A Multi-Stage Retrieval and Ensemble-Based Answer Generation Pipeline for Biomedical QA in BioASQ Task 14b

Notebook for the BioASQ Lab at CLEF 2026

Aniza Shahzadi[1,†], Muhammad Wasim[1,*,†] and Muhammad Tayyab Waqar[1,*,†]

[1]University of Management and Technology, Lahore (Sialkot Campus), Pakistan

Abstract

Automated biomedical Question Answering (QA) over large-scale PubMed literature has emerged as a critical component for supporting faster and more reliable clinical decision-making. Despite substantial progress in the field, retrieving precise evidence and generating accurate answers across different QA types, such as yes/no, factoid, list, and summary, remains a fundamental challenge. Existing systems integrate BM25-based retrieval, neural reranking, and Large Language Models (LLMs) generation; however, they suffer from degraded retrieval precision while proprietary models increase computational cost and reduce reproducibility. Furthermore, existing aggregation methods fail to effectively handle evidence conflicts, leading to a trade-off between precision and recall. To overcome these limitations, this study introduces a scalable and cost-effective multi-stage framework. In the first stage (retrieval/phase A), two independent retrieval pipelines each rerank BM25 candidates using cross-encoders (BAAI/b-reranker-v2-m3 and Alibaba-NLP/gte-reranker-modernbert-base) to improve document ranking precision. Additionally, a two-stage hybrid snippet extraction approach helps to reduce context dilution by selectively filtering passage-level evidence for downstream generation. To guide answer formatting, quality-filtered few-shot examples drawn from the BioASQ training set are injected into generation prompts to guide answer formatting across all QA types. In the generation stage (phase A+), simple consensus voting is replaced with stance-aware ensemble strategies. The system used seven weighted votes for yes/no questions with different reasoning styles, including evidence-focused, opposing, and meta-reasoning strategies. Adversarial consistency checks are used when the voting outcome becomes very close. Factoid answers are generated through Mean Reciprocal Rank (MRR)-weighted aggregation across three extraction stances, and list questions are generated using exhaustive-scan and category-guided prompts. In Phase B, the ensemble strategies are applied over gold-standard annotator-selected snippets, yielding substantial improvements, particularly for list and factoid questions. On the BioASQ Task 14b test batches, the framework achieves a yes/no Macro F1 of 1.0000 in Phase A+ and 0.9352 in Phase B, and a list F-measure of 0.5573 in Phase B, demonstrating that stance-diverse ensemble generation with adversarial consistency significantly enhances answer reliability and precision without dependence on proprietary infrastructure. The source code is publicly available at https://anonymous.4open.science/r/BioASQ_14b-571D/.

Keywords

Biomedical QA, BioASQ, RAG, Neural Reranking, Cross-Encoder, Few-Shot Prompting, Ensemble Generation, Stance-Diverse Voting, Adversarial Consistency Checking, Evidence Grounding

1. Introduction

The rapid expansion of biomedical literature has fundamentally transformed the landscape of clinical decision-making [1]. PubMed contains more than 38 million articles with an increasing demand for automated Question Answering (QA) systems [2]. It supports accurate, efficient, and evidence-based clinical practice and biomedical research [1, 3]. The BioASQ challenge has emerged as a standard evaluation benchmark to fill this gap, assessing QA systems on document retrieval, snippet extraction, and diverse type answer generation [4].

Although biomedical QA systems have improved substantially, several limitations persist, including

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

*Corresponding author.

†These authors contributed equally.

anizashahzadi12@gmail.com (A. Shahzadi); muhammad-wasim@skt.umt.edu.pk (M. Wasim); tayyab.waqar@skt.umt.edu.pk (M. T. Waqar)

vocabulary mismatch. It is a major obstacle in clinical research, particularly when similar concepts are expressed using different biomedical terminology. Furthermore, increasing context length introduces the "lost-in-the-middle" problem in LLMs, which reduces the model's ability to focus on relevant evidence [5]. Another major issue is a persistent misalignment between automated evaluation scores and expert-judged answer quality [6]. Additionally, dependence on proprietary LLMs such as GPT-4 and Claude 3.5 Sonnet, limits reproducibility and fair comparison among researchers [7]. Ensemble generation strategies have demonstrated improved performance compared to single-model biomedical QA frameworks. Majority voting strategies for yes/no QA, ensemble methods for list questions across multiple LLMs attained strong performance in the BioASQ Synergy challenge [8]. In addition, ensemble approaches with open-weight models such as Phi-4 and Qwen2.5 attain competitive performance with proprietary models at significantly reduced computational cost [7]. However, consensus voting and union-based aggregation methods perform poorly under contradictory evidence, reducing precision even when recall improves [8, 5]. Existing methods do not utilize controlled disagreement between models to enhance the grounding of answers across different BioASQ question types, even though the frequent occurrence of contradictory biomedical evidence.

