2025 DeepSeek Test Run

AI and Machine Learning

Tony Shen
Data Communications Labs
2/2/2025

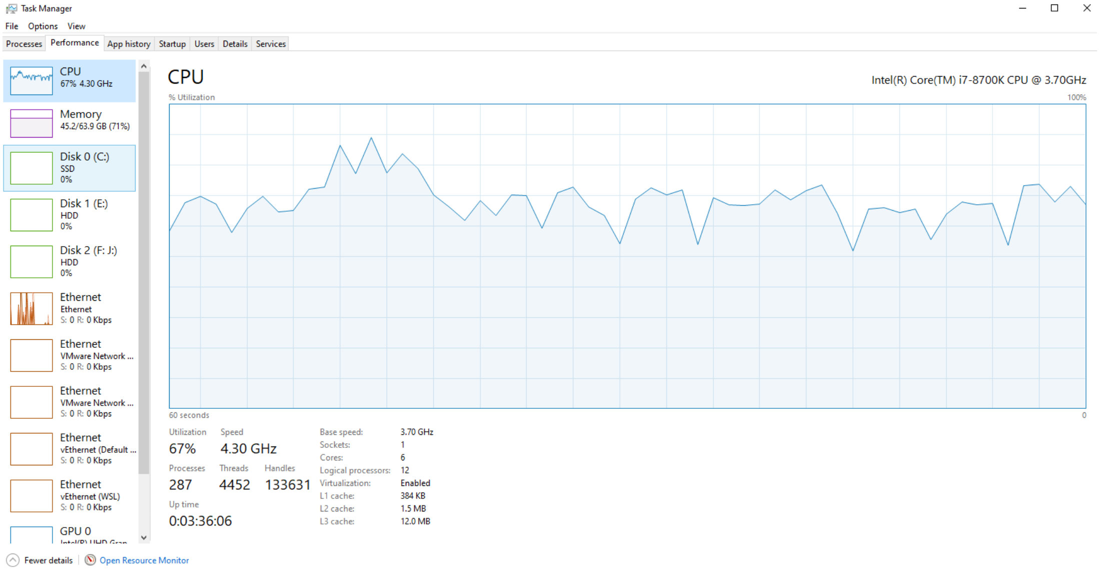
Contents

Test Environment 2
Selected Models 4
Tested Models 6
Test Method 6
Test Outcome 7
Observations 17
Model Behaviors 18
Summary 18
References 18

Test Environment

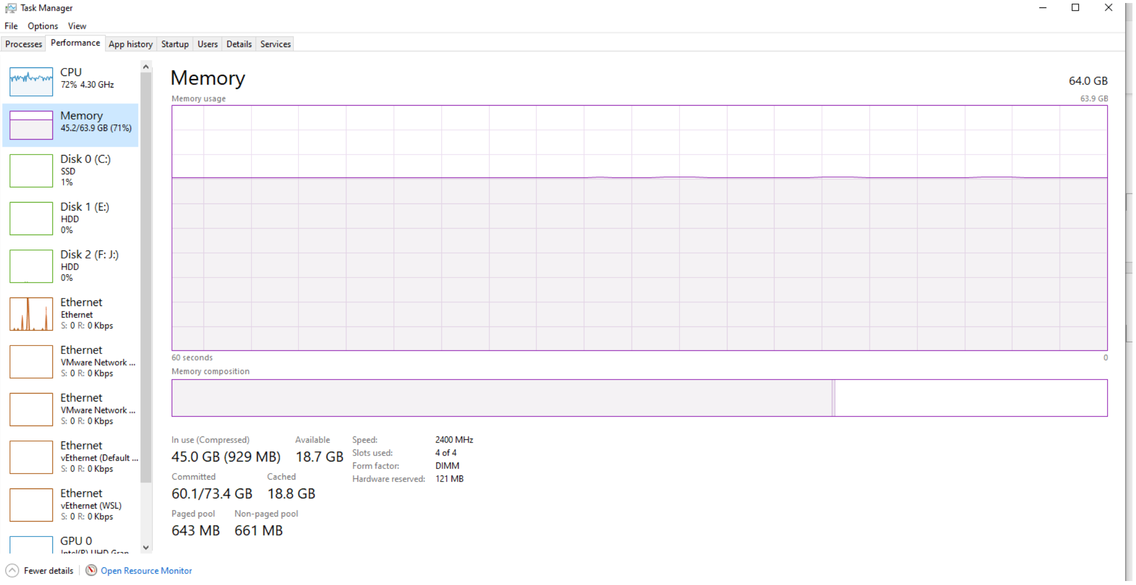
This setup involves a local test run where the four models are downloaded and installed on a standard PC. The entire test is conducted locally. The PC is equipped with 64GB of memory, an RTX-4070 (a lower-end Nvidia GPU card), and ample disk space to accommodate large models. The models are accessed via Ollama, and the user interface is provided by Open Web UI.

CPU – Intel Core i7-8700K CPU @ 3.70Ghz



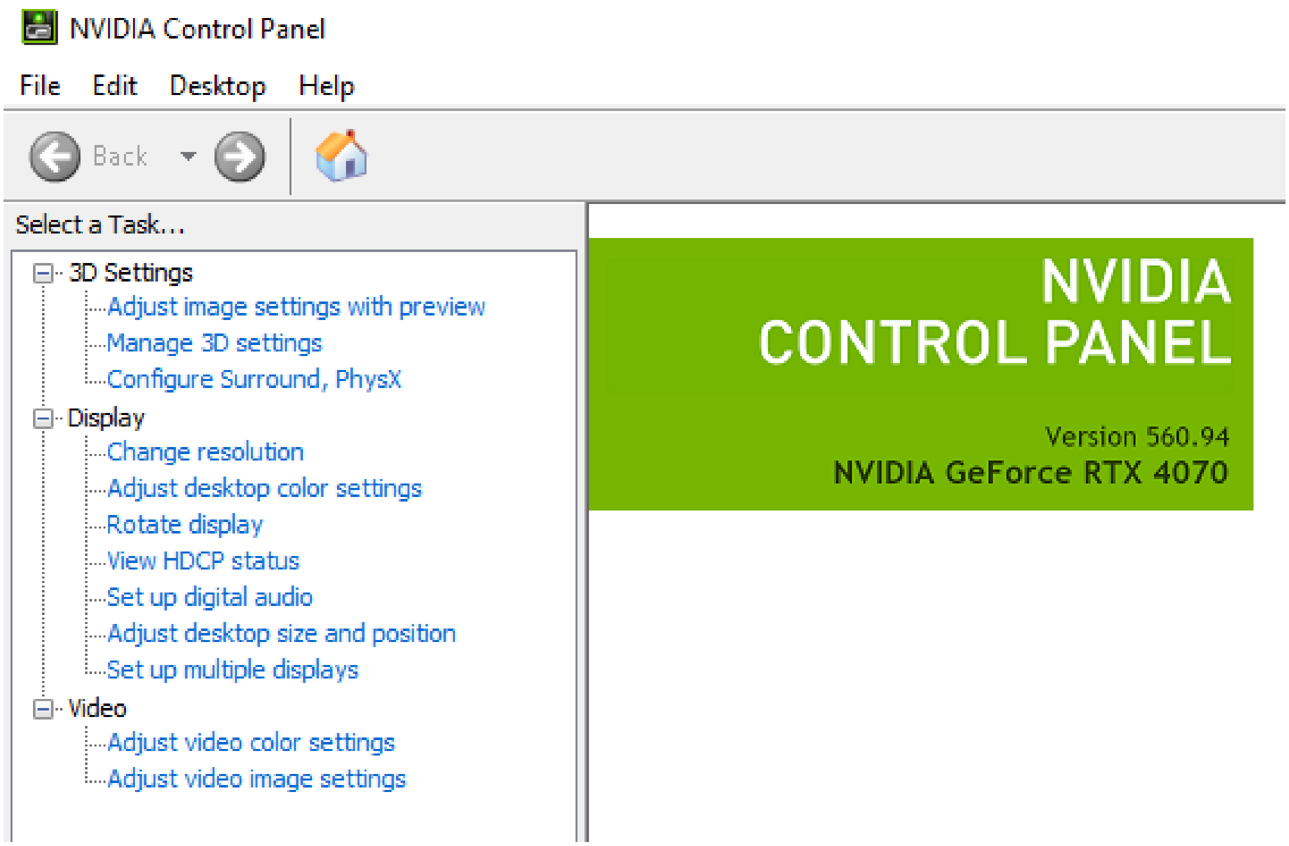
Utilization 67%
Speed 4.30 GHz
Processes 287
Threads 4452
Handles 133631
Up time 0:03:36:06
Base speed: 3.70 GHz
Sockets: 1
Cores: 6
Logical processors: 12
Virtualization: Enabled
L1 cache: 384 KB
L2 cache: 1.5 MB
L3 cache: 12.0 MB

Memory – 64GB



In use (Compressed) 45.0 GB (929 MB)
Available 18.7 GB
Committed 60.1/73.4 GB
Cached 18.8 GB
Paged pool 643 MB
Non-paged pool 661 MB
Speed: 2400 MHz
Slots used: 4 of 4
Form factor: DIMM
Hardware reserved: 121 MB

