Deloitte.

intel

Unlocking AI for government

A cost-effective, scalable solution

CPU-BASED STRATEGIES FOR MODERN GENAI DEPLOYMENT

Deloitte and Intel's **CPU-based language models offer a secure, cost-effective, and scalable approach** that meets many public sector needs. This alternative to GPU-based large language models (LLMs) makes Generative AI (GenAI) more accessible to government agencies facing budget and infrastructure limitations.

3 KEY BENEFITS

Deloitte and Intel built and tested a CPU-based solution for a state government client and found:

1. The CPU-powered small language models (SLM) matched the accuracy and speed of a GPU-based LLM, while **cutting operations and maintenance infrastructure costs by 55%.**

2. Compressing a larger language model (Llama 3.1 8B) **with Multiverse's CompactifAI tool** to run on CPUs made it faster and more memory-efficient while maintaining accuracy. Results included:
    - **Approximately 50% more cost savings**
    - **80% less memory required**
    - **4.9 billion fewer parameters**

3. With nearly **identical accuracy at a fraction of the infrastructure cost,** SLM on CPU allows clients to scale to multiple AI applications or multi agentic solutions without major infrastructure costs or changes.

*accuracy was calculated through initial benchmarking tests, not all use cases will yield the same accuracy

HAVE A GOOD USE CASE? LET'S WORK TOGETHER.

By using SLM and model compression, Government & Public Services clients can run GenAI on their current CPU or GPU infrastructure **powered by AWS cloud.** Our strategic decision-making framework helps clients choose the right hardware for their needs.

This approach broadens GenAI accessibility, reduces costs, and supports advanced use cases without sacrificing performance or data security.

CONTACT US

Doug Bourgeois
Managing Director
Deloitte Consulting LLP
dbourgeois@deloitte.com

Steven Phillips
Deloitte Alliance Manager
Intel
stephen.phillips@intel.com

Learn more about the Deloitte and Intel alliance at **https://www.deloitte.com/us/en/alliances/intel.html**