Se connecter

Proposer un contenu

Identifiant
Identifiant

Mot de passe
Mot de passe

☐ Connexion automatique

Se connecter

Pas de compte ? S'inscrire…

18 juil. 2026

3

Lien : [Clubic] Six semaines sans cloud : j'ai confié tout mon travail à une IA locale à 4 500 €

Posté par **Jona** le 18 juillet 2026 à 15:22.
Étiquettes : **intelligence_artificielle**, **clubic**, **local-first**

https://www.clubic.com/dossier-619549-six-semaines-sans-cloud-j-ai-confie-tout-mon-travail-a-une-ia-locale-a-4-500.html

(13 commentaires).

Excellent article

Posté par **orfenor** le 18 juillet 2026 à 20:44. Évalué à 3 (+3/-2).

Excellent article long et détaillé, où on apprend qu'une IA installée sur votre poste perso suffit pour 80% des gens et qu'une IA installée sur du matériel dédié (le Spark NVidia) est rentable, consomme autant qu'une ampoule et rend d'immenses services sans avoir besoin des grosses IA du Cloud.

Répondre

[^] # Re: Excellent article

Posté par **Pol' uX** (**site web personnel**) le 18 juillet 2026 à 22:06. Évalué à 10 (+11/-1).

Une ampoule moderne c'est 9W et pas 150, cependant.

--

Adhérer à l'April, ça vous tente ?

Répondre

[^] # Re: Excellent article

Posté par **gUI** (**Mastodon**) le 19 juillet 2026 à 08:16. Évalué à 6 (+3/-0).

> une IA installée sur votre poste perso suffit pour 80% des gens

Et si c'est pour demander la capitale de la Mongolie ou la recette de la pâte à crêpes on n'a même pas besoin d'IA !

Vraiment ? Il y a des gens qui demandent ça à une IA ?

--

En théorie, la théorie et la pratique c'est pareil. En pratique c'est pas vrai.

Répondre

[^] # Re: Excellent article

Posté par **barmic** 🦦 le 19 juillet 2026 à 09:27. Évalué à 1 (+0/-1).

Oui et c'est pas si bête au vu de comment l'IA est décrite pour les non initiés. 2 des points qui il me semble ont pas mal infusé dans la population c'est le fait qu'elle soit entraîné sur des données et qu'on peut commencer un prompt avec "tu es un spécialiste de gnagnagna". Ça ne paraît pas illogique de lui demander ce genre de choses. Évidemment ça ne marche pas si bien

--

https://linuxfr.org/users/barmic/journaux/y-en-a-marre-de-ce-gros-troll

Répondre

[^] # Re: Excellent article

Posté par **orfenor** le 19 juillet 2026 à 12:23. Évalué à 4 (+2/-0).

Ma copine le fait tout le temps, sur des trucs encore plus basiques. Beaucoup de personnes autour de moi font pareil. Les modèles locaux sont bons là-dessus.

Répondre

[^] # Re: Excellent article

Posté par **wilk** le 19 juillet 2026 à 14:10. Évalué à 6 (+5/-1).

Ils sont d'autant plus bons qu'ils deviennent la référence de la vérité. Tu discutes à table et quelqu'un va sortir : mais si tu vois j'ai raison je viens de vérifier sur jaipaitai.

Répondre

publi information ?

Posté par **devnewton** 🍺 (**site web personnel**) le 19 juillet 2026 à 07:22. Évalué à 5 (+2/-0).

Il y a des liens sponsorisés…

--

Ce post est offensant ? Prévenez moi sur https://linuxfr.org/board

Répondre

[^] # Re: publi information ?

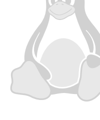
Posté par **Benoît Sibaud** (**site web personnel**) le 19 juillet 2026 à 08:59. Évalué à 9 (+6/-0). Dernière modification le 19 juillet 2026 à 09:02.

Vers la fin : *Le Spark est reparti chez NVIDIA,* il y aurait donc eu prêt de la machine à 4500€ pour le test. Ça aurait mérité d'être mentionné au début (surtout quand on commence à *jouer cartes sur table* mais l'article n'est pas dithyrambique sur la machine non plus).