GPU – Nvidia GeForce RTX 4070

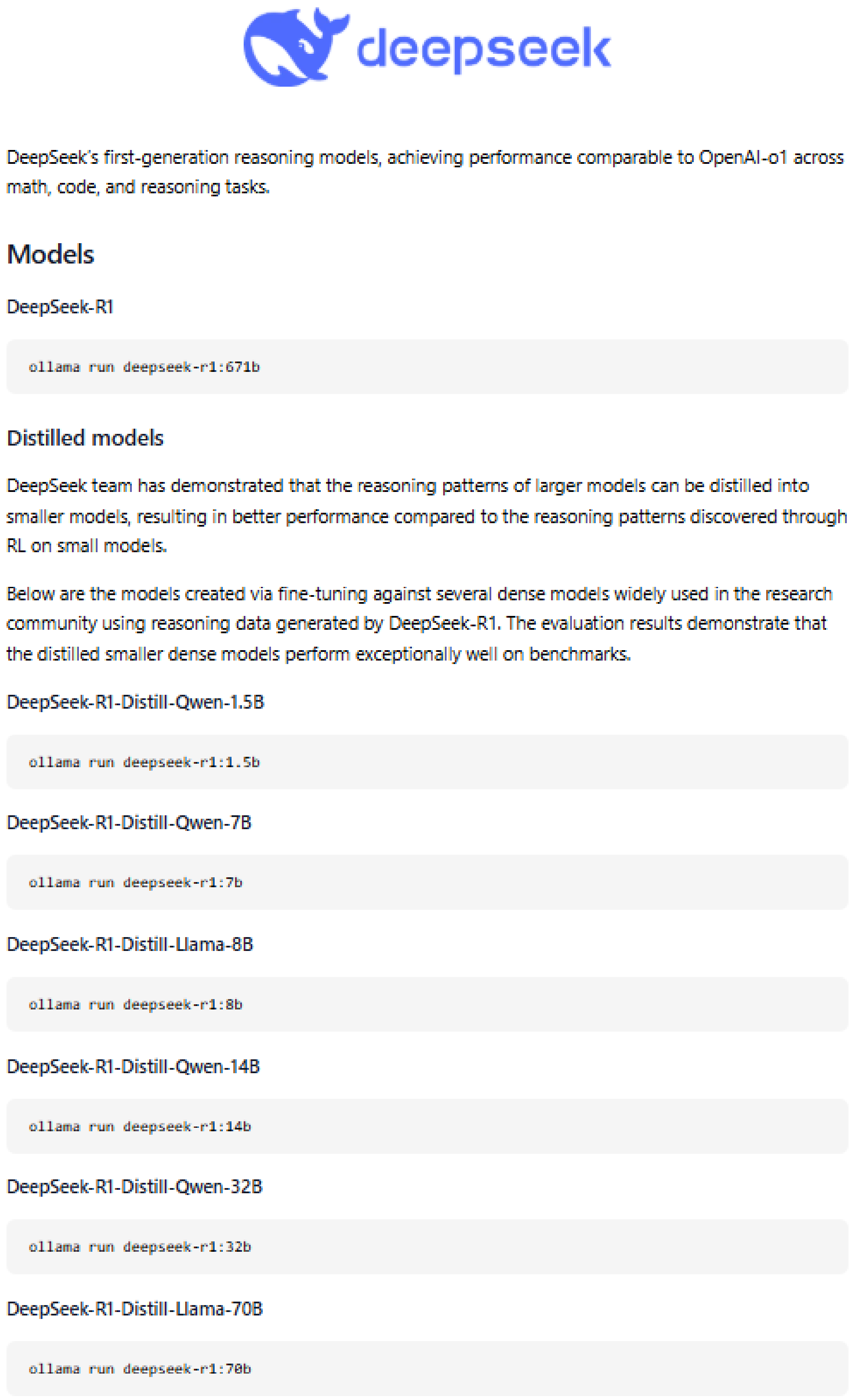


NVIDIA CONTROL PANEL

Version 560.94

NVIDIA GeForce RTX 4070

Selected Models



DeepSeek's first-generation reasoning models, achieving performance comparable to OpenAI-o1 across math, code, and reasoning tasks.

Models

DeepSeek-R1

```
ollama run deepseek-r1:671b
```

Distilled models

DeepSeek team has demonstrated that the reasoning patterns of larger models can be distilled into smaller models, resulting in better performance compared to the reasoning patterns discovered through RL on small models.

Below are the models created via fine-tuning against several dense models widely used in the research community using reasoning data generated by DeepSeek-R1. The evaluation results demonstrate that the distilled smaller dense models perform exceptionally well on benchmarks.

DeepSeek-R1-Distill-Qwen-1.5B

```
ollama run deepseek-r1:1.5b
```

DeepSeek-R1-Distill-Qwen-7B

```
ollama run deepseek-r1:7b
```

DeepSeek-R1-Distill-Llama-8B

```
ollama run deepseek-r1:8b
```

DeepSeek-R1-Distill-Qwen-14B

```
ollama run deepseek-r1:14b
```

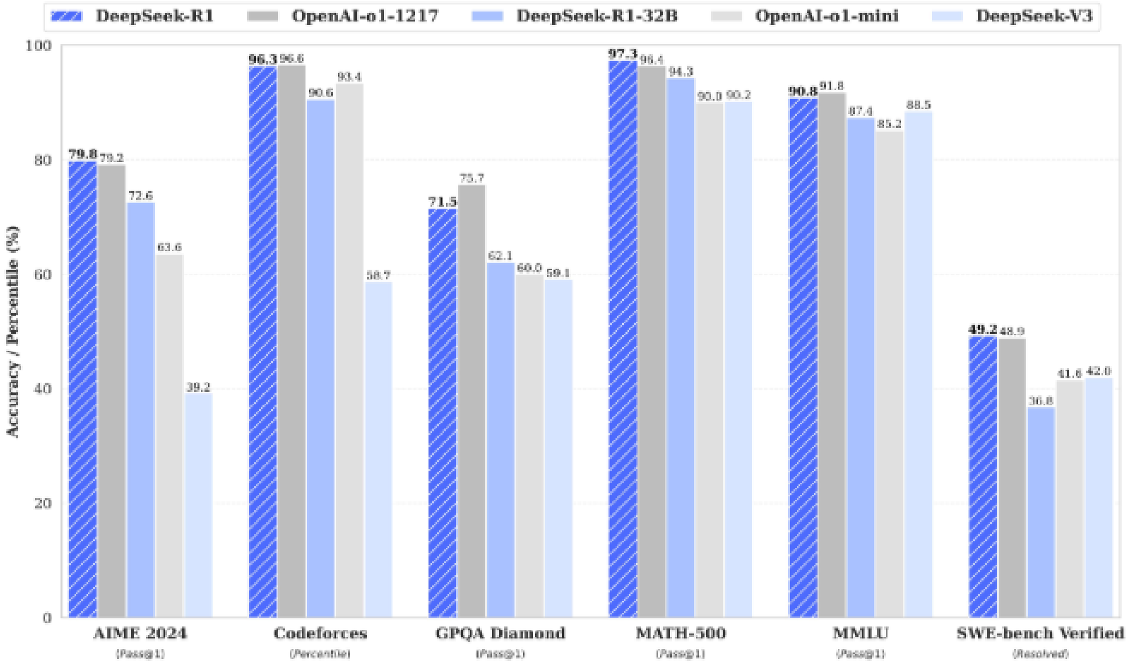
DeepSeek-R1-Distill-Qwen-32B

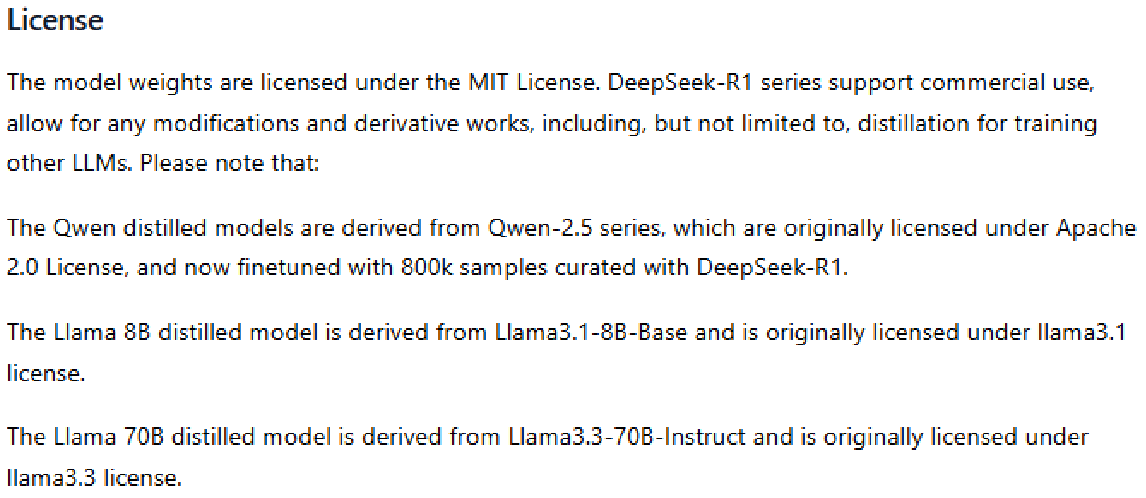
```
ollama run deepseek-r1:32b
```

DeepSeek-R1-Distill-Llama-70B

```
ollama run deepseek-r1:70b
```

DeepSeek distilled models were selected





License

The model weights are licensed under the MIT License. DeepSeek-R1 series support commercial use, allow for any modifications and derivative works, including, but not limited to, distillation for training other LLMs. Please note that:

The Qwen distilled models are derived from Qwen-2.5 series, which are originally licensed under Apache 2.0 License, and now finetuned with 800k samples curated with DeepSeek-R1.

The Llama 8B distilled model is derived from Llama3.1-8B-Base and is originally licensed under llama3.1 license.

The Llama 70B distilled model is derived from Llama3.3-70B-Instruct and is originally licensed under llama3.3 license.

