**1. Donanım Gereksinimleri**

|  |  |  |  |
| --- | --- | --- | --- |
| Model Boyutu | Minimum GPU | RAM (sistem) | Depolama |
| < 2B | CPU / 4GB GPU | 8 GB | 2-4 GB |
| 7B | 8-12 GB GPU | 16 GB | 8-16 GB |
| 13B | 24 GB GPU | 32 GB | 20+ GB |
| 30B+ | Çoklu GPU / A100 / H100 gibi | 64 GB+ | 40-80+ GB |

### ****Ollama modelleri (sıkıştırılmış) versiyonları indiriyor****

* Ollama, DeepSeek R1 8B gibi büyük modelleri.gguf formatında ve **düşük bit'li** (ör. **4-bit**, **5-bit**) sürümleriyle indiriyor.
* Bu sürümler, modelin hassasiyetini az bir miktar düşürüp **VRAM kullanımını büyük ölçüde azaltır.**

### ****Ollama'nın “offloading” sistemi****

* Ollama, GPU VRAM'i yeterli değilse bazı ağırlıkları CPU veya RAM'e offload eder.
* Örneğin 6 Gb GPU 16 GB RAM olan bir sistemde 8gb vram harcanması gerekirse VRAM eksik olduğu için eksik VRAM kısmı RAM’den karşılanıyor.
* Bu da modeli biraz yavaşlatır ama çalışmasını sağlar.

**EC2 Instance on AWS Nedir?**

**EC2 (Elastic Compute Cloud)**, Amazon Web Services (AWS)’in sunduğu **bulut tabanlı sanal sunucu** hizmetidir.

Yani **EC2 = Kiralık bilgisayar (sunucu)** demek.  
Bu bilgisayarı dilediğin gibi yapılandırırsın (Linux, Windows, GPU, CPU, RAM seçersin) ve **internetten uzaktan erişerek çalıştırırsın.**

## EC2 Instance Türleri

|  |  |  |
| --- | --- | --- |
| Tip | Kullanım Alanı | Örnek |
| t2.micro | Ücretsiz deneme (low power) | Basit testler |
| m5.large | Genel kullanım | Web sunucusu |
| c6g | Compute-intensive | Video işleme |
| g4dn.xlarge | **GPU destekli!** | Yapay zeka, ML modelleri |
| p3.2xlarge | **Yüksek performanslı GPU** | LLM'ler, eğitim, inference |

## EC2 ile Neler Yapabilirsin?

* Hugging Face modellerini çalıştırabilirsin
* Open WebUI + Ollama kurabilirsin
* Kendi API servislerini barındırabilirsin
* LLM eğitimi / ince ayar yapabilirsin (GPU’lu instance gerek)
* Web siteni yayına alabilirsin (Flask, Django vs.)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Adı | Yayınlayan | Boyut | Lisans | Yetenekler | Donanım Gereksinimi | Öne Çıkan Özellik |
| LLaMA 3 | Meta | 8B / 70B | Ticari dışı açık | Kod, dil, mantık | 8B: 12 GB VRAM | En güncel ve güçlü modellerden biri |
| Mistral | Mistral AI | 7B | Apache 2.0 (serbest) | Hızlı dil üretimi | 6-8 GB VRAM | Flash attention ile çok hızlı |
| Mixtral | Mistral AI | 12.9B (MoE) | Apache 2.0 | Gelişmiş mantık ve kod | 10-12 GB VRAM | MoE yapısıyla hafif ama güçlü |
| DeepSeek Coder | DeepSeek AI | 1.3B / 6.7B / 8B | Apache 2.0 | Kodlama odaklı | 6-8 GB VRAM | Kod üretimi ve analizinde çok iyi |
| Phi-2 | Microsoft | 2.7B | MIT | Küçük, hızlı, doğal dil | 4 GB RAM ile çalışabilir | Eğitimi temiz, verimli küçük model |
| Gemma | Google | 2B / 7B | Apache 2.0 | Genel amaçlı LLM | 4-8 GB VRAM | Google destekli, sağlam |
| Command R+ | Cohere | 104B (API), 35B (açık) | Research-only | RAG + kod | GPU sunucu gerekir | Arama destekli konuşmalar için |
| GPT-J / GPT-NeoX | EleutherAI | 6B / 20B | MIT | Genel kullanım | 12-24 GB VRAM | Eski ama sağlam klasik |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Adı** | **Parametre** | **VRAM (FP16)** | **MoE** | **Özellik** |
| **Mixtral 8x7B** | 12.9B aktif | ~24-32 GB | ✅ | RAG için çok verimli, büyük bağlam |
| **LLaMA 3 70B** | 70B | ~48-64 GB | ❌ | En güçlü açık model |
| **Command R+** | ~35B | ~24-32 GB | ✅ | RAG odaklı eğitildi |
| **Yi 34B** | 34B | ~32-40 GB | ❌ | Çok dilli, yüksek doğruluk |
| **Zephyr-β / OpenChat** | 7B | ~12 GB | ❌ | Hafif ama sohbet odaklı |

|  |  |  |
| --- | --- | --- |
| Kullanım Amacı | Tavsiye Model | Neden |
| 🔤 Genel sohbet / yazı üretimi | **LLaMA 3 (8B)** veya **Mistral 7B** | Güçlü ve hızlı |
| 💻 Kodlama / Yazılım Asistanı | **DeepSeek Coder 6.7B** veya **Code LLaMA** | Kod tabanlı eğitilmiş |
| 🧠 Düşük sistem / hızlı yanıt | **Phi-2**, **Gemma 2B**, **TinyLLaMA** | Küçük ve hafif |
| 📚 RAG + bilgi tabanlı sohbet | **Mixtral**, **Command R+** | Arama destekli yapılar |
| 🌍 Çok dilli destek | **Yi**, **Falcon**, **LLaMA 3** | Çoklu dil verileriyle eğitilmiş |
| 🖥️ Yerel inference (offline kullanım) | **GGUF quantized modeller** (ör. Mistral 7B Q4) | Ollama / llama.cpp destekli |

## 📊 Örnek:

**Mixtral 8x7B** şu demek:

* İçinde **8 tane 7B parametreli uzman var**.
* Ama bir input geldiğinde sadece **2 tanesi aktif** olur.
* Yani **8×7B = 56B parametre var ama sadece 13B’si çalışıyor.**

✅ Yani **büyük modelin gücü**, **küçük modelin hızıyla** geliyor.

MoE = “Aynı model içinde çok sayıda uzman var ama her seferinde sadece en uygun olanlar çalışıyor.”  
→ Daha güçlü sonuç, daha düşük kaynak tüketimiyle.