Du 10/03 au 11/04 **Guillaume Emorine**

UNIVERSITÉ BOURGOGNE EUROPE

Guide à destination des utilisateurs pour l'utilisation de LLMs en local avec le logiciel LM Studio

Présentation de LM Studio

LM Studio est une application disponible pour macOS, Linux, et Windows, qui permet d'exécuter des LLMs (Large Language Models) – donc des IAs – en local, sur son propre ordinateur, sans fournir de données à une compagnie tierce et sans avoir besoin d'Internet. Évidemment, l'Internet est requis pour télécharger les IAs, pour obtenir les composants requis à leur exécution, ou même pour mettre à jour l'application.

Important - Configuration requise

LM Studio est disponible sur les ordinateurs **Mac** avec les puces Apple Silicon, allant de la puce M1 à la puce M4. De plus, ils recommandent **16Go** de mémoire vive, même si 8Go peuvent suffire pour des modèles plus légers avec peu de contexte fourni.
Pour **Windows** et **Linux**, ils recommandent aussi **16Go** de mémoire vive, et un processeur avec l'instruction **AVX2** est **nécessaire**, qui est disponible sur tout processeur **Intel** sorti après **2014** excepté les Pentium et Celeron (**2020**) et sur tout processeur **AMD** sorti après **2016**. Sur Windows, on peut aussi utiliser un processeur ARM **Snapdragon X Elite**. En somme, tout ordinateur ayant 8 ans ou moins peut faire tourner LM Studio s'il répond à cette configuration requise.

Au niveau des **versions** du système, LM Studio nécessite au moins **Windows 10**, **Ubuntu 20.04** ou plus récent pour les systèmes Linux, et **macOS 13.4 (Ventura)** ou plus récent.

Installation de LM Studio

Pour installer LM Studio, il suffit de se rendre sur cette page afin de télécharger LM Studio pour son système. Il faut noter qu'il existe une traduction en français en béta, qu'on peut activer dans la section Changer le langage de l'interface.

Pour Windows :

Une fois le .exe récupéré, on peut ensuite installer LM Studio pour l'utilisateur connecté (sans droits d'administrations requis) ou pour tous les utilisateurs du poste (nécessite des droits d'administration).

Une fois l'installation terminée, on peut lancer l'application directement en fermant le programme d'installation, mais ce n'est pas obligatoire.

Pour Linux :

Pour Linux, il existe deux options : soit la distribution est basée sur Debian, et on peut récupérer le .deb ; ou alors la distribution n'est **pas** basée sur Debian, et il nous faut le AppImage. Dans les deux cas, **il faut les droits d'administration**.

Pour les distributions basées sur Debian (comme Ubuntu, Linux Mint…), il suffit de lancer la commande suivante pour installer le .deb de LM Studio :

```
sudo apt install ./<nom du fichier>.deb
```

Il faut noter que cela fonctionne si on se trouve dans le même répertoire que le fichier.

Pour les autres distributions Linux (donc Arch, Fedora, Red Hat…), on peut installer le .AppImage en lançant la commande suivante pour le rendre exécutable :

```
sudo chmod a+x <nom du paquet LM Studio>.AppImage
```

Ensuite, on peut l'exécuter depuis la ligne de commande en lançant :

```
./<nom du paquet LM Studio>.AppImage
```

Dans les deux cas, il faut être dans le même répertoire que le fichier, et il faut savoir utiliser le terminal.

Pour macOS :

