A Multistage Evidence Retrieval System for BioASQ Task 14b: Hybrid Retrieval, Reranking, and Snippet Selection

Notebook for the BioASQ lab at CLEF 2026

Yun Wang[1,*]

[1]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia

Abstract

This working note describes our system for BioASQ Task 14b, focusing on retrieval and reranking for biomedical question answering. The system follows a multi-stage retrieval-centered RAG design, combining lexical and dense retrieval with cross-encoder reranking to prepare evidence for downstream answer generation. Experiments show that hybrid retrieval and reranking improve document ranking, while diagnostic analyses highlight remaining limitations in early-rank ordering and evidence selection. These results support a conservative multi-stage design for BioASQ and motivate more adaptive evidence selection in future biomedical RAG systems. Code is available at https://github.com/fulaibaowang/BioASQ.

Keywords

BioASQ, biomedical information retrieval, reranking, reciprocal rank fusion, retrieval-augmented generation

1. Introduction

BioASQ has served for more than a decade as a shared benchmark for large-scale biomedical information retrieval and question answering [1]. Its Task b setting pairs expert-written biomedical questions with a manually curated corpus of relevant documents, supporting text snippets, and reference answers [2], making it a useful testbed for studying whether question answering systems can identify reliable biomedical evidence before generating answers.

Task b is organized into distinct phases. In Phase A, a system receives only the question and must retrieve relevant PubMed documents and extract supporting snippets from them, so evaluation targets evidence retrieval rather than the answer text. In Phase B, the system is instead given the question together with gold-standard relevant snippets and must produce exact and ideal answers. Recent editions add an end-to-end Phase A+, in which a system must return both the retrieved evidence and the corresponding answers directly from the question, without access to any gold relevant material [3, 4].

This working note describes our system for BioASQ Task 14b [5]. We focus on the retrieval side of the task—document retrieval and reranking under the Phase A/A+ setting—and use it as the main benchmark for developing and evaluating a retrieval-centered biomedical RAG system. Recent BioASQ systems have similarly adopted multi-stage pipelines that combine first-stage retrieval, neural reranking, and retrieval-augmented generation, reflecting the practical importance of evidence selection before answer generation [6, 7]. In parallel, other work has explored LLM-based query expansion and refinement for biomedical retrieval-augmented question answering [8]. Our system follows this shared-task line of work and studies how hybrid retrieval, reranking, snippet selection, and query reformulation affect evidence retrieval for biomedical questions.

The goal of this note is not to introduce a new model architecture, but to provide a compact and reproducible account of the system behavior. We therefore report the diagnostic analyses and ablation

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

yun.wang@fri.uni-lj.si (Y. Wang)

https://fulaibaowang.github.io/ (Y. Wang)

0000-0002-2441-4182 (Y. Wang)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

results that guided the final system design. Although the system includes an answer-generation stage, generation quality is not evaluated separately.

2. Methods

2.1. Pipeline overview

We use a retrieve–rerank pipeline for BioASQ Task 14b Phase A/A+ document retrieval. The system also includes document- and snippet-based evidence routes for preparing downstream generation inputs. The overall workflow is shown in Fig. 1.

For each question, the system first retrieves candidate PubMed abstracts using two first-stage retrievers: a lexical BM25 branch and a dense bi-encoder branch. The two ranked lists are combined by reciprocal rank fusion (RRF). The fused candidate pool is then reranked with a cross-encoder. Finally, the cross-encoder ranking is combined with the original retrieval ranking using a second RRF step, which we refer to as post-rerank fusion.

The final document ranking is used for document-level evaluation. The same ranked document set can also be passed to a snippet route, where local sentence windows are extracted, reranked, and fused with document-level scores to produce compact evidence for downstream generation.

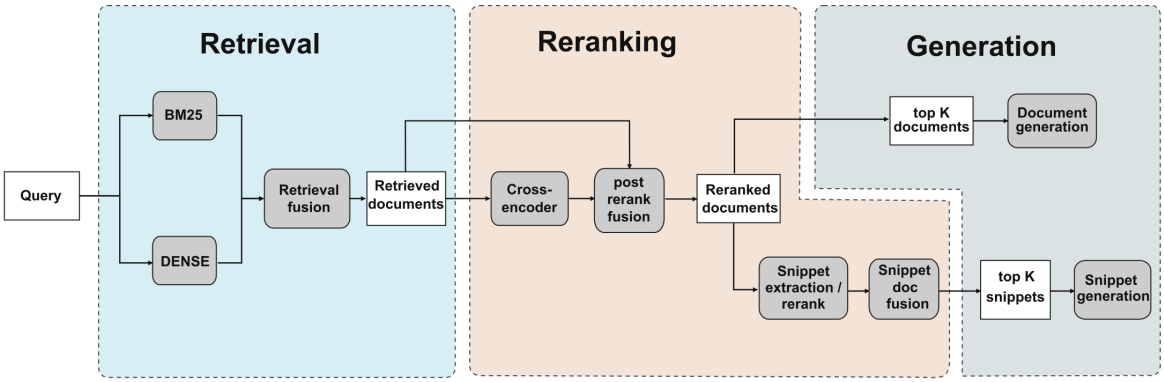

Figure 1: Pipeline overview.

2.2. First-stage retrieval

The first stage retrieves candidate abstracts with lexical and dense retrieval over the same PubMed abstract corpus.

The lexical branch uses BM25 [9] with RM3 [10] query expansion over title and abstract text. In this report, BM25 in the results refers to this BM25+RM3 setting. The dense branch uses abhinand/MedEmbed-small-v0.1 [11] as the bi-encoder and retrieves from an HNSW [12] index built over the same corpus. Each branch returns the top 5000 candidate documents per query.

The two ranked lists are fused with reciprocal rank fusion (RRF) [13]:

$$\mathrm{RRF}(d) = \sum_{i \in \mathcal{R}} \frac{w_i}{k_{\mathrm{RRF}} + \mathrm{rank}_i(d)},$$

where d is a candidate document, i indexes the retrieval branch, and w_i is the branch weight. The default retrieval fusion uses equal branch weights, $(w_{\mathrm{BM25}}, w_{\mathrm{dense}}) = (1, 1)$ and $k_{\mathrm{RRF}} = 150$.

2.3. Cross-encoder reranking and post-rerank fusion

The top 2000 candidates from first-stage fusion are passed to a cross-encoder reranker. The default reranker is BAAI/bge-reranker-v2-m3 with maximum input length 512. Each query–document pair

is formed by concatenating the question with the candidate document's title and abstract, truncated to the model's maximum input length, and is scored independently; candidates are then sorted by cross-encoder score. Most PubMed abstracts fit within the 512-token limit, so only a minority of inputs are truncated at this setting; the effect of a much shorter limit is examined separately below.

