Optimising RAG:

Detail two innovative techniques for optimising the RAG model developed in Task 1.

Optimizing a RAG (Retrieval-Augmented Generator) chatbot involves improving both the retrieval and generation components to enhance overall performance.

1. Retrieval Component:

- Optimize the retrieval model to accurately identify relevant passages. I have experiment with different models, architectures, and hyperparameters.
- Used an effective similarity metric for ranking candidate passages. Popular choices include cosine similarity and dot product.

2. Generation Component:

- Fine-tune the language generation model to improve the quality of generated responses. This may involve adjusting hyperparameters, training for longer, or experimenting with different architectures.
- Consider using techniques like reinforcement learning for fine-tuning the generation component.

3. Monitoring and Maintenance:

- Implement monitoring tools to keep track of the chatbot's performance in real-time. Regularly update and maintain the model to adapt to changes in user behavior and language patterns.

Remember that optimization is an iterative process, and it may require multiple rounds of experimentation and refinement to achieve the desired performance.