Unlocking Enterprise Innovation: Navigating the Open Model Landscape

FutureIT | NYC

Michele Rosen, PhD
Research Manager, Open GenAI, LLMs, and the Evolving Open Source Ecosystem

November 21, 2024

Agenda

- Why choose an open model?
- What are “open” models?
- Who is creating open models?
- Who is using open models?
- Choosing an open model
- Takeaways

Why choose an open model?

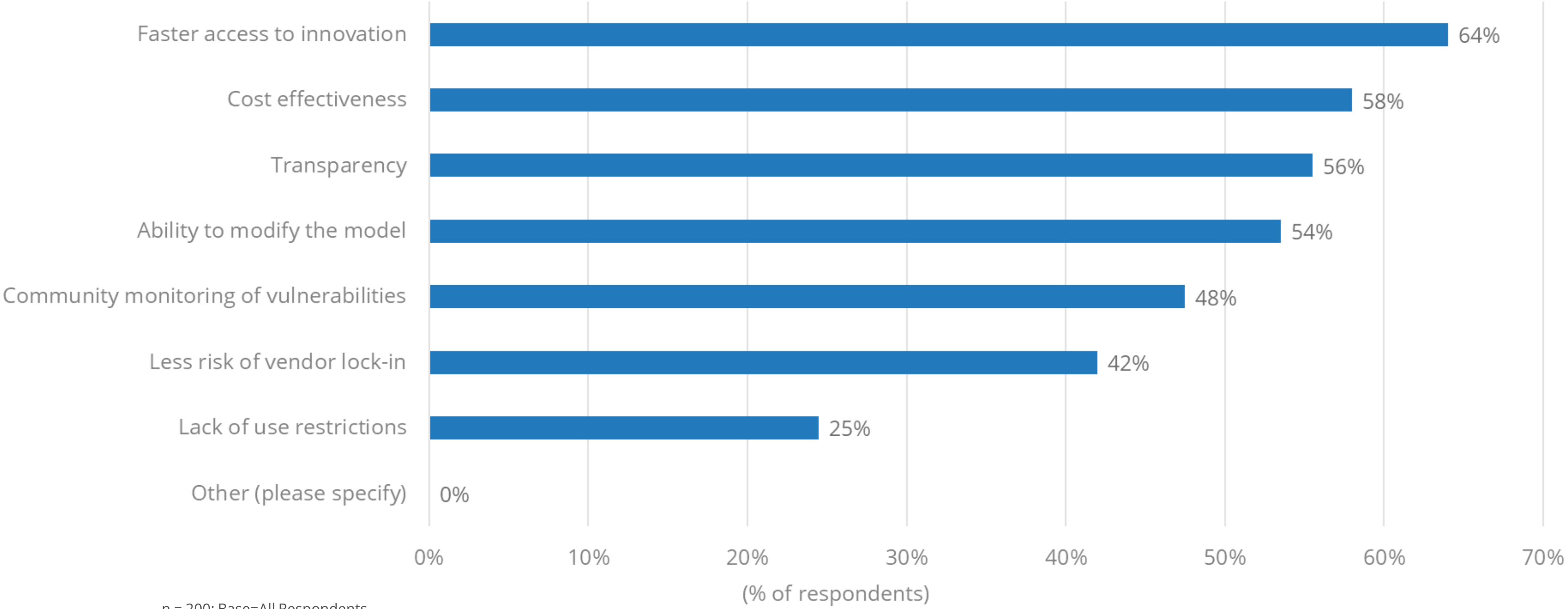
Open vs Closed Models

	PROS	CONS
Closed Models	• Top performance • Well-documented APIs • Streamlined deployment • Dedicated vendor support	• Limited transparency • Vendor lock-in • Restrictions on customization • Higher costs
Open Models	• Flexible deployment and customization • Greater transparency • Lower costs for smaller models • Potential to avoid vendor lock-in	• Requires staff expertise • Performance may not be adequate • Selecting a model is more difficult • May lack dedicated vendor support

Source: IDC, 2024

Faster access to innovation beats cost effectiveness for the most important benefit of using open models

What are the most important benefits of using open foundation models for GenAI use cases?

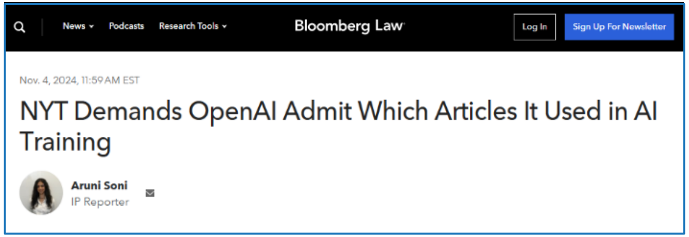


n = 200; Base=All Respondents

Notes: Managed by IDC's Global Primary Research Group.; Data Not Weighted; Multiple dichotomous table - total will not sum to 100%; Use caution when interpreting small sample sizes.

