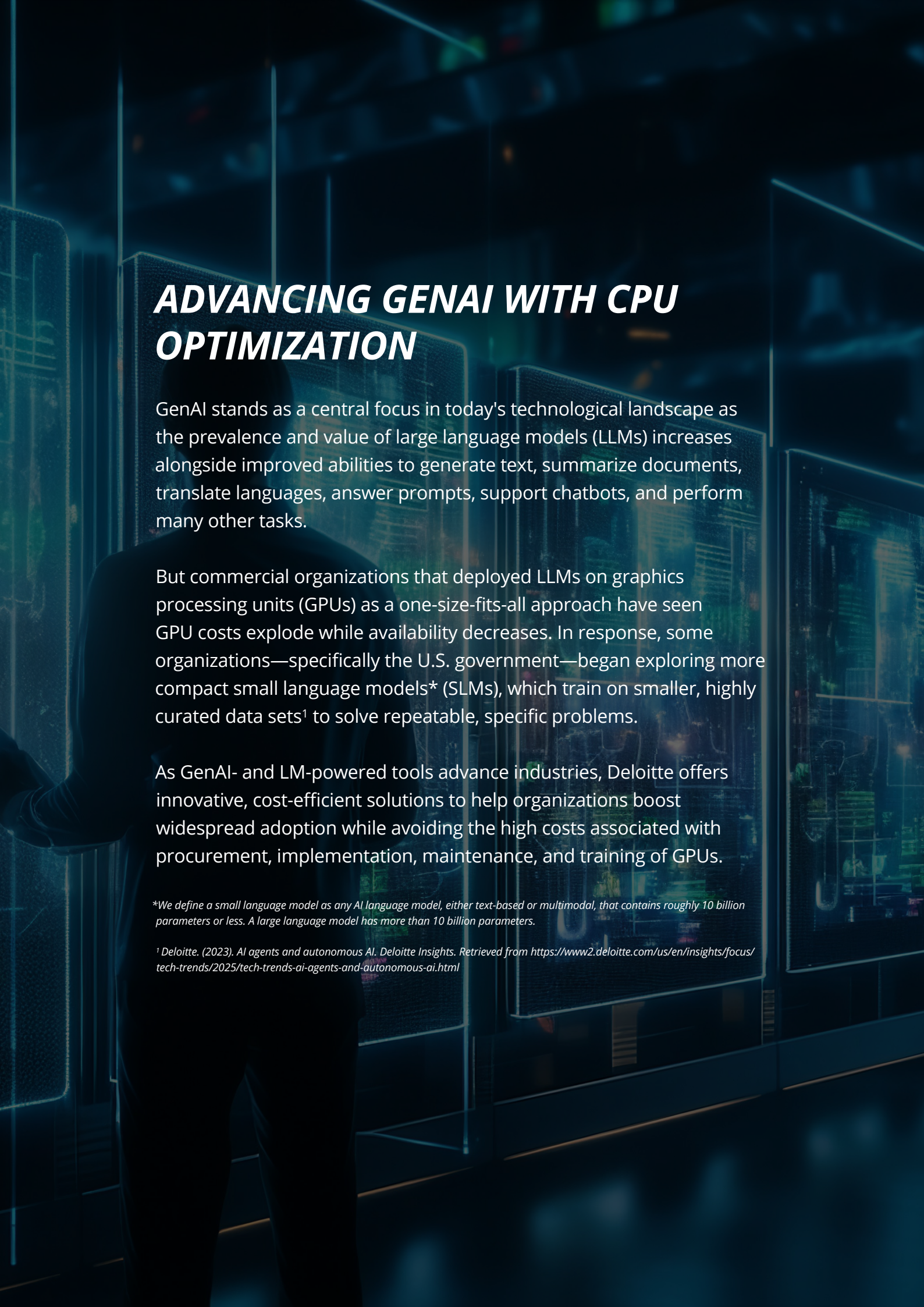
Deloitte.

INTEL GENERATIVE AI
MODEL OPTIMIZATION



ADVANCING GENAI WITH CPU OPTIMIZATION

GenAI stands as a central focus in today's technological landscape as the prevalence and value of large language models (LLMs) increases alongside improved abilities to generate text, summarize documents, translate languages, answer prompts, support chatbots, and perform many other tasks.

But commercial organizations that deployed LLMs on graphics processing units (GPUs) as a one-size-fits-all approach have seen GPU costs explode while availability decreases. In response, some organizations—specifically the U.S. government—began exploring more compact small language models* (SLMs), which train on smaller, highly curated data sets[1] to solve repeatable, specific problems.

