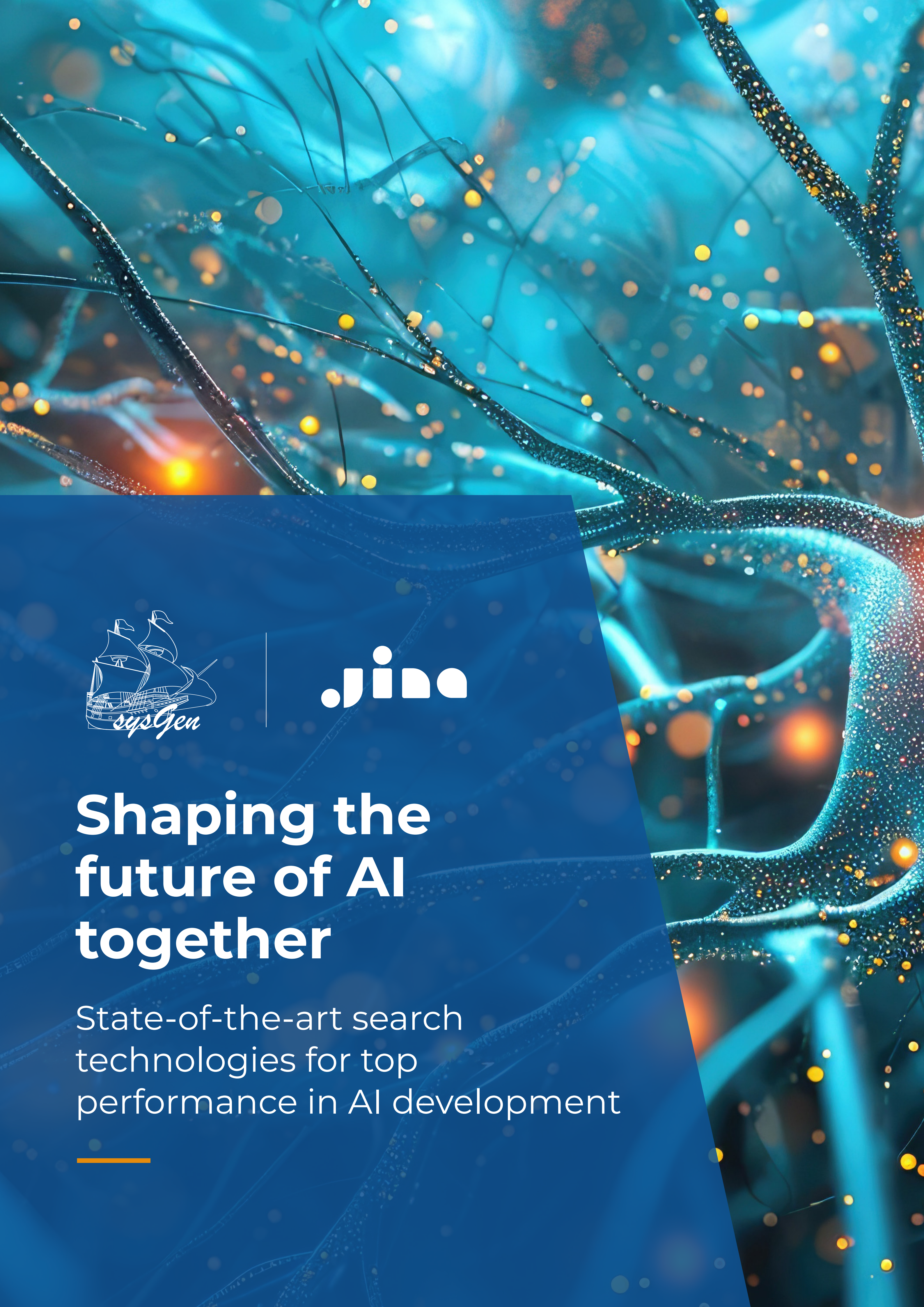


Shaping the future of AI together

State-of-the-art search technologies for top performance in AI development

Breakthrough in AI

Jina AI and the revolution of information retrieval

The problem

In an era of technological progress and ever-increasing demand for innovation, AI developers face demanding challenges. Maximilian Werk, Head of Engineering at Jina AI, a company that specializes in improving the performance of neural searches, knows these challenges first-hand. With a background in mathematics and years of experience in machine learning, he has a deep understanding of the needs of his industry.

