Run your in-house AI chatbot on an AMD EPYC 9534 processor-powered Dell PowerEdge R6615 server

Testing revealed that this server can be a great AI entry point for small organizations or departments, while upgrading to a CPU + GPU configuration could help them scale to support more users

Generative artificial intelligence (GenAI) is today's hottest topic, with decision-makers across industries wanting to take advantage of new and rapidly evolving capabilities. Many organizations are intrigued by one GenAI use case in particular: the in-house AI chatbot.[1] Combining a publicly available large language model (LLM) with an organization's own private corpus of data, these AI chatbots can provide accurate and specific customer support, help increase staff productivity with AI-powered search and content-generation capabilities, and supply valuable insights from an organization's existing data.

One of the first steps in building an in-house AI chatbot is selecting an appropriate computing solution to back it. Much of the discussion around AI in the media emphasizes how resource-intensive it can be, and some might assume that GPUs are a necessity. In fact, if only a small number of people will use your organization's chatbot simultaneously, a server platform with a powerful CPU can get the job done. Such a solution also gives you the option of scaling up by adding GPUs if your needs grow over time.

To investigate how different hardware options might support an in-house AI chatbot utilizing an LLM sized for this configuration and retrieval augmented generation (RAG), we tested two configurations of a single-socket Dell™ PowerEdge™ R6615 server. One had only a 64-core AMD EPYC™ 9534 processor and the other had this same processor plus an NVIDIA® L4 GPU.

Our testing revealed that a small organization or a department within an enterprise could use this solution in a CPU-only configuration to adopt a chatbot that supports a minimum of 9 simultaneous users at an affordable entry price. And because it is unlikely that all employees of an organization would be using the chatbot at once, this solution would effectively support larger departments comfortably. We also found that by adding a GPU to the server, the organization could scale up to support 23 or more simultaneous users as its requirements expanded.

Deploy an in-house chatbot on a server with only a CPU

Strong GenAI performance without a GPU

Support up to 9 simultaneous chatbot users

with the majority getting complete responses in under 5 seconds

Support more users with a clear and easy upgrade path

Sustain up to 23 simultaneous users by adding a GPU

Why build an in-house AI chatbot with a smaller LLM?

Publicly available GenAI chatbots have great potential to increase productivity and innovation; their usefulness, however, is limited in two key ways. First, such chatbots are trained on a large set of publicly available data rather than an organization's specific data, meaning that their answers, even when accurate, are not specific to an organization's needs. Second, the prompts that users enter into these tools typically feed back into the tools' training data, meaning that users cannot include private data in their prompts without breaking confidentiality. This is especially limiting in industries where data privacy is paramount, such as healthcare, legal services, and financial services.

In-house AI chatbots solve both of these problems. For both large businesses running chatbots at the edge and smaller organizations with just a handful of servers in a central location, a GenAI chatbot can help users find answers and create content efficiently. With a RAG architecture utilizing an organization's own private corpus of data, users can ask detailed questions and receive answers specific to their company's context. By keeping the entire solution local and private, organizations avoid feeding confidential data into a public tool.

In addition to hardware choices, organizations building their own GenAI chatbots must also determine what size LLM to use. LLMs with many parameters are likely to be highly accurate and better able to handle the biggest and most complex datasets.[2] However, they are also likely be much more resource-intensive and expensive to run; without the appropriate resources, they can be quite slow. LLMs with fewer parameters, such as Llama 3.2 1B, are much faster and can run on systems without multiple GPUs—ideal for edge devices.[3] A small model size might be the best choice for a small- to medium-sized business seeking to build an AI chatbot with a smaller deployment of servers. In addition, according to AMD, a small model size is typical for chatbot use cases.[4] For organizations using their GenAI chatbots for simple questions and conversations, a smaller LLM may be ideal.

