

Panorama des outils propriétaires et open source

ANF TDM & IA 2025

Octobre 2025

Adrian CHIFU (amU, LIS)



LABORATOIRE
D'INFORMATIQUE
& DES SYSTÈMES

UMR 7020



Qui suis-je ?

Adrian-Gabriel CHIFU

Enseignant-Chercheur (MCF)

🎓 @amU (responsable du parcours M2 MIAGE I²D, responsable communication pour la MIAGE)

🔍 @LIS (co-responsable de l'équipe R2I)

Domaine de recherche : Recherche d'Informations

Intérêts : RI, TAL, apprentissage automatique, bases de données (non-relationnelles)



Sommaire

 **Du poisson rouge à l'éléphant : l'évolution magique**

 **Bataille des titans : qui règne sur quoi ?**

 **L'ère multimodale : quand l'IA devient créative**

 **Reprendre le pouvoir : l'IA chez soi**

La révolution ChatGPT

Novembre 2022 : Le déclencheur

- **100 millions d'utilisateurs** en 2 mois
- Démocratisation de l'IA générative
- **Course effrénée** des géants tech

⚠ *"Si un produit est gratuit, c'est vous le produit !"*

the

successor_map

```
'the' → ['strange',  
         'door',  
         'lawyer',  
         'wine',  
         'after-dinner',  
         'acts',  
         ...]
```

Évolution historique - Les pionniers

Modèles n-grammes (1950-2000)

"Je mange une" → $P(\text{pomme} | \text{mange une}) = ?$

Principe : Prédiction basée sur les **mots précédents**

Analogie ludique 🎲

"Comme prédire la fin d'une phrase en ne connaissant que les 2 derniers mots"

the

successor_map

```
'the' → ['strange',  
         'door',  
         'lawyer',  
         'wine',  
         'after-dinner',  
         'acts',  
         ...]
```

Évolution historique - Les pionniers

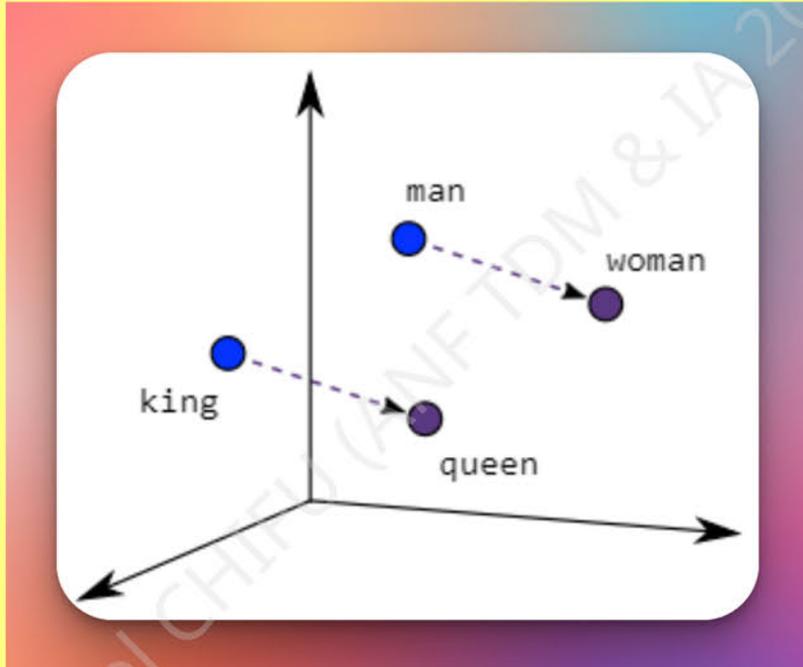
Modèles n-grammes (1950-2000)

"Je mange une" → $P(\text{pomme} | \text{mange une}) = ?$

Principe : Prédiction basée sur les **mots précédents**

Analogie ludique 🎲

"Comme prédire la fin d'une phrase en ne connaissant que les 2 derniers mots"



L'ère neuronale (2000-2017)

Révolutions successives

- **Word2Vec** : mots \rightarrow vecteurs
- **LSTM** : mémoire à long terme
- **Attention** : focus sélectif

Métaphore 🧠

"Du cerveau de poisson rouge au cerveau d'éléphant"

La révolution Transformer



"Attention is All You Need" (2017)

Innovations clés

- **BERT** (2018) : compréhension bidirectionnelle
- **GPT-1** (2018) : génération autorégressive
- **Émergence LLMs** : >1 milliard paramètres

Analogie 🙄

- *"Saisir toute la phrase d'un coup d'œil"*

ChatGPT - Le pionnier

ChatGPT (OpenAI)

- <https://chat.openai.com>

✓ **Forces** : Interface intuitive, DALL-E intégré

✗ **Faiblesses** : Hallucinations, coût 20\$/mois

🎯 **Usage** : Création contenu, programmation, assistant général



Les concurrents américains

Claude (Anthropic)

- <https://claude.ai>

✓ Sécurité, raisonnement contextuel
Gemini (Google)

