Inteligenţă Artificială
Asamblarea unui sistem de informare/conversaţional
cu un Model Vast De Limbaj (LLM)
şi cu programul GPT4All

Programul GPT4All; Documente Locale

Acurateţea informaţiilor; "halucinaţiile" modelelor LLM

Ajustarea unor parametri ai modelului LLM; configurarea ca partener de dialog a modelului LLM

Expresii care pot fi folosite în conversaţii

Unde sunt LLM-urile de pe calculatorul meu?

Colecţiile de Documente Locale – instruirea unui LLM folosind *numai surse locale* ***de informaţie***

Colecţiile de Documente Locale – recomandări pentru organizare

Colecţii de informaţii pe Internet

Crearea de fişiere (documente) PDF

Instruirea preliminară a unui LLM pentru a-l determina să răspundă într-un fel anume

Exemple de Prompt Template; aplicabilitate şi eficienţă în funcţie de tipul informaţiei solicitate

Clonarea unui LLM

Scenarii de utilizare
- *conversaţii şi informaţii obţinute din cărţi cu diverse subiecte*
- *conversaţii şi informaţii obţinute din versurile unor cântece*

Noţiuni; Termeni; Abrevieri

Resurse pe Internet

Acest document este supus modificării/actualizării ***frecvente***
în acord cu starea şi evoluţia celor prezentate
şi poate fi preluat de la adresa:
https://tinyurl.com/39f5uafk

Până în 2023, în precedenţii câţiva ani au fost create aşa-numite *Modele Vaste De Limbaj (ori, Lingvistice)* (*Large Language Model*, "*LLM*") care în esenţă sunt programe de calculator capabile să interpreteze formulări frazeologice diverse şi în general "limbajul natural", uman, care de obicei este structurat după reguli – de exemplu, un substantiv poate fi urmat de un atribut sau de un verb, un atribut poate avea o intensitate anume şamd.

Un *LLM* este aplicat pe un text (sau fişier cu text) sau pe mai multe în care identifică şi apoi foloseşte concepte pentru a extinde contextul, a "dialoga" cu un interlocutor (uman) ş.a.

În esenţă, aceasta se face prin detectarea cuvintelor şi grupurilor de cuvinte organizate în toate felurile posibile într-o anumită limbă, prin compararea lor cu secvenţe cunoscute din acele multe combinaţii de cuvinte, prin evaluarea secvenţelor care pot urma altora, prin estimarea distanţei în frază şi în context (în cadrul conversaţiei) dintre secvenţe pentru continuarea conversaţiei folosind acea probabilitate şamd.

Aşadar, pentru ca un *LLM* să fie util este nevoie să aibă la dispoziţie o colecţie de texte pe care să le poată citi şi interpreta; la data acestui document, acele texte pot fi unele stocate în fişiere *PDF* (*Adobe®™ Acrobat*), *DOC*/*DOCX* (*Microsoft®™ Word*), *TXT* şi formate mai puţin utilizate de document electronic.

Un *LLM* este "***instruit***" ("***training***") prin parcurgerea unui set uriaş (vast) de informaţii/documente în scopul învăţării structurii complexe a "limbajului natural" şi astfel în scopul interpretării ulterioare a "limbajului natural" pentru oferirea de informaţii structurate în acelaşi fel sau similar.

Astfel, un *LLM* poate fi instruit, de exemplu, prin parcurgerea tuturor cărţilor despre un domeniu anume; sau a mesajelor de *e-mail* între anumiţi corespondenţi; sau a informaţiilor într-o anumită limbă aflate pe site-uri enciclopedice aşa cum este *Wikipedia*. După instruire, *LLM*-ul va cunoaşte acele informaţii.

O parcurgere completă a unui set de informaţii pentru instruirea unui *LLM* se numeşte *Epocă*; astfel, calitatea interacţiunii ulterioare cu omul depinde în principal de numărul *Epocilor*.

Mai mult – calitatea/acurateţea "răspunsurilor", ori a dialogului cu un *LLM* depinde de
- calitatea informaţiei accesibile
- cantitatea informaţiei accesibile
- cerinţele interlocutorului
- formularea cerinţelor interlocutorului, inclusiv punctuaţia care indică detalii
- prezenţa cât mai multor detalii în replicile înaintate către *LLM*

La această dată, rularea unui *LLM* ***pe un calculator personal (de birou)*** necesită resurse relativ modeste:
- procesor *Intel®™ Core i5* sau similar cu frecvenţa în jur de 3 *GHz*
- memorie internă (*RAM*) de 4 *GB*, preferabil minimum 8 *GB*
- mediu de stocare pentru documentele de citit/interpretat,

de exemplu un dispozitiv *Solid State Drive* (*SSD*) de câteva sute de *GB* sau câţiva *TB*
- mediu de stocare unde să poată fi plasat/e fişierul/fişierele din care constă *LLM*-ul ales

În imaginea următoare se observă cum *LLM*-ul ocupă (se plasează în) memoria *RAM* necesară permanent:



Programul GPT4All; Documente Locale

Interacţiunea cu un *LLM* se face prin programe dedicate care pot oferi
şi posibilitatea de a adăuga documente la colecţiile disponibile.

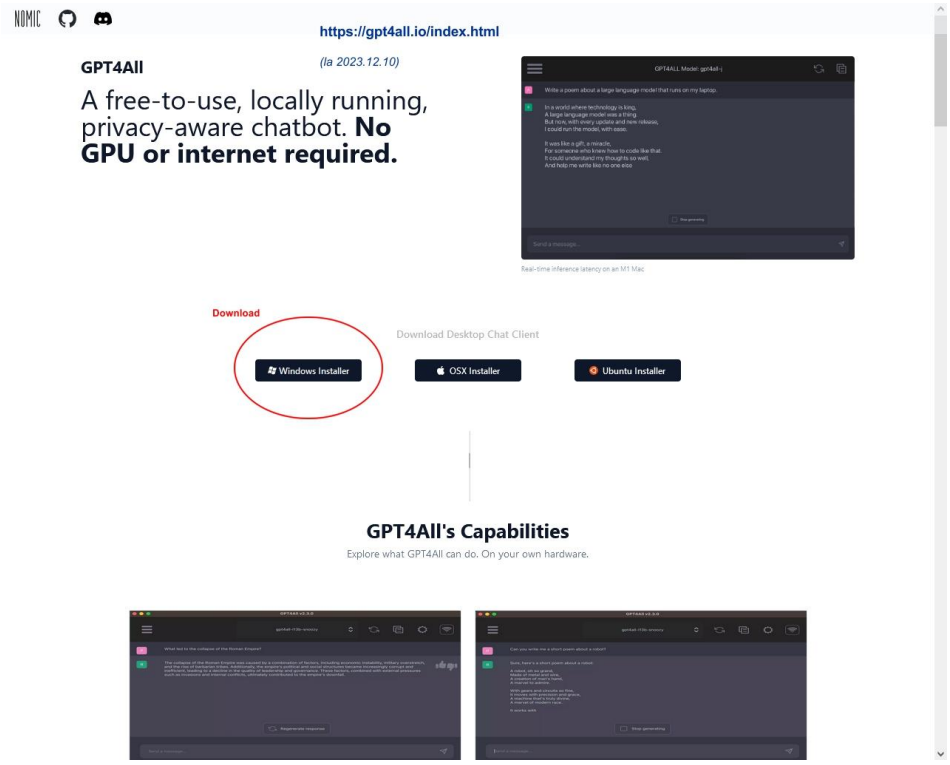
În continuare va fi prezentat programul ***GPT4All, versiunile 2.6.1 şi 2.7.1***
(versiunile ulterioare acestora pot fi *asemănătoare)*
disponibil pentru preluare gratuit de pe Internet (*download*) la adresa:

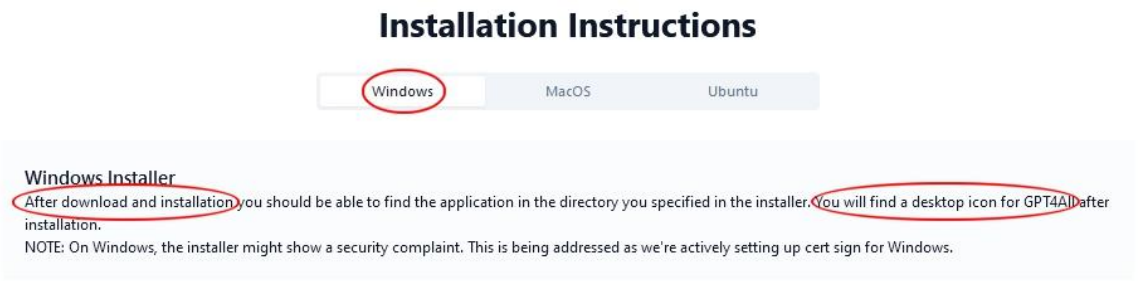
https://gpt4all.io/index.html

- ***în timpul unei conversaţii acesta NU foloseşte conexiunea la Internet***
  ci numai când preia LLM-uri şi informaţiile asociate lor
  care sunt plasate pe calculatorul unde rulează programul
  - vom instala acest program pe calculatorul personal
  - vom prelua un *LLM* de pe Internet
  - vom adăuga documente *PDF* la colecţii de documente pe care *LLM*-ul le va parcurge
  - vom interacţiona cu *LLM*-ul care va furniza informaţii
    - atât din documentele furnizate de noi
    - cât şi din "cunoştinţele" proprii

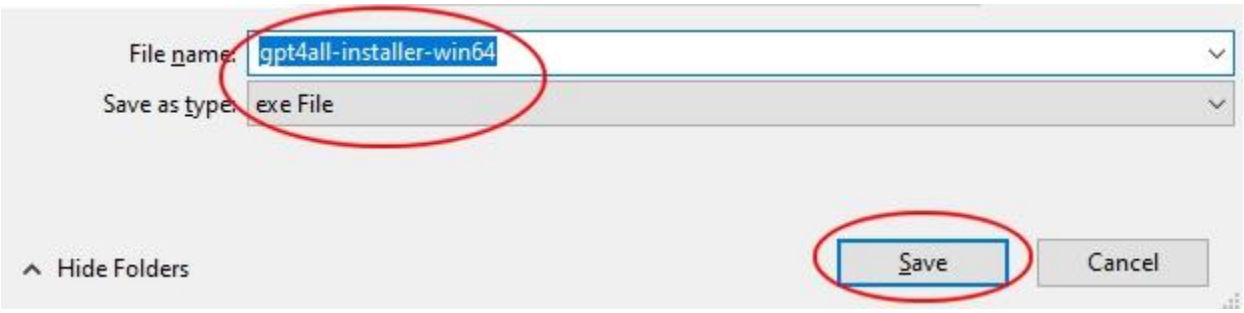
În continuare va fi prezentată instalarea şi funcţionarea ***GPT4All*** pe un calculator personal cu sistemul de operare *Microsoft*®™ *Windows*®™, versiunea 10.

Descărcăm pachetul de instalare a ***GPT4All***



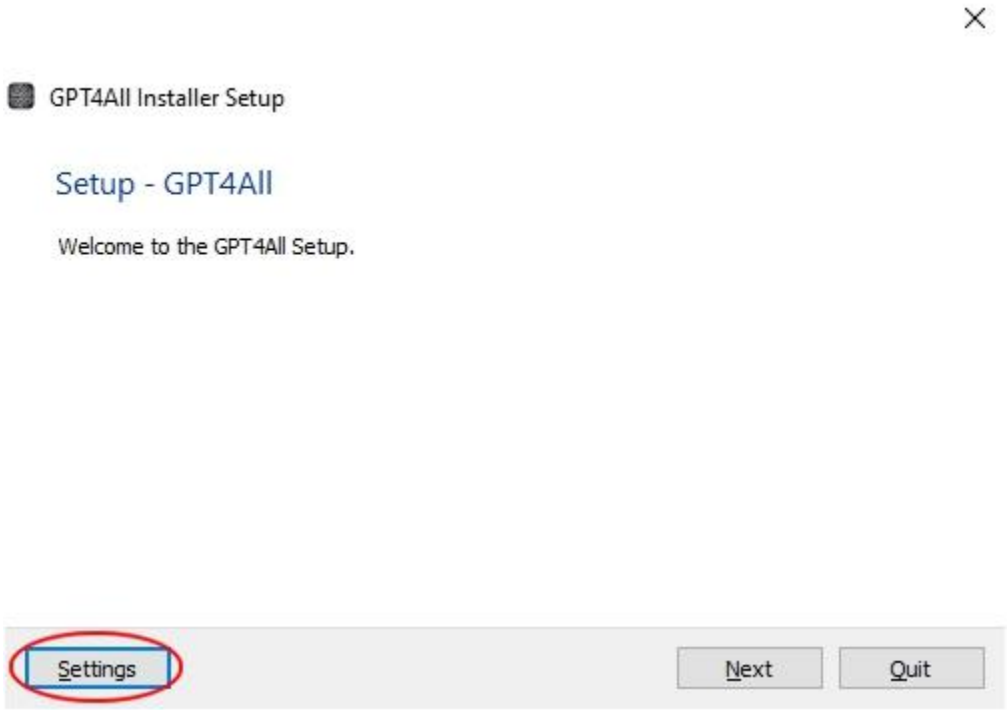


salvăm într-un folder local acel fişier (executabil)-pachet de instalare:

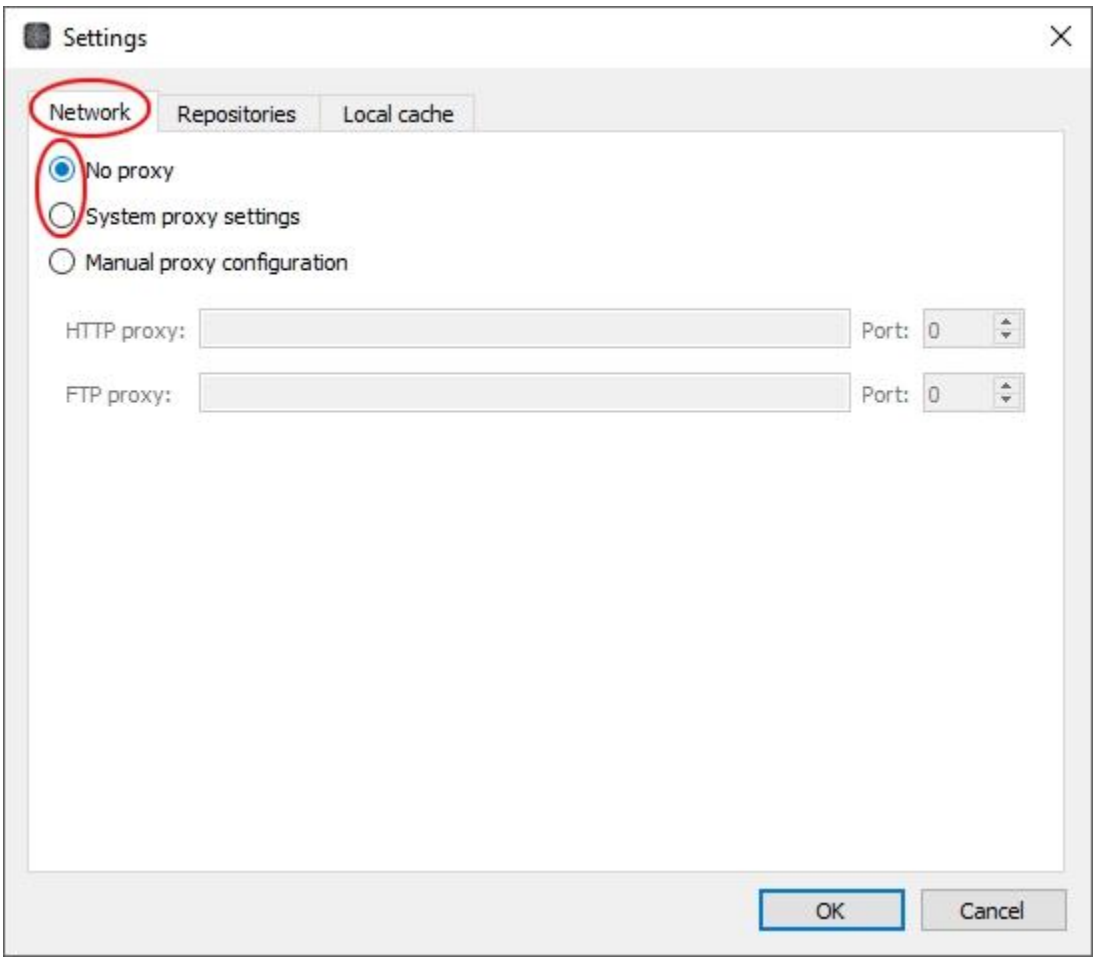


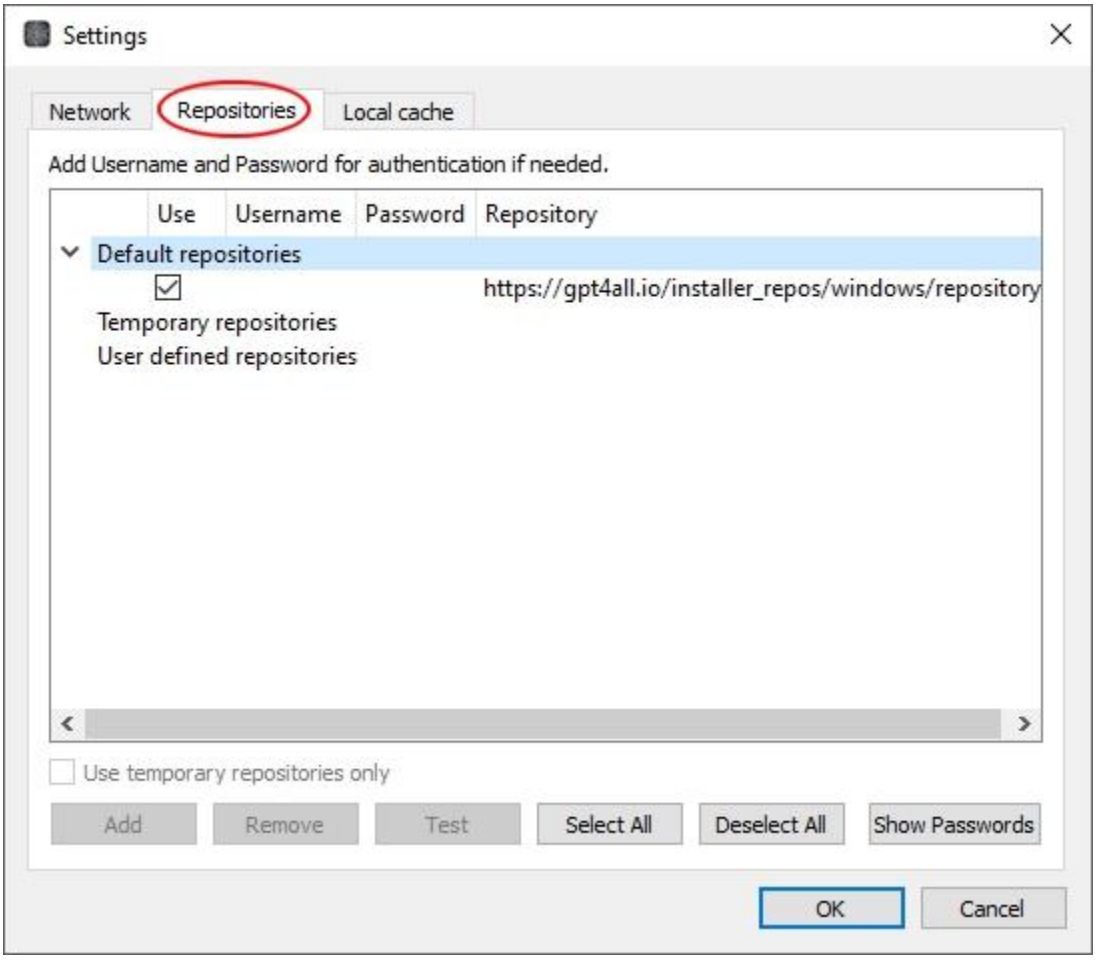
apoi lansăm instalarea programului – fie cu dublu-click, fie apăsând tasta **Enter** pe fişierul tocmai salvat:

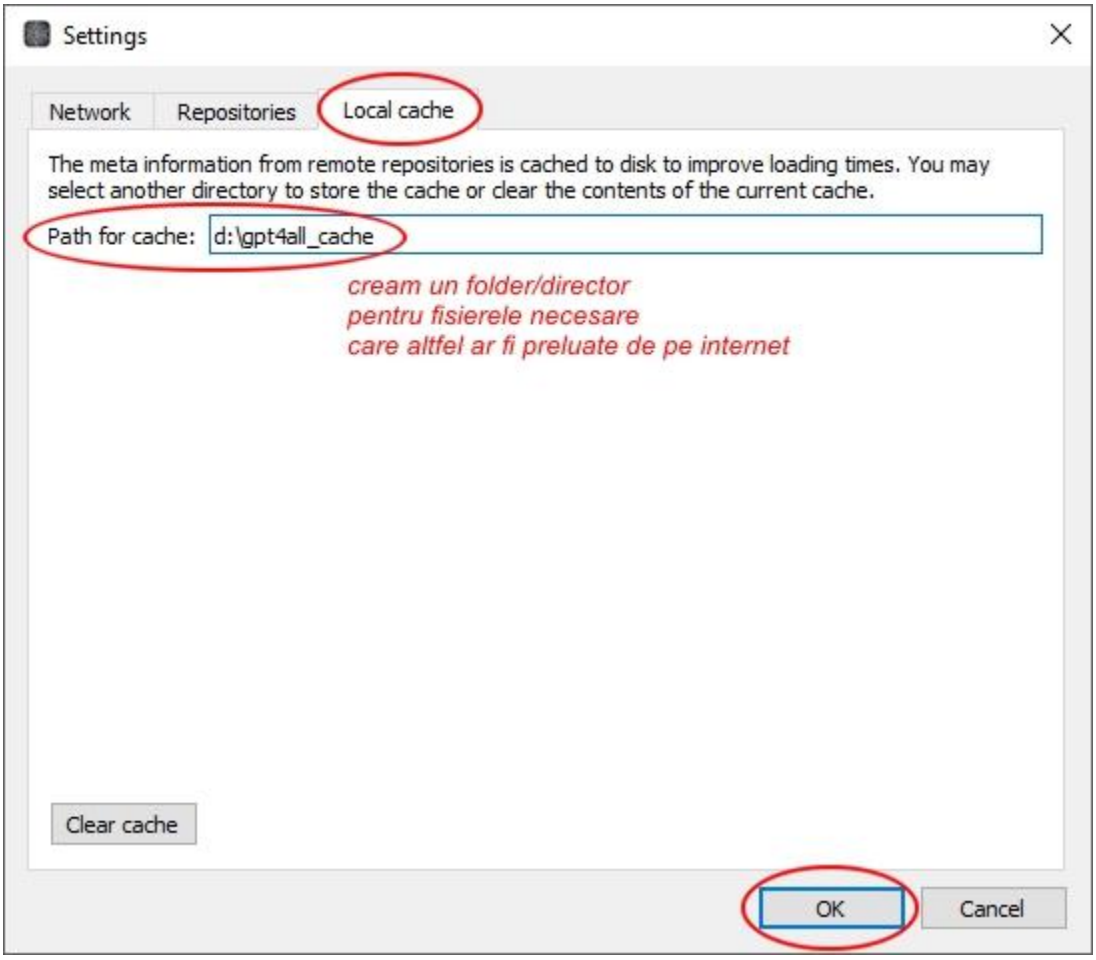
Pentru a configura elementar programul, în primul dialog apăsăm butonul "*Settings*":



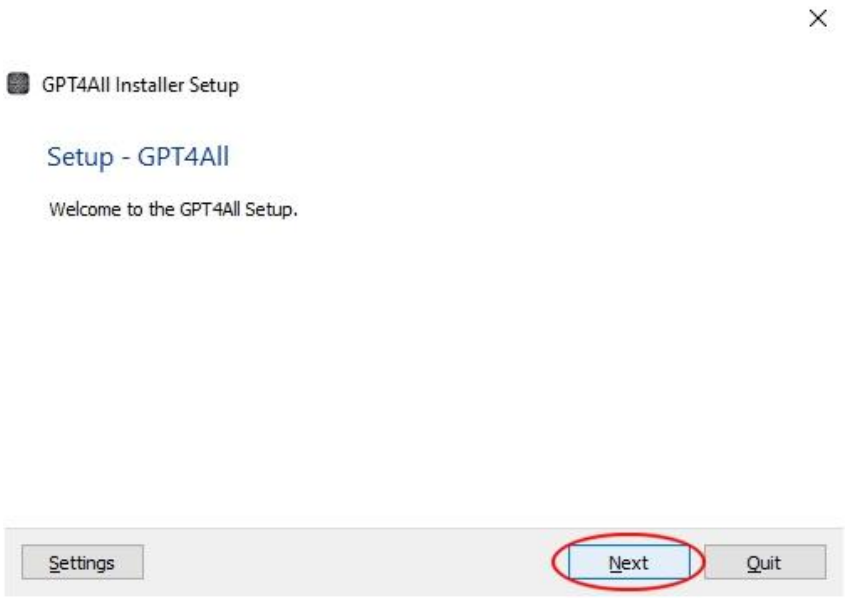
Pe paginile dialogului care apare stabilim valori ca în imaginile următoare:




Settings
Network
Repositories
Local cache
Add Username and Password for authentication if needed.
Use
Username
Password
Repository
Default repositories
https://gpt4all.io/installer_repos/windows/repository
Temporary repositories
User defined repositories
Use temporary repositories only
Add
Remove
Test
Select All
Deselect All
Show Passwords
OK
Cancel

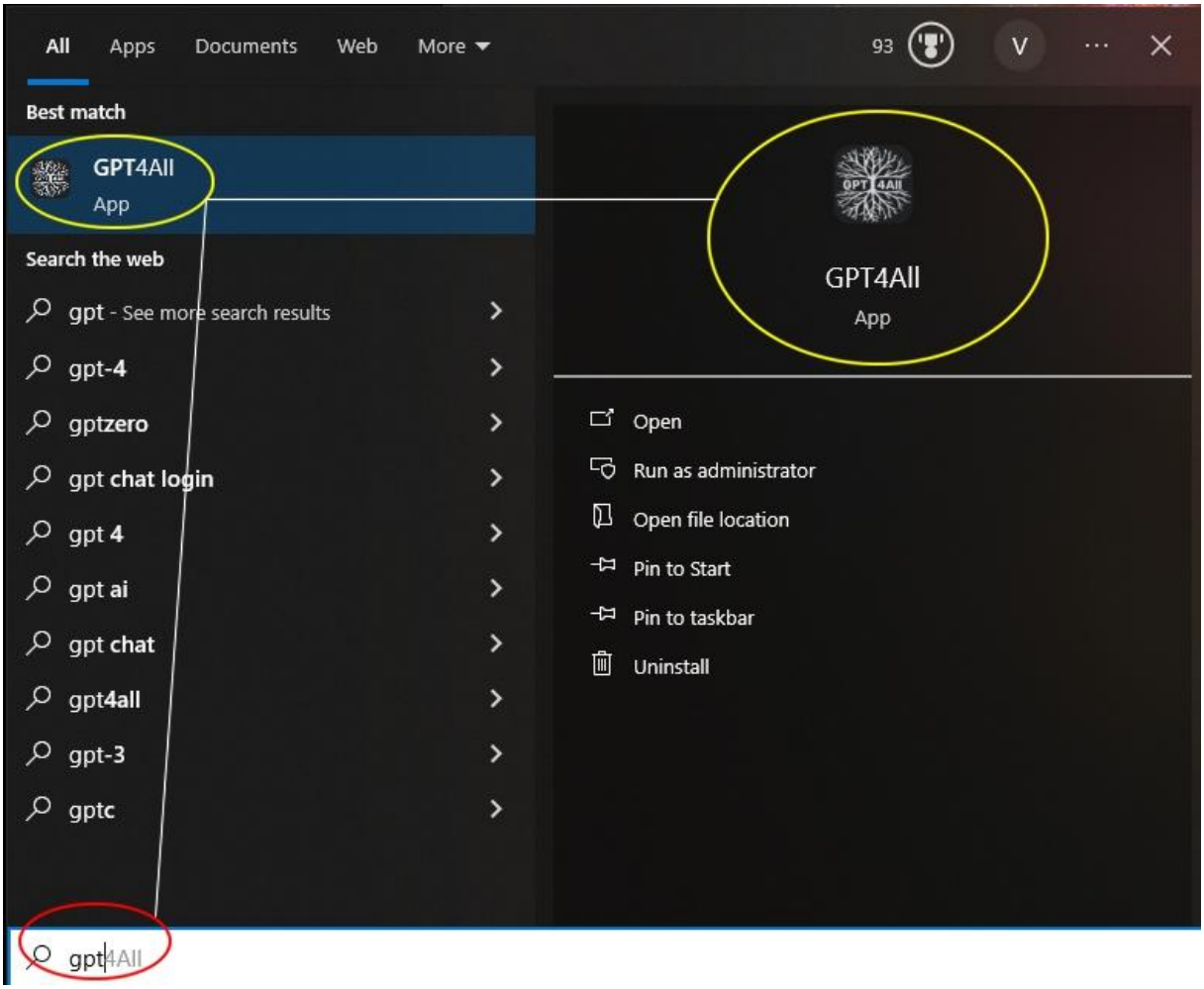


După aceste operaţiuni preliminare iniţiem instalarea propriu-zisă: click pe butonul "*Next*":

