LIVRE BLANC TECHNIQUE & MESURES PHYSIQUES

Consommation et efficience d'une IA locale

Mesures physiques en conditions réelles : puissance électrique, précision des réponses, passage à l'échelle et coûts opérationnels.

Auteur de l'étude : IDfr / Laurent Naigeon

Périmètre : 8 modèles de langage testés en local, 750 questions réparties en 7 familles métiers

Banc d'essai matériel : Serveur bi-GPU NVIDIA RTX 5090 (64 Go VRAM) et mesure électrique temps réel à la prise (500 ms)

Date de publication : Novembre 2025 / Édition actualisée 2026

Statut : Données ouvertes & reproductibles

Sommaire de l'étude

Pourquoi cette étude ?

Le déploiement de modèles de langage (LLM) en interne soulève trois questions majeures pour une organisation : la confidentialité des données, le coût énergétique et la capacité à traiter les tâches du quotidien.

Pour apporter des réponses factuelles, nous avons conçu et instrumenté notre propre serveur de calcul. L'ambition de ce travail repose sur trois piliers directeurs :

IA Souveraine	Hébergement intégral sur nos propres machines. Les données sensibles, documents clients et bases internes ne transitent par aucun prestataire tiers.
IA Frugale	Recherche du plus petit modèle capable d'accomplir la tâche avec justesse. Réduction mesurée de la dépense électrique et du matériel nécessaire.
IA Utile	Ciblage exclusif de cas d'usage à forte valeur ajoutée : analyse de documents, raisonnement logique, code source et traduction spécialisée.

1. Chiffres clés de l'étude

Ces indicateurs résument les enseignements mesurés au cours de nos 750 tests d'inférence et mesures électriques continues :

5 731 € HT

Investissement matériel total
Serveur bi-GPU 64 Go VRAM (2x RTX 5090) permettant d'exécuter des modèles jusqu'à 70 milliards de paramètres.

89 %

Taux de succès de GPT-OSS 20b
Précision constatée sur les tâches de réflexion, code, langue et documents (hors interrogation du web direct).

0,0018 kWh

Consommation par requête
Dépense mesurée pour GPT-OSS 20b, équivalant à 0,10 gCO2e par question traitée en 13 secondes.

< 60 ms

Temps de réaction en charge
Délai du premier token pour Mistral Small 24b, y compris face à 30 requêtes simultanées.

Ce que ces mesures nous apprennent

1. La compacité bat le gigantisme : Un modèle récent de 20 milliards de paramètres surpasse un modèle plus ancien de 70 milliards, tout en consommant deux fois moins d'énergie.

2. La mémoire VRAM conditionne tout : Dès lors que le modèle réside entièrement dans la mémoire vidéo des cartes graphiques, le processeur central reste au repos. En cas de débordement en mémoire vive classique, les performances s'effondrent.

3. La concision réduit la facture : Les modèles trop bavards gonflent le temps d'inférence et la consommation. Configurer un prompt système pour exiger des réponses courtes divise l'énergie consommée par deux.

2. Le banc d'essai matériel

Pour obtenir des données physiques rigoureuses, nous avons assemblé une machine dédiée et déployé une instrumentation continue de la puissance électrique appelée à la prise.

Composant	Détail technique	Prix unitaire HT
Cartes graphiques	2x NVIDIA RTX 5090 — 32 Go VRAM chacune (64 Go cumulés)	4 268,00 €
Processeur	Intel Core Ultra 9 285K — 24 cœurs (8 P-Cores + 16 E-Cores, 5,7 GHz)	604,16 €
Carte mère	ASUS Prime B860-PLUS WiFi — 4 bus PCIe 5.0	178,22 €
Mémoire vive	Kingston FURY Beast 64 Go (2x32 Go) DDR5 5200 MT/s CL40	150,00 €
Stockage	SSD NVMe P310 2 To PCIe Gen4 (7 100 Mo/s)	106,63 €
Alimentation	Corsair HX1500i modulaire 1500W certifiée Platinum	231,23 €
Ventilation & châssis	Noctua NF-F12 120 mm + châssis rackable et câblage PCIe	193,03 €
Total investissement	**Configuration serveur IA locale bi-GPU complète**	**5 731,27 €**

