UNIVERSITATEA LUCIAN BLAGA DIN SIBIU
HPI Hasso Plattner Institut Digital Engineering · Universität Potsdam
CLUJ IT
Sibiu IT Cluster
SID 2025
Sibiu Innovation Days
06-07 November, Sibiu - RO
EMERGING DISRUPTIVE TECHNOLOGIES:
Balancing Innovation, Risks, and Societal Impact

Taking Back Control: The Rise of Local AI

From cloud dependence to local autonomy

Daniel Ursu
Senior Software Engineer, BearingPoint

The Cloud Giants

OpenAI

GPT-5
GPT-5-Codex

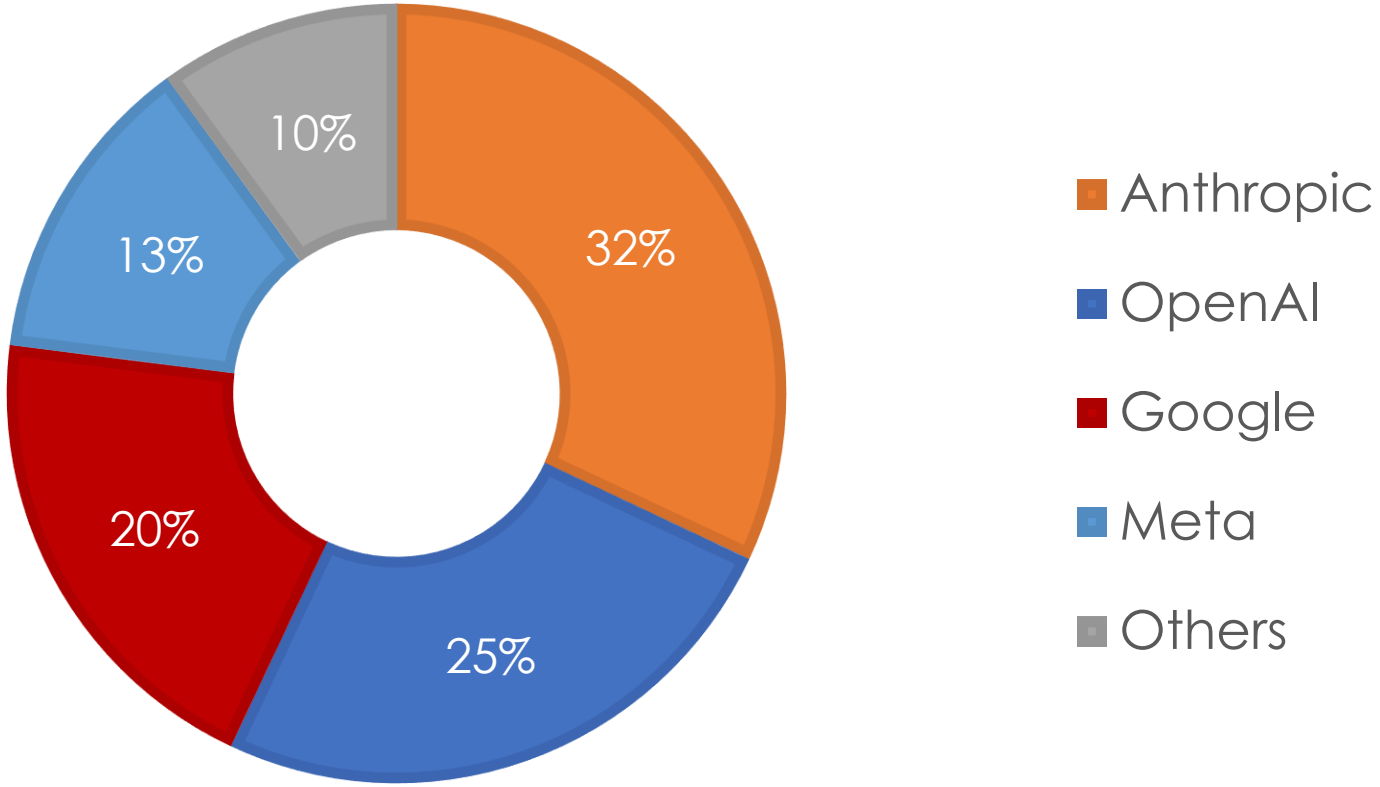
Anthropic

Claude Sonnet 4.5
Claude Opus 4.1

Google

Gemini 2.5 Pro
Gemini 2.5 Flash

Enterprise LLM API Market Share by Usage, Mid-2025



Report by Menlo Ventures

CLOUD-ONLY AI: Understanding The Trade-offs

Data Sovereignty

Your sensitive data lives on external infrastructure with complex compliance requirements

AI Transparency

Decision-making happens in opaque systems with limited visibility into logic or bias handling

Strategic Autonomy

Success depends on external providers' roadmaps, pricing, and policy changes

Shifting to Local LLMs: Why?

Privacy

Your data never leaves your device

Transparency

Understand exactly how AI makes decisions

Control

Customize models to suit your needs

No costs

No usage fees. No API subscriptions

How is it possible now?