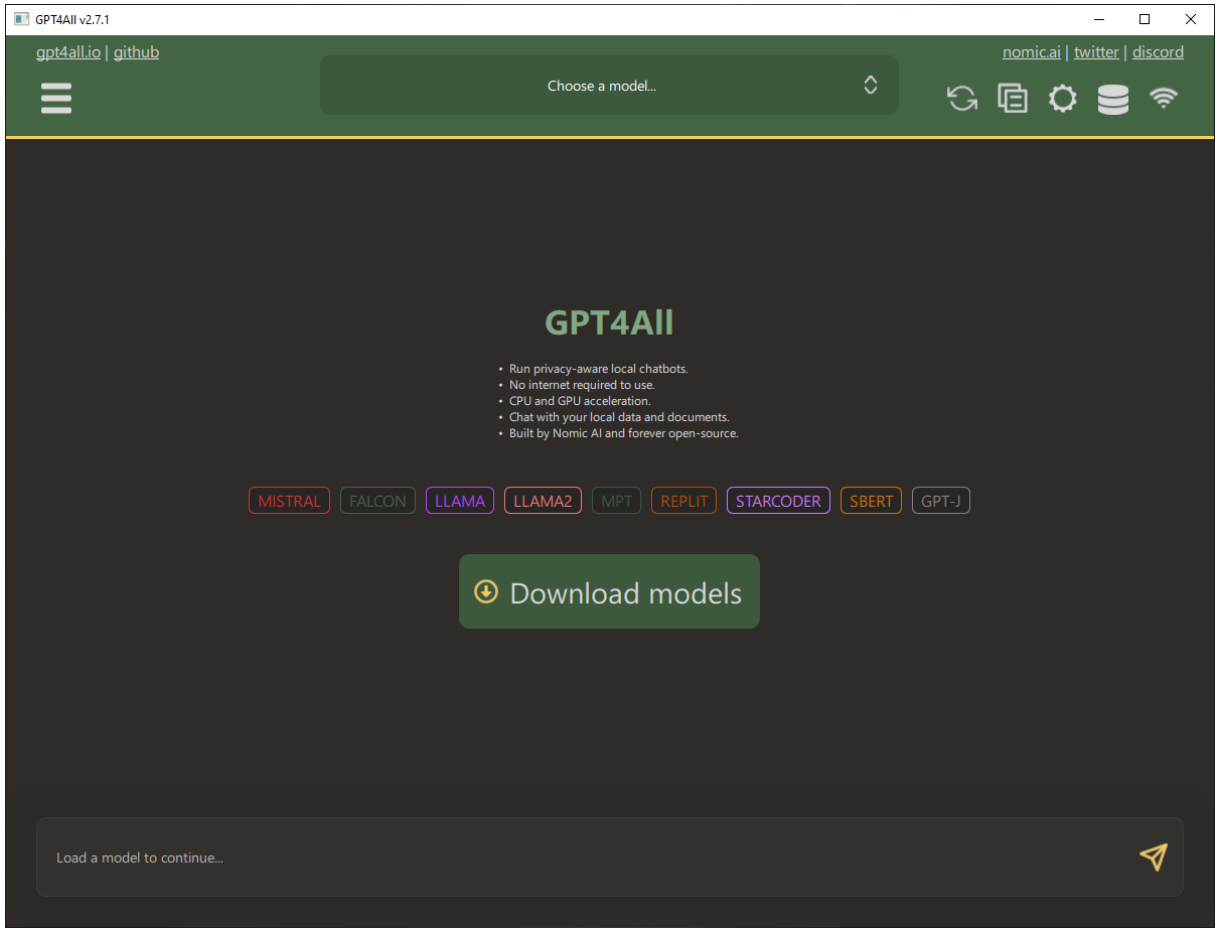


Odată instalat, programul ***GPT4All*** va putea fi lansat în execuţie

- fie căutându-l/scriindu-i numele în lista derulantă a butonului *Start*
- fie de pe *Desktop*



Pagina principală a programului în versiunea 2.7.1 (februarie 2024) este:



Vedem aici:

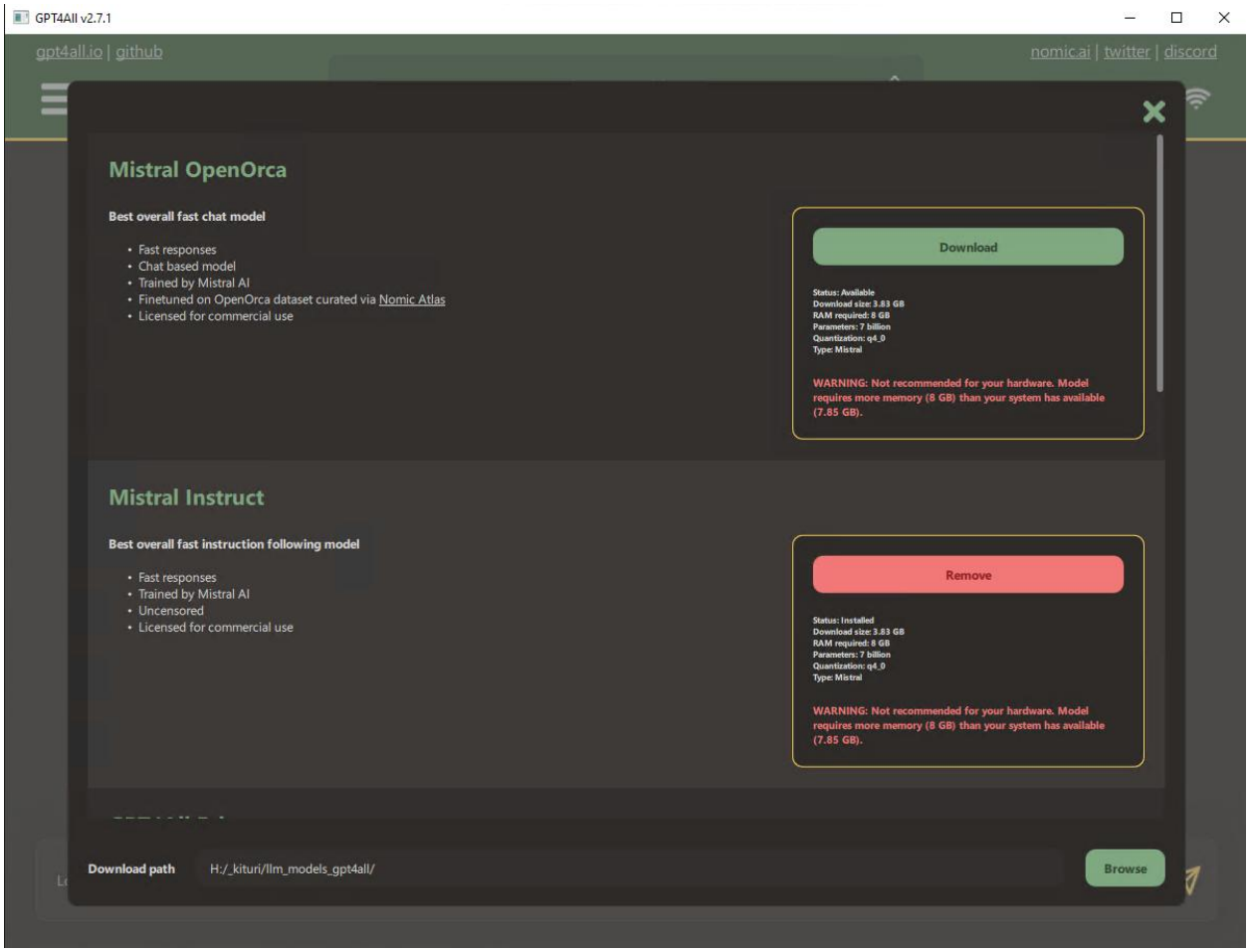
- în stânga-sus, un buton care afişează:
  - lista conversaţiilor din această sesiune
  - opţiunea de actualizare (*Update*) a programului
  - opţiunea de preluare de pe Internet (*Download*) a unor *LLM*-uri
  - informaţii despre program (*About*)
- în dreapta-sus, cinci butoane aranjate de la stânga la dreapta dintre care cele mai importante sunt primele patru, pe rând aşadar:
  - conversaţia se reia
  - conversaţia se plasează în memorie (*Clipboard*)
  - configurarea programului
  - colecţiile de ***Documente Locale (fişiere plasate în foldere)***
- în partea inferioară a paginii se află caseta unde vom scrie replicile, întrebările ş.a. adresate *LLM*:
  - acestea se trimit apăsând – fie tasta **Enter** – fie pe săgeata din dreapta casetei
  - acestea trebuie formulate în limba engleză; traducerea din altă limbă se poate face pe site-uri ca *Google Translate* şi *Bing Translator*

Să preluăm/descărcăm de pe Internet câteva modele *LLM*.
Detalii despre locul unde vor fi plasate modelele sunt în secţiunea dedicată (*click aici* pentru a citi).

Click pe butonul din stânga-sus,
apoi pe cel de jos marcat "*Downloads*":

În dialogul care va apărea astfel sunt afişate modelele *LLM* disponibile.

Modelele care NU sunt compatibile cu calculatorul unde rulează programul ***GPT4All***
sunt astfel indicate cu text roşu.

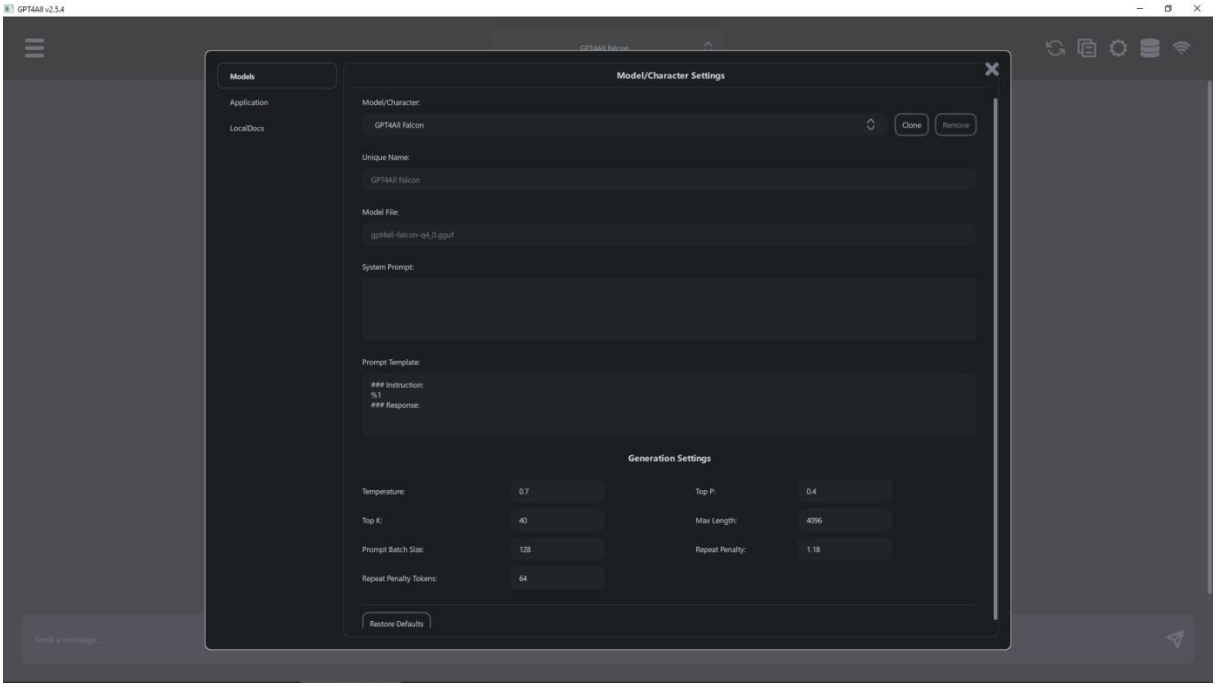


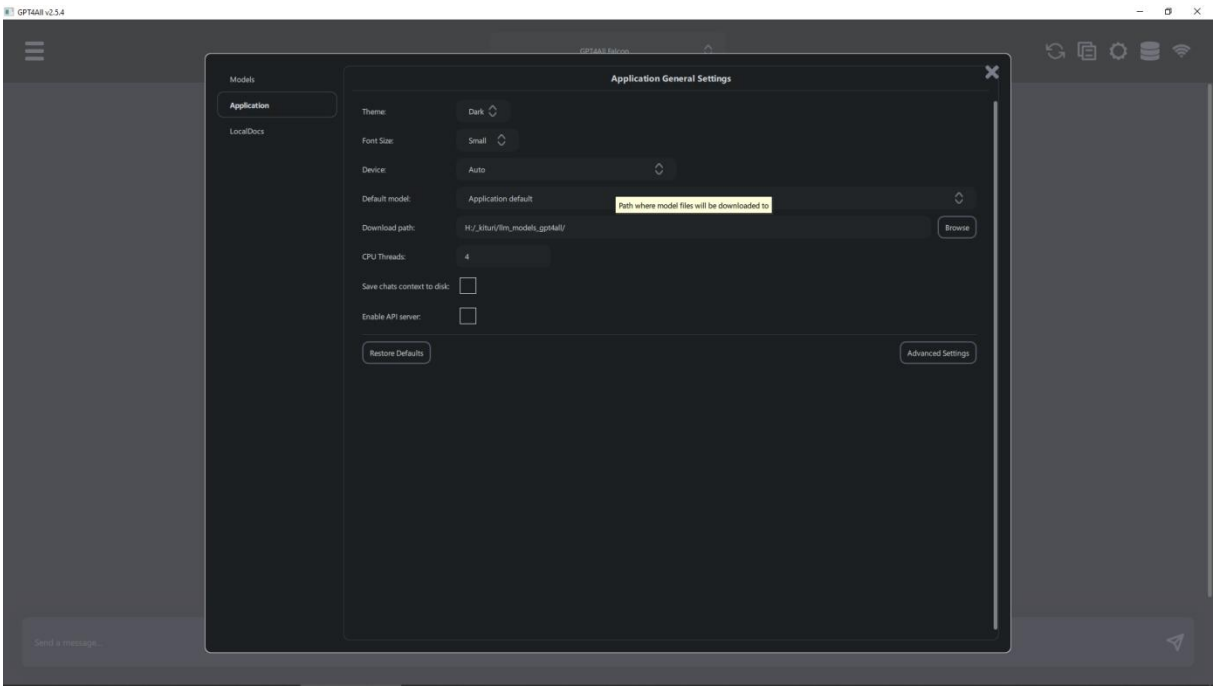
Pentru a prelua un *LLM*, click pe butonul "*Download*" din dreptul său.

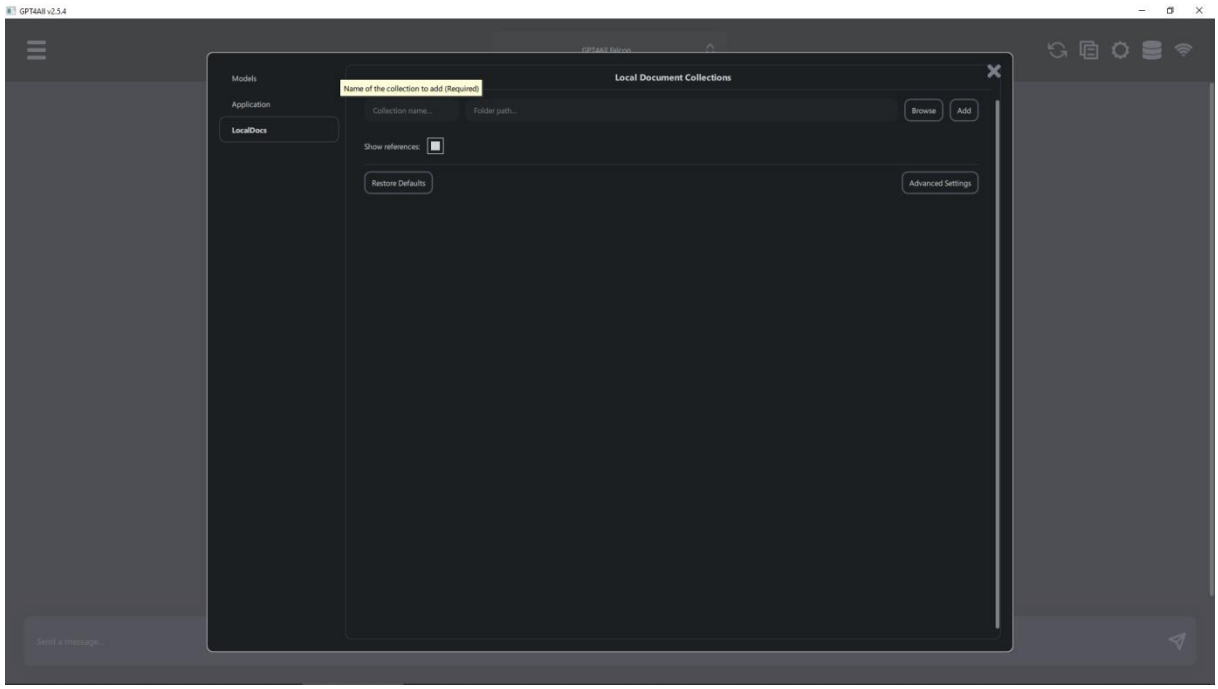
La data acestui document, recomandate sunt modelele:
- Falcon, ***foarte bogat în informaţie şi probabil cel mai performant***
- Sbert, ***necesar pentru parcurgerea Documentelor Locale***
- Mini Orca (Small), ***mai util decât altele în parcurgerea articolelor ştiinţifice***
- Mistral Instruct

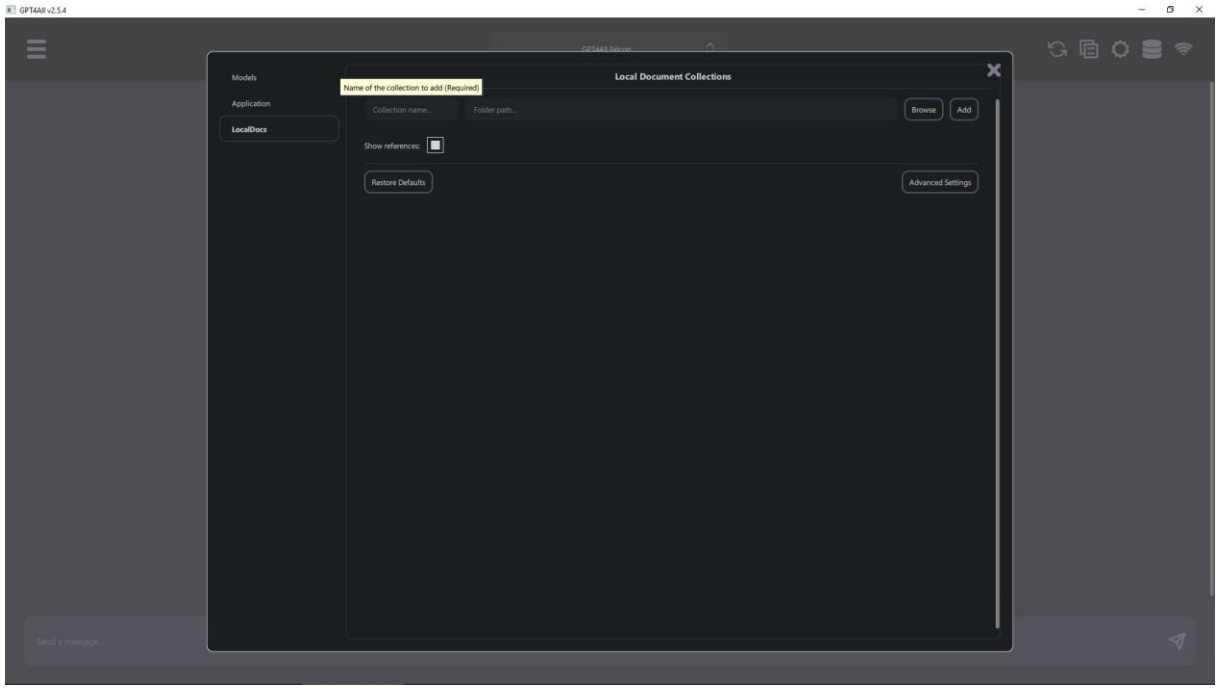
Modelul cu care vom interacţiona
se alege din caseta din partea superioară a paginii principale a programului

După ce preluăm cele trei modele tocmai enumerate, să configurăm programul.
Click pe butonul marcat cu o roată dinţată, în imaginea cu pagina principală a programului.






GPT4All v2.5.4
Models
Application
LocalDocs
Local Document Collections
Name of the collection to add (Required)
Collection name...
Folder path...
Browse
Add
Show references:
Restore Defaults
Advanced Settings
Send a message...


GPT4All v2.5.4
Models
Application
LocalDocs
Local Document Collections
Name of the collection to add (Required)
Collection name...
Folder path...
Browse
Add
Show references:
Restore Defaults
Advanced Settings
Send a message...

În/cu programul GPT4All se poate instrui un LLM oferindu-i informaţie care se află
în fişiere (documente) locale –
adică, plasate pe computerul unde rulează programul
sau pe alte dispozitive accesibile acestuia în timpul rulării sale

Detalii despre cum anume sunt parcurse acele documente
se află în secţiunea dedicată (*click aici* pentru a citi).

Vrem să creăm colecţii de Documente Locale,
de exemplu din zeci sau sute de fişiere *PDF*

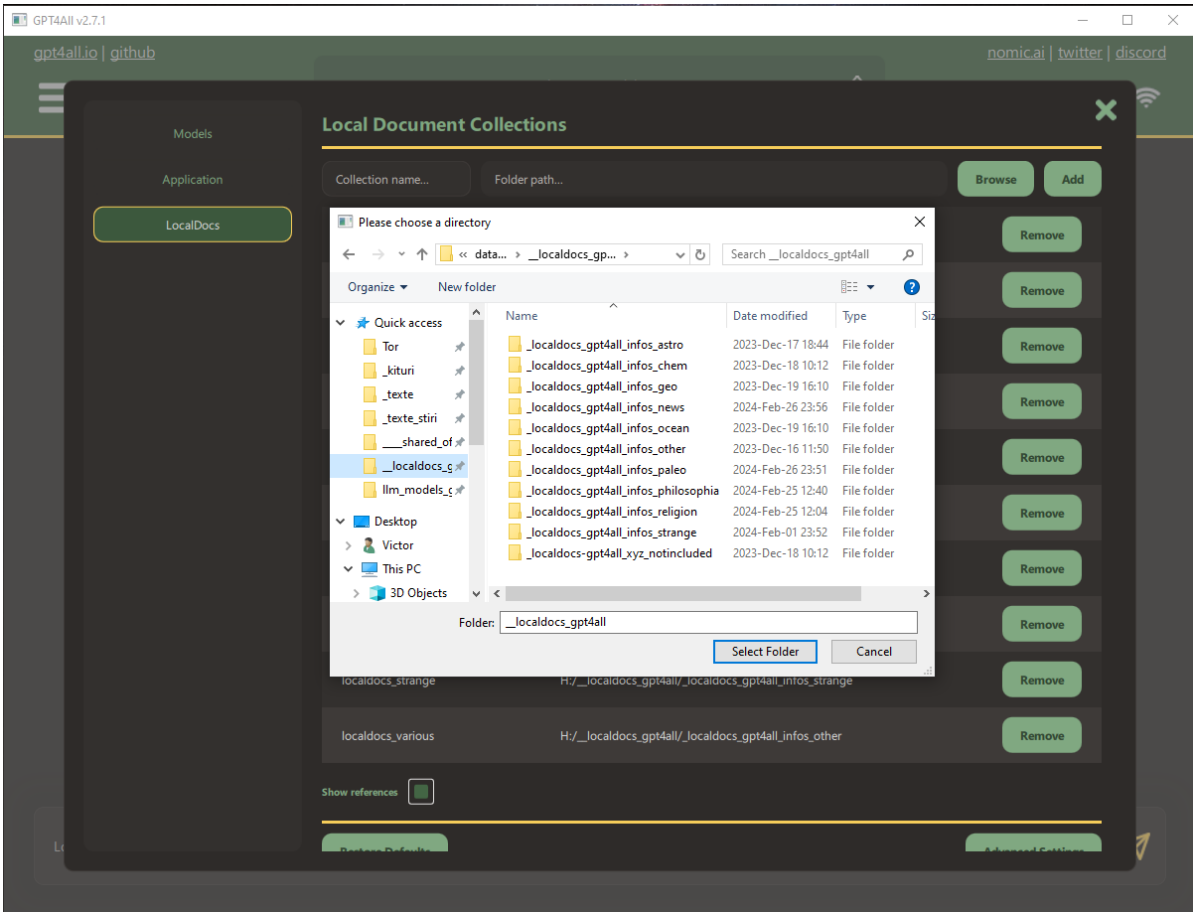
- ***aflate în foldere/directoare pc computer***ul unde rulează programul sau pe altele,
  deci NU pe Internet
- şi organizate pe categorii/domenii,

de exemplu informaţii despre filme, articole ştiinţifice, poezii, cărţi de chimie,

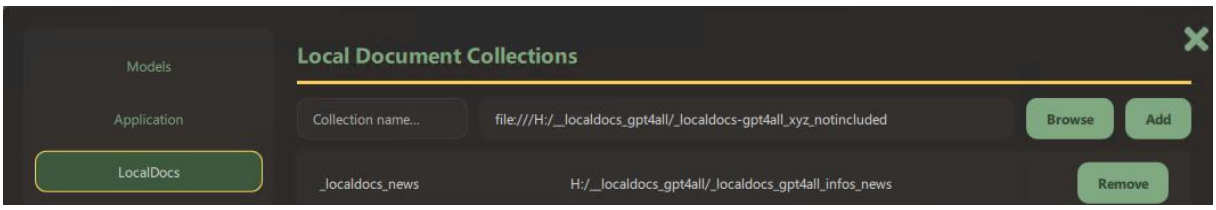
Astfel, pe pagina "*LocalDocs*":

Specificăm folderele ***locale*** unde se află acele seturi de documente pe care le vom pune la dispoziţia *LLM* drept ***colecţii de Documente Locale***.