Jina AI strives to facilitate the development of AI models that can be used in various application areas. However, like many companies, Jina AI faced the challenge of accessing powerful hardware, especially GPUs. The decision to buy GPUs instead of renting them was a significant step, yet one that came with specific obstacles.

Main research areas

Jina AI, a Berlin-based start-up founded in 2020, is a leading player in artificial intelligence. Company's focus is on improving internet search, in particular through its Retrieval Augmented Generation (RAG) system, which aims to fundamentally change the way information is searched and used. Since its inception, Jina AI has strived to improve the performance of neural searches, by developing both embedding and reranking models. These are designed to facilitate the search for relevant information and optimize the order of search results.

Areas of application explained

Neural Search

An intelligent search function that extracts relevant information from large data sets, acting like a personal assistant that finds exactly what you're looking for.

Deduplication

A duplicate detection technology that identifies and removes duplicate files or content helps maintain order and save storage space, acting like a digital detective.

Classification

An automatic classification function that sorts content into different categories, similar to how emails are automatically sorted into folders based on their content.

With these focal points, Jina AI is at the forefront of AI development and is continuously driving innovation.

Jina AI's primary customers are application developers, particularly those working on RAG programs that enhance modern chatbots.

What does RAG mean?

What is Retrieval Augmented Generation (RAG)?

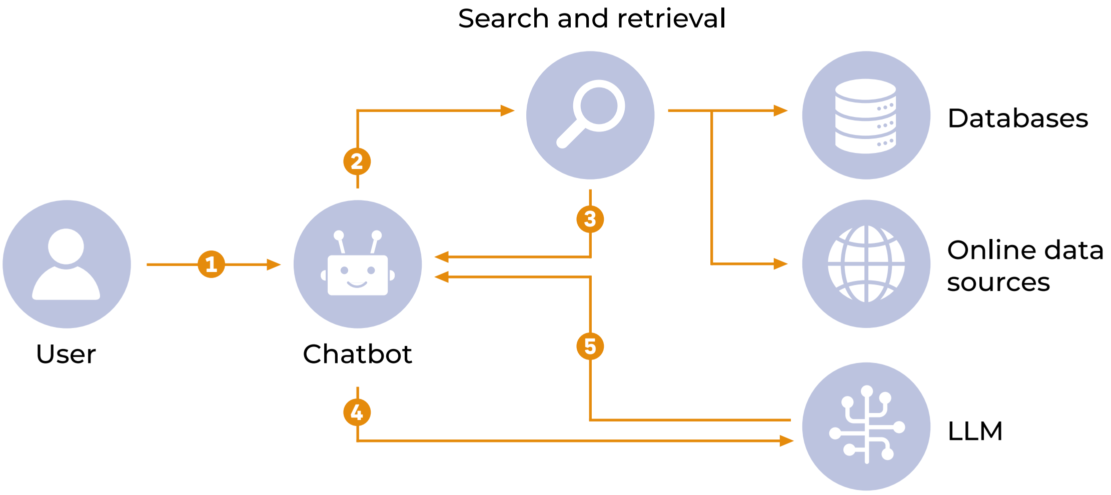
RAG combines information retrieval with artificial intelligence to provide precise answers. This technology is commonly used to efficiently search through large datasets and generate specific, well-founded responses.

Is RAG the same as generative AI?

No, RAG is not the same as generative AI. RAG is a technique that can deliver more precise results for queries than a generative language model on its own. While generative AI relies on the data it has been trained on, RAG enhances accuracy by incorporating additional external knowledge to provide more accurate answers.

How is RAG used by generative AI?

RAG is utilized by generative AI in the following way: company data is embedded in a knowledge repository and transformed into vectors stored in a vector database. When a query is made, the vector database retrieves relevant contextual information. This information is then sent along with the query to the large language model, which uses the context to generate a more timely, accurate, and contextualized response.



Advantages of RAG? ↓

What are the Advantages of RAG?

Retrieval Augmented Generation (RAG) provides several benefits that significantly enhance the quality of responses from generative AI systems. These advantages extend beyond what a large language model (LLM) can offer and include:

Timeliness of the Information

RAG has access to potentially more current data than what was used to train a large language model (LLM). This capability helps minimize the risk of „hallucinations“-incorrect or outdated answers resulting from a lack of information.