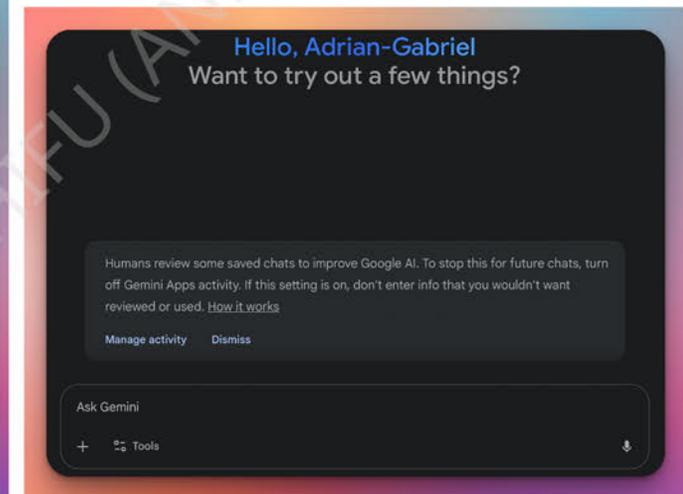
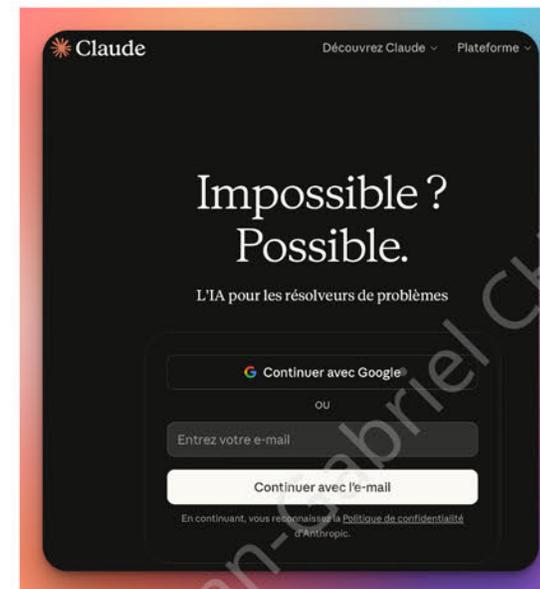
- <https://gemini.google.com>

✓ Multimodalité, intégration Google
Perplexity AI

- <https://perplexity.ai>

✓ Sources citées, recherche temps réel

💡 **Astuce** : 1 an Perplexity Pro gratuit avec mon opérateur internet !



Mistral AI – L'exception française

Mistral AI 🇫🇷

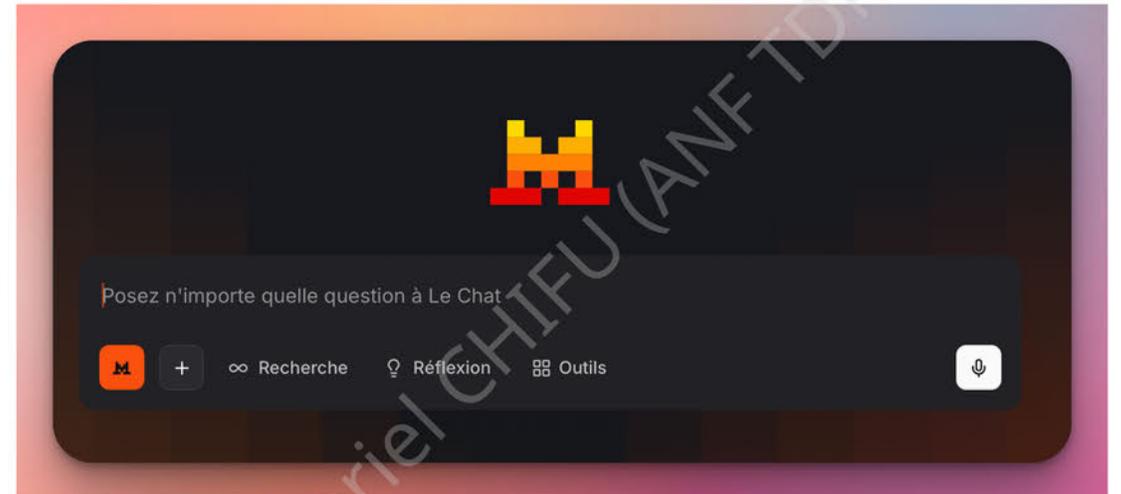
- <https://chat.mistral.ai>

🎯 **Mission** : Souveraineté numérique européenne

Modèles : Mistral 7B, Mixtral 8×7B, Mistral Large

✅ **Forces** : Open source, RGPD, multilingue

💡 **Astuce** : 1 an Mistral Pro gratuit avec mon opérateur téléphonie !