In clinical QA, incorrect answers can have a direct effect on patient safety, highlighting the need for precise and accurate evidence-based answers. Technically, retrieval and generation strategies fix the metric misalignment and precision-recall trade-off [4]. We participate in Task 14b of the BioASQ challenge [9], which is now in its fourteenth year of promoting research in biomedical indexing and question answering. Task 14b focuses on biomedical QA over English documents, requiring systems to retrieve relevant documents, snippets, and generate exact and ideal answers to biomedical questions [10]. This work presents a reproducible and cost-effective pipeline using freely available cross-encoders and API providers. To enhance resilience across BioASQ question types, the proposed system adopts a multi-stage retrieval process, evidence refinement, and type-aware response generation methods. The key contributions of this work are:

- A dual cross-encoder retrieval framework using bge-reranker-v2-m3 and gte-reranker-modernbert-base to rerank BM25 candidates, with a two-stage hybrid snippet extraction for precise and grounded evidence snippets.
- A stance-diverse seven-vote ensemble for yes/no QA with multiple reasoning perspectives using snippet-relevance weighting aggregation and adversarial consistency validation.
- Type-aware aggregation through MRR-based ranking, precision-recall filtering, and quality-controlled generation for summaries.
- Few-shot prompting with type-specific quality-filtering training examples for factoid, list, and summary questions, enforcing verbatim snippet grounding.
- Cost-efficient pipeline with multi-provider failover that sustains uninterrupted generation using freely accessible models.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the methodology. Section 4 presents results and analysis. Section 5 concludes with future directions.

2. Previous Work

Biomedical QA emerged as a benchmark task in Natural Language Processing (NLP), requiring systems to simultaneously perform precise document retrieval, extract meaningful snippets and reliable answer generation across heterogeneous QA types [1, 11]. The BioASQ challenge played a central role in evaluating systems on yes/no, factoid, list, and summary QA against large-scale PubMed corpora [1, 4]. In recent years, biomedical QA systems evolved considerably, shifting from the usual sparse retrieval and extractive pipelines toward neural retrieval, RAG, and LLM-based ensemble approaches [12, 3]. Despite these improvements, drawbacks in retrieval precision and answer reliability motivated more principled multi-stage architectures [3].

Recently, retrieval quality identified as a major bottleneck in biomedical QA frameworks [1]. Empirical evidence established that retrieval precision directly affects downstream answer generation, even high-capacity generative models cannot recover from initially poor document sets [13, 14]. Similarly, dense neural retrieval outperformed sparse BM25 baselines capturing semantic variation in biomedical questions more effectively than inverted-index methods [13]. To further improve retrieval performance, neural reranking techniques were introduced, with cross-encoder reranking demonstrated to substantially improve passage-level precision over sparse retrieval approaches. However, existing systems applied reranking either at the document or snippet level independently, instead of sequentially applying on both stages [14, 15]. This represents an important architectural limitation because snippet extraction was responsible for response generation, received less systematic attention than document retrieval [13, 16]. Furthermore, Pseudo-Relevance Feedback (PRF) showed effective re-order candidate documents within neural retrieval approaches [15].

Retrieval Augmented Generation (RAG) became a central approach in biomedical QA, replacing fine-tuned models with evidence-based retrieval methods [12, 17]. RAG-based systems surpassed fine-tuned approaches, particularly on summary-type questions that require passage-level consolidation [12]. Moreover, self-feedback loops boosted BioASQ scores, caused latency and instability on factoid questions [18]. Similarly, context length did not behave linearly with answer quality, where long inputs reduced factoid accuracy while improving summaries [16]. Nevertheless, most systems applied a fixed context budget uniformly across all types of QA [5]. Likewise, synthetic snippet enhancement helped in sparse retrieval, raising the risks of hallucinations that remained unresolved [17].

The adaptation of LLMs for biomedical QA produced variable and inconsistent results [19]. Domain-specific supervised fine-tuning generally outperformed zero-shot prompting in factoid and list questions [6], while the snippet-aware prompting approach fine-tuned performance without the high computational cost of gradient-based training [16]. Conversely, open-weight models in the 7B-13B parameter range remained reasonably strong on structured question types but trialed proprietary systems on summary generation [7]. In addition, parameter-efficient fine-tuning (PEFT) approaches such as LoRA, QLoRA provide a balance between accuracy and cost, but they lack full fine-tuning in biomedical domain adaptation tasks [19]. Therefore, no single adaptation strategy is universally optimal in all types of BioASQ questions [3, 6, 7, 19].

