Interacting with Large Language model Gemini and contrasting with Open Ai's Chat GPT

By: Parker Christenson

When thinking about how advanced the large language models, various companies and many open-source models have been developed within the past two years. Ever since the success of Chat GPT 3.0, we have seen that models are able to help with various tasks that we must do throughout the day. Chat GPT has brought machine learning to the forefront of technology as it has become very advanced in architecture, and the detail of the responses that they are able to give. With, Google has also been making various pushes for the large language models as well. Google came out with Gemini. Gemini is also a very complex large language model that does not have any sourcing on the weights and biases much like Chat GPT. Considering that Chat GPT has a free version, and a paid for version, it makes sense not to release the fine-tuning mechanisms that help achieve the tension between the transformer architecture.

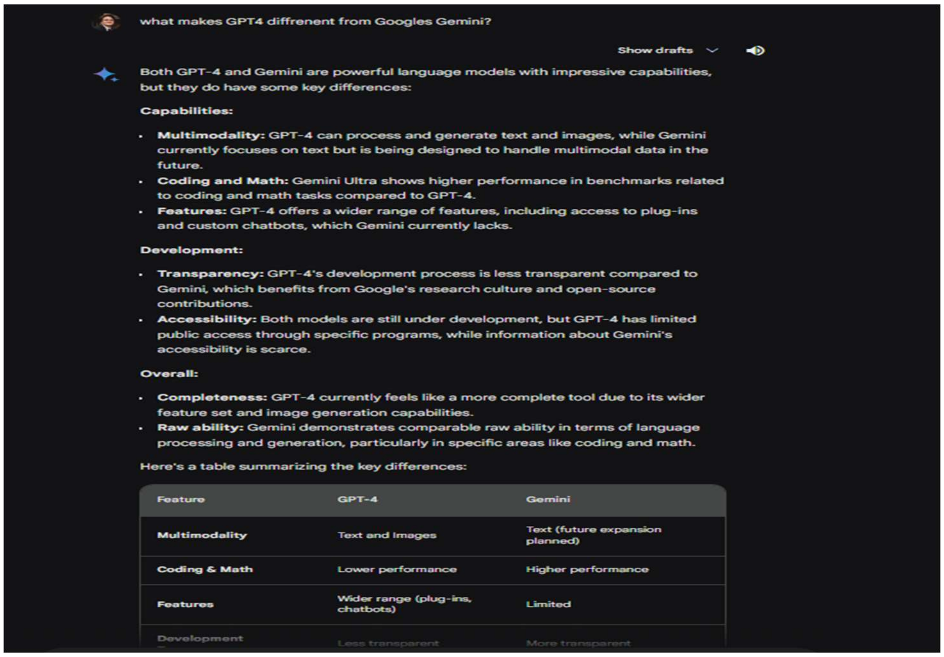
When using chat GPT you are given a very standard experience out of the box when creating a free account. The basic functions of the large language model are stored in an API, and the API is what is storing the model and giving the response back to the end user. This is the most common practice for LLM's to be deployed. The response is fast, as it is just giving text back out, which is good because it is just being asked simple questions. I have noticed that there have been lots of other applications in which chat GPT is being deployed. For instance, on that I use almost every day when I am working is Co-pilot by GitHub. Which was purchased by Microsoft, and Open Ai allowed Microsoft to use Chat GPT for investment. When using co-pilot, the AI interface is very simple, as it is a extremely good code completion tool. I use co-pilot to help me write some of the more boiler plate code, and not so much the complex code. For example, making a mat plot chart is extremely easy, as it will help you with deciding your Y axis and your X-axis.

When using the Chat GPT interface though, there are lots of limitations on what the model can say and do. But when I was using Gemini, I found that the responses are a lot more inclusive of all possibilities. I am currently trying to build a computer that is specifically for making neural networks, and the limitations of Chat GPT are very noticeable. When I would ask questions about a specific SXM4 board connector which is almost a trade secret, Chat GPT would not answer it at all. But when using Gemini, the LLM was able to direct me to the part needed, along with modules I would install when downloading Ubuntu.