We report two reranking outputs. CE rerank is the ranking produced directly by the cross-encoder. Post-rerank fusion combines the CE ranking with the original first-stage hybrid ranking using a second RRF step. The default post-rerank fusion setting is $k_{\mathrm{RRF}} = 60$, $(w_{\mathrm{rerank}}, w_{\mathrm{retrieval}}) = (0.8, 0.2)$.

Because model capacity and input length may affect cross-encoder relevance estimation, we evaluate four reranking configurations on the same first-stage candidate pool: cross-encoder/ms-marco-MiniLM-L12-v2 [14, 15] (33.4M), BAAI/bge-reranker-v2-m3 [16, 17] (568M), BAAI/bge-reranker-v2-m3 [16, 17] with maximum length 200, and BAAI/bge-reranker-v2-gemma [18, 19] (2.51B).

2.4. Snippet route

Because generation often favors compact evidence, we test whether snippet-level reranking preserves document MAP while producing shorter evidence units. The snippet route operates after document reranking. Starting from the post-rerank document shortlist, we split each abstract into overlapping sentence windows (default: 3-sentence windows, stride 1). Window selection then proceeds in two stages. In stage A, candidate windows within each abstract are scored by BM25 and by the dense bi-encoder, abhinand/MedEmbed-small-v0.1, and fused by RRF; only a few top windows per abstract are retained. This stage acts as a cheap intra-abstract filter so that the cross-encoder budget is spent on plausibly relevant windows rather than on the full sentence grid. In stage B, the preselected windows are scored with the same cross-encoder used for document reranking. For each PMID, snippet scores are aggregated by keeping the best-scoring (max-pool) window. The resulting snippet-level ranking is then fused with the document-level ranking using RRF.

The default snippet fusion is document-heavy, $(w_{\mathrm{doc}}, w_{\mathrm{snippet}}) = (0.8, 0.2)$, and we also sweep the weights from document-only to snippet-only to test the effect of snippet evidence on document MAP.

2.5. Query reformulation ablations

In addition to the original BioASQ question text, we evaluate two LLM-based query reformulation strategies as ablations. These variants test whether changing the query representation can improve retrieval or reranking under controlled settings. The exact prompts for both strategies are provided in the released repository.

The first strategy is HyDE [20]: for a given question, an LLM generates a short hypothetical biomedical answer-like passage. This passage is not treated as an answer to be evaluated. Instead, it serves as an expanded query representation that may contain terms, entities, and contextual phrasing related to relevant abstracts. Whether to apply HyDE is decided automatically for each question by an LLM, which disables HyDE for questions where an answer-like rewrite could blur a specific retrieval target, such as numeric, measurement, or exact-identifier questions.

The second strategy is direct query rewriting. Here, the LLM rewrites the original question while preserving its information. We evaluate two variants: a conservative rewrite that only corrects obvious typos, grammar, and telegraphic phrasing, and a broader rewrite that adds mild retrieval-oriented wording without adding new biomedical facts, synonym lists, or extra constraints.

For both strategies, the reformulated query text is evaluated in controlled ablations, while the rest of the pipeline is kept fixed. This allows us to distinguish gains from query reformulation from gains due to changes in retrieval depth, reranker model, or fusion settings.

2.6. Submitted system configurations

The official BioASQ Task 14b submissions correspond to four main configurations of the pipeline described above: a document route, the corresponding snippet route, a snippet route with a stronger reranker, and a snippet route that additionally applies HyDE-based dense-query reformulation (Table 1).

All four share the same retrieve–rerank pipeline, while the snippet-based configurations additionally apply snippet-window selection and snippet-level rescoring before document/snippet fusion.

Table 1
Main system configurations submitted to BioASQ Task 14b Phase A/A+. All configurations share the same BM25+dense RRF first stage and cross-encoder reranking, and differ in the final evidence route, the reranker, and the dense-query representation.

Configuration	Evidence route	Reranker	Dense query
S1	Document	BAAI/bge-reranker-v2-m3	Original
S2	Snippet	BAAI/bge-reranker-v2-m3	Original
S3	Snippet	BAAI/bge-reranker-v2-gemma	Original
S4	Snippet	BAAI/bge-reranker-v2-gemma	HyDE

3. Experimental setup

This section describes the datasets, corpus, metrics, and experiment map used for the development analyses reported below.

3.1. Evaluation splits

Most experiments are evaluated on two question sets. The internal development set contains 583 questions, sampled at approximately 10% from the available BioASQ training questions and augmented with a small number of difficult retrieval cases for fast iteration and stress testing. The second set is formed by merging the BioASQ 13B1–4 batches, with 85 questions per batch. We use the terms "dev" and "13B1–4" to denote these two evaluation sets, and use them primarily to check whether the observed retrieval and reranking behavior is consistent across question samples, rather than to imply a strict train/test split. One reformulation experiment uses smaller subsets: dev-small contains 192 questions sampled from dev, and the 13B subset contains 50 questions sampled from 13B1–4.

3.2. Corpus and indexing

We use the PubMed 2026 baseline as the retrieval corpus. Documents are represented at the PMID level using title and abstract fields, and the same title–abstract records are used by both the lexical and dense retrieval branches.

For lexical retrieval, we build a PyTerrier [21] index and apply BM25 with RM3 query expansion. For dense retrieval, we encode the same records with abhinand/MedEmbed-small-v0.1 and build an HNSW nearest-neighbor index. The same dense encoder is used at indexing and query time.

3.3. Metrics

All retrieval metrics are computed at the PMID level against the BioASQ gold document set. Candidate PMIDs are extracted from the document identifier or from the trailing component of the BioASQ document URL.

Recall@K is the fraction of a query's gold documents G_q that appear in its top-K retrieved set, averaged over all queries.

We also report MAP@K using the truncated average-precision definition used for BioASQ Phase A. For each query, average precision at K (AP@K) averages the precision taken at the ranks of the retrieved gold documents within the top K, but normalized by $\min(|G_q|, K)$ rather than by the total number of gold documents $|G_q|$; MAP@K is the mean of AP@K over all queries. Because of this truncated normalization, AP@K and MAP@K are sensitive to the number of gold documents per query, a property we revisit in the gold-set-size diagnostic.

3.4. Experiment map

Table 2 summarizes the experiments reported in the results section. The experiments are grouped into three categories. E1–E3 establish the baseline behavior of the retrieval and reranking pipeline. E4–E5 provide diagnostic analyses of BioASQ-specific retrieval behavior, focusing on gold-set size and question length. E6–E8 evaluate optional pipeline variants, including snippet-based evidence compression and query reformulation for dense retrieval and reranking.

Table 2
Overview of experiments reported in the results section.

