

Wonderful World of Open-Weights AI

Fully Local, Private, Powerful Models

Presented by: Bishal Sapkota

 bishal.io | @bishalspkt

GDG Melbourne | Build with AI



Outback Yak

Open-Weights models

Open-Weights models refers to AI models whose trained parameters (weights) are publicly accessible and available for download and modification.

This contrasts with closed, proprietary models offered via APIs, where the underlying model weights remain hidden.

Why Open-Weights models?

-  Privacy & Security
-  Customizable
-  Cost-effective
- No External Dependency



Ollama

Ollama is a simple and powerful tool that lets you run large language models (LLMs) locally, on your own hardware, with minimal setup.

```
ollama run gemma3:1b
```

Live Demo 







Open Web UI

OpenWebUI is an intuitive and user-friendly web interface designed to interact seamlessly with AI models like those served by Ollama.



Diverse Capabilities of Open Weights Models

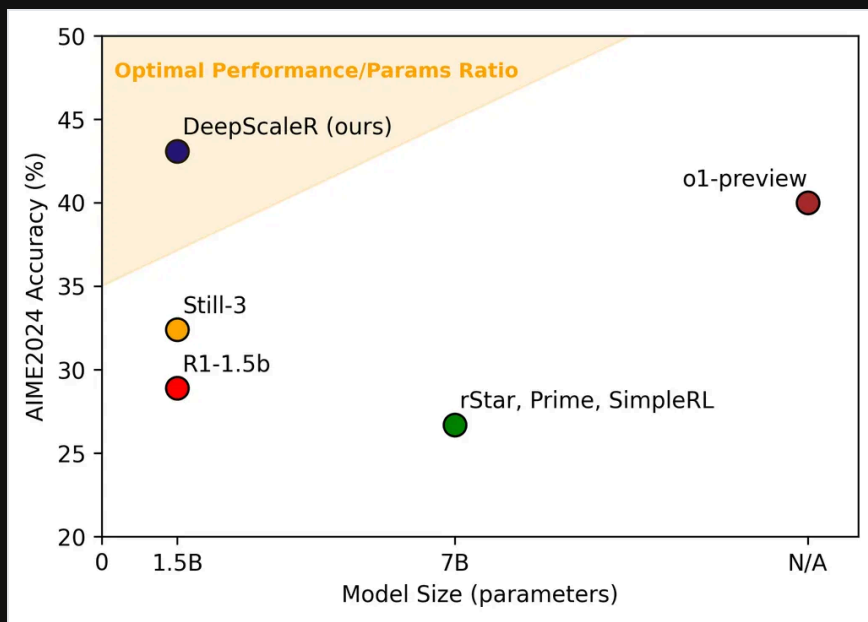
-  **DeepScaler (1.5B)**: Superior AIME performance compared to o1-preview
-  **Dolphin-Mistral**: Run uncensored models
-  **Image Processing**: Advanced visual tasks
-  **Agents with Tools**: Enhance LLMs using LangChain & LlamaIndex



Efficiency Spotlight: DeepScaler (1.5B)

■ Small yet Powerful

DeepScaler outperformed GPT-4o preview (larger model) on GSM8K mathematical reasoning benchmark



Uncensored Models

Uncensored AI models are language models (LLMs) that have minimal or no alignment or safety layers imposed on them.

Use Cases of uncensored models:

- Red teaming / security testing
- Creative Writing and Art (Explicit/Taboo Content)
- Political Analysis and Controversial Topics
- Academic Research and Unrestricted Data

```
ollama run dolphin-mistral
```



Image processing and OCR

You can run image processing and optical character recognition workloads using local models. This is something that used to take teams/organizations years to build to a degree of quality

New Chat

Workspace

Search

Chats

Today

Twelve Apostles Australia

Why the Sky is Blue

Racism & Bigotry


What are some great examples to

Yesterday

Invoice OCR Results

iPhone Evolution Timeline

gemma3:27b



What can you see in this image?

gemma3:27b

Here's a breakdown of what I can see in the image:

Landscape:




- Rock Formations:** The most prominent feature is the set of tall, weathered rock formations rising from the beach. These are the iconic "Twelve Apostles" (though there are fewer than twelve remaining) located along the Great Ocean Road in Victoria, Australia.
- Beach:** A wide, sandy beach stretches out in the foreground. The sand is a warm, reddish-brown color.
- Ocean:** The ocean is visible between the rock formations, with waves breaking on the shore.
- Sky:** The sky is cloudy and overcast, with a mix of gray and blue.

People:

- There are several people scattered along the beach, appearing quite small in comparison to the massive rock formations.



Building apps with Open Weights models




-  Private Financial Analysis
-  Local Knowledgebase Search (*Demo*)
-  File & Directory Auto-Organization



Demo time



Hands-On Demo: Deploying on Cloud Run (GPU)

-  Google Cloud Run
-  GPU: Nvidia T4/A100
-  Persistent Data with Google Cloud Storage mounted at `/root/.ollama`

```
FROM ollama/ollama:latest
ENV OLLAMA_GPU=true
CMD ["serve", "gemma:7b"]
```



Advanced Opportunities

-  **Fine-Tuning** (*LoRA*)
-  **Building Agents** (*LangChain, Langsmith, Vercel AI SDK, OpenAI Agents SDK*)



Recap & Call to Action

-  Privacy
-  Easy Setup
-  Efficient & Powerful (DeepScaler)

Start building locally!



Q&A



Let's discuss your questions!