What I have been able to find about the models and the architecture has been somewhat hidden as they are in a very highly competitive space. Open AI has been somewhat public on they have been censoring or restricting prompts while google has too. Chat GPT's architecture is based off of a large language model transformer, which relies very heavily on self-attention. Self-attention is a very complex way of having the model look back at what has been said and is able to then do some more inferencing when giving a response. GPT's model design has been uploaded to YouTube by one of the leading engineers, but did not give key specifications to exactly replicate GPT4.

Here are some of the questions I asked the two models, to show the comparison.

Gemini



what makes GPT4 diffrenent from Googles Gemini?

Both GPT-4 and Gemini are powerful language models with impressive capabilities, but they do have some key differences:

Capabilities:

- **Multimodality:** GPT-4 can process and generate text and images, while Gemini currently focuses on text but is being designed to handle multimodal data in the future.
- **Coding and Math:** Gemini Ultra shows higher performance in benchmarks related to coding and math tasks compared to GPT-4.
- **Features:** GPT-4 offers a wider range of features, including access to plug-ins and custom chatbots, which Gemini currently lacks.

Development:

- **Transparency:** GPT-4's development process is less transparent compared to Gemini, which benefits from Google's research culture and open-source contributions.
- **Accessibility:** Both models are still under development, but GPT-4 has limited public access through specific programs, while information about Gemini's accessibility is scarce.

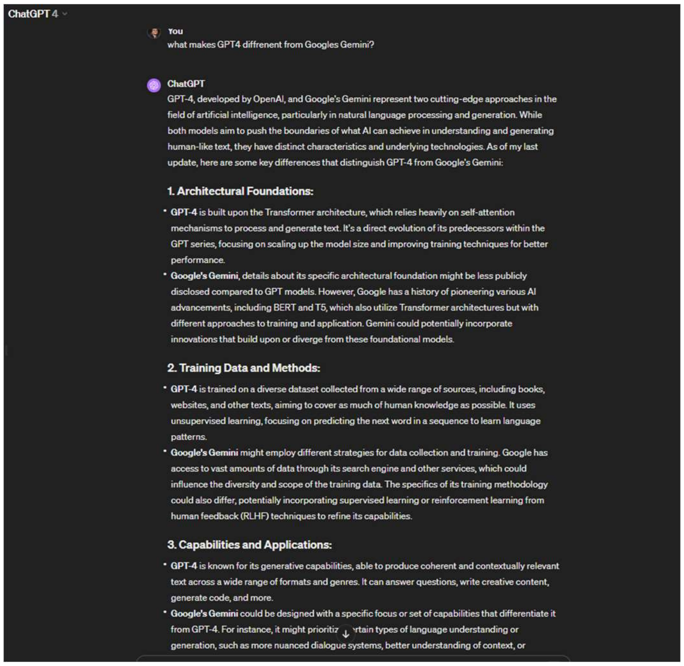
Overall:

- **Completeness:** GPT-4 currently feels like a more complete tool due to its wider feature set and image generation capabilities.
- **Raw ability:** Gemini demonstrates comparable raw ability in terms of language processing and generation, particularly in specific areas like coding and math.

Here's a table summarizing the key differences:

Feature	GPT-4	Gemini
Multimodality	Text and Images	Text (future expansion planned)
Coding & Math	Lower performance	Higher performance
Features	Wider range (plug-ins, chatbots)	Limited
Development	Less transparent	More transparent

GPT-4



You

what makes GPT4 diffrenent from Googles Gemini?

