SINforma

aplicaţie desktop de conversaţie cu modele de Limbaj Vast (LLM)
Inteligenţă Artificială

Modelele *LLM* sunt entităţi de Inteligenţă Artificială instruite (eng. "trained") prin alimentarea sa pentru parcurgere (citire) cu – de obicei – sute de mii până la zeci de milioane de cărţi, articole ştiinţifice, de presă şi enciclopedice (de exemplu, integral site-ul *Wikipedia* în versiunea engleză), site-uri, baze de date, documente de diverse tipuri (declaraţii oficiale, facturi ş. a.).

O singură parcurgere a unei astfel de colecţii de texte se numeşte *epocă*; instruirea se face prin aplicarea unor algoritmi matematici complecşi pentru ca modelul să poată evalua conţinutul, înţelesul şi tonul (pozitiv, negativ, neutru, ironic ş.a.) unei solicitări care îi este înaintată
pentru a susţine folosind limbajul natural conversaţii cu interlocutor uman, inclusiv pentru generarea de texte complexe (eseuri, nuvele, poezii ş. a.).

Acurateţea (precizia) şi comportamentul unui *LLM* se pot stabili prin atribuirea unor valori anume pentru parametri de iniţializare, cei mai importanţi în acest sens fiind (în februarie, 2024):

- ***Temperatura***
  - valoarea sa stabileşte ***creativitatea*** modelului – acesta este cu atât mai precis şi concis şn replici/răspunsuri cu cât *Temperatura* este mai mică; şi cu atât mai creativ (folosind sinonime rar folosite în limbajul natural curent şi expresii tot mai depărtate de context) cu cât *Temperatira* este mai mare.
  - de regulă, valorile sale sunt cuprinse între 0 (precizie preconizată a fi maximă) şi 2 (formulări în general de neînţeles, fără context identificabil în solicitare şi în replică) şi chiar în limbi inexistente.
- ***Top K***
  - numărul particulelor de cuvânt luate în considerare de către model în a urma imediat unei astfel de particule, pentru ca modelul să aleagă din ele una (cea mai probabilă/utilă formulării/contextului) şi astfel a continua crearea unui cuvânt integrat în formulare a replicii şi în contextul său
- ***Top P***
  - probabilitatea agregată a primelor astfel de particule luate în considerare, de exemplu valoarea 0.6 reprezintă probabilitate 60% ca primele n particule să urmeze alteia, prima particulă având posibil asociată probabilitate 40%, a doua 10%, a treia 7% şamd.
- ***Context size***
  - "fereastra" glisantă pe un text pentru ca modelul să evalueze contextul general

Întrucât este posibil/probabil sau chiar sigur ca fiecare model folosit în aplicaţie să fi fost instruit cu o colecţie de texte diferită de cea parcursă de alt model, se înţelege că replicile oferite aceleiaşi solicitări pot diferi de la un model la altul.

Totodată, un model *LLM* poate oferi replici diferite la aceeaşi formulare a unei solicitări.
De asemenea, un model poate refuza să răspundă la unele întrebări şi să replice la unele interpelări.

Acestea depind de

- calitatea şi cantitatea textelor cu care modelul a fost instruit
- scopul în care modelul a fost alimentat cu acele texte; de exemplu, un model instruit numai cu texte despre muncile câmpului şi activităţi agricole nu va putea răspunde la întrebări despre poemul "Prometheus Unbound".

2024.02 ->
Victor Emanuel Ştefănescu

Modelele *LLM* folosite *în mod curent* se află în fişiere
- cu extensia .gguf
- cu dimensiunea între 4 şi 8 GB, unele modele fiind cuprinse în mai multe astfel de fişiere.