Classements



Classements - Raisonnement & Maths

Meilleurs en raisonnement (GPQA Diamond)

- **Grok 4** : 87.5%
- **GPT-5** : 87.3%
- **Gemini 2.5 Pro** : 86.4%

Mathématiques lycée (AIME 2025)

- **GPT-5** : 100% (perfection !)
- **GPT oss 20b** : 98.7%

 **Source** : <https://www.vellum.ai/llm-leaderboard>

Classements - Code & Outils

Codage agentic (SWE Bench)

- **Grok 4** : 75%
- **GPT-5** : 74.9%
- **Claude Opus 4.1** : 74.5%

Utilisation d'outils (BFCL)

- **Llama 3.1 405b** : 81.1%
- **Llama 3.3 70b** : 77.3%
- **GPT-4o** : 72.08%

Classements - Performance & Rapidité

Classement général

- **GPT-5** : 35.2%
- **Grok 4** : 25.4%
- **Gemini 2.5 Pro** : 21.6%

Plus rapides

- **Llama 4 Scout** : 2600 tokens/s
- **Llama 3.3 70b** : 2500 tokens/s

Classements - Coût & Latence

💰 Modèles économiques (par 1M tokens)

Nova Micro : \$0.04 / \$0.14

Gemma 3 27b : \$0.07 / \$0.07

Gemini 1.5 Flash : \$0.075 / \$0.3

🚀 Latence la plus faible

Nova Micro : 0.3s

Llama 3.1 8b : 0.32s

Équation :

Performance × Vitesse ÷ Coût = Valeur

Amazon Nova Foundation Models

Amazon Nova is a new generation of foundation models with frontier intelligence and industry leading price performance. Generate text, code, and images with natural language prompts

Thank you for your interest. At this time, Amazon Nova models cannot be accessed from this website in your location.

[Learn more about accessing via Amazon Bedrock here.](#)

Nova Micro

Fast, text-only model

A text-only model that delivers the lowest latency responses.

[Learn More](#)

Nova Lite

Understands text, images and video

A multimodal model that is lightning fast for processing image, video, and text inputs.

[Learn More](#)

Nova Pro

Best combo of quality and speed

A highly capable multimodal model with the best combination of quality and speed for a wide range of tasks.

[Learn More](#)

Nova Premier

Our most capable multimodal model

Our most capable model for complex tasks and teacher for model distillation on Amazon Bedrock.

[Learn More](#)

Multimodal - Audio avec Suno AI

Suno AI

- <https://suno.com>
- "Créer de la musique en quelques secondes"

🎯 **Principe** : IA générative pour **musique complète**

🎵 **Capacités Suno 4.5+**

Chansons jusqu'à 8 minutes

Paroles + instrumentaux

100+ styles : pop, jazz, rock, métal, électro, folk, etc.

Qualité professionnelle

💰 **Prix** : 50 crédits gratuits, puis 10\$/mois (500 chansons)

🎬 **Usage** : Bandes sonores, YouTube, marketing, éducation



Multimodal - Audio avec Suno AI

Suno AI

- <https://suno.com>
- "Créer de la musique en quelques secondes"

🎯 **Principe** : IA générative pour **musique complète**

🎵 **Capacités Suno 4.5+**

Chansons jusqu'à 8 minutes

Paroles + instrumentaux

100+ styles : pop, jazz, rock, métal, électro, folk, etc.

Qualité professionnelle

💰 **Prix** : 50 crédits gratuits, puis 10\$/mois (500 chansons)

🎬 **Usage** : Bandes sonores, YouTube, marketing, éducation



Multimodal - Vidéo/image avec Nano Banana

Nano Banana AI

<https://aistudio.google.com>

- "L'éditeur d'images mystère de Google"

 **Principe** : Édition d'images par **langage naturel**

 **Capacités clés**

- **Édition locale précise**
- **Cohérence personnages >90%**
- **Génération texte→image**
- **Transfert de style** artistique

 **Exemple magique** : "Changez sa robe rouge en soie bleue étincelante, style cyberpunk, rue néon pluvieuse"

 **Intégration** : Via Gemini 2.5 Flash + Watermark SynthID

Multimodal - V

Nano Banana AI

<https://aistudio.google.com>

- "L'éditeur d'images mystère de Google"

🎯 **Principe** : Édition d'images par langage

🖼️ **Capacités clés**

- Édition locale précise
- Cohérence personnages >90%
- Génération texte→image
- Transfert de style artistique

💡 **Exemple magique** : "Changez sa robe en cuir, ses cheveux en rose, style cyberpunk, rue néon"

⚡ **Intégration** : Via Gemini 2.5 Flash +



Generate an image of me (attached image) change my instrument into a sitar, I want pink hair, punk hairstyle and cyber punk context.

Model



Nano Banana

Multimod

Nano Banana AI

<https://aistudio.google.com>

- "L'éditeur d'images mystérieux"

🎯 **Principe** : Édition d'images

🖼️ **Capacités clés**

- Édition locale précise
- Cohérence personnage
- Génération texte → image
- Transfert de style artistique

💡 **Exemple magique** : "Change mon image en style cyberpunk, cheveux roses, style punk"

⚡ **Intégration** : Via Gemini



Generate an image of me (attached image) change my instrument into a sitar, I want pink hair, punk hairstyle and cyber punk context.

