Free Weights: Open Source AI

Open source (OS) is powering Gen AI innovation. Thanks to widely available academic research and platforms like GitHub and Hugging Face, we're witnessing a boom in major projects with impressive outcomes. Despite the considerable resources—money, computing power, and data—that closed-source tech giants pour into AI, open-source initiatives are tracking their growth and performance remarkably well.

A leaked internal memo from Google went viral in 2023 for its observation that open source AI has been subtly yet effectively "eating the lunch" of big tech companies like Google and OpenAI, boasting greater speed, adaptability, privacy, and overall efficiency. Open source AI is rapidly gaining on closed source in both popularity and performance – OS models like Mistral, Llama are catching up with and even outperforming some closed-source models.

Yann LeCun
@ylecun

Open source AI models will soon become unbeatable.
Period.

> **Bindu Reddy** @bindureddy · Oct 14, 2023
> The pace of open-source **LLM innovation** and **research** is breath-taking
>
> I suspect that open-source will soon become **unbeatable** for anyone except **maybe** OpenAI
> ...
> Show more

As a result open source AI is seeing substantial interest from developers, researchers and investors alike. Github witnessed a 148% YOY increase in developer contribution to Gen AI projects in 2023. More than $8B has been invested in open source AI over the last 2 years.

OS AI Ecosystem: Substantial growth in AI projects as well as contributors

Specifically for Gen AI, the term “open source” typically implies that the source code, any applicable weights and parameters (for training models) of these components are publicly accessible, usable, modifiable, and their distribution is permitted.

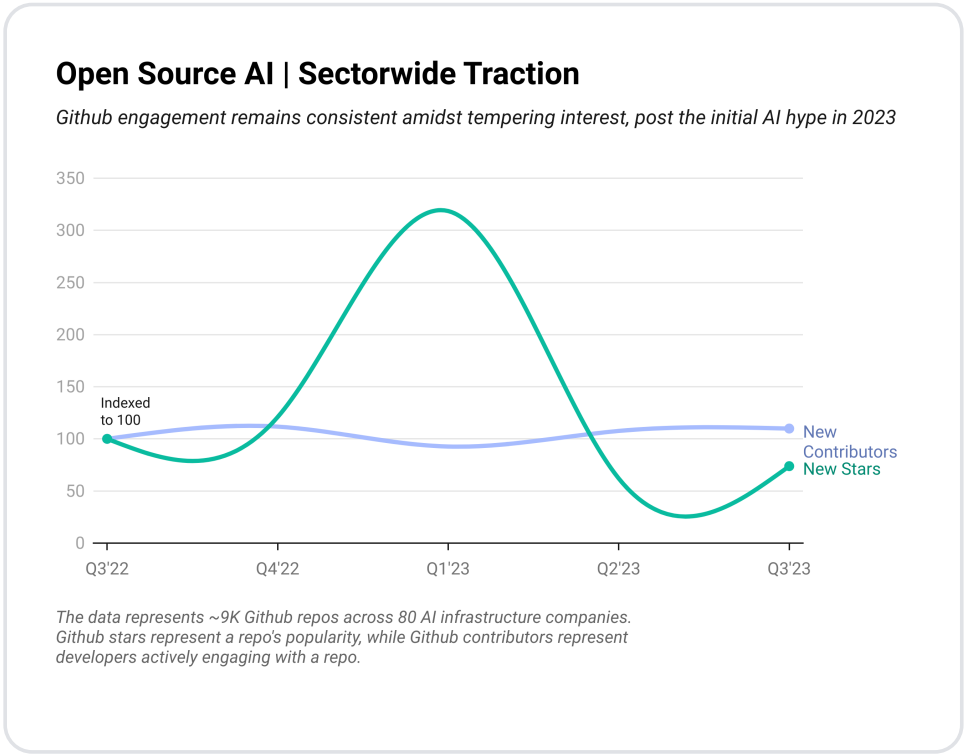
Adhering to this definition, the open source AI stack includes comprehensive set of tools to build Gen AI applications - foundational models (such as Llama, Mistral), developer tools & frameworks (such as Langchain, Fixie), model training platforms (such as Weights & Biases, Anyscale), and monitoring tools (Datadog, Seldon).

Open source AI innovation is thriving with new projects and developers:

Open source Gen AI projects are seeing significant and growing projects as well as contributors. Last year, Github witnessed 148% YOY growth in contributors and a 248% YOY growth in the total number of Gen AI projects. There are 60K Gen AI projects on Github and over 400K models on Huggingface as of 2023.

