zensar

Databricks DBRX:
Democratizing the LLM World

White paper

An RPG Company

The AI revolution has gained remarkable momentum with the growing adoption of ChatGPT, and discussions around generative artificial intelligence (GenAI) and its usage are rife worldwide.

AI has been around in one form or another for a long while, growing in leaps and bounds from the Turing Test to supervised learning models to GenAI and its evolving large language model avatars today.

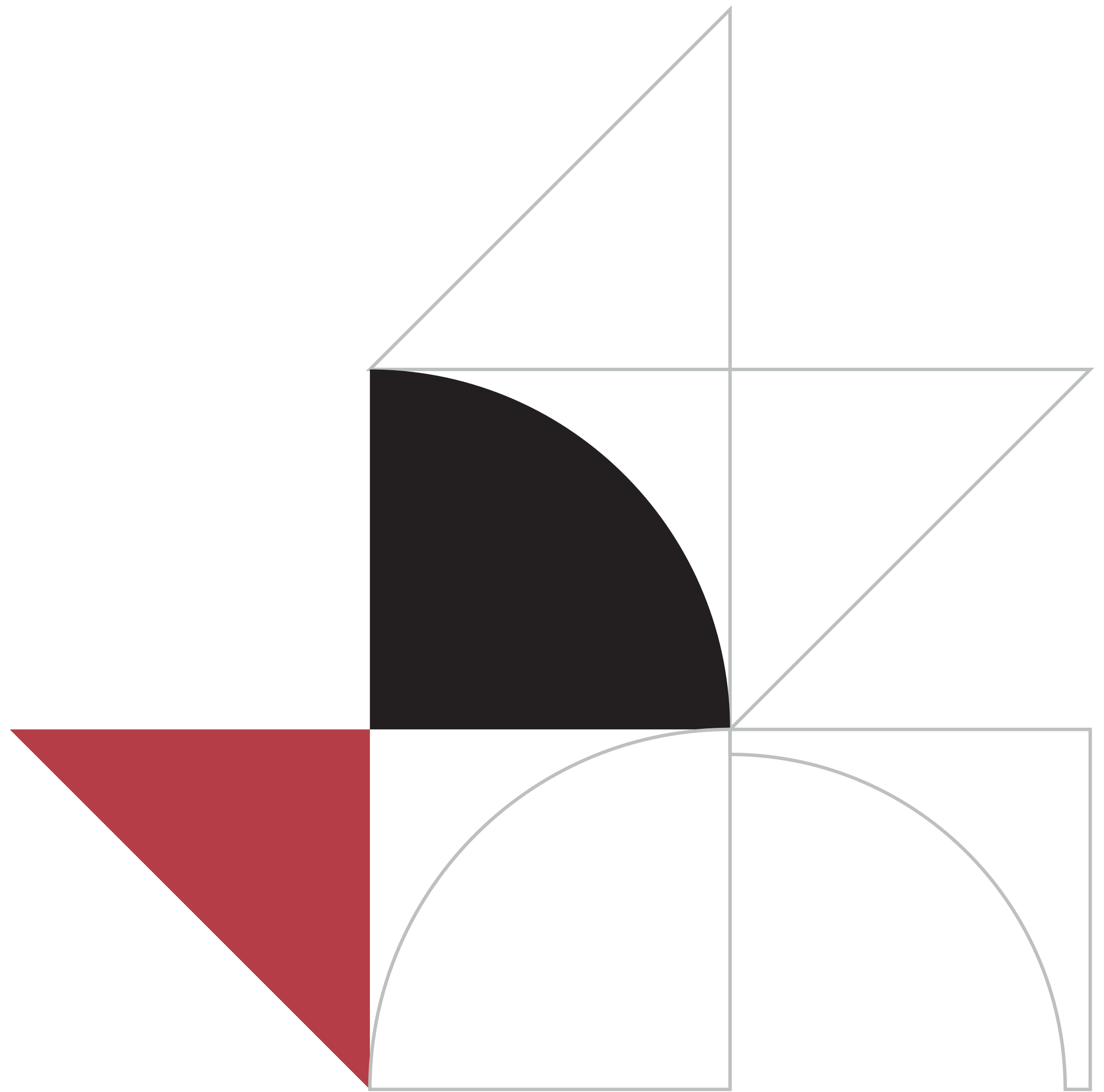
Generative artificial intelligence is artificial intelligence capable of generating text, images, videos, or other data using generative models, often in response to prompts. GenAI models learn the patterns and structure of their input training data and then generate new data with similar characteristics.