Model



Banana

NotebookLM - Assistant recherche

NotebookLM (Google)

- <https://notebooklm.google.com>

 **Principe** : IA alimentée par **VOS documents**

Fonctionnalités

- PDF, Google Docs, URLs, YouTube
- **Podcasts audio** automatiques
- Synthèses, FAQ, cartes mentales

 **Innovation** : Génération **podcast conversationnel** automatique

Limite : 50 sources max

Merlin – L'agrégateur

Merlin AI

- <https://getmerlin.in>
- "26-en-1 Chrome extension"

Capacités

- Résumés vidéos YouTube
- Chat avec sites web
- 26 outils IA intégrés
- Modèles multiples : GPT, Claude, Gemini, etc.

 **Prix** : 102 requêtes/jour gratuit, 19\$/mois Pro (j'ai trouvé une offre à 5\$/mois 🤪)

Search models

All

Pro

Reasoning

Coding

Writing

Speed



Auto Select Model

Let Merlin auto-select the best model with web search



GPT 4.1 ⚡ 15

OpenAI's most intelligent general-purpose model, best for daily complex tasks.



GPT 5 ⚡ 15

OpenAI's greatest GPT model



GPT-5 Mini ⚡ 5

OpenAI's greatest GPT model but mini



GPT-5 Nano ⚡ 1

OpenAI's greatest GPT model with speed



Gemini 2.5 Pro ⚡ 15

Google's flagship, most capable reasoning model, best for complex tasks.



Claude Sonnet 4.5 New ⚡ 60

Anthropic's latest high-intelligence thinking model, best for creatives.



Claude 4 Sonnet (Thinking) ⚡ 60

Anthropic's latest high-intelligence thinking model, best for creatives.



O4 Mini ⚡ 15



GPT 5 Pro ⚡ 60



GPT OSS 120B ⚡ 5



Gemini 2.5 Flash ⚡ 1

ComparIA – L'outil français

ComparIA

- <https://comparia.beta.gouv.fr>
-  **Mission** : Ministère Culture + DINUM
- **Principe** : Test à l'aveugle de 2 IA anonymes
- **Données** : 23 modèles, 150k+ utilisateurs
- **Objectif** : Réduire les biais culturels français

Claude 4, GLM 4.5, GPT OSS et d'autres nouveaux modèles rejoignent l'arène...à tester!

Ne vous fiez pas aux réponses d'une seule IA

Discutez avec deux IA à l'aveugle et
évaluez leurs réponses

J'accepte les [conditions générales d'utilisation](#) ↗

Vos données sont partagées à des fins de recherche

Commencer à discuter

1 Je discute avec deux IA anonymes

Échangez aussi longtemps que
vous le souhaitez

2 Je donne mon avis

Vous contribuez ainsi à
l'amélioration des modèles d'IA

3 Les modèles sont démasqués!

Apprenez en plus sur les modèles
d'IA et leurs caractéristiques

Comment puis-je vous aider aujourd'hui ?

Explique de manière vulgarisée, très brièvement et très clairement le concept de "soft state" dans un contexte de bases de données non-relationnelles.

 Sélection des modèles ▾

Envoyer

Suggestions de prompts



Prompts issus de la
consultation
citoyenne sur l'IA 



Expliquer simplement
un concept



Rédiger un document
administratif



M'exprimer dans une
autre langue

Explique de manière vulgarisée, très brièvement et très clairement le concept de "soft state" dans un contexte de bases de données non-relationnelles.

Modèle A

Un "soft state" dans le contexte des bases de données non-relationnelles, comme celles utilisées dans les systèmes distribués, fait référence à un état qui peut évoluer de manière indépendante et asynchrone. En d'autres termes, l'état d'une donnée peut changer sans que toutes les parties du système soient immédiatement au courant de cette modification. Cela permet une plus grande flexibilité et une meilleure tolérance aux pannes, mais peut aussi introduire des défis en termes de cohérence des données.



Modèle B

Un "soft state" dans le contexte de bases de données non-relationnelles (comme MongoDB ou Cassandra) désigne un document ou une entrée qui n'est pas nécessairement à jour ou cohérent avec les autres données. C'est comme une note de bas de page qui n'a pas été vérifiée ou mise à jour.

En d'autres termes, un soft state peut contenir des informations qui ne sont pas exactes ou complètes, mais qui sont encore acceptables pour le système car elles sont suffisantes pour fonctionner. Cela peut être utile pour améliorer la performance ou la disponibilité du système, mais il est important de s'en occuper pour éviter la propagation d'erreurs.



Quel modèle d'IA préférez-vous ?

Avant de découvrir l'identité des modèles, nous avons besoin de votre préférence.

Elle permet d'enrichir les jeux de données compar:IA dont l'objectif est d'affiner les futurs modèles d'IA sur le français



Modèle A



Les deux se valent



Modèle B

Quel modèle d'IA préférez-vous ?