Contributor set is becoming increasingly Global, not restricted to US and Europe

Beyond the US and Europe – where a majority of open source projects originate from – the highest number of individual contributors to open source Gen AI came from India and Japan in 2023. Developers from Hong Kong, UK, Brazil, Germany and Singapore are also making numerous contributions to open source Gen AI. By 2027, India is projected to overtake US as the largest developer community on Github.



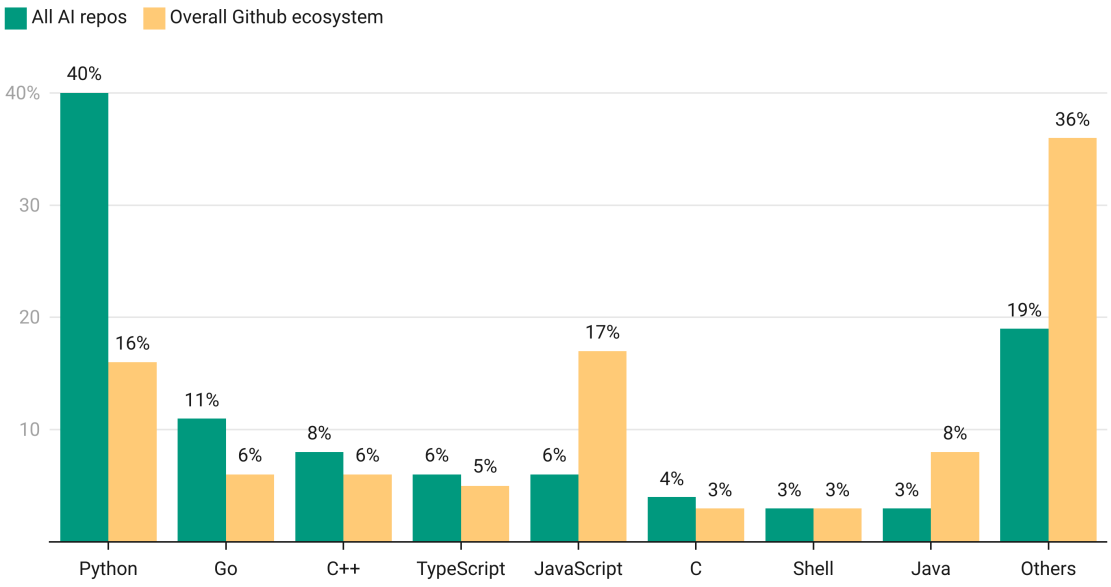
Steady increase in serious contributors, while “tourist” interest has tempered since Q1 hype

Gen AI overall has experienced a shift from initial widespread hype (peaking in Q1) to more focused and value-driven engagement - the "trough of disillusionment" phase, where initial excitement gives way to sustained, serious development.

Similar trend can be seen in # of stars across Github repos - the growth has tempered since Q1. On the other hand, serious developers (# of contributors to these projects) have grown steadily - 148% cumulatively in 2023.

Open Source AI | Programming Languages

Python stands out as the top language preference among AI developers



All AI repos includes ~9K repos across 80+ AI companies.
Overall Github ecosystem data represents all Github repos with >1K stars.

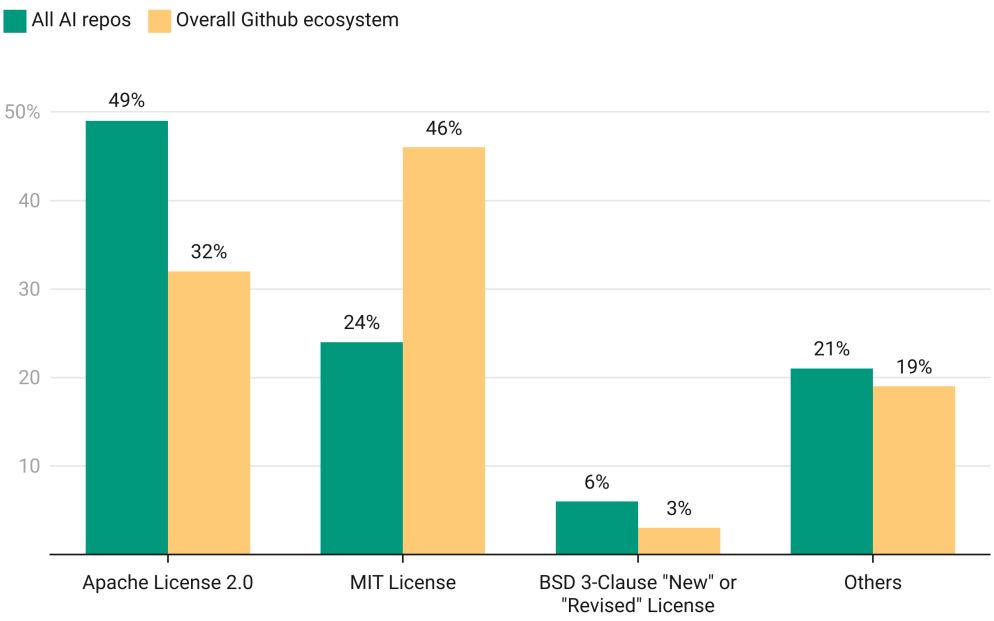
Python is the preferred language for open source AI

