Generative AI w/ DELL & PRESIDIO

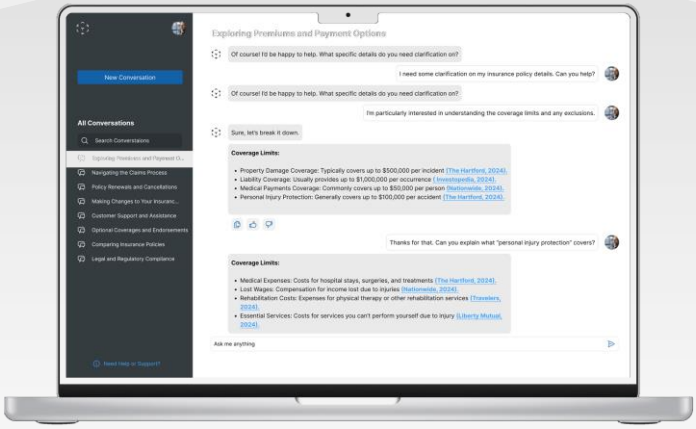
Private GPT | Secure | Modern Chat, Q/A and Summarization

THE CHALLENGE

Enterprises require secure, compliant AI that goes beyond simple chat. Presidio delivers **HAI Chat Accelerator** that safeguard data, comply with policies, and drive innovation on on-premises GPUs. Built with open-source and Nvidia NIM models, the HAI Private Chat Accelerator provides a proven path to deployment.

- Utilize the flexibility and power of state-of-the-art LLMs like LLAMA or OpenAI OSS for innovation and efficiency.
- A modern UI for Chat, QA, and Content Summarization, enhancing the user experience.
- Advanced guardrails to control model behaviors, supporting security and compliance.

DELL Technologies · NVIDIA · PRESIDIO



Chat results based on your data, ensuring relevance and accuracy. Doc. mgmt. to keep the index current.

Chat, QA, and concept summarization with conversation history for continuous context

Guardrails to enforce brand guidelines and correct behaviors, ensuring consistent outputs.

Ensures your data stays within your environment, preventing exfiltration and maintaining complete control

INFRASTRUCTURE MANAGEMENT

NVIDIA AI ENTERPRISE

DELL PRO SERVICES + PRESIDIO RAPID-AI TEAM

EQUIPMENT

- PowerEdge XE9680 with 4 x H200s Srvs + Management Node
- 4x NVIDIA GPUs: Cutting-edge GPUs optimized for AI and ML tasks
- Purpose-built for POC Workloads. Scalable.

PRESIDIO RAPID-AI TEAM

- 10-Day Implementation, customized Private GPT for your POC.
- Knowledge Transfer & Managed Services

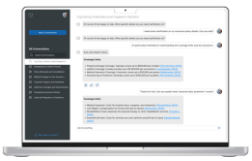
DELL Technologies
TITANIUM BLACK PARTNER

PRESIDIO®

PRIVATE AI ACCELERATOR SOLUTION SPEC SHEET

EQUIPMENT

Feature	Specification
Server	Dell PowerEdge R670 and XE9680
Processor	Dual Intel Xeon Scalable 5th Generation
Memory	64GB DDR5-5800 RDIMMs (16GB per DIMM) x 8 = 512GB total across systems
Storage	Up to 3.84TB Data Center NVMe drives, scalable across multiple slots for enhanced storage performance
GPUs	4x NVIDIA H200 (140GB) for AI/ML acceleration, designed for high-performance AI workloads
NVIDIA AI Enterprise	NVIDIA AI Enterprise Essentials Subscription per GPU for 1 year



NIM-DEPLOYED HAI CHAT ACCELERATOR

Feature	Specification
NVIDIA AI Enterprise	Full support for NVIDIA AI Enterprise suite, providing a robust foundation for enterprise AI workloads.
LLM Model	Integrated with LLAMA or OpenAI NIM, a large-scale model optimized for a wide range of GenAI/NLP tasks, including summarization, Q/A, chat, and content creation. Tested to perform at 'token per second' rates of 30+, greater for smaller models.
Application Stack	Docker-based deployment with open-source components: ▪ Frontend: React-based user interface ▪ Backend: FastAPI and Python for scalable API services ▪ Vector Database: Milvus, optimized for high-performance similarity search
Security & Compliance	Robust security features, ensuring that private data remains on-premises, fully under your control. It supports the curation of your own data and models, allowing for customized AI solutions tailored to your specific needs while maintaining strict compliance with industry standards.
Flexibility	Supports integration with other AI/GenAI workloads, including custom models and additional data sources. You can fine-tune your own models to meet domain-specific requirements. Additionally, the platform allows for easy selection from NVIDIA's extensive NIM model catalog, offering a variety of GenAI and domain-specific models across industries, including reasoning, vision retrieval, speech, biology, and more @ https://build.nvidia.com/explore/discover