Exploring real-world use cases for AI chatbots

Retail

In-house AI chatbots may be valuable for retail organizations seeking to expand—or reduce the cost of—their customer service options and explore new customer experiences. Many of us have likely already interacted with an AI-powered customer service chatbot, and training those chatbots on an organization's own data can improve their usefulness. Backed by a specific retailer's data—similar to the Airbnb dataset we used in testing—chatbots could provide more accurate and complete answers to stock queries, customer service complaints, and requests for recommendations. This isn't just to the benefit of the retailers; growing the presence of AI in shopping is what consumers want, too. In one 2024 survey, 55 percent of those surveyed wanted to use virtual assistants or bots as they shopped.[5]

Business intelligence

Individuals use public AI chatbots for everything from creative writing prompts to mapping out their families' weekly meal plans. For organizations with hundreds or thousands of employees and urgent business decisions to make every day, the stakes are considerably higher than what's for dinner. AI chatbots trained on an organization's private data can help support decision-makers by quickly offering up exactly the information they need and making connections between disparate sources of data. With the ability to draw on so many sources of data simultaneously, AI chatbots have the potential to be a smarter and faster version of institutional memory, empowering executives to use a wealth of past and current data to make informed decisions.

How we tested

To highlight how organizations could successfully utilize AMD EPYC processor-powered Dell PowerEdge servers for their GenAI chatbots, we used the PTChatterly testing service to compare two configurations of the same server:

- **Dell PowerEdge R6615 single-socket serve**r with a 64-core AMD EPYC 9534 processor
- **Dell PowerEdge R6615 single-socket server** with a 64-core AMD EPYC 9534 processor and an NVIDIA L4 GPU

About Dell PowerEdge R6615 servers

The Dell PowerEdge R6615 is a 1U, single-socket optimized rack server that Dell has designed to provide excellent performance-per-dollar value. According to Dell, this server uses an AMD EPYC 4th generation processor to deliver "up to 50% more core count per single socket platform in an innovative air-cooled blueprint."[6] It speeds access and transport of data by enabling DDR5 at 4800 MT/s memory and PCIe Gen5 with twice the speed of previous Gen4 and has room for up to two single-wide half-length GPUs.[7] Learn more at https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r6615-spec-sheet.pdf.

About AMD EPYC 9534 processors

With 64 cores, 128 threads, and a 256MB L3 cache, the AMD EPYC 9534 processor also supports AMD Infinity Guard technology. According to AMD, this 9004 Series processor can help deliver "exceptional time-to-results and energy efficiency for your business-critical applications" such as analytics, high-capacity data management, VDI, and app development.[8]

PTChatterly is a testing service that can help organizations size and understand a solution's performance for an in-house chatbot that utilizes RAG with a popular LLM and a private database of business information. In testing, we selected a number of variables, including the LLM, the corpus of data, the response time threshold, and the response time threshold percentile. We define the response time threshold as the longest acceptable amount of time the chatbot takes to fully answer a question—typically 1 to 10 seconds—and the response time threshold percentile as the proportion of questions that must complete below that threshold.

For this test, we used the following variables:

- **LLM:** We used the Llama 3.2 1B model, which "support[s] context length of 128K tokens and are state-of-the-art in their class for on-device use cases like summarization, instruction following, and rewriting tasks running locally at the edge."[9]
- **Corpus of data:** We used a text-only corpus of Airbnb rental data that includes details about home listings and reviews from customers, from which we scrubbed any obvious personal information (e.g., host names) before ingesting. This corpus is a good representation of retail-style data because it includes product descriptions, pricing, and other information to help customers make decisions.
- **Response time threshold** and **response time threshold percentile:** We chose a median response time threshold of 5 seconds, and 95th-percentile threshold of 10 seconds. This means the solution answered the majority of questions in their entirety in less than 5 seconds, and 95 percent of them in

less than 10 seconds. In all our tests, the 95th percentile threshold was never reached, and so we focus primarily on the median threshold for reporting. Note that this number of seconds is how long the complete answer takes to appear; the first characters appear in under 2 seconds, and within the total of 5 seconds, users would see a scrolling answer so they can read it as it completes.

For more detailed information on the system configurations we tested and how we utilized PTChatterly, see the science behind the report.

What we found

Support nine simultaneous users with only a CPU

We first ran PTChatterly on the Dell PowerEdge R6615 single-socket server with a 64-core AMD EPYC 9534 processor and no GPU. Figure 1 shows what we found: With one AMD processor, the PowerEdge R6615 could support up to nine simultaneous chatbot users with a 5-second response time.

