IA locale pour PME suisse

Guide décisionnel et technique

RTX Spark, DGX Spark, RTX PRO 6000, cluster HA, H100 : ce qu'il faut savoir avant d'investir

Édition août 2026

> *Ce document est rédigé à titre documentaire, sans dépendance à aucun constructeur, revendeur, éditeur ou intégrateur cité. Aucun de ces acteurs ne rémunère l'auteur, ne finance cette publication et n'a été consulté avant sa parution. Les recommandations ne dépendent d'aucun partenariat commercial et aucune prestation ne vous sera proposée à la lecture de ce guide.*

> *Les prix et benchmarks sont des estimations ou des mesures publiées par des tiers, vérifiées en août 2026. Obtenir des devis fermes avant tout engagement. Voir l'annexe C pour la méthode et les sources.*

Table des matières

Introduction

Depuis le Computex 2026, Nvidia positionne la puce RTX Spark comme une révolution pour les PC Windows : ARM, 128 Go de mémoire unifiée, 6 144 cœurs CUDA, capable de faire tourner des modèles IA en local. Les intégrateurs IT ont saisi l'opportunité et certains vendent déjà ces machines comme des « serveurs IA clé en main » pour remplacer des infrastructures cloud ou des serveurs dédiés.

Ce guide donne un cadre objectif pour évaluer ces affirmations et, plus largement, décider si un déploiement d'IA locale est pertinent pour votre organisation. Il est structuré en deux parties :

La Partie I aide à décider si l'IA locale est le bon choix et à quel usage.

La Partie II détaille les architectures, les coûts, les performances et l'ingénierie nécessaire pour choisir et déployer correctement.

> *Règle de base : une architecture matérielle ne fait pas la qualité d'un système IA. Le matériel détermine la capacité (quels modèles on peut faire tourner) et le débit (combien de tokens par seconde). La fiabilité des réponses dépend du modèle, de sa quantification et de l'ingénierie du pipeline RAG.*

Ce que les intégrateurs ne mettent pas en avant

Argument commercial fréquent	Réalité documentée
"Un serveur IA complet et souverain pour moins de 5 000 CHF."	C'est un mini-PC de bureau sous architecture ARM, pas un serveur. Aucune redondance, un seul point de défaillance. Souverain : oui. Serveur au sens production : non.
"Fait tourner les mêmes modèles que les grands datacenters."	Il charge les mêmes modèles, mais génère les réponses 4 à 6 fois plus lentement qu'une carte GPU professionnelle dédiée, faute de bande passante mémoire suffisante.
"Performant pour toute votre équipe : 6 144 cœurs CUDA."	Le nombre de cœurs CUDA ne détermine pas la vitesse de génération de tokens. Sans moteur d'inférence configuré pour la concurrence (vLLM), plusieurs utilisateurs simultanés attendent chacun leur tour.
"Clé en main, prêt à l'emploi, sans hallucinations."	Le matériel ne résout pas les hallucinations. Sans ingénierie RAG complète (reranker, prompt strict, guardrails), les taux d'erreur sur des tâches ouvertes restent significatifs, quelle que soit la carte GPU.

Partie I : Décider

1. IA locale ou LLM commercial : quand choisir quoi

Les LLM commerciaux (ChatGPT Enterprise, Claude, Gemini) et les modèles locaux open-weight ne s'opposent pas : ils répondent à des besoins différents. La confusion naît des discours marketing qui présentent l'un comme systématiquement supérieur à l'autre.

Critère	LLM commercial (cloud)	LLM local (on-premise)	Verdict
Qualité sur tâches générales	Excellent	Très bon à excellent selon modèle	Écart qui se réduit chaque année. Pour des tâches documentaires, les modèles locaux 30B+ sont compétitifs.
Souveraineté des données	Données transmises au fournisseur	Données restent dans l'entreprise	Le local s'impose dès que les données sont sensibles au sens de la nLPD.
Conformité nLPD / RGPD	Dépend du contrat et du datacenter	Totale si bien configuré	Le local élimine le risque de transfert hors juridiction suisse.
Coût à faible volume	Faible (paiement à l'usage)	Élevé (CapEx fixe)	Le cloud gagne en dessous de ~2 millions de tokens/jour.
Coût à fort volume	Élevé (facturation à l'usage)	Quasi nul (coût marginal zéro)	Le local gagne au-delà d'un seuil. Amortissement typique : 6 à 18 mois selon le volume.
Personnalisation sur données internes	RAG via API possible, données transmises	RAG complet en local, aucune donnée ne sort	Le local est presque obligatoire si les documents sources sont confidentiels (PII).
Disponibilité	Dépend de la connexion internet	Fonctionne sans internet	Le local est indispensable en environnement air-gapped.
Mise en service	Immédiate (abonnement)	4 à 8 semaines selon la complexité	Le cloud est plus rapide à démarrer. Le local nécessite une intégration.
Maintenance	Zéro (gérée par le fournisseur)	Requiert un IT interne ou un prestataire	Le cloud est plus simple à opérer.
Multilingue FR/DE/IT/EN	Excellent	Excellent (Qwen3, Llama 4)	Équivalent pour les quatre langues nationales suisses.

> *Le modèle dominant en 2026 pour les PME et ETI est l'architecture hybride : tâches non sensibles routées vers un LLM cloud (plus rapide, moins cher à faible volume), données confidentielles traitées sur un modèle local. Les deux ne s'excluent pas.*

2. Ce qu'un système IA local peut faire concrètement

Voici une liste objective de ce qui fonctionne en production en 2026, ce qui nécessite des précautions, et ce qui ne fonctionne pas encore de façon fiable. Cette liste est le point de départ pour identifier les cas d'usage pertinents dans votre organisation.

2.1 Ce qui fonctionne bien avec un pipeline RAG complet

Tâche	Description concrète	Fiabilité	Modèle recommandé
Recherche documentaire	Poser une question en langage naturel sur des contrats, procédures, historique client. Réponse sourcée avec citation du document.	Élevée avec pipeline complet	Qwen3 30B-A3B
Résumé de documents	Condenser un rapport de 50 pages. Extraire les points clés d'un compte-rendu.	Bonne, relecture recommandée	Qwen3 14B ou 30B
Extraction de données (OCR)	Lire une facture ou un formulaire et en extraire les données clés dans un format exploitable par l'ERP.	85-95% sur corpus homogène	Qwen2-VL 7B (multimodal)
Classification et triage	Classer des emails, tickets support, alertes de sécurité par catégorie et priorité.	Élevée sur catégories définies	Qwen3 8B ou 14B
Résumé de réunion Teams	À partir de la transcription automatique de Teams, générer un compte-rendu structuré avec décisions et actions.	Bonne si transcription de qualité	Qwen3 14B ou 30B
Aide au développeur	Complétion de code, revue, génération de tests, documentation automatique. Connaît la base de code interne si indexée.	Très bonne sur langages courants	Qwen2.5-Coder 14B
Traduction multilingue	FR, DE, IT, EN et 200+ langues. Adapté aux quatre langues nationales suisses.	Excellente sur langues européennes	Qwen3 (toutes tailles)
Génération de rapports	Rédiger automatiquement des rapports périodiques à partir de données ERP, en .docx avec charte graphique.	Bonne sur structure fixe	Qwen3 14B ou 30B
Support RH et helpdesk	Répondre aux questions récurrentes sur procédures, avantages, règlements internes.	Bonne si corpus RH à jour	Qwen3 8B ou 14B
Veille réglementaire nLPD	Surveiller les nouvelles publications légales et alerter si un texte impacte l'organisation.	Bonne en mode RAG sur sources officielles	Qwen3 30B

2.2 Ce qui nécessite une supervision humaine

- Génération de texte ouverte (emails à des tiers, communications officielles) : toujours relire avant envoi.
- Analyse juridique ou fiscale : le modèle peut résumer, mais ne remplace pas un juriste ou une fiduciaire pour les conclusions.
- Réponses à des clients externes : jamais sans validation humaine pour les messages engageant la responsabilité de l'organisation.

2.3 Ce qui ne fonctionne pas encore de façon fiable

- Agents autonomes multi-étapes sans supervision : taux d'erreur significatifs sur les chaînes complexes. À réserver aux tâches à faible risque ou avec validation humaine à chaque étape.
- Transcription audio en temps réel : Teams gère ça nativement. Le LLM reçoit la transcription texte, pas l'audio.
- Génération d'images et de vidéos : nécessite un pipeline séparé (Stable Diffusion, FLUX). Ce n'est pas le même type de modèle.
- Raisonnement mathématique ou comptable complexe : vérification systématique requise.
-

> *La règle des 90% : un système RAG bien configuré vise un taux de faithfulness (réponses directement traçables aux sources) supérieur à 90% sur un corpus documentaire homogène. En dessous de ce seuil, le système n'est pas prêt pour la production.*

3. Une journée type dans une PME de 15 à 20 personnes

Voici ce que fait concrètement un système IA local correctement installé. L'exemple s'appuie sur un DGX Spark comme support matériel ; le chapitre 8 explique pourquoi et quand cette option est pertinente.