Tested Models

```
llama3.3:latest a6eb4748fd29 42 GB 6 minutes ago
deepseek-r1:latest 0a8c26691023 4.7 GB 2 hours ago
deepseek-r1:32b 38056bbcbb2d 19 GB 8 hours ago
deepseek-r1:70b 0c1615a8ca32 42 GB 2 hours ago
```

Test Method

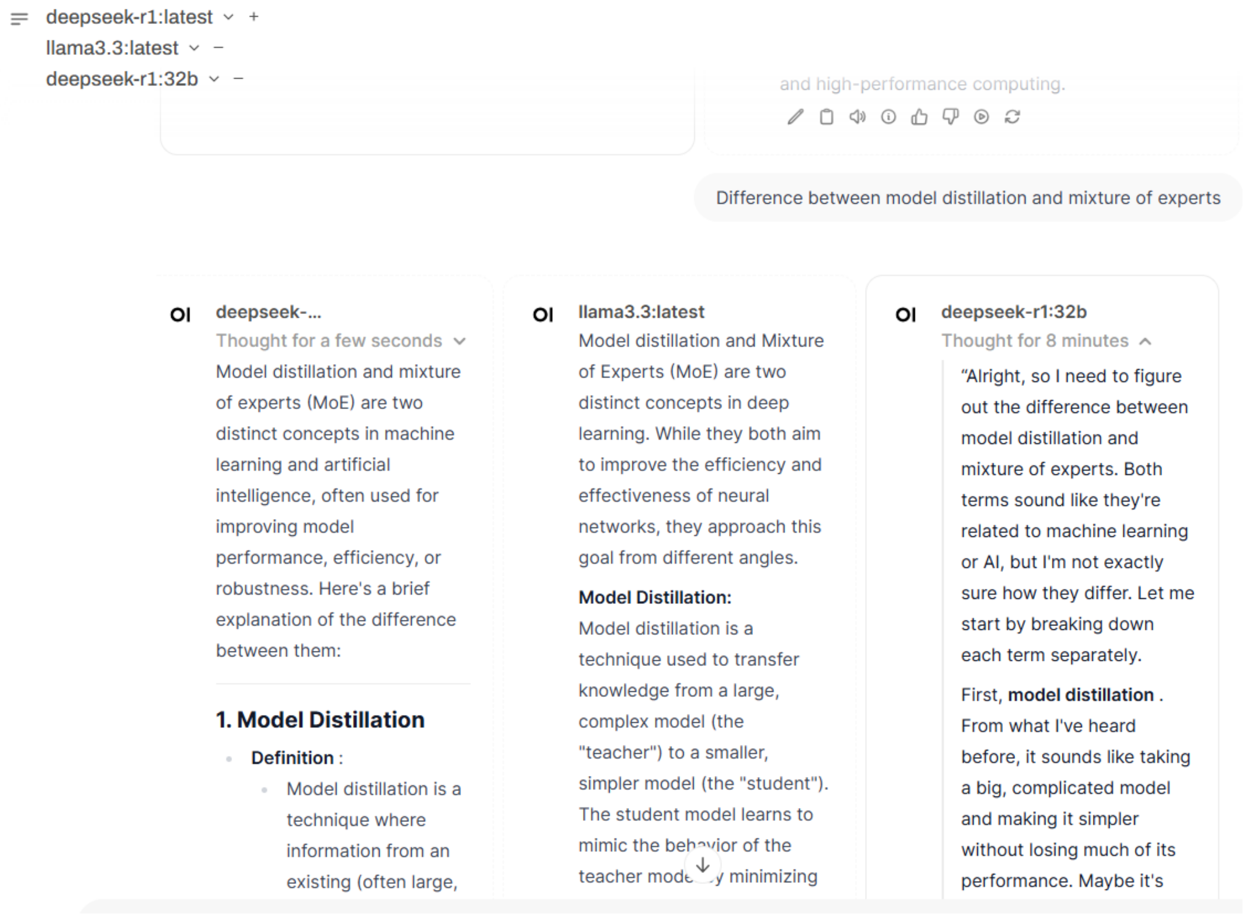
Two tests were conducted, where multiple models were run concurrently side by side to compare their real-time performance.

Test 1:
A technical question was submitted to three models: two DeepSeek R1 distilled models and one Meta LLM model. The responses from the models were observed and compared concurrently in real-time.

Test 2:
A non-technical question with two separate topics was submitted to two models, both of which were DeepSeek distilled models (one small and one medium-sized).

Test Outcome

Test 1 Output



deepseek-r1:latest

llama3.3:latest

deepseek-r1:32b

and high-performance computing.

Difference between model distillation and mixture of experts

deepseek-...

Thought for a few seconds

Model distillation and mixture of experts (MoE) are two distinct concepts in machine learning and artificial intelligence, often used for improving model performance, efficiency, or robustness. Here's a brief explanation of the difference between them:

---

1. Model Distillation

- **Definition** :
  - Model distillation is a technique where information from an existing (often large,

llama3.3:latest

Model distillation and Mixture of Experts (MoE) are two distinct concepts in deep learning. While they both aim to improve the efficiency and effectiveness of neural networks, they approach this goal from different angles.

Model Distillation:

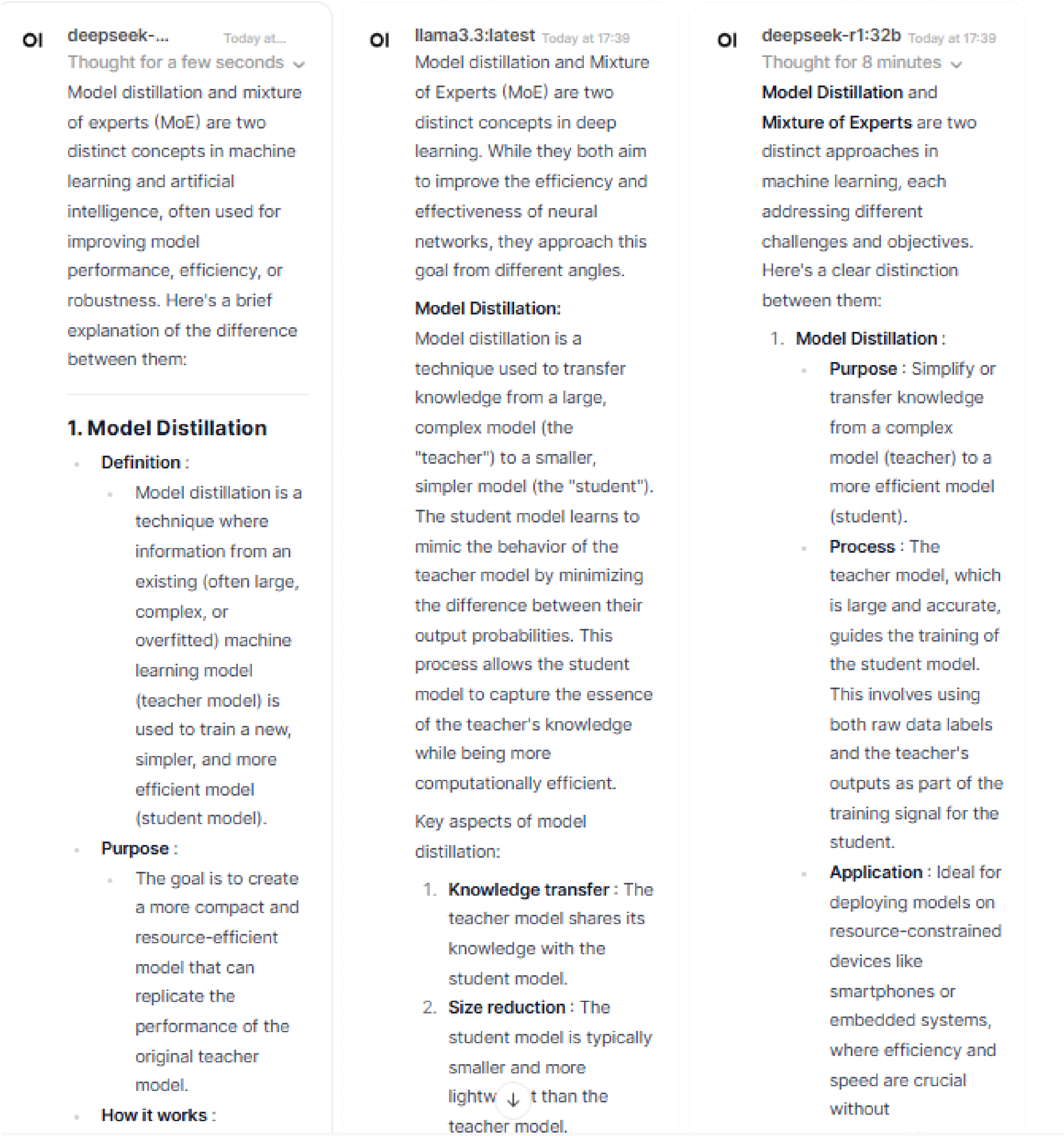
Model distillation is a technique used to transfer knowledge from a large, complex model (the "teacher") to a smaller, simpler model (the "student"). The student model learns to mimic the behavior of the teacher model by minimizing

deepseek-r1:32b

Thought for 8 minutes

“Alright, so I need to figure out the difference between model distillation and mixture of experts. Both terms sound like they're related to machine learning or AI, but I'm not exactly sure how they differ. Let me start by breaking down each term separately.

First, **model distillation** . From what I've heard before, it sounds like taking a big, complicated model and making it simpler without losing much of its performance. Maybe it's



deepseek-... Today at...

Thought for a few seconds ⌄

Model distillation and mixture of experts (MoE) are two distinct concepts in machine learning and artificial intelligence, often used for improving model performance, efficiency, or robustness. Here's a brief explanation of the difference between them:

---

1. Model Distillation

- **Definition** :
  - Model distillation is a technique where information from an existing (often large, complex, or overfitted) machine learning model (teacher model) is used to train a new, simpler, and more efficient model (student model).
- **Purpose** :
  - The goal is to create a more compact and resource-efficient model that can replicate the performance of the original teacher model.
- **How it works** :

llama3.3:latest Today at 17:39

Model distillation and Mixture of Experts (MoE) are two distinct concepts in deep learning. While they both aim to improve the efficiency and effectiveness of neural networks, they approach this goal from different angles.