Gen AI overall has experienced a shift from initial widespread hype (peaking in Q1) to more focuseWhile Javascript has been the top programming language on Github in 2023, Python is the top choice when it comes to AI repositories. Python’s preference for ML projects has carried over to Gen AI because of its comprehensive ML libraries like TensorFlow and PyTorch. Python's flexibility in data handling and its platform-independent nature make it highly adaptable for diverse AI projects.

Mojo, a variation of Python that combines the usability of Python with the performance of C++, is gaining traction as an AI-specific programming language. In Q4’23, Mojo saw a 73% MOM increase in Github stars, indicative of the repo’s popularity amongst developers.

Open Source AI | Repository License

Apache License 2.0 is the most preferred among AI repositories.



All AI repos includes ~9K repos across 80+ AI companies.
Overall Github ecosystem data represents all Github repos with >1K stars.

AI repositories favouring more protective licensing

A disproportionate share of AI repos are using the Apache License, under which developers can claim patents on derivative projects. The Apache license is known to be extensive in legal terminology and therefore offer better patent protection than other licenses. Though the open MIT license is the most popular across Github; Gen AI developers are predictably keen on securing their work with more protective licensing.

Market Map: Multiple projects /startups emerging across the Gen AI tech stack

Open Source AI | Market Map

Open source development is highly active across the Generative AI tech stack, especially in foundational models & developer tools.

	Open Source	Closed Source (Illustrative)
Foundational Models *Core AI models that form the bedrock for creating or interpreting complex data, like text, images, and audio*	MISTRAL AI_, Meta, ADEPT, stability.ai, TII, cerebras, deci, LMSYS, Bloom, mosaicML, mozilla, Luma AI, contextual.ai, together.ai, LightOn, PlayHT, Craiyon, EleutherAI	OpenAI, ANTHROPIC, cohere, Gemini, perplexity, AI21labs, IIElevenLabs, Midjourney
Model Deployment & Inference *Cloud-based services providing scalable computing resources (GPUs), essential for model hosting, AI training & performance*	Replicate, baseten, Lepton AI, Kserve, Arrikto	CoreWeave, Lambda, Modal, RunPod
Developer tools *Platforms/toolkits that support the building and operation of AI models, such as application frameworks, vector databases, and data management systems.*	Hugging Face, mongoDB., supabase, zilliz, comet, weaviate, mindsdb, HOWSO, UNSTRUCTURED.IO, Chroma, neptune.ai, allegro, LangChain, drant, BENTOML, LlamaIndex, milvus, NEUM	Pinecone, Kindo
Model Training & Finetuning *Tools for generating synthetic data for model training, labeling data, and refining pre-trained models for specific tasks, business functions or data sets.*	anyscale, Weights & Biases, DOMINO, Snorkel, gretel, TONIC, Surge AI, Synthesis.ai, hazy, OpenPipe, PyTorch	scale, parallel domain, RENDERED.AI, fireworks.ai
Monitoring & Observability *Tools for tracking AI system performance, managing LLM operations, and analyzing user engagement*	DATADOG, arize, Arthur, fiddler, truera, SELDON, iterative, WHYLABS, Aquarium, Helicone, EVIDENTLY AI, PromptLayer	Amplitude, ROBUST INTELLIGENCE, vellum, GANTRY

All tools/frameworks/models that have a publicly-available source code which can be used, modified and distributed have been categorised as "Open Source".

Synaptic

Foundational models and developer tools, the core stack of AI, are the focus areas for new startups

Over 60% of new companies in the open source AI space are focusing on foundational models and developer tools, the core elements of the AI stack. This is expected, given that these components are fundamental for building, deploying, and managing generative AI applications across various use cases. Innovation in other areas like model training, fine-tuning tools, monitoring tools, and cloud computing services primarily revolves around these core AI stack elements.

