Appendix: HQF-DE Implementation Details

A System Architecture

Figure 1 shows the overall HQF-DE system architecture with the four document variants flowing through both retrieval paths.

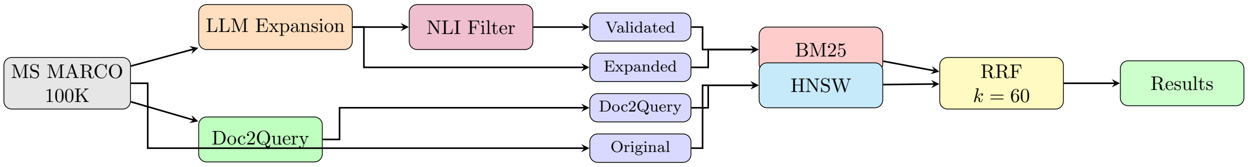


Figure 1: HQF-DE system architecture: document expansion variants flow through dual retrieval paths.

B Pipeline Stages

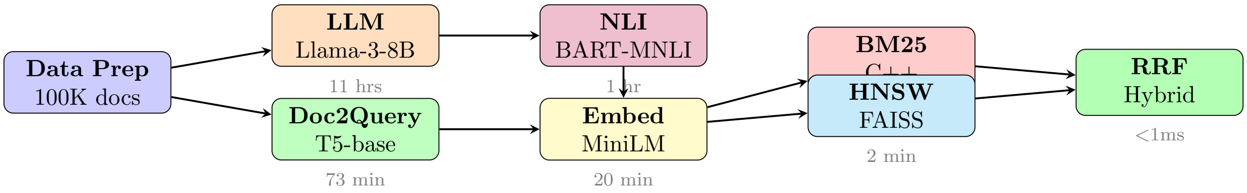


Figure 2: Pipeline stages with processing times. GPU stages run on Colab A100; CPU stages run locally.

C Infrastructure Adaptation

Due to GPU memory constraints on local hardware, we adopted a hybrid execution approach.

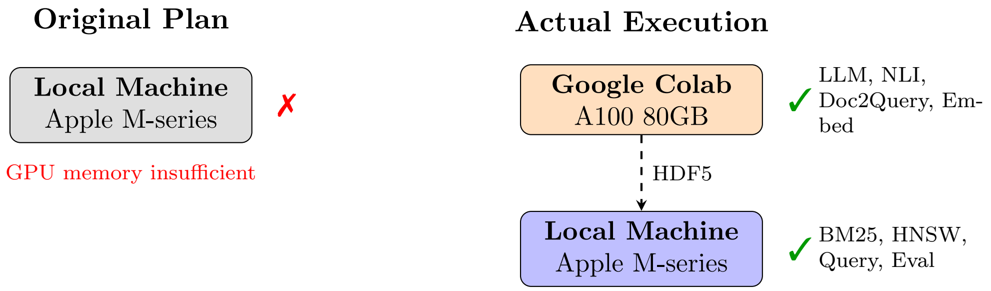


Figure 3: Infrastructure adaptation: GPU tasks on Colab, CPU tasks locally.

Table 1: Time and Space Complexity of Key Components

Component	Time	Space
BM25 Indexing	$O(N \cdot L)$	$O(V + P)$
BM25 Query	$O(\sum_t \lvert L_t \rvert)$	$O(k)$
HNSW Construction	$O(N \log N \cdot M)$	$O(N \cdot M \cdot \log N)$
HNSW Query	$O(\log N + k \cdot ef)$	$O(ef)$
RRF Fusion	$O(k_1 + k_2)$	$O(k_1 + k_2)$

N=docs, L=doc length, V=vocab, P=postings, $|L_t|$=posting list, M=connections, ef=beam width

D Complexity Summary

E Complete Results

E.1 HNSW Dense Retrieval

E.2 BM25 Sparse Retrieval

E.3 Hybrid Retrieval (RRF)

