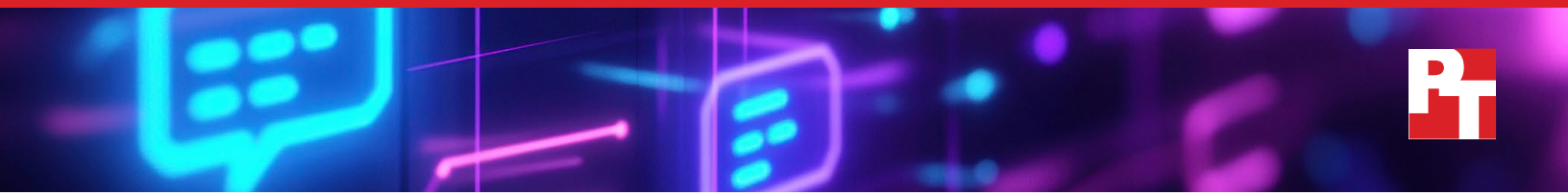


For inferencing with your in-house AI chatbot, consider the Supermicro H14 Hyper DP server powered by AMD EPYC 9965 processors

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.

Selecting the right hardware for inferencing with your in-house chatbot

While the term "AI" might bring to mind expensive GPU-filled servers, many smaller organizations and many departments of larger groups can comfortably host effective chatbots with more affordable hardware solutions using only powerful CPUs. To see how many simultaneous chatbot users your organization could expect to support using both older and current Supermicro servers powered by AMD EPYC processors, we used the PTChatterly testing service and the Llama 3.2-3B-Instruct LLM and RAG. The servers we tested were configured as follows:

4-year-old legacy server

Supermicro Ultra A+ Server AS -2124US-TNRP, part of the Ultra DP server family

- 2x AMD EPYC 7532 (32 cores, 2.4 GHz)
- 512 GB of RAM
- Ubuntu 24.10

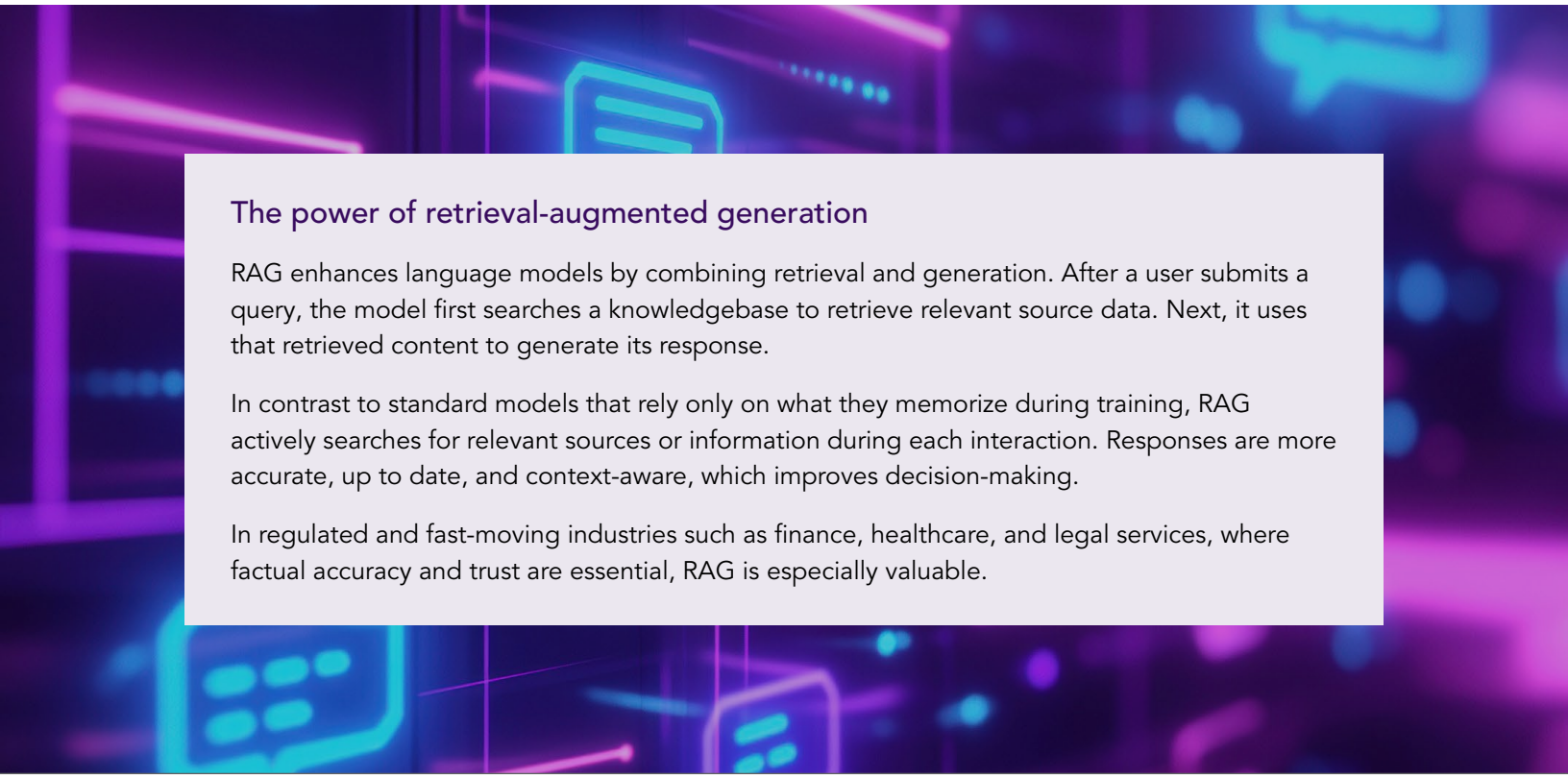
New server

Supermicro Hyper A+ Server AS -2126HS-TN, part of the H14 Hyper DP server family

- 2x AMD EPYC 9965 (192 cores, 2.25 GHz)
- 2,304 GB of RAM
- Ubuntu 24.10

The knowledgebase we used, which simulates the private data a company might include in its chatbot, consisted of text-only Airbnb rental data with details about home listings and customer reviews. This is a good representation of retail-style data because it includes product descriptions, pricing, and other information to help customers make decisions.

To learn more about how we tested, read the science behind the report.



The power of retrieval-augmented generation

RAG enhances language models by combining retrieval and generation. After a user submits a query, the model first searches a knowledgebase to retrieve relevant source data. Next, it uses that retrieved content to generate its response.

In contrast to standard models that rely only on what they memorize during training, RAG actively searches for relevant sources or information during each interaction. Responses are more accurate, up to date, and context-aware, which improves decision-making.

In regulated and fast-moving industries such as finance, healthcare, and legal services, where factual accuracy and trust are essential, RAG is especially valuable.

What our testing revealed: Upgrading to a new Supermicro H14 Hyper DP powered by AMD EPYC 9965 processors would let you support 18 simultaneous users

Our first test measured the chatbot performance of the 4-year-old Supermicro Ultra DP powered by AMD EPYC 7532 processors. This server could not support a single user with an acceptable response time. (To support even one user, it took 15 seconds to deliver a complete response.)

In our second test, we measured the performance of a new Supermicro H14 Hyper DP powered by AMD EPYC 9965 processors. We used the same testing parameters as with the legacy server and found that this server could supply complete answers to 18 simultaneous users within a median response time of 10 seconds. (See Figure 2.)