ID	Experiment	Main question	Figure
E1	First-stage retrieval	Compare the behavior of lexical BM25+RM3, dense bi-encoder retrieval, and their RRF fusion.	Fig. 2
E2	CE reranking and post-rerank fusion	Does reranking improve over retrieval, and is fusion additive?	Figs. 3–4
E3	Reranker comparison	Compare reranker model size and input-length settings.	Fig. 5
E4	Gold-set-size diagnostic	Does MAP behavior depend on the number of relevant documents?	Figs. 6–7
E5	Question-length diagnostic	Compare first-stage recall and final MAP across question-length bins.	Fig. 8
E6	Snippet route	Can snippet-level evidence preserve document MAP while compressing context?	Fig. 9
E7	HyDE query representation for retrieval	Test LLM-generated hypothetical passages as an alternative dense-query representation.	Fig. 10
E8	LLM query rewriting for reranking	Test whether rewritten question text changes CE reranking.	Fig. 11

4. Results

The following analyses characterize each stage of the pipeline and motivate the components used in the submitted system configurations (Section 2.6): hybrid first-stage retrieval, cross-encoder reranking and post-rerank fusion, the reranker choice, the snippet route, and the query-reformulation options. Unless stated otherwise, results are reported on the dev and 13B1–4 sets.

4.1. First-stage retrieval

Using the pipeline in Figure 1, we first evaluate the retrieval stage. Figure 2 shows the first-stage Recall@K for BM25, dense retrieval, and BM25+Dense fusion. BM25+Dense fusion gives the best recall on both dev and the merged 13B1–4 batches across almost all cutoffs. BM25 alone is consistently stronger than dense retrieval alone, especially at early depths, indicating that lexical matching remains a dominant signal in this task. At the same time, fusion improves over BM25, showing that dense retrieval contributes additional relevant documents that are not fully captured by lexical search.

These results support the use of hybrid RRF retrieval as the first-stage candidate generator. They also suggest that BioASQ document retrieval remains strongly lexical: biomedical entities, exact terms, and phrase-level overlap provide important retrieval signals. Dense retrieval is useful, but in this setting it is better treated as a complementary branch rather than a replacement for BM25.

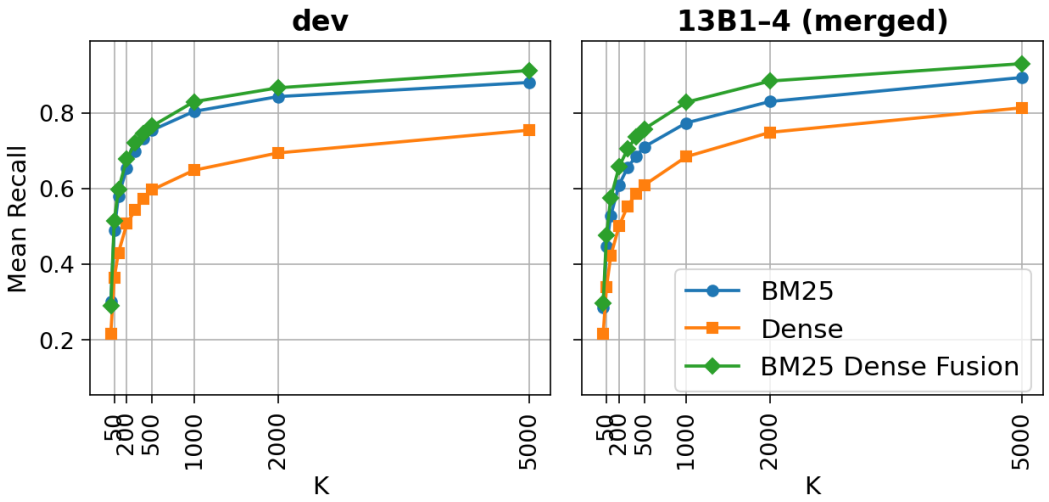

Figure 2: Stage-1 mean Recall@K for BM25, dense, and BM25+Dense RRF fusion. Results are shown on dev and the merged 13B1–4 set. BM25 uses RM3 query expansion and the dense branch uses MedEmbed-small-v0.1 with HNSW; RRF weights are $(w_{\mathrm{BM25}}, w_{\mathrm{dense}}) = (1, 1)$ and $k_{\mathrm{RRF}} = 150$.

4.2. Cross-encoder reranking

We next examine whether cross-encoder reranking improves the hybrid candidate pool. Figure 3 shows that reranking improves Recall@K over BM25+Dense fusion at all evaluated cutoffs from $K = 10$ to $K = 300$ on both dev and the merged batches. This indicates that the reranker is not only changing the order within the very top ranks, but also moves more gold documents into the evaluated top-K range.

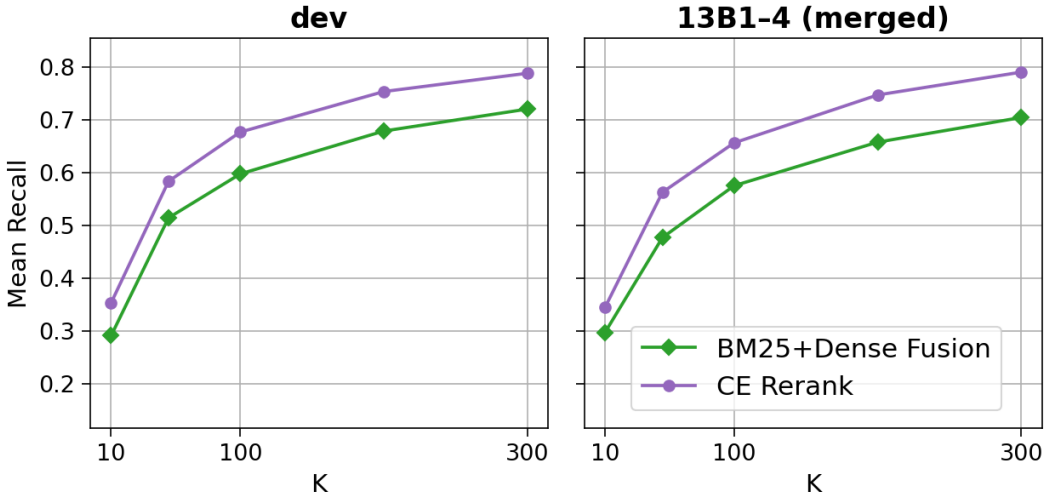

Figure 3: Stage-2 mean Recall@K for BM25+Dense fusion and cross-encoder reranking, evaluated at $K \in \{10, 50, 100, 200, 300\}$. The reranker is applied to the top-2000 retrieval candidates.

Figure 4 shows the corresponding MAP@K results. Cross-encoder reranking improves MAP over the first-stage hybrid ranking, and post-rerank fusion gives the best MAP@K on both splits. This indicates that the cross-encoder adds clear ranking value beyond candidate retrieval alone: it improves both early recall and the placement of relevant documents near the top of the list. The additional gain from post-rerank fusion suggests that the reranked list and the original retrieval list are not redundant. Reranking sharpens local relevance ordering, while the retrieval ranking preserves useful first-stage evidence and coverage.