Better Hardware

- Apple Silicon (M-series chips)
- Powerful GPUs available locally
- NVidia DGX Spark

Optimized Models

- Quantization
- Lightweight models designed for edge computing

Developer-Friendly Tools

- Ollama / LM Studio: Easy model management
- Llama.cpp / MLX: Fast, efficient inference engines

How is it possible now?

Google Gemini

How is it possible now?

Using similar architecture

- Small
- Open source/weights
- Smart enough
- Fine-tunable

Google Gemini

Google Gemma

Who's using local AI today?

Healthcare

HIPAA-compliant analysis
Protected patient data

Financial Services

Secure document processing
Internal compliance checks

Software Development

Private code review
Internal documentation

Research & Education

Sensitive data analysis
Student data privacy

Top Local LLMs You Can Use Today

Qwen

- Qwen3: 1.7B - 30B (+ Thinking)
- Qwen3 Coder 30B

Mistral

- Mistral Small 3.2
- Magistral
- Devstral

Google

- Gemma 3 (1B - 27b)
- Gemma 3n (optimized for phones)

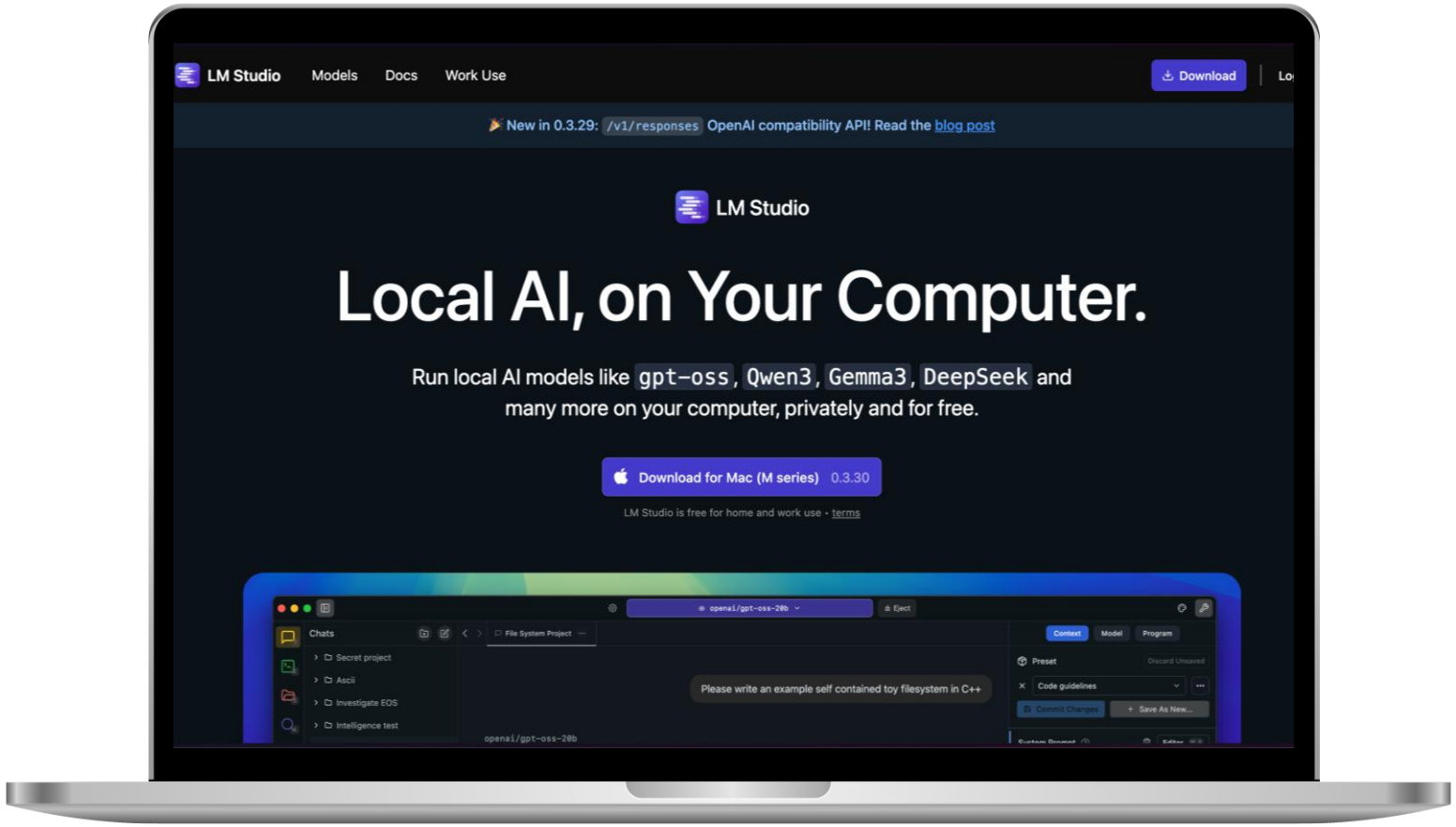
OpenAI

- gpt-oss-20b, different reasoning efforts

All models are available via HuggingFace, LM Studio or Ollama

Getting Started: Using Local AI in 5 minutes

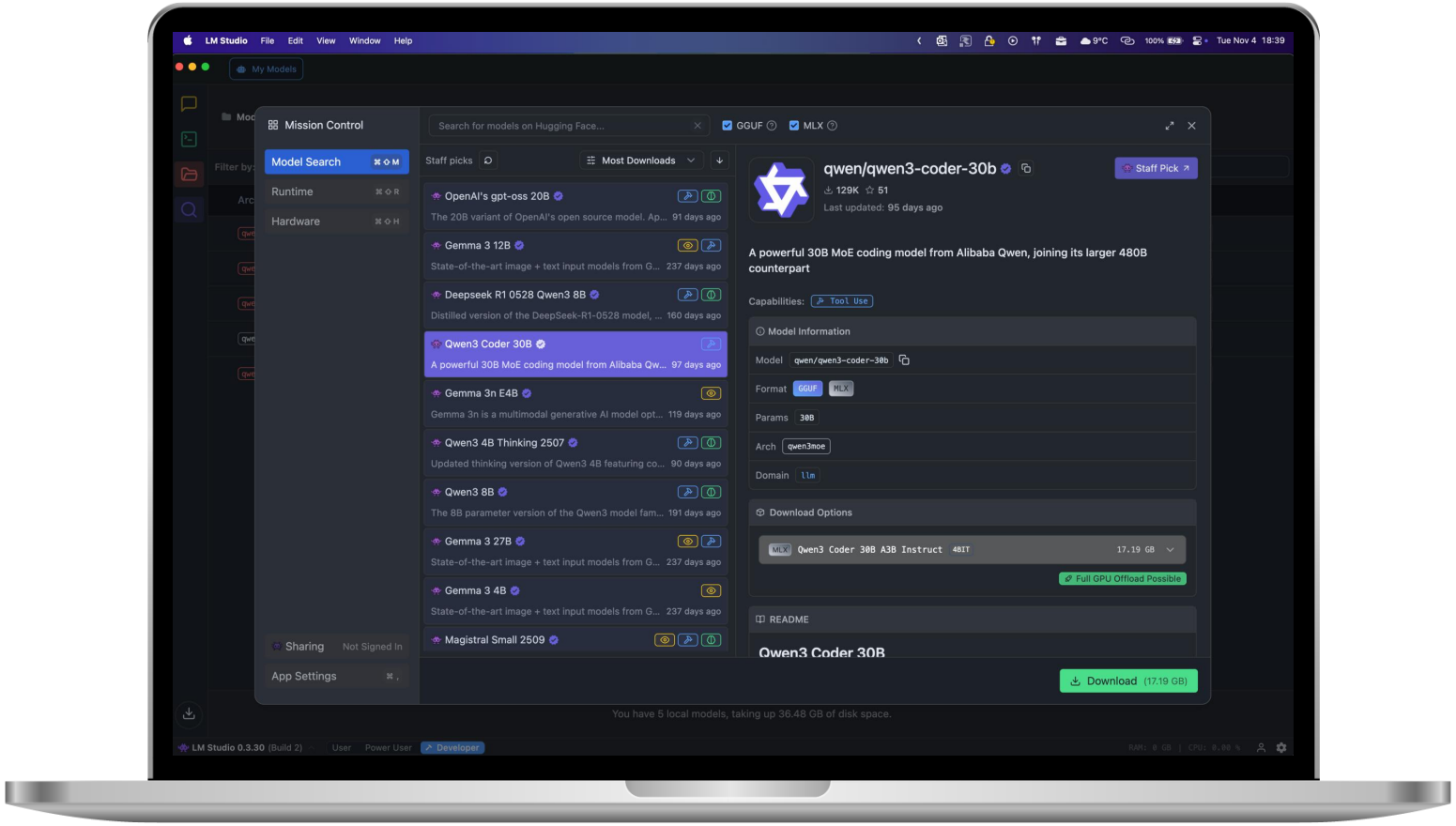
Download LM Studio app



Getting Started: Using Local AI in 5 minutes

Download model

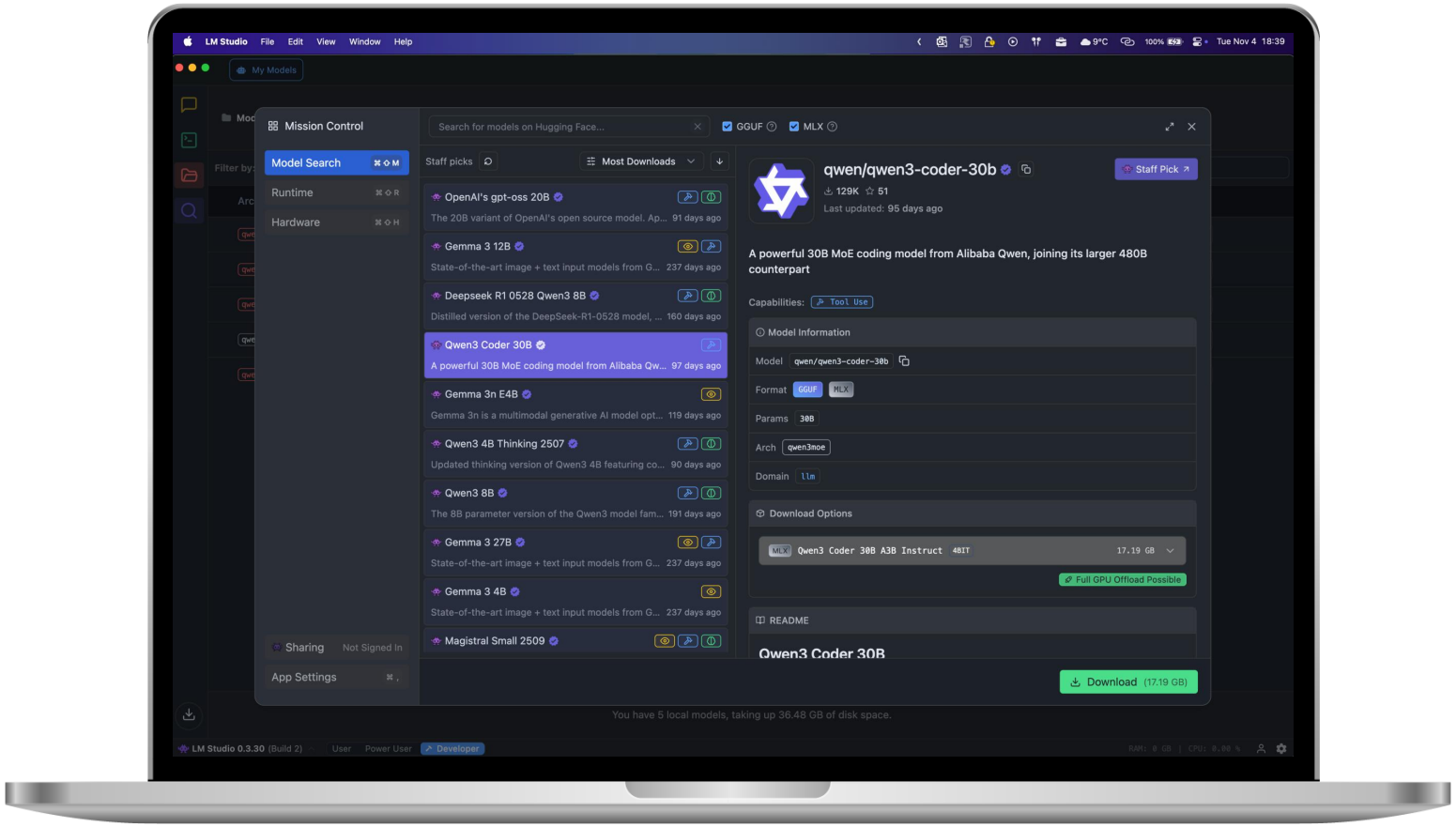
The models are fetched directly from Hugging Face