As GenAI- and LM-powered tools advance industries, Deloitte offers innovative, cost-efficient solutions to help organizations boost widespread adoption while avoiding the high costs associated with procurement, implementation, maintenance, and training of GPUs.

**We define a small language model as any AI language model, either text-based or multimodal, that contains roughly 10 billion parameters or less. A large language model has more than 10 billion parameters.*

[1] Deloitte. (2023). AI agents and autonomous AI. Deloitte Insights. Retrieved from https://www2.deloitte.com/us/en/insights/focus/tech-trends/2025/tech-trends-ai-agents-and-autonomous-ai.html

BRIDGING THE GAP

As AI grows more complex, demand is rising for SLMs that serve small teams of users and run efficiently on edge devices and compact hardware. These SLMs are ideal for on-device AI, where space and resources are limited but data security remains a top priority.

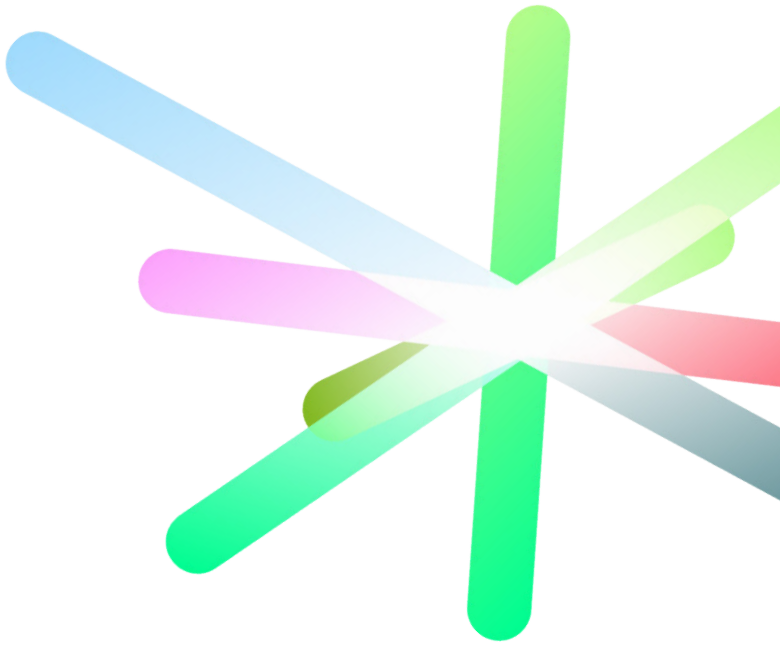
Organizations that want to keep their data on their internal network or Virtual Private Clouds (VPCs) can run AI on central processing unit (CPU)-based virtual servers. As a widely available alternative to GPUs, CPUs can help reduce hardware demands and computational requirements while maintaining security and delivering a cost savings of about 55%.

By optimizing LMs to run efficiently on existing CPU-based infrastructure, organizations can leverage their current hardware, minimize additional expenditures, and scale GenAI rapidly.

BOOSTING GENAI ACCESSIBILITY: THE POWER OF LOCAL CPU PROCESSING

Running SLMs on local CPUs offers a compelling alternative to deploying LLMs over the internet because it allows organizations to scale GenAI solutions efficiently while keeping sensitive data in secure VPC environments.

Modern CPUs, including Intel's Xeon processors, are especially well-suited for GenAI inference for three main reasons:



01 **On-chip acceleration & efficiency:**
Newer Xeon processors (starting with 4th Gen) feature Advanced Matrix Extensions (AMX), which significantly boost AI workload performance and efficiency by accelerating math operations directly on the CPU. This hardware-based acceleration delivers better performance-per-watt and reduces power consumption compared to GPUs.

02 **Simplified system design:**
Using Xeon CPUs for inference eliminates the need for separate GPUs and simplifies system architecture. Xeon processors handle both AI tasks and overall system management, leveraging their large caches and fast memory access, and supporting the wide range of software already built for x86.

03 **Lowered cost:**
Discrete GPUs are often more expensive and power-hungry than needed for most AI inference tasks. Xeon CPUs provide sufficient performance for many applications, resulting in lower hardware costs and total cost of ownership (TCO).

Advances in model optimization—such as compression and running workloads on an optimized model server—now allow SLMs and compressed LLMs to run effectively on CPU-based architecture. This makes GenAI solutions more accessible, affordable, and scalable across industries.

