Ghid de utilizare
a programului
LM Studio (v0.3.8 / ianuarie 2025)

https://lmstudio.ai/

LM Studio este un program care facilitează interacţiunea utilizatorului uman cu modele Conversaţionale şi Imagistice de Inteligenţă Artificială ("*Large Language Models*", "*Vision Models*") (în continuare, "***modele***" şi "***LLM***").

Programul este distribuit cu titlu gratuit şi poate fi folosit pe computere/laptopuri modeste (la standardele din ianuarie 2025).

Pentru a putea rula, procesorul trebuie să suporte instrucţiuni ***AVX2*** – în general, acestea sunt procesoare Intel i5/i7/i9 şi AMD (Ryzen); totodată, memoria ***RAM*** să fie de minimum ***8 GB***, iar memoria video (***VRAM – Video RAM***, sau "memoria plăcii video") să fie mai mare de ***1 GB***.

Spaţiul pe unitatea de stocare folosită de modelele preluate de pe internet ca fişiere cu extensia ".gguf" trebuie să fie de cel puţin 2 GB (pentru 1 un singur model preferat, simplu/elementar/rudimentar), ajungând în domeniul (sute de GB -> câţiva TB) pentru o colecţie de zeci sau sute de modele preferate.

În cazul modelelor care nu sunt instruite sau altfel pregătite pentru accesarea internetului, acelea sunt plasate pe o unitate locală de stocare (hard disc, SSD) şi interacţiunea (conversaţia) este de asemenea locală - fără transfer de informaţie către şi dinspre internet; astfel, conversaţiile rămân private.

În cazul celorlaltor modele, conversaţiile sau porţiuni din acestea pot fi trimise către site-uri pe internet, încât utilizatorul nu se poate aştepta la confidenţialitate.

Unii parametri de funcţionare a fiecărui model pot fi stabiliţi de către utilizator, cei mai importanţi fiind:

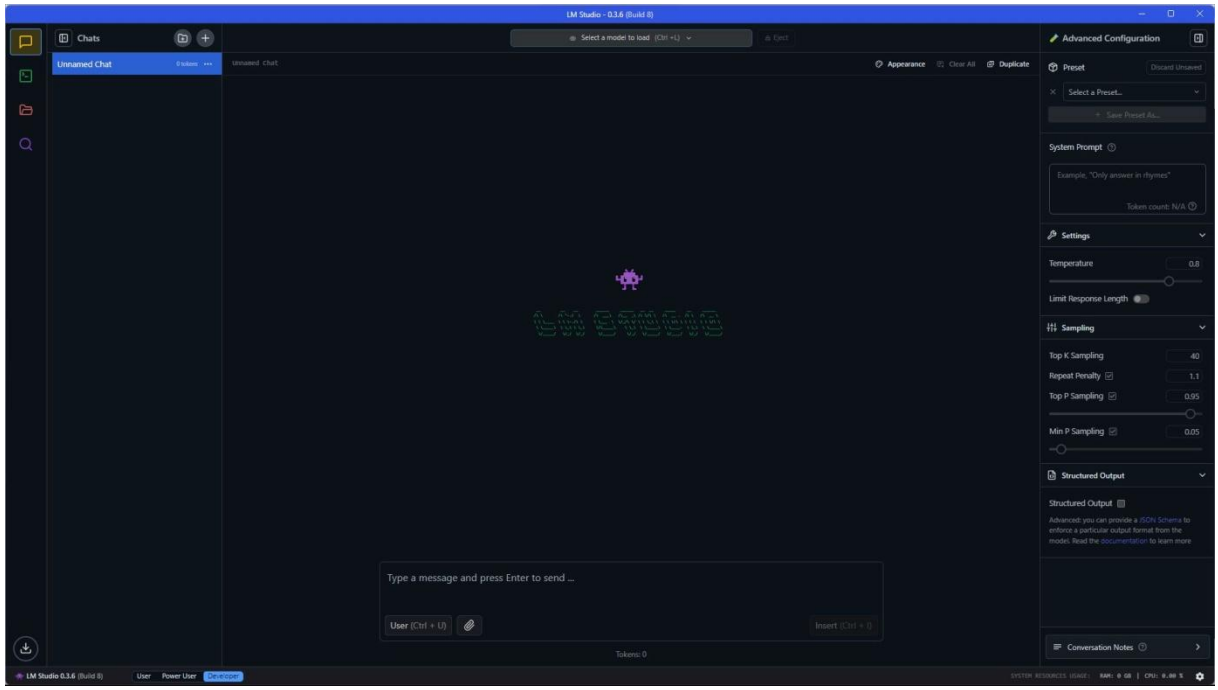
- ***Temperature*** – ghidează modelul către afirmaţii/răspunsuri/replici tot mai creative, determinând folosirea unor cuvinte/expresii rar întâlnite în limbajul natural, până la folosirea unora inexistente şi astfel cauzând *halucinaţii*
- ***Top K*** – determină alegerea din cele mai probabile K cuvinte sau particule de cuvânt care ar urma altei astfel de entităţi lingvistice ("***tokenuri***")
- ***Top P*** – probabilitatea totală (cumulată) a celor mai probabile entităţi (v. ***Top K***)
- ***Context Length*** – lungimea în tokenuri a textului care poate fi analizat într-un singur pas de către model, incluzând textul conversaţiei până în momentul calculării; cu cât este mai mare, cu atât modelul este mai performant

Interfaţa *LM Studio* este adecvată interacţiunii cu modele şi în v0.3.8 (cea curentă la data acestui document, în ianuarie 2025) este alcătuită din patru pagini sau secţiuni:

- ***Chat***:
  - se poate alege modelul cu care să se dialogheze şi i se pot stabili unii parametri de funcţionare
  - pot fi salvate seturi de parametri pentru folosire cu alte modele
  - conversaţia poate fi numită, editată şi ştearsă
  - poate fi observat gradul de solicitare a memoriei şi a procesorului ş.a.
- ***Developer*** – aici se poate lucra la nivel jos cu programul
- ***My Models*** – aici sunt afişate modelele disponibile pe unitatea locală de stocare
- ***Discover*** – aici pot fi căutate modele ce mai departe pot fi preluate în diverse versiuni

În continuare se prezintă secţiunile ***Chat***, ***My Models*** şi ***Discover***.