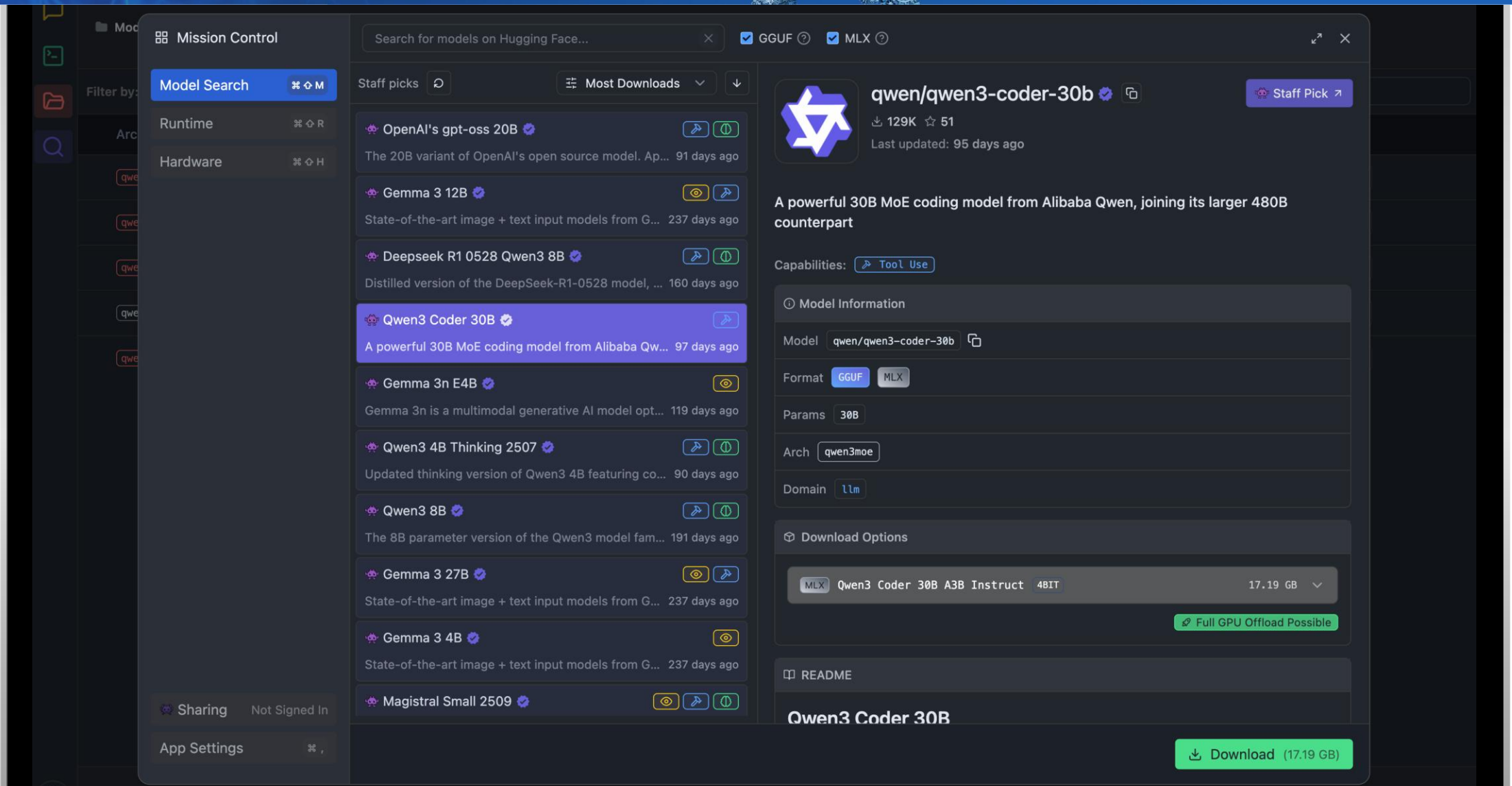


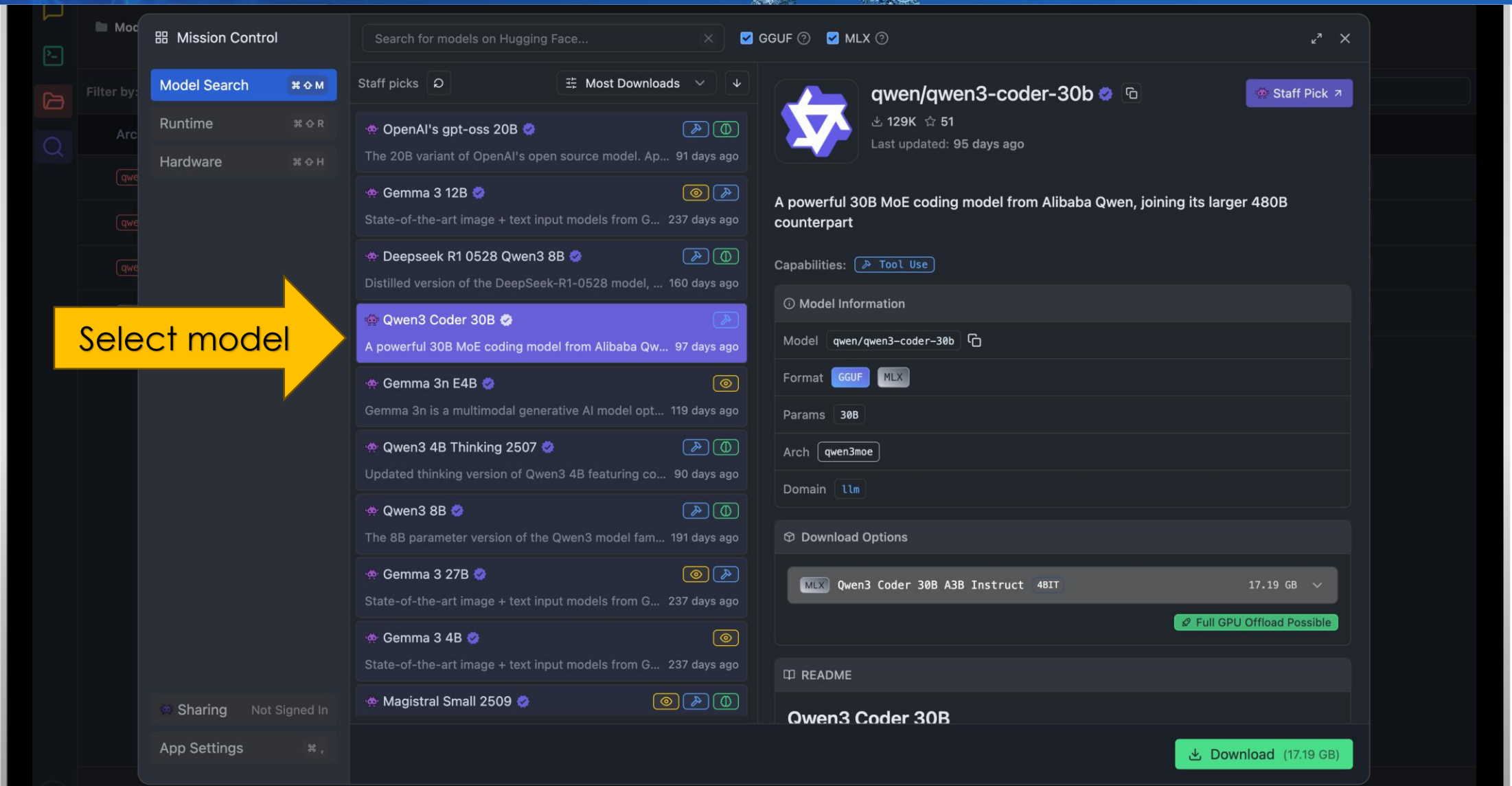
Getting Started: Using Local AI in 5 minutes