Ensemble-based approaches demonstrated consistent performance over single-model systems in biomedical QA, where combinations of models lead to more reliable results [8]. Similarly, multi-model voting outperformed individual generations when component models provide complementary retrieval biases [20]. Agent-specialized architectures also increased diversity, distributing tasks across different agents instead of relying on model variation [21]. Likewise, multi-granularity transformer representations achieved high performance across BioASQ QA types [22]. Nevertheless, most existing ensemble systems aggregate outputs through majority voting without explicitly handling disagreement between models [8]. Furthermore, the heterogeneity of BioASQ QA types remained unexplored across the literature. Some frameworks implemented type-specific ensemble methods aligned to evaluation metrics for yes/no, factoid, list, and summary QA [4]. Retrieval-based studies reported MAP and recall, making it difficult to understand how retrieval improvements influence answer quality [14]. Likewise, generation-based methods usually treat retrieved context as a fixed input, limiting meaningful comparison across systems [12].

Overall, these studies highlighted several gaps that directly motivate the proposed system. They lack dual cross-encoder reranking [15], self-feedback instability [16], ensemble methods incorporating stance diversity [8, 20], and evaluation methods that limit meaningful system comparison [12, 14]. Accordingly, the proposed pipeline integrates dual cross-encoder neural reranking, stance-diverse ensemble answer generation, and type-specific aggregation mechanisms within a unified and reproducible framework, directly addressing these gaps and offering a more robust approach for biomedical QA. Unlike dense retrieval systems such as MedCPT and hybrid BM25+dense retrieval pipelines that attained higher MAP through end-to-end fine-tuning, our system deliberately avoids task-specific fine-tuning to maintain reproducibility and cost-effectiveness.

3. Methodology

To meet the requirements of the BioASQ Task 14b challenge, we propose a modular biomedical QA framework that handles document retrieval, snippet extraction, and response generation. The system is organized into five different configurations across Phase A, A+, and B, as shown in Table 1. It combines state-of-the-art biomedical neural rerankers for efficient candidate selection with ensembles of LLMs for answer generation. Furthermore, the system employs few-shot prompting, multi-stance voting, and aggregation strategies specifically designed to accommodate the specific characteristics of different BioASQ question types.

Table 1
Overview of all submitted systems across BioASQ phases

System	Phase	Reranker	Generation Strategy
BioXR-BGE	A	BAAI/bge-reranker-v2-m3 [15]	Few-shot LLM
BioXR-GTE	A	Gte-reranker-modernbert-base [14]	Few-shot LLM ensemble
Gen-Doc	A+	BGE top-10 docs	Few-shot LLM ensemble
LLM Biomedical QA	B	Gold snippets (no reranking)	7-vote yes/no + 3-call factoid/list ensemble
Doc-gen	B	Gold snippets (no reranking)	Few-shot LLM

3.1. Corpus Preparation and Indexing

The retrieval corpus is built using PubMed 2026 baseline [1], released by the National Library of Medicine (NLM) in January 2026 and comprising over 38 million biomedical articles. Raw XML files are obtained from the National Center for Biotechnology Information (NCBI) and processed into a structured JSON format, with each record containing the article's PMID, title, abstract, keywords, and MeSH terms. These records are then indexed using the BM25 (Okapi) ranking function implemented via Pyserini [2]. BM25 parameters are tuned through grid search over the BioASQ 14b training set, maximizing MAP on a held-out validation split.

3.2. Phase A: Document Retrieval and Snippet Extraction

Phase A requires systems to retrieve up to 10 relevant articles and 10 supporting text snippets for each biomedical query. Our pipeline addresses this in two stages: sparse retrieval followed by neural reranking, with a dedicated snippet extraction module operating over the final ranked documents.

3.2.1. First-Stage Retrieval

For each test query, the top-1000 candidate documents are retrieved from the BM25 index as the initial high-recall for downstream reranking.

3.2.2. Neural Reranking

The top candidates are passed to neural rerankers for fine-grained relevance scoring. Two parallel systems are designed, including BioXR-BGE and BioXR-GTE, as discussed in Table 2.

3.2.3. Snippet Extraction

Snippet retrieval is performed over the top-10 reranked documents using a two-stage pipeline. Evidence is extracted using overlapping passage windows, where the BM25 (30%) and S-PubMedBERT-MS-

[1] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
[2] Pyserini: https://github.com/castorini/pyserini

Table 2
Neural reranking systems submitted for Phase A document retrieval.

System	Reranker Model	Architecture	Output
BioXR-BGE	BAAI/bge-reranker-v2-m3	M3-Embedding cross-encoder	Top-10 docs
BioXR-GTE	Alibaba-NLP/gte-reranker-modernbert-base	ModernBERT cross-encoder	Top-10 docs

MARCO dense embeddings (70%) ranks passages for high recall, followed by GTE ModernBERT cross-encoder reranking off-the-shelf to select the most relevant snippets. Figure 1 illustrates the complete Phase A pipeline, from raw PubMed corpus ingestion through BM25 retrieval, neural reranking, and final snippet extraction.

