#12

Uber hardware

LE PC AVEC PASSION, SANS CONCESSION

IA LOCALE

LE GRAND SAUT

INSTALLER, CONFIGURER ET OPTIMISER SON LLM DE A À Z

LLAMA, MISTRAL, PHI : QUEL MODÈLE POUR QUEL USAGE ?

VRAM : 8 GO vs 16 GO vs 32 GO

MAÎTRISER

INSTALLER WINDOWS 11 SUR UN VIEUX PC

TESTS

HAVN BF 360 FLOW, UN SUPERBE BOITIER TRÈS BIEN VENTILÉ

UGREEN DH2300 LE MEILLEUR NAS À PETIT PRIX

Uber hardware .com

12

12 | PRISE DE RISQUE

L'EDITO

Uber Hardware a toujours eu un penchant pour le gaming et le hardware pour joueurs, à tel point que vous écrivez régulièrement pour nous demander d'explorer d'autres disciplines dans lesquelles le PC excelle. Et s'il en est une qui a le vent en poupe, c'est bien l'intelligence artificielle !

Vaste domaine, s'il en est. Pour cette première approche, on vous embarque dans l'univers incroyable des LLM. Vous qui conversez déjà avec eux du matin au soir, au point d'en oublier les bons vieux moteurs de recherche, avez-vous la moindre idée des besoins matériels qui se cachent derrière ? Saviez-vous que la quantité de mémoire de votre carte graphique importe plus que la puissance de son GPU ? Qu'il était possible de faire tourner son propre modèle de langage à la maison et d'en personnaliser le fonctionnement ? On vous invite d'ailleurs à lire des imposant dossier de 51 pages, même si vous n'éprouvez pas le besoin d'avoir votre propre LLM, vous apprendrez sûrement deux ou trois trucs sur leur fonctionnement. Ce 12e opus, ce sont aussi des tests de composants très sympas, à l'image du second boîtier de la jeune marque HAVN et ses finitions exceptionnelles. On prend aussi des nouvelles d'EKWB, le leader du watercooling dans la tourmente depuis quelques années. Également dans ce numéro, un petit article des plus pratique qui vous explique comment installer Windows 11 sur un PC qui ne répond pas aux exigences matérielles, et sans compte Microsoft imposé. Allez, zou ! On embraye directement sur UH #13, on se retrouve très vite !

Thomas Olivaux

SOMMAIRE

6

DOSSIER

IA LOCALE
LE GRAND SAUT

INSTALLER, CONFIGURER ET OPTIMISER SON LLM DE A À Z LLAMA, MISTRAL, PHI : QUEL MODÈLE POUR QUEL USAGE ?

Vous êtes-vous déjà demandé ce qui se passait derrière le rideau quand vous posez une question à ChatGPT ? Saviez-vous qu'on pouvait lancer des LLM sur son propre ordinateur pour s'affranchir du cloud ? Les besoins en VRAM/RAM sont-ils vraiment si importants ? UH vous explique tout.

VRAM : 8 GO vs 16 GO vs 32 GO

76 ALIMENTATION
NZXT C750 GOLD CORE
Un bon plan Platinum... si votre PC démarre !

TESTS

60 BOÎTIER
HAVN BF 360 FLOW
Superbes finitions et ventilation au top

82 PC PORTABLE
MSI TITAN 18 HX AI A2XW
Dé-me-sure

70 AIO
EK NUCLEUS CR360 LUX DRGB
Une finition soignée, à la EK

94 NAS
UGREEN NASYNC DH2300
Le challenger est aussi le meilleur choix

Uber Geek 106

MAITRISER

102

INSTALLER WINDOWS 11 SUR UN VIEUX PC

OUTREPASSER LES LIMITES MATÉRIELLES ET LE COMPTE MICROSOFT AVEC RUFUS

Grâce à Rufus, l'indétrônable, il n'a jamais été aussi facile d'installer Windows 11 sur un PC non supporté et/ou de faire l'impasse sur le compte Microsoft. On peut même aller plus loin avec FlyOOBE.

UBER HARDWARE N° 12
FÉVRIER 2026

Prix de l'abonnement mensuel : 4,40 €
Prix de l'abonnement annuel : 45 €
Bulletin d'abonnement :
uberhardware.com/sinscrire/
L'abonnement donne accès au numéro en cours ainsi qu'aux anciens numéros.

Uber Hardware est une publication (presque) mensuelle, sur abonnement, proposée à la fois sous la forme d'un site Web et d'un ebook PDF, éditée par :
Uber Hardware
4, rue du Roudour
29410 Le Cloître Saint Thegonnec
SIRET : 98362581500011

Pour tout renseignement :
contact@uberhardware.com

Directeur de la publication :
Thomas Olivaux
Création graphique : David Benamou
Conception Web : David Benamou
Ont participé à ce numéro :
Thomas Olivaux, Manu Da Costa

Remerciements :
Guillaume, Laurent, PBG, Victor

Textes, photos : © 2026 Uber Hardware

L'envoi de tout texte, photo ou vidéo implique l'acceptation par l'auteur de leur libre publication dans le journal. Les documents ne sont pas retournés. La loi du 11 mars 1957 n'autorisant aux termes des alinéas 2 et 3 de l'article 41 d'une part que "des copies ou reproductions strictement réservées à l'usage du copiste et non destinées à une utilisation collective" et d'autre part que les analyses et courtes citations dans un but d'exemple et d'illustration, "toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite" (alinéa premier de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles 425 et suivants de l'ancien code pénal.

IA LOCALE

LE GRAND SAUT

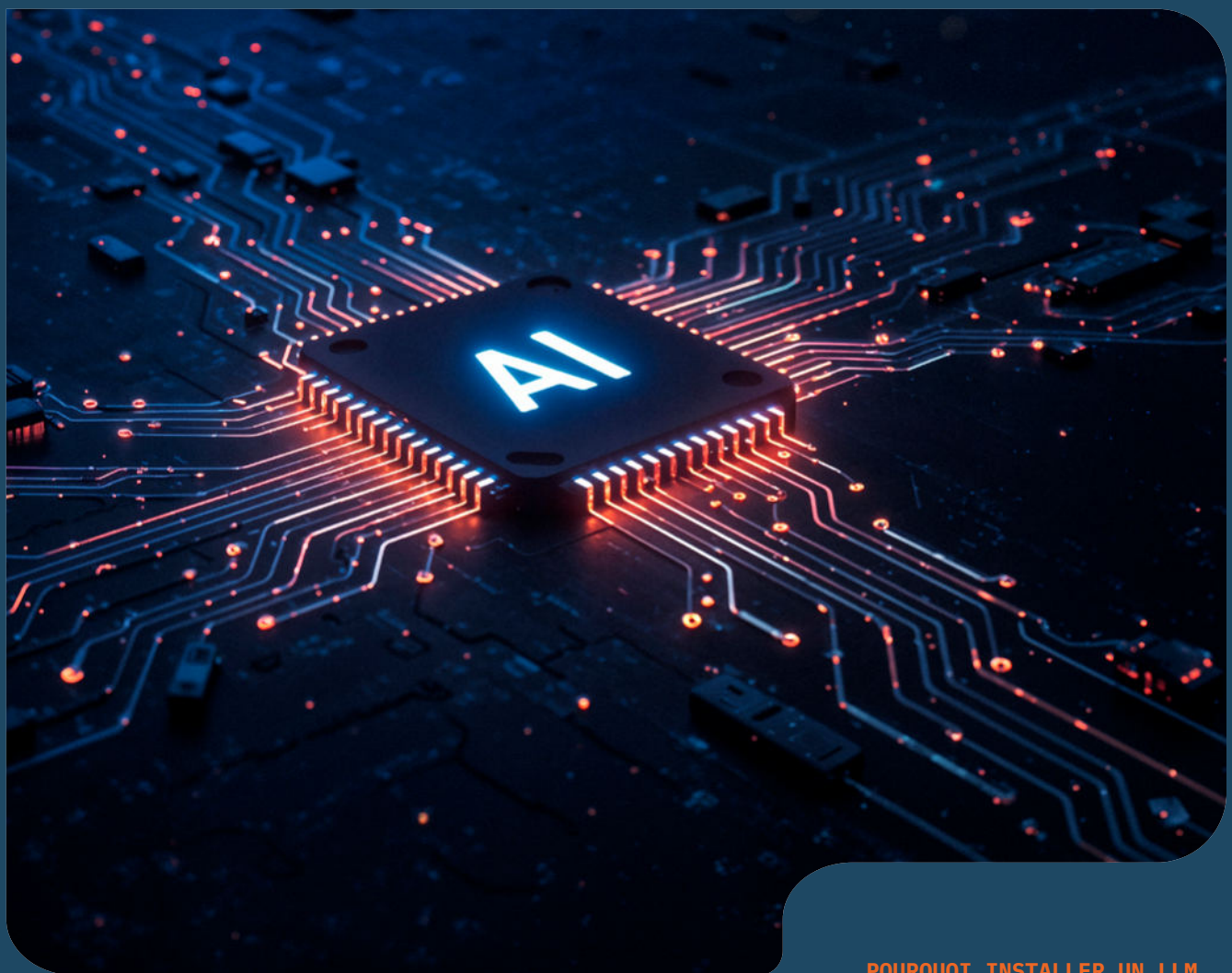
INSTALLER, CONFIGURER
ET OPTIMISER SON LLM DE A À Z

LLAMA, MISTRAL, PHI ;
QUEL MODÈLE POUR QUEL USAGE ?

VRAM :
8 GO vs 16 GO
vs 32 GO

Vous êtes-vous déjà demandé ce qui se passait derrière le rideau quand vous posez une question à ChatGPT ? Saviez-vous qu'on pouvait lancer des LLM sur son propre ordinateur pour s'affranchir du cloud ? Les besoins en VRAM/RAM sont-ils vraiment si importants ? UH vous explique tout.

par Thomas Olivaux



Quoi qu'on en pense, l'intelligence artificielle s'impose de plus en plus dans notre quotidien. En particulier les LLM, les incroyables modèles de langage tels que Gemini, Chat GPT ou Grok à qui nous demandons de plus en plus chaque jour. Au gré de leurs progrès, rapides, et des nôtres, car en tirer le meilleur requiert une certaine expertise. En quelques années à peine, l'IA est devenue une discipline à part entière de l'informatique, comme la programmation ou le réseau. Et les LLM l'une de ses principales composantes. Pour tout fan de PC, en particulier les passionnés de hardware, l'envie d'en savoir plus et de faire davantage avec sa machine est grande. C'est en réponse à de multiples mails de lecteurs, reçus à l'adresse contact@uberhardware.com, que nous vous avons concocté ce dossier spécialement consacré aux LLM locaux, de leur installation à leur optimisation en passant par l'analyse des besoins matériels. On vous invite d'ailleurs à le lire entièrement même si vous n'avez que faire de jouer et d'héberger un LLM local, rien que pour mieux comprendre le fonctionnement des grands modèles de langage que vous utilisez chaque jour dans le cloud.

POURQUOI INSTALLER UN LLM SUR PC ?

L'intérêt de basculer vers une IA locale réside avant tout dans la réappropriation de sa souveraineté numérique puisque, en traitant l'information directement sur son propre matériel, on s'assure qu'aucune donnée sensible ou privée ne part alimenter les serveurs de géants de la tech. Cette démarche garantit une confidentialité absolue, mais offre également une liberté d'usage totale en s'affranchissant des filtres de modération souvent restrictifs des services en ligne, tout en éliminant les frais d'abonnement récurrents et la nécessité d'une connexion internet

permanente. En faisant ce choix, l'utilisateur passe du statut de simple locataire d'une technologie à celui de propriétaire de son outil de travail, bénéficiant d'une puissance de calcul disponible à tout instant et totalement imperméable aux changements arbitraires ou aux pannes des fournisseurs de services cloud. Pour le particulier, c'est surtout un sujet d'étude, quoiqu'on puisse imaginer de réels usages comme la rédaction d'un roman ou aider les enfants à réviser en se basant sur leurs propres cours. Pour les professionnels, les usages sont multiples, du cabinet d'avocat qui souhaite synthétiser des centaines de pièces sans faire fuiter la moindre information au SAV de n'importe quelle marque désireuse de mettre en place un chatbot d'accueil performant afin de répondre à ses clients. Pour illustrer ce dossier, nous utiliserons cinq LLM de référence : Phi 4 Mini, Llama 3.1 et 3.3, Mistral Small et Command R+, tout en vous expliquant comment les utiliser vous-même. Milliards de paramètres, quantification, fine tuning, on vous dit tout !

UN BANC D'ESSAI À 10 PC !

Afin d'estimer au mieux les besoins des LLM locaux et leurs capacités en fonction du matériel dont vous disposez, nous avons assemblé/réuni un sacré nombre de PC ! Pas moins de dix machines en tout ! Ça va du modeste Core i5-10500 avec une ARC B580 à la toute puissante workstation IA proposée dans le dossier configs du mois dernier et son couple Ryzen 9 9900X/RTX 5090, en passant par des PC portables et même un mini PC spécial IA basé sur la puce Nvidia GB10 ! Nous n'avons pas inclus d'anciens GPU avec moins de 8 Go de VRAM car c'est la valeur minimale pour espérer faire quelque chose d'utile avec des chatbots, du moins en bénéficiant du boost de vitesse que procure la VRAM comparée à la RAM standard. Car, au cas où vous ne le sauriez pas encore, c'est avant tout et surtout la mémoire (en premier la mémoire vidéo, en second la RAM) qui conditionne le plus les capacités et performances des ordinateurs pour l'inférence. En parlant de mémoire vive, ces PC ont entre 16 et 48 Go de RAM, de la DDR4, de la DDR5 ou de la LPDDR5x selon l'ordinateur, à l'exception de la machine GB10 qui bénéficie de 128 Go de mémoire unifiée (partagée entre la partie CPU et GPU de son SoC). Nous aurions pu mettre 64 Go dans la Workstation IA, voire davantage, mais vu le contexte actuel de pénurie de RAM, ne soyons pas trop gourmands.

DE MULTIPLES TESTS, 3 SCÉNARIOS D'USAGE

Nous avons réalisé de nombreux tests, basés sur les cinq modèles cités précédemment et représentant de multiples usages, pour d'analyser les performances dans ce domaine de l'inférence. À la fin, histoire que vous puissiez vous projeter, nous détaillerons trois scénarios d'usage. Nous les avons baptisés le créatif (écrire, traduire, trouver des idées de scénarios ou de marketing), l'analyste (résumer des réunions de 2 heures, extraire des données d'un PDF de 100 pages, classer des informations) et l'expert (déboguer du Python, écrire des requêtes SQL complexes, résoudre des problèmes de maths). Le but : savoir si ce type de chatbot est réalisable avec un simple PC, avec quel matériel, quelles aptitudes et limitations, et quel confort d'usage. Plus de détails à leur sujet dans la suite du dossier.

> Entre étude technique et outil professionnel, le LLM sur PC s'affranchit du cloud. Analyse des besoins matériels et tests de performance sur les modèles de référence.

Un article de qualité ne doit pas soulever plus de questions que le nombre auquel il répond, mais sachant qu'on aborde pour la première fois un domaine aussi vaste qu'intéressant et complexe, on ne vous promet pas que cette promesse sera tenue : vous aurez certainement envie d'en savoir encore plus après sa lecture ! Gardez en tête que nous ne sommes pas non plus des professionnels de l'IA, nous avons abordé ce thème comme toujours : avec passion ! D'ailleurs, si vous appréciez cet article, n'hésitez pas à nous le faire savoir, nous pourrons vous en proposer d'autres à l'avenir ! Par exemple un spécial IA générative avec les meilleurs outils pour concevoir des images, de la musique et des vidéos, y compris en local. Ou encore un maîtriser spécial Moltbot (anciennement Clawdbot), votre assistant IA personnel super évolué qui intègre toute votre vie privée (emails, calendrier, WhatsApp et compagnie). Il ne nous reste qu'à vous souhaiter une bonne lecture et de bonnes expériences !

COMPRENDRE LA MÉCANIQUE : LE DICTIONNAIRE INDISPENSABLE DU LLM

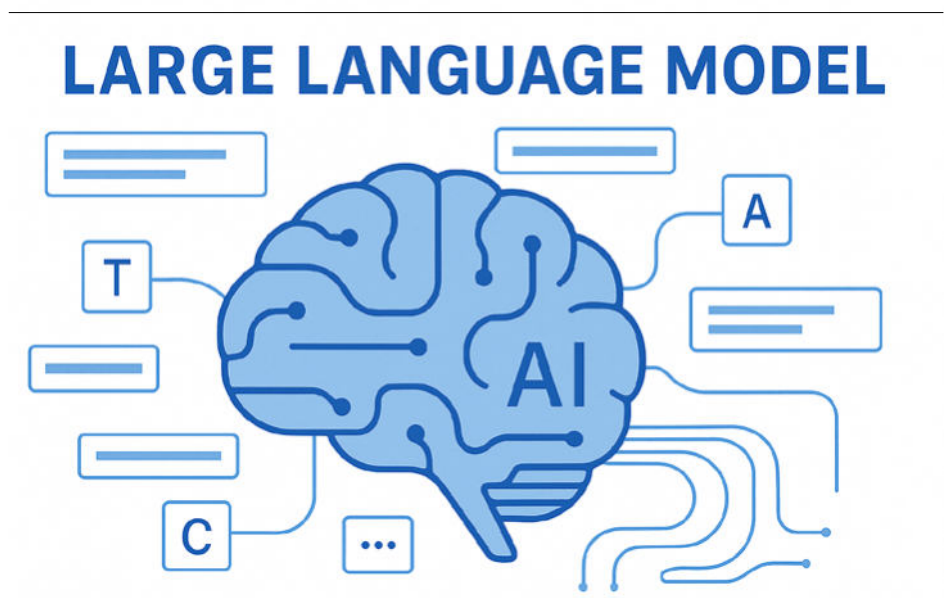
Puisqu'on aborde le sujet LLM pour la première fois, en particulier du point de vue de sa configuration, il est important d'en cerner le fonctionnement. De bien avoir en tête, notamment, qu'un LLM ne fonctionne que par probabilités et qu'il ne récite pas du texte appris par cœur. Vous pouvez poser trois fois de suite la même question au même modèle et obtenir trois réponses légèrement différentes. Mais il y a bien plus à savoir. Par exemple, à quoi correspondent 8B dans Llama 3.1 8B ? Qu'est-ce qu'un modèle "open weights" ? C'est quoi un token ? Et l'inférence ? Découvrons tout ça.

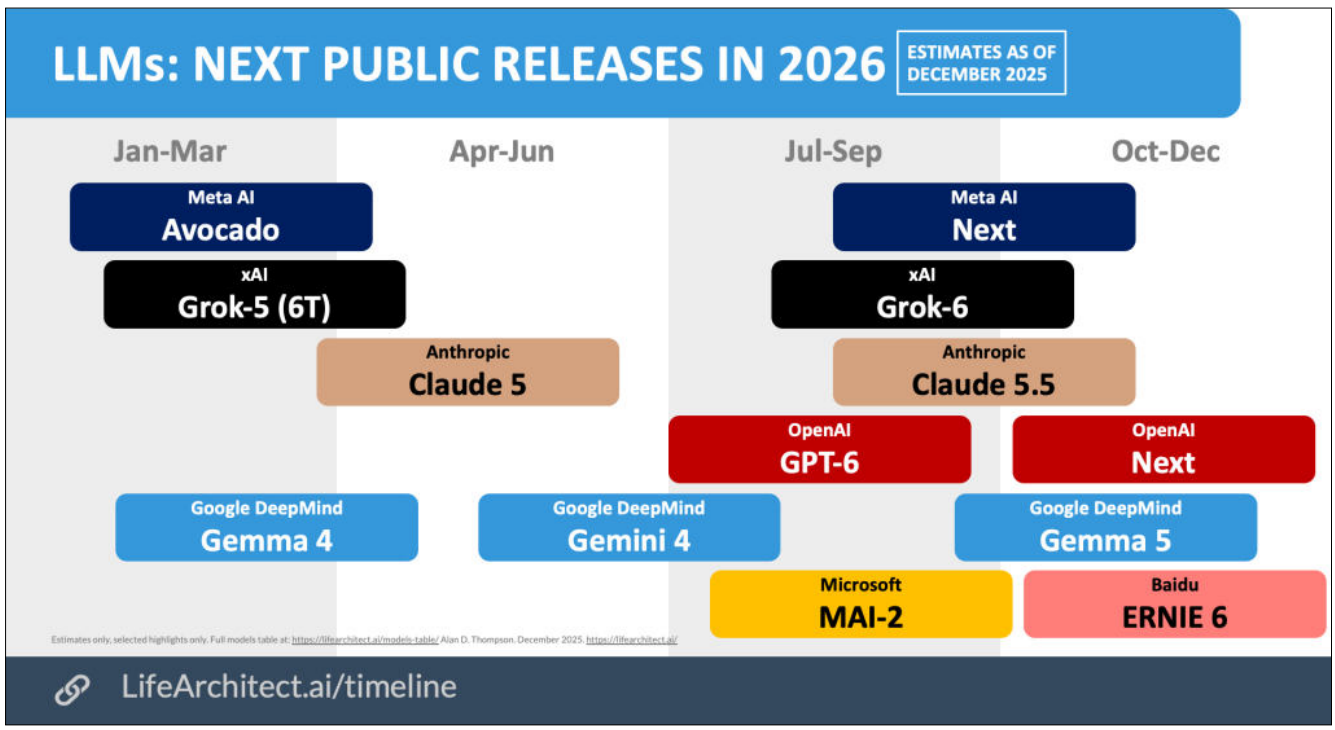
C'EST QUOI UN LLM DÉJÀ ?

Un LLM (Large Language Model), ou grand modèle de langage, est un algorithme de deep learning capable de comprendre, de générer et de manipuler du texte de manière naturelle. Contrairement aux anciens programmes basés sur des règles rigides, le LLM fonctionne par probabilités statistiques. Pour faire simple, il ne "comprend" pas le sens des mots comme un humain, mais il a été entraîné sur des volumes colossaux de données textuelles (livres, code, articles) pour prédire quel est le prochain fragment de mot (token, on va y revenir) le plus probable après une suite donnée. C'est la puissance de calcul des processeurs modernes qui permet à ces modèles de simuler une conversation ou un raisonnement complexe en calculant des milliards de paramètres simultanément. En local, le défi consiste à héberger l'intégralité de ces paramètres dans la mémoire de votre machine (VRAM ou RAM) pour s'affranchir de toute connexion internet.

CLOUD OU LOCAL : À CHAQUE USAGE SON MOTEUR

L'opposition entre les services cloud (ChatGPT, Gemini, Grok…) et l'exécution locale n'est pas une guerre de puissance pure, mais une question de compromis. Le LLM en ligne offre l'accès à des modèles XXL dont les centaines de milliards de paramètres et les fenêtres de contexte gigantesques (plus d'explications ci-après) nécessitent des infrastructures industrielles. Il excelle dans la culture générale infinie et les tâches multimodales lourdes, mais au prix d'une dépendance aux serveurs tiers et d'un coût à l'usage (abonnement mensuel pour s'affranchir des limites rapidement atteintes des services gratuits). Le LLM local, de son côté, s'impose sur les piliers de la souveraineté et de la flexibilité.





Vos données personnelles ou professionnelles ne transitent par aucun serveur externe, garantissant une confidentialité totale. Surtout, vous reprenez le contrôle ; pas de latence réseau, pas de filtres de modération arbitraires et un coût par token qui devient nul une fois le matériel amorti. Si le cloud reste la solution de facilité pour les modèles de pointe, l'IA locale est l'outil de l'indépendance, idéal pour le traitement de fichiers sensibles et la personnalisation profonde de l'outil.

DES MODÈLES PROPRIÉTAIRES ET DES MODÈLES "OPEN WEIGHTS"

La création d'un LLM est une épreuve de force technologique qui se déroule en amont de votre PC. Elle est menée par des géants de la Tech (Meta avec Llama, Google avec Gemma, Microsoft avec Phi) ou des champions de l'Open Source et de l'IA spécialisée (Mistral AI en France, DeepSeek en Chine). Le processus commence par le préentraînement, où le modèle est exposé à des pétaoctets de données textuelles. Cette phase nécessite des infrastructures titanesques ; des clusters de milliers de GPU (type Nvidia H100 ou B200) tournant pendant des mois, pour un coût de plusieurs millions, voire milliards de dollars. C'est à ce stade que les "poids" du modèle sont fixés, définissant son intelligence brute. Une fois entraîné, le modèle peut être proposé sous forme de service fermé (Cloud propriétaire) ou, de plus en plus, sous forme de modèles "Open Weights". Ces derniers permettent aux utilisateurs de télécharger

Tout est question d'échelle. Il aura fallu 100 000 Nvidia H100 rien que pour le pré-entrainement de Grok 3 fin 2024 !

la structure du modèle pour l'exécuter localement, bénéficiant ainsi de l'investissement colossal des concepteurs sans avoir à le reproduire.

DE L'ART DE MANIER LE PROMPT

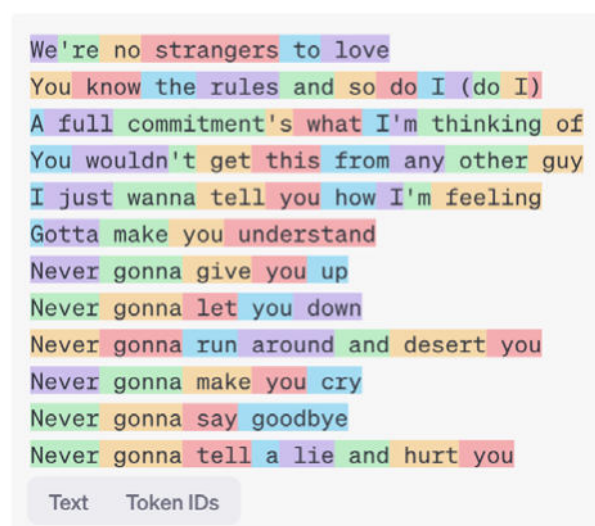
Le prompt est l'instruction, la question ou le texte que vous soumettez au modèle pour obtenir une réponse. Loin d'être une simple requête posée à un moteur de recherche, le prompt agit comme une véritable ligne de commande en langage naturel. Sa structure détermine la manière dont l'IA va naviguer dans ses milliards de paramètres pour activer les connexions neuronales les plus pertinentes. Un prompt efficace repose généralement sur trois piliers ; le contexte (qui doit être l'IA ?), la tâche (que doit-elle faire précisément ?) et les contraintes (quel format de réponse, quelle longueur, quel ton ?). On parle de Prompt Engineering pour désigner l'art d'affiner ces instructions. Pour l'utilisateur, un prompt bien conçu n'est pas seulement une question de précision sémantique, c'est aussi un gain d'efficience ; en guidant le modèle vers la bonne réponse dès le premier essai, on évite les itérations coûteuses et la génération de tokens inutiles, optimisant ainsi le temps de traitement de votre matériel.

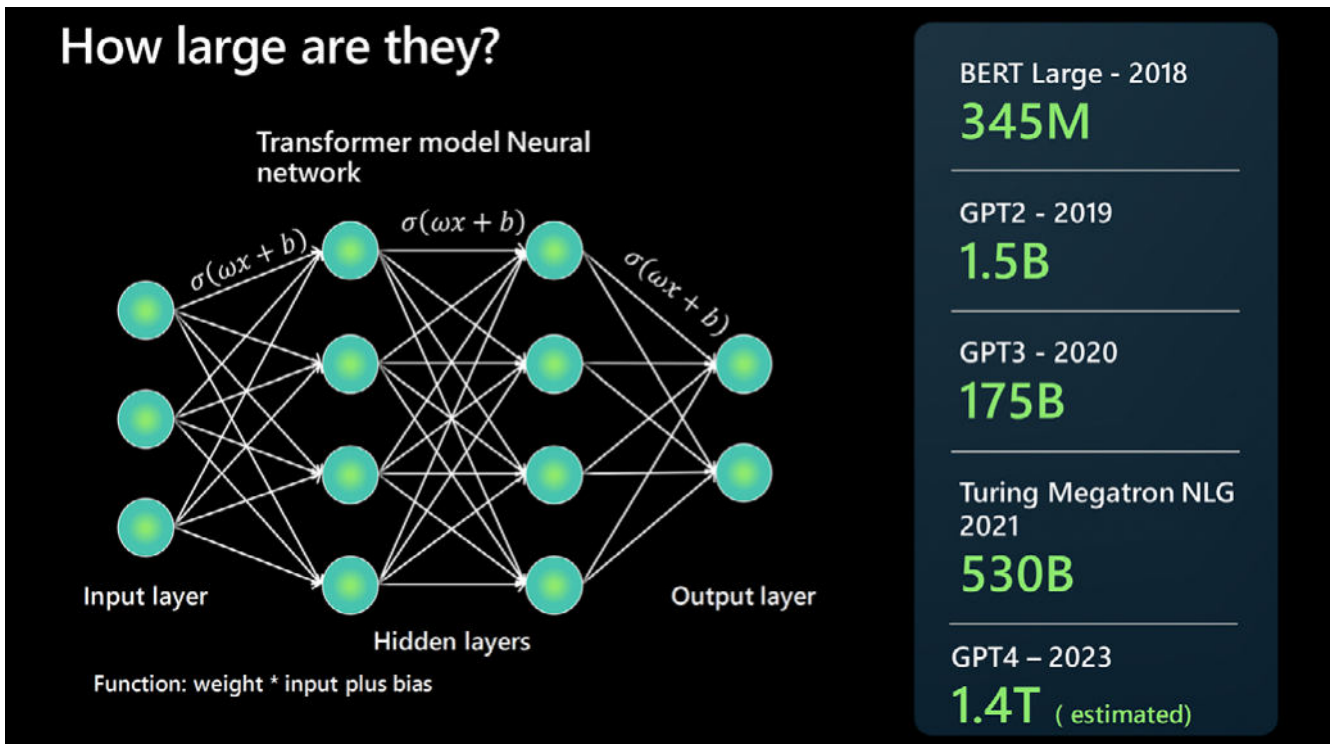
INFÉRENCE, CE MOT ÉTRANGE POUR DIRE RÉPONSE

L'inférence est le processus par lequel un modèle d'IA utilise les connaissances acquises lors de son entraînement pour répondre à une question ou accomplir une tâche. Si l'entraînement est la phase d'apprentissage (comparable à des années d'études), l'inférence est l'examen en temps réel ; c'est le moment où vous soumettez un prompt et où la machine calcule la réponse. Techniquement, c'est durant cette phase que votre matériel est mis à l'épreuve. Le processeur (GPU ou CPU) parcourt les milliards de paramètres du modèle pour prédire, les uns après les autres, les tokens qui constitueront la réponse finale. Pour le lecteur de magazine hardware, l'inférence est le seul moment qui compte vraiment au quotidien, car c'est elle qui détermine si votre PC "réfléchit" instantanément ou s'il s'essouffle à produire trois mots par minute. Sa fluidité dépend d'un équilibre critique entre la puissance de calcul brute (FLOPS) et la vitesse de transfert des données (bande passante).

LE CONCEPT DE TOKEN

Le token est l'unité de mesure fondamentale d'un LLM. Contrairement à nous, l'IA ne lit pas des mots entiers, mais découpe le texte en morceaux. Un token peut correspondre à un mot court ("chat"), une partie de mot ("biolo" et "gie"), ou même un signe de ponctuation. En moyenne, dans la langue française, on considère que 1000 tokens représentent environ 750



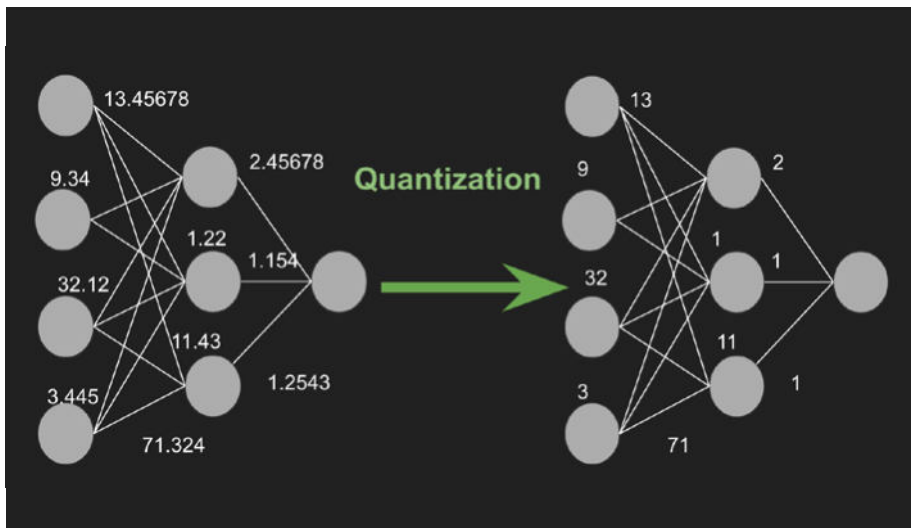


> Entre capacité de mémorisation et poids du fichier, le nombre de paramètres est le facteur limitant numéro un : un curseur qui définit la frontière entre l'IA domestique et les géants du cloud.

mots. Pourquoi est-ce crucial pour le hardware ? Parce que chaque token généré demande un passage complet de l'intégralité des paramètres du modèle à travers la bande passante de votre mémoire (VRAM). C'est pour cette raison que l'on mesure la performance d'une carte graphique ou d'un processeur en tokens par seconde (tok/s) ; plus votre mémoire est rapide, plus l'IA peut "prédire" et afficher ces fragments de mots rapidement à l'écran.

C'EST QUOI DONC CETTE HISTOIRE DE MILLIARDS DE PARAMÈTRES !?

Le chiffre qui accompagne souvent le nom d'un modèle (Llama 3 8B, Mistral 7B, etc.) désigne son nombre de paramètres. Pour schématiser, les paramètres sont l'équivalent des synapses dans un cerveau humain ; ce sont des variables mathématiques (des "poids") qui ont été ajustées durant l'entraînement pour définir la force des connexions entre les neurones du modèle. Plus un modèle possède de paramètres, plus sa capacité de mémorisation, de raisonnement et de nuance est élevée. C'est là que réside toute l'intelligence du LLM. Pour le matériel, ce nombre est le facteur limitant numéro un ; chaque paramètre doit être stocké en mémoire (VRAM ou RAM) pour que l'IA puisse fonctionner. Si un modèle de 8 milliards de paramètres peut tenir sur une carte graphique grand public, un géant de 70 ou 400 milliards exigera une architecture mémoire bien plus vaste. C'est la corrélation directe entre ce chiffre et le poids du fichier sur votre disque qui dicte, in fine, la configuration nécessaire pour le faire tourner. Nous reviendrons longuement sur cette histoire de paramètres, tant pour mesurer ce que ça permet d'un point de vue intelligence et les impacts matériels tout au long de ce dossier. Et pour info, dans les gros LLM des IA du

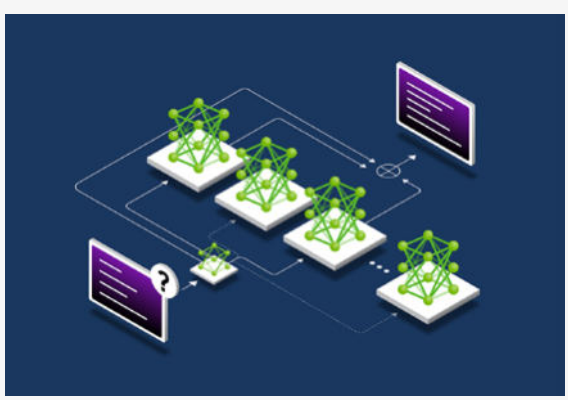


> Si la différence est invisible sur des questions simples, le volume de paramètres devient crucial dès qu'il s'agit de gérer des nuances, d'anticiper des contraintes ou de résoudre des contradictions massives.

cloud, on ne compte plus en milliards de paramètres mais en billions (trillion en anglais, soit 1000 milliards). Ce genre d'infos n'est pas publié, mais on estime que GPT-5 dépasse les 2T et que Grok-4 atteint même 3 T !

DE L'ÉTUDIANT À L'EXPERT

Pour bien visualiser cette montée en puissance, imaginons que vous demandiez à l'IA d'organiser un voyage complexe au Japon avec des contraintes de budget, de météo et de préférences culinaires : **Le modèle 8B (étudiant dégourdi)** : il vous donnera les grandes lignes (Tokyo, Kyoto, manger des sushis). C'est rapide, propre, mais il risque d'oublier que tel musée est fermé le lundi ou que votre budget ne permet pas le trajet en Shinkansen tous les jours. **Le modèle 70B (guide professionnel)** : il va structurer un itinéraire logique, vérifier les correspondances, suggérer des alternatives aux pièges à touristes et adapter le ton de ses conseils. Il "réfléchit" avant de répondre**.** **Le modèle 405B (expert local omniscient)** : on touche ici à la perfection. Il va non seulement créer l'itinéraire, mais aussi rédiger vos réservations en japonais parfait,



Les modèles MoE tourne plus vite que les autres, à nombre de paramètres identique.

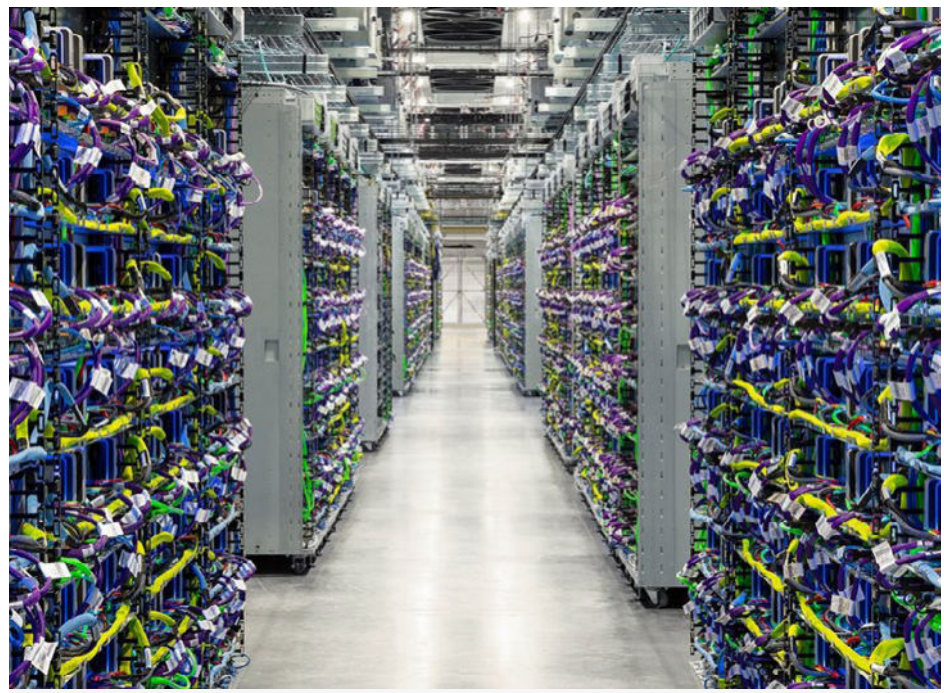
LA "MIXTURE OF EXPERTS" (MOE)

Vous allez croiser des modèles aux noms étranges, comme Mixtral 8x22B. Derrière ce jargon se cache une architecture révolutionnaire ; la Mixture of Experts. Plutôt que de construire un cerveau géant où chaque neurone travaille sur chaque mot (ce qui est très lent), on divise le modèle en plusieurs "spécialistes" (les Experts). Un modèle MoE peut avoir un très grand nombre de paramètres au total (141 milliards dans l'exemple de ce modèle de chez Mistral), mais pour chaque mot généré, un "routeur" n'active que les deux ou trois experts les plus qualifiés. L'intérêt principal est de réduire les ressources matérielles. Vous profitez de l'intelligence et des connaissances d'un modèle immense, mais avec la vitesse de calcul d'un modèle beaucoup plus petit. En revanche, si le calcul est rapide, le poids en mémoire reste celui du modèle complet. Pour un Mixtral 8x22B, il faut quand même pouvoir stocker les 141 milliards de paramètres en RAM.

anticiper les microclimats de chaque région et gérer des contradictions logiques massives sans sourciller.
En résumé ; plus il y a de paramètres, plus l'IA "comprend" la complexité du monde plutôt que de simplement "prédire" le mot suivant. Vous ne verrez pas de différence pour répondre à des questions simples, mais dès que vous voudrez des raisonnements complexes, ça change tout.

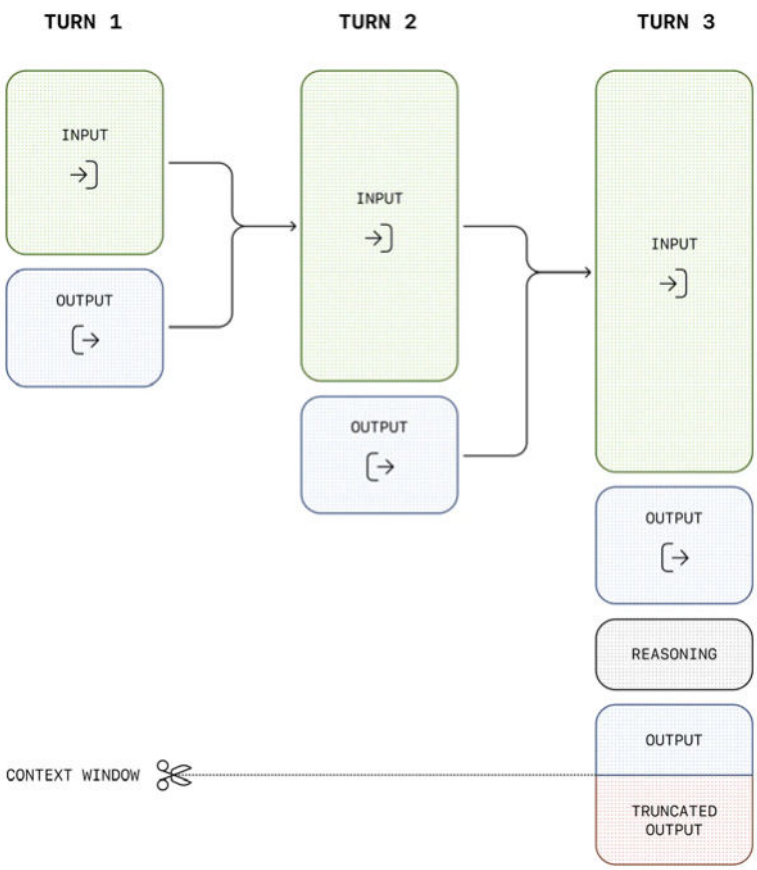
LA QUANTIFICATION À LA RESCOUSSE

La quantification (en anglais quantization) est l'étape technique indispensable qui permet de faire tourner des modèles massifs sur du matériel grand public. À l'état brut (souvent en FP16, soit 16 bits par paramètre), un modèle de 70 milliards de paramètres nécessiterait environ 140 Go de VRAM, ce qui est inaccessible hors serveurs professionnels. La quantification consiste à réduire la précision mathématique de ces paramètres (en passant de 16 bits à 8, 4 ou même 2 bits). Si l'on perd une infime fraction de "finesse" dans le raisonnement, on réduit drastiquement l'empreinte mémoire. Un modèle compressé en 4 bits (Q4_K_M) occupe quatre fois moins de place qu'à l'origine, permettant à un fleuron de l'IA de tenir sur une carte graphique de 16 ou 24 Go. Et on insiste sur la notion de compression, car la quantification n'est que la manière dont est compressé le modèle avant de le proposer au public, elle ne doit pas être confondue avec le mode de calcul en interne réalisé par un GPU ou un CPU au moment d'interagir avec le modèle (FP16, INT8, etc.). La quantification est un curseur stratégique ; il faut trouver l'équilibre entre une compression forte (pour la vitesse



IA DANS LE CLOUD : L'INFRASTRUCTURE DU GIGANTISME

Si l'IA locale se bat pour faire tenir 128 000 tokens dans 24 Go de VRAM, les modèles cloud comme Gemini 3 Pro (que votre serviteur utilise via son abonnement Google AI Plus) évoluent dans une autre dimension avec une fenêtre de 1 million de tokens ! De même, on ne parle pas de modèles de 8 ou 22 milliards de paramètres. L'IA Cloud a largement dépassé les 1 000 milliards comme Grok 3, il y a un an, qui sortait avec 1 800 milliards. On comprend pourquoi ça fonctionne si bien ! Une telle prouesse repose sur une gestion pyramidale de la mémoire à l'échelle de centres de données entiers ;
VRAM (HBM) ; le modèle cloud réside sur des clusters de puces spécialisées (très souvent des Nvidia H100, mais aussi des Google TPU v5p) embarquant chacune 80 à 95 Go de mémoire ultrarapide. Pour un contexte massif, la consommation se chiffre en téraoctets de VRAM distribués sur des dizaines de processeurs.
RAM ; pour optimiser les coûts, le KV Cache (la mémoire de votre discussion) peut être déplacé vers la RAM du serveur (jusqu'à 2 To par nœud) dès que vous marquez une pause, libérant les puces de calcul pour d'autres utilisateurs.
SSD NVMe ; les données sont "persistées" (conservées) sur des grappes de stockages flash pour les sessions de longue durée et permettre de retrouver vos historiques, même des mois plus tard et les recharger en VRAM dès que vous cliquez sur la discussion souhaitée.
Le Cloud ne gagne pas par l'élégance, mais par la force brute ; tout est réparti dynamiquement sur des milliers de serveurs interconnectés par fibre optique. Là où votre RTX 5090 est une Formule 1 isolée, l'infrastructure d'un Gemini est une usine automatisée de plusieurs hectares. On comprend aussi, sachant ça, pourquoi certains dénoncent l'empreinte énergétique du tout IA, comparé aux besoins nettement plus modestes d'une requête Google classique.



> Véritable mémoire à court terme de l'IA, la fenêtre de contexte définit la limite entre la compréhension globale d'un document et l'amnésie numérique, au prix d'une consommation accrue de ressources matérielles.

L'ALTERNATIVE NLP : QUAND LE PETIT BAS LE GÉANT

Si les LLM (Large Language Models) occupent tout l'espace médiatique, ils ne sont qu'une branche d'une discipline plus vaste ; le NLP (Natural Language Processing). Là où un LLM est un couteau suisse géant, le NLP classique regroupe des outils chirurgicaux spécialisés. Quelle différence concrète ? Un LLM (modèle généraliste) est "monstre" (milliards de paramètres) qui a tout appris pour savoir tout faire ; discuter, coder, traduire, inventer. Inconvénient ; il est lourd, lent et gourmand en VRAM. En revanc un NLP classique (modèle spécialisé) est plus compact et entraîné pour une seule tâche (par exemple en extraire uniquement les prix dans 10 000 factures). L'avantage, il est minuscule, c'est foudre de guerre en vitesse et tourne sur un simple processeur. Quand privilégier le NLP plutôt que le LLM ? Pour détecter des noms propres, des lieux ou des organisations dans un texte (tâche de Named Entity Recognition), un petit modèle NLP est souvent plus fiable et moins sujet aux hallucinations. Mais pour le traitement de masse, par exemple si vous devez classer 50 000 commentaires clients en "positif" ou "négatif", un modèle NLP fera le travail en quelques secondes là où un LLM mettra des minutes et chauffera votre GPU pour rien.

et l'accessibilité) et une compression légère (pour préserver la justesse du modèle). On peut d'ailleurs se demander à quel point jouer sur ce paramètre, car vaut-il mieux un cerveau plus gros (plus de paramètres, par exemple 70B au lieu de 8B) ou un cerveau qui ne se trompe jamais (par exemple Q8 comparé à Q2) ? Nous y reviendrons un peu plus loin durant nos multiples essais.

