



ECP · UNIVERS SPARK · GUIDE PRATIQUE

IA locale en entreprise : choisir entre RTX Spark, DGX Spark et RTX PRO 6000

Comprendre l'IA locale et choisir entre portable ASUS ProArt RTX Spark, mini-PC NVIDIA DGX Spark et station professionnelle RTX PRO 6000 Blackwell.

Guide pour revendeurs, décideurs et utilisateurs · Lecture estimée : environ 20 minutes

Édition du 23 octobre 2026 · Vérifier les références et disponibilités sur les fiches ECP.

Votre parcours de lecture

Lecture complète : environ 20 minutes. Le sommaire ci-dessous permet de choisir directement le sujet utile.

Le guide simple pour choisir votre équipement d'intelligence artificielle - environ 1 min

1. À qui s'adresse une solution d'IA locale ? - environ 1 min

2. Comprendre les différences entre portable, mini-PC et station de travail - environ 2 min

3. Pourquoi 128 Go de mémoire unifiée changent le choix des usages - environ 2 min

4. Choisir un modèle d'IA avec méthode - environ 1 min

5. Six projets concrets pour une entreprise - environ 1 min

6. Lire les performances dans le bon ordre - environ 1 min

7. Windows sur Arm : valider la chaîne de travail complète - environ 1 min

8. IA locale, cloud et confidentialité : choisir le bon périmètre - environ 1 min

9. Mobilité : distinguer puissance disponible et durée d'utilisation - environ 1 min

10. ASUS ProArt P14 ou P16 : deux formats, une sélection ECP en 128 Go - environ 1 min

11. Évaluer le retour sur investissement - environ 1 min

12. Construire un pilote avant de généraliser - environ 1 min

13. Les questions fréquentes des professionnels - environ 2 min

Préparez votre projet avec ECP - environ 1 min

Sources et périmètre du guide - environ 1 min

Guide ECP pour revendeurs, décideurs et utilisateurs • Édition du 23 octobre 2026

Le guide simple pour choisir votre équipement d'intelligence artificielle

Lecture estimée : 1 min

Vous voulez utiliser l'IA dans votre entreprise, mais les mots « modèle », « mémoire unifiée » et « GPU » vous semblent encore compliqués ? Commencez par deux questions : **que voulez-vous faire, et où utiliserez-vous votre machine ?** Ce guide vous aide à répondre sans connaissances techniques préalables.

Lecture complète : environ **20 minutes**. Le sommaire permet de rejoindre directement votre sujet. Pour approfondir un mot, le **Glossaire ECP de l'IA** reste disponible séparément en HTML et PDF : **275 termes expliqués simplement**, à consulter selon vos besoins.

En deux minutes : trois formats, trois façons de travailler

Votre priorité	La famille à regarder	Ce que vous achetez
Emporter votre environnement chez un client ou entre plusieurs lieux	ASUS ProArt P14/P16 RTX Spark	Un portable avec écran, clavier et batterie.
Installer une plateforme IA compacte sur un bureau	NVIDIA DGX Spark	Un mini-PC complet à associer à vos périphériques.
Dimensionner une station pour du calcul, de l'IA et de la création exigeants	NVIDIA RTX PRO 6000 Blackwell	Une carte graphique à intégrer dans une station compatible.

Le matériel fournit les ressources ; les logiciels transforment ces ressources en outil de travail. Il faut installer un moteur d'exécution, choisir un modèle et préparer les documents ou applications du projet. Vérifiez ce qui est réellement compris dans l'offre.

Le bon point de départ : « Je veux retrouver une information dans mes documents en moins d'une minute » aide davantage à choisir qu'une demande générale comme « je veux l'IA la plus puissante ».

1. À qui s'adresse une solution d'IA locale ?

Lecture estimée : 1 min

L'IA locale réalise tout ou partie du traitement sur votre propre matériel. Elle intéresse particulièrement les professionnels qui veulent garder la maîtrise de leur environnement, tester des modèles ou limiter la dépendance à une connexion. La mobilité devient un critère supplémentaire pour choisir un portable.