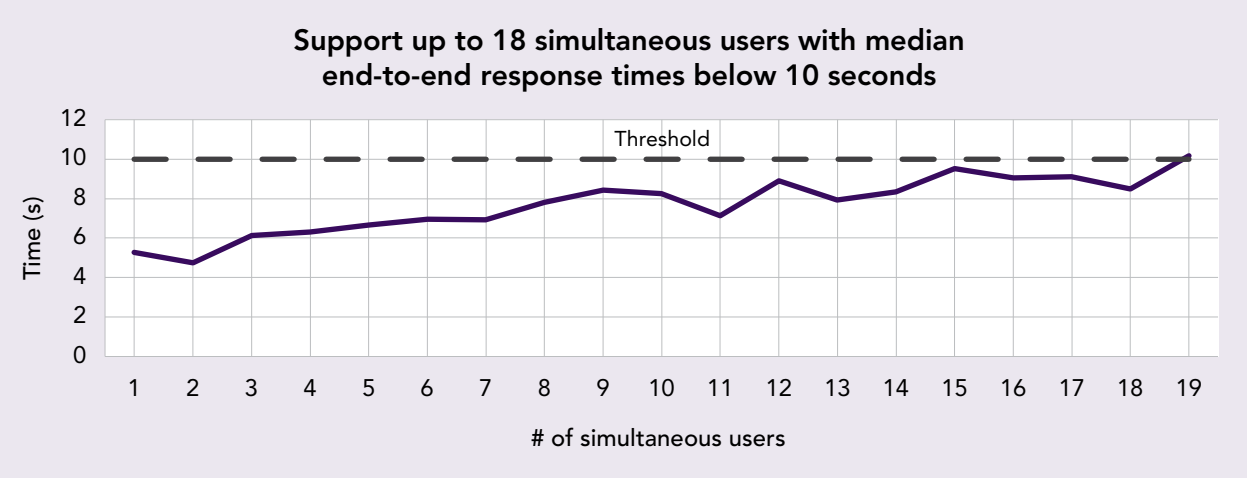
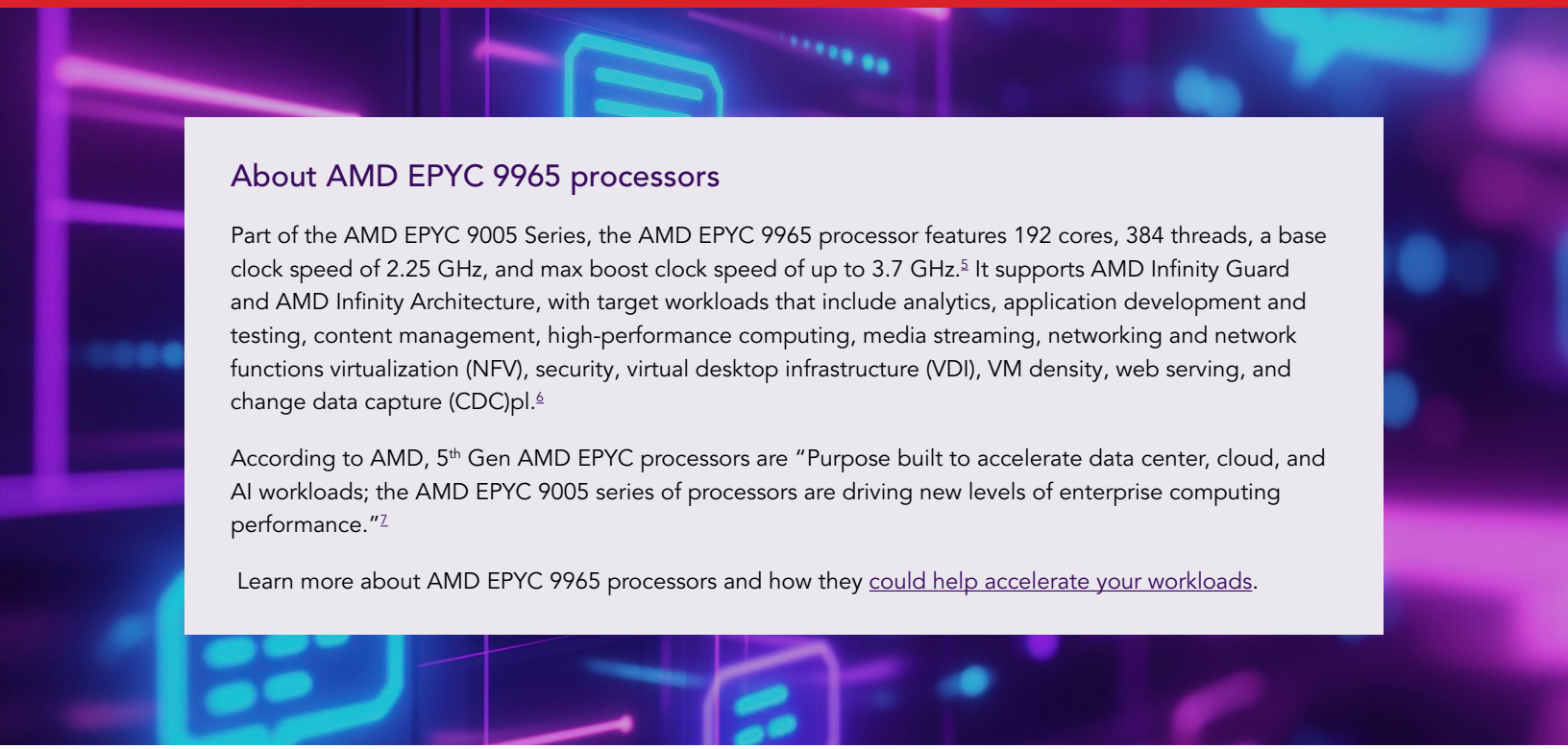