1. Secţiunea ***Chat***

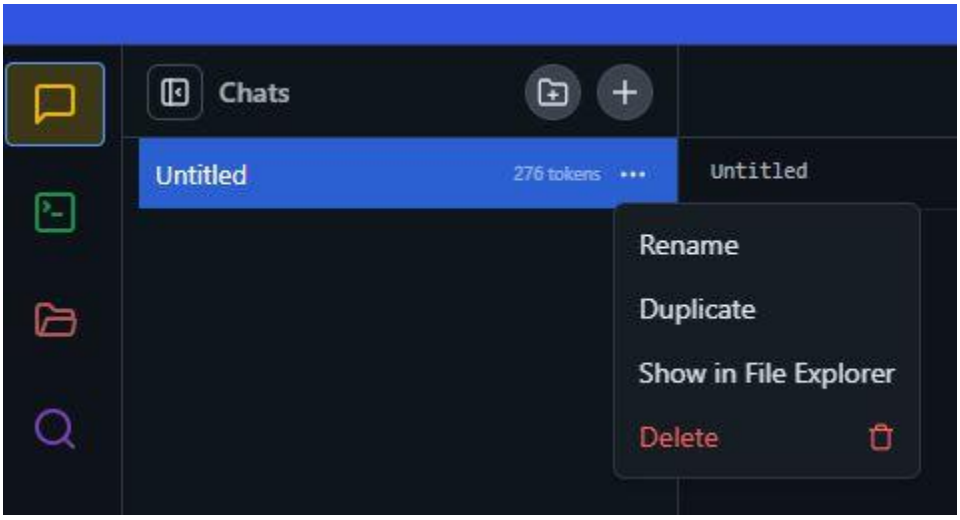


Sus, în mijloc, este lista cu modelele disponibile (existente) pe unitatea locală de stocare; acelea au fost descărcate (*download*-ed).

Din ele se poate alege câte unul pentru a susţine o conversaţie.

În stânga este lista tuturor conversaţiilor (nu cu un model anume, ci cu toate cele cu care s-a dialogat).

Click pe cele trei puncte din dreptul unei conversaţii pentru a accesa meniul de comenzi referitoare la acea.

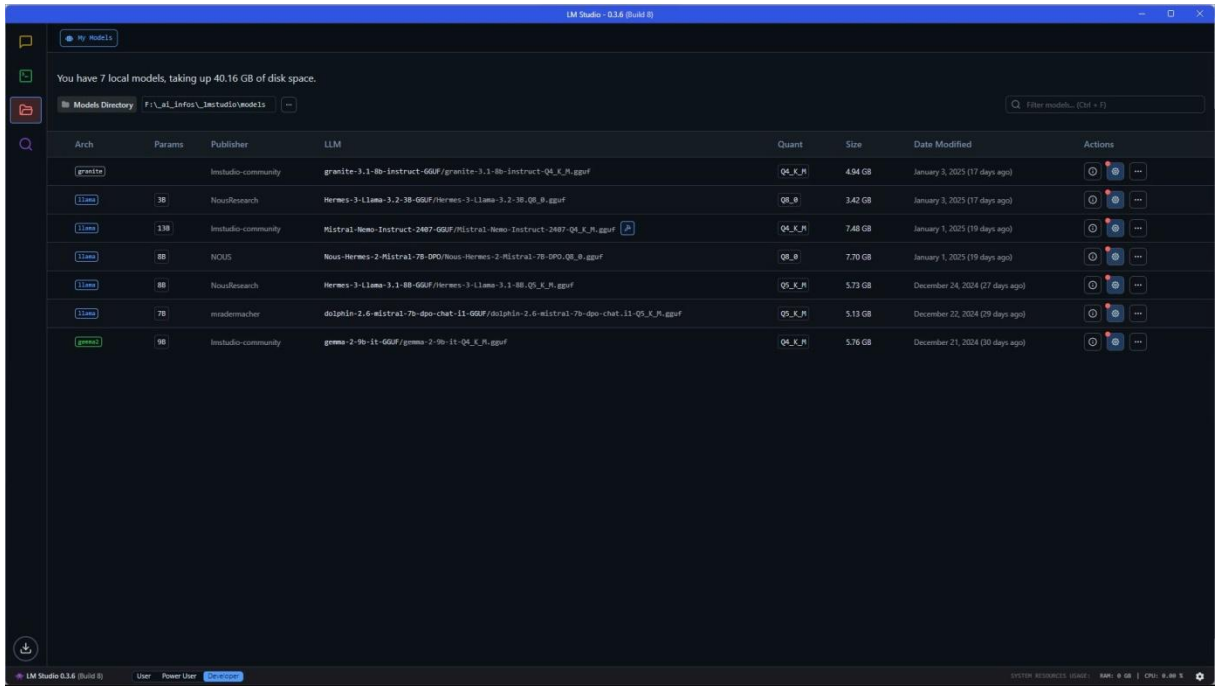


În dreapta este zona unde

- pot fi ajustaţi unii parametri ai modelului ales pentru conversaţie
- se poate defini şi/sau alege un set de astfel de parametri care astfel este aplicabil oricărui model ales
- poate fi definit aşa-numitul "***System Prompt***" – un text în care modelului i se spune cum să se comporte în dialog, cum să întocmească replicile ş.a.

Jos este caseta de text unde utilizatorul scrie întrebări/comenzi/replici înaintate către modelul curent. Un text este trimis către model apăsând fie tasta ***Enter***, fie butonul ***Send***.

