Journal : TurboQuant : enfin des LLM puissants sur votre propre machine

31 mar. 2026

6

Posté par **pas_pey** le 31 mars 2026 à 21:50. **Licence CC By-SA.**
Étiquettes : **ia_locale**, **google**

Faire tourner un grand modèle de langage chez soi, c'est souvent frustrant : la mémoire GPU s'épuise rapidement, les conversations longues ralentissent ou plantent, et les modèles vraiment capables restent réservés au cloud.

TurboQuant, publié par Google Research le 24 mars 2026, change la donne.

L'algorithme compresse la mémoire de travail des LLM par **6 fois** sans perte de précision, et accélère les calculs jusqu'à **8 fois** sur GPU haut de gamme. Concrètement, pour un utilisateur local : sur une carte graphique grand public à 12 Go, on passe de 8 000 à **40 000 tokens** de contexte utilisable. C'est la différence entre un assistant qui oublie le début de la conversation et un qui tient sur un fichier de code entier - ou un long document - sans broncher.

Des utilisateurs rapportent pouvoir tenir des conversations de **100 000 tokens** sur du matériel grand public comme un Mac Mini, sans la dégradation de qualité habituelle. Les modèles qui nécessitaient hier un abonnement cloud commencent à tourner correctement en local.

Google n'a publié aucun code, mais des développeurs indépendants ont implémenté l'algorithme à partir des seules équations du papier. En moins d'une semaine, une intégration dans **llama.cpp** était disponible. Le débit en tokens se maintient **2 à 3 fois plus élevé** dans les régimes où le KV cache saturait auparavant la mémoire GPU.

L'implémentation officielle de Google est attendue pour le Q2 2026. En attendant, les forks communautaires sont déjà utilisables.

Liens

Blog Google Research
Discussion llama.cpp #20969
turboquant_plus
Article Venturebeat

(0 commentaire). Markdown EPUB

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

- Re: Nintendo
- Re: ben faut tout lire
- code "source"
- Re: site web tout po…
- Utilisation créative …
- Re: Bizarre
- Lire ?
- Re: Selon les règles i…
- Re: La tarte au citro…
- Re: site web tout po…
- Re: Anglicismes
- Dans l'immédiat:

Étiquettes (tags) populaires

- intelligence_artificielle
- lionel_jospin
- merdification
- grands_modèles_de…
- états-unis
- sortie_version
- souveraineté_numer…
- capitalisme_de_surv…
- bigtech
- administration_fran…
- claude
- capitalisme

Sites amis

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

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