ChatGPT

GPT-4, developed by OpenAI, and Google's Gemini represent two cutting-edge approaches in the field of artificial intelligence, particularly in natural language processing and generation. While both models aim to push the boundaries of what AI can achieve in understanding and generating human-like text, they have distinct characteristics and underlying technologies. As of my last update, here are some key differences that distinguish GPT-4 from Google's Gemini:

1. Architectural Foundations:

- **GPT-4** is built upon the Transformer architecture, which relies heavily on self-attention mechanisms to process and generate text. It's a direct evolution of its predecessors within the GPT series, focusing on scaling up the model size and improving training techniques for better performance.
- **Google's Gemini**, details about its specific architectural foundation might be less publicly disclosed compared to GPT models. However, Google has a history of pioneering various AI advancements, including BERT and T5, which also utilize Transformer architectures but with different approaches to training and application. Gemini could potentially incorporate innovations that build upon or diverge from these foundational models.

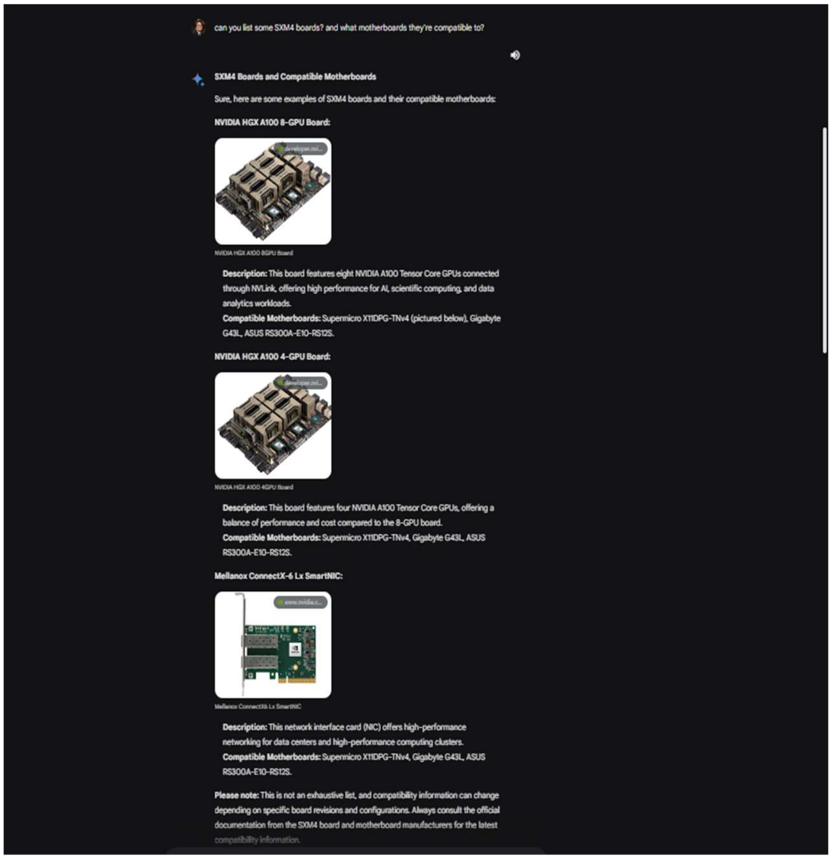
2. Training Data and Methods:

- **GPT-4** is trained on a diverse dataset collected from a wide range of sources, including books, websites, and other texts, aiming to cover as much of human knowledge as possible. It uses unsupervised learning, focusing on predicting the next word in a sequence to learn language patterns.
- **Google's Gemini** might employ different strategies for data collection and training. Google has access to vast amounts of data through its search engine and other services, which could influence the diversity and scope of the training data. The specifics of its training methodology could also differ, potentially incorporating supervised learning or reinforcement learning from human feedback (RLHF) techniques to refine its capabilities.

3. Capabilities and Applications:

- **GPT-4** is known for its generative capabilities, able to produce coherent and contextually relevant text across a wide range of formats and genres. It can answer questions, write creative content, generate code, and more.
- **Google's Gemini** could be designed with a specific focus or set of capabilities that differentiate it from GPT-4. For instance, it might prioritiz rtain types of language understanding or generation, such as more nuanced dialogue systems, better understanding of context, or

Gemini



can you list some SXM4 boards? and what motherboards they're compatible to?

