.NET Conf
TAIWAN
2024

STUDY4

為學習而生

特別感謝

Microsoft® MVP Most Valuable Professional

tsmc

好房網 HouseFun

街口支付 JKOPAY

新加坡商 鈦坦科技 TITANSOFT

STUDY4 為學習而生

以及各位參與活動的各位

.NET Conf TAIWAN 2024

LLM 初體驗

Running Microsoft Phi-3 locally

Newest SLM Microsoft Phi-4

王偉任 (Weithenn)
Micron – IT Staff Architect
MVP Microsoft® Most Valuable Professional

About Weithenn

Weithenn.org

Microsoft MVP 2012 - 2024

VMware vExpert 2012 - 2024

Nutanix Technology Champions 2025

19 IT books

Kubernetes Summit 2024, Hello World Dev Conference 2024, COSCUP 2024, DevOpsDays Taipei 2024, Cloud Summit 2024, SRE Conference 2024, DevOpsDays Tokyo 2024, Google DevFest Taipei 2023, .NET Conf Taiwan 2023, Modern Web Conference 2023, Kubernetes Summit 2023, DevOpsDays Taipei 2023, COSCUP 2023...etc.

What is Microsoft Phi-3 ?

What is Microsoft Phi-3 ?

.NET Conf TAIWAN 2024

Microsoft Phi-3 is an open-source AI model family developed by Microsoft, designed to provide efficient and cost-effective **small language models (SLMs)**.

- ✓ **Phi-3-vision:** is a 4.2B parameter multimodal model with language and vision capabilities.
- ✓ **Phi-3-mini:** is a 3.8B parameter language model, available in two context lengths (128K and 4K).
- ✓ **Phi-3-small:** is a 7B parameter language model, available in two context lengths (128K and 8K).
- ✓ **Phi-3-medium:** is a 14B parameter language model, available in two context lengths (128K and 4K).
- ✓ **Multi-platform Support:** Phi-3 models are optimized for NVIDIA GPUs, CPUs, and mobile hardware, supporting ONNX Runtime and Windows DirectML.