Source: U.S. Open Source Software Use Study, IDC, June, 2024

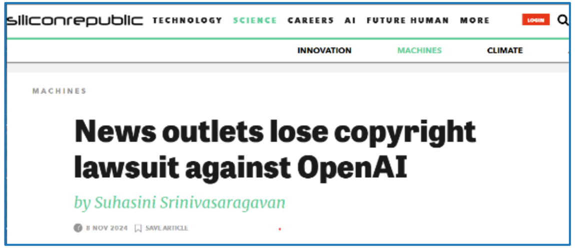
Legal and regulatory uncertainty



Nov. 4, 2024, 11:59 AM EST

NYT Demands OpenAI Admit Which Articles It Used in AI Training

Aruni Soni
IP Reporter

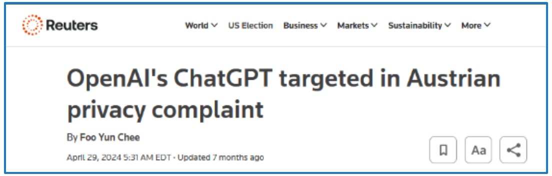


MACHINES

News outlets lose copyright lawsuit against OpenAI

by Suhasini Srinivasaragavan

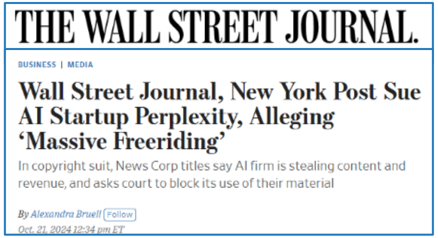
8 NOV 2024 SAVE ARTICLE



OpenAI's ChatGPT targeted in Austrian privacy complaint

By Foo Yun Chee

April 29, 2024 5:31 AM EDT · Updated 7 months ago



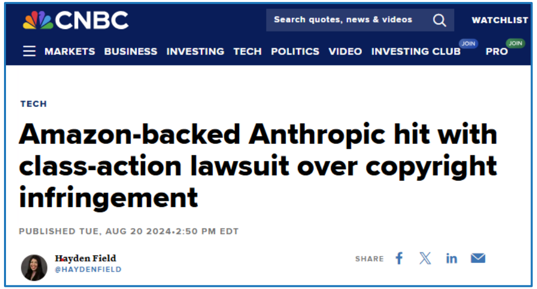
BUSINESS | MEDIA

Wall Street Journal, New York Post Sue AI Startup Perplexity, Alleging 'Massive Freeriding'

In copyright suit, News Corp titles say AI firm is stealing content and revenue, and asks court to block its use of their material

By Alexandra Bruell Follow

Oct. 21, 2024 12:34 pm ET

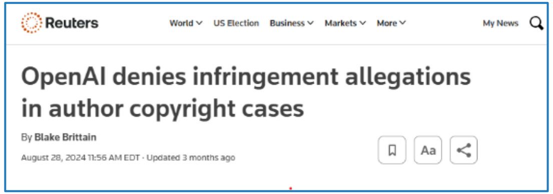


TECH

Amazon-backed Anthropic hit with class-action lawsuit over copyright infringement

PUBLISHED TUE, AUG 20 2024·2:50 PM EDT

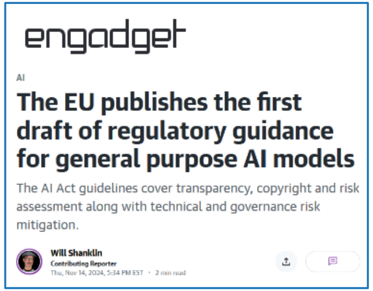
Hayden Field
@HAYDENFIELD



OpenAI denies infringement allegations in author copyright cases

By Blake Brittain

August 28, 2024 11:56 AM EDT · Updated 3 months ago



AI

The EU publishes the first draft of regulatory guidance for general purpose AI models

The AI Act guidelines cover transparency, copyright and risk assessment along with technical and governance risk mitigation.

Will Shanklin
Contributing Reporter
Thu, Nov 14, 2024, 5:34 PM EST · 2 min read

What are “open” models?

Definitions

Free/open source software licenses give users the freedom to run, copy, study, improve, and distribute code.

THE APACHE SOFTWARE FOUNDATION ESTABLISHED 1999

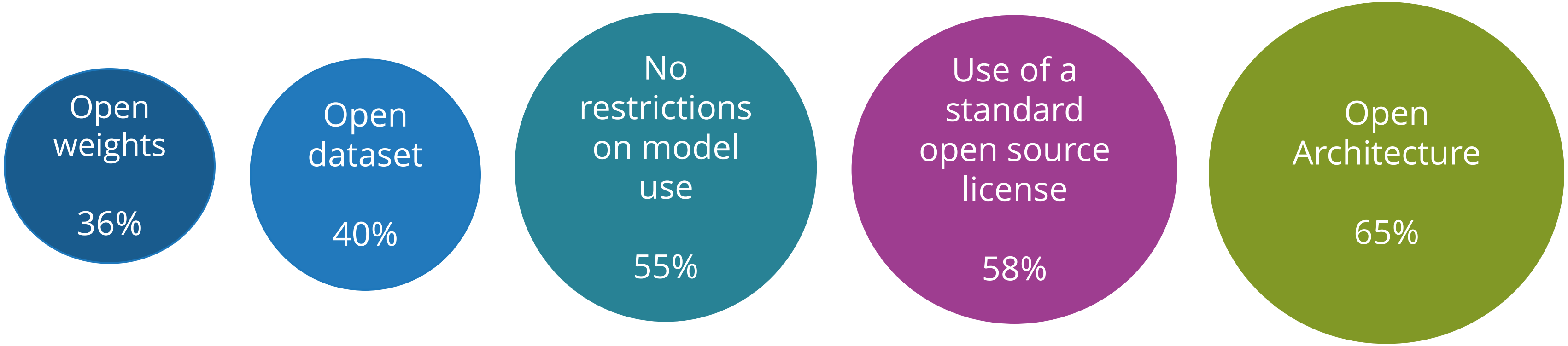
GenAI Models = Weights + Code + Data

open source initiative®

MODEL OPENNESS FRAMEWORK

Open weights model creators aren’t meeting expectations for openness

Which of the following components must be released openly for you to consider a foundation model to be open?