3.1 Nuit et tôt le matin : tâches automatiques

Tâche automatique	Ce qui se passe	Intervention humaine requise
OCR et intégration des factures	Les factures reçues par email pendant la nuit sont lues par le modèle multimodal. Les données sont extraites et envoyées à l'ERP. Les cas ambigus sont mis en file de validation.	Valider les cas en anomalie (~5-15 min/jour selon le volume)
Indexation des nouveaux documents	Les fichiers déposés sur le serveur depuis la veille sont automatiquement indexés. Les permissions d'accès sont synchronisées depuis Active Directory.	Aucune
Génération des rapports récurrents	Le rapport d'activité hebdomadaire est rédigé à partir des données ERP et envoyé au format .docx avec la charte graphique de l'organisation.	Aucune (relecture optionnelle)
Résumés des réunions Teams	Les transcriptions des réunions de la veille sont traitées. Un compte-rendu structuré (décisions, actions, responsables) est posté dans le canal Teams concerné.	Validation recommandée avant diffusion

3.2 La journée : tâches interactives en temps réel

Tâche	Ce que voit l'utilisateur	Latence typique (DGX Spark + vLLM)
Recherche documentaire (commerciaux)	« Quelles sont les conditions de paiement du contrat Dupont SA ? » → réponse en langage naturel avec citation du paragraphe exact.	3 à 8 secondes
Support RH interne	« Combien de jours de congé me reste-t-il ? » → le système consulte la politique RH indexée et répond en citant la source.	2 à 5 secondes
Copilote développeur	Le développeur tape son code dans VS Code, le modèle suggère la complétion en connaissant la base de code interne.	1 à 3 secondes
Traduction de documents	Un contrat en allemand est collé dans l'interface, la traduction en français est produite avec terminologie adaptée au domaine métier.	5 à 15 secondes selon longueur
Triage des tickets support	Un email de réclamation arrive, le modèle le classe et suggère une réponse à partir des résolutions précédentes ; le technicien valide ou modifie.	3 à 6 secondes

> *Ces latences correspondent à un DGX Spark avec vLLM, modèle Qwen3 30B-A3B en configuration FP4 Blackwell, pour 3 à 5 utilisateurs simultanés. Au-delà de 8 à 10 utilisateurs, les latences augmentent et un second DGX Spark en parallèle devient pertinent.*

3.3 La stack qui rend tout ça possible

Composant	Rôle	Ce que c'est	Coût licence
DGX Spark	Le serveur IA	Mini-PC Nvidia sous Linux (DGX OS), avec 128 Go de mémoire et accélérateur IA Blackwell	CHF 4 524 (matériel)
vLLM	Le moteur d'inférence	Logiciel qui fait tourner les modèles IA et gère les requêtes simultanées efficacement (continuous batching)	Gratuit (open source)
Qwen3 30B-A3B	Le cerveau principal	Modèle de langage open-source multilingue FR/DE/IT/EN, architecture MoE (3B de paramètres actifs sur 30B)	Gratuit (open source)
Qwen2-VL 7B	Le lecteur de documents	Modèle multimodal qui lit les images et PDF scannés pour en extraire les données	Gratuit (open source)
Qdrant	La mémoire documentaire	Base de données qui stocke les documents indexés et les retrouve selon leur sens (recherche vectorielle)	Gratuit (open source)
BGE-Reranker	Le filtre de pertinence	Modèle léger qui classe les documents trouvés par ordre de pertinence réelle pour la question posée	Gratuit (open source)
Onyx	L'interface utilisateur	Application web que les collaborateurs ouvrent dans leur navigateur pour interroger la base documentaire	Gratuit (community)
n8n	L'orchestrateur	Outil d'automatisation qui connecte tous les composants et déclenche les pipelines (OCR, résumés, rapports)	Gratuit (community)
Continue.dev	Le copilote développeur	Plugin pour VS Code qui connecte l'éditeur de code au modèle local	Gratuit

> *Tous les logiciels de cette stack sont open source et gratuits. Le seul coût fixe est le matériel. Les coûts d'intégration (configuration, connecteurs, formation) représentent l'essentiel du budget projet. Aucune redevance mensuelle.*

4. Coût de mise en service par fonctionnalité

Ce chapitre donne des ordres de grandeur pour évaluer les devis que vous recevrez. Les fourchettes de jours et de coûts sont des estimations construites par synthèse de sources publiques (appels d'offres documentés, retours d'expérience publiés sur des projets similaires en Europe occidentale). Elles peuvent varier de 30 à 50% selon la complexité réelle du projet, la maturité de l'infrastructure existante et le profil du prestataire. Voir l'annexe C pour la méthode de construction de ces estimations.

Le matériel ne représente souvent qu'une fraction minoritaire du coût total d'un projet IA qui fonctionne. L'intégration, les connecteurs, la formation et la maintenance sur 3 ans constituent l'essentiel du budget réel.

Fonctionnalité	Ce que comprend l'intégration	Jours estimés	Coût estimé (1 200 CHF/j, médiane CH observée)
RAG documentaire (base)	DGX OS + vLLM + Qdrant + Onyx + connecteur file server SMB + permissions NTFS + tests sur corpus réel	5-8 jours	6 000-10 000 CHF
OCR et intégration ERP	Pipeline n8n + modèle multimodal + connecteur API ERP + file de validation + tests sur corpus réel. Variable selon l'ERP : Abacus (API documentée) est plus simple que des ERP anciens sans API.	5-10 jours	6 000-12 000 CHF
Résumé de réunion Teams	Connecteur Graph API Teams + pipeline n8n + génération .docx + bot Teams	2-3 jours	2 400-3 600 CHF
Copilote développeur	Continue.dev configuré + indexation base de code Git + modèle Coder + tests	1-2 jours	1 200-2 400 CHF
Génération de rapports récurrents	Pipeline n8n + connecteur ERP + template .docx + planification + tests	2-3 jours	2 400-3 600 CHF
Support RH et helpdesk interne	Indexation corpus RH + Onyx configuré + tests sur questions types + formation RH	2-3 jours	2 400-3 600 CHF
Traduction multilingue	Endpoint dédié vLLM + interface simple + glossaire métier injecté en prompt	1-2 jours	1 200-2 400 CHF
Veille réglementaire nLPD	Indexation sources officielles + pipeline de surveillance + alertes email/Teams	2-3 jours	2 400-3 600 CHF
Formation utilisateurs (3 profils)	Commerciaux (Onyx), développeurs (Continue.dev), managers (rapports + Teams). **Sans formation, le ROI d'un outil IA est nul quelle que soit la qualité du déploiement**.	2 jours	2 400 CHF

Approche par phases recommandée. Phase 1 : RAG documentaire + OCR (8-12 jours, 10 000-15 000 CHF). Résultats visibles en 6 semaines. Phase 2 : résumé Teams + copilote dev + rapports (4-6 jours supplémentaires). Phase 3 : fonctionnalités avancées selon les besoins identifiés. Le connecteur ERP est le poste le plus variable : prévoir un audit technique avant de chiffrer si l'ERP est ancien ou sans API documentée.

> *Ces estimations ne constituent pas un devis. Elles servent à calibrer vos échanges avec des prestataires. La qualité du prestataire, la complexité de votre environnement et l'état de vos données documentaires influencent fortement les durées réelles.*

5. Les questions à se poser avant tout déploiement

Un projet IA local qui réussit répond à ces questions avant de choisir le matériel. Un projet qui les ignore après l'achat produit un système inutilisé dans les trois premiers mois.

5.1 Sur le problème à résoudre

- Quel processus coûte de l'argent ou du temps aujourd'hui, et combien exactement ? Si vous ne pouvez pas nommer un processus précis avec un coût estimé, le projet n'est pas prêt.
- Ce problème vient-il du manque d'accès à l'information, du manque de temps pour traiter l'information, ou du manque de personnes ? L'IA locale résout les deux premiers. Elle ne remplace pas des effectifs manquants.
- Qui utilisera le système au quotidien, et combien de fois par jour ? Un outil utilisé deux fois par semaine ne justifie pas un investissement matériel significatif.

5.2 Sur les données

- Où sont les données aujourd'hui ? File server, SharePoint, ERP, emails ? Si elles ne sont pas numérisées et accessibles, l'IA ne peut pas les exploiter.
- Les données sont-elles sensibles au sens nLPD ? Données personnelles, médicales, financières de tiers ? Si oui, le déploiement local s'impose.
- Quelle est la qualité des documents existants ? Des scans illisibles ou des formats non structurés multiplient le coût d'intégration.

5.3 Sur l'organisation

- Qui est responsable de maintenir le corpus documentaire à jour ? Sans réponse claire, le RAG se dégradera en quelques mois et les utilisateurs l'abandonneront.
- Y a-t-il quelqu'un en interne capable de gérer un serveur Linux, ou faut-il un contrat de support externe ?
- L'organisation est-elle prête à former ses utilisateurs ? Un outil IA non adopté a un retour sur investissement nul, quel que soit le matériel.

5.4 Sur le risque

- Quel est le coût d'une erreur du modèle ? Un résumé de réunion incorrect est gênant. Une erreur dans une déclaration fiscale ou un document légal est un risque réel. Le niveau de supervision humaine dépend de ce chiffre.
- L'organisation a-t-elle une politique sur l'utilisation de l'IA ? Sans cadre, le déploiement crée des risques légaux (EU AI Act, nLPD).

5.5 Sur le budget réel

- Le budget inclut-il le matériel, l'intégration, la formation et la maintenance sur 3 ans ? Le matériel représente souvent moins de 30% du coût total d'un projet qui fonctionne réellement.
- Quel est le critère de succès défini avant le démarrage ? Sans mesure claire, il n'y a pas de projet : il y a une expérimentation sans fin.

