DeepHPC

Jingjing Xu & Dr. Caesar Wu

PCOG

Self-Intro.

- 4th year PhD researcher*
- Research Intrest:
  - Transformer-based forecasting models
    - Transformer model (GPT: Generative Pre-training Transformer)
  - question answering system (a.k.a chatbot)

*Google scholar ; Github ; LinkedIn

Project - DeepHPC

- **Stage 1: Deploying DeepSeek / Mistral / Gemma2 locally** for answering questions
  - **Tools:** Models[1] & WebUI[2]

- **Stage 2: Create DeepHPC** by Fintuning DeepSeek / Mistral / Gemma2 (locally) for answering HPC questions
  - **Tools:** RAG[3] & Documents[4]

- **Stage 3: Deploying DeepHPC on HPC**

[1] Model; [2] WebUI tool; [3] RAG Tool; RAG: retrieval-augmented generation; [4] Documents

Project – Prerequisites[1]

Model Variant	Parameters (B)	VRAM Requirement (GB)	Recommended GPU Configuration
DeepSeek R1	671	~1,342	Multi-GPU setup (e.g., NVIDIA A100 80GB x16)
DeepSeek R1-Distill-Qwen-1.5B	**1.5**	**~0.7**	**NVIDIA RTX 3060 12GB or higher**
DeepSeek R1-Distill-Qwen-7B	**7**	**~3.3**	**NVIDIA RTX 3070 8GB or higher**
DeepSeek R1-Distill-Llama-8B	8	~3.7	NVIDIA RTX 3070 8GB or higher
DeepSeek R1-Distill-Qwen-14B	14	~6.5	NVIDIA RTX 3080 10GB or higher
DeepSeek R1-Distill-Qwen-32B	32	~14.9	NVIDIA RTX 4090 24GB
DeepSeek R1-Distill-Llama-70B	70	~32.7	NVIDIA RTX 4090 24GB (x2)

Reference:

[1] DeepSeek R1: Architecture, Training, Local Deployment, and Hardware Requirements

Project – Prerequisites[1]

- **Hardware Requirements:**
  - **RAM:** Minimum **16GB** recommended.
  - **Storage:** At least **20GB** of free space, preferably on an SSD for faster performance.
  - **GPU:** A dedicated GPU (e.g., NVIDIA RTX 3060 or higher) is recommended for optimal performance but not mandatory.

- **Software Requirements:**
  - **Operating System:** Linux/MacOS or Windows 10/11.
  - **Docker:** Ensure Docker is installed and running.
  - **Ollama:** A tool for managing and running AI models locally.

Reference:
[1] DeepSeek R1: Architecture, Training, Local Deployment, and Hardware Requirements