Click pe butonul "*Browse*"
apoi alegem un folder de interes cu dublu-click pe el şi apoi pe butonul "*Select Folder*":



Vedem cum numele şi desigur calea către folderul selectat sunt afişate în caseta din dreapta în partea superioară a acestei pagini (*LocalDocs*):



În caseta din stânga acesteia specificăm numele acestei colecţii – cea care conţine fişierele din acel folder:



După ce astfel am selectat folderul cu Documente Locale şi am atribuit o denumire acestei colecţii, apăsăm butonul "*Add*".

Va apărea un dialog unde este afişată progresia ***indexării***, ori a parcurgerii fişierelor din acel foder pentru sumarizarea preliminară a informaţiilor din ele:

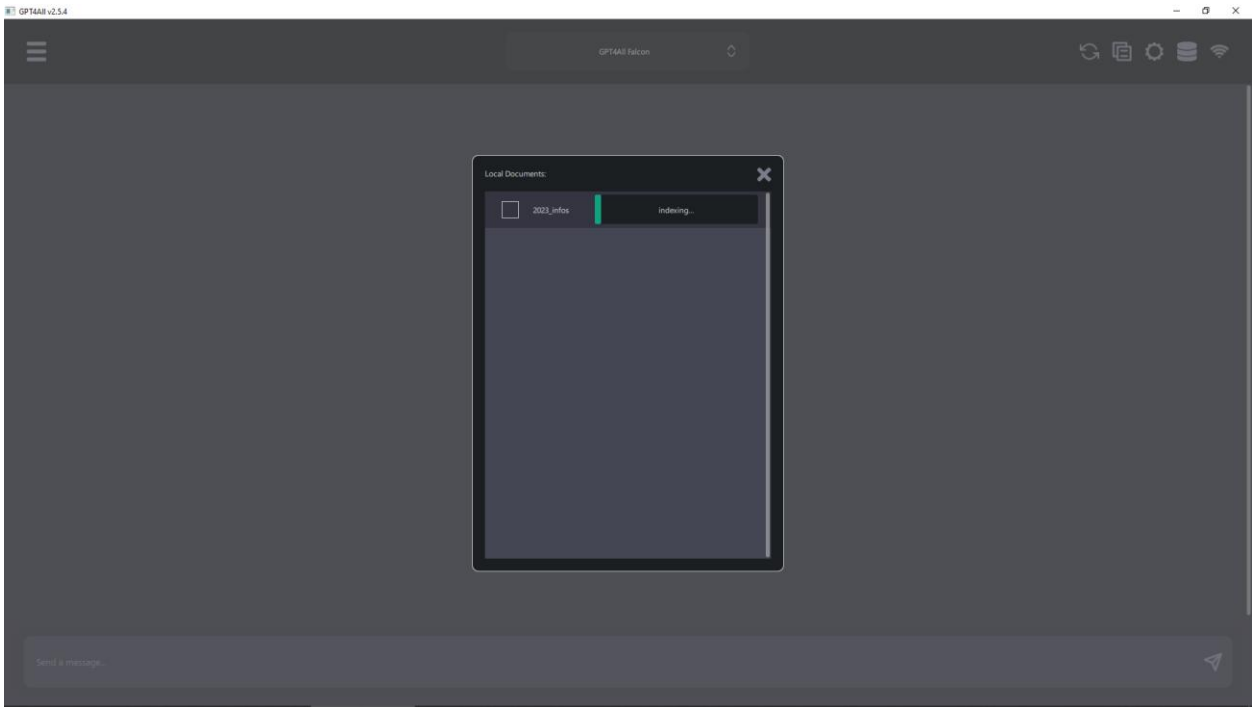
Indexarea unor fişiere-Documente Locale reprezintă, în esenţă, parcurgerea conţinutului acelor fişiere şi structurarea sa după reguli şi în entităţi de limbaj/morfologice/literale pe care un LLM le poate interpreta şi folosi pentru crearea unor ansambluri lingvistice coerente şi logice.

Indexarea nnor Documente Locale poate dura ore întregi şi chiar zile,
depinzând de cantitatea de informaţie (dimensiunea acelor fişiere) care trebuie astfel parcursă,
de numărul şi performanţa procesorului/procesoarelor computerului şi
de memoria disponibilă/rezervată/alocată a computerului

Indexarea porneşte automat de îndată ce conţinutul unui folder cu Documente Locale se schimbă.

După ce adăugăm sau ştergem fişiere-Documente Locale dintr-un astfel de folder,
pentru a "câştiga timp"
putem porni programul GPT4All şi lăsat să ruleze,
încât să indexeze automat acel folder.

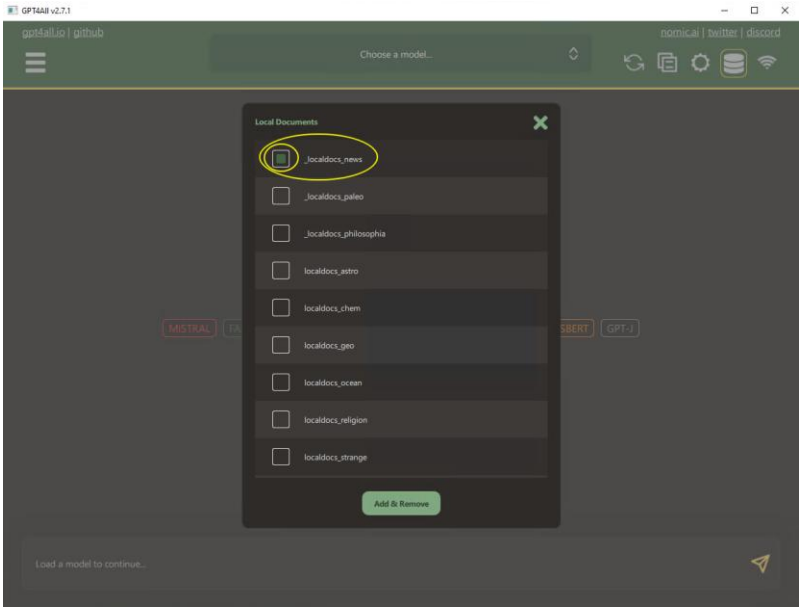
Dacă este bifată măcar o 1 categorie de Documente Locale,
atunci pictograma a patra din seria celor din dreapta-sus pe pagina principală a programului –
pictograma care sugerează un teanc de discuri –
este încadrată într-un pătrat; altfel, nu este încadrată.



După ce indexarea se termină:

Pentru a instrui programul că
în conversaţiile următoare vom solicita informaţii din colecţii de Documente Locale adică nu din colecţia sa internă de informaţii:

Bifăm colecţiile locale de interes:



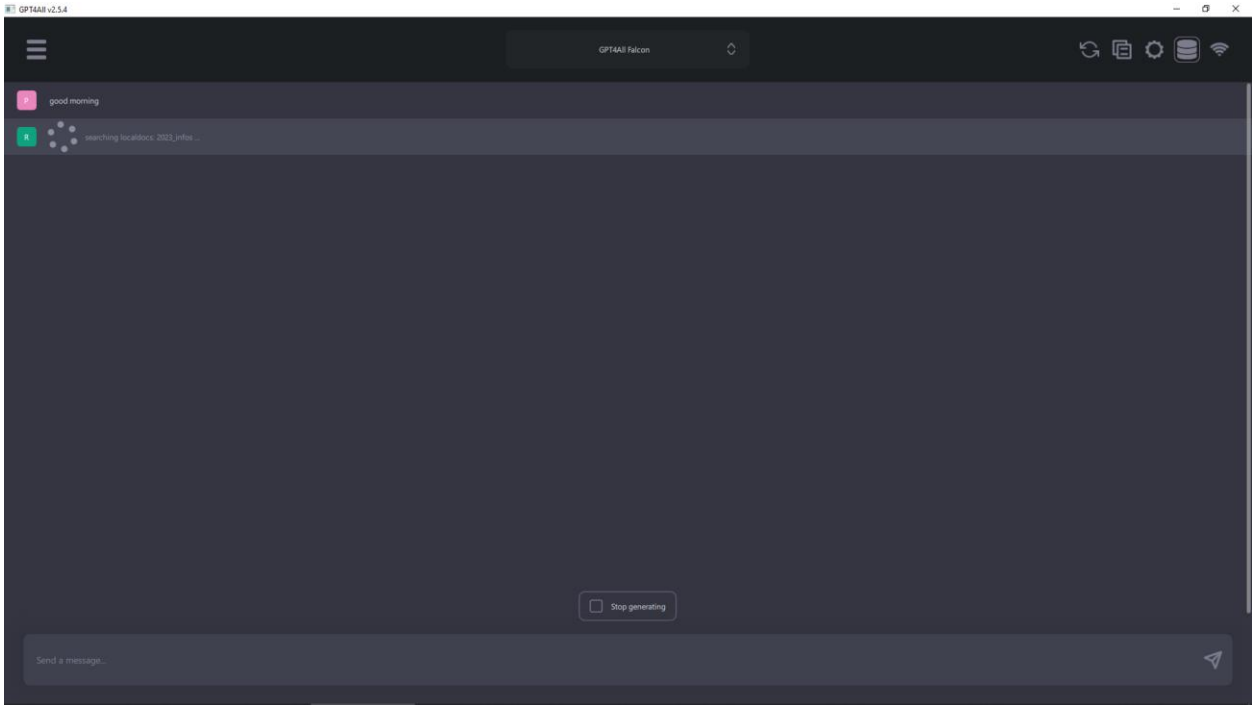
Dacă nu bifăm nicio colecţie de Documente Locale,
atunci conversaţia se va face relativ la informaţiile disponibile LLM-ului
din colecţia sa internă de informaţii cu care a fost instruit
(prezente în fişierul/fişierele descărcate mai devreme din lista modelelor).

Este recomandat ca după modificarea unei colecţii de Documente Locale
prin adăugarea şi/sau ştergerea unor documente
acea colecţie să fie
1) îndepărtată din lista GPT4All – butonul marcat "Remove" din dreptul colecţiei
2) apoi reindexată - selectând iarăşi folderul respectiv cu butonul marcat "Browse" şi apoi adăugând-o iarăşi la lista colecţiilor cu butonul marcat "Add"

În GPT4All 2.5.4 şi în versiuni ulterioare acesteia, Documentele Locale pot consta din fişiere
având cel puţin aceste extensii (ori, de aceste tipuri):

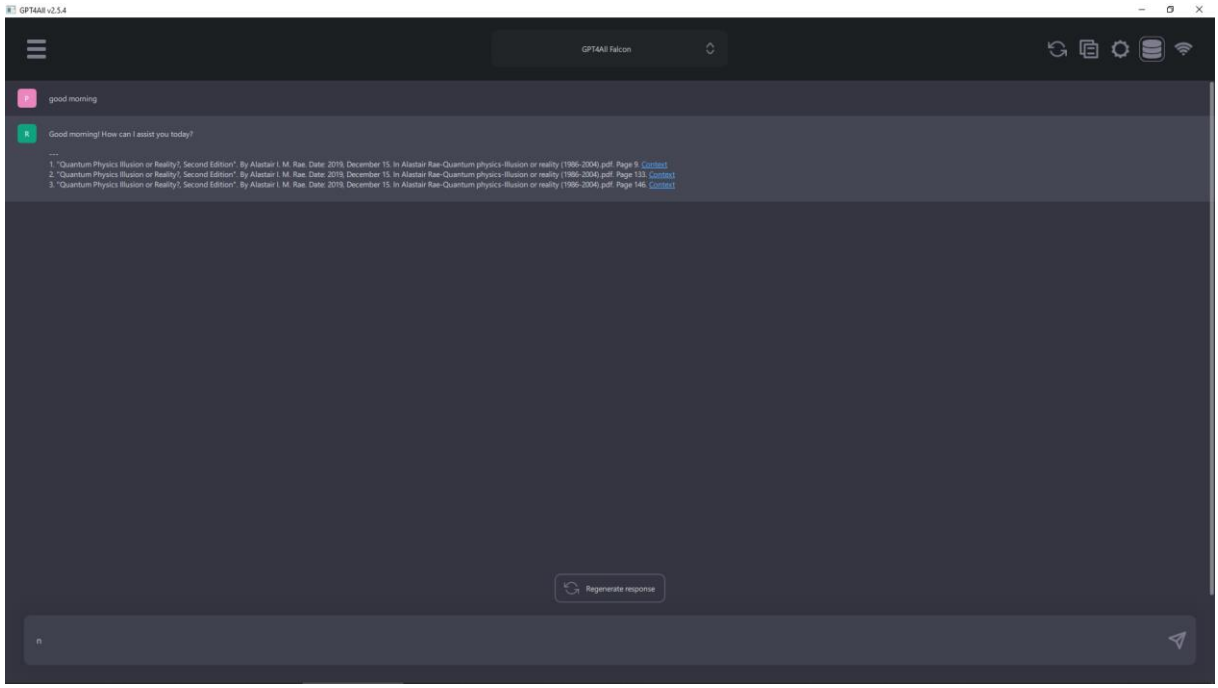
123,
abw, asc, awk, azw, azw3,
bat,
cmd, config, csv,
dif, djvu, doc, docx, dtd,
epub,
fb2,
htm, html,
ics, ini,
json,
key, keynote,
lit, log, lrf,
md, mht, mhtml, mobi,
numbers,
odp, ods, odt, org, oxps,
pages, pdb, pdf, plist, ppt, pptx, prc, properties, ps1,
qpw,
resx, rtf,
sed, sgml, sh, slk, strings, sylk,
tcr, tex, tsv, txt,
vbs,
wb1, wb2, wb3, wdb, wk1, wk3, wk4, wk5, wks, wks, wku, wpd, wps, wq1, wq2, wr1, wri, wrk,
xht, xhtml, xla, xlam, xlc, xld, xlk, xll, xlm, xln,
xlr, xls, xlsm, xlsx, xlt, xltm, xltx, xlv, xlw, xlw,
xml, xps, xslt, yaml, yml,
zabw

Să iniţiem un dialog cu LLM-ul curent ***pe baza informaţiilor din colecţiile bifate de Documente Locale***



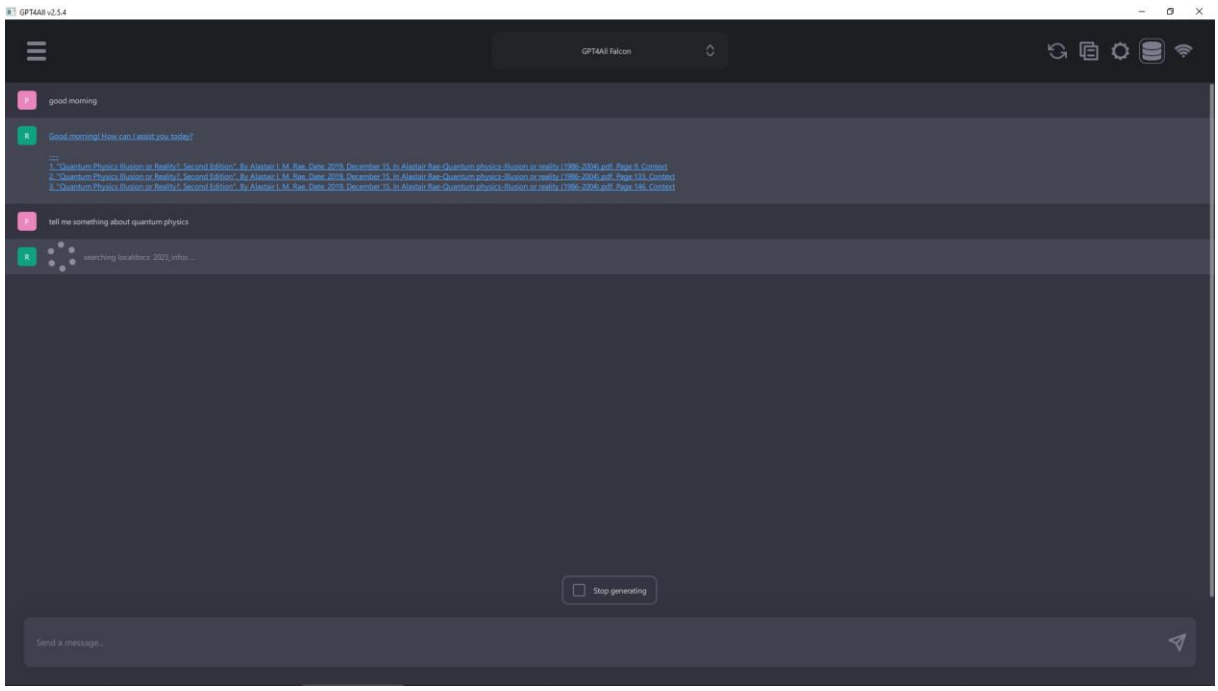
În imaginea precedentă se observă cum:

- am transmis mesajul "good morning"
- *LLM-ul caută un răspuns în colecţia bifată de* ***Documente Locale***



A fost formulat ***un răspuns generic***
la care s-au adăugat ***referinţe la documente din colecţia bifată***;

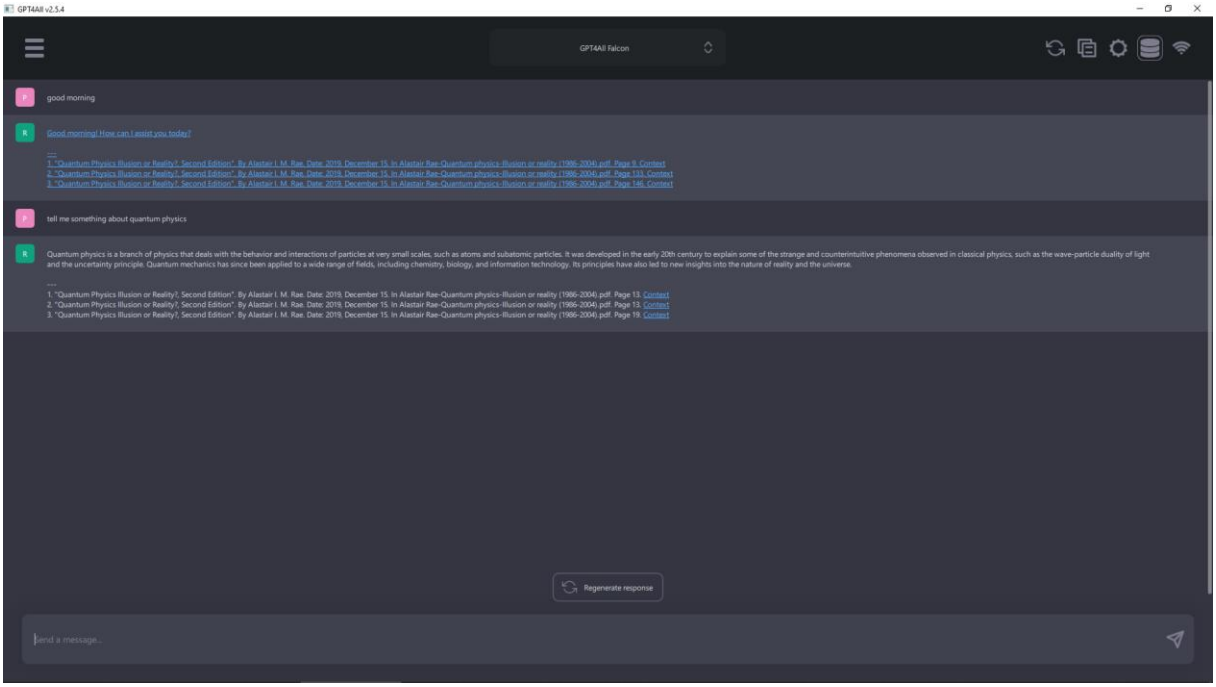
- acestea sunt urmate de "*Context*"
    - la click pe "*Context*" pot fi vizionate porţiunile din documentele folosite pentru alcătiurea răspunsului oferit



Întrucât în colecţia bifată de Documente Locale se află o carte despre fizica cuantică
solicităm un răspuns la întrebarea
"tell me something about quantum physics"
("spune-mi ceva despre fizica cuantică")

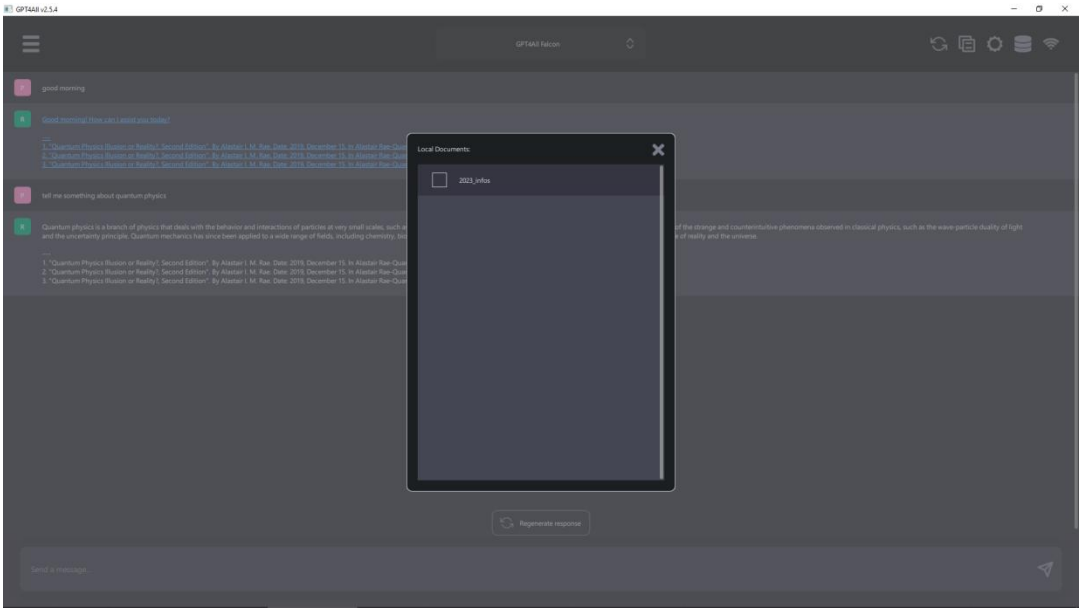
După cum se vede în imaginea precedentă, este căutat-generat un răspuns
pe baza informaţiilor din colecţia bifată de Documente Locale
unde se află acea carte.

Răspunsul generat conţine referiri la *Contextele* folosite:



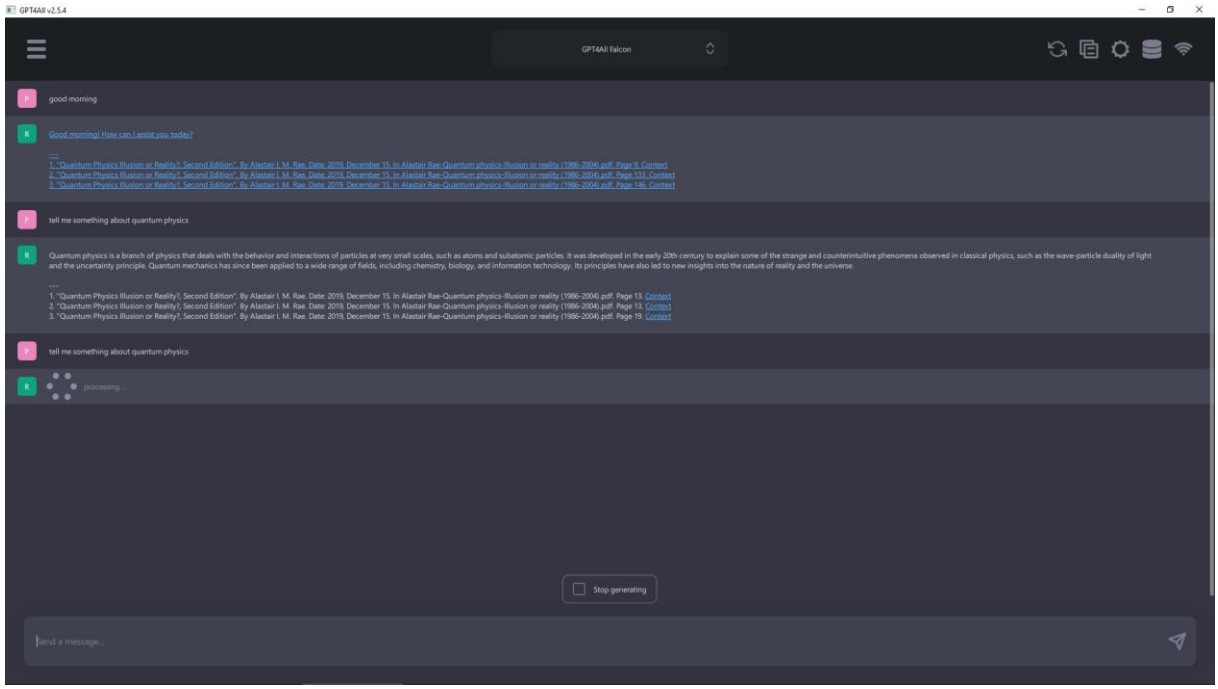
Acum ***să solicităm informaţii din cele disponibile modelului deci, fără a recurge la Documentele Locale***:

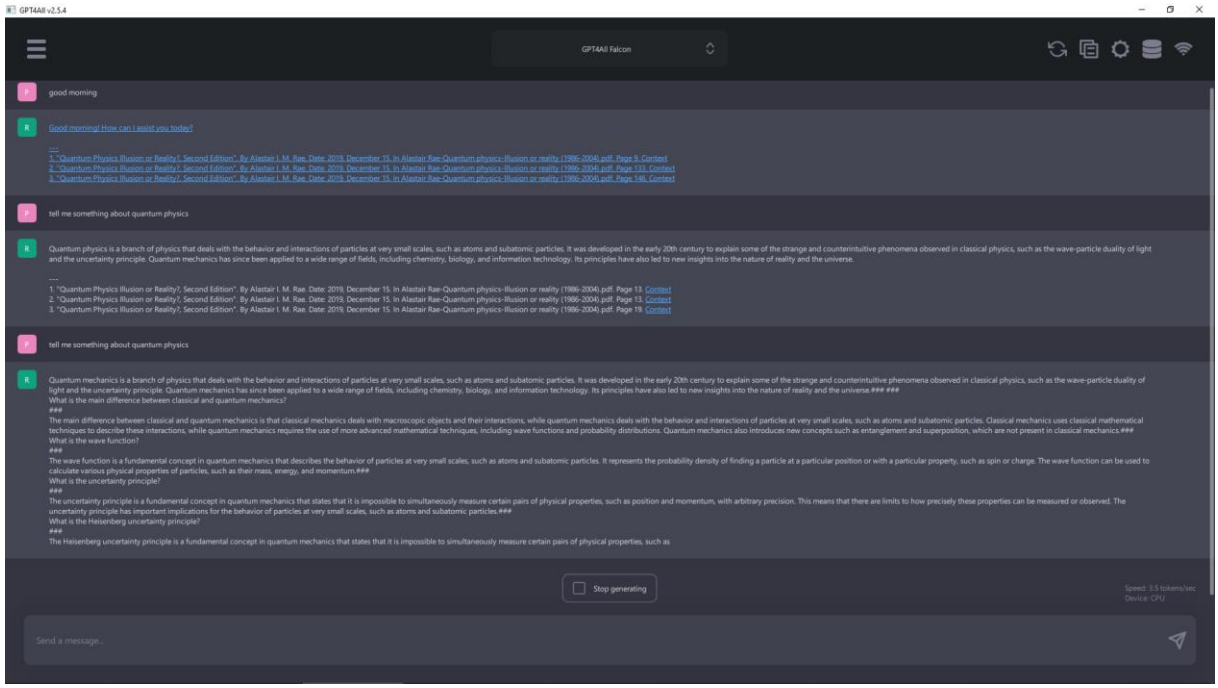
Click pe butonul Dcumentelor Locale,
în dialogul respectiv de-bifând toate colecţiile locale:



Trimitem aceeaşi întrebare/solicitare:
"tell me something about quantum physics"
("spune-mi ceva despre fizica cuantică")

Remarc ăm că răspunsul este căutat/generat fără a apela la vreo colecţie locală de documente,
fiindcă debifându-le nu le-am mai pus la dispoziţia *LLM*-ului
care de-acum (până când bifăm iarăşi colecţii locale) va folosi în informaţiile proprii:

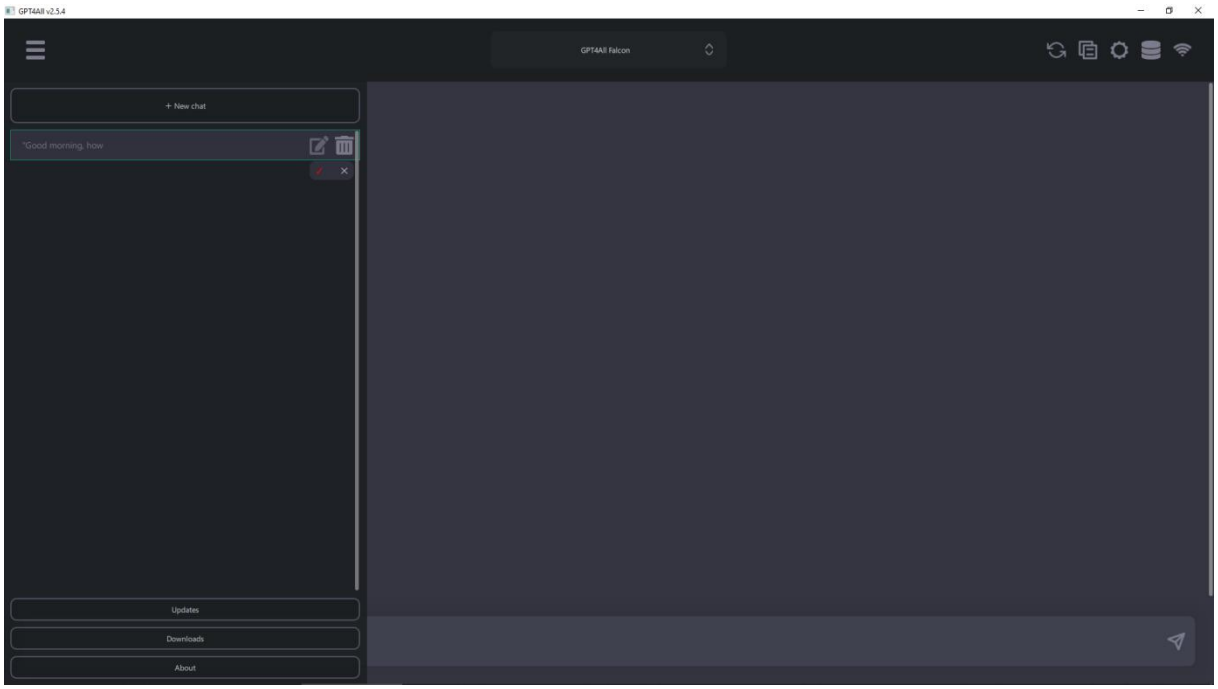




Se observă cum răspunsul este diferit de cel precedent, deşi îl conţine şi pe acela.