F Key Findings Visualization

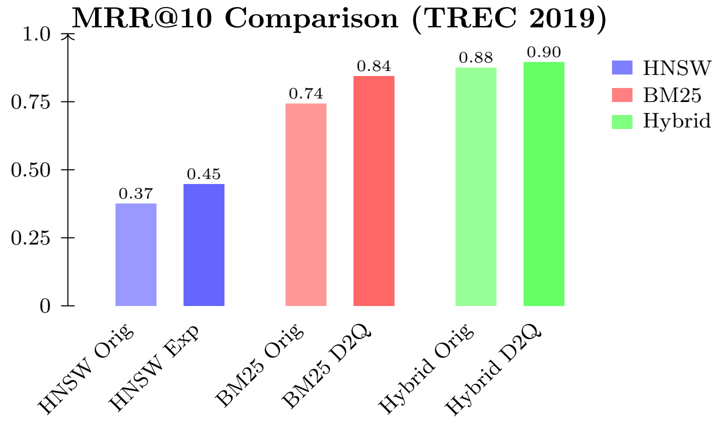


Figure 4: MRR@10 comparison. Hybrid + Doc2Query achieves best performance (0.90).

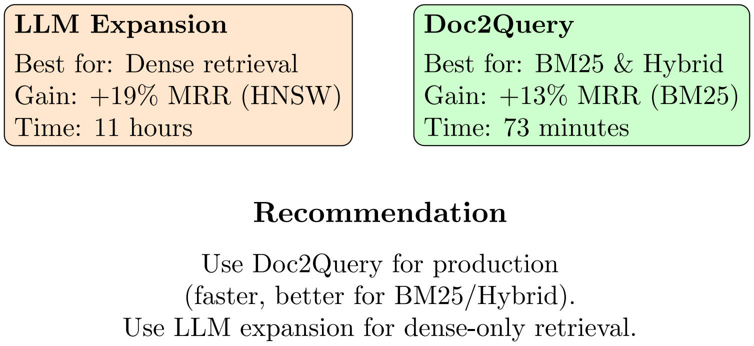
Table 2: HNSW Results (MRR@10 / nDCG@10 / Recall@100)

Variant	**TREC 2019**	**TREC 2020**	**Dev Set**
Original	0.374 / 0.154 / 0.485	0.275 / 0.105 / 0.551	0.011 / 0.002 / 0.876
Expanded	**0.447** / **0.154** / 0.471	0.273 / 0.094 / 0.550	0.012 / 0.003 / 0.866
Validated	0.381 / 0.149 / 0.472	0.267 / 0.092 / 0.551	0.013 / 0.004 / 0.869
Doc2Query	0.305 / 0.134 / 0.472	**0.307** / 0.085 / 0.540	0.013 / 0.003 / 0.865

Table 3: BM25 Results (MRR@10 / nDCG@10 / Recall@100)

Variant	**TREC 2019**	**TREC 2020**	**Dev Set**
Original	0.744 / 0.419 / 0.568	0.688 / 0.428 / 0.590	0.689 / 0.712 / 0.933
Expanded	0.752 / 0.428 / 0.579	0.713 / 0.412 / 0.602	0.658 / 0.685 / 0.908
Validated	0.736 / 0.415 / 0.578	0.703 / 0.411 / 0.601	0.662 / 0.689 / 0.911
Doc2Query	**0.844** / **0.492** / **0.596**	**0.741** / **0.465** / **0.628**	**0.746** / **0.767** / **0.944**

G Expansion Method Comparison



LLM Expansion
Best for: Dense retrieval
Gain: +19% MRR (HNSW)
Time: 11 hours

Doc2Query
Best for: BM25 & Hybrid
Gain: +13% MRR (BM25)
Time: 73 minutes

Recommendation

Use Doc2Query for production
(faster, better for BM25/Hybrid).
Use LLM expansion for dense-only retrieval.

Figure 5: Expansion method comparison with practical recommendations.

H AI Assistance

Claude and Gemini were used to assist with code development, debugging, and LaTeX formatting throughout this project.

Table 4: Hybrid Results (MRR@10 / nDCG@10 / Recall@100)

Variant	**TREC 2019**	**TREC 2020**	**Dev Set**
Original	0.877 / 0.475 / 0.594	0.757 / 0.431 / 0.641	0.701 / 0.730 / 0.945
Expanded	0.810 / 0.459 / 0.588	0.723 / 0.412 / 0.636	0.676 / 0.707 / 0.923
Validated	0.810 / 0.451 / 0.587	0.719 / 0.410 / 0.635	0.680 / 0.709 / 0.926
Doc2Query	**0.896** / **0.512** / **0.613**	**0.779** / **0.468** / **0.653**	**0.762** / **0.783** / **0.951**