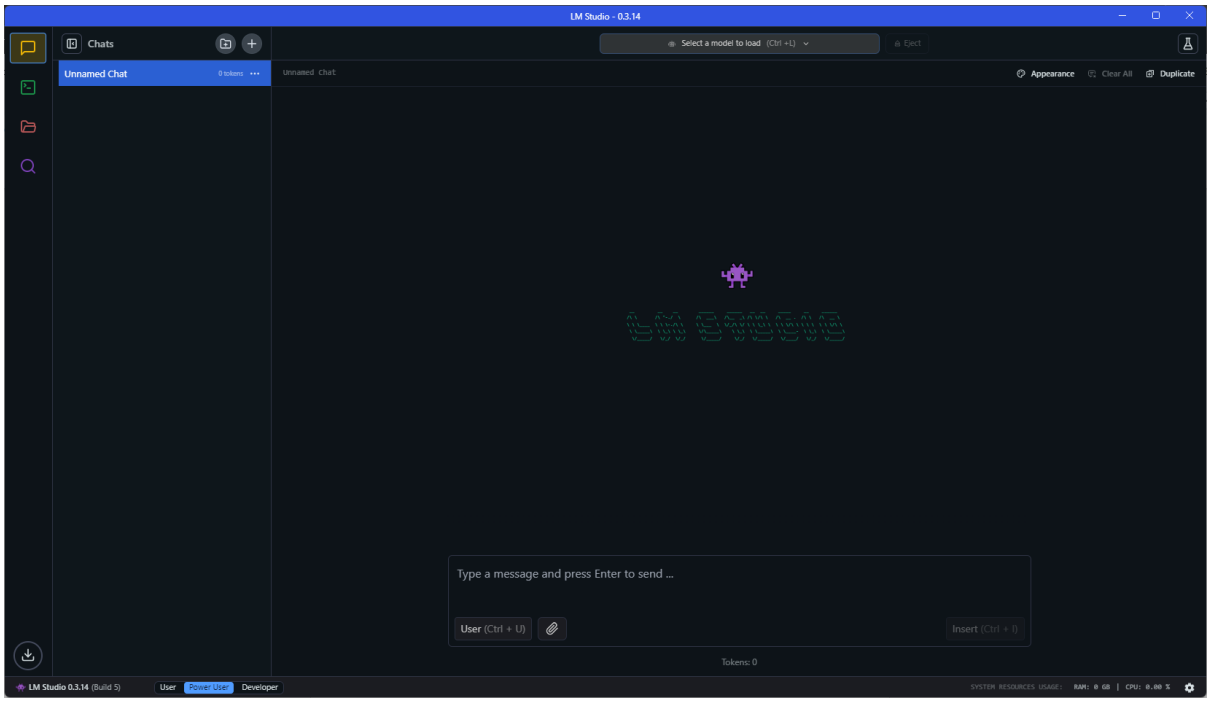
On récupère directement le .dmg depuis le site de LM Studio, et on double-clique sur celui-ci afin de le monter. Si nécessaire, les droits d'administration seront demandés. Ensuite, on ouvre le nouveau lecteur monté sur le bureau, et on glisse-dépose le dossier de l'application LM Studio dans le dossier "Applications". Une fois les fichiers copiés, on peut démonter le .dmg en faisant un clic droit et en cherchant l'option, puis on peut supprimer le .dmg.

Utilisation de LM Studio et interface

Menu Discussion

Quand l'application se lance pour la première fois, elle va nous suggérer de récupérer notre premier LLM, qu'elle va déterminer en fonction de la configuration de notre machine. Par exemple, un vieux All-In-One Dell n'aura pas forcément le même LLM suggéré qu'un tout nouveau MacBook.

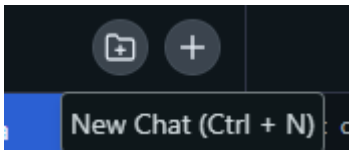
Une fois le LLM téléchargé, LM Studio va nous suggérer de commencer notre première conversation, et on se retrouve donc avec cette interface présentée :



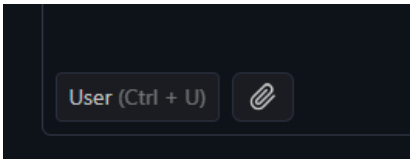
Les éléments les plus importants à retenir sont ceux-ci : la zone de discussion où on écrit notre texte est en bas, tandis que le menu pour choisir le modèle d'IA avec qui on parle est situé en haut. À gauche se situe l'historique des conversations, qu'on peut supprimer ou renommer afin de s'y retrouver, et encore plus à gauche se situe la barre latérale, qui permet de changer de menus.

Quand on envoie un message à notre assistant, il y a un menu mis à disposition juste en-dessous :

De **gauche à droite** : Régénérer le message ; Lui faire continuer son message ; Faire une branche pour pouvoir envoyer des messages différents ; Copier le message de l'assistant ; Modifier le message de l'assistant ; Et enfin, supprimer le message.

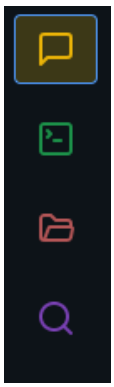


Ces boutons permettent de créer une nouvelle conversation, ou un dossier où on peut en créer.



Ces boutons dans la zone de discussion permettent de passer de l'utilisateur à l'assistant, donc d'assumer le rôle de l'IA temporairement pour mieux la guider, par exemple. Le bouton avec le trombone permet d'attacher un fichier pour que l'IA puisse travailler avec.
Le bouton à droite de la liste pour charger le modèle permet de l'éjecter de la mémoire pour pouvoir éventuellement charger un autre modèle pendant une conversation.