As GenAI adoption accelerates, leveraging widely accessible CPUs empowers private and public sector organizations to adapt quickly to evolving needs. By championing this shift, Intel helps meet the demand for more accessible AI solutions and positions itself as a key driver in the future of GenAI innovation.

OUR APPROACH

PHASE 1 OVERVIEW:

At the end of 2024, Deloitte began working with a state Medicaid department that used a GenAI chatbot for welfare policy assistance, but extremely high costs and limited access to GPU resources made the GenAI setup unsustainable. Together with Intel, Deloitte tested whether a CPU-based SLM would provide a more cost-effective alternative to the GPU-based LLM solution.

Deloitte selected the AWS cloud environment for this research because it offers a wide range of cloud instance options equipped with processors from Intel and other vendors. This made it easy to compare performance and optimize workload distribution across CPU-intensive computational tasks and GPU-accelerated machine learning, AI, and high-performance applications.

AWS is also popular with government and public sector organizations because of its robust compliance certifications, including FedRAMP and Impact Level 5 (IL5), which ensure it meets stringent security requirements for handling sensitive data.

For models, Deloitte used Mistral-7B as the small language model and Mixtral 8x7B as the large language model, both from Mistral.AI. These models often earn top global benchmark scores and, because they are open source, Deloitte and Intel could reduce costs while easily fine-tuning and customizing them for the project.

With the testing components in place, Deloitte and Intel constructed a high-level process flow for Phase 1:

With the testing components in place, Deloitte and Intel constructed a high-level process flow for Phase 1:

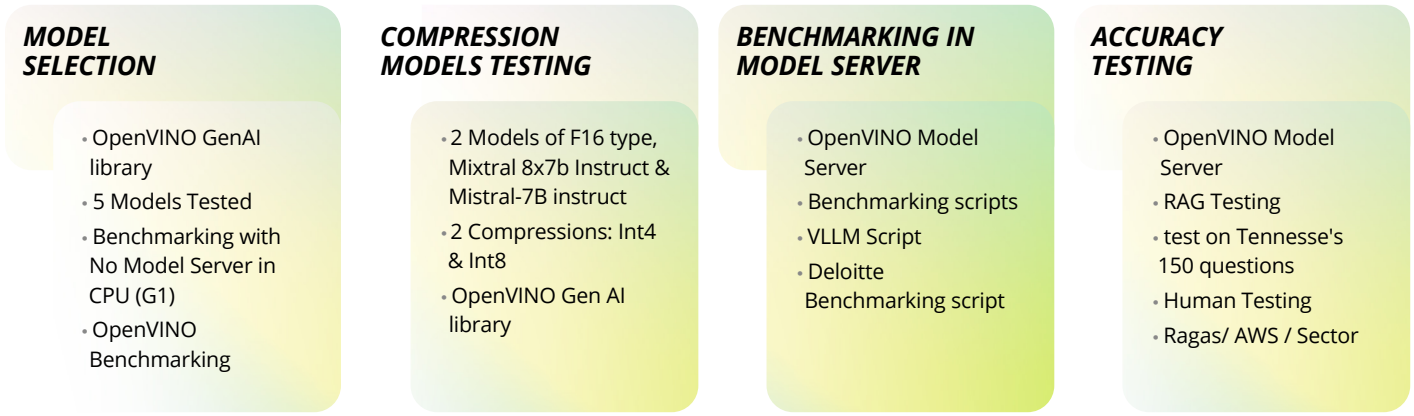


Figure 1: High-level process flow for phase 1 testing

PRELIMINARY BENCHMARKING OF PHASE 1

Preliminary benchmarking involved prompting with a single basic question such as, "What is deep learning?" For GPUs, Deloitte used HuggingFace inference, and for CPUs used Intel's open-source model server, OpenVINO, to infer SLMs. Because OpenVINO optimizes deep learning models, LLMs and SLMs can run efficiently on Intel CPUs—reducing costs and expanding AI accessibility.

Initial benchmarking revealed that in addition to the SLM (Mistral-7b), the LLM (Mixtral-8x7b) successfully ran on a CPU using the OpenVINO model server. The speed/performance metrics on the SLM were acceptable and comparable to the LLM running on the GPU and were almost identically acceptable up to 128 users. This showed that the CPU-based configuration scales much more efficiently than the GPU configuration.

When comparing cost of a G5.12x large GPU instance to the M7i-12xlarge CPU instance, organizations that opt for CPU instances can yield 56.8% cost-savings, which could be achieved with minimal or no increase in response time or user wait time.

