

INTRODUCTION TO DBRX AND IMAGE AI



Margaret Qian, Hagay Lupesko
June 13 2024

D B R X

shutterstock™

ImageAI

State-of-the-art, efficient,
open language model

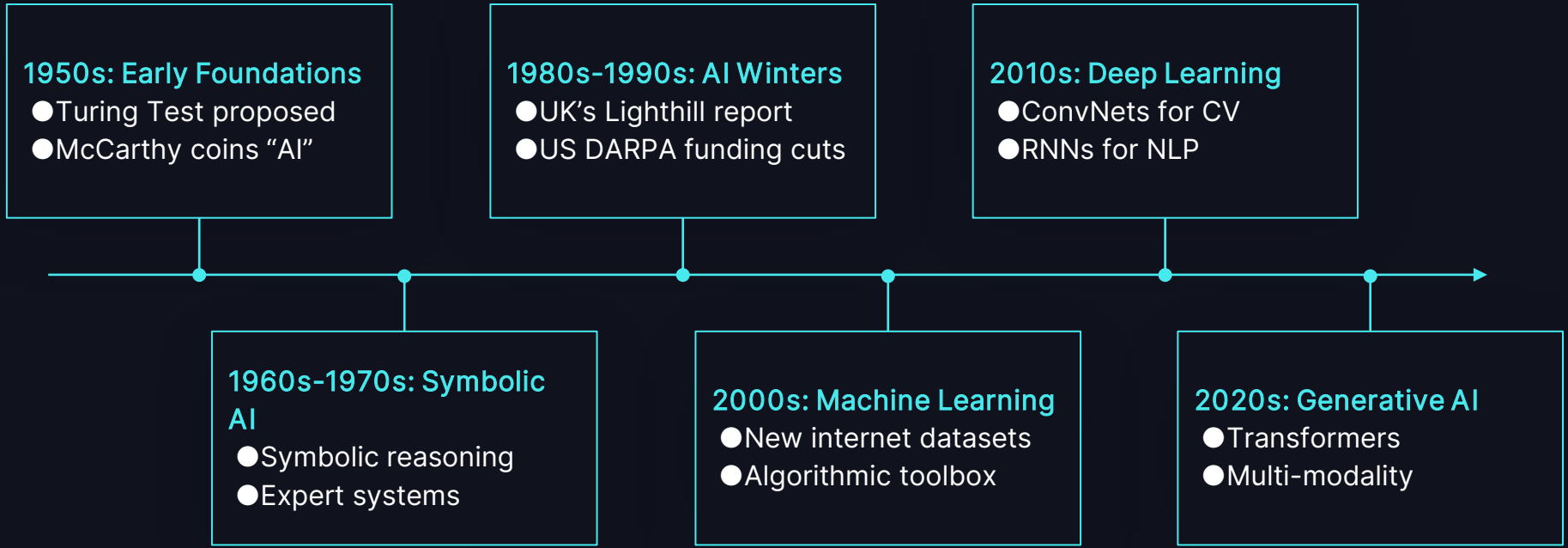
Text-to-image diffusion model,
efficient, commercially safe

Built for the enterprise by MosaicAI Research
Trained & served on Databricks MosaicAI

INTRODUCTION TO GENERATIVE AI

AI JOURNEY

From Symbolic AI to today's Generative AI



APPLICATIONS OF GENERATIVE AI

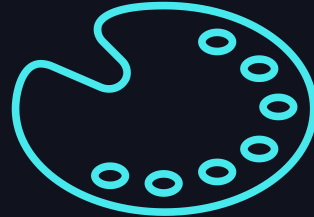
From automation to autonomous agents



E-commerce



Finance



Creative



Engineering

GENERATIVE AI ADOPTION

91%

of organizations are experimenting with GenAI¹

75%

of CEOs say companies using GenAI will have a competitive advantage²

40%

increase in performance of employees using GenAI³

1. Laying the foundation for data and AI-led growth, [MIT Technology Review](#)

2. CEO decision-making in the age of AI, [IBM Institute for Business Value](#)

3. How generative AI can boost highly skilled workers' productivity, [MIT Management Sloan School](#)



INTRODUCING IMAGE AI

INTRODUCING IMAGEAI

Shutterstock ImageAI, Powered by Databricks

Quality

- Trained on 550M+ photos
- Generates high-res, photorealistic images

Safe

- Carefully curated for corporate use
- Commercially safe, royalty free

Governance

- Governed through Databricks



LEVERAGING IMAGE AI

Shutterstock ImageAI is available on Databricks Mosaic AI

Available Now

- Available on Foundation Model API
- \$0.06 per image
- Available on AI Playground
- OpenAI SDK compatible
- Highly performant: ~3 sec/image
- Image sizes:
 - Square 1024x1024
 - Landscape 768x1280
 - Portrait 1280x768

Coming Soon

- Customize your own ImageAI model via fine-tuning or pre-training
- Deploy for high throughput inference with provisioned throughput

LEVERAGING IMAGE AI

Query ImageAI through the Foundation Model API

PYTHON

```
client = OpenAI(
    api_key=DATABRICKS_TOKEN,
    base_url='https://<workspace_id>.databricks.com/serving-endpoints',
)

response = client.images.generate(
    prompt="A cozy corner with a warm fireplace and a comfy chair",
    model="databricks-shutterstock-imageai",
    extra_body={
        "negative_prompt": "vector, illustration",
        "seed": 57,
    }
)
```