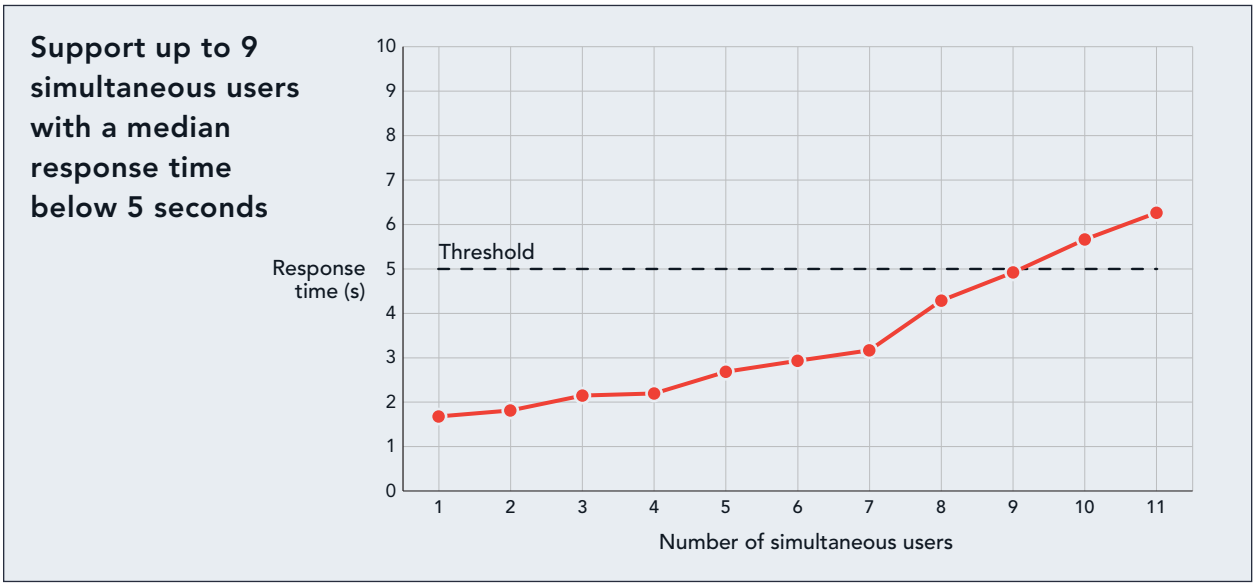

Figure 1: Response times for various numbers of simultaneous users that the CPU-only configuration of the Dell PowerEdge R6615 supported. Higher is better. Source: Principled Technologies.

This result positions this server configuration as a great choice for a small business or a department within a larger organization that wants to explore the potential of an in-house AI chatbot. For most organizations, it won't be necessary to build a chatbot that can support as many users as you have staff; to have all of your users engaging with the chatbot at the same time is unlikely.

Say your sales team wanted to access information on your different products to streamline their interactions with customers making inquiries by phone and email. If you had 12 team members, three-fourths of the group could use the chatbot at once, while the rest of the group was working on other tasks, and the chatbot would deliver complete answers to more than half of these questions within 5 seconds. Partial answers appear sooner, making the response time feel even faster.

Sustain 23 simultaneous users with a CPU and a GPU

In the second phase of our testing, we added a single NVIDIA L4 GPU to the Dell PowerEdge R6615 server and re-ran PTChatterly to measure any improvement the GPU might make, with the LLM and embedding service drawing on the GPU for most calculations. Using the same parameters as in the previous test, we found that this configuration supported 23 simultaneous users. (See Figure 2.)