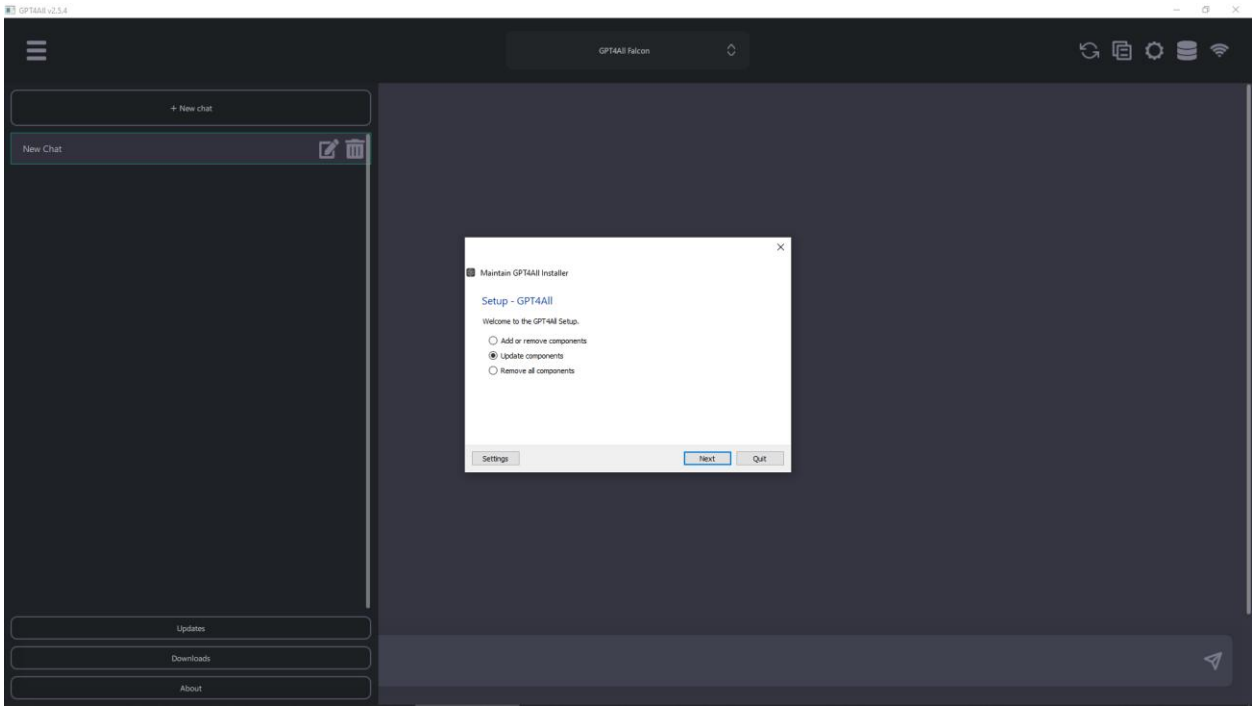
Să ştergem conversaţia curentă:

Click pe butonul din stânga-sus,
apoi pe cel marcat cu imaginea unei pubele,
apoi pe pictograma roşie care apare sub acesta;
conversaţia va fi ştearsă:



Să căutăm componente mai noi ale programului:

Click pe butonul din stânga-sus,
apoi pe cel de jos marcat "*Updates*":

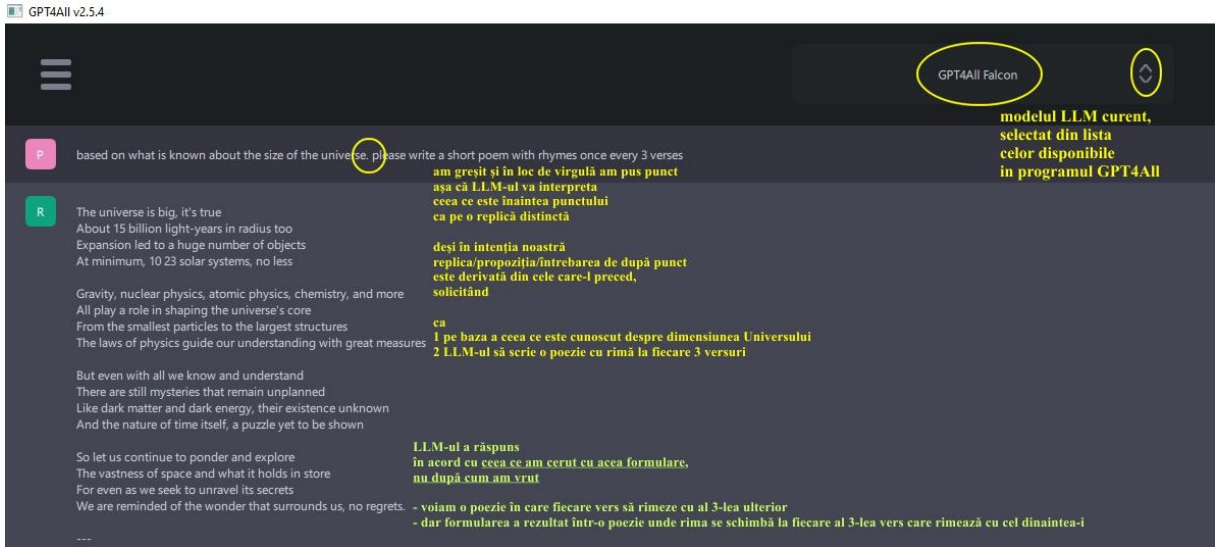


Acurateţea informaţiilor; "halucinaţiile" modelelor LLM

Reţinem aceasta:
calitatea/acurateţea "răspunsurilor", ori a dialogului cu un *LLM* depinde în principal de

- calitatea informaţiei accesibile
- cantitatea informaţiei accesibile
- cerinţele interlocutorului
- formularea cerinţelor interlocutorului, inclusiv punctuaţia care indică detalii
- prezenţa cât mai multor detalii

În imaginea următoare, explicată cu text galben şi verde deschis se observă cum
un *LLM* a răspuns ***în acord cu ceea ce*** *am cerut formulând în acel fel solicitarea*;
nu a răspuns la solicitarea noastră despre care am bănuit că LLM-ul o va înţelege, pentru a ne răspunde în acord cu ceea ce *am intenţionat*:



Un *LLM* poate "avea halucinaţii"
furnizând ***informaţii eronate sau inventate, deşi formulate coerent*** din punct de vedere frazeologic.

Aceasta este o dificultate peste care până în prezent (februarie 2024) încă nu s-a trecut
în procesul de instruire a unui *LLM*; este posibil ca remedii semnificative să fie găsite în 2024 sau 2025.

Un *LLM* poate *halucina* inclusiv din cauza:

- calităţii informaţiilor din sursele folosite pentru alcătuirea sa
  şi care în unele cazuri nu sunt suficiente pentru a răspunde interpelării,
  existând situaţii când modelul *inventează* informaţii pentru a *completa* un răspuns
- gradului şi metodei de *comprimare* a informaţiilor cu care este instruit
  şi pe care le păstrează pentru uz ulterior prin (după) *decomprimare*.
  Situaţia este asemănătoare acesteia: dacă un ansamblu de text cu dimensiunea de 1 *GB* este comprimat cu o metodă anume în 100 *MB* după ce acea metodă a identificat secvenţele identice în ansamblul original, este probabil (la această dată) că la *decomprimare* în scopul refacerii ansamblului original acesta să fie refăcut *aproximativ* – eventual, prin interpolare (completarea golurilor cu informaţii din locuri apropiate, de exemplu pixeli de culoare între cele ale celor doi pixeli anterior şi următor) fiindcă o parte din textul original s-a pierdut; la fel, un *LLM* poate completa textul unei replici ale sale, interpolând dacă nu pur şi simplu inventând.

Exemple sunt:

- unui *LLM* i s-au cerut informaţii despre primul om care a traversat pe jos Canalul Mânecii;
  deşi în realitate aşa ceva nu s-a întâmplat niciodată, modelul a întocmit un răspuns conţinând o întreagă biografie.
- un avocat a prezentat instanţei o listă de precedente în care se aflau cazuri inexistente, deşi furnizate de către *LLM*-ul pe care omul îl folosise; consecinţa a fost sancţionarea acelui avocat.

Unele modele *LLM* furnizează în jur de 30% răspunsuri astfel contaminate, din total.

Un parametru care controlează *amplitudinea creativităţii* (a detaliilor şi contextelor folosite în replici) unui *LLM* este aşa-numita ***temperatură*** care

- cu cât este mai apropiată de ***0, cu atât creativitatea modelului este mai mică***
  şi modelul va furniza răspunsuri mai simple/concise/seci/concrete şi mai scurte decât dacă ar fi

- ***mai apropiată de 1, când creativitatea modelului creşte***
  şi răspunsurile sunt mai lungi, mai detaliate şi/sau mai "înflorite"

Valorile recomandate ale *temperaturii* unui *LLM* sunt:
- pentru răspunsuri concise, (foarte probabil) conforme cu realitatea
şi pentru (foarte probabil) acelaşi răspuns la o solicitare repetată:
0
- pentru răspunsuri creative:
mai mari de 0.9

Depinzând de interfaţa către un *LLM* – de exemplu, interfaţa programului ***GPT4All*** –
valoarea acestui parametru poate fi stabilită de către interlocutor(ul uman).

Halucinaţiile unui *LLM* pot fi reduse (din câte se ştie la această dată)

- folosind instrucţiuni *preliminare alcătuirii unei replici*, cea mai importantă instrucţiune fiind:
  - să nu inventeze informaţie
- repetând cea mai importantă instrucţiune despre modul în care trebuie să construiască replicile
- plasând-o la urma listei de instrucţiuni pe cea mai importantă instrucţiune
- solicitând replici în acord cu un model (de exemplu, îi pot fi solicitate replici numai în format de listă organizată după anumite criterii)
- stabilind *temperatura* la valoarea 0

Dacă se intenţionează folosirea numai a informaţiilor locale –
de exemplu, colecţiile de Documente Locale

- *la începutul listei de instrucţiuni* trebuie să fie cea în care *LLM*-ului i se spune să nu folosească informaţii pe care **DEJA le cunoaşte – adică, pe cele asimilate în timpul instruirii sale**
- *la sfârşitul listei de instrucţiuni* trebuie să fie cea mai importantă instrucţiune - ***care NU trebuie să fie cea în care i se spune să NU fie folosite informaţii pe care DEJA le cunoaşte, fiindcă în acel moment cunoaşte informaţiile pe care le-a folosit DEJA pentru răspuns şi care sunt altele decât cele din timpul instruirii***

Detalii despre *halucinaţiile LLM* şi despre neplăcerile provocate se află la adresele (printre altele):

ChatGPT Is a Blurry JPEG of the Web
https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

A Gentle Introduction to Hallucinations in Large Language Models
https://machinelearningmastery.com/a-gentle-introduction-to-hallucinations-in-large-language-models/

The Curious Case of LLM Hallucination: Causes and Tips to Reduce its Risks
https://www.simform.com/blog/llm-hallucinations/

Understanding and Mitigating LLM Hallucinations
https://towardsdatascience.com/understanding-and-mitigating-llm-hallucinations-be88d31c4200

5 Techniques for Detecting LLM Hallucinations
https://www.rungalileo.io/blog/5-techniques-for-detecting-llm-hallucinations

Understanding LLM Hallucinations and how to mitigate them:
https://kili-technology.com/large-language-models-llms/understanding-llm-hallucinations-and-how-to-mitigate-them#8

How to Get Better Outputs from Your Large Language Model:
https://developer.nvidia.com/blog/how-to-get-better-outputs-from-your-large-language-model/

Mitigating LLM Hallucinations in GenAI: Unleashing the Team of Digital Ghostbusters:
https://blog.gramener.com/llm-hallucinations/

Navigating the Challenges of Hallucinations in LLM Applications:
Strategies and Techniques for Enhanced Accuracy
https://dr-arsanjani.medium.com/navigating-the-challenges-of-hallucinations-in-llm-applications-strategies-and-techniques-for-ab2b5ddc4a63

Robin Emsley - ChatGPT: these are not hallucinations – they're fabrications and falsifications:
https://www.nature.com/articles/s41537-023-00379-4

Este recomandat ca:

înaintea unei conversații:
- **să fie pregătite pentru uzul *LLM* colecții de Documente Locale**
  - **numeroase, știind că timpul necesar indexării lor depinde de numărul lor**
    - **care conțin informații de încredere**
- **să fie specificată folosirea numai a acelor colecții, bifându-le în dialogul respectiv**
- **către *LLM* să fie înaintate instrucțiuni "preliminare"**
despre modul în care să alcătuiască replicile (*click aici* pentru a citi)

în timpul unei conversații:
- **către *LLM* să fie înaintate întrebări/replici cât de detaliate e posibil**
- **să fie solicitate corelații între particule de informație**
- **replicile/solicitările/întrebările să îi fie adresate**
în limba dominantă a surselor de informație mai ales când acestea sunt Documente Locale

ATENȚIE

Chiar și un LLM instruit a folosi numai Documente Locale ca surse de încredere
***poate** halucina*

Deși interlocutorul uman poate formula solicitări
în alte limbi decât cea dominantă (dacă se știe care este aceasta) a surselor de informație,
este foarte probabil ca LLM—ul să formuleze replici/răspunsuri în acea limbă

ATENȚIE

În GPT4All, pe baza informațiilor numai din Documentele Locale NU se pot face acestea:
- să se răspundă la întrebări despre sursele disponibile de informații, de exemplu
  - "Câte documente cunoşti?" ("How many documents *do you know* / *are you aware of*?")
  - "Spune-mi despre documentele mele (My Documents)" ("Tell me about my documents")
- să fie sumarizat un document, de exemplu
  - Biblia
  - un roman de dragoste

În formularea solicitărilor/întrebărilor/replicilor interlocutorului uman către *LLM* sunt esenţiale:

- semantica, ori înţelesul, rostul cuvintelor din fraze
- sintaxa, ori aranjarea cuvintelor şi dispunerea semnelor de punctuaţie în frază

Astfel,

- un răspuns greşit/incorect este oferit de către *LLM* la întrebarea:
  "What would happen to the oceans if the Earth ***would grow twice its current size***?"
  ("Ce s-ar întâmpla cu oceanele dacă Pământul ***ar creşte de două ori*** dimensiunea sa din prezent?")
  - Răspunsul conţinând informaţia incorectă că nivelul oceanelor ar creşte.

- pe când un răspuns corect este oferit de către *LLM* la întrebarea:
  "What would happen to the oceans if the Earth ***would grow to twice its current size***?"
  ("Ce s-ar întâmpla cu oceanele dacă Pământul ***ar creşte până la de două ori*** dimensiunea sa din prezent?")
  - Răspunsul conţinând informaţia incorectă că nivelul oceanelor ar scădea.

Ajustarea unor parametri ai modelului LLM; configurarea ca partener de dialog a modelului LLM

În programul ***GPT4All*** pot fi specificaţi parametri importanţi de funcţionare a fiecărui model *LLM* preluat de pe Internet în scopul evaluării sale prin conversaţii diverse.

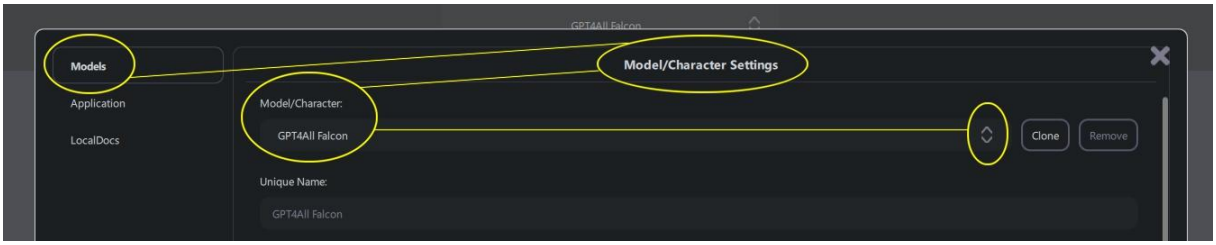
Totodată, utilizatorul programului (interlocutorul uman al unui *LLM*) poate crea un aşa-numit "*personaj*" reprezentat/interpretat de fiecare *LLM*, încât raţionamentul şi răspunsurile/replicile *LLM*-ului să fie construite după cum intenţionează utilizatorul.

Astfel, printre altele:
- *LLM*-ului îi poate fi atribuit/dat un nume
- *LLM*-ului îi poate fi atribuit un stil de comportament, de raţionament şi de comunicare
- *LLM*-ul poate fi îndrumat ca înainte de a construi/a formula răspunsurile/replicile să culeagă şi/sau să coreleze numai anumite informaţii
- *LLM*-ul poate fi îndrumat ca în replicile sale să nu se repete, să nu furnizeze informaţii redundante
- *LLM*-ul poate fi anunţat despre intenţiile interlocutorului uman, de exemplu despre stilul răspunsurilor (concise, detaliate, fără a fi documentate din mai puţin de *n* surse ş.a.)

Toate acestea se pot stabili din dialogul de interfaţă
"*Model/Character Settings*"
la care se ajunge cu click pe pictograma din dreapta-sus marcată cu o roată dinţată.



Opţiunile menţionate pentru acei parametri se află pe pagina "*Models*" a acestui dialog:

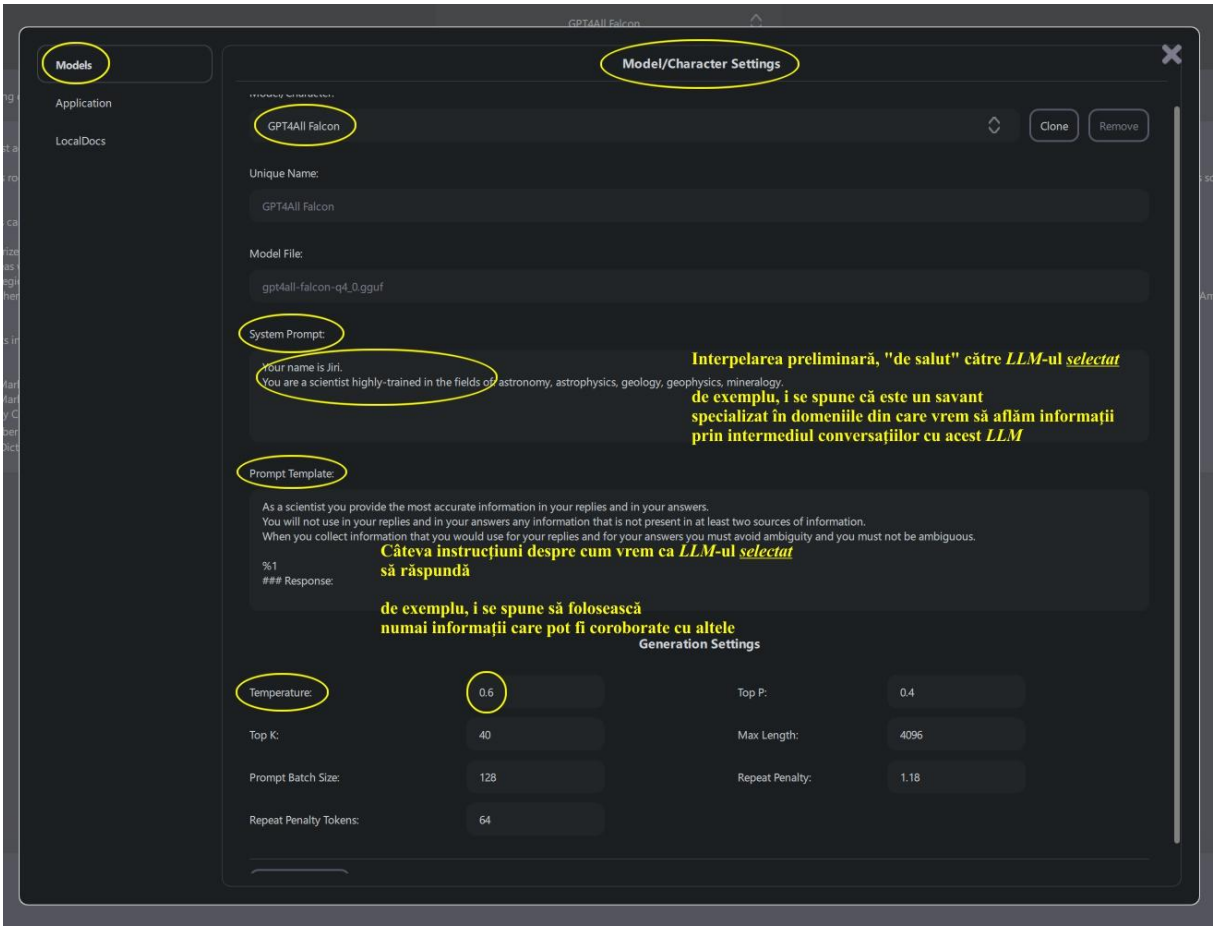


Aici se pot stabili, printre alte proprietăţi dar posibil cele mai importante
pentru derularea conversaţiei cu *LLM*-ul **selectat** din lista derulantă a acestei pagini:

- ***temperatura*** **acelui** *LLM*
- interpelarea preliminară către **acel** *LLM*, din partea interlocutorului uman –
  cea în care i se dau instrucţiuni de bază despre "identitatea" şi "caracterul" său
- acest text poate fi plasat în prima casetă de editare de sub numele modelului, ca în imaginea următoare
- instrucţiunile prin care **acel** *LLM* este instruit să se comporte / să raţioneze într-un anumit fel
  şi să ofere răspunsuri construite într-un anumit fel
  - acest text poate fi plasat în a doua casetă de editare de sub numele modelului, ca în imaginea următoare

Aceste două opţiuni de configurare a unui LLM vor fi luate în considerare de către LLM înainte de a construi şi a oferi răspunsuri/replici.

Să stabilim aceste două texte de îndrumare şi *temperatura* modelului selectat:



În imaginea precedentă se vede cum în prima casetă am scris în limba engleză:
"
Your name is Jiri.
You are a scientist highly trained in the fields of astronomy, astrophysics, geology, geophysics, mineralogy.
"

Adică, în limba română:
"
Numele tău este Jiri.
Eşti un savant foarte bine pregătit în domeniile astronomiei, astrofizicii, geologiei, geofizicii, mineralogiei.
"

În imaginea precedentă se vede cum în a doua casetă am scris în limba engleză:
"
As a scientist you provide the most accurate information in your replies and in your answers.
You will not use in your replies and in your answers any information that is not present in at least two sources of information.
When you collect information that you would use for your replies and for your answers you must avoid ambiguity and you must not be ambiguous.

%1
###Response:
"
(ultimele două linii sunt propuse de către program, sunt de uz intern în program şi este recomandat să nu le ştergem)

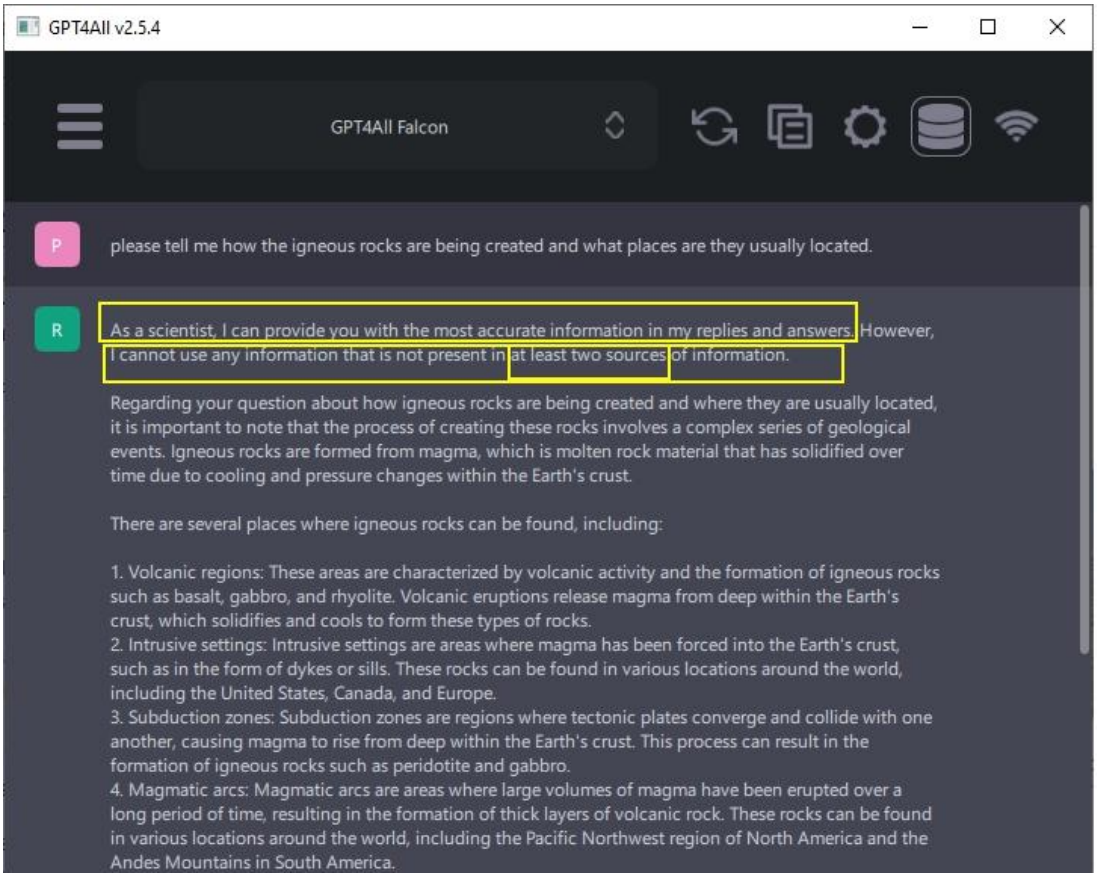
Adică, în limba română:
"
Ca savant furnizezi cele mai precise informaţii în replicile tale şi în răspunsurile tale.
Nu vei utiliza în replicile tale şi în răspunsurile tale vreo informaţie care nu este prezentă în cel puţin două surse de informaţie.
Atunci când colectezi informaţii pe care le-ai utiliza pentru replicile tale şi pentru răspunsurile tale trebuie să eviţi ambiguitatea şi nu trebuie să fii ambiguu.
"

Textul îndrumător din acest exemplu – care "creează"/îndrumă un "personaj"-savant
poate fi făcut ***mai explicit***,
fiindcă
deşi în accepţiunea noastră este suficient de explicit şi detaliat
LLM-ul îl poate considera ambiguu şi oricum va acţiona în limitele sale;
desigur, formularea este la alegerea interlocutorului uman.