DEMO



IMAGE AI: ARCHITECTURE

ImageAI is similar to the Stable Diffusion architecture

“A cozy corner with
a warm fireplace and
a comfy chair”

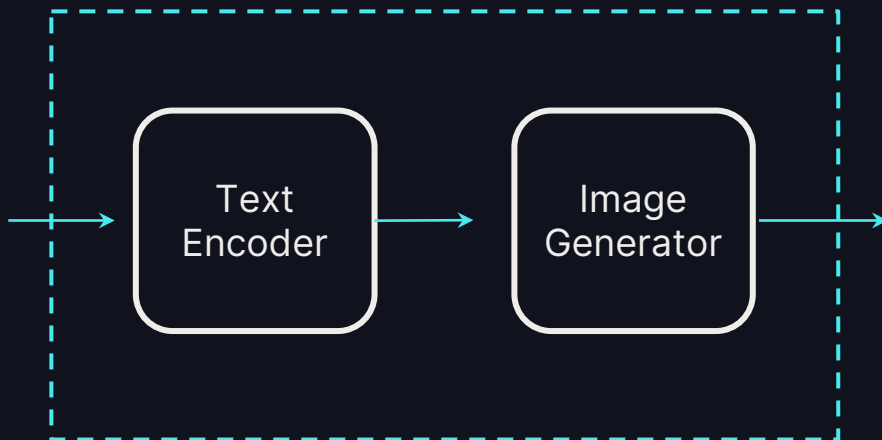
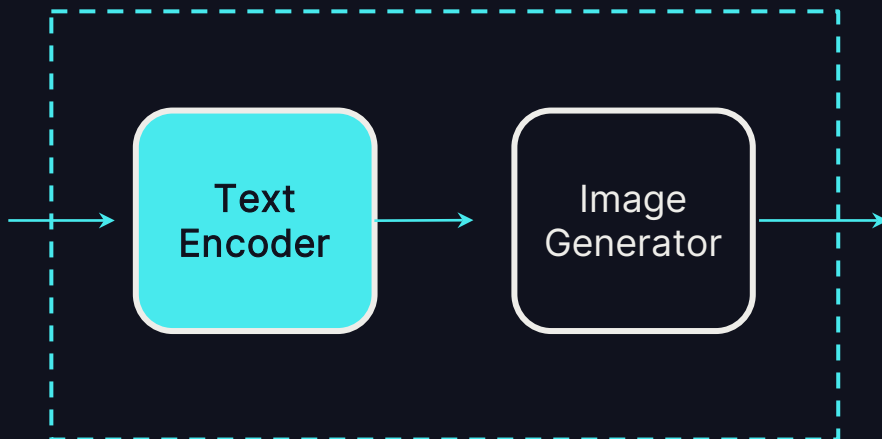


IMAGE AI: ARCHITECTURE

ImageAI is similar to the Stable Diffusion architecture

“A cozy corner with
a warm fireplace and
a comfy chair”



IMAGEAI: ARCHITECTURE

ImageAI is similar to the Stable Diffusion architecture

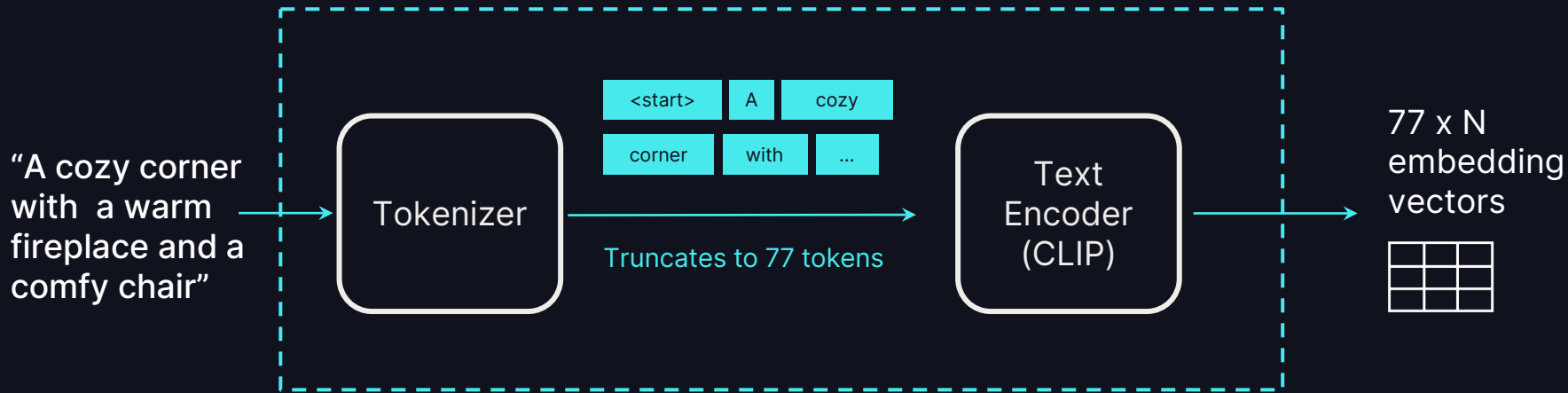


IMAGE AI: ARCHITECTURE

ImageAI is similar to the Stable Diffusion architecture

“A cozy corner with
a warm fireplace and
a comfy chair”