Industries ranging from healthcare and pharma to advertising and software are increasingly finding practical applications of GenAI to improve the range, quality, and speed of essential tasks. Whether it is generating code using Devin AI, predicting weather in agriculture, superior customer experience in insurance, or drug discovery, GenAI is playing a significant role in enhancing productivity and spurring innovation across the globe. GenAI is positioned at the Peak of Inflated Expectations on the Gartner Hype Cycle for Emerging Technologies, 2023, and is projected to reach transformational benefit within two to five years. The optimism around GenAI's capability is also reflected in the share price of the world leader in AI computing, NVIDIA.

GenAI's impact on world economy and productivity

An Accenture study found that artificial intelligence could add $14 trillion to the global economy by 2035, with the most significant gains in China and North America. The study also predicted that AI could increase labor productivity by up to 40 percent in some industries.

A McKinsey report found that AI could automate up to 45 percent of the tasks currently performed by retail, hospitality, and healthcare workers. While this could lead to job displacement, the report also noted that just because AI could automate a job doesn't necessarily mean that it will, as cost, regulations, and social acceptance can also be limiting factors.



Large language models – open vs closed

The latest innovations in GenAI—large language models (such as OpenAI’s ChatGPT)—are gaining huge traction.

These innovations are attributed to advancements in machine learning and the vast increase in computational power, which enable these models to process and learn from billions of words and text on the Internet.

A large language model (LLM) is an artificial intelligence program that can recognize and generate text, among other tasks. LLMs are trained on huge data sets—hence the name "large." LLMs are built on machine learning – specifically, a type of neural network called a transformer model. They have the same relationship with data as any engine has with oil, and good quality oil at that.

In late 2022, Microsoft-backed OpenAI shaped the landscape of LLMs with the introduction of GPT-3.5, marking a pivotal moment in the AI world. GPT-3.5 was not fully open-sourced, which gave rise to closed-source large language models.

Closed-source LLMs, such as GPT 3.5 by OpenAI and Gemini by Google, treat model architecture and weights as their internal assets. There is no access to code or design-level details, and they have limited customizations.

Advantages of closed LLMs

- Controlled quality and consistencyVendor support and proprietary
- innovations
- Smooth integration

On the other hand, open-source LLMs such as Databricks DBRX have publicly accessible model architectures, source code, and weight parameters, which enables the open-source community to analyze the models, assess their quality, and customize them. DBRX by Databricks (which we will examine in detail) and LLaMA 2 by Meta are prime examples of open LLMs that are democratizing the LLM space.

Advantages of open LLMs

- Innovation
- Broad participation and diverse contribution
- Continuous improvement through constant feedback
- Knowledge sharing

Image courtesy of Databricks

Databricks DBRX

Databricks DBRX is an open, general-purpose large language model developed by Databricks. It has set new standards and is considered a state-of-the-art open-source LLM.

DBRX provides the open community and enterprises building their own LLMs with capabilities that were previously limited to closed model APIs, surpassing GPT-3.5 and rivaling Gemini 1.0 Pro. DBRX excels in various benchmarks, including language understanding, programming, and mathematics.
It is trained using next-token prediction with a fine-grained mixture-of-experts (MoE) architecture, resulting in significant improvements in training and inference performance. The model is available for Databricks customers via APIs, and Databricks customers can pre-train their own DBRX-class models from scratch or continue training on top of one of its checkpoints using the same tools and science used to build it. Its efficiency is highlighted by the training and inference performance, surpassing other established models while being smaller in size than similar models.

DBRX is a game changer in terms of Databricks' next generation of Gen-AI products, designed to empower enterprises and the open community.

Unique architecture

DBRX is a transformer-based decoder-only large language model that was trained using next-token prediction. Its innovative MoE architecture utilizes 132 billion total parameters, of which 36B parameters are active on any input. This focus on active parameters significantly improves efficiency compared to other models. DBRX boasts inference speeds that are up to twice as fast as LLaMA2-70B. Additionally, it boasts a compact size, being roughly 40 percent smaller than Grok-1 in both total and active parameter counts. When hosted on Mosaic AI Model Serving, DBRX delivers text generation speeds of up to 150 tokens per second per user. The training process for DBRX demonstrates significant improvements in compute efficiency.

Databricks DBRX vs leading open LLMs

DBRX Instruct is the leading model on composite benchmarks, programming, and mathematics benchmarks, and MMLU. It surpasses all chat or instruction fine-tuned models on standard benchmarks.

Databricks DBRX vs leading closed models

DBRX Instruct surpasses or - at worst - matches GPT-3.5. DBRX Instruct outperforms GPT-3.5 on general knowledge and commonsense reasoning.

DBRX Instruct especially shines on programming and mathematical reasoning.