OUR APPROACH

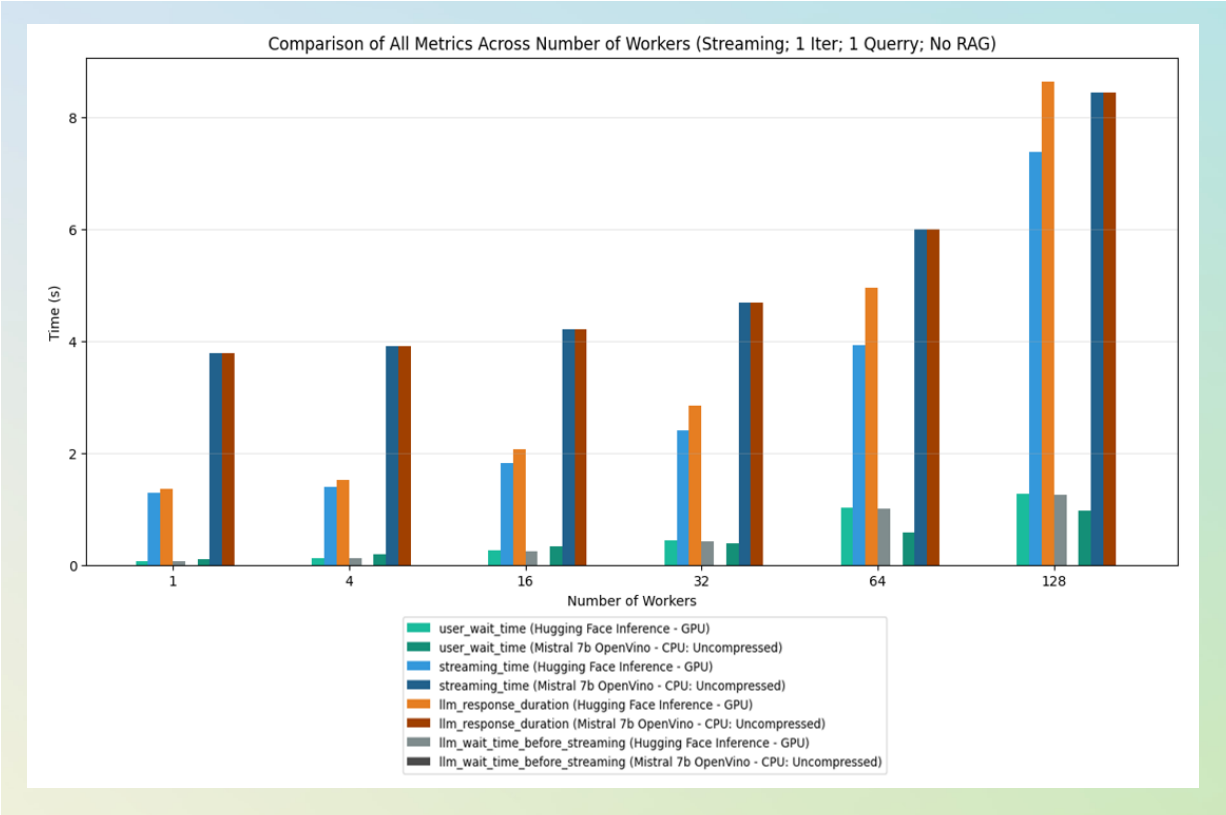


Figure 2: Phase 1 benchmarking metrics

US STATE GOVERNMENT: TEST USE CASE

Using AIdvisor, Deloitte's LLM-enabled Medicaid chatbot built for state government, Deloitte tested and compared the accuracy of the SLM and LLM on CPU and GPU. The AIdvisor team developed a set of 150 golden questions, each with agreed-upon 'accurate' ground truth responses and 'relevant' reference texts. These testing sets span Medicaid eligibility policy, a domain known for its high complexity, which allows it to effectively highlight accuracy issues in LLMs.