Model Distillation:

Model distillation is a technique used to transfer knowledge from a large, complex model (the "teacher") to a smaller, simpler model (the "student"). The student model learns to mimic the behavior of the teacher model by minimizing the difference between their output probabilities. This process allows the student model to capture the essence of the teacher's knowledge while being more computationally efficient.

Key aspects of model distillation:

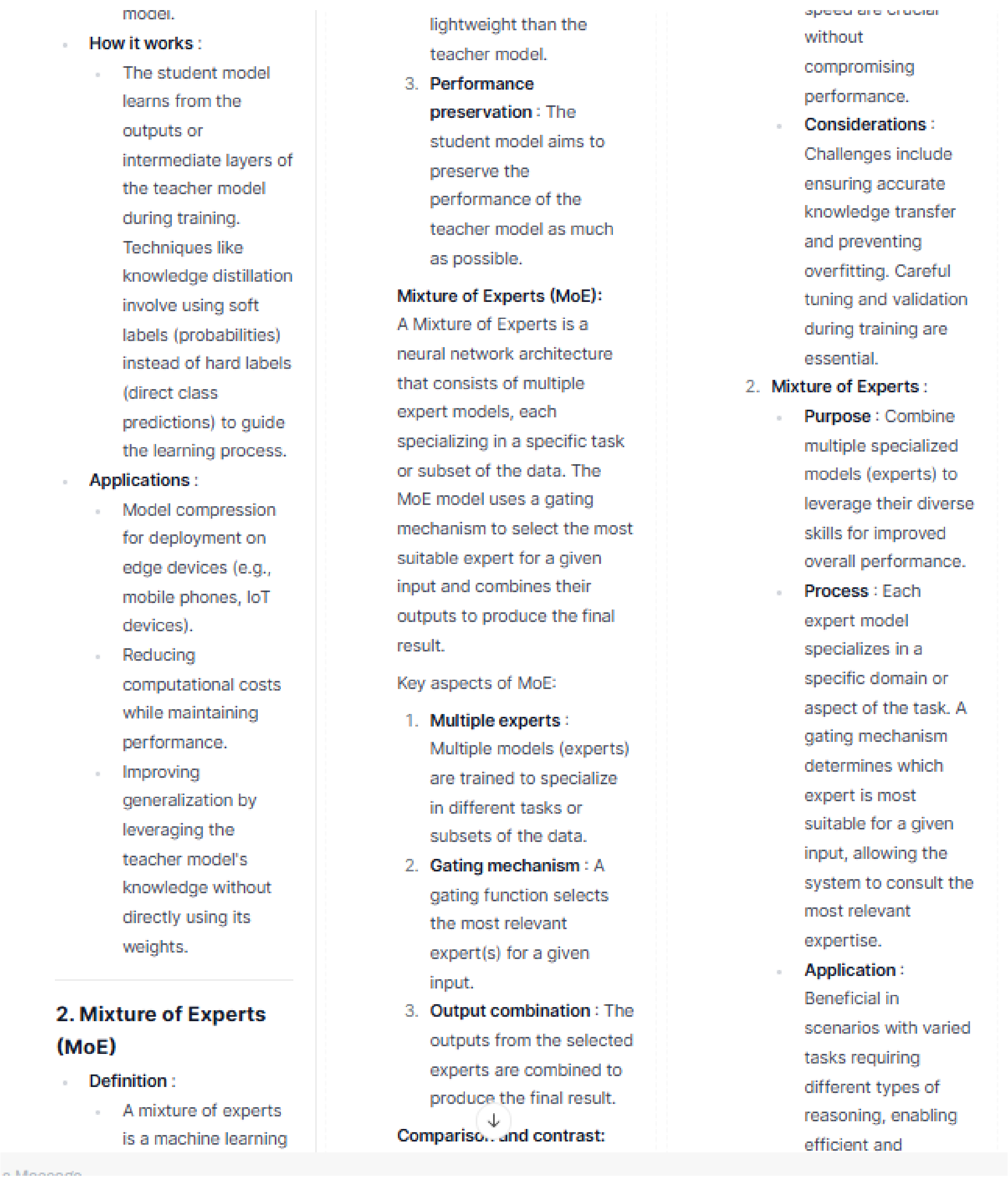
1. **Knowledge transfer** : The teacher model shares its knowledge with the student model.
2. **Size reduction** : The student model is typically smaller and more lightw ↓ t than the teacher model.

deepseek-r1:32b Today at 17:39

Thought for 8 minutes ⌄

Model Distillation and **Mixture of Experts** are two distinct approaches in machine learning, each addressing different challenges and objectives. Here's a clear distinction between them:

1. **Model Distillation** :
   - **Purpose** : Simplify or transfer knowledge from a complex model (teacher) to a more efficient model (student).
   - **Process** : The teacher model, which is large and accurate, guides the training of the student model. This involves using both raw data labels and the teacher's outputs as part of the training signal for the student.
   - **Application** : Ideal for deploying models on resource-constrained devices like smartphones or embedded systems, where efficiency and speed are crucial without



model.

- **How it works** :
    - The student model learns from the outputs or intermediate layers of the teacher model during training. Techniques like knowledge distillation involve using soft labels (probabilities) instead of hard labels (direct class predictions) to guide the learning process.
- **Applications** :
    - Model compression for deployment on edge devices (e.g., mobile phones, IoT devices).
    - Reducing computational costs while maintaining performance.
    - Improving generalization by leveraging the teacher model's knowledge without directly using its weights.

---

2. Mixture of Experts (MoE)

- **Definition** :
    - A mixture of experts is a machine learning

lightweight than the teacher model.

3. **Performance preservation** : The student model aims to preserve the performance of the teacher model as much as possible.

Mixture of Experts (MoE):
A Mixture of Experts is a neural network architecture that consists of multiple expert models, each specializing in a specific task or subset of the data. The MoE model uses a gating mechanism to select the most suitable expert for a given input and combines their outputs to produce the final result.

Key aspects of MoE:

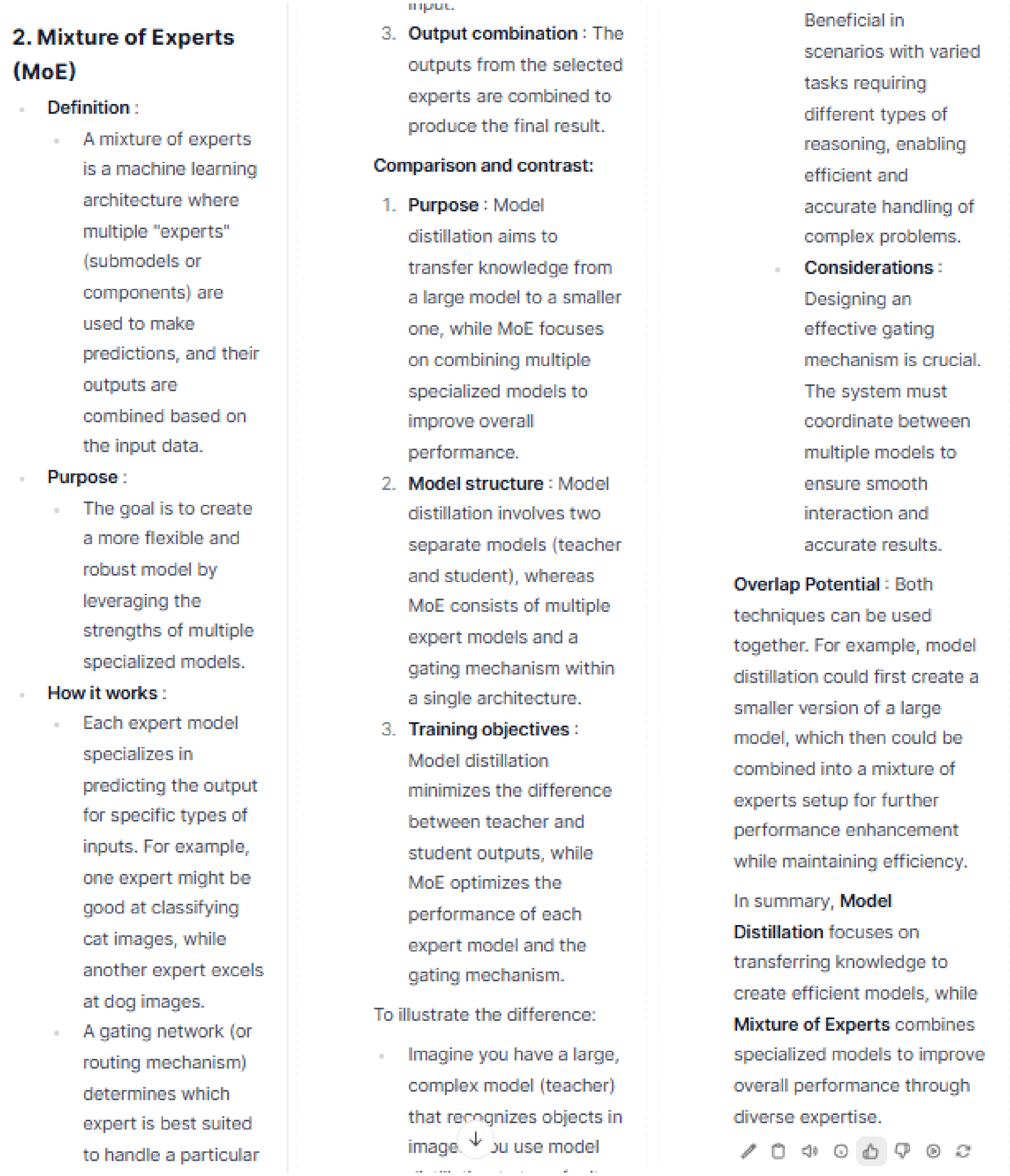
1. **Multiple experts** : Multiple models (experts) are trained to specialize in different tasks or subsets of the data.
2. **Gating mechanism** : A gating function selects the most relevant expert(s) for a given input.
3. **Output combination** : The outputs from the selected experts are combined to produce the final result.