2. Secţiunea ***My Models***

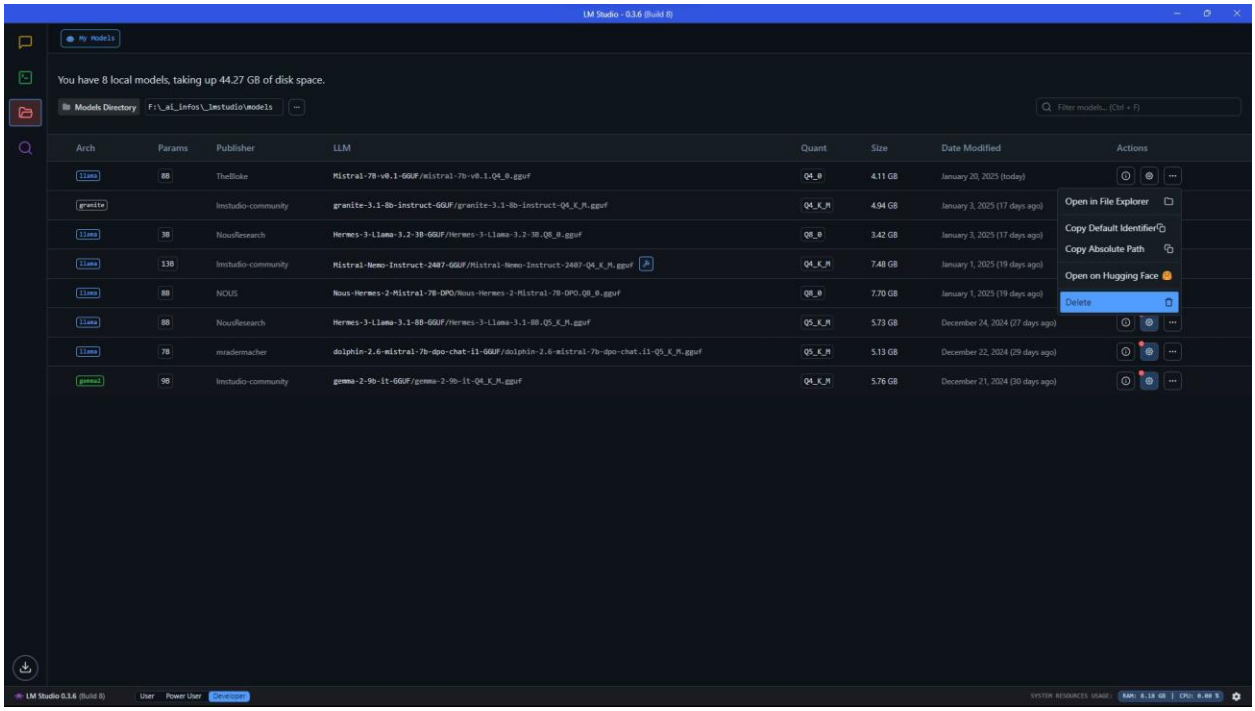


Aici este afişată lista modelelor preluate de pe internet, plasate disponibile pe unitatea locală de stocare şi astfel disponibile pentru conversaţii.

Deasupra listei, în stânga-sus, este afişat numărul modelelor disponibile şi spaţiul total ocupat pe acea unitate.

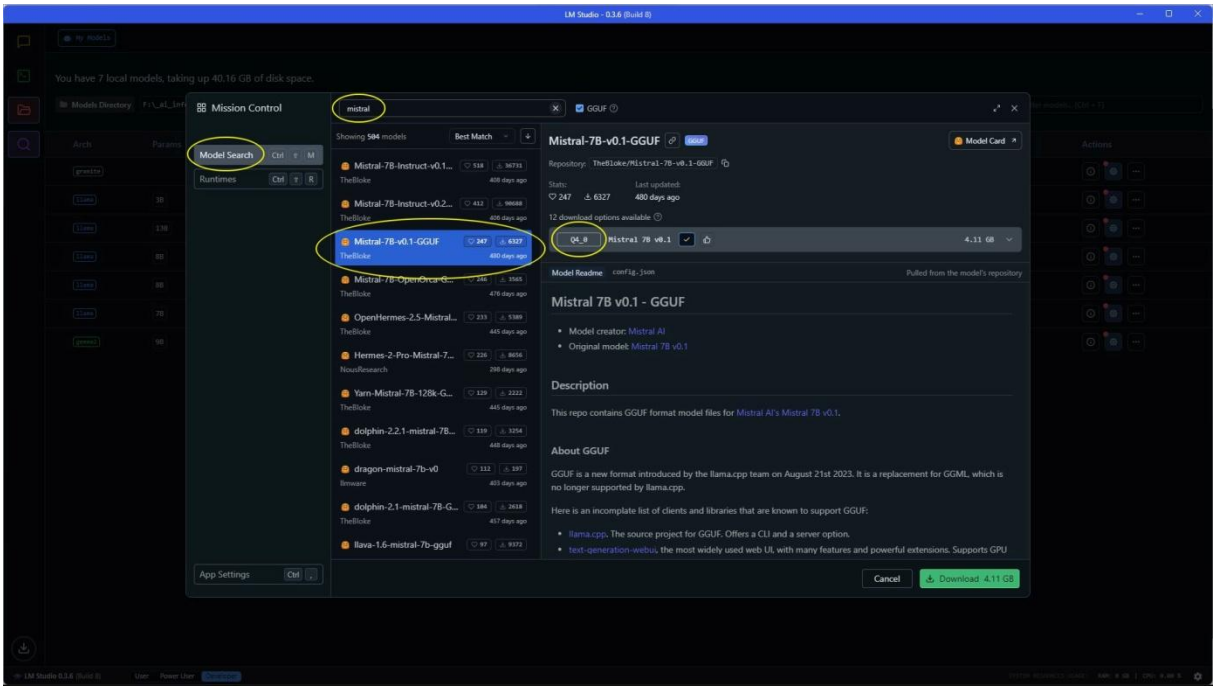
Pentru a elibera spaţiu pe acea unitate sau pentru a păstra numai modele "bune", un model se şterge astfel:

- click pe butonul din dreapta denumirii sale
- selectarea comenzii de ştergere
- confirmarea ştergerii





3. Secţiunea ***Discover***



Aici pot fi căutate modele de interes, plasate pe internet – site-ul *HuggingFace* – şi din care se pot alege pentru preluare (*download*) pe unitatea locală de stocare în scopul interacţiunii prin *LM Studio*.