6. Les signaux d'alerte : quand ne pas se lancer maintenant

Ces signaux n'indiquent pas que l'IA locale est une mauvaise idée. Ils indiquent que le projet échouera dans sa forme actuelle. Mieux vaut corriger ces points avant d'acheter le matériel qu'après.

- Le déclencheur du projet est une démonstration commerciale ou un article de presse, pas un problème opérationnel identifié.
- Personne ne peut nommer le processus précis qui sera automatisé ni estimer le temps qu'il coûte actuellement.
- Les documents qui serviront de corpus RAG ne sont pas numérisés, pas organisés, ou accessibles sans politique claire de gestion documentaire.
- Il n'y a pas de responsable désigné pour maintenir le système après le déploiement.
- Le budget ne couvre que le matériel et pas l'intégration ni la formation.
- L'objectif déclaré est de « faire la même chose que ChatGPT mais en local » sans cas d'usage métier défini.

> *Repère statistique : selon McKinsey State of AI 2025, seulement 5,5% des organisations qui ont déployé l'IA observent un impact mesurable supérieur à 5% sur leur résultat opérationnel. La première cause d'échec identifiée est l'absence de cas d'usage clairement défini avant le déploiement. Source citée en annexe C.*

Partie II : Choisir et déployer

7. Les architectures disponibles en Suisse

7.1 Vue d'ensemble

Cinq catégories d'infrastructure sont pertinentes pour une PME souhaitant faire tourner des LLM en local, complétées par les offres cloud GPU avec résidence des données en Suisse.

Solution	Format	GPU / Puce	Mémoire	Bande passante	TDP système	Disponibilité CH
RTX Spark laptop	Portable	GB10 ARM SoC	64 ou 128 Go LPDDR5X unifiés	300 GB/s	110-140 W	Automne 2026, pas encore vendu en CH
DGX Spark / Ascent GX10	Mini PC bureau	GB10 ARM SoC	128 Go LPDDR5X unifiés	273-300 GB/s	~170 W système	Disponible, dès CHF 4 524
RTX PRO 6000 Blackwell	Station travail	GB202 PCIe 5	96 Go GDDR7 ECC	1 792 GB/s	600 W GPU + hôte	Disponible, prix en forte hausse (+87% depuis lancement)
Cluster 2 nœuds RTX PRO 6000 + Proxmox HA	Salle technique	2× GB202 PCIe 5	2× 96 Go GDDR7 ECC	2× 1 792 GB/s	~1 200-1 400 W total	Sur commande, infra locale requise
H100 PCIe 80 Go	Serveur rack	GH100 Hopper	80 Go HBM2e	2 000 GB/s	350 W GPU + serveur	Via intégrateurs, infra requise
SwissGPU H100 (cloud)	Cloud Zurich	H100 80 Go HBM2e	80 Go partagés	2 000 GB/s	N/A	On-demand, 3,11 USD/h
SwissGPU RTX 6000 PRO (cloud)	Cloud Zurich	RTX 6000 PRO Blackwell	96 Go GDDR7	1 792 GB/s	N/A	On-demand, 1,50 CHF/h + 0,30 CHF/kWh

RTX Spark laptop et DGX Spark reposent sur la même puce GB10. La différence est uniquement le format : portable versus mini-PC de bureau. Le cluster 2 nœuds est la seule option locale offrant un failover automatique et un SLA défendable contractuellement. H100 PCIe : 350 W et HBM2e (à ne pas confondre avec le H100 SXM qui est à 700 W et HBM3). Sources : Nvidia Computex 2026, toppreise.ch, Thunder Compute H100 Specs Guide 2026.

7.2 La bande passante mémoire : le critère décisif

Le chiffre que les intégrateurs ne mentionnent pas est la bande passante mémoire. C'est elle, et non le nombre de cœurs CUDA, qui détermine la vitesse de génération des tokens.
Une analogie pour les décideurs : un modèle 70B, c'est environ 40 Go de données à faire circuler entre la mémoire et les unités de calcul pour chaque token généré.

La mémoire, c'est l'entrepôt. La bande passante, c'est le tuyau de livraison. Le RTX Spark a un grand entrepôt (128 Go) mais un tuyau étroit (300 GB/s). La RTX PRO 6000 a un tuyau six fois plus large (1 792 GB/s). La taille de l'entrepôt ne compense pas la largeur du tuyau quand il faut livrer vite.

Architecture	Bande passante	Rapport vs RTX Spark	Débit 70B, 1 utilisateur (vLLM Q4_K_M, t/s par utilisateur)	Débit 70B, 5 utilisateurs simultanés (vLLM, t/s par utilisateur)
RTX Spark / DGX Spark	300 GB/s	Référence (×1)	~5-8 t/s	~7-9 t/s/user (vLLM) ou attente séquentielle (Ollama)
RTX PRO 6000 Blackwell	1 792 GB/s	×6	~24-31 t/s	~12-15 t/s/user
H100 PCIe (HBM2e)	2 000 GB/s	×6,7	~95 t/s	~30-40 t/s/user

Seuil de lisibilité : ~15 t/s en streaming. Seuil d'acceptabilité minimale : ~5 t/s. Les chiffres multi-utilisateurs supposent un continuous batching actif (vLLM) et représentent le débit par utilisateur individuel. Sans batching correctement configuré (Ollama par défaut), les 5 utilisateurs attendent chacun leur tour. Sources : LMSYS DGX Spark benchmark, AITOT Inference Benchmark juillet 2026.

7.3 Ollama vs vLLM : la distinction que personne n'explique

La performance d'un système RAG d'entreprise dépend autant du moteur d'inférence que du matériel. C'est la distinction la plus fréquemment absente des offres commerciales.

Critère	Ollama	vLLM	Impact PME
Installation	2 commandes, tout OS	Linux uniquement, configuration dédiée	Ollama s'installe en 20 min. vLLM demande une journée de travail qualifié.
Traitement multi-utilisateurs	Séquentiel : file d'attente	Concurrent : continuous batching	Au-delà de 3 utilisateurs simultanés, seul vLLM est viable en production.
Débit mesuré à 10 utilisateurs	~148 t/s total (benchmark Red Hat 2026)	~485 t/s total (×3,3)	L'écart se creuse avec la concurrence.
Gestion mémoire KV cache	Allocation fixe	PagedAttention : fragmentation < 4%	vLLM fait tenir plus de contextes en VRAM simultanément.
Prefix caching	Non	Oui : prompt système calculé une fois	Gain de latence sur les RAG avec prompt système long.
Complexité opérationnelle	Faible : idéal pour le développement	Élevée : monitoring et tuning requis	vLLM exige un IT capable de gérer un serveur Linux et des paramètres d'inférence.
Usage recommandé	Prototypage, dev solo, démonstrations	Production multi-utilisateurs, RAG d'équipe	

> *Règle pratique : une offre qui ne précise pas quel moteur d'inférence est installé et configuré est une offre incomplète. « Clé en main avec vLLM » et « clé en main avec Ollama » ne donnent pas les mêmes performances pour une équipe. C'est la première question à poser. Sources : Red Hat vLLM benchmark 2026, Ki-Mittelstand benchmark juillet 2026.*

7.4 DGX Spark vs laptop RTX Spark : deux situations différentes

- DGX Spark : tourne sous DGX OS, distribution Linux basée sur Ubuntu ARM64. CUDA 13, drivers pré-installés, vLLM nativement supporté. Tout pipeline RAG Linux (Qdrant, LlamaIndex, LangChain, vLLM) tourne sans couche d'émulation. C'est un vrai serveur Linux de bureau.
- Laptop RTX Spark : tourne sous Windows on ARM. La compatibilité logicielle est à valider application par application. vLLM n'a pas de support officiel Windows en production à ce jour. Les outils métier (antivirus, MDM, ERP) peuvent poser des problèmes de pilotes.

> *Même puce GB10, mais deux machines très différentes en production. Un DGX Spark avec vLLM est un serveur RAG d'équipe viable. Un laptop RTX Spark sous Windows avec Ollama est un outil de prototypage individuel.*

7.5 DGX Spark bien configuré vs mal configuré

Configuration	Débit 70B	Multi-users	Verdict
Laptop RTX Spark + Windows + Ollama	~5-8 t/s	Séquentiel, file d'attente	Poste individuel. Ne pas vendre comme serveur d'équipe.
DGX Spark + DGX OS + Ollama	~5-10 t/s	Séquentiel, file d'attente	Prototypage correct. Pas de concurrence réelle.
DGX Spark + vLLM (Q4_K_M)	~20-30 t/s	Concurrent, batching actif	Serveur RAG viable, 3-5 utilisateurs, 70B en local.
DGX Spark + vLLM (FP4 natif Blackwell)	~35-45 t/s	Concurrent, débit optimisé	Meilleure configuration GB10. Exige vLLM v0.24+ et kernels CUDA 13.
2× DGX Spark indépendants + Nginx LB	Modèles MoE 30B ×2	16-24 utilisateurs simultanés	Meilleure option pour la concurrence. Plus simple et plus efficace qu'un cluster tensor parallel.