Système de mesure et supervision

Nous avons écarté les simples approximations logicielles pour mesurer la dépense électrique réelle :

• **Prise connectée AwoX & Bluetooth :** Mesure de la puissance instantanée toutes les 500 millisecondes directement au niveau de la prise murale 230V. Un agent logiciel interne transmet ces relevés vers notre serveur de métriques.

• **Télémétrie GPU NVIDIA-SMI :** Relevé de la puissance appelée par chaque processeur graphique et de l'occupation mémoire VRAM.

• **Scaphandre (Hubblo) :** Estimation de la charge processeur CPU en complément.

• **Moteur d'inférence :** Ollama sous Linux Ubuntu, interfacé avec OpenWebUI.

3. Protocole d'évaluation des réponses

Pour mesurer objectivement les aptitudes de chaque modèle, nous avons bâti un banc de test automatisé sous Python. Les questions sont injectées une à une dans l'API d'Ollama, en consignant pour chaque requête la durée d'exécution, la consommation électrique et le volume de tokens produits.

7 familles de tests métiers

Domaine	Objectif testé	Exemple concret
Langage	Synthèse, reformulation et clarté rédactionnelle	Résumer un rapport d'audit en 5 points
Raisonnement	Logique déductive et résolution de problèmes	Énigme d'inclusion-exclusion sportive
Traduction	Traduction technique français-anglais fidèle	Expressions idiomatiques de la presse économique
Mathématiques	Calcul arithmétique et respect des règles opératoires	Multiplication à 3 chiffres, fractions
Code informatique	Développement d'algorithmes et détection de bugs	Fonctions de traitement de données sous Python
Extraction PDF	Compréhension de documents longs et tableaux	Repérage d'un chiffre clé dans un rapport financier
Recherche Web	Synthèse d'actualités et faits récents	Dernières annonces institutionnelles

Arbitrage impartial et vérification humaine

Chaque réponse retournée par l'IA locale a été confrontée à un corrigé type rédigé par nos soins. Pour écarter tout biais d'interprétation :

1. Un modèle de référence distant note la concordance stricte entre la réponse locale et le corrigé attendu.

2. Un contrôle manuel exhaustif a ensuite été mené sur l'intégralité des 750 évaluations pour corriger d'éventuels faux positifs ou négatifs.

4. Résultats comparatifs des 8 modèles

Moyenne constatée sur 67 questions par modèle, exécutées dans des conditions identiques :

Modèle	Taille	Note	Vitesse	Volume	Temps	Énergie / q.	CO2e / q.
Qwen 3	30b (Q4)	**81 %**	164 t/s	3 777 t	32 s	0,0045 kWh	0,25 g
OpenAI gpt-oss	20b (MXFP4)	**79 %**	**208 t/s**	2 079 t	**13 s**	**0,0018 kWh**	**0,10 g**
Qwen 3	32b (Q4)	75 %	59 t/s	1 938 t	39 s	0,0087 kWh	0,49 g
Mistral Small 3.2	24b (Q4)	73 %	80 t/s	1 183 t	29 s	0,0065 kWh	0,36 g
Google Gemma 3	27b (Q4)	67 %	69 t/s	601 t	10 s	0,0021 kWh	0,12 g
Meta Llama 3.3	70b (Q4)	61 %	34 t/s	425 t	12 s	0,0028 kWh	0,15 g
Meta Llama 3.1	8b (Q4)	51 %	161 t/s	426 t	3 s	0,0006 kWh	0,03 g
Mistral Nemo	12b (Q4)	45 %	142 t/s	903 t	16 s	0,0035 kWh	0,20 g

Le cas particulier de GPT-OSS 20b

Ce modèle ressort comme le meilleur arbitrage global de notre banc d'essai :

• **Précision de 89 % hors recherche Web :** Les modèles locaux échouent tous sans moteur de recherche connecté. Sur les 6 autres catégories métiers, ce modèle frôle les 90 % de réussite.