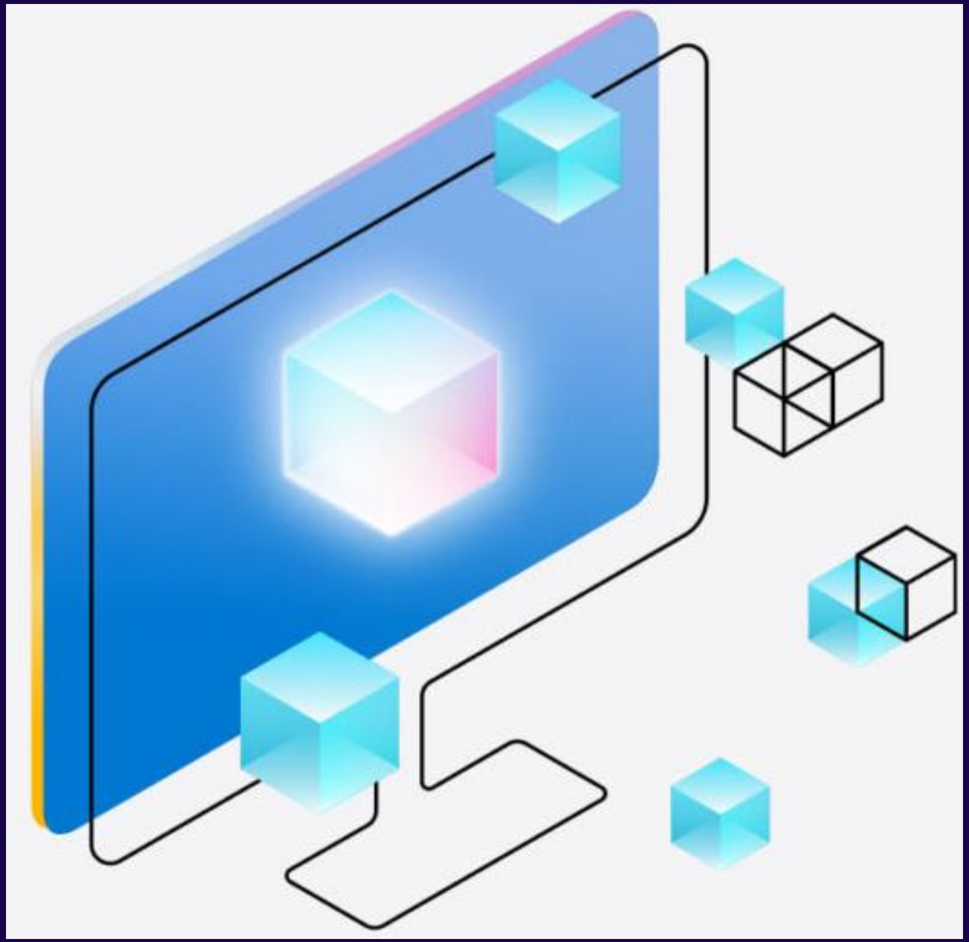


Image From: [Introducing Phi-3: Redefining what's possible with SLMs | Microsoft Azure Blog]