COMPRENDRE LA FENÊTRE DE CONTEXTE

La fenêtre de contexte représente la « mémoire à court terme » du modèle lors d'une session donnée. Elle définit la quantité totale de texte (mesurée en tokens) que l'IA peut garder à l'esprit en une seule fois, incluant vos instructions, ses réponses précédentes et les documents que vous lui soumettez. Si vous demandez à un LLM de résumer un livre, la taille de cette fenêtre est cruciale ; si elle est trop petite (par exemple 8 000 tokens), l'IA subira une forme d'amnésie et oubliera le début de l'ouvrage avant d'en avoir atteint la fin. Aujourd'hui, les modèles modernes

À RETENIR

- Fonctionnement probabiliste : un LLM ne récite pas, il prédit statistiquement le fragment de mot (token) suivant.
- Local vs Cloud : l'IA locale privilégie la confidentialité, la souveraineté des données et l'absence de coût à l'usage.
- Le token : unité de base du texte (environ 750 mots pour 1000 tokens) qui dicte la charge de travail de la mémoire vive.
- Paramètres (8B, 70B...) : représentent la "force de connexion" du cerveau de l'IA ; plus ils sont nombreux, plus le modèle est intelligent mais exigeant en matériel.
- Inférence : phase d'exécution où votre matériel (GPU/CPU) calcule la réponse en temps réel.
- Quantification : compression de la précision des paramètres pour faire tenir de gros modèles sur des cartes graphiques grand public.
- Fenêtre de contexte : capacité de "mémoire à court terme" du modèle lors d'une conversation, limitée par la quantité de VRAM disponible.

(ceux qu'on peut installer sur nos PC) affichent des fenêtres allant de 32 000 à plus de 128 000 tokens. Cependant, pour le matériel, le contexte n'est pas gratuit. Plus vous remplissez cette fenêtre, plus la consommation de VRAM augmente de manière significative pour stocker les données de calcul intermédiaires (le KV Cache). Une fenêtre de contexte immense sur une carte graphique limitée en mémoire peut ainsi saturer votre GPU et ralentir drastiquement la vitesse de génération. D'ailleurs, le logiciel Ollama qui est une des références pour faire tourner un LLM en local utilise un réglage par défaut de 4 096 tokens qu'il faut manuellement modifier si on veut plus et exploiter le maximum dont est capable le modèle.

DES MODÈLES QUI RAISONNENT

Sur les grandes plateformes cloud comme avec Gemini 3, vous avez peut-être l'habitude de basculer entre plusieurs modes ; "Rapide" (Flash) pour la vitesse, "Pro" pour la polyvalence, ou "Raisonnement" (Thinking) pour les tâches complexes. Dans les deux derniers modes, l'IA ne répond pas immédiatement ; elle prend un temps de pause pour décortiquer le problème, vérifier ses hypothèses et simuler une réflexion interne avant de formuler sa réponse finale. C'est ce qu'on appelle le chain of thought (chaîne de pensée). En local, sur votre PC, ce n'est généralement pas un simple bouton ou une option à cocher dans LM Studio ou Ollama (des logiciels dont nous allons reparler d'ici peu, qui permettent d'exécuter des modèles). La capacité de raisonnement est intrinsèque au fichier du modèle que vous téléchargez. Les modèles classiques (de type "instruct"), comme Llama 3, sont entraînés pour répondre directement. Ils sont rapides, parfaits pour du chat ou de la rédaction, mais peuvent "halluciner" sur des énigmes logiques. Les modèles de raisonnement existent aussi. Inspirés par des architectures comme o1 ou DeepSeek-R1, ces modèles sont spécifiquement entraînés pour générer des balises de pensée (souvent affichées entre <think> et </think>) avant de donner la réponse. Concrètement, si vous voulez ce comportement de raisonnement en local, vous ne changez pas de mode, vous changez de modèle. Vous devrez télécharger une version spécifique (souvent suffixée par reasoning ou des modèles comme Qwen-Math). Attention ; cette réflexion consomme des ressources ! L'IA va générer des milliers de tokens invisibles (sa pensée) avant de vous répondre, ce qui rend l'affichage de la réponse finale beaucoup plus lent et monopolise votre carte graphique plus longtemps.

L'ÉTABLI LOGICIEL : INSTALLER ET CONFIGURER SON PREMIER MODÈLE

À **présent que les bases sont posées, entrons dans la pratique**. Nous allons voir comment s'y prendre pour très facilement installer et essayer des modèles de langage sur votre PC Windows du quotidien.

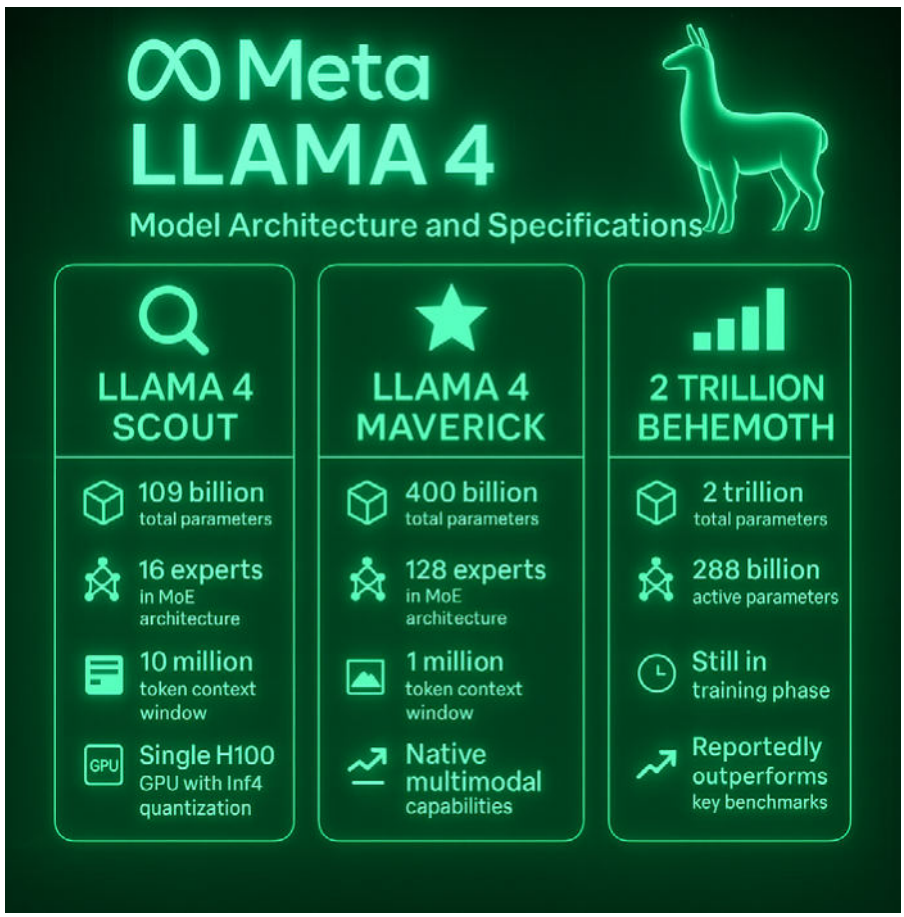
FACE AU CLOUD, LES MODÈLES "OPEN WEIGHTS"

Première chose à faire : choisir un modèle. Or il y en a des dizaines (pour ne pas dire des centaines), chacun étant lui-même décliné en de multiples versions. Nous avons vu un peu plus tôt qu'on distingue deux grandes familles de modèles : d'un côté, les modèles propriétaires ou fermés (closed weights), tels que GPT-4 d'OpenAI ou Claude d'Anthropic. Ces modèles sont accessibles uniquement via une interface web ou une API ; vous ne pouvez pas télécharger leurs fichiers, et leur fonctionnement interne reste le secret industriel de l'entreprise qui le développe et le promeut. À l'opposé, les modèles dits "open weights" (poids ouverts) mettent à disposition du public l'intégralité de leurs fichiers de paramètres. Bien que souvent qualifiés d'Open Source par abus de langage, la nuance est importante ; si le code de l'architecture est libre, les données d'entraînement exactes ne sont pas toujours divulguées. Cependant, pour l'utilisateur final, l'avantage est colossal ; le modèle est téléchargeable et peut s'exécuter à 100 % localement, sans dépendre d'un serveur tiers ni payer d'abonnement à l'usage. Parmi les figures de proue de cette catégorie, on trouve la famille Llama de Meta (Facebook), qui a véritablement lancé la révolution de l'IA locale, mais aussi les modèles Mistral de la pépite française Mistral AI, réputés pour leur efficacité, ou encore la gamme Phi de Microsoft, optimisée pour offrir de hautes performances sur des configurations matérielles modestes. Utiliser un modèle open weights, c'est garantir la confidentialité de ses données et la pérennité de ses outils.

OpenAI propose depuis peu un modèle utilisable à domicile, baptisé GPT-OSS.

À LA DÉCOUVERTE DES PRINCIPAUX MODÈLES

Face à l'explosion du nombre de modèles disponibles, le choix peut sembler vertigineux. Pour ce dossier, nous avons donc pris le parti de concentrer nos tests sur les trois piliers que nous venons d'évoquer ; Llama, Mistral et Phi. Ce choix n'est pas arbitraire ; il représente les trois forces



motrices de l'IA locale actuelle. Llama (Meta) est le standard industriel, celui qui bénéficie du plus large écosystème de compatibilité. Mistral (Mistral AI) incarne l'excellence européenne avec un rapport performance/taille souvent imbattable, particulièrement pertinent pour le français. Enfin, Phi (Microsoft) prouve qu'avec un entraînement de haute qualité, un modèle minuscule peut rivaliser de logique avec des géants, comprendre par là qu'il peut donner des réponses aussi bonnes qu'un modèle plus lourd (avec plus de paramètres). Mais ils ne sont pas seuls et on trouve de nombreux candidats intéressants, par exemple le chinois DeepSeek. Voici une brève analyse des principaux modèles disponibles (sans entrer dans leurs multiples sous-versions) :

Llama (Meta) : le standard incontournable de l'Open Source
o *Avantages ;* c'est la famille de modèles la plus populaire au monde, ce qui garantit une compatibilité parfaite avec tous les logiciels et optimisations existantes (comme Ollama ou LM Studio). Les versions 3.1 et 3.2 sont extrêmement polyvalentes, excellentes en rédaction, en synthèse et en conversation naturelle. C'est le choix de la sécurité ; il fait tout très bien.
o *Inconvénients ;* bien que très performant, il peut s'avérer moins pointu que DeepSeek sur le code pur ou moins optimisé que Command pour le RAG professionnel. Sa version la plus lourde (405B) est totalement hors de portée des stations de travail classiques, frustrant ceux qui voudraient goûter au sommet de la gamme.

Mistral & Mixtral (Mistral AI) : la pépite française qui défie les géants
o *Avantages ;* réputés pour leur efficacité redoutable, les modèles de Mistral AI offrent souvent un rapport performance/taille imbattable. L'architecture "Mixtral" (Mixture of Experts) permet de simuler un très grand cerveau tout en n'activant qu'une fraction de ses neurones à chaque mot, ce qui booste la vitesse d'inférence. Ils sont particulièrement bons pour le raisonnement logique et le respect des consignes complexes.
o *Inconvénients ;* les versions les plus massives (comme le 8x22B) ont une gestion du contexte très gourmande, ce qui peut saturer la RAM même sur des configurations haut de gamme. De plus, leur style rédactionnel est parfois jugé plus "sec" ou plus direct que celui des modèles Llama, privilégiant l'efficacité à la rondeur de la conversation.

Phi (Microsoft) : le prodige des "petits modèles de langage" (SLM)
o *Avantages ;* une prouesse d'ingénierie où la qualité des données d'entraînement prime sur la quantité. Les modèles Phi-3 et Phi-4 sont incroyablement compacts, ce qui les rend d'une rapidité fulgurante même sur du matériel modeste. Ils sont capables de prouesses logiques et mathématiques qui rivalisent avec des modèles dix fois plus gros. C'est le candidat idéal pour une IA ultra-réactive et intégrée.
o *Inconvénients ;* en raison de sa petite taille, il possède une "culture générale" plus limitée que les mastodontes comme Llama ou Command. Il peut manquer de nuances dans les rédactions créatives ou les débats philosophiques complexes, et

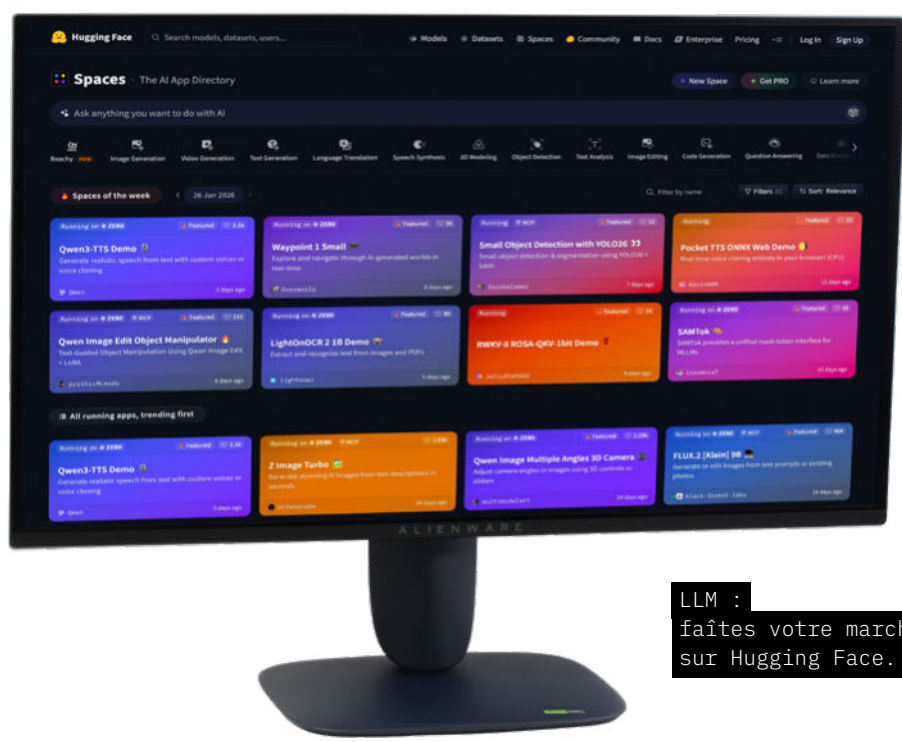
sa fenêtre de contexte est souvent plus réduite, limitant l'analyse de documents très volumineux.

Qwen (Alibaba) : le génie du code et des maths.
o *Avantages ;* une famille de modèles qui domine actuellement les classements Open Source (Open LLM Leaderboard). Qwen est d'une polyvalence rare ; il est aussi à l'aise pour coder et résoudre des problèmes mathématiques complexes que pour discuter de manière fluide. Sa version 2.5 est particulièrement optimisée pour la vitesse, offrant un débit de tokens impressionnant même sur des architectures imposantes.
o *Inconvénients ;* comme pour DeepSeek, son origine (Chine) peut freiner certains usages en entreprise pour des raisons de souveraineté. De plus, il a tendance à être très "verbeux" (il parle beaucoup), ce qui peut consommer inutilement de la fenêtre de contexte si on ne bride pas ses réponses via le prompt.
Gemma (Google) : l'héritier de la technologie Gemini.
o *Avantages ;* bénéficiant des mêmes avancées architecturales que le célèbre Gemini, Gemma se distingue par une compréhension sémantique très fine et une grande rigueur dans le suivi des instructions. C'est un modèle particulièrement "propre" et sûr, idéal pour les tâches de rédaction soignée, de traduction ou de synthèse où la nuance est primordiale. Sa version 2 (9B et 27B) offre un rapport performance/taille parmi les meilleurs du marché.
o *Inconvénients ;* en raison des filtres de sécurité intégrés par Google, Gemma peut parfois se montrer excessivement prudente (refusant de répondre à certaines requêtes pourtant anodines). Son architecture est également un peu plus spécifique à charger pour certains outils, ce qui peut demander des mises à jour régulières de vos logiciels (Ollama/LM Studio) pour garantir des performances optimales.

DeepSeek (DeepSeek-AI) : le nouveau champion de l'efficience.
o *Avantages ;* une architecture innovante (MoE - Mixture of Experts) qui permet d'obtenir des performances de niveau GPT-4 avec une consommation de ressources optimisée. Ils sont particulièrement réputés pour leurs capacités en Mathématiques et en Programmation (Coding), dépassant souvent Llama sur ces terrains.
o *Inconvénients ;* des modèles parfois plus complexes à configurer pour en tirer le maximum de vitesse. Leur origine peut également soulever des questions de confidentialité pour certaines entreprises occidentales, bien que leurs poids soient librement accessibles.

Command (Cohere) : l'expert de la recherche et du multilingue.
o *Avantages ;* une famille de modèles (Command R, R+ et Command A) pensée dès le départ pour la productivité et l'automatisation. Ils excellent dans le "RAG" (analyse de documents) et se distinguent par leurs capacités "agentiques" (capacité à utiliser des outils logiciels). La gestion native de nombreuses langues est l'une des meilleures du marché.
o *Inconvénients ;* des modèles globalement plus gourmands en mémoire que leurs équivalents chez Mistral ou Meta. La version la plus performante (104B) impose une barrière matérielle haute, rendant l'usage de stations de travail musclées quasiment obligatoire pour ne pas sacrifier la vitesse.

GPT-OSS (OpenAI) : le petit dernier… et non des moindres !
o *Avantages ;* un modèle d'une efficacité redoutable qui condense le savoir-faire d'OpenAI dans un format "poids ouverts". Il est extrêmement rapide, très peu gourmand par rapport



LLM : faîtes votre marché sur Hugging Face.

à son intelligence, et bénéficie d'une optimisation parfaite pour les architectures GPU les plus récentes. C'est l'outil idéal pour ceux qui veulent la "logique GPT" avec une confidentialité totale. Pour ceux qui vivent dans une grotte : OpenAI est l'entreprise derrière ChatGPT !

o *Inconvénients ;* avec 20B de paramètres, il reste un cran en dessous des modèles géants (100B+) pour les tâches de création littéraire très longues ou les synthèses de documents de 500 pages. Sa fenêtre de contexte est plus courte que celle d'un Command R+.

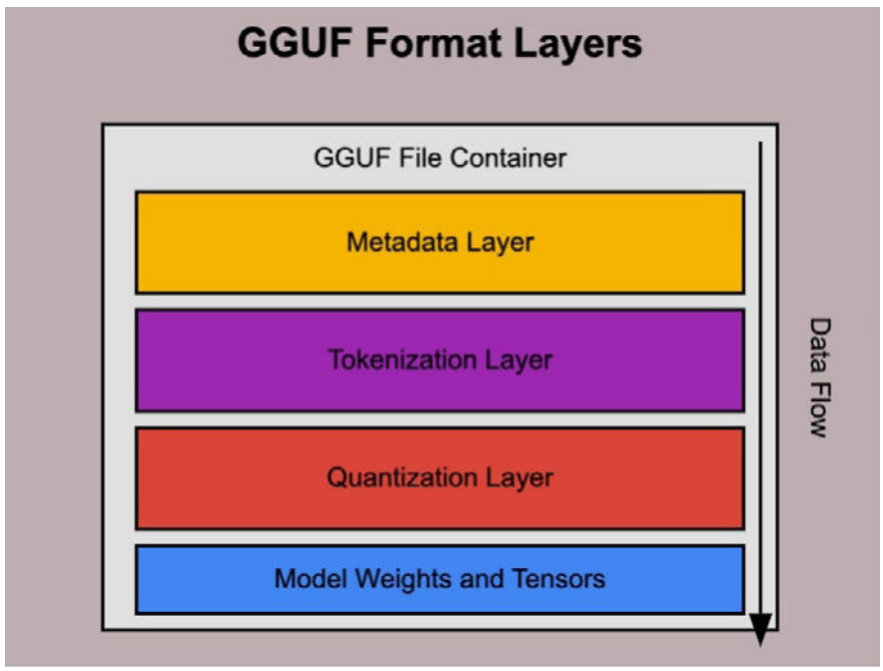
LE "GITHUB DE L'IA" : HUGGING FACE

Si vous vous lancez dans l'IA locale, Hugging Face est votre passage obligé. Véritable plaque tournante de l'écosystème open source, cette plateforme est à l'intelligence artificielle ce que GitHub est au développement logiciel ; un immense répertoire collaboratif où chercheurs, entreprises (Meta, Microsoft, Mistral AI) et passionnés partagent leurs travaux. Pour l'utilisateur de LLM sur PC, Hugging Face est avant tout le coffre-fort où sont stockés les fameux poids des modèles. C'est ici que vous irez chercher les fichiers nécessaires à votre assistant local. Chaque modèle possède sa propre "Model Card", une fiche d'identité technique détaillant son architecture, ses données d'entraînement, ses performances sur les benchmarks et, point crucial, sa licence d'utilisation. Au-delà du simple stockage, la plateforme héberge des milliers de "Spaces" (des démos interactives), des jeux de données (datasets) pour le fine-tuning, et surtout des versions optimisées des modèles de référence. C'est grâce à la communauté Hugging Face que les modèles massifs sont convertis en formats exploitables par nos machines grand public, comme le GGUF ou l'EXL2, rendant l'IA de pointe accessible à n'importe quel possesseur d'une carte graphique moderne. Vous ne téléchargez pas les modèles sur le site de leur concepteur, mais via Hugging Face.

LM STUDIO, L'INTERFACE DE BORD POUR TESTER DES LLM

Pour qui ne souhaite pas s'aventurer dans les méandres des lignes de commande, LM Studio (Windows, Linux, macOS) s'est imposé comme l'outil de référence. C'est le cockpit de l'IA locale ; une interface graphique léchée qui transforme l'installation d'un modèle complexe en un simple clic. Mais ne vous y trompez pas, sous ses airs de logiciel grand public, c'est un véritable instrument de précision pour le fan de hardware. L'une de ses forces réside dans son intégration native avec Hugging Face. Depuis la barre de recherche, vous accédez directement aux versions quantifiées des modèles, y compris les plus récents, comme Llama 4 Scout ou GPT-OSS. L'application analyse votre matériel en temps réel et vous indique, via un code couleur, si le modèle choisi logera intégralement dans votre VRAM ou s'il devra déborder sur la RAM système — un indicateur précieux pour éviter les effondrements de performance et donc nous allons longuement reparler un peu plus loin. Pour les possesseurs de configurations musclées, LM Studio propose désormais des contrôles granulaires pour le multi-GPU. Il est possible de choisir précisément la répartition des couches du modèle entre deux cartes (par exemple une RTX 5090 et une 4090 de génération précédente) ou de forcer l'utilisation exclusive de la mémoire dédiée pour garantir un débit de tokens maximal. Une fois le modèle chargé, l'onglet *Server* permet de transformer votre PC en un point d'accès local compatible avec l'API d'OpenAI, permettant

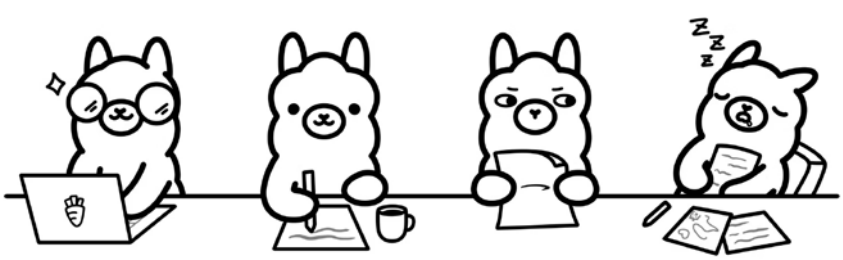
> De la fiche d'identité technique au téléchargement des fichiers, Hugging Face est le passage obligé de tout utilisateur d'IA locale.



de brancher n'importe quel logiciel tiers sur votre propre IA privée. C'est l'outil parfait pour prototyper, benchmarker et valider la stabilité de votre build avant de passer à des solutions plus automatisées pour l'exécution quotidienne.

LE FORMAT GGUF À PRIVILÉGIER

Si vous téléchargez un modèle pour le faire tourner sur votre PC, vous tomberez inévitablement sur des fichiers portant l'extension. gguf. Successeur du format GGML et porté par la communauté autour du projet llama.cpp, le GGUF (GPT-Generated Unified Format) est devenu le standard incontournable de l'IA locale. Son génie réside dans deux points clés. D'abord, c'est un format « tout-en-un » ; un seul fichier contient à la fois les poids du modèle et toutes les métadonnées nécessaires à son exécution (architecture, tokenizer, paramètres de contexte). Ensuite, il est conçu pour la flexibilité matérielle. Contrairement à d'autres formats purement pensés pour le GPU, le GGUF permet de répartir les couches du modèle entre la VRAM de votre carte graphique et la RAM de votre système. C'est cette technologie qui permet de faire tourner un modèle trop gros pour votre GPU en acceptant simplement une baisse de vitesse, là où d'autres formats auraient purement et simplement refusé de se lancer. Pour l'utilisateur, c'est la garantie d'une compatibilité universelle, quel que soit l'équilibre de sa configuration.



Malgré le nom ressemblant, Ollama est un logiciel open source qui n'a rien à voir avec les modèles Llama de Meta.

OLLAMA, LE MOTEUR SOUS LE CAPOT POUR UTILISER LES LLM AU QUOTIDIEN

Si LM Studio est le plus pratique pour exécuter un modèle et en modifier quelques paramètres à l'aide d'une belle interface graphique, Ollama (Windows, Linux, macOS) est plus efficace pour lancer un LLM en production et le laisser tourner au quotidien. Contrairement à une application classique que l'on ouvre pour une tâche précise, Ollama s'exécute comme un service (daemon) en arrière-plan. Il démarre avec votre système d'exploitation et reste tapi dans l'ombre, prêt à mobiliser votre GPU dès qu'une requête lui parvient, que ce soit via le terminal ou une application tierce. Cette architecture « headless » (sans interface, quoiqu'il existe désormais une mini interface) est le secret de sa réactivité. Pour déployer un modèle comme Llama 4 Scout ou Mistral Nemo, il suffit d'une ligne de commande ; ollama run llama4. Ollama gère alors l'intégralité du pipeline ; téléchargement automatique des poids quantifiés depuis sa propre bibliothèque, allocation intelligente des couches du modèle sur la VRAM et exposition d'une API locale. C'est cette approche de « gestionnaire de paquets » pour l'IA qui en a fait le standard de fait pour l'usage quotidien. Mais la véritable force d'Ollama réside dans sa fonction Modelfile. Ce petit fichier de configuration permet de créer des instances personnalisées en quelques secondes. Vous pouvez y graver un System Prompt permanent, ajuster la température ou modifier

les paramètres de contexte. Par exemple, vous pouvez configurer un Phi-4 spécifiquement pour la correction de code technique et l'appeler d'un simple *ollama run mon-expert-code*. Enfin, son API locale permet de connecter l'IA à tout votre écosystème ; extensions d'éditeurs de code (IDE), plugins de prise de notes ou interfaces web avancées comme Open WebUI. C'est l'outil de choix pour celui qui veut une IA intégrée à son flux de travail, et non une simple fenêtre de chat isolée. Notons que nous allons nous concentrer sur LM Studio pour cette première approche et essayer des LLM, Ollama étant comme vous l'aurez compris l'outil à choisir pour mettre un modèle de langage en production une fois que vous avez arrêté vos choix.

INSTALLEZ VOTRE PREMIER LLM PAS À PAS

Pour ce premier contact, nous allons privilégier la méthode la plus visuelle en utilisant LM Studio, qui simplifie drastiquement la gestion des modèles. Voici la marche à suivre pour transformer votre PC en station d'IA en moins de cinq minutes. Tout commence par le téléchargement du logiciel. Rendez-vous sur le site officiel de LM Studio, qui est en version 0.4.0 il y a quelques jours à peine, et récupérez l'installateur pour Windows. L'application est un exécutable classique ; installez-la et lancez-la. Au premier démarrage, zappez le téléchargement du premier modèle d'exemple et cochez la case pour activer les fonctions développeur (vous pourrez aussi le faire après coup dans les paramètres, roue dentée tout en bas à gauche de la fenêtre principale).

LM STUDIO ET L'ACCÉLÉRATION MATÉRIELLE

Une fois le logiciel chargé, cliquez sur la roue dentée tout en bas à gauche (et non plus à droite, comme sur la 0.3.39 qui vient d'être remplacée), rendez-vous dans l'onglet *Hardware* (invisible si vous n'avez pas coché les options développeur justement) pour vous assurer que la carte graphique est bel et bien détectée et activée. Ainsi que la case *Offload KV Cache to GPU Memory*. Dans la foulée, dans l'onglet *Runtime* vous pouvez mettre à jour et le " moteur de fonctionnement" de LM Studio. C'est à adapter à votre ordinateur. Là où un laptop sans GPU choisira probablement CPU, un possesseur de GeForce choisira en

OLLAMA, LES COMMANDES ESSENTIELLES

La suite de ce tutoriel se fait sur LM Studio. Si toutefois vous souhaitez exécuter Ollama, voici les principales commandes à connaître :

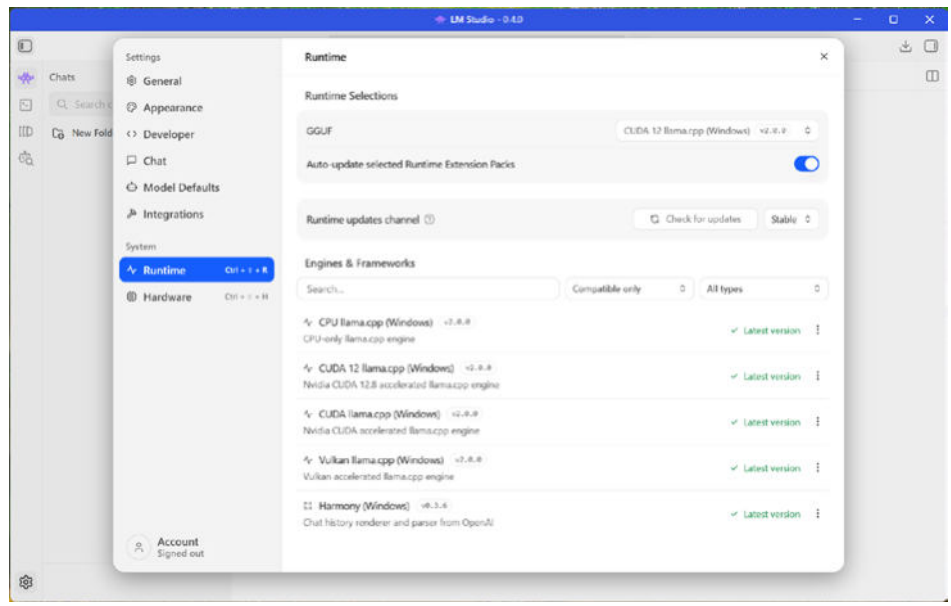
Commandes système (depuis PowerShell sous Windows ou Terminal sous Linux) :

Commande	Action	Notes
ollama run <modèle>	Lance un modèle	Charge un modèle et ouvre immédiatement une session de chat (téléchargement automatique s'il n'est pas installé).
ollama pull <modèle>	Télécharge un modèle	Prépare le modèle sans le lancer (idéal pour charger le SSD avant un bench).
ollama list	Liste les modèles	Vérifie ce qui est installé et la taille réelle sur le disque.
ollama ps	Modèles actifs	Affiche quel modèle est actuellement chargé en (V)RAM.
ollama stop	Décharger de la mémoire	Libère la VRAM/RAM sans supprimer le modèle du SSD.
ollama rm <modèle>	Supprime un modèle	Nettoie le SSD après une batterie de tests.
ollama show <modèle>	Infos détaillées	Affiche l'architecture, la licence et les paramètres par défaut du modèle.
ollama serve	Démarre le serveur	Utile si le service automatique ne s'est pas lancé (ou si tu l'as tué).

Commandes de session (à l'intérieur même du chat Ollama) :

Commande	Action	Notes
/set verbose	Mode Stats	INDISPENSABLE. Affiche le débit de tokens/s après chaque réponse.
/set parameter <clé> <val>	Réglage à la volée	Permets de changer la température ou le num_ctx sans relancer le modèle.
/show info	Détails techniques	Affiche le "système de prompt" et les paramètres actuels du modèle.
/show system	Voir le Prompt Système	Affiche les instructions "cachées" (ex: "Tu es une IA utile...").
/bye	Quitter	Ferme la session de chat et retourne au terminal Windows.
/?	Aide	Liste toutes les commandes de session disponibles.

priorité CUDA 12 tandis qu'un propriétaire de Radeon privilégiera Vulkan, qui contre toute attente, se révèle plus rapide que ROCm (le GPGPU d'AMD, équivalent à CUDA pour Nvidia) qui est encore assez jeune. Vous pouvez d'ailleurs vous amuser à passer de l'un à l'autre (seuls ceux qui sont compatibles avec votre matériel sont sélectionnables) pour comparer les performances.

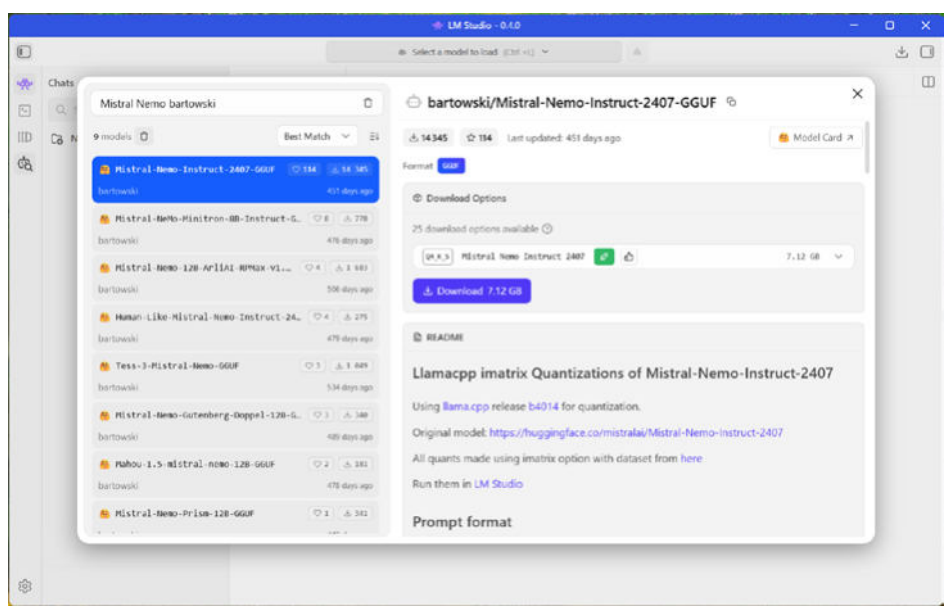


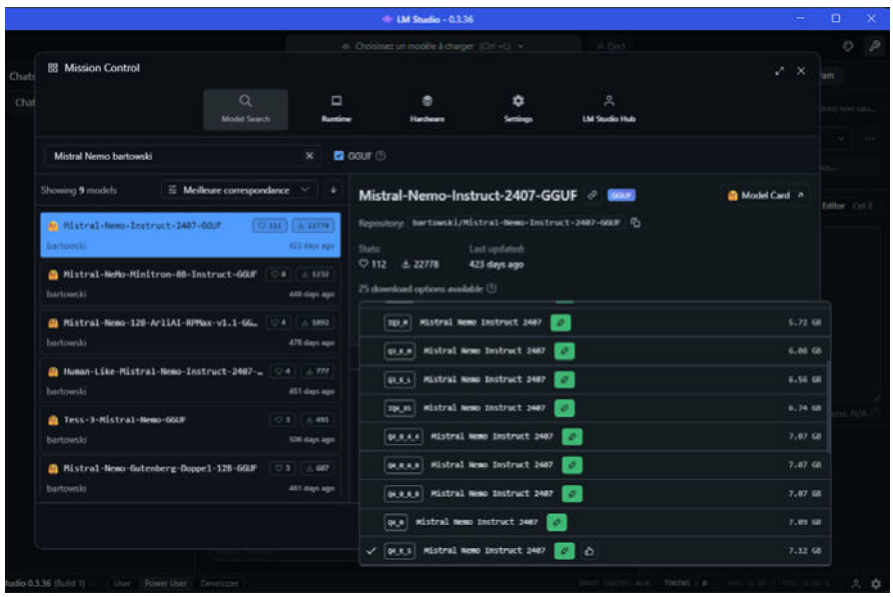
LA QUÊTE DU MODÈLE

Cliquez sur l'icône de la loupe dans la barre latérale gauche. Dans le champ de recherche, tapez *Llama 3 8b* ou *Mistral Nemo*. LM Studio interroge alors en temps réel les serveurs de Hugging Face et vous affiche une liste de résultats. Par exemple, sur nos captures, vous pouvez voir que nous sommes partis sur Mistal Nemo. Vous constaterez, en dessous du nom de chaque modèle, qu'il y a le nom de l'auteur. Ce n'est pas l'auteur du modèle en lui-même, mais celui qui a réalisé le port du modèle dans un format GGUF et créé différents niveaux de quantification pour l'adapter aux capacités hardware de chacun. Bartowski est un utilisateur réputé pour fournir des fichiers de qualité, c'est donc sur l'un de ses ports de Mistal Nemo que nous nous sommes orientés pour ce premier contact.

QUELLE QUANTIFICATION CHOISIR ?

C'est l'étape cruciale pour adapter le modèle à votre matériel. Une fois le modèle sélectionné dans la moitié gauche de la fenêtre popup, une liste de fichiers apparaît sur la droite (souvent marqués Q4_K_M, Q8, etc.). Regardez la colonne Compatibility ; LM Studio affiche une pastille verte si le modèle tient entièrement dans votre VRAM. Privilégiez si vous pouvez Q4_K_M ou Q4_K _S qui sont de bons compromis poids/intelligence (plus bas, Q3 et surtout Q2), l'IA hallucine beaucoup. À l'inverse, si Q8 apporte plus de précision et donc plus "d'intelligence", ça reste peu remarquable et au-dessus vous ne ferez même pas la différence. Il y a donc deux façons de voir les choses : soit vous adaptez la quantification du modèle souhaité à votre GPU, soit vous partez sur une quantification donnée (telle que Q4_K_M) et vous pourrez charger des modèles plus ou moins gros (avec plus ou moins de paramètres) dans la VRAM de votre carte graphique. Pour l'illustration, nous avons choisi le Mistral Nemo Q4_K_S de Bartowski qui occupe 7,12 Go et tient donc dans





n'importe quel GPU accompagné d'au moins 8 Go de mémoire. Cliquez, en bas à droite, sur *Download*.

CHARGEMENT ET PREMIER PROMPT

Une fois le téléchargement terminé (quelques Go tout de même), rendez-vous sur l'icône AI Chat (la bulle de discussion). En haut de la fenêtre, cliquez sur Select a model to load et choisissez votre fichier. Observez la jauge circulaire en bas à droite, c'est l'occupation de la fenêtre de contexte. Il est temps d'essayer ! Dans le panneau de droite, vérifiez que le paramètre GPU Offload est au maximum pour garantir que c'est bien votre carte graphique qui travaille, et non votre processeur. Tapez votre question dans la barre de chat. Si vous voyez le texte s'afficher instantanément, félicitations ; vous venez de réaliser votre première inférence locale. Si la génération est trop lente (moins de 5 tokens/seconde), c'est que le modèle est trop lourd pour votre VRAM et déborde sur votre RAM système. Essayez alors une version plus compressée (un chiffre de quantification plus bas, comme Q3 ou Q2) ou un modèle plus petit comme Phi-4, mais nous allons parler de tout ça plus longuement d'ici peu, avec des nombreux benchs à la clé.

QUEL MODÈLE POUR QUEL USAGE ?

Maintenant que votre moteur est prêt, il reste la question cruciale ; quelle « personnalité » allez-vous télécharger ? Si la technique impose des limites de taille (la VRAM), le besoin métier dicte la spécialisation. Pour vous aider à choisir, reprenons nos trois profils types :

Le Créatif (priorité à la nuance et au style) : votre but est de rédiger des articles, des scripts ou de brainstormer. Vous avez besoin d'un modèle qui possède de la « rondeur » et une grande fluidité narrative. Le candidat idéal ; Llama 3.1 (Meta) ou Gemma 2 (Google). Ils sont réputés pour leur capacité à ne pas être trop « robotiques » et à respecter un ton spécifique.

L'Analyste (priorité à la logique et à la rigueur) : vous traitez des données, vous voulez des résumés factuels ou de l'aide sur des fichiers Excel complexes. Ici, on cherche la précision avant le style. Le candidat idéal ; Mistral ou Qwen. Ces modèles sont souvent plus « secs », mais extrêmement performants pour structurer l'information et éviter les envolées lyriques inutiles.

L'Expert (priorité au code et à la technique) : votre IA est votre copilote de développement ou votre assistant en mathématiques. Vous avez besoin d'un modèle entraîné sur des langages de programmation. Le candidat idéal ; DeepSeek ou la famille CodeLlama. Leurs performances en Python, C++ ou SQL surpassent souvent des modèles généralistes pourtant plus volumineux.

TESTEZ VOTRE FUTUR CERVEAU

Au-delà de nos trois profils types, comment savoir si tel ou tel nouveau modèle est fait pour vous ? Ne vous fiez pas uniquement aux benchmarks officiels (souvent flatteurs). Adoptez une approche de testeur. Premièrement, posez-lui une question piège pour tester sa qualité de raisonnement. Ne demandez pas ; « Qui est Napoléon ? » (trop facile, c'est du savoir encyclopédique). Posez plutôt une question de logique pure ou de raisonnement

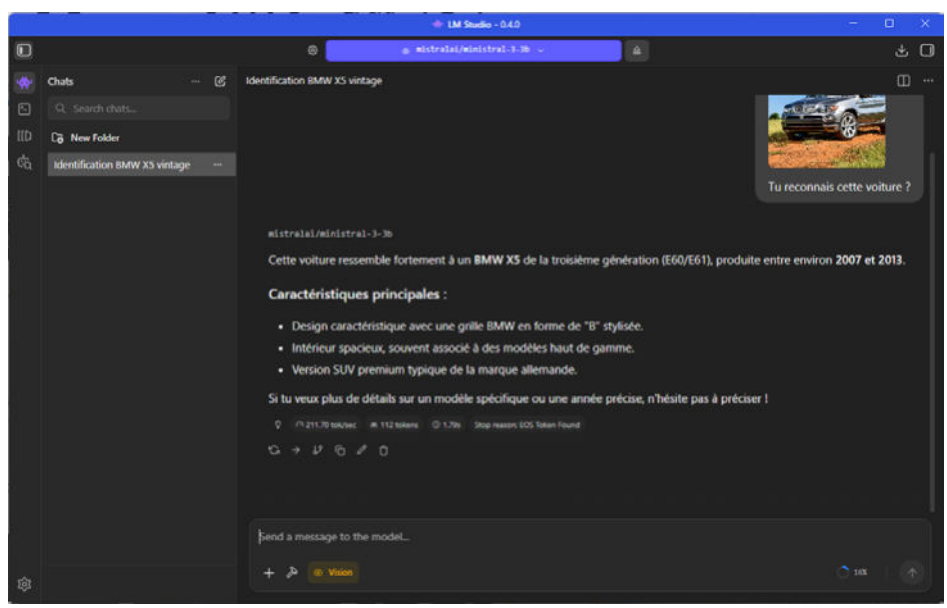
> Passer de locataire à propriétaire de son intelligence artificielle pour garantir une confidentialité absolue.

déductif ; « Si j'ai trois chemises et qu'il faut 3 heures pour les faire sécher au soleil, combien de temps faut-il pour en faire sécher six ? » Si le modèle répond 6 heures, sa logique est faible. S'il répond 3 heures, il "réfléchit". Deuxièmement, le stress-test du contexte pour juger de sa capacité d'analyse. Copiez-collez un long article technique et demandez-lui d'en extraire trois points clés qui contredisent une idée reçue. Si le modèle commence à inventer des faits (hallucination) ou s'il perd le fil après 2000 mots, c'est qu'il n'est pas adapté à vos besoins. En trois, faites un bon vieux comparatif en aveugle (méthode LMSYS locale). Ouvrez deux instances (par exemple LM Studio et Ollama) avec deux modèles différents. Posez la même question complexe et comparez. La pertinence ; la réponse est-elle juste ? La structure ; est-ce bien mis en page (listes à puces, gras) ? La vitesse (tokens/s) ; le confort de lecture est-il acceptable ? Enfin, en quatrième option, vous pouvez demander au modèle ce qu'il pense de lui-même et de ses capacités par rapport à vos besoins. Par exemple « Je suis un développeur Python travaillant sur des bases de données SQL. Quelles sont tes spécificités d'entraînement qui te rendent plus ou moins efficace que Llama 3 pour m'aider à coder ? » Un bon modèle (comme Qwen ou DeepSeek) saura vous citer ses forces et ses faiblesses techniques.

Le meilleur modèle est celui qui offre le meilleur compromis entre votre patience (vitesse de génération) et votre exigence (justesse de la réponse).

ON RÈGLE LA TEMPÉRATURE

Une fois votre modèle chargé dans LM Studio, vous remarquerez dans les réglages (dans la barre latérale de droite après avoir cliqué sur la clé à molette, puis dans l'onglet Model) un curseur nommé *Temperature*. Si le terme évoque la physique, il désigne ici un paramètre purement mathématique qui agit sur la créativité du modèle. Pour comprendre son impact, il faut se rappeler qu'une IA ne choisit pas le mot suivant par certitude, mais par probabilité. Si vous réglez la température au plus bas (entre 0.1 et 0.3), l'IA devient rigoureuse et factuelle ; elle choisit systématiquement le mot ayant la probabilité la plus élevée. C'est le réglage idéal pour du code informatique, de l'analyse de données ou tout travail de rédaction



En optant pour un modèle Vision, vous pouvez joindre une image à vos prompt. Bravo Mistral d'avoir identifié un X5, en revanche c'est un E53 (1999 à 2006).

technique où la précision prime sur le style. Les réponses sont stables et reproductibles, mais peuvent paraître froides ou répétitives. D'ailleurs, en mettant 0, vous obtiendrez toujours la même réponse (sans changer de modèle, bien entendu). À l'inverse, en poussant le curseur au-delà de 0.8, vous autorisez l'IA à piocher dans des mots moins probables, ce qui fluidifie son style et lui permet de faire des liens plus originaux. C'est parfait pour du brainstorming ou de l'écriture créative. Attention toutefois ; au-delà de 1.2, l'IA devient "fiévreuse" ; elle perd le fil de sa propre logique et commence à inventer des faits ou à dériver vers des réponses incohérentes. Attention aux réglages par défaut ; ne vous y trompez pas, l'absence de réglage manuel ne signifie pas l'absence de température. Chaque logiciel applique sa propre "recette" par défaut (0.7 sur Ollama et 0.8 sur LM Studio). Dans LM Studio, les réglages de sampling placés juste sous la température permettent un ajustement plus précis de celle-ci, mais n'y touchez pas pour le moment. Pour l'anecdote, la température des gros modèles commerciaux, comme Gemini de Google, est dynamique. Le modèle adapte automatiquement la température au type de question posée par l'utilisateur !

PIÈCES JOINTES

Sur des services comme ChatGPT ou Gemini, on a pris l'habitude de glisser n'importe quel fichier (photo, PDF, document Word) dans la conversation pour que l'IA l'analyse. En local, cette "magie" demande un peu plus de doigté car elle dépend de deux mécaniques distinctes ; la vision (pour les images) et l'ingestion de texte (pour les documents). Pour les images, cela ne fonctionne pas avec n'importe quel LLM. Il faut impérativement charger un modèle dit "Multimodal" ou "Vision" (comme LLaVA, BakLLaVA, Qwen-VL ou les versions "Vision" de Llama 3.2). Sans cela, l'IA est aveugle. Pour les documents (Word, PDF), l'IA ne "lit" pas le fichier binaire ; votre logiciel d'interface doit d'abord extraire tout le texte du document, le nettoyer, et l'injecter dans la fenêtre de contexte (le prompt). Attention donc ; un gros PDF peut saturer instantanément votre mémoire à court terme (voir partie précédente sur la fenêtre de contexte).