> *La configuration FP4 natif Blackwell exige vLLM v0.24+ avec support Blackwell activé et les kernels CUDA 13 du DGX OS. Sans ces éléments, vLLM tombe sur des kernels génériques moins performants. C'est le premier point à vérifier lors d'une installation.*

8. Prix et coût total de possession sur 3 ans

8.1 Avertissement sur les prix actuels

> *Le prix de la RTX PRO 6000 a subi deux hausses successives : 8 565 USD au lancement (mars 2025), 13 250 USD en juin 2026, puis 16 000 USD en août 2026 (+87% en 16 mois), principalement due à la pénurie de mémoire GDDR7. Le H100 PCIe consomme 350 W (et non 700 W qui correspond au H100 SXM). Les laptops RTX Spark n'ont pas encore de prix officiels en Suisse. Obtenir des devis fermes avant tout engagement.*

8.2 Tableau comparatif TCO 3 ans

Solution	CapEx estimé	OpEx / an	TCO 3 ans estimé	Infra requise
RTX Spark laptop	~2 500-4 000 CHF	~150 CHF	~3 000-5 000 CHF	Aucune
DGX Spark (Ascent GX10)	~4 500-6 000 CHF	~300 CHF	~5 500-8 000 CHF	Prise 16 A, bureau
RTX PRO 6000 + station	~22 000-28 000 CHF	~4 000 CHF	~34 000-40 000 CHF	Bureau, pas de rack
H100 PCIe + serveur rack	~40 000-60 000 CHF	~7 000 CHF	~61 000-81 000 CHF	Salle serveur, rack, refroid.
SwissGPU RTX 6000 PRO (40h/sem)	0	~3 400 CHF	~10 200 CHF	Aucune, données en CH
SwissGPU H100 (40h/sem)	0	~6 500 CHF	~19 500 CHF	Aucune, données en CH

OpEx RTX Spark laptop : électricité négligeable (~110 W, usage partiel). DGX Spark : ~170 W continu, ~300 CHF/an. RTX PRO 6000 : ~900 W total système, maintenance incluse. H100 PCIe : 350 W GPU + ~500 W serveur = ~850 W total. Ne pas confondre avec le H100 SXM (700 W, HBM3) qui nécessite une plateforme spécialisée.

8.3 Détail du TCO H100 PCIe on-premise

Poste de coût	Montant estimé	Remarque
GPU H100 PCIe 80 Go (HBM2e, 350 W)	25 000-30 000 USD	Prix marché août 2026. Ne pas confondre avec le H100 SXM (700 W, HBM3, plateforme dédiée).
Serveur hôte (Dell/HPE)	8 000-15 000 USD	CPU, RAM ECC, NVMe, alimentation redondante
Rack, alimentation, réseau	3 000-8 000 USD	PDU, switch 10 GbE, câblage
Refroidissement (350 W GPU + ~500 W serveur = ~850 W)	2 000-8 000 USD	Moins contraignant que le SXM. Refroidissement standard suffisant dans beaucoup de cas.
CapEx total estimé	**~40 000-60 000 USD**	Hors intégration et formation
Électricité (350 W GPU, 75% utilisation, 0,22 CHF/kWh)	~450-600 CHF/an	H100 PCIe = 350 W (HBM2e). Ne pas confondre avec le H100 SXM (700 W, HBM3).
Maintenance et support	~2 500-5 000 CHF/an	Contrat HW + firmware
Admin IT (fraction ETP)	~3 000-8 000 CHF/an	Coût souvent ignoré
OpEx total estimé / an	**~6 000-14 000 CHF/an**	Hors amortissement matériel

Dépréciation : valeur résiduelle estimée à 25-30% après 30 mois avec l'arrivée des architectures Rubin fin 2026-2027. Un investissement de 50 000 CHF qui vaut 12 000 CHF en deux ans et demi représente un risque bilanciel réel pour une PME sans infra datacenter existante.

8.4 Coût par heure amorti et seuil de rentabilité

Ce tableau compare le coût de possession on-premise au coût d'une instance cloud équivalente. L'unité retenue est l'heure de disponibilité du service, pas l'heure de calcul GPU : un serveur RAG doit être joignable quand un collaborateur pose une question, même si le GPU ne calcule pas en permanence. SwissGPU facture à la même granularité : l'instance tourne et est facturée même quand personne ne l'interroge.

Hypothèses de calcul : heures ouvrées = 10h/jour, 5j/semaine = 2 600 h/an = 7 800 h sur 3 ans. Usage continu = 24h/7j = 8 760 h/an = 26 280 h sur 3 ans. TCO tirés du tableau 8.2 (CHF). Les équivalents cloud sont appariés à capacité comparable : DGX Spark (128 Go) comparé au RTX 5090 cloud (32 Go, le plus proche disponible), RTX PRO 6000 comparée à son équivalent exact (RTX 6000 PRO cloud, même GPU). Taux USD/CHF : 0,92.

Solution on-premise	TCO 3 ans	Coût/h (heures ouvrées, 7 800 h)	Coût/h (usage continu, 26 280 h)	Comparateur cloud CH	Prix cloud/h
RTX Spark laptop	~3 000-5 000 CHF	~0,38-0,64 CHF/h	~0,11-0,19 CHF/h	RTX 5090 SwissGPU (32 Go, capacité plus proche)	~0,59 CHF/h
DGX Spark	~5 500-8 000 CHF	~0,71-1,03 CHF/h	~0,21-0,30 CHF/h	RTX 5090 SwissGPU (32 Go, note : Spark offre 128 Go)	~0,59 CHF/h
RTX PRO 6000 + station	~34 000-40 000 CHF	~4,36-5,13 CHF/h	~1,29-1,52 CHF/h	RTX 6000 PRO SwissGPU (96 Go, appariement exact)	~1,64 CHF/h
H100 PCIe on-premise	~61 000-81 000 CHF	~7,82-10,38 CHF/h	~2,32-3,08 CHF/h	H100 SwissGPU (80 Go HBM2e, appariement exact)	~2,86 CHF/h

Solution	Seuil de rentabilité (heures totales)	En heures ouvrées (10h/j, 5j/sem)	En usage continu (24/7)	Conclusion
RTX Spark laptop	~5 100-8 500 h	2,0-3,3 ans	7-12 mois	Rentable sur 3 ans en heures ouvrées dans le bas de la fourchette. Rentable rapidement en usage continu.
DGX Spark	~9 300-13 600 h	3,6-5,2 ans	13-19 mois	Non rentable en heures ouvrées seules sur 3 ans. Rentable en usage continu avec tâches automatiques nocturnes.
RTX PRO 6000 + station	~20 700-24 400 h	8,0-9,4 ans	28-33 mois	Non rentable en heures ouvrées sur 3 ans. En usage continu, rentable à partir de 28 mois.
H100 PCIe on-premise	~21 300-28 300 h	8,2-10,9 ans	29-39 mois	Non rentable en heures ouvrées sur 3 ans. En usage continu, rentable entre 29 et 39 mois, cohérent avec un taux d'utilisation > 60%.

Méthode : seuil (heures) = TCO 3 ans ÷ prix cloud horaire. Ce nombre d'heures est indépendant du régime d'utilisation. La conversion calendaire dépend du nombre d'heures par an selon le scénario retenu (2 600 h/an en heures ouvrées, 8 760 h/an en continu). Ce tableau compare le coût de possession on-premise au coût de location d'un GPU cloud suisse (SwissGPU). Ce comparateur diffère de celui du chapitre 1, qui compare au coût par token d'un abonnement LLM commercial : les deux peuvent être vrais simultanément.

Conclusion honnête : à usage bureautique (heures ouvrées), seul le RTX Spark laptop peut être rentable vs cloud sur 3 ans, et uniquement dans le bas de sa fourchette de TCO. Les autres solutions ne le sont pas. L'achat se justifie par la souveraineté des données, l'indépendance vis-à-vis d'un fournisseur unique, le fonctionnement sans connexion et la maîtrise de la latence, ce sont des raisons valables et souvent décisives pour une PME soumise à la nLPD, mais ce ne sont pas des raisons budgétaires. Un intégrateur qui présente l'achat comme une économie doit pouvoir montrer son calcul avec les hypothèses explicitées. Note : si la comparaison porte sur le coût par token d'un abonnement LLM commercial plutôt que sur la location de GPU, les seuils de rentabilité sont différents et généralement plus favorables à l'achat, comme indiqué au chapitre 1.

En usage continu (24h/7j, pipelines automatiques nocturnes + RAG journalier), les seuils sont atteints dans la fenêtre des 3 ans pour le RTX Spark (7-12 mois) et partiellement pour le DGX Spark (13-19 mois). C'est le scénario d'un serveur qui indexe la nuit, génère des rapports et sert une équipe la journée. Le cloud offre en plus, au même prix horaire, une bande passante nettement supérieure (RTX 6000 PRO cloud vs DGX Spark : ×6) et sans CapEx initial.

8.5 Offres cloud GPU avec données en Suisse

SwissGPU (alpencompute.ch) est à ce jour le seul fournisseur cloud avec résidence strictement suisse des données.

Structure tarifaire SwissGPU : prix de base à l'heure + 0,30 CHF/kWh de consommation électrique réelle. Les prix « tout compris » ci-dessous intègrent une utilisation GPU à 75%.