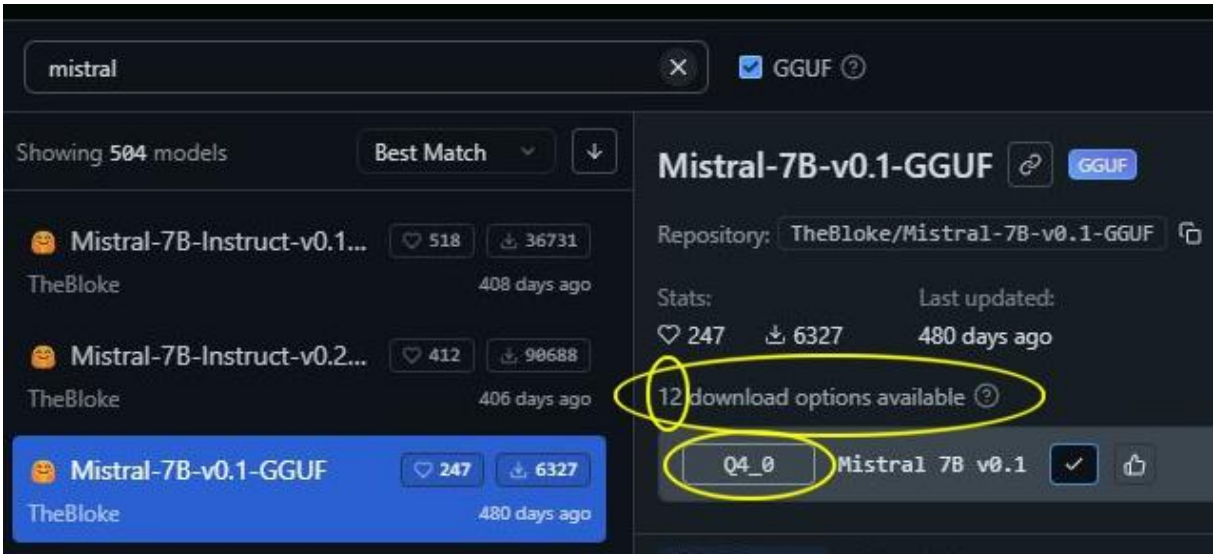
Modelele create de comunitatea *LM Studio* apar primele în listă şi sunt recomandate.
Dacă dorim a prelua un model care nu apare implicit în listă, atunci deasupra acesteia scriem (o parte din) denumirea acelui model sau a familiei de modele din care face parte (de exemplu, familiile sau seriile *Mistral*, *Hermes*, *Llama*, *Gemma*, *Qwen*).

În imaginea precedentă se observă cum:
- în stânga-sus am scris "mistral" pentru a căuta modele din familia Mistral
- am ales modelul evidenţiat cu albastru
- este preselectată versiunea "Q4_0" a acestui model

Un model poate avea mai multe versiuni – fiecare versiune având o anumită precizie, ori performanţă şi solicitând în mică, moderată sau mare măsură resursele computerului; o versiune se numeşte îndeobşte "*Quant*" – termen ce se referă la modul de stocare în model (în fişierul modelului) a legăturilor (ca numere reale) între particulele de cuvinte sau între cuvintele care fac modelul să recunoască şi să folosească limbajul natural; de exemplu

- **versiunea 4_K_M este mai performantă şi mai accesibilă** decât ***4_0***; acestea sunt versiunile recomandate pentru computere modeste/obişnuite
- **versiunea 8_0 este mai performantă, dar mai solicitantă** decât ***4_0***; este o versiune recomandată pentru computere cu minimum 8 GB de RAM
- în denumirea unui model, particula alcătuită dintr-uin număr şi litera B – de exemplu, 7B – indică
    - performanţa modelului: cu cât acel număr este mai mare, cu atât modelul este (teoretic) mai performant – astfel, modelele 70B sunt mai "bune" decât modelele 7B
    - spaţiul de memorie (RAM, VRAM) necesar modelului: un model 7B poate fi integrat în 8 GB de RAM, dar un model 70B necesită cel puţin 64 GB de RAM şi 8 GB de VRAM.
      Pentru context – în ianuarie 2025, plăcile de bază obişnuite suportă 128 GB de RAM; iar placa video cu cea mai mare capacitate are 80 GB VRAM şi este prohibitivă pentru utilizatorul comun.

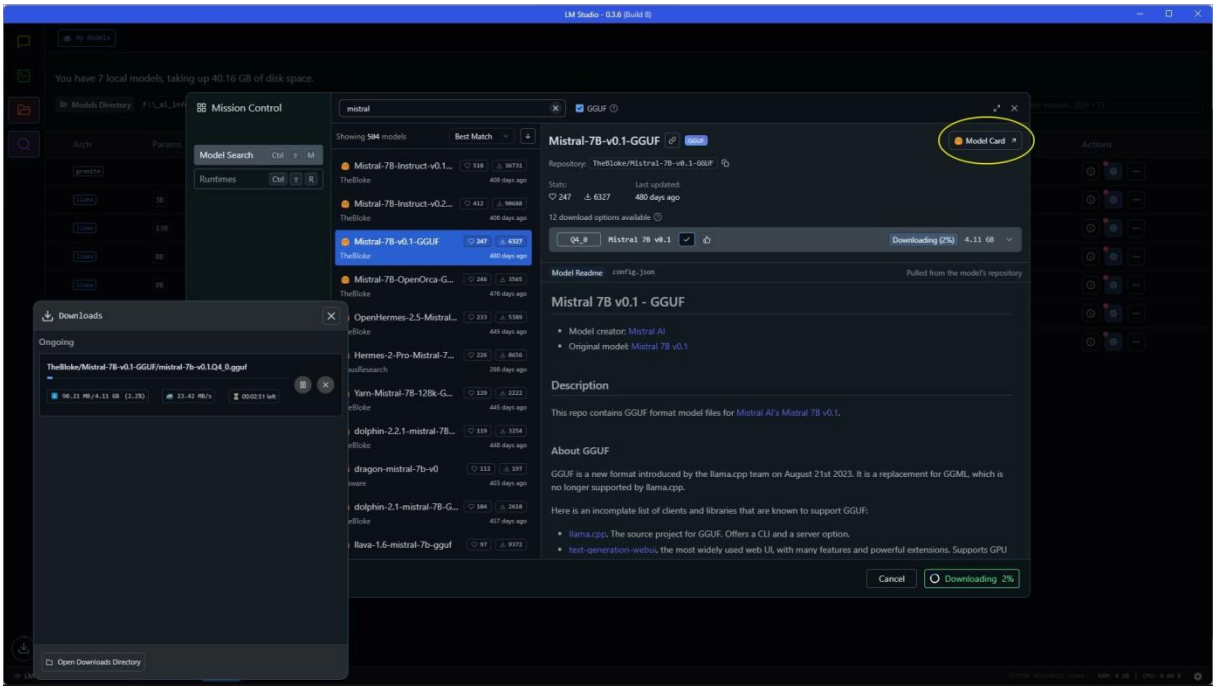
Odată ce în lista modelelor găsite a căror denumire conţine particula pe care am scris-o pentru a căuta modele s-a selectat unul – nu neapărat pentru *download*, ci probabil pentru informare – în dreapta sunt afişate informaţii despre acel model, printre care numărul versiunilor acelui model:

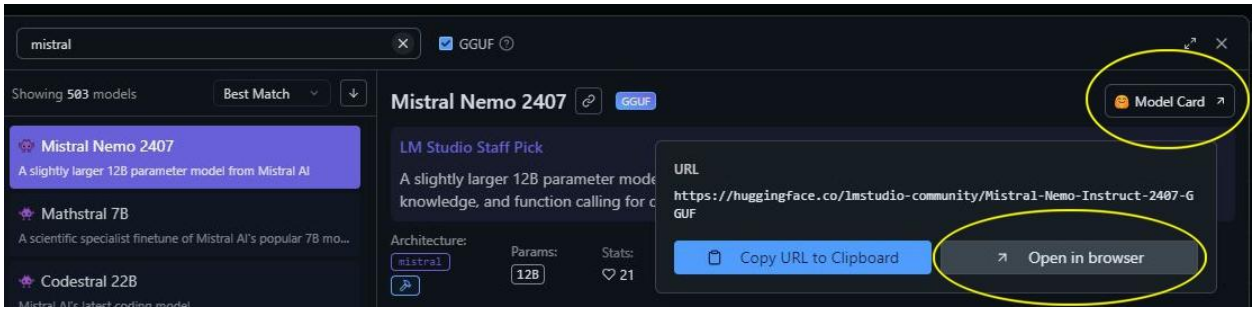


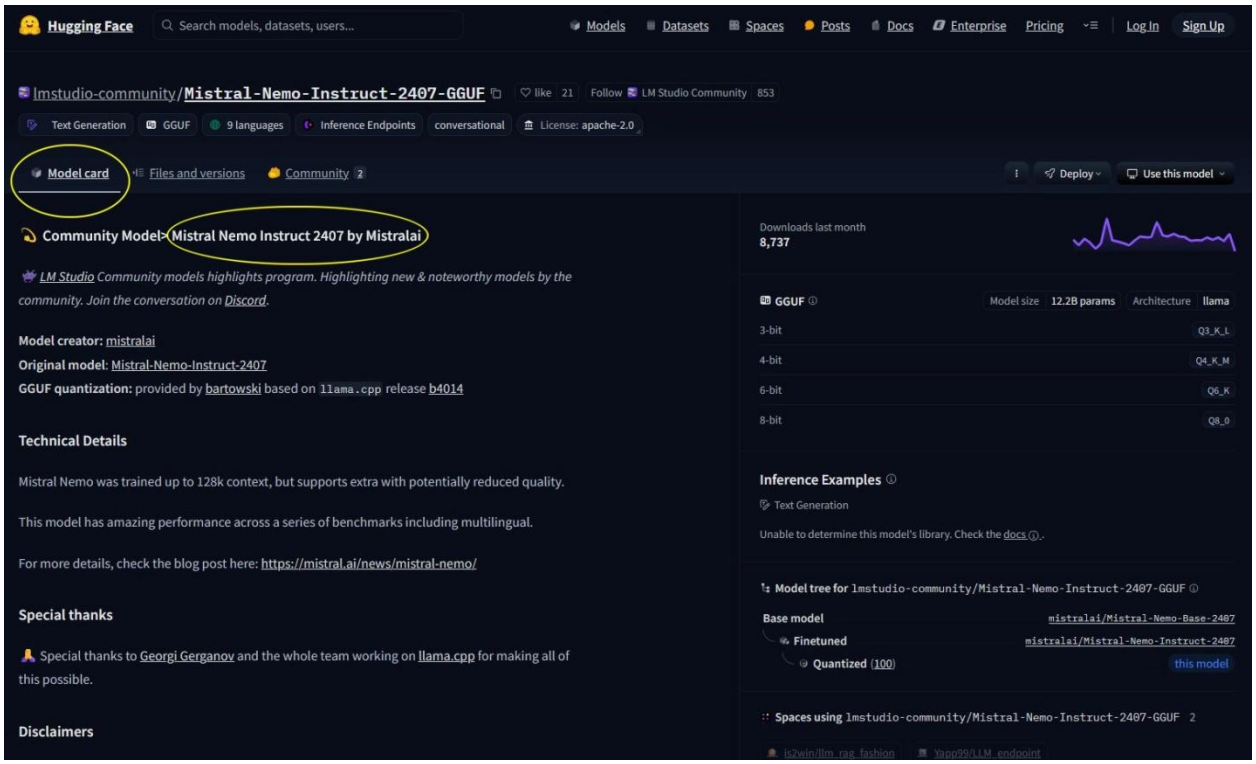
Rezonabil este ca fiecare model să aibă un aşa-numit prospect de prezentare, "*model card*", întocmit de către creatorii săi.

Acolo, ideal, modelul este prezentat în privinţa performanţei în raport cu alte modele, în privinţa condiţiilor în care se poate interacţiona (optim) cu el.