Open weights 36%

Open dataset 40%

No restrictions on model use 55%

Use of a standard open source license 58%

Open Architecture 65%

n = 200; Base=All Respondents
Notes: Managed by IDC's Global Primary Research Group.; Data Not Weighted; Multiple dichotomous table - total will not sum to 100%; Use caution when interpreting small sample sizes.
Source: U.S. Open Source Software Use Study, IDC, June, 2024

Open language models by license type

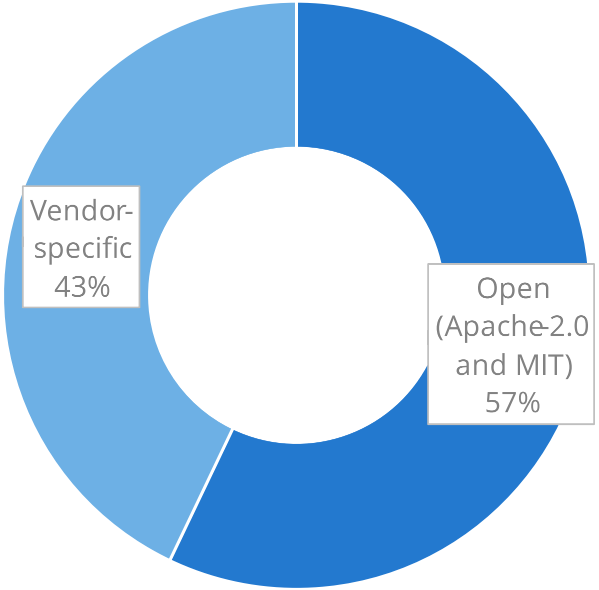
What license does the model use?