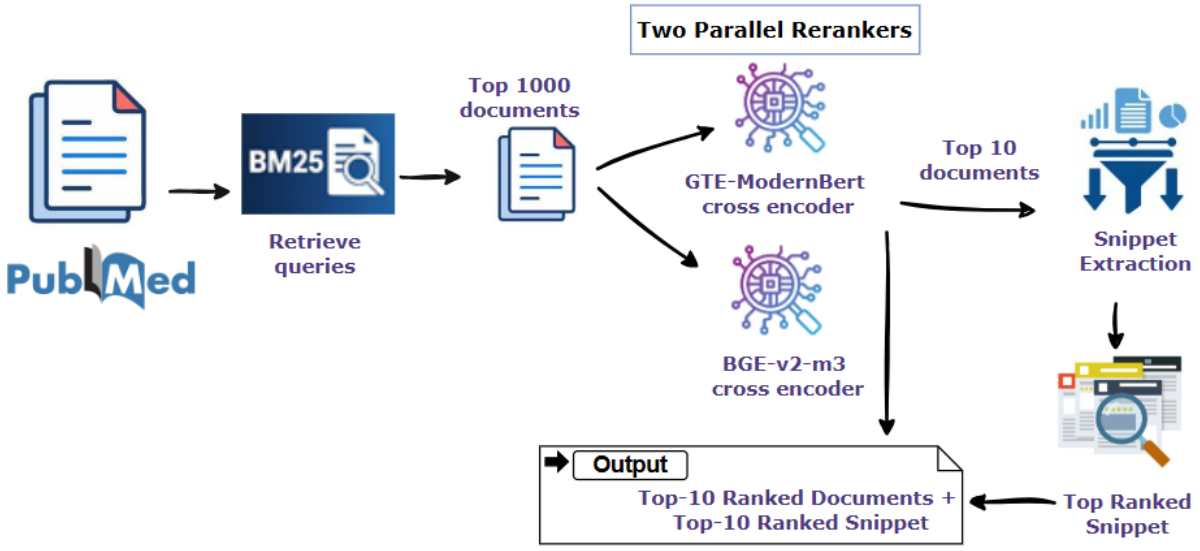

Figure 1: Phase A document retrieval and snippet extraction pipeline.

3.3. Phase A+: Answer Generation from Retrieved Evidence

Phase A+ extends the previous phase A, requiring systems to generate answers directly from extracted evidence. For each question, the system generates either an exact answer for yes/no, factoid, and list questions, or an ideal answer for summary questions.

3.3.1. Context Construction and Few-Shot Learning

For each question, the top retrieved snippets are numbered and concatenated into a single context block, with each snippet capped at 400 characters. A few-shot example pool is curated from BioASQ 14b training set (5,729 QA pairs) [3], and selected according to strict question type-specific quality criteria. For list questions, the answer requires at least three items, of which at least 50% appear verbatim in the snippets. For factoid questions, the answer is to appear explicitly within snippets and not exceed six words. For yes/no questions, examples are balanced between positive and negative scenarios, and summary examples must be at least 40 words long and have a fair amount of overlap with the snippet. To assist the model in comprehending the anticipated output format, each generation prompt contains up to three samples per question type.

[3] https://participants-area.bioasq.org/datasets/

3.3.2. System Configurations

Table 3 summarizes the three Phase A+ system configurations, highlighting differences in the retrieval source and generation strategy.

Table 3
Phase A+ system configurations.

System	Retrieval Source	Documents	Generation Strategy	Purpose
Gen-Doc	BGE-reranked	Top-10 + snippets	Full ensemble + few-shot	Primary submission
BioXR-GTE	GTE-reranked	Top-10 + snippets	Full ensemble + few-shot	Reranker comparison
BioXR-BGE	BGE-reranked	Top-10 + snippets	Single-call, no ensemble	Ablation baseline

3.3.3. Question-Type Specific Generation Strategies

Answer generation is performed through two API providers, Cerebras (LLaMA-3.3-70B, LLaMA-3.1-70B, LLaMA-3.1-8B) and OpenRouter (LLaMA-3.3-70B-Instruct, Mistral-7B-Instruct, Qwen-2.5-72B-Instruct), with automatic failover between providers and models. Each question type follows a dedicated prompt design and ensemble strategy, as summarized in Table 4. For yes/no questions, each vote is weighted

Table 4
Prompt design and ensemble strategy by question type across Phase A+ and Phase B systems.