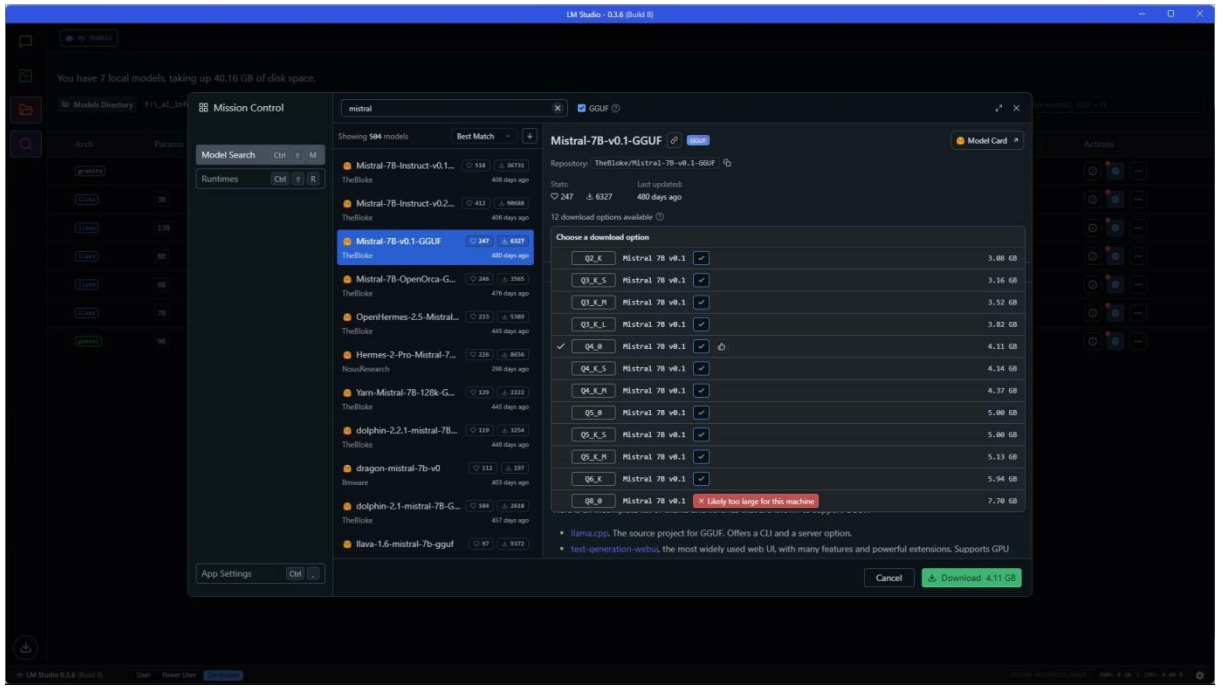
Prospectul unui model poate fi accesat pe site-ul *HuggingFace* după click pe butonul asociat din dreapta-sus, apoi selectând opţiunea "*Open in browser*" sau similar:





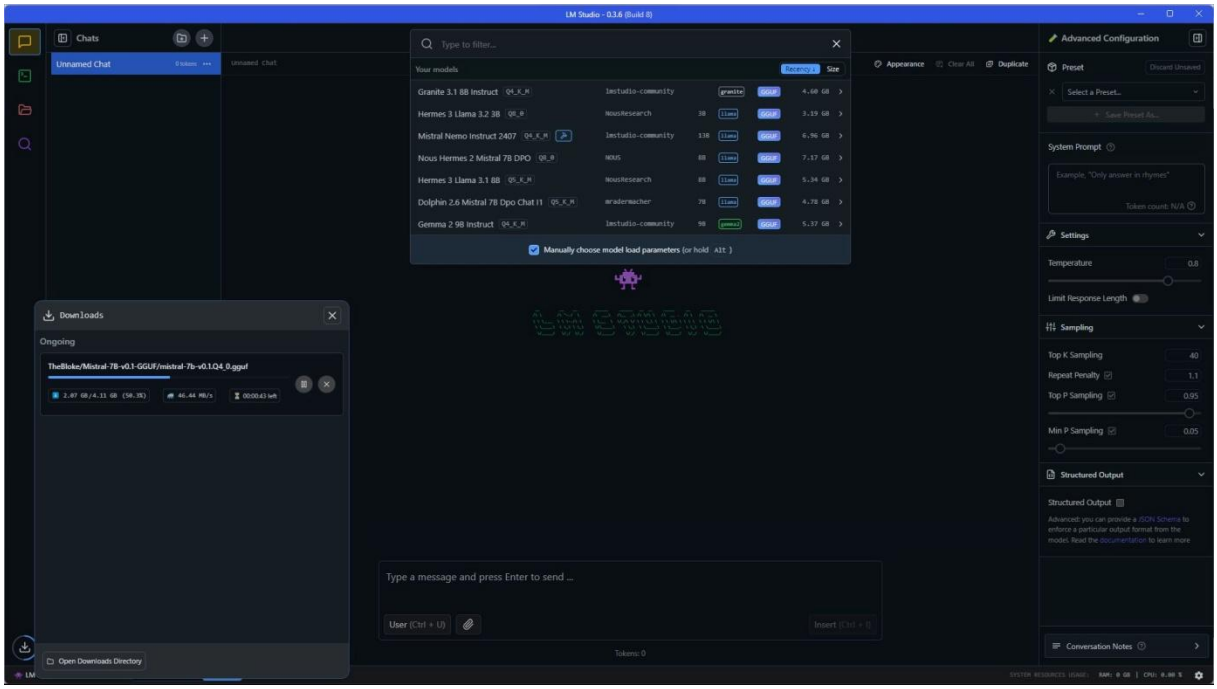


Pentru a alege o versiune de model (4_0, 4_K_M, 8_0 şamd)
- click pe butonul pe care scrie versiunea implicită de preluat
- din lista versiunilor se poate alege una diferită de cea implicită