AI21 labs

databricks

Google

Technology Innovation Institute



ALLEN INSTITUTE

IBM

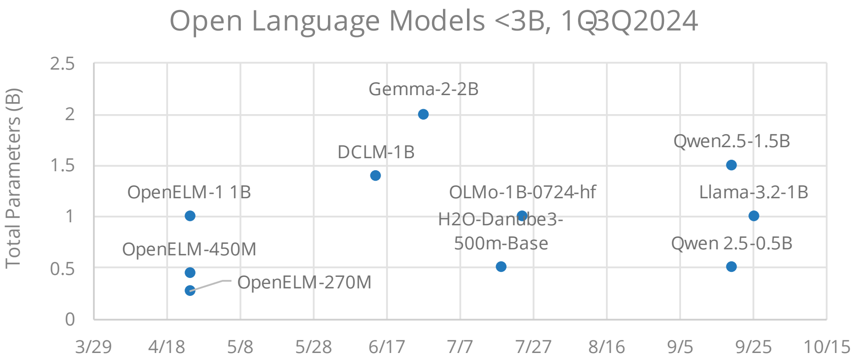
Microsoft

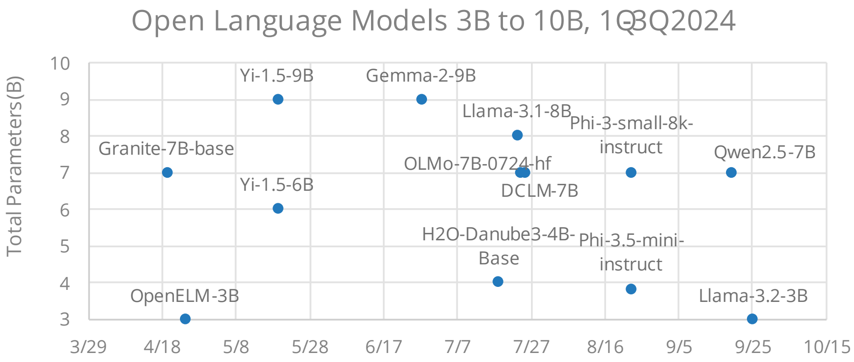
snowflake

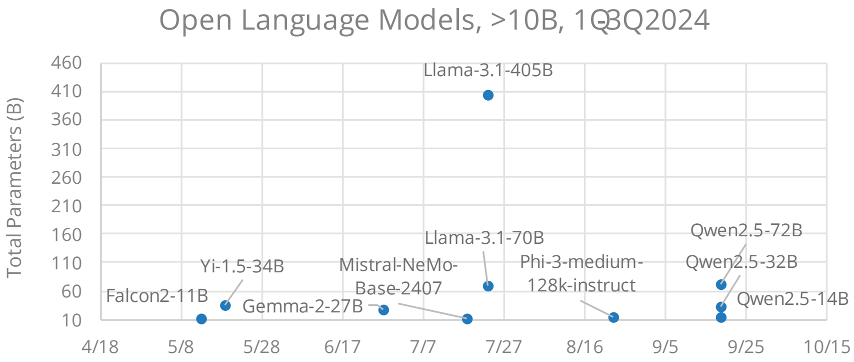
Source: IDC, 2024

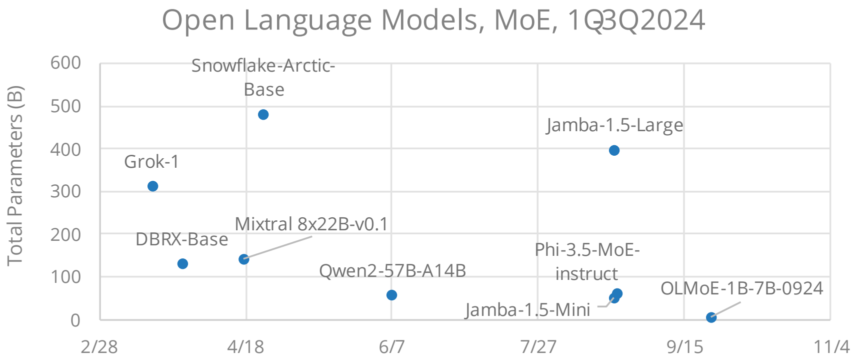
Who is creating open models?

