Choosing the Right Model

•••

Krzysztof Nowak (Institute of Polish Language, Kraków)

Outline

- model types
- architecture
- cost
- computational resources
- performance

LLM Ecosystem

Qwen

ANTHROP\C

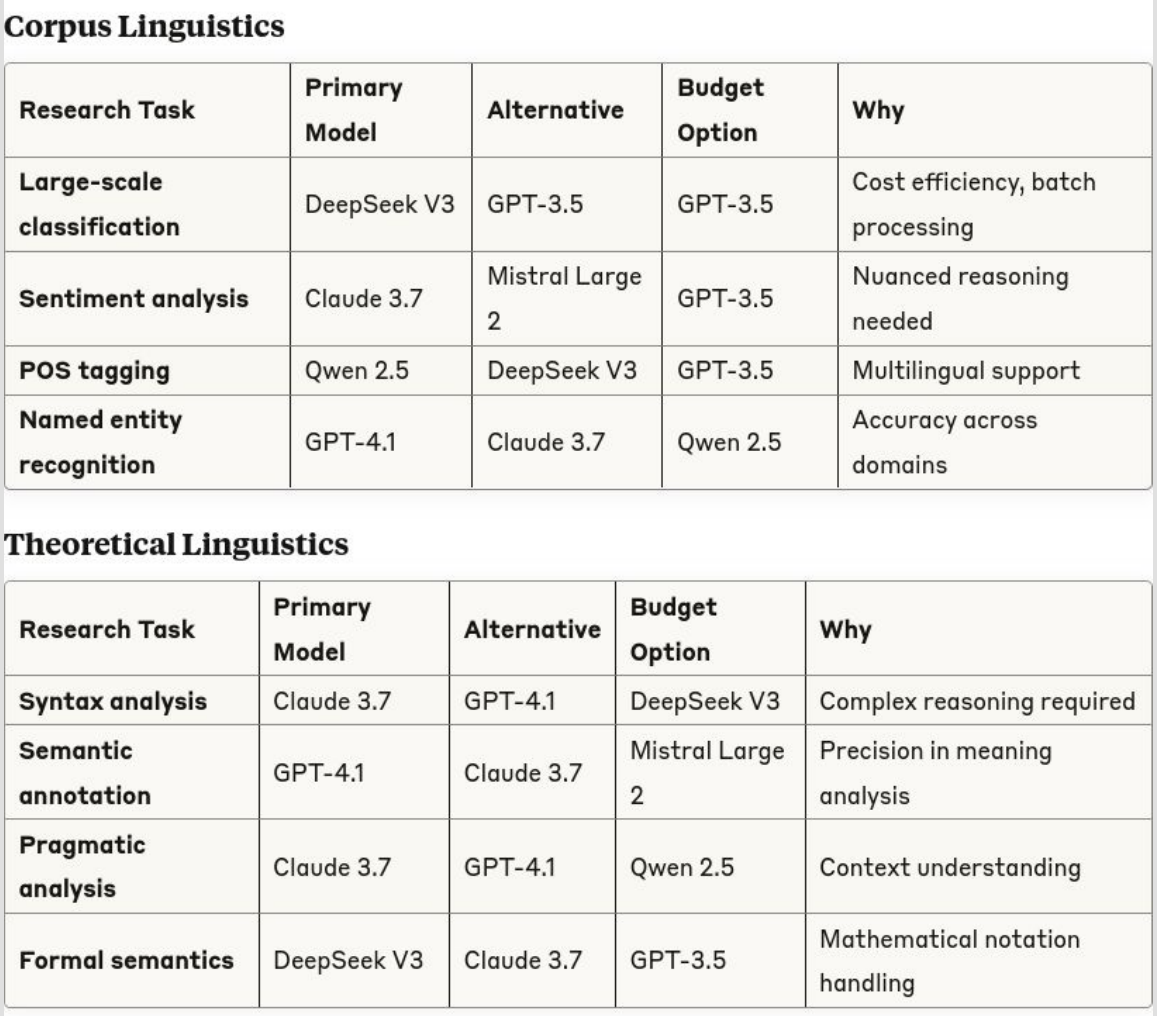
deepseek

LLaMA

perplexity

Focus, Licensing, Cost

Model	Provider	Type	Cost Range ($/1M)	Access
Claude 3.7 Sonnet	Anthropic	Proprietary	$1.5-7.5	API/Web
GPT-4.1	OpenAI	Proprietary	$2.0-8.0	API/ChatGPT
Gemini Ultra	Google	Proprietary	$1.0-2.0	AI Studio
Qwen 2.5 72B	Alibaba	Open/Hosted	$0.5-1.5	API/Local
Llama 3.1 405B	Meta	Open Source	Free	Local only
GPT-3.5 Turbo	OpenAI	Proprietary	$0.5-1.0	API/Web
Mistral Large 2	Mistral AI	Proprietary	$1.0-3.0	API/Web
DeepSeek V3	DeepSeek	Open/Hosted	$0.3-1.0	API/Local

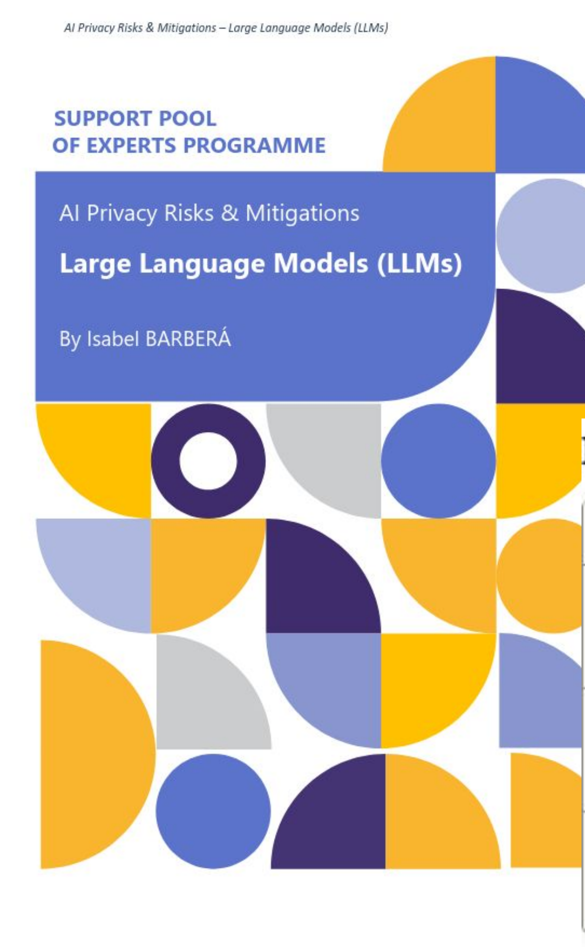


Corpus Linguistics

Research Task	Primary Model	Alternative	Budget Option	Why
Large-scale classification	DeepSeek V3	GPT-3.5	GPT-3.5	Cost efficiency, batch processing
Sentiment analysis	Claude 3.7	Mistral Large 2	GPT-3.5	Nuanced reasoning needed
POS tagging	Qwen 2.5	DeepSeek V3	GPT-3.5	Multilingual support
Named entity recognition	GPT-4.1	Claude 3.7	Qwen 2.5	Accuracy across domains

Theoretical Linguistics

Research Task	Primary Model	Alternative	Budget Option	Why
Syntax analysis	Claude 3.7	GPT-4.1	DeepSeek V3	Complex reasoning required
Semantic annotation	GPT-4.1	Claude 3.7	Mistral Large 2	Precision in meaning analysis
Pragmatic analysis	Claude 3.7	GPT-4.1	Qwen 2.5	Context understanding
Formal semantics	DeepSeek V3	Claude 3.7	GPT-3.5	Mathematical notation handling



AI Privacy Risks & Mitigations – Large Language Models (LLMs)