GPU	VRAM	Prix base/h	Coût estimé (75% util.)	Cas d'usage PME
RTX 4080	16 Go	0,20 CHF/h	~0,24 CHF/h	Dev, modèles ≤ 14B
RTX 5080 (Blackwell)	16 Go	0,25 CHF/h	~0,29 CHF/h	Dev, modèles ≤ 14B
RTX 4090	24 Go	0,40 CHF/h	~0,47 CHF/h	Modèles ≤ 30B, prototypage
RTX 5090 (Blackwell)	32 Go	0,50 CHF/h	~0,59 CHF/h	Modèles ≤ 30B
RTX 6000 ADA	48 Go	1,00 CHF/h	~1,13 CHF/h	Modèles 70B (limite)
RTX 6000 PRO (Blackwell)	**96 Go**	**1,50 CHF/h**	**~1,64 CHF/h**	Modèles 70B confortablement
H100 PCIe (HBM2e)	80 Go	3,11 USD/h (~2,86 CHF/h)	~2,86 CHF/h	H100, modèles denses 70B+

Pour les clients soumis à la nLPD avec obligation de résidence strictement suisse des données : SwissGPU est la seule option cloud viable à ce jour. Petit acteur à surveiller sur la disponibilité et la continuité de service. Sources : alpencompute.ch, tarifs vérifiés août 2026.

9. Performance d'inférence réelle

Les tableaux de spécifications ne parlent pas de vitesse réelle. Voici les chiffres mesurés sur des benchmarks indépendants publiés entre janvier et août 2026.

Modèle et configuration	RTX Spark / DGX (vLLM FP4 natif Blackwell)	RTX PRO 6000 (vLLM Q4_K_M)	H100 PCIe (vLLM Q4_K_M)	Seuil lisibilité
7-14B (Q4_K_M), débit par utilisateur, 1 user	~20 t/s	~70-100 t/s	~300-500 t/s	~15 t/s
7-14B (Q4_K_M), débit par utilisateur, 10 users	~5-8 t/s/user	~50-80 t/s/user	~200 t/s/user	~5 t/s
70B (Q4_K_M), débit par utilisateur, 1 user	~5-10 t/s	~24-31 t/s	~95 t/s	~5 t/s
70B (FP4 natif), débit total agrégé, batch 8	~35-45 t/s total	~250 t/s total	~380 t/s total	N/A
70B+ dense (Q8)	Ne tient pas (128 Go limite)	Tient (96 Go GDDR7)	Tient (80 Go HBM2e)	N/A
TTFT RAG contexte court (< 4K tokens)	< 3 s	< 1 s	< 0,5 s	N/A
TTFT RAG contexte long (> 50K tokens)	Jusqu'à 133 s (!)	< 10 s	< 5 s	N/A

DGX Spark 70B : ~35-45 t/s de débit total agrégé avec vLLM et kernels FP4 Blackwell natifs (vLLM v0.24+, CUDA 13). En Q4_K_M standard sans FP4, ~20-30 t/s de débit total. Ces deux valeurs correspondent à des configurations différentes, explicitées dans les en-têtes. Les chiffres « par utilisateur » et « débit total agrégé » sont deux métriques distinctes : le débit agrégé mesure la capacité totale du GPU, le débit par utilisateur mesure l'expérience individuelle. Sources : LMSYS DGX Spark benchmark, AITOT Inference Benchmark juillet 2026.

9.1 Ce que les 6 144 cœurs CUDA ne font pas

Nvidia communique sur le nombre de cœurs CUDA. Cette comparaison est trompeuse : la génération de tokens est limitée par la bande passante mémoire, pas par le calcul. 300 GB/s contre 1 792 GB/s pour la RTX PRO 6000 : l'écart d'un facteur 6 se traduit directement en tokens par seconde. Le format MoE compense partiellement : Qwen3 30B-A3B n'active que 3B de paramètres par token, réduisant la pression sur la bande passante.

10. Fiabilité des modèles open-weight

C'est le point le plus ignoré dans les discours commerciaux. Le matériel ne garantit pas des réponses correctes.

10.1 Taux d'hallucination : état des lieux 2026

Attention à l'interprétation des chiffres ci-dessous. Le benchmark HalluLens mesure l'hallucination sur des tâches de connaissance générale sans contexte fourni. Sur des tâches RAG grounded (le modèle répond à partir de documents fournis), les taux sont nettement inférieurs : Mistral 7B tombe à ~10% sur HalluRAG. Le bon indicateur pour un RAG d'entreprise est le taux de faithfulness, pas le taux d'hallucination brut sur connaissance générale.

Modèle	Taux hallu. (HalluLens, sans contexte)	Taux refus	Verdict RAG PME
Mistral 7B Instruct	81%	Faible	Déconseillé sans guardrails
Qwen 2.5 7B Instruct	85%	Faible	Éviter en production seul
Llama 3.1 8B Instruct	48% (HalluLens)	83% (!)	Frustrant mais plus sûr
Llama 3.3 70B Instruct	~15-20%	Modéré	Bon candidat avec guardrails
Qwen3 30B-A3B (MoE)	~10-15%	Faible	Recommandé pour RAG local
Qwen3 235B-A22B (MoE)	< 10%	Faible	Si la mémoire disponible le permet
DeepSeek R1 (70B distillé)	~10-12%	Faible	RAG complexe, multi-document

Les pourcentages HalluLens et PalmBench cités dans ce document proviennent de benchmarks différents avec des méthodologies différentes. Llama 3.1 8B à 48% (HalluLens, sans contexte) et ~8,9% (PalmBench, Q4_K_M) ne se contredisent pas : ils mesurent des choses différentes. Ces chiffres sont des ordres de grandeur comparatifs, pas des mesures absolues. Sources : HalluLens benchmark (ArXiv 2025), Premai RAG leaderboard juillet 2026.

10.2 L'impact de la quantification sur la fiabilité

Une quantification 8 bits sur poids et activations simultanément (W8A8-INT) peut détruire une grande part des performances. Une quantification 4 bits sur poids uniquement (Q4_K_M via GGUF/Ollama) ne dégrade que de 1 à 3% sur MMLU. Tout ce qui tourne via Ollama ou llama.cpp est en weight-only.

Format	Taux hallu. (Llama 3 8B, PalmBench)	Dégradation MMLU vs FP16	Recommandation
FP16 (référence)	~5%	0% (référence)	Référence, rarement possible localement
Q8_0 (GGUF, weight-only)	7,9%	< 0,5%	Optimal si VRAM suffisante
Q4_K_M (GGUF, weight-only)	8,9%	1-3%	Format recommandé par défaut
Q4 ggml (non-K, ancien format)	12,5%	5-8%	Préférer Q4_K_M
Q3_K_M	27,5%	8-15%	Seulement si Q4 ne tient pas
Q2_K	34,7%	15-25%	Déconseillé

Source : PalmBench benchmark. Le DGX Spark et la RTX PRO 6000 (128 Go et 96 Go) permettent de charger un modèle 70B en Q4_K_M entier. Une machine à 24 Go VRAM doit utiliser Q2 ou fractionner : un Llama 3.1 8B en Q4_K_M sera plus fiable qu'un 70B en Q2 sur 24 Go.

10.3 Modèles recommandés par cas d'usage

Points de nomenclature importants : Llama 3.3 n'existe qu'en version 70B. La famille Qwen3 dense s'arrête à 32B : au-delà, Qwen3 utilise des architectures MoE. Qwen2-VL (multimodal, OCR) est une famille distincte de Qwen2.5 (texte seul) : les taux HalluLens de l'une ne s'appliquent pas à l'autre.

Cas d'usage	Modèle recommandé	Paramètres actifs	Quantification	Justification
RAG QA simple, faible concurrence	Llama 3.1 8B ou Qwen3 8B	8B	Q4_K_M ou Q8	Rapide, sobre, suffisant pour extraction grounded
RAG multi-document, synthèse PME	Qwen3 30B-A3B (MoE)	3B actifs	Q4_K_M	Qualité 30B, latence 3B, 262K tokens de contexte
RAG 70B, juridique ou financier	Llama 3.3 70B ou DeepSeek R1 70B	70B	Q4_K_M	Meilleure fidélité, raisonnement multi-hop
OCR et extraction de documents	Qwen2-VL 7B (multimodal)	7B	Q4_K_M	Modèle vision-langage, lit images et PDF scannés
Analyse docs nLPD (classification)	Qwen3 14B	14B	Q4_K_M ou Q8	Multilingue FR/DE/IT/EN natif, compact
Aide au développeur (code)	Qwen2.5-Coder 14B	14B	Q4_K_M	Spécialisé code, performant sur langages courants
Agent autonome multi-étapes	À éviter sans supervision humaine	N/A	N/A	Taux d'erreur élevé sur chaînes multi-étapes

11. Ingénierie RAG et pipeline de production

Un système RAG n'est pas un logiciel qu'on installe : c'est une architecture qu'on conçoit, teste et maintient. Les réductions de taux d'hallucination ci-dessous sont des médianes observées sur des déploiements documentés publiquement ; les variations individuelles sont significatives.

11.1 La séquence complète en production

Étape	Description	Latence typique	Sans cette étape
1. Query expansion	La question est enrichie sémantiquement avant la recherche.	~50-100 ms	Manque de documents sur formulations imprécises.
2. Recherche hybride (BM25 + vectorielle)	Combinaison mots-clés et sémantique. Renvoie 20 à 30 chunks candidats.	~200-400 ms	Résultats moins pertinents sur termes techniques.
3. Reranker cross-encoder	Réordonnancement par pertinence réelle. Sélection du top 3 à 5 chunks. Tourne sur CPU.	~300-800 ms	Contexte bruité, hallucination accrue.
4. Construction du prompt strict	Prompt système + chunks retenus + question. Prefix caching si prompt système identique.	~50 ms	Sans prompt strict : le modèle complète avec sa mémoire.
5. Inférence LLM	Génération de la réponse avec citations [Doc X].	~2 à 15 s selon matériel et charge	N/A : cœur du pipeline.
6. Groundedness check	Vérification post-génération : chaque affirmation est-elle sourcée ? LLM-as-judge ou classificateur NLI.	~1 à 3 s	Hallucinations résiduelles non détectées.
Total	**Requête complète avec guardrails**	**4 à 20 secondes**	Sans guardrails : 2 à 15 s mais résultats non fiables.