• **Débit très élevé (208 tokens/s) :** Grâce à la quantification MXFP4 tirant parti des Tensor Cores de la RTX 5090.

• **Bilan énergétique sobre :** Seulement 0,0018 kWh par question, soit 2,5 fois moins que Qwen 3 30b et 4,5 fois moins que Qwen 3 32b.

5. Consommation électrique et bilan carbone

La facture électrique d'un serveur d'IA ne dépend pas uniquement de ses pointes d'activité : la consommation au repos constitue la base incontournable de son empreinte.

Régime d'utilisation	Puissance	Jour	Mois	Année	Coût estimé
Serveur au repos (2x RTX 5090 sans requête)	85,8 W	2,1 kWh	62,6 kWh	751 kWh	151 €/an (42 kgCO2e)
Activité IA (10h/jour) (gpt-oss 20b en charge)	300 à 420 W	5,0 kWh	152,0 kWh	1 825 kWh	368 €/an (102 kgCO2e)

Mise en perspective avec les études Boavizta et ADEME

L'étude Boavizta (200 serveurs physiques observés pendant 6 mois) : Un serveur d'hébergement classique consomme 10 à 60W au repos, et 80 à 200W en charge maximale. Notre machine se situe dans la moyenne haute au repos (85,8W avec deux cartes graphiques), mais grimpe à 300-420W en charge d'inférence, soit **2 fois plus qu'un serveur web traditionnel**.

La base Empreinte ADEME / NegaOctet : Un serveur physique complet génère en moyenne 3 215 kWh/an en prenant en compte son cycle de vie complet (fabrication, transport, fin de vie). L'énergie d'usage direct de notre serveur (1 825 kWh/an) représente donc une part conséquente de l'impact global.

Comparatif avec les modèles cloud géants : Selon les publications officielles, une requête vers Mistral Large 2 émet environ 1,14 gCO2e, tandis que notre requête sur GPT-OSS 20b en local ne rejette que **0,10 gCO2e** (11 fois moins).

6. Tests de montée en charge et simultanéité

Un serveur de production doit encaisser les sollicitations de plusieurs collaborateurs simultanés. Nous avons mesuré le temps avant l'arrivée du premier octet (TTFT) en faisant varier le nombre d'utilisateurs simultanés (1, 5, 10, 30 VUs) sur une fenêtre de 180 secondes.

Modèle testé	Threads Ollama	1 utilisateur	5 utilisateurs	10 utilisateurs	30 utilisateurs	Verdict
Mistral Small 24b	3 requêtes //	0,041 s	0,049 s	0,058 s	0,059 s	**Excellente tenue**
Mistral Small 24b	6 requêtes //	0,041 s	0,056 s	0,057 s	0,065 s	**Parfaitement stable**
Qwen 3 30b	3 requêtes //	4,0 s	20,0 s	26,0 s	32,0 s	Dégradation marquée
Qwen 3 30b	6 requêtes //	8,0 s	20,0 s	23,0 s	28,0 s	Latence excessive
Google Gemma 3 27b	3 requêtes //	5,0 s	11,0 s	18,0 s	32,0 s	Plafond atteint
Google Gemma 3 27b	6 requêtes //	16,0 s	22,0 s	30,0 s	34,0 s	Surcharge serveur

Enseignements sur le parallélisme

• **Ollama par défaut :** Le moteur ne traite qu'une seule requête à la fois s'il n'est pas paramétré. Les demandes s'empilent dans une file d'attente, provoquant des erreurs de délai dépassé (504 Gateway Timeout).