Question Type	Strategy	Key Design	Ensemble / Retry Logic
Yes/No	7-Vote Weighted	Three system prompts: evidence-first, contrarian, meta-reasoning. Diverse snippet orderings (relevance, reverse, shuffled) per vote.	Relevance-weighted vote aggregation + adversarial consistency check on narrow majorities ($\leq$4/7).
Factoid	3-Call Ensemble	Three specialist prompts: *exact_extractor* (frequency-based), *context_reasoner* (type-aware), *frequency_ranker* (recall-maximizing). Temperature: 0.0 / 0.2 / 0.3.	MRR-weighted aggregation (rank-1=3pts, rank-2=2pts, rank-3=1pt). Near-duplicate merging. Last-resort retry.
List	3-Call Ensemble	Three specialist prompts: standard (full extraction), *forced_scan* (per-snippet exhaustive), *category_guided* (entity type filter). Temperature: 0.0 / 0.2 / 0.3.	Item included if present in $\geq$2 calls, or in *forced_scan* only. Top-10 returned.
Summary	Single Call + Retry	Verbatim-copy prompt: exact biomedical terms from snippets; 3–5 sentences; ROUGE-optimised phrasing. Few-shot injected.	Retry if word count $<$40 or snippet-term overlap $<$25%. Best candidate selected by overlap score.

by $w = 1 + \bar{s}$, where $\bar{s}$ is the mean relevance score of snippets seen by that vote, computed using TF-IDF-inspired keyword overlap with bigram bonuses, and when the majority is narrow (≤ 4 out of 7 votes), an adversarial check is triggered using the contrarian prompt on the top-5 most relevant snippets to challenge the majority answer. For factoid questions, answers are combined by giving higher-ranked contenders greater weight throughout calls. Near-duplicate entities are merged using substring containment: If entity A is a substring of entity B, their scores are summed, and the higher-scoring canonical form is retained. If an item appears in the exhaustive per-snippet scan call alone, or in at least two of the three calls, it is included in the final response for list questions.

3.4. Phase B: Answer Generation from Gold Evidence

Phase B provides systems with gold-standard annotator-selected documents and snippets. Table 5 summarizes the two submitted systems and their configurations.

In order to maintain annotator document ordering, the Phase B context is built solely from gold snippets provided in the testset. Unlike Phase A+, which relies on retrieved evidence, the Phase B system specifically mentions the gold-standard character of snippets. This pushes the models to extract responses directly from the given evidence rather than depending on its parametric knowledge. Figure 2 shows the complete generation pipeline of Phase A+ and Phase B.

Table 5
Phase B system configurations.

System	Input Evidence	Snippet Ordering	Generation Strategy	Prompt Adaptation
LLM Biomedical QA	Gold snippets	Annotator document priority + character offset.	Full ensemble + few-shot (same as Phase A+)	LLM informed evidence is expert-curated.
Doc-Gen	Gold snippets	Reranked by Sentence-BERT similarity.	Few-shot prompting	Same as LLM Biomedical QA.