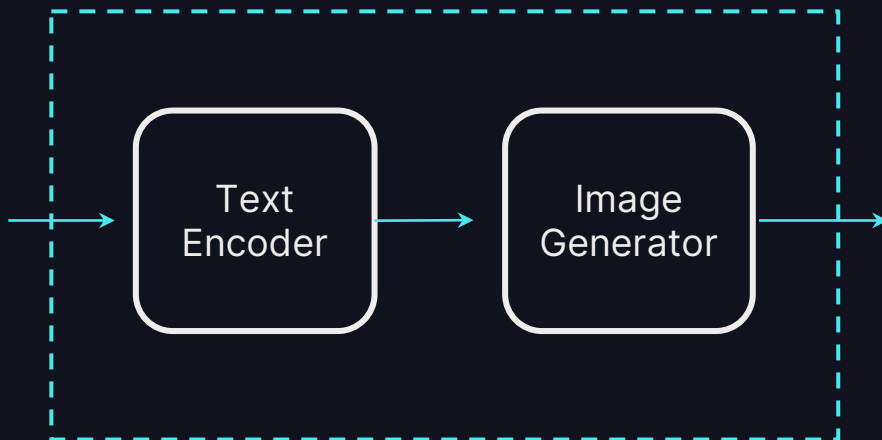


IMAGE AI: ARCHITECTURE

ImageAI is similar to the Stable Diffusion architecture

“A cozy corner with
a warm fireplace and
a comfy chair”

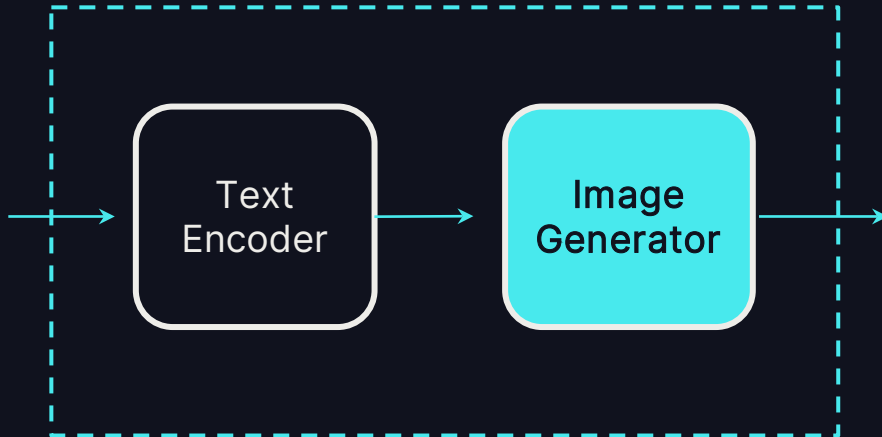
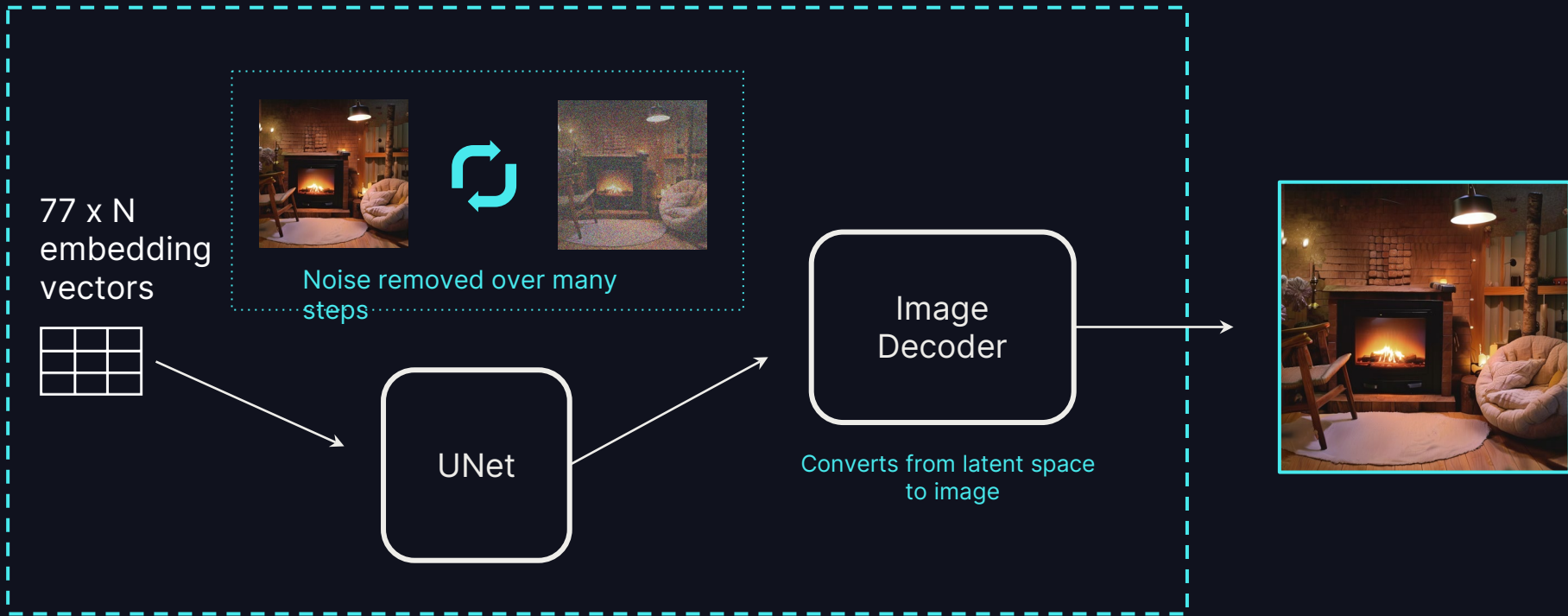