High-quality open source AI reducing reliance on proprietary big tech AI, but data is key

The volume and quality of open source AI is sufficiently robust, enabling developers and startups to effectively compete with proprietary solutions. OS model Mixtral 8×7B surpassed closed source GPT 3.5 in chatbot as well as holistic performance. Other OS models like Llama, Yi are not far behind.

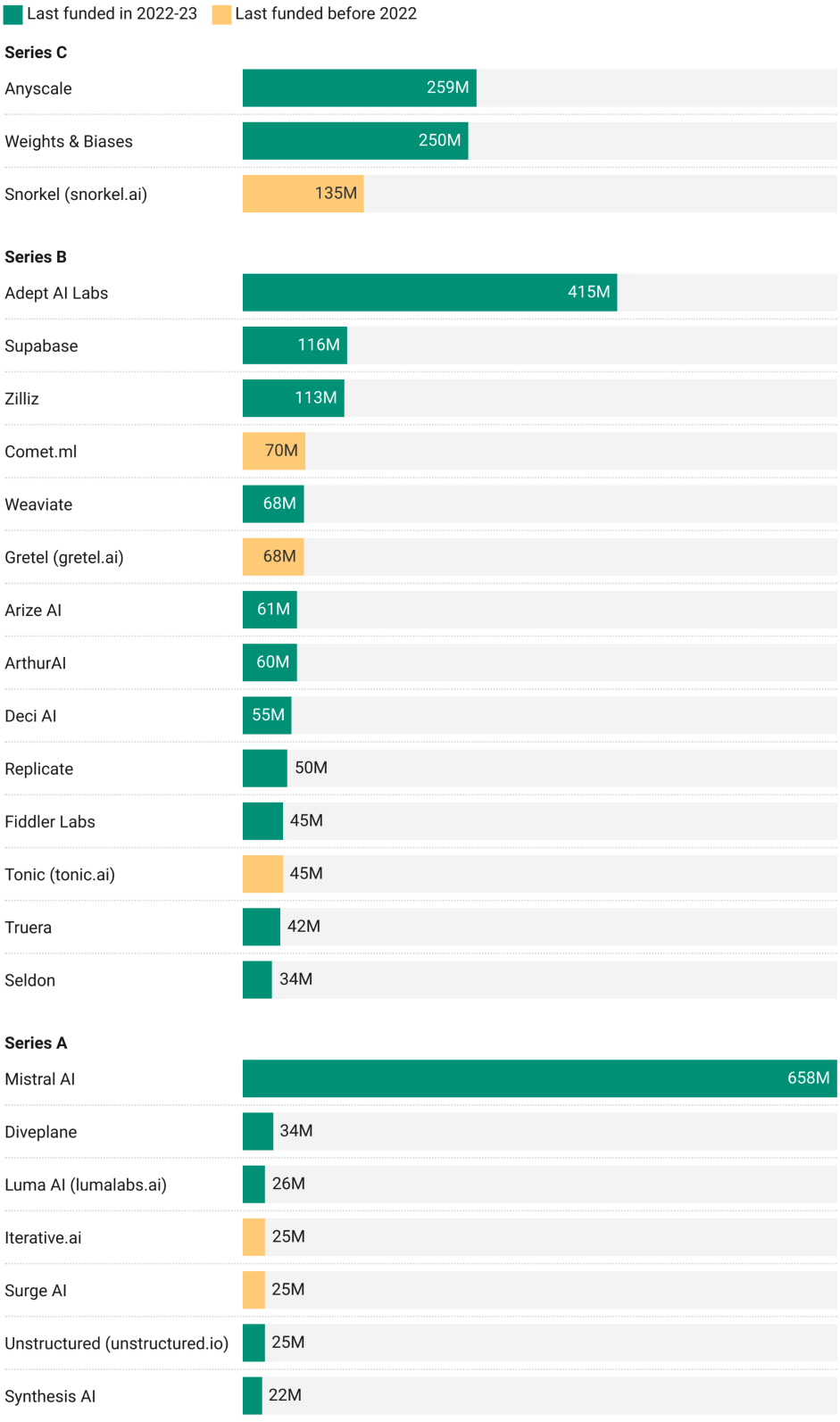
However, a crucial advantage that big tech firms with closed systems hold is their access to extensive data resources. This is evident in the fact that some recent OS models such as Llama-2 or Mistral 7B do not open source their training data. Data is likely to be the key proprietary element in the space.