Open Language Models, 1Q-3Q2024, by release date and model size


Open Language Models <3B, 1Q3Q2024

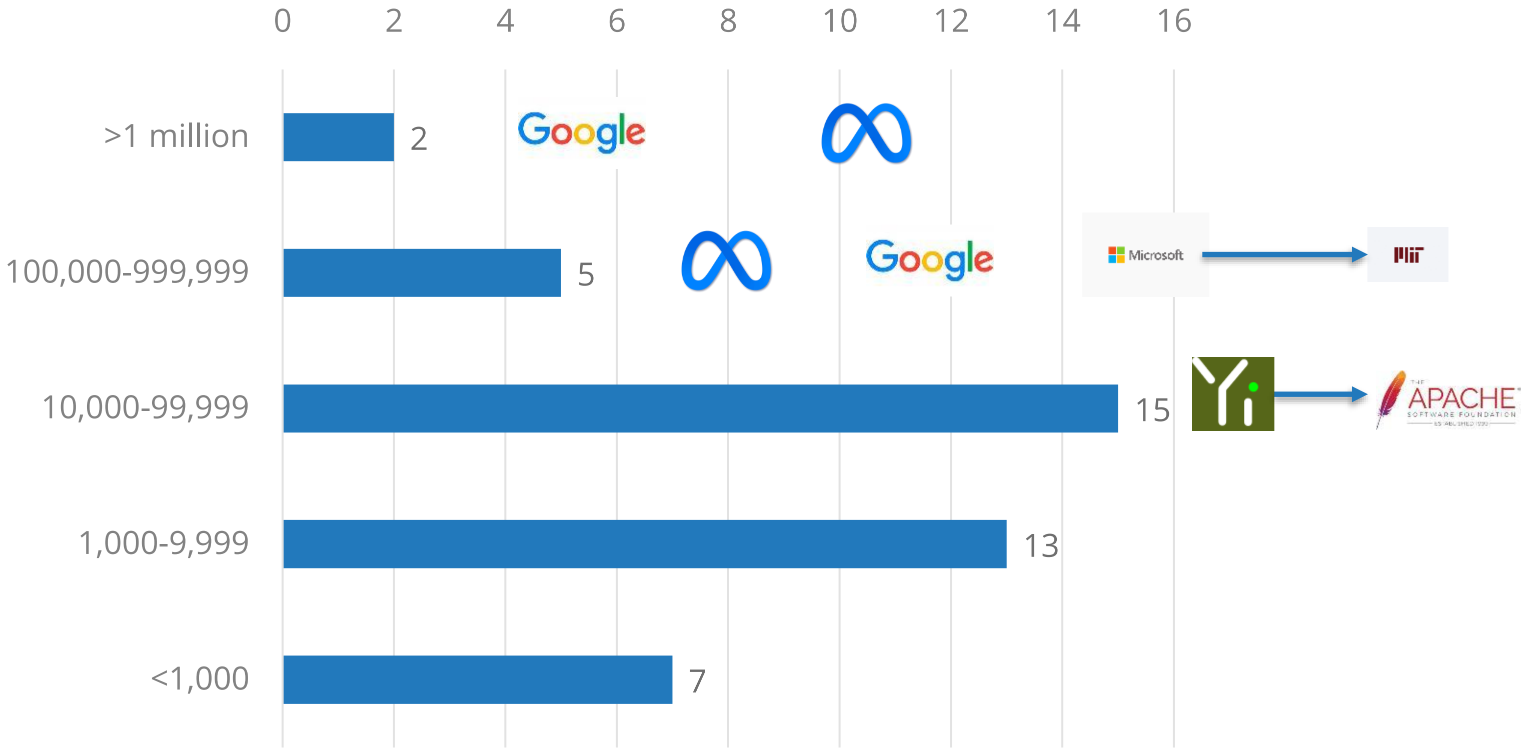

Open Language Models 3B to 10B, 1Q3Q2024


Open Language Models, >10B, 1Q3Q2024


Open Language Models, MoE, 1Q3Q2024

Source: IDC, 2024

Open language models by downloads

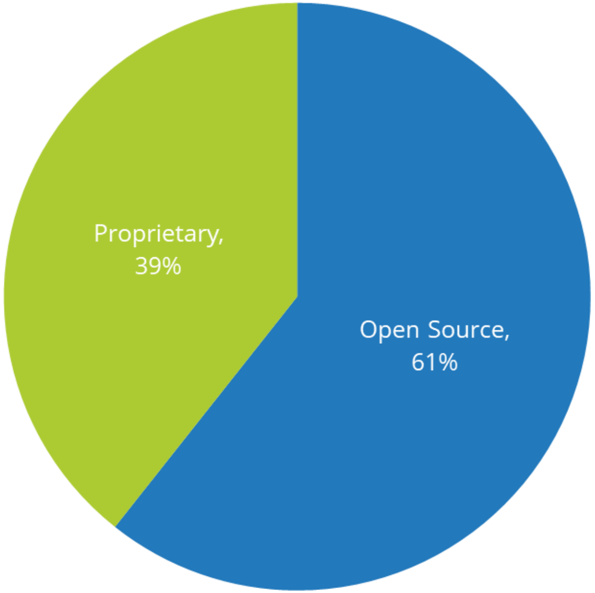


Source: IDC, 2024

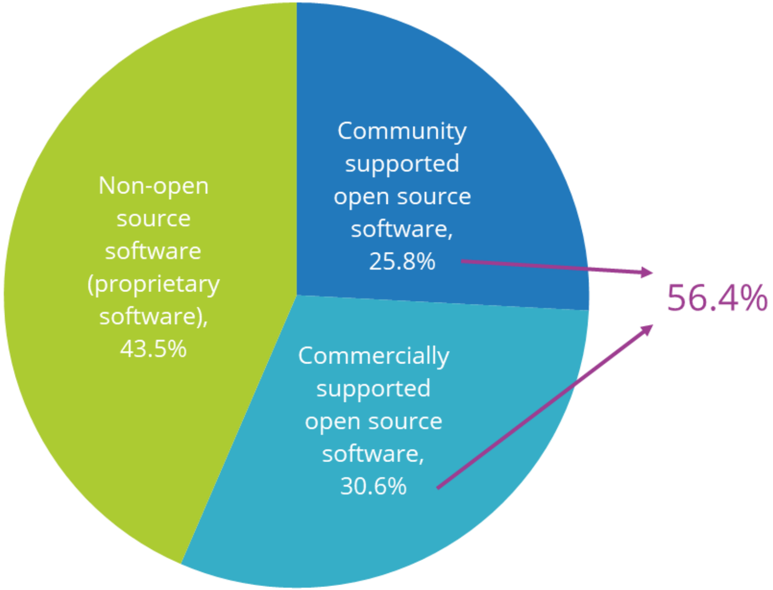
Who is using open models?

Foundation model use is split between open and closed models

Plans for open vs. closed model use by percentage of generative AI **use cases**



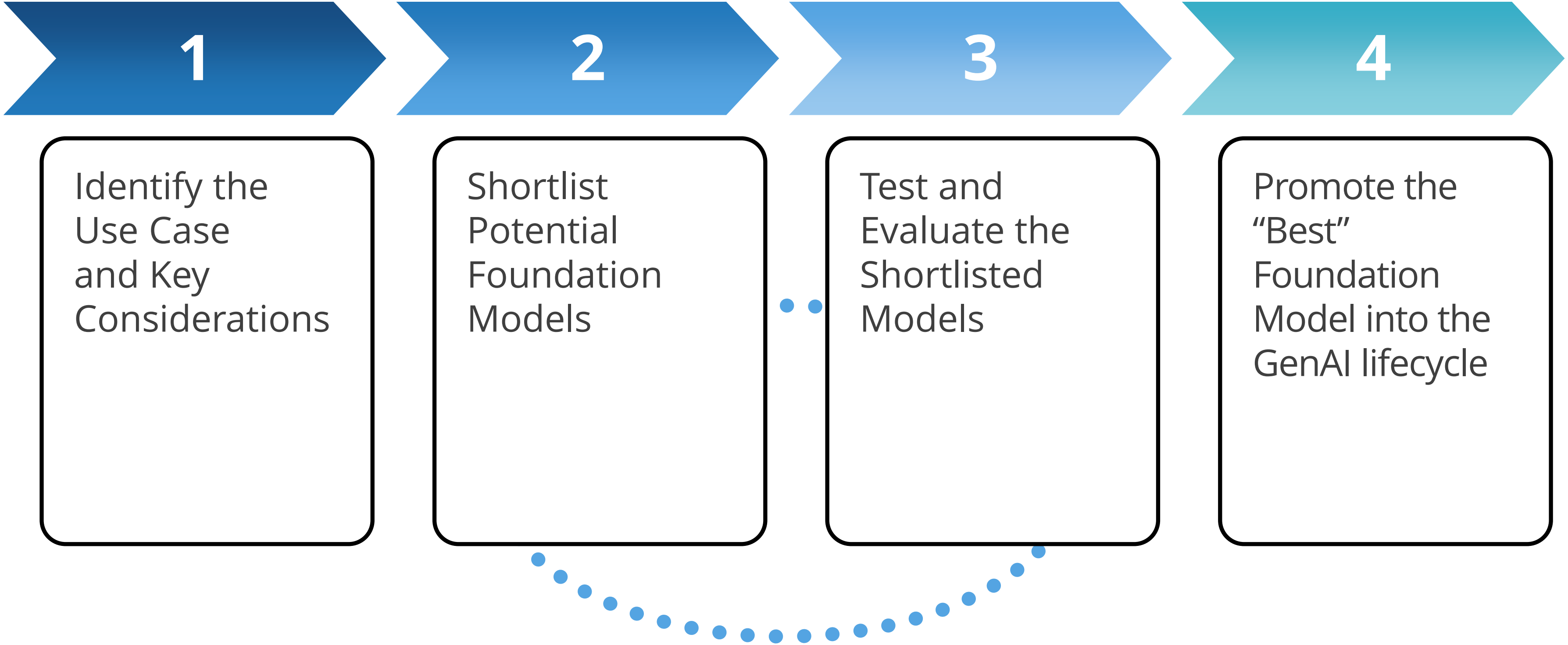
Percentage of foundation models in use on company servers and cloud instances



