BIZ & IT / INFORMED TECHNOLOGY

Exponential growth brews 1 million AI models on Hugging Face

Hugging Face cites community-driven customization as fuel for diverse AI model boom.

Benj Edwards - Sep 26, 2024 7:50 pm UTC

Enlarge

Hugging Face / anucha sirivisansuwan via Getty Images

On Thursday, AI hosting platform Hugging Face surpassed 1 million AI model listings for the first time, marking a milestone in the rapidly expanding field of machine learning. An AI model is a computer program (often using a neural network) trained on data to perform specific tasks or make predictions. The platform, which started as a chatbot app in 2016 before pivoting to become an open source hub for AI models in 2020, now hosts a wide array of tools for developers and researchers.

The machine-learning field represents a far bigger world than just large language models (LLMs) like the kind that power ChatGPT. In a post on X, Hugging Face CEO Clément Delangue wrote about how his company hosts many high-profile AI models, like "Llama, Gemma, Phi, Flux, Mistral, Starcoder, Qwen, Stable diffusion, Grok, Whisper, Olmo, Command, Zephyr, OpenELM, Jamba, Yi," but also "999,984 others."

The reason why, Delangue says, stems from customization. "Contrary to the '1 model to rule them all' fallacy," he wrote, "smaller specialized customized optimized models for your use-case, your domain, your language, your hardware and generally your constraints are better. As a matter of fact, something that few people realize is that there are almost as many models on Hugging Face that are private only to one organization—for companies to build AI privately, specifically for their use-cases."

Model Growth Monthly
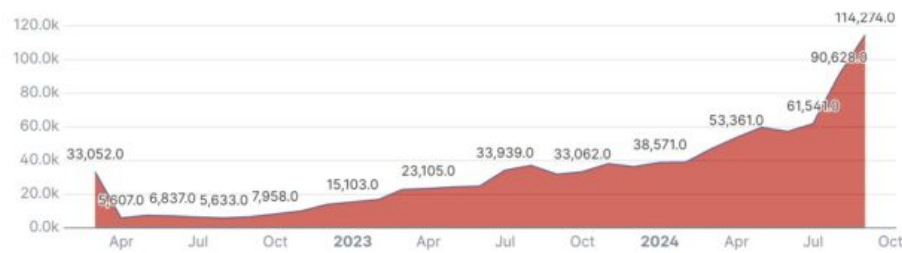
/ A Hugging Face-supplied chart showing the number of AI models added to Hugging Face over time, month to month.
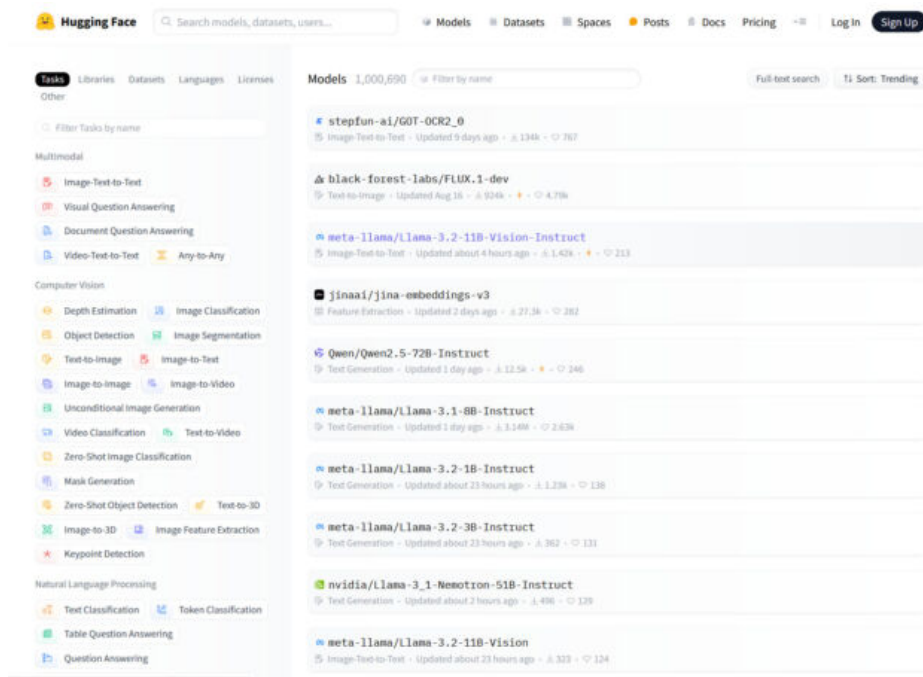
Hugging Face

Hugging Face's transformation into a major AI platform follows the accelerating pace of AI research and development across the tech industry. In just a few years, the number of models hosted on the site has grown dramatically along with interest in the field. On X, Hugging Face product engineer Caleb Fahlgren posted a chart of models created each month on the platform (and a link to other charts), saying, "Models are going exponential month over month and September isn't even over yet."

The power of fine-tuning

As hinted by Delangue above, the sheer number of models on the platform stems from the collaborative nature of the platform and the practice of fine-tuning existing models for specific tasks. Fine-tuning means taking an existing model and giving it additional training to add new concepts to its neural network and alter how it produces outputs. Developers and researchers from around the world contribute their results, leading to a large ecosystem.

For example, the platform hosts many variations of Meta's open-weights Llama models that represent different fine-tuned versions of the original base models, each optimized for specific applications.

Hugging Face's repository includes models for a wide range of tasks. Browsing its models page shows categories such as image-to-text, visual question answering, and document question answering under the "Multimodal" section. In the "Computer Vision" category, there are sub-categories for depth estimation, object detection, and image generation, among others. Natural language processing tasks like text classification and question answering are also represented, along with audio, tabular, and reinforcement learning (RL) models.

Enlarge / A screenshot of the Hugging Face models page captured on September 26, 2024.

Hugging Face

When sorted for "most downloads," the Hugging Face models list reveals trends about which AI models people find most useful. At the top, with a massive lead at 163 million downloads, is Audio Spectrogram Transformer from MIT, which classifies audio content like speech, music,

and environmental sounds. Following that, with 54.2 million downloads, is BERT from Google, an AI language model that learns to understand English by predicting masked words and sentence relationships, enabling it to assist with various language tasks.

Rounding out the top five AI models are all-MiniLM-L6-v2 (which maps sentences and paragraphs to 384-dimensional dense vector representations, useful for semantic search), Vision Transformer (which processes images as sequences of patches to perform image classification), and OpenAI's CLIP (which connects images and text, allowing it to classify or describe visual content using natural language).

No matter what the model or the task, the platform just keeps growing. "Today a new repository (model, dataset or space) is created every 10 seconds on HF," wrote Delangue. "Ultimately, there's going to be as many models as code repositories and we'll be here for it!"

Benj Edwards / Benj Edwards is Ars Technica's Senior AI Reporter and founder of the site's dedicated AI beat in 2022. He's also a widely-cited tech historian. In his free time, he writes and records music, collects vintage computers, and enjoys nature. He lives in Raleigh, NC.
@benjedwards