Microsoft

Introduce
Phi-4-mini & Phi-4-multimodal

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
- NEW! Phi-4-multimodal-5.6B (text only)

Audio

- NEW! Phi-4-multimodal-5.6B

Advanced Reasoning

- NEW! Phi-4-reasoning-14B
- NEW! Phi-4-mini-reasoning-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-multimodal's groundbreaking performance

Phi-4-multimodal brings together speech, vision, and text for enhanced efficiency

Category		Benchmark	Phi-4-Mini-MM	Qwen2-Audio	WhisperV3	SeamlessM4T-V2-Large	Gemini-1.5-Flash	Gemini-1.5-Pro	GPT-4o-RT-preview-10-01-2024
Speech recognition (lower is better)		CommonVoice	**6.8**	8.6	8.1	8.5	14.9	8.9	18.1
		FLEURS	**4.0**	8.3	4.6	7.3	5.2	3.6	5.4
		OpenASR	**6.1**	7.4	7.4	20.7	13.3	9.5	15.8
Speech translation	X->En	CoVoST2	**40.8**	34.8	33.3	37.5	23.6	35.9	37.1
		FLEURS	**32.3**	23.7	25.8	28.9	20.6	32.3	32.6
	En->X	CoVoST2	**38.7**	34.0	N/A	32.8	15.2	20.7	37.2
		FLEURS	**33.6**	23.2	N/A	30.4	30.8	35.5	36.8
Speech QA		MT Bench	**7.0**	4.9	N/A	N/A	8.3	8.5	8.1
		MMMLU-8L	**38.5**	15.5	N/A	N/A	65.3	72.1	72.6
		MMMLU-EN	**54.3**	16.0	N/A	N/A	73.0	79.1	78.8
Audio understanding		AIRBench Chat	**7.0**	6.9	N/A	N/A	6.9	7.1	6.5
		MMAU	**55.6**	52.5	N/A	N/A	43.4	46.9	53.3
Speech summarization		Golden3	**6.3**	2.3	N/A	N/A	6.7	6.7	6.8
		AMI	**6.3**	1.3	N/A	N/A	6.7	6.6	6.5

Speech:
The Phi-4-multimodal has demonstrated exceptional capabilities in speech-related tasks, emerging as a groundbreaking open model in multiple areas. It especially outperforms specialized models in both automatic speech recognition (ASR) and speech translation (ST).

Support 20+ languages:
Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Ukrainian

Current vocabulary bank is 200K words (previous models were 32K words)

Phi-4-multimodal's groundbreaking performance

Phi-4-multimodal brings together speech, vision, and text for enhanced efficiency

Category	Benchmark	Phi-4-Mini-MM-Ins	Qwen2-VL-2B-Ins	InternVL 2.5-4B	Qwen2-VL 7B-Ins	InternVL 2.5-8B	Gemini-1.5-Flash	Gemini-1.5-Pro	Claude-3.5-Sonnet-2024-10-22	Gpt-4o-2024-11-20
Popular aggregated benchmark	MMMU	**55.1**	38.2	48.3	50.1	50.6	49.3	54.1	55.8	61.7
	MMBench (dev-en)	**86.7**	79.4	86.8	85	88.2	85.7	87.9	86.7	89
	MMMU-Pro (standard / vision)	**38.5**	23.2	32.4	31.4	34.4	37.1	51.3	54.3	53
Visual science reasoning	ScienceQA Visual (img-test)	**97.5**	77	96.2	85	97.3	84.5	86	81.2	88.2
Visual math reasoning	MathVista (testmini)	**62.4**	32.8	51.2	55.9	56.7	55.3	57.4	56.9	56.1
	InterGPS	**48.6**	37.4	53.7	44.4	54.1	39.4	58.2	47.1	49.1
Chart & table reasoning	AI2D	**82.3**	69.7	80	80.1	83	78.4	75.6	70.6	83.8
	ChartQA	**81.4**	71.1	79.1	79.6	81	57.6	68.2	78.4	75.1
	DocVQA	**93.2**	90.1	91.6	94.5	93	82.6	93.1	95.2	80.4
	InfoVQA	**72.7**	65.5	72.1	76.5	77.6	65.2	81	74.3	65.1
Document Intelligence	TextVQA (val)	**75.6**	78.4	70.9	81.7	74.8	67.4	64.5	58.6	73.1
	OCR Bench	**84.4**	78.5	71.6	84.2	74.8	75	74.5	77	77.7
Object visual presence verification	POPE	**85.6**	87.3	89.4	88.4	89.1	86.1	89.3	82.6	86.5
Multi-image perception	BLINK	**61.3**	41.2	51.2	51.3	52.5	45.8	61	56.9	62.4
	Video MME 16 frames	**55**	51.5	57.3	57.6	58.7	62.3	62.6	60.2	68.2
Average		**72**	**61.4**	**68.8**	**69.7**	**71.1**	**64.8**	**71**	**69.1**	**71.3**

Vision reasoning:
Phi-4-multimodal is comprised of 5.6B active parameters and on average outperforms competitor models of the same size. The vision capabilities make this model competitive against much bigger models with multi-frame capabilities across various benchmarks, most notably achieving strong performance on mathematical and science reasoning.