LM Studio : l'outil a fait d'énormes progrès. Vous trouverez une icône "trombone" ou "Import" à côté de la barre de saisie. Si vous chargez une image, assurez-vous d'avoir sélectionné un modèle compatible (souvent tagué "Vision" dans la recherche). Pour les fichiers texte (PDF, .docx), LM Studio extrait automatiquement le contenu et vous indique combien de tokens cela va consommer. C'est du "clic et marche". **Ollama** : c'est bien plus rustique. En ligne de commande pure, Ollama gère très bien les images si vous lancez un modèle vision (ex: ollama run llava), il suffit de coller le chemin du fichier image. En revanche, pour les documents Word ou PDF, Ollama ne sait pas les ouvrir nativement. Vous devrez soit copier-coller le texte brut dans le terminal, soit passer par une interface graphique tierce (comme Open WebUI ou Page Assist) qui se connecte à Ollama et ajoute cette fonctionnalité de gestion de fichiers.

À RETENIR

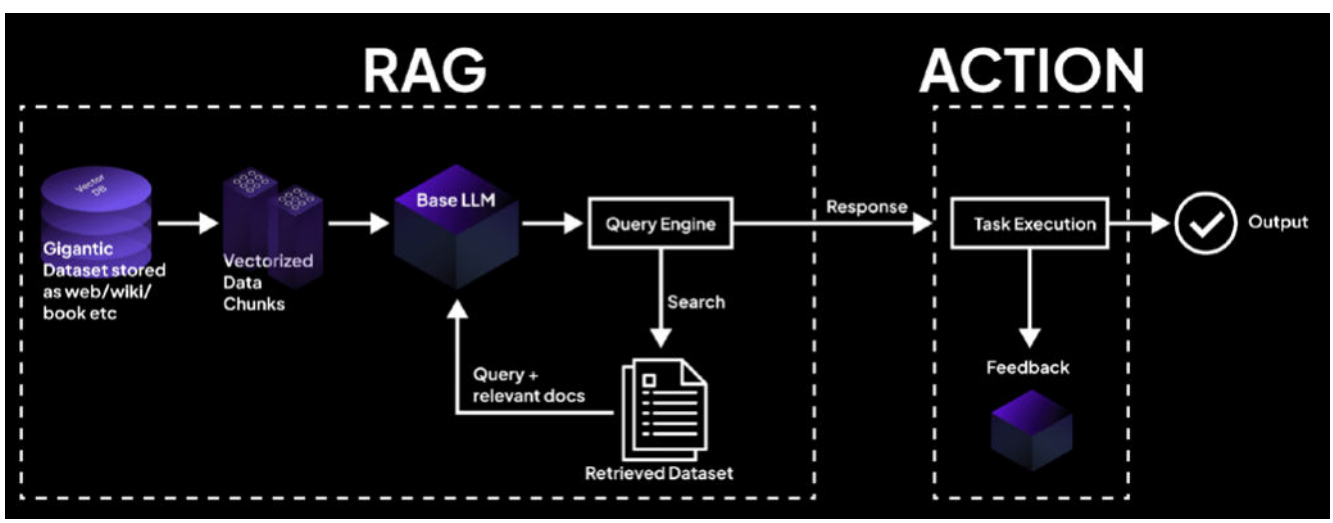
- Modèles "Open Weights" : contrairement aux modèles fermés (GPT-4), leurs fichiers sont téléchargeables pour une exécution 100 % locale et privée.
- Hugging Face : la plateforme incontournable (le "GitHub de l'IA") pour trouver et télécharger les modèles.
- LM Studio : logiciel avec interface graphique idéal pour débuter, tester la compatibilité avec votre matériel et gérer le multi-GPU.
- Ollama : outil performant qui fonctionne en arrière-plan, idéal pour une utilisation quotidienne et l'intégration dans d'autres applications.
- Format GGUF : le standard à privilégier pour sa flexibilité, permettant de répartir le modèle entre la carte graphique (VRAM) et la mémoire système (RAM).
- Choix de la quantification : le réglage Q4_K_M est le meilleur compromis entre intelligence et légèreté ; évitez de descendre sous Q3 pour limiter les erreurs.
- La Température : curseur de créativité (0 pour la précision technique/code, 0.8+ pour la rédaction créative).
- Multimodalité : pour analyser des images, il faut impérativement choisir un modèle spécifique portant la mention "Vision".

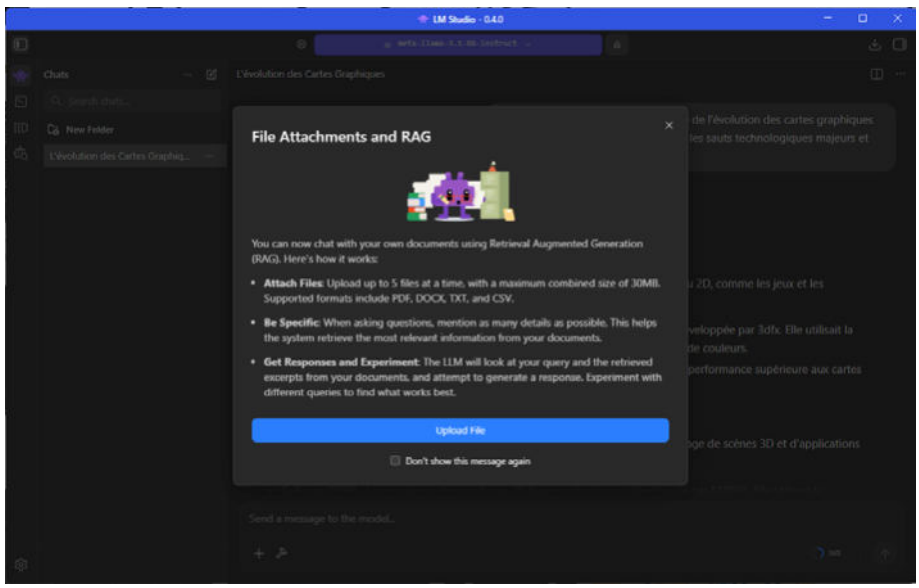
EXPERTISE : PERSONNALISER L'IA AVEC LE RAG ET LE FINE-TUNING

Attention à ne pas confondre les usages. Si Ollama et LM Studio permettent d'exploiter la puissance des LLM, ils ne permettent pas d'en améliorer l'usage en profitant de vos propres documents (RAG). Vous pourrez faire ça avec d'autres outils tels qu'Anything LLM (qui en plus s'interface aisément avec Ollama) ou Jan.ai. Ollama et LM Studio ne sont pas non plus capables d'entraîner les modèles. Pour le Fine-tuning (la personnalisation profonde du modèle), il faudra sortir des sentiers battus et utiliser des outils spécialisés comme Unsloth ou un script Python dédié.

LE RAG, OU COMMENT DONNER UNE MÉMOIRE À COURT TERME À VOTRE IA

Pour comprendre l'intérêt du RAG (Retrieval-Augmented Generation), il faut accepter une limite fondamentale des LLM ; leur savoir est figé à l'instant où leur entraînement s'est terminé. Si vous demandez à un modèle classique de résumer une réunion qui a eu lieu ce matin ou d'analyser un contrat que vous venez de signer, il sera incapable de vous répondre, ou pire, il inventera une réponse plausible, mais totalement fausse, ce qu'on appelle l'hallucination. Le RAG change la donne en agissant comme une mémoire de travail externe. Imaginez un étudiant brillant passant un examen ; sans RAG, il doit compter uniquement sur ses révisions pour répondre. Avec le RAG, il est autorisé à ouvrir son classeur personnel. Lorsqu'une question lui est posée, l'IA commence par parcourir les documents que vous lui avez fournis, identifie les passages les plus pertinents, puis utilise ces informations fraîches pour construire sa réponse. Cette technique offre deux avantages majeurs pour l'utilisateur local. D'abord, elle garantit une fiabilité exemplaire puisque l'IA ne "devine" plus, elle "cite" vos propres sources. Ensuite, elle préserve une confidentialité totale ; vos fichiers sensibles restent sur votre machine et ne sont jamais envoyés dans le cloud





pour être ingérés par des géants de la tech. C'est la solution idéale pour transformer un modèle généraliste en un assistant spécialisé sur vos données propres. Le RAG, c'est ce que vous faites à chaque fois que vous joignez une pièce jointe (un fichiers PDF ou Word) dans ChatPT par exemple.

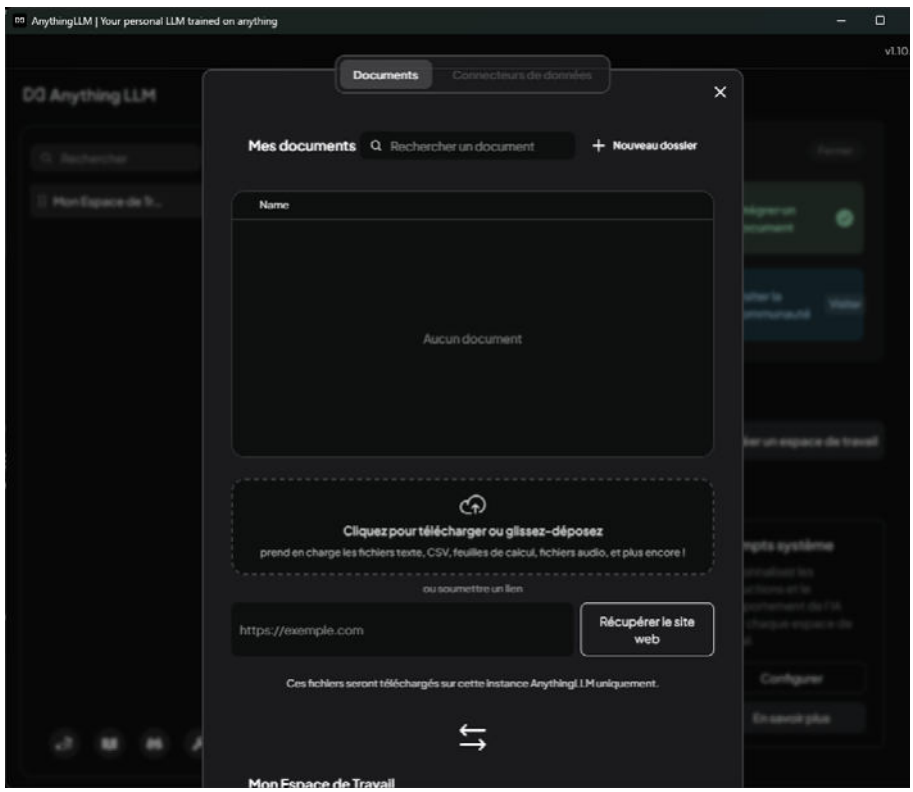
EN PRATIQUE, TRANSFORMEZ VOS DOSSIERS EN BASE DE CONNAISSANCES AVEC ANYTHINGLLM

Passer de la théorie à la pratique nécessite un outil capable de faire le pont entre vos fichiers et le moteur de l'IA. Si Ollama s'occupe de faire tourner le modèle en arrière-plan, AnythingLLM (Windows, Linux, macOS) s'impose comme l'interface idéale pour orchestrer vos documents. Une fois l'application lancée, la première étape indispensable est le jumelage avec votre moteur local (en l'occurrence Ollama). Allez dans les réglages (Settings), puis dans la section « AI Providers ». Sous l'onglet « LLM », sélectionnez *Ollama* comme fournisseur. Si Ollama tourne déjà sur votre machine, le logiciel devrait remplir automatiquement l'adresse du serveur (généralement http://127.0.0.1:11434). Cliquez sur *Check connection* ; si le lien est établi, une liste déroulante apparaîtra, vous permettant de choisir parmi les modèles que vous avez déjà téléchargés (Llama, Mistral, etc.). Répétez l'opération dans l'onglet *Embedding* en choisissant également *Ollama*, ce qui permettra à l'IA d'utiliser la même logique pour indexer vos fichiers. Le cœur de l'outil repose ensuite sur la création de « Workspaces » (espaces de travail). Vous pouvez en créer un pour vos factures, un autre pour vos projets pros. À l'intérieur de ces espaces, vous trouverez une zone de dépôt de fichiers. Glissez-y vos PDF, documents Word ou fichiers texte, puis cliquez sur le bouton *Move to Workspace* suivi de *Save and Embed*. C'est à cet instant que la magie opère ; le logiciel découpe vos textes en petits morceaux et les transforme en vecteurs numériques. Une fois cette indexation terminée, le chat change de nature. En mode « Query », l'IA ne répondra qu'en utilisant vos documents. Elle affichera même des citations cliquables ; un clic sur une source ouvre directement le document à la page exacte où l'IA a trouvé

> Entre le moteur et vos documents, AnythingLLM offre l'interface indispensable pour transformer vos fichiers bruts en une base de connaissances interactive.

LES MODÈLES PEUVENT-ILS CHERCHER DES INFORMATIONS SUR INTERNET ?

Dans leur forme brute et locale, les LLM sont totalement déconnectés du réseau. Contrairement aux versions cloud de ChatGPT ou de Gemini qui disposent d'outils de navigation intégrés, votre modèle installé via Ollama ou LM Studio ne connaît du monde que ce qu'il a appris lors de son entraînement. Pour lui donner accès au Web, il faut utiliser des interfaces spécifiques (comme AnythingLLM ou Perplexity-Llama) capables de jouer le rôle d'intermédiaire. Ces outils effectuent une recherche sur Google ou DuckDuckGo à votre place, récupèrent le contenu des sites pertinents et l'injectent dans le contexte du modèle via la technique du RAG. C'est une solution puissante, mais qui brise la règle du 100 % local et de la confidentialité totale si jamais.



l'information. C'est ici que votre matériel est sollicité ; l'étape d'indexation profite directement du nombre de cœurs de votre processeur et de la vitesse de votre RAM pour traiter des centaines de pages en quelques secondes.

ET LE GRAPHRAG, C'EST QUOI ?

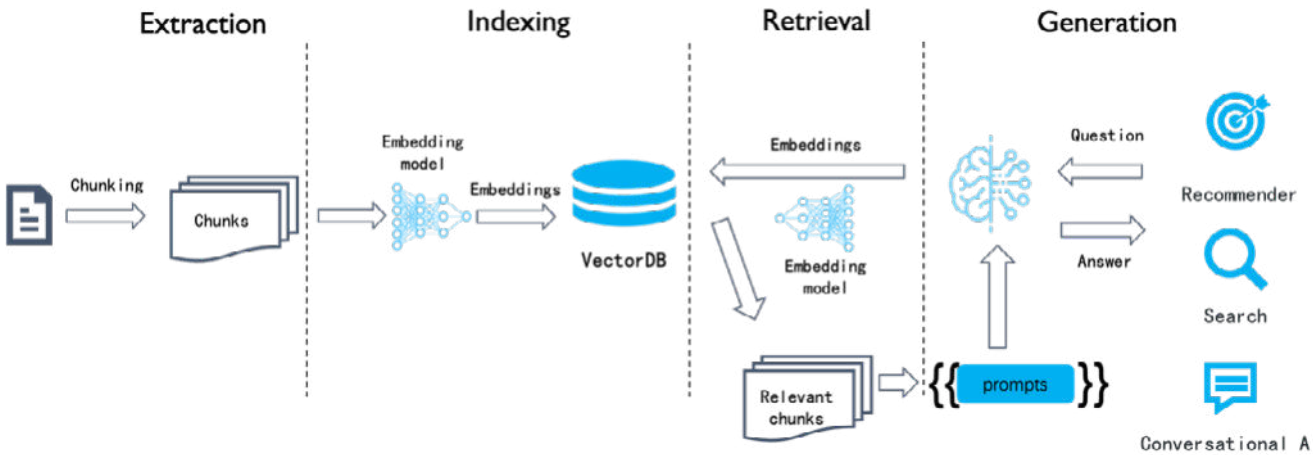
Si le RAG classique est une révolution, il possède une limite ; il est "aveugle" aux relations globales. Imaginez que vous demandiez à votre IA de résumer les points communs entre dix tests de boîtiers différents. Un RAG vectoriel classique va chercher des morceaux de texte isolés contenant le mot « boîtier », mais il aura du mal à synthétiser une vue d'ensemble. C'est là qu'intervient le GraphRAG. Au lieu de simplement stocker des blocs de texte, cette méthode utilise le LLM pour extraire des entités (des noms, des composants, des technologies) et, surtout, pour cartographier les relations qui les unissent. On ne cherche plus dans une bibliothèque de pages, mais dans une toile d'araignée géante où chaque concept est relié aux autres. Le résultat ? L'IA devient capable de répondre à des questions transversales complexes (« Quelles sont les tendances de refroidissement sur les modèles testés en 2025 ? ») avec une pertinence bien supérieure. C'est l'outil ultime pour l'Analyste, mais il a un coût ; il demande une puissance de calcul et une mémoire vive colossales lors de la création de cet index "intelligent". Un terrain de jeu idéal pour nos stations de travail les plus musclées. Si des outils populaires comme AnythingLLM hésitent encore à intégrer le GraphRAG, c'est justement parce qu'il demande une puissance machine que peu de PC possèdent. Faire du GraphRAG en local, c'est entrer dans le monde de la station de travail (type GX10) où les 128 Go de mémoire permettent enfin d'envisager ces structures de données complexes sans que le PC ne mette trois jours à indexer un dossier. Si vous voulez explorer cette technique, intéressez-vous à RAGFlow ou Kotaemon.

LE FINE-TUNING : QUAND L'IA CHANGE DE PERSONNALITÉ

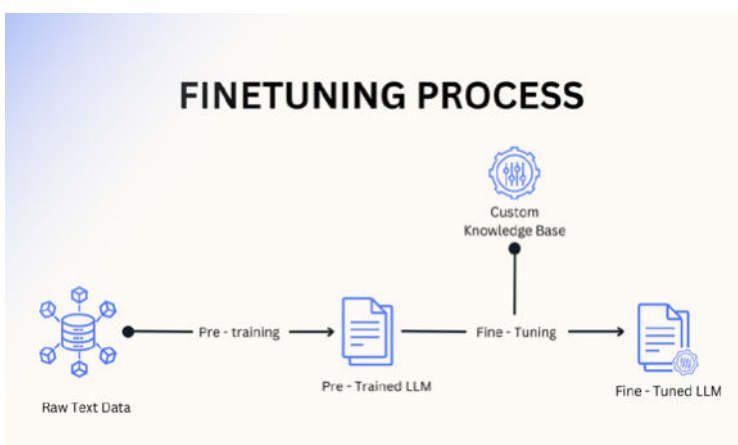
Si le RAG s'apparente à donner un manuel à un étudiant, le fine-tuning correspond à une spécialisation chirurgicale de son cerveau. Ici, on ne se contente plus de fournir des documents en annexe ; on modifie physiquement les poids du modèle (ses connexions neuronales) pour qu'il adopte un comportement, un style ou un savoir-faire spécifique. C'est l'étape ultime pour ceux qui veulent

POURQUOI NE PAS TOUT FAIRE AVEC ANYTHINGLLM ?

Notez qu'AnythingLLM propose son propre moteur de chat interne, c'est-à-dire qu'il peut exécuter de lui-même des modèles comme le font LM Studio et Ollama. Mais nous vous conseillons vivement de conserver Ollama comme fournisseur et de n'utiliser AnythingLLM que pour le RAG. En séparant le moteur (Ollama) de la carrosserie (AnythingLLM), vous gagnez en performances pures, notamment sur la gestion de la mémoire vidéo, et vous gardez la liberté d'utiliser vos modèles dans d'autres applications sans avoir à les télécharger plusieurs fois.



que leur IA ne se contente pas de répondre juste, mais qu'elle le fasse avec une expertise métier indétectable. L'intérêt du fine-tuning est avant tout une question de forme et de méthode. Vous pouvez, par exemple, entraîner un modèle sur des milliers de lignes de code informatique respectant strictement les normes de votre entreprise, ou sur l'intégralité de vos écrits passés pour qu'il adopte votre plume, vos tournures de phrases et votre structure de pensée. À l'issue du processus, le modèle n'a plus besoin qu'on lui rappelle ses consignes à chaque question ; ces directives sont désormais inscrites dans son ADN numérique. Toutefois, cette puissance vient avec une exigence technique et matérielle d'un tout autre niveau. Contrairement au RAG, qui est une opération de lecture, le fine-tuning est une opération d'écriture massive. Il nécessite une phase d'entraînement lourde où l'IA doit traiter des jeux de données de haute qualité (le fameux *Dataset*). C'est ici que la frontière entre le simple utilisateur et l'expert se dessine ; là où l'inférence (le chat) demande de la rapidité, le fine-tuning exige une endurance et une quantité de mémoire vidéo qui mettront à genoux n'importe quelle configuration, même puissante.



Le fine tuning consiste à adapter personnaliser un LLM selon ses besoins.

LORA ET QLORA : L'ENTRAÎNEMENT POUR TOUS (OU PRESQUE)

Pendant longtemps, le fine-tuning est resté une citadelle imprenable pour le commun des mortels. Modifier les milliards de paramètres d'un modèle réclame une puissance de calcul et une quantité de VRAM colossales, souvent réservées aux supercalculateurs. Mais l'arrivée de techniques d'optimisation comme le LoRA (Low-Rank Adaptation) a changé la donne, rendant l'entraînement accessible sur des

> Le fine-tuning ne se contente pas de donner un manuel à l'IA : il modifie son ADN numérique pour lui offrir une expertise métier chirurgicale.

	Pretrained Weight W	Pretrained Weight W ❄ B A LoRA (Hu et al., 2021)	Quantized Pretrained Weight ❄ Q B A QLoRA (Dettmers et al., 2023)
Memory of the model	7B x 16bit = 14GB	7B x 16bit = 14GB	7B x 4bit = 3.5GB
Memory of grad and optim states	7B x 32bit x 3 = 84GB	5% x 7B x 32bit x 3 = 4.2GB	5% x 7B x 32bit x 3 = 4.2GB
GPU Print	108GB	27GB	14.5GB

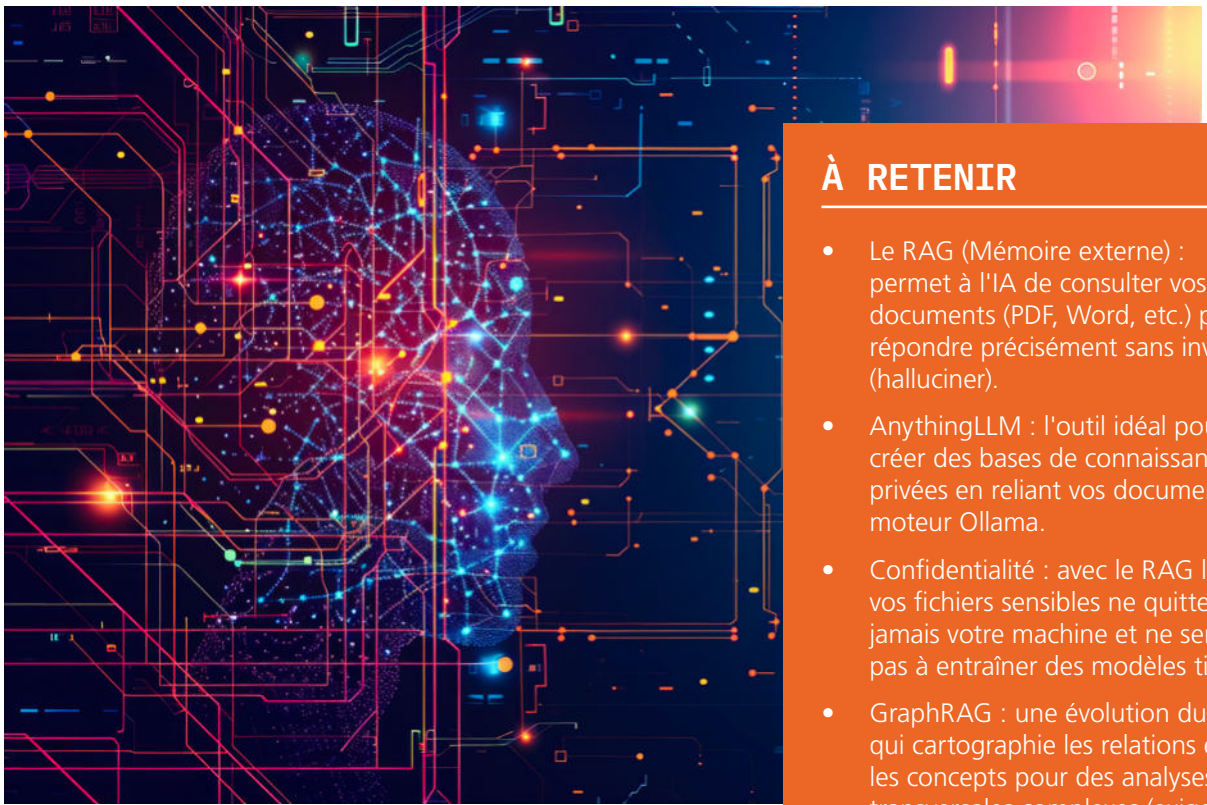
QLoRA rend le fine-tuning pas bien plus accessible aux quantités de mémoire de nos PC.

stations de travail domestiques musclées. Le principe du LoRA est d'une élégance rare ; au lieu de tenter de modifier tous les poids du modèle original (ce qui serait trop lourd), on fige le modèle de base et on lui ajoute de toutes petites couches de paramètres supplémentaires, comme des "adaptateurs". On n'entraîne que ces couches légères, qui viennent se greffer sur le cerveau de l'IA pour en modifier le comportement. C'est un peu comme si, au lieu de réécrire tout un livre pour y ajouter vos notes, vous vous contentiez d'y coller des Post-it aux endroits stratégiques. Le résultat est quasiment identique, mais l'effort de calcul est divisé par dix. Le QLoRA pousse le concept encore plus loin en combinant cette technique avec la quantification (le passage en 4 bits). Cette prouesse permet de fine-tuner des modèles imposants, comme un Llama de 70 milliards de paramètres, sur une configuration matérielle qui n'aurait normalement pu gérer qu'un petit modèle de 8 milliards. Ça prend quand même quelques heures, mais le simple fait que ce soit possible à l'échelle d'un particulier ou d'une PME est assez génial.

> Paramètres, contexte et entraînement : les trois piliers qui définissent votre expérience de l'IA, du simple chatbot au moteur professionnel spécialisé.

AXOLOTL, UNSLOTH : QUEL OUTIL POUR FRANCHIR LE PAS ?

Si le RAG est désormais à la portée d'un clic, le fine-tuning reste une discipline qui demande de mettre un peu plus les mains dans le cambouis. Ici, point d'interface graphique léchée à la LM Studio ; on entre dans le royaume des scripts et des lignes de commande. Pour l'utilisateur sous Windows ou Linux, deux noms dominent actuellement la scène de l'optimisation locale ; Axolotl et Unsloth. Axolotl est devenu le couteau suisse incontournable pour les experts. C'est un orchestrateur qui permet de configurer l'entraînement de presque tous les modèles existants via un simple fichier de configuration (YAML). Il gère nativement le LoRA et le QLoRA,



ainsi que des techniques plus pointues comme le Multi-GPU. C'est l'outil de référence pour ceux qui veulent un contrôle total sur chaque paramètre de l'apprentissage. Cependant, la véritable révolution de ces derniers mois s'appelle Unsloth. Cette bibliothèque s'est imposée comme la star du Fine-tuning local en réécrivant les algorithmes de calcul mathématique pour les rendre beaucoup plus légers. Les résultats sont spectaculaires ; Unsloth permet de fine-tuner un modèle jusqu'à deux fois plus vite tout en consommant 70 % de mémoire vidéo en moins. Sur une configuration musclée, cela signifie que vous pouvez entraîner un modèle de taille moyenne en moins de temps qu'il n'en faut pour déjeuner. Pour franchir le pas, il vous faudra sortir de l'environnement Windows classique pour installer un sous-système Linux (par exemple WSL 2 sous Windows) ou passer par des conteneurs Docker. C'est à l'intérieur de cet environnement que vous configurerez Python, chargerez les bibliothèques nécessaires et préparerez votre jeu de données (souvent au format JSON). C'est un cap technique certain, mais c'est à ce prix que vous passerez du statut de consommateur d'IA à celui de créateur. Une fois votre "adaptateur" LoRA créé, il ne vous restera plus qu'à l'injecter dans votre moteur habituel pour profiter d'une IA qui ne ressemble à aucune autre. Nous pensons que pour un premier contact avec les LLM, il n'est pas nécessaire d'entrer plus dans la pratique sur des sujets aussi avancés, mais les plus motivés d'entre vous auront au moins les pistes à suivre ;)

À RETENIR

- Le RAG (Mémoire externe) : permet à l'IA de consulter vos documents (PDF, Word, etc.) pour répondre précisément sans inventer (halluciner).
- AnythingLLM : l'outil idéal pour créer des bases de connaissances privées en reliant vos documents au moteur Ollama.
- Confidentialité : avec le RAG local, vos fichiers sensibles ne quittent jamais votre machine et ne servent pas à entraîner des modèles tiers.
- GraphRAG : une évolution du RAG qui cartographie les relations entre les concepts pour des analyses transversales complexes (exige beaucoup de RAM).
- Le fine-tuning (ADN numérique) : processus consistant à modifier les poids du modèle pour lui apprendre un style, une plume ou une expertise métier précise.
- LoRA et QLoRA : techniques d'optimisation permettant d'entraîner une IA sur un PC domestique en n'ajoutant que de légères couches de neurones ("post-it").
- Unsloth : la solution logicielle de référence pour le fine-tuning rapide, réduisant drastiquement la consommation de mémoire vidéo (VRAM).
- Utilisez le RAG pour le savoir (faits, documents) et le fine-tuning pour la forme (style, comportement, logique métier).

100 % HARDWARE : QUELLE CONFIGURATION POUR QUEL USAGE ?

Uber Hardware, c'est avant tout un lien privilégié avec le matériel. On va donc rentrer dans le vif du sujet avec une batterie de mesures de performances comme on aime. Dans un premier temps, on va se concentrer sur les performances à proprement parler, c'est-à-dire la vitesse d'exécution des LLM. Pour ça, il existe plusieurs métriques. La plus importante, c'est le débit de l'inférence qui s'exprime en tokens par secondes (tok/s). Mais il faut aussi considérer le TTFT, Time To First Token, c'est le temps que l'IA met à interpréter votre prompt avant de débuter sa réponse. Enfin, nous mesurons aussi le temps que met le modèle à se charger initialement, avant même que vous ne puissiez entrer votre premier prompt.

LES TOKENS/S

La vraie valeur qui compte, c'est le débit de parole, non ? Personne n'a envie d'attendre 107 ans derrière son écran que son chatbot crache sa réponse. Un être humain est capable de lire, en général, entre 200 et 300 mots par minutes. Ce qui veut dire grosso modo 3,5 à 5 mots par secondes. Autrement dit, sachant que 1000 tokens représentent environ 750 mots, entre 4 à 6 tok/s. Jusqu'à 10 tok/s si on lit rapidement, en diagonale comme on dit. Partant de ce constat, il est donc agréable pour utiliser un LLM régulièrement qu'il débite au moins à 10/15 tok/s. S'il fait 4 ou 5 tok/s, c'est encore utilisable, mais c'est nettement moins agréable. S'il tombe à 0,5 tok/s (et ça n'a rien d'impossible), on n'a plus envie de discuter avec.

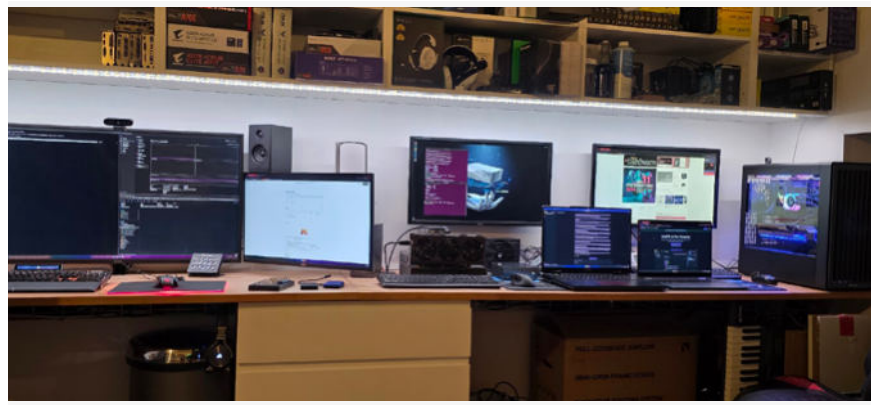
À titre de comparaison, la majorité des gros LLM du cloud débitent entre 50 et 80 tok/s ; on voit bien d'ailleurs que le texte s'affiche beaucoup plus rapidement qu'on n'est capable de le lire. La vitesse d'inférence d'une IA cloud est extrêmement variable car elle dépend de la charge des serveurs à un instant T. Car vous n'êtes pas seul dessus. En parlant de ça, c'est aussi valable pour les LLM locaux. Vous verrez que certaines cartes graphiques récentes sont capables de dépasser les 200 tok/s dans certains contextes. Ça semble totalement superflu de prime abord, sauf si on a des besoins de scalabilité. Par exemple mettre en place un chatbot SAV dans son entreprise capable de répondre à plusieurs clients à la fois. Avec 200 tok/s, vous pouvez répondre à 10 clients simultanés (OK, la scalabilité n'est pas parfaitement linéaire, c'est pour illustrer) tandis que le service s'effondrerait rapidement sur un PC ne débitant que 30 ou 40 tok/s.

5 ÉPREUVES DE VITESSE

Pour se rendre compte des possibilités selon le hardware dont vous disposez, nous avons mené cinq tests de performance différents sur chacune des machines que nous venons de citer. Chaque test a par ailleurs été lancé 3x et nos graphiques de performances sont la moyenne de ces trois mesures. Nous étions partis pour vous proposer les tests Llama en utilisant les tout

Épreuve	Modèle (Taille/sortie)	Poids sans/avec KV Cache (Quantification)	Enjeu Technique
L'efficience	Phi-4 Mini (3.8B/ janvier 2025)	2,5/3,3 Go (Q4_K_M)	Plafond de réactivité du bus et des drivers et adapté aux GPU modestes n'ayant que 6 Go.
Le sprint	Llama 3.1 (8B/juillet 2024)	4,9/5,5 Go/ (Q4_K_M)	Fluidité réelle sur le modèle de référence pour une majorité de GPU.
L'endurance	Mistral Small 3 (22B/septembre 2024)	12/13 Go (Q4_0)	Sanction immédiate pour les GPU de 8 Go, met en valeur les cartes avec 16 Go.
Le mur	Llama 3.3 (70B/décembre 2024)	26/27 Go (Q2_K)	Le maximum lançable sur un super PC équipé d'une RTX 5090 et ses 32 Go.
La montagne	Command R+ (104B/avril 2024)	59/61 Go (Q4_0)	Le territoire réservé des machines aux 128 Go de mémoire unifiée.

nouveaux modèles Llama 4, mais il s'avère qu'ils n'existent pas dans des tailles qui correspondent aux besoins de cet article ! Nous utilisons donc deux modèles Llama 3.x (3.1 8B et 3.3 70B), qui ont pour avantage d'être de véritables références depuis leur sortie en 2024 (respectivement en juillet et décembre) et dont on trouve partout des chiffres sur Internet. Comme nous l'avons exprimé plus tôt, une quantification Q4 est idéale pour réduire la taille d'un modèle sans visiblement altérer des capacités de réflexion. Pour Llama 3.3 70B, on a volontairement choisi de l'utiliser e, Q2 car c'est la seule manière qu'il entre sur un PC "classique" équipé d'une GeForce RTX 5090 et ses 32 Go ; de plus, avec un modèle aussi évolué, la compression Q2 reste parfaitement exploitable. Nous avons aussi choisi Mistral Small 3 22B, plutôt que son remplaçant Mistral Small 3.2 24B tout récent, afin qu'il puisse tenir à 100 % en VRAM sur nos PC de test équipés d'un GPU avec 16 Go, ça passe tout juste ! Et encore, s'il tient dans les 16 Go d'une GeForce… ce ne fût pas le cas sur le PC avec la Radeon RX 9070 XT ! De peu, mais nous étions en mode 6% CPU / 94 % GPU. Nous reviendrons plus longuement dans notre analyse des performances sur ce qu'il se passe quand on dépasse la capacité mémoire de la carte graphique, mais vous verrez que l'impact sur les performances est immédiat et important. Dans le tableau ci-contre, le poids avec KV Cache comprend simplement la fenêtre de contexte par défaut d'Ollama, à savoir 4 096 tokens. Le besoin en mémoire augmente en même temps que vous souhaitez profiter d'un contexte plus important. D'ailleurs, pour le test d'endurance, dans lequel l'IA est supposée résumer en 10 points



NOS CONDITIONS DE TEST

Avant de publier nos résultats, voici le détail de nos conditions de test. Il y a 10 PC en tout, mais on débute par la configuration logicielle utilisée :

CONFIGURATION LOGICIELLE
OS : Windows 11 25H2 (sauf Asus Ascent GX10 sous DGX OS)
Pilotes Nvidia : 591.86
Pilotes AMD : 26.1.1
Pilotes Intel : 32.0.101.8425
LM Studio : 0.4.0
Ollama : 0.15.2

DESKTOP 1
CPU : Core i5-10500
RAM : 32 Go (4* 8 Go) DDR4-2666 (Corsair Dominator 4000)
GPU : ARC B580 12 Go (Intel de référence)

DESKTOP 2
CPU : Ryzen 5 7600X
RAM : 32 Go (2* 16 Go) DDR5-6400 (Lexar Ares 7200)
GPU : GeForce RTX 3070 8 Go (Asus TUF Gaming OC)

DESKTOP 3
CPU : Intel Core i9-14900K
RAM : 32 Go (2* 16 Go) DDR5-5600 (Lexar Ares 7200)
GPU : Radeon RX 9070 XT 16 Go (Predator BiFrost OC)

DESKTOP 4
CPU : Intel Core Ultra 7 265K
RAM : 32 Go (2* 16 Go) DDR5-7200 (Lexar Ares 7200)
GPU : GeForce RTX 5070 Ti 16 Go (Zotac AMP Extreme Infinity)

DESKTOP 5
CPU : Intel Core Ultra 9 285K
RAM : 32 Go (2* 16 Go) DDR5-8000 (Lexar Ares 8000)
GPU : GeForce RTX 5080 16 Go (Gigabyte Gaming OC)

DESKTOP 6
CPU : AMD Ryzen 9 9900X
RAM : 48 Go (2* 24 Go) DDR5-5600 (Kingston FURY Renegade 8200)
GPU : GeForce RTX 5090 32 Go (Zotac AMP Extreme Infinity)

LAPTOP 1 : SAMSUNG GALAXY BOOK5 PRO 360
CPU : Core Ultra 7 256V
RAM : 16 Go LPDDR5x-8533
GPU : Arc 140V

LAPTOP 2 : RAZER BLADE 16 :
CPU : Ryzen AI 9 365
RAM : 32 Go LPDDR5x-8000
GPU : GeForce RTX 5080 Laptop 16 Go

LAPTOP 3 : MSI TITAN 18 HX AI (2025)
CPU : Intel Core Ultra 9 285HX
RAM : 64 Go DDR5-6400
GPU : GeForce RTX 5090 Laptop 24 Go

WORKSTATION IA À MÉMOIRE UNIFIÉE : ASUS ASCENT GX10
CPU/GPU : Nvidia GB10
Mémoire unifiée : 128 Go LPDDR5x-8533

clés un de nos tests de boîtier, on a dû augmenter très légèrement le contexte (4 500 tokens) pour que ça passe.

LES PROMPTS UTILISÉS

Pour que vous jugiez de nos tests et que vous puissiez les reproduire chez vous, voici les prompts que nous avons utilisé pour chacun de ces cinq tests :

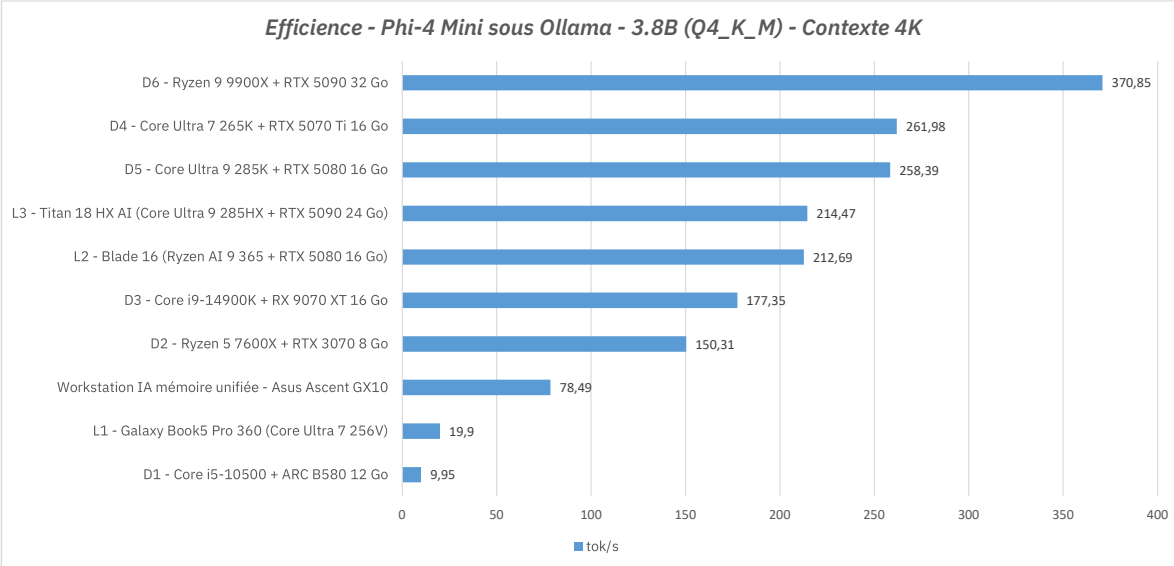
- **L'efficience** : raconte-moi une courte histoire de 10 lignes sur un robot qui découvre la pluie.
- **Le sprint** : peux-tu me faire un historique détaillé de l'évolution des cartes graphiques de 1995 à 2026, en te concentrant sur les sauts technologiques majeurs et l'arrivée de l'IA ?
- **L'endurance** : Liste-moi 10 points clés qui caractérisent ce boîtier (suivi du test du HAVN BF 360 Flow publié dans ce même numéro, sans la conclusion).
- **Le mur** : analyse l'impact d'une transition mondiale vers l'hydrogène vert d'ici 2040, mais confronte spécifiquement le paradoxe suivant ; comment les pays du Golfe peuvent-ils maintenir leur influence géopolitique tout en perdant leur rente pétrolière ? Propose trois scénarios de transition réalistes et critique les limites de l'infrastructure européenne actuelle.
- **La montagne** : rédige une analyse approfondie comparant deux époques charnières ; l'ère des systèmes experts (1956-1990) et l'ère de l'apprentissage profond (2012-2026). Explique comment le passage des processeurs symboliques à la puissance de calcul des GPU (de l'architecture Tesla à Blackwell) a permis de passer d'une IA qui 'suit des règles' à une IA qui 'apprend par elle-même'. Conclue sur l'émergence des modèles de raisonnement actuels et leur impact sur la créativité humaine.

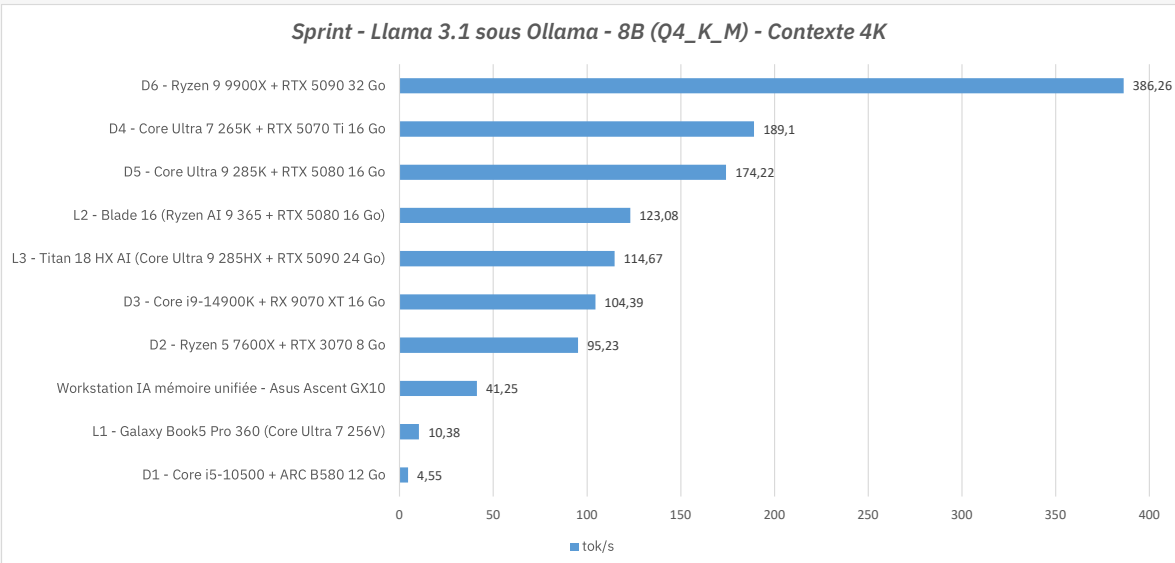
ENFIN LE VERDICT !

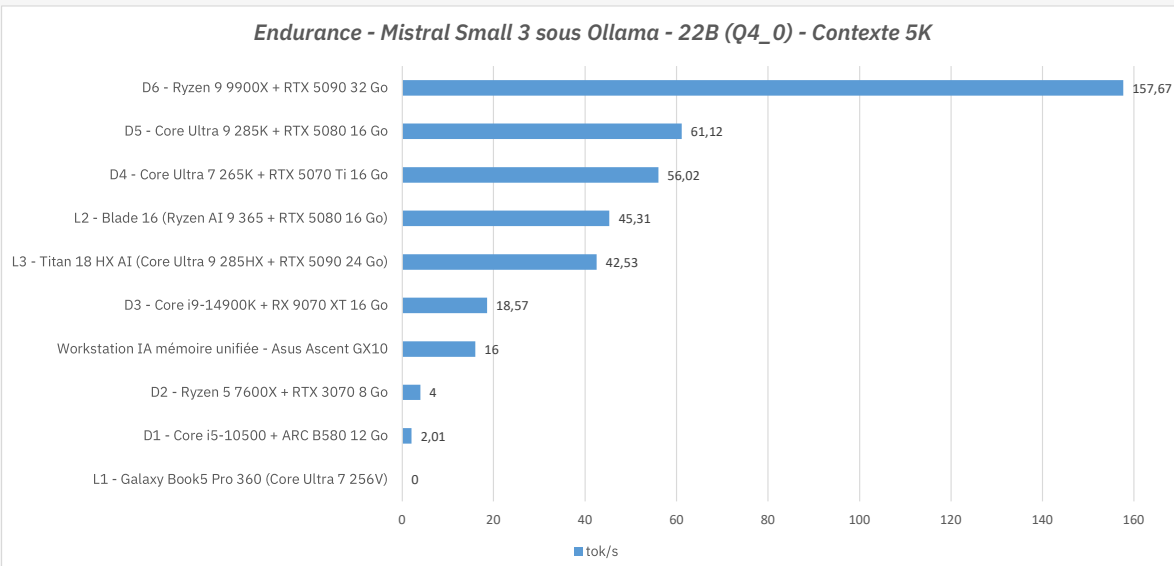
Le moins qu'on puisse dire est qu'on constate des différences phénoménales suivant les PC. Jugez plutôt en parcourant les graphiques de performance. Précisons qu'on traite exclusivement de vitesse pour le moment, on traitera de la qualité des réponses, de "l'intelligence" de votre LLM, dans la partie suivante de ce dossier.

