Technology | Italy

Deploying AI agents anywhere

Memori

Memori's AISURU platform enables companies to build conversational AI agents with ease. Working with Araneum Group, Memori created a reference architecture for secure on-premises deployments of AISURU based on Lenovo ThinkSystem SR675 V3 servers and NVIDIA® L40S Tensor Core GPUs.

1 Customer background

Who is Memori?

Memori is an Italian company specializing in technologies for the development of AI-enhanced conversational interfaces and experiences. Founded in 2017, its mission is to improve people's lives through AI. Memori places the human dimension at the heart of AI creation, enabling companies and new generations of creators to build their own unique AIs, quickly and easily.

2 The challenge

Memori's flagship solution is AISURU, a platform to create and manage conversational AI agents whose purpose is to interact with users and offer personalized information, support, and assistance.

AISURU integrates custom and open-source large language models (LLMs) with Memori's proprietary natural language processing (NLP) technologies, empowering clients to create and manage their own AI team. AISURU can personalize content based on the user, their specific role, the time and place of interaction, as well as the context and initial questions posed to it. That way, the AI agent provides a truly unique and engaging experience for each user.

What's more, AISURU's innovative 'Deep Thinking' feature provides the AI with persistent memory, making the agents capable of remembering user preferences and habits to offer a more tailored experience over time.

2 The challenge

To date, banks, engineering companies, publishing groups, hotel chains, pharmaceutical companies, manufacturers, and educational institutions have all adopted AISURU to create AI agents for a wide range of use cases. These include virtual tour guides, sales assistants, training advisors, product configurators, document compilation assistants, and many more.

Developed as a cloud-based framework, AISURU was initially only available as a Platform as a Service (PaaS) solution running on AWS. For many companies, this proved to be a barrier to entry, with concerns about data sovereignty in the public cloud.

Nunzio Fiore, Founder & CEO of Memori, begins: “The first question on every prospective client’s lips is ‘where is my data stored?’ For many companies in highly regulated industries, public cloud deployments are simply not an option, so we looked to make AISURU available for private cloud and on-premises platforms.”

“We want to give clients opportunity to **deploy AISURU however best fits their needs**, whether that’s in the public or private cloud, or on-premises in their own data center.”

Nunzio Fiore

Founder & CEO, Memori

3 The solution

Partnering for success

Memori teamed up with Lenovo to offer a dedicated hardware solution for on-premises implementations, optimized to run both the AISURU platform and the client's chosen LLM models locally.

The company engaged trusted technology partner Araneum Group to help design, configure, and test a reference architecture based on a Lenovo ThinkSystem SR675 V3 server equipped with four NVIDIA® L40S Tensor Core GPUs.

Hardware
Lenovo ThinkSystem SR675 V3
NVIDIA® L40S Tensor Core GPUs

Software
AISURU

3 The solution

Secure, powerful, on-premises architecture

The Araneum Group team was granted access to the Lenovo Innovation Center for one month to run tests on a Lenovo ThinkSystem SR675 V3 server.

Fabio Lecca, CTO at Araneum Group, recalls: “We built custom software to enable communication between the AISURU platform and the Lenovo hardware, and deployed three existing AI agents to the on-premises infrastructure for stress testing.”

Equipped with four NVIDIA L40S Tensor Core GPUs dedicated to AI compute, the Lenovo ThinkSystem SR675 V3 is the ideal server for accelerated performance and efficient execution of generative AI models and NLP.

ARANEUM

4 The results

By partnering with Lenovo, Memori can offer clients a pre-configured, dedicated hardware solution for a fully on-premises AISURU deployment, which seamlessly integrates with on-premises data sources.

There is even support for air-gapped environments for companies with very strict security requirements. “There is zero dependence on external cloud services, guaranteeing total data sovereignty and control,” confirms Nunzio Fiore.

The partnership with Lenovo offers organizations the possibility to keep the entire conversational AI infrastructure completely in-house, ideal for sectors with requirements for stringent data sovereignty and regulatory compliance, and for deployments in locations with limited internet connectivity.

- Certified reference architecture for on-premises deployments
- GPU-accelerated AI performance
- Total data sovereignty and control

“The **collaboration with Lenovo marks the beginning of a new chapter** for Memori. By offering on-premises deployments of the AISURU platform, we hope to help more companies in more industries **harness the power of conversational AI agents**.”

Nunzio Fiore

Founder & CEO, Memori

Why Lenovo and NVIDIA?

A chance meeting at an industry event led Nunzio Fiore to introduce the AISURU platform to a Lenovo rep. “We got to talking about AISURU and conversational AI agents, and I opened up about our plans to offer a dedicated on-premises solution,” says Fiore. “Before long, we were working on a reference architecture together. Lenovo’s NVIDIA-accelerated, AI-optimized servers are the ideal foundation for our on-premises AISURU offering.”

How can companies in regulated industries make the most of AI?

Memori enables secure on-premises deployments of its innovative AISURU platform with a reference architecture based on Lenovo and NVIDIA technology.

Explore Lenovo AI Solutions