Microsoft

Tech Talk Fridays

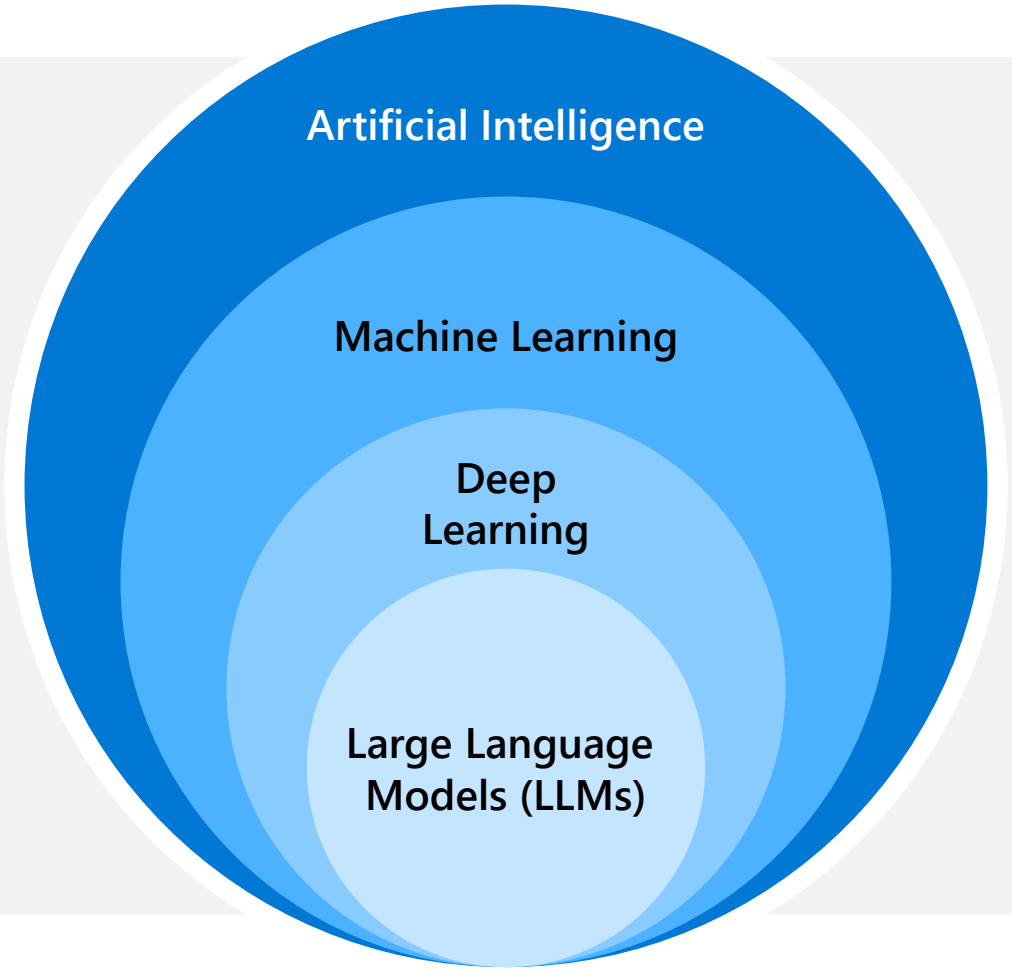
A Look at Phi3 and the Growing S/LLM Landscape in Azure

Florian Follonier
CSAs for Data & AI

Agenda

- Our journey so far
- Advent of SLMs
- Phi model series and Phi-3 Family
- Azure AI Hugging Face

Our journey so far...



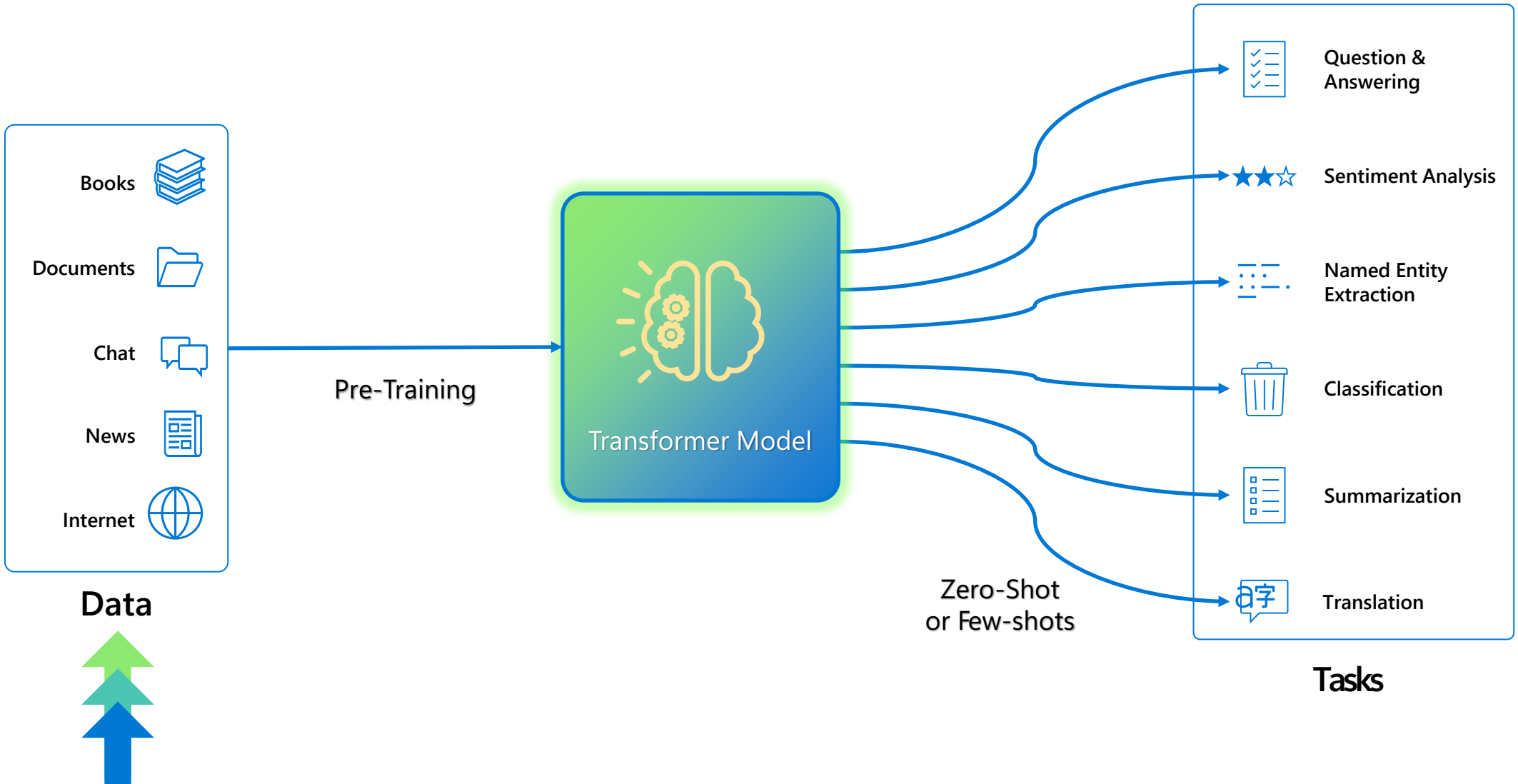
1956 **Artificial Intelligence** The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997 **Machine Learning** Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012 **Deep Learning** A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