Phi-4-multimodal's groundbreaking performance

Phi-4-multimodal brings together speech, vision, and text for enhanced efficiency

Benchmarks	Phi-4-Mini-MM-Ins	InternOmni-7B	Gemini-1.5-Flash	Gemini-1.5-Pro
s_AI2D	68.9	53.9	69.5	67.7
s_ChartQA	69.0	56.1	36.2	39.6
s_DocVQA	87.3	79.9	76.5	78.2
s_InfoVQA	63.7	60.3	62.4	66.1
Average	72.2	62.6	61.2	62.9

Audio + text reasoning:
Phi-4-multimodal is also able to process both visual and audio in one input query. The Phi-4-multimodal model achieves stronger performance across multiple benchmarks when the model quality when the input query for vision content is synthetic speech on chart/table understanding and document reasoning tasks

Phi-4-mini's enhanced performance

Phi-4-mini outperforms language models of the same size and larger

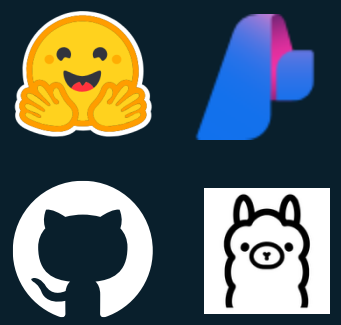
Category	Benchmark	Phi-4-Mini-Ins	Phi-3.5-Mini-Ins	Llama-3.2-3B-Ins	Ministral-3B	Qwen2.5-3B-Ins	Qwen2.5-7B-Ins	Ministral-8B-2410	Llama-3.1-8B-Ins	Gemma 2-9B-It	GPT-4o-mini-2024-07-18
Popular aggregated benchmarks	MMLU (5-shot)	67.3	34.4	61.8	60.8	65	72.6	63	68.1	71.3	77.2
	MMLU-Pro (0-shot, CoT)	52.8	63.1	39.2	35.3	44.7	56.2	36.6	44	50.1	62.8
	Arena Hard	32.8	65.5	17	26.9	32	55.5	37.3	25.7	43.7	75
	BigBench Hard CoT (0-shot)	70.4	47.4	55.4	51.2	56.2	72.4	53.3	63.4	65.7	80.4
Reasoning	ARC Challenge (10-shot)	83.7	84.6	76.1	80.3	82.6	90.1	82.7	83.1	89.8	93.5
	BoolQ (2-shot)	81.2	77.7	71.4	79.4	65.4	80	80.5	82.8	85.7	88.7
	GPQA (0-shot, CoT)	30.4	25.2	26.6	24.3	24.3	30.6	26.3	26.3	31	41.1
	HellaSwag (5-shot)	69.1	72.2	69	77.2	74.6	80.1	80.9	73.5	80.9	87.1
	OpenBook QA (10-shot)	79.2	81.2	72.6	79.8	77.6	86	80.2	84.8	89.6	90
	PIQA (5-shot)	77.6	78.2	68.2	78.3	77.2	80.8	76.2	81.2	83.7	88.7
	Social IQA (5-shot)	72.5	75.1	68.3	73.9	75.3	75.3	77.6	71.8	74.7	82.9
	TruthfulQA (MC2) (10-shot)	66.4	65.6	59.2	62.9	64.3	69.4	63	69.2	76.6	78.2
	WinoGrande (5-shot)	67	72.2	53.2	59.8	63.3	71.1	63.1	64.7	74	76.9
Multilingual	Multilingual MMLU (5-shot)	49.3	51.8	48.1	46.4	55.9	64.4	53.7	56.2	63.8	72.9
	MGSM (0-shot CoT)	63.9	47	49.6	44.6	53.5	64.5	58.3	56.7	75.1	81.7
Math	GSM8K (8-shot, CoT)	88.6	76.9	75.6	80.1	80.6	88.7	81.9	82.4	84.9	91.3
	MATH (0-shot, CoT)	64	49.8	46.7	41.8	61.7	60.4	41.6	47.6	51.3	70.2
Long context	Qasper	40.4	41.9	33.4	35.3	32.1	38.1	37.4	37.2	13.9	39.8
	SQuALITY	22.8	25.3	25.7	25.5	25.3	23.8	24.9	26.2	23.6	23.8
Instruction follow	IFEval	70.1	50.6	68	47.5	59	69.5	52.5	74.1	73.2	80.1
Function call	BFCL	70.3	66.1	78.6	61.4	74.2	81.3	74	77	59.9	83.3
Code generation	HumanEval (0-shot)	74.4	70.1	62.8	72	72	75	70.7	66.5	63.4	86.6
	MBPP (3-shot)	65.3	70	67.2	65.1	65.3	76.3	68.9	69.4	69.6	84.1
Overall		63.5	60.5	**56.2**	**56.9**	**60.1**	**67.9**	**60.2**	**62.3**	**65**	**75.5**

Phi-4-multimodal is comprised of 3.8B active parameters outperforms various other models in reasoning, language understanding, and math

Support 20+ languages:

Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Ukrainian

Microsoft Phi Family in Edge AI



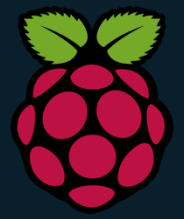
Phi Family

Microsoft Olive

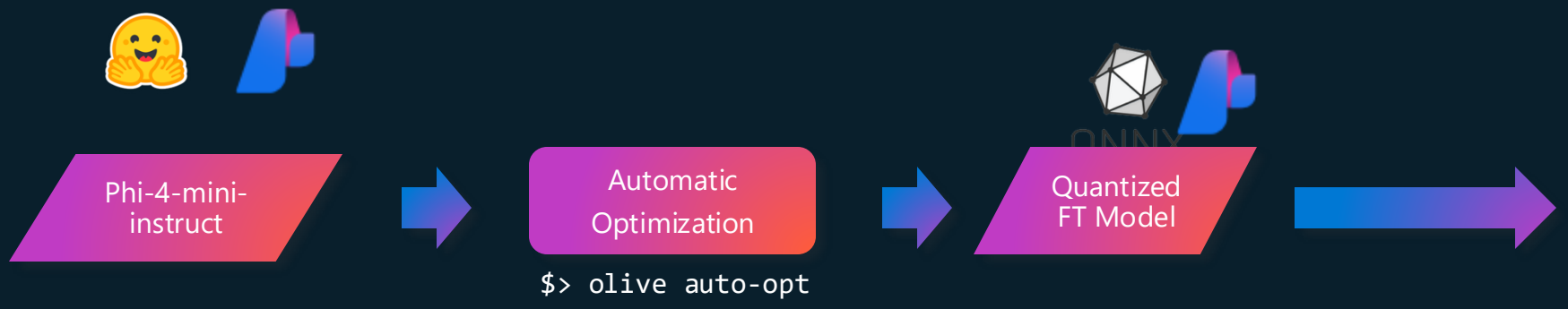
Intel OpenVINO

Quantization and conversion

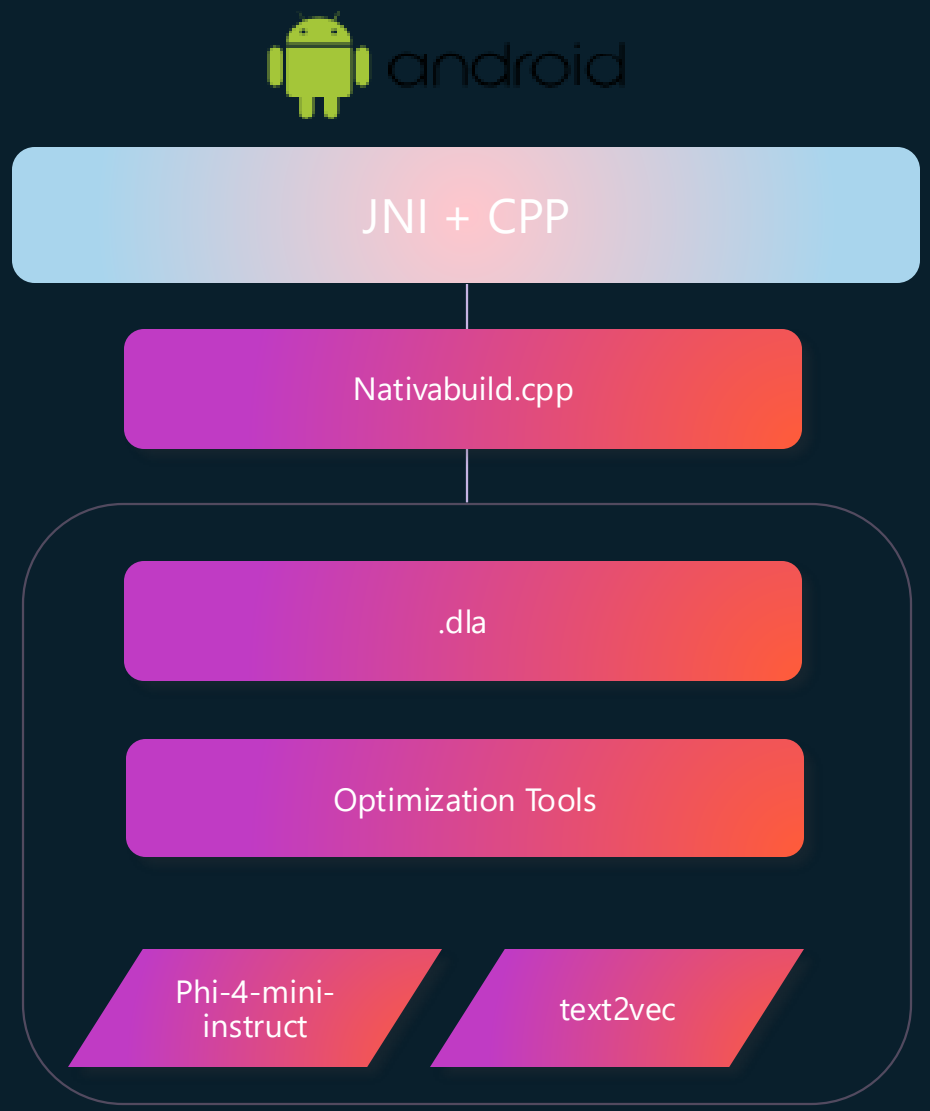
Deployment

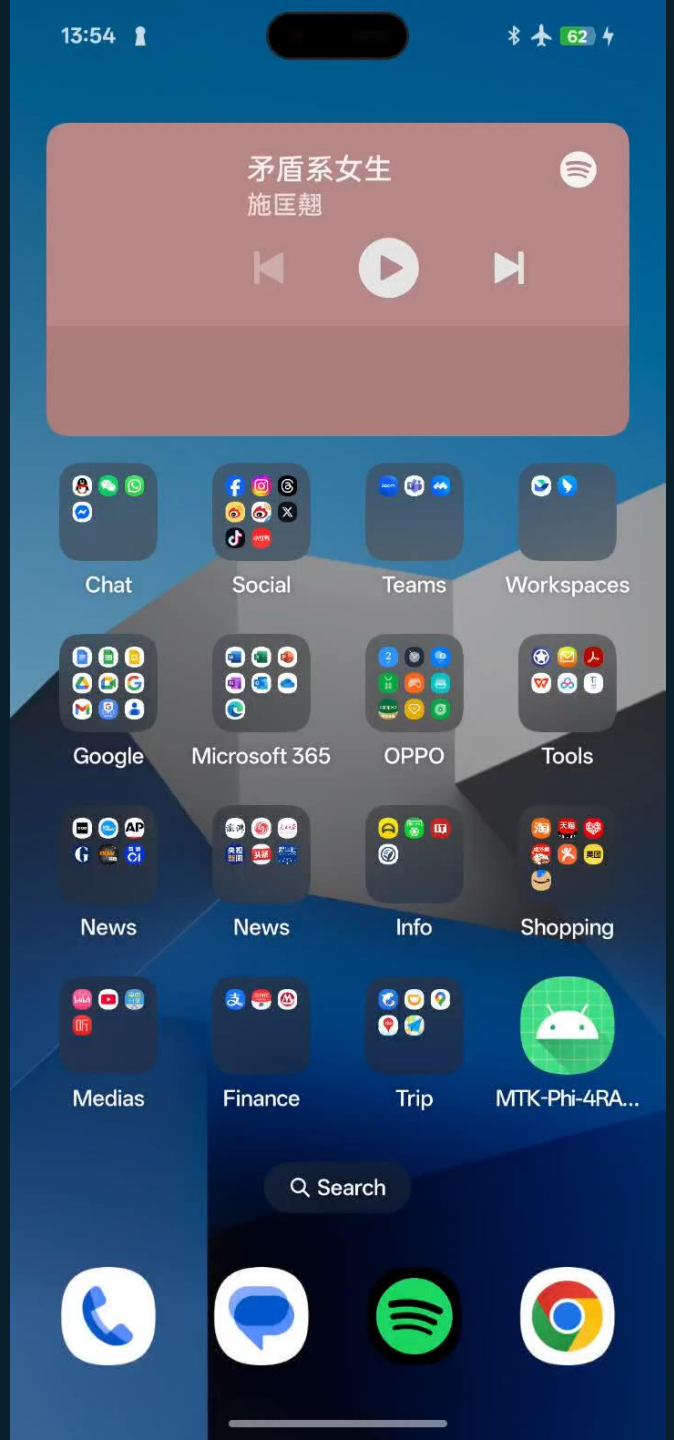


Phi-4-mini-onnx run in iPhone 12 Pro



Cases – RAG@EdgeAI with MTK





Phi-4-mini
Function Calling

1. init function calling

Tool Definitions + Binding Tools JSON

get_match_result(match)

```
tools = [
    {
        "name": "get_match_result",)
        "description": "get match result",
        "parameters": {
        "match": {
            } "description": "The name of the match",
            "type": "str",
            "default": "Arsenal vs ManCity"
        }

    },
]
```

+

2. Messages

Messages

```
{
    "role": "system",
    "content": "You are a helpful assistant",
    `"tools": json.dumps(tools),
},
```

"What is the result of Arsenal vs ManCity today?

3. Call

Tool call get_match_result('Arsenal vs ManCity')

4. Result

"Arsenal vs ManCity": "1:1"

NEW! Phi-4-mini-3.8B

Build AI Agent with Phi-4-mini (With Function Calling)

Running Jupyter Notebook by AI Agent(By Phi-4-mini)

{ Travel Agent }

Booking fights / hotels

```
AGENT_TOOLS = {
      "booking_fight": {
          "name": "booking_fight",
          "description": "booking fight",
          "parameters": {
              "departure": {
                  "description": "The name of Departure airport code",
                  "type": "str",
              },
              "destination": {
                  "description": "The name of Destination airport code",
                  "type": "str",
              },
              "outbound_date": {
                  "description": "The date of outbound flight",
                  "type": "str",
              },
              "return_date": {
                  "description": "The date of return flight",
                  "type": "str",
              }
          }
      },
      "booking_hotel": {
          "name": "booking_hotel",
          "description": "booking hotel",
          "parameters": {
              "query": {
                  "description": "The name of the city",
                  "type": "str",
              },
              "check_in_date": {
                  "description": "The date of check in",
                  "type": "str",