Profil	Exemple de projet	Premier format à étudier
Revendeur, intégrateur, consultant	Démontrer un assistant chez un client	Portable RTX Spark

Profil	Exemple de projet	Premier format à étudier
Développeur, éditeur de logiciels	Tester du code, des modèles et des agents	Portable ou DGX Spark selon la mobilité
Équipe innovation, formation	Comparer des modèles et organiser des ateliers	DGX Spark ou portable
Studio, agence, bureau d'études	Combiner création et traitements IA compatibles	Portable ou station RTX PRO
Équipe avec calculs réguliers ou plusieurs utilisateurs	Construire un service ou une station spécialisée	Station dimensionnée selon la charge

Ces orientations sont des pistes de qualification ECP, pas des garanties de performance. Pour la bureautique et quelques questions à un service en ligne, un ordinateur conventionnel peut suffire. Pour un service partagé, il faut aussi prévoir la disponibilité, les accès et le nombre de demandes simultanées.

2. Comprendre les différences entre portable, mini-PC et station de travail

Lecture estimée : 2 min

Le **CPU** est le processeur qui fait fonctionner le système et les tâches générales. Le **GPU** accélère les calculs compatibles, notamment certaines opérations d'IA, d'image et de 3D. **Blackwell** désigne une génération d'architecture NVIDIA ; ce nom seul ne donne pas la vitesse d'un travail réel.

Les portables RTX Spark associent un CPU Arm et un GPU Blackwell RTX ; le DGX Spark utilise la superpuce GB10 et DGX OS. Une RTX PRO 6000 est un GPU dédié, installé dans une station avec son propre CPU, sa RAM et son système. Ces familles partagent des technologies NVIDIA, mais leurs logiciels et leurs contraintes diffèrent. ^{[2][3][18]}

Portable RTX Spark : 5 bonnes raisons de le choisir

- 1. Faire suivre votre IA dans vos déplacements.** Conservez vos modèles et votre démonstration sur la même machine, du bureau au rendez-vous client.
- 2. Disposer d'une grande réserve mémoire.** Les versions ECP en 128 Go laissent de la place pour des modèles, leur contexte et les applications ouvertes, dans les limites du scénario testé.
- 3. Réunir travail et présentation.** Écran, clavier et batterie permettent de montrer un résultat sans installer une station complète sur place.
- 4. Associer IA et création.** Explorez des traitements d'image, de vidéo ou de 3D dans les logiciels compatibles avec cette plateforme. Validez aussi leurs extensions et exports.
- 5. Préparer un environnement local réutilisable.** Un consultant ou un formateur peut retrouver ses outils entre deux lieux et limiter sa dépendance au réseau, après installation et essai hors ligne.

À privilégier si vous devez emporter le calcul. L'autonomie sous forte charge et la compatibilité Windows sur Arm restent à vérifier. Les cinq arguments décrivent des bénéfices possibles, pas des performances promises. ^{[2][4][5]}

Le DGX Spark : un ordinateur de bureau très compact

La référence NVIDIA DGX Spark proposée par ECP associe GB10, **128 Go de mémoire unifiée LPDDR5x, un SSD de 4 To et DGX OS**. Son boîtier mesure environ 15 × 15 × 5 cm pour 1,2 kg. Prévoyez écran, clavier, souris et connexion réseau selon l'installation. ^{[12][13]}

DGX Spark : 5 bonnes raisons de le choisir

- 1. Installer une plateforme IA dédiée au bureau.** Vos essais disposent de leur propre machine, distincte du poste utilisé pour les tâches courantes.
- 2. Garder un grand espace de travail mémoire.** Les 128 Go partagés entre CPU et GPU facilitent l'exploration de modèles exigeants en capacité, selon leur format et leur contexte.
- 3. Éviter l'encombrement d'une tour.** Le petit boîtier trouve sa place près d'un écran dans un bureau, un laboratoire ou une salle de formation.
- 4. Développer dans l'environnement NVIDIA.** DGX OS et la pile logicielle NVIDIA constituent une base pour l'inférence, les prototypes et l'adaptation de modèles compatibles.
- 5. Choisir une plateforme compacte en puissance électrique.** GB10 est annoncé à 140 W de TDP et l'appareil dispose d'une alimentation de 240 W. C'est une autre approche qu'une station équipée d'un GPU pouvant atteindre 600 W ; les performances et les mesures ne sont pas équivalentes. ^{[13][18]}

À privilégier si la machine reste au bureau et que DGX OS convient au projet. Le traitement local peut éviter certaines facturations à la requête, mais matériel, logiciels, énergie et maintenance ont toujours un coût. ^{[12][13]}

