insight.tech

Simplify Data Retrieval with GenAI Tools and AI GPUs

When business customers need answers to their questions, they usually need them fast, especially with product designers and developers looking for critical information needed to complete a project. This is a challenge that sales reps often face.

More often than not, core technical know-how and product data is in the hands—and on the servers—of product managers and sales engineers. This means salespeople must make requests to get the data their customer wants, which often results in time wasted with a lot of back and forth between both parties, not to mention unhappy customers.

Removing IT/OT information silos, commonly found in manufacturing, engineering, and similar organizations, helps solve this challenge. No-code generative AI (GenAI) tools, large language models (LLMs), and chatbot solutions make it possible, by putting data in the hands of those who need it when they need it. Technical teams are freed up to focus on their work while sales can respond to customers faster and more effectively.

A collaboration between Tallgeese AI, an on-premises, privacy-first GenAI software company, and hardware manufacturer ADLINK Technology Inc. shows how it's done.

Together, the two companies make a "perfect match," says Jeffrey Lai, cofounder and CEO of Tallgeese AI. "We can run our software on ADLINK's robust hardware smoothly, and together provide our customers full control of the data and AI capability on-premises, reducing cybersecurity risks."

No-Code AI Speeds Chatbot Training

The ADLINK hardware and Tallgeese AI software form a cohesive orchestration where even nontechnical users can ask an AI chatbot to find information that exists locally and get a response within moments.

The Tallgeese AI turnkey workstation solution performs the data ingestion, converts documents stored on the server, and puts them in a vector database. The LLMs access the files from the database and run AI inference and training on-site, while creating an offline record system for data retrieval.

"We take a no-code AI approach. You turn on your computer, select the file directory, click 'train the chatbot', and you've basically gone through the entire chatbot training process," says Lai. "We parse PDF, Excel, PowerPoint, and other file types, turning all the tags, elements, and images into vectors. We store it locally, essentially preparing it for AI access—all in a matter of minutes."

Authorized users simply login to the Tallgeese AI portal, choose the on-premises server, and see all AI chatbots created on that machine. They can drag and drop files from databases, file directories, and even web pages for external crawling.

With a plug-and-play model, no IT support is needed. As a rule, Tallgeese AI software comes pre-loaded and ready to run on a network-connected PC. In this case, it's the ADLINK AI GPU Server—a new product based on 4th Generation Intel® Xeon® processors and Intel® ARC A770 GPUs **(Video 1)**.



Video 1. ADLINK AI GPU Server overview. (Source: ADLINK)

Secure, Fast Data Access with AI GPU Servers

ADLINK knows firsthand about this challenge. It faced the same knowledge management challenges as its own customers, making the company a great example of how deploying a GenAI-powered solution makes mundane tasks easier.

The development of almost any product requires a massive amount of documentation throughout its lifecycle—from concept to end-of-life. Product requirements documents, component specifications, design documentation, engineering change notices (ECNs), and much more are created and updated continually. Many documents contain proprietary IP and need to be securely accessible for a variety of purposes by a range of non-engineering personnel. But gaining access to this information is time-consuming and frustrating.

Hank YH Lin, Product Manager in ADLINK's Networking and Communication Department, describes a typical problem: "Our customers have a huge number of regulatory compliance data, and if their quality team goes to the factory for an on-site audit, they need very experienced subject matter experts to answer questions."

The Tallgeese GenAI solution, running on an ADLINK AI GPU Server, can improve the quality of customer interactions—even with new and less knowledgeable employees.

Another challenge for engineering companies is tracking historical documentation such as the many product engineering change notices (ECN) and associated FAQs. "The problem is once an ECN is issued, people will forget about it," says Lman Chu, CSO of Tallgeese AI. "I think the key is that AI doesn't forget. All the records are inside the AI's head, so to speak, helping everyone do a better job."

The Tallgeese no-code AI solution enables project teams to quickly, seamlessly, and securely integrate GenAI into their existing workflow—without having to rely on the IT department. They can benefit from new productivity gains almost immediately. And when information is easily found and shared, product development can be more agile, creating the potential to bring new solutions to market faster.

GenAI Tools Democratize AI

It's inevitable that businesses across every vertical will invest in GenAI-powered transformation to increase operational efficiency, lower costs, and stay ahead of the competition.

"From the top down, organizations that are adopting AI help their employees focus more on meaningful work, maximize their productivity, and allow for a better work-life balance," says Lai. "I think it's executives' responsibility and the outcome I see for the future."

Technologies like GenAI, LLMs, and chatbots are in use today and are where we'll see true democratization of AI for the future.

intel partner alliance

© 2024 insight.tech. Sponsored by Intel®. Trademarks.

LRI Distribuída no Brasil pela LRI | vendas@lri.com.br | 51 2129-3000