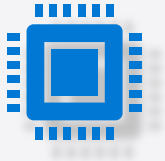
2022 **Large Language Models** For the first time we are able capture and **model knowledge**. Further, we observe **emergent behaviors** as we scale up.

Large Language Modelling



Our learnings from building copilots

1 LLMs area great for building copilots, and everyone wants to build one… **HOWEVER!**



Resources

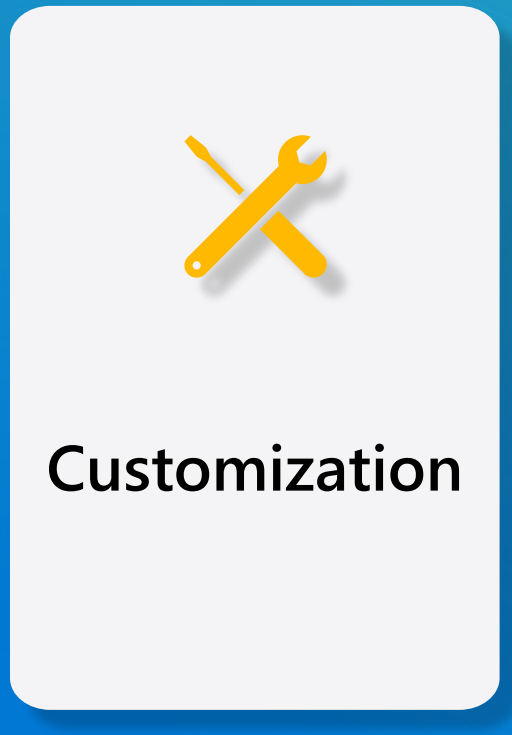
High Latency

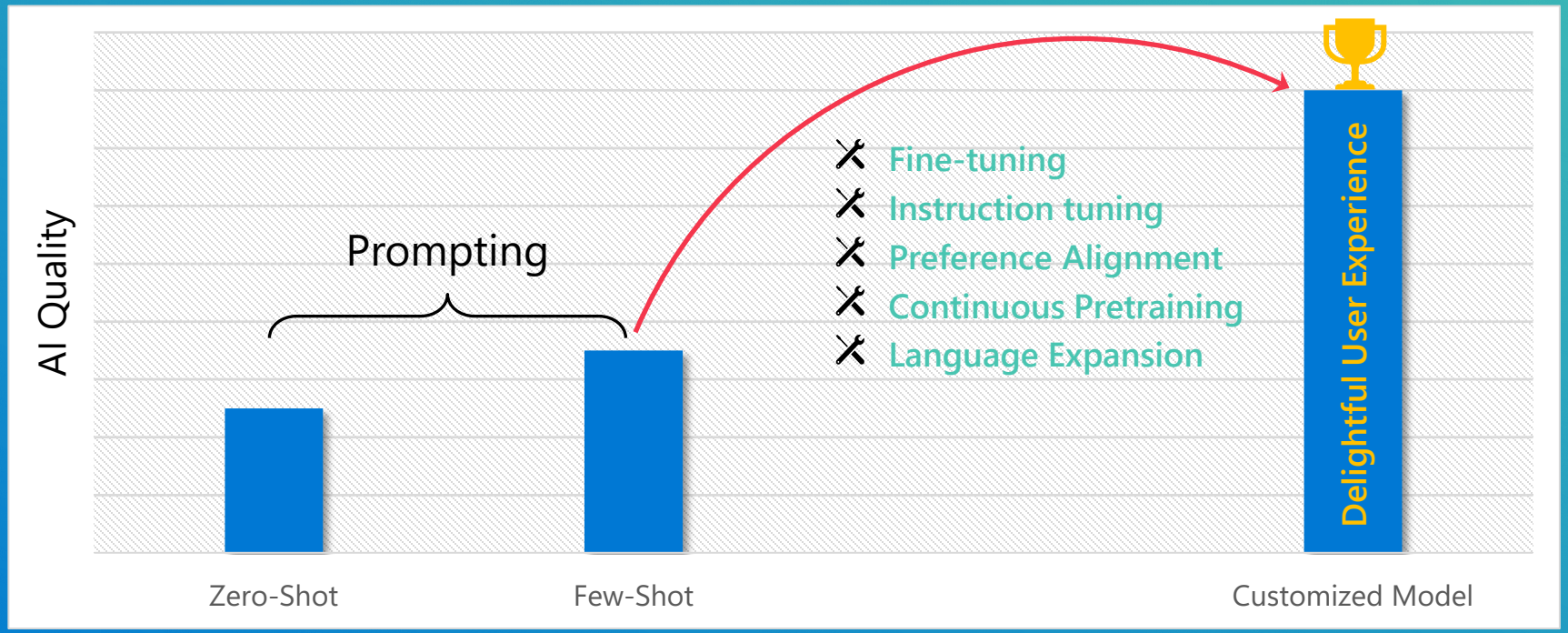
High Cost

Deployment

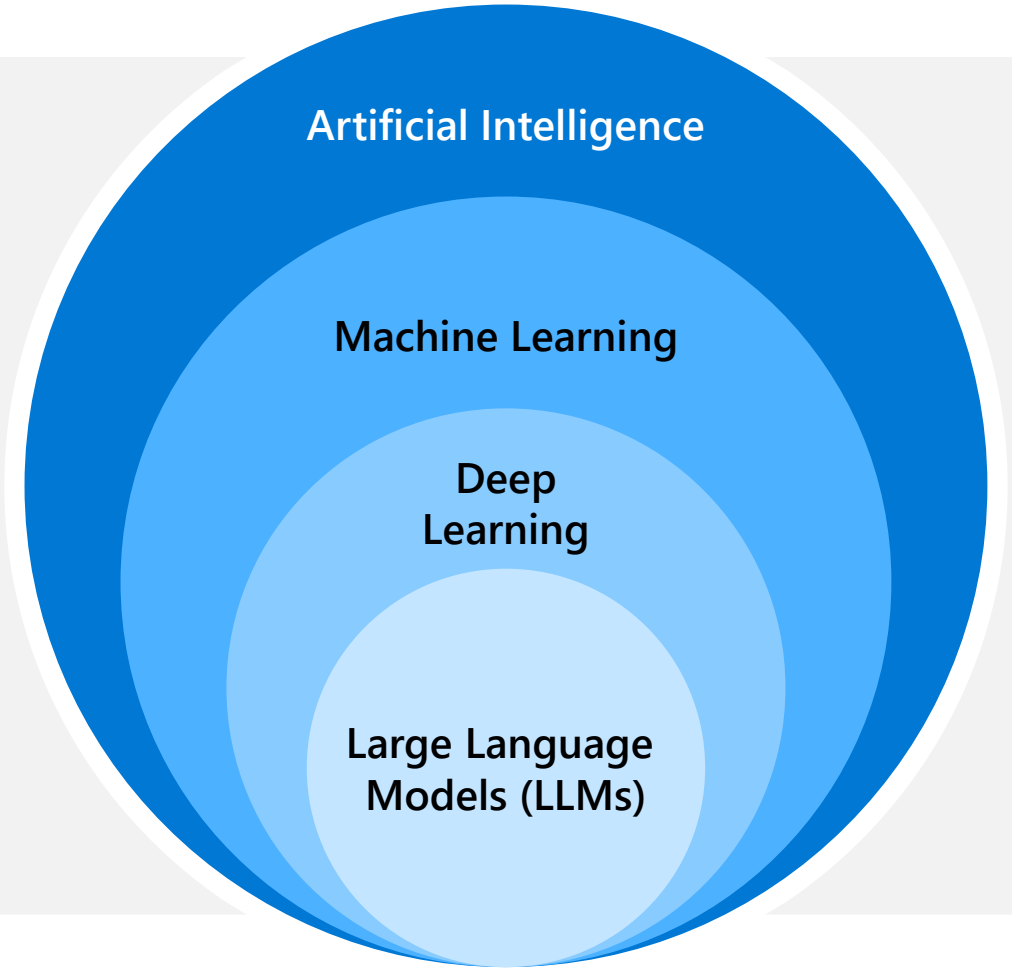
Our learnings from building copilots

2 To deliver an AI feature at a user expected high-quality bar needs **Customization**





Our journey so far...



