Microsoft

Getting started with Phi-4 and Chainlit

Sergio Zenatti Filho

Sr Cloud Solution Architect,
Microsoft

Connect

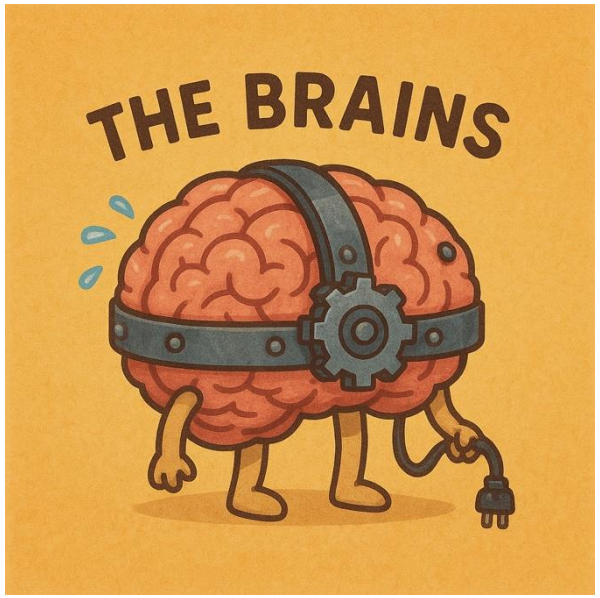
in /sergiozenatti

Today’s Presentation Story

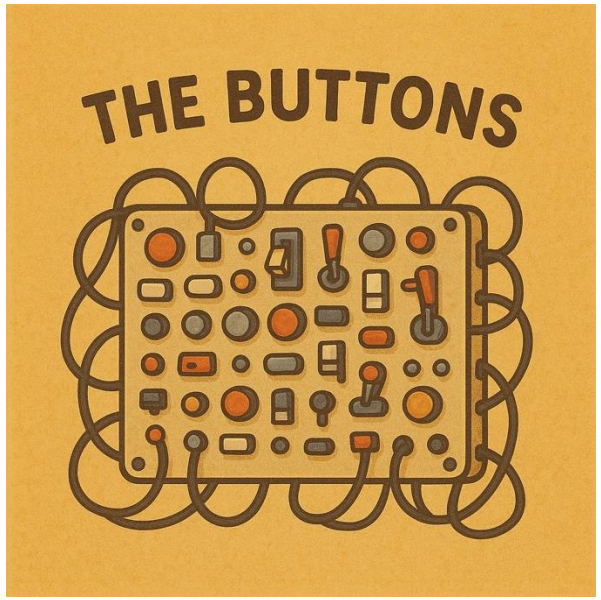
We own a **toy workshop**, and our **dream** is to build **smart**, **talking** toys for every child in the world.

However, we face two big challenges:

The Brains – Most talking-toy brains are huge and power hungry. Fitting them into a tiny teddy bear? Impossible!

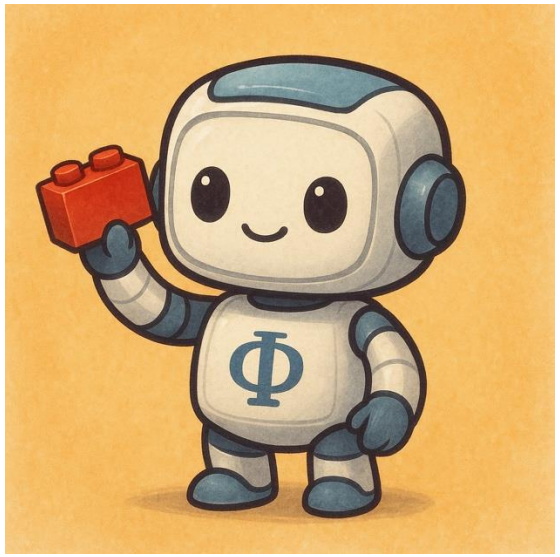


The Buttons – Wiring every button and switch by hand will take forever. By the time we finished, the holidays are over!