Avant de découvrir l'identité des modèles, nous avons besoin de votre préférence.

Elle permet d'enrichir les jeux de données compar:IA dont l'objectif est d'affiner les futurs modèles d'IA sur le français

Modèle A

Les deux se valent

Modèle B

Modèle A

Comment qualifiez-vous ses réponses ?



Utile

Complète

Créative

Mise en forme claire



Incorrecte

Superficielle

Instructions non suivies

Modèle B

Comment qualifiez-vous ses réponses ?



Utile

Complète

Créative

Mise en forme claire



Incorrecte

Superficielle

Instructions non suivies

Ajouter des détails

Mistral AI/Mistral Saba

PROPRIÉTAIRE

TAILLE ESTIMÉE (M)

SORTIE 02/2025

Modèle de taille moyenne conçu pour une compréhension linguistique et culturelle fine des langues du Moyen-Orient et d'Asie du Sud, notamment l'arabe, le tamoul et le malayalam.

Impact énergétique de la discussion

35 milliards param. (est.)
taille du modèle

×

177 jetons
taille du texte

=

 1.34wh
énergie conso.

Ce qui correspond à :

 0.82g
CO₂ émis

 16min
ampoule LED

 1min
vidéos en ligne

Voir plus

Meta/Llama 3.1 8B

Votre vote

SEMI-OUVERT

8 MDS DE PARAMÈTRES

SORTIE 07/2024

Petit modèle conçu pour un usage local sur ordinateur portable, tout en offrant de bonnes capacités pour la synthèse de texte et les réponses simples.

Impact énergétique de la discussion

8 milliards param.
taille du modèle

×

204 jetons
taille du texte

=

 0.77wh
énergie conso.

Ce qui correspond à :

 0.47g
CO₂ émis

 9min
ampoule LED

 53s
vidéos en ligne

Voir plus

Applications par spécialité

Codage → Grok 4, GPT-5, Claude

- Génération code complexe
- Debug automatique

Mathématiques → GPT-5, GPT oss 20b

- Résolution équations
- Preuves mathématiques

Outils/API → Llama 3.1 405b

- Intégration systèmes

Limites - Attention aux pièges !

Hallucinations

- *"Informations inventées mais plausibles"*

Exemple concret

- Demande : *"Comment utiliser `ollama export` ?"*
- IA : **"Utilisez ollama export model_name output.tar"**
- Réalité : **Cette commande n'existe pas !**

 **Piège** : Les IA inventent des fonctionnalités logiques

Solutions open source

Modèles stars

- **LLaMA 3 (Meta)** : 1B→405B paramètres
- **Mistral 7B** : Champion français
- **Qwen (Alibaba)** : Bilingue chinois
- **DeepSeek** : Alternative chinoise

😊 **Hugging Face** : Plateforme majeure d'hébergement, déploiement, et leaderboard public de modèles open source https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Avantages

- Gratuité après achat matériel
- Respect de la confidentialité des données
- Facilité de personnalisation et fine-tuning
- Accès à un large écosystème de modèles et outils via Hugging Face



Open LLM Leaderboard Archived

Comparing Large Language Models in an open and reproducible way

Search by model name · try "meta @architecture:llama @license:mit"

4576 / 4576

Advanced Filters

Supports strict search and regex · Use semicolons for multiple terms

Quick Filters

For Edge Devices · 786

For Consumers · 430

Mid-range · 3185

For the GPU-rich · 165

Only Official Providers · 470

table options

column visibility

	Rank	Type	Model	Aver...	IFEval	BBH	MATH	GPQA	MUSR	MML...	CO ₂ ...
🏆	1	🔹	MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %	66.01 kg
🏆	2	🗨️	MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %	64.44 kg
🏆	3	🗨️	dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %	25.99 kg
🏆	4	🗨️	MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %	25.95 kg
🏆	5	🔹	huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %	50.41 %	76.77 kg

Afficher un menu

Quantification - La magie

Principe : Réduire précision sans perdre performance

Transformation

- **FP32** (32 bits) → **INT8** (8 bits)
- **Taille ÷ 4 !**
- **Vitesse × 3 !**

Analogie 📷 : *"Photo 4K → HD : 4x plus petite, toujours belle"*

LM Studio - Interface graphique

LM Studio

- <https://lmstudio.ai>

Installation : 3 étapes simples

- Télécharger LM Studio
- Choisir modèle
- Charger & utiliser

 **Parfait débutants**

- **Configuration** : 16GB RAM minimum

Check for updates...



llama 3

Clear Results

Go

Supports any

Llama

Mistral

Phi-3

Falcon

StarCoder

StableLM

GPT-NeoX

gguf



model file on Hugging Face

Welcome to LM Studio!

[Release Notes \(v0.2.31\)](#)

LM Studio enables you develop and experiment with Large Language Models (LLMs) in your local computer environment, fully offline.