              },
              "check_out_date": {
                  "description": "The date of check out",
                  "type": "str",
              }
          }
      },
}
```

I have found several hotel options for your stay in New York from March 21 to March 27, 2025. Here are some of the best ones:\n\n1. **The Fifty Sonesta Hotel New York**: A 4-star hotel located at 51 St, with an overall rating of 4.1. It offers stylish rooms and suites, a lounge, bar, and free wine hours. The hotel is located near St. Patrick's Cathedral and John F. Kennedy International Airport. The rate per night is $236, with a total rate of $1,417. The hotel is pet-friendly and offers free Wi-Fi, parking, and a fitness center.\n\n2. **The Manhattan at Times Square Hotel**: A 4-star hotel located at 7 Av, with an overall rating of 3.8. It offers modern rooms and suites, a marble-clad lobby lounge, and a gym. The hotel is located near Grand Central Terminal and John F. Kennedy International Airport. The rate per night is $218, with a total rate of $1,307. The hotel is pet-friendly and offers free Wi-Fi, parking, and a restaurant.\n\n3. **The Manhattan by Marriott**: A 4-star hotel located at 7 Av, with an overall rating of 4.1. It offers contemporary rooms and suites, a rooftop lounge, and a fitness center. The hotel is located near Grand Central Terminal and John F. Kennedy International Airport. The rate per night is $238, with a total rate of $1,242. The hotel is pet-friendly and offers free Wi-Fi, parking, and a restaurant.\n\n4. **The Park Hyatt New York**: A 5-star hotel located at 5 Av, with an overall rating of 4.9. It offers luxurious rooms and suites, a rooftop terrace, and a spa. The hotel is located near Grand Central Terminal and John F. Kennedy International Airport. The rate per night is $398, with a total rate of $2,191. The hotel is pet-friendly and offers free Wi-Fi, parking, and a fitness center.\n\n5. **The Langham, New York**: A 5-star hotel located at 7 Av, with an overall rating of 4.9. It offers elegant rooms and suites, a rooftop terrace, and a restaurant. The hotel is located near Grand Central Terminal and John F. Kennedy International Airport. The rate per night is $398, with a total rate of $2,191. The hotel is pet-friendly and offers free Wi-Fi, parking, and a fitness center.\n\nPlease let me know which hotel you would like to book, and I will proceed with the booking process.