Barre latérale

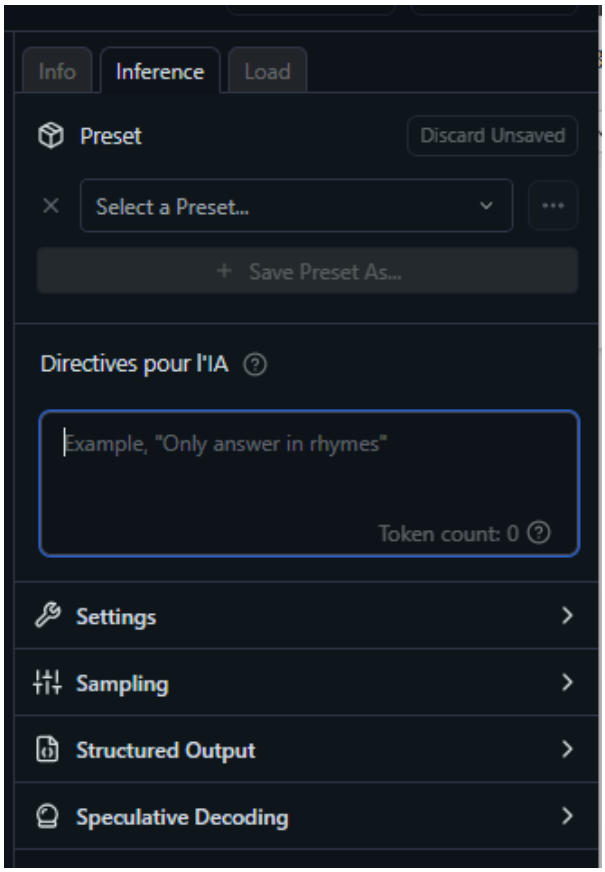


La barre latérale permet donc de basculer entre les différents menus offerts dans l'application. Le menu jaune est le menu de chat, de conversation avec l'IA. L'icône rouge de dossier permet de voir quels modèles d'IA sont présents sur notre machine. L'icône de terminal vert permet d'accéder au menu de développeur, où on peut exécuter un serveur avec lien API en local.
Enfin, le menu avec la loupe violette permet d'obtenir et de rechercher d'autres modèles IA pour les exécuter en local. Tous ces menus vont être abordés dans leur propre section.

Menu Développeur

Ici, on peut trouver tout ce qu'il faut pour modifier un modèle, exécuter un serveur avec endpoint API en local, ou même modifier comment l'IA va se comporter et répondre. Ce menu n'est pas nécessaire ni utile dans la majorité des cas pour les utilisateurs standards, mais il mérite quand même d'être évoqué :

On peut donc y voir les journaux d'exécution de notre IA sur le serveur local s'il est lancé et actif, modifier les paramètres de ce serveur (comme le numéro de port), et modifier comment l'IA va répondre, avoir des précisions sur le modèle, et voir la structure des liens API.



Ici, on peut modifier les directives données à l'IA, modifier les paramètres de température pour varier les réponses, modifier la structure de sa réponse...

Dans l'onglet "Load" se trouve des paramètres concernant le chargement du modèle dans la mémoire vive, et d'autres paramètres expérimentaux.

Dans l'onglet "Info" se trouve des informations sur le modèle, tout simplement.

Menu Mes modèles

Ici, on peut voir le répertoire où se situent nos modèles, et on peut le changer. On peut aussi voir la liste des modèles installés localement, ainsi que l'espace disque qu'ils prennent individuellement et au total.

C'est là qu'on peut modifier les paramètres globaux pour l'IA, car les paramètres vus précédemment dans le menu développeur ne s'appliquent qu'à l'IA lorsqu'elle est activée et à l'écoute sur le port réseau spécifié.

On peut aussi ouvrir les modèles afin de les voir sur un site comme HuggingFace, qui répertorie un grand nombre de LLMs (et qui peut servir de source), ou encore supprimer les modèles d'IA qui ne sont pas (ou plus) utilisés.