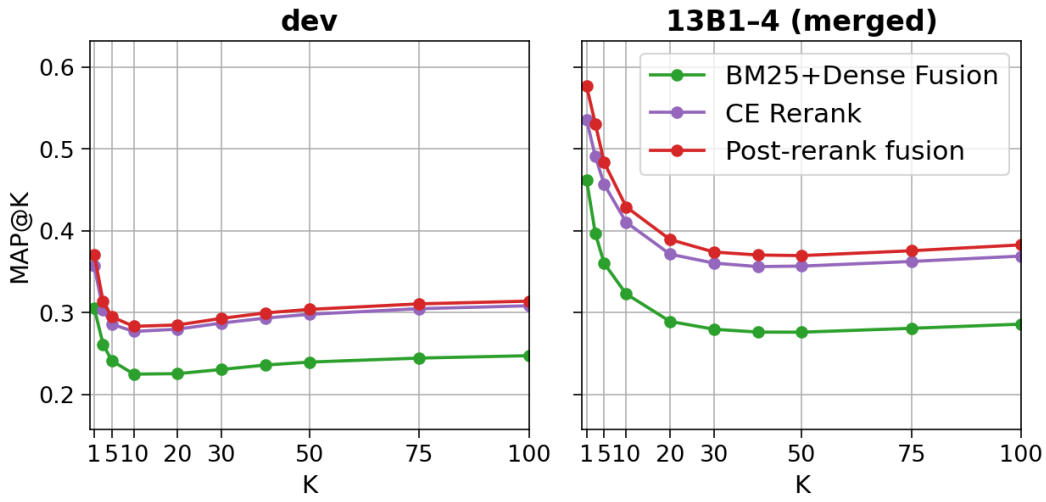

Figure 4: Stage-2 MAP@K for BM25+Dense fusion, cross-encoder reranking, and post-rerank fusion. Post-rerank fusion combines reranker and retrieval rankings with $k_{\mathrm{RRF}} = 60$ and $(w_{\mathrm{rerank}}, w_{\mathrm{retrieval}}) = (0.8, 0.2)$.

We compare rerankers of different sizes on the same BM25+Dense top-2000 candidate pool (Fig. 5). The large bge-reranker-v2-gemma model gives the highest MAP@K across all cutoffs on both dev and 13B1–4. The bge-reranker-v2-m3 is consistently second, while MiniLM-L12 and the truncated bge-m3 setting with maximum length 200 are weaker. The ordering between models is stable across K: the curves drop sharply from very small K to around $K = 10$–20 and then flatten, but the relative model ranking does not change.

These results show a clear reranker-capacity effect in this candidate-pool setting. Once candidate retrieval is fixed, stronger cross-encoders provide better relevance estimation. The advantage of bge-gemma is not limited to top-1 decisions, but persists across the full cutoff range. The drop from full-length bge-m3 to the 200-token version further suggests that aggressive input truncation removes useful evidence and materially hurts document-level ranking.

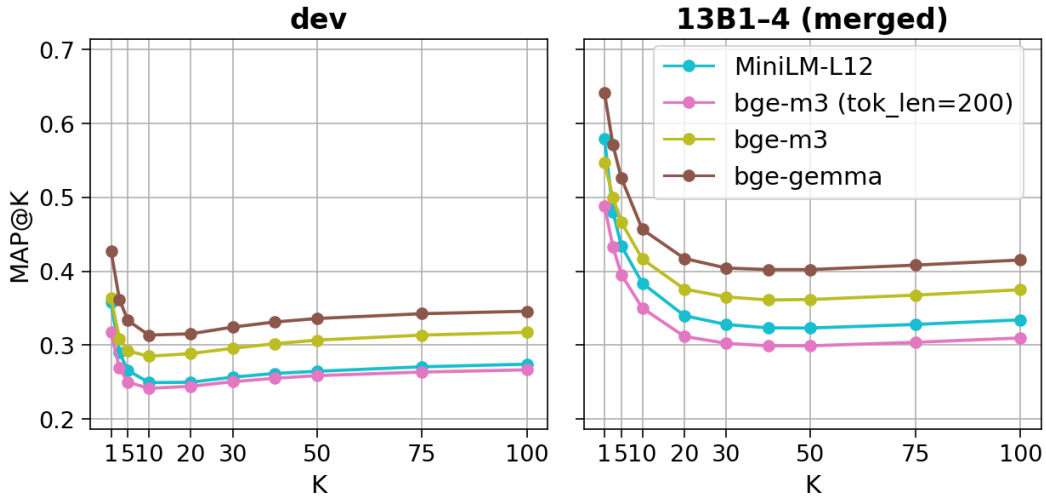

Figure 5: Reranker comparison by MAP@K: cross-encoder/ms-marco-MiniLM-L12-v2 (33.4M), BAAI/bge-reranker-v2-m3 (568M; maximum length 200 or 512), and BAAI/bge-reranker-v2-gemma (2.51B), each applied to the same hybrid stage-1 top-2000 candidates.

4.3. Diagnostics: gold-set size and question length

The aggregate MAP@K curves reveal a consistent pattern across systems: strong performance at rank 1, a drop at small K, and only limited recovery at larger K. This shape is partly influenced by the BioASQ MAP@K calculation, because AP is affected by the number of gold documents available for each query. We therefore analyze performance by gold-set size.

Figure 6 shows that the relevance-set size is highly skewed. There is a visible spike at $|G| = 1$, but the distribution has a long tail, and many queries contain more than five relevant documents. We therefore group queries into four buckets: $|G| = 1$, $|G| = 2$, $|G| = 3$–5, and $|G| > 5$, and compare MAP@K and Recall@K within each bucket in Fig. 7.

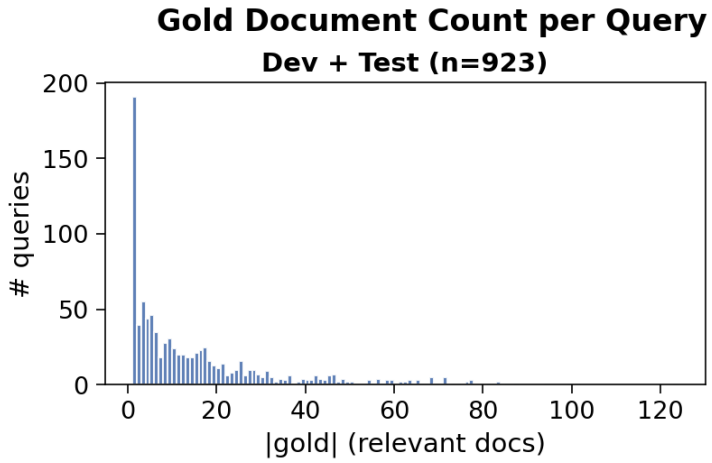

Figure 6: Distribution of the number of gold documents per query on dev plus merged 13B1–4 ($n = 923$).