Phi-3

Quality vs. size in SLM

Model quality measured on MMLU benchmark

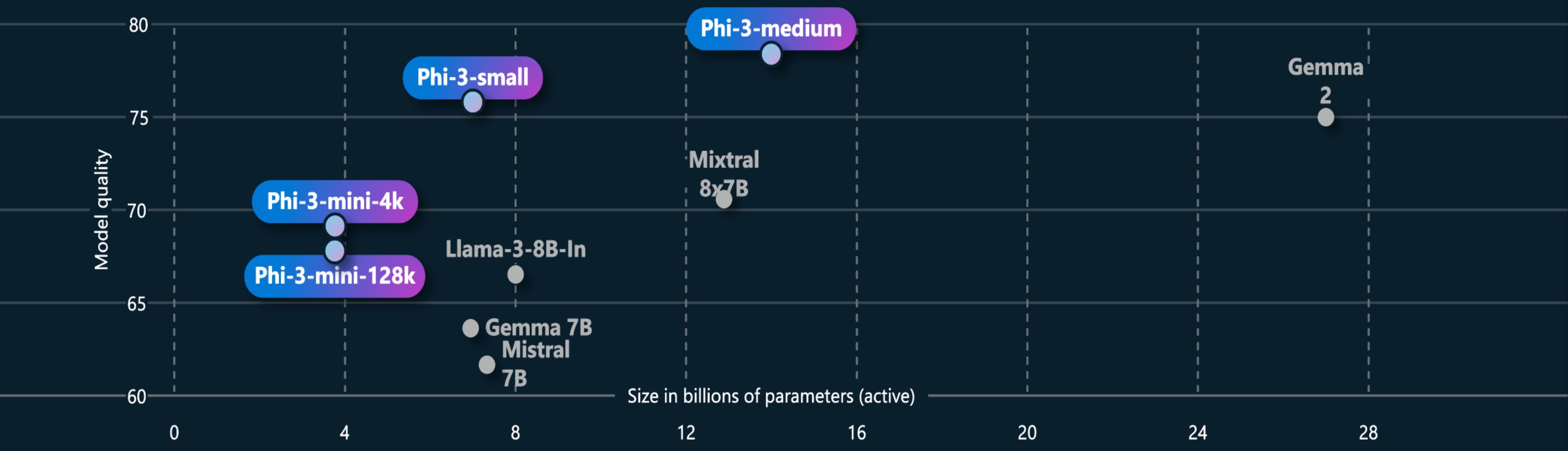
