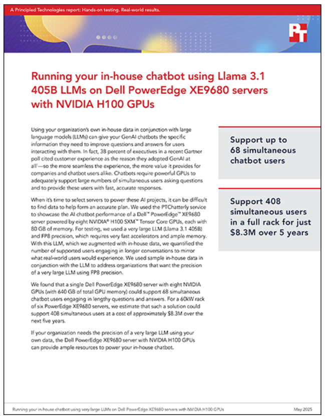
A Principled Technologies report: Hands-on testing. Real-world results.



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

Running your in-house chatbot using Llama 3.1 405B LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs

Using your organization's own in-house data in conjunction with large language models (LLMs) can give your GenAI chatbots the specific information they need to improve questions and answers for users interacting with them. In fact, 38 percent of executives in a recent Gartner poll cited customer experience as the reason they adopted GenAI at all[1]—so the more seamless the experience, the more value it provides for companies and chatbot users alike. Chatbots require powerful GPUs to adequately support large numbers of simultaneous users asking questions and to provide these users with fast, accurate responses.

When it's time to select servers to power these AI projects, it can be difficult to find data to help form an accurate plan. We used the PTChatterly service to showcase the AI chatbot performance of a Dell™ PowerEdge™ XE9680 server powered by eight NVIDIA® H100 SXM™ Tensor Core GPUs, each with 80 GB of memory. For testing, we used a very large LLM (Llama 3.1 405B) and FP8 precision, which requires very fast accelerators and ample memory. With this LLM, which we augmented with in-house data, we quantified the number of supported users engaging in longer conversations to mirror what real-world users would experience. We used sample in-house data in conjunction with the LLM to address organizations that want the precision of a very large LLM using FP8 precision.

We found that a single Dell PowerEdge XE9680 server with eight NVIDIA GPUs (with 640 GB of total GPU memory) could support 68 simultaneous chatbot users engaging in lengthy questions and answers. For a 60kW rack of six PowerEdge XE9680 servers, we estimate that such a solution could support 408 simultaneous users at a cost of approximately $8.3M over the next five years.

If your organization needs the precision of a very large LLM using your own data, the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs can provide ample resources to power your in-house chatbot.

Support up to 68 simultaneous chatbot users

Support 408 simultaneous users in a full rack for just $8.3M over 5 years

Running your in-house chatbot using very large LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs May 2025

The science behind the report:

Running your in-house chatbot using very large LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs

This document describes what we tested, how we tested, and what we found. To learn how these facts translate into real-world benefits, read the report Running your in-house chatbot using very large LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs.

We concluded our hands-on testing on March 23, 2025. During testing, we determined the appropriate hardware and software configurations and applied updates as they became available. The results in this report reflect configurations that we finalized on March 22, 2025 or earlier. Unavoidably, these configurations may not represent the latest versions available when this report appears.

Our results

To learn more about how we have calculated the wins in this report, go to http://facts.pt/calculating-and-highlighting-wins. Unless we state otherwise, we have followed the rules and principles we outline in that document.

Table 1: Results of our PTChatterly testing.

	Dell PowerEdge XE9680 (8x NVIDIA H100 GPUs)
Results	
Max users reached	68
Cut off on question response time (s)	30
Threshold percentage	95%
Median threshold (s)	15
CPU AVG %UTIL @ max users	8.6
GPU AVG %UTIL @ max users	99.6
Idle power (W)	4,802
Active power (W)	6,534.1

	Dell PowerEdge XE9680 (8x NVIDIA H100 GPUs)
Test parameters	
LLM version	nvcr.io/nim/meta/llama-3.1-405b-instruct:1.3.0
LLM model	“meta/llama-3.1-405b-instruct(tensorrt_llm-h100-fp8-tp8-pp1-latency)”
LLM instances	1
Corpus	airbnb
Corpus # of listings	487974
Corpus total text size (MB)	1952.69 MB
# of questions in conversation	1-9
Think time (s)	2.5
Think time range (s)	1
Ladder delay(s) (divided by number of users for each run)	10
User delay basis (s)	30
Max user delay (s)	0.1
Maximum # of tokens per prediction	128
Prompt context size	8,192
Threads (CPU only)	Default
Full options for LLM service	“NIM_ENABLE_KV_CACHE_REUSE=1NIM_MAX_MODEL_LEN=8192--shm-size 16G”
Benchmark mode	Timed Ladder
Begin run with # users	1
Increment by # users	1
Broker endpoint	http://100.67.190.137:8072/prompts
Broker version	v2025.03.2.10 (22 March 2025)
Client version	v2025.03.2.09 (18 March 2025)