Figure 7 shows that MAP@K depends strongly on gold-set size. For $|G| = 1$, MAP behaves almost like a single-hit metric: once the relevant document is found, increasing K does not reduce the score. For $|G| = 2$ and $|G| = 3$–5, MAP drops at small K and then stabilizes, while Recall@K continues to increase. This suggests that the first relevant document is often ranked early, but additional relevant documents are spread farther down the list. The effect is strongest for $|G| > 5$, the largest bucket. In this group, Recall@K increases steadily, but MAP remains much lower, showing that the main difficulty is not simply finding any relevant document, but ranking many relevant documents near the top.

Thus, the aggregate MAP@K shape reflects both metric behavior and model behavior. The MAP definition makes the curve sensitive to gold-set size, especially for single-gold versus multi-gold queries, but the bucketed Recall@K results show that the effect is not purely an artifact of the formula. For queries with multiple gold documents, relevant documents continue to be recovered at larger K while MAP remains low, indicating that the main remaining difficulty is ranking several relevant documents early and compactly.

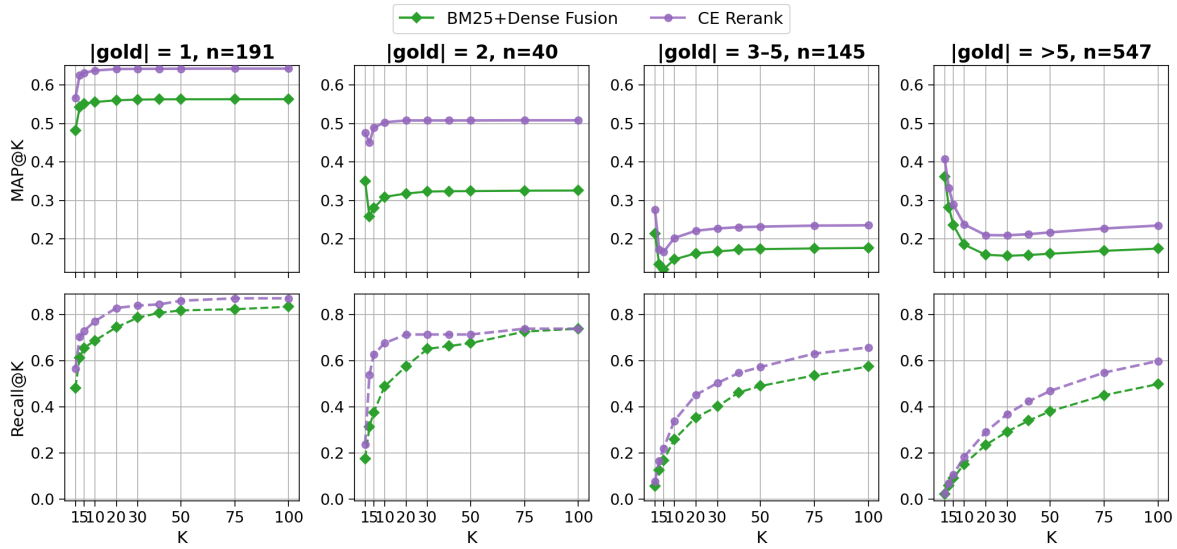

Figure 7: MAP@K (top row) and Recall@K (bottom row) for BM25+Dense fusion and CE reranking, by gold-document-count bucket. Queries are grouped by $|G| = 1$, $|G| = 2$, $|G| = 3$–5, and $|G| > 5$.

We also examine question length as a simple proxy for query specificity. Figure 8 groups questions into short, mid-length, and long bins based on whitespace-tokenized question-body length. In both dev and the 13B1-4, long questions achieve the highest MAP@K, mid-length questions are generally intermediate, and short questions perform worst across most cutoffs. CE reranking improves over BM25+Dense fusion in all length bins, but it does not remove the gap between short and long questions.

The stage-1 recall curves in Fig. 8 show a related but weaker pattern. Short questions have lower first-stage Recall@K than longer questions, especially at small candidate depths, suggesting that they provide fewer lexical and semantic constraints for retrieval. However, the recall gap narrows at larger K. Therefore, the persistent MAP gap cannot be explained only by missing candidates in the first stage. Relevant documents are often present somewhere in the candidate pool, but they are not placed densely enough near the top of the final ranking.

Together, these diagnostics suggest that short questions mainly suffer from weak early ranking rather than complete retrieval failure. Longer questions likely contain more specific lexical and semantic signals, benefiting both hybrid retrieval and CE reranking.