Funding Landscape: Robust funding in 2022-23; foundational models & training tools secure maximum dollars

Gen AI infrastructure, due to its heavy reliance on vast amounts of data, extensive research, and substantial compute power, requires significant capital investment, has led to larger funding rounds compared to typical enterprise solutions.

Open Source AI | Funding Landscape

Robust funding activity in 2022-23; Mistral AI and Adept AI have received substantial deals.



Companies that are public, pre-seed, seed, merged, Series D+ or have >$1B funding have not been included in this chart.

Robust funding activity in 2022-23; foundational models and model training software secured maximum dollars

75% of open source AI startups secured funding in 2022-23. Foundational models and model training/fine-tuning software have attracted >70% of the investment dollars.

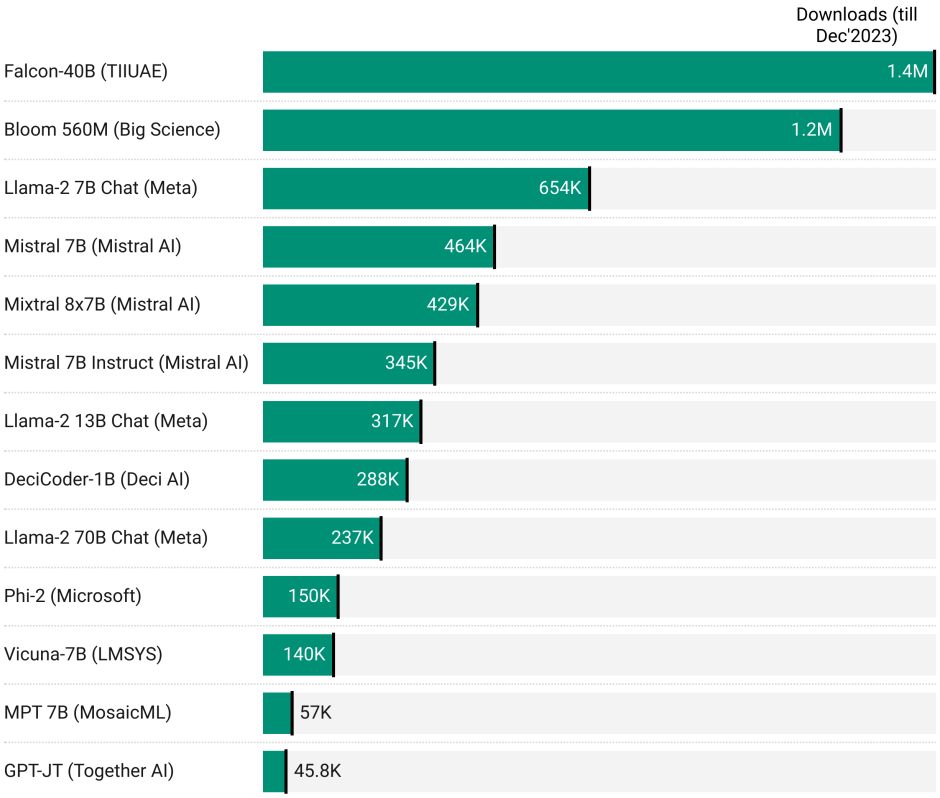
Nvidia, a leading graphics chips manufacturer for AI, has been a strategic investor in this space, with investments in top startups like Mistral AI and Adept AI.

Foundational Models: Open source models are catching up with closed source in popularity and performance

Gen AI infrastructure, due to its heavy reliance on vast amounts of data, extensive research, and substantial compute power, requires significant capital investment, has led to larger funding rounds compared to typical enterprise solutions.