Cost analysis

In Tables 2 through 7, we detail the various costs and assumptions we used to create our example 5-year TCO for 6x Dell PowerEdge XE9680 servers each with eight NVIDIA H100 SXM GPUs. As with any TCO calculation, your expected costs will vary depending on several factors.

Total systems required for equivalent performance

Table 2: The results of our PTChatterly testing extrapolated to the number of users a rack of six servers could support.
Source: Principled Technologies.

	6x Dell PowerEdge XE9680 *with NVIDIA H100 GPUs*	**Notes**
Servers		
Total number of systems in one rack	6	Number of servers that fit in a 60kW rack.
Total number users in one rack	408	68 simultaneous AI chatbot users per server times 6 servers

Hardware costs

Table 3: Estimated hardware costs for 6x Dell PowerEdge XE9680 with NVIDIA H100 GPUs. Source: Principled Technologies.

	6x Dell PowerEdge XE9680 *with NVIDIA H100 GPUs*	**Notes**
Configuration details		
System	1x PowerEdge XE9680	
CPU	2x Intel Xeon Platinum 8468	
Memory	32x 64GB DDR5	
OCP NIC	1x Broadcom Gigabit Ethernet BCM5720	
Power supply	6x 2800W PSU	
NVMe	8x 3.84 NVMe U.2	
Boss	2x 480GB M.2	
GPU	NVIDIA H100 8-way	
Total cost for one system	**$1,343,725.49**	
Total cost for six systems (1 rack)	**$8,062,352.94**	

Power and cooling costs

Table 4: Power and cooling cost estimates over five years for 6x Dell PowerEdge XE9680 with NVIDIA H100 GPUs. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 *with NVIDIA H100 GPU*	**Notes**
Cost per kWh	0.1701	Avg price of electricity Nov 2024: https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_es1a
Hours in year	8,760	
Percentage time under load/active (remainder idle)	0.50	Assume 50% active (12 hours per day), but this can be set to any percentage of total operating time we feel comfortable assuming the servers would be under full load.
Power usage per server (watts-active)	6,534	
Power usage per server (watts-idle)	4,802	
Typical watts	5,668	Calculation: (% time active * watts-active) + (% time idle * watts-idle)
Annual kWh per server	49651.68	
Total annual energy cost per server	$8,445.75	
Total 5y energy cost per server	**$42,228.75**	
Total 5y cost for six servers	**$253,372.52**	

Datacenter space costs

Table 5: Data center space cost estimates over five years for 6x Dell PowerEdge XE9680 with NVIDIA H100 GPUs. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 *with NVIDIA H100 GPU*	**Notes**
Rack units per server (u)	6	
Annual data center costs per rack (42u)	$2,000.00	PT estimate. Could range from $1,000-$3,000+ depending on location, bandwidth, infra, and other factors. See https://cyfuture.cloud/kb/colocation/how-much-does-renting-rack-space-cost-key-factors-to-consider. Including infrastructure requirements, a full cabinet is assumed.
Annual cost for all required systems	**$2,000.00**	
Total 5y cost for all required systems	**$10,000.00**	

Maintenance and administration costs

Table 6: Maintenance and administration cost estimates over five years for 6x Dell PowerEdge XE9680 with NVIDIA H100 GPUs. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 *with NVIDIA H100 GPU*	**Notes**
Number of servers per IT admin	100	PT estimate. Could vary significantly based on management/infrastructure factors.
Number admins needed for all required systems	0.06	Total required systems divided by number of servers per IT admin.
Average salary of an administrator	$100,580.00	Average for network and computer systems administrator, BLS May 2023 https://www.bls.gov/oes/current/oes_nat.htm
Burden rate	0.2962	Burden rate for private industry workers, BLS Sept 2024 https://www.bls.gov/news.release/ecec.nr0.htm.
Average burdened salary	$130,368.90	Average salary * (1 + burden rate)
Annual administration cost	$7,822.13	
Total 5y administration cost	**$39,110.67**	

Total operational costs

Table 7: Operational cost estimates over five years for 6x Dell PowerEdge XE9680 with NVIDIA H100 GPUs. Source: Principled Technologies.

