Phase 4: Parameter Experimentation Report

Experimental Design

This phase systematically evaluated embedding model selection and retrieval depth through 12 experimental configurations. Three embedding models were tested: all-MiniLM-L6-v2 (384d), all-MiniLM-L12-v2 (384d deeper), and all-mpnet-base-v2 (768d). Each was evaluated with four retrieval strategies: top-1, top-3, top-5, and top-10 concatenated passages. All experiments used basic prompting with Flan-T5-base on 100 QA pairs, measuring F1-score and Exact Match.

Results

Embedding Model	Dim	Retrieval	F1	EM	Retrieval Score
MiniLM-L6-v2	384	top-10	62.83	57.0	0.6775
MPNet-base	768	top-10	61.58	57.0	0.6751
MiniLM-L12-v2	384	top-10	61.48	57.0	0.6698
MPNet-base	768	top-5	61.27	56.0	0.6751
MiniLM-L6-v2	384	top-5	60.61	55.0	0.6775
MiniLM-L12-v2	384	top-3	59.98	56.0	0.6698
MiniLM-L12-v2	384	top-5	59.71	56.0	0.6698
MPNet-base	768	top-3	58.74	54.0	0.6751
MiniLM-L6-v2	384	top-3	58.11	52.0	0.6775
MPNet-base	768	top-1	57.09	53.0	0.6751
MiniLM-L12-v2	384	top-1	55.89	52.0	0.6698
MiniLM-L6-v2	384	top-1	53.59	49.0	0.6775

Analysis

Embedding Comparison: The 768-dimensional MPNet model underperformed expectations, averaging 59.67 F1 versus 384d models (58.78-59.26 F1). The baseline MiniLM-L6-v2 achieved the highest peak performance (62.83 F1) while requiring only 2 minutes for embedding generation compared to MPNet's 15 minutes. Comparing MiniLM-L6 versus L12 at identical dimensions showed minimal advantage (0.48 F1) for the deeper architecture, suggesting model depth provides diminishing returns for this task.

Retrieval Strategy Impact: Performance scaled consistently with retrieval depth: top-1 averaged 55.52 F1, top-3 reached 58.94 F1, top-5 achieved 60.53 F1, and top-10 peaked at 61.96 F1. This 6.44-point improvement demonstrates that multiple passages significantly enhance answer quality by providing diverse information sources. Despite exceeding Flan-T5's 512-token limit, concatenated contexts continued improving performance through top-10, indicating effective utilization of truncated multi-passage input.

Key Finding: The optimal configuration (MiniLM-L6-v2 + top-10) achieved 62.83 F1, representing a 9.24-point improvement over the Phase 3 baseline (53.59 F1)—a 17% relative gain. Critically, this improvement stems entirely from retrieval strategy optimization rather than embedding sophistication. Retrieval depth (top-1 to top-10) delivered 9.24 F1 improvement while upgrading embeddings (384d to 768d) provided less than 2 points, demonstrating that retrieval strategy dominates performance outcomes.