The results of this assignment demonstrate the performance of different AI models in answering a variety of questions. The outputs show responses from four different systems: Base LLM, Base RAG, Agentic RAG, and Agentic RAG (Fine-tuned), with a clear a clear progression in quality and accuracy from the Base LLM to the more advanced Agentic RAG systems.

**Specific Observations**

**Base LLM**

The Base LLM provides general, somewhat outdated information. This is expected. For example, it incorrectly states that AlphaFold 2 is the latest version.

**Base RAG**

The Base RAG system shows improvement by incorporating more recent and specific information. It correctly identifies AlphaFold3 as the latest version and provides more detailed, contextual responses.

**Agentic RAG**

The Agentic RAG system demonstrates further improvement, offering more comprehensive and nuanced answers. It often provides a balanced view of topics, including potential challenges and limitations.

**Agentic RAG (Fine-tuned)**

The Agentic RAG with Finetuned simplifier system takes more comprehensive and accuracy answers of the Agentic RAG system and makes slight improvements in readability by making the wording of its responses more concise. However, even with this improvement in readability, the finetuning is unstable, introducing inaccuracies. For example, in the question about OpenAI's Sora, it is wrong, as it never mentions Sora as a revolutionary text-to-video system, rather it discusses Sora’s application on text-to-text LLM processing tasks.

**Query-Specific Analysis**

1. **OpenAI's Sora**: The responses show a progression from general AI applications to specific text-to-video capabilities, with the fine-tuned model unfortunately being wrong.
2. **COVID-19**: All models provide accurate basic information, but the RAG and Agentic RAG models offer more recent and detailed insights into research and global impact. The finetuned Agentic Model specifically generalizing research into more broad groups rather than focusing on specific research from the abstracts.
3. **AlphaFold**: The Base LLM provides outdated information, while the RAG and Agentic RAG models correctly identify AlphaFold3 and discuss recent developments like HelixFold3. Here, the Agentic RAG model provided more of a readable summary on HelixFold3 and AlphaFold3 whereas the Finetuned Agentic RAG model provided more of a general view on the progression from AlphaFold to Alphafold3 and HelixFold3. I would say that the Agentic RAG model was superior in this case.
4. **AI in Medical Pathology**: The advanced models provide more specific examples of AI applications and ongoing initiatives in pathology, showing a deeper understanding of the field's current state. The finetuned Agentic RAG model provides the best answer to this question.

**Conclusion**

The Agentic RAG model consistently provide more accurate, up-to-date, and comprehensive responses. The fine-tuned model generally gives more readable responses, but has some instability, where in some cases it answers wrong. This is most likely because some information is lost when simplifying documents and ChatGPT3.5 Turbo has a difficult time piecing together the simplified and original abstracts. Overall, I would say to remove the Finetuned LLM and replace it with a Chatgpt3.5 Turbo query into simplifying the abstracts as it’s clear that ChatGpt3.5 is better.

Agentic RAG AI with finetuned abstract simplifier