Our Heroes

Tiny Phi (Phi-4) – a pocket-sized genius. Think of him as a Lego-brick brain: small, light, but surprisingly clever.



Magic Chain (Chainlit) – a super-fast conveyor belt that snaps those Lego brains into toys with just a tap.

"Tiny Phi + Magic Chain = lightning-fast, high-performance AI fun!"

What is Phi-4

State-of-the-art open model trained on synthetic datasets, filtered public-domain web content, academic books, and Q&A datasets.

Focused on high-quality, advanced reasoning in a compact 14B-parameter architecture.

Model: 14 billion-parameter, dense decoder-only Transformer

Context Window: 16 000 tokens

Input: Text prompts (chat format)

Training Details

Data: 9.8 trillion tokens (cutoff June 2024)

Compute: 1 920 NVIDIA H100-80 GB GPUs

Duration: 21 days (Oct 2024 – Nov 2024)

Release Date: December 12, 2024

License: MIT

Downloads last month
422,302

		Small models				Large models		
		phi-4 14b	**phi-3** 14b	**Qwen 2.5** 14b instruct	**GPT** 4o-mini	**Llama-3.3** 70b instruct	**Qwen 2.5** 72b instruct	**GPT** 4o
simple-evals	MMLU	84.8	77.9	79.9	81.8	86.3	85.3	**88.1**
	GPQA	**56.1**	31.2	42.9	40.9	49.1	49.0	50.6
	MATH	**80.4**	44.6	75.6	73.0	66.3[1]	80.0	74.6
	HumanEval	82.6	67.8	72.1	86.2	78.9[1]	80.4	**90.6**
	MGSM	80.6	53.5	79.6	86.5	89.1	87.3	**90.4**
	SimpleQA	3.0	7.6	5.4	9.9	20.9	10.2	**39.4**
	DROP	75.5	68.3	85.5	79.3	**90.2**	76.7	80.9
	MMLUPro	70.4	51.3	63.2	63.4	64.4	69.6	**73.0**
	HumanEval+	82.8	69.2	79.1	82.0	77.9	78.4	**88.0**
	ArenaHard	75.4	45.8	70.2	76.2	65.5	**78.4**	75.6
	LiveBench	47.6	28.1	46.6	48.1	**57.6**	55.3	**57.6**
	IFEval	63.0	57.9	78.7	80.0	**89.3**	85.0	84.8
	PhiBench (internal)	56.2	43.9	49.8	58.7	57.1	64.6	**72.4**

Phi Family

Open Source With MIT License

Language

- Phi-1.5-1.3B
- Phi-2-2.7B
- Phi-3-mini-3.8B
- Phi-3-small-7B
- Phi-3-medium-14B
- Phi-3.5-mini-3.8B
- NEW! Phi-4-14B
- NEW! Phi-4-mini-3.8B
- NEW! Phi-4-multimodal-5.6B

Coding

- Phi-1-1.3B
- Phi-1.5-1.3B
- Phi-2-2.7B
- Phi-3-mini-3.8B
- Phi-3-small-7B
- Phi-3-medium-14B
- Phi-3.5-mini-3.8B
- NEW! Phi-4-14B
- NEW! Phi-4-mini-3.8B
- NEW! Phi-4-multimodal-5.6B

Vision

- Phi-3-VISION-4.2B
- Phi-3.5-VISION-4.2B
- NEW! Phi-4-multimodal-5.6B

Function calling

- NEW! Phi-4-mini-3.8B

Audio

- NEW! Phi-4-multimodal-5.6B

Advanced Reasoning

- NEW! Phi-4-14B
- NEW! Phi-4-mini-3.8B

MoE

- Phi-3.5-MoE-42B (Active params is 6.6B)