Continuous Updating

RAG knowledge repository can be continuously updated at minimal cost, ensuring that the information remains current and relevant.

Contextualization

The data in the RAG knowledge repository can be more specific and contextual compared to that in a generalized LLM, resulting in more accurate and precise answers.

Error Correction

The sources of information in the RAG vector database are identifiable, allowing for the correction or removal of any incorrect data. This enhances the reliability of the generated responses.

Guardrails

Guardrails can be established to ensure that the generated responses adhere to the desired linguistic style and avoid inappropriate or incorrect content.

The challenges

Transition from Renting to Owning
GPU Compute Power

The demands on AI developers such as Jina AI are therefore challenging and diverse, especially at a time of rapid technological progress and increasing demand for innovation and ever higher standards. Access to powerful hardware such as GPUs is of crucial importance.

The previous method of renting GPUs was not sufficient for Jina AI and was not sustainable in the long term. The rental costs were much higher than the purchase costs, especially when fully utilized for at least 12 months. In addition, there were often problems renting certain GPU models at short notice.

Jina AI therefore decided to buy the GPUs. This was not only more economical, but also strategically better for long-term planning and cost efficiency.

When purchasing the GPUs, Jina AI faced a major challenge: the desired NVIDIA H100s GPUs were difficult to obtain. They also lacked experience in setting up and maintaining their own hardware. To overcome these hurdles, Jina AI brought sysGen on board. sysGen is an expert in hosting and configuring high-performance hardware.

Innovative IT Partnership

Jina AI and sysGen

sysGen is a Bremen-based company, located in beautiful northern Germany. Equipped with the latest technology and driven by a highly motivated and international team, which is always up to date. According to the premise – work must be fun – sysGen works effectively in a team with a lot of motivation and respectful interaction with each other.

As a solution-oriented and manufacturer-independent IT supplier for industry, trade, research, and education, sysGen specializes in the development, production, and distribution of high-quality system solutions, servers, workstations, and components. Their offerings include both system and application software, as well as professional services.

The extensive IT portfolio includes revolutionary and future-oriented solutions for the areas of artificial intelligence (AI/deep learning), high-performance computing, high-availability computing (HA), software defined storage and networks, virtualization and cloud computing. Additional services such as on-site installation with integration into existing IT infrastructures and employee training for hardware and software are offered for all solutions and system deliveries. The offers are also preceded by qualified consulting, and a fee-based, Europe-wide IT service rounds off the range of services.

As a partner of renowned companies such as NVIDIA, Supermicro, Gigabyte, Intel, DDN, GRAIDtech, sysGen offers innovative IT solutions.

The partners at a glance:

- Company: sysGen GmbH
  Head office: Bremen, Germany
  Partner: NVIDIA and Supermicro Elite Partner
  Industry: Multi-industry
  Focus area: developing, producing and distributing high-quality system solutions, servers, workstations and components

- Company: Jina AI GmbH
  Foundation: 2020
  Head office: Berlin, Germany
  Industry: software development, artificial intelligence
  Focus area: multimodal AI, neural search

The solution

Availability as a Lifeline

The availability of NVIDIA GPUs was crucial for Jina AI, presenting a significant challenge. As an NVIDIA Elite Partner, sysGen facilitated access to the NVIDIA H100 GPUs. With efficient warehousing, sysGen was able to deliver the systems promptly, which was essential for the project‘s success.

Although Jina AI initially had no experience with hosting and hardware configuration, this proved to pose no obstacle. sysGen provided comprehensive support with the pre-installation of the software. Since Jina AI had no experience with its own bare-metal systems, sysGen provided extensive support for the pre-installation of the software applications.

Jina AI first tested the systems at sysGen, which gave them the necessary assurance that the servers would run flawlessly in the target data center. The remote configuration greatly simplified commissioning.

