Microsoft

Generative AI with JavaScript

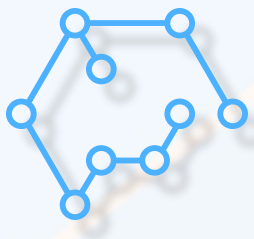
Run AI models on your local machine with Ollama

Why local models?

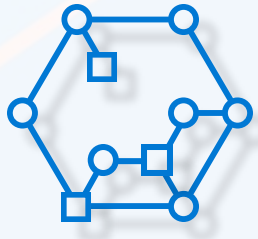
- **Experimentation at no cost**
- **Faster inner loop**
- **Offline development**
- **Testing and evaluation**

Phi-3 models

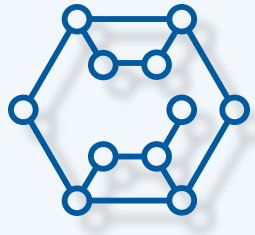
Groundbreaking performance for size, with frictionless availability



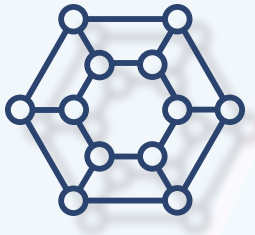
Phi-3-mini
(3.8B)



Phi-3-vision
(4.2B)



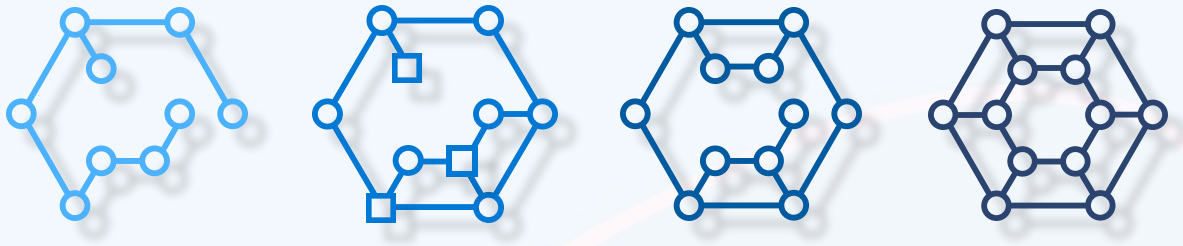
Phi-3-small
(7B)



Phi-3-medium
(14B)

Phi-3 models

Groundbreaking performance for size, with frictionless availability



Phi-3-mini
(3.8B)

Phi-3-vision
(4.2B)

Phi-3-small
(7B)

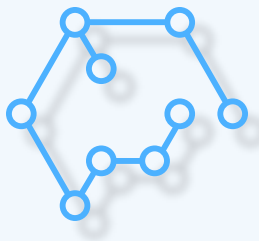
Phi-3-medium
(14B)

Instruction Tuned

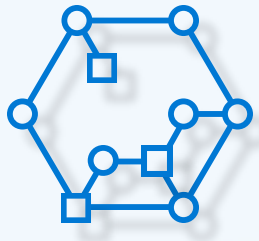
RAI Safety Aligned

Phi-3 models

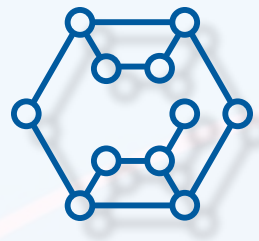
Groundbreaking performance for size, with frictionless availability



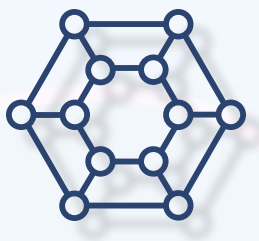
Phi-3-mini
(3.8B)



Phi-3-vision
(4.2B)



Phi-3-small
(7B)



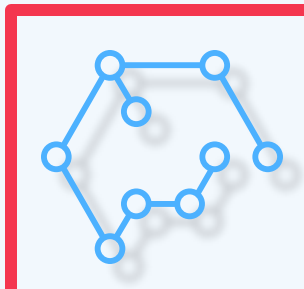
Phi-3-medium
(14B)

Instruction Tuned

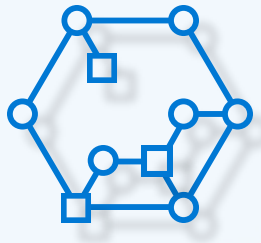
RAI Safety Aligned

Phi-3 models

Groundbreaking performance for size, with frictionless availability

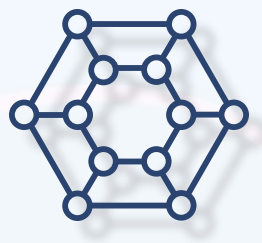


Phi-3-mini
(3.8B)



Phi-3-vision
(4.2B)

Phi-3-small
(7B)



Phi-3-medium
(14B)

Instruction Tuned

RAI Safety Aligned

Available on

Azure AI
Model Catalog

Hugging Face

ONNX Runtime

NVIDIA NIM

Ollama

LM Studio

VS Code AI Toolkit

Ollama

Open-source tool based on llama.cpp server

Run LLMs/SLMs on your local machine

Download models directly

OpenAI-compatible API

Get started with Ollama + Phi-3

Demo

Microsoft

Resources

- Phi-3 cookbook aka.ms/phi3/cookbook
- AI JavaScript playground aka.ms/ai/js/play
- Ollama website ollama.com