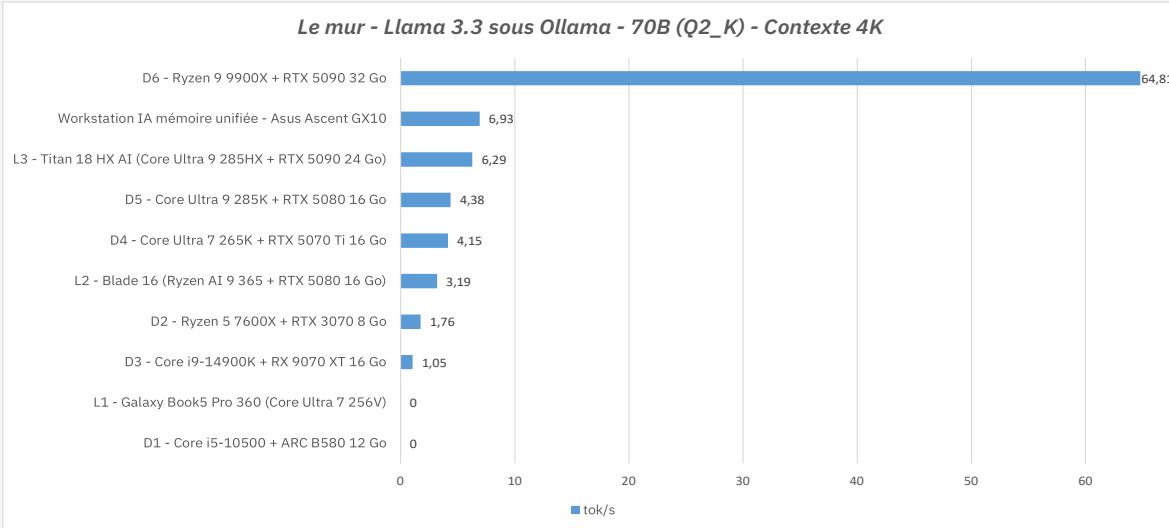
On constate de prime abord que le premier PC assemblé s'avère extrêmement lent dans les 5 graphiques réalisés sous Ollama, mais qu'il s'en sort mieux sur le graphique du test complémentaire fait avec LM Studio. Ceci s'explique par le fait qu'Ollama ne prend pas en charge l'accélération matérielle du GPU Intel ; il existe bien des tutoriels pour y arriver, en partant d'une image Docker fourni par Intel, mais c'est une complication sans fin. En revanche, la B580 est correctement exploitée sous LM Studio. Et comme vous pouvez le voir, entre la bande passante de cette dernière et la DDR4-2666 de la carte mère, c'est le jour et la nuit. Il a failli se produire la même chose avec le PC équipé de la Radeon RX 9070 XT, car là aussi Ollama se lance par défaut en mode 100 % CPU, sans exploiter l'accélération du GPU, mais tout se passe bien sous LM Studio. Toutefois, le 26 janvier, AMD a sorti un nouveau package AI Bundle (c'est une case optionnelle à cocher durant l'installation des pilotes Adrenaline) qui installe les librairies ROCm sous Windows et des logiciels optimisés pour bien exploiter les puces RDNA. Verdict ? Aucune différence pour Ollama, toujours bloqué en mode CPU. On a donc désinstallé notre Ollama 0.15.2 installé précédemment pour installer Ollama cette fois via l'interface AMD Install Manager (depuis les pilotes Adrenaline), qui installe une version 0.13.5 d'Ollama, mais ce n'est pas mieux. Enfin, après avoir de nouveau mis à jour Ollama pour repasser en 0.15.2, on a d'abord mis comme variable *$env:OLLAMA_VULKAN="1"* puis relancé le serveur via la commande *ollama serve* et, miracle, c'est bon ! On est bien en mode 100 % GPU ! Notre expérience sous LM Studio, bien plus limpide, montre de toute façon que Vulkan est pour le moment bien plus performant (plus rapide et plus économe en énergie) que ROCm, l'API officielle AMD pour le GPGPU. Tout ça démontre bien l'avance dont Nvidia dispose sur ses concurrents, une fois encore, car CUDA est développé et supporté depuis bien des années. Nous n'avons rien eu à faire pour qu'Ollama soit en 100 % GPU sur l'ensemble des PC équipés d'une GeForce ou sur la station de travail GB10.

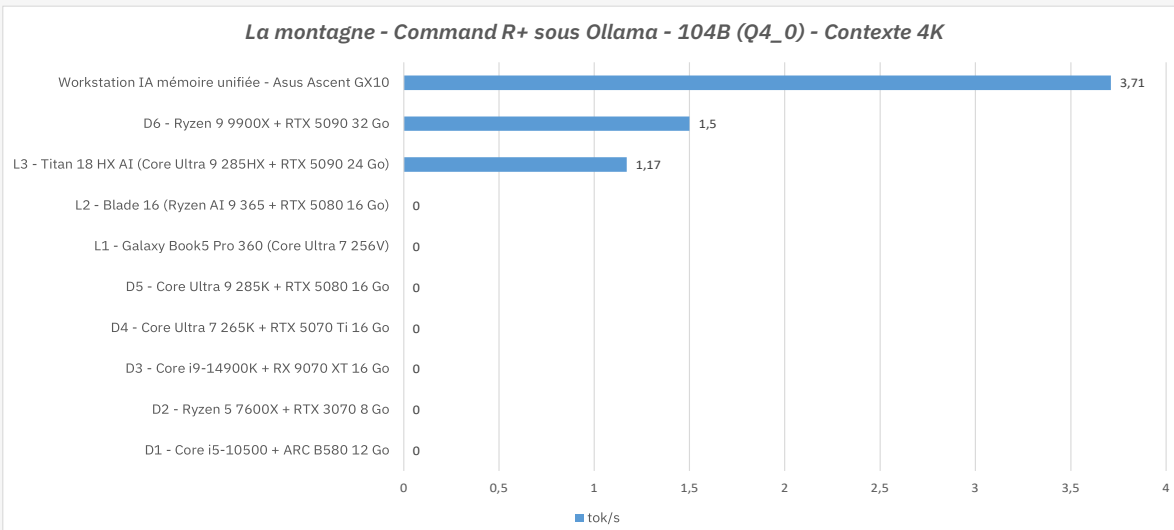
> Des différences phénoménales : nos tests révèlent que la vitesse d'une IA locale dépend autant du pilote installé que de la puissance brute de la carte graphique.

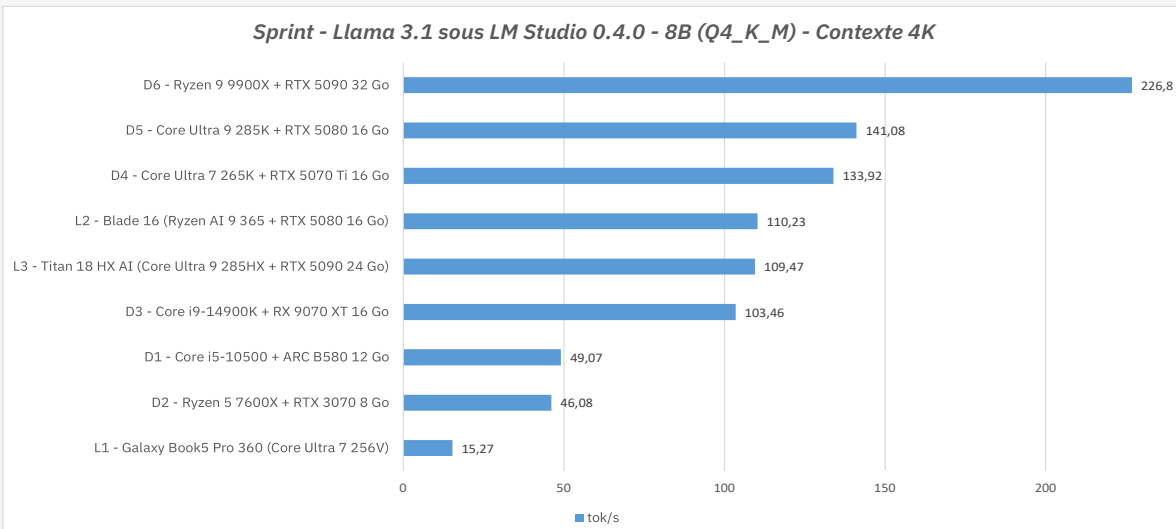

Efficience - Phi-4 Mini sous Ollama - 3.8B (Q4_K_M) - Contexte 4K
D6 - Ryzen 9 9900X + RTX 5090 32 Go
370,85
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
261,98
D5 - Core Ultra 9 285K + RTX 5080 16 Go
258,39
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
214,47
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
212,69
D3 - Core i9-14900K + RX 9070 XT 16 Go
177,35
D2 - Ryzen 5 7600X + RTX 3070 8 Go
150,31
Workstation IA mémoire unifiée - Asus Ascent GX10
78,49
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
19,9
D1 - Core i5-10500 + ARC B580 12 Go
9,95
0
50
100
150
200
250
300
350
400
tok/s

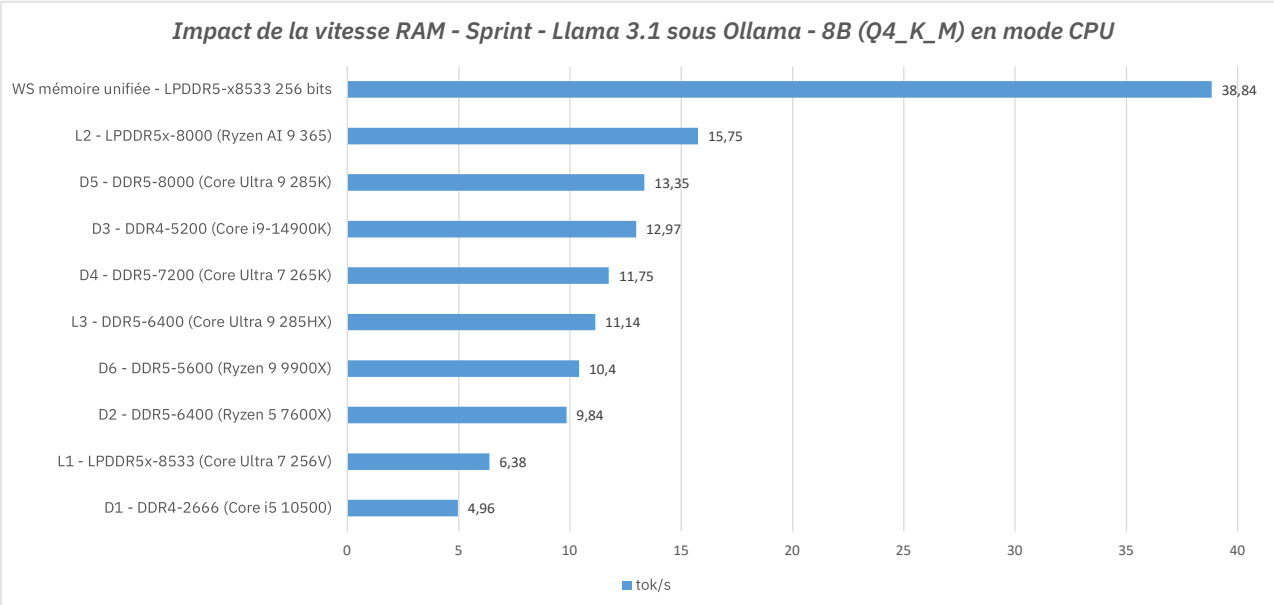

Sprint - Llama 3.1 sous Ollama - 8B (Q4_K_M) - Contexte 4K
D6 - Ryzen 9 9900X + RTX 5090 32 Go
386,26
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
189,1
D5 - Core Ultra 9 285K + RTX 5080 16 Go
174,22
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
123,08
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
114,67
D3 - Core i9-14900K + RX 9070 XT 16 Go
104,39
D2 - Ryzen 5 7600X + RTX 3070 8 Go
95,23
Workstation IA mémoire unifiée - Asus Ascent GX10
41,25
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
10,38
D1 - Core i5-10500 + ARC B580 12 Go
4,55
0
50
100
150
200
250
300
350
400
tok/s


Endurance - Mistral Small 3 sous Ollama - 22B (Q4_0) - Contexte 5K
D6 - Ryzen 9 9900X + RTX 5090 32 Go
157,67
D5 - Core Ultra 9 285K + RTX 5080 16 Go
61,12
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
56,02
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
45,31
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
42,53
D3 - Core i9-14900K + RX 9070 XT 16 Go
18,57
Workstation IA mémoire unifiée - Asus Ascent GX10
16
D2 - Ryzen 5 7600X + RTX 3070 8 Go
4
D1 - Core i5-10500 + ARC B580 12 Go
2,01
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
0
0
20
40
60
80
100
120
140
160
tok/s


Le mur - Llama 3.3 sous Ollama - 70B (Q2_K) - Contexte 4K
D6 - Ryzen 9 9900X + RTX 5090 32 Go
64,81
Workstation IA mémoire unifiée - Asus Ascent GX10
6,93
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
6,29
D5 - Core Ultra 9 285K + RTX 5080 16 Go
4,38
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
4,15
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
3,19
D2 - Ryzen 5 7600X + RTX 3070 8 Go
1,76
D3 - Core i9-14900K + RX 9070 XT 16 Go
1,05
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
0
D1 - Core i5-10500 + ARC B580 12 Go
0
0
10
20
30
40
50
60
tok/s


La montagne - Command R+ sous Ollama - 104B (Q4_0) - Contexte 4K
Workstation IA mémoire unifiée - Asus Ascent GX10
3,71
D6 - Ryzen 9 9900X + RTX 5090 32 Go
1,5
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
1,17
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
0
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
0
D5 - Core Ultra 9 285K + RTX 5080 16 Go
0
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
0
D3 - Core i9-14900K + RX 9070 XT 16 Go
0
D2 - Ryzen 5 7600X + RTX 3070 8 Go
0
D1 - Core i5-10500 + ARC B580 12 Go
0
0
0,5
1
1,5
2
2,5
3
3,5
4
tok/s


Sprint - Llama 3.1 sous LM Studio 0.4.0 - 8B (Q4_K_M) - Contexte 4K
D6 - Ryzen 9 9900X + RTX 5090 32 Go
226,8
D5 - Core Ultra 9 285K + RTX 5080 16 Go
141,08
D4 - Core Ultra 7 265K + RTX 5070 Ti 16 Go
133,92
L2 - Blade 16 (Ryzen AI 9 365 + RTX 5080 16 Go)
110,23
L3 - Titan 18 HX AI (Core Ultra 9 285HX + RTX 5090 24 Go)
109,47
D3 - Core i9-14900K + RX 9070 XT 16 Go
103,46
D1 - Core i5-10500 + ARC B580 12 Go
49,07
D2 - Ryzen 5 7600X + RTX 3070 8 Go
46,08
L1 - Galaxy Book5 Pro 360 (Core Ultra 7 256V)
15,27
0
50
100
150
200
tok/s



L'IMPACT MONSTRE DE LA BANDE PASSANTE VRAM (ET RAM)

Dans le monde du hardware traditionnel, nous avons l'habitude de surveiller la puissance brute de calcul (les TFLOPS). Mais en IA locale, le véritable juge de paix, c'est la bande passante mémoire. Pour comprendre pourquoi, il faut visualiser le fonctionnement d'un LLM ; à chaque fois que l'IA génère un mot (un token), elle doit « lire » l'intégralité des paramètres du modèle. Si vous utilisez un modèle de 8 milliards de paramètres (8B) quantifié en 4-bits, cela représente environ 5 Go de données à déplacer de la mémoire vers les cœurs de calcul... pour un seul et unique token. Pour atteindre une vitesse fluide de 100 tok/s, votre système doit donc être capable de brasser 500 Go de données par seconde (5 Go x 100). C'est là que le goulot d'étranglement apparaît. C'est ce qui explique la hiérarchie brutale de nos tests : **Les GPU dédiés** : avec de la GDDR7 ultra rapide et des bus mémoire allant jusqu'à 512 bits dans le cas de la RTX 5090, les cartes graphiques offrent des bandes passantes énormes, certaines dépassant même les 1 To/s. Le modèle est chargé et lu presque instantanément, ce qui permet des débits dépassant les 150 tok/s. **Les puces à mémoire unifiée (Apple M4 Ultra, Nvidia GB10, AMD Strix Halo)** : en soudant la RAM au plus près du processeur et en élargissant les bus de communication, ces solutions offrent des bandes passantes de 250 à 800 Go/s. Elles permettent de faire tourner des modèles énormes (70B) avec une fluidité déconcertante, là où un PC classique s'effondre. **Le passage en RAM (GGUF sur CPU)** : c'est le scénario catastrophe. Même avec de la DDR5 rapide en bicanal, on plafonne souvent entre 50 et 100 Go/s. Résultat ; dès que le modèle ne tient plus en VRAM (mémoire vidéo) et déborde sur la RAM du PC, les performances sont divisées par 10 ou 20. On passe d'une conversation fluide à un affichage de texte "mot à mot" pénible à suivre ; tout dépend, comme vous le voyez dans les graphiques, de la taille du modèle.

Le graphique ci-dessus, dédié à la RAM, montre l'impact de la vitesse d'un même modèle exécuté exclusivement en RAM (mode CPU 100 %) ; nous n'utilisons plusieurs plateformes, ce qui fausse un peu les résultats, mais la vitesse de la RAM est quand même le facteur limitant, peu importe la puissance du CPU pour le nombre de tok/s avec un seul utilisateur.

Malgré les heures passées à faire toutes ces mesures, nous avons choisi de ne pas multiplier les graphiques en nous concentrant sur l'essentiel. Vous pouvez néanmoins retrouver l'intégralité de nos résultats, y compris les TTFT ou la consommation électrique, sur ce fichier Google Sheet hébergé sur Google Drive (il faut être identifié avec son compte Google).

COMPRENDRE LE RÔLE DU GPU

Pour bien comprendre comment votre matériel influence l'IA, il faut distinguer deux composants qui travaillent main dans la main au sein de la carte graphique ; les cœurs de calcul (le GPU proprement dit) et la mémoire vidéo (VRAM). Mais attention aux idées reçues ; ici, le plus rapide n'est pas forcément celui qu'on croit. Le GPU, avec ses milliers de cœurs spécialisés (les cœurs CUDA chez Nvidia), est le muscle. Il réalise les opérations mathématiques pour prédire chaque token. Cependant, dans le cas particulier des LLM, le GPU est généralement sous-exploité. D'ailleurs on a eu le loisir de le constater durant nos tests, nous avons rarement vu le GPU dépasser quelques pour cent d'utilisation, même sur les cartes les plus modestes. Pourquoi ? Parce qu'un modèle de langage passe son temps à déplacer des

données. Pour générer un seul mot, le système doit lire l'intégralité des paramètres du modèle stockés en VRAM comme nous l'avons dit il y a peu. C'est là que la VRAM et surtout sa bande passante (sa vitesse de transfert) deviennent le véritable acteur. En pratique, c'est elle qui dicte la vitesse de génération (les tokens par seconde). Même avec le GPU le plus puissant du monde, si votre bande passante mémoire est faible, votre IA sera lente. La mémoire n'est donc pas qu'un simple réservoir, c'est elle qui régule le flux :

La capacité (Go) : elle définit la taille du cerveau. C'est la limite physique ; soit le modèle "tient" en mémoire, soit il ne tient pas. Si vous manquez de place, vous devrez utiliser des modèles plus petits ou plus compressés.

La bande passante (Go/s) : c'est elle qui définit la vitesse réelle de discussion. C'est le véritable goulot d'étranglement de l'IA locale. Une carte avec une bande passante monstre fera défiler le texte

LA BANDE PASSANTE MÉMOIRE

Composant	Bande passante	Notes
ARC B580 12 Go	456 Go/s	GDDR6 128 bits
GF RTX 3070 8 Go	448 Go/s	GDDR6 256 bits
GF RTX 5070 Ti 16 Go	896 Go/s	GDDR7 256 bits
GF RTX 5080 16 Go	960 Go/s	GDDR7 256 bits
GF RTX 5090 32 Go	1 790 Go/s	GDDR7 512 bits
GF RTX 5080 Laptop 16 Go	896 Go/s	GDDR7 256 bits
GF RTX 5090 Laptop 32 Go	896 Go/s	GDDR7 256 bits
Radeon RX 9070 XT 16 Go	644,6 Go/s	GDDR6 256 bits
DDR4-3200	51,2 Go/s	Dual channel (128 bits)
DDR5-5600	89,6 Go/s	Dual channel (128 bits)
DDR5-6400	102,4 Go/s	Dual channel (128 bits)
DDR5-7200	115,2 Go/s	Dual channel (128 bits)
DDR5-8000	128 Go/s	Dual channel (128 bits)
LPDDR5x-8000	128 Go/s	Dual channel (128 bits)
LPDDR5x-8533	136,6 Go/s	Dual channel (128 bits)
LPDDR5x-8533 unifié GB10	273,1 Go/s	Quad channel (256 bits)

Si vous désirez calculer la bande passante de votre RAM, n'hésitez pas à utiliser une calculatrice en ligne telle que celle-ci.

instantanément, car elle sature moins vite le GPU en données.
En résumé, si le GPU est le moteur, la VRAM est à la fois le réservoir et la pompe à essence. Vous pouvez avoir 20 000 cœurs de calcul, si la mémoire ne peut leur fournir les données qu'au compte-gouttes, votre débit de tokens restera désespérément bas. C'est pour cela qu'une carte "pro" ou une puce à mémoire unifiée (comme la GB10) peut parfois surpasser des cartes gaming surpuissantes ; leur architecture privilégie la vitesse de circulation des données plutôt que la force brute de calcul. Précisons quand même que le GPU joue un rôle prépondérant pour le TTFT (l'interprétation de votre prompt), mais ça ne dure qu'une fraction de seconde comparé à la génération du texte qui, là, est conditionné par la vitesse de votre VRAM (si le modèle tourne à 100 % dans cette dernière !). Toutefois et pour nuancer ce que nous venons d'évoquer, quand la VRAM dépote, le GPU n'est pas neutre, comme nous l'avons parfois constaté via l'usage GPU dans le gestionnaire des tâches et sur la mesure de consommation. Alors que dans la plupart des cas le GPU reste quasiment au repos, sur le PC le plus puissant avec la RTX 5090 (et sa fameuse VRAM à 1 790 Go/s), son chip est proche de 100 % d'utilisation durant toute l'inférence et le PC réclame plus de 700 W.

> Du repos forcé à la saturation totale : sur les configurations extrêmes, la vitesse de la VRAM parvient enfin à libérer toute la puissance brute du GPU.

QUE SE PASSE-T-IL SI LE MODÈLE NE TIENT PAS EN VRAM ?

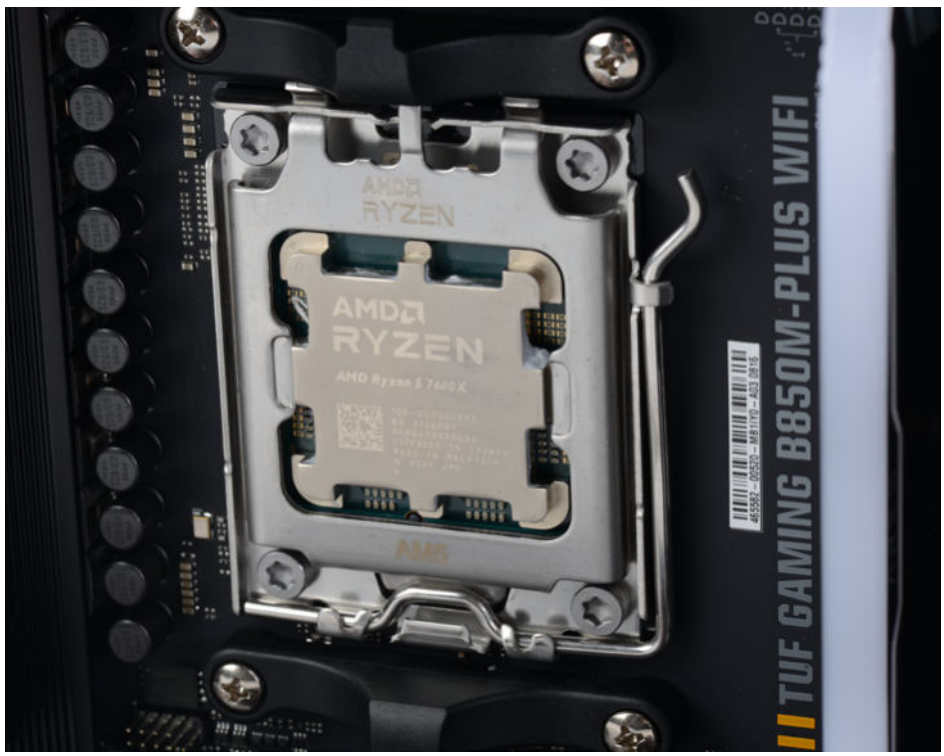
C'est le scénario que redoute (ou devrait redouter) tout utilisateur d'IA locale ; vouloir lancer un modèle trop gros pour sa carte graphique. Contrairement à un jeu vidéo qui pourrait simplement planter ou refuser de se lancer, les outils de LLM (comme Ollama ou LM Studio) sont plus résilients, mais le prix à payer est lourd. Si votre modèle pèse 12 Go et que votre carte n'en possède que 8, le logiciel va procéder à un offloading (déchargement) partiel. En effet, un LLM n'est pas un bloc monolithique, mais un empilement de couches successives (layers). Si votre carte graphique ne peut pas supporter l'intégralité du poids, le logiciel va placer autant de couches que possible dans la VRAM et déverser le reste dans la RAM système. Ce mélange des genres est une bouée de sauvetage, mais elle transforme votre foudre de guerre en escargot (OK, on exagère un peu). Le problème est mathématique ; pour générer un mot, le système doit lire l'intégralité du modèle. Si une partie est dans la VRAM (disons à 500 Go/s) et l'autre dans la RAM système (à 50 Go/s), l'ensemble du processus est bridé par le maillon le plus faible. C'est l'effet goulot d'étranglement ; la communication entre le processeur et la carte graphique est infiniment plus lente que la communication interne à la carte. Pour comprendre pourquoi la saturation arrive plus vite qu'on ne le croit, il faut regarder comment est grignotée votre VRAM. Elle est occupée par trois acteurs ; **Le poids du modèle** : c'est la base. Un modèle occupe une place donnée, dépendant principalement de son nombre de paramètres (par exemple 8B) et de sa quantification (par exemple 4-bits) et occupe un certain espace, fixe. **La fenêtre de contexte (KV Cache)** : c'est l'invité surprise. Plus votre discussion est longue ou plus vous fournissez de documents à l'IA, plus l'espace nécessaire pour stocker cette mémoire immédiate grandit. Sur de gros contextes, cela peut représenter plusieurs Go supplémentaires qui s'ajoutent au poids du modèle. **La taxe de l'OS** : on a tendance à l'oublier, mais votre VRAM n'est jamais totalement dédiée au LLM. Même sans jeu qui tourne, Windows, votre navigateur web avec ses 50 onglets et l'affichage de votre écran grignotent déjà leur part. Durant nos tests, on a relevé entre 1,5 et 1,8 Go d'occupé avant même de lancer le moindre LLM.
Le résultat concret ? Votre débit de tokens s'effondre. Vous passez d'une lecture fluide et confortable à une génération ralentie, voire saccadée, où chaque mot semble être une épreuve pour la machine.

> Poids du modèle, fenêtre de contexte et taxe de l'OS : un trio gourmand qui grignote votre mémoire vidéo et menace la fluidité de votre IA.

Tant que la partie GPU reste majoritaire, ou que votre RAM est rapide pour réduire le problème, ça va, mais comme vous pouvez le voir lancer un modèle en RAM seulement fait chuter certains modèles en dessous d'1 tok/s (ça devient inutilisable). Si on prend l'exemple de notre Config 5, celle avec la RTX 5080 16 Go, on remarque que Mistral Small 3 22B tourne à 61 tok/s. En 100 % GPU. Mais si on élargit la fenêtre de contexte et qu'on empiète sur la RAM (85 % GPU / 15 % CPU), on n'est plus qu'à 28,7 tok/s. Et comme vous pouvez le voir sur un gros modèle 70B, pourtant réduit à mort grâce à une quantification agressive (2 bits), on passe à 51 % GPU / 49 % RAM et le débit n'est plus que de 4,38 tok/s. On vous laisse imaginer ce même modèle sur un PC en DDR4 et avec seulement 8 Go de VRAM…

QUAND IL N'Y A PAS ASSEZ DE RAM NON PLUS

On précise aussi que même si l'offloading est une solution, encore faut-il qu'il y ait suffisamment de RAM ! Les LLM sont déjà assez lents en RAM pour ne pas du tout exploiter la mémoire virtuelle (sur le SSD) quand il n'y a plus d'espace. Ainsi, quand on essaie de lancer Command R+ 104B sur notre toujours sur le Desktop 5 (32 Go de RAM et 16 Go de VRAM), on obtient alors le message "Error: 500 Internal Server Error: model requires more system memory (46.9 GiB) than is available (44.5 GiB)". Il aurait sans doute été possible de récupérer les 2 Go manquant en fermant 100 % des programmes ouverts, mais nous avons préféré conserver un navigateur avec quelques onglets, Word et WhatsApp pour mieux se rapprocher d'un usage réel.



Finalement, c'est dans le cas des PC les plus modestes, sans dGPU, que le processeur joue un rôle plus important.

ET LE CPU ALORS, IL SERT À QUOI POUR UN LLM

Puisque nous avons établi que la VRAM est le réservoir et le débit de données son débit, on pourrait être tenté de voir le processeur central (CPU) comme un simple spectateur. Est-il utile d'investir dans un Core i9 ou un Ryzen 9 pour faire tourner son IA ? La réponse courte est ; pas forcément, mais il ne faut pas pour autant le négliger. Le rôle du CPU intervient principalement à deux moments clés ; **L'orchestration et le "Prompt Processing"** : avant que le GPU ne commence à générer du texte, il y a une phase d'ingestion. Le CPU s'occupe de la tokenisation (transformer vos mots en chiffres que l'IA comprend) et gère l'envoi des données vers le GPU. Un CPU trop lent peut créer un léger décalage (latence) entre le moment où vous appuyez sur "Entrée" et le début de la réponse. C'est le chef d'orchestre ; s'il est mou, les musiciens (les cœurs CUDA) attendent son signal. **Le filet de sécurité (l'offloading)** : comme nous l'avons vu, si le modèle déborde de la VRAM, c'est le CPU qui prend le relais pour calculer les couches restantes en s'appuyant sur la RAM système. C'est ici que sa puissance brute (nombre de cœurs et fréquence) joue un rôle. Mais attention ; même le meilleur CPU du monde reste bridé par la lenteur de la RAM comparée à la VRAM. Un gros CPU pourra certes traiter ces couches un peu plus vite qu'un petit, mais vous resterez de toute façon dans une expérience "dégradée".

Faut-il craquer pour un gros CPU ? Pour un PC dédié aux LLM, la règle d'or est l'équilibre. Un CPU de milieu de gamme moderne (type Ryzen 7 ou Core i5/i7) est largement suffisant pour ne pas brider une RTX 5080. Investir 300 € de plus dans un processeur haut de gamme n'augmentera pas votre nombre de tokens par seconde. Cet argent serait bien mieux investi dans une carte graphique disposant de plus de VRAM ou dans de la RAM système plus rapide (DDR5 à haute fréquence) pour limiter la casse en cas d'offloading. Ou tout simplement plus de RAM pour se permettre de lancer de plus gros modèles (ou avec des fenêtres de contextes plus importantes). Comme pour les GPU, nous avons surveillé l'occupation CPU avec le gestionnaire des tâches durant nos inférences. La majorité du temps, il ne fait pas grand-chose, surtout quand il attend la RAM. Mais il ne fait pas rien quand même. On remarque par exemple sur le laptop sans dGPU, le Samsung, que 4 des 8 cores sont à 100 %, autrement dit les 4 P-cores dont dispose ce dernier.

LE BUS PCIE JOUE AUSSI UN RÔLE

Il ne faudrait pas oublier les tuyaux qui relient ces composants au reste de la machine. C'est ici qu'entrent en scène le bus PCI Express (PCIe) et votre stockage. Le bus PCIe n'intervient pas dans la discussion entre le CPU et votre RAM (qui communiquent directement via le bus mémoire de la carte mère). Son rôle est ailleurs ;

il est l'autoroute qui relie la carte graphique au reste du système. C'est par lui que transite le modèle de l'IA lors de son chargement. Plus votre bus PCIe est moderne (Gen 4 ou Gen 5) et dispose de lignes (x16), plus le modèle sera transféré rapidement du stockage/RAM vers la VRAM. Mais ça compte aussi en cas d'offloading. Comme nous l'avons vu, si une partie du modèle déborde en RAM, le GPU doit récupérer les données manquantes à chaque token généré. Ces données doivent alors quitter la RAM, passer par le CPU, puis traverser le bus PCIe pour atteindre le GPU. Dans ce scénario, le PCIe devient un maillon de la chaîne de ralentissement, surtout si vous utilisez une carte bridée en "x8" (comme les RTX 4060) ou un vieux port PCIe Gen 3. Toutefois, nous n'avons pas pris le temps de réaliser des tests spécifiques pour illustrer ce point (merci de ne pas nous jeter des tomates !). Quant au SSD, son rôle est souvent surestimé par les débutants. Pour être direct ; il n'influence pas la vitesse de génération (les tokens par seconde). Une fois que l'IA a commencé à répondre, le modèle est en mémoire et le SSD ne travaille plus. Son impact se limite au temps de chargement. Un modèle de 30 Go se chargera logiquement plus vite avec un SSD PCIe 5.0 qui débite à 14 Go/s que sur un vieux SSD SATA bridé à 550 Mo/s ; selon cet exemple, on passe quand même d'un chargement en à peine 2 secondes dans le meilleur des cas à pas loin d'une minute ! Le SSD reprend toutefois une importance capitale pour le RAG (lorsque l'IA doit fouiller dans vos documents personnels) ; une indexation et une lecture rapide des fichiers sont alors indispensables pour ne pas faire attendre l'IA avant qu'elle ne commence à traiter votre question.

IA " RÉFLEXE" VS IA "CERVEAU"

Au fil de nos tests, une distinction majeure a fini par s'imposer, au-delà des simples chiffres de performance. On ne choisit pas son matériel uniquement pour « aller vite », mais pour le type d'intelligence que l'on souhaite solliciter. C'est ici qu'apparaît la fracture entre ce que nous appelons l'IA « réflexe » et l'IA « cerveau ».

L'IA « réflexe » : la vitesse avant tout Il s'agit des modèles dits légers, généralement compris entre 3B et 14B paramètres (comme Phi-4 Mini, Llama 8B ou Mistral Nemo). Ces modèles sont les champions de l'efficience. Sur une configuration moderne, ils débitent du texte bien plus vite que vous ne pouvez le lire (souvent plus de 100 tok/s).

- L'usage : c'est l'IA du quotidien. Elle est parfaite pour résumer un mail, corriger du code, ou répondre à des questions factuelles simples.
- L'impact hardware : elle se contente de peu. Avec 8 à 12 Go de VRAM, ces modèles tiennent entièrement en mémoire et offrent une interactivité instantanée. C'est le choix de la fluidité ; on interagit avec elle comme on poserait une question à un collègue au détour d'un couloir.

L'IA « cerveau » : la profondeur au prix de la patience. Ici, on change de dimension. On parle des modèles massifs, de 70B à 400B paramètres (comme Llama 3.1 70B ou les gros modèles MoE). Ces géants possèdent une finesse de raisonnement, une compréhension des nuances et une capacité de synthèse qu'aucun petit modèle ne peut égaler. Mais cette « sagesse » a un coût.

- L'usage : c'est l'IA des tâches complexes. Elle est là pour analyser des documents juridiques, concevoir une architecture logicielle de zéro ou philosopher sur des concepts abstraits.
- L'impact hardware : la VRAM, encore et toujours la VRAM. Pour faire tourner un modèle 70B sans qu'il s'effondre à 1 tok/s, il faut soit une débauche de GPU (plusieurs cartes en parallèle), soit une solution à mémoire unifiée (Apple M4 Ultra, Nvidia GB10). Même sur ces machines de guerre, le débit est souvent plus proche des 10 à 20 tok/s. On n'est plus dans le réflexe, mais dans la réflexion.

À RETENIR

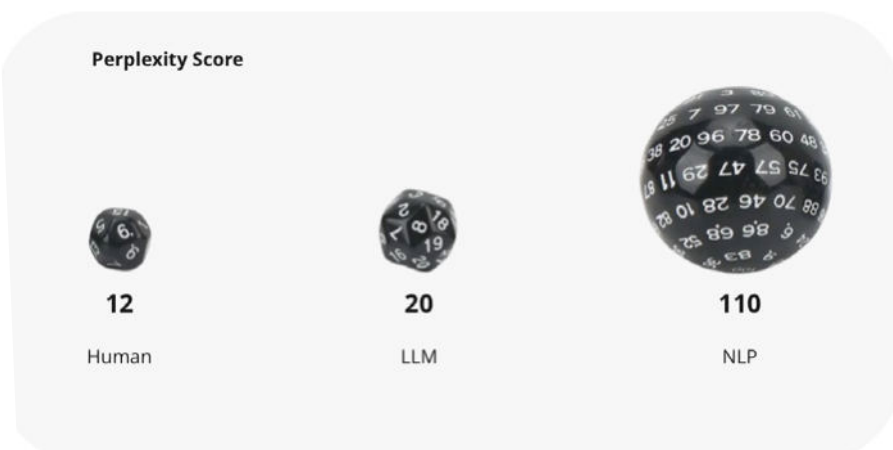
- Le débit (tok/s) : c'est la métrique de confort. Un débit de 10 à 15 tok/s est idéal pour une lecture fluide ; en dessous de 5 tok/s, l'expérience devient pénible.
- Bande passante mémoire : c'est la clé. La vitesse de l'IA dépend plus de la rapidité de la VRAM (Go/s) que de la puissance brute du GPU (TFLOPS).
- Capacité VRAM : elle définit la taille du "cerveau" utilisable. Si le modèle dépasse la mémoire vidéo, il déborde sur la RAM système (offloading), divisant la vitesse par 10 ou 20.
- KV Cache : Attention, la "mémoire à court terme" de votre discussion consomme aussi de la VRAM en plus du poids du modèle !

LE TEST DE VÉRITÉ : MESURER LA JUSTESSE ET LA PERPLEXITÉ

C'est bien beau d'être rapide, encore faut-il apporter des réponses pertinentes ! Mais juger de "l'intelligence" n'a rien d'évident. D'ailleurs, comment on fait ?

JUSTESSE : L'IA DIT-ELLE LA VÉRITÉ ?

La justesse — c'est-à-dire la capacité de l'IA à donner une réponse factuellement correcte, logique et bien structurée — ne dépend pas de la puissance de calcul brute de votre GPU. Que vous fassiez tourner Llama 3.1 8B sur une modeste ARC B580 ou sur une monstrueuse RTX 5090, si le fichier du modèle et sa quantification sont rigoureusement identiques, la réponse sera de qualité égale au mot près (à condition de fixer la "graine" de génération). La carte graphique est le lecteur, pas l'auteur. Autrement dit, un PC à 5 000 € n'est pas forcément plus intelligent qu'un autre à 1 000 €. Cependant, le matériel joue un rôle de « portier » indirect sur l'intelligence. Tout d'abord, c'est la quantité de VRAM qui décide si vous avez le droit de charger un modèle « intelligent » (comme un 70B ou 104B) ou si vous devez vous contenter d'un modèle plus modeste (8B). Et c'est aussi la VRAM qui détermine si vous pouvez charger ce modèle dans une version fidèle (Q4 voire Q8) ou si vous devez le compresser de façon extrême (Q2) pour qu'il rentre, au risque de dégrader la cohérence des réponses.



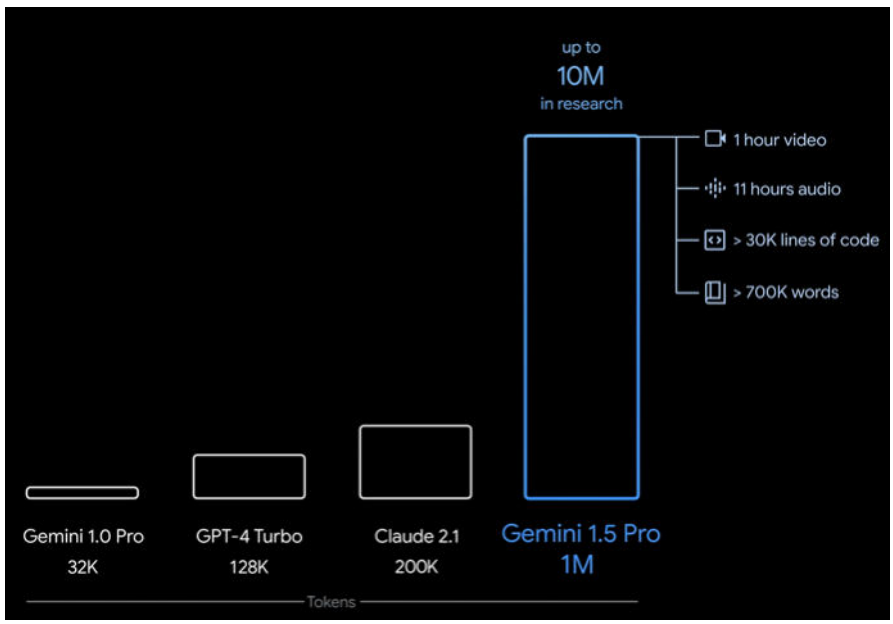
PERPLEXITÉ : L'INDICE DE CONFIANCE DE L'IA

Si la justesse est le verdict de l'utilisateur, la perplexité est le thermomètre interne des chercheurs en IA. Pour faire simple ; c'est la mesure de l'hésitation de la machine. En informatique, on cherche souvent le chiffre le plus haut (FPS, Score 3DMark), mais ici, c'est l'inverse ; plus la perplexité est basse, meilleur est le modèle. Imaginez que l'IA doive compléter la phrase ; « Le ciel est... ». Si elle est certaine à 99 % que le mot suivant est « bleu », sa perplexité est très basse. Elle est "lucide". Si elle commence à hésiter entre « bleu », « banane », « 42 » ou « demain », sa perplexité explose. C'est cet indice qui nous permet de savoir si un modèle a été trop compressé. Une perplexité qui grimpe, c'est le signe que l'IA commence à "douter" de ses propres connaissances, ce qui mène inévitablement aux hallucinations.

QUE SE PASSE-T-IL QUAND ON SATURE LA FENÊTRE DE CONTEXTE ?

Si le poids du modèle est la taille du « cerveau » de l'IA, la fenêtre de contexte est sa mémoire à court terme. C'est l'espace dans lequel elle stocke votre discussion actuelle et les documents fournis. En technique, on appelle cela le KV Cache. Et attention, cet espace est un mangeur de VRAM insatiable. Sur un outil comme Ollama, cette fenêtre est fixée par défaut à 4 096 tokens (4k), c'est peu, environ 3 000 mots. Pour vous donner une idée, sur Gemini vous avez une fenêtre de contexte d'1 million de tokens ! Que se passe-t-il quand vous dépassez cette limite ?

L'effet « Alzheimer » immédiat : par défaut, pour ne pas saturer votre mémoire vive, l'outil utilise une fenêtre glissante. Dès que vous entrez



Les modèles qu'on exécute localement permettent désormais d'utiliser des fenêtres de contexte aussi importantes que les géants du cloud d'il y a un deux ans à peine.

le 4 097ème token, l'IA oublie purement et simplement le tout début de la conversation. Si vous lui aviez donné des instructions cruciales au premier message, elles disparaissent. C'est une cause majeure d'hallucinations ; l'IA commence à inventer car elle n'a plus accès aux faits cités plus haut. Par exemple, quand nous avons demandé pour le test endurance de résumer le contenu de deux tests de boîtiers, le copier/coller de nos deux documents Word dépassant la taille du KV Cache, non seulement le modèle ne parlait pas du tout du premier des deux (déjà oublié), mais en plus au lieu de lister 10 points clés du second, on en avait que 6 ou 7. Ou encore l'IA se met à débiter du texte en boucle, à l'infini ! C'est rentré dans l'ordre en se contentant de résumer un seul boîtier (on voit dans Ollama tout en bas, en mode /set verbose, la taille du contexte à la fin d'une inférence), mais aussi avec les deux boîtiers après avoir doublé le KV Cache à 8K. Mais… **Le crash de performance** : c'est le piège. Si votre modèle occupe déjà presque toute votre VRAM, l'augmentation du contexte va pousser le KV Cache à déborder dans la RAM système. Résultat ? Votre débit de tokens s'effondre à mesure que la discussion s'allonge. On a vu des configurations passer de 50 tok/s en début de chat à seulement 5 tok/s après quelques échanges, simplement parce que la "mémoire de travail" avait fini par saturer le GPU.

Comme nous venons de l'évoquer, on peut on peut manuellement augmenter la fenêtre de contexte dans Ollama. Du moins dans la limite de ce que permet le modèle, jusqu'à 32k voire 128k sur la plupart des modèles âgés de moins de deux ans, et votre matériel (VRAM disponible). Pour ça, vous devez taper dans le pompt d'Ollama la commande /set parameter num_ctx 8000 ou toute autre valeur que vous souhaitez expérimenter. Sur certains gros modèles, passer de 4k à 32k de contexte peut demander plusieurs Go de VRAM supplémentaires. C'est pour cette raison que pour nos tests de vitesse, nous sommes restés sur le réglage par défaut (sauf pour le test endurance, une très légère augmentation à 4500 pour passer le pompt sans déborser sur les GPU de 16 Go). Cela garantit que nous mesurons la performance du "cerveau" sans être pollués par la gestion de sa "mémoire immédiate". Et bien sûr, la fenêtre étant tout juste taillée pour une réponse, inutile de vouloir continuer à converser avec sans tout réinitialiser (ou augmenter davantage le KV Cache).

LA RÈGLE D'OR

Pour de la production (travail, rédaction sans erreur), privilégiez la stabilité (modèle moyen en Q4 ou Q6). Pour de la recherche ou de la résolution de problèmes complexes, tentez le "Vieux Sage" (gros modèle en Q2), mais gardez un œil critique sur ses éventuelles sautes d'humeur.

NOS RÉSULTATS, VOTRE INTERPRÉTATION

Pour ceux qui souhaitent se rendre compte par eux-mêmes des aptitudes des modèles, nous avons exécuté les prompts de deux

> La fenêtre de contexte définit la limite de compréhension de l'IA : trop petite, elle oublie le début du livre ; trop grande, elle sature votre matériel et paralyse la génération.

de nos cinq tests sur l'ensemble des modèles pour les comparer entre eux. À noter que nous l'avons fait sur l'ensemble des machines, mais sachant que nous n'avons jamais saturé la fenêtre de contexte (volontairement), publier les résultats des 10 machines n'apporterait rien ; seule la vitesse de génération change. Voici un lien vers le Google Doc contenant les réponses (et nos résultats des tests ayant servis aux graphiques). D'un côté, sur le test « Sprint » (Llama 8B), la cohérence est quasi parfaite entre toutes les machines. Les dates de l'histoire des GPU sont respectées, les sauts technologiques sont bien identifiés. C'est l'IA "réflexe" dans toute sa splendeur ; elle est stable et fiable sur toutes les plateformes. De l'autre, le test « La Montagne » (Command R+ 104B) montre les premières fêlures. Sur les machines manquant de mémoire, nous avons dû forcer des quantifications très agressives. Le résultat ? Si le fond reste globalement correct, on perd en subtilité. L'IA devient plus répétitive, elle oublie parfois une partie de la consigne (comme l'analyse du paradoxe des pays du Golfe) ou produit des phrases à la syntaxe plus lourde. C'est la preuve par l'exemple ; la vitesse (tok/s) ne sert à rien si le cerveau est trop compressé pour réfléchir correctement. Ceci étant dit, ça reste très bon.