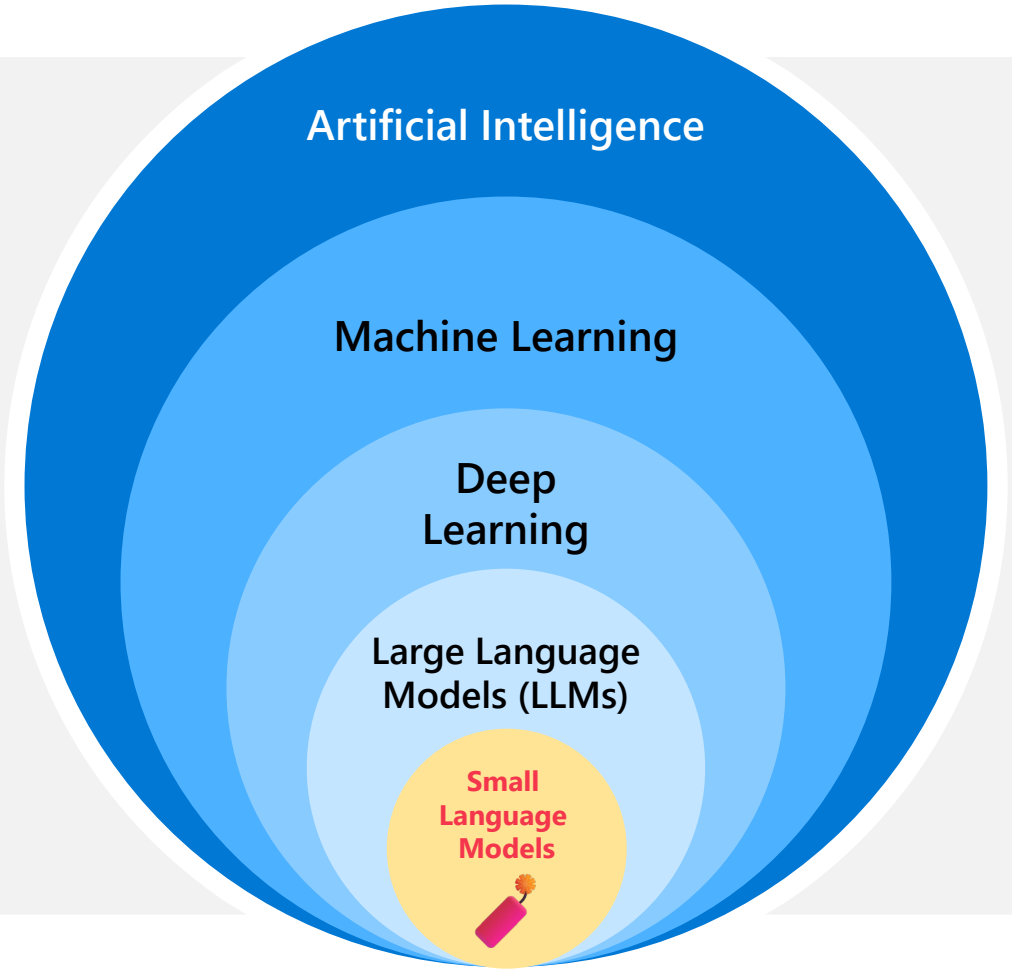
1956 **Artificial Intelligence** The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997 **Machine Learning** Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012 **Deep Learning** A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

2022 **Large Language Models** For the first time we are able capture and **model knowledge**. Further, we observe **emergent behaviors** as we scale up.

Our journey so far...



1956 **Artificial Intelligence** The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997 **Machine Learning** Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012 **Deep Learning** A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

2022 **Large Language Models** For the first time we are able capture and **model knowledge**. Further, we observe **emergent behaviors** as we scale up.

2023 **Advent of Phi SLMs,** Tiny but mighty language models that challenge status quo!

SLMs are natural progressions to LLMs

Harward Mark-1

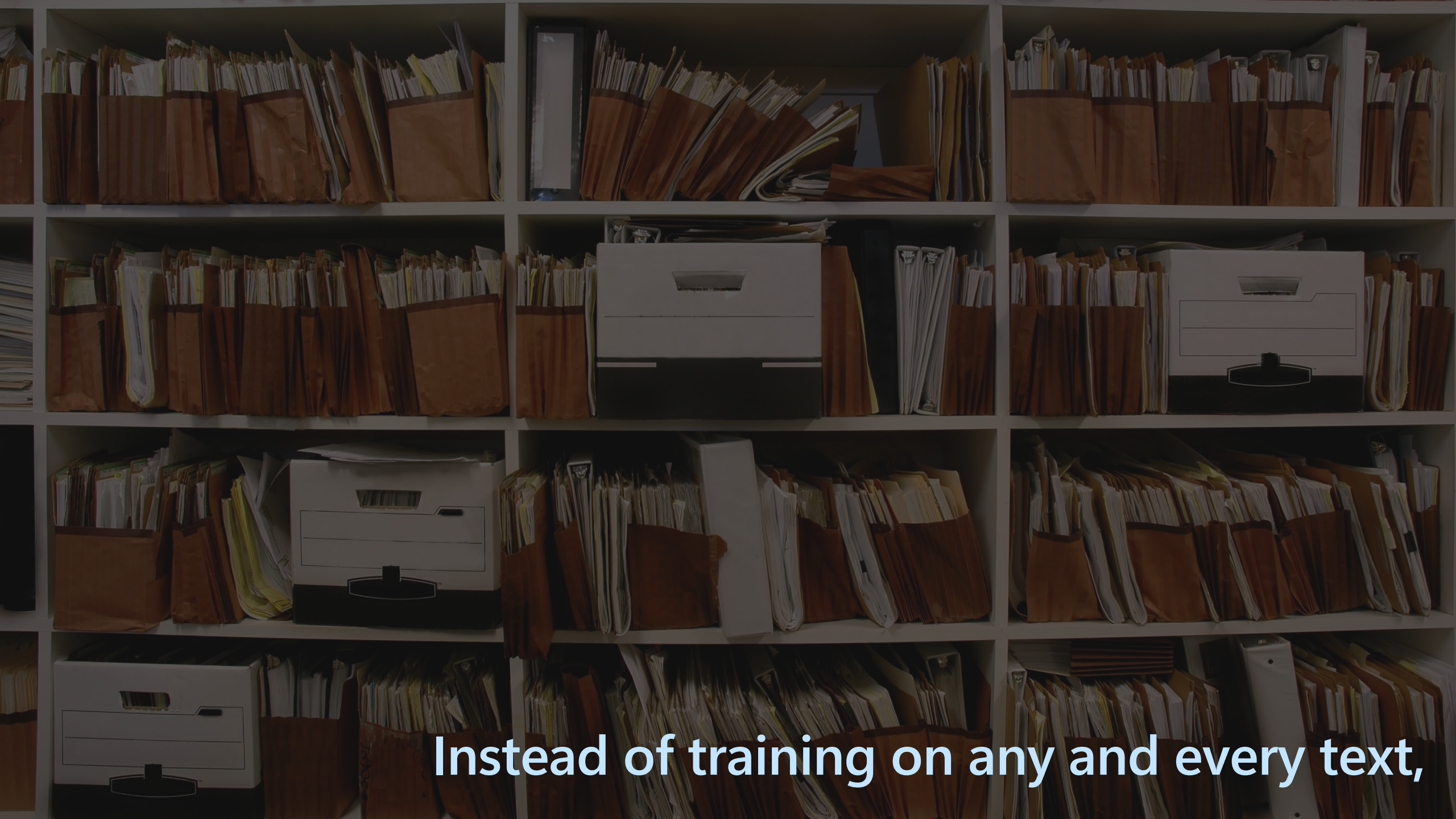
Conceived by Harvard physics professor Howard Aiken, and designed and built by IBM, the Harvard Mark 1 is a room-sized, relay-based calculator. The machine had a fifty-foot long camshaft running the length of machine that synchronized the machine's thousands of component parts and used 3,500 relays. The Mark 1 produced mathematical tables but was soon superseded by electronic stored-program computers.