Phi-4-mulitimodal

Audio

OCR

Multi-Language

Vision

Vision + Audio - Phi-4-mulitimodal

https://github.com/kinfey/PhiCookbook/blob/main/md/02.Application/08.Multimodel/Phi4/TechJournalist/phi_4_mm_audio_text_publish_news.ipynb

Audio - Phi-4-mulitimodal

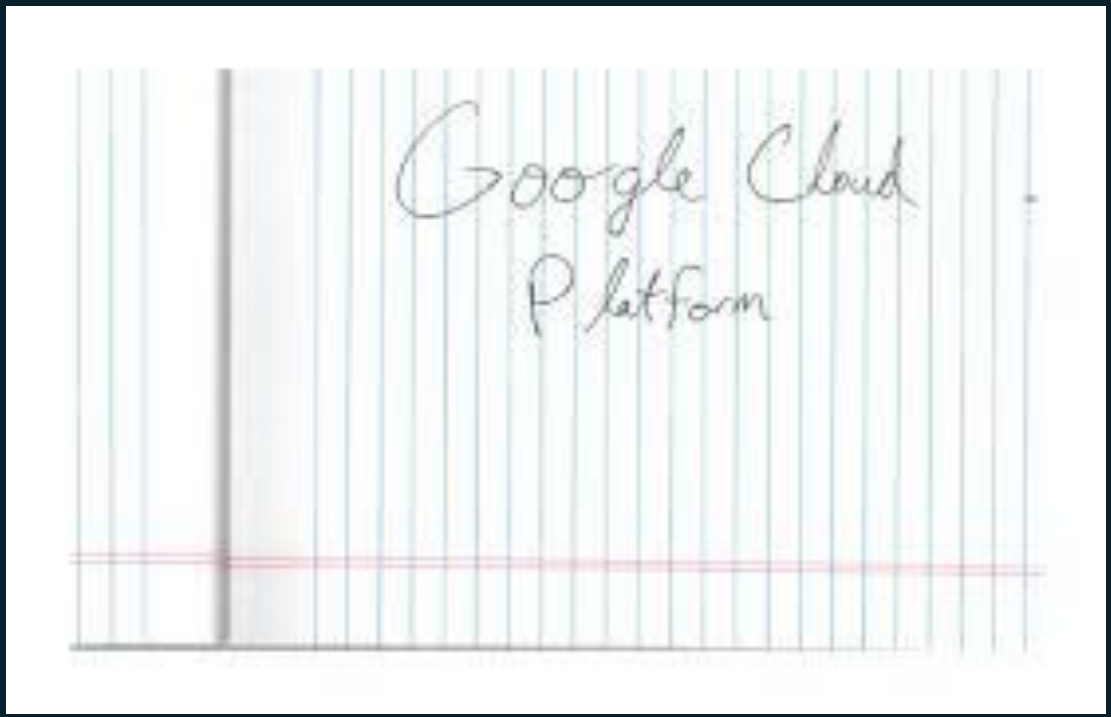
Speech to Text

```
[ ] import soundfile

[ ] speech_prompt = "Based on the attached audio, generate a comprehensive text transcription of the spoken content."
    prompt = f'{user_prompt}<|audio_1|>{speech_prompt}{prompt_suffix}{assistant_prompt}'

[ ] audio = soundfile.read('./satya.wav')

[ ] inputs = processor(text=prompt, audios=[audio], return_tensors='pt').to('cuda:0')

[ ] generate_ids = model.generate(
        **inputs,
        max_new_tokens=1200,
        generation_config=generation_config,
    )

/usr/local/lib/python3.10/dist-packages/torch/utils/checkpoint.py:87: UserWarning: None of the inputs have requires_grad=True. Gradients will be None
  warnings.warn(

[ ] generate_ids = generate_ids[:, inputs['input_ids'].shape[1] :]

    response = processor.batch_decode(
        generate_ids, skip_special_tokens=True, clean_up_tokenization_spaces=False
    )[0]

[ ] response

'Welcome to Ignite. Today I want to focus on AI and this transformational power as it drives growth in business. It improves efficiency, it improves operating leverage. And
to do that, we are building out three platforms, Co-Pilot, Co-Pilot Devices, and Co-Pilot and AI Stack. Co-Pilot is the UI for AI. It's rapidly becoming an organizing layer
for work and how work gets done. Every employee will have a Co-Pilot that knows them, their work, helping them unlock productivity, enhancing creativity, and saving time. A
nd Co-Pilot Studio will allow you to create agents that automate business processes.'
```