QUANTIFICATION, DU GAIN D'ESPACE À LA LOBOTOMIE

Nous l'avons vu plus haut, la quantification est le régime minceur de l'IA. Mais comme tout régime extrême, il y a un moment où l'on commence à perdre du muscle, puis de la jugeote. Si le Q4 (4-bits) est devenu le standard par défaut, l'offre s'est considérablement étoffée. Voici notre guide pour ne pas vous tromper de curseur. La Hiérarchie de l'intelligence :

Le Luxe (Q8_0 / Q6_K) : c'est le top. La perte de précision par rapport au modèle original est virtuellement indécelable. C'est le choix idéal pour les modèles 8B si vous avez de la

> Le Q4_K_M reste le meilleur rapport intelligence/poids, mais la nouvelle garde IQ3 permet désormais de faire tenir des modèles massifs dans des configurations grand public.

VRAM à revendre. Au-delà (FP16), c'est du gaspillage de mémoire.

Le "Sweet Spot" (Q4_K_M / Q5_K_M) : sans hésiter le meilleur rapport intelligence/poids. C'est le standard utilisé par la communauté. Si vous hésitez, choisissez toujours le Q4_K_M (le "M" pour Medium, une méthode qui préserve mieux les couches critiques).

La nouvelle garde (IQ3_M / IQ3_S) : les "IQ" (Importance Quantization) sont les nouveaux champions. Grâce à une matrice d'importance, un IQ3_M (3-bits) est souvent plus "intelligent" qu'un vieux Q4 tout en étant plus léger. C'est la solution miracle pour faire tenir un modèle de 30B dans 16 Go de VRAM.

La zone de danger (Q2_K / IQ2_M) : ici, l'IA commence à sérieusement "plisser les yeux". Elle perd sa superbe, bégaye et oublie les nuances. C'est le dilemme du "vieux sage". Vaut-il mieux un "petit" modèle (8B ou 22B) utilisé à pleine puissance (Q8), ou un "géant" (70B) compressé à l'extrême (Q2) ? Le 70B en Q2 (le vieux sage) : malgré la compression, il garde souvent une capacité de raisonnement et une culture générale supérieures. Il comprend mieux les nuances complexes, mais il est sujet à des bévues soudaines, des répétitions de mots ou des erreurs de grammaire. C'est un génie qui a des absences. Le 22B en Q6/Q8 (l'étudiant brillant) ; Il est d'une stabilité exemplaire. Il ne fera jamais de faute de français et suivra les consignes au pied de la lettre. En revanche, il sera plus vite limité sur des sujets très pointus ou des raisonnements logiques complexes.

La mauvaise idée (Q1.5 ou Q1) : il existe des quantifications expérimentales encore plus basses (1.5 bit ou 1 bit). Fuyez-les. À ce stade, ce n'est plus de la compression, c'est de la destruction. Le modèle perd la structure même du langage ; il se met à répéter le même mot à l'infini, génère des symboles aléatoires ou oublie simplement comment construire une phrase. C'est le stade de la lobotomie totale. Même avec un modèle de 400 milliards de paramètres, une fois réduit à 1.5 bit, devient inutile alors qu'il s'en sort encore pas mal en Q2.

À RETENIR

- Vitesse vs Intelligence : la carte graphique ne rend pas l'IA plus maligne, elle la rend seulement plus rapide. L'intelligence dépend uniquement du modèle choisi et de sa compression.
- Perplexité : c'est l'indice d'hésitation de l'IA. Plus elle est basse, plus le modèle est sûr de lui et cohérent. Une perplexité élevée est synonyme d'hallucinations.
- L'effet "Alzheimer" : par défaut, Ollama et LM Sutdio limitent le contexte à 4 096 tokens. Si vous dépassez le contexte, l'IA oublie le début de la discussion (consignes, faits) et commence à inventer.
- Le sweet spot de la quantification (Q4/Q5) : la quantification 4-bits ou 5-bits est le réglage idéal, offrant le meilleur compromis entre gain de place et fidélité du raisonnement.
- Le dilemme du "Vieux Sage" : il vaut mieux parfois un très gros modèle (70B) compressé en Q2 pour sa sagesse, qu'un petit modèle (8B) en Q8 qui sera certes parfait sur la forme, mais limité sur le fond.
- Évitez les quantifications sous les 2 bits (Q1.5, Q1) qui "lobotomisent" littéralement l'IA, la rendant incapable de structurer une phrase.

SOBRIÉTÉ ET EFFICIENCE : LA SUPER PUCE NVIDIA GB10 CHANGE LA DONNE

Pour ceux qui veulent aller plus loin dans leurs expérimentations des modèles de langage, notamment tester des modèles de 70B et plus sans trop sacrifier la justesse et la perplexité, d'autres solutions s'offrent à vous où l'on privilégie la mémoire vive plutôt que la mémoire vidéo. On parle plus précisément de mémoire unifiée, puisque des puces de mémoire standard, mais très rapide (avec des bus de mémoire plus larges et de la LPDDR5X à haute fréquence) servent à la fois de RAM et de mémoire vidéo à des SoC sophistiqués capables de charger des modèles massifs. Il ne faut certes pas s'attendre a des performances aussi bonnes qu'en utilisant de vrais GPU avec leur VRAM encore plus véloce, mais c'est tout à fait adapté aux développeurs IA, ingénieurs IA et même les passionnés qui veulent expérimenter. Avec 128 Go de mémoire unifiée, on peut envisager de faire tourner des modèles jusqu'à 128B sans trop perdre de précision, par exemple Llama 4 Scout 109B avec une précision de 4 bits) ou tout simplement exécuter des modèles de 32B en FP16. Ce genre de quantité de mémoire ouvre également la porte à des fenêtres de contexte bien plus importantes, par exemple 256 000 tokens qui s'avèrent nécessaires pour bosser sur de très longs documents ou des raisonnements en plusieurs étapes. Accessoirement, ces solutions sont beaucoup plus sobres, tant en énergie qu'en encombrement : on parle généralement de mini PC à peine plus imposants qu'un NUC !

APPLE, NVIDIA ET AMD, 3 APPROCHES DIFFÉRENTES

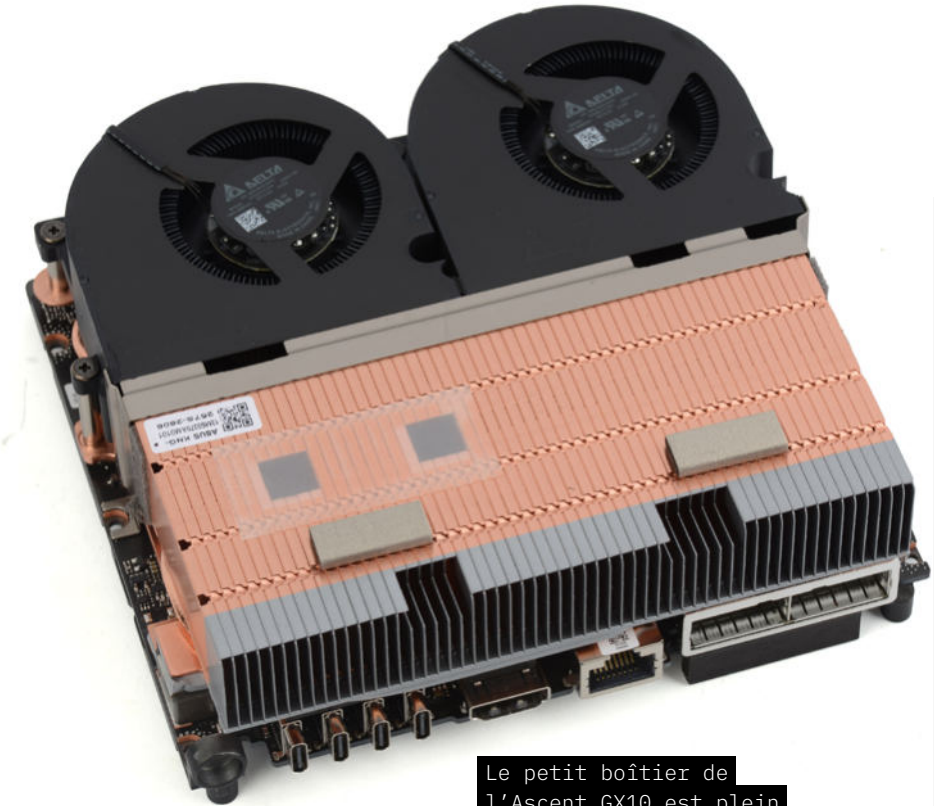
Sur ce terrain de la mémoire unifiée, trois philosophies s'affrontent pour répondre aux besoins des utilisateurs exigeants. Apple a ouvert la voie avec ses puces maison (Apple silicon), dont les déclinaisons Max proposent déjà 128 Go ultras rapides (400 Go/s sur M3 Max et 546 Go/s grâce au bus mémoire large de 512 bits), tandis que les versions Ultra peuvent grimper même au-delà et bénéficier d'une bande passante mémoire phénoménale allant jusqu'à 819 Go/s (le bus mémoire est encore doublé, on profite de 1 024 bits !) pour presque rivaliser avec ce dont sont capables les dernières cartes graphiques ! Mais, en outre, elles imposent de quitter l'écosystème Windows, ce qui représente un frein majeur pour de nombreux utilisateurs habitués à la flexibilité du PC. À l'opposé, NVIDIA propose avec sa plateforme GB10 une solution radicale pour le monde de la station de travail. Son point fort ne réside pas uniquement dans sa capacité mémoire de 128 Go, qui grâce à un bus de 256 bits et une vitesse

de 8533 MT/s offre une bande passante de 273 Go/s, mais dans la puissance brute de ses cœurs Tensor, taillés pour l'entraînement, et le fine-tuning. C'est la machine de l'ingénieur par excellence ; elle est un peu moins véloce qu'un Mac pour générer du texte (LLM), mais bien plus robuste pour traiter des calculs complexes. Mais elle aussi impose de s'éloigner de Windows pour utiliser Linux. Enfin, AMD s'engouffre dans une voie intermédiaire avec ses processeurs Ryzen AI Max (Strix Halo), le plus puissant étant le Ryzen AI Max+ 395, gérant grâce lui aussi jusqu'à 128 Go de mémoire unifié. Vous allez dire, à raison, que n'importe quel processeur AMD Ryzen ou Intel Core gère aujourd'hui davantage de mémoire (192 Go chez AMD, 256 Go chez Intel), mais ils se contentent de DDR5 classique et en dual channel. Ça veut dire, un bus large de 128 bits (2* 64 bits) et à des fréquences pas forcément très élevées, surtout quand on cumule 4 barrettes. En particulier chez AMD où il est déjà difficile d'atteindre 5600 MHz, ce qui donne une bande passante de 89,6 Go/s. Les Ryzen AI Max et AI Max+ exploitent quant à eux quatre canaux de LPDDR5x-8000, ce qui leur confère une bande passante de 256 Go/s ! Ne confondez pas non plus les processeurs Strix Halo avec les Strix Point, comme le Ryzen AI 9 HX 370 qui équipe l'Asus ROG Zephyrus G14 qu'on a testé il y a quelques mois et qui eux aussi se contente d'un bus mémoire de 128 bits. Ce n'est pas tout à fait aussi bien que Nvidia et bien loin de l'option d'Apple, mais c'est aussi la plus abordable et la seule qu'on peut envisager sous Windows et dont on peut même profiter sous la forme d'un vrai PC de bureau assemblé.

Framework est le seul constructeur, à notre connaissance, à proposer un tour basée sur le Ryzen AI Max+ 395 (2 769 € en version 128 Go, 1 879 € en 64 Go).

COMBIEN ÇA COÛTE ?

Ces solutions ne sont globalement pas données. Il faut compter un ticket d'entrée proche de 3 000 €, pour une solution Strix Halo complète tandis que les solutions mini PC à base de Nvidia GB10 sont entre 4 000 et 5 000 €. On ne parle même pas des Mac, surtout les modèles avec 256 voire 512 Go. Un Mac Studio en M4 Max avec 128 Go est facturé 4 424 €. Un Mac Studio en M3 Ultra avec 512 Go atteint 11 674 €. Bon, dans tout ça vous vous dites peut-être qu'on a parlé là que de machines vendues au moins 3 000 € et qu'on est bien gentils. Mais comparez plutôt ! Le moindre PC équipé d'une RTX 5090 coûte lui aussi plus de 4 000 €, sachant que la carte graphique seule vaut déjà pas loin de 3 500 € de nos jours. Et même en étant beaucoup plus sage, avec une RTX 5070 Ti par exemple, on atteint vite entre 2 000 et 3 000 €, sachant que cette machine sera globalement moins utile pour les LLM qu'une machine à mémoire unifiée, quoi que plus rapide dans certaines circonstances. Et surtout plus polyvalente et notamment bien meilleure pour jouer. Il ne faut pas oublier qu'on est entré depuis quelques mois dans une période de crise et d'incertitudes quant aux prix et à la disponibilité de la mémoire et que ça n'arrange pas les choses, c'est le moins qu'on puisse dire.



Le petit boîtier de l'Ascent GX10 est plein à craquer !

> Crise des composants et pénurie de mémoire : dans un contexte de prix incertains, la construction d'une station IA devient un véritable défi stratégique et financier.

DES LAPTOPS POUR LES INGÉNIEURS IA

Peut-on développer sur les LLM avec un PC portable ? Oh que oui ! Et on ne parle pas de lâcher 6 500 € dans le MSI Titan 18 HX AI et sa RTX 5090 qui n'a que 16 Go de VRAM. Les solutions à mémoire unifiée que nous venons d'évoquer existent pour partie en version mobile ! En effet, s'il n'y a pas encore de puce Nvidia GB10 à emporter, mais il en existe en AMD Strix Halo. Mais attention, il ne suffit pas d'avoir un Ryzen AI Max+ 395 pour avoir 128 Go de RAM. Par exemple, l'Asus ROG Flow Z13 sorti l'an dernier à 2800 € et actuellement en promo à 2000 € n'a que 32 Go et cette capacité n'est pas extensible. Pour bénéficier de 128 Go, le maximum pour cette plateforme, les choix sont finalement assez restreints avec le Zbook Ultra G1a de chez HP (un peu plus de 4000 € en 128 Go, ouch) et les futurs ProArt PX13 et PX16 d'Asus qui débarqueront d'ici peu. On sait déjà qu'un modèle 13" avec 128 Go de RAM a été annoncé pour l'Angleterre (2 999 £). Chez Apple, les MacBook Pro existent en M4 Pro et M4 Max, tant en 14" qu'en 16". Un 14" en M4 Max (bande passante de 410 Go/s) avec 128 Go de mémoire unifiée vous coûtera quand même 5 674 €, avec seulement 1 To de stockage. Ça pique, mais en prenant un peu de recul c'est largement plus performant pour essayer et développer sur des modèles d'IA tout en coûtant moins cher que le MSI Titan qui nous a servi pour les tests. Et moins bruyant. Et largement plus autonome.

ESSAI GB10 : ASUS ASCENT GX10

Pour repousser un peu les limites, on a demandé à nos amis de chez Asus s'ils pouvaient nous prêter leur dernier joujou, l'Ascent GX10. Si la plupart de nos configurations de test sont des assemblages classiques, l'Asus Ascent GX10 joue dans une autre cour. Ce n'est pas un PC de jeu qui fait de l'IA par accident, c'est une station de travail IA miniaturisée. Armé de la puce Nvidia GB10 dont on a parlé un peu plus haut, ce petit boîtier de moins de 2 litres est tout mignon. À l'usage, l'Ascent GX10 offre une expérience qui rappelle celle d'un Mac Studio, mais sous un environnement ouvert. En l'occurrence, il tourne sous DGX OS de Nvidia, une distribution dérivée de Ubuntu avec quelques outils déjà en place pour exploiter l'intelligence artificielle. Grâce à ses 128 Go de mémoire unifiée, nous avons pu charger le modèle Command R+ (104B) en Q4_K_M sans la moindre hésitation. Là où une RTX 5090 (24 ou 32 Go) commence à montrer ses limites et doit déléguer une partie du travail à la RAM système (faisant s'effondrer les performances), le GX10 maintient une vitesse de croisière constante, même si elle est plutôt moyenne. Et oui, il a beau avoir de la LPDDR5x ultra rapide, ça ne reste jamais que 273 Go/s. La page Web du produit indique qu'il est capable d'exécuter des modèles jusqu'à 200B, mais c'est quand même très optimiste. Enfin pas plus que notre

volonté de rentrer un modèle 70B dans une RTX 5090 de 32 Go nous direz-vous. En pratique, et sans lâcher la quantification Q4, on a galéré à lancer Mixtral 8x22B, un modèle maous costaud de 141B qui occupait littéralement les 128 Go de RAM. Bien sûr, ce n'est pas le genre de machine qu'on achète sans réfléchir. Non seulement elle coûte 4 000 €, mais elle n'est pas upgradable et surtout d'usage vraiment très orienté IA locale.

DES MACHINES PLUS COMPACTES ET EFFICIENTES

L'un des chocs les plus brutaux quand on passe d'une tour de bureau survitaminée à une machine comme l'Asus Ascent GX10, c'est le rapport entre le volume et l'intelligence.

Là où notre Desktop 6 (le duo Ryzen 9950X et RTX 5090) nécessite un boîtier de 60 litres, une alimentation de 1200W et une armée de ventilateurs pour ne pas fondre, ces nouvelles solutions unifiées tiennent dans un format "NUC" d'à peine 1,5 kg. On ne parle plus de "tour", mais d'un petit cube que l'on glisse entre deux livres ou que l'on emporte en déplacement. Et quelle efficience ! L'intégration du CPU, du GPU et de la RAM sur le même package (comme sur la puce GB10) limite les pertes et on remarque aussi que le CPU ARM est plus efficace que n'importe quelle puce x86. Et ce qui permet de s'affranchir des lenteurs de traduction entre les architectures quand on travaille sous Linux avec les outils Nvidia (puisque leurs plus gros serveurs Grace-Hopper sont aussi en ARM). Alors que le PC avec la

Nvidia prévoit qu'on puisse combiner la puissance de deux puces GB10 via son interface maison NVLink, plus efficace que le bus PCIe.

```
tolivaux@gx10-e2ee: ~
Toutes les 1,0s: nvidia-smi; free -h                    gx10-e2ee: Wed Jan 28 11:45:49 2026

Wed Jan 28 11:45:49 2026
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.126.09             Driver Version: 580.126.09     CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GB10                     On |   0000000F:01:00.0  On |                  N/A |
| N/A   45C    P0              12W /  N/A | Not Supported          |      3%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A            3916      G   /usr/lib/xorg/Xorg                      237MiB |
|    0   N/A  N/A            4091      G   /usr/bin/gnome-shell                    261MiB |
|    0   N/A  N/A            4847      G   .../7715/usr/lib/firefox/firefox        270MiB |
|    0   N/A  N/A           12385      C   /usr/local/bin/ollama                 58266MiB |
+-----------------------------------------------------------------------------------------+
              total        utilisé      libre     partagé tamp/cache   disponible
Mem:          119Gi         63Gi       1,5Gi       112Mi        55Gi        56Gi
Échange:       15Gi        880Mi        15Gi
```

Vous pensiez être larges avec 128 Go ? Ça se remplit pourtant si vite…

> L'architecture ARM et la mémoire partagée : le futur de l'IA "Desktop" se dessine avec des machines sobres, discrètes et dotées de réservoirs de RAM colossaux.

RTX 5090 grimpe à plus de 600 W en pleine inférence, on n'a jamais dépassé 180 W avec l'Ascent GX10, la plupart des modèles se contentant de tirer dans les 140 W. Même les plus gros qu'un PC surpuissant n'est pas capable de lancer (ou de façon très lente) ! Pour un utilisateur qui fait tourner des modèles de langage toute la journée, la différence sur la facture d'électricité est colossale, mais c'est surtout le confort sonore qui change la vie. On peut enfin faire réfléchir son IA dans le silence, sans avoir l'impression qu'un avion de chasse décolle sous son bureau. C'est sans doute là le futur de l'IA "Desktop" ; des machines discrètes, sobres, mais dotées d'un réservoir de mémoire (128 Go) qui ferait pâlir n'importe quel PC de jeu traditionnel. Mais elles s'adressent avant tout aux ingénieurs IA. Nos PC suffisent à expérimenter.

À RETENIR

- Mémoire Unifiée : une architecture où la RAM est partagée entre le processeur et la partie graphique, permettant de charger des modèles géants (70B à 141B) inaccessibles aux cartes graphiques classiques.
- Apple (Série M Max/Ultra) : le pionnier du secteur. Offre les bandes passantes les plus élevées (jusqu'à 819 Go/s), mais impose l'écosystème macOS.
- Nvidia GB10 (Asus GX10) : une solution miniaturisée (format NUC) dédiée à l'IA sous Linux, offrant 128 Go de mémoire pour un usage "ingénieur" très sobre (moins de 180W).
- AMD Strix Halo (Ryzen AI Max et AI Max+) : la voie du compromis. Permet de bénéficier de 128 Go de mémoire unifiée tout en restant sous Windows, avec des performances supérieures à la DDR5 classique.
- Capacité vs vitesse : ces solutions sont moins rapides en débit pur (tok/s) qu'une RTX 5090, mais elles permettent d'exécuter des modèles beaucoup plus complexes sans erreurs.
- Sobriété : ces puces consomment 3 à 4 fois moins d'énergie qu'une tour PC haut de gamme pour une tâche d'IA équivalente.
- Idéal pour les développeurs, les chercheurs ou ceux qui travaillent sur de très longs documents nécessitant une énorme fenêtre de contexte (256k tokens).

LE MOT DE LA FIN : VOTRE PC, VOTRE IA

Si vous êtes arrivé jusqu'ici, le fonctionnement des LLM devrait être beaucoup plus clair pour vous. Au début de ce dossier, l'IA pouvait sembler être une magie lointaine, réservée aux serveurs de la Silicon Valley. Vous savez désormais qu'il s'agit d'une mécanique de précision faite de poids, de matrices et de probabilités, que l'on peut démonter et remonter chez soi. Vous ne vous contentez plus d'utiliser l'intelligence artificielle, vous l'administrez. Voilà typiquement un paragraphe de texte qu'une IA pourrait générer !

LA VICTOIRE DE LA SOUVERAINETÉ

Au-delà de la performance technique, faire tourner un LLM en local est une déclaration d'indépendance. C'est la garantie absolue que vos données — qu'il s'agisse de bilans financiers, de codes sources propriétaires ou de journaux intimes — ne quitteront jamais le châssis de votre PC. C'est aussi l'assurance de ne jamais dépendre du bon vouloir d'une API, d'une augmentation tarifaire ou d'un changement de politique de modération. Sur votre machine, le modèle est à vous. Il ne changera pas demain matin, sauf si vous le décidez.

MATÉRIEL ; VERS UNE NOUVELLE ÈRE D'EFFICIENCE

Ce voyage au cœur du hardware nous a aussi enseigné une leçon cruciale sur l'évolution de nos machines. Nous avons vu que la puissance brute ne fait pas tout. Si les cartes graphiques classiques (nos fidèles GeForce) restent les reines de la vitesse, la consommation gargantuesque d'une RTX 5090 pour faire tourner un gros modèle pose question. Comme nous l'avons vu dans la partie 7, l'avenir de l'IA "Desktop" pourrait bien ressembler à l'Ascent GX10 ; des machines sobres, silencieuses, privilégiant une énorme quantité de mémoire unifiée de bonne fréquence qui offrent plus de possibilités qu'un PC classique avec son GPU, même si c'est un peu moins performant. L'IA n'a pas besoin de faire du bruit pour être intelligente ; elle a besoin d'espace pour "penser". Certes, on joue sans se demander si notre GPU est gourmand en énergie, mais c'est absolument nécessaire pour faire tourner les jeux beaux et fluides. Pour l'IA, et en particulier les LLM, nous avons bien compris que le nerf de la guerre c'est la mémoire. Cette fameuse mémoire que tous les géants du cloud spécialisés dans l'IA achètent à tour de bras et qui explique la fameuse crise dans laquelle nous sommes empêtrés depuis quelques mois d'ailleurs.

Les stations à mémoire unifiée consomment beaucoup moins qu'une configuration gamer pour jouer avec des LLM.

QUE PEUT FAIRE VOTRE PC ?

Pour conclure concrètement, voici ce que vous pouvez attendre de votre matériel selon votre profil. Ne cherchez pas à tout faire avec la même configuration ; à chaque usage son goulot d'étranglement.

Le créatif (rédaction, traduction, roleplay, brainstorming) : c'est, contrairement à ce qu'on pourrait croire, le terrain de jeu le plus accessible. L'intelligence "conversationnelle" ne demande pas une mémoire titanesque. Une carte graphique de 8 à 12 Go de VRAM s'en sort déjà très bien et vous pourrez même essayer sur un GPU avec seulement 6 Go en visant des modèles 7B/8B comme Llama 3.1 8B ou Mistral Nemo. Ce qui est possible ; générer du texte à plus de 50 mots/seconde (plus vite que vous ne pouvez lire). Vous pouvez traduire des emails, réécrire des chapitres entiers ou discuter avec un personnage fictif sans aucune latence. La limite ; La mémoire à court terme. Avec 8 Go, votre IA "oubliera" le début de la conversation après environ 4 000 à 6 000 mots. Suffisant pour un article, trop court pour un roman.

L'analyste (résumé de documents, RAG, analyse de PDF) : ici, le piège est de privilégier l'intelligence du modèle au détriment de la mémoire de contexte. Un GPU avec 16 Go de VRAM (RTX 4070 Ti Super, RTX 5070 Ti) est un MINIMUM. Avec ce genre de carte, restez sur un modèle pas trop gros, du 8B par exemple. Pourquoi ? Car un "petit" modèle ne prend que 4 à 6 Go de VRAM, laissant 10 à 18 Go libres pour charger le contexte (vos documents). Ce qui est possible ; Ingérer et interroger un PDF de 100 pages ou le transcript complet d'une réunion

> La leçon de ce dossier est claire : pour les LLM, l'avenir appartient aux machines capables de concilier une énorme quantité de mémoire et une efficience énergétique radicale.

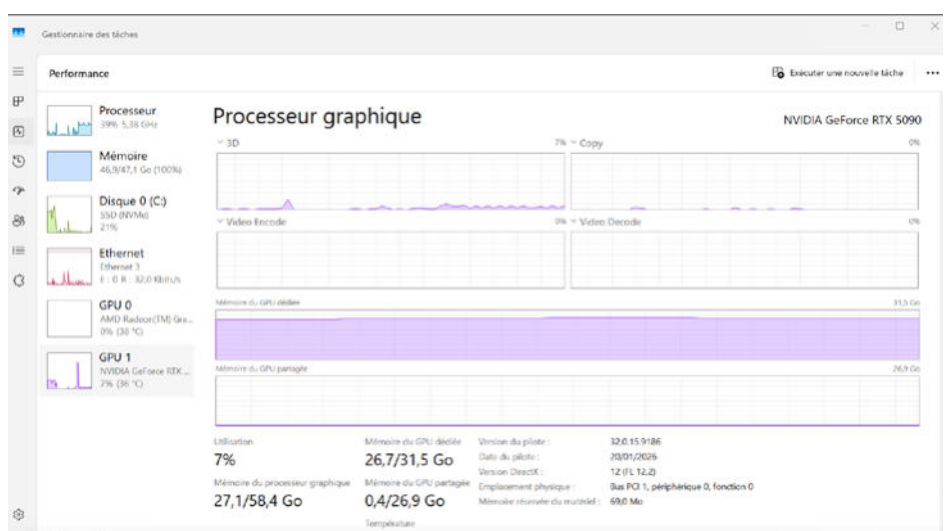
Si vous n'avez pas un GPU avec au moins 6 ou 8 Go, n'espérez pas faire grand-chose avec un LLM en local.

de 3 h sans que l'IA ne sature. Ce qui est impossible ; Faire la même chose avec un modèle "moyen" (22 ou 24B) sur une carte de 16 Go. Le modèle prendrait toute la place, ne laissant aucune mémoire pour lire votre fichier, il faudrait alors un GPU de 24 ou 32 Go.

L'expert (code complexe, mathématiques, raisonnement logique) : là, ça fait mal. Les petits modèles échouent souvent sur la logique complexe. Il faut du "lourd" (70B+, et encore on est très loin des capacités de l'IA cloud), et c'est là que le PC grand public souffre. C'est le territoire des configurations extrêmes ; dual GPU (2x RTX 4090/5090 avec NVLink pour 48/64 Go VRAM) ou les nouvelles puces à mémoire unifiée (Nvidia GB10 / Mac Studio avec 64 Go+). Le modèle idéal ; Llama 3.3 70B, Qwen 2.5 72B ou Command R+. Ce qui est possible ? Obtenir une qualité de réponse équivalente à GPT-4 ou presque ! Déboguer du code complexe ou résoudre des problèmes logiques ardus. La solution "système D" ; si vous n'avez pas ce matériel, ce qui se comprend aisément, vous pouvez faire tourner ces modèles sur votre processeur (CPU) et votre RAM système (à condition d'avoir au moins 48 voire 64 Go de DDR5). Ça fonctionnera, mais à la vitesse d'un escargot (1 à 3 mots/seconde).

L'APPEL À L'EXPÉRIMENTATION

Maintenant, c'est à vous de jouer. N'ayez pas peur de télécharger des modèles, de saturer votre VRAM, de voir votre interface planter, et de recommencer. Le monde de l'IA Open Source bouge à une vitesse vertigineuse. Ce qui est vrai aujourd'hui sera (sans doute) obsolète dans six mois, mais les fondamentaux que vous avez acquis grâce à ce dossier — la quantification, le contexte, l'offloading — resteront. Votre PC est désormais prêt. Il n'attend plus que vos prompts pour s'éveiller. Et après ? La prochaine étape sera sans doute de lui donner des yeux (vision), des oreilles (voice-to-text) et de l'autonomie (agents). Mais ça, c'est une autre histoire, pour un prochain numéro.



Comme nous, amusez-vous a explorer les limites de votre matériel !

UBERHARDWARE
.COM

FOR THOSE WHO DARE
START
REPUBLIC OF GAMERS
PUSH
ZOTAC GAMING

60 HAVN BF 360 Flow

70 EK Nucleus CR360 Lux DRGB

76 NZXT C750 Gold Core

82 MSI Titan 18 HX AI A2XW

94 Ugreen NASync DH2300

Tests 12

BOÎTIER

HAVN BF 360 Flow

SUPERBES FINITIONS ET VENTILATION AU TOP

Entre son design raffiné, ses matériaux haut de gamme et sa conception innovante, le BF 360 Flow confirme le savoir-faire de HAVN, un an après la sortie du superbe HS 420.

Prix constaté : 190 €

→ par Thomas Olivaux

C'est toujours la même chose. À chaque fois qu'on reçoit un nouveau boîtier, on est super emballés et on a l'impression de vous proposer une nouvelle pépite. Est-ce qu'on choisit particulièrement bien les modèles qu'on teste ? Est-ce que les constructeurs ne produisent plus que des super boîtiers ? Sommes-nous simplement un peu trop à fond ? Allez savoir. En tout cas, on est heureux de vous faire découvrir aujourd'hui le HAVN BF 360 Flow, un boîtier haut de gamme tant dans sa conception que son finish et qui ne coûte pas une fortune pour autant. Comptez un poil moins de 190 € pour vous offrir cette beauté ATX, disponible soit en noir, soit en blanc. À savoir qu'il existe aussi un BF 360 « tout court », sans ventilateurs, à 160 €.

BIENVENUE HAVN !

Les lettres du logo HAVN sont en alliage de zinc encastré dans une façade plastique au look très minéral.

Vous ne connaissez peut-être pas HAVN, et c'est bien normal. Cette jeune marque est née fin 2024, en même temps qu'Uber Hardware, et le BF 360 n'est jamais que leur second boîtier. Il compète l'incroyable HS 420, un boîtier aquarium superbe et original sorti il y a un peu plus d'un an et dont nous vous proposerons également le test si vous le souhaitez. HAVN est une entreprise constituée de vétérans du hardware PC et en particulier d'anciens de chez Fractal Design.

FICHE TECHNIQUE

Marque : HAVN

Modèle : BF 360 Flow

Type : tour moyenne à double compartiment

Support carte mère : E-ATX, ATX, micro-ATX, mini ITX

Compatibilité BTF : non

Ports d'extension : 7*

Matériaux : Acier + plastique + verre trempé

Emplacements 2,5" : 5* (dont 4* partagés avec les emplacements 3,5")

Emplacements 3,5" : 2* (partagés avec les emplacements 2,5")

Emplacements pour ventilateurs : 2* 180 mm en façade, 3* 120 ou 2* 140 mm ou 2 * 180 mm en haut, 120/140 mm à l'arrière

Ventilateurs fournis : 2* 180 mm (PWM) + 1* 140 mm (PWM)

Emplacements pour radiateurs : jusqu'à 360 mm en haut

Hauteur max ventirad CPU : 19,5 cm

Longueur max GPU : 41 cm

Prises en façade : USB-C, 2* USB-A, jack 3,5 mm combo

Dimensions : 51,5 x 25,4 x 52,3 cm (68,4 litres)

Poids : 14,79 kg

Versions : noir (Black), blanche (White)

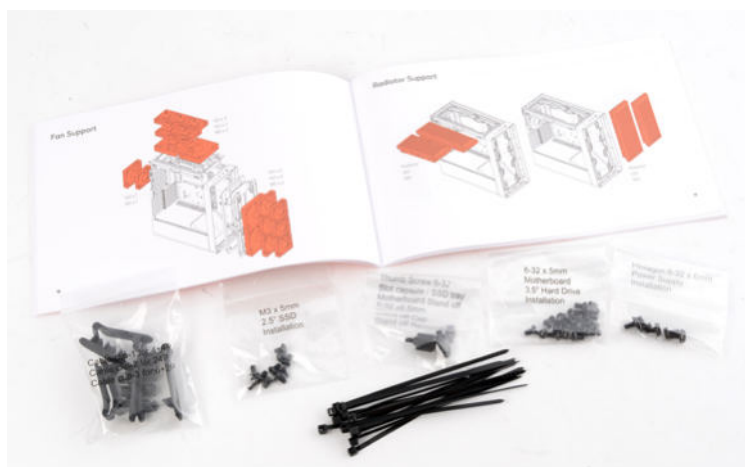
Ce n'est pas qu'un « pousseur de cartons » comme on dit parfois dans le jargon, ils conçoivent intégralement leurs boîtiers. Et s'ils ont les moyens de le faire, c'est parce que HAVN est une marque du groupe PGG (Pro Gamers Group), la maison mère qui possède également Casking (le célèbre revendeur allemand), Overclockers UK ou encore la marque de sièges Noblechairs.

C'EST DU LOURD

Le BF 360 Flow est à la fois très classique dans son format, c'est une tour assez imposante, et original sur de multiples points de design sur lesquels nous allons revenir. Mesurant 25,4 cm de large par 51,5 cm de profondeur et 52,3 cm de haut, ce n'est pas le plus imposant boîtier jamais produit (tant s'en faut) mais il est quand même imposant (68,4 litres). À titre de comparaison, le bequiet! Light Base 500 testé dans UH #11, pourtant beaucoup plus large (30,5 cm), est 17 % moins volumineux. Ça se remarque aussi sur le poids de la bête. Tandis que le Light Base 500, un boîtier pourtant robuste et livré avec quatre ventilateurs, se contente de 10,91 kg, le BF 360 Flow pèse 14,79 kg ! C'est plus pour le boîtier seul que pas mal d'unités centrales complètes !

LE BOÎTIER AUX 7 MATÉRIAUX

Ce poids ne vient pas uniquement des dimensions imposantes, c'est



aussi une conséquence directe de la qualité de fabrication. La plupart du temps, les fiches techniques des boîtiers mentionnent un métal (acier le plus souvent, parfois en aluminium), du verre trempé quand il y a une vitre latérale et du plastique, en général de l'ABS. Sur le site Web HAVN, on constate une liste de sept matériaux pour ce boîtier : SGCC sheet metal, tempered glass, ABS, Nylon, Silicon, Zinc alloy, NdFeB magnets. À vos souhaits. Le SGCC sheet metal, c'est l'ensemble des tôles qui forment le châssis. Sauf qu'au lieu d'être de l'acier bas de gamme (SPCC, acier laminé à froid « brut »), ni même de l'acier de qualité (SECC, acier recouvert d'une fine couche de zinc déposée par électrolyse) comme on trouve dans la majorité des boîtiers des grandes marques (bequiet!, Corsair, NZXT...), HAVN utile de l'acier encore plus robuste ! L'acier SGCC (Steel, Galvanized, Cold-rolled, Coated) est en fait une version musclée du SECC, sauf que la couche de protection

> L'acier SGCC, galvanisé par immersion, offre une rigidité et une résistance à la corrosion bien supérieures aux standards SPCC ou SECC du marché.

n'est pas obtenue par un dépôt électrolytique mais par trempage dans de grands bains, formant ainsi une couche plus épaisse et résistante, ce qui le rend plus rigide, plus résistant aux rayures profondes et à la corrosion.

Pour ne rien gâcher, c'est de l'acier de 1 mm d'épaisseur là où les boîtiers moins haut de gamme se contentent généralement de 0,8 mm, voire moins pour les tours à petit prix. Du moins on suppose que ça fait 1 mm d'épaisseur, car on mesure avec notre pied à coulisse autour de 1,2 mm desquels il faut retirer l'épaisseur de la peinture. Là aussi, signalons le choix de la qualité fait par HAVN, c'est de la peinture époxy, qu'on appelle également powder coating en référence à son procédé d'application (de la poudre « aimanté » à l'acier puis cuite au four). Ce type de peinture résiste mieux aux rayures et aux chocs que la peinture liquide classique.

ON RESTE SAGE POUR LA PORTE EN VERRE ET LES FILTRES À POUSSIÈRE

Le paragraphe précédent ne traite que d'un seul des sept matériaux cités ! Il y a bien sûr le verre trempé qui forme la vitre latérale. Nous mesurons ce dernier, qui est transparent, à 3,5 mm d'épaisseur, ce qui est assez standard. L'ABS, c'est le plastique utilisé pour la façade. Mais là aussi HAVN innove. Alors que le plastique est rarement associé à des éléments qualitatifs, le choix des pigments et même la texture de cette façade rappellent la pierre. Ça donne un finish minéral très réussi, tout en restant léger, qui tranche avec le noir du reste de la tour. Vient ensuite le nylon, qui constitue les grilles, les filtres à poussière ; une grille recouvre l'ensemble de la face intérieure de la façade avant et du capot supérieur, sans oublier un filtre à poussière placé sous la prise d'air de l'alimentation tout en bas.

CES PETITS DÉTAILS QUI FONT PLAISIR

On remarque le silicone sur certaines zones de préhension, invisibles quand le boîtier est fermé. Par exemple en deux endroits sur la porte vitrée,

Les capots avant et du dessus embarquent tous deux un filtre à poussière.

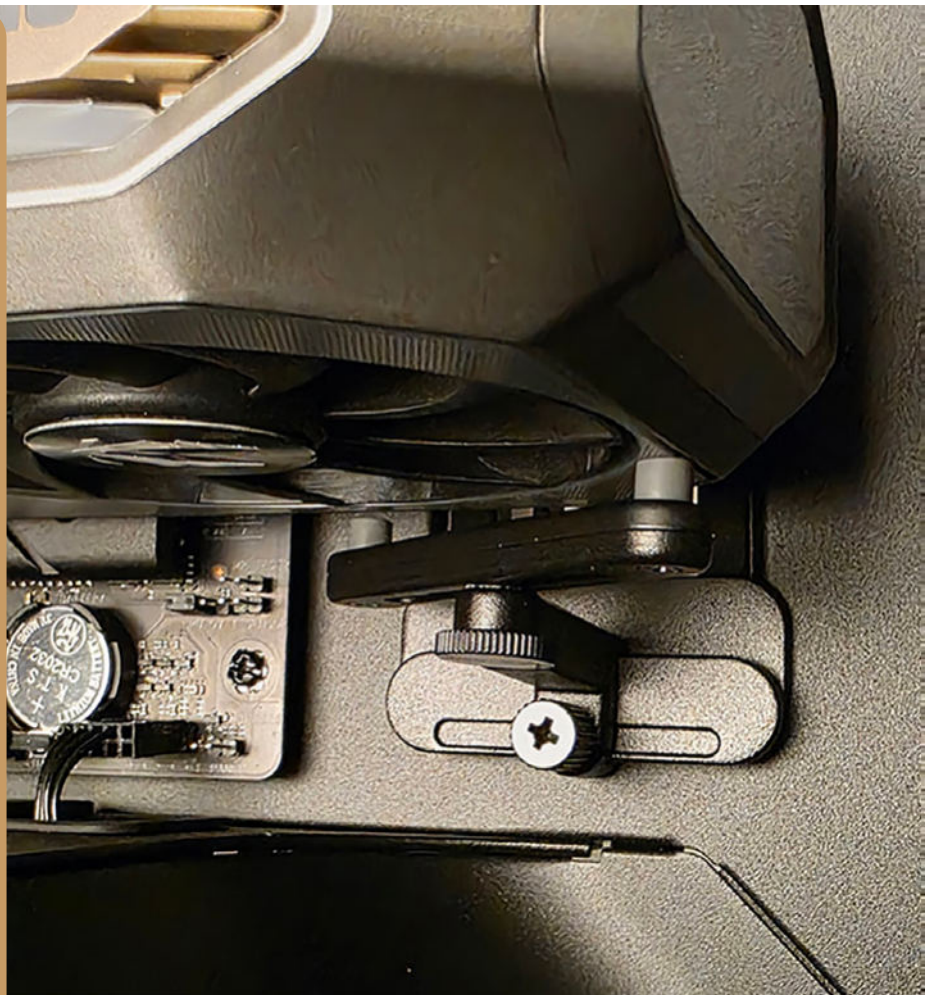
> HAVN révolutionne l'ABS : une façade plastique imitant la pierre, verre trempé à 3,5 mm et nylon filtrant – sept matériaux pour un boîtier premium.

correspondant aux encoches par lesquelles on glisse les doigts pour attraper ce panneau, ceci afin d'éviter qu'il ne vous glisse des mains. Sans oublier les patins qui réduisent la propagation des vibrations des deux supports de disques durs. Les deux derniers matériaux cités sont également très discrets. L'alliage de zinc correspond, d'après nous, au logo HAVN écrit en quatre lettres majuscules tout en bas de la façade. Celui-ci est pourtant affleurant, mais au lieu de se contenter de lettres en plastique recouvert d'un finish brillant et collées, comme on voit régulièrement, HAVN a réellement moulé des lettres en alliage de zinc qu'il a encastré dans la façade. Rien à dire, c'est beau. Enfin les aimants, ils sont là pour maintenir la façade en place sur le châssis. Signalons aussi la petite lanière tressée située en haut du boîtier, tout à l'arrière, pour le porter. La matière de celle-ci ne semble pas indiquée, à moins qu'il s'agisse là aussi de nylon.

SUPPORT GPU ET CABLE MANAGEMENT AVANCÉ

Vous aurez compris que l'esthétique et la finition du BF 360 nous plaisent beaucoup, c'est un doux mélange de sobriété (d'ailleurs il n'y a pas la moindre LED !) et d'originalité, même s'il faut s'en approcher pour mieux s'en rendre compte et que, comme toujours, les formes d'un boîtier plaisent ou non selon les goûts de chacun.

HAVN a aussi travaillé l'intérieur de son boîtier. Du côté visible, on remarque des passages de câbles assez originaux, ce sont des lanières en silicone qui bouchent partiellement les ouvertures par lesquelles ont passe les différents câblages, on aime ou on n'aime pas. On remarque aussi que HAVN a intégré un support GPU totalement réglable qui viendra soutenir votre puissante et lourde GeForce afin qu'elle reste bien droite et ne force pas trop sur le port PCI-Express de la carte mère. Au dos, on retrouve des chemins de câbles avec des bandes de type velcro pour tout maintenir en place, un peu comme chez NZXT. HAVN est assez sympa pour nous donner des recommandations écrites, mais



avec une faute d'orthographe sur le sticker (« Recommendned » ou lieu de « Recommended ») reprise à plusieurs endroits. C'est ballot. Autant le fond de panier est capable de recevoir des cartes mères imposantes, jusqu'à 27,7 cm de large en E-ATX, autant il n'y a pas de découpes pour permettre l'usage des cartes à connecteurs arrière (type BTF). Dommage. C'est un boîtier qualifié de semi-ouvert car, après avoir retiré le couvercle supérieur, on peut aussi retirer le support pour ventilateurs et radiateur afin de bénéficier d'un meilleur accès aux connectiques situées en haut de la carte mère, notamment. Le montage est globalement très aisé, notre seul bémol concerne les deux vis qui tiennent chacune des portes (une vis par porte) et les deux vis du capot supérieur qui sont logées dans un petit puit : c'est élégant, mais peu pratique pour récupérer la vis qui s'y trouve.

UNE ÉCOPE POUR DIRIGER LE FLUX D'AIR VERS LE GPU

Outre les boîtiers, HAVN fabrique aussi ses propres ventilateurs. On trouve trois variantes, les H12, H14 et H18, de très beaux moulins dont on retrouve deux exemplaires en 18 cm sur la face avant, qui aspirent de l'air frais, et un 14 cm placé en extraction à l'arrière du boîtier comme de coutume. Ce sont des versions « case » (boîtier) légèrement différentes des modèles vendus au détail, mais leurs performances de refroidissement sont inchangées ; la seule différence est l'absence de la prise mâle permettant de chaîner plusieurs ventilateurs entre eux. Précisons que ces ventilateurs sont particulièrement épais. Le 14 cm mesure 30 mm (au lieu de 25 mm habituellement) et les 18 cm font carrément 40 mm ! Voilà qui permet un angle plus prononcé des pales afin de brasser plus d'air,

ce qui est rendu possible par des moteurs coupleux à roulement FDB (fluid dynamic bearing à triple rainure).
Tous trois sont régulés par PWM, avec une vitesse maximale de 1 600 tpm pour le 14 cm et 1 300 tpm pour les 18 cm. D'ailleurs, ce boîtier embarque un petit HUB PWM pouvant recevoir jusqu'à six ventilateurs, c'est-à-dire le nombre maximal qu'on peut installer dans ce boîtier (jusqu'à 3x 120 mm en haut en plus des trois ventilateurs présents). Signalons, avant de débuter les tests, une particularité unique de ce boîtier. Comme bon nombre de tours contemporaines, c'est un boîtier dit à double compartiments, à savoir qu'il y a un espace en bas dans lequel se loge l'alimentation et les câbles en excédant qui est séparé du compartiment principal dans lequel sont exposés les principaux composants électroniques. Mais dans le cas du BF 360, la séparation inférieure ne va pas jusqu'à l'avant du boîtier. L'espace est un peu réduit (rassurez-vous, on y loge quand même des blocs d'alimentation jusqu'à 20 cm !) et on remarque une écope qui permet de remonter l'air frais aspiré par le ventilateur le plus bas de la façade directement sous la carte graphique, en direction de ses ventilos à elle. Nous allons voir de ce pas si ça fonctionne.

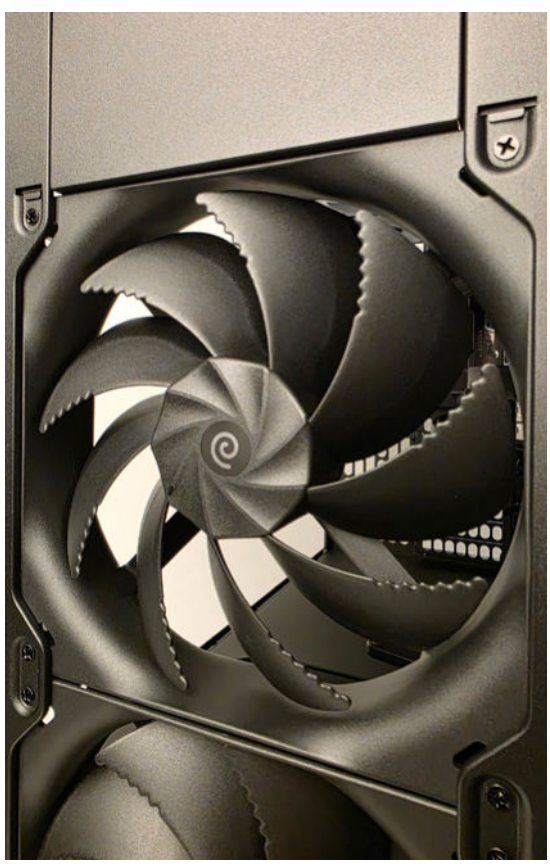
SUPER COOLING… MAIS EN SILENCE ?