SUPPORT POOL
OF EXPERTS PROGRAMME

AI Privacy Risks & Mitigations

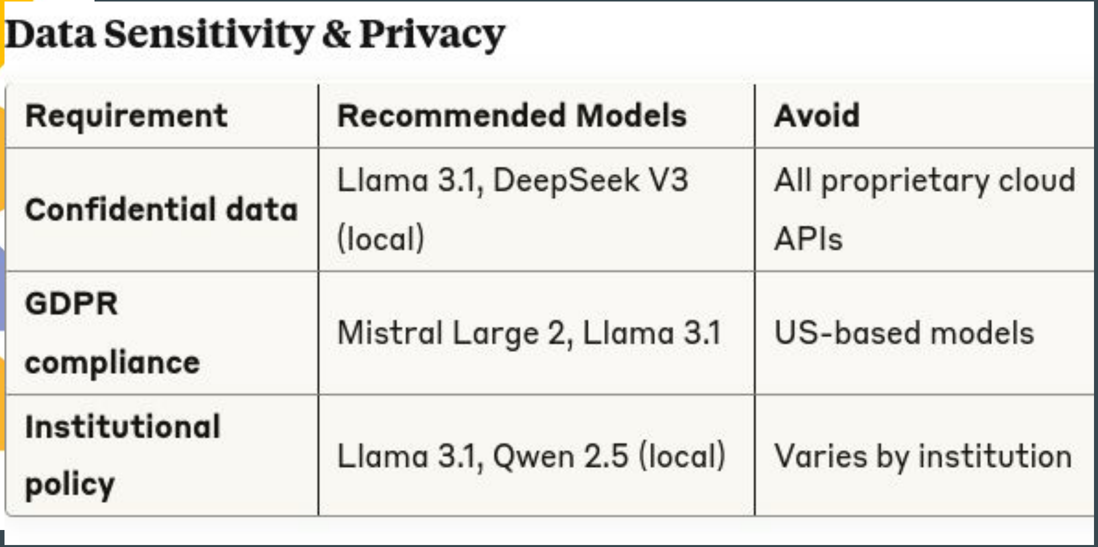
Large Language Models (LLMs)

By Isabel BARBERÁ

Data sensitivity

- confidential interviews
- sensitive information
- proprietary corpora

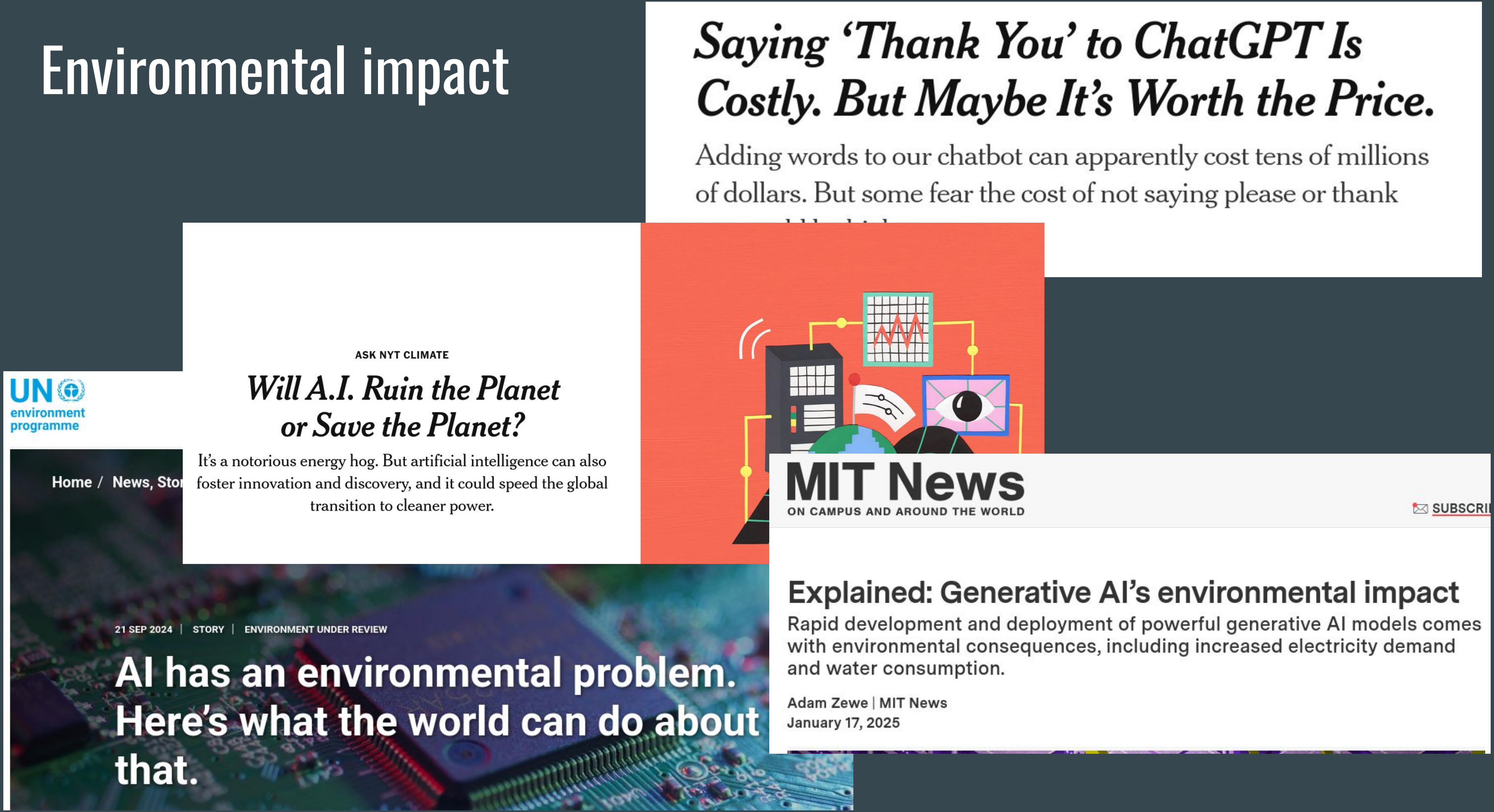
GDPR compliance



Data Sensitivity & Privacy

Requirement	Recommended Models	Avoid
Confidential data	Llama 3.1, DeepSeek V3 (local)	All proprietary cloud APIs
GDPR compliance	Mistral Large 2, Llama 3.1	US-based models
Institutional policy	Llama 3.1, Qwen 2.5 (local)	Varies by institution

Environmental impact



Saying 'Thank You' to ChatGPT Is Costly. But Maybe It's Worth the Price.

Adding words to our chatbot can apparently cost tens of millions of dollars. But some fear the cost of not saying please or thank

ASK NYT CLIMATE

Will A.I. Ruin the Planet or Save the Planet?

It's a notorious energy hog. But artificial intelligence can also foster innovation and discovery, and it could speed the global transition to cleaner power.

Home / News, Sto

21 SEP 2024 | STORY | ENVIRONMENT UNDER REVIEW

AI has an environmental problem. Here's what the world can do about that.

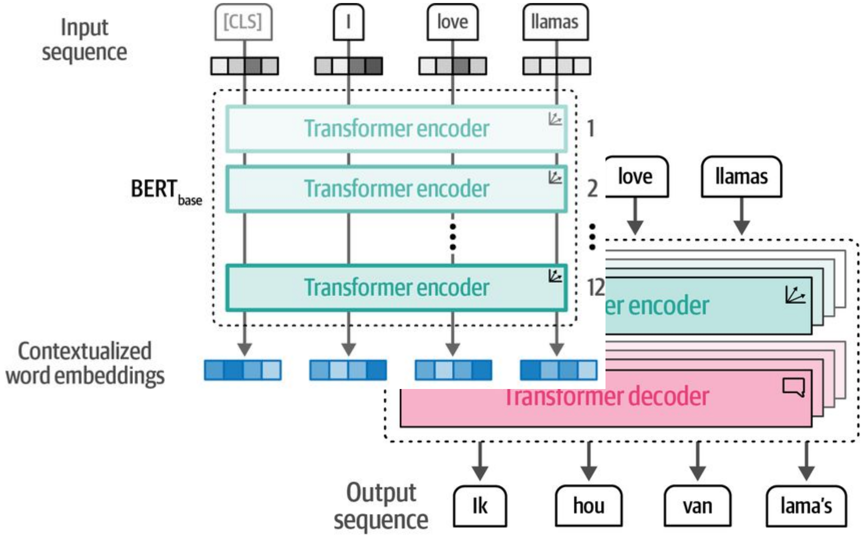
Explained: Generative AI's environmental impact

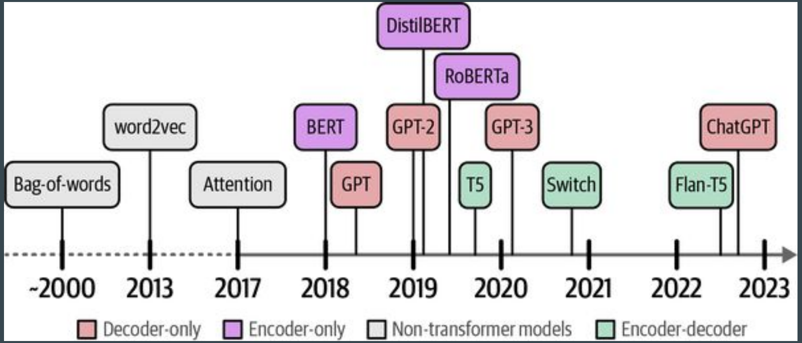
Rapid development and deployment of powerful generative AI models comes with environmental consequences, including increased electricity demand and water consumption.