Répondre

GTT

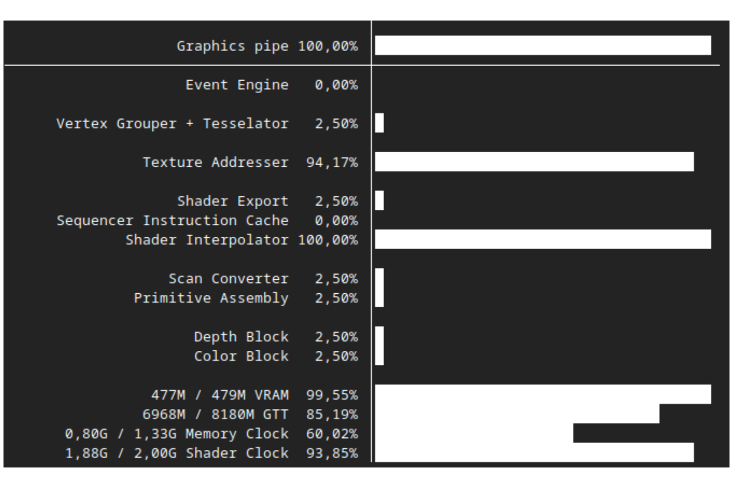
Posté par **bubar**🦥 le 19 juillet 2026 à 14:28. Évalué à 3 (+0/-0). Dernière modification le 19 juillet 2026 à 14:31.



Et sinon n'oubliez pas que Linux gère remarquablement bien le "débordement GTT" ;-)

Certes la ram n'est pas la vram et certes les bus vers la ram n'est pas le bus vers la dram.
Certes, gnian gnian blabla, gnian gnian ok ok …

Sur une Radeon ayant 512 faméliques Mega de Vram, se configurer un GTT sur 8G, c'est avoir **16** fois plus de ram dispo pour le GPU.



```
                Graphics pipe 100,00%
                 Event Engine   0,00%
   Vertex Grouper + Tesselator   2,50%
             Texture Addresser  94,17%
                 Shader Export   2,50%
   Sequencer Instruction Cache   0,00%
          Shader Interpolator 100,00%
                Scan Converter   2,50%
            Primitive Assembly   2,50%
                   Depth Block   2,50%
                   Color Block   2,50%
             477M / 479M VRAM  99,55%
            6968M / 8180M GTT  85,19%
    0,80G / 1,33G Memory Clock  60,02%
    1,88G / 2,00G Shader Clock  93,85%
```

Et làvoilà : gemma-4-e4b envoie du 30tokens/s …

Vous pouvez moinsser.

Répondre

[^] # Re: GTT

Posté par **bubar**🦥 le 19 juillet 2026 à 15:10. Évalué à 4 (+1/-0). Dernière modification le 19 juillet 2026 à 15:15.

Quelques détails :

- ce laptop est un pc d'entrée de gamme (de 450 à 500€ selon les vendeurs) : Un Acer Aspire GO15 AG15-42P-R38D.
    - Ryzen 7 5825U
    - 16G RAM
    - Un écran moins que moyen, il faut jouer avec l'inclinaison pour voir correctement
    - Un ssd lent (pas emmc, un ssd remplaçable)

C'est un "ordinateur bureautique qui ne convient pas pour le jeu vidéo", selon l'appellation commerciale. Et clairement, en configuration par défaut, un avance rapide sur une lecture vidéo en fullhd et au dessus : ça saccade, il n'y arrive pas, l'usage est plus que désagréable : il faut oublier l'avance rapide vidéo (et toutes velléités d'édition vidéo), ce n'est pas utilisable.

La faute à une CG Radeon amputée : une Bracelo rc1 Renoir n'ayant que 512m de vram ..

Pourtant … ces points faibles sont compensés par son point fort : 16G de ram qui épaulent un Ryzen 7. Miam miam.

Alors, pour éviter au lectorat de se perdre dans les méandres d'un web où les documents sont si nombreux qu'il est difficile de trouver le convenable dans cette botte de f_oin_, l'option correcte de boot est : `amdgpu.gttsize=8192`

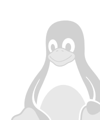
C'est tout.
16 fois plus de ram pour le gpu. This is Linux.
Non seulement l'avance rapide vidéo fonctionne normalement, mais vos poils sont plus soyeux, votre bureau plus fluide … et surtout vous pouvez jouer avec une IA locale, gemma-4-e4b envoie vraiment du 30/s et saura répondre à haute voix (via kokoro) à la phrase prononcez aussi à haute-voix (via whisper c++) «fait un bilan résumé de la situation des nappes phréatiques en Occitanie et un focus sur le débit de Aire Sur l'Adour»
Il faudra patienter 4mn pour avoir la réponse, c'est là où il n'y a pas de magie, bien sûr. Mais vous venez de passer d'un pc entrée de gamme pas foutu de lire du 4K ou d'avancer rapide sur du fHD, à un pc capable d'exécuter des modèles IA correctement configurés.