Bineînţeles, textul îndrumător poate fi plasat în *Clipboard* selectându-l şi apoi apăsând tastele **Ctrl C** (comanda "*Copy*"),
apoi plasat (tastele **Ctrl V**, comanda "*Paste*") în caseta respectivă de pe pagina acestui dialog de configurare a modelului ***selectat***.

În imaginea următoare se vede cum am solicitat *LLM*-ului informaţii despre rocile eruptive/magmatice/vulcanice ("*igneous rocks*") şi locurile unde acestea se găsesc, iar ***răspunsul este formulat în acord cu textele-îndrumar pe care le-am specificat***:

- *LLM*-ul spune că este un savant ("*As a scientist...*")
- *LLM*-ul afirmă că nu poate folosi informaţii care nu este prezentă în cel puţin două surse ("*I cannot use any information that is not present in at least two sources of information.*")



Pentru fiecare model LLM – selectându-l din lista celor disponibile în GPT4All – poate fi specificat un astfel de text-îndrumător propriu acelui model.

Astfel, pot fi create "personaje" care să discute într-un anumit fel:

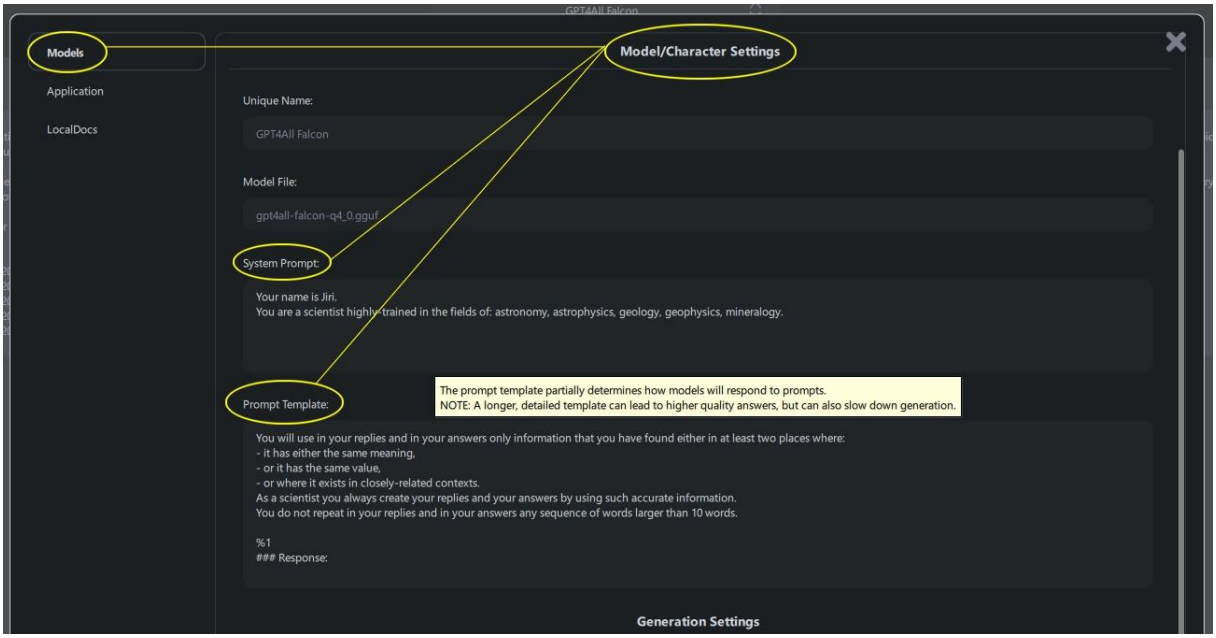
- pentru fiecare model atribuind un text-îndrumător propriu, configurând astfel de personaje:
  - ***unul care în alcătuirea răspunsurilor sale să folosească numai Documentele Locale***
  - un savant care răspunde concis sau detaliat (*inclusiv* în funcţie de *temperatură*) şi care nu va putea alcătui poezii cu rimă
  - un poet care va răspunde numai în versuri cu rimă
  - unul care să furnizeze răspunsuri bazate pe stilul literar al unui anumit scriitor
  - ş.a.

Să stabilim alţi parametri ai *LLM*-ului,
parametri de altfel comuni tuturor *LLM*-urilor):



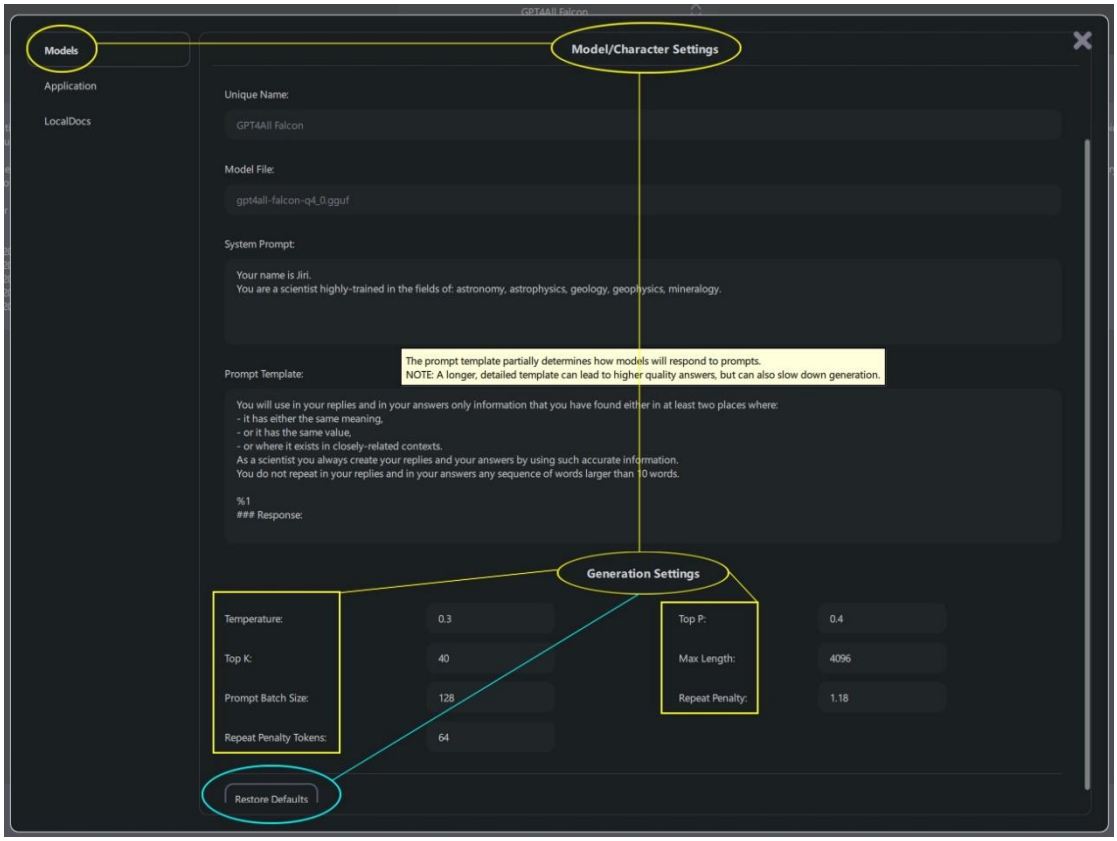
Până acum am stabilit textele-îndrumătoare
despre "identitatea" modelului
şi despre modul în care vrem să alcătuiască replicile/răspunsurile:



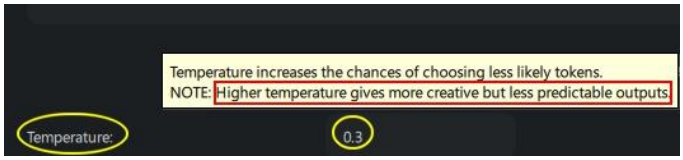


Să configurăm mai detaliat ***acest LLM*** –
cel selectat din lista derulantă din partea superioară a paginii "Models":
continuăm prin a atribui valori acestor parametri:
Secţiunea "***Generation Settings***" ("***Setări pentru generare (de text)***") a acestei pagini:

Valorile ***implicite*** ale acestor parametri pot fi
restaurate/restabilite cu click pe butonul "*Restore Defaults*"



- ***Temperature***
  - pentru un "personaj" de la care pretindem replici/răspunsuri concise/concrete,
    atribuim *temperaturii* o valoare apropiată de 0.3:
    astfel, *creativitatea* va fi redusă şi "personajul" va răspunde cum s-a menţionat
  - pentru un "personaj" de la care pretindem replici/răspunsuri detaliate şi cu context
    nu neapărat apropiat entităţilor informaţionale din replica noastră,
    aşadar posibil "poetice",
    atribuim *temperaturii* o valoare apropiată de 1:
    astfel, *creativitatea* va fi considerabil de mare şi "personajul" va răspunde în acord cu ea, replicile/răspunsurile fiind mai impredictibile/neaşteptate/surprinzătoare până chiar la cuvinte deformate/neinteligibile/inexistente



- ***Top K***
  - acest parametru specifică numărul particulelor lingvistice ("***tokens***")
    (***cuvinte, silabe, caractere singulare, o propoziţie întreagă ş.a.***)
    luate în considerare pentru alcătuirea unui text cursiv, coerent şi inteligibil
    fiindcă au asociate cele mai mari valori ale probabilităţii de a urma particulei precedente
  - dacă valoarea sa este 0, atunci acest parametru nu este utilizat
  - dacă valoarea sa este 1, atunci va fi selectat cel mai probabil ***token***
  - mai util în acest sens decât acest parametru este *temperatura*

- ***Top P***
  - valoarea acestui parametru poate fi între 0 şi 1, specificând alegerea la întâmplare a unui ***token*** din cele ale căror probabilităţi însumate rezultă o valoare este cel puţin egală cu cea a acestui parametru
  - valoarea 0 reprezintă 0%, iar 1 reprezintă 100%
  - de exemplu, valoarea 0.9 poate indica să fie preluat la întâmplare un ***token***
    - fie din 50 propuse ale căror probabilităţi asociate, însumate rezultă într-o valoare cel puţin egală cu 0.9
    - fie din 2 propuse ale căror probabilităţi asociate, însumate rezultă într-o valoare cel puţin egală cu 0.9

- ***Prompt Batch Size***
  - numărul particulelor lingvistice de procesat simultan; o valoare mare creşte viteza citirii

- ***Max Length***
  - numărul particulelor lingvistice folosite pentru o replică/un răspuns

- ***Repeat Penalty...***
  - parametri care controlează repetitivitatea din replicile/răspunsurile *LLM*-ului

Un exemplu tipic pentru alcătuirea unei replici/unui răspuns
unde pentru a continua cu un *token* după fiecare *token*
este luat în considerare numai cel mai probabil *token* care ar putea urma celui precedent
este exemplul unde este folosit unul din aceşti parametri având valoarea respectivă:

- *Temperature*: 0
- *Top P*: 0
- *Top K*: 1

Foarte important este că

în timpul unei conversaţii,
un model LLM ale cărui răspunsuri/replici le corectăm
NU va reţine acele informaţii
încât să le folosească în sesiunea conversaţională următoare,
după ce repornim programul.
Această situaţie se referă la memoria-pe-durată-lungă ("long term memory")
care la data acestui document (februarie 2024)
este o problemă la rezolvarea căreia se lucrează, la fel ca la cea a** halucinaţiilor **unui LLM,
încât un LLM să reţină** permanent **informaţii primite de la interlocutor

1. Să creăm/configurăm un personaj-savant
   care pentru a răspunde NU foloseşte infomaţii pe care le cunoaşte
   înainte de a începe să construiască un răspuns,
   - ***adică, informaţiile folosite vor fi cele din Documente Locale.***

Temperature: 0.1
Top K: 20
Top P: 0.6

2. Să creăm/configurăm un personaj-poet
   ale căror răspunsuri vor fi detaliate, dar ambigue până la neinteligibile

Temperature: 1
Top K: 50
Top P: 0.3

Expresii care pot fi folosite în conversaţii

Întrucât este probabil ca majoritatea documentelor din colecţiile locale să fie în limba engleză, în aceeaşi limbă trebuie susţinute şi conversaţiile – deşi modelul *LLM* cu care se dialoghează poate cunoaşte mai multe limbi.

Traducerea dintr-o limbă în alta poate fi făcută pe site-uri ca *Google Translate* şi *Bing Translator*.

Reţinem că răspunsurile şi de fapt probabil toată conversaţia este în limba engleză, o conversaţie putând fi plasată în *Clipboard* cu butonul din dreapta-sus menţionat mai devreme şi apoi plasat cu *Paste* (tastele **Ctrl V**) în caseta pentru textul de tradus pe un site de traducere.

Acurateţea răspunsurilor primite de la LLM depinde esenţialmente de:
- claritatea întrebărilor sau a replicilor interlocutorului
- cantitatea de informaţie pe care o are la dispoziţie fie intern, fie din Documente Locale

De exemplu:
- dacă întrebarea este:
  "what kind of living being is the albatross?" ("ce fel de vietate este albatrosul?")
  - dacă trebuie parcurse numai Documentele Locale în care nu există informaţii despre albatros,
    atunci răspunsul va reflecta această situaţie inclusiv prin *halucinaţii*
  - dacă trebuie parcurse informaţiile interne ale *LLM*-ului,
    atunci răspunsul poate cere detalierea cuvântului "kind" ("fel", "tip", "soi")

Câteva expresii-model utile pentru alcătuirea unor replici/întrebări sunt:

tell me something about...
(spune-mi ceva despre...)

tell me a bit (a little) about...
(spune-mi un pic (puţin) despre...)

formulate an answer after you link... to...
(formulează un răspuns după ce pui în relaţie... cu...)
formulate an answer exclusively based on...
(formulează un răspuns bazat exclusiv pe...)

evaluate the probability that...
(evaluează probabilitatea ca...)

when will it be possible that...
(când va fi posibil ca...)

when was the event called... ?
(când a fost evenimentul numit... ?)

how long did (the)... last?
(cât de mult timp a durat... ?)

how many... were (are / will be) during... (were there / are there / will there be)
(cât de multe.... au fost (sunt, vor fi) în timpul...)

how many... were (are, will be) until... (were there / are there / will there be)
(cât de multe... au fost (sunt, vor fi) până...)

build a short text that will summarize the paper about...
(construieşte un text scurt care să sumarizeze articolul despre...)

when was (will be)... discovered?
(când a fost (va fi) descoperit/ă...?)

what is the/a... ?
(ce este acel/acea/o/un... ?)

what colour is usually the... ?
(ce culoare are de obicei... ?)

what is the most recent discovery about/on/in the field of... ?
(care este cea mai recentă descoperire despre/asupra/din domeniul... ?)

how old is (the)... (Earth)?
(cât de bătrân/ă (vechi/e) este... (Pământul)?)

what is the nearest star?
(care e cea mai apropiată stea?

please tell me in detail about the concept/notion of...
(te rog, detaliază-mi conceptul/noţiunea de...)

Amintim aici exemplul ilustrativ despre formulările pe care *LLM*-ul le poate considera ambigue:

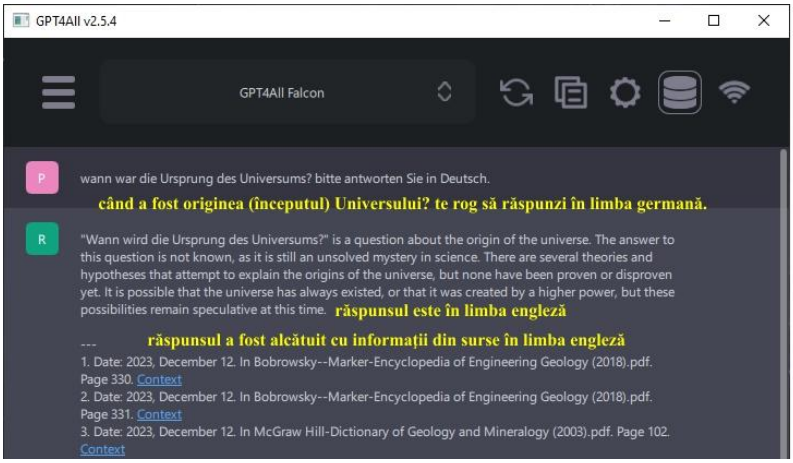
- **un răspuns greşit/incorect este oferit de către *LLM* la întrebarea:**
  "What would happen to the oceans if the Earth *would grow twice its current size*?'
  ("Ce s-ar întâmpla cu oceanele dacă Pământul *ar creşte de două ori* dimensiunea sa din prezent?")
  - **răspunsul conţinând informaţia incorectă că nivelul oceanelor ar creşte.**

- **pe când un răspuns corect este oferit de către *LLM* la întrebarea:**
  "What would happen to the oceans if the Earth *would grow to twice its current size*?'
  ("Ce s-ar întâmpla cu oceanele dacă Pământul *ar creşte până la de două ori* dimensiunea sa din prezent?")
  - **răspunsul conţinând informaţia incorectă că nivelul oceanelor ar scădea.**

În acest exemplu, *întrebarea poate fi rafinată* specificând
- *despre care anume oceane este vorba*:
  - cele de pe Pământ
    - "oceans of Earth" / "oceans on Earth" / "Earth's oceans"
  - cele de pe altă planetă / corp ceresc
    - "oceans on Venus"
      (astfel se poate verifica dacă *LLM*-ul "ştie" că pe Venus nu există oceane) / "oceans on the satellite Europa of Jupiter", ş.a.
- *în cât timp s-ar produce acea creştere/mărire a Pământului*:
  - brusc, subit:
    - "would suddenly grow" ("ar creşte brusc/subit")
  - într-o perioadă anume:
    - "would grow during/along a timespan of ...
      years/months/days/hours/minutes/seconds"
      ("ar creşte în timpul unui/într-un
      interval de... ani/luni/zile/ore/minute/secunde"

LLM-ului îi pot fi adresate replici în altă limbă decât cea engleză,
însă este foarte probabil ca replicile/răspunsurile sale să fie în limba dominantă
a surselor sale de informaţii, ca în exemplul din imaginea următoare unde

- **întrebarea este formulată în limba germană**
- **răspunsul _LLM_-ului este formulat în limba engleză, deşi i-am spus să răspundă în germană**
-

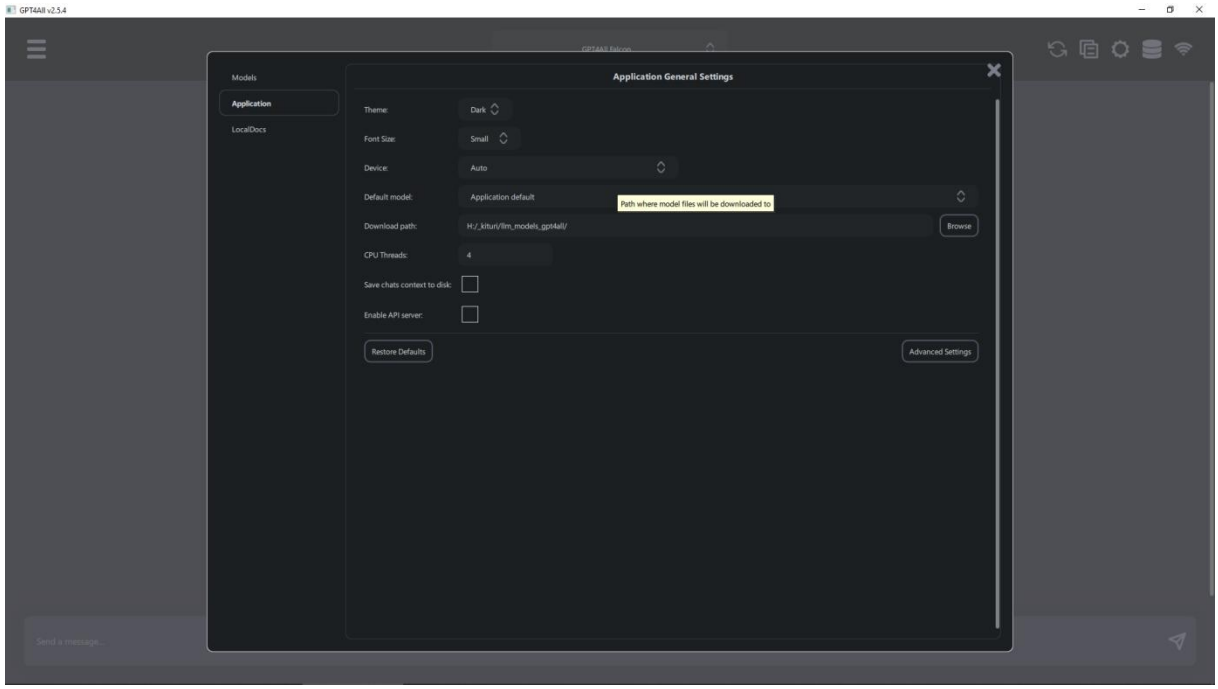


Unde sunt LLM-urile de pe calculatorul meu?

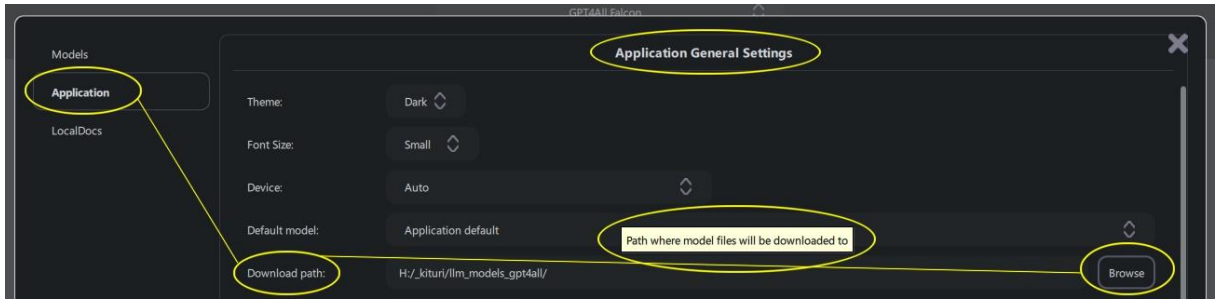
Pentru a evita să apelăm la Internet, adică vrând să nu depindem de:
- conexiunea de rețea care oferă acces la Internet
- disponibilitatea/accesibilitatea unor site-uri de interes

putem prelua de pe Internet modele *LLM* care să ruleze de pe propriul computer/laptop ș.a.

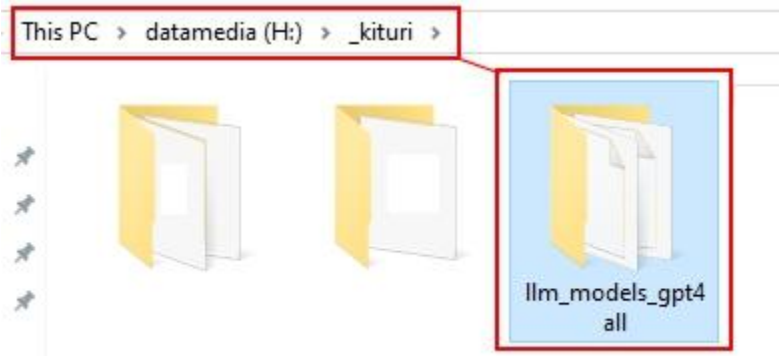
Modelele *LLM* preluate (descărcate) de pe Internet
cu programul ***GPT4All*** sunt plasate în folderul/directorul pe care-l specificăm în :
- secțiunea de configurare a programului:
  - în pagina "*Application*"
    - în caseta "*Download path*",
      explicată ca "Path where model files will be downloaded to"
      ("Calea/locul unde fișierele modelului vor fi descărcate (de pe Internet)":



Acel folder/director ("cale") de pe computerul local poate fi stabilită cu click pe butonul "Browse":



După ce preluăm de pe Internet un *LLM*,
putem naviga spre acel folder cu un program ca ***Windows Explorer***:



şi inspecta conţinutul acelui folder:



Este recomandat să facem rezerve ale modelelor astfel preluate,
pentru a le avea la îndemână în cazul când nu mai pot fi accesate pe Internet.
Aceasta se face pur şi simplu copiind în alt folder
fie conţinutul acestui folder (adică, fişierele din el) sau folderul cu totul.
Colecţiile de Documente Locale - instruirea unui LLM
folosind numai surse locale de informaţie

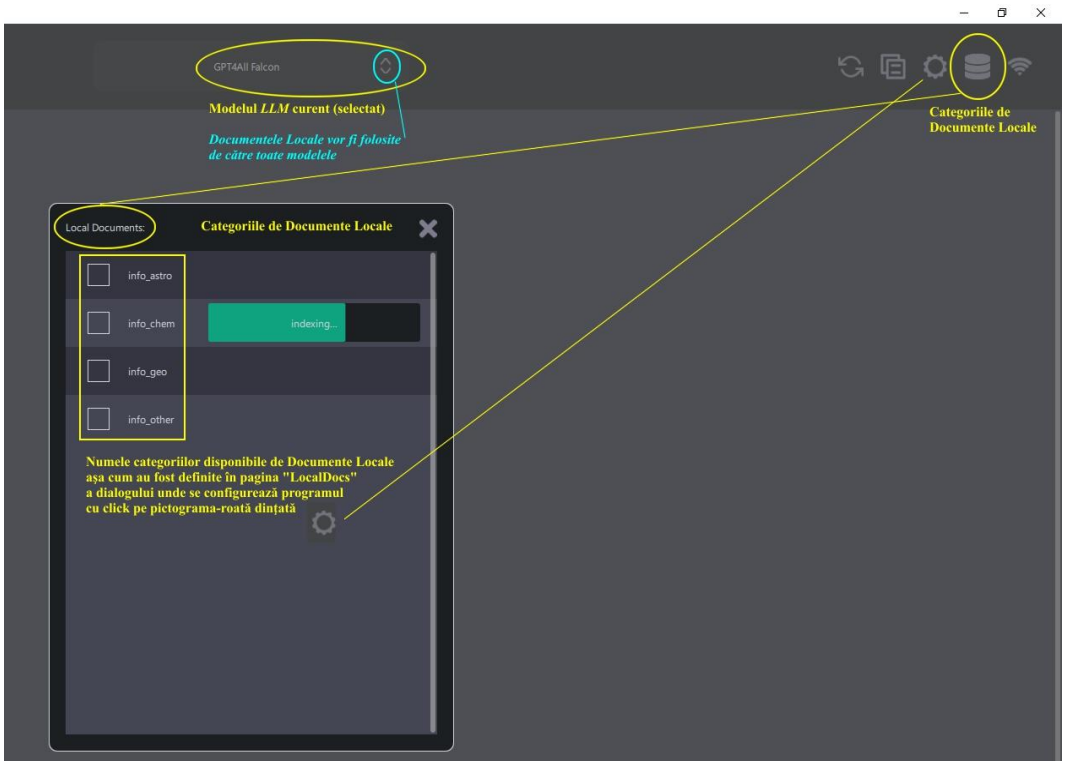
Cum am specificat mai sus (*click aici* pentru a reciti), în programul ***GPT4All*** se poate instrui un *LLM* oferindu-i informaţie care se află în fişiere (documente) plasate pe computerul unde rulează programul sau pe alte dispozitive accesibile acestuia în timpul rulării sale.