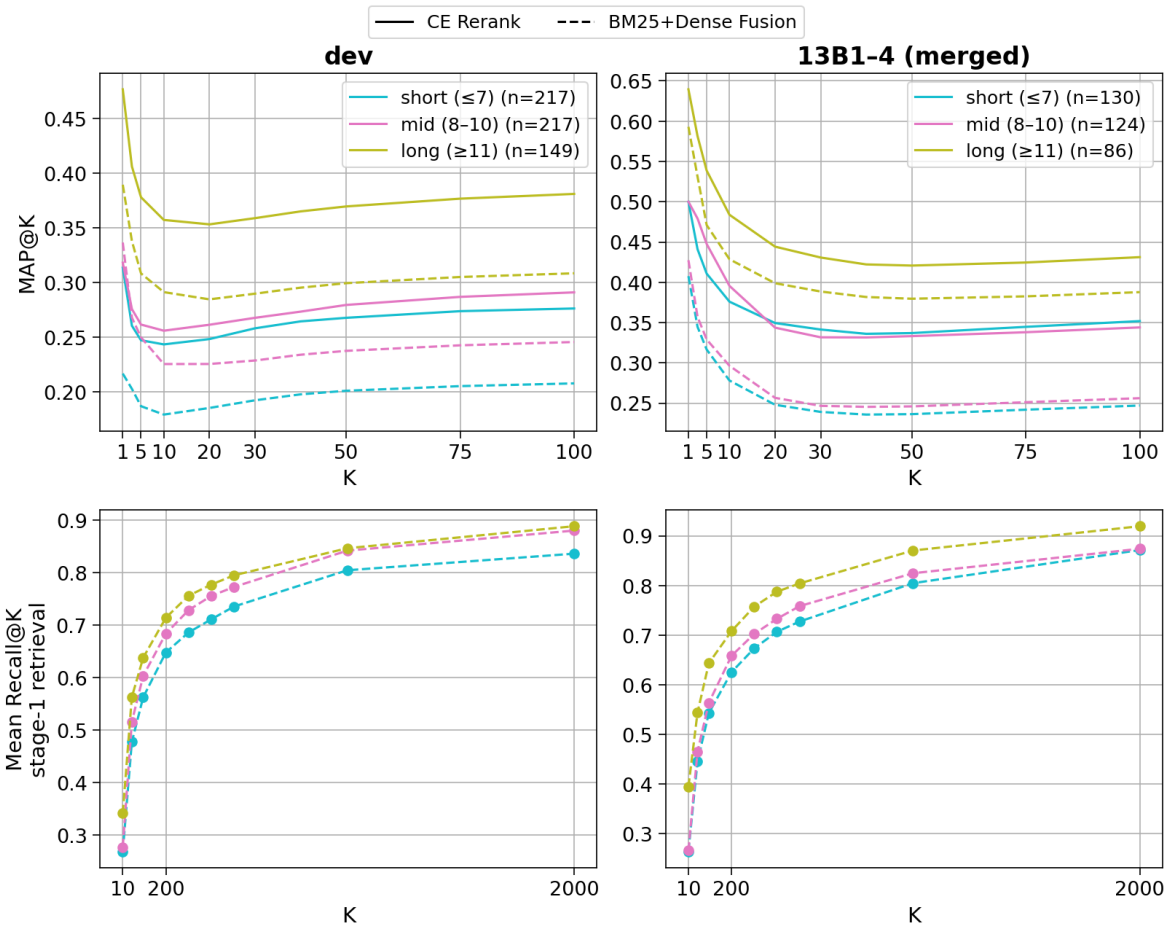

Figure 8: MAP@K and stage-1 Recall@K by question length. Questions were grouped by question-body length: short ($\leq$ 7 words), mid-length (8–10 words), and long ($\geq$ 11 words). The top row compares MAP@K for CE reranking and BM25+Dense RRF fusion. The bottom row shows stage-1 BM25+Dense RRF Recall@K.

4.4. Snippet route

We next compare full-abstract document scoring with snippet-level scoring within the same reranked shortlist (Fig. 9). Snippet scoring is close to full-abstract document scoring on the development set, but is consistently lower on the merged 13B1–4 set. This suggests that local snippets can capture useful evidence, but may lose broader context needed for reliable document-level relevance estimation. The RRF weight sweep therefore mainly serves to identify a safe fusion setting between document and snippet rankings. Across splits, document-heavy mixtures are preferable: they preserve the stronger full-abstract ranking signal while allowing snippet-level evidence to contribute. In contrast, snippet-heavy settings are less stable and degrade more clearly on the 13B1-4 set.

Overall, the snippet route should be interpreted as a context-compression module rather than a method for improving document MAP. It does not reliably outperform full-abstract reranking, but it can provide localized evidence with limited loss when fused conservatively with document scores. This trade-off is useful for the generation stage, where the constraint is not only ranking quality but also how much evidence can fit into the LLM context window.

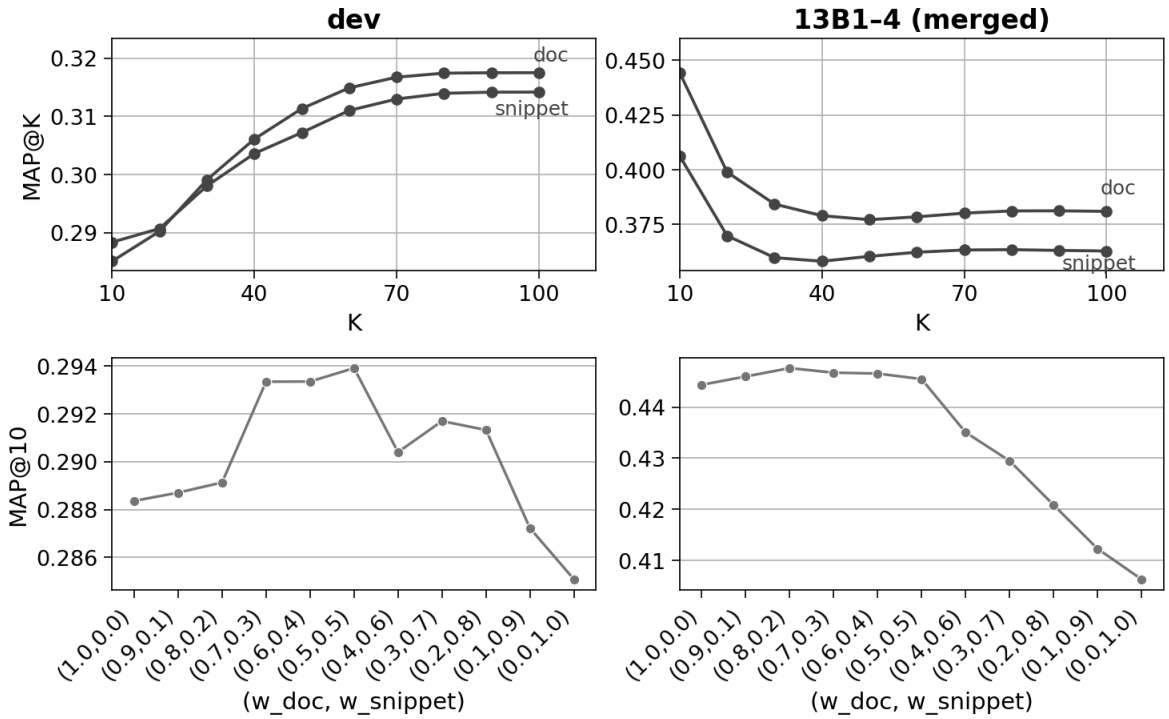

Figure 9: Snippet-route ablation. Top: MAP@K for full-abstract document scoring (*doc*) and snippet scoring (*snippet*) using the same cross-encoder within the post-rerank document shortlist. Snippets are local sentence windows, represented by the best-scoring snippet per document. The bottom row shows MAP@10 under document/snippet RRF weight sweeps, where $(w_{\mathrm{doc}}, w_{\mathrm{snippet}})$ are the document and snippet fusion weights.

4.5. Query reformulation

The previous experiments suggest two query-side limitations. Dense retrieval is weaker than BM25 but complementary after fusion, motivating HyDE for semantic retrieval. In addition, the question-length diagnostic suggests that short questions often suffer from ranking or disambiguation problems even when relevant documents are present, motivating direct query rewriting for cross-encoder reranking.

4.5.1. HyDE on the dense branch

HyDE tests whether a generated evidence-like query representation can strengthen dense retrieval. Figure 10 shows that HyDE consistently improves dense Recall@K over the original query on both dev-small and the 13B subset. The gain appears already at small and intermediate candidate depths and remains positive up to $K = 5000$. This suggests that HyDE helps the dense retriever place relevant documents higher in its candidate list, rather than only adding recall at very large K.