	6 x Dell PowerEdge XE9680 *with NVIDIA H100 GPU*	**Notes**
Number of systems required	6	
Total system cost	$8,062,352.94	
Total 5y power cost	$253,372.52	
Total 5y data center space cost	$10,000.00	
Total 5y maintenance cost	$39,110.67	
Total 5y costs	**$8,364,836.13**	

System configuration information

Table 8: Detailed information on the systems we tested.

System configuration information	Dell PowerEdge XE9680 server	Dell PowerEdge R660 client
BIOS name and version	Dell 2.4.4	Dell 2.4.4
Non-default BIOS settings	N/A	N/A
Operating system name and version/build number	Ubuntu 22.04.5 LTS Kernel version 5.15.0-133-generic	Ubuntu 22.04.5 LTS Kernel version 5.15.0-133-generic
Date of last OS updates/patches applied	2/22/25	2/24/25
Power management policy	Performance	Performance
Processor		
Number of processors	2	1
Vendor and model	Intel® Xeon® Platinum 8468	Intel Xeon Gold 6448Y
Core count (per processor)	48	32
Core frequency (GHz)	2.1	2.1
Max turbo frequency (GHz)	3.8	4.1
Stepping	6	8
Memory module(s)		
Total memory in system (GB)	2048	512
Number of memory modules	32	16
Vendor and model	Hynix® HMCG94AEBRA109N	Hynix HMCG88AEBRA107N
Size (GB)	64	32
Type	PC5-38400	PC5-38400
Speed (MT/s)	4,800	4,800
Speed running in the server (MT/s)	4,400	4,800
Storage controller		
Vendor and model	N/A	PERC H965i Front (Embedded)
Cache size (GB)	N/A	8GB
Firmware version	N/A	8.8.0.0.18-31
Driver version	N/A	8.0.0.69.0
Local storage		
Number of drives	1	4
Drive vendor and model	Dell NVMe® P5600 MU U.2 3.2TB	Samsung® PM1645a MZILT800HBHQ0D3
Drive size (GB)	2,980	800
Drive information (speed, interface, type)	PCIe gen4x4, NVMe, SSD	12Gbps, SAS, SSD
Purpose	OS, application	OS, application

System configuration information	Dell PowerEdge XE9680 server	Dell PowerEdge R660 client
Network adapter		
Vendor and model	Broadcom® BCM5720 Gigabit Ethernet	Broadcom BCM5720 Gigabit Ethernet
Number and type of ports	2 x 1GbE	2 x 1GbE
Driver version	tg3 5.15.0-133-generic	tg3 5.15.0-133-generic
Accelerators		
Vendor and model	NVIDIA® H100	N/A
Number	8	N/A
Memory type	HBM3	N/A
Memory size	80GB	N/A
Memory bandwidth	3.35 TB/s	N/A
Power cap limit	700W	N/A
Connection	SMX5, PCIe gen5, 16 lanes	N/A
Driver	570.86.15	N/A
Cooling fans		
Vendor and model	Dell Gold	Dell Gold
Number of cooling fans	32	16
Power supplies		
Vendor and model	Dell 0FX9WJA00 Dell 01PDR6A00	Dell 07DWXYA01
Number of power supplies	5 1	2
Wattage of each (W)	2,800	1,400

