Implémenter l’IA dans le navigateur en respectant la vie privée, facile ?

BDX I/O
Bordeaux - Novembre 2024

https://www.nicolas-hoffmann.net/BDX-IO-2024

« Mission Difficile serait un jeu d’enfant pour vous M. Hunt, vous êtes dans Mission Impossible. »

Proton | Privacy by default

Présentation : votre serviteur

- Nico(-las Hoffmann)
- UX-Engineer (un FE sous CSStéroïdes)
- Aime aussi l'A11Y, la L10N/I18N, Opquast, etc.
- 6 ans de Proton, et 20 années de métier

  #teamDéambulateur

Présentation : Proton AG

- Des contraintes propres à Proton (E2E, sécurité, etc.)
- Et des contraintes « normales » (bande passante, etc.)

⚠️ Disclaimer ⚠️

Je ne suis pas DU TOUT expert LLM.

Ni chatGPT fanboy.

Une image de l'IA

Demande business sur l'IA

Réelle demande des clients (productivité).

Le challenge :

Comment concilier vie privée/sécurité avec IA ?

Première étape : laisser le choix à l'utilisateur

- de s'en servir… ou pas
- sur ses données personnelles pour l'entrainement

LinkedIn Is Quietly Training AI on Your Data—Here's How to Stop It

Seconde étape, les LLMs

Boites noires ?

- Ouverture du modèle
- Les données d’entrainement
- Et les requêtes utilisateur, pour l’entrainement aussi ?

Seconde étape, les LLMs

Questions éthiques et business

- Chiffrement E2E, et pas pour le plaisir
- Demande business sur IA générative

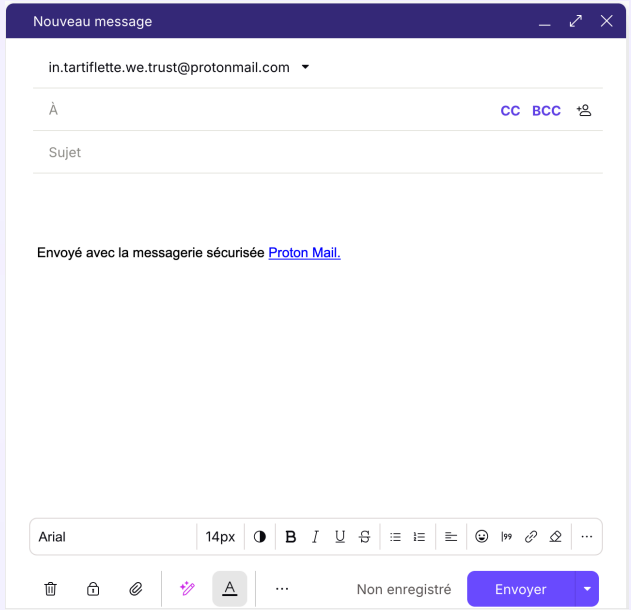
Aller sur une version locale :

→ une **version sans compromis !**

Faire tourner un LLM en local

L’idée : dans le **Composer**.

⚠️ élément **critique** pour une application mail !



Pour faire tourner un LLM en local

- Une lib pour le LLM : web-llm
- Une lib maison qui sert « d'intermédiaire »
- L'utilisateur qui envoie les demandes

⚠️ Alerte peaux de bananes ⚠️

Personne n’avait tenté cela (à notre échelle et en prod) !

Il a fallu quasiment **TOUT** découvrir.

Spoiler : on a pris cher.

La découverte du POC

— On a un POC qui tourne en local ! 🎉

— 🧐 Ah ouais ?

— Il faut télécharger 4Go.

— 🤨 Pardon ?

— Et avoir WebGPU sinon ça rame. Chromium-only au fait.

— WHAT ? 😤😤

— Et faut de la RAM, etc.

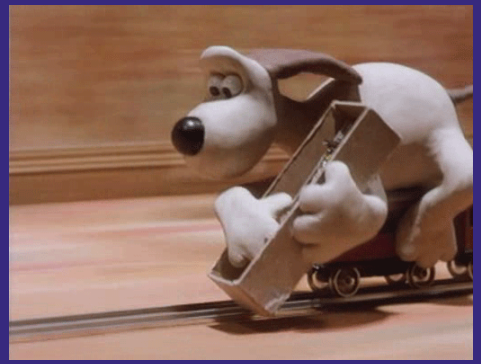
— 🤬🤬 Ta GU**********

⚠️ Idée de l’état d’esprit des devs ⚠️

Mais c’est quoi ce **délire** ?