Source: 1944 | Timeline of Computer History | Computer History Museum

Apple Watch 9 (2023)

has a new 4-core Neural Engine that can process machine learning tasks up to twice as fast, when compared with Apple Watch Series 8. The power efficiency of the S9 SiP allows Apple Watch Series 9 to maintain all-day 18-hour battery life.

Source: 1944 | Timeline of Computer History | Computer History Museum



Instead of training on any and every text,

What if we use highly educational Textbook like content?

Just like we were educated!

Customization
(achieved via CPT, SFT, DPO)

Journalist, Engineer,
Doctor, Scientist,
Product Manager, etc.

Small Language Models

Foundation
(achieved via Pretraining)

Analytical & Application

Literature, Math, Science

Language & Grammar

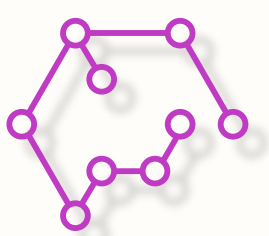
Reading, Writing & Speaking

Alphabets & Numbers

Image by macrovector on Freepik

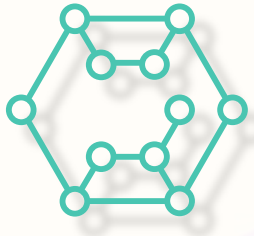
Phi-3

Groundbreaking performance for size, with frictionless availability



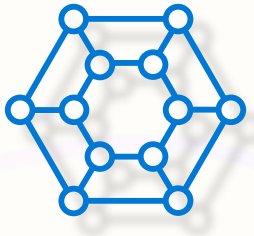
Phi-3-mini
(3.8B)

Available today



Phi-3-small
(7B)

Coming soon



Phi-3-medium
(14B)

Coming soon

Instruction Tuned

RAI Safety Aligned

Available on

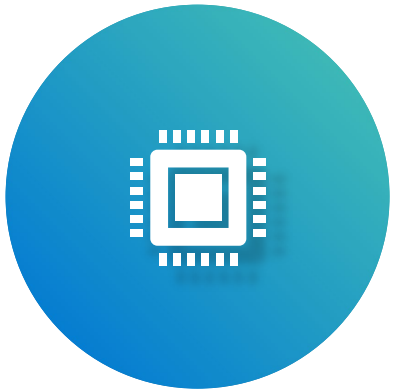
Azure AI
Model Catalog

Hugging Face

Ollama

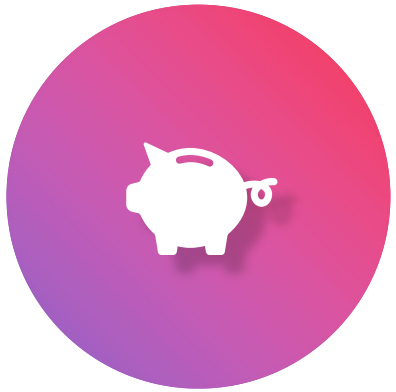
ONNX

Benefits of *phi-3-mini* (3.8B)



Low compute footprint and can run on older GPUs

Ultra Low Latency thanks to its small size



Easy on your wallet, and hence business viable

Can be deployed on-prem or on-edge devices

Easier & Affordable to customize (fine-tuning)

Stop | phi-3-mini | Save

If Paris were not the capital of France, what would be some good alternatives?