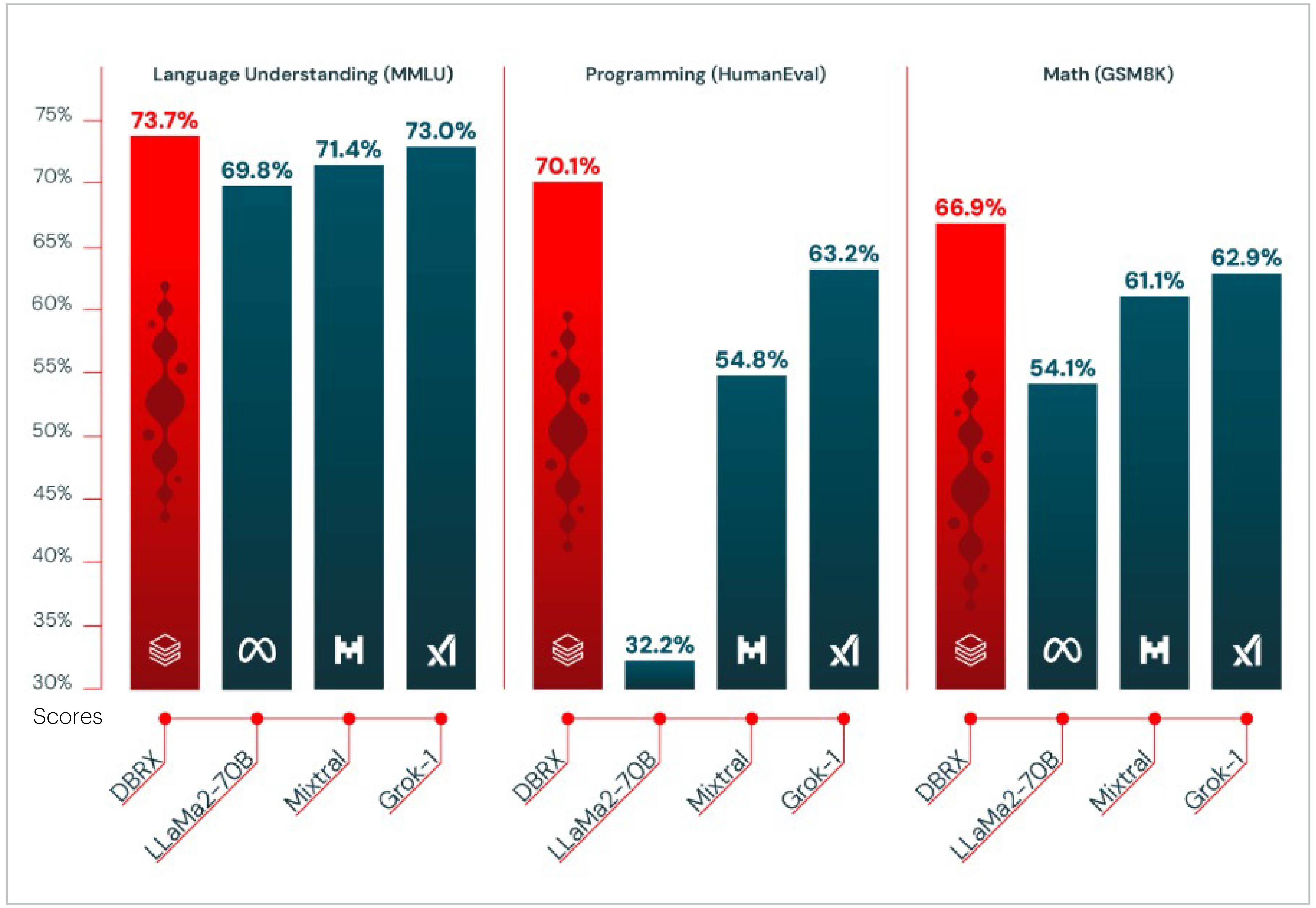


Figure: DBRX outperforms established open source models on language understanding (MMLU), Programming (HumanEval), and Math (GSM8K)

Accelerating the shift to open LLMs

DBRX is only available for enterprises whose data platform is hosted on Databricks, which is a major drawback for data platforms hosted elsewhere. DBRX seems to be a good competitor to the existing closed-source models, at the least because it gives every enterprise the ability to control its data and its destiny in the emerging world of GenAI.

DBRX sets a new standard for open-source AI models, offering customizable and transparent GenAI solutions for enterprises. There is growing interest among AI industry leaders in open-source adoption as fine-tuned models approach closed-source performance levels.

Databricks expects DBRX to accelerate the shift from closed to open-source models.

Author:

Atul Kumar
Solution Architect - Data & AI

zensar
An RPG Company