La RTX PRO 6000 : une carte graphique, pas un ordinateur

La RTX PRO 6000 Blackwell possède **96 Go de mémoire GDDR7 dédiée au GPU, avec ECC**, un mécanisme de correction de certaines erreurs mémoire. Les éditions Workstation et Max-Q n'ont pas la même puissance maximale : respectivement 600 W et 300 W. ECP référence plusieurs éditions ; faites confirmer la version retenue. ^{[14][18]}

RTX PRO 6000 Blackwell : 5 bonnes raisons de la choisir

- 1. Accueillir des charges volumineuses dans la mémoire GPU.** Les 96 Go dédiés sont utiles pour certains modèles d'IA, jeux de données et scènes 3D qui dépassent la capacité de cartes plus petites.
- 2. Alimenter rapidement les calculs.** La GDDR7 offre une bande passante élevée : les données circulent vite entre le GPU et sa mémoire. Le gain réel dépend de l'application.
- 3. Réunir plusieurs usages professionnels.** IA, rendu 3D, visualisation et vidéo peuvent profiter du même GPU lorsque les logiciels prennent en charge ses accélérations.
- 4. Construire une station autour de vos besoins.** CPU, RAM système, stockage et éventuellement plusieurs GPU se dimensionnent ensemble. L'évolution dépend du châssis et de la plateforme choisie.
- 5. Disposer de fonctions destinées aux environnements professionnels.** Mémoire ECC, pilotes professionnels et certifications d'applications peuvent compter pour un poste de production ; vérifiez les versions exactes de vos logiciels. ^[18]

À privilégier si le projet justifie une station spécialisée. Il faut ajouter CPU, RAM, SSD, boîtier, alimentation et refroidissement. Deux cartes de 600 W peuvent représenter jusqu'à 1 200 W pour les GPU seuls, avant les autres composants. Ce maximum théorique n'est pas leur consommation permanente.

Plusieurs GPU peuvent aider à traiter des modèles plus grands ou davantage de demandes, **si le logiciel sait répartir le travail**. Leurs mémoires ne fusionnent pas automatiquement : deux cartes de 96 Go ne deviennent pas simplement une carte de 192 Go.

3. Pourquoi 128 Go de mémoire unifiée changent le choix des usages

Lecture estimée : 2 min

Imaginez deux organisations. Dans un PC classique, le CPU possède sa table de travail, la RAM, et le GPU possède la sienne, la VRAM. Les données circulent entre les deux. Dans une plateforme Spark à mémoire unifiée, **CPU et GPU utilisent une grande réserve commune**. Cela réduit le besoin de copier les données entre deux mémoires distinctes. ^{[11][13]}

128 Go : une réserve commune

Schéma de principe — aucune répartition fixe n'est représentée.

CPU Arm

Système, applications et tâches générales

GPU Blackwell RTX

Traitements graphiques et IA compatibles

128 Go de mémoire unifiée au total

Modèles + contexte + applications + système

Plus de capacité ouvre des usages ; cela ne garantit pas une réponse plus rapide.

Mémoire unifiée : le CPU et le GPU utilisent une réserve commune de 128 Go au total

Station RTX PRO : plusieurs réserves séparées

Schéma de principe - les tailles et débits ne sont pas à l'échelle.

RAM système

Mémoire de travail du processeur
Capacité choisie pour la station

CPU de la station

Système, applications et orchestration
SSD, alimentation et refroidissement à prévoir

GPU 1 - RTX PRO 6000 Blackwell

96 Go de GDDR7 dédiés à ce GPU
Mémoire distincte de la RAM système

GPU 2 éventuel - même principe

96 Go propres à cette seconde carte
Installation et logiciel multi-GPU compatibles

Deux cartes de 96 Go ne créent pas automatiquement une mémoire unique de 192 Go.

Un moteur compatible peut répartir un modèle ou les tâches entre plusieurs GPU.

Station classique : CPU et RAM système, GPU et mémoires dédiées séparées

Les 128 Go sont le total partagé par le modèle, son contexte, le système et les applications. Ils ne correspondent donc pas à 128 Go entièrement disponibles pour l'IA. Sur une station RTX PRO 6000, les 96 Go sont dédiés au GPU, tandis que le CPU dispose en plus de RAM système.

128 Go partagés et 96 Go dédiés ne se comparent pas par une simple soustraction. La place réellement disponible, le modèle et le logiciel déterminent ce qui tient en mémoire. Une grande capacité permet d'ouvrir certains usages ; elle ne garantit pas une réponse plus rapide.