In the full pipeline, this gain may be partly attenuated after BM25+Dense fusion, because BM25 already contributes many overlapping relevant candidates. Nevertheless, HyDE still improves the dense branch and can increase the retrieval pool before reranking, especially for questions where semantic expansion helps recover relevant abstracts not well matched by the original wording.

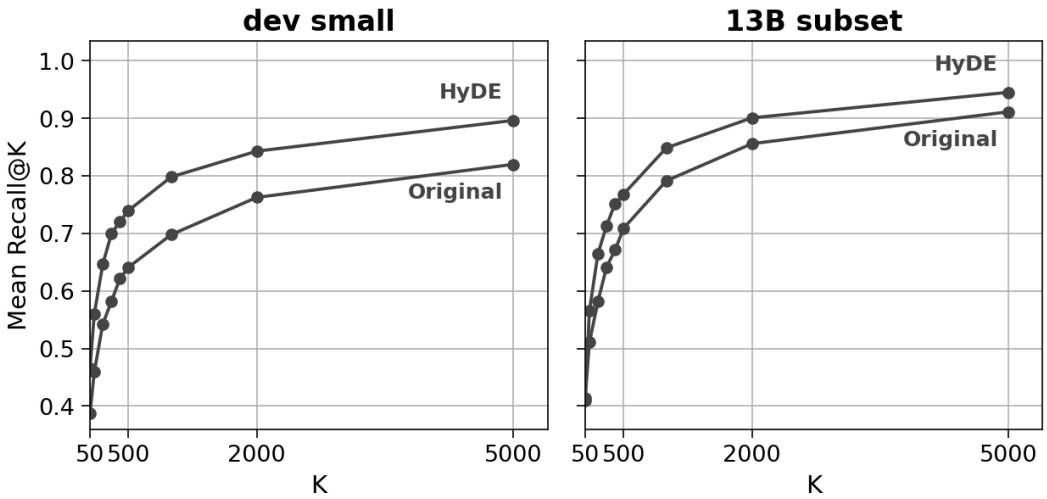

Figure 10: HyDE versus original-query dense retrieval. Mean Recall@K is shown for dev-small ($n = 192$) and the 13B subset ($n = 50$) using the same dense encoder and HNSW index.

4.5.2. LLM query rewriting in reranking: no improvement

The second ablation addresses the ranking-side difficulty suggested by the question-length diagnostic in Fig. 8. We test whether rewriting the query text improves cross-encoder query–document matching while keeping the retrieved candidate pool fixed. This isolates the effect of the query text used by the reranker from changes in first-stage retrieval.

We evaluated two rewriting variants as described in the methods section: Variant A for conservative normalization and Variant B for generic enrichment. Figure 11 shows that query rewriting does not improve reranking performance. The no-rewrite baseline and Variant A are nearly identical on both dev and the merged 13B1–4 set, with only negligible differences across K. This suggests that the CE reranker is already robust to minor surface noise, so typo correction and light grammatical normalization add little once the candidate set is fixed.

Variant B is consistently worse, even though the enrichment is deliberately mild. This suggests that generic additions can dilute the original information need or weaken exact biomedical query–document matching. The contrast with HyDE is therefore informative: HyDE helps when applied as a targeted dense-retrieval intervention, but generic query rewriting at the reranking stage does not help and can be harmful.

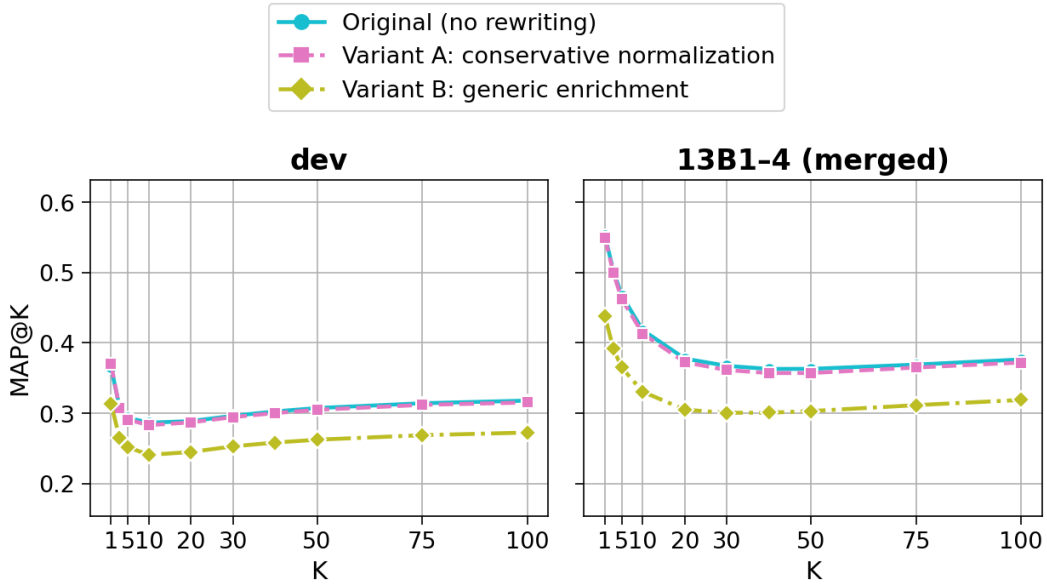

Figure 11: Query rewriting for reranking. MAP@K is shown for the original question, Variant A with conservative normalization only, and Variant B with normalization plus minimal generic retrieval-friendly enrichment. The retrieved candidate pool is fixed; only the query text passed to the cross-encoder reranker is changed.

5. Discussion

This working note mainly evaluates a document retrieval and reranking pipeline for BioASQ Phase A/A+. The results show that a conservative multi-stage design—combining lexical retrieval, dense retrieval, cross-encoder reranking, and post-rerank fusion—is effective for document-level MAP. A central design choice is the repeated use of reciprocal rank fusion: first to combine lexical and dense retrieval, and later to combine the reranked list with the original retrieval ranking. This helps retain complementary signals, since lexical retrieval remains important for exact biomedical entities and terminology, dense retrieval can add semantically related candidates, and the cross-encoder improves local relevance estimation. At the same time, the reported system uses fixed retrieval depths, fixed fusion settings, and a fixed reranking stage. These choices are configurable in the released workflow, but they are fixed across queries in the reported experiments. The system is therefore best understood as a strong MAP-oriented retrieval pipeline, rather than as a fully adaptive or deployment-oriented biomedical QA system.

The diagnostic results suggest, but do not fully prove, that remaining errors often involve early-rank ordering rather than only missing candidates. For multi-gold questions, recall can continue to increase while MAP remains low, indicating that additional relevant documents are often found only deeper in the ranking. For short questions, the weaker MAP may reflect both reduced candidate recall and weaker ranking signals. Since we did not manually inspect failed questions or classify errors by whether gold documents were absent from the candidate pool, these conclusions should be treated as aggregate tendencies rather than complete failure-mode explanations.

