. C'est le seul modèle de cette liste reproductible de bout en bout, et donc le seul qui réponde à la définition stricte de l'open source telle qu'on la distingue en introduction.

13. Qwen3-8B (Alibaba — 8B paramètres)

Modèle dense de 8,2 milliards de paramètres qui tourne sur une carte milieu de gamme avec 5 à 6 Go de VRAM en quantification Q4_K_M, à plus de 40 tokens par seconde — une simple commande Ollama suffit (ollama run qwen3:8b). Sa signature est un double mode commutable à la demande : « thinking » pour le raisonnement pas-à-pas (math, code, logique) et « non-thinking » pour la réponse rapide, le tout dans un seul modèle. Multilingue natif sur 119 langues, français inclus, il est excellent en codage pour sa catégorie — environ 1810 d'Elo CodeForces en mode raisonnement, au-dessus de GPT-4o, et il soutient la comparaison avec le Qwen2.5-14B de la génération précédente, deux fois plus gros. Fenêtre de contexte de 32K tokens, extensible à 131K par YaRN. Licence Apache 2.0. C'est le point de départ pratique par excellence pour qui veut un assistant local de qualité avant de passer à des configurations plus lourdes.

14. Gemma 3 27B (Google DeepMind — 27B paramètres)

Il tourne sur une seule RTX 4090 en quantification Q4, sans multi-GPU ni matériel serveur — juste une carte gaming classique, la quantification QAT préservant une qualité proche du bf16. Sorti en mars 2025, il accepte texte et image, couvre plus de 140 langues, et offre une fenêtre de contexte de 128K tokens, généreuse pour un modèle de cette taille. Une précision toutefois, qui corrige une idée reçue tenace : ce modèle n'est pas sous licence Apache 2.0, mais sous la licence « Gemma Terms of Use », assortie d'une politique d'usages interdits et soumise à l'acceptation des conditions de Google. C'est un cas d'école d'open weights — poids librement téléchargeables — sans être de l'open source au sens strict. Cela n'enlève rien à ses qualités : il offre le meilleur équilibre capacité/accessibilité de la liste — du moins pour sa génération, car sa relève est déjà là : Gemma 4 (cf. n°6) a depuis repris le flambeau sous une vraie licence Apache 2.0, son MoE de 26B offrant une qualité comparable tout en

tenant sur 8 Go de RAM.

15. Mistral Small 4 (Mistral AI — 119B total / 6B actifs)

Architecture Mixture-of-Experts de 119 milliards de paramètres totaux (128 experts, 4 actifs par token) pour seulement 6 milliards activés à chaque inférence : en vitesse de calcul, il se comporte comme un dense d'environ 22B. Présenté en mars 2026 à la GTC de NVIDIA, il a pour véritable ambition d'unifier en un seul checkpoint quatre produits jusque-là distincts — l'instruct (Small), le raisonnement (Magistral), la vision (Pixtral) et le codage (Devstral). De là son atout distinctif : un curseur de raisonnement configurable par requête (reasoning_effort), qui bascule à la demande entre réponse légère et analyse pas-à-pas, sans maintenir plusieurs modèles. Il accepte texte et image, offre 256K tokens de contexte, et affiche 40 % de latence en moins et un débit trois fois supérieur à Small 3. Licence Apache 2.0.

Une réserve essentielle, qui tempère sa place dans cette section : « Small » est ici un nom trompeur. S'il est frugal en calcul, il ne l'est pas en mémoire — ses 119 milliards de paramètres doivent tous résider en VRAM, soit plus de 60 Go en Q4 et, en pratique, du matériel de datacenter. Ce n'est donc pas le choix des équipes « à ressources limitées » au sens matériel, mais celui des équipes européennes qui veulent un modèle unique, multimodal et librement licencié pour consolider toute une pile applicative.

Tableau récapitulatif

État des modèles abordés dans le guide, au 21 juin 2026. Coût d'inférence indicatif et relatif.

Modèle	Paramètres	Contexte	Licence	Coût	Inférence locale
DeepSeek-V4-Pro (DeepSeek)	1,6 T / 49 B (MoE)	1 M	MIT	**E**	**NON**

DeepSeek-V4-Flash (DeepSeek)	284 B / 13 B (MoE)	1 M	MIT	**B**	**GPU HAUT DE GAMME**
MiMo-V2.5-Pro (Xiaomi)	1,02 T / 42 B (MoE)	1 M	Permissive	**D**	**NON**
Kimi-K2.6 (Moonshot AI)	1 T / 32 B (MoE)	256 K	MIT mod.	**D**	**NON**
GLM-5.2 (Z.ai)	744 B / 40 B (MoE)	1 M	MIT	**D**	**NON**
Qwen3.5-397B-A17B (Alibaba)	397 B / 17 B (MoE)	256 K	Apache 2.0	**C**	**GPU HAUT DE GAMME**
Gemma 4 (Google DeepMind)	31 B dense • 26 B/4 B (MoE)	128–256 K	Apache 2.0	**A**	**GPU CLASSIQUE**
MiniMax-M2.7 (MiniMax)	230 B / 10 B (MoE)	205 K	Open weights *	**B**	**GPU HAUT DE GAMME**
DeepSeek-R1 (DeepSeek)	671 B / 37 B (MoE)	128 K	MIT	**C**	**NON (distill. : classique)**