LPDDR5x et GDDR7 : capacité partagée et débit dédié

La **LPDDR5x** est adaptée à une mémoire commune dans un système compact, avec des contraintes d'énergie et de chaleur. La **GDDR7** vise un débit très élevé vers un GPU dédié. La GDDR7 de la RTX PRO 6000 offre une bande passante nettement supérieure à celle de la mémoire du DGX Spark. ^{[13][18]}

En pratique, Spark privilégie une grande réserve partagée dans un petit format ; une station RTX PRO consacre davantage de ressources au calcul GPU et à l'alimentation rapide de ce GPU. Aucun chiffre isolé ne suffit à prédire la vitesse de votre assistant.

Pourquoi ECP privilégie les versions 128 Go

Sur les P14/P16 présentés ici, la mémoire est intégrée : **on ne pourra pas ajouter des barrettes plus tard**. ECP choisit donc 128 Go dès le départ pour offrir davantage de marge aux projets IA. ^{[1][4][5]}

Une machine de 48 Go peut parfaitement répondre à un besoin défini. Elle pourra accueillir différents modèles compatibles, mais restera limitée à cette capacité matérielle. Ce sont surtout les modèles plus volumineux, les contextes longs et les usages simultanés qui peuvent devenir difficiles. Choisissez selon votre besoin réel et sa probable évolution.

Énergie : comparer les bonnes mesures

Le **TDP** indique une enveloppe thermique de conception ; la puissance d'une alimentation indique ce qu'elle peut fournir. La consommation réelle varie avec la charge. Les 140 W de TDP de GB10, l'alimentation 240 W du DGX Spark et les 300/600 W maximaux des cartes RTX PRO mesurent des choses différentes. ^{[13][18]}

Pour estimer le coût, mesurez le système complet à la prise pendant votre tâche. Prenez aussi en compte sa durée : une machine plus puissante peut terminer plus vite. Le choix porte sur le travail obtenu, l'énergie, le bruit, la place et le budget.

4. Choisir un modèle d'IA avec méthode

Lecture estimée : 1 min

Un **LLM** est un modèle entraîné à produire et interpréter du texte. Ses **paramètres** sont des réglages appris pendant l'entraînement ; ils ne sont pas une liste de faits vérifiés. Davantage de paramètres peuvent aider sur certaines tâches, sans garantir davantage de justesse.

L'**inférence**, c'est le moment où le modèle utilise ce qu'il a appris pour produire une réponse. L'entraînement ressemble à l'apprentissage d'un élève ; l'inférence ressemble à l'exercice qu'il fait ensuite. Sur un service distant, le calcul s'effectue chez le fournisseur. Avec Ollama et un modèle local, il s'effectue sur votre machine.

La **quantification** réduit la précision de représentation du modèle pour diminuer ses besoins de mémoire. Il faut vérifier son effet sur la qualité. Un modèle **MoE**, ou mélange d'experts, n'active qu'une partie de ses paramètres à chaque étape : sa taille totale ne suffit pas à prévoir sa vitesse.

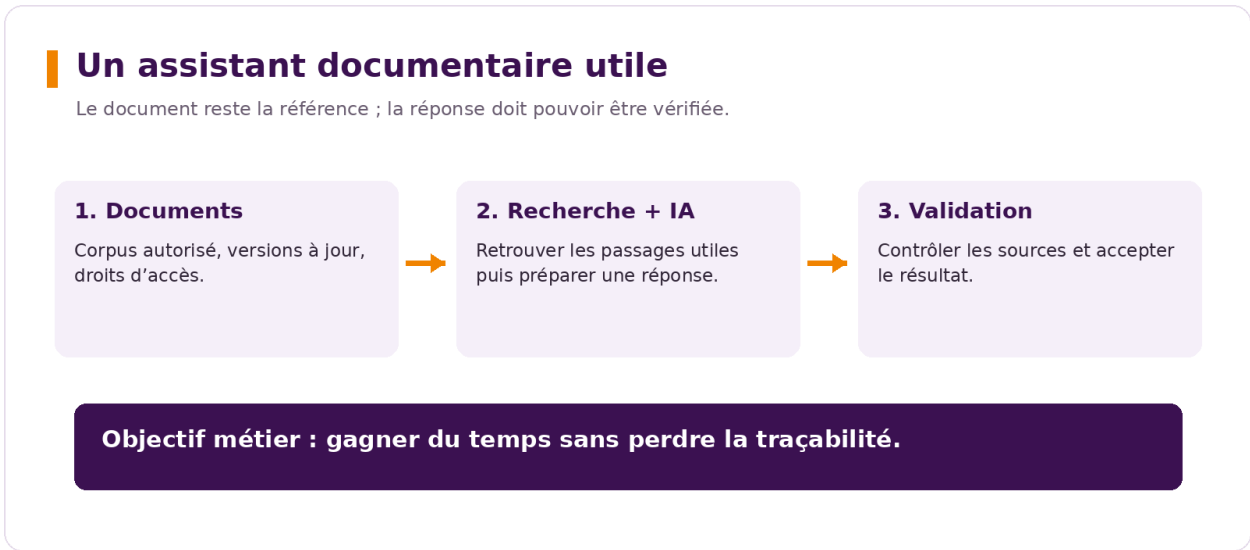
Commencez par un modèle compact pour résumer, classer ou reformuler. Comparez ensuite une variante plus grande si la qualité est insuffisante. Pour dimensionner la mémoire, additionnez :