Azure AI Foundry

Hugging Face

GitHub Models

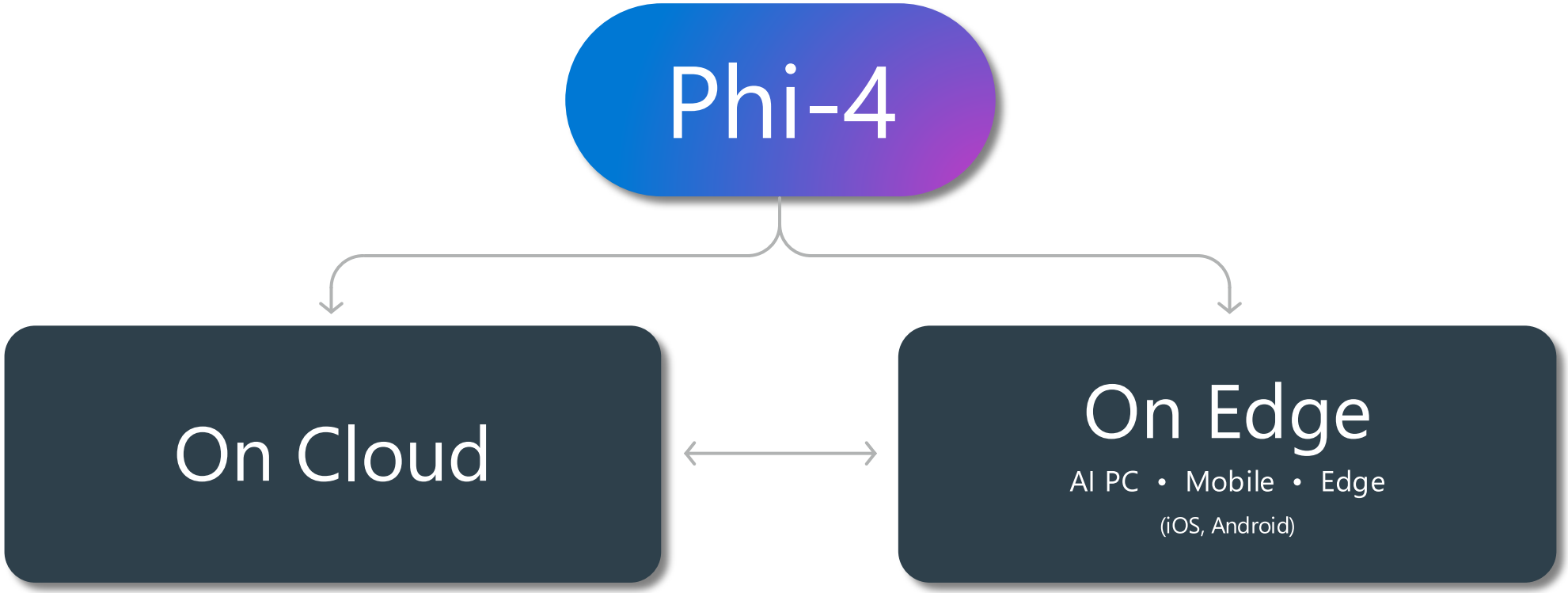
Available on (HF,ONNX,GGUF)

NVIDIA NIM

Ollama

AITK

LM Studio


Phi-4
On Cloud
On Edge
AI PC • Mobile • Edge
(iOS, Android)

Demo – Model Details

What is Chainlit

Open-source Python framework for building and sharing LLM-powered chat apps in minutes.

Provides an instant ChatGPT-style UI, step-by-step reasoning visualisation, data logging, and first-class integrations with LangChain, LlamaIndex, Haystack and any custom Python code.