> ***“sysGen has always provided professional and quick support. We made it clear from the start that we had no experience with servers and received excellent advice accordingly.”***
>
> Maximilian Werk | Jina AI

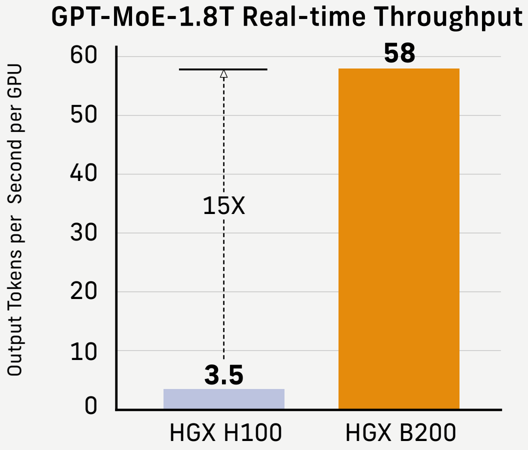
Outcome

Purpose-Built for AI and High-Performance Computing

After professional consultation, Jina AI decided to purchase its own GPU Compute power. These systems are each equipped with eight NVIDIA® HGX™ H100 GPUs, which enable high GPU computation through NVIDIA® NVLINK™ and NVIDIA® NVSwitch™ and are perfect for high-performance computing (HPC), deep learning training, industrial automation, retail and climate and weather modeling.

AI, complex simulations, and massive datasets require multiple GPUs with extremely fast interconnections and a fully accelerated software stack. The NVIDIA HGX™ AI supercomputing platform brings together the full power of NVIDIA GPUs, NVIDIA NVLink™, NVIDIA networking, and fully optimized AI and high-performance computing (HPC) software stacks to provide the highest application performance and drive the fastest time to insights.

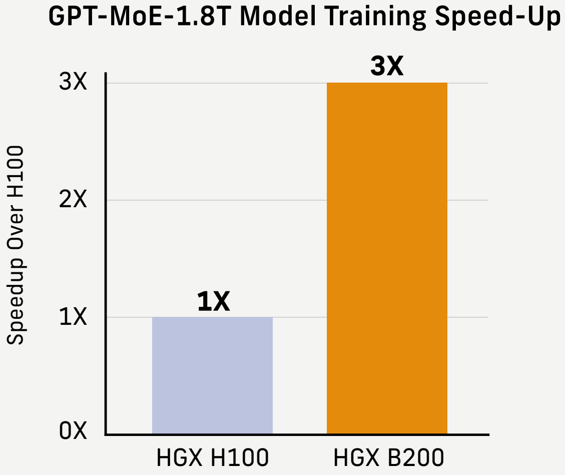
Deep Learning Inference: Performance and Versatility



Real-Time Inference for the Next Generation of Large Language Models

HGX B200 achieves up to 15X higher inference performance over the previous NVIDIA Hopper™ generation for massive models such as GPT-MoE-1.8T. The second-generation Transformer Engine uses custom Blackwell Tensor Core technology combined with TensorRT™-LLM and Nemo™ Framework innovations to accelerate inference for large language models (LLMs) and Mixture-of-Experts (MoE) models.

Deep Learning Training: Performance and Scalability



Next-Level Training Performance

The second-generation Transformer Engine, featuring 8-bit floating point (FP8) and new precisions, enables a remarkable 3X faster training for large language models like GPT-MoE-1.8T. This breakthrough is complemented by fifth-generation NVLink with 1.8TB/s of GPU-to-GPU interconnect, InfiniBand networking, and NVIDIA Magnum IO™ software. Together, these ensure efficient scalability for enterprises and extensive GPU computing clusters.

Source: https://www.nvidia.com/en-us/high-performance-computing/

Use Case

The Centerpiece in Action

By leveraging the NVIDIA H100 GPUs provided by sysGen, cutting-edge models were trained that significantly enhanced the relevance of search results for a client—a young startup. This startup assists sales representatives in generating precise response scripts during phone calls with potential customers. The powerful GPUs drastically reduced training times and significantly boosted the accuracy of these models. As a result, the startup was able to provide quicker and more accurate responses during customer interactions, leading to improved customer satisfaction and ultimately driving business success.

> ***“We teamed up with sysGen to train embedding and reranking models for our customers’ RAG systems,and the results were excellent.”***
>
> Maximilian Werk | Jina AI

RAG use case

Industry Analysis

Preparation of market reports with RAG on the basis of industry data.

Customer Service

Development of chatbots for reliable assistance, such as a retailer’s bot for delivery and return policies.

Content Creation

Use of RAG for customized content such as articles and newsletters.