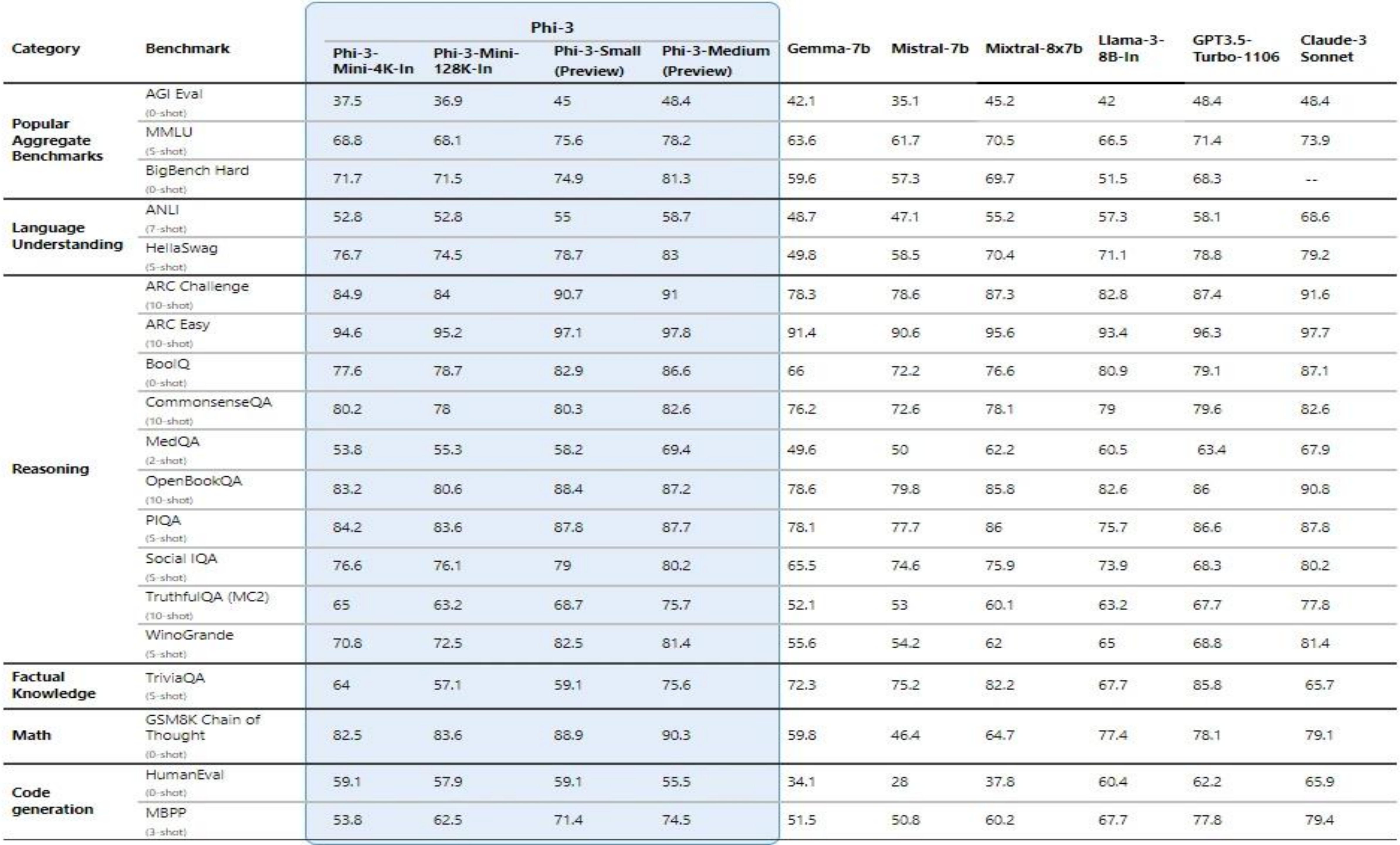
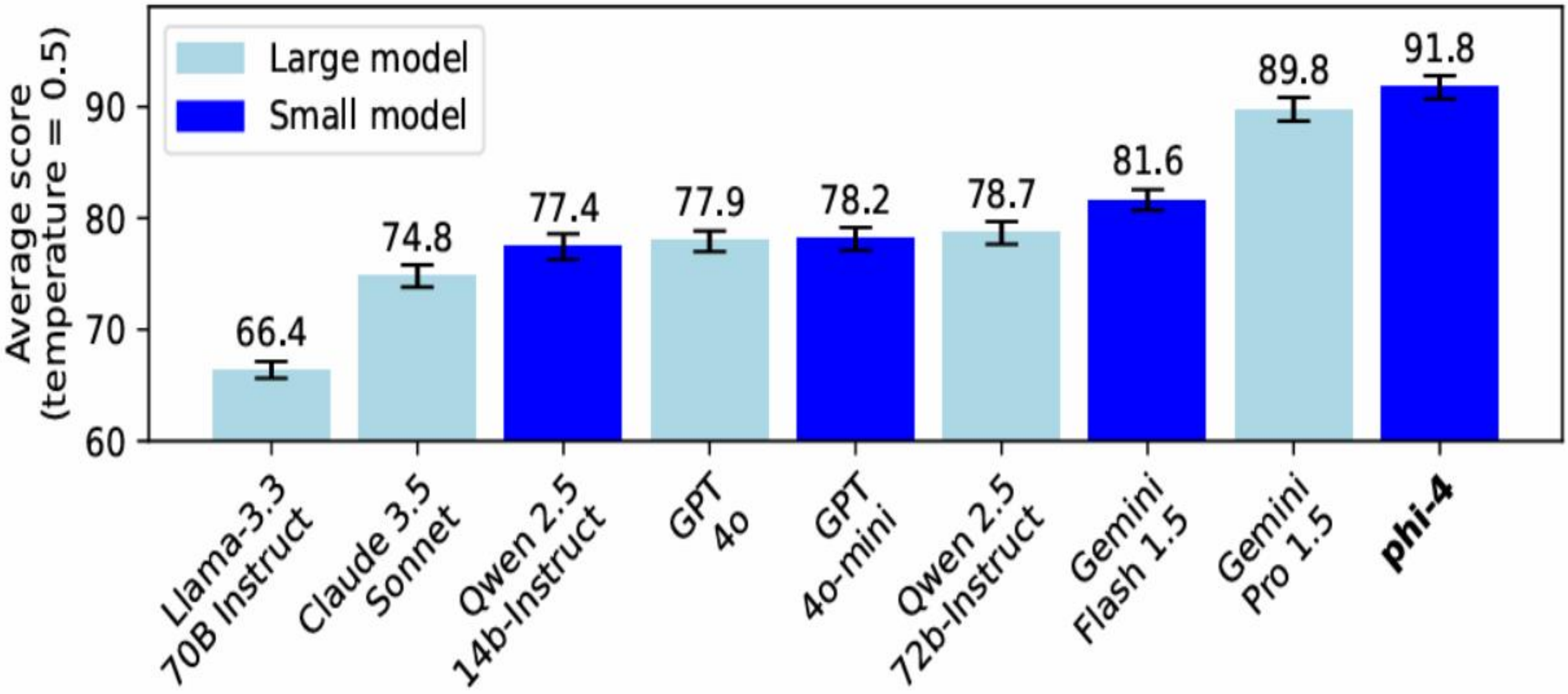