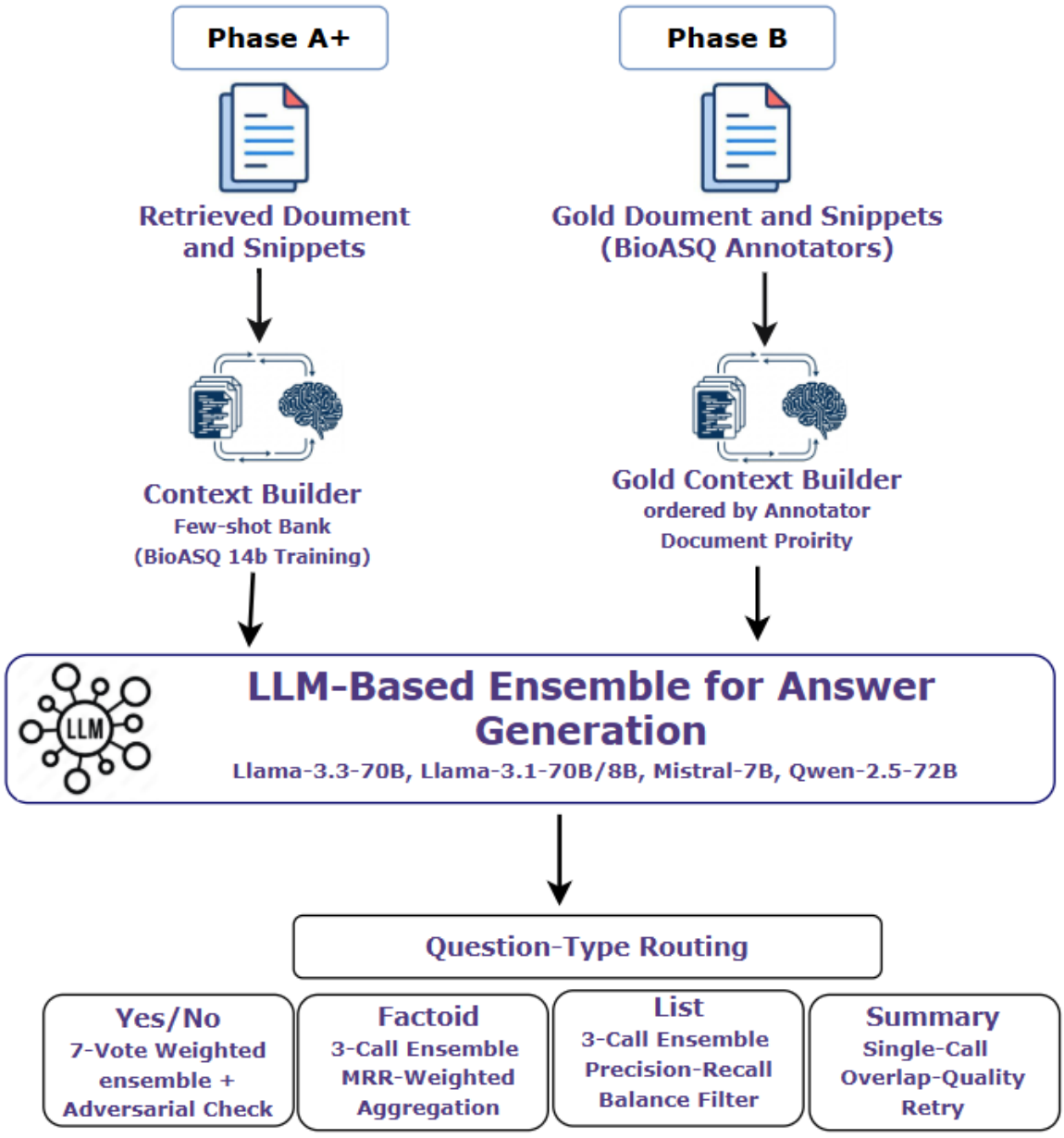

Figure 2: Phase A+ and B answer generation pipeline.

4. Results and Discussion

In this section, we report preliminary results across different batches of the BioASQ 14b challenge. Final rankings may change following updates to the gold-standard annotations and the completion of human evaluation. The reported results are based on the automatic evaluation metrics provided by the BioASQ organizers: MAP and F-measure for Phase A retrieval, and Macro F1, MRR, F-measure, ROUGE-2, and

ROUGE-SU4 for Phase A+ and B generation.

4.1. Phase A results

This section reports the results of Phase A across Batch 3 and 4, presenting document and snippet retrieval for both systems (BioXR-GTE and BioXR-BGE), as shown in Table 6. Both systems constantly outperformed the competition median, with BioXR-GTE achieving the stronger retrieval results in both batches. BioXR-GTE consistently performed better than BioXR-BGE across both evaluation batches,

Table 6
Phase A document retrieval and snippet retrieval performance. Bold values indicate our best results. Best = best competitor in the leaderboard.

Batch	System	Documents					Snippets
		Mean Prec	MAP	GMAP	Recall	F-Measure	MAP
3	BioXR-GTE	**0.0604**	**0.1674**	0.0010	0.2776	0.0948	—
	BioXR-BGE	0.0485	0.1423	0.0007	0.2413	0.0783	—
	Best (B3)	—	*0.3236*	—	*0.4322*	—	*0.4322*
4	BioXR-GTE	**0.0717**	**0.1949**	0.0015	0.3022	0.1067	**0.0004**
	Best (B4)	—	*0.2310*	—	—	—	*0.1634*

achieving a MAP improvement of approximately 0.02-0.05 points. This highlights that the ModernBERT-based reranker provides stronger biomedical relevance modeling compared to the BGE-M3 architecture. Although both systems performed above the competition median but remained below the leaderboard best across both batches. The improvement from Batch 3 to Batch 4 for both configurations further demonstrates the stability and robustness of the proposed retrieval framework across different test sets. Despite these improvements, the relatively low GMAP scores across both systems highlight the inherent difficulty of achieving strong early-precision ranking in large-scale biomedical Information Retrieval (IR). BioXR-GTE obtained a low snippet MAP score of 0.0004 in Batch 4, which remained lower than the leaderboard's best score of 0.1634. These findings demonstrate that snippet-level retrieval remains a weak component of our Phase A pipeline and a clear direction for future improvement.

4.2. Phase A+ Results

The Phase A+ evaluation demonstrates that generation approaches are particularly effective for binary Biomedical QA, while remaining reasonably competitive for factoid and list-type questions, as shown in table 7.

Table 7
Phase A+ results across Batches 3 and 4. ⋆ = matches best competitor.