0.1 sec | 0.0 tokens/sec | 0 tokens

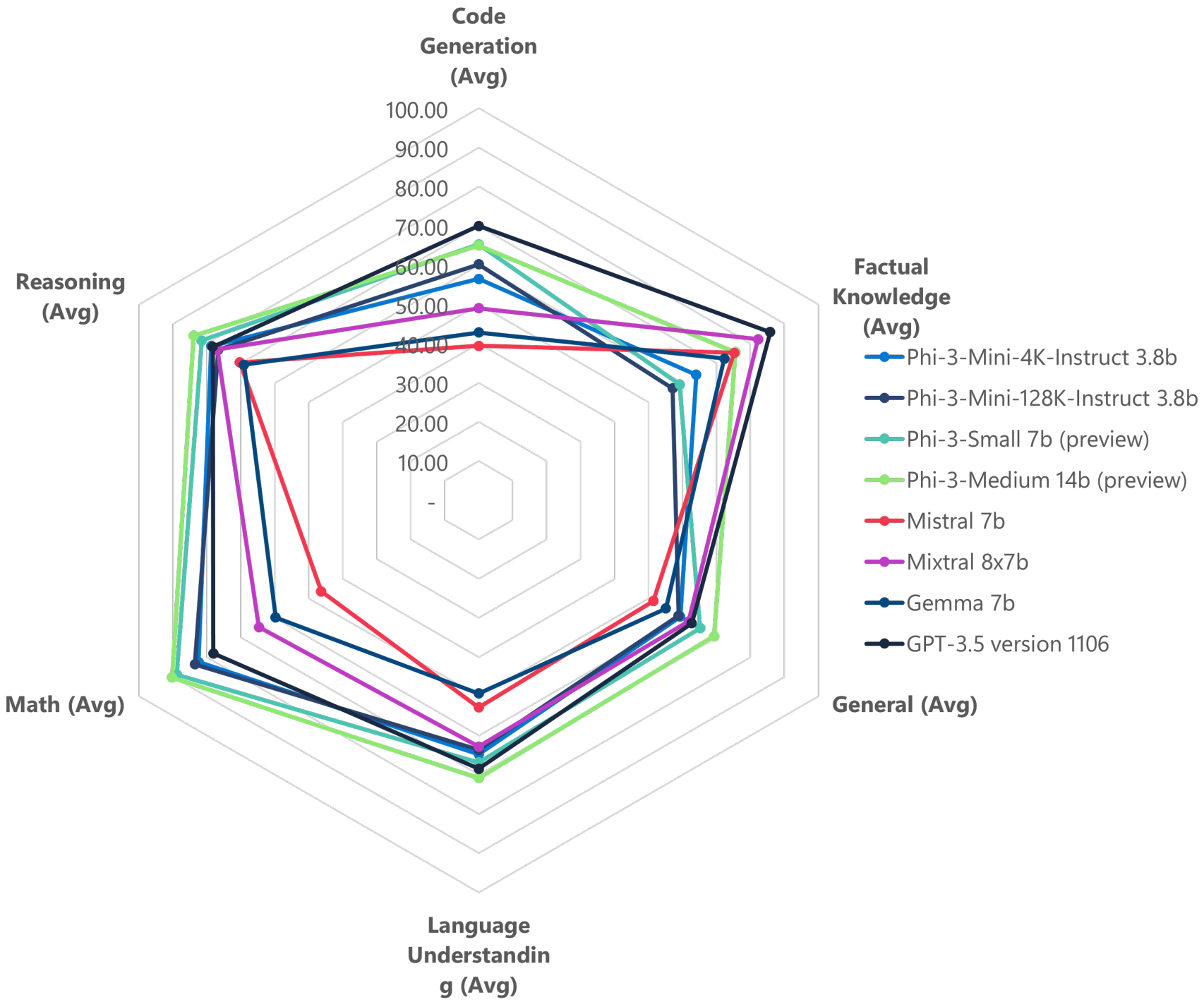
Stop | GPT 4 Paid (1106 | Save

If Paris were not the capital of France, what would be some good alternatives?