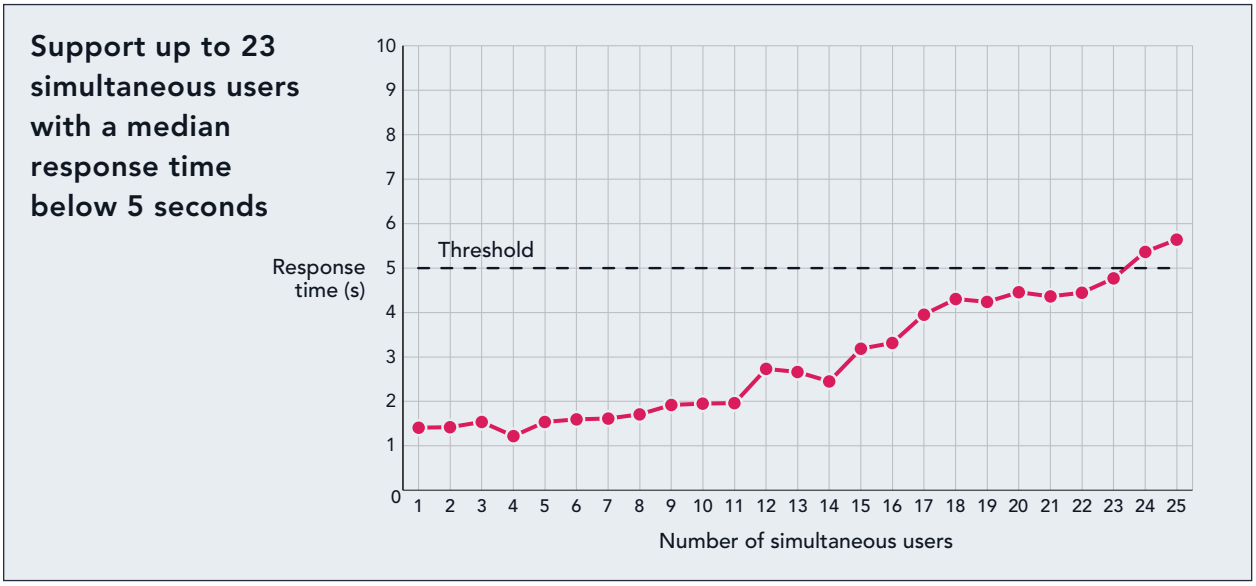

Figure 2: Response times for various numbers of simultaneous users that the CPU-plus-GPU configuration of the Dell PowerEdge R6615 supported. Higher is better. Source: Principled Technologies.

Figure 3 compares the two solutions' performance. The CPU+GPU configuration supported 23 simultaneous users, 2.5 times as many as the CPU-only configuration.

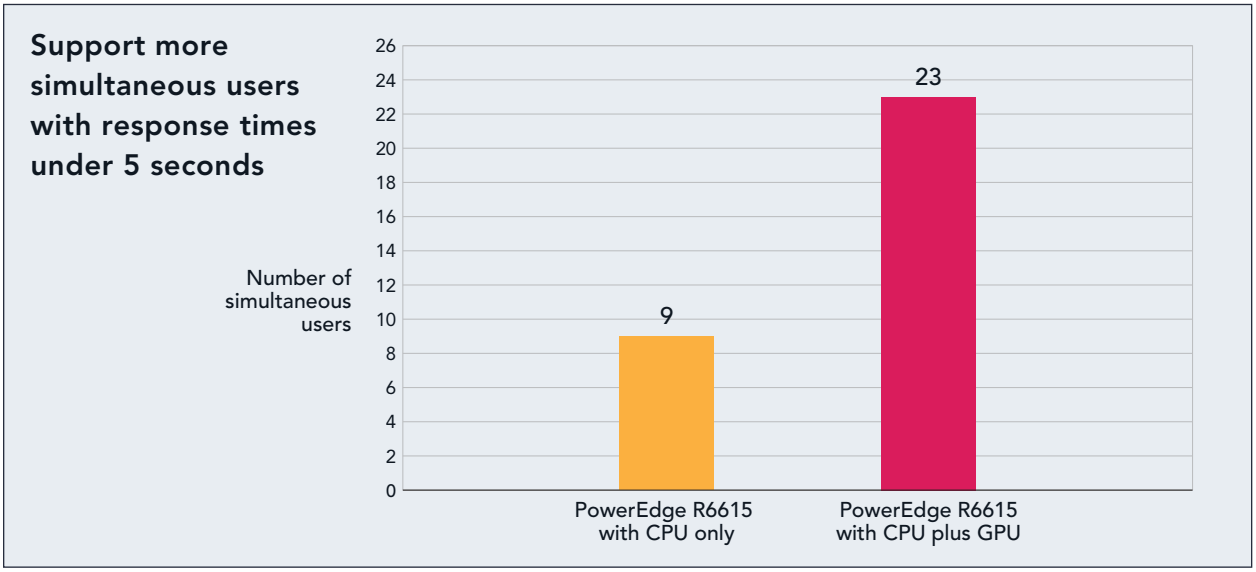

Figure 3: Number of simultaneous users the CPU-only and CPU-plus-GPU configurations of the Dell PowerEdge R6615 supported. Higher is better. Source: Principled Technologies.