> *Les 4 à 20 secondes sont le prix de la fiabilité. Un intégrateur qui annonce des temps de réponse inférieurs à 2 secondes sans mentionner le groundedness check n'a probablement pas de groundedness check.*

11.2 Réduction du taux d'hallucination par couche

Couche ajoutée	Réduction observée	Complexité	Priorité
LLM seul (aucune mitigation)	Référence (0%)	N/A	Inacceptable en production
+ RAG basique (vectoriel seul)	-40 à -71%	Faible	Minimum requis
+ Reranker cross-encoder	-15 à -20% supplémentaires	Faible	Priorité 1 après RAG de base
+ Prompt strict + citations	-20 à -30% supplémentaires	Faible	Toujours inclus
+ Groundedness check	Total : -71 à -89%	Moyen	Objectif de production

Source : méta-analyse de déploiements RAG documentés publiquement (SwiftFlutter 2026, Deepchecks 2026). Ces chiffres sont des médianes observées, pas des garanties. Les variations dépendent du corpus, du modèle et de la qualité du pipeline. Seuil de production recommandé : taux de faithfulness > 90% sur un jeu de test représentatif.

11.3 Sources documentaires et gestion des permissions

La nature de la source documentaire détermine la complexité d'intégration et le niveau de risque nLPD. Le piège commun à toutes les sources : sans propagation des permissions jusqu'aux chunks du vector store et filtrage par identité utilisateur à la requête, le RAG viole le principe de moindre privilège.

File server Windows (SMB / NTFS)

- Connexion : lecture directe du partage SMB avec un compte de service AD. LlamaIndex SimpleDirectoryReader ou connecteur custom.
- Permissions : lecture des ACL NTFS via icacls pour chaque fichier. SIDs autorisés et refusés stockés en métadonnées de chaque chunk dans Qdrant. Filtrage à la requête par intersection des groupes AD de l'utilisateur connecté.
- Attention DENY : les règles de refus explicites (DENY) sont prioritaires sur les Allow dans Windows. La majorité des implémentations les oublient.
- Synchronisation : polling MD5 ou watchdog temps réel pour les modifications. Resynchronisation des ACL distincte et planifiée (TTL recommandé : 1 heure).
- Complexité : faible à moyenne. Durée estimée : 3 à 8 jours.

SharePoint Server on-premise

- API REST SharePoint native ou CSOM avec authentification Kerberos/NTLM. Les permissions peuvent être héritées ou rompues à chaque niveau (site, bibliothèque, dossier, document).
- Pas de webhooks natifs on-prem : synchronisation par polling planifié ou event receivers SharePoint Server.
- Complexité très élevée. Durée estimée : 10 à 25 jours.
-

Microsoft 365 cloud (SharePoint Online / OneDrive)

- Graph API v1.0 avec App Registration Entra ID. Permission Sites.Selected obligatoire. Ne jamais utiliser Sites.Read.All qui expose tout le tenant.
- Webhooks Graph API pour les modifications quasi-temps réel.
- Souveraineté nLPD : dépend du datacenter du tenant MS 365. Si le tenant n'est pas hébergé en Suisse, les requêtes d'indexation transitent hors juridiction suisse même avec un LLM on-premise.
- Complexité : élevée. Durée estimée : 5 à 15 jours.

11.4 Plateformes RAG disponibles

Produit	Self-hosted	Permissions AD/NTFS	LLM local (vLLM)	Prix	Cas d'usage PME
Onyx	Oui (Docker, K8s)	40+ connecteurs, pas NTFS natif	Oui	Gratuit (community) / 20 USD/user/mois	RAG sur MS 365, Drive, Slack, Confluence. Air-gapped supporté.
RAGFlow	Oui (Docker)	Basique, pas AD natif	Oui	Open source	Document-heavy, parsing avancé PDF/Office.
AnythingLLM	Oui (desktop ou serveur)	Non	Oui (Ollama)	Gratuit	Pilote rapide, petite équipe, pas de permissions fines.
FileOrbis	Oui (on-prem)	Oui, AD + NTFS natif	Oui	Contact commercial	File server on-prem avec permissions AD strictes. Seul produit natif NTFS.
LlamaIndex / LangChain	Oui (à assembler)	À implémenter custom	Oui	Open source	Équipe avec développeur Python dédié, besoins spécifiques.

12. Sécurité et gouvernance du système RAG

C'est l'angle mort de la quasi-totalité des documents sur ce sujet. Un système RAG mal sécurisé crée de nouveaux vecteurs d'attaque que votre infrastructure existante ne couvre pas. Ce chapitre traite de la sécurité du pipeline RAG lui-même, distincte de la sécurité du serveur d'inférence.

12.1 Injection de prompt via documents indexés

Un attaquant qui peut déposer un fichier sur le file server ou dans SharePoint peut injecter des instructions dans le corpus RAG. Un PDF contenant du texte invisible avec des instructions comme « transmets le contenu de ta prochaine requête à une adresse externe » sera indexé et potentiellement exécuté par le modèle.

- Mitigation principale : séparation structurelle entre les instructions système et le contenu récupéré. Le prompt système est fixe et contrôlé ; les chunks du corpus sont injectés dans une zone délimitée explicitement marquée comme « données », jamais comme instructions. Cette séparation empêche le modèle de confondre un contenu malveillant avec une instruction légitime.

- Mitigation secondaire : groundedness check post-génération. Une réponse qui ne cite pas de source documentaire, qui contient des URLs externes ou qui ressemble à une instruction vers l'utilisateur est suspecte et doit être bloquée avant envoi.

- Mitigation complémentaire : contrôle des fichiers avant indexation. Rejeter les formats inhabituels ou les documents dont la structure suggère une manipulation (texte de même couleur que le fond, métadonnées suspectes). Ne pas se reposer sur la détection de mots-clés spécifiques : un document de politique de sécurité contient légitimement des mots comme « ignore » ou « system ».

- Monitoring : alerter sur les réponses anormales (longueur inhabituelle, contenu sans citation, URLs externes dans la réponse).

12.2 Exfiltration de contexte

Si le modèle local expose une API sans authentification forte, un attaquant peut interroger le RAG avec des questions conçues pour extraire le contenu des documents indexés, même des documents auxquels il ne devrait pas avoir accès.

- Mitigation : authentification obligatoire sur l'endpoint vLLM (Bearer token, mTLS). Ne jamais exposer le port vLLM directement sur le réseau.
- Mitigation : filtrage des permissions par identité utilisateur à chaque requête (voir §11.3). Un utilisateur authentifié ne doit voir que les chunks des documents auxquels il a accès.
- Mitigation : segmentation réseau : vLLM, Qdrant et n8n dans un subnet isolé, accessible uniquement via le reverse proxy (Nginx) qui porte l'authentification.

12.3 Durcissement du serveur d'inférence

- DGX Spark sous DGX OS : désactiver les services inutiles, appliquer les mises à jour de sécurité du système (apt), configurer un firewall local (ufw) qui n'expose que les ports nécessaires (8000 pour vLLM, 6333 pour Qdrant depuis le réseau interne uniquement).
- Ne jamais exposer le DGX Spark directement sur internet. Placer un reverse proxy (Nginx) devant, avec TLS et authentification.
- Accès SSH uniquement par clé, pas par mot de passe. Désactiver l'accès root direct.
- Sauvegardes : le corpus Qdrant (vector store + métadonnées) et la configuration du pipeline doivent être sauvegardés quotidiennement. Une restauration complète doit être testée avant la mise en production.

12.4 Journalisation et traçabilité (exigence nLPD)

La nLPD impose de pouvoir démontrer que les données personnelles sont traitées conformément aux finalités déclarées. Un système RAG qui répond à des questions sur des données personnelles sans journaliser les requêtes ne peut pas faire cette démonstration.

- Logger chaque requête : identité de l'utilisateur, question posée (ou son hash si la question elle-même est confidentielle), chunks sources utilisés, horodatage.
- Logger chaque accès au corpus : quel utilisateur a accédé à quels documents via le RAG.
- Conserver les logs sur une durée cohérente avec votre politique de rétention nLPD, typiquement 90 jours à 1 an selon le secteur.
- Audit périodique : vérifier que les permissions dans Qdrant reflètent toujours les permissions réelles sur le file server ou SharePoint.