Figure 1: Response times for various numbers of simultaneous users that a new Supermicro Hyper DP H14 server powered by AMD EPYC 9965 processors supported with a median response time less than 10 seconds. Source: PT.

Depending on your number of prompts per day for your organization and user level of activity, the number of actual users the server could support would be much higher, perhaps over 100, assuming those 100 users had intermittent usage through a workday and the total number of simultaneous users never exceeded 18.



About AMD EPYC 9965 processors

Part of the AMD EPYC 9005 Series, the AMD EPYC 9965 processor features 192 cores, 384 threads, a base clock speed of 2.25 GHz, and max boost clock speed of up to 3.7 GHz.[5] It supports AMD Infinity Guard and AMD Infinity Architecture, with target workloads that include analytics, application development and testing, content management, high-performance computing, media streaming, networking and network functions virtualization (NFV), security, virtual desktop infrastructure (VDI), VM density, web serving, and change data capture (CDC)pl.[6]

According to AMD, 5th Gen AMD EPYC processors are "Purpose built to accelerate data center, cloud, and AI workloads; the AMD EPYC 9005 series of processors are driving new levels of enterprise computing performance."[7]

Learn more about AMD EPYC 9965 processors and how they could help accelerate your workloads.

AI by day and batch workloads by night: The Supermicro H14 Hyper DP featuring AMD EPYC 9965 processors as a double-duty solution

As companies adopt AI technologies to remain competitive, their existing workload requirements continue. To maximize the return on their investment in servers featuring powerful 5th Gen AMD EPYC processors, companies can adopt a 24/7 approach. During the workday, the servers can support prioritize chatbots and other AI-driven activities that require low latency. At night and on weekends, the servers can run batch data workloads that are resource-intensive but not time-sensitive.

Generative AI at work is here to stay

Many professionals are recognizing the power of AI to help them carry out routine and time-consuming tasks, and that number is only increasing. One survey of 2,000 U.S. knowledge workers found that 62 percent were interested in using AI for drafting emails, organizing spreadsheets, and taking meeting notes.[8] By setting up in-house chatbots powered by their own private, company-specific data—such as product details, service records, and operational manuals—small businesses or departments can capitalize on employees' desire to work smarter. Chatbots can also answer employee questions about HR policies, onboarding procedures, and IT concerns, improving access to information and reducing dependency on management or support teams.

Unlike public AI tools, in-house chatbots safeguard sensitive information by keeping it within the business's secure systems. They also offer customization options to reflect the brand's voice and operational nuances. As AI platforms become more accessible, even small businesses can implement secure, scalable solutions to enhance both team productivity and customer experiences—making them more agile and competitive in an increasingly digital marketplace.

In-house AI chatbots in action across industries

Law

To streamline operations and reduce administrative load, law firms are training in-house AI chatbots on their internal case files, legal templates, and procedural manuals. The chatbots, hosted securely on premises, help paralegals as they draft standard legal documents, answer common procedural questions, and locate relevant case law—potentially reducing research time substantially. Attorneys can use the tools to prepare client intake summaries and clarify filing requirements. By keeping chatbots in-house, firms can ensure confidentiality and compliance with data privacy standards, which are key concerns in the legal industry.

Chatbots can also have a big payoff. According to a 2025 Reuters report, "AI could free up 4 hours of a legal professional's time per week. For U.S. lawyers alone, the savings could translate into 266 million hours of increased productivity—approximately $100,000 in new billable time per lawyer each year."[9]

Real estate

To facilitate lead management and client interactions, real estate agencies are launching in-house AI chatbots trained on internal MLS data (to which only MLS members have access), property listings, client communications, and local zoning laws.[4] The chatbots automatically answer buyer and seller questions about listings, mortgage basics, open house schedules, and local market trends—freeing agents to focus on closing deals. Internally, staff use the chatbots to generate property descriptions, summarize contracts, and onboard new agents. Using a private, in-house system helps keep sensitive client and transaction data secure and lets the agency tailor the chatbot's responses to reflect local knowledge.