Adam Zewe | MIT News

January 17, 2025

Do I really need an LLM?



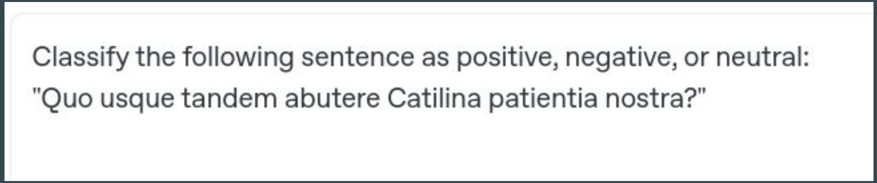


- cost
- speed
- interpretability
- specific tasks

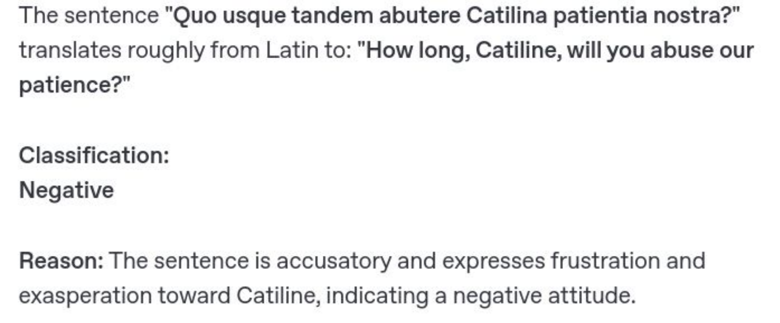
Practical Considerations

Access and Running Models Locally

Web Interface



Classify the following sentence as positive, negative, or neutral: "Quo usque tandem abutere Catilina patientia nostra?"

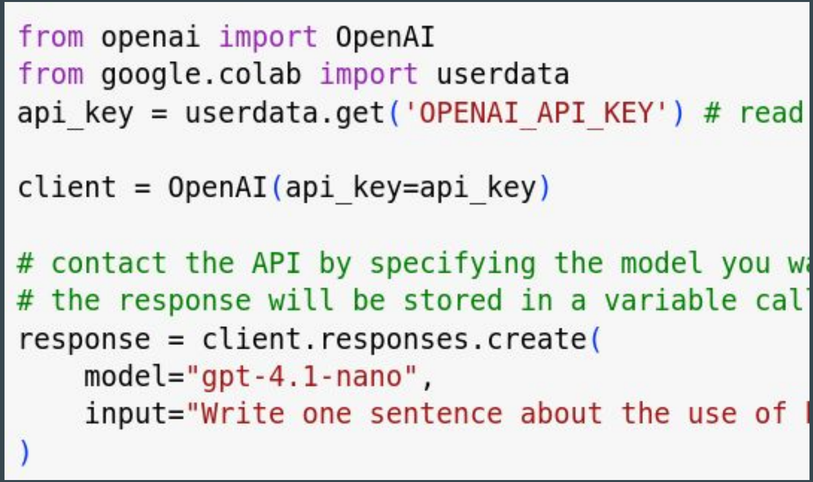


The sentence **"Quo usque tandem abutere Catilina patientia nostra?"** translates roughly from Latin to: **"How long, Catiline, will you abuse our patience?"**

Classification:
Negative

Reason: The sentence is accusatory and expresses frustration and exasperation toward Catiline, indicating a negative attitude.

API



```
from openai import OpenAI
from google.colab import userdata
api_key = userdata.get('OPENAI_API_KEY') # read

client = OpenAI(api_key=api_key)

# contact the API by specifying the model you wa
# the response will be stored in a variable cal
response = client.responses.create(
    model="gpt-4.1-nano",
    input="Write one sentence about the use of
)
```

Running Models Locally

Servers

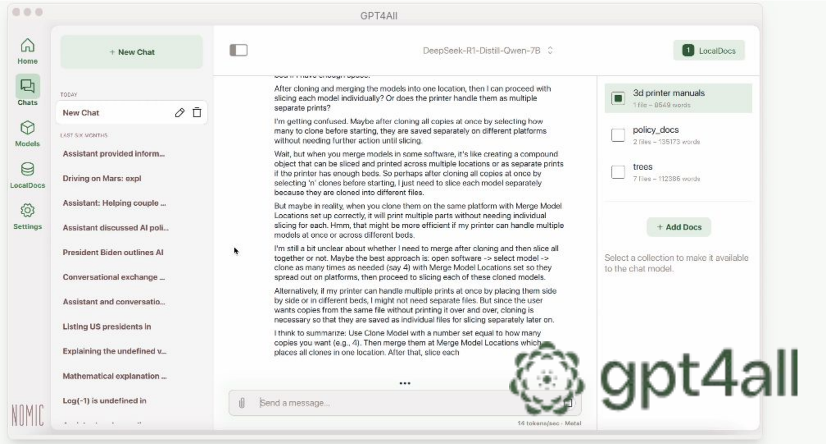
LLaMA C++

ollama

LM Studio

Running Models Locally

Clients



After cloning and merging the models into one location, then I can proceed with slicing each model individually? Or does the printer handle them as multiple separate prints?

I'm getting confused. Maybe after cloning all copies at once by selecting how many to clone before starting, they are saved separately on different platforms without needing further action until slicing.

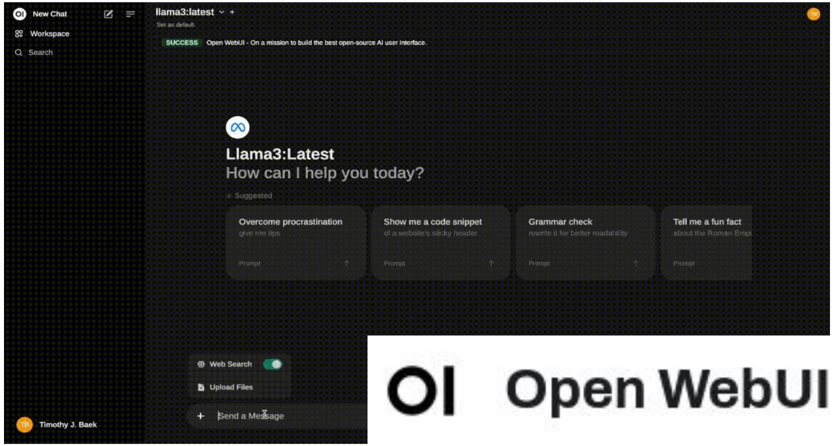
Wait, but when you merge models in some software, it's like creating a compound object that can be sliced and printed across multiple locations or as separate prints if the printer has enough beds. So perhaps after cloning all copies at once by selecting 'n' clones before starting, I just need to slice each model separately because they are cloned into different files.

But maybe in reality, when you clone them on the same platform with Merge Model Locations set up correctly, it will print multiple parts without needing individual slicing for each. Hmm, that might be more efficient if my printer can handle multiple models at once or across different beds.

I'm still a bit unclear about whether I need to merge after cloning and then slice all together or not. Maybe the best approach is: open software -> select model -> clone as many times as needed (say 4) with Merge Model Locations set so they spread out on platforms, then proceed to slicing each of these cloned models.

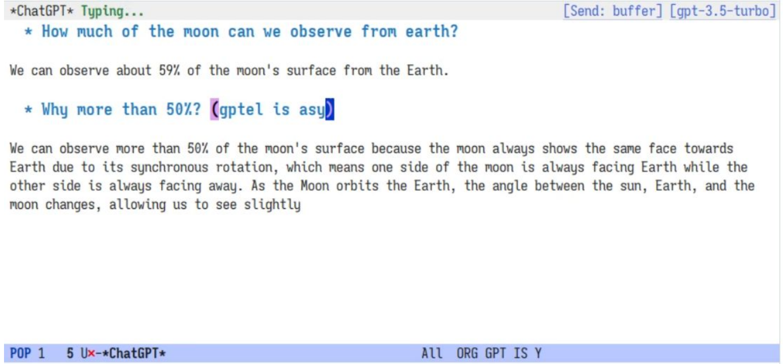
Alternatively, if my printer can handle multiple prints at once by placing them side by side or in different beds, I might not need separate files. But since the user wants copies from the same file without printing it over and over, cloning is necessary so that they are saved as individual files for slicing separately later on.