IMAGE AI: ARCHITECTURE

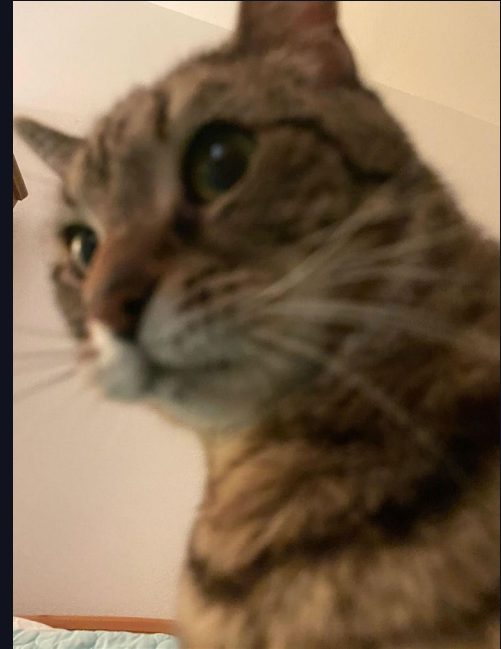
ImageAI is similar to the Stable Diffusion architecture



SCENARIO: MARKETING BROCHURE

Prompt engineering tips and tricks

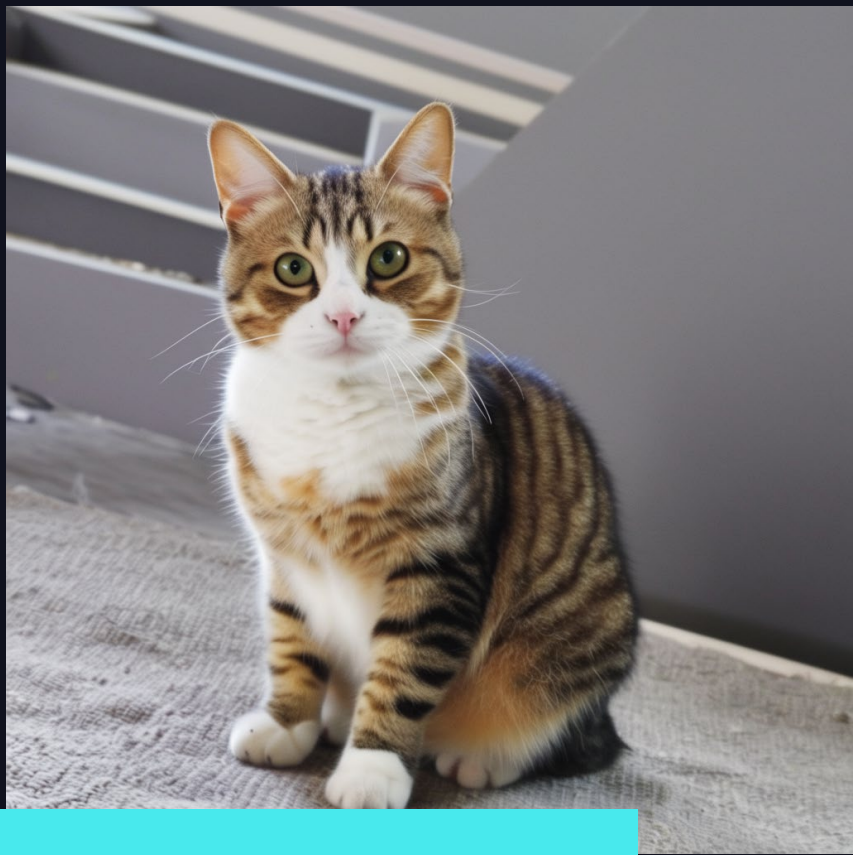
- I'm trying to create a brochure for an animal adoption center
- Hard to get professional quality photos of animals
- Let's try to generate a high quality, photo-realistic hero image!



EXAMPLE: A CUTE CAT

Prompt engineering tips and tricks

- Anatomically correct, but composition could be improved
- Background is messy
- Can we make this better with prompting?



Prompt: "a cute cat"

EXAMPLE: A CUTE CAT

Prompt engineering tips and tricks

- Tip: Add photorealistic modifiers to prompt
 - Quality: “portrait photograph”, “prime lens”
 - Lighting: “studio lighting”
 - Composition: “on a white background”
- Result: cleaner background, consistent lighting, less blurry
- Can we get more creative?



Prompt: “A high quality portrait photograph of a cute sitting cat, on a white background, prime lens, studio lighting”

EXAMPLE: A CUTE CAT

Prompt engineering tips and tricks

- Tip: Add keywords to negative prompt to specify things you don't want in the photo
 - Blurry, ugly, deformed
 - Amateur
 - Vector, illustration
- We have our unique hero image!