Sur nos photos, vous voyez le BF 360 Flow monté avec notre

LES DIFFÉRENCES ENTRE BF 360 ET BF 360 FLOW

On pourrait penser que seule la présence des trois ventilateurs (1* H14 + 2* H18) caractérise le BF 360 Flow par rapport au BF 360 vendu 30 € de moins. Mais en y regardant de plus près, on constate que la face avant du châssis n'est pas identique. Du moins la façade est visuellement la même, mais derrière elle, au lieu du support pour deux ventilateurs de 18 cm terminé en haut par un petit cache venant combler l'espace restant, on se trouve face à support plus classique qui peut accueillir jusqu'à trois moulins de 120 mm, et donc pourquoi pas un radiateur de watercooling ; avec le BF 360 Flow, le refroidissement liquide est limité à un radiateur placé tout en haut. La fiche technique précise d'ailleurs un nombre maximal de ventilateurs qui passe de six sur notre exemplaire à sept.

> 75 °C CPU, 59 °C GPU à fond, 38,6 dBA en PWM : HAVN BF 360 Flow domine le refroidissement, silencieux et rigide, loin devant NZXT H5 Flow ou Phanteks XT View.

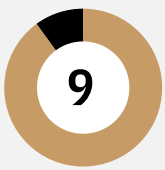


configuration workstation spéciale IA montée spécialement pour le dossier principal de ce même numéro, avec un Ryzen 9 9900X et une GeForce RTX 5090 Zotac. Toutefois, pour les mesures de performances, nous avons utilisé notre plateforme de test habituelle afin que les résultats soient comparables aux autres boîtiers testés dans nos colonnes, à savoir un Core i7-14700K et une GeForce RTX 4080 Super Trinity Black, elle aussi prêtée par Zotac. Le refroidissement dont est capable le boîtier HAVN est tout simplement le meilleur que nous ayons eu depuis la mise en place du protocole actuel fin 2024. Avec les ventilateurs à fond, on relève 75 °C sur le CPU (dans Cinebench en boucle) et seulement 59 °C sur la carte graphique, c'est tout bonnement excellent. Toutefois ça se paye sur le niveau sonore puisqu'en l'absence de régulation, on relève un bruit assez élevé de 45,3 dBA. Avec la résultation PWM en place, à savoir le profil silencieux du BIOS de notre carte mère Asus ROG Maximus Z690 Extreme, le bruit chute à 38,6 dBA ce qui fait de lui un boîtier finalement très discret, pour ne pas dire silencieux. Et pourtant, ses aptitudes de refroidissement restent supérieures à ses concurrents. Notamment on relève 86 °C sur le CPU en charge dans Cinebench, mais seulement 65 °C sur le même CPU en jeu et 65 °C sur la carte graphique également. À titre de rappel, les c'est mieux que le NZXT H5 Flow (86/67/69 °C), que le Phanteks XT View (89/68/70 °C) ou encore que le bequiet! Light Base 500 (88/69/70 °C) qui produisent respectivement 40, 38,9 et 40,6 dBA. En deux mots : c'est excellent ! Ajoutons à ça que grâce à sa rigidité exemplaire et ses patins en silicone (il y en a aussi aux quatre coins des ventilateurs), on ne souffre d'aucune vibration, peu importe le scénario.

CONCLUSION

Une fois de plus on termine un test de boîtier avec les yeux qui brillent. Certes, à 190 €, on pouvait s'attendre à ce que ce boîtier soit mieux fini que ceux vendus à moitié prix. Mais tout de même, soulignons les efforts fournis par HAVN pour rendre une copie aussi parfaite alors qu'il ne s'agit jamais que de son second modèle. On pourra critiquer l'absence de support GPU vertical (il en existe un, universel, en option à 90 €, mais on perdrait l'utilité du flux d'air dirigé sous la carte) et l'incompatibilité avec les cartes mères BTF, certains n'apprécieront peut-être pas les passages de câbles en bandelettes de silicone (nous, si), mais globalement la copie est très bonne. Nous n'avons pas en tête de boîtier à ce prix offrant un tel finish. Le Phanteks Evolv X² par exemple, qui nous a tapé dans l'œil il y a quelques mois, est proposé sans ventilateur au même prix que le BF 360 « nu », mais si son design est encore plus original, et la finition plus que correcte, c'est un net cran en dessous quand on se penche sur les détails. Celui qui s'en approche le plus est peut-être le Fractal Design Torrent (180 €), avec cinq ventilateurs dont 2* 18 cm et un HUB 9 ports PWM, mais lui non plus ne semble pas tout à fait aussi robuste et bien fini que le HAVN. Les jolis Asus ProArt PA401 Wood Edition (125 €) et Antec Flux Pro (190 €) se rapprochent un peu de l'idée aussi, avec leurs 2x 16 cm et HUB 5 ports PWM (Asus) ou 2x 12 cm + 4x 14 cm et HUB 5 ports PWM aussi (Antec), mais là aussi la finition quoi qu'assez bonne est un net cran en dessous.

HAVN BF 360 FLOW



10	Prix
9,5	Finitions
10	Refroidissement
8	Silence
9	Montage

Points forts

- Look !
- Finitions !
- Robustesse !
- Refroidissement
- Silence en usage normal (ventilation régulée)
- Façade aimantée

Points faibles

- Assez bruyant avec les ventilateurs à fond
- On aurait apprécié 0,5 cm d'épaisseur en plus au dos
- Les vis des panneaux amovibles sont difficiles à attraper
- Incompatible BTF

AIO

EKWB Nucleus AIO CR360 Lux D-RGB

UNE FINITION SOIGNÉE À LA EK

Vous pensiez EKWB morts et enterrés ? Que nenni. L'entreprise repart de plus belle, on vous résume cette histoire. L'occasion de vous proposer l'essai de leur Nucleus AIO CR360 Lux RGB, un kit sorti il y a un peu plus de deux ans qui est toujours en vente et qui n'a pas à rougir face à la concurrence, même récente.

Prix officiel et constaté : 181,50 €

→ par Thomas Olivaux

Il n'est pas rare chez Uber Hardware qu'on vous propose un test des mois après la sortie du produit, à l'image du MSI Titan 18 HX AI publié dans ce même numéro. Mais là, c'est une tout autre histoire. Le célèbre fabricant de composants pour watercooling DIY, conscient de la généralisation des watercooling AIO depuis la fin des années 2010, s'est mis en tête de rejoindre ce marché il y a quelques années. Après des mois de gestation, de prototypes et de retards, sa gamme EK Nucleus a finalement vu le jour et on a reçu, fin 2023, un mail nous proposant d'en essayer un modèle. Les premiers tests ont été publiés début 2024, précisément au moment où les magazines PC Update et Hardware Mag pour lesquels votre serviteur officiait à l'époque ont cessé de

paraître. Quelques mois plus tard, Uber Hardware est né, mais EKWB était alors dans la tourmente. Les choses allant mieux à présent, et ce kit AiO étant vraiment très sympa, on s'est décidé à vous en proposer notre étude. Il existe en 240 et 360 mm, avec éclairage (LUX D-RGB) ou sans (Dark), en noir ou en blanc. Le 360 mm éclairé, que nous testons, est à 181,50 € sur le site officiel d'EK (contre 210 à 220 € chez les revendeurs), contre 145,20 € pour la version discrète.

CHAOS ET RENAISSANCE CHEZ EKWB

L'année 2024 restera comme l'année noire d'EK Water Blocks. Tout a basculé au printemps, suite aux enquêtes retentissantes de médias comme Gamers Nexus, révélant une gestion financière catastrophique : salaires impayés, dettes massives auprès des fournisseurs et retards de livraison chroniques. La marque, autrefois intouchable leader du watercooling premium, était

alors au bord du précipice. Ce naufrage a provoqué une fuite des cerveaux sans précédent : des piliers historiques ont rejoint la concurrence, à l'image d'Attila Gobor (Responsable Marketing Produit) parti chez l'autrichien Noctua, ou de Joe Robey (Designer Produit Principal) devenu responsable de la R&D mécanique chez Thermal Grizzly. Pourtant, contre toute attente, le navire n'a pas sombré. En mars 2025, la survie de la marque a nécessité une manœuvre chirurgicale : le transfert des activités vers une nouvelle entité baptisée LM TEK d.o.o. Ce n'est pas un rachat par un fonds de pension anonyme, mais une reprise en main interne menée par Edvard König, le fondateur historique de la marque. Conscient que l'image d'EKWB était devenue toxique, König a repris les commandes en écartant l'ancienne direction — notamment le PDG de l'époque, Matjaž Krč. Pour financer ce nouveau départ, il s'est entouré d'investisseurs locaux slovènes et a convaincu un noyau dur d'ingénieurs de rester pour sauver leur outil de travail. En ce début d'année 2026, le bilan est étonnamment positif. Ce « nouveau » EKWB, plus humble et discipliné financièrement, a réussi à stabiliser sa logistique et son support client. La marque a rationalisé son catalogue, mettant de côté les projets ultra-niche pour se concentrer sur ses fondamentaux. Si le fantôme de 2024 plane encore parfois dans les esprits, l'entreprise actuelle a

FICHE TECHNIQUE

Marque : EKWB

Modèle : Nucleus AIO CR360 Lux D-RGB

Compatibilité : AMD AM4/AM5, Intel LGA 1150/1151/1200/1700/1851/2011/2011-3/2066

Dimensions de la pompe : 82,3 x 69,2 x 61,6 mm

Longueur des tubes : 400 mm

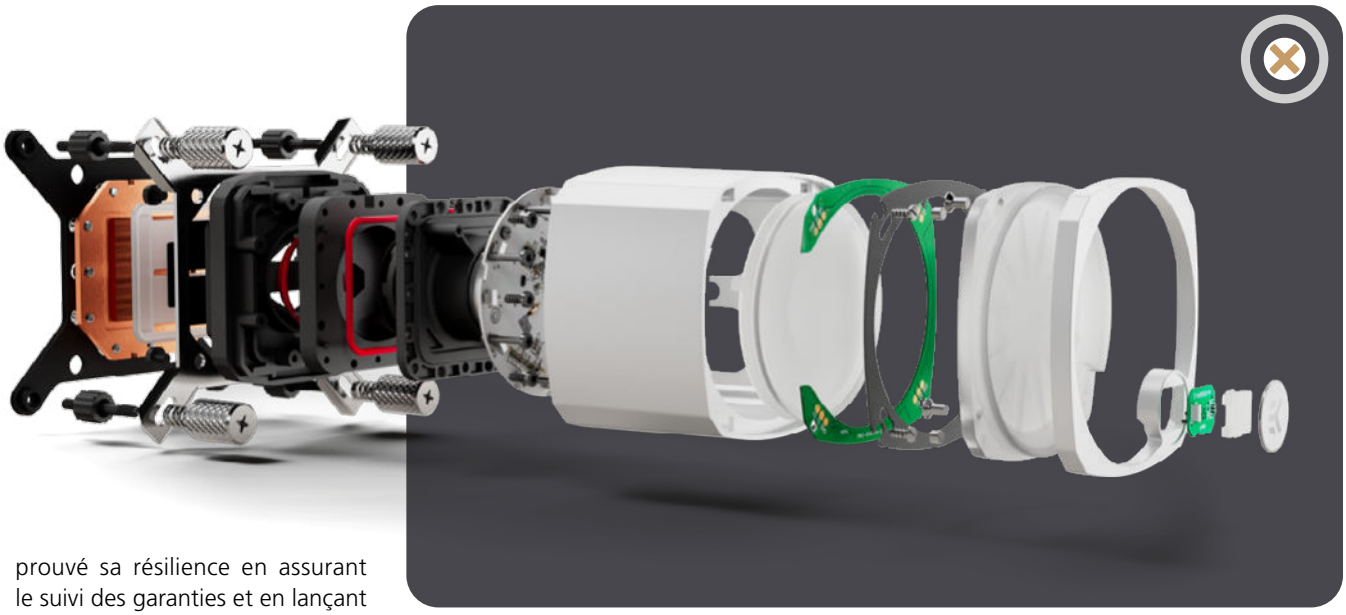
Dimensions du radiateur : 400 x 124 x 27 mm

Ventilateurs : 3* 120 mm PWM (max 2 300 tpm)

RGB : pompe et ventilateurs

Garantie : 5 ans

Divers : pompe PWM (max 3 100 tpm)



prouvé sa résilience en assurant le suivi des garanties et en lançant avec succès ses blocs pour la nouvelle génération de GPU. C'est dans ce contexte de « réveil » que nous avons finalement choisi de publier notre test du CR360 Lux D-RGB.

WATERBLOCK + POMPE CONÇUS EN INTERNE

Contrairement à une grande partie du marché qui se contente de rebadger des solutions tierces (souvent signées Asetek ou CoolIT), EKWB a pris le pari risqué, mais payant, de concevoir sa propre plateforme. Le bloc de ce CR360 Lux cache une pompe PWM capable de monter jusqu'à 3 100 tpm. Cette vitesse élevée est l'une des clés de ses performances, mais grâce à la gestion PWM, elle sait se faire oublier au repos. L'installation est un modèle du genre : EK a simplifié son système de fixation au fil des années pour offrir quelque chose de robuste et rapide, que ce soit sur le socket AM5 ou le socket LGA 1851 avec lequel il est compatible bien qu'il soit sorti avant que ne débarque cette plateforme fin 2024. Petit gadget sympa, le top de la pompe est rotatif. Peu importe le sens dans lequel vous installez le bloc, vous pourrez toujours remettre le logo à l'endroit d'un simple geste. C'est simple, mécanique, et ça évite les usines à gaz logicielles.

UN DESIGN RAVAGEUR

Le CR360 Lux porte bien son nom. Là où beaucoup de concurrents proposent des radiateurs basiques aux bords arrondis et à la peinture fragile, EK soigne la finition. On retrouve un radiateur aux bords « diamond-cut » (angles vifs et brossés) qui lui donne un aspect industriel très haut de gamme.

Le kit EKWB installé dans un PC tout blanc… dommage que nous ayons reçu la version noire !

> Pompe et flux à 69 °C, murmure PWM à 38,2 dBA : EK dompte la chaleur en silence souverain, battant tous ses rivaux sur la plateforme 2025.

Ses dimensions sont tout à fait standards, c'est un modèle de 27 mm d'épaisseur. Les tuyaux en caoutchouc à faible évaporation sont gainés d'un tressage en nylon dense qui ne « s'effiloche » pas au moindre frottement contre le boîtier. Mais le clou du spectacle reste la tête de pompe : un anneau de diffusion D-RGB parfaitement uniforme qui entoure un centre noir profond. C'est sobre, efficace.

DES VENTILOS SEXY AUSSI

Pour accompagner ce radiateur, EK fournit ses ventilateurs EK-FPT (Full Pressure Technology). Conçus spécifiquement pour le refroidissement liquide, ils affichent une plage de rotation allant de 550 à 2 300 tpm, ce qui leur permet d'être aussi à l'aise dans un build silencieux que dans une machine overclockée. Côté câblage, le Nucleus utilise le système EK-OmniLink. Les ventilateurs se chaînent les uns aux autres pour ne laisser sortir qu'un seul faisceau de câbles vers la carte mère.

C'est propre et cela réduit considérablement le temps de montage. Cependant, en 2026, on sent ici l'âge de la conception : face aux systèmes à contact aimanté sans fil comme les Lian Li Uni Fan ou le système iCUE LINK de Corsair, le chaînage par câbles d'EK paraît un peu daté. Cela reste très efficace, mais moins révolutionnaire qu'à sa sortie début 2024. Au final, sur la carte mère, vous avez besoin d'une prise pour alimenter les ventilateurs (4 broches PWM) et une prise pour piloter l'éclairage (D-RGB 5V 3 pins).

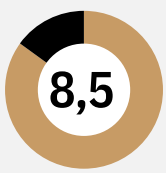
IL EN A SOUS LES AILETTES

On a soumis cet AiO à des tests de performances en l'installant sur notre plateforme d'essai 2025, déjà utilisée pour les autres essais de kits AiO ces derniers mois, et qui repose sur un Core Ultra 7 265K. En laissant tant la pompe que les ventilos à fond, l'EK se fait clairement entendre avec 44,7 dBA indiqués sur notre sonomètre. Toutefois avec le profil de ventilation silencieux activé dans l'UEFI de la carte mère Gigabyte Z890 AI TOP, c'est très bon. Pour ainsi dire inaudible au repos (37,1 dBA) et très silencieux en charge (38,2 dBA). Et tout cela en maintenant des températures à un très bon niveau. En effet, il maintient le CPU à 76 °C avec la régulation PWM et seulement 69 °C si on le laisse à fond. Il fait donc mieux que tous les AiO testés précédemment dans UH !

CONCLUSION

Deux ans après sa véritable arrivée sur nos étals, le EK-Nucleus AIO CR360 Lux D-RGB n'a absolument rien perdu de sa superbe. Si l'on pouvait craindre en 2024 que le produit ne soit qu'un baroud d'honneur avant le naufrage, il est finalement toujours en vente. D'un point de vue purement matériel, c'est un sans-faute. Sa pompe maison et ses ventilateurs FPT offrent un équilibre performances/bruit qui tutoie encore les sommets, même face aux sorties les plus récentes. Certes, son système de câblage OmniLink accuse un léger retard ergonomique face aux solutions magnétiques sans fil de la concurrence, mais c'est un détail que l'on oublie vite une fois le montage terminé. Côté portefeuille, avec un prix affiché à 181,50 € sur le store officiel (désormais géré par LM TEK), il se positionne comme une alternative très sérieuse aux géants du secteur. Il est à la fois mieux fini et un poil plus performant que les références bon marché telles que l'Arctic Liquid Freezer III Pro 360 A-RGB (125 €) ou que le be quiet! Pure Loop 3 LX 360 (120 €), pour comparer des modèles similaires (360 mm avec éclairage) et bien plus abordable que les modèles à écrans LCD de chez Corsair ou ASUS. On remarque qu'il est même un peu plus efficace que le be quiet! Silent Loop 3 360 testé dans UH #8 à 150 €, qui lui n'offre pas de ventilateurs RGB ni de câblage en daisy-chain et dont le radiateur est d'une finition plus basique. EKWB est de retour, et avec un tel produit au catalogue, on ne peut que s'en réjouir. Pour l'anecdote, EK a même proposé pendant un moment une version « direct die » de cet AiO, s'adressant à ceux qui veulent un refroidissement ultime en retirant l'IHS de leur processeur (delid). Et pour ceux d'entre vous qui sont anglophones et qui souhaitent en savoir plus sur ce qu'il se passait en coulisse chez EKWB en 2023/2024, on vous invite à visionner les enquêtes de Gamer Nexus sur YouTube.

EKWB NUCLEUS AIO CR360 LUX D-RGB



7	Prix
9	Silence
9	Refroidissement
10	10 RGB

Points forts

- Performances thermiques de premier ordre.
- Qualité de finition exceptionnelle (Radiateur diamond-cut).
- Installation simple et robuste (compatible LGA 1851).
- Top rotatif
- 5 ans de garantie

Points faibles

- Le système OmniLink un peu daté face au « sans-fil » de 2026.
- Pompe un peu audible à plein régime (3 100 tpm).
- Écart de prix important entre le site officiel et les revendeurs

ALIMENTATION

NZXT C750 Gold Core

UN BON PLAN PLATINUM... SI VOTRE PC DÉMARRE

La nouvelle C750 Gold Core est une des alimentations 750 W ATX 3.1 les plus abordables, mais son câble 12V-2×6 limité à 300 W en restreint grandement l'usage. Dommage, car elle est belle, silencieuse et efficiente.

Prix constaté : 99 €

→ par Thomas Olivaux

NZXT a sorti il y a peu, en même temps que les AiO Kraken Core (le 240 mm est en test dans UH #11), de nouvelles alimentations baptisées Gold Core. Sont proposées les C750, C850 et C1000, dont le nom indique la puissance, aux prix respectifs de 100 €, 130 € et 190 €. Nous testons la version 750 W, une valeur théoriquement a priori pour la majorité des PC.

UN C GOLD PEUT EN CACHER UN AUTRE

La série C n'est pas nouvelle chez NZXT. Déjà, en 2020, nous avions la C750 (modèle dont la référence débute par NP). Alimentation qui s'est vue remplacée en 2022 par la C750 Gold (cette fois les références débutent par PA). NZXT a également sorti des C Gold ATX 3.1 en 2024, adoptant le nouveau connecteur GPU (à l'époque en

12VHPWR), mais à partir de 850 W. Cette fois, c'est la C750 Gold Core, un modèle qui se veut un peu plus abordable tout en offrant les dernières normes et, donc, un câble pour cartes graphiques 12V-2×6. Pour le moment, les deux dernières séries cohabitent et on a donc le choix, pour les modèles 850 et 1 000 W, entre les Gold ATX 3.1 et les Gold Core.

UNE ALIM PLATINUM QUI NE DIT PAS SON NOM

Sachant ça, on poursuit par une drôle d'anomalie : la série C Gold Core, malgré ce que son nom suggère, est certifiée Platinum pour ce qui est du rendement électrique par le laboratoire Cybenetics ! En voilà une heureuse surprise, surtout à ce prix. Mais pourquoi donc les responsables produit et marketing de NZXT ont-ils fait ce choix-là ? Tout d'abord parce qu'il existe chez ce constructeur un modèle très haut de gamme baptisé C1500 Platinum qui, au-delà de sa puissance, est réellement supérieur en tout point et NZXT ne souhaite donc pas assimiler sa nouvelle alimentation à cette dernière. Sans oublier ce que nous venons d'évoquer, les C Gold ATX 3.1 de 2024 restent au catalogue et elles coûtent un

peu plus cher que les C Gold Core. D'autre part, pour une raison qui nous échappe, elle n'a été certifiée qu'en Gold chez 80 Plus, le label historique, dont le protocole de test est pourtant moins exigeant. Soit.

UN BLOC TRÈS COMPACT ET DES CÂBLES DE TOUTE BEAUTÉ

La C750 Gold Core est une alimentation au look soigné, toute noire et avec une grille en alvéoles assez classique pour laisser passer le flux d'air brassé par son ventilateur de 135 mm. Ventilateur qui propose un mode semi-fanless, débrayable via un bouton placé à l'arrière comme ça se fait sur pas mal d'alimentations. C'est un bloc très compact qui facilitera le rangement des câbles dans les boîtiers exigus ; elle ne mesure que 15 cm de long, ce qui est aussi le cas des versions de 850 W et même de 1 000 W ! Cette alimentation 100 % modulaire est fournie avec de très jolis câbles dont chaque fil est recouvert individuellement d'une gaine de protection noire et assez souple.

On bénéficie même, sur les principaux câbles, de peignes pour maintenir les fils bien rangés. Pour votre carte graphique, vous trouverez un câble ATX 12V-2×6 d'une puissance de 300 W (450 W sur les versions 850 et 1 000 W) ; c'est un détail de première importance, on va y revenir. Pour rappel, c'est la prise utilisée depuis 2020 par la majorité des GeForce RTX 40 et 50 à partir du milieu de gamme (xx70). Pour les GPU plus modestes, plus anciens ou encore ceux d'AMD/Intel, le bundle comprend deux câbles PCIe 6+2 pins classiques (il y en a trois avec les 850 et 1 000 W). On dispose aussi de deux câbles pour CPU (un 4+4 pins et un 8 pins).

SEASONIC, CWT ET FINALEMENT HEC

À l'intérieur du boîtier on trouve un électronique qui bénéficie d'une bonne qualité de fabrication malgré le peu d'espace. C'est une topologie Half-Bridge et DC-to-DC, comme on aurait pu s'en douter à ce niveau de rendement. NZXT, qui rédige ses propres cahiers des charges, ne fabrique pas lui-même ses alimentations. Historiquement, la prestigieuse marque Seasonic était l'OEM derrière les premiers blocs série C, tandis que CWT fabriquant les modèles plus récents. Mais l'usine à encore changé, et bien que la présence de condensateurs taïwanais (principalement de marque Teapo) et la certification Platinum puissent faire penser à FSP, derrière la be quiet! Power Zone 2 qu'on a testée il y a quelques mois, il n'en est rien. C'est HEC qui produit la C750 Gold Core et ses sœurs plus puissantes. Si ce nom ne vous dit rien, peut-être connaissez-vous sa marque grand public Cougar, même si on

FICHE TECHNIQUE

Marque : NZXT

Modèle : C750 Gold Core

Puissances disponibles : 750 W, 850 W, 1 000 W

Modèle testé : 750 W

Format : ATX (15 cm de profondeur)

Rails 12 V : 1 (62,5 A)

Connecteur(s) pour CPU : *2 (1* 8 pins + 1* 4+4 pins)

Connecteur(s) pour GPU : *3 (1* 12V-2x6 300 W + 2* 6+2 pins)

SATA : *13

Molex : *2

Normes respectées : ATX 3.1 et PCIe 5.1

Certifications : 80 Plus Gold, Cybenetics ETA Platinum, Cybenetics Lambda A++

Garantie : 7 ans

UN CHOIX TECHNIQUE QUI INTERROGE

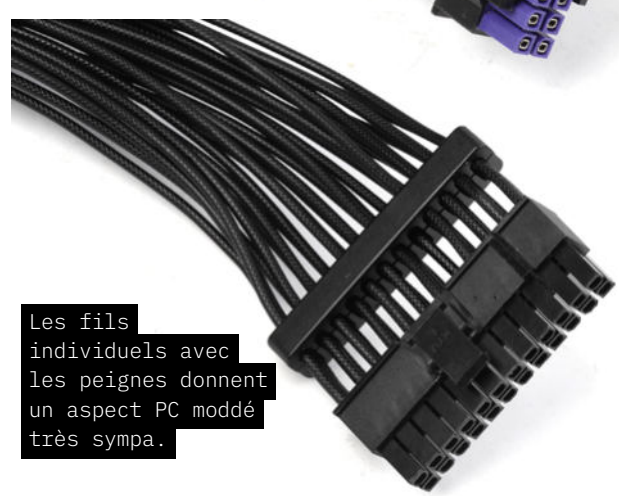
Le problème viendrait-il d'un problème de puissance de l'alimentation ? On n'a même pas essayé les RTX 5080 et RTX 5090, mais tant les spécifications officielles Nvidia que celles des fiches techniques de Zotac indiquent une puissance conseillée de 750 W pour ces deux cartes graphiques. Mais, ce que tout le monde ne sait pas, c'est qu'il existe trois versions du câble ATX 12V-2×6, qui peuvent délivrer soit 300 W, soit 450 W, soit 800 W. Or, pour la RTX 4080 Super, Nvidia indique bel et bien qu'il faut un câble de 450 W ou plus. Soit. Toutefois cette même page Nvidia suggère qu'une RTX 5070 Ti fonctionne avec un câble

ne les trouve pas très facilement en France. Le refroidissement est assuré par un ventilateur Hong Hua HA13525H12SF-Z, c'est-à-dire un 135 mm sur roulement FDB, qui grimpe jusqu'à 2 200 tpm et qui bénéficie d'une régulation PWM.

UNE 750W… QUI N'ALIMENTE MÊME PAS UNE RTX 5070 TI !

Il est temps de mettre la bête à l'épreuve. Et là, nous nous sommes retrouvé face à un problème de taille : impossible de démarrer notre configuration de bench classique, celle qu'on utilise pour les tests boîtiers avec le Core i7-14700K et la Zotac GeForce RTX 4080 Super Trinity Black. Les choses ne se sont pas arrangé quand nous avons constaté que notre configuration avec le Core Ultra 7 265K et la Zotac GeForce RTX 5070 Ti AMP Extreme Inifinity ne démarrait pas non plus ! Plus précisément, le PC se met en route, en observant les LED de diagnostic de démarrage de la carte mère Gigabyte, CPU passe, puis RAM, et ça bloque sur VGA. Pourtant, la LED du connecteur ATX 12V-2×6 sur la carte graphique, qui est là pour indiquer si la prise est mal branchée, reste éteinte. Quand on retire le GPU et qu'on branche l'écran sur la prise HDMI de la carte mère (iGPU), ça fonctionne. Quand on alimente, comme sur la photo ci-contre, la carte graphique par une autre alimentation (la be quiet! Dark Power 14 850 W pour ne pas la nommer), ça fonctionne aussi. La sortie ATX 12V-2×6 de l'alimentation ou son câble serait-il HS ? Non, car en branchant la Gigabyte GeForce RTX 5070 Eagle OC, le PC démarre.

Le coupable le voilà, le câble (et la prise dans l'alimentation) limité à 300 W.



Les fils individuels avec les peignes donnent un aspect PC moddé très sympa.

On a été obligé de brancher une autre alimentation sur la Zotac RTX 5070 Ti AMP Extreme Infinity pour qu'elle démarre sur un PC alimenté par la NZXT C750 Gold Core.

de 300 W, ce que nous avions également en tête. Toutefois, la présence d'un adaptateur 3x 8 pins vers ATX 12V-2×6 dans le bundle Zotac suggère qu'elle aussi réclament 450 W. Pourtant, quand nous avions testé cette carte, nous avions constaté un TBP réglé à 300 W dans son BIOS, mais avec une marge de manœuvre de 15%, ce qui suggère un maximum permis logiciellement de 345 W. quoi qu'il en soit, une carte graphique qu'on imagine utiliser avec une alimentation de ce niveau de puissance ne daigne même pas démarrer. Et n'allez pas croire qu'elle ne soit pas réellement capable de délivrer la puissance annonce (un rail 12 V de 62,5 A, soit 750 W). On a branché, à l'aide des deux câbles 6+2 pins, une bonne vieille RTX 3080 et son TBP de 350 W et tout fonctionne très bien ! C'est vraiment rageant, car il ne fait aucun doute que ce bloc est suffisamment puissant et bien conçu pour alimenter des GPU du calibre des RTX 5070 Ti. Dommage, nous n'avons pas à la rédaction de RTX 5070 Ti sans le moindre overclocking d'usine pour essayer.

PERFORMANTE ET SILENCIEUSE À PART ÇA

Quand on utilise une carte graphique compatible, l'alimentation se comporte très bien. Les tensions sont d'une stabilité exemplaire, il n'y a aucun bruit parasite électronique et la ventilation est très discrète. C'est même un progrès notable comparé aux premières séries C que nous avions eu entre les mains il y a 5 ou 6 ans. Non seulement grâce au mode semi-passif qui fait que l'alimentation n'émet strictement aucun bruit tant qu'on ne joue pas, mais même en charge maximale, notre sonomètre indique 38,5 dBA ce qui, selon nos repères habituels, est très peu audible. Ce n'est d'ailleurs pas pour rien qu'elle a aussi été certifiée Lambda A++ par Cybenetics.

LES SENSE PINS

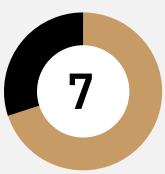
Il faut savoir qu'avec la nouvelle connectique, en plus des 12 pins servant à l'alimentation à proprement parler, il y a 4 petites pins de détection sur le connecteur, dont deux sont dédiées à la notion de puissance. Si ces dernières n'échangent pas de données au même titre qu'un câble USB pourrait le faire, elles permettent à la carte graphique, durant sa phase d'initialisation, d'interroger l'alimentation pour savoir comme sa prise ATX 12V-2×6 est configurée (300 W, 450 W ou 600 W). Et, si ça ne respecte pas le minimum imposé par le fabriquant, ça ne démarre pas.

État S3 (Sense0)	État S4 (Sense1)	Puissance maximale autorisée
Ouvert	Ouvert	0 W
Masse	Ouvert	300 W
Ouvert	Masse	450 W
Masse	Masse	600 W

CONCLUSION

NZXT signe une alimentation au rendement impressionnant et assez silencieuse pour moins de 100 €. Ses câbles tressés individuellement lui donnent un look premium irrésistible. Cependant, le bridage électronique de son connecteur 12V-2×6 à seulement 300 W est un sacré point noir : elle refusera de démarrer les cartes graphiques les plus musclées, comme notre Zotac GeForce RTX 5070 Ti AMP Extreme Infinity qu'on s'attend à alimenter sans problème avec un bloc de 750 W. C'est un excellent choix pour une configuration sobre, mais un piège pour les gamers exigeants. Sachant que les RTX 5060 n'utilisent même pas cette prise pour la plupart, son câble ATX12V-2×6 est, en résumé, limité aux RTX 4070, RTX 4070 Super, RTX 5070 et quelques RTX 5070 Ti potentiellement. La solution ? Optez pour la MSI MAG A750GL PCIE5 que nous avons recommandée dans le dossier config d'UH #11, elle propose un câble 12V-2×6 de 450 W. Ou pour une Corsair RMe (2025), la version 750 W vendue officiellement à 124,90 € est trouvable sur Amazon quasiment au prix de la NZXT, à savoir 105 €. Sinon, il y a bien la variante 850 W de la NZXT Gold Core, avec son rendement certifié Cybenetics Platinum, mais elle coûte quasiment 130 €. Ces deux dernières ont carrément un câble 12V-2×6 de 600 W.

NZXT C750 GOLD CORE



8,5	Prix
8,5	Performances/stabilité
10	Rendement
8,5	Bruit
5	Câblage

Points forts

- Normes les plus à jour (ATX 3.1 / PCIe 5.1)
- Rendement
- Semi fanless
- Câbles gainés individuellement

Points faibles

- Câble ATX12V-2×6 limité à 300 W
- Seulement deux sorties PCIe 8 pins classiques

PC PORTABLE

MSI Titan 18 HX AI (A2XWJG)

DÉ-ME-SURE

Le MSI Titan 18 HX AI porte bien son nom. Avec sa RTX 5090 24 Go et son Core Ultra 9, il écrase tous les benchmarks. Mais cette puissance brute a un prix, un poids, et un volume sonore. Est-ce l'arme ultime pour remplacer votre PC fixe ?

Configuration testée (prix constaté) : autour de 6 500 €

→ par Thomas Olivaux

Il y a des produits comme ça qui traînent à la rédaction pendant des mois. C'est notamment le cas d'un PC portable, totalement hors normes, que nous avons entre les mains depuis le printemps 2025. Il s'agit du MSI Titan 18 HX AI, dans une version quasiment maxée avec, notamment, une GeForce RTX Laptop 5090. Nous avions accepté le prêt à l'époque justement faire connaissance avec ce GPU hors normes, mais devant le tarif prohibitif de la bête (entre 6 500 et 7 000 € avec notre configuration), il aura finalement fallu que nous vous publiions un dossier spécial IA début 2026 et dans lequel nous nous en sommes servi, pour finalement vous proposer son test complet. Si vous êtes un lecteur fidèle, vous l'avez déjà entre aperçu car il nous a servi

il y a peu pour essayer le SSD externe Thunderbolt 5 de LaCie. Nous avions été impressionnés par l'Asus ROG Strix Scar 18 (G835) et sa RTX 5080 Laptop, en test dans UH #6, pensant avoir atteint un plafond de verre en termes de performances mobiles. MSI vient briser cette illusion. Le Titan ne cherche pas seulement à rivaliser, il veut dominer. C'est le remplaçant spirituel et physique de votre tour de bureau. OK, on exagère, car le Strix Scar 18 existe également en version RTX 5090 ainsi que d'autres monstres du même acabit, par exemple le Razer Blade 18. Cet ordinateur ne cherche pas à se glisser dans un sac à dos, mais à écraser la concurrence par la puissance brute. Il s'adresse aux joueurs fortunés, aux créateurs de contenu exigeant une station de travail transportable et aux passionnés de hardware qui veulent simplement « le meilleur, mais transportable ». Notre version de test (A2XWJG-618US) représente le sommet de la chaîne alimentaire : Core Ultra 9 285HX, la GeForce RTX 5090 Laptop avec ses 24 Go de VRAM, 64 Go de DDR5 et un stockage délirant de 6 To en RAID 0. C'est parti !

FICHE TECHNIQUE

Marque : MSI

Modèle :
Titan 18 HX (A2XWJG)

Écran : 18" Mini-LED (16:10 3840x2400 120 Hz)

Processeur :
Intel Core Ultra 9 285HX (8 P-cores + 16 E-cores + NPU, jusqu'à 5,5 GHz)

Mémoire vive :
64 Go DDR5-6400 (2 slots SO-DIMM)

Carte graphique :
Nvidia GeForce RTX 5090 Laptop 24 Go 175 W (150 W + 25 W Dynamic Boost)

SSD : 6 To RAID 0 : 1* M.2 2280 NVMe 2 To (PCIe 5.0) + 2* M.2 2280 NVMe 2 To (PCIe 4.0) ; 1 port M.2 libre

Connectique (gauche) :
2* USB-A (10 Gb/s), lecteur SD

Connectique (droite) :
2* USB-C (Thunderbolt 5, 80 Gb/s, DP Alt mode, PD 3.0), USB-A (10 Gb/s), jack 3,5 mm combo

Connectique (arrière) :
port de charge, HDMI 2.1, 2,5 GbE

Webcam :
FullHD compatible Windows Hello

Réseau :
Wi-Fi 7 (5,8 Gb/s, Intel Killer BE1750x)

Batterie : 99,9 Wh

Dimensions :
40,4 x 30,7 x 3,2 cm

Poids : 3,6 kg

Divers : souris sans-fil (2,4 GHz + BT) MSI M18 en bundle

Le monstre, à peine caché derrière le Zephyrus G14.

UN CHÂSSIS TITANESQUE QUI IMPOSE LE RESPECT

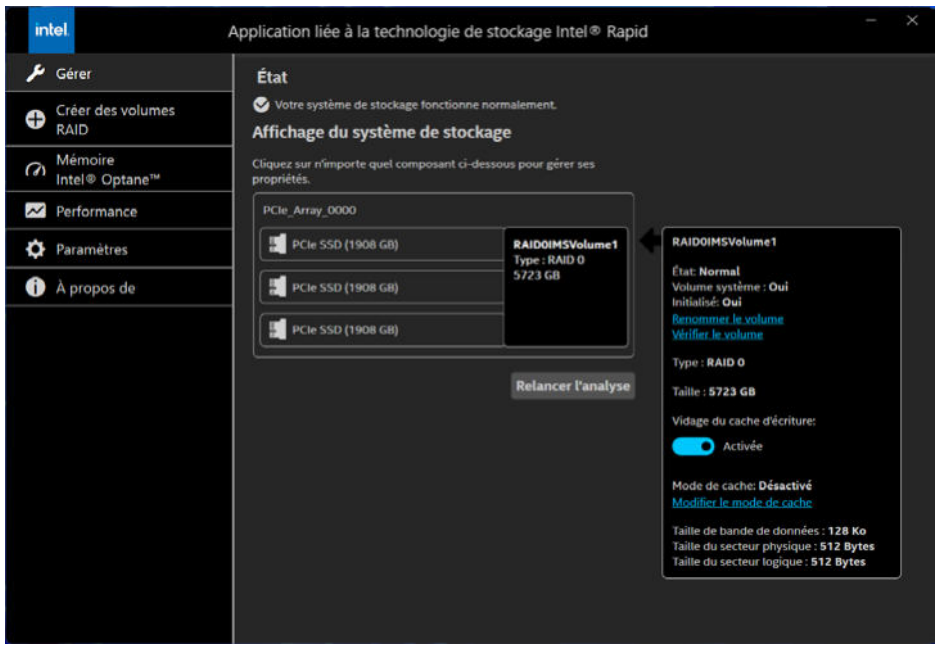
Quand on sort le Titan de son carton, la première chose qui frappe, c'est la densité de l'objet. Avec 3,6 kg sur la balance (sans compter l'immense bloc d'alimentation de près d'un kilo), nous sommes face à un « desktop replacement » pur jus. Le châssis mélange un alliage magnésium aluminium pour le capot et le plateau clavier, avec un dessous en composite haute densité. La rigidité est absolue : contrairement à beaucoup de 18 pouces qui ploient légèrement sous la contrainte, le Titan est une brique. C'est d'ailleurs sur la finition que MSI marque sa différence avec l'Asus ROG Strix Scar 18. Là où le Scar jouait la carte du « Gamer assumé » avec ses plastiques translucides laissant entrevoir les composants (un choix qui divise à plus de 4000 €), le Titan opte pour une approche plus industrielle. Les surfaces métalliques sont douces au toucher, moins sensibles aux traces de doigts que le revêtement « soft touch » d'Asus, et l'assemblage est chirurgical. Esthétiquement, MSI ne fait pas dans la dentelle non plus. Le célèbre dragon sur le capot est désormais illuminé en RGB, tout comme le logo MSI à l'arrière. Mais l'élément le plus distinctif de ce châssis reste l'intégration du pavé tactile : il est totalement invisible, fondu dans le repose-poignet en verre. Une prouesse de design épuré qui contraste avec l'agressivité des énormes évents d'aération arrière, nécessaires pour dissiper les 270W de TDP cumulé (CPU + GPU). En termes d'encombrement, prévoyez un sac à dos large, très large. L'arrière du PC dépasse largement derrière l'écran (à la manière des Alienware ou du Scar), ce qui éloigne l'écran de l'utilisateur et occupe une profondeur conséquente sur le bureau. C'est un transportable, conçu pour voyager d'une prise de courant à une autre, pas pour être utilisé sur les genoux dans un TGV.

La dalle Mini-LED de ce laptop n'a rien à envier aux écrans OLED à la mode.

MINI-LED PLUTÔT QU'OLED

Pour son vaisseau amiral, MSI écarte la technologie OLED qui se répand pourtant comme une traînée de poudre ces derniers temps. Ce choix du Mini-LED (technologie AmLED d'AU Optronics) sur ce format 18 pouces est avant tout dicté par la réalité du marché : à l'heure actuelle, les dalles OLED 4K capables de grimper à 120 Hz n'existent tout simplement pas dans cette diagonale de 18 pouces. MSI a donc privilégié la surface et la définition plutôt que de redescendre sur du 16 pouces pour céder aux sirènes de l'OLED. C'est un mal pour un bien, car le Mini-LED est particulièrement cohérent avec la cible « station de travail » du Titan. Si l'OLED offre le contraste absolu, le Mini-LED permet d'atteindre des pics de luminosité bien supérieurs sans aucun risque de marquage (burn-in), un argument vital pour les créatifs qui laissent des barres d'outils statiques affichées toute la journée sur Premiere Pro ou Blender. Et à vrai dire, on a bien du mal à trouver une différence de noir et de contraste avec une vraie dalle OLED, c'est tout bonnement excellent. La dalle affiche une définition 4K UHD+ (3840 x 2400 pixels) au format 16:10. La finesse d'affichage est chirurgicale, rendant le moindre texte ou texture d'une netteté absolue. Par défaut, le zoom de Windows est à 200 %, mais il nous arrive de descendre à 175 voire 150 % pour travailler en affichant plus de contenu. Côté rétroéclairage, la technologie s'appuie ici un local dimming très abouti de 1008 zones, ce qui est énorme pour une si petite surface. Le résultat en HDR est percutant : avec un pic de luminosité annoncé à 1000 nits (certification HDR 1000), les effets de lumière sont littéralement éblouissants. D'ailleurs, on relève durant nos tests une luminosité qui grimpe à 708 cd/m² en SDR, c'est un record ! Et même sur batterie on profite de 452 cd/m² avec le réglage poussé au maximum. C'est très appréciable. Dommage

Cette modeste souris gamer, avec quelques LED, est livrée avec le PC.



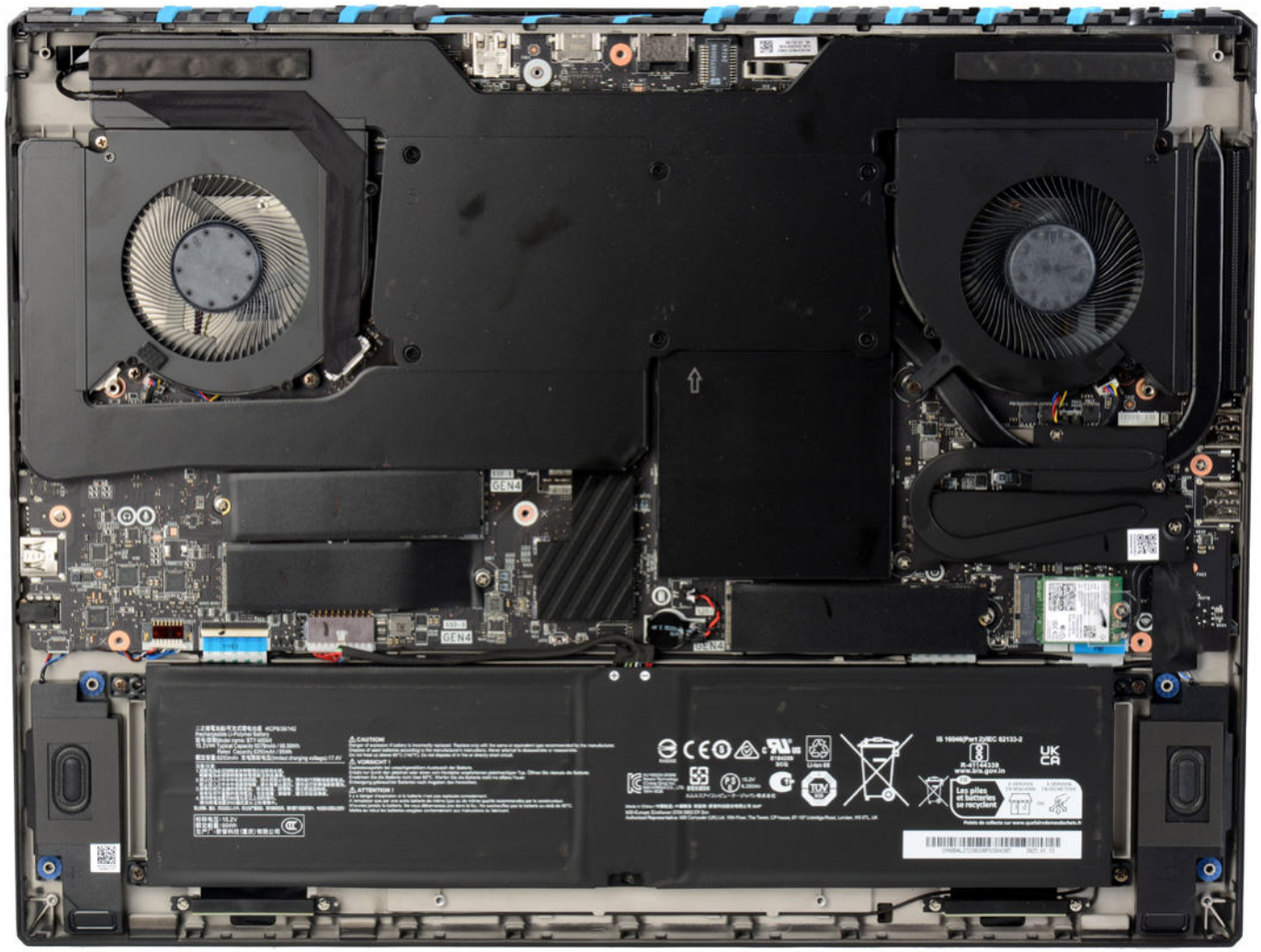
Le RAID est en fait géré par le contrôleur de stockage et les pilotes d'Intel.

que ce soit un PC portable qu'on n'utilise justement jamais en extérieur… Et en HDR, on peut atteindre un peu plus de 1 100 cd/m² sur 10 % de blanc, mais en plein écran c'est à peine plus qu'en SDR (et c'est déjà pas mal). Le blooming (halo lumineux autour des objets clairs sur fond noir) est remarquablement maîtrisé pour une dalle de cette taille. On ne le perçoit que dans des conditions extrêmes, comme un curseur blanc sur un fond noir total. Le taux de rafraîchissement de 120 Hz pourra sembler « juste » pour les puristes de l'eSport, mais en 4K natif, même une RTX 5090 aura fort à faire pour saturer cette fréquence sur les derniers titres AAA. Enfin, la colorimétrie est gérée par la suite logicielle True Color. En sortie de boîte, le profil DCI-P3 est couvert à 99,7 %, avec un Delta E moyen de 3,8 (on a vu mieux sur ce dernier point). Contrairement aux dalles OLED souvent brillantes, MSI applique ici un traitement mat efficace qui élimine les reflets gênants, permettant de profiter de la puissance de la machine même dos à une fenêtre, d'autant plus grâce à sa luminosité.

