ICI Overview of AI/LLM Solutions, Dec 2024

AI and LLM

Artificial Intelligence (AI) encompasses a broader concept than Large Language Models (LLMs). LLMs gained significant attention in late 2022 with the debut of ChatGPT (GPT-3.5) on November 2022, followed by GPT-4's release in March 2023. For the average user, an LLM can be thought of as a prodigy with pre-trained, high-quality human knowledge. While this "prodigy" can quickly answer most questions, it cannot directly engage with additional knowledge, information, instructions, or regulations without specific enhancements.

Enhancing LLM Performance (2023-2024)

- **Increased Data and Hardware**: Utilizing more data, larger neural networks, and more parameters with additional hardware/AI cards for pre-training. This approach is nearing its limits and has already incorporated substantial synthetic data.
- **Chain of Thoughts**: Implementing step-by-step quality assessment using smaller parameter models, allowing LLMs to reason progressively and self-correct. Examples include OpenAI's GPT-4o, Kimi's reasoning model, and Perplexity's Pro mode.
- **RAG (Retrieval-Augmented Generation)**: Enabling LLMs to access external data in real-time, including internet search results or specific stored/incremental data (public or private). This approach aims to provide more targeted, real-time thinking/reasoning services for specific scenarios.
- **Fine-tuning/Post-training**: Optimizing pre-trained neural networks for specific situations, scenarios, industries, or domains. This process typically uses common inference devices/hardware and can be completed in a few days.

Optimal AI Utilization for Enterprises

For most businesses, the most practical AI implementation involves branching from mainstream AI engines, integrating local/specific data, and performing RAG/Fine-tuning operations to maximize LLM effectiveness.

Key Considerations for RAG and Fine-tuning

- RAG systems rely on data cleaning, standardization, and vector database construction.
- Fine-tuning requires preparing/purchasing specific post-training datasets. Specialized companies like Glaive, which created HyperWrite Reflection 70B, offer such datasets.

Recommended LLM Implementation Process for Enterprises

- Purchase public services (e.g., OpenAI, Perplexity advanced accounts) to familiarize with basic LLM usage.
- Integrate APIs into non-sensitive business systems to adapt to business interactions gradually.
- Set up RAG on cloud platforms (e.g., vast.ai) using non-sensitive data without modifying the original large model.
- For sensitive data or regulatory compliance, establish RAG in a private environment without modifying the original large model.
- Once the programming team is familiar with data cleaning and private dataset preparation, experiment with fine-tuning.

Cooling Considerations for High-Performance GPUs

Consider water/liquid cooling solutions for high-performance GPUs, especially in multi-GPU setups or for GPUs with very high-power consumption (e.g., over 400W). The decision should be based on specific models, usage scenarios, and overall system thermal design.

GPU Model Considerations for Inference and Fine-Tuning – Performance of Modern GPUs

The latest generation Tensor cores in GPUs, like the RTX 4090 with 4th generation Tensor cores, offers performance improvements 5 times over 3rd generation Tensor cores like in RTX 2080 series cards. While exact comparisons depend on specific use cases, these newer GPUs can provide substantial speedups for AI, machine learning, scientific calculation, and computer vision tasks. When choosing GPUs for AI tasks, always first consider late models like the RTX 4090, L40, L40S, or RTX 6000 Ada, all carrying the same AD102 AI/graphic chip, to achieve better tensor performance per watt and dollar.

4090 Power Module Caveat

Early RTX 4090 models experienced some power connector issues, which were addressed in later revisions.