Prompt: "A cute cat jumping in a colorful cyberpunk setting. Blurry neon signs in the background. Professional quality good photo"

Negative prompt: "blurry, ugly, deformed, amateur, vector, illustration"

OTHER EXAMPLES



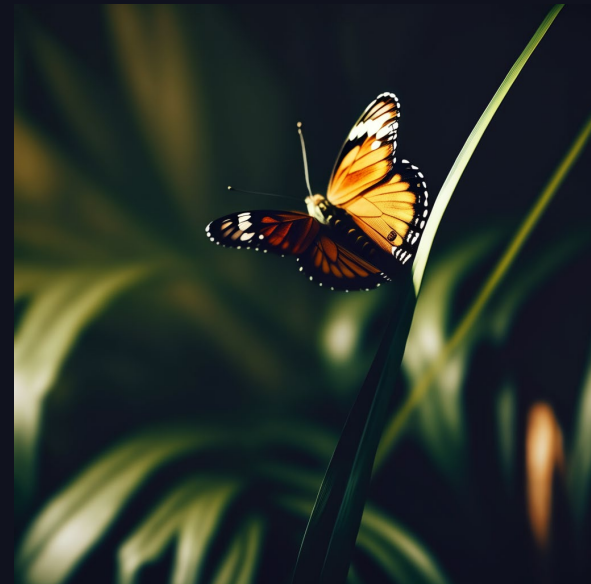
Prompt: "Back of cinematic young man in winter coat walking outside in futuristic city on a moody, foggy, winters night. tilt shift"

Negative prompt: "illustration, typography, deformed, bad anatomy, mutation, bad proportions, disfigured, blurry"



Prompt: "A photo of a woman working on her laptop in a coffee shop, a blurred crowd in the background, natural skin, close-up portrait, side view"

Negative prompt: "Deformed, blurry, blur, out of focus, bad anatomy, bad eyes, crossed eyes, disfigured, extra limb, ugly, missing limb, blurry, floating limbs, disconnected limbs, malformed hands, cropped"



Prompt: "butterfly on a leaf, dark, cinematic, tilt shift lens, close up"

Negative prompt: "illustration, typography"

INTRODUCING DBRX

INTRODUCING DBRX

DBRX Base

Functions like a smart autocomplete.
Great as a baseline for fine-tuning on your
data

DBRX Instruct

Designed to follow instructions.
Fine-tuned from DBRX for instruction
following

DBRX

The model **your data**
has been waiting for

INTRODUCING DBRX

Open

Open source architecture and weights

Control

Full model ownership and customizability

Quality

Performant across benchmarks

Performance

Optimized for efficient serving at scale

DBRX

The model **your data**
has been waiting for

DBRX - OPEN

DBRX is open -source from architecture to weights

- Freely available on GitHub and HuggingFace for research and commercial use
- Leverage DBRX architecture to train your own custom LLM
- You own the model and the weights



DBRX - CONTROL

Customize DBRX on your custom data

~~Fine-tune DBRX in order to...~~

- ... Specialize on a specific task
- ... Control model behavior
- ... Reduce inference cost and latency
- ... Own model trained on your data

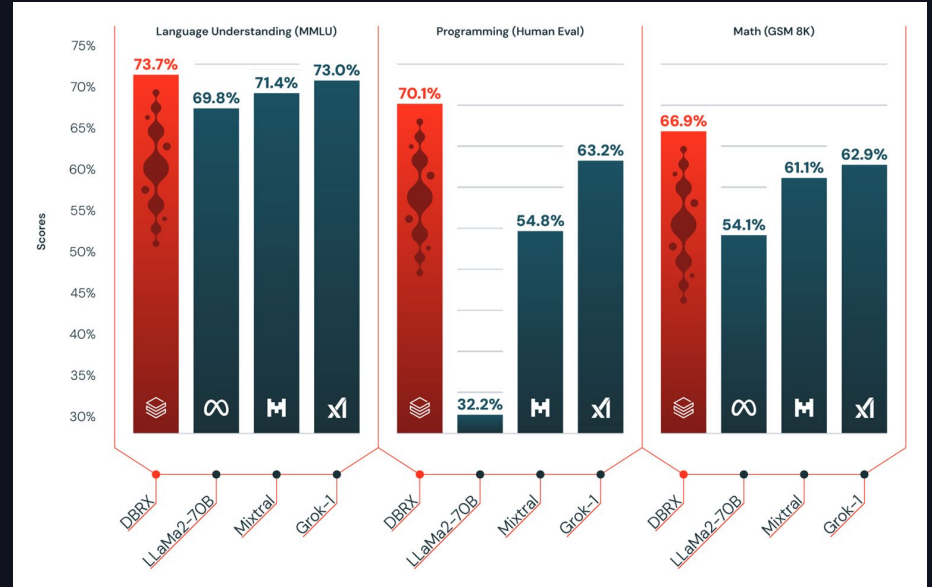
~~Pre-train DBRX in order to...~~

- ... Fully tailor to a new language or a unique domain
- ... Fully control the training data
- ... Train a highly cost-efficient custom model

DBRX - QUALITY

DBRX is one of the highest quality open models available today

- Pre-trained on 12T tokens of carefully curated text and code
- Context length of up to 32K
- Competitive performance across benchmarks