Assistants for Document Research

Creation of chatbots for HR, compliance and security queries from company documents.

Health Advice

Provision of medical information and support through RAG-controlled chatbots for 24/7 patient care.

Track record

Jina AI and sysGen: A success story

Thanks to the NVIDIA GPU systems provided by sysGen, Jina AI was able to significantly increase the efficiency and accuracy of its neural search systems. Purchasing its own hardware enabled a better cost-benefit calculation, as the GPU power no longer needs to be rented. Owning this hardware enables both costs and work processes to be optimized. In addition, the training data can be permanently stored locally, which means less organizational work and less time wasted on data transfer. Updates, maintenance and error analyses are carried out remotely by sysGen, which further reduces the organizational effort.

The acquisition of bare-metal systems proved to be less complicated than expected, not least thanks to the personal advice provided by sysGen in advance.

> ***"For model training on single-node GPU instances, the solution is perfect for us. I consider to purchase more GPU machines in the second half of 2024."***
>
> Maximilian Werk | Jina AI

Would you also like to optimize your IT infrastructure and benefit from tailor-made solutions? Get in touch with your contact at sysGen and find out more about the individual consulting and service offers from sysGen.

Your contacts:

Gabriele Nikisch

CEO / Sales management

☎ Phone: +49 421 409 66 21

✉ gnikisch@sysgen.de

Sergius Siczek

CTO / Technical management

☎ Phone: +49 421 409 66 32

✉ ssiczek@sysgen.de

📍 sysGen GmbH, Am Hallacker 48, 28327 Bremen

Visit our Website:

COMING SOON!

New

NVIDIA DGX B200

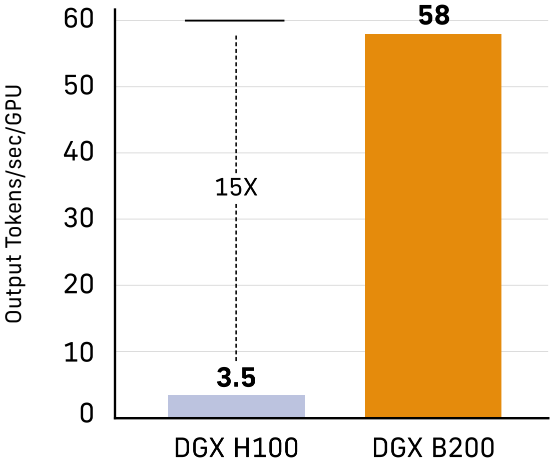
The foundation for your AI center of excellence.

NVIDIA DGX™ B200 is equipped with eight NVIDIA Blackwell GPUs interconnected with fifth-generation NVIDIA® NVLink®, DGX B200 delivers leading-edge performance, offering 3X the training performance and 15X the inference performance of previous generations. Leveraging the NVIDIA Blackwell GPU architecture, DGX B200 can handle diverse workloads—including large language models, recommender systems, and chatbots—making it ideal for businesses looking to accelerate their AI transformation.

NVIDIA DGX™ B200 is an unified AI platform for develop-to-deploy pipelines for businesses.

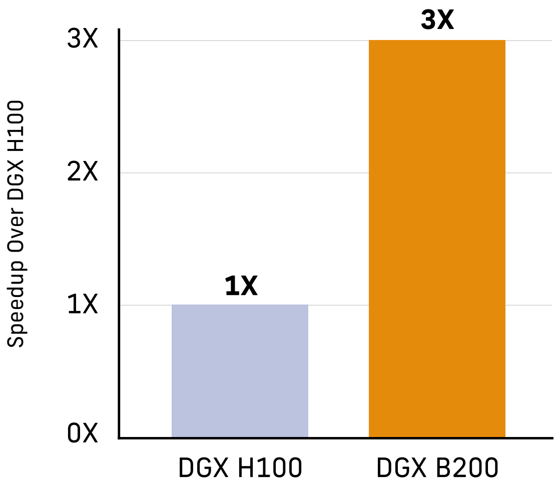
NVIDIA DGX systems set the standard for AI performance and efficiency. Click here

Real-time inference for large language models



GPT-MoE-1.8T Real-time Throughput

New standards in AI training performance



GPT-MoE-1.8T Model Training Speed-Up

Source: https://www.nvidia.com/en-us/data-center/dgx-b200/#referrer=vanity