Open Source AI | Foundational Models - Huggingface traction

Falcon and Bloom are seeing remarkable engagement on Huggingface



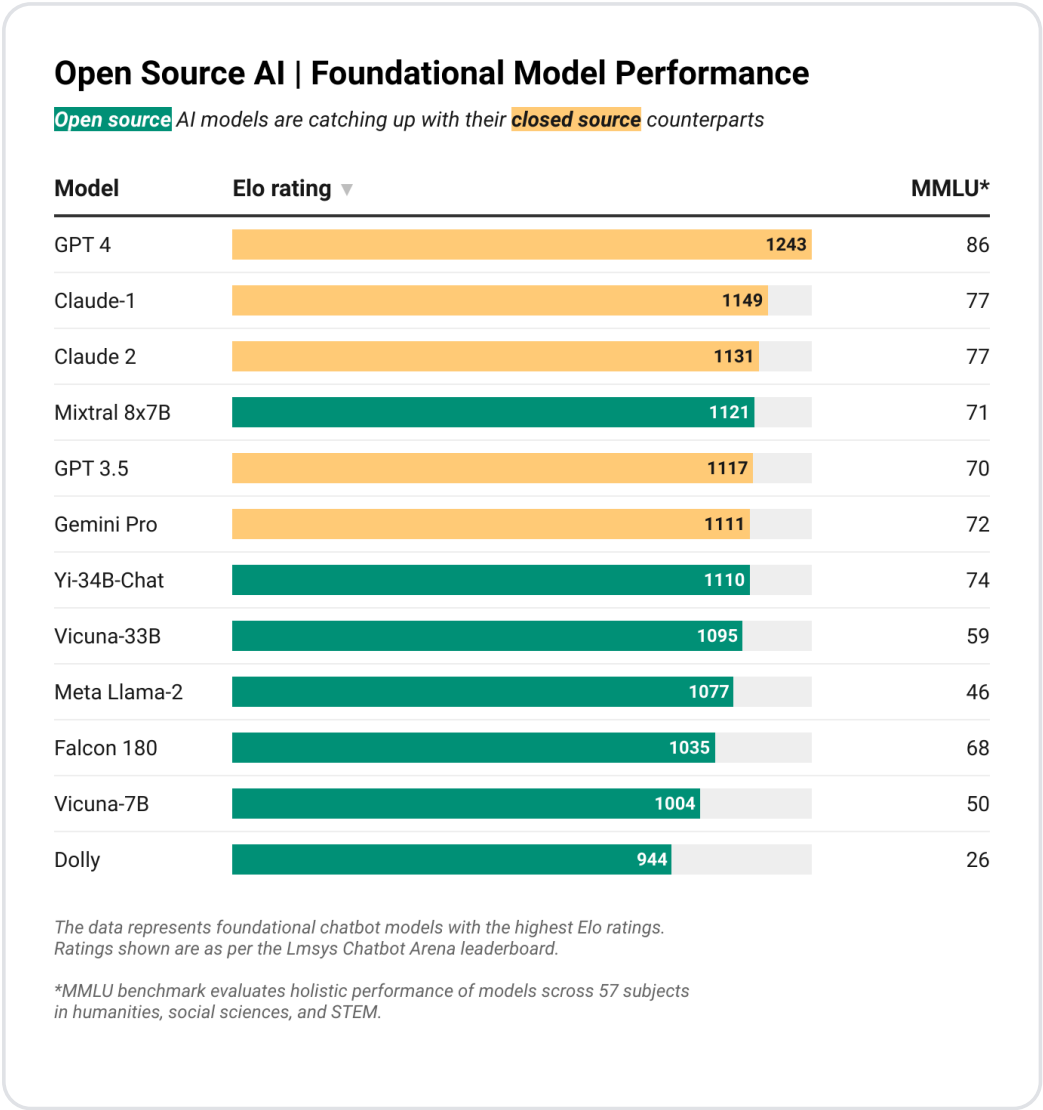
The data represents open source, text-generation foundational models on Huggingface with the highest downloads (as of Dec'23) after GPT 2.

Model versions fine-tuned by other organisations/individuals have not been shown.

Open source LLMs Falcon and Bloom have received significant engagement

Falcon, a large language model (LLM) developed by Abu Dhabi's Technology Innovation Institute, and BLOOM, created by the collaborative research organization BigScience, have recorded the highest downloads on Huggingface – surpassing Meta's Llama-2.

Launched recently, Mistral AI's models Mistral 7B and Mixtral 8×7B have gained significant popularity, surpassing many established models on Huggingface in terms of traction.



Open source AI models are not far behind closed source models

Although closed source big models like GPT4 and Claude are at the top of the chatbot leaderboard, open source models like Mistral, Vicuna, Yi, Llama are catching up – which bodes well for the ecosystem.

However, closed source models are still a step ahead according to the MMLU benchmark, which tests knowledge and problem-solving skills across 57 subjects in humanities, social sciences, and STEM. MMLU measures a model's comprehensive performance, and in this context, closed source models like GPT and Gemini continue to outperform open source alternatives.

Open source development is leading to higher efficiency in models

Startups working with open source AI, who don't posses the extensive data resources or compute power of major tech firms, are motivated to create more efficient models that deliver high-quality results with less computational demand.