Search

Search and download compatible model files

AI Chat

Chat with local LLMs fully offline

Multi Model

Load and prompt multiple local LLMs simultaneously

Local Server

Run an OpenAI-like HTTP server on localhost

My Models

Manage your downloaded models

• Join [LM Studio's Discord Server](#) to discuss models, prompts, workflows and more.

Meta AI

8B

Llama

Requires 8GB+ RAM

Llama 3.1 8B Instruct

Llama 3.1 is a dense Transformer with 8B, 70B, or 405B parameters and a context window of up to 128K tokens trained by Meta.

File Size 4.92 GB

Small & Fast

Q4_K_M

Download

Published by lmstudio-community on Hugging Face

Microsoft Research

3B

Phi-3

Requires 8GB+ RAM

Phi 3 mini 4k Instruct

Phi-3-Mini-4K-Instruct is a 3.8B parameters, lightweight, state-of-the-art open model trained with the Phi-3 datasets that includes both synthetic data and the filtered publicly available websites... [Show all](#)

File Size 2.39 GB

Small & Fast

Q4_K_M

Download

Published by lmstudio-community on Hugging Face



Chats

+ New Chat (⌘N)

Export

Llama

7B

Q4_K_M

GGUF

plaintext

markdown

monospace

Settings

Reset to Default Settings

Preset

Llama 3 V2

Danger Zone

Discard Changes

Override Preset

System Prompt

You are a helpful, smart, kind, and efficient AI assistant. You always fulfill the user's requests to the best of your ability.

Ask before overriding

28 tokens

Advanced Configuration

Context Length

n_ctx 2048

Tools

Model Inspector

Context Overflow Policy

Behavior for when the generated tokens length exceeds the context window size.

- Stop at limit
- Keep the system prompt and the first user message, truncate middle
- Maintain a rolling window and truncate past messages (default)

Localization

Conversation Notes

USER

Qu'est-ce que tu sais faire ?

Je suis un assistant virtuel très polyvalent ! Voici quelques-unes des choses que je peux faire :

- Répondre à des questions** : Je peux répondre à des questions sur une large gamme de sujets, tels que la science, l'histoire, la géographie, les mathématiques, la culture générale, etc.
- Traduire du texte** : Je peux traduire le texte du français en d'autres langues (et inversement) pour vous aider à communiquer avec des personnes de différentes origines.
- Gérer les tâches administratives** : Je peux vous aider à créer des rappels, à planifier des événements, à calculer des horaires et à gérer vos agendas.
- Suggérer des idées créatives** : Si vous avez besoin d'idées pour un projet ou un problème que vous rencontrez, je peux vous proposer des suggestions et des solutions innovantes.
- Fournir des informations sur les nouvelles technologies** : Je suis au courant des dernières tendances en matière de technologie, y compris l'intelligence artificielle, le machine learning, la cybersécurité, etc.

AI

6. **Aider à la rédaction de textes** : Si vous avez besoin d'aide pour écrire un texte, je peux vous suggérer des idées, des structures et des arguments.

Regenerate

Continue

USER

Enter a user message ...

0 tokens

↵ to send, shift + ↵ for new line

time to first token:	gen t:	speed:	stop reason:	gpu layers:	cpu threads:	mlock:	token count:
0.40s	14.63s	27.81 tok/s	eosFound	33	4	true	453/2048

LM Studio - Serveur local

Lancement serveur

- Charger modèle dans LM Studio
- Onglet "Local Server" → Start
- Serveur disponible sur localhost:1234
- **Code Python exemple**

```
import requests  
  
response = requests.post("http://localhost:1234/v1/chat/completions",  
json={"messages": [{"role": "user",  
"content": "Qu'est-ce que le soft state en NoSQL?"}]})  
print(response.json())
```

RAM Usage: 4.55 GB
CPU: 0.00 %

Local Inference Server

Start a local HTTP server that mimics select OpenAI API endpoints.

Supported endpoints: `/v1/chat/completions`, `/v1/embeddings`, `/v1/moderations`

API Documentation: [Local Server API \(↗ open in browser\)](#)

Configuration

Server Port:

Cross-Origin-Resource-Sharing (CORS) ON OFF

Verbose Server Logs ON OFF

Server logs Server running on port 1234 (logs are saved into /tmp/lmstudio-server-logs)

```

{
  "index": 0,
  "message": {
    "role": "assistant",
    "content": "Le \"soft state\" est un concept utilisé dans les bases de données NoSQL pour décrire une situation où la cohérence des données n'est pas garantie après une opération de mise à jour ou d'insertion. En effet, les bases de données NoSQL sont souvent conçues pour offrir de faibles latences et un bon rendement en échange d'une certaine flexibilité sur la cohérence des données. Cela signifie que dans certains cas, une opération peut être considérée comme réussie avant que les données ne soient complètement synchronisées avec toutes les autres instances de la base de données. Le \"soft state\" se caractérise par le fait qu'une instance de donnée peut connaître un état temporairement incohérent, mais qui sera rapidement corrigé lors d'un prochain cycle de cohérence. Cela peut se produire après une opération telle que l'insertion ou la mise à jour d'une ressource en parallèle. Il existe deux types de \"soft state\" :\n\n- **Under replica** : c'est lorsque le nombre de répliques disponibles pour une donnée est inférieur au nombre requis pour garantir la cohérence.\n- **Over replica** : c'est lorsque le nombre de répliques disponibles pour une donnée dépasse le nombre requis.\n\nLes systèmes qui utilisent le \"soft state\" doivent être conçus pour gérer les situations d'incohérence temporaire des données et assurer que l'état cohérent soit récupéré dans un délai raisonnable."
  },
  "usage": {
    "prompt_tokens": 54,
    "completion_tokens": 341,
    "total_tokens": 395
  }
}

