28/04/2025

Llamafile – Exécutez des modèles de langage en un seul fichier !

Llamafile – Exécutez des modèles de langage en un seul fichier !

llamafile est un projet complètement barré qui va vous permettre de transformer des modèles de langage en exécutables. Derrière se cache en fait la fusion de deux projets bien badass : llama.cpp, un framework open source de chatbot IA, et Cosmopolitan Libc, une libc portable pour compiler des programmes C multiplateformes. En combinant astucieusement ces deux technos, les petits gars de Mozilla ont réussi à pondre un outil qui transforme les poids de modèles de langage naturel en binaires exécutables. Imaginez un peu, vous avez un modèle de langage qui pèse dans les 4 gigas, dans un format .gguf (un format couramment utilisé pour les poids de LLM). Et bien avec llamafile, vous pouvez le transformer en un exécutable standalone qui fonctionnera directement sur le système sur lequel il est sans avoir besoin d'installer quoi que ce soit. Ça va permettre de démocratiser l'utilisation et la diffusion des LLM. Et niveau portabilité, c'est le feu puisque ça tourne sur six OS, de Windows à FreeBSD en passant par macOS. Les devs ont bien bossé pour que ça passe partout, en résolvant des trucs bien crados comme le support des GPU et de dlopen() dans Cosmopolitan et croyez-moi (enfin, croyez-les) ça n'a pas été une mince affaire ! Niveau perf aussi c'est du brutal ! Sur Linux llamafile utilise pledge() et SECCOMP pour sandboxer le bousin et empêcher les accès fichiers non désirés et avec les derniers patchs de Justine Tunney, la perf CPU pour l'inférence en local a pris un boost de malade du genre 10 fois plus rapide qu'avant. Même sur un Raspberry Pi on peut faire tourner des petits modèles à une vitesse honnête. Allez, assez parlé, passons à la pratique ! Voici comment tester vous-même un llamafile en un rien de temps :

Téléchargez l'exemple de llamafile pour le modèle LLaVA (licence : LLaMA 2, OpenAI) : llava-v1.5-7b-q4.llamafile (3,97 Go). LLaVA est un nouveau LLM qui peut non seulement discuter, mais aussi analyser des images que vous uploadez. Avec llamafile, tout se passe en local, vos données ne quittent jamais votre PC. Ouvrez le terminal de votre ordinateur. Si vous êtes sous macOS, Linux ou BSD, vous devrez autoriser l'exécution de ce nouveau fichier. (À faire une seule fois) : chmod +x llava-v1.5-7b-q4.llamafile Sous Windows, renommez simplement le fichier en ajoutant « .exe » à la fin. Lancez le llamafile, par exemple : ./llava-v1.5-7b-q4.llamafile Votre navigateur devrait s'ouvrir automatiquement sur une interface de chat. (Sinon, ouvrez-le et allez sur http://localhost:8080) Quand vous avez fini, retournez dans le terminal et faites Ctrl-C pour arrêter llamafile. Évidemment, Mozilla ne compte pas s'arrêter là et continue de bosser comme des dingues pour suivre le rythme des nouveaux modèles qui sortent et avec le support des dernières architectures dès leur sortie. Il est même prévu qu'on puisse bientôt générer nos propres llamafiles en un seul clic ! D'ailleurs, Hugging Face est déjà dans la boucle pour héberger tout ce petit monde. Bref, je vous le dis, les amis, llamafile est un projet à suivre absolument ! Alors on dit merci qui ? Merci Mozilla !
https://korben.info/llamafile-executez-modeles-langage-fichier.html