Image From: Introducing Phi-3: Redefining what's possible with SLMs | Microsoft Azure Blog



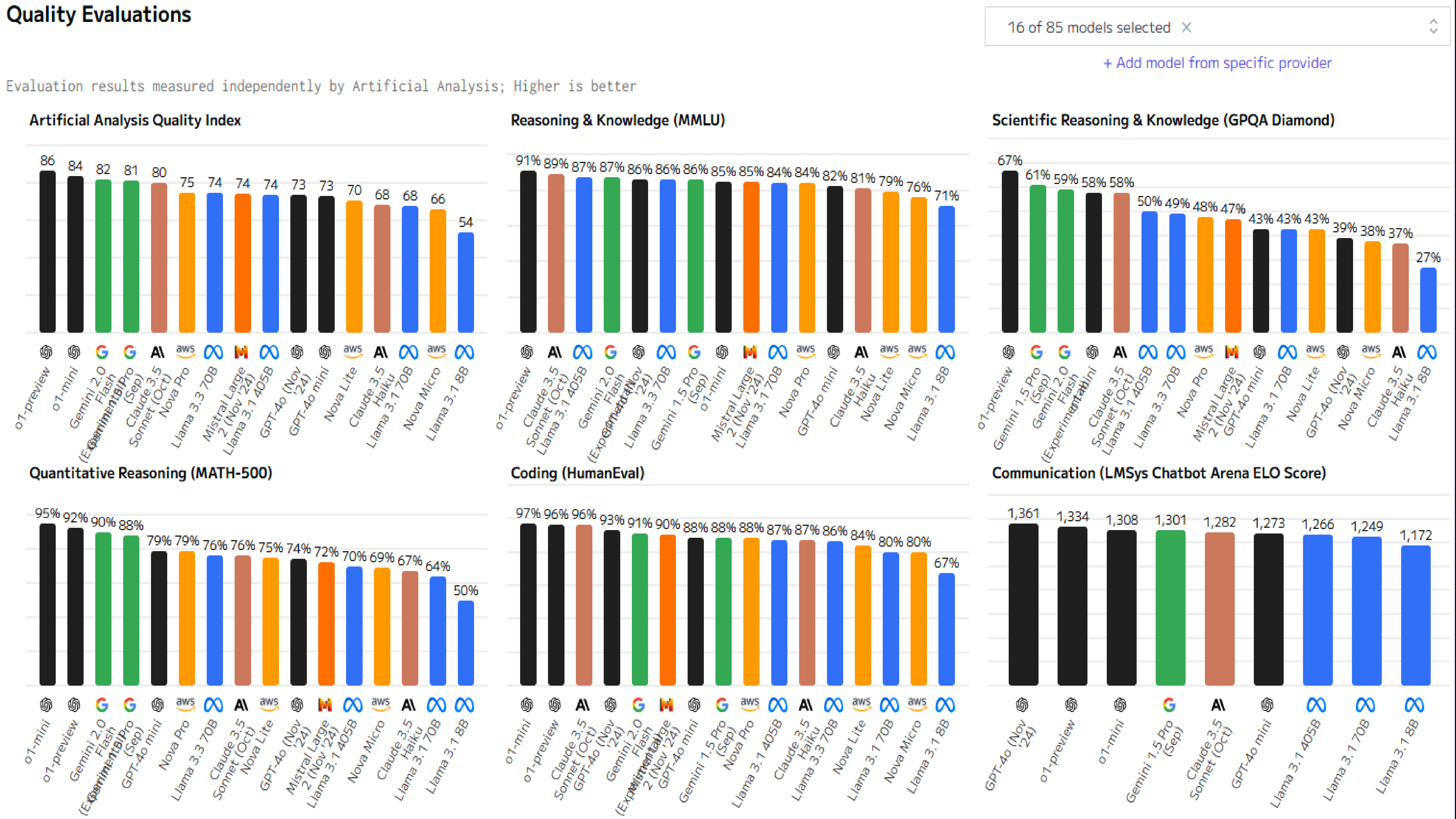
Category	Benchmark	Phi-3				Gemma-7b	Mistral-7b	Mixtral-8x7b	Llama-3-8B-In	GPT3.5-Turbo-1106	Claude-3 Sonnet
		Phi-3-Mini-4K-In	Phi-3-Mini-128K-In	Phi-3-Small (Preview)	Phi-3-Medium (Preview)						
Popular Aggregate Benchmarks	AGI Eval (0-shot)	37.5	36.9	45	48.4	42.1	35.1	45.2	42	48.4	48.4
	MMLU (5-shot)	68.8	68.1	75.6	78.2	63.6	61.7	70.5	66.5	71.4	73.9
	BigBench Hard (0-shot)	71.7	71.5	74.9	81.3	59.6	57.3	69.7	51.5	68.3	--
Language Understanding	ANLI (7-shot)	52.8	52.8	55	58.7	48.7	47.1	55.2	57.3	58.1	68.6
	HellaSwag (5-shot)	76.7	74.5	78.7	83	49.8	58.5	70.4	71.1	78.8	79.2
Reasoning	ARC Challenge (10-shot)	84.9	84	90.7	91	78.3	78.6	87.3	82.8	87.4	91.6
	ARC Easy (10-shot)	94.6	95.2	97.1	97.8	91.4	90.6	95.6	93.4	96.3	97.7
	BoolQ (0-shot)	77.6	78.7	82.9	86.6	66	72.2	76.6	80.9	79.1	87.1
	CommonsenseQA (10-shot)	80.2	78	80.3	82.6	76.2	72.6	78.1	79	79.6	82.6
	MedQA (2-shot)	53.8	55.3	58.2	69.4	49.6	50	62.2	60.5	63.4	67.9
	OpenBookQA (10-shot)	83.2	80.6	88.4	87.2	78.6	79.8	85.8	82.6	86	90.8
	PIQA (5-shot)	84.2	83.6	87.8	87.7	78.1	77.7	86	75.7	86.6	87.8
	Social IQA (5-shot)	76.6	76.1	79	80.2	65.5	74.6	75.9	73.9	68.3	80.2
	TruthfulQA (MC2) (10-shot)	65	63.2	68.7	75.7	52.1	53	60.1	63.2	67.7	77.8
	WinoGrande (5-shot)	70.8	72.5	82.5	81.4	55.6	54.2	62	65	68.8	81.4
Factual Knowledge	TriviaQA (5-shot)	64	57.1	59.1	75.6	72.3	75.2	82.2	67.7	85.8	65.7
Math	GSM8K Chain of Thought (0-shot)	82.5	83.6	88.9	90.3	59.8	46.4	64.7	77.4	78.1	79.1
Code generation	HumanEval (0-shot)	59.1	57.9	59.1	55.5	34.1	28	37.8	60.4	62.2	65.9
	MBPP (3-shot)	53.8	62.5	71.4	74.5	51.5	50.8	60.2	67.7	77.8	79.4