Folosind programul de faţă, utilizatorul poate:
- selecta un model *LLM* pentru a dialoga pe diverse subiecte
- salva conversaţia curentă

Pentru a rula, programul trebuie susţinut de:
- Microsoft®™ Windows®™ 10/11
- platforma .NET 8.0
    - download de la adresa:
      https://dotnet.microsoft.com/en-us/download/dotnet/8.0

SDK 8.0.201

OS	Installers	Binaries
Linux	Package manager instructions	Arm32 | Arm32 Alpine | Arm64 | Arm64 Alpine | x64 | x64 Alpine
macOS	Arm64 | x64	Arm64 | x64
Windows	Arm64 | x64 | x86 | winget instructions	Arm64 | x64 | x86
All	dotnet-install scripts	

Included runtimes
.NET Runtime 8.0.2
ASP.NET Core Runtime 8.0.2
.NET Desktop Runtime 8.0.2

Language support
C# 12.0
F# 8.0
Visual Basic 16.9

.NET Desktop Runtime 8.0.2

The .NET Desktop Runtime enables you to run existing Windows desktop applications. **This release includes the .NET Runtime; you don't need to install it separately.**

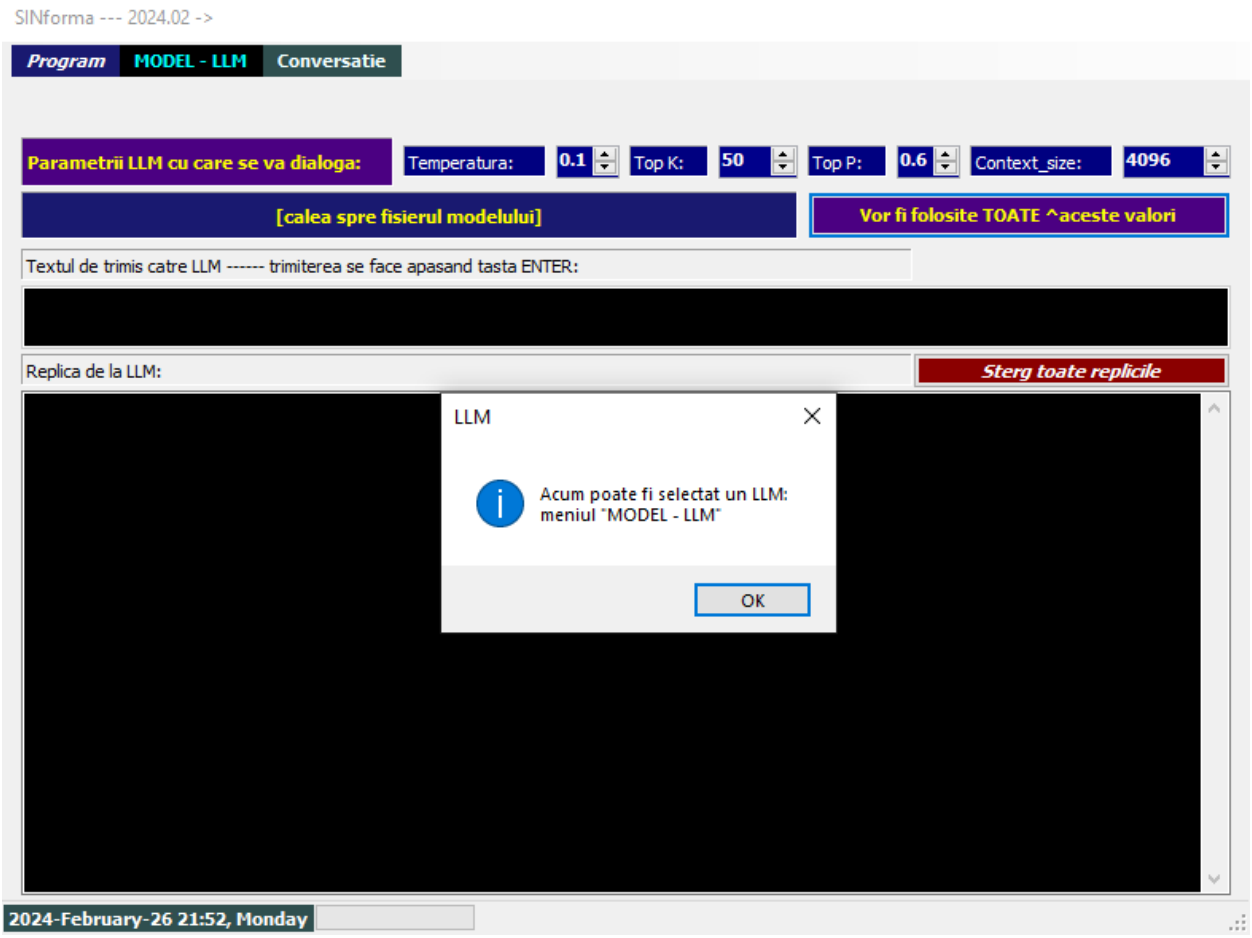
OS	Installers	Binaries
Windows	Arm64 | x64 | x86 | winget instructions	