Menu Découvrir

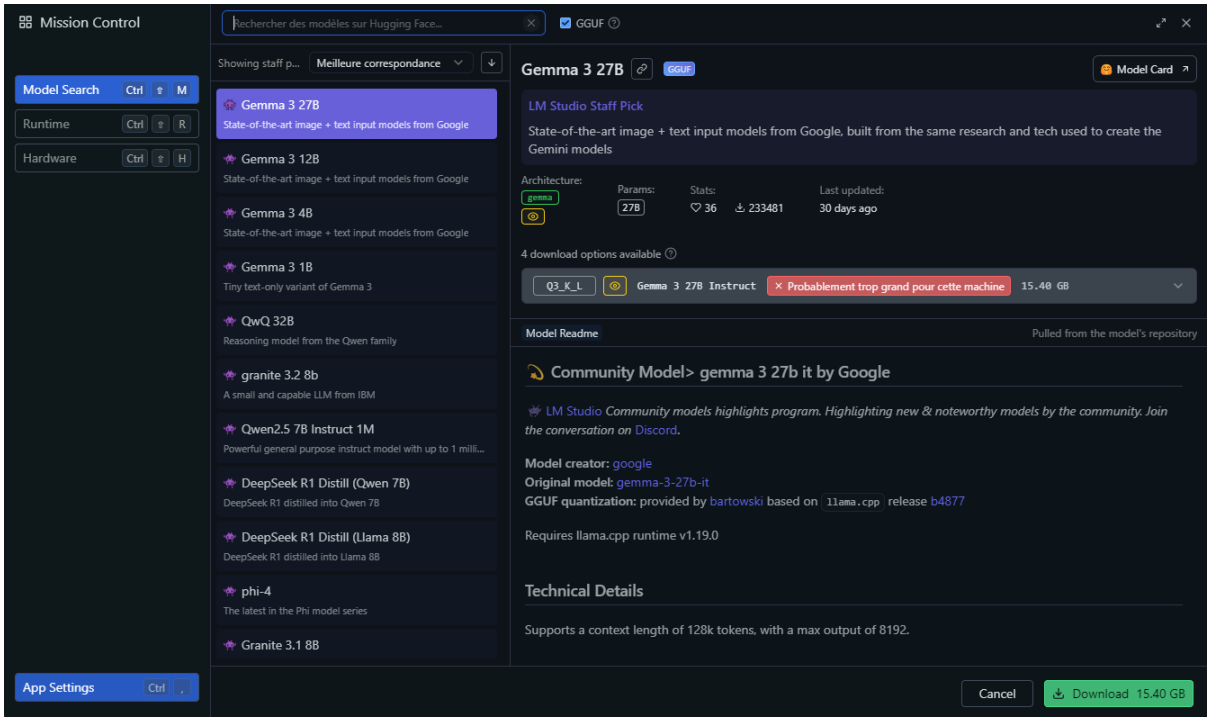
C'est dans ce menu qu'on va pouvoir obtenir des nouveaux modèles d'IA pour LM Studio.
Ici, on obtient une grande liste de LLMs mis à disposition pour une utilisation libre et non-commerciale dans la grande majorité des cas, et on peut aussi voir des informations par rapport au modèle, comme l'espace qu'il prendra et s'il va pouvoir s'exécuter en local sur notre machine ou pas (car certains modèles sont plus "gourmands" que d'autres).

Ici, on peut aussi mettre à jour et télécharger les "runtimes" pour modèles larges, qui permettent donc d'exécuter les LLMs, ainsi qu'une section "Hardware" pour voir des informations sur notre matériel. Tout en bas, on peut accéder aux paramètres de l'application en cas de besoin.

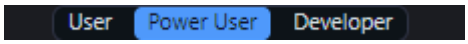
Changer le langage de l'interface

Par défaut, LM Studio possède une interface utilisateur en anglais, mais il est possible de changer la langue utilisée, même si ces traductions ne seront pas intégrales, car elles sont encore en béta. Pour cela, il faut se rendre dans les paramètres de l'application (**Ctrl + , sur Windows**), puis on peut choisir le langage voulu dans une liste déroulante sous la section "Language". On peut aussi accéder aux paramètres de l'application depuis le menu "Découvrir" (surligné en bleu dans la capture d'écran

suivante) :



Interface “simplifiée” :

Une fois qu’on a obtenu tous les modèles d’IA souhaités, il est possible de simplifier l’interface pour n’avoir que l'interface “Discussion”, et faire disparaître la barre latérale. Pour cela, il faut simplement cliquer sur “User” dans ce menu, qui est situé tout en bas de l’application, vers la gauche : 

On se retrouve donc avec une interface simplifiée, qui n’inclut que l'interface de discussion. On peut bien sûr rebasculer vers le profil “Power User” pour avoir accès à la barre latérale en cliquant sur celui-ci à tout moment.

Points importants :

Il faut noter qu’exécuter un LLM en local sur sa machine est assez lourd en termes de ressources, d’où leurs recommandations par rapport à la mémoire. De plus, un LLM qu’on exécute en local sera toujours – ou du moins très souvent – plus lent qu’une IA utilisée sur Internet, même si l’utiliser en local possède quelques avantages, comme la confidentialité de ses données ou la capacité d’utiliser des documents en local sans se soucier qu’OpenAI (l’entreprise derrière ChatGPT) y ait accès, par exemple.

Malgré ces avantages présentés par l'exécution d'un LLM en local, il existe néanmoins des inconvénients. Comme cité précédemment, la vitesse de la réponse sera plus ou moins variable en fonction des capacités de votre matériel. Si l'ordinateur possède 8Go de mémoire vive, il prendra plus de temps à répondre que sur un ordinateur identique avec le double. De plus, la qualité de la réponse de l'IA peut varier si on ne lui donne pas d'instructions spécifiques : il faut savoir guider l'IA afin de la garder entre certains "rails" afin qu'elle puisse fournir exactement ce qu'on attend d'elle.