Comparison and contrast:

speed are crucial without compromising performance.

- **Considerations** : Challenges include ensuring accurate knowledge transfer and preventing overfitting. Careful tuning and validation during training are essential.

2. **Mixture of Experts** :
    - **Purpose** : Combine multiple specialized models (experts) to leverage their diverse skills for improved overall performance.
    - **Process** : Each expert model specializes in a specific domain or aspect of the task. A gating mechanism determines which expert is most suitable for a given input, allowing the system to consult the most relevant expertise.
    - **Application** : Beneficial in scenarios with varied tasks requiring different types of reasoning, enabling efficient and



2. Mixture of Experts (MoE)

- **Definition** :
  - A mixture of experts is a machine learning architecture where multiple "experts" (submodels or components) are used to make predictions, and their outputs are combined based on the input data.
- **Purpose** :
  - The goal is to create a more flexible and robust model by leveraging the strengths of multiple specialized models.
- **How it works** :
  - Each expert model specializes in predicting the output for specific types of inputs. For example, one expert might be good at classifying cat images, while another expert excels at dog images.
  - A gating network (or routing mechanism) determines which expert is best suited to handle a particular input.
  3. **Output combination** : The outputs from the selected experts are combined to produce the final result.

Comparison and contrast:

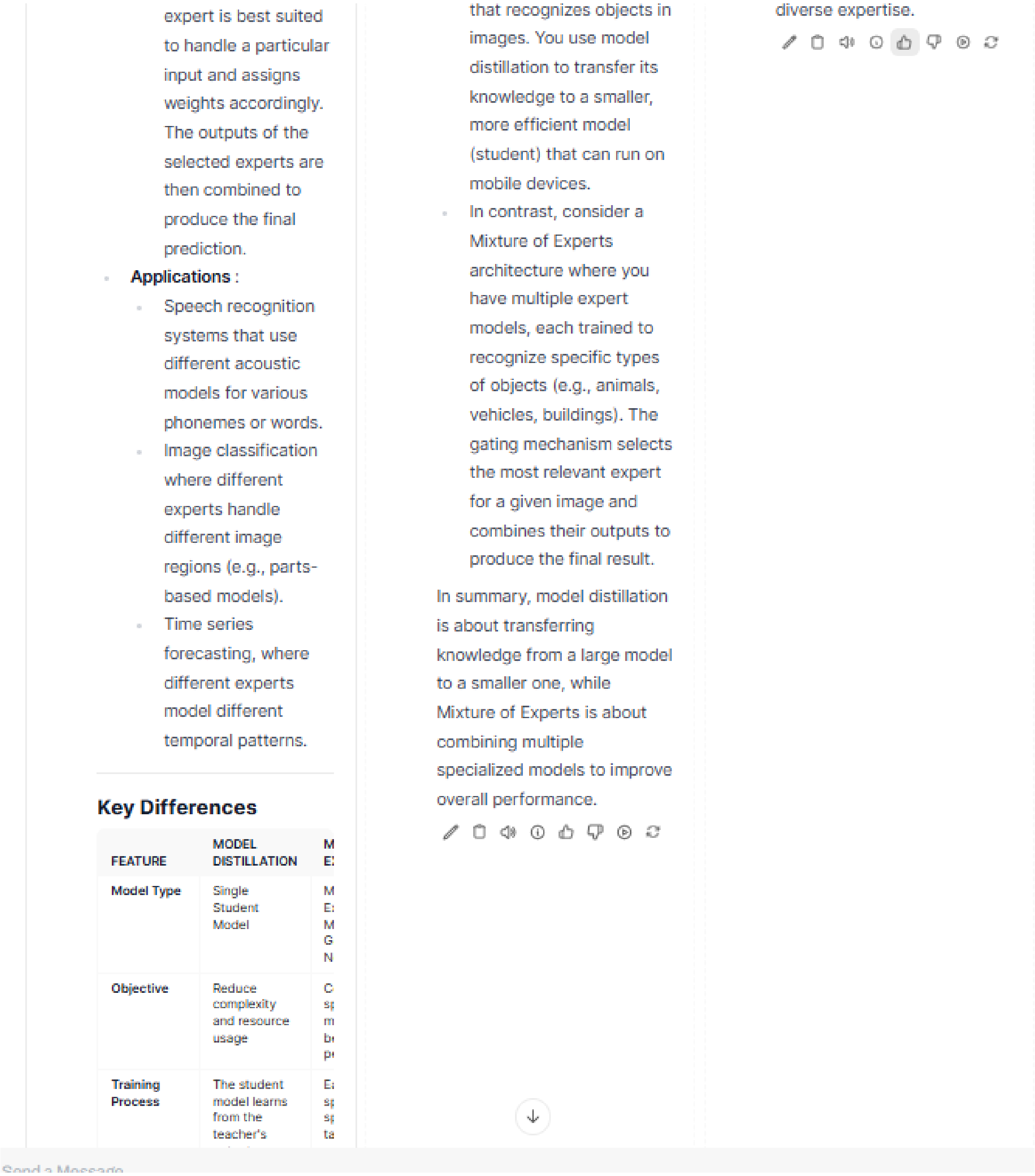
1. **Purpose** : Model distillation aims to transfer knowledge from a large model to a smaller one, while MoE focuses on combining multiple specialized models to improve overall performance.
2. **Model structure** : Model distillation involves two separate models (teacher and student), whereas MoE consists of multiple expert models and a gating mechanism within a single architecture.
3. **Training objectives** : Model distillation minimizes the difference between teacher and student outputs, while MoE optimizes the performance of each expert model and the gating mechanism.

To illustrate the difference:

- Imagine you have a large, complex model (teacher) that recognizes objects in image. ou use model
  - Beneficial in scenarios with varied tasks requiring different types of reasoning, enabling efficient and accurate handling of complex problems.
  - **Considerations** : Designing an effective gating mechanism is crucial. The system must coordinate between multiple models to ensure smooth interaction and accurate results.

Overlap Potential : Both techniques can be used together. For example, model distillation could first create a smaller version of a large model, which then could be combined into a mixture of experts setup for further performance enhancement while maintaining efficiency.

In summary, **Model Distillation** focuses on transferring knowledge to create efficient models, while **Mixture of Experts** combines specialized models to improve overall performance through diverse expertise.



expert is best suited to handle a particular input and assigns weights accordingly. The outputs of the selected experts are then combined to produce the final prediction.

- **Applications** :
  - Speech recognition systems that use different acoustic models for various phonemes or words.
  - Image classification where different experts handle different image regions (e.g., parts-based models).
  - Time series forecasting, where different experts model different temporal patterns.

Key Differences

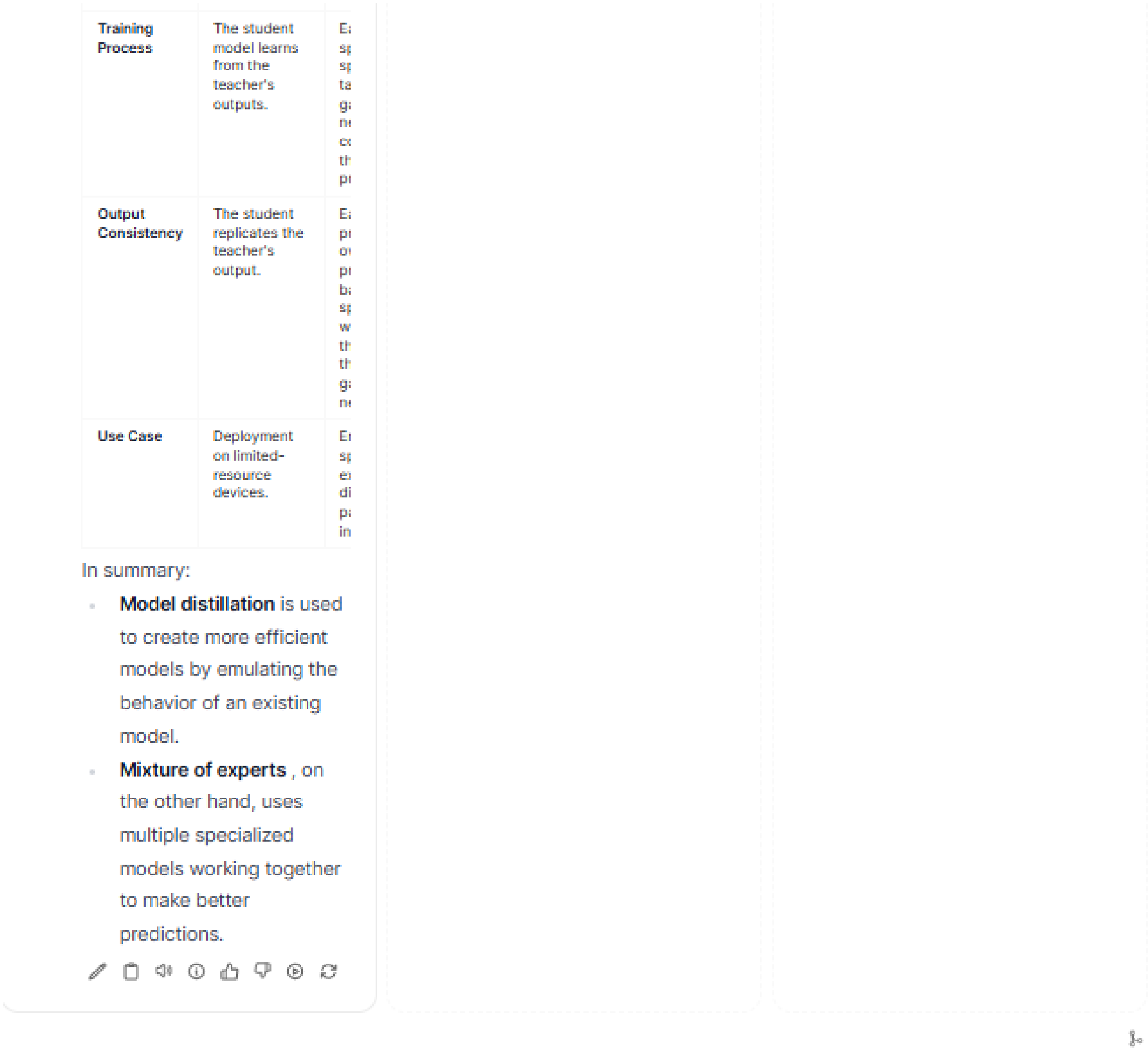
FEATURE	MODEL DISTILLATION	M E
Model Type	Single Student Model	M E M G N
Objective	Reduce complexity and resource usage	C sp m b p
Training Process	The student model learns from the teacher's	E sp sp ta