I think to summarize: Use Clone Model with a number set equal to how many copies you want (e.g., 4). Then merge them at Merge Model Locations which places all clones in one location. After that, slice each



Running Models Locally

Editors
Libraries



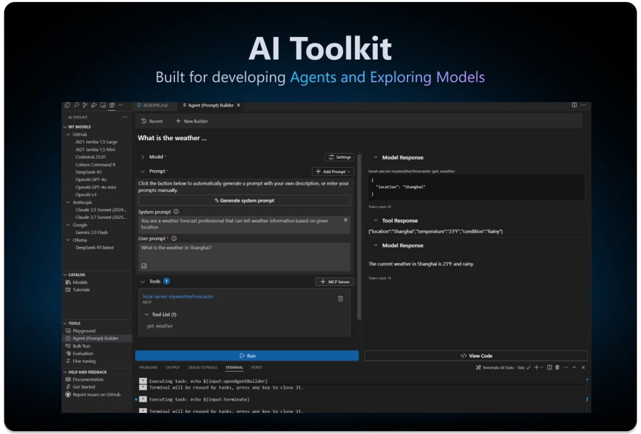
```
*ChatGPT* Typing...                                        [Send: buffer] [gpt-3.5-turbo]
 * How much of the moon can we observe from earth?

We can observe about 59% of the moon's surface from the Earth.

 * Why more than 50%? (gptel is asy)

We can observe more than 50% of the moon's surface because the moon always shows the same face towards
Earth due to its synchronous rotation, which means one side of the moon is always facing Earth while the
other side is always facing away. As the Moon orbits the Earth, the angle between the sun, Earth, and the
moon changes, allowing us to see slightly

POP 1   5 U×-*ChatGPT*                                     All  ORG GPT IS Y
```



More: https://github.com/ollama/ollama

Practical Considerations

Computational Resources

Hardware Requirements

GeForce RTX 4090

WIĘCEJ NIŻ SZYBKA

NVIDIA® GeForce RTX™ 4090 to absolutnie najlepsza karta graficzna GeForce. Zapewnia ogromny skok w zakresie wydajności, efektywności i grafiki opartej na SI. Ciesz się ultrawydajnym gamingem, niezwykle szczegółowymi wirtualnymi światami, niespotykaną produktywnością i nowymi sposobami tworzenia. Oparta jest na architekturze NVIDIA Ada Lovelace i została wyposażona w 24 GB pamięci G6X, aby zapewnić graczom i twórcom absolutnie najwyższy komfort.

Z ceną od 7.749 zł

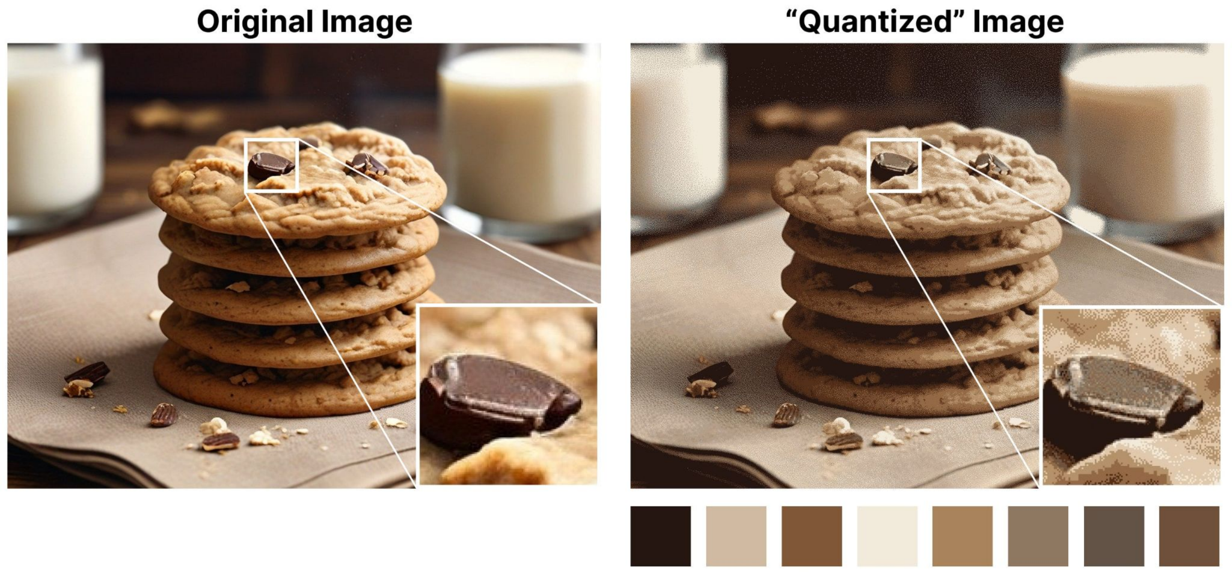
7B model → 14GB VRAM

70B model → 140GB VRAM

Consumer GPUs (RTX 4090)

- quantized models
- quality compromises

Quantization The Quality vs. Accessibility Trade-off



Source: Maarten Grootendorst, *A Visual Guide to Quantization*

Practical Considerations

Cost

Cost

Counting tokens

Cost Management

Context Window Limits

Latency

Performance Optimization

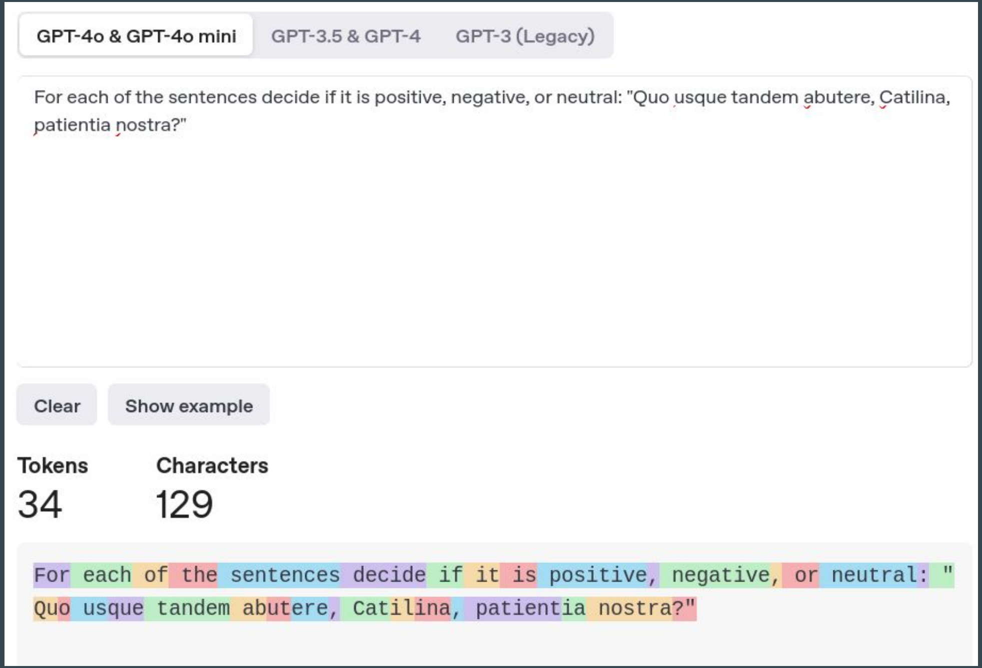
Why Token Counting Matters

- Subword units that LLMs process (not always = words)

```
"unbelievable" → [un|believ|able]
     1 word           3 tokens
```

- Punctuation, spaces, and special characters

```
Sentence: "Hello, world!"
Tokens: [Hello][,][ world][!]
          (4 tokens)
```



For each of the sentences decide if it is positive, negative, or neutral: "Quo usque tandem abutere, Catilina, patientia nostra?"

Tokens
34

Characters
129

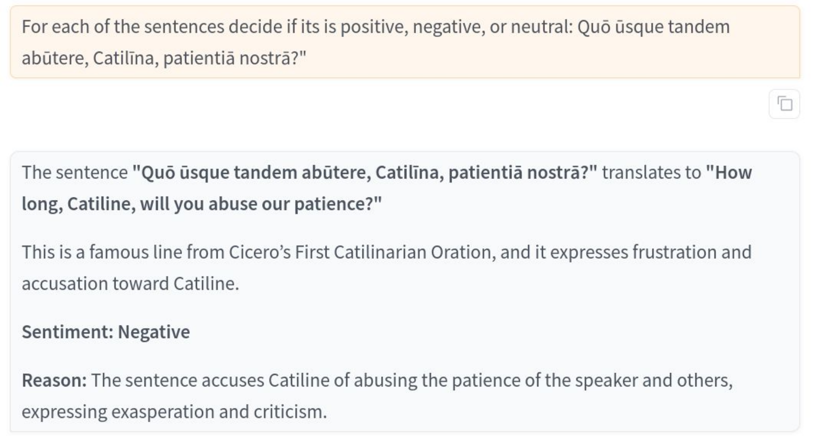
For each of the sentences decide if it is positive, negative, or neutral: "Quo usque tandem abutere, Catilina, patientia nostra?"

https://platform.openai.com/tokenizer

Input: ~32 tokens

(18 + 14 tokens)

Output: ~95-105 tokens



For each of the sentences decide if its is positive, negative, or neutral: Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?"