⚠️ Second point important ⚠️

Sommet de l’iceberg 😅



4 Go pour Mistral 7B

- Bande passante côté client ET infra
- Téléchargement « pausable »
- Éviter de le refaire… voire même de le faire ?

WebGPU

- Le LLM, ça demande BCP de ressources/puissance
- Impact direct sur le design des étapes

Chromium-only ?

Pas exactement.

C’est une histoire de **hardware**,

d’**accélération graphique**

et de **navigateurs**.

#çaiCompliquaÿ

Côté hardware

- Mac : un solide M1/M2/équivalent → ✅
- Windows : ça dépend → accélération graphique
- Linux → bazar.

Accélération graphique ?

- GPU intégré sur CPU ou *discrete* GPU ?
- Sur le CPU : 0,5w en veille, *discrete* 9w !
- Pas vraiment de moyen de choisir ¯_(ツ)_/¯

Côté navigateurs

- Chromium → ✅
- Firefox → Flags + sombre histoire de *Shader-F16*
- Safari → Flags aussi.

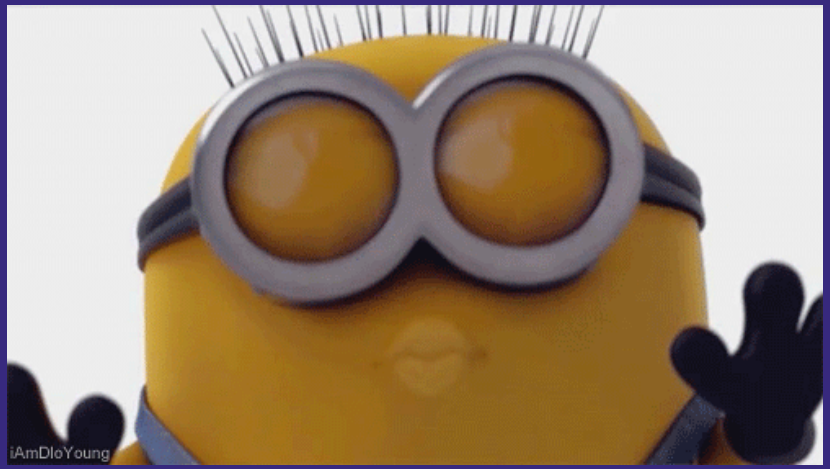
Shader-F16 VS Shader-F32 ?

- LLM « quantisé » en 16 bits pour les perfs
- 30% plus rapide en F16 qu’en F32

Adapter : demande support F16.

❤️ BRAVO ! 🫶

C’était le passage le plus relou.



⚠️ C’est pas fini ! ⚠️

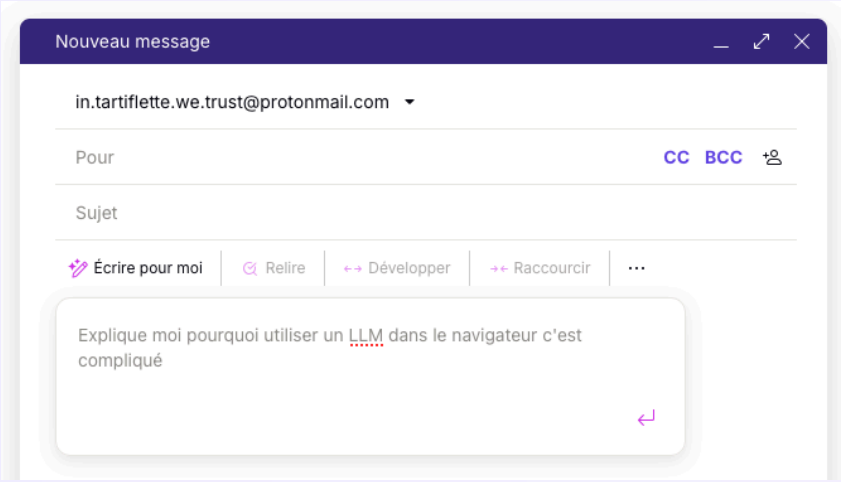
On a encore des trucs à régler !