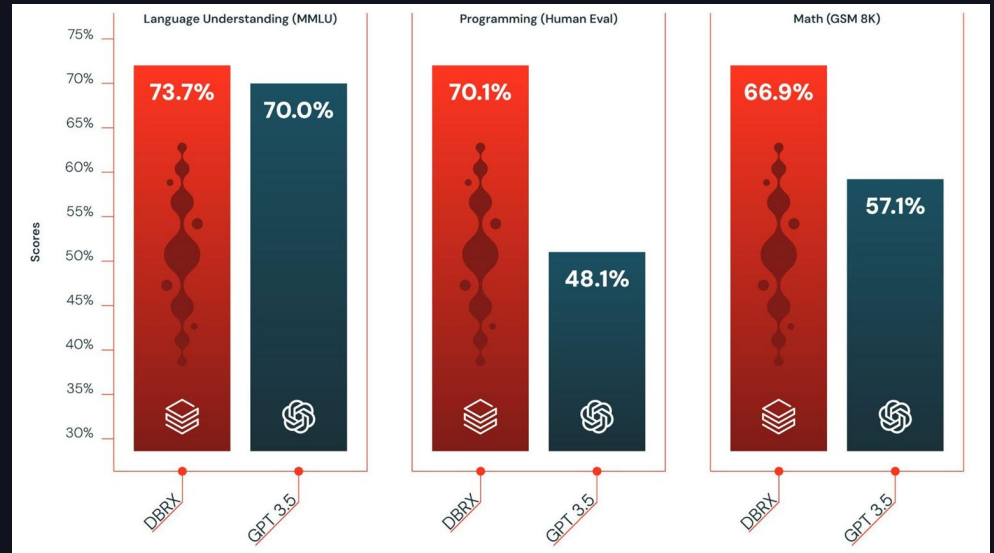


<https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm>

DBRX - QUALITY

DBRX is one of the highest quality open models available today


- Pre-trained on 12T tokens of carefully curated text and code
- Context length of up to 32K
- Competitive performance across benchmarks



<https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm>

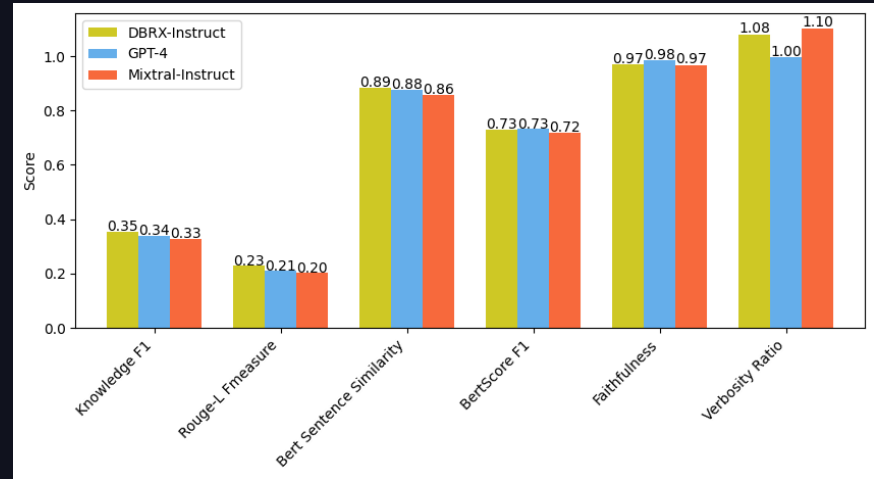
DBRX - QUALITY

DBRX is one of the highest quality open models available today

 **Julia Neagu**
@JuliaANeagu

@DbrxMosaicAI DBRX outperforms @OpenAI GPT-4 on realistic, domain-specific benchmark datasets. For example, on a customer support summarization use-case 🙄🙄🙄

Still neck and neck but it shows that open models can be the no-brainer choice for actual enterprise applications.

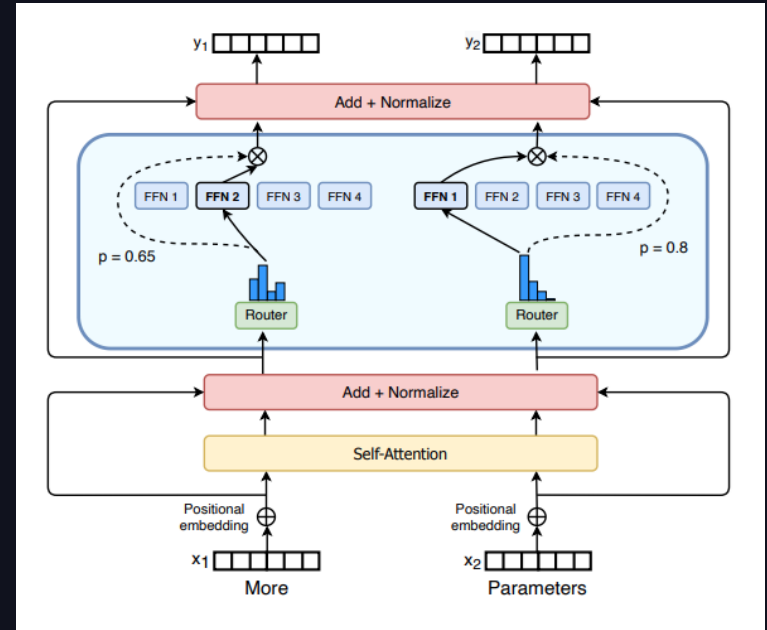


<https://x.com/JuliaANeagu/status/1773831864660189294>

DBRX - PERFORMANCE

DBRX was designed for optimized inference performance and cost


- Fine-grained MoE architecture
- 4 active experts out of 16
- 36B active parameters out of 132B
- 2x faster than dense models
- ~150 tok/sec with 8-bit inference



MoE layer from the Switch Transformers paper (<https://arxiv.org/abs/2101.03961>)

DBRX - PERFORMANCE

DBRX offers a great tradeoff between cost and performance

 **virat** ✓
@virattt

Friday is LLM battle day.