Download model

The models are fetched directly from Hugging Face




Mission Control
Model Search ⌘⇧M
Runtime ⌘⇧R
Hardware ⌘⇧H
Sharing Not Signed In
App Settings ⌘,
Search for models on Hugging Face...
GGUF
MLX
Staff picks
Most Downloads
OpenAI's gpt-oss 20B
The 20B variant of OpenAI's open source model. Ap... 91 days ago
Gemma 3 12B
State-of-the-art image + text input models from G... 237 days ago
Deepseek R1 0528 Qwen3 8B
Distilled version of the DeepSeek-R1-0528 model, ... 160 days ago
Qwen3 Coder 30B
A powerful 30B MoE coding model from Alibaba Qw... 97 days ago
Gemma 3n E4B
Gemma 3n is a multimodal generative AI model opt... 119 days ago
Qwen3 4B Thinking 2507
Updated thinking version of Qwen3 4B featuring co... 90 days ago
Qwen3 8B
The 8B parameter version of the Qwen3 model fam... 191 days ago
Gemma 3 27B
State-of-the-art image + text input models from G... 237 days ago
Gemma 3 4B
State-of-the-art image + text input models from G... 237 days ago
Magistral Small 2509
qwen/qwen3-coder-30b
129K 51
Last updated: 95 days ago
Staff Pick
A powerful 30B MoE coding model from Alibaba Qwen, joining its larger 480B counterpart
Capabilities: Tool Use
Model Information
Model qwen/qwen3-coder-30b
Format GGUF MLX
Params 30B
Arch qwen3moe
Domain llm
Download Options
MLX Qwen3 Coder 30B A3B Instruct 4BIT 17.19 GB
Full GPU Offload Possible
README
Qwen3 Coder 30B
Download (17.19 GB)


Mission Control
Model Search ⌘⇧M
Runtime ⌘⇧R
Hardware ⌘⇧H
Sharing Not Signed In
App Settings ⌘,
Search for models on Hugging Face...
GGUF
MLX
Staff picks
Most Downloads
OpenAI's gpt-oss 20B
The 20B variant of OpenAI's open source model. Ap... 91 days ago
Gemma 3 12B
State-of-the-art image + text input models from G... 237 days ago
Deepseek R1 0528 Qwen3 8B
Distilled version of the DeepSeek-R1-0528 model, ... 160 days ago
Qwen3 Coder 30B
A powerful 30B MoE coding model from Alibaba Qw... 97 days ago
Gemma 3n E4B
Gemma 3n is a multimodal generative AI model opt... 119 days ago
Qwen3 4B Thinking 2507
Updated thinking version of Qwen3 4B featuring co... 90 days ago
Qwen3 8B
The 8B parameter version of the Qwen3 model fam... 191 days ago
Gemma 3 27B
State-of-the-art image + text input models from G... 237 days ago
Gemma 3 4B
State-of-the-art image + text input models from G... 237 days ago
Magistral Small 2509
Select model
qwen/qwen3-coder-30b
129K 51
Last updated: 95 days ago
Staff Pick
A powerful 30B MoE coding model from Alibaba Qwen, joining its larger 480B counterpart
Capabilities: Tool Use
Model Information
Model qwen/qwen3-coder-30b
Format GGUF MLX
Params 30B
Arch qwen3moe
Domain llm
Download Options
MLX Qwen3 Coder 30B A3B Instruct 4BIT 17.19 GB
Full GPU Offload Possible
README
Qwen3 Coder 30B
Download (17.19 GB)

Mission Control

Model Search ⌘⇧M

Runtime ⌘⇧R

Hardware ⌘⇧H

Sharing Not Signed In

App Settings ⌘,

Search for models on Hugging Face...

GGUF MLX

Staff picks

Most Downloads

OpenAI's gpt-oss 20B
The 20B variant of OpenAI's open source model. Ap... 91 days ago

Gemma 3 12B
State-of-the-art image + text input models from G... 237 days ago

Deepseek R1 0528 Qwen3 8B
Distilled version of the DeepSeek-R1-0528 model, ... 160 days ago

Qwen3 Coder 30B
A powerful 30B MoE coding model from Alibaba Qw... 97 days ago

Gemma 3n E4B
Gemma 3n is a multimodal generative AI model opt... 119 days ago

Qwen3 4B Thinking 2507
Updated thinking version of Qwen3 4B featuring co... 90 days ago

Qwen3 8B
The 8B parameter version of the Qwen3 model fam... 191 days

Gemma 3 4B
State-of-the-art image + text input models from G... 237 days ago

Magistral Small 2509

Click on Download Options

qwen/qwen3-coder-30b

Staff Pick

129K ☆ 51

Last updated: 95 days ago

A powerful 30B MoE coding model from Alibaba Qwen, joining its larger 480B counterpart