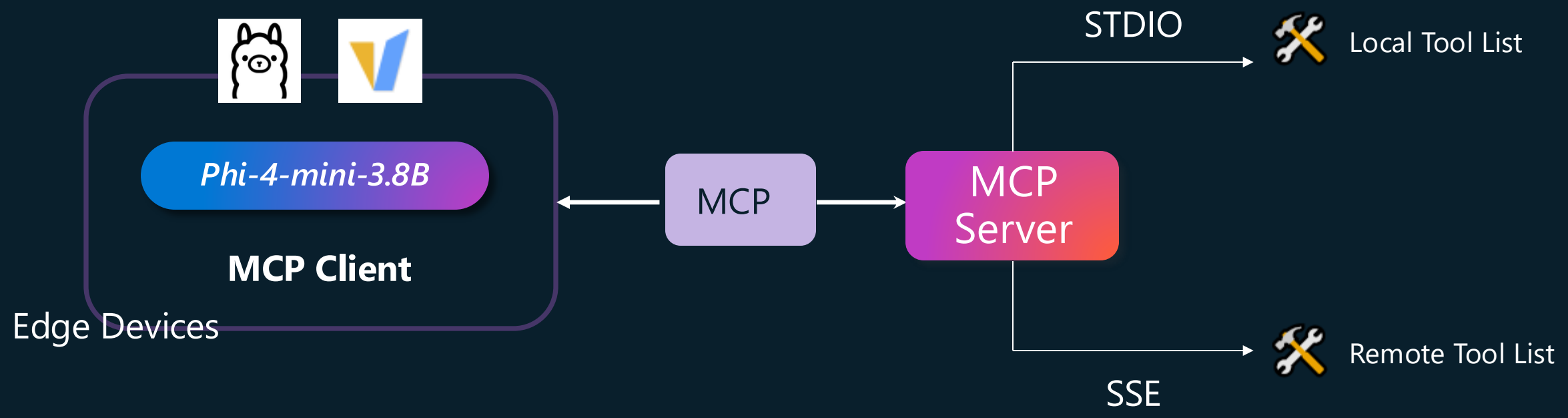
Multi Language - Phi-4-mulitimodal

https://github.com/kinfey/PhiCookbook/blob/main/md/02.Application/05.Audio/Phi4/Translate/demo.ipynb

OCR - Phi-4-mulitimodal

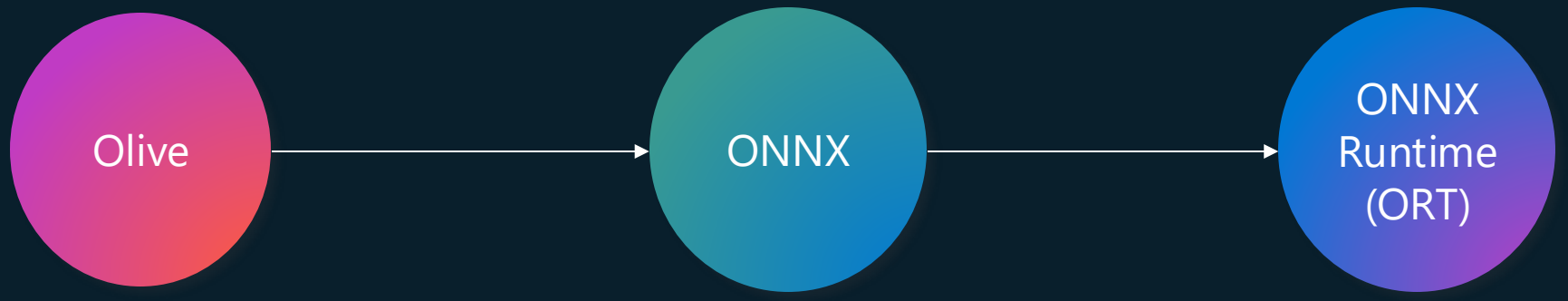


Phi-4-mini MCP Client in Edge AI



The ONNX Trilogy

an E2E solution for On-Device AI



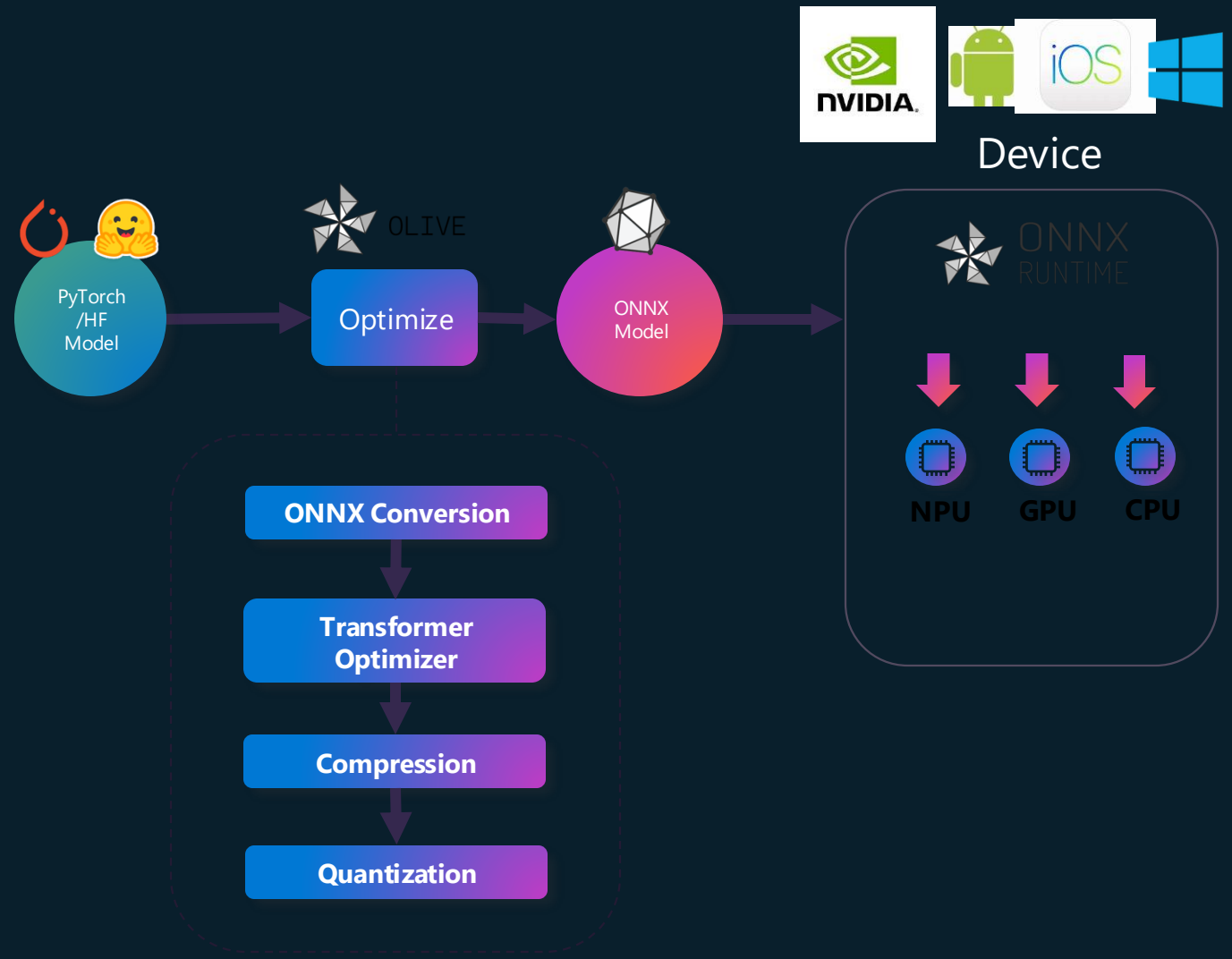
Toolchain to create optimized ONNX models