PHASE 1 ACCURACY

Accuracy testing of the LLM from Phase 1 revealed that the model performed with similar accuracy on the CPU in addition to the GPU, despite performance slowdown. Further testing using OpenVINO quantization on the LLM showed improved performance. Initial accuracy testing of the SLM initially showed a decrease in accuracy on CPU. Averaging 28.8 seconds per query, the SLM on CPU did show viability for numerous use-cases, however, especially those that do not require overly complex responses. **Overall, Phase 1 showed that SLMs and compressed LLMs can run on a CPU, but more complex use cases could improve output accuracy.**

MODEL TYPE	CPU OR GPU	MODEL SERVER	HUMAN ACCURACY	AVG TOTAL RESPONSE TIME (SEC) PER QUERY
Mixtral8x7b	GPU	Mixtral8x7b	85.33%	16.77
Mixtral8x7b	CPU	Mixtral8x7b	83.33%	92.66
Mistral 7b	GPU	Mistral 7b	72.00%	5.61
Mistral 7b	CPU	Mistral 7b	66.67%	28.84

Figure 3: Phase 1 accuracy & response time results

The Deloitte team evaluated accuracy consistency by querying the SLM ten times for each of the 150 questions in the benchmark dataset and measuring the RAGAS metrics of faithfulness, correctness, and semantic similarity. These results showed wide variation, underscoring that the same model can alternate between strong and weak answers. This indicates a need for consistency-oriented interventions and more robust evaluation pipelines before relying on SLMs in production.

OUR APPROACH

PHASE 2 OVERVIEW:

After Phase 1 proved that SLMs could run well on CPU architecture and performed comparably to a GPU-based AI solution, Deloitte explored running even smaller, "compressed" language models on CPUs.

To do this, Deloitte collaborated with the startup Multiverse, which offers a tool called CompactifAI. This tool uses quantum-inspired tensor-based compression to decrease the size of a language model while maintaining 95%-97% accuracy. While Phase 1 focused on proving SLMs could work on CPUs, Phase 2 aimed to make CPU-based AI even faster and more practical by reducing inference time.

The AWS cloud environment, testing process, and state government use case remained the same in Phase 2. The major difference in this phase was the model: Deloitte and Intel used Meta's Llama 3.1-8B, which Multiverse had successfully compressed. Phase 2 tested and compared compressed and uncompressed (original) versions of this model, which is also open-source.

PRELIMINARY PHASE 2 BENCHMARKING:

Deloitte planned to continue using the HuggingFace model server for GPUs and OpenVINO for CPUs to run the LLMs and SLMs. However, when testing the compressed model revealed challenges with HuggingFace on GPUs, they switched to the vLLM model server for GPU inference.

Phase 2 benchmarking showed extremely impressive results: the compressed model on CPU was faster and more efficient than the uncompressed model, and it also handled more users with less increase in response time. While CPUs initially garnered longer wait times than GPUs, they stayed efficient as user numbers grew, whereas GPUs struggled to keep response time low as user volume increased. Using Intel Xeon CPUs with OpenVINO and CompactifAI yielded a response time of less than ten seconds regardless of user count—something the GPU and uncompressed testing configurations could not match.

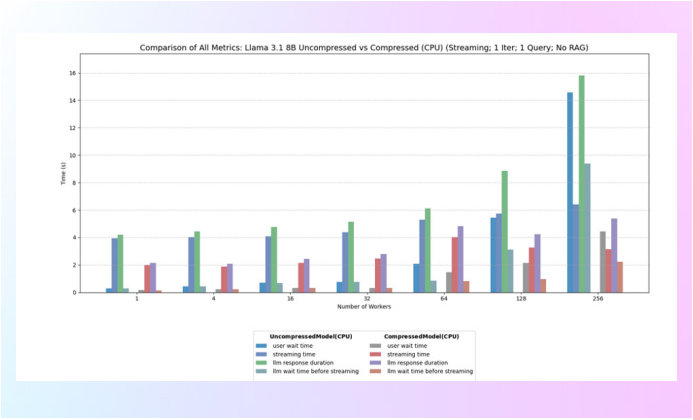


Figure 4: Phase 2 benchmarking metrics: Uncompressed model vs. compressed model on CPU