SXM4 Boards and Compatible Motherboards

Sure, here are some examples of SXM4 boards and their compatible motherboards:

NVIDIA HGX A100 8-GPU Board:

NVIDIA HGX A100 8GPU Board

Description: This board features eight NVIDIA A100 Tensor Core GPUs connected through NVLink, offering high performance for AI, scientific computing, and data analytics workloads.
Compatible Motherboards: Supermicro X11DPG-TNv4 (pictured below), Gigabyte G43L, ASUS RS300A-E10-RS12S.

NVIDIA HGX A100 4-GPU Board:

NVIDIA HGX A100 4GPU Board

Description: This board features four NVIDIA A100 Tensor Core GPUs, offering a balance of performance and cost compared to the 8-GPU board.
Compatible Motherboards: Supermicro X11DPG-TNv4, Gigabyte G43L, ASUS RS300A-E10-RS12S.

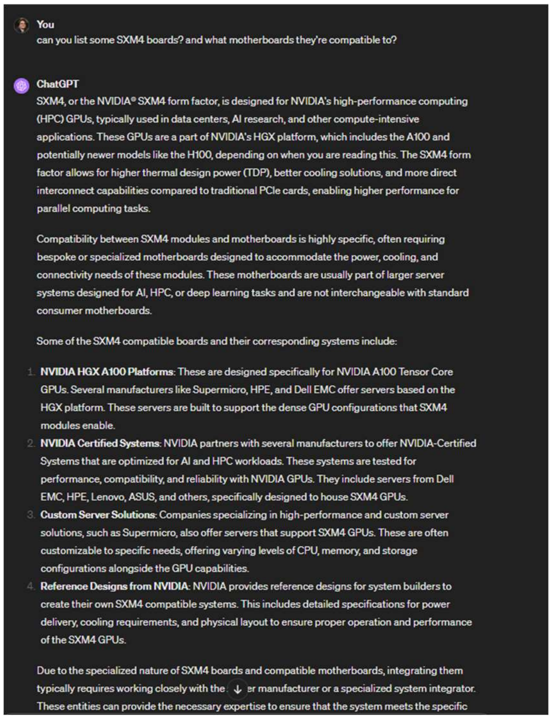
Mellanox ConnectX-6 Lx SmartNIC:

Mellanox ConnectX6 Lx SmartNIC

Description: This network interface card (NIC) offers high-performance networking for data centers and high-performance computing clusters.
Compatible Motherboards: Supermicro X11DPG-TNv4, Gigabyte G43L, ASUS RS300A-E10-RS12S.

Please note: This is not an exhaustive list, and compatibility information can change depending on specific board revisions and configurations. Always consult the official documentation from the SXM4 board and motherboard manufacturers for the latest compatibility information.

GPT-4



You
can you list some SXM4 boards? and what motherboards they're compatible to?

ChatGPT
SXM4, or the NVIDIA® SXM4 form factor, is designed for NVIDIA's high-performance computing (HPC) GPUs, typically used in data centers, AI research, and other compute-intensive applications. These GPUs are a part of NVIDIA's HGX platform, which includes the A100 and potentially newer models like the H100, depending on when you are reading this. The SXM4 form factor allows for higher thermal design power (TDP), better cooling solutions, and more direct interconnect capabilities compared to traditional PCIe cards, enabling higher performance for parallel computing tasks.