SWITCHES CHERRY ET TOUCHPAD INVISIBLE

C'est l'un des points de différenciation majeurs de la gamme Titan : l'intégration de véritables contacteurs mécaniques sous les doigts. MSI poursuit son partenariat avec Cherry pour proposer des switches MX Ultra Low Profile. Contrairement aux claviers à dômes classiques (ou « chiclet ») que l'on trouve sur 99 % des portables du marché, ces contacteurs utilisent un mécanisme à ressort et des contacts en or. La sensation est incomparable : la course est de 1,8 mm avec un point d'activation très franc et un retour tactile marqué. Pour le jeu, c'est d'une précision chirurgicale ; pour la saisie de texte, c'est un pur bonheur, à condition d'aimer le « clic » sonore caractéristique, assez aigu, qui pourrait agacer vos voisins de bureau. On regrettera toutefois que ce traitement de faveur ne concerne que les touches principales : le pavé numérique et les touches de fonction restent sur des membranes classiques, créant une disparité de sensation un peu déroutante au début. Le pavé tactile, quant à lui, est une véritable pièce de design. MSI a opté pour une surface en verre totalement lisse et « seamless », c'est-à-dire sans aucune délimitation physique avec le repose-poignet. Lorsqu'il est éteint, il est virtuellement invisible. Une fois la machine sous tension, une zone de rétroéclairage RGB délimite ses contours. Mais la vraie prouesse est technologique : il s'agit d'un pavé haptique. À l'instar de ce que propose Apple sur ses MacBook, il n'y a pas

Profil d'énergie	TDP CPU (PL1)	TDP CPU (Max Turbo Power PL2)	Cinebench 2024 (multi)	TDP GPU total (Dynamic Boost inclus)	3DMark Time Spy (Graphics)	Bruit maximum (dBA)
ECO-Silent	15 W	15 W	394	96 W	17 493	38,9
Équilibré	140 W	220 W	2 130	150 W	23 154	48,1
Performances extrêmes	200 W	220 W	2 132	175 W	24 987	51,2
MSI AI Engine	Variable	Variable	2 134	Variable	23 182	Variable
Sur batterie	55 W	55 W	688	24 W	8 416	41,5



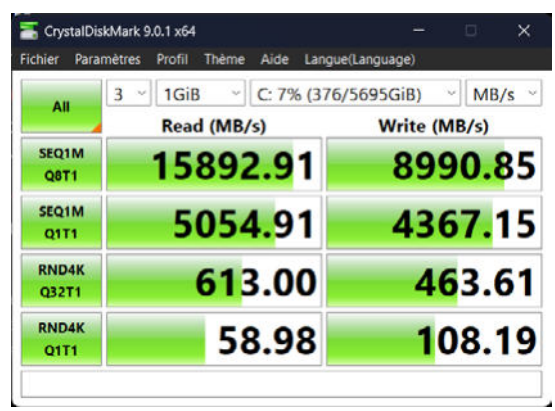
de mouvement mécanique de « clic ». Des moteurs de vibration (actuateurs) simulent la sensation d'enfoncement sous le doigt. L'avantage est double : on peut cliquer sur toute la surface de la zone avec la même force, et la fiabilité est accrue puisqu'il n'y a pas de pièce mobile susceptible de s'encrasser. En pratique, la glisse est bonne, mais l'absence de rebords physiques demande un certain temps d'adaptation pour ne pas « sortir » de la zone par erreur. De toute façon, vous utiliserez ce PC sur un bureau avec une souris, sachant que MSI en fournit même une en bundle ; une M18 et son capteur 5 000 ppp qui fonctionne, au choix, en 2,4 GHz avec son dongle, en BT et en filaire USB. L'ensemble est piloté par la suite SteelSeries Engine, qui permet une personnalisation touche par touche du rétroéclairage. On peut ainsi créer des profils spécifiques pour chaque jeu, mettant en avant les raccourcis essentiels. C'est complet, puissant, et cela contribue à l'aspect vitrine technologique de la machine. On notera aussi que le repose-poignet en verre est particulièrement agréable car il reste frais, même lorsque les composants internes commencent à chauffer sérieusement.

CPU DIGNE D'UN DESKTOP ET 6 TO DE SSD EN RAID

Sous le capot, MSI a choisi le Core Ultra 9 285HX, fleuron de l'architecture Arrow Lake. Ce processeur n'a de mobile que le nom : avec ses 24 cœurs (8 P-cores orientés performance et 16 E-cores pour l'efficacité), il est conçu pour encaisser des charges de travail colossales. Un CPU avec un TDP de 140 W (bien au-delà des 55 W de la spécification Intel !) et un PL2 de 220 W digne d'un PC de bureau ! Le Core Ultra 9 285K, avec le même nombre de cores, c'est 125 W en PL1 et 250 W en PL2. Précisons quand même

que le cumul CPU + GPU ne peut excéder 270 W (technologie MSI Overboost Ultra) ; quand le GPU est sollicité, le CPU tombe à 95 W. En pratique, cela signifie que le processeur peut maintenir des fréquences très élevées sur de longues périodes sans s'effondrer. Bien que cette puce intègre un NPU pour les tâches liées à l'IA, c'est bien sa puissance brute en calcul pur qui impressionne, venant taquiner sans complexe les configurations de bureau les plus sérieuses. Un seul chiffre pour illustrer cela : 2 132 points dans Cinebench 2024, tant avec le profil Performances extrêmes que le profil Equilibré. Et pas sur une seule passe, mais sur une session de 10 minutes non-stop. Pour vous donner une idée, le Core Ultra 9 285K, le CPU de bureau le plus puissant d'Intel, atteint 2 450, soit seulement 18,5 % de mieux. Le Ryzen 9 9900X et ses 1 812 points est battu.

Quant au Core Ultra 9 275HX sur le ROG Strix Scar 18, on avait fait 2 218 sur ce même test. Ce sont deux CPU virtuellement identiques (seule la fréquence Turbo passe de 5, 4 à 5,5 GHz), c'est le TDP et le cooling du MSI qui font la différence. C'est impressionnant. Et bruyant, on en reparlera. À la prise, rien qu'en Cinebench (donc sans que le GPU ne fasse quoi que ce soit), on est entre 270 et 280 W. La partie mémoire vive est à l'avenant avec 64 Go de DDR5-6400. C'est un point à souligner, car MSI exploite ici une fréquence plus élevée que bon nombre de laptops gamers concurrents qui plafonnent souvent à 5200



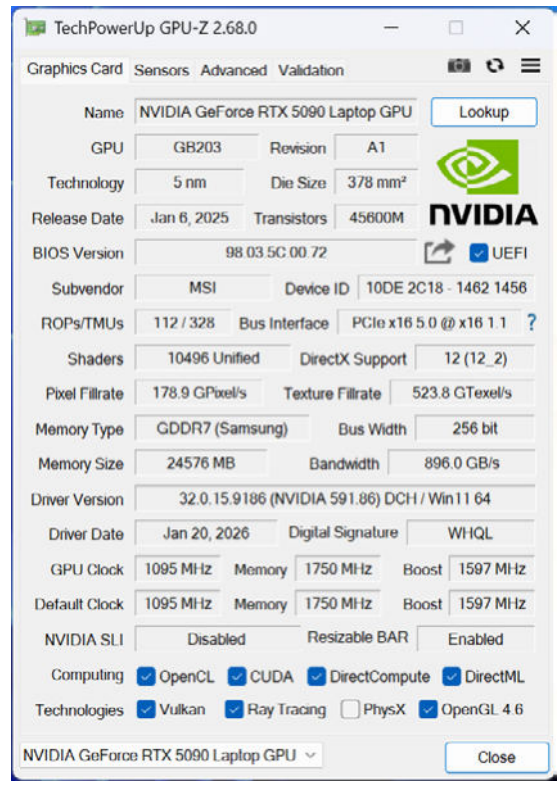
ou 5600 MHz. Cette mémoire est installée via deux barrettes SO-DIMM de 32 Go chacune, ce qui signifie que la machine n'est pas figée dans le temps : vous pourrez monter jusqu'à 2 x 48 Go (96 Go au total) si vos besoins en création 3D ou en virtualisation l'exigent. En revanche, c'est une régression

si l'on compare au châssis 2024 du Titan 18 2024 qui disposait de 4 slots SO-DIMM. Mais c'est le stockage qui constitue la véritable démesure de cette version. MSI a opté pour des composants de premier choix avec des SSD Samsung PM9E1 (la version OEM du célèbre 990 Pro) pour le slot PCIe 5.0 et des PM9A1 (équivalents aux 980 Pro) pour les deux autres en PCIe 4.0, sachant qu'il y a même un 4e port M.2 (PCIe 4.0 aussi) de libre ! Cette configuration de 6 To au total est livrée configurée d'usine en RAID 0, ce qui explique les débits séquentiels stratosphériques que nous avons pu observer. Si ce choix maximise la vitesse pure pour le montage vidéo lourd ou les chargements instantanés, il multiplie aussi les risques : si un seul des trois disques lâche, l'intégralité des 6 To est perdue. Que les plus prudents se rassurent : il reste tout à fait possible de « casser » ce volume RAID pour repasser à un fonctionnement classique avec trois lecteurs distincts, moyennant une réinstallation complète du système.

RTX 5090 LAPTOP, UNE 5070 TI AVEC 24 GO

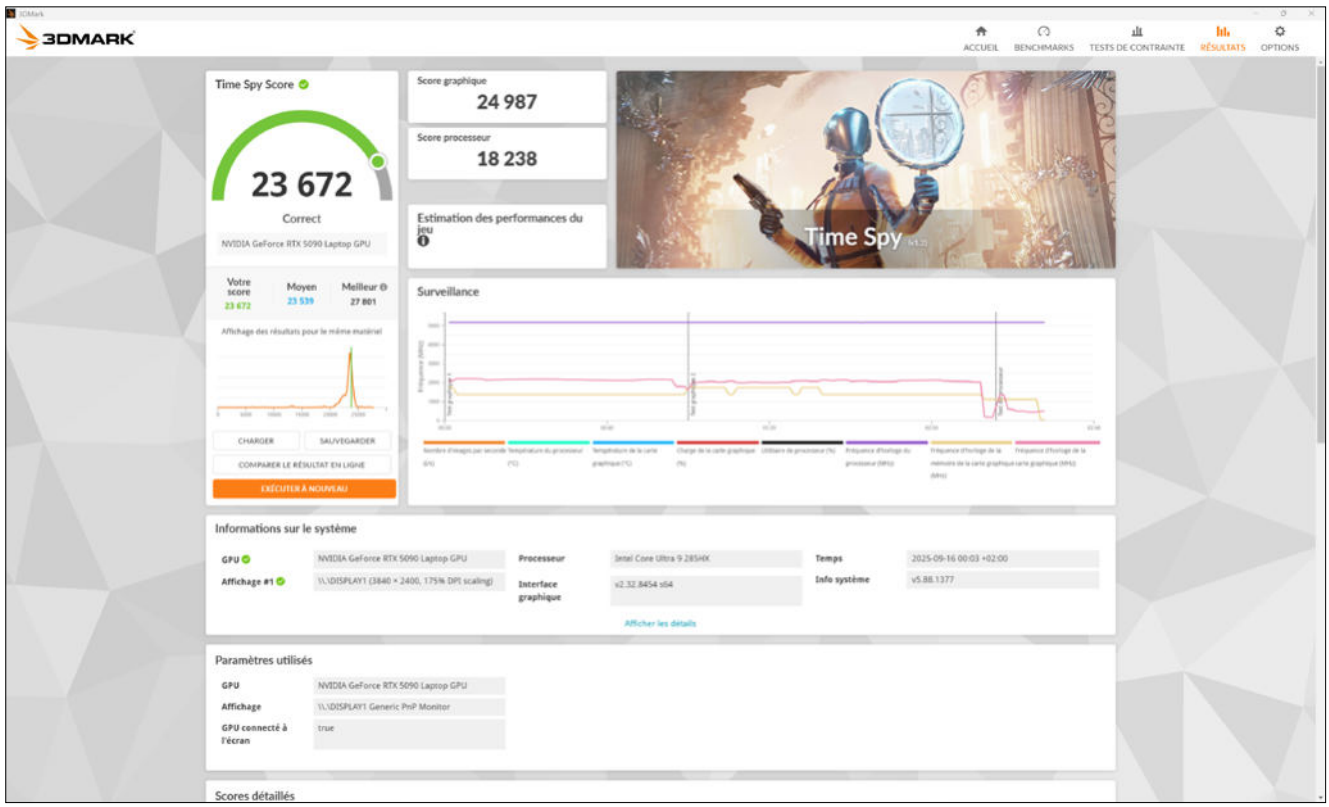
On ne va pas se mentir, le nom RTX 5090 sur un ordinateur portable relève toujours d'une subtile gymnastique marketing de la part de Nvidia. Sous le capot, nous ne sommes pas face au monstre de puissance qui équipe les PC de bureau les plus extrêmes. Précisément, exit la puce GB202 et ses 21 760 CUDA cores, c'est une GB203 comme sur les RTX 5080 et RTX 5070 Ti de PC assemblés, mais avec ici 10 496 unités, au lieu de 10 752 et 8 960 sur ces deux dernières. Et compte tenu du TDP inférieur aux versions de bureau, on se trouve, comme vous le verrez un peu plus loin, plus ou moins au niveau de puissance d'une RTX 5070 Ti. Cependant, et c'est là que MSI et Nvidia frappent un grand coup, cette version mobile est dotée d'un argument massue : 24 Go de mémoire vidéo dédiée (VRAM). C'est une première historique sur PC portable. Si cette débauche de mémoire profite évidemment au confort de jeu en 4K UHD+ (évitant tout swap de textures), elle ouvre surtout à ce laptop des perspectives qui dépassent largement le cadre du divertissement.

Comme nous l'expliquons dans notre dossier spécial LLM publié dans ce même numéro, la quantité de VRAM est le facteur limitant numéro un pour l'intelligence artificielle générative locale. Avec 24 Go, le Titan 18 HX AI change de dimension : il permet de faire tourner sans mal un modèle de taille moyenne (un 24B quantifié en Q4) tout en conservant une belle fenêtre de contexte. C'est une prouesse pour une machine transportable, qui devient de fait une plateforme de développement IA redoutable. Cette architecture Blackwell est ici poussée à son maximum avec un TGP de 175 W (plus Dynamic Boost). La puissance brute de calcul reste certes contrainte par le format, mais les 24 Go de VRAM agissent comme un véritable poumon. Pour les créateurs ou les chercheurs en IA, c'est l'assurance de pouvoir manipuler des jeux de données ou des scènes 3D complexes que même le récent Scar 18 et ses 16 Go de VRAM commençaient à trouver trop étroits. MSI exploite parfaitement cette débauche de silicium grâce à sa chambre à



DE L'IMPORTANCE DU DLSS (DANS CYBERPUNK 2077)

Réglages du jeu	DLSS (modèle Transformer)	FPS moyen
2560×1440 RT Ultra	Auto + FG x2	151,84
3840×2400 Ultra	Qualité + FG x2	93,95
3840×2400 Ultra	Qualité + FG x4	162,13
3840×2400 Ultra	Performance + FG x2	123,29
3840×2400 Ultra	Performance + FG x4	210,50
3840×2400 Ultra	Performance ultra + FG x4	288,29
3840×2400 RT Ultra	Désactivé	20,57
3840×2400 RT Ultra	Qualité (modèle Transformer)	36,73
3840×2400 RT Ultra	Performance (modèle Transformer)	54,03
3840×2400 RT Ultra	Auto (modèle Transformer) + FG x2	100,31



vapeur massive, qui évite au GPU de voir ses fréquences s'effondrer après dix minutes de charge intensive. Hélas, là aussi le label RTX 5090 est trompeur, car les 24 Go sont câblés en 256 bits (comme sur les RTX 5070 Ti et RTX 5080) et non en 512 bits (telle la RTX 5090). Résultat des courses, une bande passante de 896 Go/s identique à celle de… la GeForce RTX 5070 Ti 16 Go.

PERFORMANCES EN JEU : 4K, AVEC DES COMPROMIS

Avec une telle fiche technique, l'objectif est clair : jouer en 4K UHD+ native (3840 x 2400) sans avoir à sacrifier les détails graphiques. Dans les faits, la RTX 5090 Laptop et ses 24 Go de VRAM permettent effectivement d'atteindre un niveau de confort inédit sur portable. Sur des titres comme Cyberpunk 2077 ou le plus récent S.T.A.L.K.E.R. 2, la fluidité est au rendez-vous et la quantité de mémoire vidéo garantit l'absence totale de saccades liées au chargement des textures, un problème qui commence à toucher les cartes de 12 ou 16 Go dans cette résolution. Cependant, le titre de cette section ne ment pas : le « natif » a ses limites dès que l'on active les technologies de rendu les plus gourmandes. Dans Cyberpunk 2077 RT Ultra, et encore, sans Path Tracing, on n'obtient que 20,57 FPS en natif. Vous avez bien lu. Si vous comptez jouer avec le Path Tracing (le Ray Tracing intégral) au maximum en 4K, la puce Blackwell, aussi puissante soit-elle, finit par montrer ses limites de débit pur. Pour maintenir un framerate digne de l'écran 120 Hz du Titan, le passage par le DLSS 4 et la Frame Generation reste indispensable. C'est là que le compromis intervient : on ne joue plus vraiment en 4K « réelle », mais on profite d'une image reconstruite par l'IA. Heureusement, à ce niveau de densité de pixels sur un écran de 18 pouces, la différence est virtuellement imperceptible à l'œil nu. Et on profite d'un excellent compromis qualité/performances avec la FG x2 qui permet d'atteindre 100 FPS de moyenne sans aucune perte de qualité ou de fluidité (si ce n'est quelques effets de tearing en l'absence de VRR). Dans notre panel de tests, le Titan 18 HX AI s'impose logiquement comme la nouvelle référence. Dans 3DMark Time Spy Extreme, il décroche un score de 23 484 points, creusant un écart significatif avec les machines équipées de la génération précédente ou même avec le Scar 18 et sa 5080. En jeu, sur Shadow of the Tomb Raider ou Forza Motorsport, on dépasse allègrement les 100 FPS en réglages maximum. Mais

dès que l'on attaque des « gros morceaux » comme Alan Wake 2 avec tous les curseurs à droite, on réalise que la physique reste la physique : une puce de 175 W, aussi optimisée soit-elle, ne remplacera jamais totalement une carte de bureau de 450 W. Le Titan est le roi des portables, certes, mais un roi qui doit encore composer avec les outils d'upscaling pour offrir l'expérience « parfaite ». Et surtout, alors que la RTX 5090 est presque deux fois plus performante qu'une RTX 5080, la RTX 5090 Laptop ne va même pas 15 % plus vite qu'une RTX 5080 Laptop. Et à peu près autant comparé à une RTX 4090 Laptop (sans prendre en compte le MFG, absent de cette dernière).

REFROIDISSEMENT : LA CHAMBRE À VAPEUR À L'ÉPREUVE DES WATTS

Dissiper la chaleur dégagée par un Core Ultra 9 et une RTX 5090 dans un châssis de portable, même au format 18 pouces, relève de la gageure physique. Pour relever le défi, MSI s'appuie sur une chambre à vapeur (Vapor Chamber) surdimensionnée, couvrant à la fois le CPU et le GPU. Contrairement aux systèmes classiques à base de caloducs (heatpipes) qui transportent la chaleur vers des points précis, la chambre à vapeur répartit les calories sur une surface plane beaucoup plus vaste avant de les envoyer vers les radiateurs. C'est ce dispositif qui autorise la technologie MSI Overboost Ultra, capable de maintenir une enveloppe thermique cumulée de 270 W. En clair, même lorsque les 24 cœurs du CPU et les unités de calcul du GPU sont sollicités simultanément, le système évite le « thermal throttling » et maintien des fréquences stables. Le flux d'air est assuré par deux ventilateurs massifs qui rejettent l'air chaud via quatre sorties (deux à l'arrière, deux sur les côtés). En mode « Extreme Performance », le Titan ne fait pas de miracle : le souffle est puissant et clairement audible. On a mesuré 51,2 dBA à 50 cm de l'écran.

C'est bruyant, c'est indéniable, et le port du casque sera fortement recommandé pour de longues sessions de jeu. Toutefois, on notera une différence notable avec les châssis plus compacts : la signature sonore est ici beaucoup plus grave et sourde, ce qui la rend moins fatigante à l'oreille que les sifflements aigus que l'on peut croiser sur un Zephyrus G14 en plein effort. Un point particulièrement satisfaisant concerne la gestion des températures de surface. Grâce à l'épaisseur du châssis

PAS DE G-SYNC SUR UN PC GAMER À PLUS DE 6 000 € ?

C'est la mauvaise surprise. Malgré une fiche technique à faire pâlir une station de travail et un tarif stratosphérique, la dalle Mini-LED du Titan 18 HX AI fait l'impasse sur la technologie Nvidia G-Sync. Un comble, alors que d'autres constructeurs parviennent à proposer du rafraîchissement adaptatif sur leurs dalles Mini-LED, même si c'est en QHD. Le problème ne vient pas de l'absence de MUX Switch (il est bien présent et permet de connecter la dalle directement au GPU Nvidia), mais plus probablement du TCON (Timing Controller) de cet écran 4K 120 Hz. Synchroniser en temps réel les milliers de zones de rétroéclairage Mini-LED avec une fréquence de rafraîchissement variable (VRR) est un défi technique complexe. Bon, on dit ça, mais le site d'AUO et en particulier la page dédiée à la technologie AmLED évoquent une compatibilité avec G-Sync, FreeSync et VESA Adaptive Sync. Peut-être MSI a-t-il simplement souhaité jouer la prudence. Ou s'est raté. En l'absence de G-Sync, si votre framerate chute sous les 120 FPS en 4K, vous devrez choisir entre le tearing (déchirure d'image) ou l'input lag de la V-Sync classique. Pour une machine qui se veut la Rolls du gaming, c'est une lacune difficile à justifier, surtout quand on sait que la RTX 5090 aura fort à faire pour maintenir un 120 FPS constant en 4K native sur les titres les plus lourds.

et à l'efficacité du bloc de refroidissement, la chaleur est très bien isolée du plateau clavier. La zone ZQSD ne surchauffe pas trop (44,5 °C en jeu selon nos mesures), et le repose-poignet en verre contribue à cette sensation de confort thermique pour les mains. En usage bureautique ou multimédia, le MSI Center gère intelligemment la ventilation à condition d'activer le profil MSI AI Engine : les ventilateurs peuvent se couper totalement, rendant la machine parfaitement silencieuse tant que l'on ne sollicite pas le GPU Blackwell. C'est une machine de l'extrême, mais une machine civilisée quand il le faut. Mais si vous optez pour le profil Equilibré, les ventilateurs s'activent très vite, même en surfant. Pour l'anecdote, il y a aussi un heatpire qui transporte la chaleur du SSD PCIe 5.0 vers la chambre à vapeur.

DEUX PORTS THUNDERBOLT 5

Sur une machine de ce gabarit, on attend une connectique qui ne souffre d'aucun compromis, et MSI ne déçoit pas. Le passage à la plateforme Core Ultra 9 de 2026 apporte enfin le support du Thunderbolt 5 sur les deux ports USB-C dédiés. Avec une bande passante pouvant grimper jusqu'à 120 Gbps en mode « boost » (80 Gb/s en mode normal), c'est la garantie de pouvoir piloter plusieurs écrans 8K ou des solutions de stockage externes ultrarapides sans aucun goulot d'étranglement. Ces derniers permettent notamment de copier des fichiers à un peu plus de 6 Go/s depuis le SSD externe LaCie Rugged SSD Pro5 vers notre grappe RAID. Quel confort. Le reste de la dotation est à l'avenant : trois ports USB 3.2 Gen2 Type-A, un port HDMI 2.1 (FRL 48 Gbps) et un port réseau RJ-45. On remarque la présence d'un lecteur de carte SD Express, capable d'exploiter les débits des cartes les plus rapides du moment, un outil indispensable pour les photographes et vidéastes qui tournent en RAW. Hélas nous n'avons pas pu le tester car, si nous possédons une carte microSD Express, nous n'avons pas de carte SD Express en plein format.

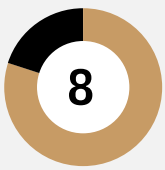
Côté réseau, outre le port 2,5 GbE, le Titan 18 HX AI embarque une puce Wi-Fi 7 Intel Killer BE1750 (une BE200 avec un logiciel un peu amélioré). Avec un routeur compatible, les débits sans fil n'ont plus rien à envier au filaire, avec une latence réduite au minimum pour le jeu en ligne. On transfère à 196 Mo/s sur notre Netgear Orbi 970. La Webcam est surtout pratique car elle est compatible Windows Hello (par reconnaissance faciale IR), mais sa qualité en visio-conférence est assez honteuse pour une machine de ce calibre. Enfin, abordons le sujet qui fâche : l'autonomie. Avec une batterie de 99,9 Wh (le maximum autorisé pour voyager en avion), le Titan dispose de la plus grosse réserve d'énergie possible. Mais face à un écran 4K Mini-LED de 18 pouces et un processeur HX, les miracles n'existent pas. En usage bureautique léger (mode MSHybrid, luminosité à 50 %, rafraîchissement à 60 Hz), nous avons péniblement atteint les 4 heures d'utilisation. Dès que l'on sollicite le GPU ou que l'on pousse la luminosité de la dalle HDR, l'autonomie fond comme neige au soleil. Considérez cette batterie comme un onduleur intégré vous permettant de déplacer la machine d'une pièce à l'autre sans l'éteindre, plutôt que comme une source d'énergie pour une journée de travail nomade. Notez d'ailleurs que les performances sont extrêmement bridées lorsque le PC n'est pas relié à son massif bloc d'alimentation, vous pouvez vous en rendre compte avec les scores Cinebench 2024 et 3D Mark TimeSpy (Graphics) du tableau un peu plus haut.

L'alimentation de 400 W à elle seule pèse aussi lourd qu'un ultrabook.

CONCLUSION

Le MSI Titan 18 HX AI est un produit d'exception qui ne s'adresse qu'à une niche d'utilisateurs. Contrairement au Zephyrus G14 qui tentait le grand écart entre ultrabook et gaming, le Titan choisit son camp : la puissance absolue, sans l'ombre d'un compromis. Nous avons cet exemplaire entre les mains depuis le printemps 2025, et après presque un an de recul, le constat reste le même : il n'y a toujours pas plus puissant sur le marché mobile, même en 2026. Il y a quelques mois, nous accordions notre Award Black au ROG Strix Scar 18 (G835). Ce dernier reste une machine formidable, peut-être plus « raisonnable » avec sa RTX 5080 et son écran QHD+ très rapide. Mais le Titan joue dans une autre cour. Face aux ténors de 2025 que sont le Razer Blade 18, l'Alienware m18 R2 ou le Lenovo Legion 9i, MSI garde une longueur d'avance sur deux points critiques : son clavier mécanique Cherry MX, indétrônable en termes de sensation, et son stockage RAID 0 de 6 To qui humilie n'importe quelle autre configuration d'usine. Surtout, c'est l'un des rares laptops à transformer l'essai de l'intelligence artificielle locale. Là où les modèles équipés de 12 ou 16 Go de VRAM commencent à montrer leurs limites, les 24 Go de sa RTX 5090 lui ouvrent les portes des LLM les plus gourmands. C'est une vitrine technologique, mais peut-on vraiment justifier son tarif stratosphérique ? Pas si sûr. On adore l'utiliser, mais même si nous avions un énorme budget nous ne l'achèterions sans doute pas. Point d'award, donc.

MSI TITAN 18 HX AI (A2XWJG)



2	Prix
1	Portabilité
4	Autonomie
9,5	Écran
10	Performances Internet/multimédia
9,5	Performances en jeu
8	Bruit Internet/multimédia
5	Bruit en jeu
10	Connectique

Points forts

- Performances CPU/GPU au sommet de la hiérarchie
- Les 24 Go de VRAM (un "game changer" pour l'IA locale)
- Écran Mini-LED 4K somptueux et très lumineux
- Le confort incomparable du clavier Cherry MX
- Stockage 6 To en RAID 0 (Samsung PM9E1/PM9A1) foudroyant
- Connectique complète avec deux ports Thunderbolt 5

Points faibles

- Tarif élitiste dépassant l'entendement
- Encombrement et poids (machine + bloc d'alimentation)
- Nuisances sonores importantes
- Autonomie anecdotique, même pour un 18 pouces
- Le RAID 0 d'usine : un risque pour les données sans sauvegarde

NAS

Ugreen NASync DH2300

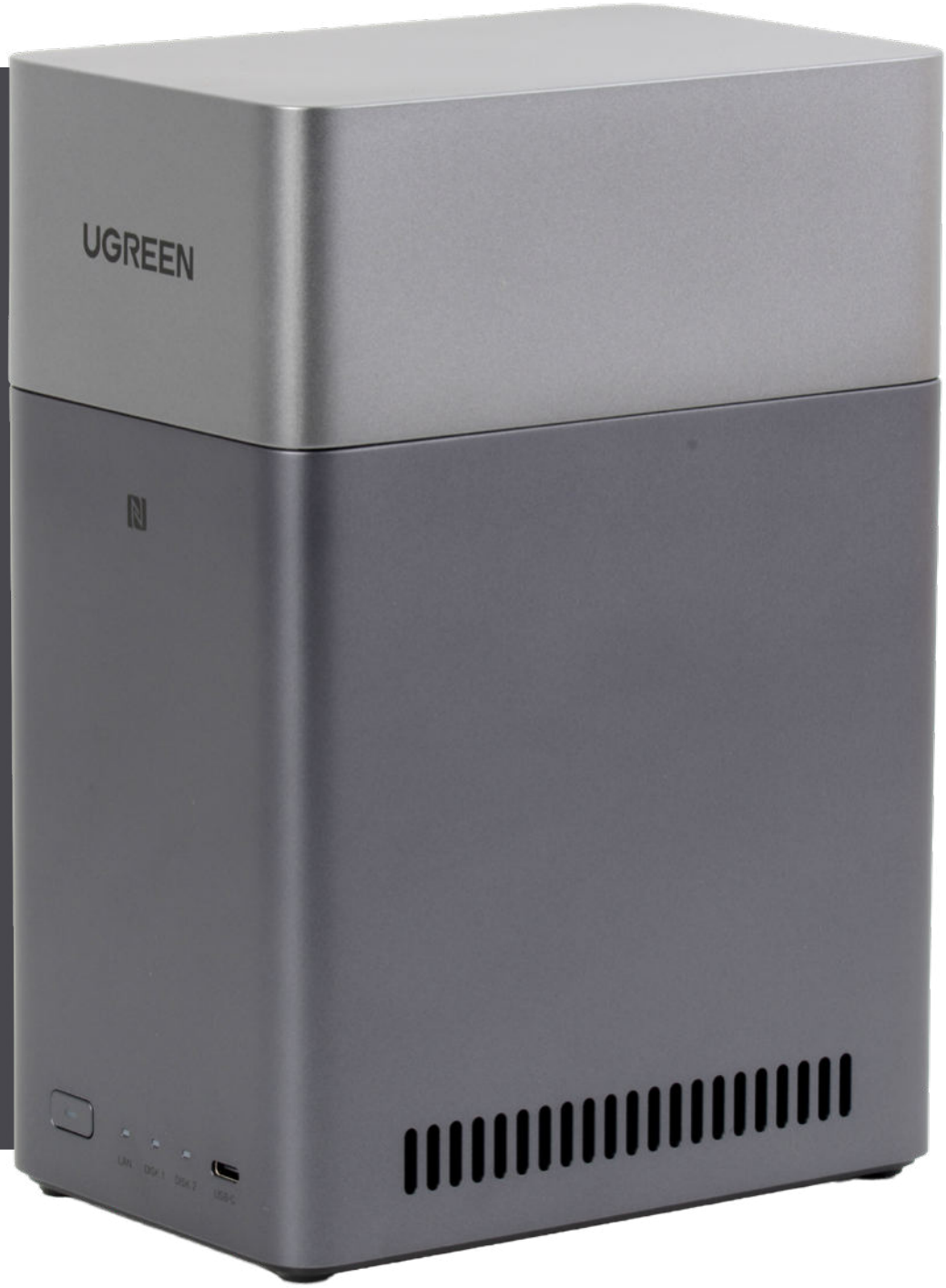
LE CHALLENGER EST AUSSI LE MEILLEUR CHOIX

Face à l'hégémonie des services de stockage en ligne, le Ugreen NASync DH2300 offre une réponse locale et accessible. Compact et intuitif, ce NAS à deux baies permet aux particuliers comme aux TPE de bâtir leur propre cloud privé pour gérer et protéger leurs fichiers essentiels à un tarif particulièrement agressif.

Prix constructeur et constaté : 188,99 €

→ par Thomas Olivaux

Ugreen, que l'on connaît bien pour son expertise dans le domaine de la connectique et des accessoires mobiles, fabrique aussi des NAS depuis quelques temps. Une gamme baptisée NASync et que nous n'avions jamais essayée jusqu'ici. La sortie récente de deux nouveaux modèles d'entrée de gamme en deux et quatre baies nous a donné envie d'en savoir plus. Nous essayons aujourd'hui le DH2300. Ce serveur de stockage en réseau (NAS), typé entrée de gamme, a pour ambition de transformer votre foyer en un véritable cloud personnel de confiance. Doté de deux baies SATA pouvant accueillir un total de 60 To de données, il permet de regrouper au même endroit vos clichés, vos films et



vos documents divers. La sécurité n'est pas oubliée avec la présence d'un chiffrement AES et d'un système d'authentification à deux facteurs. Sous le capot, le constructeur a misé sur un processeur Rockchip RK3576 pourvu de huit cœurs, épaulé par 4 Go de mémoire vive, ce qui lui confère une belle réactivité pour les opérations de routine, les sauvegardes automatiques ou la diffusion de flux 4K sur un téléviseur grâce à son port HDMI. À la fois discret acoustiquement et sobre en consommation électrique, ce modèle se paramètre en à peine 10 minutes via l'application dédiée, mettant ainsi la souveraineté numérique à la portée de tous, loin des services cloud payants traditionnels. Parfait pour un usage familial, il profite de fonctions d'intelligence artificielle pour l'organisation de vos photos par visages ou par lieux, tout en garantissant un accès à distance fluide via web ou mobile. Proposé initialement à 209,99 €, son tarif a déjà glissé à 188,99 € sur la boutique officielle, prix que l'on peut encore réduire de 5 % en s'abonnant simplement à la newsletter. Les acteurs historiques du secteur ont du souci à se faire !

LE TOUR DU PROPRIÉTAIRE

Le DH2300 est livré avec tout le nécessaire : son bloc d'alimentation, les vis pour fixer des supports 3,5" ou 2,5" (HDD comme SSD) et même un petit tournevis pour l'installation. Son esthétique tout en hauteur détonne dans l'univers des NAS ; il rappelle davantage le look de certaines box Internet. Sur la partie basse de la façade, on retrouve l'interrupteur d'allumage, les indicateurs LED (état du réseau et des deux unités de stockage) ainsi qu'une connectique USB-C à 5 Gb/s. Le panneau arrière regroupe quant à lui deux ports USB-A (5 Gb/s), la prise réseau RJ-45 (Gigabit Ethernet), une sortie HDMI assez inhabituelle sur ce segment et l'entrée pour l'alimentation électrique. L'appareil repose sur quatre pieds antidérapants qui créent un espace de quelques millimètres en dessous du châssis, permettant ainsi à l'air frais de pénétrer par la base. Le circuit de refroidissement est complété par des entrées d'air latérales et une large grille d'extraction située tout en haut, au dos du boîtier.

Vous pouvez aisément brancher, et même retirer à chaud, des disques et SSD externes USB.

FICHE TECHNIQUE

Marque : Ugreen

Modèle : NASync DH2300

Baies de stockage SATA : *2 (2,5 et 3,5")

Emplacements SSD M.2 : N/A

Processeur : Rockchip RK3576 (ARM 8 cores à 2,2 GHz + NPU)

Mémoire vive : 4 Go LPDDR4X, non extensible

Réseau : 1 GbE

USB : USB-C en façade (3.2 Gen1 5 Gb/s) + 2* USB-A au dos (3.2 Gen1 5 Gb/s)

RAID : JBOD, RAID 0, RAID 1

OS : Ugreen UGOS Pro

Systèmes de fichiers : Ext4, Btfrs

Dimensions : 15,1, x 19,8 x 21,4 cm

Garantie : 2 ans

Divers : eMMC 32 Go pour l'OS, sortie vidéo HDMI 4K, NFC pour simplifier la configuration

QUELQUES VIS ET PUIS C'EST TOUT

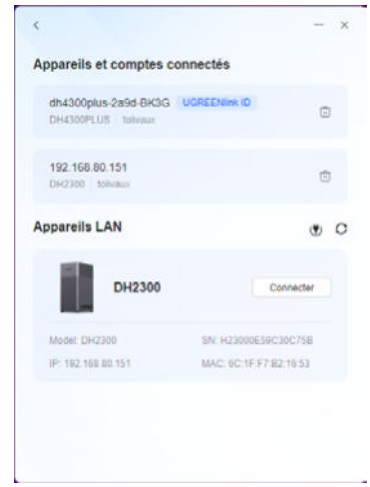
Pour accéder aux entrailles de la machine, il suffit de retirer le panneau supérieur. D'un gris plus clair que le reste du boîtier et floqué du nom de la marque, ce couvercle est simplement aimanté : il se retire en le tirant vers le haut, de préférence en maintenant la base de l'autre main pour plus de stabilité. Une fois ouvert, on accède aux deux paniers amovibles destinés à recevoir vos disques SATA. Si l'usage le plus courant sera d'y loger des disques durs 3,5", le système accepte sans broncher des unités 2,5" (HDD ou SSD). Comme c'est la règle dans ce domaine, le NAS est vendu « nu », sans stockage préinstallé. Pour les besoins de ce banc d'essai, nous l'avons équipé de deux unités Seagate IronWolf de 4 To, des modèles spécifiquement optimisés pour fonctionner dans un NAS. L'extraction des berceaux

se fait en pinçant les deux ergots situés à leur sommet avant de les tirer verticalement. Les supports de stockage doivent être vissés à l'intérieur de ces paniers (ou « caddys »).

Si le guide d'utilisation préconise l'usage de six vis par disque, quatre suffisent largement à assurer leur maintien. On pourrait s'étonner de ne pas trouver de rondelles de découplage en caoutchouc pour amortir les vibrations mécaniques au niveau des vis, mais Ugreen a résolu le problème autrement : les bases des tiroirs sont équipées d'un revêtement spécial absorbant sur les zones de contact avec le châssis interne. Et le résultat est là : dès la mise sous tension, on est frappé par la discrétion de l'ensemble. Le léger ronronnement des disques durs est à peine perceptible et ne se propage pas sur notre bureau en bois, pourtant réputé pour faire office de caisse de résonance.

QUELQUES MINUTES SUFFISENT

La mise en place logicielle brille par sa flexibilité et sa simplicité. La méthode la plus intuitive consiste à utiliser son smartphone et l'application Ugreen NAS (disponible sur iOS et Android). Pas besoin de la chercher sur les stores : il suffit de passer son téléphone sur le logo NFC présent sur la façade du NAS pour que le lien de téléchargement apparaisse. Ugreen a également eu la bonne idée de décliner son logiciel sur PC (Windows 32/64 bits) et Mac (processeurs Intel ou puces Apple M), et même sur les systèmes TV comme Android TV ou l'Apple tvOS. Pour ceux qui rechignent à installer quoi que ce soit, une simple fenêtre de navigateur web pointée sur l'adresse find.ugnas.com permet de repérer l'appareil sur le réseau local. L'interface, tout comme la documentation papier, est intégralement traduite en français. La configuration de base ne demande que quelques instants. Après avoir nommé votre serveur (nom qui apparaîtra sur votre réseau), vous devrez créer le compte administrateur et renseigner une adresse email. Cette dernière sera utilisée pour le service d'accès distant simplifié et pour recevoir des notifications importantes (comme une coupure de courant ou une indisponibilité système).



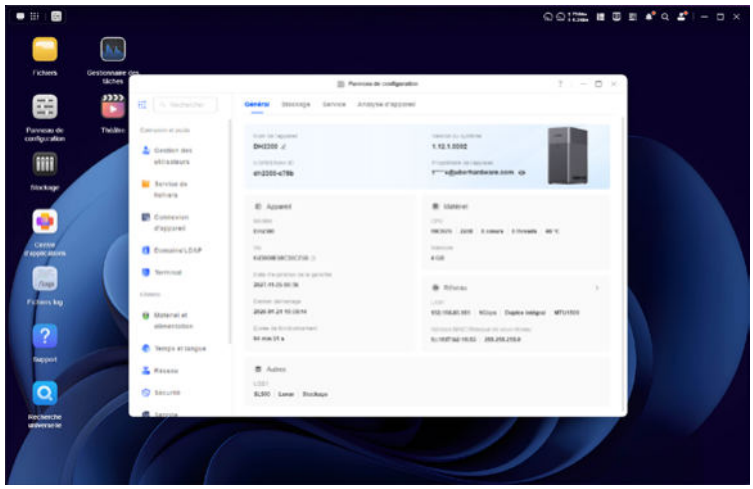
Le système d'exploitation s'installe ensuite de manière autonome. Il faut noter une particularité bienvenue : contrairement à beaucoup de NAS qui utilisent une partie de vos disques de stockage pour héberger leur système, Ugreen a intégré une puce eMMC de 32 Go dédiée exclusivement à l'OS. C'est une excellente initiative qui booste la réactivité (l'eMMC fonctionnant comme un petit SSD) tout en préservant l'intégralité de l'espace de vos disques durs.

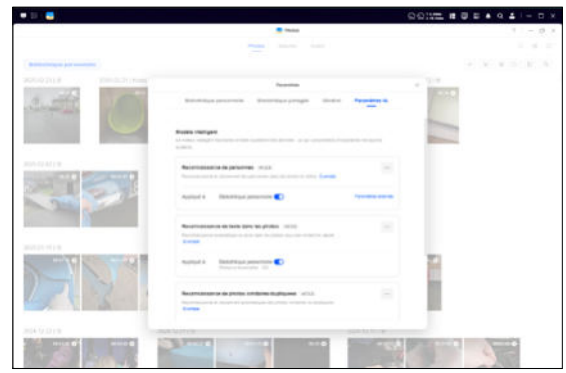
PREMIERS PAS SOUS UGOS

Une fois la phase d'installation terminée, on découvre une interface graphique soignée, agrémentée d'un fond d'écran élégant et d'icônes pour les différentes applications. Des widgets permettent de surveiller d'un coup d'œil les ressources système et la charge matérielle. Un petit guide interactif prend alors le relais pour vous apprendre à créer votre premier espace de stockage via l'utilitaire dédié. La procédure consiste à choisir vos disques et à définir le type de « grappe » souhaité. Sur ce modèle à deux baies, vous avez le choix entre plusieurs configurations : le RAID 0 (qui cumule les performances mais double le risque de perte de données), le RAID 1 (qui clone les données d'un disque sur l'autre pour une sécurité maximale), le mode Basic (gestion indépendante) ou encore le JBOD. Nous avons opté pour le RAID 1, ce qui nous laisse une capacité utile de 3,7 To (celle d'un seul disque). Le formatage a été effectué en ext4, un système de fichier robuste et performant sous Linux. Le Btrfs est également proposé pour ceux qui veulent profiter des « instantanés », même si l'ext4 reste une valeur sûre en termes de stabilité. Une fois la grappe validée, elle est opérationnelle, bien qu'il faille compter environ une journée pour que l'initialisation complète se termine et offre les meilleures performances.

UN LOGICIEL FLUIDE ET INTUITIF

On poursuit par la création de dossiers pour organiser et partager vos contenus. Le système UGOS gère intelligemment les répertoires personnels, créant automatiquement un espace privé pour chaque utilisateur, à l'instar du dossier « Mes documents » sur un ordinateur. Bien entendu, vous pouvez créer des « Dossiers partagés » accessibles à plusieurs personnes, en définissant les droits d'accès de chacun depuis le Panneau de configuration. L'application « Fichiers » centralise toute la gestion des documents. Enfin, le tutoriel se termine sur la découverte de « Ugreenlink », un service d'accès distant très simple. Il permet de se connecter à son NAS depuis n'importe où via un lien unique, évitant ainsi de plonger dans les réglages complexes de redirection de ports sur votre box internet.



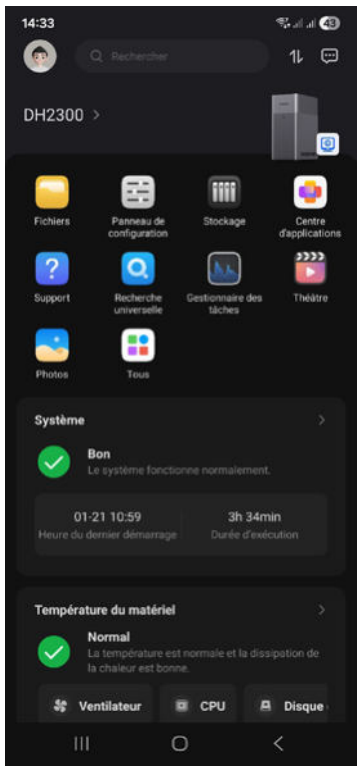


Les fonctions IA appliquées à la bibliothèque de photos exploitent le NPU du SoC Rockchip RK3576.

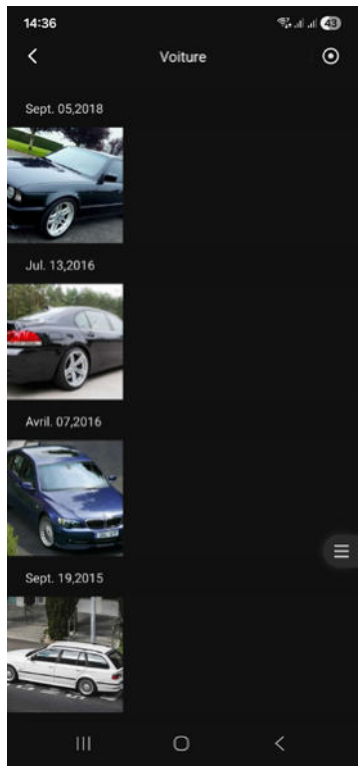
EXPLOITER TOUT LE POTENTIEL DU NAS

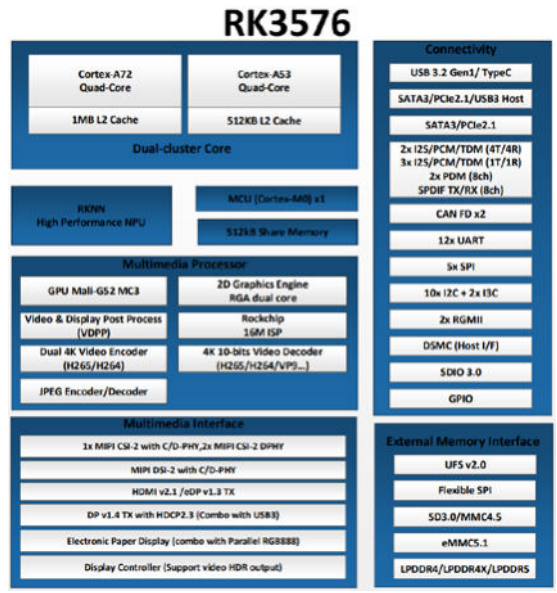
On peut notamment y retrouver sa bibliothèque de photo avec reconnaissance des images par IA. Pour voir vos dossiers partagés directement dans l'explorateur Windows, il suffit d'activer le protocole SMB dans le Panneau de configuration (onglet Service de fichiers). Après un redémarrage, le NAS devient une extension de votre PC. Vous pouvez également activer des protocoles plus experts comme le FTP ou Rsync. Le système regorge d'options pratiques, comme la programmation du démarrage automatique après une coupure de courant ou la gestion de la mise en veille des disques. Le « Centre d'applications » permet d'étendre les capacités de la machine. On y trouve par exemple « Cloud Drives », qui permet de synchroniser ses fichiers entre différents appareils et le NAS, créant ainsi votre propre version de OneDrive ou Google Drive, mais sans abonnement mensuel et avec vos données stockées physiquement chez vous. On a aussi particulièrement apprécié Photos, une des rares applications capables d'exploiter le NPU intégré au processeur Rockchip, d'une puissance modeste de 6 TOPS. Voilà qui permet, sans même faire ramer le NAS, des recherches IA performantes. À conditions d'installer manuellement des extensions dans l'application Photo, le NAS est capable d'identifier des personnes, des animaux, des objets (voitures, ordinateurs) et même du texte dans vos photos, comme ce que propose Google Photos. Et ça marche bien ! Côté performances pures, l'expérience est excellente : l'interface ne souffre d'aucune latence et les transferts de fichiers plafonnent à 113 Mo/s, que ce soit en lecture ou en écriture. C'est la limite maximale d'une prise Ethernet Gigabit ; les disques durs ne sont donc pas le frein, c'est le réseau qui sature. C'est toutefois déjà très confortable, dépassant souvent la vi-

Le test du DH4300 Plus, avec un CPU plus puissant et du 2,5 GbE, sera publié dans UH #13.



