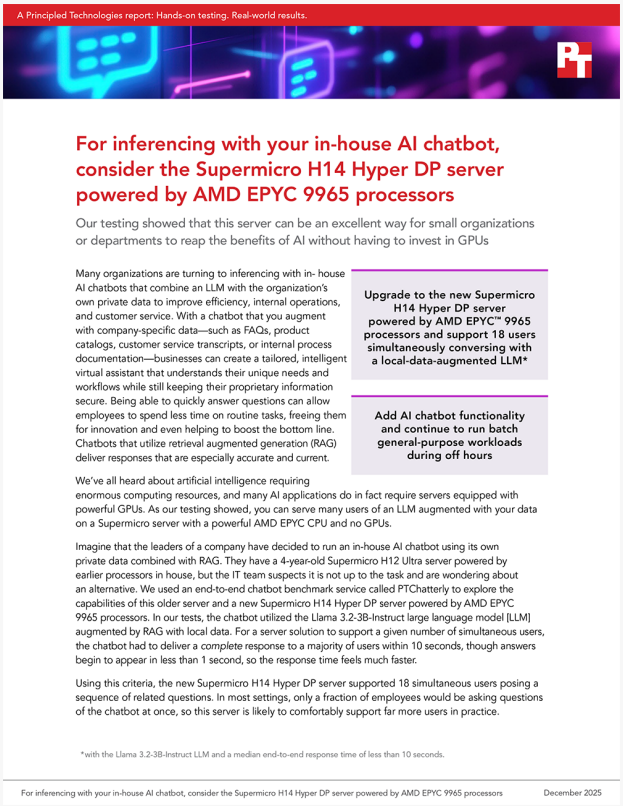
New study finds that a Supermicro H14 Hyper DP server powered by AMD can support a local-data-augmented LLM

The study, released by Principled Technologies (PT), highlights the performance of the Supermicro H14 Hyper DP server compared to a 4-year-old server.

SAN JOSE, CA, UNITED STATES, December 5, 2025 /EINPresswire.com/ -- Organizations of all sizes are considering in-house artificial intelligence (AI) chatbots that combine a large language model (LLM) with their own data, offering the opportunity to improve operations without risking the exposure of private data. Chatbots that utilize retrieval augmented generation (RAG) deliver responses that are especially accurate and current. For these organizations, it's critical to choose a server solution that can deliver quick response times for multiple users conversing with the LLM simultaneously.

Third party Principled Technologies compared the LLM performance of a new Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors to that of a 4-year-old server. They recently released a test report with their findings.


A Principled Technologies report: Hands-on testing. Real-world results.

For inferencing with your in-house AI chatbot, consider the Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors

Our testing showed that this server can be an excellent way for small organizations or departments to reap the benefits of AI without having to invest in GPUs

Many organizations are turning to inferencing with in- house AI chatbots that combine an LLM with the organization's own private data to improve efficiency, internal operations, and customer service. With a chatbot that you augment with company-specific data—such as FAQs, product catalogs, customer service transcripts, or internal process documentation—businesses can create a tailored, intelligent virtual assistant that understands their unique needs and workflows while still keeping their proprietary information secure. Being able to quickly answer questions can allow employees to spend less time on routine tasks, freeing them for innovation and even helping to boost the bottom line. Chatbots that utilize retrieval augmented generation (RAG) deliver responses that are especially accurate and current.

Upgrade to the new Supermicro H14 Hyper DP server powered by AMD EPYC™ 9965 processors and support 18 users simultaneously conversing with a local-data-augmented LLM*

Add AI chatbot functionality and continue to run batch general-purpose workloads during off hours

We've all heard about artificial intelligence requiring enormous computing resources, and many AI applications do in fact require servers equipped with powerful GPUs. As our testing showed, you can serve many users of an LLM augmented with your data on a Supermicro server with a powerful AMD EPYC CPU and no GPUs.

Imagine that the leaders of a company have decided to run an in-house AI chatbot using its own private data combined with RAG. They have a 4-year-old Supermicro H12 Ultra server powered by earlier processors in house, but the IT team suspects it is not up to the task and are wondering about an alternative. We used an end-to-end chatbot benchmark service called PTChatterly to explore the capabilities of this older server and a new Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors. In our tests, the chatbot utilized the Llama 3.2-3B-Instruct large language model [LLM] augmented by RAG with local data. For a server solution to support a given number of simultaneous users, the chatbot had to deliver a *complete* response to a majority of users within 10 seconds, though answers begin to appear in less than 1 second, so the response time feels much faster.

Using this criteria, the new Supermicro H14 Hyper DP server supported 18 simultaneous users posing a sequence of related questions. In most settings, only a fraction of employees would be asking questions of the chatbot at once, so this server is likely to comfortably support far more users in practice.

*with the Llama 3.2-3B-Instruct LLM and a median end-to-end response time of less than 10 seconds.

For inferencing with your in-house AI chatbot, consider the Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors December 2025

For inferencing with your in-house AI chatbot, consider the Supermicro Hyper DP H14 server powered by AMD EPYC 9965 processors

"Imagine that the leaders of a company have decided to run an in-house AI chatbot using its own

private data combined with RAG," explains the test report. "They have a 4-year-old Supermicro H12 Ultra server powered by earlier processors in house, but the IT team suspects it is not up to the task and are wondering about an alternative. We used an end-to-end chatbot benchmark service called PTChatterly to explore the capabilities of this older server and a new Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors. In our tests, the chatbot utilized the Llama 3.2-3B-Instruct large language model [LLM] augmented by RAG with local data. For a server solution to support a given number of simultaneous users, the chatbot had to deliver a complete response to a majority of users within 10 seconds, though answers begin to appear in less than 1 second, so the response time feels much faster."

The report continues: "The new Supermicro H14 Hyper DP server supported 18 simultaneous users posing a sequence of related questions. In most settings, only a fraction of employees would be asking questions of the chatbot at once, so this server is likely to comfortably support far more users in practice."

To learn more, read the full report at https://facts.pt/5kPexfD

About Principled Technologies, Inc.
Principled Technologies, Inc. is the leading provider of technology marketing and learning & development services.

Principled Technologies, Inc. is located in Durham, North Carolina, USA. For more information, please visit www.principledtechnologies.com.

Sharon Horton
Principled Technologies, Inc.
press@principledtechnologies.com
Visit us on social media:
LinkedIn
Facebook
YouTube
X

This press release can be viewed online at: https://www.einpresswire.com/article/872923996

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2025 Newsmatics Inc. All Right Reserved.