The sentence **"Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?"** translates to **"How long, Catiline, will you abuse our patience?"**

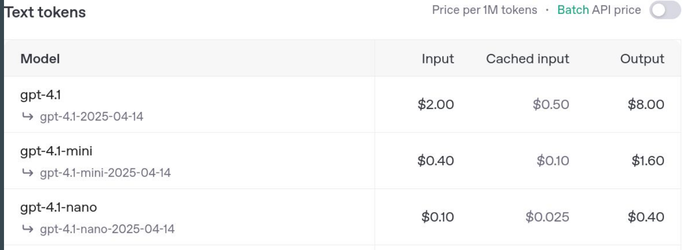
This is a famous line from Cicero's First Catilinarian Oration, and it expresses frustration and accusation toward Catiline.

Sentiment: Negative

Reason: The sentence accuses Catiline of abusing the patience of the speaker and others, expressing exasperation and criticism.

	1 token	Request	1000 Requests
Input	$0.000002	$0.000064	$0.064
Output	$0.000008	$0.0008	$0.8
Total		**$0.000864**	**$0.864**

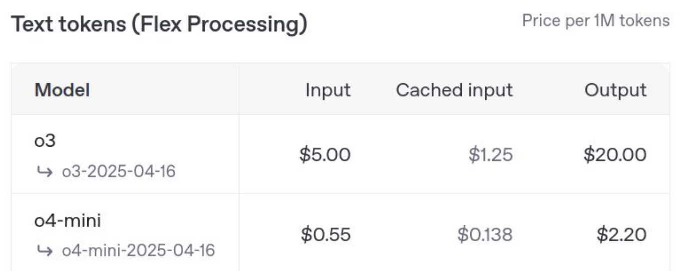
Model Selection



Text tokens

Price per 1M tokens · Batch API price

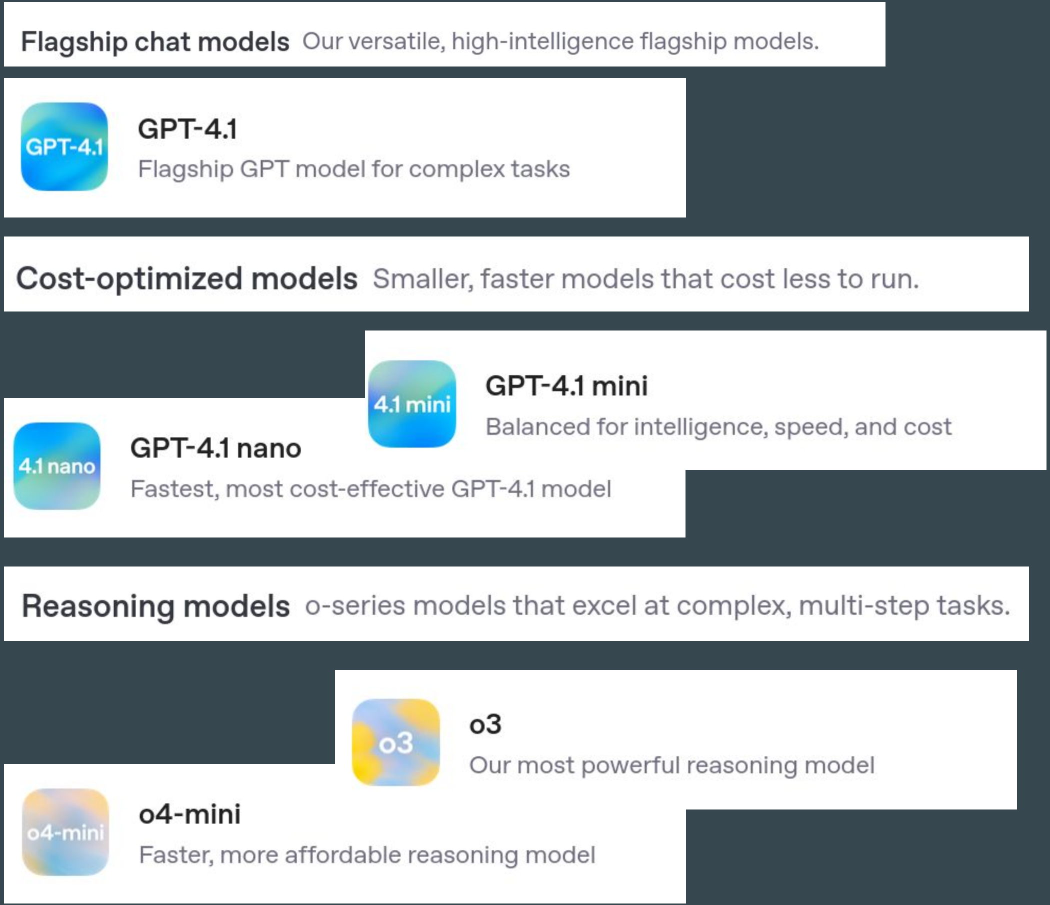
Model	Input	Cached input	Output
gpt-4.1 ↳ gpt-4.1-2025-04-14	$2.00	$0.50	$8.00
gpt-4.1-mini ↳ gpt-4.1-mini-2025-04-14	$0.40	$0.10	$1.60
gpt-4.1-nano ↳ gpt-4.1-nano-2025-04-14	$0.10	$0.025	$0.40



Text tokens (Flex Processing)

Price per 1M tokens

Model	Input	Cached input	Output
o3 ↳ o3-2025-04-16	$5.00	$1.25	$20.00
o4-mini ↳ o4-mini-2025-04-16	$0.55	$0.138	$2.20



Flagship chat models Our versatile, high-intelligence flagship models.

GPT-4.1
Flagship GPT model for complex tasks

Cost-optimized models Smaller, faster models that cost less to run.

GPT-4.1 mini
Balanced for intelligence, speed, and cost

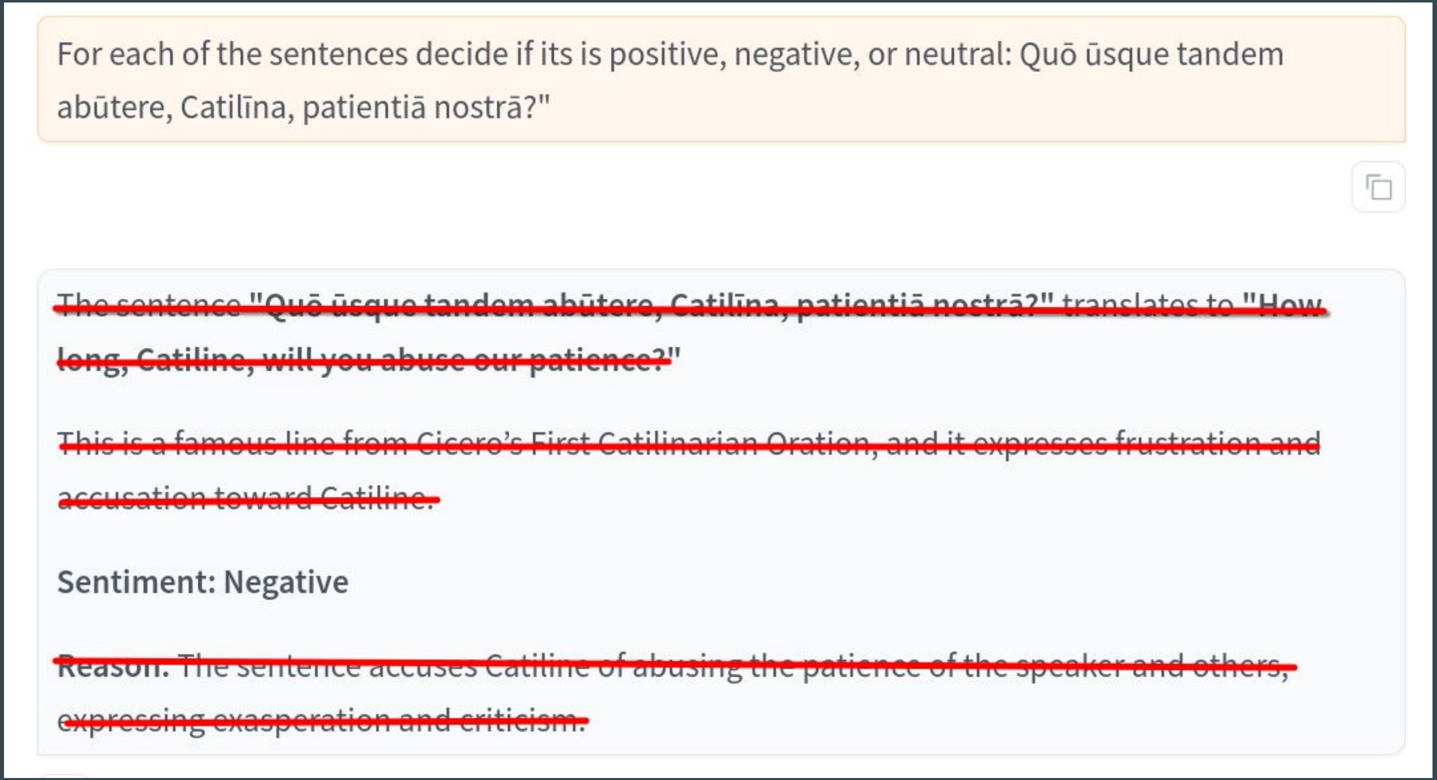
GPT-4.1 nano
Fastest, most cost-effective GPT-4.1 model