• **Mistral Small 24b est taillé pour la charge :** C'est le seul modèle qui ne subit aucun ralentissement perceptible même avec 30 utilisateurs simultanés, avec une latence restant sous les 65 millisecondes.

• **Le goulot d'étranglement :** Au-delà de 10 utilisateurs simultanés sur les modèles denses (Qwen, Gemma), la machine sature et les temps de réponse dépassent 30 secondes, ce qui impose une mise à l'échelle horizontale (plusieurs serveurs) pour ce type de modèles.

7. Benchmarks publics face aux mesures locales

Les classements en ligne (Chatbot Arena, MMLU, GSM8K, AGIEval) orientent souvent les choix techniques. Pourtant, s'appuyer uniquement sur ces données publiques induit en erreur.

Critère	Benchmarks en ligne (HuggingFace, Arena)	Notre banc d'essai local
Précision de calcul	Modèles testés en pleine précision (FP16 ou FP32) sur des serveurs industriels (clusters H100).	Modèles quantifiés (Q4_K_M, MXFP4) adaptés au matériel réellement accessible en entreprise.
Données d'évaluation	QCM académiques en anglais, souvent mémorisés lors de l'entraînement préliminaire du modèle.	Cas concrets en français : analyse de contrats, code sur mesure, logique métier et raisonnement.
Mesure d'énergie	Absent ou théorique. Ne prend pas en compte le matériel client ni les phases de repos.	Mesure physique au wattmètre (500 ms) incluant GPU, CPU et repos machine.
Comportement en charge	Requêtes unitaires isolées sans prise en compte des files d'attente d'un serveur d'équipe.	Épreuve de montée en charge jusqu'à 30 utilisateurs simultanés sous contrainte de VRAM.

Pourquoi tester sur votre propre configuration ?

Tester en conditions réelles valide directement la précision de la quantification sur votre vocabulaire métier, et vérifie la vitesse de génération permise par vos bus PCIe et votre mémoire VRAM.

C'est la méthode directe pour dimensionner vos achats matériels sans surinvestir dans des capacités inutiles.

8. Recommandations concrètes pour l'entreprise

Sur la base de nos mesures physiques et des retours d'expérience du banc de test, voici la feuille de route préconisée pour intégrer une IA locale avec succès :

1. Privilégier les modèles de 20 à 24 milliards de paramètres	Ils apportent 80 à 90 % de la précision des très grands modèles tout en divisant la consommation électrique par trois et en restant réactifs sous charge.
2. Dimensionner la VRAM avec marge	Le modèle et sa fenêtre de contexte doivent tenir à 100 % dans la mémoire des cartes graphiques. Tout débordement sur la RAM de la carte mère détruit le débit de traitement.
3. Imposer la concision via le prompt système	Chaque token généré consomme de l'énergie. Demander des réponses synthétiques et structurées réduit immédiatement l'empreinte carbone et divise le temps d'attente.
4. Ajuster le parallélisme Ollama pour la production	Fixer la variable d'environnement `OLLAMA_NUM_PARALLEL` entre 4 et 8 selon les modèles pour éviter les blocages de requêtes dès que l'équipe grandit.
5. Découpler la recherche d'actualités	Les modèles locaux excellent dans la synthèse et le code, mais échouent sur l'actualité brute. Pour ces besoins, coupler le LLM à un module de recherche documentaire interne (RAG).

Conclusion générale

Cette étude prouve qu'une IA locale souveraine, frugale et performante est aujourd'hui une réalité technique accessible pour les entreprises.

Avec un investissement matériel maîtrisé (moins de 6 000 € HT) et un coût d'électricité annuel inférieur à 400 €, une organisation gagne en indépendance, protège ses données confidentielles et réduit considérablement son empreinte environnementale comparé aux géants du cloud.

Besoin d'échanger sur cette étude ou d'étudier votre configuration locale ?
Contact : Laurent Naigeon — laurent@wysistat.com — IDfr / Wysistat