that recognizes objects in images. You use model distillation to transfer its knowledge to a smaller, more efficient model (student) that can run on mobile devices.

- In contrast, consider a Mixture of Experts architecture where you have multiple expert models, each trained to recognize specific types of objects (e.g., animals, vehicles, buildings). The gating mechanism selects the most relevant expert for a given image and combines their outputs to produce the final result.

In summary, model distillation is about transferring knowledge from a large model to a smaller one, while Mixture of Experts is about combining multiple specialized models to improve overall performance.

diverse expertise.

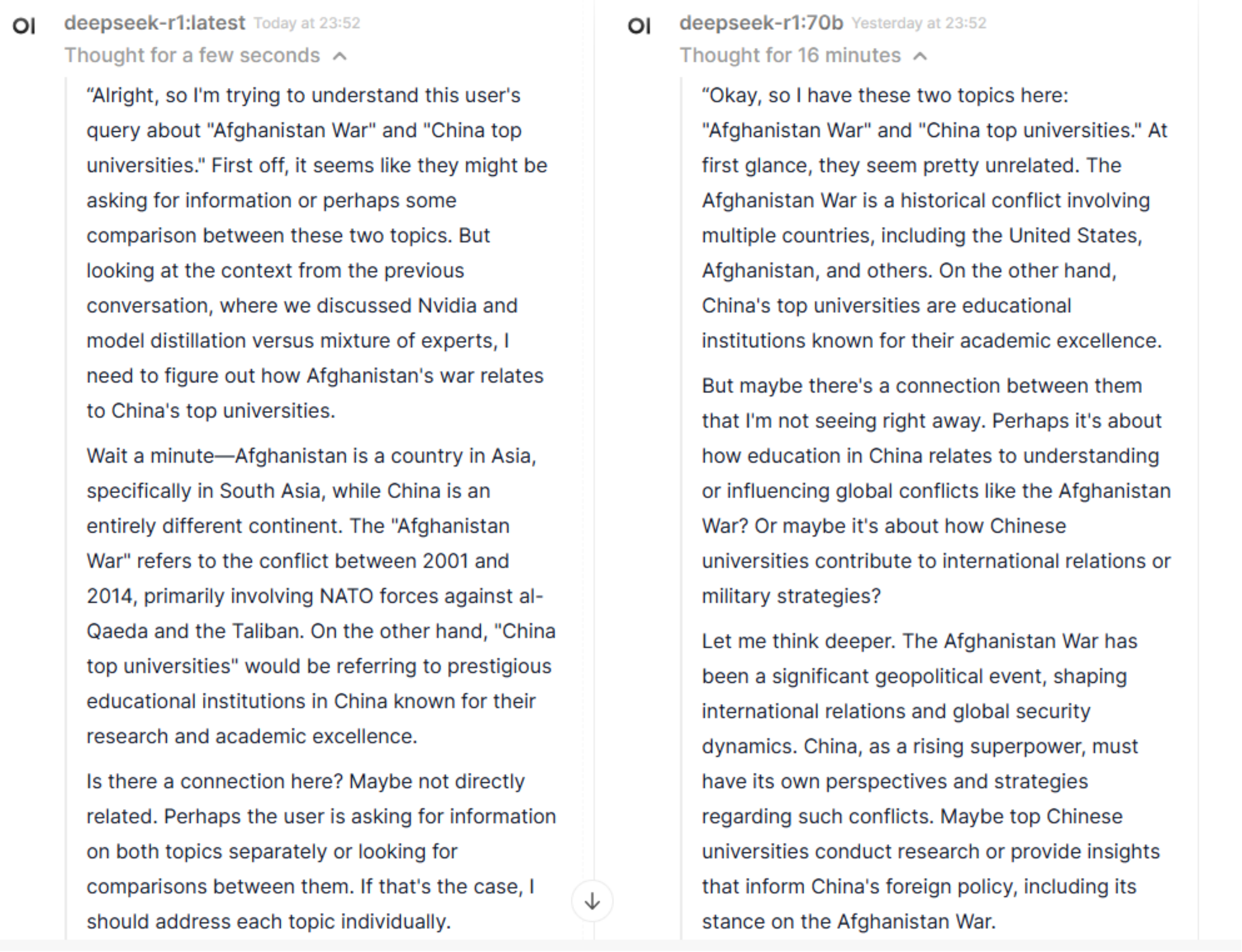


Training Process	The student model learns from the teacher's outputs.	Ea sp sp ta ga ne co th pr
Output Consistency	The student replicates the teacher's output.	Ea pr ou pr ba sp w th th ga ne
Use Case	Deployment on limited-resource devices.	En sp en di pa in

In summary:

- **Model distillation** is used to create more efficient models by emulating the behavior of an existing model.
- **Mixture of experts** , on the other hand, uses multiple specialized models working together to make better predictions.

Test Output 2



deepseek-r1:latest Today at 23:52

Thought for a few seconds

"Alright, so I'm trying to understand this user's query about "Afghanistan War" and "China top universities." First off, it seems like they might be asking for information or perhaps some comparison between these two topics. But looking at the context from the previous conversation, where we discussed Nvidia and model distillation versus mixture of experts, I need to figure out how Afghanistan's war relates to China's top universities.

Wait a minute—Afghanistan is a country in Asia, specifically in South Asia, while China is an entirely different continent. The "Afghanistan War" refers to the conflict between 2001 and 2014, primarily involving NATO forces against al-Qaeda and the Taliban. On the other hand, "China top universities" would be referring to prestigious educational institutions in China known for their research and academic excellence.

Is there a connection here? Maybe not directly related. Perhaps the user is asking for information on both topics separately or looking for comparisons between them. If that's the case, I should address each topic individually.

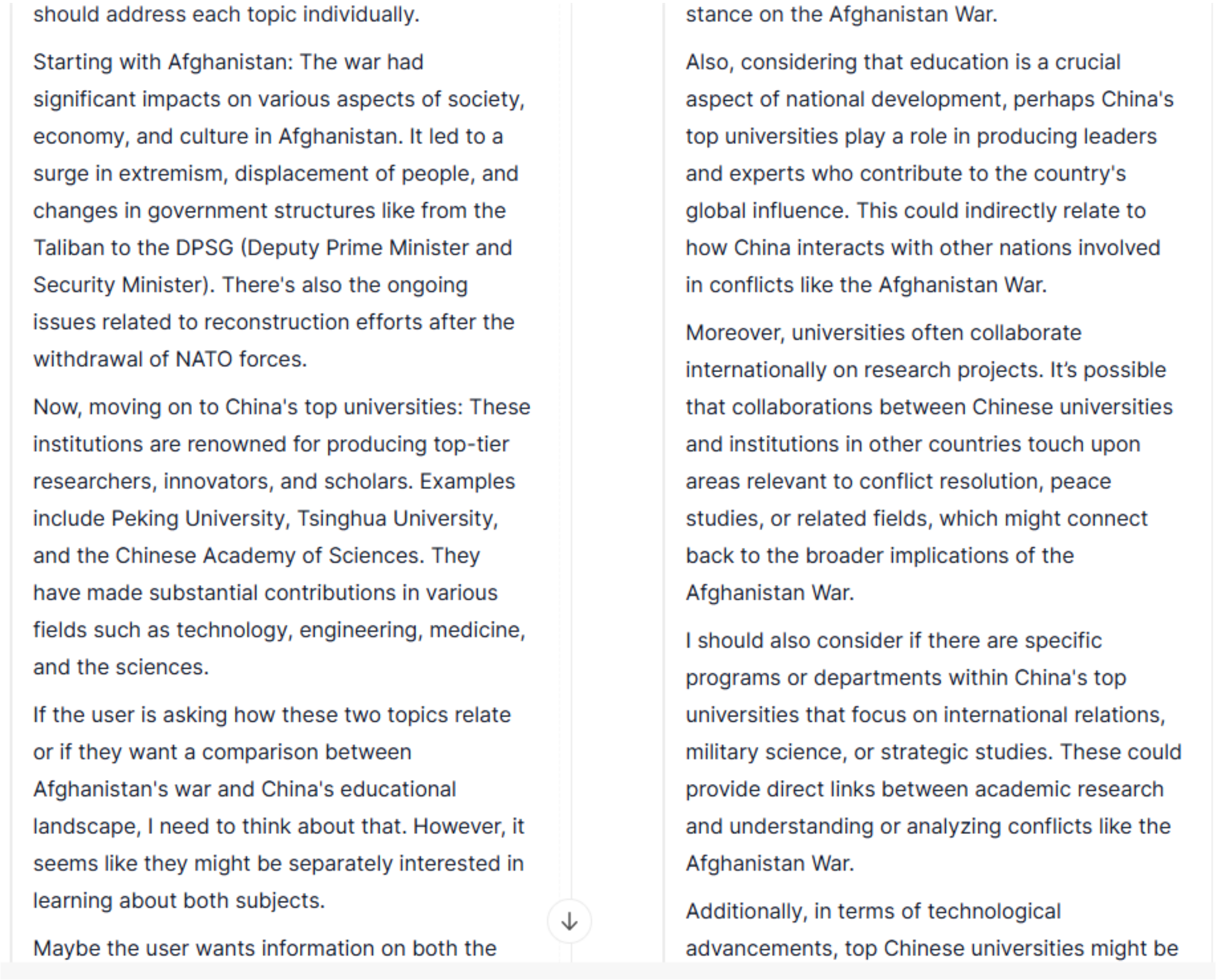
deepseek-r1:70b Yesterday at 23:52

Thought for 16 minutes

"Okay, so I have these two topics here: "Afghanistan War" and "China top universities." At first glance, they seem pretty unrelated. The Afghanistan War is a historical conflict involving multiple countries, including the United States, Afghanistan, and others. On the other hand, China's top universities are educational institutions known for their academic excellence.