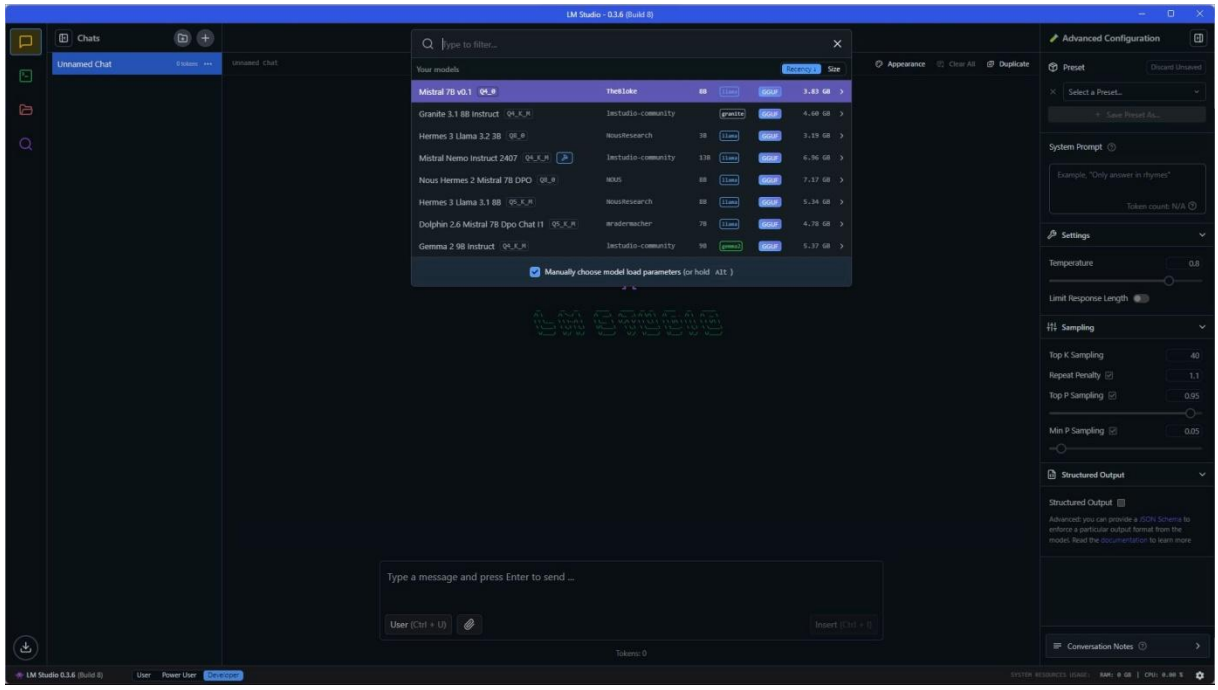


Pentru a prelua (*download*) versiunea aleasă

- în dreapta-jos, click pe butonul "***Download***"; va apărea un dialog care înfăţişează progresia preluării modelului; procesul poate fi întrerupt şi oprit; se observă lipsa modelului din listă:

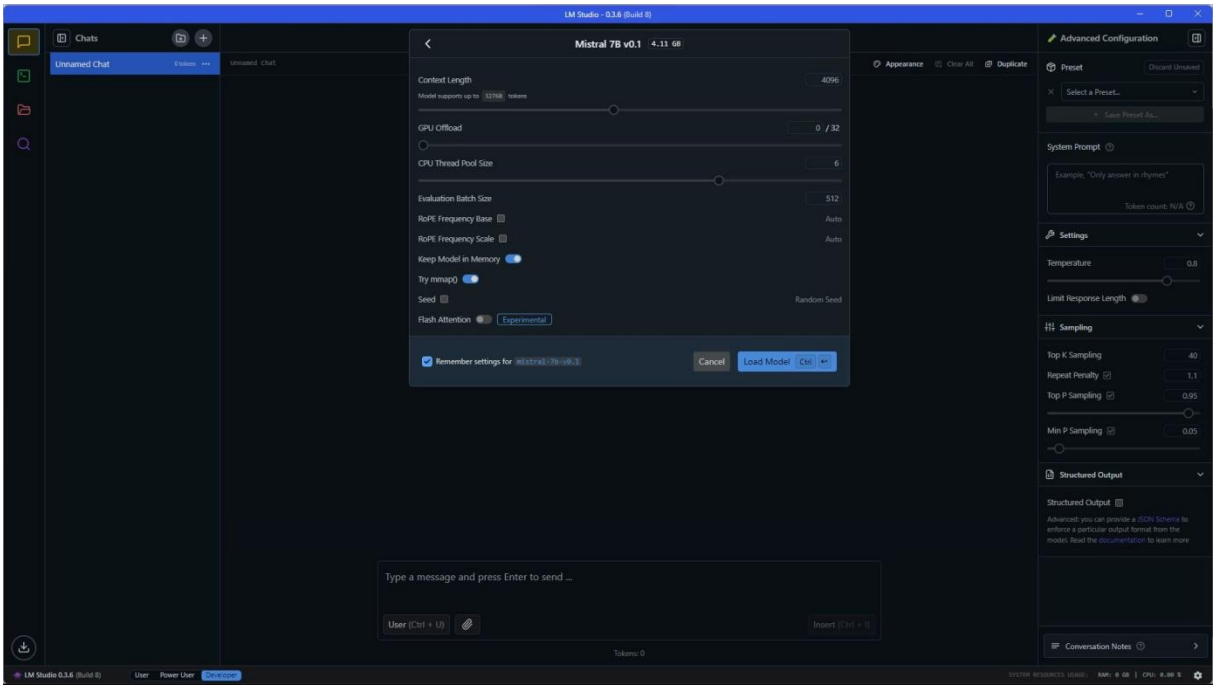


După preluarea unui model, acela va apărea în lista din secţiunea Chat şi se va putea alege de acolo pentru a dialoga cu el



Înainte de a dialoga cu modelul ales din listă e bine să stabilim parametrii accesibili de funcţionare a acelui model.