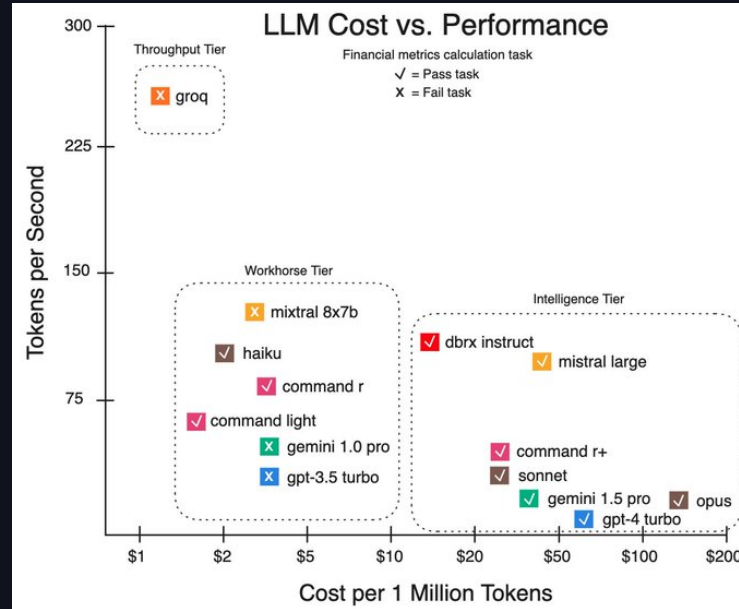
I added DBRX to the financial metrics challenge.

Overall, very impressed with DBRX.

Main takeaways:

- correctly calculated metrics
- ranked top 4 fastest models
- competitive pricing

DBRX was +50% cheaper and +100% faster than models in its tier.



<https://x.com/virat/status/1778828787951546382>

LEVERAGING DBRX

LEVERAGING DBRX

DBRX is available on Databricks Mosaic AI

Serving

Foundation Model API

\$0.75/M input tokens

\$2.25/M output tokens

Provisioned Throughput:

\$12/Hour for 600 tokens/sec

Fine-tuning

Foundation Model Training

\$0.65/DBU

Cost varies with data scale

10M tokens ~\$193

500M tokens ~\$9300

Pre-training

Multi Cloud Training

\$13.27/hour H100 GPU

Cost varies with model and data scale

Starting at ~\$40K for DBRX-9B

SERVING DBRX

Mosaic AI Foundation Model API offers DBRX inference endpoints, with Provisioned Throughput and Pay-per-token pricing

PYTHON

```
client = OpenAI(
    api_key=DATABRICKS_TOKEN,
    base_url='https://<workspace_id>.databricks.com/serving-endpoints',
)

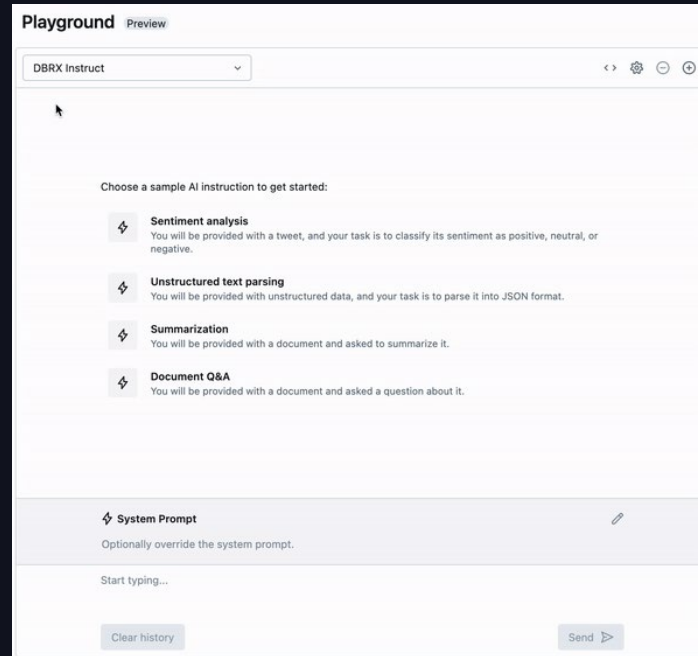
chat_completion = client.chat.completions.create(
    messages=[
        {
            "role": "user",
            "content": "Give me the character profile of a knight in JSON.",
        }
    ],
    model="databricks-dbrx-instruct"
)
```

DEMO



SERVING DBRX

Mosaic AI Foundation Model API offers DBRX inference experimentation through the AI Playground



FINE-TUNING DBRX

Mosaic AI Foundation Model Training offers various model tuning through a simple SDK and GUI.