An open and interoperable file format for ML and DNN models.

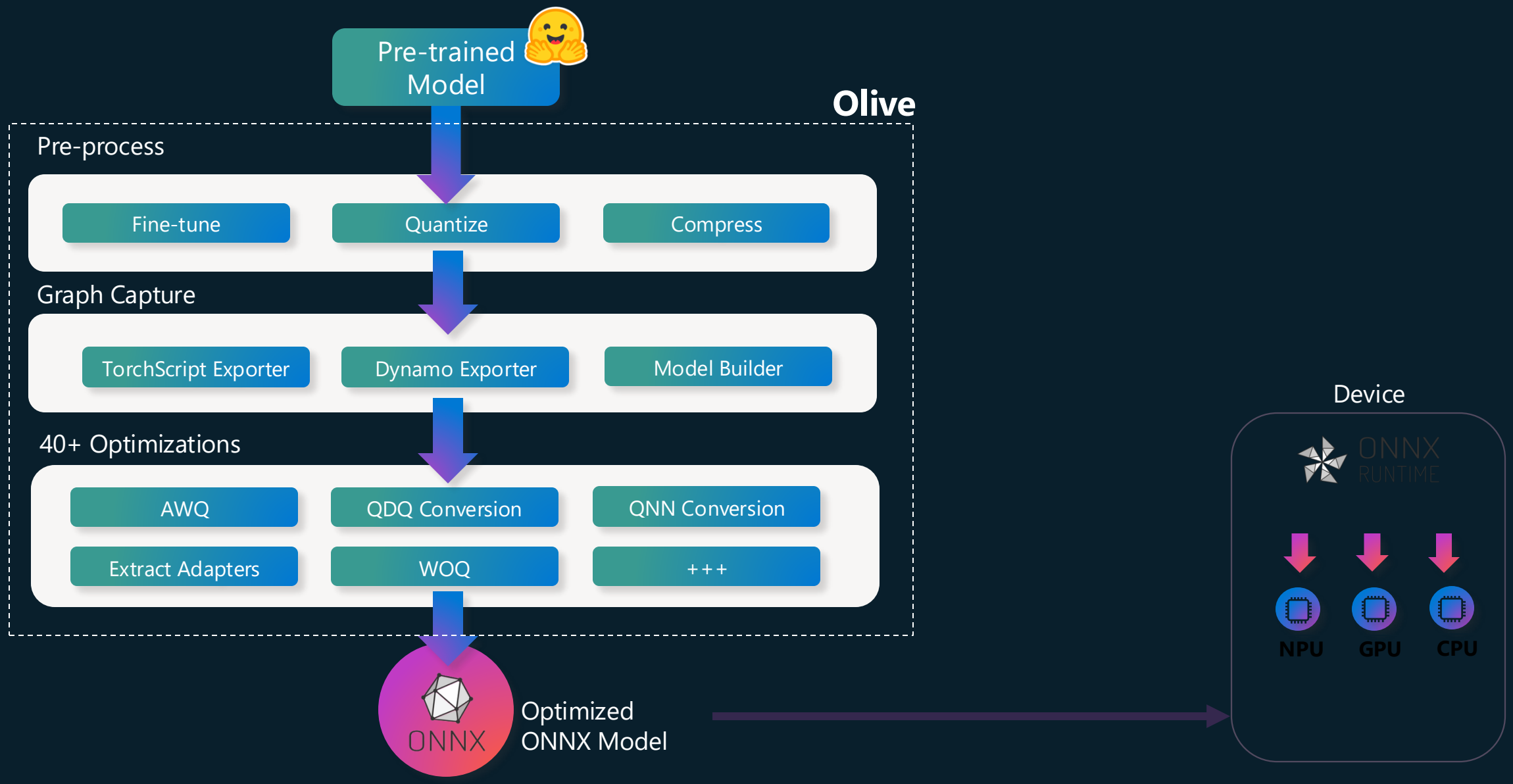
Fast and efficient model inference engine that runs on a variety of hardware accelerators and platforms.

What is Olive?

- **O(NNX)Live** –AI model optimization toolkit for ONNX Runtime.
- Provide AI engineers with an easy-to-use and integrated toolchain where the following tasks can be combined into workflows (pipelines):
  - Finetuning (QLoRA/LoRA/LoftQ)
  - Graph Capture (Dynamo Exporter/Model Builder)
  - Model Optimization (CPU/GPU/NPU/DirectML)
  - Quantization (GPTQ, AWQ, WOQ)
  - Runtime tuning
  - Creates models for Multi-LoRA serving
  - Deployment
- **Automatically** Automatically find the optimized model for a given hardware target without expertise and manual work.