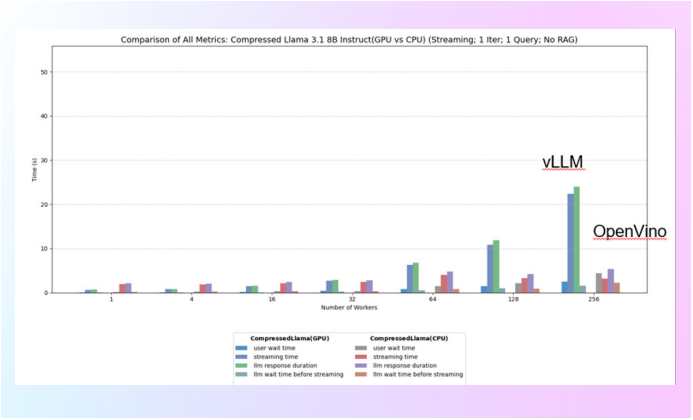


Figure 5: Phase 2 benchmarking results: CPU vs. GPU with compressed model

OUR APPROACH

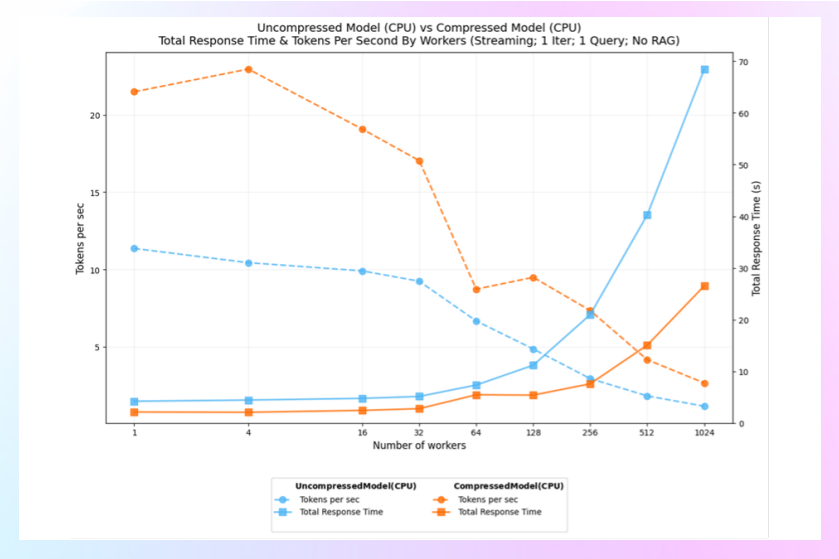


Figure 6: Phase 2 speed deep-dive benchmarking results: Uncompressed model vs. compressed model on CPU

PHASE 2 ACCURACY:

Using the same state government use case from Phase 1, Deloitte found that running compressed Llama 3.1 on CPU with OpenVINO achieved a human accuracy level of 84%—the highest CPU accuracy yet.

The response time for retrieval-augmented generation (RAG) averaged 36.26 seconds per query, almost identical to Mistral 7b on CPU from Phase 1. When testing at scale with 256 workers, compressed Llama 3.1 on CPU using OpenVINO performed much better than the same model on GPU with VLLM. The compressed models maintained similar accuracy and response time metrics as the originals while reducing size and memory use by 80%—showing they can handle real-world AI workloads efficiently.

MODEL TYPE	*CPU OR GPU*	*MODEL SERVER*	*HUMAN ACCURACY*	*AVG TOTAL RESPONSE TIME (SEC) PER QUERY*
Llama 3.1-8b	GPU	HF / vLLM	78.67%	11.82
Llama 3.1-8b	CPU	Open Vino	84.67%	29.51
Llama 3.1-8b (compressed)	GPU	HF / vLLM	68%	8.08
Llama 3.1-8b (compressed)	CPU	Open Vino	84%	36.26

Figure 7: Phase 2 accuracy & response time results

OUR APPROACH

MULTI-AGENTIC USE CASE

Building on results from Phases 1 and 2, Deloitte developed a real-world application for SLM and compression technology. The team chose a complex, multi-agentic use case for a federal government agency, which involved extracting entities from text documents and classifying which ones to mask based on the U.S. government's sensitivity hierarchy. While conventional LLMs can handle the receipt and reasoning of a user's prompt, SLMs or compressed models can efficiently tackle simpler tasks such as generating SQL commands. Because these compressed models are open source, Deloitte could develop the solution in a closed, more secure environment that met the client's strict data security needs.

In the example use case architecture below, an Intel CPU instance on AWS runs a multi-agentic entity extraction application. This tool uses small or compressed language models to identify sensitive entities in a text document, cross-reference them against a list of sensitive words, and redact them if necessary. The orchestrator model, which instructs the smaller models, can either be a conventional LLM or SLM depending on response time requirements.

