Towards Compliant and Sovereign LLM Inference in Enterprise Environments

François Costa

Abstract

Despite impressive progress in the last few month, Compliant and Sovereign LLM Inference in Enterprise Environments is yet not possible. This paper first describes the main challenges toward democratizing LLMs in corporate worlds and suggests a new paradigm: *the microdatacenter*

1 LLM: The Ongoing Revolution

Despite impressive progress in recent years, there are two challenges to the adoption of LLMs in corporations: *the need for compliance* and *the need for a specific hardware-software stack*.

The Need for Compliance: LLMs suffer from a *compliance crisis*. Large companies such as OpenAI and DeepSeek run language models on *private untrusted clouds* located in specific countries. Many nations have strict compliance regulations, such as `PHIPA` in Canada, which mandates that hospitals storing patient records must use data centers located within Canada. Similarly, `GDPR` restricts the storage of certain EU data to European locations, while the `Swiss Federal Act on Data Protection` requires companies to store health data within Switzerland. This means that **a vast amount of data remains ineligible for LLM-based processing and RAG pipelines**. Even for American companies, the issue of trust remains a significant challenge. Many high-value industries, such as hedge funds and consulting firms, cannot use LLMs on *untrusted cloud providers* due to privacy concerns. This prevents many employees from leveraging LLMs to boost their productivity. In one of the companies where I worked, my two desk colleagues constantly used their private phones to ask ChatGPT about Python scripts.

The Need for a Specific Hardware-Software Stack: LLMs suffer from an *infrastructure crisis*. GPU-centric data centers are widely adopted in big tech companies such as Google and Meta. However, most companies still operate CPU-centric racks, **investing heavily in expensive and power-hungry CPUs costing up to $20,000, leaving less budget for high-performance GPU compute, which represents an inefficient allocation of resources**. Additionally, specific hardware stacks introduce the possibility of a CUDA/PTX runtime *à la carte*, improving LLM inference. GPU code is highly sensitive to many parameters. While a specific implementation might perform well on one GPU, it may perform worse on another due to microarchitectural differences. Finally, users often overprovision resources. For example, if employees are given the choice between an average model and a premium one, everyone will opt for the premium model, even for simple queries. Since LLMs are expensive, there is a pressing need for a globally adopted system that efficiently allocates resources and directs only complex queries to complex models.

2 Towards the Distributed Microdatacenter

The future trend is clear: LLMs will follow a trend we can call *inverse Moore's law for LLMs*, where inference size will shrink by a factor of two over a fixed period while chip capacity continues to increase in TFLOP/s. LLMs are a very recent technology, and inefficiency at this stage is expected—every new technology undergoes massive improvements over time, and the same must apply to LLMs. DeepSeek demonstrated a significant improvement over previous work, and within two years, new models will likely relegate DeepSeek to history. We are approaching a period where large-scale local LLM inference will be feasible and cost-efficient.

There is a unique opportunity for systems programmers to build an hardware/software infrastructure towards efficient inference systems to start democratizing LLM inference for enterprise environments. As mentioned earlier, local inference is the unique solution to *the need for compliance*, which is currently the most significant obstacle to full LLM integration in corporate settings. Moreover, while LLMs generate general-purpose content based on publicly available internet data, compliant sovereign LLMs can answer many company-specific queries using a technique called retrieval-augmented generation (RAG), since they are unrestricted by privacy concerns. For example, it is quite common in companies for employees to ask company-specific questions such as: "Who can help me with this specific problem?", "Who can grant me access to a new software license?", or "How do I set up this specific database?". LLMs with company-specific knowledge can navigate these complex queries and help employees find answers without disturbing busy colleagues.

3 Conclusion

I illustrated that running LLMs locally to comply with strict regulations is crucial in many sectors. Today, there is a unique opportunity to rethink LLM inference through a concept I name *the microdatacenter*, where a small local server rack handles LLM inference. This becomes feasible with the *inverse Moore's law for LLMs*. Systems engineers have a rare window of opportunity to design a cost-efficient, optimized local hardware stack to execute LLM inference on-premises. Employees want to use LLMs, and companies increasingly recognize that LLMs are crucial to their competitiveness.