PYTHON

```
from databricks.model_training import foundation_model as fm

# UC Volume with JSONL formatted fine-tuning data
train_data_path = 'dbfs:/Volumes/main/mydirectory/ift/train.jsonl'

run = fm.create(
    model='databricks/dbrx-instruct',
    train_data_path=train_data_path,
    register_to='main.mydirectory',
)
```

Pick your tuning task:

- Supervised Fine Tuning (SFT)
- Continued Pre Training (CPT)
- Chat Completion

DEMO



FINE-TUNING DBRX

Mosaic AI Foundation Model Training offers various model tuning

The screenshot displays the 'Experiments' page in the Databricks interface. At the top, there is a search bar and a 'Compare (0)' button. Below the header, the 'Create experiment' section offers three options:

- AutoML:** Input your dataset and create a regression or classification model using AutoML. Includes a 'Create AutoML experiment' button.
- Traditional:** Track traditional ML and deep learning models and find the best model. Includes a 'Create Traditional Experiment' button.
- Foundation Model Training:** Train a foundation model to work on your dataset. Includes a 'Create Foundation Model Experiment' button.

Below the creation options is a 'Filter experiments' section with a search icon and a checkbox for 'Only my experiments'. The main area contains a table with the following columns: Name, Created by, Last modified (with a sort icon), Location, and Description. The table is currently empty. At the bottom right of the table area, there are navigation controls: '< Previous', 'Next >', and '25 / page'.

PRE-TRAINING DBRX

Mosaic AI Multi Cloud Training offers compute, orchestration, full software stack to pre-train your own DBRX from scratch

SHELL

```
> cat config.yaml
model: databricks/dbrx-9b
train_data: s3://mybucket/exampledataset
save_folder: s3://mybucket/saved_model
experiment_tracker:
  mlflow:
    experiment_path: /Users/example@domain.com/my_experiment
    model_registry_path: catalog.schema.model_name
compute:
  gpus: 128
```

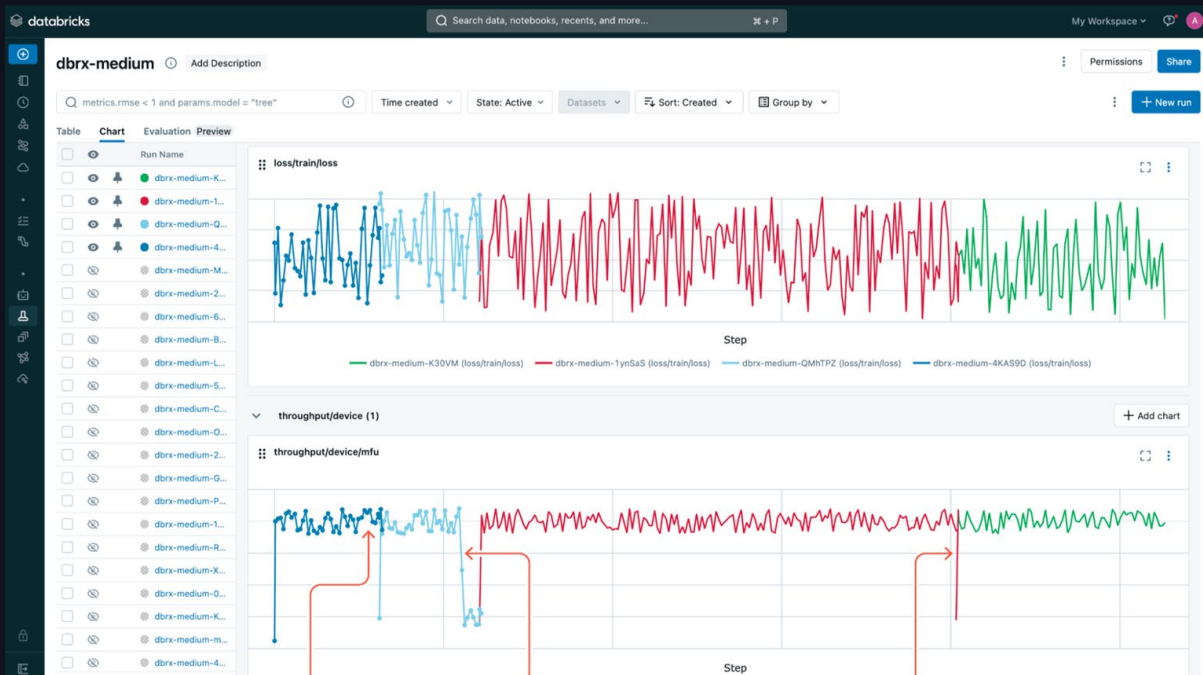
```
> mcli train -f config.yaml
✓ Run pretrain-dbrx-9b-aq7pd2 submitted.
```

To see the training run's progress, use:
`mcli describe run pretrain-dbrx-9b-aq7pd2`

Choose your model size:

- DBRX-9B (2.6B active)
- DBRX-19B (5.2B active)
- DBRX-36B (10B active)
- DBRX-73B (20B active)

PRE-TRAINING DBRX



1. Identified bad node with xid error
2. Cordoned node and resumed

1. Detected abnormal MFU
2. Paused run and swept cluster
3. Cordoned two bad nodes and resumed

1. Failure (transient network error)
2. Resumed



LET'S RECAP

D B R X

shutterstock™

ImageAI

State-of-the-art, efficient,
open language model

Text-to-image diffusion model,
efficient, commercially safe

Available for builders on Mosaic AI

THANK YOU