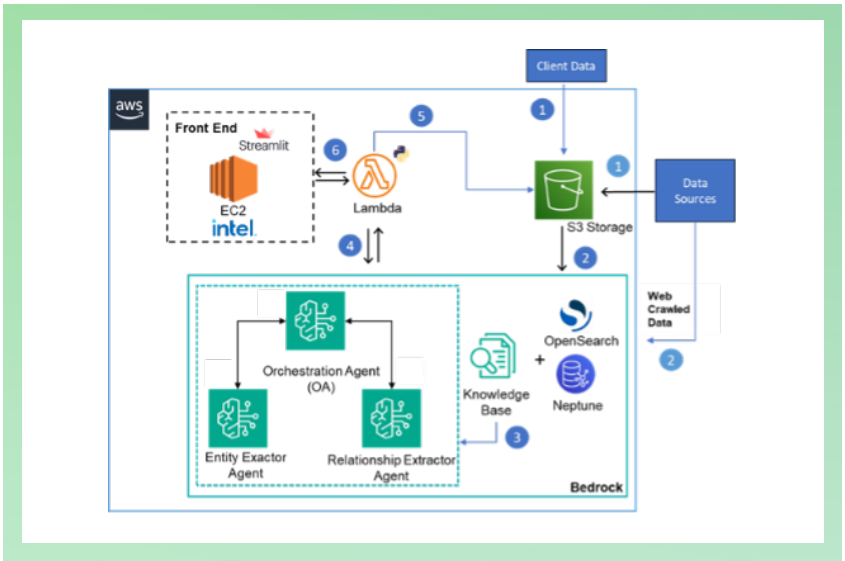


Figure 8: Multi-agentic use case example architecture

1. *User makes a query*
2. *Flask application is deployed on EC2 to handle frontend queries, which in turn triggers Lambda*
3. *Lambda invokes the Orchestrator Agent*
4. *Orchestrator Agent checks for Redshift entry for a date. If found, it will run the Summary Agent, or else runs the Extractor Agent*
5. *Search Agent invokes Lambda search and stores the result in Redshift.*
6. *Summarizer Agent retrieves data from Redshift*
7. *The Flask application retrieves the agent's response and displays it to the user.*

This successful use case demonstrates that smaller, compressed models can handle simpler tasks within a multi-agent system. This approach makes it possible to support more complex workflows without substantially increasing inference cost or response time.

OUR APPROACH

DELOITTE'S CHART/DECISION MATRIX:

Through Phases 1 and 2, we found many use cases previously thought suitable only for GPUs are actually suitable for CPU by using either SLMs, compressed models, or both models in tandem with a conventional LLM. The chart below shows how Deloitte determined hardware requirements (CPU or GPU) and model selection based on use case complexity and existing infrastructure.

Use Case Category	Use Case Example	Multi Modal?	Latency Requirement	Sustainable Model(s)	Hardware w/o Compression	Hardware with Compression	Concurrent Users
Text Classification	Sentiment analysis of social media posts	No	Low (<30s)	SLM, LLM (if high accuracy needed)	CPU	CPU	200
Image Classification	Identify objects in an image	No	Low (<30s)	SLM, LLM (for complex scenes)	CPU or GPU	CPU	200
Question Answering	Answer factual questions based on a knowledge base	No	Low (<30s)	SLM, LLM (for complex or open-ended questions)	CPU or GPU	CPU	200
Machine Translation	Translate news articles from English to Spanish	No	Moderate (<30-100s)	SLM, LLM (for higher fluency)	CPU or GPU	CPU	100
Summari-zation	Summarize long documents	No	Moderate (<30-100s)	SLM, LLM (for more nuanced summaries)	CPU or GPU	CPU	100
Object Detection	Locate and identify multiple objects in an image	No	Moderate (<30-100s)	SLM, LLM (for higher accuracy)	CPU or GPU	CPU	100
Text Generation & Tex to SQL	Knowledge mining and Data Analytics	No	Moderate (30-100s)	SLM, LLM (Multi-Agentic)	GPU	CPU	50-100
Visual Question Answering	Answer questions about the content of an image	Yes	Moderate (30-100s)	LLM Multimodal	GPU	CPU?	50?
Document Question Answering	Extract information from a document and answer questions	Yes	Moderate (30-100s)	LLM Multimodal	GPU	CPU?	50?

Figure 9: Example AI decision matrix

As Deloitte and Intel continue to test, we hope to incorporate bigger and more complex use cases on CPU architecture.

ENABLING GENAI AT SCALE

Scaling GenAI in government is challenging, despite strong interest and demand[2]—often higher than in the commercial sector[3]. Many government employees lack access to GenAI tools: a Deloitte survey found 75% of government respondents reported that less than two out of five workers have access. The Phase 2 results show that running LLMs on CPUs can make GenAI more accessible to organizations of all sizes and open to more use cases.