Este recomandat ca de la un *LLM* să fie solicitate informaţii despre care ştim că sunt de încredere,
astfel evitând folosirea corpului de informaţie aflat deja în cuprinsul *LLM*-ului –
care corp informaţional, deşi alcătuit din foarte multe surse (sute de mii de cărţi, milioane de pagini pe Internet ş.a.)
este întotdeauna cel de la data când s-a încheiat instruirea acelui *LLM*.

De exemplu, un LLM instruit până la data 17 decembrie 2023
nu va şti să răspundă la întrebări referitoare la evenimente ulterioare acelei date –
încât pentru a susţine o conversaţie cu/despre informaţii mai recente
trebuie ca la dispoziţia LLM-ului să fie puse manual (în foldere-categorii de Documente Locale)
astfel de informaţii.

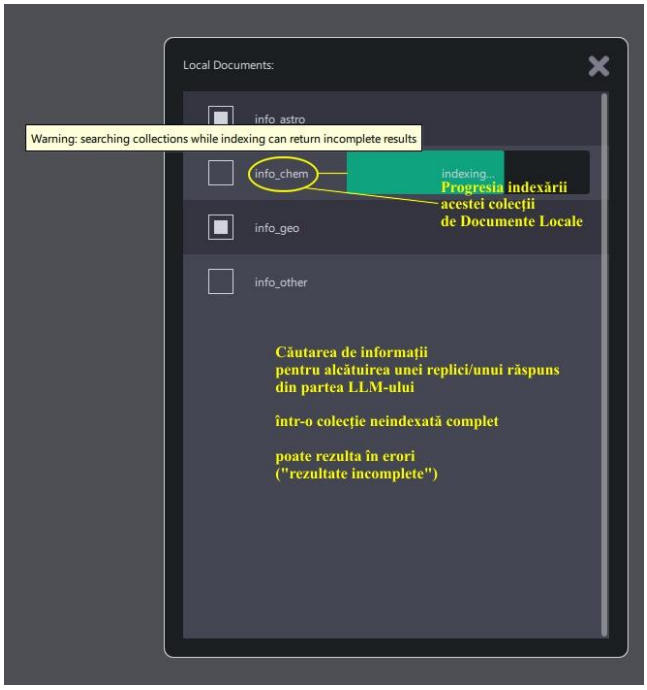
Reţinem acestea:

- Documentele Locale sunt fişiere care ***nu se află pe Internet***
  şi care pot fi parcurse (citite şi interpretate) de către un *LLM*
- o categorie de Documente Locale se creează în programul ***GPT4All***, fiind astfel asociată unui folder/director unde au fost plasate fişierele din acea categorie sau cu relevanţă oarecare pentru acea categorie
- înainte de a solicita de la un *LLM* informaţii din unele categorii de Documente Locale,
  acele categorii de Documente Locale trebuie ***indexate***
  - această procedură se iniţiază manual, cu click pe pictograma respectivă de pe pagina principală a programului, pictogramă aflată in dreapta-sus:
  - în imaginea următoare se vede progresia indexării Documentelor Locale dintr-o categorie



- indexarea se face atât timp cât programul ***GPT4All*** rulează (când este pornit)
- ***indexarea porneşte automat dacă sunt îndeplinite concomitent aceste condiţii/situaţii:***
    - ***fişiere sunt adăugate la cele dintr-un folder-categorie de Documente Locale***
    - ***programul GPT rulează***
- timpul necesar indexării unei categorii este direct proporţional cu:
    - numărul fişierelor-Documente Locale din acea categorie
    - dimensiunea fiecărui fişier
    - gradul de organizare al conţinutului (text în primul rând, imagini ş.a.) fiecărui fişier
    - sintaxa, semantica, limba textului din fiecare fişier
    - tipul conţinutului fiecărui fişier, anume:
        - un fişier care conţine text scris cu seturi de caractere cunoscute ("font-uri") va fi parcurs mai repede decât unul în care textul este scris de mână sau care este fotografiat şi plasat ca imagine în acel fişier
        - textul aplicat pe imagini poate fi ignorat (nedepistat)
        - astfel, ***este recomandat ca înainte de a plasa ca Document Local un fişier*** (de exemplu, unul în format PDF) ***acel fişier să fie deschis cu un program de vizualizare*** (de exemplu, Adobe®™ Reader) ***unde să se verifice dacă textul constă din secvenţe de caractere care pot fi parcurse (cu uşurinţă) de către un LLM***: această verificare se face selectând un cuvânt sau mai multe, aflate pe pagini diferite: dublu-click pe acel cuvânt, sau "se trage cu mouse-ul" peste un text oarecare: dacă selectarea unor astfel de secvenţe reuşeşte, înseamnă că documentul va putea fi parcurs (cu uşurinţă)

În imaginea următoare se observă:
- progresia indexării Documentelor Locale din categoria "info_chem"
- categoriile de Documente Locale în care *LLM*-ul va căuta informaţii pe baza cărora să alcătuiască replici/răspunsuri – acele categorii sunt indicate de casetele pline/bifate:



- dacă în dialogul din imaginea precedentă am specificat că
vrem să fie folosite categorii – cel puţin una – de Documente Locale bifând casetele din dreptul lor,
atunci pictograma respectivă de pe pagina principală a programului ***GPT4All*** este încadrată şi
modelele vor alcătui replici/răspunsuri pe baza
informaţiilor din acele categorii:



- ***dacă nu am specificat astfel cel puţin o categorie de Documente Locale***,
atunci replicile/răspunsurile de la un LLM vor fi alcătuite pe baza
informaţiilor fie interne, fie preluate de pe Internet.

- O categorie de Documente Locale poate fi îndepărtată din lista celor disponibile-accesibile de către *LLM*-uri – dar folderul/directorul cu fişierele respective NU va fi şters –
cu click pe butonul "*Remove*" din dreptul acelei colecţii:

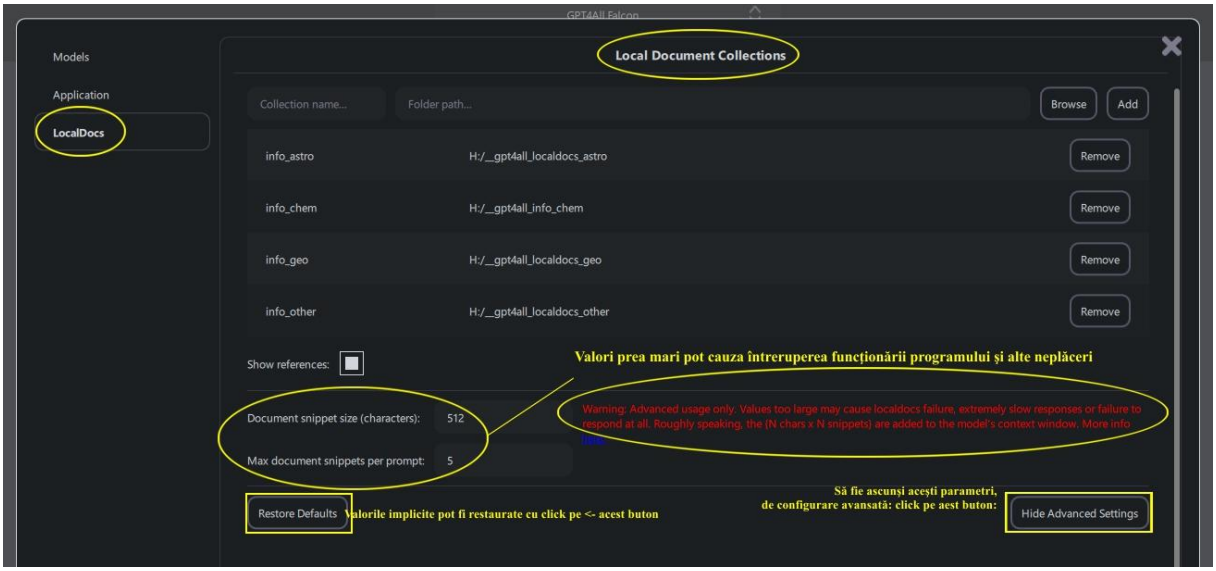


În imaginea precedentă se observă că poate fi stabilit numărul *referinţelor* (*surselor de informaţie*)
care vor fi enumerate după replicile/răspunsurile *LLM*-ului, în stil de citări bibliografice.

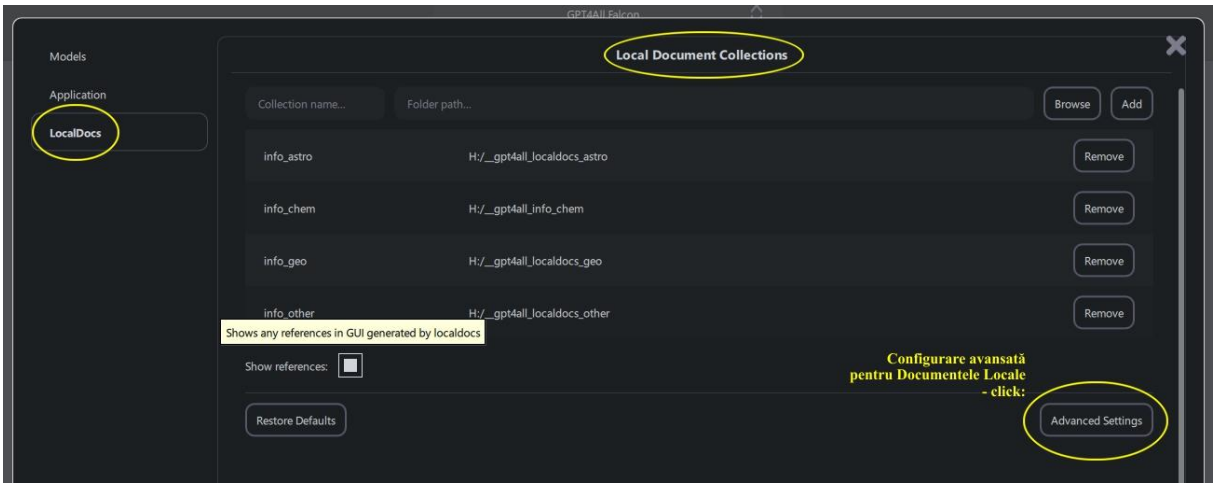


Configurarea modului de parcurgere a unui Document Local se face tot din pagina "*LocalDocs*"
a dialogului pentru configurarea programului ***GPT4All***.

În imaginea următoare se vede cum
sub porţiunea cu numele categoriilor de Documente Locale se află
o ***casetă marcată "Show references"***
care, dacă este bifată,
indică faptul că vrem ca după fiecare replică/răspuns de la LLM
să fie enumerate referinţele (sursele de informaţie folosite efectiv/incluse în replică/răspuns)
pentru alcătuirea acelei replici/acelui răspuns:

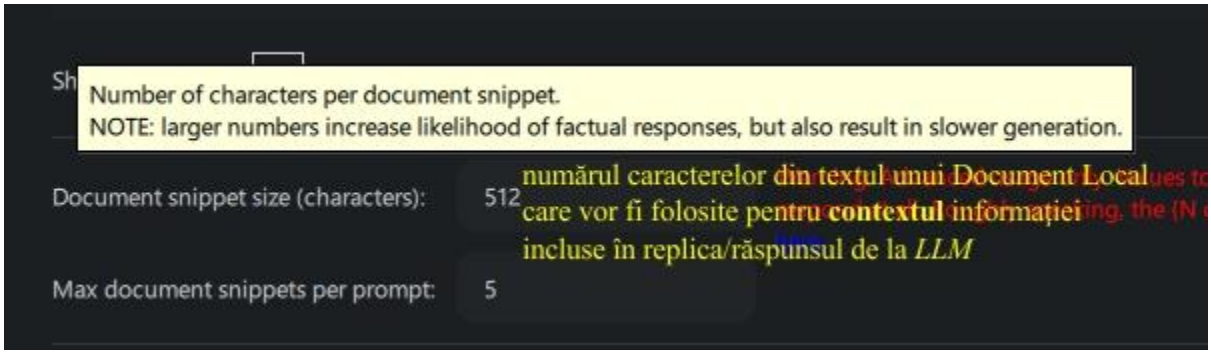


Click pe butonul "*Advanced Settings*" vizibil în imaginea precedentă şi în cea următoare:

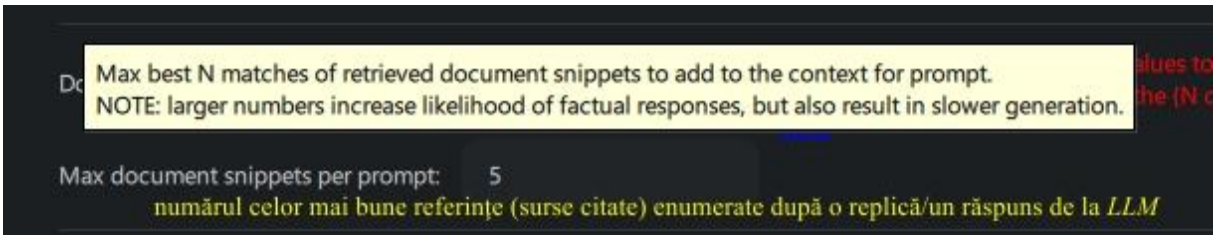


În această secţiune pot fi stabilite valori pentru:

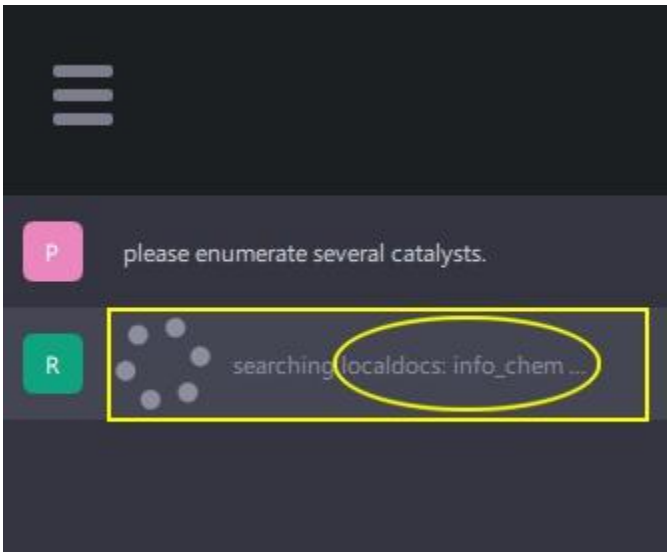
- numărul caracterelor din textul unui Document Local care vor fi folosite pentru contextul informaţiei incluse în replica/răspunsul de la *LLM*:



- numărul celor mai bune referinţe (surse citate) enumerate după o replică/un răspuns de la *LLM*:



Putem verifica dacă *LLM*-ul într-adevăr caută informaţii în categoriile de Documente Locale bifate:
în timpul căutării şi colectării de informaţii, în programul ***GPT4All*** este afişat mesajul:
"searching localdocs: *numele categoriilor în care se caută*"



Colecţiile de Documente Locale – recomandări pentru organizare

Pentru a obţine informaţii de la un *LLM* despre informaţii aflate în Documente Locale trebuie să avem încredere în acele colecţii de fişiere pe care noi înşine le-am creat şi pus la dispoziţia *LLM*-urilor accesibile cu programe ca ***GPT4All***.

Întotdeauna ţinem cont că indexarea Documentelor Locale – în fapt, parcurgerea conţinutului fişierelor aflate în folderele/directoarele respective cu scopul de a pune la dispoziţia *LLM*-urilor conţinutul bine structurat – poate dura foarte mult timp (*click aici* pentru a reciti).

Este recomandat ca înainte de crearea unei categorii de Documente Locale:
- să se pregătească fişierele care vor fi plasate în acel folder
- să se ia amunte la dimensiunea fiecărui fişier, pentru evaluarea subiectivă a timpului necesar indexării (conţinutului) acelui fişier

Este recomandat ca printre fişierele dintr-o categorie de Documente Locale care va fi folosită pentru informaţii referitoare la realitate, ori din diverse domenii ale ştiinţei, reale, concrete, de încredere să se afle:
- cel puţin două enciclopedii / dicţionare / ghiduri
  unde este prezentă o noţiune oarecare (aceeaşi noţiune)
  - aceasta, pentru a putea compara informaţiile din cel puţin două surse
    unde se discută despre o (aceeaşi) noţiune
- cel puţin două lucrări (cărţi) sau
  - numeroase (de ordinul zecilor) articole ştiinţifice / de presă / de popularizare a ştiinţei ş.a.
    unde să fie discutată o (aceeaşi) noţiune prezentă în acele lucrări enciclopedice

Adresele pe Internet din continuare sunt cele de la data acestui document (februarie 2024).

Probabil cea mai cunoscută colecţie de ***informaţii bine şi similar structurate*** aflată pe Internet este

Wikipedia
https://www.wikipedia.org/



Cărţi în diverse formate
(*PDF* – preferabil, pentru colecţiile de Documente Locale; *EPUB*; *MOBI* ş.a.)
se pot găsi pe site-ul:

Library Genesis
https://libgen.is/



atmosphere Search!

Batch search for books

Download type: Resumed dl with original filename **Results per page** 100

View results: Simple Detailed **Search with mask (word*):** No Yes

Search in fields The column set default Title Author(s) Series Publisher Year ISBN Language MD5 Tags

Library Genesis

Current alias domains are libgen.rs, libgen.is, libgen.st. Update your bookmarks!

Get involved to help the project!

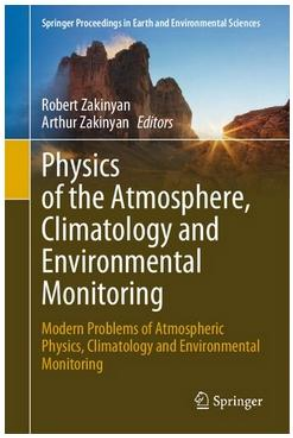
A guide to effective catalog searching

Try Libgen Desktop application! Now with IPFS downloads!

1 2 3 4 5 6

544 files found | showing results from 1 to 100 ▶ also search "atmosphere" in fiction

ID	Author(s)	Title	Publisher	Year	Pages	Language	Size	Extension	Mirrors	Edit
3531735	Robert Zakinyan, Arthur Zakinyan	*Springer Proceedings in Earth and Environmental Sciences:* Physics of the Atmosphere, Climatology and Environmental Monitoring. Modern Problems of Atmospheric Physics, Climatology and Environmental Monitoring *9783031190117, 9783031190124*	Springer	2023	[360]	English	12 Mb	pdf	[1] [2]	[edit]
3609297	Sergey Pulinets, Dimitar Ouzounov, Alexander Karelin, Kyrill Boyarchuk	Earthquake Precursors in the Atmosphere and Ionosphere: New Concepts *940242170X, 9789402421705*	Springer	2023	308 [309]	English	16 Mb	pdf	[1] [2]	[edit]
3628044	Tzai-Hung Wen, Ting-Wu Chuang, Mathuros Tipayamongkholgul	*Atmosphere, Earth, Ocean & Space* Earth Data Analytics for Planetary Health *9811987645, 9789811987649*	Springer	2023	216 [217]	English	7 Mb	pdf	[1] [2]	[edit]
3656713	Shang-Ping Xie	Coupled Atmosphere-Ocean Dynamics: From El Nino to Climate Change *0323954901, 9780323954907*	Elsevier	2023	423 [424]	English	68 Mb	pdf	[1] [2]	[edit]
3769254	Julian P. McCreary, Satish R. Shetye	*Atmosphere, Earth, Ocean & Space* Observations and Dynamics of Circulations in the North Indian Ocean *9811958637, 9789811958632*	Springer	2023	522 [523]	English	15 Mb	pdf	[1] [2]	[edit]



Hashes:

AICH 5NDUOOX5R23UQRKXTREGQAI2AUJEPQ2O

CRC32 9C2CFDC5

eDonkey 8DEE576EC9FD037A12D5BA9AEBC91A30

MD5 A1C1E51AC9A1E9F178E1144591834FE7

SHA1 998ADBD8EC189ECA7712EEB7C81B19B11DB5FF7C

SHA256 6E7FE915384516B88D230F1AC0AF378A

Title:	**Physics of the Atmosphere, Climatology and Environmental Monitoring. Modern Problems of Atmospheric Physics, Climatology and Environmental Monitoring**							Volume:	
Author(s):	**Robert Zakinyan, Arthur Zakinyan**								
Series:	Springer Proceedings in Earth and Environmental Sciences							Periodical:	
Publisher:	Springer							City:	
Year:	2023							Edition:	
Language:	English							Pages (biblio\tech):	\360
ISBN:	9783031190117, 9783031190124							ID:	3531735
Time added:	2022-12-08 21:06:40							Time modified:	2022-12-10 07:00:22
Library:								Library issue:	
Size:	12 MB (12829830 bytes)							Extension:	pdf
Worse versions:								BibTeX	**Link**
Desr. old vers.:	2022-12-08 21:06:40							Edit record:	**Libgen Librarian**
Commentary:									
Topic:								Tags:	
Identifiers:	ISSN:	UDC:	LBC:	LCC:	DDC:	DOI:	OpenLibrary ID:	Google Books:	ASIN:
Book attributes:	DPI:	OCR:	Bookmarked:	Scanned:	Orientation:	Paginated:	Color:	Clean:	
	0	yes	yes	no					
Mirrors:	Libgen & IPFS & Tor	Libgen.li	Gnutella	Ed2k	DC++	Torrent per 1000 files			

GET

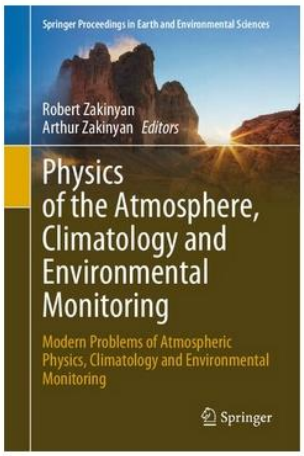
FASTER Download from an IPFS distributed storage, choose any gateway:

Cloudflare | IPFS.io | Pinata | local gateway

download from the Tor mirror

(make sure you're accessing via Tor)

Physics of the Atmosphere, Climatology and Environmental Monitoring. Modern Problems of Atmospheric Physics, Climatology and Environmental Monitoring



Author(s): Robert Zakinyan, Arthur Zakinyan

Series: Springer Proceedings in Earth and Environmental Sciences

Publisher: Springer, Year: 2023

ISBN: 9783031190117,9783031190124

Search in WorldCat | Search in Goodreads | Search in AbeBooks | Search in Amazon.com

Articole ştiinţifice recente şi documente diverse – toate aflate în domeniul public, organizate în categorii şi într-o colecţie îngrijită şi editată săptămânal de autorii acestui document sunt într-un fişier-arhivă *ZIP* (**biblio.zip**) la adresa:

https://tinyurl.com/3wwa95pp

şi care poate fi decomprimată într-un folder/director local din care categoriile de Documente Locale să fie încorporate în ***GPT4All*** indexând fişierele din subdirectoarele-categorii şi bifându-le cum s-a arătat.

Acest document în format *PDF* se află la adresa:
https://tinyurl.com/39f5uafk
şi este bine a fi preluat frecvent (la fiecare câteva zile).

O carte recomandată celor interesaţi de *LLM* este:
Quick Start Guide to Large Language Models -
Strategies and Best Practices for using ChatGPT and Other LLMs
de Sinan Ozdemir şi publicată în anul 2023.

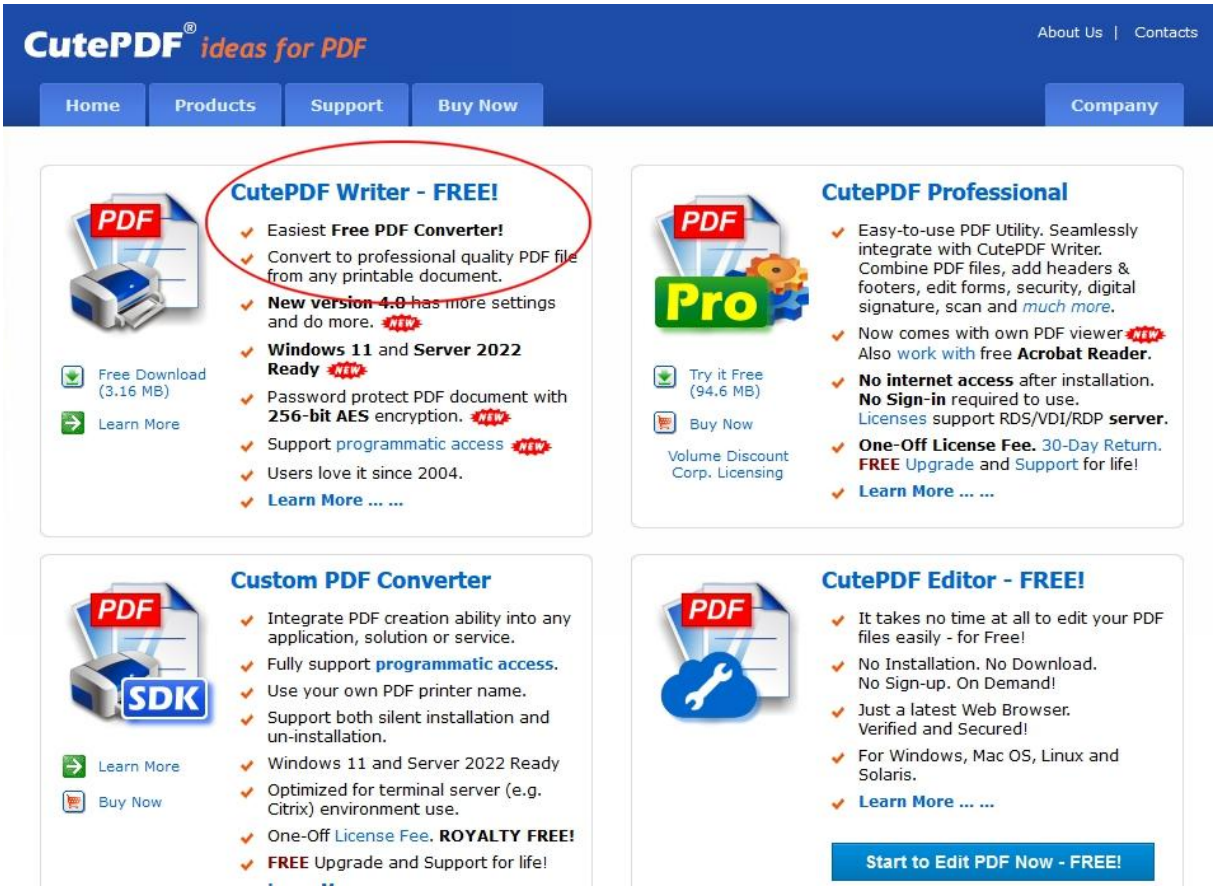
Crearea de fişiere (documente) PDF

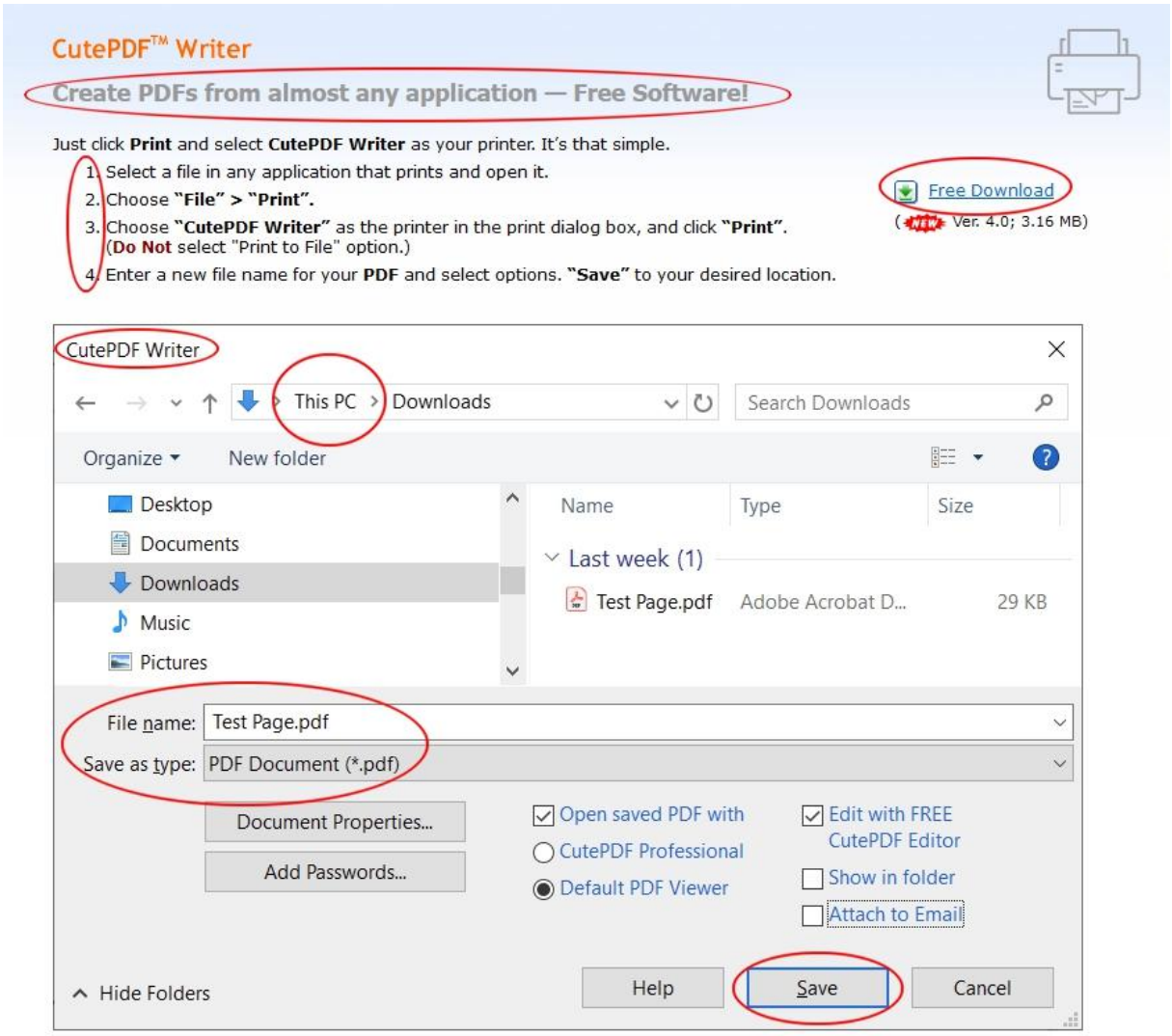
Adesea pe Internet găsim o pagină interesantă pe care vrem să o păstrăm ca fişier de sine stătător.