Batch	System	Yes/No (Macro F1)		Factoid (MRR)		List (F-Measure)		Summary (R-SU4 F1)	
		Best	Ours	Best	Ours	Best	Ours	Best	Ours
3	Gen-Doc	1.0000	**1.0000**⋆	0.5294	**0.3824**	0.3258	**0.2785**	0.1537	**0.0318**
	BioXR-GTE	1.0000	**1.0000**⋆	0.5294	0.3235	0.3258	0.2773	0.1537	**0.0318**
	BioXR-BGE	1.0000	0.8952	0.5294	0.3529	0.3258	0.2722	0.1537	0.0243
4	Gen-Doc	0.9352	0.7922	0.4091	—	0.4840	**0.3123**	0.1699	0.0305

The best performance is observed for yes/no questions, where both Gen-Doc and BioXR-GTE achieved a perfect Macro F1 score of 1.0000 in Batch 3. This demonstrates that the seven-vote weighted ensemble with diverse reasoning stances and relevance-weighted aggregation accurately resolves binary biomedical questions even from imperfectly retrieved evidence.

In the factoid questions, Gen-Doc obtained the highest MRR among our systems in Batch 3, reasonably close to the top-performing leaderboard systems. The three-pass ensemble approach with task-specific

prompts effectively reduced single-call variance, especially for precise biomedical entities such as gene names and drug identifiers. By using a single call strategy, BioXR-BG performed slightly below the ensemble systems, demonstrating the effectiveness of multi-call aggregation.

The performance of the list questions remained relatively consistent across all three systems, but below the competition best, which is expected since the list-type questions depend heavily on comprehensive evidence coverage, a condition not always achievable from the retrieved snippets. Similarly, ideal answer ROUGE-SU4 F1 scores were modest across all submissions, reflecting the inherent complexity of open-ended summarization from imperfect retrieved evidence. In Batch 4, overall performance dropped slightly, likely due to increased question complexity, with Gen-Doc remaining the strongest submission.

The lower factoid performance of BioXR-GTE's ensemble (MRR=0.3235) relative to BioXR-BGE's single call (MRR=0.3529) is due to two interacting factors. First, the comparison between Gen-Doc and BioXR-GTE, differing only in reranker selection, identifies a 0.0589 MRR loss demonstrating the critical role of reranking. Although BioXR-GTE attains higher document-level MAP, this metric reflects overall topical relevance rather than the answer-bearing snippets ranking. Factoid question answering instead requires precise entity-level ranking, where BAAI/bge-reranker-v2-m3 appears more effective than GTE-ModernBERT at prompting snippets containing the correct answer. Second, when retrieval includes partially relevant evidence, the ensemble prompts are more likely to extract diverse probable entities. MRR-weighted aggregation distributes confidence among them rather than converging on the correct answer. In contrast, BioXR-BGE's single-call strategy avoids this effect by selecting a single answer from the highest-ranked evidence. The comparison between Gen-Doc and BioXR-BGE further shows that the ensemble yields a 0.0295 MRR gain when retrieval quality is high, suggesting that ensemble answer generation improves performance only when supported by accurate evidence retrieval.

4.3. Phase B Results

The Phase B evaluation further reports the importance of evidence quality for effective answer generation. Systems provided with manually expert-annotated gold snippets significantly achieved improved performance compared to the corresponding Phase A+ systems, especially for factoid and list questions. Table 8 reports the Phase B results for all submitted systems in Batches 2, 3, and 4.

Table 8
Phase B results. ⋆ = best across our submissions.

Batch	System	Yes/No (Macro F1)		Factoid (MRR)		List (F-Measure)		Ideal (R-SU4 (F1))	
		Best	**Ours**	**Best**	**Ours**	**Best**	**Ours**	**Best**	**Ours**
2	LLM Biomedical QA	1.0000	0.8329	0.4000	0.1500	0.4720	0.0735	0.2310	0.0939
3	LLM Biomedical QA	1.0000	**0.8952**	0.5882	**0.4118**	0.5175	**0.3836**	0.2547	**0.0530**
	Gen-Doc	1.0000	**0.8952**	0.5882	**0.4118**	0.5175	**0.3865**	0.2547	0.0508
4	LLM Biomedical QA	1.0000	**0.9352**⋆	0.5758	0.2121	0.6543	**0.5573**	0.2699	0.0560

The most notable results among all evaluations is the yes/no Macro F1 of **0.9352** obtained by LLM Biomedical QA in Batch 4, representing the highest yes/no score across all our submissions and close to the top competition results. Both systems achieved 0.8952 in Batch 3, consistently surpassing the competition median and demonstrating that the seven-vote ensemble remains robust when supported by high-quality gold evidence.

In factoid questions, both systems achieved an MRR of **0.4118** in Batch 3, improving substantially over the phase A+ results and competitive with top-performing leaderboard systems. The improvements directly reflects the benefit importance of gold snippets, which often explicitly contain the required answer entities. Factoid scores decreased in Batch 4, likely because of higher question complexity, while Batch 2 produced the lowest scores results due to initial system configuration before later-stage optimization.