n = 200, Base=All Respondents
Notes: Managed by IDC's Global Primary Research Group.; Data Not Weighted; Managed by IDC's Global Primary Research Group.
Source: U.S. Open Source Software Use Study, IDC, June, 2024

Choosing an open model

Model Selection Framework



Identify the Use Case and Key Considerations

Shortlist Potential Foundation Models

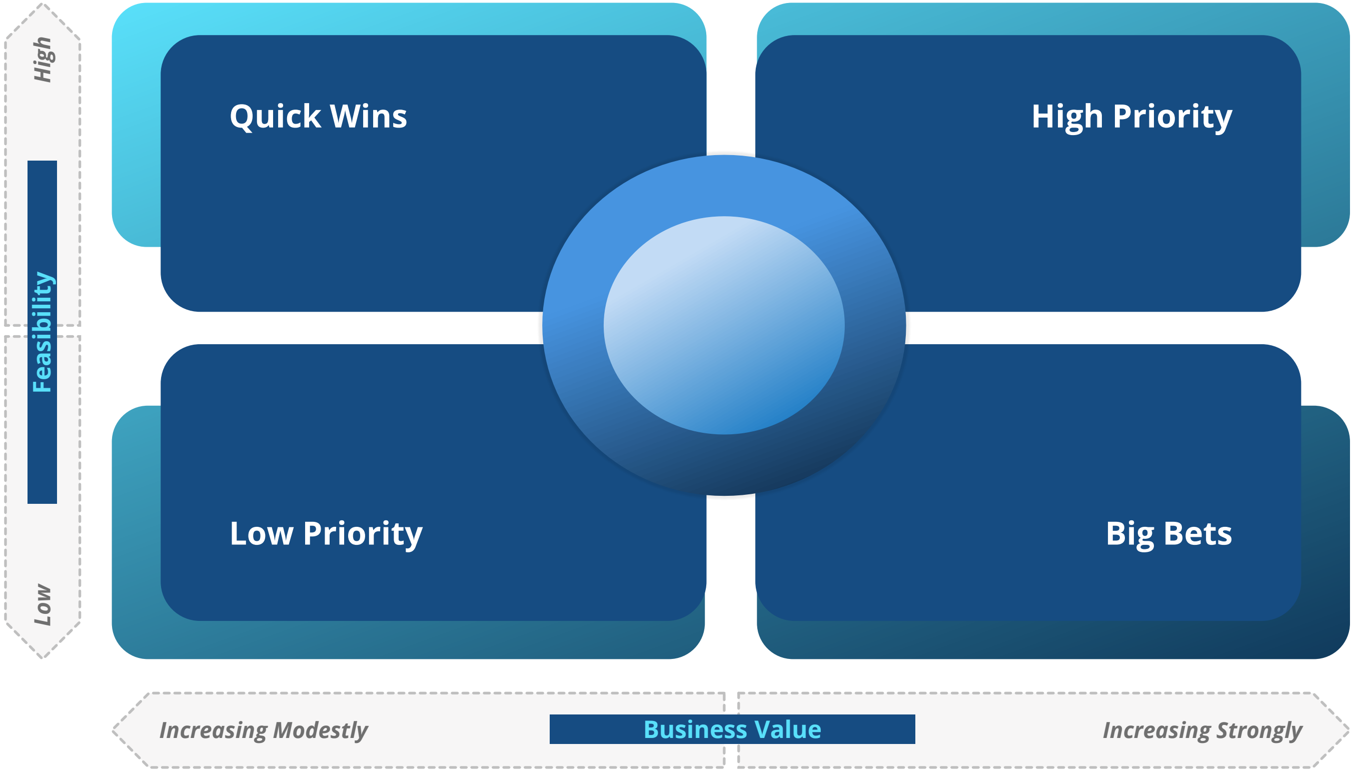
Test and Evaluate the Shortlisted Models