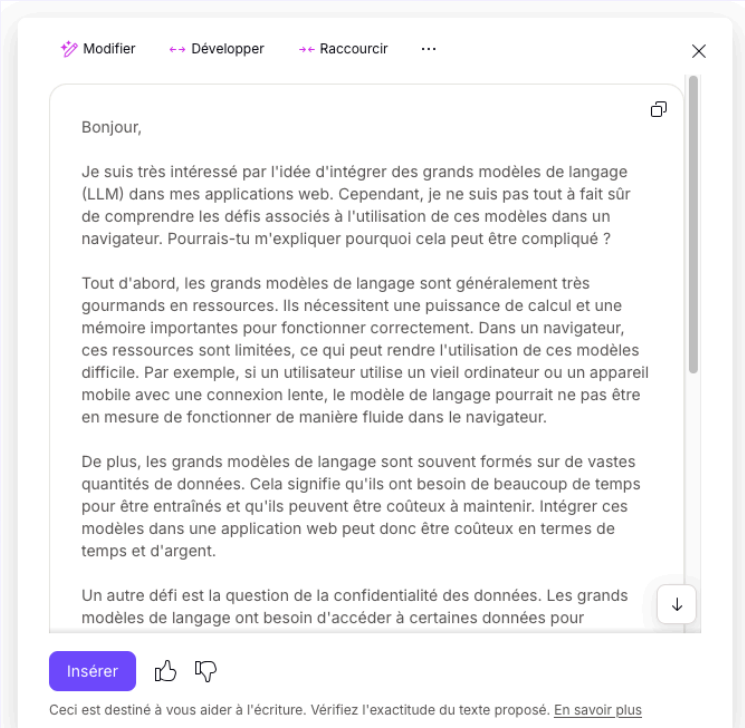
D’autres soucis/questionnements

- CSP et WebAssembly (utilisé par Web-LLM)
- Iframe pour sandboxer/mutualiser le téléchargement

Choix pour le Composer

- Écrire ET générer ?





La version serveur

La version serveur

- Envoyer des requêtes au serveur, qui répond
- Serveur = environnement maîtrisé
- Modèle plus puissant ? (Mistral Nemo 12B model)

Comment garantir la vie privée ?

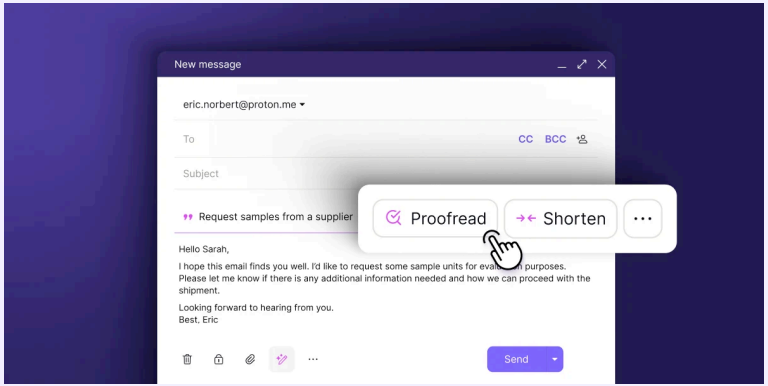
- No-logs (c’est écrit dans la privacy policy)
- PAS d’entrainement avec les données utilisateur.

Quelques couacs

- Stream API, génération… JSON pas complet ? 😅
- Côté Front-End → presque une balade de santé car + simple

Toilettage/redesign et derniers ajouts

- Commencer par le + dur → code spaghetti
- Redesign/refacto rapides
- Actions de *refine* ajoutée en cours (*proofread*/etc.)



Conclusion

Un **LLM en local,** de l'**IA générative respectueuse de la vie privée** en production, possible ?

C'est faisable ! On l'a fait.

Mais c'est pas facile.

Cependant…

Vers l'infini et au-delà

- Des modèles + petits/rapides/spécialisés
- Chiffrement homomorphe
- Groupes de travail au W3c pour des LLMs en local

Ressources

- How to build privacy-protecting AI, Introducing Proton Scribe
- Tracking openness of instruction-tuned LLMs
- Encrypted LLM
- Web-LLM, API WebGPU, Nouveautés de WebGPU (Chrome 120)
- Content security policy (PW), CSP (Smashing Magazine)
- Slides

Merci pour votre attention ❤️

Des questions ?

Merci pour votre attention ❤️

Des questions ?

https://www.nicolas-hoffmann.net/BDX-IO-2024

Proton | Privacy by default