Capabilities: Tool Use

Model Information

Model qwen/qwen3-coder-30b

Format GGUF MLX

Params 30B

Arch qwen3moe

Domain llm

Download Options

MLX Qwen3 Coder 30B A3B Instruct 4BIT 17.19 GB

Full GPU Offload Possible

README

Qwen3 Coder 30B

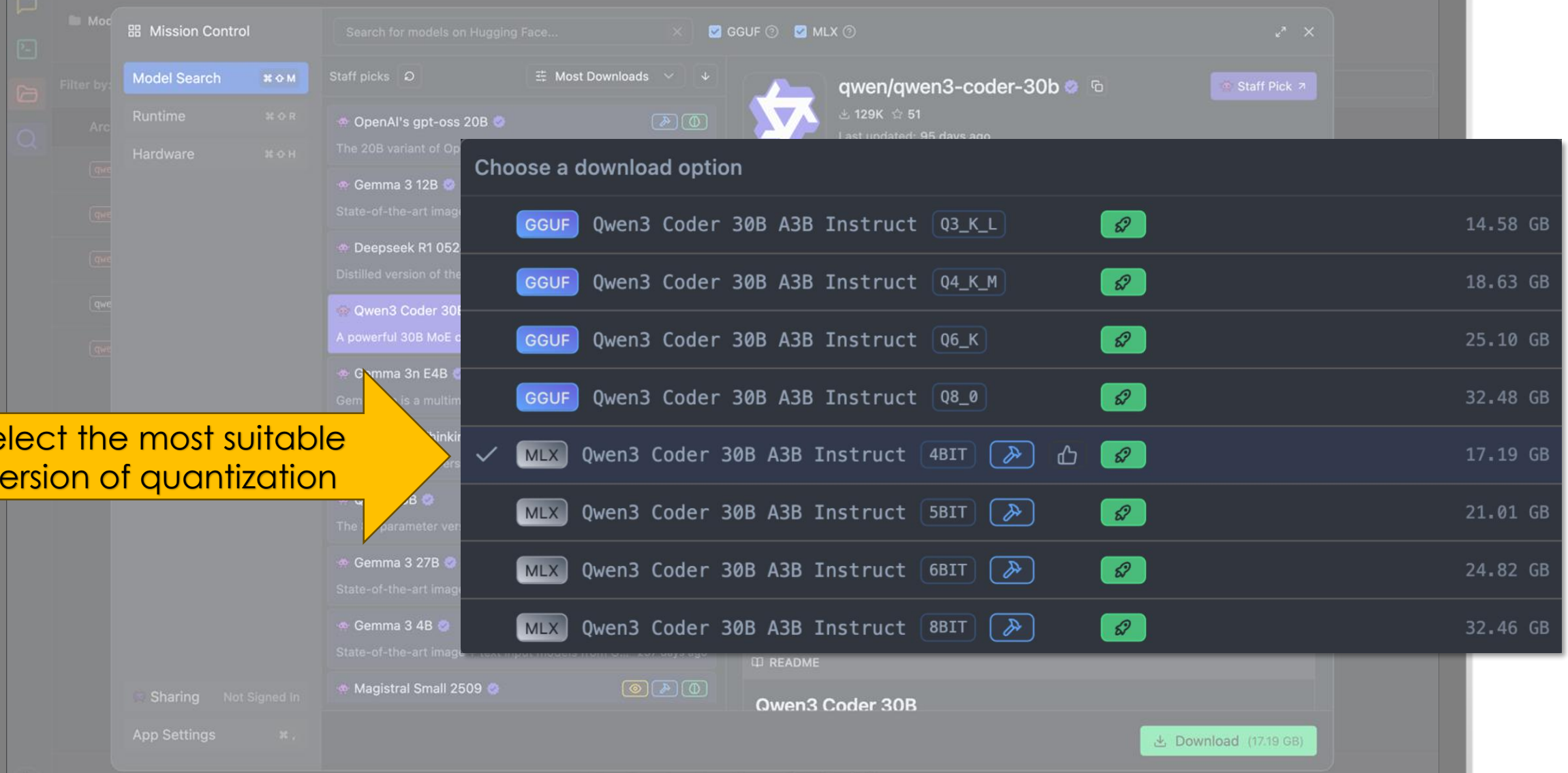
Download (17.19 GB)

SID 2025

Sibiu Innovation Days

06-07 November, Sibiu - RO

EMERGING DISRUPTIVE TECHNOLOGIES:
Balancing Innovation, Risks, and Societal Impact



Select the most suitable version of quantization

Mission Control

Model Search ⌘⇧M

Runtime ⌘⇧R

Hardware ⌘⇧H

Sharing Not Signed In

App Settings ⌘,

Search for models on Hugging Face...

GGUF MLX

Staff picks | Most Downloads

- OpenAI's gpt-oss 20B — The 20B variant of OpenAI's open source model. Ap... 91 days ago
- Gemma 3 12B — State-of-the-art image + text input models from G... 237 days ago
- Deepseek R1 0528 Qwen3 8B — Distilled version of the DeepSeek-R1-0528 model, ... 160 days ago
- Qwen3 Coder 30B — A powerful 30B MoE coding model from Alibaba Qw... 97 days ago
- Gemma 3n E4B — Gemma 3n is a multimodal generative AI model opt... 119 days ago
- Qwen3 4B Thinking 2507 — Updated thinking version of Qwen3 4B featuring co... 90 days ago
- Qwen3 8B — The 8B parameter version of the Qwen3 model fam... 191 days ago
- Gemma 3 27B — State-of-the-art image + text input models from G... 237 days ago
- Gemma 3 4B — State-of-the-art image + text input models from G... 237 days ago
- Magistral Small 2509

qwen/qwen3-coder-30b

Staff Pick

129K ☆ 51

Last updated: 95 days ago

A powerful 30B MoE coding model from Alibaba Qwen, joining its larger 480B counterpart