Promote the "Best" Foundation Model into the GenAI lifecycle

Use Case Prioritization Matrix

Identify the Generative AI Use Case

1

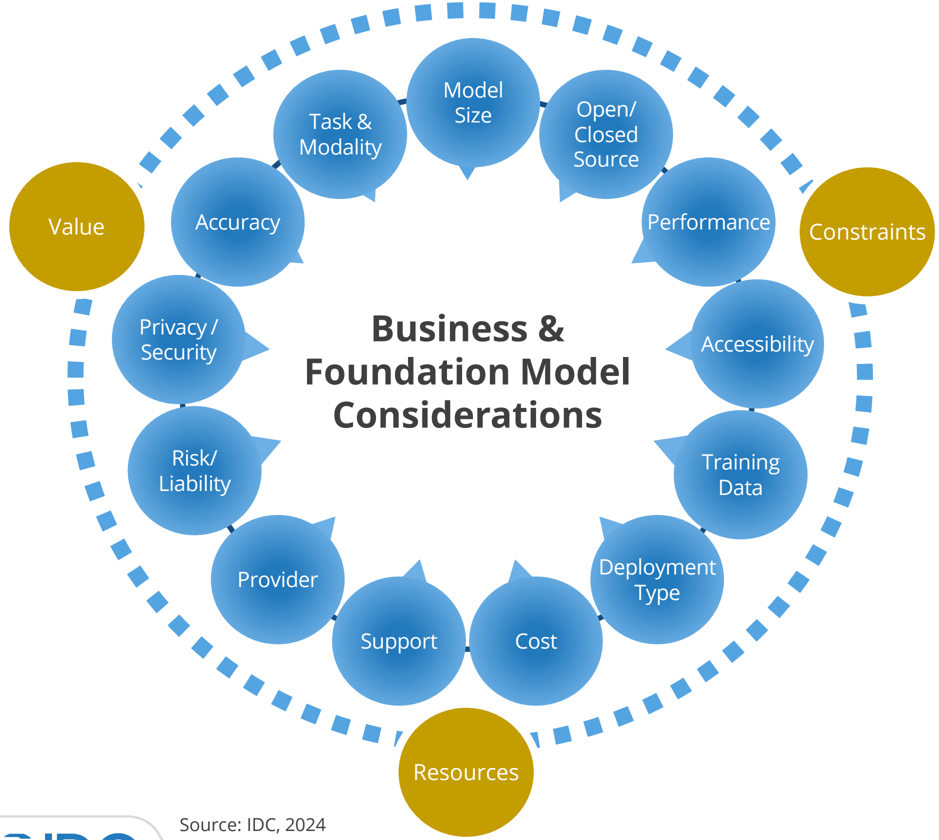


Source: Doc: # EUR150282023; June 2024

Prioritize Key Business and Foundation Model Considerations

Identify the Generative AI Use Case

1



Source: IDC, 2024

Top Criteria Influencing Foundation Model Choice	
Performance	41.1%
Cost	35.3%
Computational efficiency	29.0%
Training data size/quality	28.9%
Policy compliance	28.2%

Source: Future Enterprise Resiliency & Spending Survey Wave 7, IDC, July, 2024, N=891

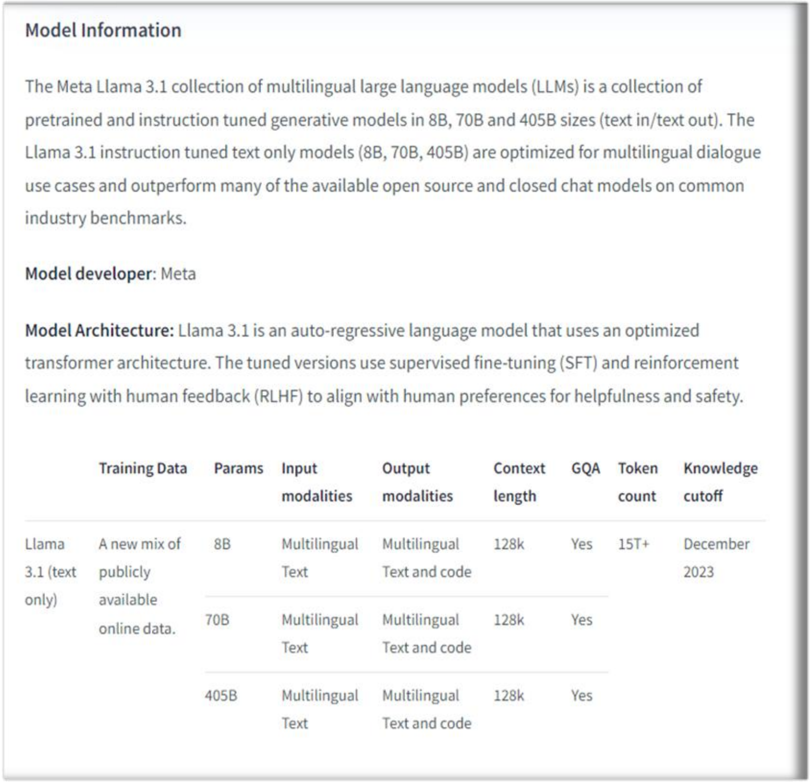
Use Model Cards to Filter on Key Model Attributes

Shortlist Potential Foundation Models

2

- Model type
- Model sizes
- Language capabilities

- Algorithm
- Fine-tuning methods



Model Information

The Meta Llama 3.1 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction tuned generative models in 8B, 70B and 405B sizes (text in/text out). The Llama 3.1 instruction tuned text only models (8B, 70B, 405B) are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

Model developer: Meta

Model Architecture: Llama 3.1 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