```

Model Downloads 0 downloading · 2 completed

test_LM_studio.py ×

```

test_LM_studio.py > ...
1 import requests
2 response = requests.post("http://localhost:1234/v1/chat/completions",
3 json={"messages": [{"role": "user",
4 "content": "Qu'est-ce que le soft state en NoSQL?"]})
5 print(response.json())

```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** PORTS

```

source /Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test/bin/activate
/Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test/bin/python /Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test_LM_studio.py
(base) adrian.chifu@Host-014 ANF-TDM % source /Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test/bin/activate
(test) (base) adrian.chifu@Host-014 ANF-TDM % /Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test/bin/python /Users/chifua/cloud.LIS/MCF/Presentations/2025/ANF-TDM/test_LM_studio.py
{"id": "chatcpl-6c0v3bwv1mn02tt647qlczx", "object": "chat.completion", "created": 1759186461, "model": "lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF/Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf", "choices": [{"index": 0, "message": {"role": "assistant", "content": "Le \"soft state\" est un concept utilisé dans les bases de données NoSQL pour décrire une situation où la cohérence des données n'est pas garantie après une opération de mise à jour ou d'insertion. En effet, les bases de données NoSQL sont souvent conçues pour offrir de faibles latences et un bon rendement en échange d'une certaine flexibilité sur la cohérence des données. Cela signifie que dans certains cas, une opération peut être considérée comme réussie avant que les données ne soient complètement synchronisées avec toutes les autres instances de la base de données. Le \"soft state\" se caractérise par le fait qu'une instance de donnée peut connaître un état temporairement incohérent, mais qui sera rapidement corrigé lors d'un prochain cycle de cohérence. Cela peut se produire après une opération telle que l'insertion ou la mise à jour d'une ressource en parallèle. Il existe deux types de \"soft state\" :\n\n- **Under replica** : c'est lorsque le nombre de répliques disponibles pour une donnée est inférieur au nombre requis pour garantir la cohérence.\n- **Over replica** : c'est lorsque le nombre de répliques disponibles pour une donnée dépasse le nombre requis.\n\nLes systèmes qui utilisent le \"soft state\" doivent être conçus pour gérer les situations d'incohérence temporaire des données et assurer que l'état cohérent soit récupéré dans un délai raisonnable."}, "finish_reason": "stop"}], "usage": {"prompt_tokens": 54, "completion_tokens": 341, "total_tokens": 395}}
(test) (base) adrian.chifu@Host-014 ANF-TDM %

```

Ln 4, Col 49 Spaces: 4 UTF-8 LF {} Python test (3.12.2)

Ollama - Ligne de commande

Ollama

- <https://ollama.com>

Installation express

```
bash
```

```
curl -fsSL https://ollama.com/install.sh | sh
```

```
ollama run mistral
```