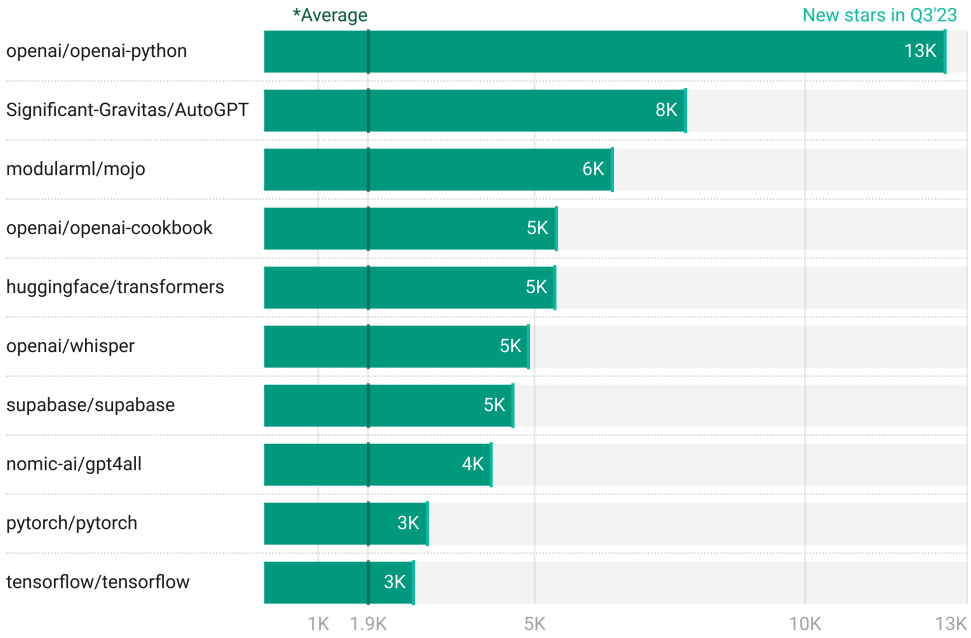
Mixtral 8×7B, an 85B-parameter 'mixture of experts' model that operates with the compute power of just a 14B model. It has outperformed all other open source models, including the larger Llama-2 70B, in terms of efficiency and effectiveness. This will be crucial in making these models more accessible to local applications (e.g. voice assistants on mobile).

Github Traction: AutoGPT, Mojo attracting significant developer interest

Gen AI infrastructure, due to its heavy reliance on vast amounts of data, extensive research, and substantial compute power, requires significant capital investment, has led to larger funding rounds compared to typical enterprise solutions.

Open Source AI | Github Traction

AutoGPT (AI agent) and Mojo (programming language) have seen significant traction in Q3'23.



The data represents the top 10 repos with the highest new Github stars in Q3'23, from our dataset of ~9K AI repos.

**The "Average" line depicts average new stars in Q3'23 for all repos with >10K cumulative stars.*

AutoGPT, Modular's Mojo are witnessing high developer traction

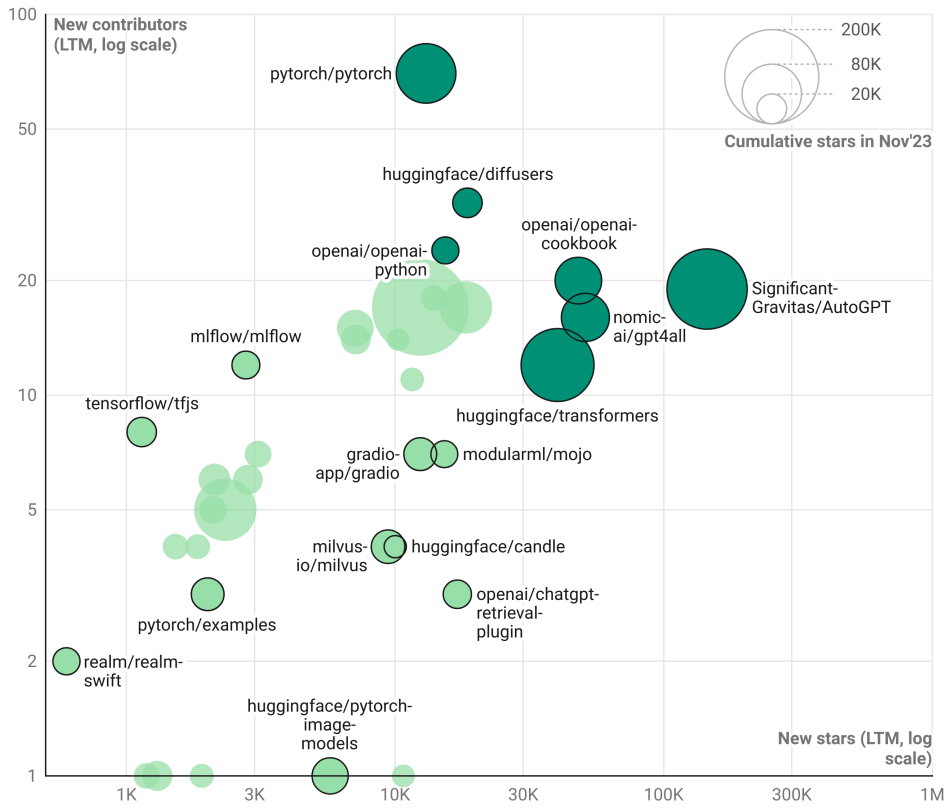
As a primary platform for developers to interact with and contribute to open source AI projects, GitHub activity tends to be a strong indicator of traction. GitHub stars (similar to a "follow" on social media) are a direct indicator of a project's popularity on GitHub.