The snippet route should also be interpreted within this scope. It is not primarily intended to improve document MAP. Its purpose is to turn retrieved abstracts into more compact local evidence windows for downstream generation. The current evaluation only tests whether this compression substantially damages document-level ranking; it does not evaluate whether the selected snippets improve answer generation, evidence coverage, or factual grounding. Because generation is outside the scope of the present evaluation, future work should evaluate the snippet route under fixed context budgets and with generation-oriented metrics. This distinction is important because a ranking that is optimal for MAP@10 is not necessarily optimal for producing grounded answers.

The negative result for direct query rewriting shows that conservative rewriting does not improve

reranking, while broader rewriting can hurt. This does not mean that all query-side interventions are unhelpful. HyDE can still be useful when targeted at dense retrieval, where the problem is semantic candidate generation. The broader lesson is that query reformulation should be matched to a specific failure mode. Generic rewriting at the reranking stage appears less promising than targeted interventions for candidate generation, short-question disambiguation, or query-aware fusion.

There are also several system-level limitations. The current architecture is not latency-aware, does not optimize GPU cost, and does not use parallel or conditional execution to reduce unnecessary computation. Every question is processed through the same general route, although different BioASQ questions may require different behavior. Some questions may only need lexical retrieval, some may benefit from dense expansion, some may require stronger reranking, and some may require snippet-level evidence selection. A future system could therefore use question-adaptive routing or learned fusion to decide which components are needed for each question.

From an engineering standpoint, the pipeline is released as a Docker-based workflow with a bash orchestrator and configuration templates. Major stage-level choices, including retrieval depth, RRF parameters, reranker model and input length, snippet settings, and document/snippet route selection, are exposed as configuration variables rather than hard-coded. The LLM prompts used in the pipeline—query parsing and HyDE generation, query rewriting, and answer generation, together with the per-type answer schemas—are also included in the released repository. The orchestrator also supports step-wise resume, so individual components can be changed without rerunning the full pipeline. This improves reproducibility and practical experimentation, but does not imply full portability: the numerical results still depend on the PubMed baseline snapshot, BM25 and dense index artifacts, model checkpoints, and available GPU resources. Reproducing the reported numbers therefore requires rebuilding the same corpus and model state, not only running the container. The exact code used for the reported experiments is archived as release v0.1.0.[1]

Overall, this working note shows that a conservative multi-stage retrieval and reranking pipeline is effective for BioASQ Phase A/A+, while also clarifying the limits of optimizing a fixed document-level MAP pipeline. Future work should connect these retrieval gains more directly to the final goal of biomedical RAG: selecting compact, reliable evidence that supports accurate and grounded answer generation. This requires generation-oriented evaluation of snippet-selected evidence, together with more adaptive ranking and evidence-selection strategies for different question types.

Acknowledgments

The author thanks Blaž Zupan for helpful discussions and guidance on this work. This publication is co-funded by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie COFUND Postdoctoral Programme grant agreement No. 101081355–SMASH, and by the Republic of Slovenia and the European Union from the European Regional Development Fund. We also thank the FRIDA cluster at the University of Ljubljana and HPC Vega at IZUM for providing the computational resources used in this work.

Disclaimer. Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to check grammar and spelling and to paraphrase and reword. The author also used Cursor and Claude to assist with code refactoring and debugging for the experimental pipeline. After using these tools and services, the author reviewed and edited the content and code as needed and takes full responsibility for the publication's content.

[1] https://github.com/fulaibaowang/BioASQ/releases/tag/v0.1.0

References

[1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A.-C. N. Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138. URL: https://doi.org/10.1186/s12859-015-0564-6. doi:10.1186/s12859-015-0564-6.

[2] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for biomedical question answering, Scientific Data 10 (2023) 170. URL: https://doi.org/10.1038/s41597-023-02068-4. doi:10.1038/s41597-023-02068-4.

[3] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ tasks 13b and synergy13 in CLEF2025, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1–18. URL: https://ceur-ws.org/Vol-4038/paper_1.pdf.

[4] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ tasks 14b and synergy14 in CLEF2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, 2026.

[5] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[6] R. A. A. Jonker, T. Almeida, J. R. Almeida, S. Matos, BIT.UA at BioASQ 13b: Revisiting evaluation, DPRF-enhanced retrieval and fine-tuned LLMs, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 328–342. URL: https://ceur-ws.org/Vol-4038/paper_22.pdf.

[7] J.-C. Han, B.-C. Chih, H.-C. Hung, R. T.-H. Tsai, NCU-IISR: A retrieval-augmented generation approach for BioASQ 13b phase a and a+, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 292–301. URL: https://ceur-ws.org/Vol-4038/paper_19.pdf.

[8] S. Ateia, U. Kruschwitz, Can language models critique themselves? investigating self-feedback for retrieval augmented generation at BioASQ 2025, 2025. URL: https://arxiv.org/abs/2508.05366. arXiv:2508.05366.

[9] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[10] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, C. Wade, UMass at TREC 2004: Novelty and HARD, in: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), 2004. URL: https://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf.

[11] K. S. Abhinand, abhinand/MedEmbed-small-v0.1, Hugging Face model card, 2024. URL: https://huggingface.co/abhinand/MedEmbed-small-v0.1, accessed: 2026-05-21.

[12] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 824–836. URL: https://doi.org/10.1109/TPAMI.2018.2889473. doi:10.1109/TPAMI.2018.2889473.

[13] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[14] Sentence Transformers, cross-encoder/ms-marco-MiniLM-l12-v2, Hugging Face model card, 2026.

URL: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L12-v2, accessed: 2026-05-21.

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, pp. 5776–5788. URL: https://dl.acm.org/doi/10.5555/3495724.3496209.

[16] BAAI, BAAI/bge-reranker-v2-m3, Hugging Face model card, 2026. URL: https://huggingface.co/BAAI/bge-reranker-v2-m3, accessed: 2026-05-21.

[17] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[18] BAAI, BAAI/bge-reranker-v2-gemma, Hugging Face model card, 2026. URL: https://huggingface.co/BAAI/bge-reranker-v2-gemma, accessed: 2026-05-21.

[19] C. Li, Z. Liu, S. Xiao, Y. Shao, Making large language models a better foundation for dense retrieval, 2023. URL: https://arxiv.org/abs/2312.15503. doi:10.48550/arXiv.2312.15503. arXiv:2312.15503.

[20] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1762–1777. URL: https://aclanthology.org/2023.acl-long.99/. doi:10.18653/v1/2023.acl-long.99.

[21] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative experimentation in python from BM25 to dense retrieval, in: Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021. URL: https://doi.org/10.1145/3459637.3482013. doi:10.1145/3459637.3482013.