But maybe there's a connection between them that I'm not seeing right away. Perhaps it's about how education in China relates to understanding or influencing global conflicts like the Afghanistan War? Or maybe it's about how Chinese universities contribute to international relations or military strategies?

Let me think deeper. The Afghanistan War has been a significant geopolitical event, shaping international relations and global security dynamics. China, as a rising superpower, must have its own perspectives and strategies regarding such conflicts. Maybe top Chinese universities conduct research or provide insights that inform China's foreign policy, including its stance on the Afghanistan War.



should address each topic individually.

Starting with Afghanistan: The war had significant impacts on various aspects of society, economy, and culture in Afghanistan. It led to a surge in extremism, displacement of people, and changes in government structures like from the Taliban to the DPSG (Deputy Prime Minister and Security Minister). There's also the ongoing issues related to reconstruction efforts after the withdrawal of NATO forces.

Now, moving on to China's top universities: These institutions are renowned for producing top-tier researchers, innovators, and scholars. Examples include Peking University, Tsinghua University, and the Chinese Academy of Sciences. They have made substantial contributions in various fields such as technology, engineering, medicine, and the sciences.

If the user is asking how these two topics relate or if they want a comparison between Afghanistan's war and China's educational landscape, I need to think about that. However, it seems like they might be separately interested in learning about both subjects.

Maybe the user wants information on both the

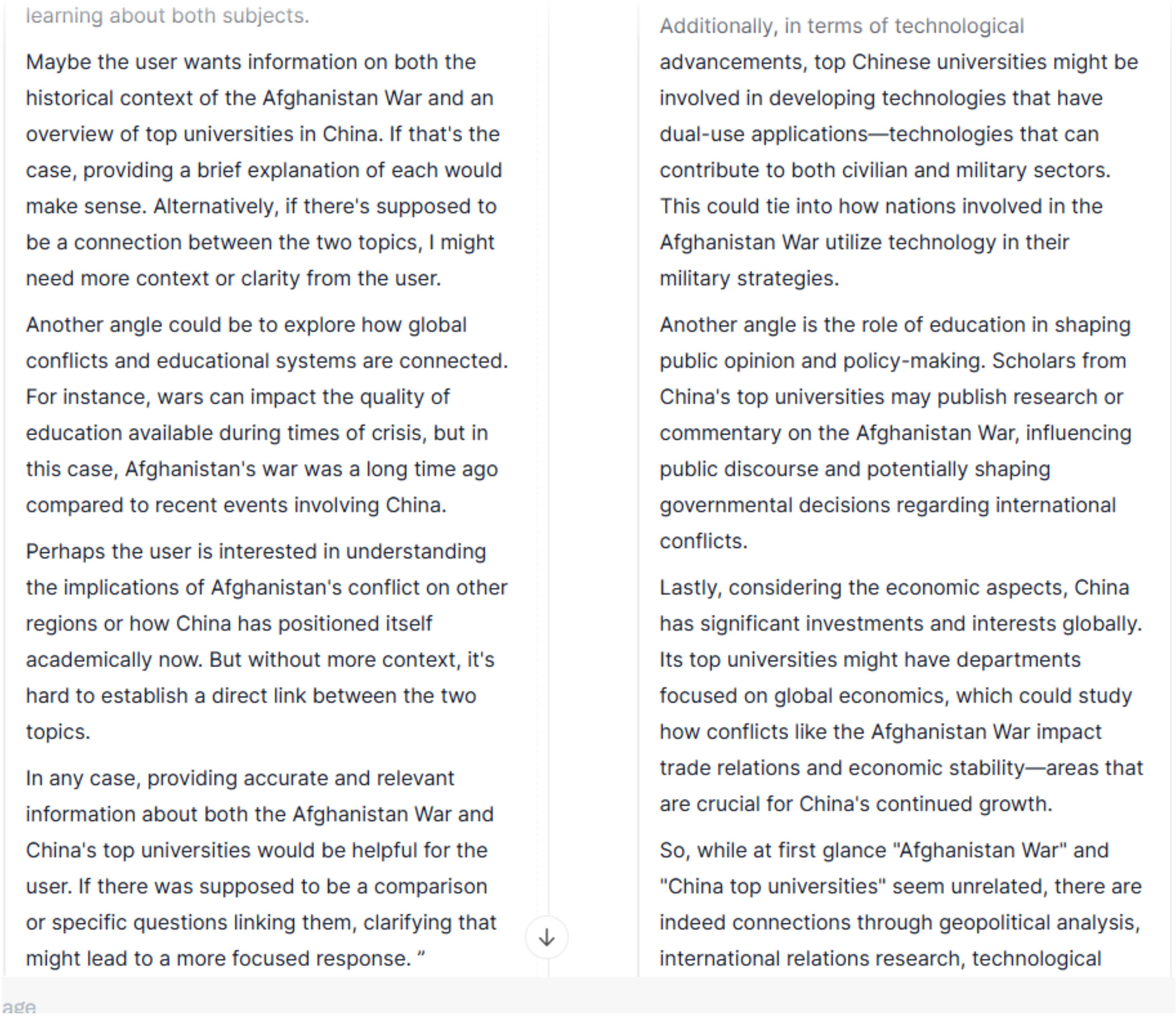
stance on the Afghanistan War.

Also, considering that education is a crucial aspect of national development, perhaps China's top universities play a role in producing leaders and experts who contribute to the country's global influence. This could indirectly relate to how China interacts with other nations involved in conflicts like the Afghanistan War.

Moreover, universities often collaborate internationally on research projects. It’s possible that collaborations between Chinese universities and institutions in other countries touch upon areas relevant to conflict resolution, peace studies, or related fields, which might connect back to the broader implications of the Afghanistan War.

I should also consider if there are specific programs or departments within China's top universities that focus on international relations, military science, or strategic studies. These could provide direct links between academic research and understanding or analyzing conflicts like the Afghanistan War.

Additionally, in terms of technological advancements, top Chinese universities might be



learning about both subjects.

Maybe the user wants information on both the historical context of the Afghanistan War and an overview of top universities in China. If that's the case, providing a brief explanation of each would make sense. Alternatively, if there's supposed to be a connection between the two topics, I might need more context or clarity from the user.

Another angle could be to explore how global conflicts and educational systems are connected. For instance, wars can impact the quality of education available during times of crisis, but in this case, Afghanistan's war was a long time ago compared to recent events involving China.

Perhaps the user is interested in understanding the implications of Afghanistan's conflict on other regions or how China has positioned itself academically now. But without more context, it's hard to establish a direct link between the two topics.

In any case, providing accurate and relevant information about both the Afghanistan War and China's top universities would be helpful for the user. If there was supposed to be a comparison or specific questions linking them, clarifying that might lead to a more focused response. "