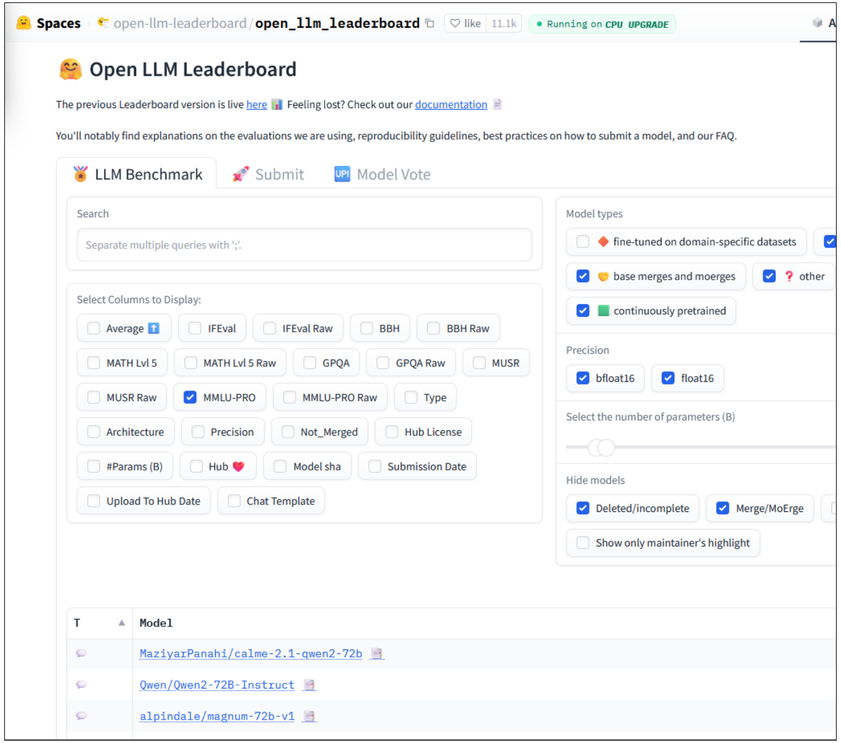
	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cutoff
Llama 3.1 (text only)	A new mix of publicly available online data.	8B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023
		70B	Multilingual Text	Multilingual Text and code	128k	Yes		
		405B	Multilingual Text	Multilingual Text and code	128k	Yes		

- Context length
- Token count
- Knowledge cutoff date

Evaluate model performance using third-party benchmarks

Shortlist Potential Foundation Models

2



Spaces · open-llm-leaderboard / open_llm_leaderboard ♡ like 11.1k ● Running on CPU UPGRADE

Open LLM Leaderboard

The previous Leaderboard version is live here 📊 Feeling lost? Check out our documentation 📄

You'll notably find explanations on the evaluations we are using, reproducibility guidelines, best practices on how to submit a model, and our FAQ.

LLM Benchmark | Submit | Model Vote

Search

Separate multiple queries with ';'.

Select Columns to Display:

☐ Average ⬆ ☐ IFEval ☐ IFEval Raw ☐ BBH ☐ BBH Raw
☐ MATH Lvl 5 ☐ MATH Lvl 5 Raw ☐ GPQA ☐ GPQA Raw ☐ MUSR
☐ MUSR Raw ☑ MMLU-PRO ☐ MMLU-PRO Raw ☐ Type
☐ Architecture ☐ Precision ☐ Not_Merged ☐ Hub License
☐ #Params (B) ☐ Hub ❤ ☐ Model sha ☐ Submission Date
☐ Upload To Hub Date ☐ Chat Template

Model types

☐ 🔶 fine-tuned on domain-specific datasets ☑
☑ base merges and moerges ☑ ❓ other
☑ 🟩 continuously pretrained

Precision

☑ bfloat16 ☑ float16

Select the number of parameters (B)

Hide models

☑ Deleted/incomplete ☑ Merge/MoErge
☐ Show only maintainer's highlight

T	Model
	MaziyarPanahi/calme-2.1-qwen2-72b
	Qwen/Qwen2-72B-Instruct
	alpindale/magnum-72b-v1

Commonly used benchmarks

MMLU: Measuring Massive Multitask Language Understanding (2020): 16,000 multiple choice questions spanning 57 academic subjects

HumanEval (2021): 164 original programming problems to evaluate models trained on code

Hellaswag (2019): Tests commonsense natural language inference by completing video captions

GSM-8k (2021): 8,500 grade school math problems to test multistep mathematical reasoning

GPQA (2023): 448 graduate-level multiple choice science questions

MATH (2021): 12,500 competition-level math problems

Compare, Test, and Evaluate Model Output in a Playground

A playground is a secure sandbox environment where developers and AI engineers can:

Test prompts with one or more models	**Compare** model output, performance, and cost
Evaluate prompt/model combinations (groundedness, context relevance, safety)	**Tune** prompts and parameters (temperature, max tokens, Top P, etc.)
Measure the quality and effectiveness of GenAI applications	**Prototype** AI applications

Takeaways

Takeaways

- Open models represent a significant business opportunity for technology vendors and their customers.
- Organizations are adopting these models for faster access to innovation, cost effectiveness, greater transparency, and flexibility.
- Before pursuing an open source AI strategy, organizations should consider their staff's expertise and other technical constraints.
- Beyond open weights, there are several openness factors to consider when selecting an open model:
    - license type (standard vs. vendor-specific)
    - components released (weights, code, data)
    - comprehensiveness of model documentation
    - access and use restrictions
    - openness of the training dataset
- The open model ecosystem is complex, but it can be navigated by taking an iterative, stepwise approach to model selection, beginning with identifying a specific use case and the key relevant model evaluation criteria.
- Vendors that help customers adopt open models have a major opportunity to gain market share.

IDC

Michele Rosen
mrosen@idc.com
https://www.linkedin.com/in/mfrosen/

IDC.com

linkedin.com/company/idc

twitter.com/idc

blogs.idc.com