Comparison Report

Model Information:

Model name: Alibaba-NLP/gte-Qwen2-1.5B-instruct

Max sequence length: 512 (Max: 8192)

Model Size: 1.5B

Embedding Dimension: 1536

Max Input Tokens: 32k

Size on disk: 6.62GB

Testing Platform:

Platform name: Google Colab

Device: CPU

Test Basis:

1. Using llama.cpp to convert tensors to .gguf file format
2. Using default model tensors

Dataset Description:

Number of queries: 10

Average query length: 65 characters

Number of documents: 10

Average document length: 410 characters

Language: Nepali

Results:

Latency results

	Using llama.cpp	Without llama.cpp
Query Embedding	62.399462s	304.110372s
Document Embedding	341.39s	485.287s

Similarity results

Fig: Without llama.cpp

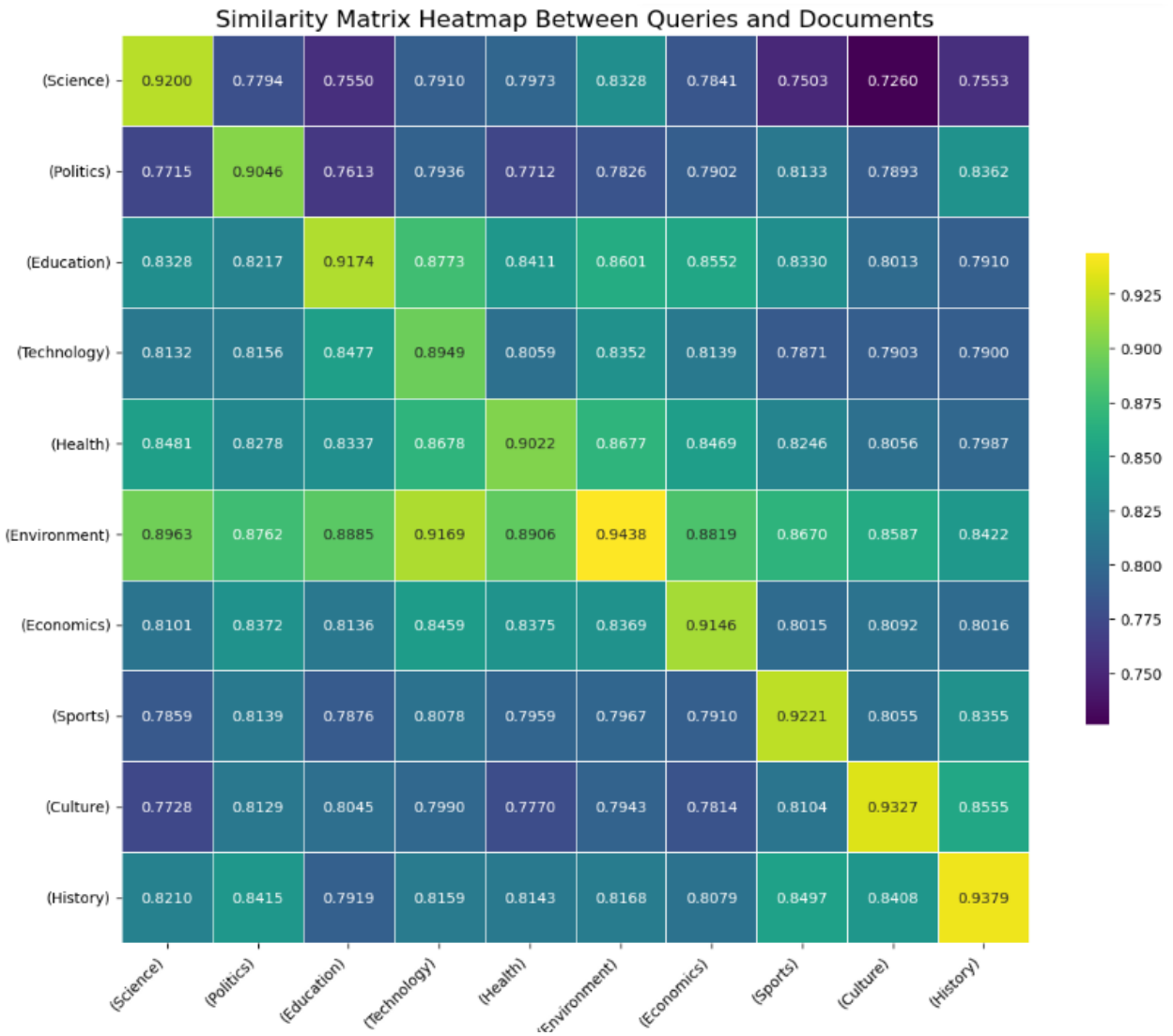

Fig: With llama.cpp (Using Mean pooling)