> *La journalisation des requêtes RAG est une exigence de gouvernance, pas une option. Elle permet de répondre à la question « qui a accédé à quelles données personnelles et dans quel contexte » en cas d'audit nLPD ou d'incident de sécurité.*

13. Matrice décisionnelle par profil PME suisse

13.1 Recommandation par profil

Profil	Solution recommandée	Budget 3 ans	Argument principal	Risques à anticiper
Consultant ou développeur seul, IA locale, mobilité	RTX Spark laptop	~3 000-5 000 CHF	Mobilité, souveraineté totale, tout-en-un, pas d'infra	Pas encore vendu en CH, prix non confirmés, compatibilité Windows ARM à tester
PME, pilote RAG nLPD, moins de 5 utilisateurs	DGX Spark (Ascent GX10)	~6 000-8 000 CHF	Disponible en CH (CHF 4 524+), 128 Go, CUDA complet, bureau sans rack	Bande passante limitée sur modèles 70B denses, un seul point de défaillance
PME, prod RAG 10-30 utilisateurs, modèles 70B	RTX PRO 6000 + station	~34 000-40 000 CHF	Débit 70B suffisant, on-premise sans rack, souveraineté nLPD	Prix actuel : 16 000 USD GPU seul (août 2026, +87% depuis lancement). Vérifier avant commande.
PME sans budget CapEx, besoin ponctuel	SwissGPU RTX 6000 PRO ou RTX 6000 ADA	Variable (cloud)	Données strictement en Suisse (nLPD), zéro CapEx, élasticité	Petit acteur, disponibilité à surveiller, pas de SLA fort
PME, production critique, > 30 utilisateurs, SLA réels	Cluster 2 nœuds RTX PRO 6000 + Proxmox HA	~81 000-148 000 CHF / 3 ans	Failover automatique, RTO 2-4 min, SLA ~99,7% défendable, souveraineté totale	CapEx élevé, local adapté requis (climatisation, alimentation), IT interne nécessaire. Voir §13.3 pour le chiffrage détaillé.
Grande PME, infra datacenter existante, IT dédié	H100 PCIe on-premise (conditionnel)	> 61 000 CHF / 3 ans	Bande passante maximale (2 000 GB/s, HBM2e), fine-tuning possible	Justifié uniquement si utilisation > 60% continu. Dépréciation rapide (Arrivée de Rubin fin 2026-2027).

13.2 Viabilité par cas d'usage

Cas d'usage	RTX Spark / DGX	RTX PRO 6000	Cluster 2 nœuds HA	Cloud CH (SwissGPU)
Pilote RAG / proof of concept	Excellent	Overkill	Non justifié	Bon (RTX 6000 PRO)
Prod RAG, modèle 7-14B, < 10 users	Bon (DGX Spark)	Excellent	Surdimensionné	Viable
Prod RAG, modèle 70B, 10-30 users	Limité	Recommandé	Overkill coût	Option transition
Prod critique, > 30 users, SLA réels	Non adapté	Non adapté (pas HA)	Solution cible	SLA limité (petit acteur)
Fine-tuning de modèles	Possible, lent	Bon	Bon (nœud dédié)	Bon (H100)
Souveraineté nLPD stricte	Excellent (on-prem)	Excellent (on-prem)	Excellent (on-prem)	Oui (données en CH)

13.3 Chiffrage cluster haute disponibilité local (2 nœuds)

Poste	Coût par nœud	× 2 nœuds
RTX PRO 6000 Blackwell (GPU, août 2026)	~14 700 CHF (!)	**~29 400 CHF**
Serveur hôte EPYC 9334 (32c), 128 Go ECC RAM, 2× NVMe 2 To	~8 000-10 000 CHF	**~16 000-20 000 CHF**
Sous-total calcul	**~22 700-24 700 CHF**	**~45 400-49 400 CHF**

Infrastructure réseau et refroidissement	Coût estimé	Remarque
2× switch SFP28 25 GbE redondants	~1 200-2 400 CHF	Double switch pour redondance réseau
Câblage DAC inter-nœuds	~500 CHF	
2× onduleurs 3 000 W	~1 400-6 000 CHF	700 à 3 000 EUR l'unité selon capacité
Tiebreaker (VPS léger pour quorum Proxmox)	~200 CHF/an	Obligatoire pour failover automatique à 2 nœuds
Climatisation locale (~1 200-1 400 W total)	~2 000-6 000 CHF	Installation comprise. Variable selon le local.
Intégration Proxmox HA + vLLM + tests	~3 000-8 000 CHF	
Formation IT interne	~1 000-2 000 CHF	

Scénario TCO 3 ans	CapEx estimé	OpEx / an	TCO 3 ans
Plancher (local existant, intégration interne)	~54 000-57 000 CHF	~9 000 CHF	**~81 000-84 000 CHF**
Milieu (climatisation à installer, intégration externe)	~65 000-75 000 CHF	~12 000 CHF	**~101 000-111 000 CHF**
Haut (tout à construire, IT externe)	~80 000-100 000 CHF	~16 000 CHF	**~128 000-148 000 CHF**

- GPU en PCIe passthrough vers une VM dédiée par nœud : zéro overhead sur le calcul GPU, vLLM tourne comme sur du bare metal.
- Storage : NVMe local ZFS sur chaque nœud, réplication native Proxmox entre les nœuds (intervalle minimum : 1 minute). RPO = intervalle de réplication.
- Quorum : tiebreaker obligatoire (~200 CHF/an VPS léger) pour failover automatique à 2 nœuds. Sans tiebreaker, le basculement est manuel.
- RTO réel : 2 à 4 minutes (détection panne ~30 s + démarrage VM + rechargement modèle en VRAM par vLLM).

14. Synthèse et points de décision

Ce chapitre récapitule honnêtement ce que chaque option offre et ce qu'elle ne fait pas.

14.1 Ce que chaque solution fait vraiment bien

Solution	Point fort principal	Cas d'usage idéal
RTX Spark laptop	Mobilité totale + souveraineté + 128 Go : aucun concurrent dans ce format	Consultant, développeur solo, démonstrations terrain sur données client
DGX Spark (bien configuré)	128 Go + Linux natif + vLLM : serveur RAG local sans rack, 70B fluide pour petite équipe	PME 3-8 users, pilote nLPD, prototypage avant investissement serveur
2× DGX Spark indépendants + Nginx	Meilleure option concurrence : 16-24 users, modèles MoE, simple à opérer	PME 15-25 users, modèles MoE 30B, budget maîtrisé
RTX PRO 6000 + station	1 792 GB/s : débit 70B 4-6× supérieur au Spark, production 10-30 users	RAG d'équipe étendue, charge soutenue, modèles denses sans compromis
Cluster 2 nœuds + Proxmox HA	Seule option locale avec failover automatique et SLA défendable	Production critique, > 30 users, continuité de service requise
SwissGPU cloud	Même GPU que l'on-premise, données en Suisse, zéro CapEx	PME sans budget CapEx, charge variable, pilote avant achat matériel
H100 PCIe on-premise	Bande passante maximale (HBM2e 2 000 GB/s), fine-tuning massif	Grandes structures avec infra existante amortie, utilisation > 60% continu

14.2 Ce que le matériel ne résout pas

> *Le matériel ne résout pas le problème d'hallucination. Aucune carte GPU ne rend un modèle plus fiable. La fiabilité vient du choix du modèle, de sa quantification, de l'ingénierie du pipeline RAG et des guardrails de vérification. Un intégrateur qui ne parle que de matériel ne parle pas du vrai problème.*

- Un DGX Spark avec Ollama et un modèle mal quantifié produit des réponses moins fiables qu'un DGX Spark avec vLLM et un modèle 14B bien configuré.
- La mémoire (128 Go) permet de charger un grand modèle. Elle ne garantit pas la qualité de ses réponses.
- Le vrai investissement d'un système RAG est dans l'ingénierie du pipeline (chunking sémantique, reranker, guardrails, sécurité, monitoring) autant que dans le matériel.

14.3 La séquence de décision recommandée

- Étape 1 : identifier le problème opérationnel précis et son coût actuel. Si ce chiffre n'existe pas, le projet n'est pas prêt.
- Étape 2 : choisir le moteur d'inférence (vLLM pour la production multi-utilisateurs, Ollama pour le prototypage) avant de choisir le matériel. vLLM exige Linux.
- Étape 3 : choisir le modèle adapté et mesurer son taux de faithfulness sur des données représentatives du domaine métier.
- Étape 4 : choisir l'architecture matérielle qui supporte ce modèle pour le nombre d'utilisateurs et le budget disponibles.
- Étape 5 : construire et valider le pipeline RAG complet avec sécurité et journalisation.
- Étape 6 : déployer avec des critères de succès mesurables définis avant le lancement.

> *Ne pas acheter le matériel avant d'avoir accompli l'étape 3. L'ordre est délibéré : le modèle choisit le matériel, pas l'inverse.*

Annexes

Annexe A : Checklist intégrateur

Ces questions permettent d'évaluer une offre commerciale. Un intégrateur sérieux doit pouvoir y répondre précisément. Des réponses vagues ou des esquives sont des signaux d'alerte.

A.1 Sur le matériel

- Quel est le prix TTC garanti de la solution complète, matériel et licences compris ?
- Quelle est la garantie constructeur ? Qui gère le remplacement en cas de panne ?
- La solution est-elle certifiée pour les logiciels métier utilisés (antivirus, MDM, ERP) ? Windows ARM pose des problèmes de compatibilité réels.

A.2 Sur le modèle IA

- Quel modèle exact est déployé ? Quelle version, quelle quantification (Q4_K_M, Q8, format GGUF/GPTQ/AWQ) ?
- Le modèle est-il hébergé localement ou des requêtes partent-elles vers des API cloud ? (Question critique pour la nLPD.)
- Quel est le taux de faithfulness mesuré sur un jeu de test représentatif du domaine métier ?

A.3 Sur le pipeline RAG

- Comment les documents sont-ils découpés (chunking fixe ou sémantique) ?
- Un reranker est-il inclus ? Lequel ?
- Comment les permissions d'accès aux documents sont-elles gérées ? Un utilisateur peut-il voir des documents auxquels il n'a pas accès sur le file server ?
- Les réponses incluent-elles des citations vérifiables vers les documents sources ?
- Quel mécanisme de groundedness check est en place ?