modèle + contexte + outils + système + marge de travail. Le contexte est l'ensemble des éléments que le modèle traite au cours de l'échange.

Exemple : la hauteur de la tour Eiffel. Un petit modèle peut donner la bonne réponse et un grand modèle peut se tromper. Pour une mesure exacte, datée ou au centimètre près, il faut une source fiable ou un outil de recherche. La taille du modèle ne transforme pas sa mémoire en base de données à jour.

5. Six projets concrets pour une entreprise

Lecture estimée : 1 min



6. Lire les performances dans le bon ordre

Lecture estimée : 1 min

Démarrage

Ouvrir et charger

➡

Première réponse

Traiter le contexte

➡

Résultat final

Exécuter la tâche

➡

Validation

Vérifier et corriger

Mesurer la qualité et la durée totale, pas uniquement les tokens par seconde.

Mesurer la qualité, le délai, la mémoire et les conditions d'utilisation

Testez le parcours complet : chargement du modèle, première réponse, résultat final et correction éventuelle. Les **tokens par seconde** indiquent un débit de fragments de texte ; ils ne mesurent ni la justesse ni le temps total gagné.

Ce qu'il faut mesurer	Pourquoi cela compte
Première réponse et durée totale	Évaluer l'attente réelle de l'utilisateur
Résultats acceptés après relecture	Vérifier la qualité utile au métier
Mémoire libre pendant le travail	Garder de la marge pour les autres applications
Plusieurs tâches successives	Observer chaleur, stabilité et bruit
Secteur, batterie ou plusieurs utilisateurs	Reproduire les conditions prévues

Ce que les lectures de presse apportent

Clubic explique les choix d'architecture et la maturité de l'écosystème. Frandroid a testé un ProArt P16 128 Go : son protocole rapporte 63,3 tokens/s avec gpt-oss-120b MXFP4 sous llama.cpp/CUDA, en mode Performance sur secteur. Il signale aussi une limite de chargement d'un très gros modèle liée à l'espace SSD disponible pour le fichier d'échange. ^{[6][7]}

Ce sont les résultats de cet exemplaire et de ces réglages, pas des mesures ECP ni une promesse pour chaque P14/P16. Le chiffre « un pétaflop FP4 » est lui aussi théorique : il ne donne directement ni la vitesse d'un assistant ni celle d'un logiciel de création. ^[2]

7. Windows sur Arm : valider la chaîne de travail complète

Lecture estimée : 1 min

Les portables ASUS concernés utilisent Windows sur Arm. Des applications natives Arm64 fonctionnent directement ; de nombreuses applications x86/x64 passent par l'émulation Prism. Les pilotes doivent

disposer d'un support adapté à Arm64. ^{[8][9]}

Avant l'achat, essayez vos logiciels métiers avec leurs extensions, votre VPN, votre protection du poste, vos périphériques et vos exports. Pour un développeur, ajoutez bibliothèques et compilateurs ; pour un créateur, codecs et plugins. Une application qui s'ouvre ne prouve pas que tout le travail sera possible.

Le DGX Spark utilise **DGX OS**, un environnement Linux pour cette plateforme. Une compatibilité sous Windows, Linux x86 ou une autre machine NVIDIA ne garantit pas la même expérience sur DGX OS Arm64. Faites confirmer les versions et l'accélération GPU de vos outils.