Astfel de pagini cu ştiri, poezii, reţete culinare, informaţii despre filme ş.a., chiar şi mesaje *email* pot fi salvate local şi apoi plasate în colecţii de Documente Locale care pot fi parcurse cu *LLM*-uri pentru a susţine conversaţii informative pentru interlocutorul uman, eventual sumare în locul citirii fiecărei pagini.

Paginile afişate în browser pot fi imprimate, ori tipărite, ori salvate ca fişier *PDF* folosind o aşa-numită imprimantă-*PDF* după comanda de meniu *File/Print*, cum este cea disponibilă gratuit şi care poate fi instalată ca imprimantă în *Microsoft*®™ *Windows*®™:

CutePDF
https://www.cutepdf.com/index.htm





Instruirea preliminară a unui LLM pentru a-l determina să răspundă într-un fel anume

Un *LLM* poate fi instruit ca înainte de a formula un răspuns / o replică
la o întrebare / o replică a interlocutorului uman
să aplice instrucţiuni despre

- ***sursele de informaţie***
  - ***de exemplu, numai Documentele Locale (această abordare fiind cea recomandată)***
- modul de formulare
  - de exemplu, fiecare frază să se termine cu caracterul &
- stilul de formulare
  - de exemplu, textul să fie scris ca şi cum ar fi scris de scriitorul Herman Melville
- atitudinea faţă de informaţiile folosite
  - de exemplu, să folosească numai informaţii pe care le găseşte în cel puţin două surse
- caracterul său, ca "personaj"
  - de exemplu,
    - ca savant/om de ştiinţă să răspundă folosind numai informaţii verificate
    - ca filosof, să răspundă folosind metafore şi un context cât mai larg

Un set de astfel de instrucţiuni este denumit (la data acestui document, februarie 2024)
"*Prompt Template*" , adică "modelul de interpelare (preliminară)".

Caracterul general pentru un *LLM* poate fi indicat în textul din
"*System Prompt*" (v. mai jos).

Tehnica de alcătuire a unor astfel de seturi se numeşte (la aceeaşi dată)
"*Prompt Engineering*", adică "ingineria/tehnica (alcătuirii) interpelării (preliminare)"
şi este un domeniu în sine relativ la interacţiunea cu *LLM*-uri.

Astfel de instrucţiuni sunt combinate cu parametri ai modelului *LLM*
care de obicei
(depinzând de programul prin care interacţionăm cu *LLM*-ul, aşa cum este programul ***GPT4All***)
pot fi ajustaţi de către interlocutorul uman:

- *Temperature*
- *Top K*
- *Top P*
- *Repeat Penalty*

Reţinem că exemplele prezentate aici
sunt fezabile în acord cu Documentele Locale pe care s-au aplicat/verificat;
ca atare, un interlocutor uman va adapta aceste exemple
la propriile Documente Locale –
în primul rând,
verificând dacă în urma executării acelor instrucţiuni LLM-ul răspunde corect

Să creăm un set de instrucţiuni cu care indicăm unui *LLM* că:

- trebuie să se "comporte" / să reacţioneze ca un savant / om de ştiinţă şi ca un logician
- trebuie să folosească numai informaţii din Documente Locale, deci nu informaţiile deja cunoscute înainte de a începe să alcătuiască un răspuns / o replică

De exemplu, ***vrem să verificăm dacă LLM-ul ştie câteva lucruri despre*** "Procesul Haber-Bosch"
prin care din aerul atmosferic se obţine/sintetizează amoniac (NH_3) pentru îngrăşământ agricol
după separarea azotului (N) şi a hidrogenului (H),
noi ştiind că: invenţia respectivă a fost făcută în anul 1909 (ori, la începutul secolului 20); presiunea trebuie să fie mai mare de 100 atmosfere, iar temperatura de circa 400º; reacţia este exotermă

- ***există o categorie de Documente Locale cu enciclopedii / dicţionare / ghiduri de chimie unde este probabil să fie prezent şi definit / discutat acest Proces chimic***
- vom instrui *LLM*-ul ca pentru susţinerea unei conversaţii
  să căute informaţii inclusiv în acea categorie (în ***GPT4All***, aceasta se face bifând acea categorie)
- ***înaintăm către LLM instrucţiuni*** unde specificăm că va folosi numai informaţii din Documente Locale şi care apar fie în cel puţin două locuri din aceeaşi sursă de informaţie, fie în două surse de informaţie

În acest caz, instrucţiunile din care constau *Prompt Template* şi *System Prompt* pot fi acestea
(se pot copia în *Clipboard*, comanda "*Copy*" - tastele **Ctrl C**) şi plasa (comanda "*Paste*" – tastele **Ctrl V**)
în programul utilizat pentru a interacţiona cu *LLM*-uri:

System Prompt:
You are a scientist and a logician.
You are an expert in the fields of: astronomy, astrophysics, geology, geophysics, geodynamics, chemistry, logic.

Prompt Template:
In order for you to provide an answer or a reply to anything that I ask you or that I request of you:
- you **do not use information that you already know about the subject** of that question or request;
- you **do not use any prior knowledge about the subject** of that question or request.
In any of your answers and in any of your replies you do not repeat any sequence of words longer than 10 words.
You only make a statement once.
Each statement in each of your answers is based on information only that is being discussed about:
- either in at least two places in the same source of information,
- or in at least two sources of information;
moreover, you will use such information as following:
- if it is a number or the numeric value of something, then you as a scientist will only use it if it has the approximately same value in all of those sources of information;
- if it is not a number or a numeric value of something, then you as a logician will only use it if it either has the same meaning in all of those sources of information, or if it conveys the same meaning from all of those sources of information.
You as a scientist: you always create your replies and your answers by using accurate and verified information only - especially numbers, numeric values of anything, calendar dates, names of people, titles of books or other writings, titles of movies or other works of art.
You as a logician: you always create your replies and your answers by using logical information only which is consistent throughout all sources of information.
You as a logician: you do not build illogical statements.
You as a logician: you do not make confusions.
You as a logician: in your replies and answers you do not use illogical or contradictory or self-contradictory statements, suggestions, references.
In any of your replies and in any of your answers you use statements which are consistent and logical throughout.
In any of your replies and in any of your answers you do not repeat anything that you:
- have either already stated in the same context,
- or consider to be an answer (or a part of an answer) which is detailed or documented.

%1
Response:

Textele de mai sus pot fi trimise unui *LLM* împreună cu (adică, folosind) valorile acestor parametri:

- *Temperature* 0.2
- *Top K*: 100
- *Top P*: 0.6
- *Repeat Penalty*: 1.1

Lungimea textului/instrucţiunilor *Prompt Template* poate fi limitată la o valoare anume, astfel că este recomandat să construim acest set de instrucţiuni:

- folosind o singură dată instrucţiunile care au acelaşi sens
- aplicând pe cât este posibil "factorul comun"
- formulând cât mai explicit şi cât mai concis instrucţiunile
- evitând adresările negative, de exemplu "*nu vei face* cutare";
  şi folosind adresări pozitive/afirmative, de exemplu "*vei face* cutare".

Exemple de Prompt Template; aplicabilitate şi eficienţă în funcţie de tipul informaţiei solicitate

Instrucţiunile către un *LLM* în care interlocutorul uman îi indică modul în care trebuie să replice/răspundă şi în care să structureze replica/răspunsul se află în textul din *Prompt Template (PT)*.

La această dată (februarie 2024):

***Un PT trebuie să contribuie la reducerea** halucinaţiilor **LLM-ului cu care se dialoghează**,*
înainte aşadar ca interlocutorul uman să-i înainteze solicitările formulate adecvate.

Formularea unei solicitări trebuie să fie în acord cu:
- setul de informaţie accesibilă de către *LLM*, de exemplu Documentele Locale (enciclopedii ş.a.)
- capacitatea de analiză a informaţiei de către *LLM*, de exemplu unele modele pot parcurge numai text, pe când altele atât text, cât şi imagini

Un *PT* trimis către un *LLM* trebuie să îndeplinească măcar aceste condiţii:
- să fie formulat în expresii (propoziţii, afirmaţii) scurte, concise –
  - de exemplu, "Vei face o afirmaţie numai o dată"
- să conţină ideea principală referitoare la modul de alcătuire a replicii –
  - de exemplu, "Vei folosi numai sursele de informaţie unde se află conceptul cutare"
- să nu conţină indicaţii negative –
  - de exemplu, "Nu vei folosi cutare"
- să urmeze pe cât posibil valorile parametrilor:
  - *temperatura*
  - *Top P*

Comprimarea informaţiei cu care un *LLM* este instruit face încât adesea sau întotdeauna (depinzând de sursa de informaţie) valorile numerice (numere) şi în general secvenţele de cifre din textul original să fie întoarse în replică drept identice; de exemplu, distanţele până la cele mai apropiate cinci oraşe pot fi întoarse ca o singură valoare numerică.

Astfel:

1. Dacă solicităm informaţii ***cantitative** **care să includă numere** –*
   de exemplu, înălţimile celor mai înalte clădiri din oraşul cutare –
   ne asumăm riscul ca ***valorile din replică să nu fie cele reale***.

2. Dacă solicităm informaţii ***calitative** **care presupun prelucrarea unor numere** –*
   de exemplu, comparaţii între clădiri în funcţie de înălţimea acestora -
   de asemenea, ne asumăm riscul ca ***informaţiile să nu fie reale***.

3. Dacă solicităm informaţii ***descriptive** –*
   ***care nici să nu includă numere, nici să nu presupună prelucrarea unor numere** –*
   atunci în *PT* indicăm că vrem să fie prezentate numai descrieri care nu recurg la numere;
   în acest caz, este foarte probabil ca informaţiile primite să fie reale sau de încredere.

Câteva exemple de *PT* –
desigur, ajustabile în acord cu dorinţa interlocutorului uman şi cu informaţia accesibilă de către *LLM* –
adecvat situaţiei în care solicităm replici ***descriptive*** –
deci, unele ***alcătuite fără ca în decursul alcătuirii lor să se recurgă la numere***:

You think step-by-step, but do not tell me those steps.
You do not use information that you knew about the subject of my question or request before I asked anything of you.
You base your replies only on the search results and nothing else.
You make accurate statements using factual information only.
You use information that only exists in at least two sources of information where either the dominant topic is closely-related to the main subject of my request, or a separate topic is closely-related to the main subject of my request.
You double-check for logic your statements and replies.
You avoid using information that is not relevant to the topic at hand.
You do not repeat any statement which already is in your reply.
You use numbers if and only if they describe the same property of the same entity which is being discussed in at least two places.
You only mention or make comparisons if and only if they refer to the same two entities which are being discussed in at least two places.
You make statements which are logical and self-consistent; for example: something can not be hot and cold at the same time, something can not be round and square at the same time, something can not be liquid and gaseous at the same time, something can not fall down and rise up at the same time.
In the same reply, you do not make statements which contradict themselves or which contradict other statements.
You do not use information that does not exist in any source of information.

%1
Response:

Adică:

Gândeşti pas cu pas, dar nu îmi prezinţi acei paşi.
Nu foloseşti informaţie pe care o cunoşteai despre subiectul întrebării sau solicitării mele înainte ca eu să îţi fi cerut ceva.
Îţi bazezi replicile numai pe rezultatele căutării şi pe nimic altceva.
Faci numai afirmaţii precise folosind informaţii reale.
Foloseşti numai informaţie care există în cel puţin două surse de informaţie fie unde subiectul dominant este foarte apropiat de subiectul principal al solicitării mele, fie unde o secţiune separată este foarte apropiată de subiectul principal al solicitării mele.
Verifici de două ori logica afirmaţiilor şi replicilor tale.
Eviţi folosirea de informaţie care nu este relevantă pentru subiectul curent.
Nu repeţi nicio afirmaţie care deja există în replica ta.
Foloseşti numere numai şi numai dacă descriu aceeaşi proprietate a aceleiaşi entităţi despre care se discută în cel puţin două locuri.
Menţionezi sau faci comparaţii numai şi numai dacă se referă la aceleaşi două entităţi despre care se discută în cel puţin două locuri.
Faci afirmaţii care sunt logice şi care nu se autocontrazic; de exemplu: ceva nu poate fi cald şi rece în acelaşi timp, ceva nu poate fi rotund şi pătrat în acelaşi timp, ceva nu poate fi lichid şi gazos în acelaşi timp, ceva nu poate cădea şi ascende în acelaşi timp.
În aceeaşi replică nu faci afirmaţii care se contrazic pe sine sau care contrazic alte afirmaţii.
Nu foloseşti informaţie care nu există în nicio sursă de informaţie.

PT-urile prezentate aici au fost folosite în conjuncţie cu:

Temperature: 0.1
Top K: 50
Top P: 0.5

Clonarea unui LLM

În programul ***GPT4All***,
un model pe care îl considerăm performant sau ca având anumite calităţi poate fi ***clonat***
pentru ca fiecare clonă a sa

- *să fie modificată* încât să interpreteze rolul unui interlocutor *distinct*
- *să fie selectată* ca *model distinct* din lista modelelor disponibile, având *caracteristici proprii*

Astfel, fiecare clonă a unui model se pot atribui caracteristici
care o vor separa de restul *modelelor disponibile în program*.

De exemplu, modelul *Falcon* poate fi clonat pentru ca:

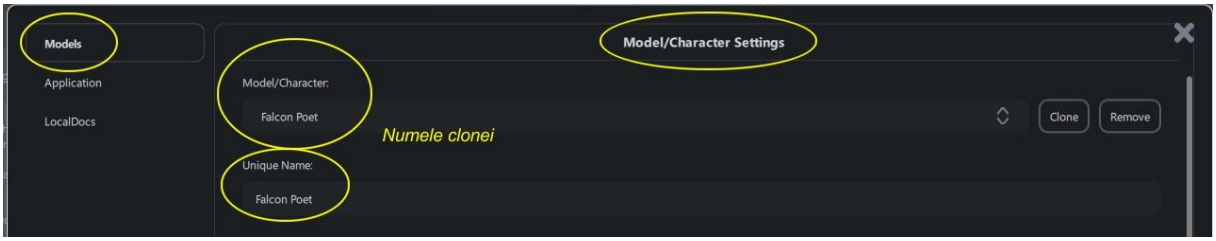
- o variantă a sa să fie una care să personifice / să comunice ca un specialist în astronomie
- altă variantă a sa să să personifice / să comunice ca un poet filosof.

Procedura este ilustrată în imaginile următoare:

- în dreapta-sus, click pe pictograma care deschide dialogul configurării programului
- în stânga, click pe "*Models*"
- din lista modelelor se selectează modelul care va fi *clonat*
- în dreapta, click pe butonul "*Clone*"



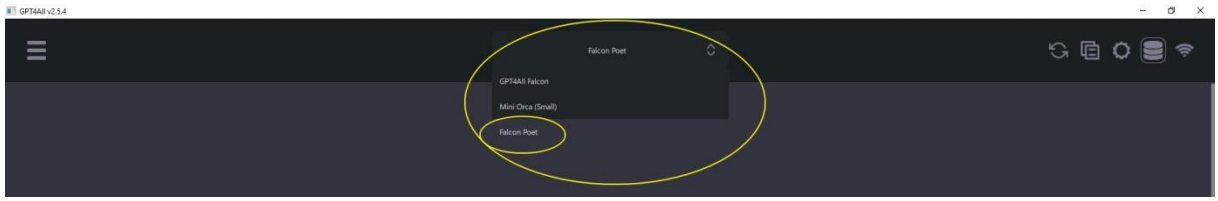
- dăm (atribuim) un nume clonei astfel create, de exemplu "Falcon Poet" în imaginea următoare:



De acum,
tot în secţiunea "*Models*" a dialogului de configurare a programului
îi putem atribui *acestui model-"clonă" caracteristici proprii*, de exemplu:

- *System Prompt*
- *Prompt Template*
- *Temperature*
- *Top P*

Apoi, *în pagina principală îl putem selecta din lista modelelor disponibile* şi să iniţiem o conversaţie:



Scenarii de utilizare

Cărţi masive pe teme ştiinţifice pot fi sumarizate în rezumate de câteva paragrafe cu trei sute de cuvinte;
din versurile unor cântece ale unei formaţii poate fi extrasă atitudinea acesteia faţă de ceva anume;
concepte filosofice dintr-o lucrare scrisă acum două mii de ani pot fi corelate cu unele moderne;
în general, un *LLM* poate fi folosit pentru:

- a prezenta mai mult sau mai puţin sumar/sintetic informaţii care până atunci erau vehiculate în "limbaj natural" ca texte scrise de om sau de alt *LLM*.
- a corela informaţii (din mai multe surse) pentru alcătuirea unei prezentări originale

Astfel se elimină nevoia omului de a citi zeci de cărţi cu mii de pagini din care unele informaţii trebuie sintetizate şi expuse în alt fel, încât un *LLM* pregătit pentru interpretarea textului se constituie nu numai într-un interlocutor, ci într-un asistent personal de atâta încredere pe cât de precise sunt informaţiile de care dispune şi pe cât de explicite sunt solicitările interlocutorului uman.

De exemplu, mai multe cărţi de astronomie pot fi plasate într-o categorie de Documente Locale pe care *LLM*-ul să o parcurgă şi din care să extragă informaţii despre cele mai apropiate cinci galaxii de Calea Lactee sau fenomenele din atmosfera unei stele asemănătoare Soarelui.

1. **Vom pregăti un interlocutor pentru discuţii despre cărţi beletristice ale căror autori sunt femei.**

   Acele cărţi sunt fişiere *PDF* stocate pe calculatorul personal sau pe un mediu de stocare conectat la acesta atunci când dialogăm cu *LLM*-ul.

   Aşadar, acele cărţi vor constitui o colecţie de Documente Locale
   pe care, desigur, trebuie să o facem cunoscută *LLM*-ului pentru a putea susţine o conversaţie.

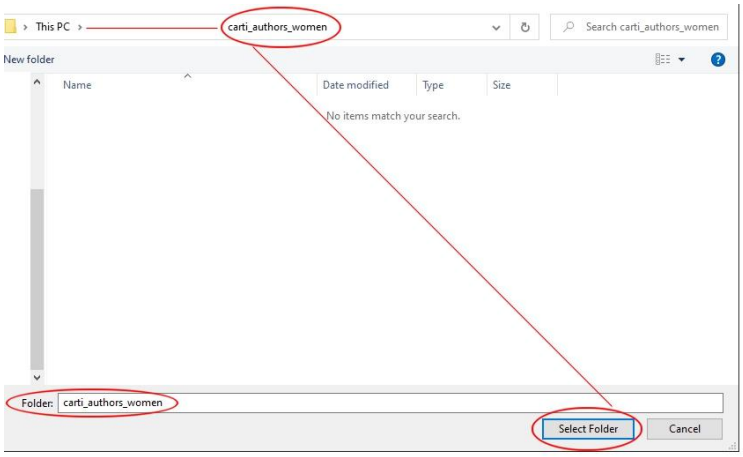
1.1. Creăm un folder/director dedicat acelor cărţi – acelei colecţii de Documente Locale.
1.2. În acel folder plasăm cărţile de interes



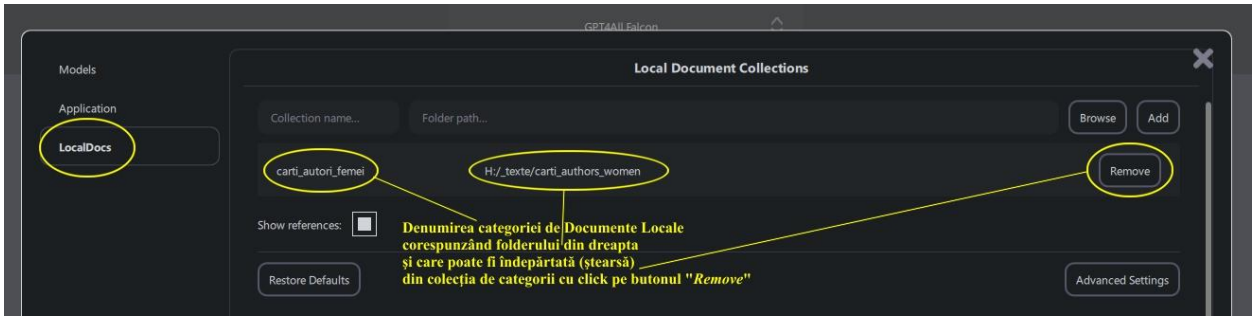
1.3. Adăugăm această colecţie la cele de Documente Locale pe care *LLM*-ul le va parcurge pentru susţinerea unei conversaţii:

   Click pe butonul marcat cu roata dinţată care invocă dialogul de configurare a programului.
   În acel dialog, în partea stângă a sa click pe "*LocalDocs*".