Llama 4 Scout (Meta)	109 B / 17 B (MoE)	10 M	Llama 4 (non-OSI)	**C**	**GPU HAUT DE GAMME**
Mistral Large 3 (Mistral AI)	675 B / 41 B (MoE)	256 K	Apache 2.0	**C**	**GPU HAUT DE GAMME**
Phi-4-mini-instruct (Microsoft)	3,8 B (dense)	128 K	MIT	**A**	**GPU CLASSIQUE / CPU**
SmolLM3-3B (Hugging Face)	3 B (dense)	64–128 K	Apache 2.0 (full OSS)	**A**	**GPU CLASSIQUE**
Qwen3-8B (Alibaba)	8,2 B (dense)	32–131 K	Apache 2.0	**A**	**GPU CLASSIQUE**
Gemma 3 27B (Google DeepMind)	27 B (dense)	128 K	Gemma ToU	**B**	**GPU CLASSIQUE**
Mistral Small 4 (Mistral AI)	119 B / 6 B (MoE)	256 K	Apache 2.0	**C**	**GPU HAUT DE GAMME**

Légende

Coût d'inférence (indicatif, relatif) : A = très faible · B = faible · C = modéré · D = élevé · E = très élevé. L'échelle combine le prix par token via API et le coût matériel d'auto-hébergement.

Inférence locale :

NON — nécessite plusieurs GPU de datacenter ; pas de chemin réaliste sur une machine individuelle.

GPU HAUT DE GAMME — une seule machine haut de gamme : un H100 80 Go, un Mac Studio à mémoire unifiée (128–256 Go), ou un GPU 24 Go + forte RAM (offloading MoE).

GPU CLASSIQUE — une carte grand public (type RTX 4090), voire 6–8 Go de VRAM ou un simple processeur pour les plus petits.

Abréviations : MIT mod. = MIT modifiée (mention obligatoire au-delà de 100 M d'utilisateurs) · Gemma ToU = Gemma Terms of Use · full OSS = open source intégral (poids + données + recette) · * MiniMax-M2.7 : poids ouverts, mais usage commercial soumis à autorisation écrite.

Conclusion — Le luxe de pouvoir choisir

Au terme de ce parcours, une chose s'impose : la question posée en introduction — à quel modèle aurai-je le droit d'accéder demain ? — a cessé d'être angoissante. Non que la menace ait disparu, l'épisode Fable 5 étant là pour rappeler qu'un accès peut se retirer d'un simple trait de plume ; mais parce que la réponse, désormais, ne dépend plus d'un seul acteur. Une quinzaine de modèles, issus de laboratoires chinois, américains et européens, couvrant tout l'éventail du géant de mille milliards de paramètres au modèle de trois milliards qui tient sur un ordinateur ordinaire, sont aujourd'hui téléchargeables, hébergeables et modifiables par quiconque le décide. La dépendance à une poignée d'acteurs fermés n'est plus une fatalité technique : elle est devenue un choix — et un choix que rien n'oblige à faire.

Encore faut-il choisir avec discernement. C'est tout le sens de la distinction, posée dès les premières pages, entre l'open source véritable et les poids simplement ouverts : entre la liberté et la permission. Nous l'avons vue à l'œuvre tout au long de cette liste — ici une licence Apache 2.0 sans la moindre condition, là une autorisation commerciale à solliciter avant tout usage, ailleurs un verrou régional qui prive l'utilisateur européen d'une fonctionnalité. Lire ce que l'on signe, ou ce que l'on ne signe pas, demeure le premier réflexe de l'esprit prudent ; car la puissance d'un modèle ne vaut rien si les conditions de son usage peuvent se refermer un matin, sans préavis.

Reste l'essentiel, qui n'est pas technique mais stratégique. Cette liste est un instantané : dans six mois, certains de ces modèles auront été dépassés, d'autres seront nés — le mouvement, lui, ne se renversera pas. L'intelligence artificielle de pointe est en train de descendre vers le plus grand nombre, vers l'indépendant, la PME, l'administration, le chercheur, sans rien perdre de sa capacité. À celui qui s'inquiétait de sa souveraineté numérique, ce guide aura voulu adresser un message simple : les outils existent, ils sont libres, et ils tournent, ici et maintenant, sur sa propre machine. Le reste ne tient plus

qu'à une décision.