8. IA locale, cloud et confidentialité : choisir le bon périmètre

Lecture estimée : 1 min

Local, cloud ou hybride ?

Choisir où chaque traitement s'exécute — et quelles données peuvent sortir.

LOCAL

Modèle sur le poste. Tester le scénario complet hors ligne.

HYBRIDE

Règle explicite de routage. Autoriser ou refuser chaque transfert.

CLOUD

Service distant. Vérifier données transmises et conditions d'usage.

La confidentialité dépend du logiciel, des accès et des flux configurés.

IA locale et cloud : identifier où le calcul a lieu et où les données circulent

Le traitement local donne davantage de maîtrise sur l'exécution et les données, à condition de configurer l'ensemble. Une application peut utiliser un modèle local tout en envoyant certains éléments vers une API, une sauvegarde ou un service connecté.

Demandez simplement : **qu'est-ce qui reste sur la machine, qu'est-ce qui sort et qui peut y accéder ?** Pour travailler hors ligne, téléchargez modèles et logiciels à l'avance, puis testez toutes les étapes sans réseau. Pour un fonctionnement hybride, précisez les documents autorisés dans le cloud.

Un **agent IA** peut lire des fichiers ou appeler des outils pour accomplir une tâche. Accordez-lui les droits nécessaires et prévoyez une validation avant les actions engageantes. La présence d'une IA locale ne remplace ni les sauvegardes ni la revue des résultats.

9. Mobilité : distinguer puissance disponible et durée d'utilisation

Lecture estimée : 1 min

Un portable capable de calculer sur batterie peut l'épuiser rapidement sous forte charge. Dans son essai P16, Frandroid observe des performances de texte proches du secteur dans certaines conditions et estime environ une heure de génération continue : ce n'est pas une journée de calcul intensif. ^[7]

Pour une démonstration, testez la durée réelle avec votre luminosité et vos applications. Pour les traitements longs, prévoyez le secteur. Vérifiez aussi l'équipement livré : **les références ECP présentées sont indiquées sans chargeur** ; faites confirmer le chargeur et le câble USB-C adaptés. ASUS mentionne 140 W selon configuration. ^{[1][4][5]}

10. ASUS ProArt P14 ou P16 : deux formats, une sélection ECP en 128 Go

Lecture estimée : 1 min

P14 ou P16 : choisissez votre façon de travailler

Sélection ECP : 128 Go sur les deux modèles. Schémas non contractuels.

P14 · 14 pouces

128 Go

1 To SSD · mobilité fréquente

P16 · 16 pouces

128 Go

2 To SSD · confort sur écran intégré

Comparer le même workflow : le format seul ne prouve pas une différence de vitesse.

Choisir entre le format P14 et le format P16, représentation schématique

Critère	ASUS ProArt P14	ASUS ProArt P16
Modèle ECP	H7407BA-PROART-DISM04X	H7607BA-PROART-DISE07X
Référence ECP	ASU00054057	ASU00054059
Référence constructeur	90NB16H1-M00EVO	90NB16A1-M00R20
Plateforme au catalogue	RTX Spark N1X, 20 cœurs	RTX Spark N1X, 20 cœurs
Mémoire unifiée	128 Go	128 Go
SSD	1 To	2 To
Écran	OLED tactile 14 pouces	OLED tactile 16 pouces
Masse annoncée pour la gamme	1,48 kg	1,77 kg
Batterie annoncée pour la gamme	90 Wh	99,9 Wh
Chargeur au catalogue ECP	Non fourni	Non fourni

Références et configuration : ECP ; données de gamme : ASUS. La fiche de la référence commandée fait foi pour l'écran, les accessoires et les caractéristiques contractuelles. ^{[1][4][5]}

Choisissez le P14 si vous transportez souvent votre poste et utilisez un écran externe une fois installé. Son SSD de 1 To demande d'organiser les modèles et projets conservés en mobilité.

Choisissez le P16 si vous travaillez longtemps sur l'écran intégré et souhaitez davantage de stockage dès l'achat. Son SSD de 2 To facilite la conservation de plusieurs environnements. Le format 16 pouces ne suffit pas à prouver une supériorité de calcul : comparez le même scénario si la vitesse est décisive.