Reasoning models o-series models that excel at complex, multi-step tasks.

o3
Our most powerful reasoning model

o4-mini
Faster, more affordable reasoning model

Prompt Engineering



For each of the sentences decide if its is positive, negative, or neutral: Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?"

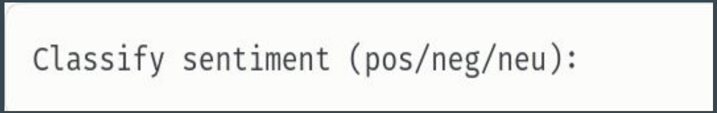
~~The sentence **"Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?"** translates to **"How long, Catiline, will you abuse our patience?**~~**"**

~~This is a famous line from Cicero's First Catilinarian Oration, and it expresses frustration and accusation toward Catiline.~~

Sentiment: Negative

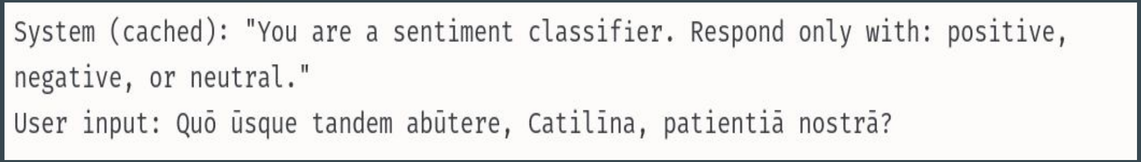
~~**Reason:** The sentence accuses Catiline of abusing the patience of the speaker and others, expressing exasperation and criticism.~~

Strategies



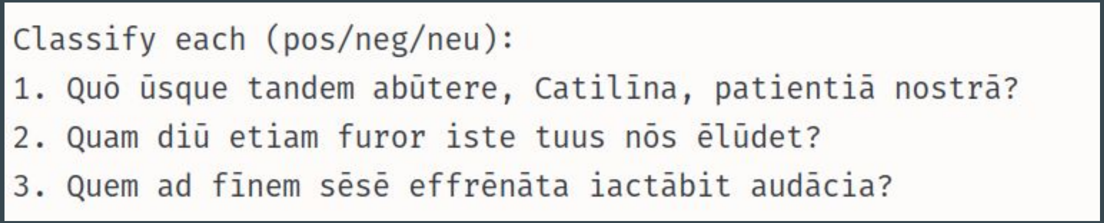
```
Classify sentiment (pos/neg/neu):
```

Smart Truncation



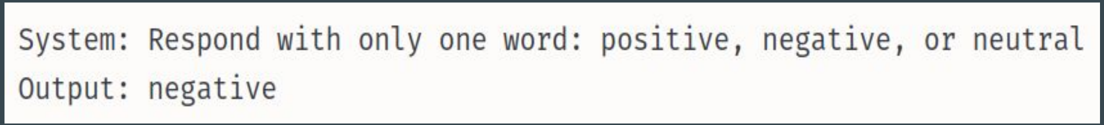
```
System (cached): "You are a sentiment classifier. Respond only with: positive,
negative, or neutral."
User input: Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?
```

Caching



```
Classify each (pos/neg/neu):
1. Quō ūsque tandem abūtere, Catilīna, patientiā nostrā?
2. Quam diū etiam furor iste tuus nōs ēlūdet?
3. Quem ad fīnem sēsē effrēnāta iactābit audācia?
```

Batching

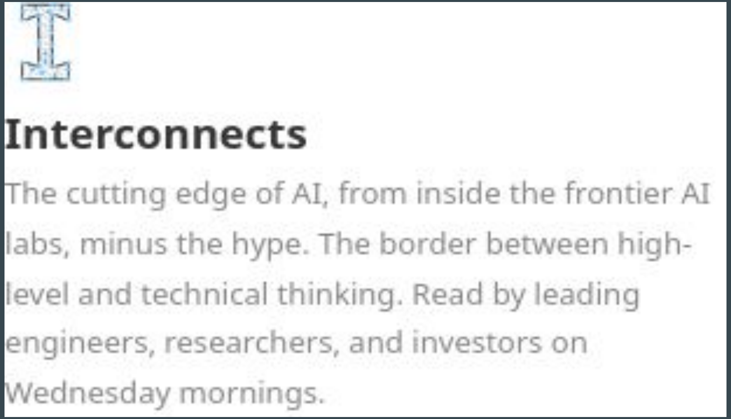


```
System: Respond with only one word: positive, negative, or neutral
Output: negative
```

Output Control

Resources

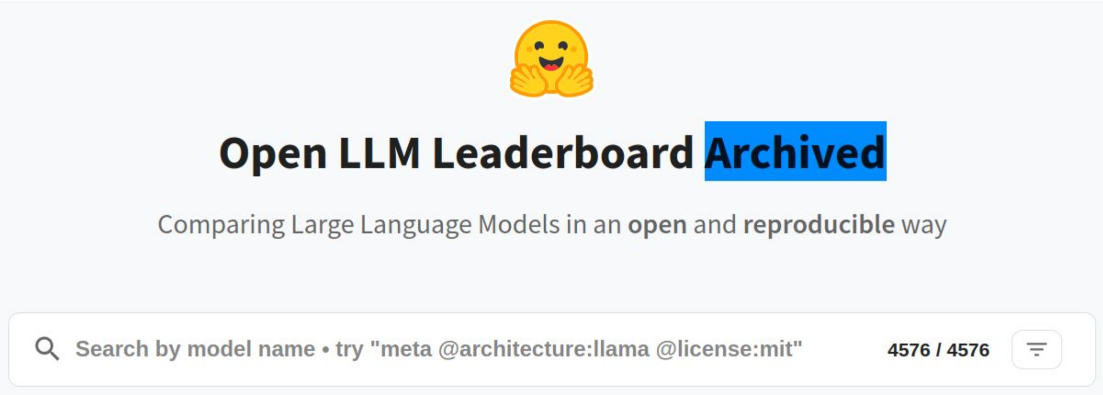
- Model documentation
- Papers
- Blogs, Newsletters, Reddit
- Substack



Interconnects

The cutting edge of AI, from inside the frontier AI labs, minus the hype. The border between high-level and technical thinking. Read by leading engineers, researchers, and investors on Wednesday mornings.