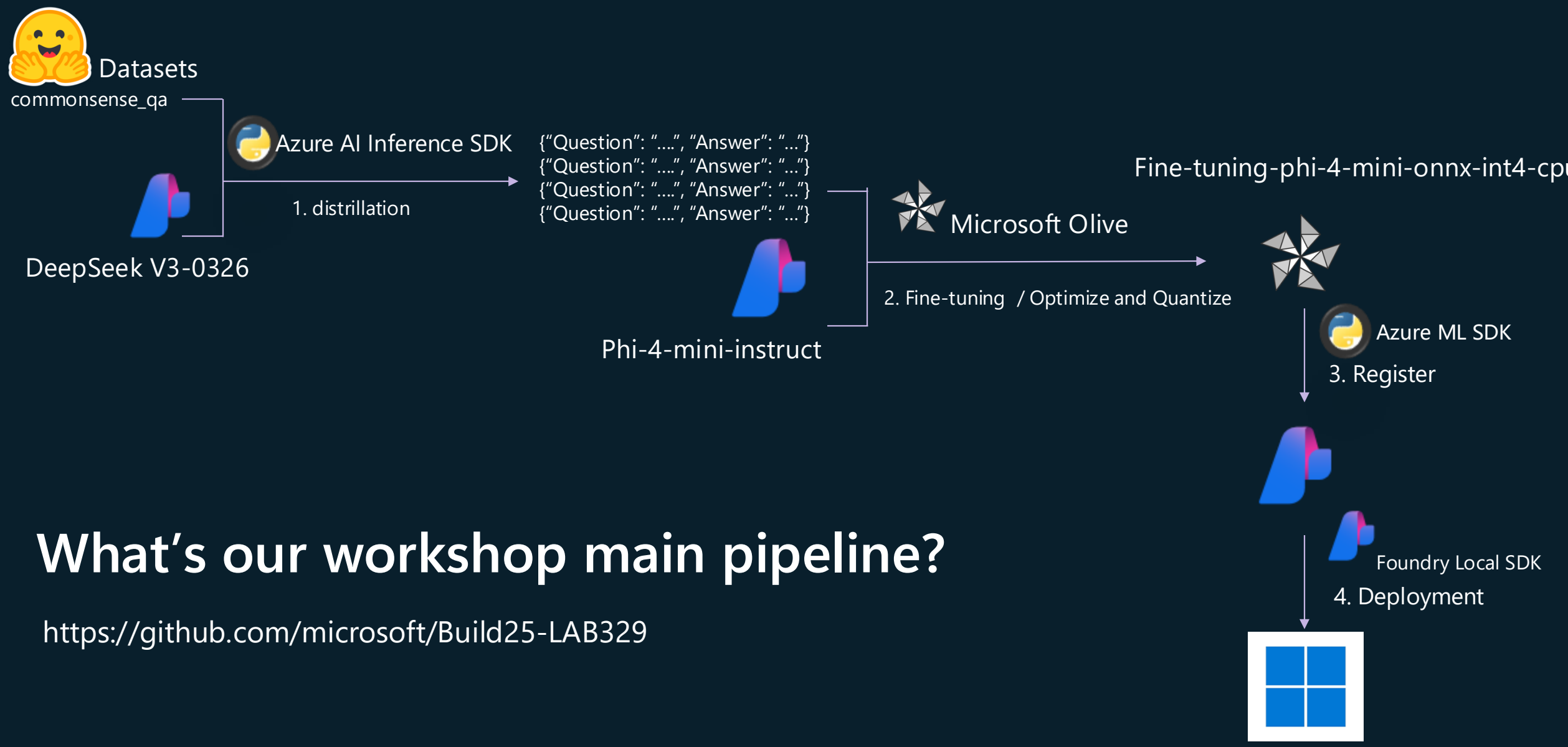
The anatomy of an Olive workflow



$> Olive

Simplified CLI for common model optimization tasks

Command	Description	Executed Olive Workflow
`auto-opt`	Automatically optimize a PyTorch model into ONNX with optional quantization.	Model ➡️ OnnxConversion | ModelBuilder ➡️ OrtTransformersOptimization ➡️ [Quantization] ➡️ OrtPerfTuning
`quantize`	Quantize a PyTorch or ONNX model using algorithms such as AWQ, QuaRoT, GPTQ, RTN and more.	Model ➡️ [OnnxConversion | ModelBuilder] ➡️ Quantization
`finetune`	Finetune a model on a dataset using techniques like LoRA and QLoRA.	Model ➡️ LoRA | QLoRA
`capture-onnx-graph`	Capture the ONNX graph from a Hugging Face or PyTorch model.	Model ➡️ OnnxConversion | ModelBuilder
`generate-adapter`	Extract the adapter weights from an ONNX model and store as an external weights file for ORT.	Model ➡️ ExtractAdapters
`convert-adapters`	Convert existing (Q)LoRA adapter weights to a weights file for ORT.	Adapter Weights ➡️ ORT Adapter Weights
`run`	Run the supported 40+ optimization passes in the sequence you wish.	Defined by AI Engineer in YAML/JSON.



What's our workshop main pipeline?

https://github.com/microsoft/Build25-LAB329

Resources

Microsoft Phi Cookbook
https://aka.ms/Phicookbook

Microsoft Phi-4-multimodal techreport
https://aka.ms/phi-4-multimodal/techreport

Microsoft Phi-4 Paper
https://arxiv.org/abs/2412.08905

Microsoft

Thank you!