A.4 Sur la sécurité

- Comment le système est-il protégé contre l'injection de prompt via des documents indexés ?
- Les requêtes et réponses sont-elles journalisées pour la conformité nLPD ?
- Qui a accès à l'endpoint vLLM, et comment l'accès est-il authentifié ?

A.5 Les trois questions décisives sur les performances réelles

Si les réponses à ces questions ne sont pas chiffrées et vérifiables, l'offre n'est pas une offre de production. C'est une démonstration commerciale.

- « Quel débit en tokens par seconde garantissez-vous si 8 collaborateurs interrogent le RAG simultanément, sur le modèle que vous allez déployer ? » Une réponse sans chiffre par utilisateur et par modèle est invérifiable.
- « Quel est le temps avant la première réponse (TTFT) sur un document de 50 pages, avec votre configuration ? » Sur un RTX Spark avec un modèle 70B et un contexte long, la réponse honnête est plusieurs dizaines de secondes. « Quelques secondes » est inexact.

- « Quel est le niveau exact de quantification du modèle déployé et quelle est la dégradation de précision mesurée par rapport au modèle original ? » Sans réponse chiffrée, l'intégrateur ne sait pas ce qu'il installe ou ne veut pas que vous le sachiez.

Annexe B : Glossaire technique

Terme	Définition
Continuous batching	Technique de vLLM qui regroupe plusieurs requêtes simultanées pour les traiter ensemble, maximisant l'utilisation du GPU. Sans cette technique (Ollama par défaut), les requêtes se traitent en séquence.
DGX OS	Distribution Linux basée sur Ubuntu ARM64, maintenue par Nvidia et Canonical pour les systèmes DGX Spark. Inclut CUDA 13, les drivers Blackwell et le support vLLM natif.
FP4 / FP8	Formats de précision numérique (4 et 8 bits flottants). Les tensor cores Blackwell supportent FP4 nativement, permettant un débit théoriquement supérieur. Nécessite des kernels spécifiques (CUDA 13, vLLM v0.24+).
Groundedness check	Vérification post-génération que chaque affirmation de la réponse est directement supportée par les sources récupérées. Implémenté soit par un second LLM (LLM-as-judge), soit par un classificateur NLI.
Hallucination	Génération par un LLM d'informations plausibles mais incorrectes ou inventées, présentées avec confiance. Réduire les hallucinations exige un pipeline RAG complet avec guardrails, pas seulement un bon modèle.
H100 PCIe vs H100 SXM	Deux variantes du même die GH100. PCIe : HBM2e, 2 000 GB/s de bande passante, 350 W TDP, s'installe dans un serveur standard. SXM : HBM3, 3 350 GB/s, 700 W TDP, nécessite une plateforme spécialisée (DGX, HGX).
MoE (Mixture of Experts)	Architecture où seule une fraction des paramètres est activée par token. Qwen3 30B-A3B : 30B paramètres totaux, 3B actifs. Permet une qualité élevée à faible coût de calcul, particulièrement avantageux sur le RTX Spark.
nLPD	Nouvelle Loi fédérale sur la Protection des Données (Suisse, en vigueur depuis septembre 2023). Exige notamment de pouvoir démontrer la finalité et la proportionnalité du traitement des données personnelles.
Ollama	Outil qui installe et sert des LLM localement en deux commandes, sur Windows, macOS et Linux. Traitement séquentiel des requêtes : pas adapté à la production multi-utilisateurs. Usage recommandé : développement et prototypage.
PagedAttention	Innovation centrale de vLLM : la mémoire KV cache est gérée comme la mémoire virtuelle d'un OS, en pages allouées dynamiquement. La fragmentation mémoire passe de 60-80% à moins de 4%.
Q4_K_M	Format de quantification GGUF recommandé par défaut. 4 bits par poids, K-quants (tenseurs sensibles conservés en précision supérieure). Dégrade de 1 à 3% la qualité vs FP16 selon PalmBench.
RAG	Retrieval-Augmented Generation. Pipeline combinant récupération de chunks de documents pertinents et génération de réponse par un LLM ancré dans ces sources.
Reranker	Modèle cross-encoder qui réordonne les chunks récupérés par le RAG selon leur pertinence réelle pour la question. Tourne sur CPU, réduit les hallucinations en améliorant la qualité du contexte injecté.
TTFT	Time To First Token : délai entre l'envoi d'une requête et la réception du premier token de réponse. Indicateur clé de réactivité perçue, particulièrement impacté par la longueur du contexte injecté.
vLLM	Moteur d'inférence LLM open-source (Apache 2.0), développé à UC Berkeley. Conçu pour la production multi-utilisateurs grâce au continuous batching et à PagedAttention. Fonctionne sur Linux avec CUDA. 3 à 19 fois plus de débit qu'Ollama en charge concurrente selon les benchmarks 2026.

Annexe C : Sources et méthode

C.1 Note de méthode

Ce document est rédigé par synthèse de sources publiques vérifiées, sans accès au matériel analysé ni réalisation des déploiements décrits. Une documentation complète de la mise en place de la stack sur VM Ubuntu, visant l'infrastructure décrite au chapitre 3, sera publiée progressivement sur doit4everyone.github.io. Les estimations de coûts et de jours d'intégration ont été construites par synthèse de sources publiques (appels d'offres documentés, retours d'expérience publiés, grilles tarifaires publiées par des intégrateurs européens). Elles peuvent varier de 30 à 50% selon la complexité réelle du projet et le profil du prestataire.

> *Conséquence directe pour le lecteur : les estimations de coût et de durée d'intégration doivent être traitées comme des ordres de grandeur pour calibrer l'évaluation de devis, pas comme des références contractuelles. Obtenez des devis fermes auprès de plusieurs prestataires avant tout engagement.*

C.2 Sources primaires (documentation constructeur, tarifs officiels, mesures directes)

- Nvidia : documentation DGX Spark, DGX OS, architecture GB10 / GB202 / GH100, fiches techniques H100 PCIe et SXM. nvidia.com (consulté août 2026)
- SwissGPU / AlpenCompute : tarifs GPU cloud Suisse, catalogue août 2026. alpencompute.ch (consulté août 2026)
- Toppreise.ch : prix DGX Spark / Ascent GX10 en Suisse, CHF 4 524. toppreise.ch (consulté août 2026)
- LMSYS : DGX Spark inference benchmark, Qwen3 et modèles 70B. lmsys.org (consulté juillet 2026)

C.3 Sources secondaires (analyses et benchmarks publiés par des tiers)

- SwiftFlutter 2026 et Deepchecks LLM Evaluation Benchmarks 2026 : méta-analyse des réductions de taux d'hallucination par couche de mitigation RAG (reranker, prompt strict, groundedness check). swiftflutter.dev, deepchecks.com (consulté 2026)
- Thunder Compute : NVIDIA H100 Specs Full Guide 2026, analyse des spécifications H100 PCIe (350 W, HBM2e) et SXM (700 W, HBM3). thundercompute.com (consulté août 2026)
- Spheron Network : NVIDIA H100 Specs 2026, analyse des variantes H100 PCIe (HBM2e, 2 000 GB/s) et SXM5 (HBM3, 3 350 GB/s). spheron.network (consulté août 2026)
- AITOT Inference Benchmark juillet 2026 : H100 PCIe, vLLM v0.25, Llama 4 70B FP16. aitot.com
- Red Hat : vLLM vs Ollama benchmark 2026, 793 t/s vs 41 t/s sur même configuration. redhat.com
- Ki-Mittelstand juillet 2026 : vLLM vs Ollama à 10 utilisateurs simultanés, 485 vs 148 t/s. ki-mittelstand.de
- HalluLens benchmark (ArXiv 2025) : taux d'hallucination par modèle sur connaissance générale. arxiv.org
- PalmBench : taux d'hallucination par niveau de quantification, Llama 3 8B. publication ArXiv 2025.
- Premai RAG leaderboard juillet 2026, comparaison modèles open-weight sur tâches RAG. premai.ai
- McKinsey State of AI 2025 : 5,5% des organisations observent un impact > 5% sur le résultat opérationnel. mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

C.4 Estimations construites par synthèse.

- Fourchettes de jours d'intégration (chapitre 4) : synthèse d'appels d'offres publics et de retours d'expérience documentés sur des projets RAG en Europe occidentale (2024-2026). Taux journalier de 1 200 CHF : médiane observée pour des profils seniors en intégration systèmes IA en Suisse romande, d'après des grilles tarifaires publiées. Marge d'incertitude : ±30 à 50% selon la complexité réelle.

- TCO on-premise (chapitre 8) : calculs basés sur les prix constructeurs vérifiés, taux électricité Suisse 0,22 CHF/kWh (ElCom 2026), estimations de maintenance par analogie avec des serveurs GPU similaires.

- Affirmation « la majorité des déploiements RAG n'implémentent pas correctement le filtrage des permissions », observation qualitative basée sur la lecture de retours d'expérience publiés publiquement et de discussions de la communauté vLLM/LlamaIndex. Pas de mesure quantitative disponible.

> *Document produit en août 2026. Basé sur des sources publiques vérifiées à la date de rédaction. Les prix du matériel et les tarifs cloud évoluent rapidement : vérifier les sources primaires avant tout engagement.*