Aceasta se poate face de îndată ce alegem modelul, cu click pe denumirea sa sub care apare dialogul de stabilire/ajustare a acelor parametri:

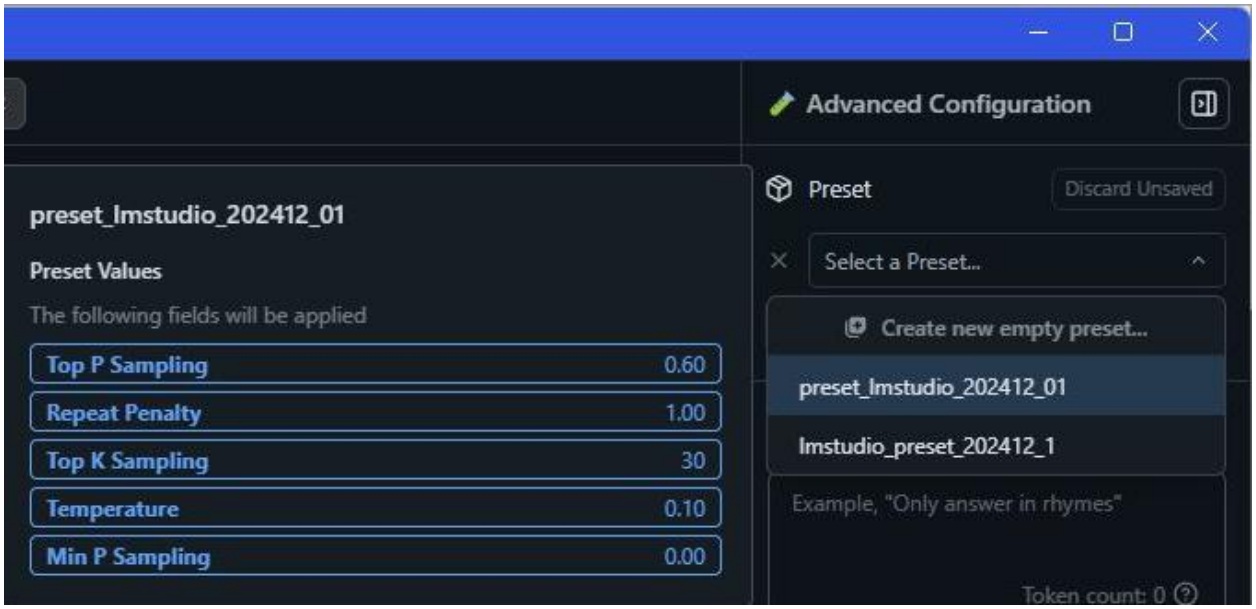


În imaginea precedentă se observă şi informaţia foarte importantă sau esenţială despre modelul ales – anume, cea despre:

- **Context Length**

Valoarea numerică respectivă nu poate fi modificată, fiind stabilită la crearea modelului şi fixă.

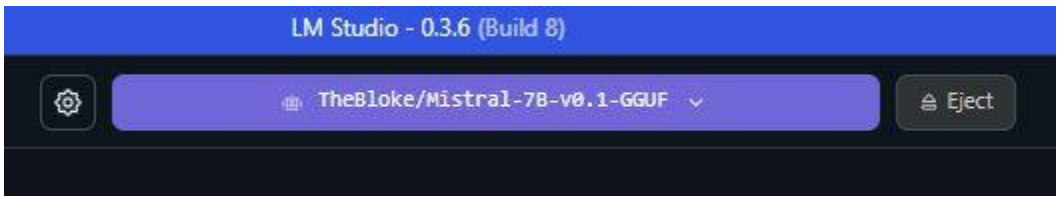
Dacă sesizăm că un model sau mai multe modele se comportă "bine" în raport cu solicitările noastre putem salva setul parametrilor care au făcut ca modelul sau modelele să se comporte astfel – aceasta se face definind – în dreapta-sus, zona "*Advanced Configuration*" aşa-numite "***preseturi***" de parametri din care câte unul poate fi ales înainte de a încărca modelul asociat sau unul pe care se vrea a se testa acel preset:



Valori de parametri care ***aplicate împreună*** (deci care pot fi salvate într-un preset pentru toate modelele) forţează un model să răspundă cât mai precis îi este posibil sunt:

- Temperature: 0.1
- Top K: 30
- Top P: 0.6

După ce stabilim acei parametri accesibili sau după ce (recomandat) alegem un preset unde câţiva parametri sunt deja stabiliţi, încărcăm efectiv modelul pentru a putea dialoga cu el: click pe butonul "***Load model***" sau similar; se urmăreşte progresia încărcării – la început rapid, apoi tot mai lent până când bara de progresie dispare şi denumirea modelului apare ca text al listei modelelor disponibile:



De-acum, modelul ales e încărcat în memorie (RAM/VRAM/ambele) unde ocupă câţiva GB; memoria poate fi eliberată apăsând butonul "***Eject***" pentru a elimina modelul din memorie

- fie înainte de o conversaţie
- fie în timpul unei conversaţii

Eliminarea cu ***Eject*** din memorie a unui model **nu şterge** nici conversaţia curentă, nici alte conversaţii cu acel model.

Tot de acum se poate dialoga cu modelul ales-încărcat: în caseta de text scriem şi aşteptăm replica modelului:





Întrucât majoritatea textelor cu care au fost instruite majoritatea modelelor sunt în limba engleză, se subînţelege că un dialog de calitate se recomandă a fi condus în acea limbă; aceasta depinde pur şi simplu de *corpus*-ul cu care a fost instruit un model – există modele la fel de "bune" la mai multe limbi, de exemplu *ChatGPT* şi unele din familia *Qwen*.

Exemple de solicitări/"***prompt***"-uri în limba engleză de trimis către un model:

- You write a concise and precise history of the Roman Empire.
- You write an essay about autumn.
- You tell me which icecream is better-tasting: vanilla, or strawberry.
- You write the plot of a movie involving two bank robbers and one cop.
- You write a poem in the style of Keats.
- Tell me the conditions that favoured the French Revolution that ended monarchy in that country close to the end of the eighteenth Century.
- You make a detailed comparison between the Battle of Waterloo and the Battle of Stalingrad.
- Why is the sky blue?
- Are you alive?
- What time is it now?

Se poate comunica şi în alte limbi decât engleza, însă riscul apariţiei unor cuvinte inexistente şi cel al identificării incorecte a unor cotntexte este mai mare decât în cazul aceleia.