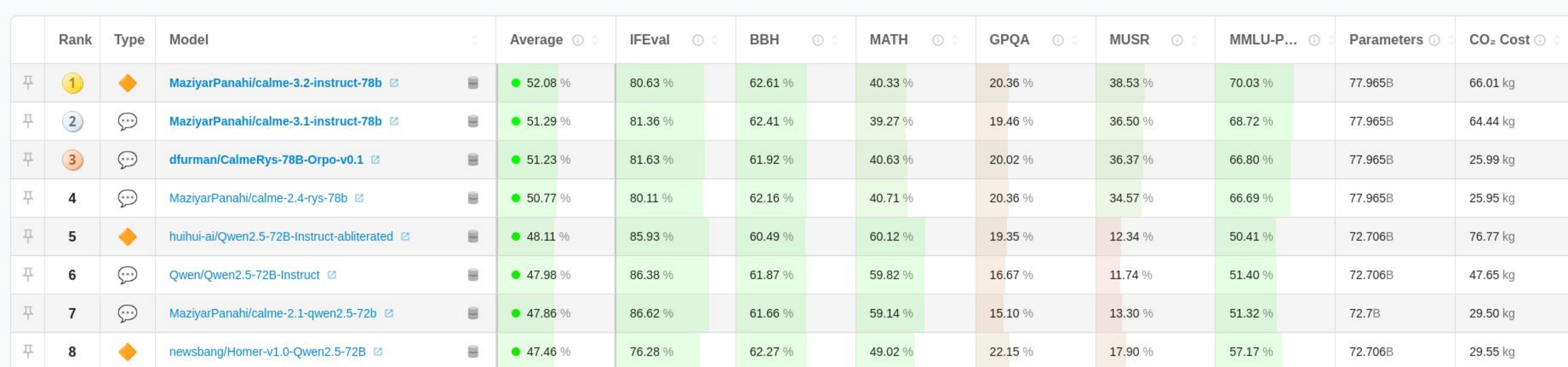
Benchmarks, Leaderboards



Open LLM Leaderboard Archived

Comparing Large Language Models in an **open** and **reproducible** way

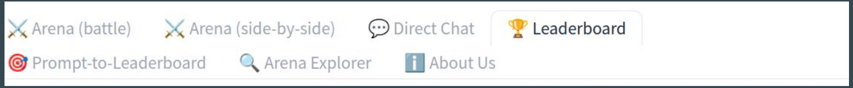
Search by model name • try "meta @architecture:llama @license:mit"
4576 / 4576

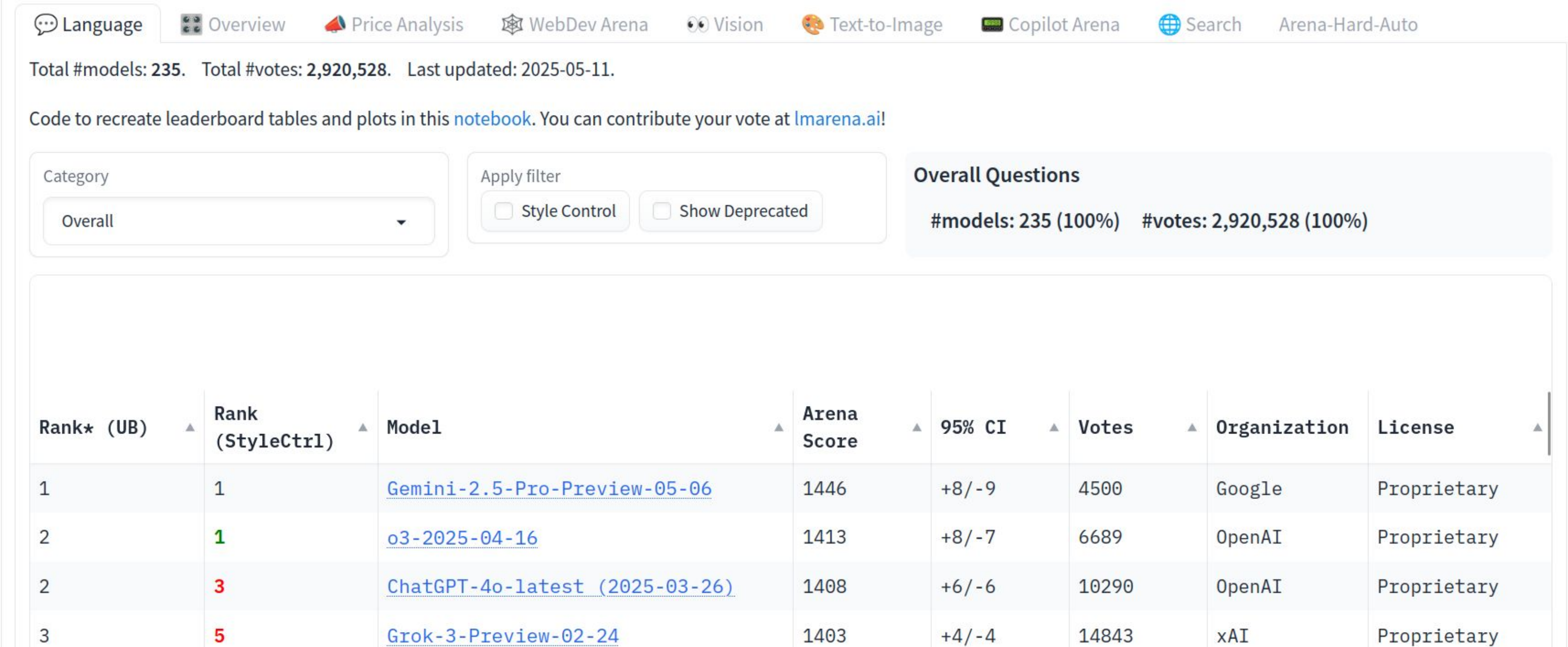


Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU-P...	Parameters	CO₂ Cost
1	🔶	**MaziyarPanahi/calme-3.2-instruct-78b**	● 52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %	77.965B	66.01 kg
2	💬	**MaziyarPanahi/calme-3.1-instruct-78b**	● 51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %	77.965B	64.44 kg
3	💬	**dfurman/CalmeRys-78B-Orpo-v0.1**	● 51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %	77.965B	25.99 kg
4	💬	MaziyarPanahi/calme-2.4-rys-78b	● 50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %	77.965B	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 %	72.706B	76.77 kg
6	💬	Qwen/Qwen2.5-72B-Instruct	● 47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %	72.706B	47.65 kg
7	💬	MaziyarPanahi/calme-2.1-qwen2.5-72b	● 47.86 %	86.62 %	61.66 %	59.14 %	15.10 %	13.30 %	51.32 %	72.7B	29.50 kg
8	🔶	newsbang/Homer-v1.0-Qwen2.5-72B	● 47.46 %	76.28 %	62.27 %	49.02 %	22.15 %	17.90 %	57.17 %	72.706B	29.55 kg

https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

Chatbot Arena (https://lmarena.ai)





Total #models: **235**. Total #votes: **2,920,528**. Last updated: 2025-05-11.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

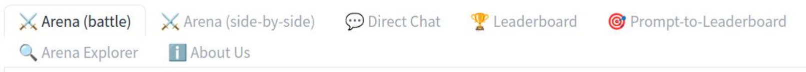
Apply filter: ☐ Style Control ☐ Show Deprecated

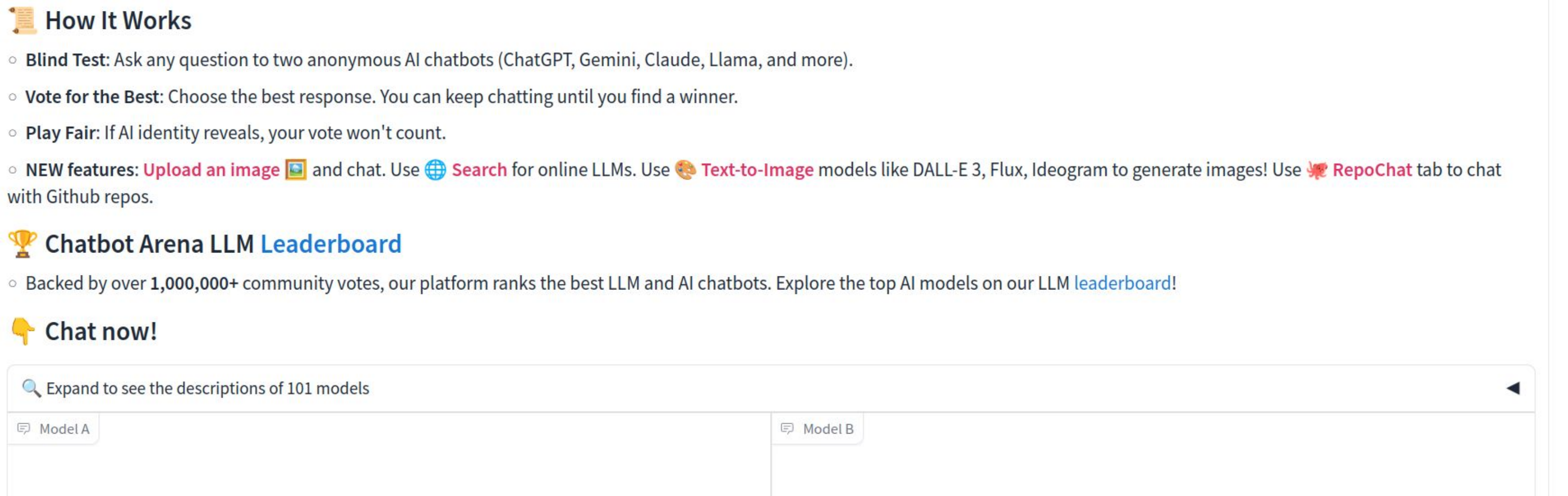
Overall Questions

#models: 235 (100%) **#votes: 2,920,528 (100%)**

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Preview-05-06	1446	+8/-9	4500	Google	Proprietary
2	1	o3-2025-04-16	1413	+8/-7	6689	OpenAI	Proprietary
2	3	ChatGPT-4o-latest (2025-03-26)	1408	+6/-6	10290	OpenAI	Proprietary
3	5	Grok-3-Preview-02-24	1403	+4/-4	14843	xAI	Proprietary

