Overview of Leading Open Source AI Models

Open Source AI Models Overview

- Meta's Llama series highlighted as leading open source models
- Meta open sourced Llama to differentiate from OpenAI and Anthropic
- Llama 4 is Meta's largest open source model; Llama 3.2 notable for small, local versions (1B, 3B parameters)
- SLMs (small language models) term used for smaller models like Llama 3.2
- Mistral (French company) offers Mixtral, a mixture of experts model, routes queries to specialist sub-models
- Alibaba Cloud's Qwen models praised for power, less known than Llama, recommended for use
- Gemma is Google's open source cousin to Gemini, available in very small sizes (down to 270M parameters)
- Microsoft's Phi series (Phi-4 latest) noted for tool use and commercial applications

Deepseek and OpenAI's Open Source Moves

- Deepseek AI's model gained attention for high capability at low training cost ($4M vs. OpenAI's $100M+)
- Deepseek main model has 671B parameters, too large for local use; smaller variants available
- Smaller Deepseek variants are actually distilled versions of Llama and Qwen, trained with synthetic data from Deepseek
- Distillation process used to create efficient, smaller models for local use (e.g., via Ollama)
- OpenAI recently released GPT OSS (open source GPT), possibly in response to Deepseek
- GPT OSS available in 20B and 120B parameter sizes; 20B can be run locally, 120B is much larger

Ways to Use Language Models

- Three main usage methods: packaged products, cloud APIs, and local inference
- Packaged products (e.g., ChatGPT, Claude) offer user interfaces and built-in features
- Cloud APIs allow direct calls to models hosted remotely (OpenAI, Amazon Bedrock, Google Vertex, Azure ML)
- Platforms like OpenRouter route requests to multiple model providers
- Some APIs (e.g., Groq with a Q) offer fast cloud inference using proprietary hardware

Running Open Source Models Locally

- Open source models can be downloaded and run directly on personal computers
- Hugging Face Transformers library: run models in Python, direct access to code and weights
- Ollama: packaged product for easy, efficient local model inference, uses optimized code and compressed weights (GGUF files)
- Ollama supports a limited set of models, provides a local API for easy integration
- Main difference: Hugging Face is more flexible, Ollama is more user-friendly and efficient for supported models

Next Steps

- Plan to practice local inference using Ollama and Hugging Face
- Upcoming hands-on lab to reinforce concepts and explore model behavior in practice