PTChatterly

PTChatterly helps you size and understand a solution's performance for an in-house chatbot that supplements a popular LLM with a private dataset. 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 people the chatbot can support. Rather than reporting results in technical measurements that few users would understand, it provides a metric that is meaningful and simple to grasp: For example, it might say that the server under test supports 32 people having simultaneous conversations with a response time of 10 seconds or less. Learn more at PTChatterly.ai.

To return to our hypothetical example, imagine that the sales team found the product chatbot so useful that they wanted to expand access to the customer support team. By adding a GPU to the Dell PowerEdge R6615 server, the organization would enable the chatbot to support up to 14 additional users with the same response time. (As we have noted, given the unlikelihood that everyone would be asking questions of the chatbot at once, the actual number of users supported would probably be higher.)

Conclusion

GenAI is here to stay: 67 percent of business leaders in one Deloitte survey said that their organizations planned to invest further in GenAI initiatives,[10] and GenAI chatbots are an interesting possibility for companies of all sizes. While larger companies with dozens or hundreds of users probably will opt for solutions involving GPUs, small and medium businesses and groups within larger organizations can enjoy the benefits of GenAI using a server with just a CPU—while retaining the option of upgrading to a GPU when and if the need arises.

To demonstrate this, we tested two configurations of single-socket Dell PowerEdge R6615 servers for an in-house GenAI chatbot use case. The configuration with only a 64-core AMD EPYC 9534 processor supported nine simultaneous users with a response time threshold of 5 seconds and a response time threshold percentage of 95. When we added an NVIDIA L4 GPU, this server was able to support 23 simultaneous users. (Because a typical user is likely to engage with the chatbot for only part of their workday, these configurations are likely to support even larger numbers of effective users.)

These results demonstrate that the Dell PowerEdge R6615 server powered by a 64-core AMD EPYC 9534 processor is a strong choice for small and medium organizations and departments in enterprises who want to develop in-house GenAI chatbots at an affordable price point. When the number of likely simultaneous users is in the single digits, the CPU-only configuration of this server provides a good user experience. As the number of users grows, companies can easily scale the server's capabilities by adding a GPU and support more than twice as many users.

1. Samriddhi Srivastava, "12 companies that rolled out internal AI tools for employees," accessed January 23, 2025, https://www.peoplematters.in/article/technology/12-companies-that-rolled-out-internal-ai-tools-for-employees-40958.
2. Sean Michael Kerner, "What are large language models (LLMs)?" accessed January 14, 2025, https://www.techtarget.com/whatis/definition/large-language-model-LLM.
3. Meta, "Introducing quantized Llama models with increased speed and a reduced memory footprint," accessed January 14, 2025, https://ai.meta.com/blog/meta-llama-quantized-lightweight-models.
4. AMD, "Accelerate your Enterprise AI Inference Deployments with AMD EPYC™ Processors," accessed October 15, 2024, https://www.amd.com/en/products/processors/server/epyc/ai/9004-inference.html#tabs-66ffb7f055-item-f323db1036-tab.
5. IBM, "2024 Consumer Study: Revolutionize retail with AI everywhere" accessed October 16, 2024, https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-retail.
6. Dell Technologies, "PowerEdge R6615 Specification Sheet," accessed October 7, 2024, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r6615-spec-sheet.pdf.
7. Dell Technologies, "PowerEdge R6615 Specification Sheet."
8. AMD, "AMD EPYC™ 9534," accessed October 7, 2024, https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9534.html.
9. Meta, "Llama 3.2: Revolutionizing edge AI and vision with open, customizable models," accessed October 9, 2024, https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.
10. Deloitte, "Now decides next: Moving from potential to performance," accessed October 15, 2024, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q3.pdf.

Read the science behind this report at https://facts.pt/mlJino2 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.