0.1 sec | 0.0 tokens/sec | 0 tokens

Benchmark Numbers

Phi3 excels in particular in Reasoning, Math and Language Understanding Tasks



Benchmark Numbers

Category Dataset	Phi-3-Mini-4K-Instruct 3.8b	Phi-3-Mini-128K-Instruct 3.8b	Phi-3-Small 7b (preview)	Phi-3-Medium 14b (preview)	Mistral 7b	Mixtral 8x7b	Gemma 7b	GPT-3.5 version 1106
Code Generation (Avg)	**56.45**	**60.20**	**65.25**	**65.00**	**39.40**	**49.00**	**42.80**	**70.00**
HumanEval (0-Shot)	59.10	57.90	59.10	55.50	28.00	37.80	34.10	62.20
MBPP (3-Shot)	53.80	62.50	71.40	74.50	50.80	60.20	51.50	77.80
Factual Knowledge (Avg)	**64.00**	**57.10**	**59.10**	**75.60**	**75.20**	**82.20**	**72.30**	**85.80**
TriviaQA (5-Shot)	64.00	57.10	59.10	75.60	75.20	82.20	72.30	85.80
General (Avg)	**59.33**	**58.83**	**65.17**	**69.30**	**51.37**	**61.80**	**55.10**	**62.70**
AGIEval (0-Shot)	37.50	36.90	45.00	48.40	35.10	45.20	42.10	48.40
BigBench-Hard (0-Shot)	71.70	71.50	74.90	81.30	57.30	69.70	59.60	68.30
MMLU (5-Shot)	68.80	68.10	75.60	78.20	61.70	70.50	63.60	71.40
Language Understanding (Avg)	**64.75**	**63.65**	**66.85**	**70.85**	**52.80**	**62.80**	**49.25**	**68.45**
ANLI (7-Shot)	52.80	52.80	55.00	**58.70**	47.10	55.20	48.70	58.10
HellaSwag (5-Shot)	76.70	74.50	78.70	83.00	58.50	70.40	49.80	78.80
Math (Avg)	**82.50**	**83.60**	**88.90**	**90.30**	**46.40**	**64.70**	**59.80**	**78.10**
GSM-8K (0-Shot; CoT)	82.50	83.60	88.90	90.30	46.40	64.70	59.80	78.10
Reasoning (Avg)	**78.71**	**76.72**	**81.56**	**83.96**	**70.33**	**76.96**	**69.14**	**78.32**
Arc-C (10-Shot)	84.90	84.00	90.70	91.00	78.60	87.30	78.30	87.40
Arc-E (10-Shot)	94.60	95.20	97.10	97.80	90.60	95.60	91.40	96.30
BoolQ (0-Shot)	77.60	78.70	82.90	86.60	72.20	76.60	66.00	79.10
CommonSenseQA (10-Shot)	80.20	78.00	80.30	82.60	72.60	78.10	76.20	79.60
MedQA (2-Shot)	70.00	55.30	58.20	69.40	50.00	62.20	49.60	63.40
OpenBookQA (10-Shot)	84.20	80.60	88.40	87.20	79.80	85.80	78.60	86.00
PIQA (5-Shot)	83.20	83.60	87.80	87.70	77.70	86.00	78.10	86.60
SociQA (5-Shot)	76.60	76.10	79.00	80.20	74.60	75.90	65.50	68.30
TruthfulQA (10-Shot)	65.00	63.20	68.70	75.70	53.00	60.10	52.10	67.70
WinoGrande (5-Shot)	70.80	72.50	82.50	81.40	54.20	62.00	55.60	68.80
Average	**71.26**	**70.11**	**74.91**	**78.16**	**61.23**	**69.76**	**61.73**	**74.32**

Where are some use cases where Phi 3 shines?

- Resource constrained environments where local inference may be needed
- Latency bound scenarios where fast response times are critical
- Cost constrained use cases particularly those with simpler tasks of summarization, analysis and logical reasoning.

Customers using Phi-3

Phi-3 applications deliver accurate results at a small size, making it possible to run on phones and devices

ITC Limited

"Our goal with the Krishi Mitra copilot is to improve efficiency while maintaining the accuracy of a large language model. We are excited to partner with Microsoft on using fine-tuned versions of Phi-3 to meet both our goals - efficiency and accuracy!"

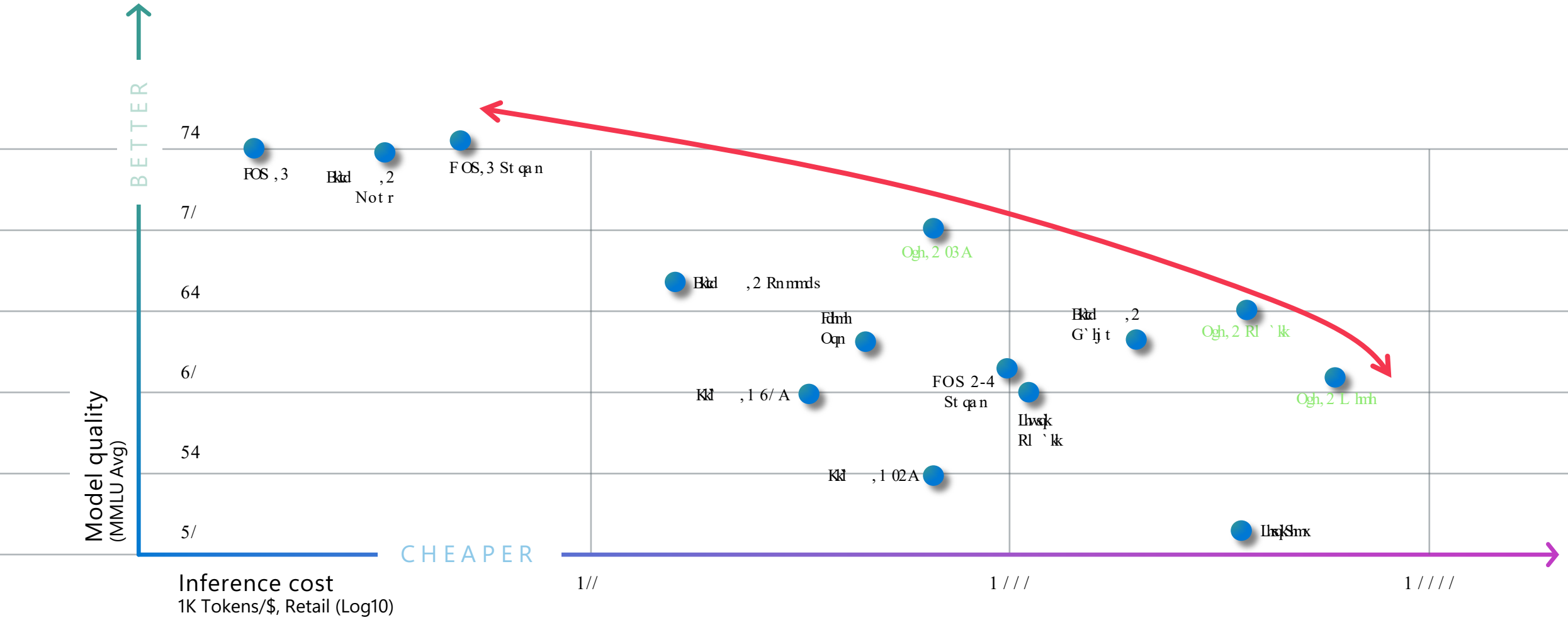
—Saif Naik, Head of Technology, ITCMAARS

Digital Green

"Farmers and frontline government workers in rural communities around the world often incur high data costs and have intermittent connectivity with the Internet. We are excited about our collaboration with Microsoft Research to develop a Phi-3 based copilot to run offline on phones instead of requiring constant connection to the cloud."