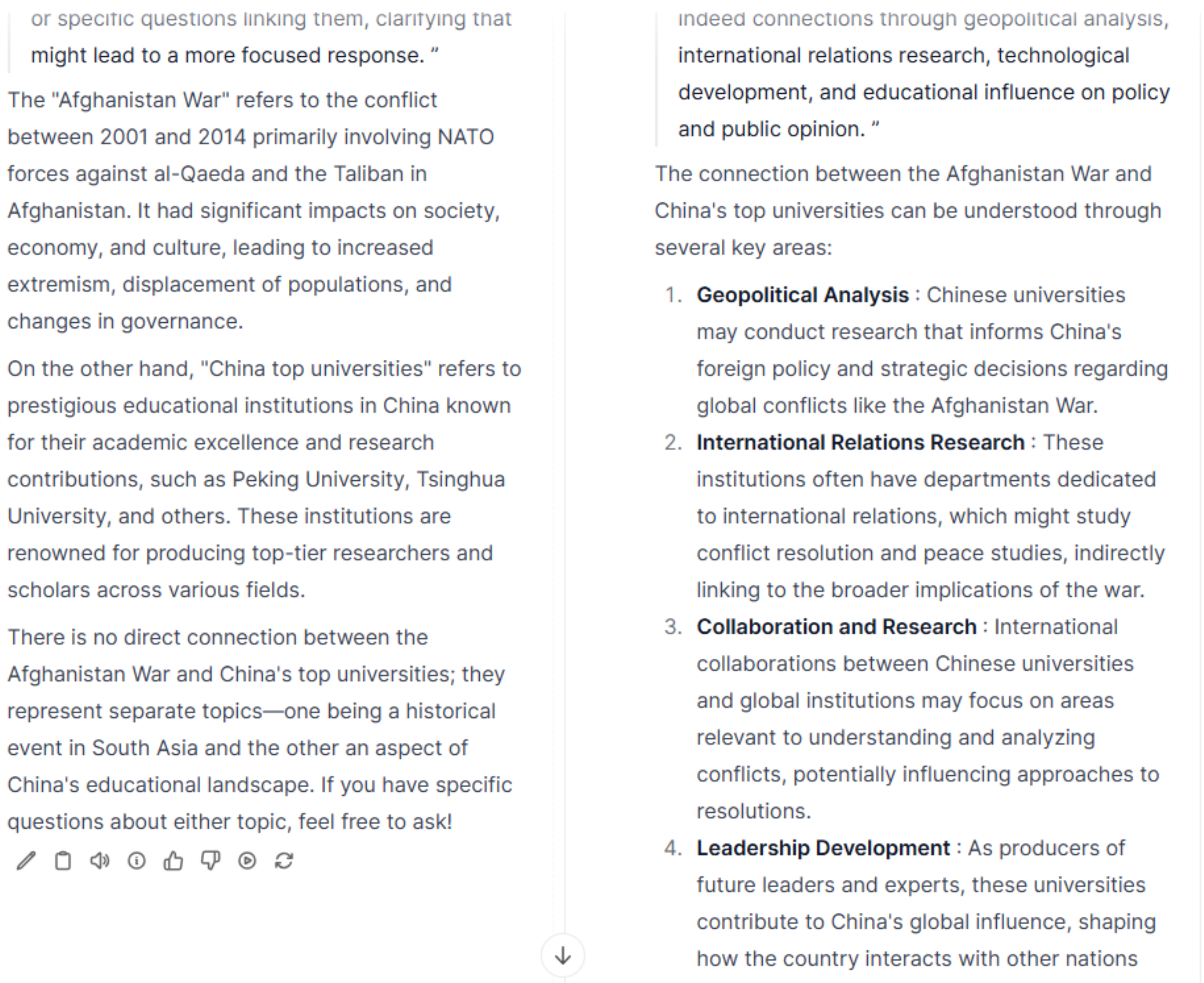
Additionally, in terms of technological advancements, top Chinese universities might be involved in developing technologies that have dual-use applications—technologies that can contribute to both civilian and military sectors. This could tie into how nations involved in the Afghanistan War utilize technology in their military strategies.

Another angle is the role of education in shaping public opinion and policy-making. Scholars from China's top universities may publish research or commentary on the Afghanistan War, influencing public discourse and potentially shaping governmental decisions regarding international conflicts.

Lastly, considering the economic aspects, China has significant investments and interests globally. Its top universities might have departments focused on global economics, which could study how conflicts like the Afghanistan War impact trade relations and economic stability—areas that are crucial for China's continued growth.

So, while at first glance "Afghanistan War" and "China top universities" seem unrelated, there are indeed connections through geopolitical analysis, international relations research, technological

age



> or specific questions linking them, clarifying that might lead to a more focused response. ”

The "Afghanistan War" refers to the conflict between 2001 and 2014 primarily involving NATO forces against al-Qaeda and the Taliban in Afghanistan. It had significant impacts on society, economy, and culture, leading to increased extremism, displacement of populations, and changes in governance.

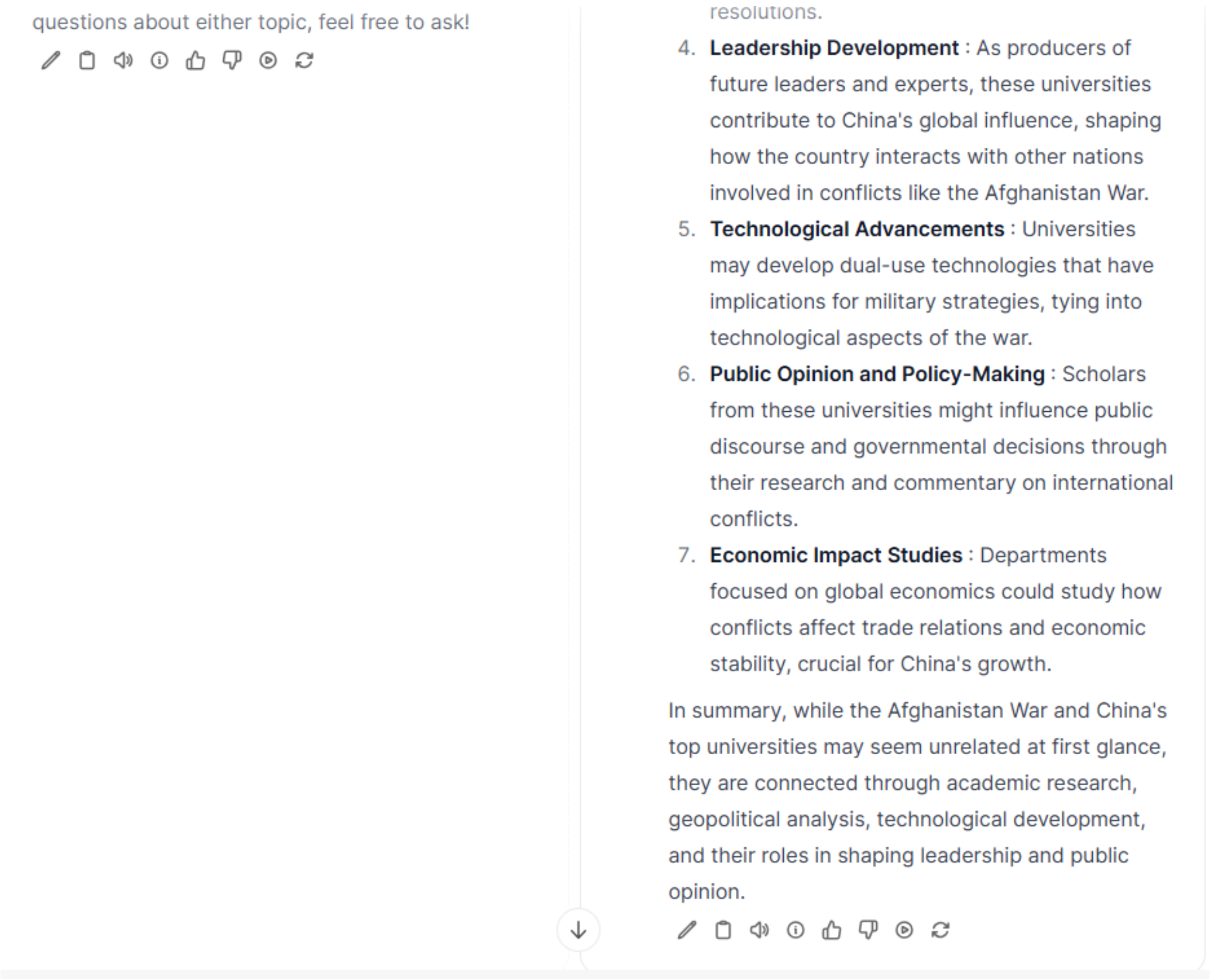
On the other hand, "China top universities" refers to prestigious educational institutions in China known for their academic excellence and research contributions, such as Peking University, Tsinghua University, and others. These institutions are renowned for producing top-tier researchers and scholars across various fields.

There is no direct connection between the Afghanistan War and China's top universities; they represent separate topics—one being a historical event in South Asia and the other an aspect of China's educational landscape. If you have specific questions about either topic, feel free to ask!

> indeed connections through geopolitical analysis, international relations research, technological development, and educational influence on policy and public opinion. ”

The connection between the Afghanistan War and China's top universities can be understood through several key areas:

1. **Geopolitical Analysis** : Chinese universities may conduct research that informs China's foreign policy and strategic decisions regarding global conflicts like the Afghanistan War.
2. **International Relations Research** : These institutions often have departments dedicated to international relations, which might study conflict resolution and peace studies, indirectly linking to the broader implications of the war.
3. **Collaboration and Research** : International collaborations between Chinese universities and global institutions may focus on areas relevant to understanding and analyzing conflicts, potentially influencing approaches to resolutions.
4. **Leadership Development** : As producers of future leaders and experts, these universities contribute to China's global influence, shaping how the country interacts with other nations



questions about either topic, feel free to ask!

resolutions.

4. **Leadership Development** : As producers of future leaders and experts, these universities contribute to China's global influence, shaping how the country interacts with other nations involved in conflicts like the Afghanistan War.
5. **Technological Advancements** : Universities may develop dual-use technologies that have implications for military strategies, tying into technological aspects of the war.
6. **Public Opinion and Policy-Making** : Scholars from these universities might influence public discourse and governmental decisions through their research and commentary on international conflicts.
7. **Economic Impact Studies** : Departments focused on global economics could study how conflicts affect trade relations and economic stability, crucial for China's growth.

In summary, while the Afghanistan War and China's top universities may seem unrelated at first glance, they are connected through academic research, geopolitical analysis, technological development, and their roles in shaping leadership and public opinion.

Observations

- **Memory Usage:** High, averaging at 50% or more
- **GPU Usage:** Ranged from low to high, occasionally reaching 100% but not constantly
- **CPU Usage:** Averaging at 50%
- **Disk I/O:** Low

Model Behaviors

- **Overthinking in Larger Models:** In Test 2, the question with two unrelated topics was intended to test how the models would handle reasoning. The smaller model correctly identified that the topics were unrelated. However, the larger model attempted to establish connections between the topics, demonstrating its artificial intelligence at work.

Summary

In this test run, the DeepSeek R1 models demonstrated the practicality of using AI LLM models in a local environment with total security, without the need for expensive hardware resources.

References

DeepSeek-R1 Release | DeepSeek API Docs