Chatbot Arena (https://lmarena.ai)





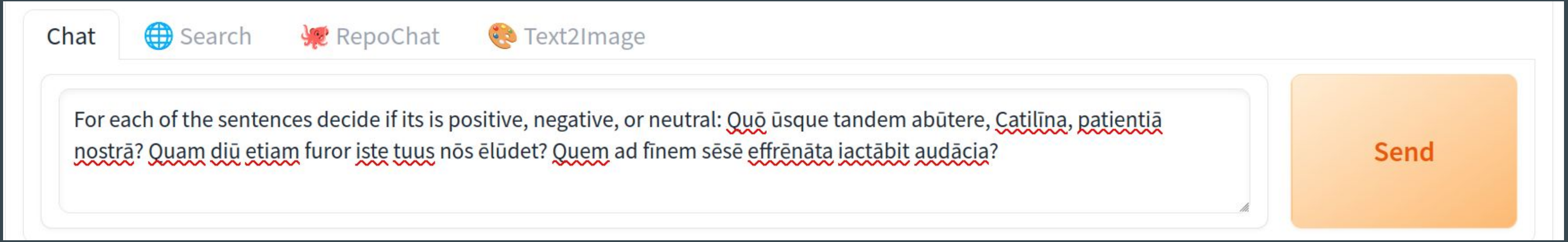
📜 How It Works

- **Blind Test**: Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).
- **Vote for the Best**: Choose the best response. You can keep chatting until you find a winner.
- **Play Fair**: If AI identity reveals, your vote won't count.
- **NEW features**: Upload an image 🖼️ and chat. Use 🌐 Search for online LLMs. Use 🎨 Text-to-Image models like DALL-E 3, Flux, Ideogram to generate images! Use 🐙 RepoChat tab to chat with Github repos.

🏆 Chatbot Arena LLM Leaderboard

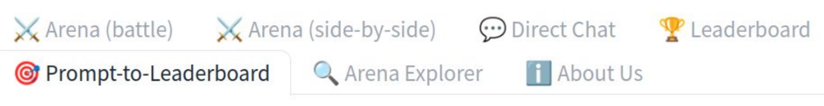
- Backed by over **1,000,000+** community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!

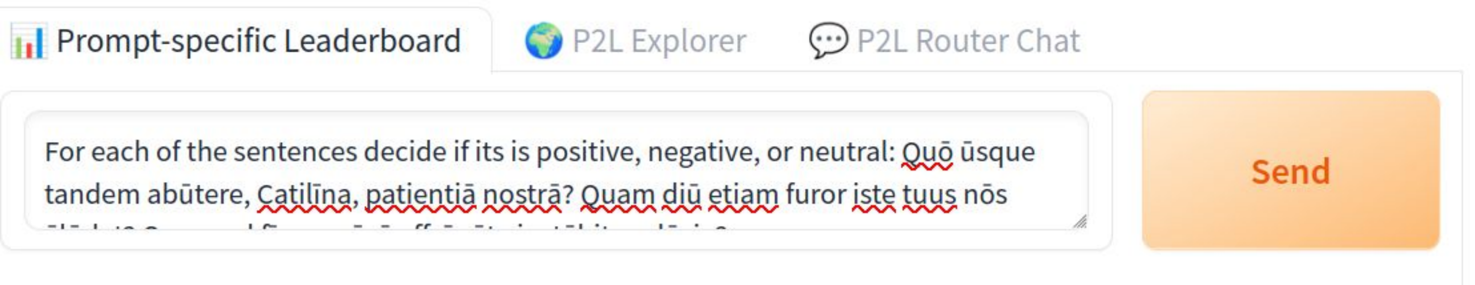
👇 Chat now!



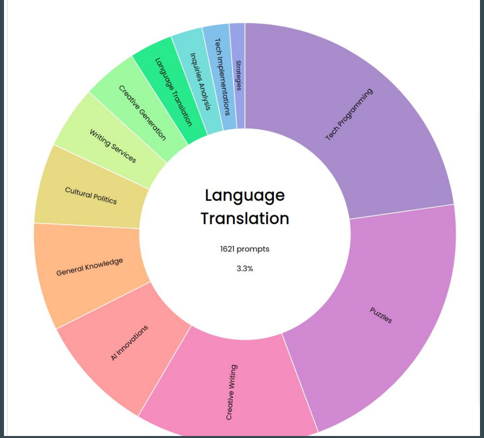
For each of the sentences decide if its is positive, negative, or neutral: Quō ūsque tandem abūtere, Catilīna, patientiā nostrā? Quam diū etiam furor iste tuus nōs ēlūdet? Quem ad fīnem sēsē effrēnāta iactābit audācia?

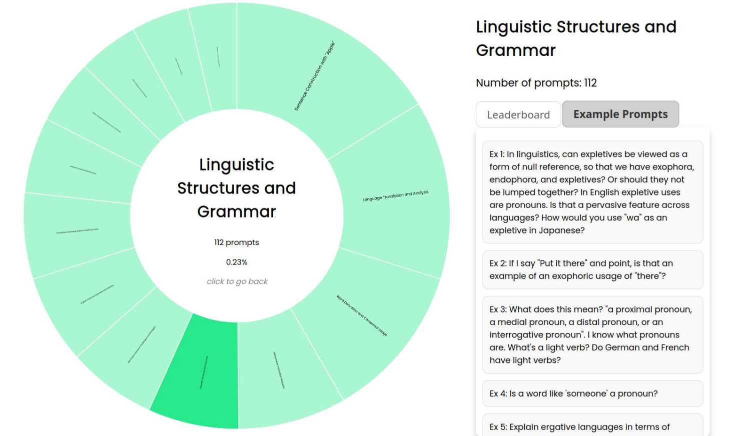
Chatbot Arena (https://lmarena.ai)





For each of the sentences decide if its is positive, negative, or neutral: Quō ūsque tandem abūtere, Catilīna, patientiā nostrā? Quam diū etiam furor iste tuus nōs




Linguistic Structures and Grammar
112 prompts
0.23%
click to go back
Sentence Construction with "Apple"
Language Translation and Analysis
Word Derivation and Contextual Usage
Leaderboard
Example Prompts

Linguistic Structures and Grammar

Number of prompts: 112

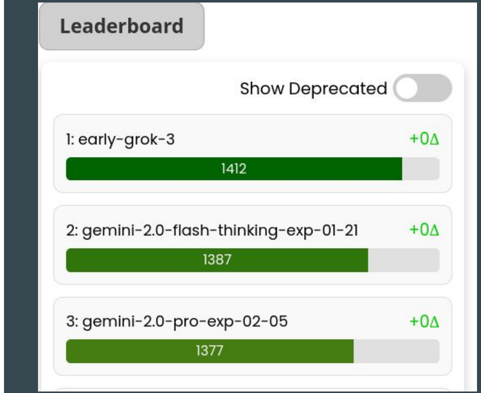
Ex 1: In linguistics, can expletives be viewed as a form of null reference, so that we have exophora, endophora, and expletives? Or should they not be lumped together? In English expletive uses are pronouns. Is that a pervasive feature across languages? How would you use "wa" as an expletive in Japanese?

Ex 2: If I say "Put it there" and point, is that an example of an exophoric usage of "there"?

Ex 3: What does this mean? "a proximal pronoun, a medial pronoun, a distal pronoun, or an interrogative pronoun". I know what pronouns are. What's a light verb? Do German and French have light verbs?

Ex 4: Is a word like 'someone' a pronoun?

Ex 5: Explain ergative languages in terms of



Model	Change	Score
1: early-grok-3	+0Δ	1412
2: gemini-2.0-flash-thinking-exp-01-21	+0Δ	1387
3: gemini-2.0-pro-exp-02-05	+0Δ	1377

Advice? Experiment!