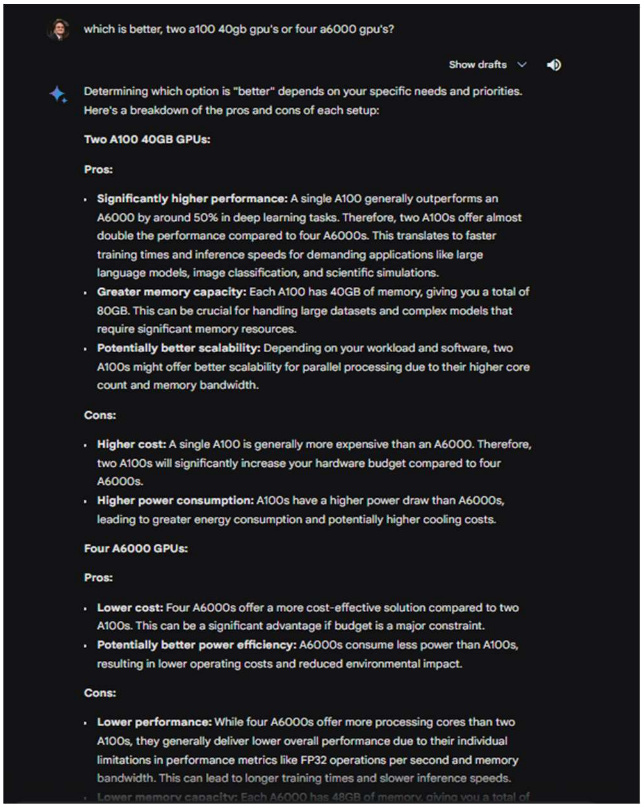
Compatibility between SXM4 modules and motherboards is highly specific, often requiring bespoke or specialized motherboards designed to accommodate the power, cooling, and connectivity needs of these modules. These motherboards are usually part of larger server systems designed for AI, HPC, or deep learning tasks and are not interchangeable with standard consumer motherboards.

Some of the SXM4 compatible boards and their corresponding systems include:

1. **NVIDIA HGX A100 Platforms**: These are designed specifically for NVIDIA A100 Tensor Core GPUs. Several manufacturers like Supermicro, HPE, and Dell EMC offer servers based on the HGX platform. These servers are built to support the dense GPU configurations that SXM4 modules enable.
2. **NVIDIA Certified Systems**: NVIDIA partners with several manufacturers to offer NVIDIA-Certified Systems that are optimized for AI and HPC workloads. These systems are tested for performance, compatibility, and reliability with NVIDIA GPUs. They include servers from Dell EMC, HPE, Lenovo, ASUS, and others, specifically designed to house SXM4 GPUs.
3. **Custom Server Solutions**: Companies specializing in high-performance and custom server solutions, such as Supermicro, also offer servers that support SXM4 GPUs. These are often customizable to specific needs, offering varying levels of CPU, memory, and storage configurations alongside the GPU capabilities.
4. **Reference Designs from NVIDIA**: NVIDIA provides reference designs for system builders to create their own SXM4 compatible systems. This includes detailed specifications for power delivery, cooling requirements, and physical layout to ensure proper operation and performance of the SXM4 GPUs.

Due to the specialized nature of SXM4 boards and compatible motherboards, integrating them typically requires working closely with the ↓ er manufacturer or a specialized system integrator. These entities can provide the necessary expertise to ensure that the system meets the specific

Gemini



which is better, two a100 40gb gpu's or four a6000 gpu's?

Determining which option is "better" depends on your specific needs and priorities. Here's a breakdown of the pros and cons of each setup:

Two A100 40GB GPUs:

Pros:

- **Significantly higher performance:** A single A100 generally outperforms an A6000 by around 50% in deep learning tasks. Therefore, two A100s offer almost double the performance compared to four A6000s. This translates to faster training times and inference speeds for demanding applications like large language models, image classification, and scientific simulations.
- **Greater memory capacity:** Each A100 has 40GB of memory, giving you a total of 80GB. This can be crucial for handling large datasets and complex models that require significant memory resources.
- **Potentially better scalability:** Depending on your workload and software, two A100s might offer better scalability for parallel processing due to their higher core count and memory bandwidth.

Cons:

- **Higher cost:** A single A100 is generally more expensive than an A6000. Therefore, two A100s will significantly increase your hardware budget compared to four A6000s.
- **Higher power consumption:** A100s have a higher power draw than A6000s, leading to greater energy consumption and potentially higher cooling costs.