✓ API intégrée

✓ Automation avancée

Ollama

Ollama

- <https://ollama.com>

Installation

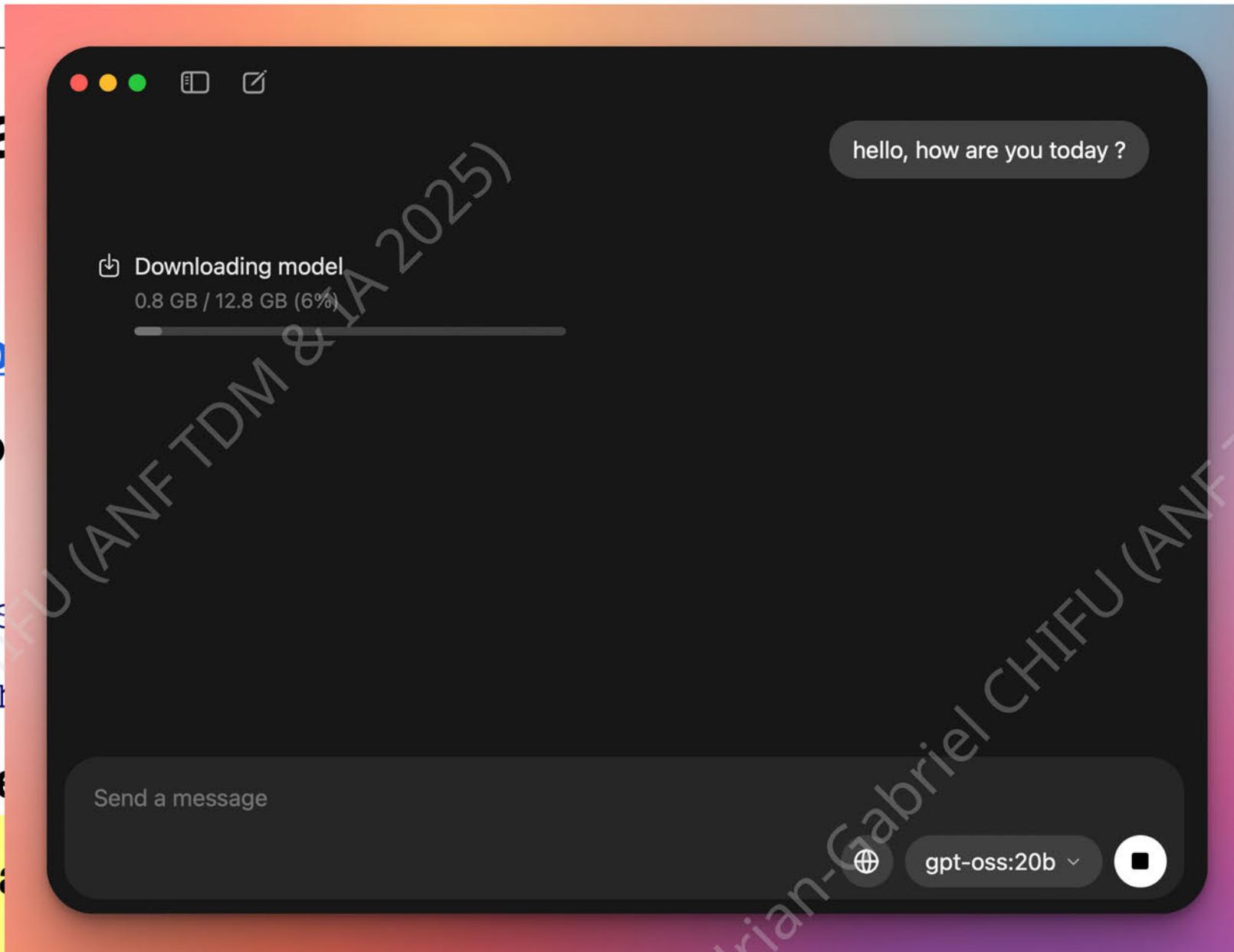
```
bash
```

```
curl -fsS
```

```
ollama
```

✓ API interface

✓ Automatic



Choix matériel

Critères VRAM

VRAM	Modèles possibles
4GB	Gemma 2B, Phi 3 Mini
8GB	Llama 3 8B (Q8)
16GB	Gemma 27B (Q4)
24GB	Llama 3 70B (Q2)

Règle : *1 milliard paramètres \approx 2GB en Q4*

Matrice de décision

Choix par usage

Usage	Modèle	Score	Coût
Maths	GPT-5	100%	\$\$\$
Code	Grok 4	75%	\$\$\$
Temps réel	Nova Micro	-	\$0.04
Local	Llama 3.1 405b	81.1%	Gratuit

Règle : *Choisir selon la tâche, pas la notoriété*

Tendances 2025-2026

Évolutions majeures

- **Spécialisation par tâches** : Fini l'IA "couteau suisse"
- **Agents autonomes** : Workflows complets automatisés
- **Vitesse** : Latence < 1s standard
- **Local** : Modèles puissants sur laptop

Ère des agents

- Agents spécialisés par domaine
- Orchestration multi-agents
- *Exemple* : "Analysez + créez + programmez"

Recommandations

🎯 Stratégies gagnantes

- **Hybride** : Combiner outils selon contexte
- **Agents** : Automatiser tâches répétitives
- **Test & Learn** : Expérimenter régulièrement
- **Local + Cloud** : Équilibrer sécurité/performance

💡 Règles d'or

- *"Critique → Vérifier"*
- *"Répétitif → Déléguer aux agents"*
- *"Former > Subir"*

Ressources essentielles

Comparaison

- **ComparIA** : comparia.beta.gouv.fr 
- **Vellum** : vellum.ai/llm-leaderboard
- **Hugging Face** : modèles et leaderboard open source

Outils clés

- **Cloud** : ChatGPT, Claude, Mistral, Perplexity
- **Local** : LM Studio, Ollama
- **Agents** : GPTs, Claude Projects

L'avenir radieux !?

 **Pourquoi être optimiste ?**

 **Démocratisation** : Expertise accessible à tous

 **Agents intelligents** : Collaborateurs digitaux

 **Créativité** : L'IA libère l'imagination

 **Souveraineté** : L'Europe reprend la main

 **Message final**

« L'ère des agents IA commence. L'avenir appartient à ceux qui sauront orchestrer ces nouveaux collaborateurs. » (Perplexity)

 **L'aventure ne fait que commencer !**

 **MERCI!** 

? Questions ?  

 **Contact :** adrian.chifu@univ-amu.fr