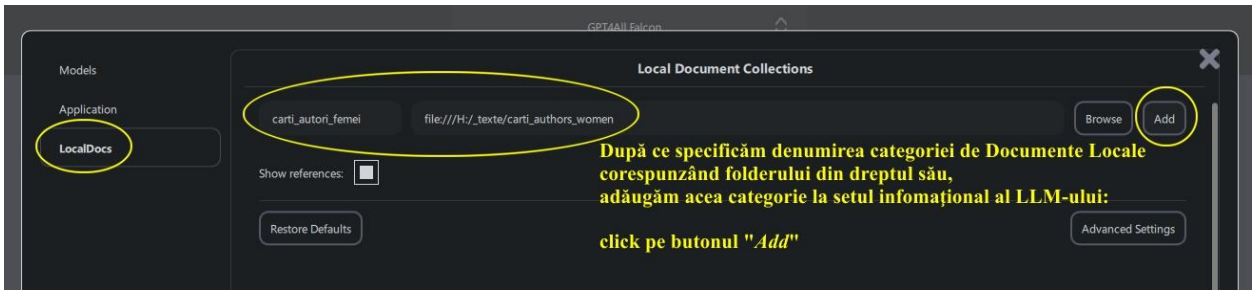
Pe această pagină
- specificăm folderul unde se află Documentele Locale din noua categorie:
  - click pe butonul "*Browse*"
  - dublu-click pe numele folderului respectiv
  - click pe butonul "*Select Folder*"



Apoi specificăm denumirea (numele) acestei (noii) categorii de Documente Locale:



- apoi apăsăm butonul "*Add*"



1.4. Închidem acest dialog şi astfel revenim la pagina principală a programului

1.5. Instruim *LLM*-ul curent (cel selectat) să parcurgă noua colecţie (sau pe oricare le-am fi adăugat):

- click pe butonul "*Local Doc(ument)s*":



1.6. Aşteptăm până când colecţia sau colecţiile de Documente Locale sunt indexate de către *LLM*:



1.7. Tot în acest dialog ("*Local Documents*")
bifăm categoriile de Documente Locale pe care vrem ca *LLM*-ul să le folosească în conversaţie:

Vrem să dialogăm despre cărţile scrise de femei, aşa că bifăm acea categorie pe care desigur am denumit-o în acord cu ce anume documente conţine:

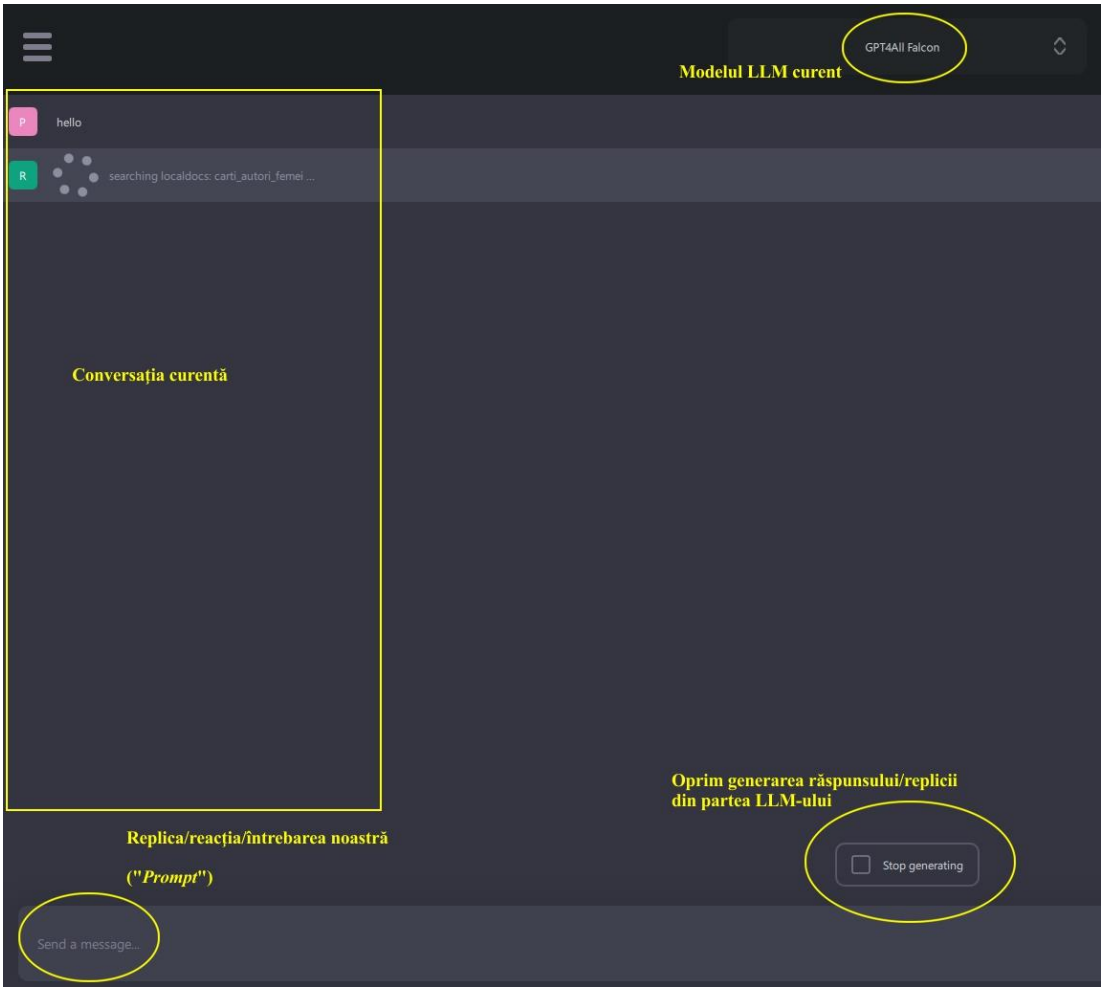


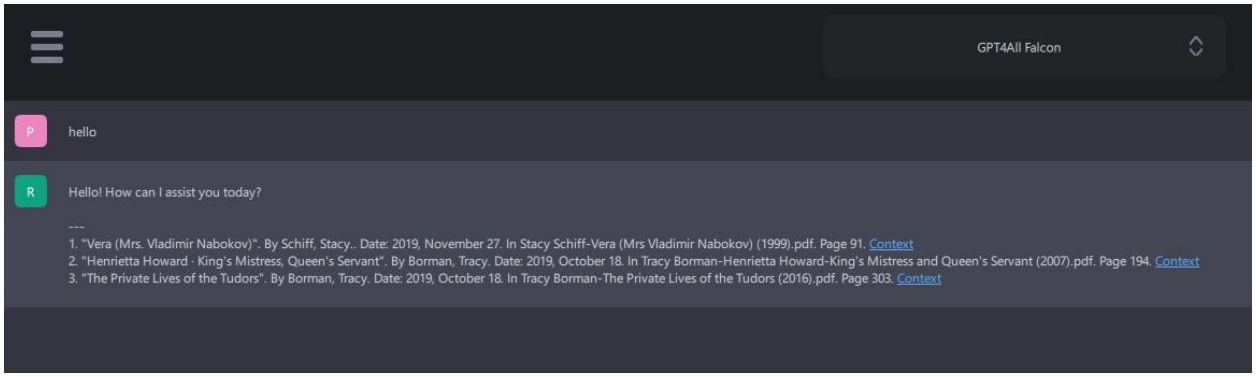
După ce bifăm categoriile de Documente Locale din care *LLM*-ul curent (cel selectat, în acest exemplu fiind modelul *Falcon*) va extrage informaţiile pentru susţinerea convesaţiei închidem acest dialog şi

1.8. Iniţiem dialogul cu *LLM*-ul.

Începem cu un salut simplu:
"hello"

LLM-ul va încerca să replice pe baza informaţiilor din Documentele Locale:
Se observă cum se caută un răspuns pe baza acelora:

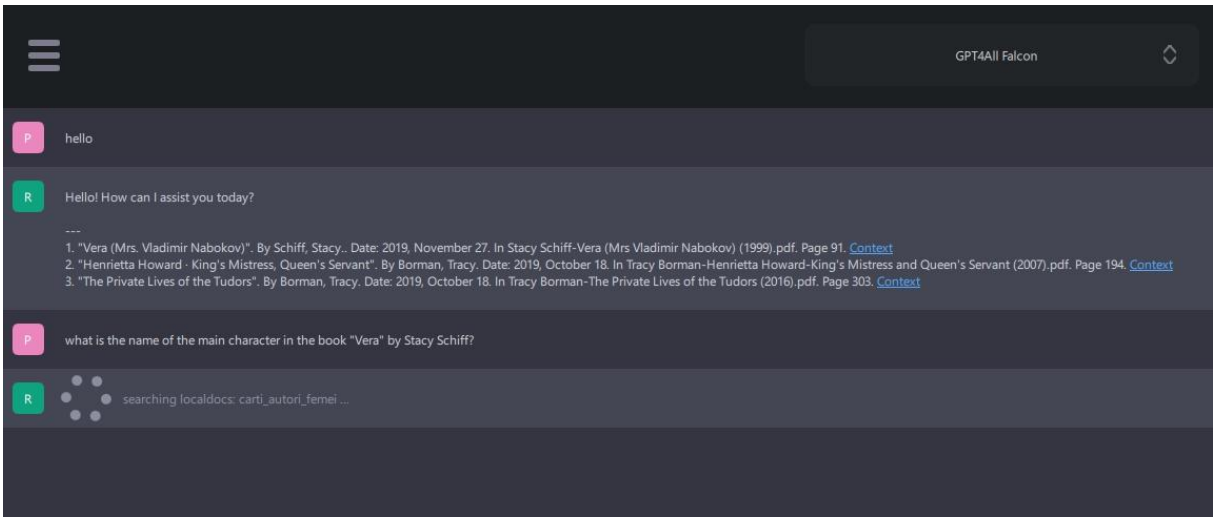




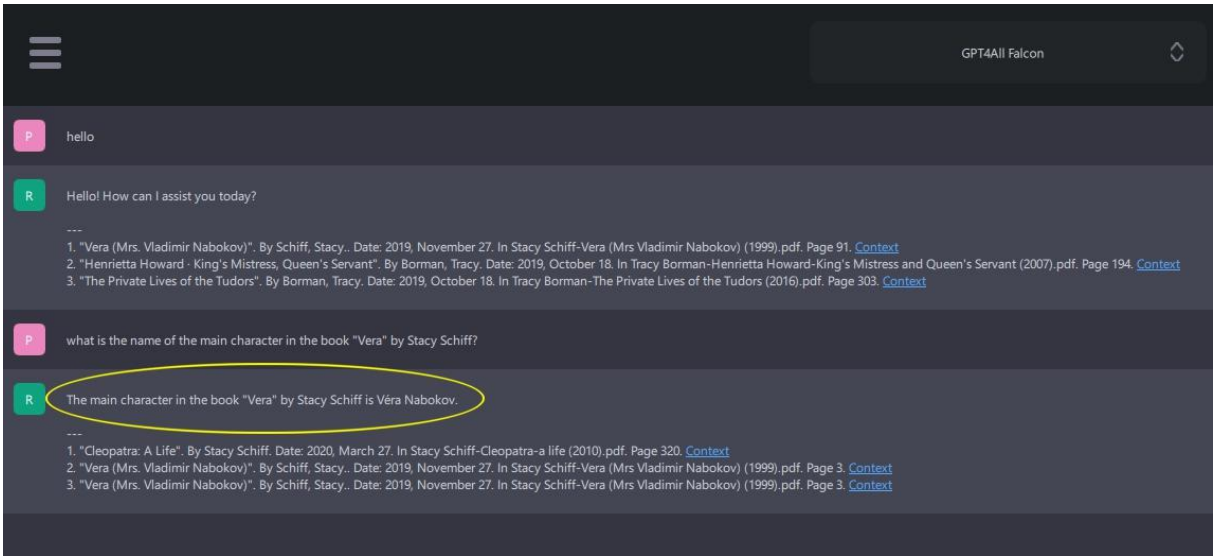
După schimbul de amabilităţi întrebăm ceva concret,
ştiind că detaliile întrebărilor/replicilor noastre sunt importante pentru obţinerea unui răspuns decent:

Vrem să întrebăm *LLM*-ul despre:
Care este personajul principal din cartea "Vera" scrisă de Stacy Schiff?
(What is the main character in the book "Vera" by Stacy Schiff?)

Mai explicit, mai detaliat, putem întreba:
Care este **numele** personajului principal din cartea "Vera" (scrisă) de Stacy Schiff?
What is **the name of** the main character in the book "Vera" by Stacy Schiff?



Răspunsul *LLM*-ului:



De acum putem solicita sumarul unei cărţi, subiectul/tema, ideea principală, lista personajelor din carte, replica unui personaj la ceea ce a spus altul şi diverse alte informaţii despre care bănuim sau despre care ştim că se află în colecţiile bifate de Documente Locale.

Se subînţelege că la fel putem proceda cu colecţii de alte feluri de cărţi,
încât de exemplu să avem la dispoziţie
un savant expert în meteorologie, astronomie, chimie, fizică, arhitectură ş.a.

2. **Vom pregăti un interlocutor pentru discuţii despre versurile cântecelor unei formaţii.**

Textele respective vor fi în fişiere *PDF* plasate într-un director dedicat.

Pentru a salva ca *PDF* astfel de texte, apelăm la site-uri care prezintă versurile unor cântece – de exemplu,

AZLyrics
https://www.azlyrics.com/



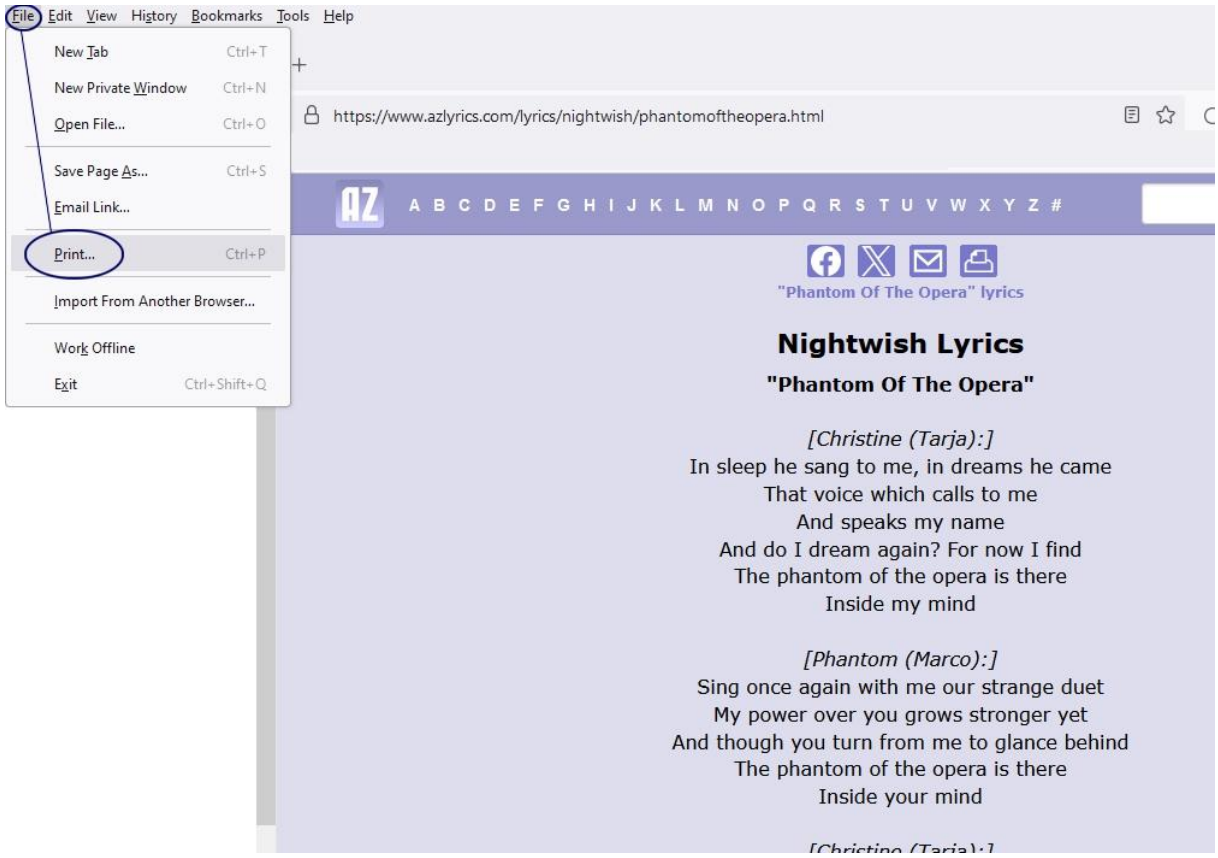
În acest exemplu, căutăm un cântec al formaţiei *Nightwish*.
Am *scrisNightwish* in caseta de căutare şi am apăsat butonul care porneşte căutarea cântecelor acestora.
Rezultatul posibil:



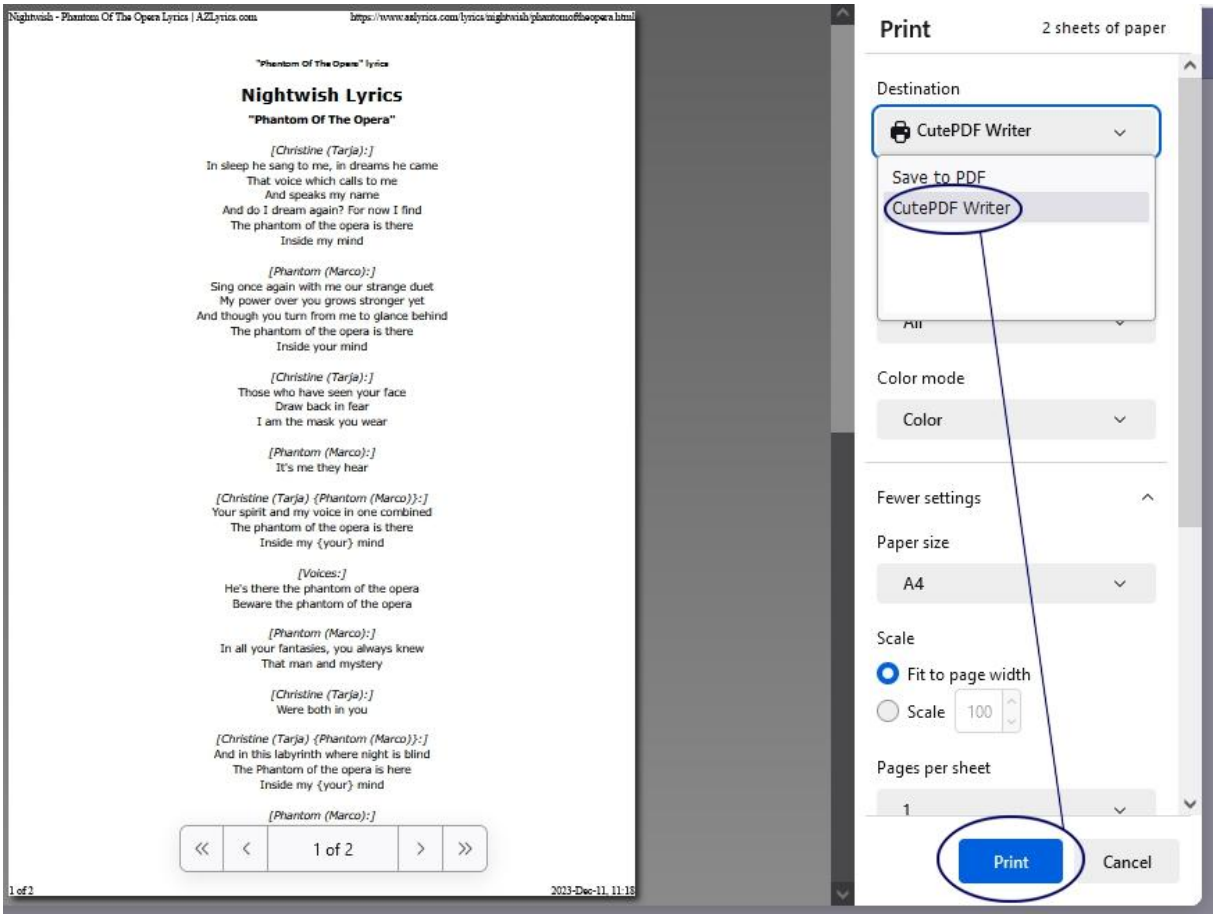
Click pe linkul spre versurile de interes, în acest exemplu cele ale cântecului "*Phantom Of The Opera*":

Salvăm această pagină ca fişier *PDF*:

Pentru aceasta putem folosi
"imprimanta" *CutePDF* prezentată mai devreme care "tipăreşte" în fişiere *PDF*:
"Tipărim" (salvăm ca fişier *PDF*) pagina la această "imprimantă":
Meniul *File -> Print*:



Selectăm "imprimanta" *PDF*:



Salvăm acest fişier *PDF* într-un folder pe care-l vom folosi drept categorie de Documente Locale,

apoi eventual repetăm această procedură cu versurile altor cântece,

apoi urmăm aceiaşi paşi ca înainte pentru a susţine o conversaţie cu *LLM*-ul care va asimila informaţiile din acele pagini-fişiere *PDF* după ce le va indexa.

Noţiuni; Termeni; Abrevieri

Adresele pe Internet prezentate în această secţiune erau accesibile în anul 2024, luna februarie.

AI / IA

Inteligenţă **A**rtificială / **A**rtificial **I**ntelligence

NLP

Natural **L**anguage **P**rocessing

Tehnica de Procesare a Limbajului Natural

Ramura *AI* preocupată de a oferi computerelor capacitatea de a înțelege textul și cuvintele vorbite (Limbajul Natural, uman) în același mod în care pot oamenii.

NLP combină lingvistica computațională – modelarea bazată pe reguli a limbajului uman – cu modele statistice, de învățare. Împreună, aceste tehnologii permit computerelor să proceseze Limbajul Natural sub formă de text sau date vocale și să înțeleagă sensul său complet, inclusiv intenția (și eventual) sentimentul scriitorului sau vorbitorului.

- https://www.ibm.com/topics/natural-language-processing

LLM

Large **L**anguage **M**odel

Model **V**ast de **L**imbaj / **M**odel de **L**imbaj de **M**ari **D**imensiuni

Ansamblu *software* care poate recunoaşte, sumariza, traduce, prevede/preconiza text şi alte tipuri de *conţinut*. Termenul "vast" provine din aceea, că modelele sunt *instruite* pe/cu cantităţi uriaşe de informaţii (de exemplu: biblioteci întregi, colecţii anuale de reviste ştiinţifice, ediţiile zilnice dintr-o lună ale unui ziar/jurnal ş.a.).

- https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part1

RAG

Retrieval **A**ugmented **G**eneration

Tehnică din domeniul *IA* care combină *preluarea* de text cu *generarea* de text, astfel permiţând modelelor *IA* să preia

informaţie din surse de cunoştinţe separate de colecţia pe care a fost instruit

şi să o *încorporeze în textul generat*.

- https://www.datastax.com/guides/what-is-retrieval-augmented-generation
- https://aws.amazon.com/what-is/retrieval-augmented-generation/

Training / Instruire

Procesul prin care un *LLM* este alimentat cu informaţii de regulă în format text (pentru interpretarea limbajului natural), de exemplu: cărţi/lucrări literare/articole ştiinţifice ş.a. în format electronic (*PDF*, *DOCX* ş.a.) –

proces în urma căruia acel *LLM* va recunoaşte multitudinea de cuvinte, expresii şi particule de expresii şi combinaţiile tuturor acestora din Limbajul Natural, încât pe baza comparării replicilor interlocutorului (uman) cu expresiile cunoscute va putea recunoaşte, extrage şi asambla informaţii din cele cu care a fost alimentat – încât astfel să poată formula expresii (afirmaţii, replici, răspunsuri) tot în limbaj natural.

Token

Particulă de text: un caracter , o silabă, un cuvânt, două sau mai multe cuvinte alăturate, o propoziţie, o frază şamd. În practica obişnuită, un *token* constă din 3-4 litere.
Un text parcurs de un *LLM* este secţionat în *token*-uri care sunt stocate în *LLM* pentru a le compara cu replicile interlocutorului uman.

Chunking / Secţionare

Procedura prin care un text este împărţit în secţiuni (*chunks*, bucăţi) constând în caractere literale sau în *token*-uri. Separarea se poate face atât în raport cu caractere sau litere comune, cât şi în raport cu *token*-uri mai complexe (cuvinte, expresii, propoziţii, fraze ş.a.).
Secţionarea este aplicată/folosită în timpul instruirii unui *LLM* care parcurgând texte construieşte lista *token*-urilor întâlnite în acele texte pentru ca apoi să poată evalua probabilitatea ca un *token* să-i urmeze altuia, iar astfel să înţeleagă replicile interlocutorului uman şi să le compare cu cele cunoscute din acea listă încât să formuleze replici în Limbaj Natural şi în funcţie de contextul derivat din rezultatele secţionării premergătoare.
Tehnica implică stabilirea judicioasă (depinzând de sursa de informaţii) a doi parametri:

1. dimensiunea/lungimea şirului de caractere parcursă/citită/interpretată la o trecere, de exemplu textul este secţionat în bucăţi de câte 1000 caractere.
2. suprapunerea ("*overlapping*") şirurilor de caractere astfel parcurse şi interpretate; parametru util pentru identificarea *contextului* unui *token* – rostul său în contextul lingvistic imediat, de exemplu valoarea 200 indică *LLM*-ului că în timpul parcurgerii unei surse de informaţie trebuie să suprapună (*overlap*) 200 caractere din (de la sfârşitul) fiecărei bucăţi de text pe o bucată de text alăturată.

Conţinut

Entitate informaţională care poate fi stocată (şi eventual modificată) şi eventual prezentată, de exemplu: text, imagini, fişiere.

Software

Programe pentru/de calculator/computer/dispozitive de calcul;
fie entitatea executabilă, fie instrucţiunile (codul) din care aceea este alcătuită.

UX

User E**x**perience

Calitatea interacţiunii utilizatorului cu un ansamblu *software*.
Include (în primul rând) calitatea interfeţei care face posibilă comunicarea bidirecţională.

Resurse pe Internet

Adresele pe Internet prezentate în acest document erau accesibile în anul 2024, luna februarie.

1. Carte în format *PDF*, autor: **NVIDIA**

   A Beginner's Guide to Large Language Models
   https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part1

   How LLMs are Unlocking New Opportunities for Enterprises
   https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part2

2. Pagină a **IBM**:

   What is Natural Language Processing?
   https://www.ibm.com/topics/natural-language-processing

3. Pagină a **DATASTAX**

   Retrieval Augmented Generation (RAG) - A Comprehensive Guide
   https://www.datastax.com/guides/what-is-retrieval-augmented-generation

4. Pagină a **Vectara**:

   Retrieval Augmented Generation (RAG) Done Right: Retrieval
   https://vectara.com/retrieval-augmented-generation-rag-done-right-retrieval/

10 Best AI Apps (December 2023)
https://www.unite.ai/10-best-ai-apps/

Prompt Engineering Guide
https://www.promptingguide.ai/

The Curious Case of LLM Hallucination: Causes and Tips to Reduce its Risks
https://www.simform.com/blog/llm-hallucinations/

Informaţii detaliate precum cele enumerate aici despre *LLM, NLP* şi alte concepte din domeniul *IA* se află pe site-ul **Medium** (unele pagini pot fi citite numai după autentificare ca abonat la site sau ca utilizator al *GMail* / *Facebook* ş.a.):

What is LLM? Understanding with Examples
https://medium.com/@seaflux/what-is-llm-understanding-with-examples-da971392fb21

Introduction to LLMs and their Applications
https://medium.com/ml-and-ai-for-non-data-scientists/introduction-to-llms-and-their-applications-f68d31c8d0fd

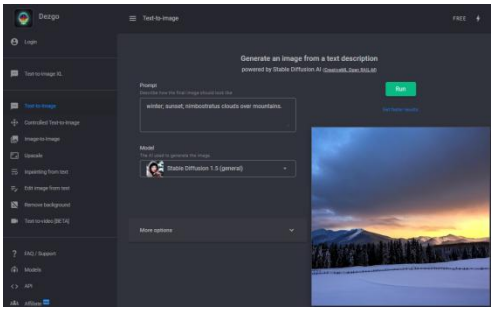
6 ways to reduce hallucination using prompt engineering
https://medium.com/@masteringllm/6-ways-to-reduce-hallucination-using-prompt-engineering-b30eb244b9b2

Understanding Retrieval-Augmented Generation: A Simple Guide
https://medium.com/@amodwrites/understanding-retrieval-augmented-generation-a-simple-guide-d638ac92c123

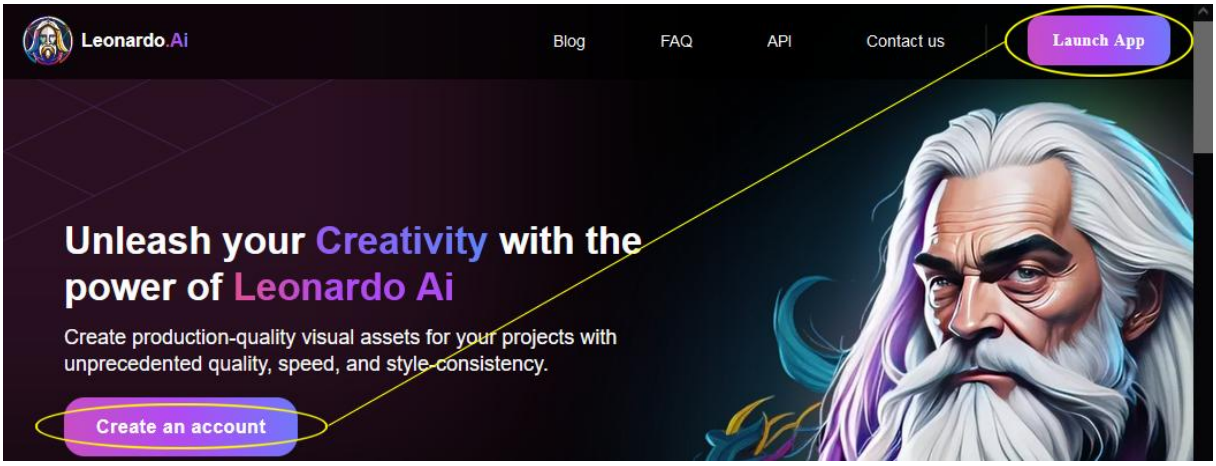
Spread Your Wings: Falcon 180B is here
https://huggingface.co/blog/falcon-180b

Generare de imagini din text (din descrieri):

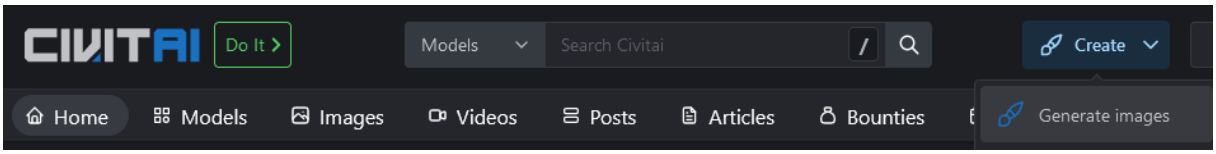
DEZGO
https://dezgo.com/



Leonardo.ai
https://leonardo.ai/



CIVITAI
https://civitai.com/



*Acest document este supus modificării/actualizării **frecvente***
în acord cu starea şi evoluţia celor prezentate
şi poate fi preluat de la adresa:
http://tinyurl.com/4h3td2n3