Average performance on November 2024 AMC 10/12 tests

Language Model Comparison Highlights

.NET Conf TAIWAN 2024



Quality Evaluations

16 of 85 models selected ×

+ Add model from specific provider

Evaluation results measured independently by Artificial Analysis; Higher is better

Artificial Analysis Quality Index

Reasoning & Knowledge (MMLU)

Scientific Reasoning & Knowledge (GPQA Diamond)

Quantitative Reasoning (MATH-500)

Coding (HumanEval)

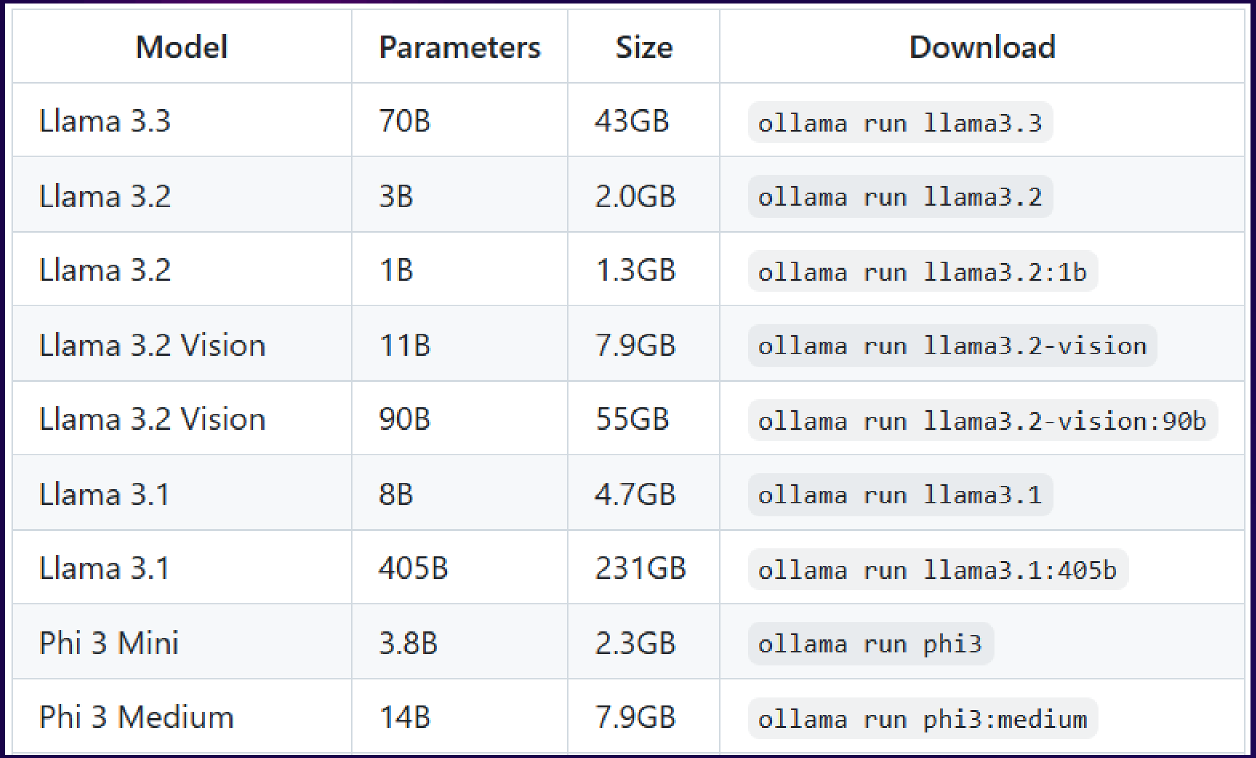
Communication (LMSys Chatbot Arena ELO Score)

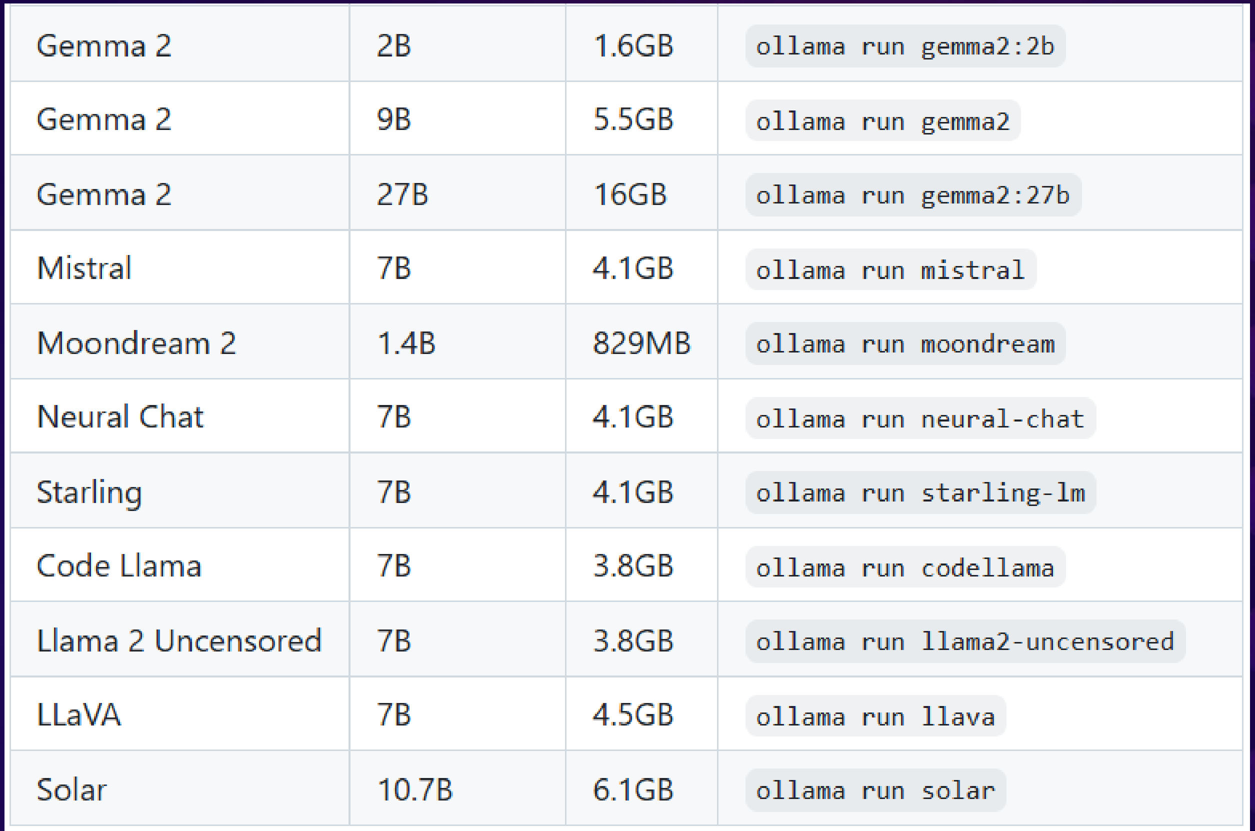
What is Ollama ?