About the Supermicro H14 Hyper DP server family

The Supermicro H14 Hyper DP server family includes 2U, dual-socket servers featuring 5th Generation AMD EPYC 9005 Series processors. They has space for 24 DIMMs of DDR5-6400 memory and support NVMe®, SAS, SATA3, and M.2 drive options. According to Supermicro, "With the H14 Hyper systems, the flexible selection of density and storage capacity gives you a high-performance server for every purpose, including:

- Virtualization and cloud, including virtual desktop
- Infrastructure with GPU acceleration
- Scale-out, clustered software-defined storage"[10]

To learn more about the Supermicro H14 Hyper DP server family, visit https://www.supermicro.com/en/products/hyper.

About PTChatterly

PTChatterly is a benchmark service that measures a solution's performance for an in-house chatbot. It couples a full-stack AI implementation of an LLM, augmented with in-house data, with a testing harness that lets you determine how many simultaneous users the chatbot can support.

AI-assisted chatbots are by nature very complex, consisting of a variety of components working together. PTChatterly measures the entire pipeline of activity, from beginning to end. This is exactly the type of software stack you would need to augment an LLM with your own organization's data.

In contrast, most other LLM inference benchmarks do not simulate complete chatbots. For example, both the MLPerf Inference benchmark[1] and the Google inference benchmark[2] measure only LLM inference. The HuggingFace inference benchmarker is more complete; it uses real-world inputs for three real-world inferencing use cases and one type of synthetic input—but it lacks the critical in-house data component of PTChatterly.[3]

Figure 1 shows the test architecture of PTChatterly and the box on the following page explains its primary components. Learn more at PTChatterly.ai.

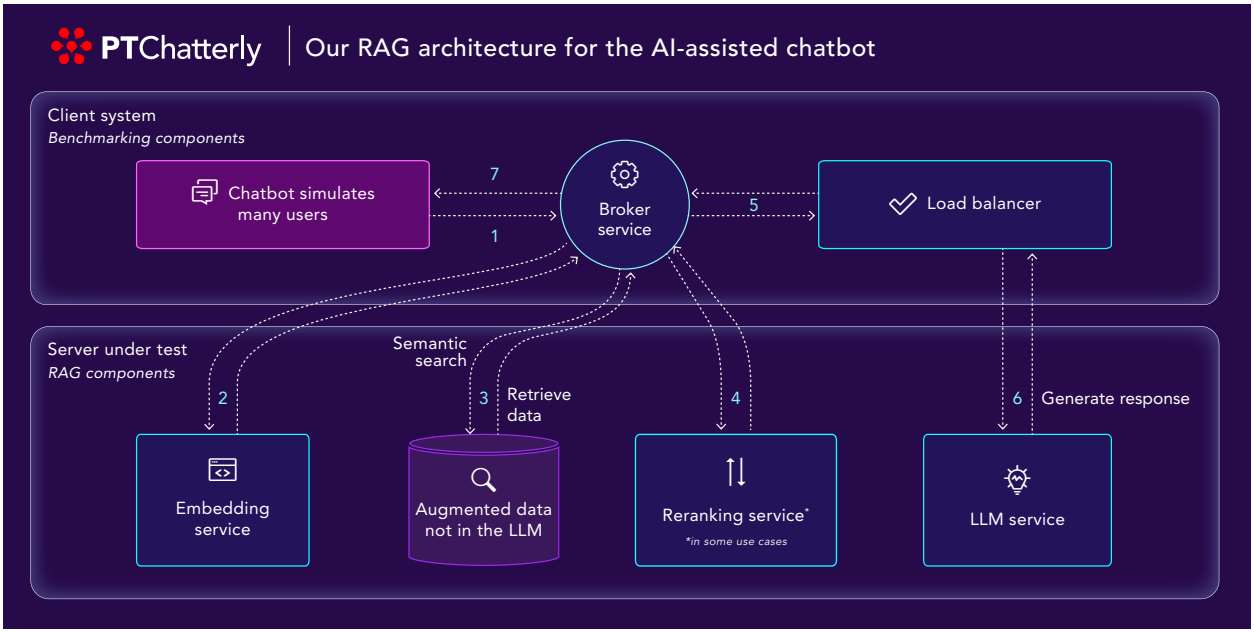


Figure 3: Architecture of the PTChatterly benchmark service. Source: PT.