Four A6000 GPUs:

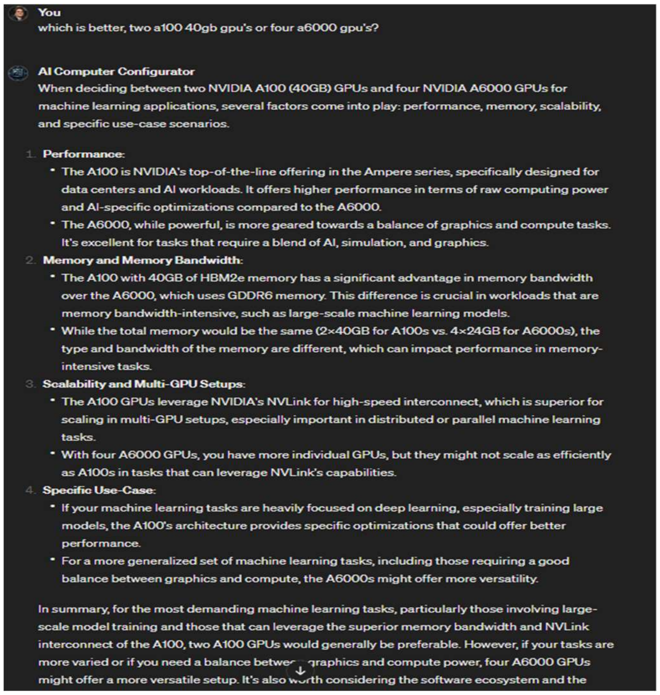
Pros:

- **Lower cost:** Four A6000s offer a more cost-effective solution compared to two A100s. This can be a significant advantage if budget is a major constraint.
- **Potentially better power efficiency:** A6000s consume less power than A100s, resulting in lower operating costs and reduced environmental impact.

Cons:

- **Lower performance:** While four A6000s offer more processing cores than two A100s, they generally deliver lower overall performance due to their individual limitations in performance metrics like FP32 operations per second and memory bandwidth. This can lead to longer training times and slower inference speeds.
- **Lower memory capacity:** Each A6000 has 48GB of memory, giving you a total of

GPT-4



You

which is better, two a100 40gb gpu's or four a6000 gpu's?

AI Computer Configurator

When deciding between two NVIDIA A100 (40GB) GPUs and four NVIDIA A6000 GPUs for machine learning applications, several factors come into play: performance, memory, scalability, and specific use-case scenarios.

1. **Performance:**
   - The A100 is NVIDIA's top-of-the-line offering in the Ampere series, specifically designed for data centers and AI workloads. It offers higher performance in terms of raw computing power and AI-specific optimizations compared to the A6000.
   - The A6000, while powerful, is more geared towards a balance of graphics and compute tasks. It's excellent for tasks that require a blend of AI, simulation, and graphics.
2. **Memory and Memory Bandwidth:**
   - The A100 with 40GB of HBM2e memory has a significant advantage in memory bandwidth over the A6000, which uses GDDR6 memory. This difference is crucial in workloads that are memory bandwidth-intensive, such as large-scale machine learning models.
   - While the total memory would be the same (2×40GB for A100s vs. 4×24GB for A6000s), the type and bandwidth of the memory are different, which can impact performance in memory-intensive tasks.
3. **Scalability and Multi-GPU Setups:**
   - The A100 GPUs leverage NVIDIA's NVLink for high-speed interconnect, which is superior for scaling in multi-GPU setups, especially important in distributed or parallel machine learning tasks.
   - With four A6000 GPUs, you have more individual GPUs, but they might not scale as efficiently as A100s in tasks that can leverage NVLink's capabilities.
4. **Specific Use-Case:**
   - If your machine learning tasks are heavily focused on deep learning, especially training large models, the A100's architecture provides specific optimizations that could offer better performance.
   - For a more generalized set of machine learning tasks, including those requiring a good balance between graphics and compute, the A6000s might offer more versatility.

In summary, for the most demanding machine learning tasks, particularly those involving large-scale model training and those that can leverage the superior memory bandwidth and NVLink interconnect of the A100, two A100 GPUs would generally be preferable. However, if your tasks are more varied or if you need a balance betwe graphics and compute power, four A6000 GPUs might offer a more versatile setup. It's also worth considering the software ecosystem and the