Language Stack: Python (≥ 3.9) back-end & React/TypeScript front-end

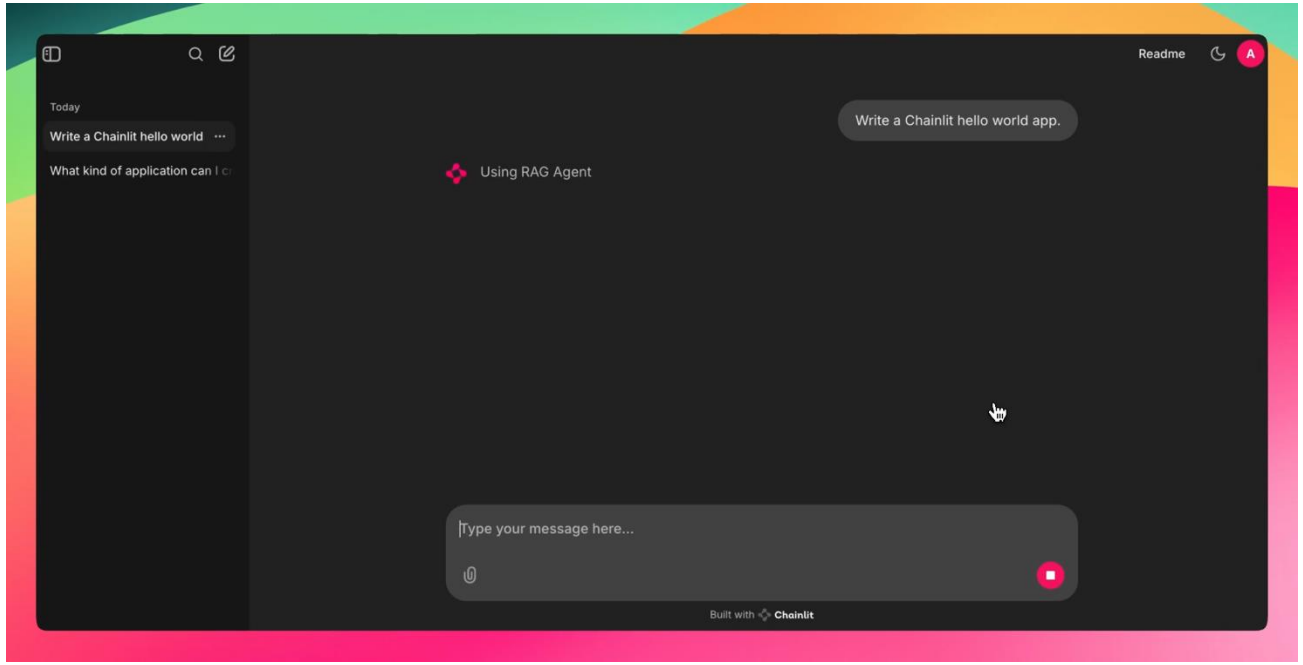
CLI: chainlit run my_app.py (live-reload dev server)

UI Widgets: Markdown, code blocks, images, audio, file-upload, forms

Dev Tools: trace viewer, prompt/state diff, telemetry & analytics dashboard

Deployment: local dev, Docker, static export, or managed Chainlit Cloud

Integrations: OpenAI, Anthropic, Hugging Face, Azure OpenAI, Vertex AI, etc.



Current Stable: Version 2.5.5 (released April 2025)

Initial Release: v0.1.0 – July 20, 2023

GitHub Stats: ~9.4k stars, ~1.3k forks

License: Apache 2.0

https://docs.chainlit.io/

https://github.com/Chainlit/chainlit

Demo

Chainlit + Phi4 + Phi4 Multimodel

Get started with Phi-4 and Chainlit

- Read blogpost on Azure http://aka.ms/phi4
- Chainlit + Phi-4 (Ollama) Quick-start https://github.com/szenatti/phi4-chainlit-demo-ai-bootcamp
- Installing & Running Microsoft Phi-4 Multimodal Instruct Locally with Chainlit https://github.com/szenatti/phi4-multimodel
- Chainlit documentation https://docs.chainlit.io/get-started/overview

Q&A

Thanks!

Connect

/sergiozenatti