11. Évaluer le retour sur investissement

Lecture estimée : 1 min

Sur une même durée, additionnez matériel, accessoires, installation, licences, énergie, maintenance et accompagnement. Comparez ce coût au temps réellement récupéré après relecture et aux dépenses effectivement évitées. Le local peut compléter le cloud ; tous les abonnements ne disparaîtront pas forcément.

Exemple illustratif, hors taxes : un surcoût de projet de 3 000 € et un gain net de 150 € par mois donnent un retour simple de 20 mois. Avec 50 € par mois, il atteint 60 mois. Ce ne sont ni des tarifs ECP ni des gains annoncés.

Pour un revendeur, précisez ce que couvre l'offre : matériel seul, installation, choix des modèles, documentation et accompagnement éventuel. Un périmètre clair aide le client à comparer et à prévoir le budget complet.

12. Construire un pilote avant de généraliser

Lecture estimée : 1 min

Du premier essai au déploiement

Un pilote documenté permet de décider sur des résultats reproductibles.

01

Cadrer

Une tâche et un besoin précis.

02

Accepter

Des critères de qualité et de délai.

03

Installer

Des versions et paramètres identifiés.

04

Éprouver

Les conditions de travail réelles.

05

Décider

Déployer, ajuster ou changer.

À conserver : configuration, tests, limites connues et procédure de restauration.

Un pilote : définir la tâche, installer, tester, mesurer et décider

- 1. Choisir une tâche précise.** Par exemple, retrouver une caractéristique dans une documentation autorisée.
- 2. Fixer le résultat attendu.** Délais, qualité, sources et validation humaine.
- 3. Installer une configuration identifiée.** Conserver les versions, le modèle et les réglages.
- 4. Tester les conditions réelles.** Documents longs, autres applications, secteur/batterie et réseau prévu.
- 5. Décider sur les résultats.** Déployer, ajuster ou choisir une autre architecture.

Prévoyez une sauvegarde et une procédure de restauration. Pour un service d'équipe, ajoutez des essais de demandes simultanées et désignez la personne chargée des mises à jour.

13. Les questions fréquentes des professionnels

Lecture estimée : 2 min

Puis-je installer Ollama et choisir Qwen ?

Oui : Ollama est le logiciel qui lance le modèle ; Qwen est une famille de modèles que vous pouvez choisir. Téléchargez la version d'Ollama adaptée au système, puis une variante compatible du modèle. Le catalogue propose plusieurs tailles. Sur Windows Arm ou DGX OS, vérifiez le support GPU de la version utilisée et testez votre scénario. ^{[16][17]}

Quelle taille de Qwen faut-il choisir ?

Commencez par une variante qui tient confortablement en mémoire et comparez ses réponses à vos attentes. Les noms 8B, 32B ou 70B expriment des milliards de paramètres, pas des gigaoctets. Format, quantification et contexte déterminent la mémoire nécessaire. Un modèle plus grand peut aider sur certaines tâches, sans être automatiquement plus fiable ou plus rapide.

Est-ce que 128 Go permettent de charger un fichier modèle de 128 Go ?

Non : système, applications, contexte et calcul utilisent aussi de la mémoire. Il faut garder de la marge. Demandez le modèle exact, son format, le contexte choisi et une mesure de mémoire pendant la tâche.

Pourquoi ne pourrai-je pas augmenter la RAM plus tard ?

La LPDDR5x est intégrée à la plateforme ; les P14/P16 ne prévoient pas l'ajout de barrettes. ECP retient 128 Go pour laisser davantage de possibilités dès l'achat. Une machine 48 Go peut suffire à votre usage ; elle restera toutefois limitée à cette capacité, même si vos besoins augmentent. ^{[1][4][5]}

Un DGX Spark sera-t-il mieux pour moi qu'un portable ?

Si vous restez au bureau et souhaitez une plateforme IA dédiée sous DGX OS, c'est une option pertinente. Si vos démonstrations et vos données doivent vous suivre, regardez les portables. Comparez les logiciels et le résultat réel : la même capacité mémoire n'implique pas les mêmes performances.

Une RTX PRO 6000 de 96 Go accepte-t-elle de plus gros modèles ?

Cela dépend du modèle. Ses 96 Go sont dédiés au GPU ; les 128 Go Spark sont partagés avec le système. La RTX PRO offre aussi un débit mémoire élevé. Pour dépasser une seule carte, un moteur compatible peut répartir un modèle sur plusieurs GPU, avec des contraintes d'échange et d'intégration.