What is Ollama ?

- ✓**OpenAI Compatibility:** Ollama is compatible with OpenAI's API, making it easy to manage models.
- ✓**Model Customization:** Allows for model customization with prompts.
- ✓**Model Library:** Provides a library of pre-built models. For example, Llama 3.1, Phi 3 Mini, Gemma 2, Mistral...etc.
- ✓**System Requirements:** Requires at least 8 GB of RAM available to run the 7B models, 16 GB to run the 13B models, and 32 GB to run the 33B models.

Ollama – Model Library





Model	Parameters	Size	Download
Llama 3.3	70B	43GB	`ollama run llama3.3`
Llama 3.2	3B	2.0GB	`ollama run llama3.2`
Llama 3.2	1B	1.3GB	`ollama run llama3.2:1b`
Llama 3.2 Vision	11B	7.9GB	`ollama run llama3.2-vision`
Llama 3.2 Vision	90B	55GB	`ollama run llama3.2-vision:90b`
Llama 3.1	8B	4.7GB	`ollama run llama3.1`
Llama 3.1	405B	231GB	`ollama run llama3.1:405b`
Phi 3 Mini	3.8B	2.3GB	`ollama run phi3`
Phi 3 Medium	14B	7.9GB	`ollama run phi3:medium`
Gemma 2	2B	1.6GB	`ollama run gemma2:2b`
Gemma 2	9B	5.5GB	`ollama run gemma2`
Gemma 2	27B	16GB	`ollama run gemma2:27b`
Mistral	7B	4.1GB	`ollama run mistral`
Moondream 2	1.4B	829MB	`ollama run moondream`
Neural Chat	7B	4.1GB	`ollama run neural-chat`
Starling	7B	4.1GB	`ollama run starling-lm`
Code Llama	7B	3.8GB	`ollama run codellama`
Llama 2 Uncensored	7B	3.8GB	`ollama run llama2-uncensored`
LLaVA	7B	4.5GB	`ollama run llava`
Solar	10.7B	6.1GB	`ollama run solar`

Image From: ollama/ollama: Get up and running with Llama 3.3, Mistral, Gemma 2, and other large language models.

Live Demo

Resource & More

Introducing Phi-3: Redefining what's possible with SLMs | Microsoft Azure Blog

New models added to the Phi-3 family, available on Microsoft Azure | Microsoft Azure Blog

ollama/ollama: Get up and running with Llama 3.1, Mistral, Gemma 2, and other large language models. (github.com)

open-webui/open-webui: User-friendly WebUI for LLMs (Formerly Ollama WebUI) (github.com)

AI Model & API Providers Analysis | Artificial Analysis

Get .NET 9

Download .NET 9

aka.ms/get-dotnet-9

Thank you