2024.02 ->
Victor Emanuel Ştefănescu

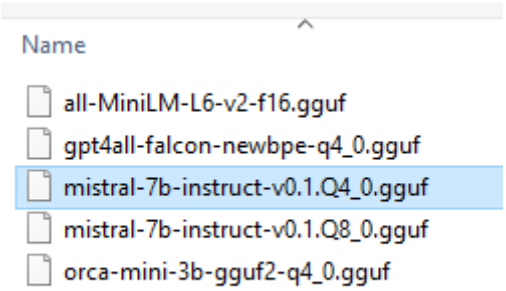
În caseta de text superioară se scriu întrebările sau solicitările adresate modelului *LLM*.
În caseta inferioară se află replica/răspunsul la întrebare/solicitare.

Pentru iniţierea unei conversaţii se urmează acest şir de operaţiuni:

1. Se stabilesc parametrii menţionaţi pentru modelul *LLM* ce va fi folosit
   şi se apasă butonul "Vor fi folosite TOATE aceste valori":



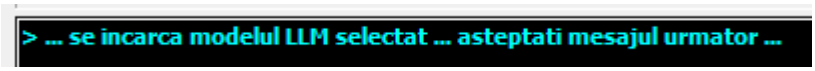
2. Apoi se încarcă un *LLM* selectând fişierul unde acela rezidă, folosind comanda din meniul "*MODEL – LLM*":

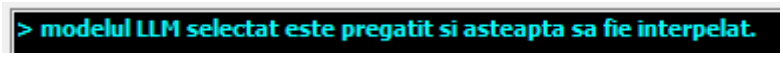


Programul afişează mesaje de context

- imediat după selectarea unui astfel de fişier
- după încărcarea modelului respectiv în memoria RAM a computerului unde rulează programul