How we tested

Software versions

- OS: Ubuntu Server 22.04.5 LTS (kernel version 5.15.0-133-generic)
- NVIDIA driver: 570.86.15
- CUDA toolkit: 12.8.57-1
- Docker: docker-ce 5:28.0.0-1~ubuntu.22.04~jammy
- HAproxy: 2.4.24-0ubuntu0.22.04.1
- VectorDB: qdrant/qdrant:v1.13.4
- VectorDB model: sentence-transformers/msmarco-distilbert-cos-v5
- VectorDB precision: FP32
- Embedding server: michaelf34/infinity:0.0.67
- Embedding server model: sentence-transformers/msmarco-distilbert-cos-v5
- Embedding server precision: FP32
- Go Lang version: 1.23.1
- PTChatterly client: v2025.03.2.09 (18 March 2025)
- PTChatterly broker: v2025.03.2.10 (22 March 2025)
- LLM server (NVIDIA GPU): nvcr.io/nim/meta/llama-3.1-405b-instruct:1.3.0
- LLM model (NVIDIA GPU): meta/llama-3.1-405b-instruct (tensorrt_llm-h100-fp8-tp8-pp1-latency)
- LLM model precision: FP8
- Nmon: 16q

Configuring the system under test (SUT)

1. Install Ubuntu Server 22.04.5 LTS, making sure that sshd is included and running.
2. Configure Ubuntu OS following the instructions in the section Configuring Ubuntu 22.04.
3. Install Docker following the instructions in the section Installing Docker on Ubuntu.
4. Install NVIDIA drivers and CUDA Toolkit following the instructions in the section Installing NVIDIA drivers.
5. Install the nmon resource monitoring tool on the SUT. The resulting binary file(s) can be copied to the remaining systems:

```
# Remove any preinstalled versions of nmon
sudo apt remove -y nmon
# Build requirements:
#  ncurses-dev
#  NVIDIA GPU drivers for SUTs with GPUs (already installed)
#  gcc, make, and wget (already installed)
sudo apt install ncurses-dev
# Download source files
mkdir nmon_build
cd nmon_build
wget 'https://sourceforge.net/projects/nmon/files/lmon16q.c/download' -O lmon.c
wget 'https://sourceforge.net/projects/nmon/files/makefile/download' -O Makefile
```

   a. For SUTs with NVIDIA GPUs, build and install nmon with GPU support:

```
# patch the Makefile so that it finds and uses the NVIDIA GPU management library

make gpu
sudo install nmon_X86_Ubuntu22_16q_gpu /usr/local/sbin/nmon
```

   b. For SUTs without NVIDIA GPUs, build and install nmon without GPU support:

```
make
sudo install nmon_X86_Ubuntu22_16q /usr/local/sbin/nmon
```

6. Install `nmonchart`, the `nmon` parser and chart creator, on the client system:

```
# Set up nmonchart in a convenient directory on the client system
# The script uses the Korn shell
sudo apt install ksh
wget 'https://sourceforge.net/projects/nmon/files/nmonchart42.tar/download' -O nmonchart42.tar
tar -xvf nmonchart42.tar ./nmonchart
# This fix is only needed for parsing GPU data from systems with more than 2 GPUs
cp nmonchart nmonchat-orig
sed -i 'sZ/,0,0/Z/,0,0,0,0,0,0,0,0,0/Z' nmonchart
# Install nmonchat

sudo install nmonchart /usr/local/sbin/nmonchart
```

Ingesting Airbnb data into the vector database

1. Create a new directory on the SUT:

```
mkdir ingest
cd ingest
```

2. Create and activate a new virtual Python environment:

```
sudo apt install python3-venv
python3 -m venv .ingest
. .ingest/bin/activate
```

3. Copy the following files to the directory above:

- Files
    - Property listings file dataset: `AirbnbProps-20240830.json.gz`
    - Python ingestion script: `ingestAirBnB.py`

4. Create symbolic link to property listings file dataset:

```
ln -s AirbnbProps-20240830.json.gz AirbnbProps.json.gz
```

5. Add Python packages:

```
pip3 install -U torch --index-url https://download.pytorch.org/whl/cu126
pip3 install -U sentence-transformers
pip3 install -U qdrant-client
```

Configuring the associated client / test harness server

1. Install Ubuntu Server 22.04.5 LTS, making sure that `sshd` is included and running.
2. Configure Ubuntu OS following the instructions in Appendix 1.
3. Install Docker following the instructions in Appendix 2.
4. Install HAProxy:

```
sudo apt update
sudo apt install haproxy
sudo systemctl disable haproxy
```

5. Install the PTChatterly broker and client.

Configuring Ubuntu 22.04

After installation, perform these configuration steps. We assume the login for the non-root user is ptuser.