Deux cartes donnent-elles automatiquement 192 Go pour un seul modèle ?

Non. Chaque GPU conserve sa mémoire. Le logiciel doit savoir distribuer le modèle ou les tâches, et la station doit fournir les connexions, la RAM système, l'alimentation et le refroidissement nécessaires. Une édition Server demande une infrastructure adaptée à cette édition.

Mes documents restent-ils privés et utilisables sans Internet ?

C'est possible avec un fonctionnement entièrement local, après installation. Testez le scénario déconnecté et contrôlez les fonctions cloud, la télémétrie, les sauvegardes et les droits d'accès. Fournir des documents au modèle ne lui donne pas automatiquement des réponses exactes : demandez les sources et relisez.

Peut-on entraîner une IA et servir plusieurs collaborateurs ?

Ces plateformes peuvent servir à l'inférence, au développement et à certaines adaptations de modèles. Entraîner un grand modèle depuis zéro relève d'une autre échelle. Pour plusieurs utilisateurs, mesurez les demandes simultanées et prévoyez une machine disponible : un portable fermé ou emporté ne peut plus rendre le même service.

Le stockage peut-il évoluer ?

ASUS indique un emplacement M.2 sur le P14 et deux sur le P16, emplacements utilisés inclus. Faites confirmer l'occupation, les SSD compatibles et les conditions d'intervention. Gardez aussi de l'espace pour les modèles, les fichiers temporaires et les mises à jour. ^{[4][5]}

Quelles informations transmettre à ECP ?

Précisez votre tâche, vos logiciels et versions, le modèle envisagé, les documents à traiter, le nombre d'utilisateurs, la mobilité, le budget et l'échéance. Indiquez si vous cherchez uniquement le matériel ou un projet à construire avec votre partenaire.

Préparez votre projet avec ECP

Lecture estimée : 1 min

Votre IA vous suit : explorez les portables ASUS ProArt P14/P16 RTX Spark.

Votre IA reste sur votre bureau : découvrez le mini-PC NVIDIA DGX Spark.

Votre projet demande une station spécialisée : consultez les cartes graphiques RTX PRO 6000 Blackwell.

Contactez votre interlocuteur ECP avec votre scénario et vos applications. Références, stocks et caractéristiques commerciales évoluent : vérifiez les fiches produits au moment de votre projet. Les conseils de ce guide ne constituent pas une prestation incluse avec le matériel.

Sources et périmètre du guide

Lecture estimée : 1 min

Ce guide est une rédaction originale orientée vers la compréhension et la décision B2B. Il ne constitue pas un test matériel ECP. Les données constructeur, observations de presse et recommandations éditoriales sont distinguées. Socle documentaire vérifié le 22 octobre 2026 ; références RTX PRO 6000 complétées le 23 octobre 2026 ; les logiciels, références, stocks et offres commerciales peuvent évoluer.

- [1] **ECP** - Sélection SPARK Notebook.
- [2] **NVIDIA** - Présentation et caractéristiques RTX Spark.
- [3] **NVIDIA** - DGX Spark.
- [4] **ASUS** - Caractéristiques ProArt P14 H7407.
- [5] **ASUS** - Caractéristiques ProArt P16 H7607.
- [6] **Clubic** - Décryptage de la plateforme RTX Spark.
- [7] **Frandroid** - Test NVIDIA RTX Spark.
- [8] **Microsoft Learn** - Fonctionnement de l'émulation sur Arm.
- [9] **Microsoft Support** - FAQ des PC Windows Arm.
- [10] **NVIDIA** - Annonce de la plateforme RTX Spark et de la collaboration Microsoft.
- [11] **ASUS** - Présentation des ProArt P14 et P16 RTX Spark.
- [12] **ECP** - NVIDIA DGX Spark proposé par ECP.
- [13] **NVIDIA** - Caractéristiques DGX Spark.
- [14] **ECP** - Sélection de cartes graphiques professionnelles RTX PRO 6000.
- [15] **NVIDIA / PNY** - RTX PRO 6000 Blackwell Workstation Edition et RTX PRO 6000 Blackwell Max-Q.
- [16] **Ollama** - Catalogue Qwen3.
- [17] **Ollama** - Téléchargements et versions et versions du projet.
- [18] **NVIDIA** - RTX PRO 6000 Workstation et RTX PRO 6000 Max-Q.