Government teams, such as defense, intelligence, and law enforcement, are often small due to security needs. These small- and medium-sized teams can benefit from SLMs and compressed LLMs running on their existing CPU-based virtual servers.

Increases Use Case Feasibility

- Running GenAI models on CPUs quickly and efficiently extends the scope of AI use cases across industries.
- CPUs effectively support LLMs and SLMs for training, testing, and inference, which increases use cases and removes barriers related to GPU costs and availability.
- CPUs allow secure, local data processing in edge environments, such as remote healthcare diagnostics or smart manufacturing, where latency and data privacy are critical.
- Organizations can optimize workflows and scale AI pipelines for continuous model improvement.
- CPUs are widely available and don't require specialized hardware and expertise, which simplifies and accelerates adoption for organizations with limited resources.

Open-Source Framework

- Publicly available language models and Intel's OpenVINO model server enable organizations to configure their GenAI stack on CPU architecture.
- Teams can fully customize their language model and infrastructure based on use case.
- Open-source framework means organization owns and governs data—no uploading to a cloud or public infrastructure.
- No recurring subscription or licensing costs to run language models on local CPUs.

Agentic AI Deployment

- CPU-based AI supports the development of agentic AI—systems capable of reasoning, planning, and decision-making.
- Connecting enterprise data to AI models on CPUs unlocks advanced AI-driven automation and innovation.
- CPUs act as the operations workhorse, while agentic AI leverages these capabilities to perform complex tasks autonomously.
- Agentic AI is in high demand for many organizations.

Decreases Costs, optimizes TCO

- Model compression enhances the feasibility of running GenAI models on lower-cost architectures using CPUs, ensuring competitive performance while reducing computational overhead.
- Reduced complexity enables organizations transitioning to GenAI to deploy AI solutions with existing IT infrastructure.
- CPUs eliminate the need for costly GPU instance usage, infrastructure overhauls, and specialized IT support—so industries with legacy systems can embrace GenAI.
- Inclusive AI adoption accelerates innovation across sectors that couldn't previously invest in advanced technology.

Secure On-Premises AI Operations

- Running AI on CPUs lets organizations keep data secure on their own servers or private clouds, helping protect sensitive information and meet regulatory requirements.
- CPU-based GenAI supports customized security protocols tailored to specific use cases, protecting AI deployments from vulnerabilities.
- Smaller teams can run curated SLMs in their own environments, enhancing data security.
- Adaptability builds trust in AI systems, promoting greater adoption in sectors like defense, healthcare, finance, and public administration.
- By balancing scalability, security, and flexibility, CPUs provide a robust foundation for confidently adopting GenAI, even in high-risk environments.

[2] Deloitte. (2023). Government faces challenges with generative AI adoption. Deloitte Insights. Retrieved from https://www2.deloitte.com/us/en/insights/industry/public-sector/government-faces-challenges-with-generative-ai-adoption.html?id=us:2sm:3ab:4diUS187755:5awa:6di:102424&pkid=1012823

[3] 77% of employees and 96% of technical leaders are eager to use GenAI compared to commercial's 39% and 85%

CONCLUSION

GenAI is revolutionizing industries and reshaping how work gets done. For government, the democratization of GenAI is more achievable than ever, thanks to the ability to run advanced models on existing CPU infrastructure.

Deloitte and Intel empower organizations to deploy highly accurate yet efficient models without complex hardware, so public and private sector organizations can maximize GenAI value. We also help develop strategic decision-making frameworks that help government agencies and other clients allocate resources effectively, balancing performance requirements and budget constraints while achieving desired outcomes.

By bringing together Deloitte's deep industry leadership with global technology leaders such as Intel Corp and AWS, we make it easier to leverage the cutting-edge technologies of today to build solutions for tomorrow.

MEET THE THOUGHT LEADERS

Key points of contact for any questions regarding the content of this paper.

Doug Bourgeois
Managing Director
Deloitte Consulting LLP
GPS CBO Cloud Engineering

Saaket Varma
Specialist Senior
Deloitte Consulting LLP
GPS CBO Cloud Engineering

Arthi Submaranian
Specialist Master
Deloitte Consulting LLP
GPS AI & Data

Matt Sheerin
Consultant
Deloitte Consulting LLP
GPS CBO Hybrid Cloud Infrastructure

Robert Simmons
Specialist Leader
Deloitte Consulting LLP
GPS Analytics and Cognitive

Burnie Legette
Director
Intel Corporation
Public Sector Sales, Artificial Intelligence