1. Enable password-less sudo:

```
echo "$USER ALL=(ALL:ALL) NOPASSWD: ALL" | sudo tee /etc/sudoers.d/$USER
sudo chmod 640 /etc/sudoers.d/$USER
```

2. Set time zone; e.g.,

```
sudo timedatectl set-timezone America/New_York
```

3. Extend the root LVM filesystem, if necessary:

```
sudo lvextend -r -l +100%FREE /dev/ubuntu-vg/ubuntu-lv
```

4. Disable unattended package updates:

```
sudo systemctl stop unattended-upgrades.service
sudo systemctl disable unattended-upgrades.service
```

5. Modify the value of the unattended-upgrade variable to 0 in file /etc/apt/apt.conf.d/20auto-upgrades:

```
APT::Periodic::Update-Package-Lists "1";
APT::Periodic::Unattended-Upgrade "0";
```

6. Install the latest updates:

```
sudo apt update
sudo apt upgrade
```

7. Install standard Ubuntu packages you will need in subsequent steps:

```
sudo apt install ca-certificates curl wget lsb-release sysstat smartmontools vim nmon numactl
```

8. Reboot the system:

```
sudo shutdown -r now
```

Installing Docker on Ubuntu 22.04

1. Remove any previous Docker packages:

```
for pkg in docker.io docker-doc docker-compose docker-compose-v2 \
    podman-docker containerd runc; do \
  sudo apt remove $pkg
done
```

2. Add Docker's official GPG key:

```
sudo install -m 0755 -d /etc/apt/keyrings
sudo curl -fsSL https://download.docker.com/linux/ubuntu/gpg -o /etc/apt/keyrings/docker.asc
sudo chmod a+r /etc/apt/keyrings/docker.asc
```

3. Add the Docker repository to the system:

```
echo \
  "deb [arch=$(dpkg --print-architecture) signed-by=/etc/apt/keyrings/docker.asc] https://download.
docker.com/linux/ubuntu \
  $(. /etc/os-release && echo "$VERSION_CODENAME") stable" | \
  sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
sudo apt update
```

4. Install Docker CE:

```
sudo apt update
sudo apt install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin
```

5. Add current user to Docker permissions group, and add docker group to the user’s current session:

```
sudo usermod -aG docker $USER
newgrp docker
```

6. Confirm Docker function under non-root user:

```
docker run --rm hello-world
```

7. Enable Docker services:

```
sudo systemctl enable --now docker.service
sudo systemctl enable --now containerd.service
```

Installing NVIDIA drivers and CUDA Toolkit

1. Add the NVIDIA repo to the system:

```
wget https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64/cuda-
keyring_1.1-1_all.deb
sudo dpkg -i cuda-keyring_1.1-1_all.deb
sudo apt update
sudo apt install -y nvidia-headless-570
sudo apt install -y nvidia-utils-570 nvidia-container-toolkit nvtop
```

2. Reboot the system, which is necessary because the kernel's default module list has been updated:

```
sudo shutdown -r now
```

3. Validate GPU functionality (optional):

```
docker run --rm --gpus all nvcr.io/nvidia/k8s/cuda-sample:nbody nbody -gpu -benchmark
```

LLM parameters:

```
# Adjust --tensor-parallel-size  to the number of GPUs you wish to allocate to your vLLM instance
--env NIM_ENABLE_KV_CACHE_REUSE=1
--env NIM_MAX_MODEL_LEN=8192
--shm-size 16G
```

Read the report at https://facts.pt/8Dabhdx ▶

This project was commissioned by Dell Technologies.

Principled Technologies®

Facts matter.®

Principled Technologies is a registered trademark of Principled Technologies, Inc.
All other product names are the trademarks of their respective owners.

DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:
Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.'s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.'s testing. Customer's sole and exclusive remedies are as set forth herein.