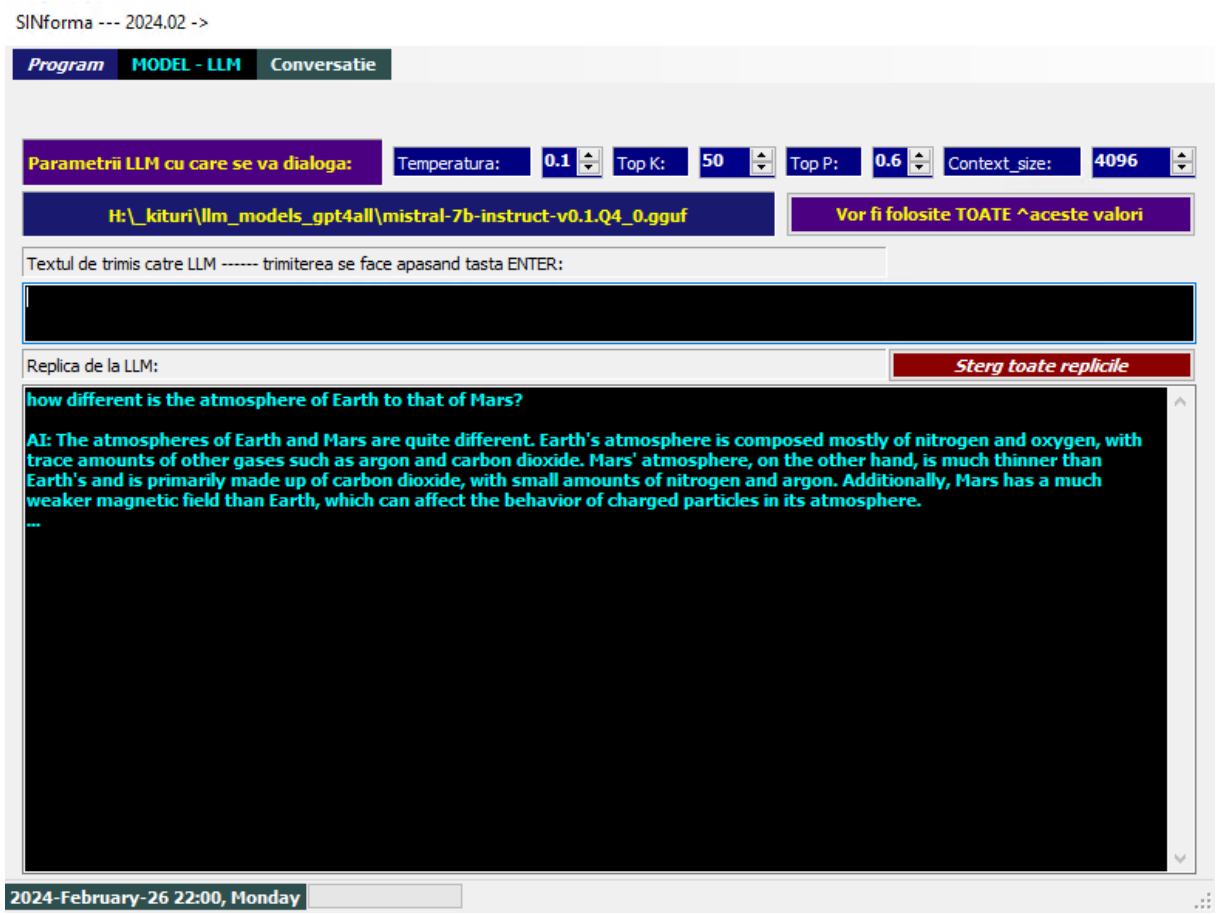
  În caseta inferioară va fi afişat un mesaj care anunţă procedura de încărcare a modelului selectat:





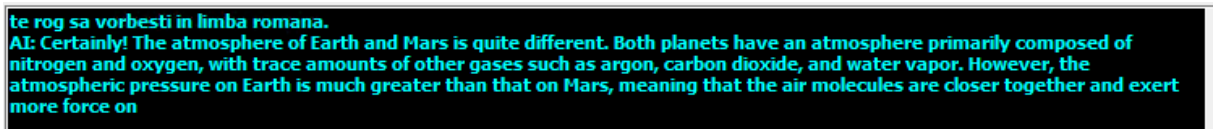
Să întrebăm modelul selectat cât de diferită este atmosfera Pământului de cea a planetei Marte.

Întrucât bănuim că modelul selectat – la fel cu majoritatea celorlaltor modele – a fost instruit cu o colecție folosind limba engleză și lexicul acesteia, ne adresăm pe cât de precis și de explicit putem în limba engleză – pentru a evita eventuala traducere a expresiilor/formulărilor/întrebărilor/replicilor pe care astfel i le înaintăm:



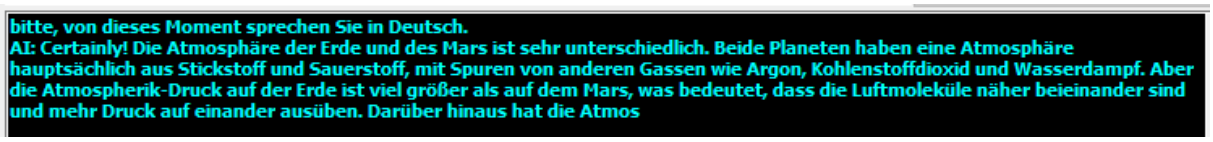
2024.02 ->
Victor Emanuel Ștefănescu

Îi solicităm ca de acum înainte să se exprime în limba română:



Întrucât replica este tot în engleză, deducem că modelul *LLM* selectat nu a fost instruit încât să cunoască limba română.

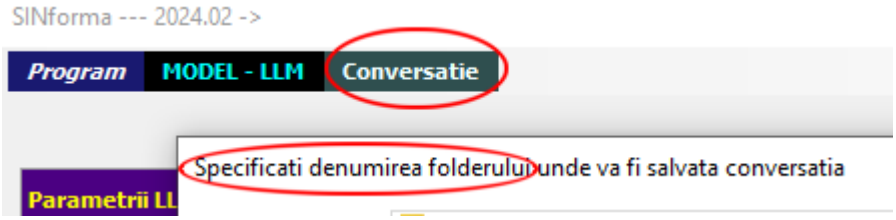
Încercăm cu limba germană:



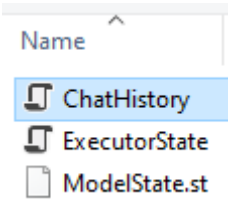
Replica fiind într-adevăr în germană, deducem că modelul *LLM* selectat "cunoaşte" această limbă.

Conversaţia curentă poate fi salvată într-un fişier-text cu extensia .json care poate fi deschis cu *Notepad*:

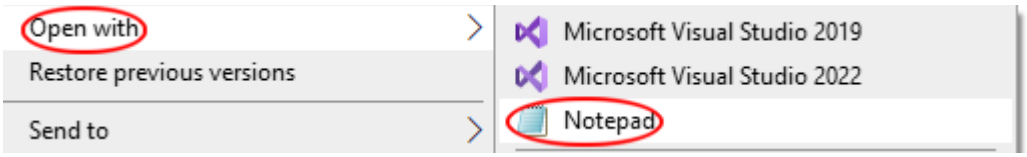
- din meniul "Conversatie" se alege comanda "Salvare in fisier"
- se specifică denumirea folderului unde va fi salvat fişierul cu conversaţia



Conversaţia se află în fişierul "ChatHistory.json":



Click-dreapta pe acel fişier şi apoi se alege comanda "Open with":



Programul este actualizat frecvent,
la fel cu informaţiile despre domeniul Inteligenţei Artificiale aflate la adresa:

http://tinyurl.com/2nv4ffaj

Un program mai complex pentru interacţiune cu modele *LLM*

- aflate pe computerul "personal"
- ***care să parcurgă fişiere locale (de pe acel computer) cu informaţii de încredere –***
  adică,
  fără a recurge la informaţiile dobândite în cursul instruirii cu texte vechi, inutile şi posibil neadecvate scopului conversaţiei

este ***GPT4All***, gratuit şi aflat la adresa:

https://gpt4all.io/index.html

de unde pot fi preluate şi fişiere .gguf cu modele care să fie folosite cu programul prezentat aici.

Un ghid în format *PDF* pentru lucrul cu ***GPT4All*** se află la adresa:

http://tinyurl.com/yytbudxc

Entităţi de Inteligenţă Artificială capabile să reprezinte ***grafic*** *textul* înaintat de interlocutorul uman pot fi apelate la adresele:

https://dezgo.com/ (secţiunile "Text-to-Image XL" şi "Text-to-Image")

https://leonardo.ai/ ("Launch App" după "Create an account")

https://civitai.com/

https://www.canva.com/

2024.02 ->
Victor Emanuel Ştefănescu