L'application Ugreen Nasync existe aussi sur smartphone.



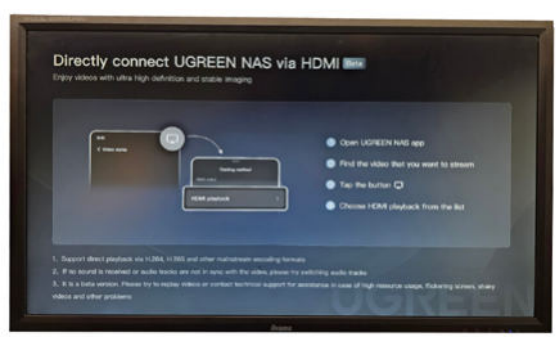


Les capacités de décodage du RK3576 sont bonnes mais font l'impasse sur le format AV1.

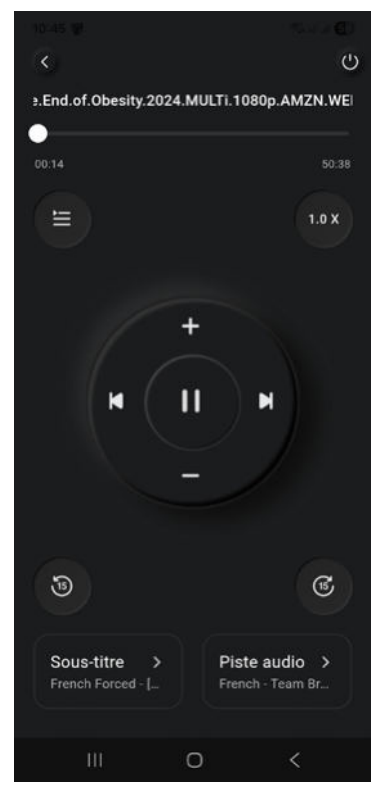
tesse de nombreuses connexions internet, même en fibre.

DES CAPACITÉS MULTIMÉDIA SURPRENANTES

Un bon NAS se doit de briller sur le terrain du multimédia. Ugreen propose pour cela trois applications dédiées : Photos, Musique et Théâtre. Chacune offre une interface simplifiée pour profiter de ses médias. Le DH2300 se démarque surtout par sa sortie HDMI 4K @60Hz, une option rarissime sur cette gamme de prix. En reliant directement le NAS à votre télévision, vous pouvez diffuser vos films sans passer par le réseau. En l'absence de télécommande physique, le smartphone sert de centre de contrôle. Bien que le service soit encore identifié comme étant en version bêta, nos tests ont été concluants : lecture fluide de films en 1080p et 4K, support des formats H.264 et H.265 (même avec des débits dépassant les 120 Mb/s), gestion des pistes audio et des sous-titres… Le processeur Rockchip démontre ici toute sa puissance en matière de décodage vidéo !



En branchant le DH2300 sur votre téléviseur, il peut lire les vidéos H.264 et H.265 qu'il stocke…

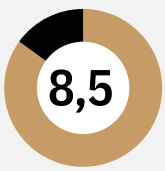


…à condition de piloter la lecture par votre smartphone.

CONCLUSION

Pour un coup d'essai, c'est un coup de maître : le Ugreen NASync DH2300 nous a totalement convaincus ! Il s'impose d'emblée avec un rapport qualité/prix qui semble aujourd'hui sans équivalent sur le marché. Proposé au même tarif qu'un Asustor Drivestor 2 Lite (AS1102TL) ou qu'un QNAP TS-233, et même moins cher qu'un Synology DS223j, il les éclipse pourtant techniquement : on profite ici d'un processeur à 8 cœurs (contre 4 chez ses trois rivaux) et surtout de 4 Go de RAM, là où la concurrence se contente de 1 Go (Asustor/Synology) ou 2 Go (QNAP). Certes, les systèmes DSM de Synology, QTS de QNAP ou même ADM d'Asustor restent les références absolues en termes de maturité, avec des catalogues d'applications très fournis. Toutefois, le système UGOS Pro de Ugreen nous a agréablement surpris. S'il fait l'impasse sur certaines fonctionnalités avancées — comme un équivalent à Surveillance Station pour la gestion des caméras —, il assure l'essentiel avec brio. Le logiciel est fluide, réactif, parfaitement traduit et, surtout, d'une stabilité exemplaire. Malgré la jeunesse de la marque sur ce secteur, le produit respire la finition et le sérieux : aucun plantage n'est venu entacher notre test. Que pourrait-on lui reprocher au final ? Bien peu de choses, si ce n'est l'absence d'un port Ethernet à 2,5 Gb/s qui aurait permis de doubler les débits pour atteindre les limites physiques des disques durs. Mais c'est là l'exclusivité d'une autre nouveauté Ugreen, le NASync DH4300 Plus à 4 baies, dont nous vous proposerons le test dès le prochain numéro. En attendant, avec la compatibilité pour les nouveaux disques durs allant jusqu'à 30 To, ce petit DH2300 a tous les atouts pour séduire une immense majorité d'utilisateur, privés et pros. Bravo Ugreen !

UGREEN NASYNC DH2300



9	Prix
7,5	Performances et débits
8,5	OS et interface utilisateur
8	Connectique
9	Consommation

Points forts

- Prix
- Sortie HDMI 4K
- OS sur puce eMMC 32 Go
- Installation Express
- NPU dédiée à l'IA…

Points faibles

- … mais NPU peu puissant et à limité à Ugreen Photos
- Port Ethernet limité à 1 Gb/s
- RAM non extensible
- Pas d'emplacement M.2
- Ecosystème logiciel encore jeune
- Incompatible avec Docker

INSTALLER WINDOWS 11 SUR UN VIEUX PC

OUTREPASSER LES LIMITES MATÉRIELLES ET LE COMPTE MICROSOFT AVEC RUFUS

→ par Thomas Olivaux

Grâce à Rufus, l'indétrônable, il n'a jamais été aussi facile d'installer Windows 11 sur un PC non supporté et/ou de faire l'impasse sur le compte Microsoft. On peut même aller plus loin avec FlyOOBE.

Depuis son lancement, Windows 11 s'est forgé une réputation de système exclusif, érigeant des barrières techniques souvent jugées arbitraires par la communauté hardware. Entre l'exigence d'une puce TPM 2.0 (Trusted Platform Module), le support du Secure Boot et une liste de processeurs compatibles laissant sur le carreau d'excellentes puces d'ancienne génération, comme les Ryzen 1000 ou les Intel Core de 7e génération, Microsoft a sciemment restreint son parc. Pourtant, le noyau de Windows 11 reste structurellement proche de celui de Windows 10, ce qui rend ces limitations essentiellement logicielles. L'objectif de ce tutoriel est d'utiliser le bon vieil utilitaire Rufus, dont les versions récentes, et en particulier les 4.x ont été pensées pour Windows 11, pour créer un média d'installation "augmenté", capables d'ignorer ces vérifications et de vous redonner le contrôle sur votre compte utilisateur.

CE QU'IL FAUT VRAIMENT SOUS LE CAPOT

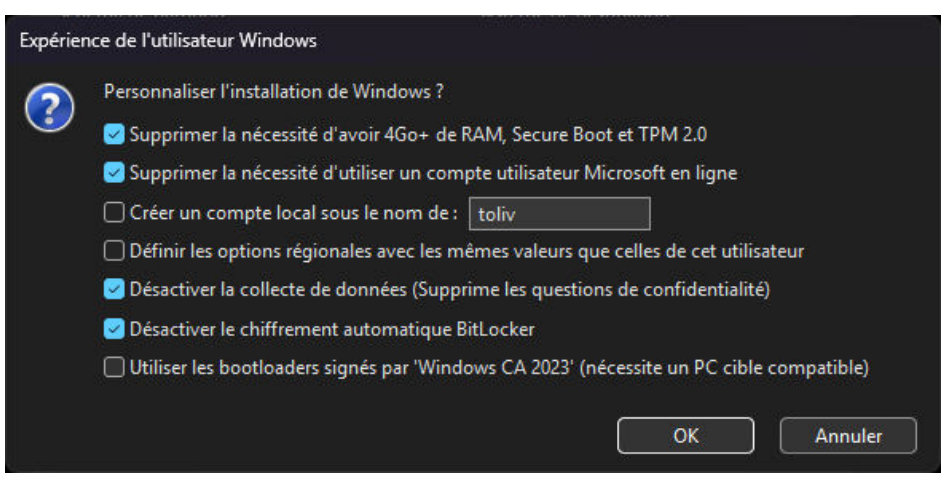
Avant de forcer le passage, un point sur le hardware s'impose. Si Microsoft impose des barrières strictes, la réalité technique est plus souple, mais pas infinie. Vous pouvez toujours tenter de faire tourner Windows 11 sur un CPU dual core (64 bits obligatoires) et 4 Go de RAM, mais vous aurez envie de jeter l'ordinateur par la fenêtre dans la journée. Pour un tant soit peu de confort, visez un quad core, au moins de génération Sandy Bridge (Core de 2e génération), l'idéal étant un Core de 4e génération (Haswell) ou un AMD Ryzen de 1ère génération, accompagnés de 8 Go de RAM. Mais le véritable enjeu, c'est le stockage : l'installation sur un disque dur mécanique (HDD) est à proscrire. Windows 11 multiplie les petits accès disques pour ses services d'arrière-plan et la télémétrie ; sans un SSD (même un modèle SATA d'entrée de gamme), l'expérience sera entachée de lags insupportables, quel que soit votre processeur.

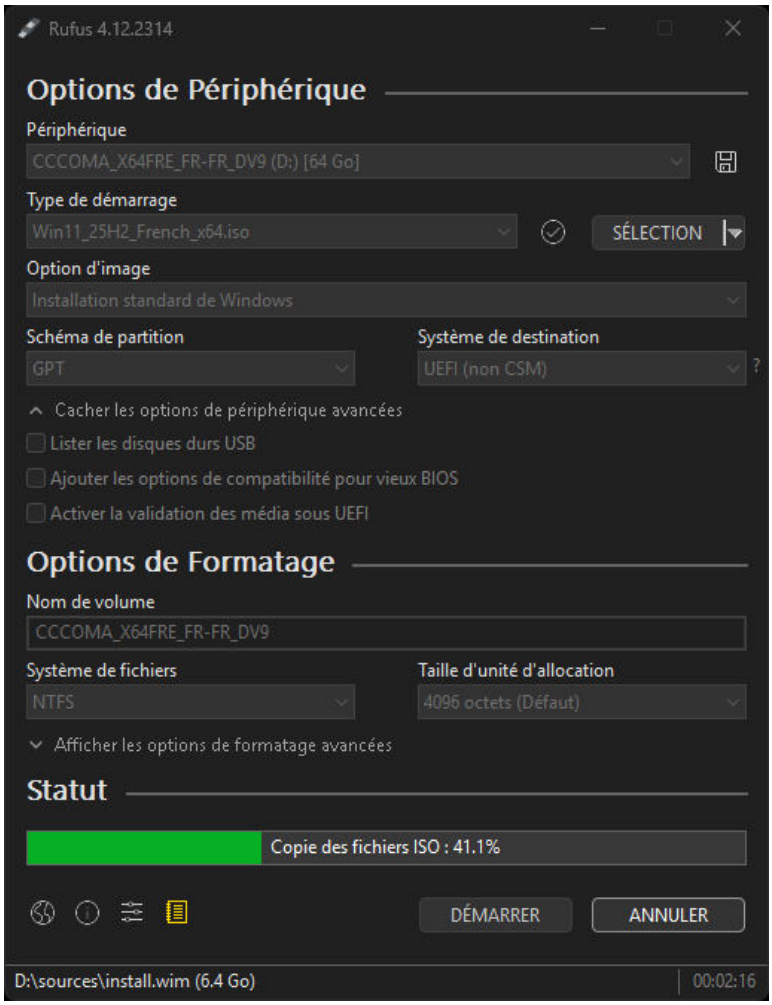
> Ce qui compte vraiment pour un usage agréable, outre un CPU 64 bits et au moins 8 Go de RAM, c'est d'avoir un SSD (même un vieux SATA).

RUFUS : UN HACK CHIRURGICAL EN TROIS CLICS À PEINE

La première étape consiste à récupérer l'image disque (ISO) officielle de Windows 11. Pas besoin de lancer votre client torrent, ça se passe directement chez Microsoft, sans clé ni carte bancaire. Une fois muni de ce fichier et d'une clé USB de 8 Go minimum, téléchargez Rufus, ouvrez-le et sélectionnez votre périphérique. Assurez-vous de régler le schéma de partition sur GPT pour les bios UEFI, ou MBR si vous tentez l'aventure sur une machine très ancienne dotée d'un BIOS traditionnel (les premiers UEFI sont apparus en 2011, mais certaines marques ont mis quelques années à franchir le pas).

En cliquant sur Démarrer, une

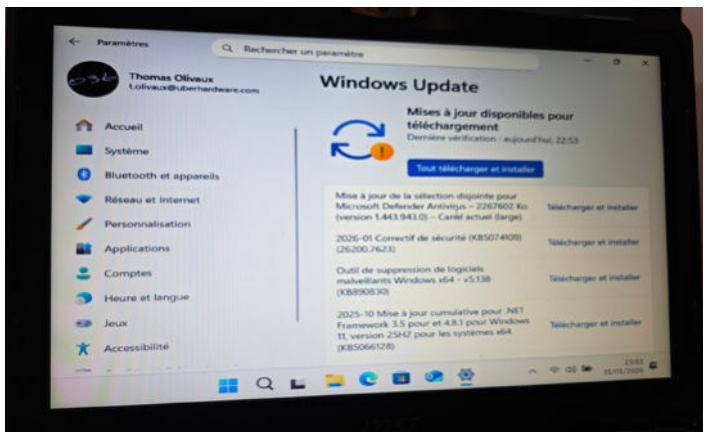




fenêtre cruciale nommée Windows User Experience apparaît. C'est ici que vous neutralisez les verrous. En cochant Remove requirement for 4GB+ RAM, Secure Boot and TPM 2.0, Rufus injecte des scripts dans le registre de l'installateur (via les clés AllowUpgradesWithUnsupportedTPMOrCPU et LabConfig), trompant ainsi l'utilitaire AppraiserRes.dll. Parallèlement, cocher Remove requirement for an online Microsoft account est l'option salvatrice pour la vie privée : elle automatise le contournement du script de connexion obligatoire (ce qu'on faisait il y a quelque temps avec OOBE\BYPASSNRO durant l'installation, mais ça ne fonctionne plus). Rufus peut même pré-créer votre compte local, vous épargnant les questions intrusives et les multiples pages de réglages de confidentialité au premier démarrage.

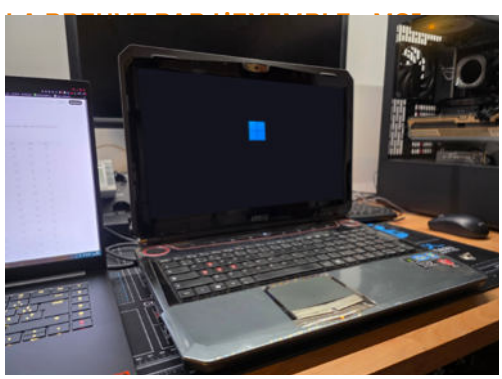
L'ALTERNATIVE FLYOOBE : LE « LIVE-PATCHING » SANS PRÉPARATION

Si Rufus reste la référence pour préparer sa clé en amont, FlyOOBE (pour Fly Out Of Box Experience) propose une approche radicalement différente et particulièrement élégante : le contournement "à la volée". Imaginez que vous ayez sous la main une clé USB d'installation Windows 11 tout à fait officielle, créée avec l'outil de Microsoft, et que vous fassiez face au message fatidique : "Ce PC ne peut pas exécuter Windows 11". C'est ici que FlyOOBE entre en scène. En pressant simplement Shift + F10 pour ouvrir l'invite de commande, il suffit de taper une ligne de commande PowerShell (irm flyoobe.com | iex) pour lancer un script qui va modifier le registre en temps réel. Ce script ne se contente pas de faire sauter les verrous TPM 2.0, Secure Boot ou les limites de RAM ; il intercepte le processus d'installation pour vous offrir un menu complet. Contrairement à Rufus qui agit de manière statique, FlyOOBE est dynamique : il permet de désactiver les publicités du menu Démarrer, de supprimer les widgets inutiles, d'empêcher l'installation automatique de Microsoft Teams ou encore d'activer d'emblée le mode sombre. C'est l'outil parfait pour ceux qui ne veulent pas s'embêter à préparer une clé spécifique, mais préfèrent hacker l'installateur officiel une fois celui-ci lancé. Nous le gardons en alternative car il nécessite une connexion internet (ou le téléchargement du script au préalable) pour être exécuté via l'invite de commande, là où Rufus fonctionne en mode "hors-ligne" total dès le boot, et qu'il représente un peu plus d'étapes, mais au profit de bien plus de fonctionnalités.



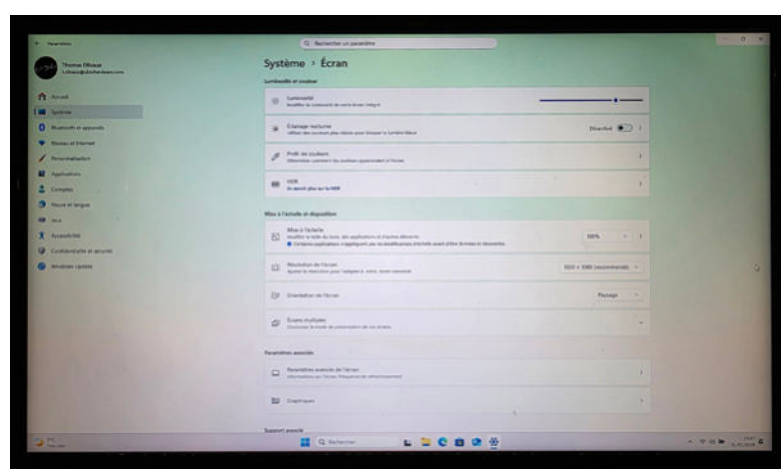
APRÈS L'INSTALLATION, DOMPTER LES DRIVERS RÉCALCITRANTS

Une fois Windows 11 installé, le défi principal reste celui des pilotes. Pour le matériel "non supporté", Windows Update fait souvent des miracles en piochant dans la base de données de Windows 10, mais des points d'interrogation peuvent subsister dans le Gestionnaire de périphériques. La règle d'or : un pilote Windows 10 fonctionne à 99% sur Windows 11. Si votre constructeur s'arrête à Windows 7 ou 8 (notamment pour de vieux chipsets audio ou des lecteurs de cartes), tentez une installation manuelle en mode de compatibilité. Pour les cas désespérés, oubliez les sites de drivers douteux et tournez-vous vers Snappy Driver Installer Origin (SDIO). Cet outil open source est une mine d'or pour le hardware vintage : il analyse l'ID matériel de vos composants et trouve le pilote le plus proche dans sa base de données massive. Enfin, gardez en tête que sur les vieux GPU (type Intel HD 4000 ou vieilles cartes NVIDIA), l'absence de pilotes WDDM 2.0 peut désactiver certaines transparences de l'interface, mais n'empêchera pas le système d'être parfaitement fonctionnel pour de la bureautique ou du web.



GT683DX DE 2011

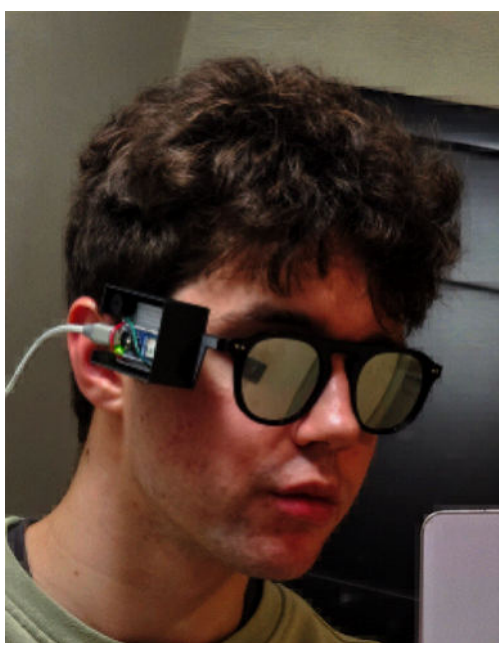
Cet article devrait avant tout séduire les propriétaires de CPU tels que le Core i7 7700K, encore parfaitement utilisables au quotidien, et que Microsoft a décidé d'exclure de sa liste de compatibilité. Mais, comme nous l'avons évoqué en début d'article, on peut tout à fait pousser l'expérience un peu plus loin. Pour illustrer nos propos, nous avons installé, grâce à une clé crée sous Rufus en 5 minutes à peine, Windows 11 25H2 sur un vieux laptop gamer MSI GT683DX de 2011, sorti du temps de Windows 7 ! Sachant qu'on a laissé à l'intérieur son disque dur de 500 Go, et que son modeste Core i7-2630QM n'est épaulé que par 8 Go de RAM, il a fallu jouer de patience, mais ça fonctionne tout à fait bien. On a même poussé le vice à l'installer avec un vrai compte Microsoft, ce qui était facilité par la prise en charge native du Wi-Fi. L'installation s'est faite dans une résolution ridicule faute de pilote, mais, une fois arrivé sous Windows et après un tour sous Windows Update, la GeForce GTX 570M est reconnue et on peut passer en 1920x1080. Une fois sous Windows, le CPU n'est pas aussi sollicité qu'on pourrait le croire ! En remplaçant le HDD par un SSD SATA, ce PC pourrait tout à fait servir à surfer et même lancer quelques vieux jeux. Après tout, en 2011, on jouait déjà à Minecraft !



Uber Geek

Bienvenue dans cette nouvelle rubrique Uber qui célèbre les objets qui font vibrer notre fibre geek, qui déclenchent un sourire immédiat, un “wow” silencieux ou une envie irrépressible de démonter, bidouiller, brancher, explorer… et dépenser. Entre pop culture iconique, retrogaming qui sent la cartouche chaude, glitchs fossilisés, gadgets high-tech un peu fous, casques VR/AR à construire soi-même, Lego qui réveillent le collectionneur qui sommeille en vous et même un livre pour apprivoiser le langage de programmation le plus utilisé dans le monde vidéoludique, cette rubrique est un terrain de jeu assumé pour nos coups de cœur.

→ par Manu Da Costa

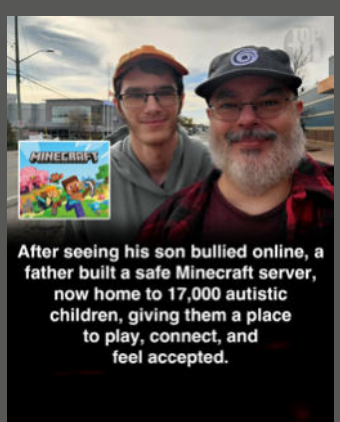


COMMENT NEUTRALISER LES LUNETTES INTELLIGENTES QUI VOUS ESPIONNENT

Si vous avez vécu dans une grotte ces derniers temps, vous n’aurez peut-être pas entendu parler des smart glasses qui sont des lunettes intelligentes intégrant des capteurs, caméras, micros, haut-parleurs et de l’IA pour enrichir notre perception. Parmi les modèles les plus populaires et les plus vendus, citons les Ray-Ban Meta Smart Glasses nées d’une collaboration entre Meta et EssilorLuxottica. Si les applications sont nombreuses en pratique, les risques inquiètent tout autant avec l’enregistrement ambigu des personnes filmées dans la rue sans consentement, la collecte des données par Meta et les vulnérabilités du système exposant l’utilisateur en cas de piratage, une manipulation des informations contextuelles. Ce qui inquiète bien sûr les régulateurs européens et les défenseurs de la vie privée. C’est à partir de ce constat alarmant que Peter Whiting, aka NullPxl, a choisi de créer le projet open source Ban-Rays dont le but est de créer des lunettes capables de repérer cette forme d’espionnage moderne menée par une bande de petits curieux équipés de Ray-Ban Meta ou de modèles concurrents. Une expérimentation qui s’appuie pour l’instant sur un Arduino Uno, des LED infrarouges, une photodiode et un transistor (évolutif), et qui utilise deux approches : Optics et Networking. La première est une méthode de classification de la caméra à l'aide des reflets lumineux (l'effet "cat-eye"), alors que la seconde analyse les réseaux Bluetooth et Wi-Fi mais nous vous laissons de loisir de découvrir en détail le projet si vous êtes curieux, voire d’y contribuer afin de l’améliorer vu que le projet en est à ses débuts, tout le code étant en C/C++ et Python.

TETRIS AU POIGNET

Après avoir lancé les montres connectées ***Space Invaders*** et ***Atari 2600***, My Play Watch revient avec une troisième version. En collaboration avec The Tetris Company, la Tetris: My Play Watch ressemble à une Apple Watch avec son cadran en acier inoxydable au design carré et ses coins arrondis, sa couronne fonctionnelle et son écran tactile couleur de 1,91 pouce. Mais la comparaison s'arrête là puisque la montre est livrée avec un bracelet interchangeable et un cordon magnétique de recharge, n'offre aucune connectivité smartphone. Mais qu'importe, on adore cette smartwatch qui offre un peu plus que les fonctionnalités de base d'horloge et de calendrier, puisqu'en plus d'être classe, elle permet de jouer à Tetris, embarque une série de cadrans personnalisables sur le thème de Tetris et permet de suivre son rythme cardiaque, ses pas et vos calories brûlées grâce au moniteur d'activité physique intégré. La Tetris: My Play Watch paraîtra le 20 février 2026 et sera disponible sur Amazon au prix de 79,99 €.



UN MONDE VIRTUEL SANS PRESSION SOCIALE

En 2013, Stuart Duncan, un développeur web et père célibataire de Timmins (Ontario au Canada) connu sous le pseudo d'AutismFather, voit son fils Cameron harcelé sur les serveurs publics Minecraft. Un phénomène encore trop répandu, soulignant combien la bêtise humaine peut être démesurée. Alors qu'il aurait pu abandonner face aux vagues d'intimidations et d'insultes, il choisit au contraire de lancer Autcraft : un serveur privé destiné aux autistes et même à leurs parents, modéré 24/7, avec des règles strictes de respect et une aide pour résoudre les conflits. Initialement prévu pour une douzaine d'enfants, le serveur est victime de son succès avec pas moins de 750 inscriptions dès le lancement ! Autcraft est administré par des adultes atteints ou non de troubles mentaux, ainsi que par leurs amis ou leur famille et en décembre 2025, le serveur comptait plus de 19 000 joueurs inscrits sur la liste blanche ! Une aventure humaine et un exemple que nous ne pouvions pas taire d'autant plus que Mojang en personne en avait profité pour rencontrer Stuart et faire une interview.

TIKI KIWI

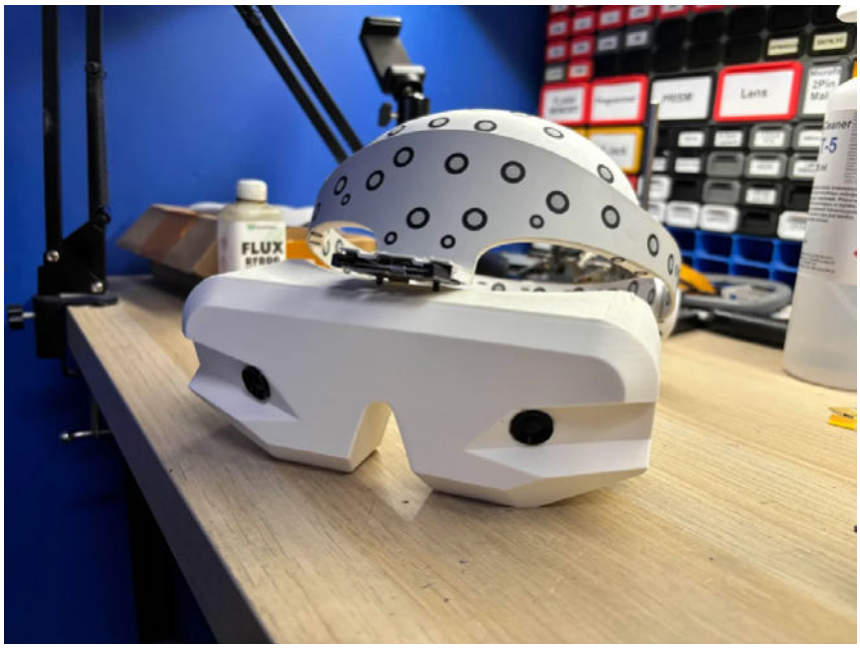
Entreprise japonaise mythique, Taïto incarne l'âge d'or des jeux d'arcade dans les années 80 avec le cultissime Space Invaders (1978) puis des jeux d'arcade très populaires comme Qix, Elevator Action, Arkanoid, Bubble Bobble, Rainbow Islands: The Story of Bubble Bobble 2, Operation Wolf, Chase H.Q. Avec son petit côté 2 flics à Miami, Liquid Kids, ou encore Puzzle Bobble (Bust-a-Move). Des titres incontournables sur lesquels nous avons passé d'innombrables heures, à l'image de Bubble Bobble, un petit jeu coopératif en local au gameplay simple mais addictif où on incarnait deux petits dragons, Bub et Bob, crachant des bulles pour capturer et éclater des ennemis sur des plateformes colorées avec des pièges et des ennemis différents. Et malgré les remasters et réinterprétations modernes de classiques intemporels tels qu'Operation Wolf First Mission, Bubble Bobble 4 Friends ou Bubble Bobble Sugar Dungeons pourtant ultra séduisants sur le papier pour tout geek ayant fréquenté les salles d'arcade, la réalisation a manqué de magie laissant une impression un peu fade là où on espérait retrouver le frisson des bornes d'origine. Alors ne retenant pas les erreurs du passé, nous restons malgré tout excités à l'idée de redécouvrir dès le mois de février sur Steam, The New Zealand Story Untold Adventure, le remake officiel du fameux jeu de plates-formes tout mignon sorti en 1988, dans lequel on incarne Tiki, un petit kiwi (l'oiseau pas le fruit) devant sauver ses amis et sa compagne Phee-Phee enlevés par un léopard de mer à travers différents niveaux représentant l'archipel Néo-Zélandais. Une anecdote officiellement rapportée par un développeur de Taïto raconte que les niveaux auraient été inspirés d'un voyage en Nouvelle-Zélande (mais attendez, c'est ce qui explique le nom du jeu ?) et c'est l'éditeur Commodore Industries (qui propose aussi des PC Portables Commodore tel que l'ORION Slim 15) qui est derrière ce remake qui promet de respecter l'œuvre originale avec son level design non linéaire, sa difficulté parfois punitive, des améliorations artistiques et techniques mais aussi de nouveaux niveaux, de nouveaux ennemis et des boss inédits. Reste à espérer que la qualité sera au rendez-vous.

DES JUMELLES INTELLIGENTES ET UN HUD DE RÊVE

Ah, les macrojumelles, ces instruments qui transforment chaque détail en une expérience gigantesque et où l'infime devient spectaculaire. On pense inévitablement à Iron Man où Tony Stark teste une vision HUD dans son casque, Minority Report et Tom Cruise utilisant une interface gestuelle et des écrans interactifs transparents, ou encore le costume high tech de Spider-Man contenant un HUD qui permet de repérer ennemis et objets via vision augmentée. Des gadgets qui font rêver et qui ne sont pas seulement les fruits d'une imagination débordante à l'image des jumelles intelligentes Envision développées par Unistellar, une entreprise française spécialisée dans les télescopes intelligents qui s'est associée à la NASA et à l'Institut SETI en plus d'un partenariat avec Nikon. Ces dernières intègrent une optique de haute qualité ainsi que des fonctionnalités de réalité augmentée avancées qui délivrent quasi instantanément des informations contextuelles sur la terre comme dans le ciel. Utilisables de jour comme de nuit, ces jumelles intelligentes permettent par exemple en randonnée de lire les terrains, d'identifier des sentiers, des repères (sources d'eau, refuges, etc), d'afficher en temps réel de l'altitude et de l'azimut et même de verrouiller une cible dans le champ de vision. De nuit, vous pourrez suivre et observer les astres célestes tout en profitant d'informations comme leur nom, leur appartenance à la constellation, leur distance en années-lumière par rapport à la terre, etc. Mais ce n'est pas tout car grâce à son système de mises à jour, ces jumelles sont évolutives et nous avons hâte de découvrir les prochaines fonctionnalités de ces jumelles Envision primées au CES Las Vegas en 2025 et dont la commercialisation est prévue en 2027. N'hésitez pas à vous inscrire pour être averti de l'ouverture des prochaines précommandes, une première séance ayant déjà été bouclée au prix de 899€ au lieu de 1 499 €.

DIY : UN CASQUE XR MODULAIRE ET RÉPARABLE

Il existe parfois des situations injustes, des inventions technologiques pourtant géniales aussi bien sur le papier qu'en réalité et qui tombent malgré tout aux oubliettes, car boudées par les utilisateurs finaux, les développeurs et parfois même les industriels. Les casques de réalité virtuelle en sont une illustration parfaite. On pourrait croire que leur prix onéreux justifie un tel désamour de la part du public, mais alors comment expliquer l'engouement et les volumes de vente des smartphones très haut de gamme, certes incroyables techniquement mais proposant au final trop peu de fonctionnalités exploitant réellement ces capacités hardware. L'Apple Vision Pro semblait pourtant avoir offert un nouveau souffle dans le domaine des casques de réalité virtuelle, mais cette flamme passionnée s'est très rapidement éteinte. Pour autant, nous continuons de croire dans cette technologie et nous ne sommes pas les seuls. En témoigne cet énième projet open source pensé pour les makers désirant fabriquer leur propre casque AR/VR milieu de gamme tout en contenant au maximum le budget dans une fourchette de 800 à 850 dollars. Le projet est en perpétuelle évolution, mais s'appuie sur les processeurs RK3576 pour l'AR/XRUIOS, CIX P1 pour le gaming autonome, et carte SLAM dédiée (option G100 via GravityXR), un écran principal d'une résolution de 2160×2160 pixels à 90Hz (TFT LCD), une vision périphérique de 144 LED WS2812B (expérimentation FOV étendue), le suivi du mouvement des yeux et de la bouche, des haut-parleurs Tectonic BMR amovibles pour un son spatial, une connectivité Wi-Fi 6 / Bluetooth 5.4, etc. Les gants et contrôleurs open source sont en développement et vous pourrez suivre le projet sur le serveur discord du projet.



Que vous réalisiez votre casque VR vous-même ou non, vous pourrez toujours découvrir le travail de LukeRoss qui propose une application adaptant la VR à plusieurs dizaines de jeux, même si ce dernier s'est souvent retrouvé sous le feu des critiques, les studios de développement n'appréciant pas particulièrement le fait de proposer une application payante permettant de jouer en réalité virtuelle à leurs jeux. Ce qui a été récemment le cas avec le studio CD Project et son jeu Cyberpunk 2077.

DÉBLOQUER DES ÉMOTIONS LÉGENDAIRES

Maintes fois nous avons proclamé que le succès d'un jeu ne reposait pas uniquement sur son gameplay, ses graphismes et son scénario. La bande son originale d'un jeu est tout aussi important puisqu'elle rythme l'aventure et participe à créer une immersion dans un monde imaginaire et même plus, puisqu'il n'est pas nécessaire de jouer pour revivre les moments d'anthologie qui vous ont fait vibrer et cela, où que vous vous trouviez. Commençons par la bande originale de Hadès II composée par Darren Korb. Un cocktail de styles mélangeant du rock alternatif, de l'électro organique, des percussions tribales et des mélodies obsédantes amplifiant ce que nous avons pu ressentir à la fois dans l'urgence du combat, l'intimité des récits des personnages et l'euphorie. C'est comme si chaque morceau était un artefact musical prêt à exploser dans ton cerveau. Il faut dire que Darren Korb n'en est pas à son coup d'essai puisqu'il a également composé les BO du premier volet ainsi que Bastion et Transistor.

Non moins connu, Christopher Larkin qui a signé l'OST du premier Hollow Knight et de ses DLC (et scoop, il compose la bande son du prochain Moonlighter 2), a su une nouvelle fois mêler élégance acoustique et intensité orchestrale avec la bande originale de Hollow Knight: Silksong. De quoi revivre l'exploration et la beauté des lieux du royaume de Pharloom et susciter à travers ses 53 pistes, un poème orchestral et viscéral dont vous ne ressortirez pas indemne.

Terminons cette pause musicale en beauté avec des airs légendaires de The Legend of Zelda sublimés par le compositeur Rozen avec ce troisième opus de Ballads of Hyrule qui offre là aussi, une promenade imaginaire faisant ressortir un déluge d'émotions diverses où chaque morceau est comme un hymne racontant une légende oubliée. Et si vous n'avez pas encore découvert Ballads of Hyrule I & II et que vous ne souhaitez pas laisser cette quête secondaire inachevée ou être couverts de honte, vous savez ce qu'il vous reste à faire...

À LA RECHERCHE DES GLITCHS PERDUS

Le glitch désigne à l'origine un dysfonctionnement ou bug informatique, mais caractérise depuis quelques années déjà, une anomalie ou plutôt une interférence graphique. Un phénomène très répandu et populaire dans le monde vidéoludique d'autant plus que nombre de joueurs se livrent à une exploration intense et profonde dans le but de révéler au monde des réalités cachées. Et en ce sens, l'Archéologie du Glitch est dans ce sens une excellente surprise, un sujet fascinant et une exploration structurée basée sur cinq années de recherches menées par Hugo Montembeault dans le cadre de ses études doctorales. Il nous dévoile ainsi les coulisses du glitch au sein de nos jeux vidéo préférés, une réflexion économique, culturelle, artistique et parfois même politique. Parmi les glitchs les plus courants, citons le Wall-breach ou comment le joueur traverse un mur ou un sol censé être solide et se retrouve "hors map" (Blue Hell). Le Zip dans lequel le personnage se retrouve coincé entre deux objets et où le moteur physique panique et éjecte parfois le personnage à une vitesse supersonique. Mais le glitch est également utilisé comme un outil par les speedrunners : le speedrun, où comment sauter un niveau entier ou obtenir un objet clé sans faire la quête associée, effectuer des actions précises pour se téléporter directement au dernier niveau d'un jeu...Bref, un sujet passionnant pour lequel vous n'aurez besoin de débourser que la modique somme de 14,90 €.

LE MONDE NE VA PAS ÊTRE RÉÉCRIT EN RUST

C'est le débat le plus brûlant de la décennie dans le monde du développement système. Rust est effectivement brillant : il offre la sécurité mémoire garantie à la compilation (adieu les segmentation fault et les data races), possède un outillage moderne (Cargo) et les vidéos de louange sur ce nouveau langage ne manquent pas sur YouTube. Cependant, même face à la montée en puissance de Rust, apprendre le C++ reste à notre avis un investissement stratégique loin d'être inutile pour plusieurs raisons. "Too Big to Fail" : le monde ne va pas être réécrit en Rust, c'est une réalité économique et une grande partie de l'infrastructure mondiale tourne sous C++ à l'image des navigateurs web (Google Chrome, Edge, Safari et son moteur de rendu webkit, Firefox même si une partie migre vers Rust), des moteurs de jeux (Unreal, Unity), les systèmes d'exploitation (Windows, une partie de Linux, Android), la suite Adobe, Autodesk Maya / 3ds Max, Blender...Mais surtout, apprendre le C++ qui offre un écosystème de bibliothèques inégalé, c'est une aventure technique et créative qui pourrait transformer votre façon de penser et de coder en comprenant la machine, la mémoire et en maîtrisant les performances. D'autant plus que paradoxalement, cela vous prépare à devenir un excellent développeur Rust si vous souhaitez migrer plus tard vers ce langage. Par où commencer ? Par l'incontournable Guide du C++ moderne – de débutant à développeur (C++23), un livre pédagogique, didactique et complet qui vous accompagnera avec les bases (syntaxe, compilation, pointeurs, références, gestion de la mémoire), la STL : Conteneurs (vector, map, unordered_map), algorithmes, itérateurs, la Programmation Orienté Objet moderne (classes, héritage, polymorphisme), mais aussi les concepts introduits par C++23 qui est la norme actuelle sans oublier les bonnes pratiques de développement. Une référence incontournable dans le domaine qui existe en version papier et numérique. Mais dans le dernier cas, assurez-vous d'utiliser un lecteur compatible avec le MathML afin d'afficher correctement les expressions mathématiques de l'ouvrage.

PIXELS NINJA ET CHATS SAMOURAÏS

Suite à cette fabuleuse odyssée de l'arcade, Sega et Nintendo s'affrontent avec deux générations de consoles qui se chevauchent : NES/SuperNES vs Master System/Mega Drive. Mais un challenger tout droit sorti du Japon tente de se frayer discrètement un chemin audacieux avec une console qui surpassait techniquement Nintendo et Sega : la ***PC Engine***, fruit de l'alliance entre NEC et Hudson Soft. Grâce à son format HuCard, rapide et quasi instantané, ainsi que ses capacités graphiques supérieures à la 8-bit classique, la PC Engine, renforcée plus tard par l'arrivée du Super CD-ROM², profite d'une créativité foisonnante de la part des développeurs et impose une incroyable maîtrise du shoot'em up et du jeu d'arcade-like avec un catalogue remarquable dont The New Zealand Story qui était l'adaptation la plus réussie de la borne d'arcade, mais aussi des titres comme R-Type, PC Genjin aka Bonk's Adventure, Dungeon Explorer, Bomberman, Ninja Spirit, Gate of Thunder, Ys I & II, Valis II, III et IV, Record of Lodoss War, Legend of Xanadu I & II et Star Parodier, Gate of Thunder, Castlevania: Rondo of Blood avant Ghouls'n Ghosts, Aldynes, 1941, Granzort ou Battle Ace portés par une évolution plus puissante de la console arrivée trop tardivement : la SuperGrafx. Et malgré une stratégie commerciale mal maîtrisée faisant que ces deux fleurons techniques n'ont bénéficié d'aucune distribution officielle en Europe, elles continuent d'occuper une grande place dans notre cœur. Pas seulement pour nous d'ailleurs puisque un nouvel homebrew PC Engine/TurboGrafx-16 Super CD-ROM² est en cours de développement : Nyanja!, un jeu d'action/plateforme à écrans fixes, inspiré de classiques comme Bubble Bobble et Snow Bros, financé via Kickstarter et prévu aussi sur Steam. On y incarne deux chats ninjas, ***Nyanmaru*** et ***Nekomi***, ayant pour objectif de récupérer l'idole « Jade Paw » volée par le Shadow/Dark Kitsune Clan afin de sauver la ville d'Edo-Kyo. Le jeu offre une direction artistique se voulant la plus fidèle et authentique possible de la PC Engine et un gameplay d'arcade classique comme on les aime. Une version Windows/Steam est également prévue, la bande-son originale est également disponible (idéal pour ceux qui aiment le Japon, les ninjas et les chats), et une démo est d'ores et déjà disponible. Une nouvelle campagne Kickstarter est également prévue afin de porter le jeu sur d'autres consoles telles que la Mega Drive et la Sega Dreamcast.

ATTRAPEZ TOUTES LES BRIQUES

Il faut croire que si on ne visite pas régulièrement la boutique officielle Lego, et à moins d'être un membre du programme de fidélité, on passera inévitablement à côté de sets incroyables à l'image du set Pokémon 72153 Venusaur, Charizard and Blastoise (649,99 €) qui permettait d'obtenir un exemplaire du set promotionnel 40892 Kanto Region Badge Collection à l'intérieur duquel se trouvaient les 8 badges de la région de Kanto Pokémon (Roche, Cascade, Foudre, Prisme, Âme, Marais, Volcan et Terre) que les passionnés de Pokémon reconnaîtront immédiatement. À peine quelques heures après l'ouverture des précommandes, ce sont 54560 exemplaires qui ont trouvé preneur ! La bonne nouvelle, c'est que devant ce succès inattendu, le set devrait être à nouveau disponible dès le 27 février 2026 à l'occasion du lancement officiel des trois nouveautés de la gamme LEGO Pokémon qui coïncide avec le 30ème anniversaire de Pokémon. Toutefois tenez-vous prêts car il y a fort à parier que les quantités resteront une nouvelle fois insuffisantes et entraîneront l'inflation immédiate du prix sur le web. Sans surprise, ces trois sets sont destinés aux collectionneurs adultes et les deux derniers sets seront proposés à des prix plus friendly : Pikachu et Poké Ball (199,99 €), et Évoli (59,99€). La lutte sera malgré tout sans pitié !

BRANCHEZ UNE « VRAIE » SUPER NINTENDO SUR VOTRE PC

Le SN Operator est un dock USB-C pensé comme une interface matérielle pour ta collection de cartouches SNES/Super Famicom qui fonctionne sur PC, Mac, Linux, Steam Deck, Asus ROG Ally et même Raspberry Pi. Ce périphérique, développé par Epilogue, permet d'insérer une cartouche originale et de lire à la fois le code du jeu et les données de sauvegarde (SRAM / EEPROM / batterie). Outre la lecture directe des jeux via l'émulateur maison Playback, il génère des sauvegardes de ROM, synchronise tes progrès entre appareils, et même remplace ta souris ou ton clavier pour émuler des accessoires cultes comme la SNES Mouse ou le Super Scope. En prime, le SN Operator est même capable d'authentifier la cartouche pour détecter les contrefaçons. Un vrai plus pour vérifier qu'une cartouche récupérée est bien authentique, surtout avec des titres rares dont la cote flambe sur le marché de l'occasion. De plus, grâce à la lecture directe du bus matériel, la compatibilité avec les cartouches SNES est quasi totale avec un support de toutes les puces d'amélioration : Super FX, SA-1, DSP, CX4...Le côté génial c'est que ça évite de devoir ressortir une vraie console SNES et son écran cathodique qui prend de la place, pour profiter de sa collection. Plus encore si votre SNES est en panne. Malheureusement, à l'heure où vous lirez ces lignes, le SN Operator est déjà en rupture de stock, preuve que le retrogaming est très populaire. En espérant un réapprovisionnement rapide, le prochain Operator pour lequel nous n'avons aucune information est d'ores et déjà en développement pour une commercialisation prévue fin 2026.

#12

Uber hardware .com

IA LOCALE

LE GRAND SAUT

12