1. Analyze your results, when does it make sense to use the various approaches?

Use the CPU method for very small matrices or when GPU resources are unavailable. For larger matrices, cuBLAS is optimal for performance, while Tiled GPU is a good balance between learning and efficiency.

2. How did your speed compare with cuBLAS?

cuBLAS was consistently the fastest, outperforming all other methods due to its highly optimized implementation, especially for larger matrices.

3. What went well with this assignment?

Implementing and comparing the various approaches allowed for a clear understanding of GPU optimization techniques like tiling and shared memory.

4. What was difficult?

Managing GPU kernel launches and optimizing shared memory usage for the Tiled GPU method was challenging.

5. How would you approach differently?

Focus more on understanding shared memory and GPU optimization techniques before implementation.

6. Anything else you want me to know?

The logarithmic scale was used to distinguish the four methods.