Capabilities: Tool Use

Model Information

Model	qwen/qwen3-coder-30b
Format	GGUF MLX
Params	30B
Arch	qwen3moe
Domain	llm

Download Options

MLX Qwen3 Coder 30B A3B Instruct 4BIT 17.19 GB

Full GPU Offload Possible

README

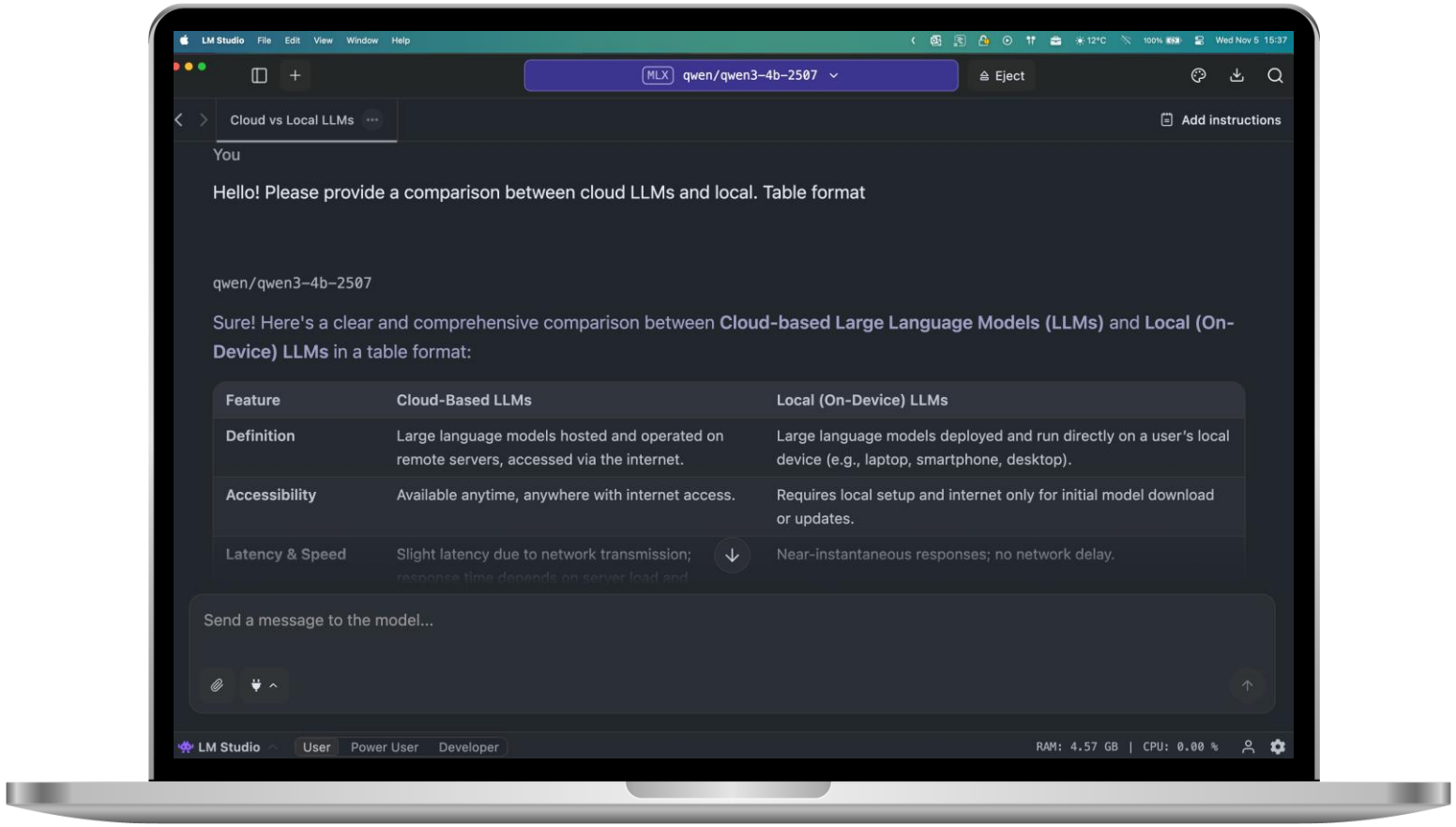
Download

Download (17.19 GB)

Getting Started: Using Local AI in 5 minutes

Chat with the model

You do not need internet connection



Challenges and The Future Ahead

Current Challenges

- Resource constraints (memory, GPU compute)
- Quality limitations compared to largest cloud models
- Fragmentation of ecosystems

Future Opportunities

- Hybrid architectures (edge + selective cloud usage)
- Growing open-source community innovations

Vision

- Democratization of AI: more accessible, customizable, ethical.

Key Takeaways

- Local LLMs increase privacy, fairness, transparency.
- Technology advancements make local AI practical today.
- Hybrid AI architectures will combine the best of both worlds.

> "Local AI shifts power from big tech to individual creators, developers, and organizations."

Thank You!

Let's Discuss.

Daniel Ursu
Senior Software Engineer, BearingPoint