The primary components of PTChatterly

Corpus or knowledgebase

The in-house dataset that the PTChatterly AI-assisted chatbot uses, along with the LLM, to answer questions

Bulk loader

A set of custom Python modules that reads and parses the corpus, as a set of documents or dataset; stores it in a vector database; and creates a vector index for semantic searches

Multi-threaded benchmark harness

Custom client harness written in Go that simulates multiple users having simultaneous conversations with the AI chatbot, provides corpus-specific conversations to the broker, and collects user-experience response times

Broker service

Custom Go code that orchestrates the data flow, receiving queries from clients and using the framework's services to generate a response

Embedding service

Handles calls to the embedding model via an API and enables efficient and more accurate semantic searching of the corpus

Vector database

An information store that provides efficient search for structured and unstructured documents, storing the original data, its embedding, and a vector index

Semantic search

Performs a non-keyword, sentence-based query of the documents in the vector database via an API endpoint

Reranking service (for some use cases)

Uses AI to compare the results of the semantic search of the corpus to the original query and to form a better context for answering the query; note that we did not use the reranking service for this testing

Large language model

An AI model that recognizes and generates text, utilizing the specific information the search gathered from the corpus and forming a response as a chatbot

Conclusion

Our testing demonstrates that a new Supermicro H14 Hyper DP server powered by two AMD EPYC 9965 processors can effectively meet the in-house AI chatbot needs of small organizations and departments within larger companies. We found a 4-year-old legacy Supermicro Ultra DP server with two AMD EPYC 7532 processors could not support a single user within an acceptable response time. In contrast, the new Supermicro H14 Hyper DP server solution, powered by two AMD EPYC 9965 processors, simultaneously provided 18 users with end-to-end responses in a median of 10 seconds, with answers beginning to appear within 1 second. Choosing this new Supermicro server solution lets you harness AI benefits securely and cost-effectively without having to invest in expensive GPU-based solutions. Speedy in-house AI chatbots that provide tailored, company-specific answers to common questions can help improve productivity, streamline operations, and maintain data privacy.

---

1. MLCommons, "MLPerf Inference 5.1: Benchmarking Small LLMs with Llama3.1-8B," accessed September 25, 2025, https://mlcommons.org/2025/09/small-llm-inference-5-1.
2. Github, "AI-Hypercomputer/inference-benchmark," accessed September 25, 2025, https://github.com/AI-Hypercomputer/inference-benchmark.
3. Github, "huggingface/inference-benchmarker," accessed September 25, 2025, https://github.com/huggingface/inference-benchmarker.
4. John Miley, "The New AI Agents Will Tackle Your To-Do List," accessed June 24, 2025, https://www.kiplinger.com/business/the-new-ai-agents-will-tackle-your-to-do-list.
5. AMD, "AMD EPYC™ 9965," accessed June 30, 2025, https://www.amd.com/en/products/processors/server/epyc/9005-series/amd-epyc-9965.html.
6. AMD, "AMD EPYC™ 9965."
7. AMD, "5th Generation AMD EPYC™ Processors," accessed June 24, 2025, https://www.amd.com/en/products/processors/server/epyc/9005-series.html.
8. SWNS, "Many are turning to AI to escape from repetitive tasks in the workplace, new study reveals," accessed June 24, 2025, https://nypost.com/2025/06/19/lifestyle/many-are-turning-to-ai-to-escape-from-repetitive-tasks-in-the-workplace-new-study-reveals/.
9. Thomson Reuters, "How AI is transforming the legal profession (2025)," accessed June 24, 2025, https://legal.thomsonreuters.com/blog/how-ai-is-transforming-the-legal-profession/.
10. Supermicro, "H14 Hyper Systems," accessed June 24, 2025, https://www.supermicro.com/datasheet/h14/datasheet_H14_Hyper.pdf.

Read the science behind this report at https://facts.pt/87Jj77P ▶

Principled Technologies®

Facts matter.®

This project was commissioned by AMD and Supermicro.