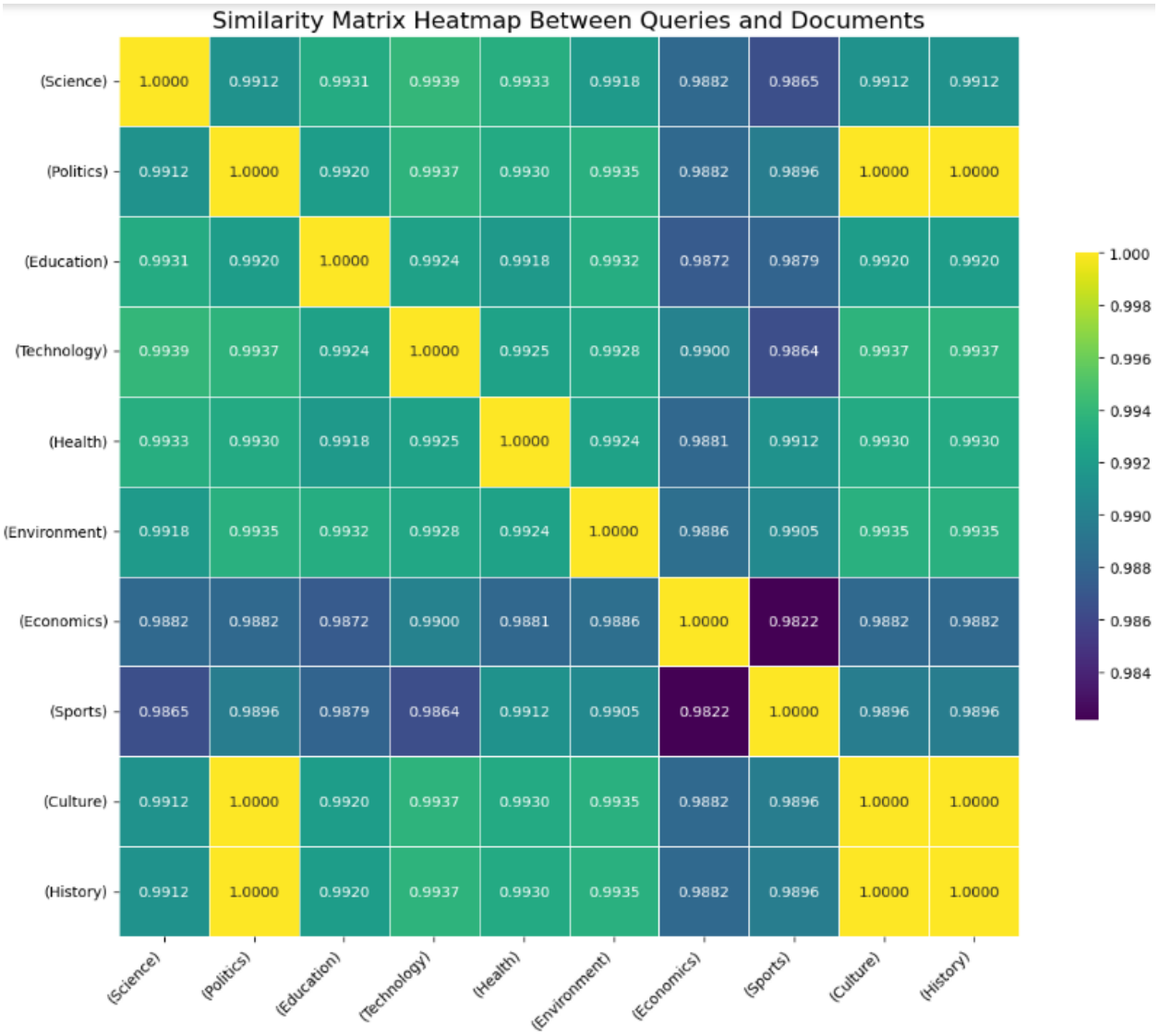

Fig: With llama.cpp (Using CLS pooling)