Enjoy.

Répondre

[^] # Re: GTT

Posté par **bubar**🦥 le 19 juillet 2026 à 15:24. Évalué à 3 (+0/-0).



Et pour l'autre lectorat, il s'agit en fait d'un bug dans le calcul par défaut, pourtant simple, de la taille du débordement. En attendant sa correction, ouplutôt son amélioration pour avoir un tableau plutôt q'un seul calcul, fixer l'allocation au boot fait des merveilles. Mais c'est une longue histoire, ça *(et de toutes façons sur ce type de config il est plus convenable de la fixer pour balancer plein de ram au gpu*)

Répondre

[^] # Re: GTT

Posté par **fearan** le 19 juillet 2026 à 15:46. Évalué à 3 (+0/-0).

en fait je suis intéressée par ton setup; quel moteur pour faire tourner le llm ? (et quel utilitaire pour voir les infos que tu as donné)

--

Il ne faut pas décorner les boeufs avant d'avoir semé le vent

Répondre

[^] # Re: GTT

Posté par **bubar** le 19 juillet 2026 à 17:00.
Évalué à 3 (+0/-0). Dernière modification le 19 juillet 2026 à 17:04.

> Dans les grandes lignes :
> Le sucre autour, c'est Newell qui s'en charge. Et bien qu'il apporte un *overhead* si on compare à LMstudio, ce dernier ~~pue fort~~ n'est pas libre et d'autre part je suis un fervent supporter de Newell du début et tjs à ce jour :-)
>
> Pour la conf :
>
> - llama-server v9748 (livré par l'équipe Newell)
>   - reasoning-format désactivé
>   - reasoning-budget désactivé
>   - semantic-memory activée
> - gemma-4-e4b.gguf
>   - température à 0
>   - Pénalité de fréquence à 2
>   - Native tools calling
>   - désactivation du TextStreamer pour éviter le "think leak" sans désactiver le "think"
>   - Mmproj activé pour la reconnaissance d'image
>   - accélération matérielle activée
> - RAG actif / accès à certains dossiers de mes documents
>
> Reconnaissance vocale :
>
> - Whisper C++ (ggml-small-bin)
>   - model small
>   - VAD activé (silence à 1 / speech-duration à 5)
>   - accélération matérielle activée
>
> Synthèse vocale :
>
> - Kokoro (82m / voice v1)
>
> Utilitaire d'affichage gpu :
>
> - radeontool
>
> En espérant avoir répondu à tes questions :)

Répondre

Envoyer un commentaire

Suivre le flux des commentaires

Note : les commentaires appartiennent à celles et ceux qui les ont postés. Nous n'en sommes pas responsables.

Revenir en haut de page

Derniers commentaires

- **déjà vu**
- **Re: Instruction à do…**
- **Re: Instruction à do…**
- **Re: Instruction à do…**
- **Re: Instruction à do…**
- **Re: Instruction à do…**
- **Re: Instruction à do…**
- **Re: GTT**
- **Re: Anecdotique, au…**
- **Re: Instruction à do…**
- **Densité des tables S…**
- **Re: Anecdotique, au…**

Étiquettes (tags) populaires

- **intelligence_artificielle**
- **le_monde_diplomati…**
- **merdification**
- **grands_modèles_de…**
- **états-unis**
- **canicule**
- **capitalisme**
- **datacenter**
- **réchauffement_clim…**
- **logiciel_libre**
- **france**
- **administration_fran…**

Sites amis

- **Agenda du Libre**
- **April**
- **Éditions D-BookeR**
- **Éditions Diamond**
- **Éditions Eyrolles**
- **Éditions ENI**
- **En Vente Libre**
- **Framasoft**
- **La Quadrature du Net**
- **Lea-Linux**
- **Open Source Initiative**
- **Imprimerie Grafik Plus**

À propos de LinuxFr.org

- **Mentions légales**
- **Faire un don**
- **L'équipe de LinuxFr.…**
- **Informations sur le s…**
- **Aide / Foire aux que…**
- **Suivi des suggestion…**
- **Wiki du site**
- **Règles de modération**
- **Statistiques**
- **API pour le dévelop…**
- **Code source du site**
- **Plan du site**