AutoGPT, an autonomous AI assistant built on GPT4, has received significant developer traction. The model is capable of acting as an AI agent, breaking a large task into various sub-tasks without the need for user input, which are then chained together and performed sequentially to yield a larger result. AutoGPT is also capable of connecting to the internet, thereby allowing for up-to-date information retrieval for its tasks.

ModularML's Mojo is a variation of Python tailored for high-performance AI applications, balancing the efficiency of languages like C++ and Rust with Python's simplicity. Mojo's core goals are to streamline AI development, integrate AI/ML infrastructure seamlessly, and deliver robust performance.

Open Source AI | Github Engagement

AutoGPT, Pytorch repos have witnessed significant growth in both stars and contributors



The data represents the growth in Github stars and Github contributors from November'22 to '23.
The chart includes all AI repos with >10K stars on Github.

Pytorch, Huggingface, AutoGPT, and Supabase stand out with the significant engagement on Github

Github contributors are developers who make changes (known as “commits”) to the code, actively engaging with the repo to improve it. Contributors are indicative of serious developer activity on repos.

The year-over-year GitHub analysis for 2022-23 highlights a notable uptick in both interest and active engagement with various repositories among the developer community.

Though AI agents are still in an experimental phase from a customer-facing POV, GitHub data reveals substantial developer activity in this domain, which could likely lead to some agent-based AI apps emerging soon. AutoGPT, an AI agent repo, is experiencing significant developer activity on GitHub. Other AI agent repos like Bloop, X-Agent are also seeing similar interest from the developer community.

Looking ahead

- **Open source isn’t merely a playground for Gen AI, its at the forefront of innovation**
  Open source AI is seeing active innovation - Github saw a 148% annual growth in contributors and a 248% annual growth in Gen AI projects in 2023, HuggingFace has 400k+ models. Open Source stack for Gen AI is competitive or better than proprietary products across categories - from foundational models to infra & tooling.

- **Open Source models are not far from flagship proprietary models in performance and are leading in efficiency, achieving this performance with lower compute & data**
  Open source models like Mistral, Vicuna, Yi, and Llama are rapidly catching up to closed source leaders like GPT4 and Claude, with Mixtral 8×7B even surpassing GPT3.5 in Elo and MMLU ratings. Open source development is fostering more compute-efficient models, which will be crucial for deploying AI locally on devices (e.g. mobile phones).

- **Access to high quality, abundant data will be the limiting function for OS AI models**
  Data will be a key battleground for the development of large models. Recent models, such as Llama-2, Mistral 7B, which we released as “open source”, have chosen not to make their model training data publicly accessible. Big Tech, of course, will have a significant advantage on data. Synthetic data platforms (like Gretel) can potentially augment training and fine-tuning, but expect data-protectionism to increase (NYT vs GPT is a case in point).

- **AI agents are seeing significant developer activity, expect killer agent-based applications on the market soon**
  While AI agents are still largely experimental & nascent in customer-facing applications (see our article on productivity tools), Github data indicates serious & continuing developer interest in agents. There are 70+ AI Agent repos on Github as of today, with repos like AutoGPT, Bloop, XAgent getting significant traction (8-10K+ stars) and engagement (30+ contributors). Definitely an area to keep an eye out for.

- **Expect standout open source AI projects to attract big rounds in 2024**
  Startups in open source AI have seen some extraordinarily large deals and active funding rounds across stages. Mistral AI obtained its unicorn status after a recent $487M deal. AutoGPT, Supabase and DeciAI are poised for future funding rounds in the next 1-2 years.

About Synaptic

Synaptic is an alternative data and intelligence platform that helps investors, data teams and research analysts leverage alternative datasets across their investment processes.

For more information, visit us at synaptic.com or write to us at contact@synaptic.com