—Rikin Gandhi, CEO, Digital Green

Ogh2 * Nodm@H nc dkr gdko ot r g sgd dmudknod



Microsoft

The Growing Model Space

Stronger Azure AI Model Portfolio

Offering the widest collection of frontier and open-source models

Azure OpenAI Service

- GPT-4-Turbo
- GPT-4
- GPT-4V
- Text-embedding-ada-002
- GPT-3.5-Turbo

Meta

- Llama-2-70b/70b-chat*
- Llama-2-13b/13b-chat*
- Llama-2-7b/7b-chat*
- Llama-3*
- CodeLlama

Mistral AI

- Mistral Large*
- Mistral 7b
- Mixtal 7b*8—Mixture of Experts

cohere

- Cohere R*
- Cohere R+*
- Embed v3—Multilingual*
- Embed v3—English*

Microsoft's SLMs

Phi

- Phi-1
- Phi-1.5
- Phi-2
- Phi-3

Orca

- Orca 1
- Orca 2

Hugging Face

- Falcon/TII
- Stable Diffusion/Stability AI
- Dolly/Databricks
- CLIP/OpenAI

NVIDIA

- Nemotron-3-8B-4k
- Nemotron-3-8B-Chat-SFT/RLHF/SteerLM
- Nemotron-3-8B-QA

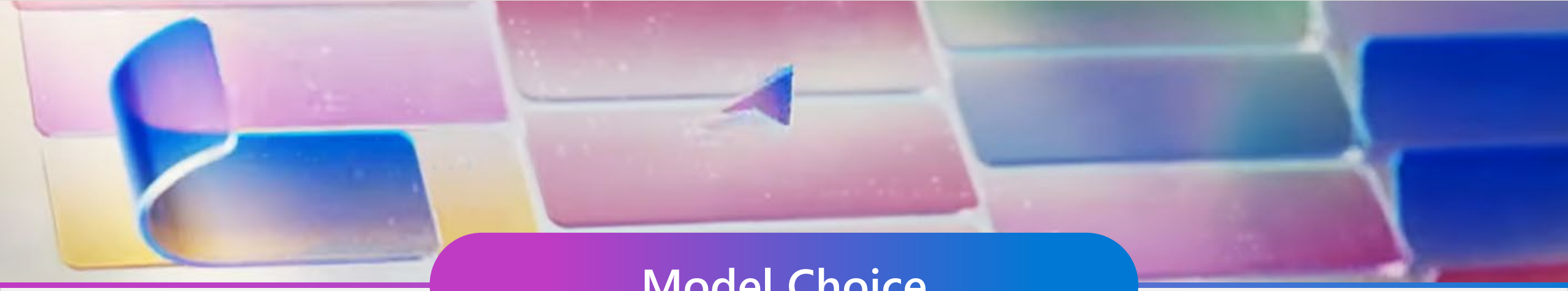
Databricks

- Databricks/dbrx-base
- Databricks/dbrx-instruct

G42

- Jais*

* Available via MaaS



Model Choice

Identify the optimal models for your use case

Choose from the foundation and open-source models in the **Model Catalog**, including OpenAI GPT-4 Turbo with Vision

Leverage **Model as a Service (MaaS)**, including Llama 2, and pay as you go for convenient and cost-effective infrastructure

Experiment and evaluate models with intuitive **benchmarking** based on industry-leading data or your own

Wide model choice & evaluation tools

Content Generation

Empower your users
with AI-generated content based on natural prompt commands

Summarization
Text generation
Image generation

+

Multimodality

Increase user engagement
with media-rich interactions that understand diverse data types

Image to text
Text to image
Video to text

+

Customization

Strengthen model performance
with grounding in your own data and simplified tailoring techniques

RAG
Prompt engineering
Fine-tuning

Microsoft Model Family
Azure OpenAI Model Family
Mistral AI Model Family
Meta Llama 2 Model Family
Databricks Model Family
Cohere Model Family
Hugging Face Model Family
NVIDIA Model Family
Deci AI Model Family

It's all in Azure AI Studio

Microsoft

Thank you

Let’s stay connected!

Florian Follonier
CSA Data & AI

Q&A

Next Sessions

CALENDAR

Intro to Azure AI Studio
April 19, 2024
REGISTER NOW

Intro to Copilot Studio
April 26, 2024
REGISTER NOW

GitHub Copilot Enterprise
May 3, 2024
REGISTER NOW

Fabric Copilot Demo & News
May 17, 2024
REGISTER NOW

Copilot for Security
May 24, 2024
REGISTER NOW

Hugging Face and Small Language Models
May 31, 2024
REGISTER NOW

PowerApps Premium Benefits
June 7, 2024
REGISTER NOW

Data governance with Microsoft Purview and Fabric
June 14, 2024
REGISTER NOW

Copilot for Microsoft 365
June 21, 2024
REGISTER NOW

LLM Performance
June 28, 2024
REGISTER NOW

Register here


Microsoft
GPS Switzerland
TECH
TALK
FRIDAYS

Running Phi models on mobile device

https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/bringing-genai-offline-running-slm-s-like-phi-2-phi-3-and/ba-p/4128056

Coming soon

Build and filter a custom category