Performance on list questions also demonstrates a notable gain compared to Phase A+. LLM Biomedical QA achieves a list F-measure of **0.5573** in Batch 4, the strongest list-question performance among our

submissions, compared to 0.27 in Phase A+. These results confirm that list question performance relies heavily on evidence recall, as gold snippets provide comprehensive entity enumeration that retrieved snippets cannot reproduce. ROUGE-SU4 F1 scores for the ideal answer remained modest throughout all batches, emphasizing the challenge of biomedical summarization even when gold-standard evidence is available. The verbatim-copy method, the prompting strategy, and the overlap-quality retry process collectively contributed to stable consistent grounding across all submitted systems.

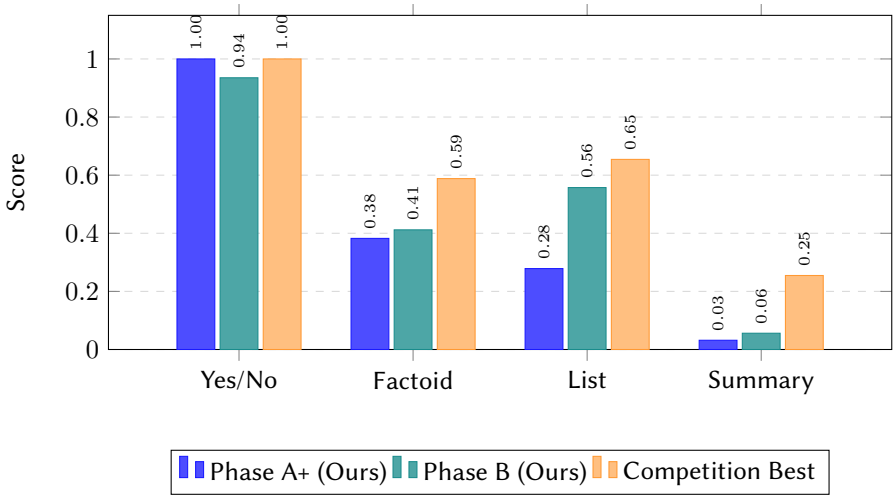

Figure 3: Comparison of our best system performance across Phase A+ and Phase B against the competition.

Figure 3 illustrates performance across Phase A+ and Phase B against the competition. Yes/No scores are strong in both phases, approaching the competition best. List questions show the largest improvement from Phase A+ to Phase B (0.28 to 0.56), confirming that evidence quality is the primary bottleneck for entity enumeration. Factoid and summary scores remain below the competition best across both phases.

4.4. Cross-Phase Analysis

A comparison between Gen-Doc (full ensemble) and BioXR-BGE (single-call) in Batch 3 provides partial ablation evidence regarding the contribution of ensemble answer generation. Across the evaluation, the multi-call generation strategy outperformed single-call generation approaches. Overall, the results indicate that the answer generation strategy is more influential than retrieval quality for most question types, as ensemble-based answer generation reduces the performance gap caused by weaker upstream retrieval. However, the list questions are highly dependent on evidence completeness, indicating retrieval improvement as the highest priority direction for future work. Table 9 summarizes the key findings across all phases.

4.5. Error Analysis and Limitations

Ideal answer generation remains the most challenging task, with R-SU4 F1 scores (<0.06) consistently below across all systems and evaluation batches. The primary factor is low snippet retrieval performance, with Batch 4 achieving a snippet MAP of only 0.0004 compared with 0.1634 for the leading system. This reflects the difficulty of accurately retrieving sentence-level evidence from more than 39 million PubMed abstracts without dedicated snippet extraction training. Consequently, incomplete or suboptimal evidence constrains the quality of the generated summaries. A second factor is the R-SU4 metric that evaluates similarity to expert-written summaries using skip-bigram overlap, favoring concise, human-authored synthesis over evidence-grounded reproduction. While the proposed verbatim-grounding strategy enforces a 25% snippet-term overlap threshold, it improves factual grounding but cannot reproduce concise, synthesised language characteristic of expert-written summaries. This evaluation

Table 9
Cross-phase key findings.

Finding	Observation
Reranker vs Generation	In Phase A MAP, BioXR-GTE looks better than BioXR-BGE, by something like 0.02 to 0.05 points, but later in Phase A+ generation, that distance starts to shrink a bit, suggesting the ensemble partially compensates for retrieval quality differences.
Yes/No Ensemble	The seven-vote ensemble attains Macro F1 of 1.0000 in Batch 3 (Phase A+) and 0.9352 in Batch 4 (Phase B), showing robustness across both retrieved and gold evidence settings.
List Evidence Bottleneck	A 15–28 point F-measure gap between Phase A+ and Phase B list results confirms that retrieved evidence cannot reliably enumerate all the targets entities, motivating future work on dense retrieval and pseudo-relevance feedback.
ROUGE Limitation	Moderate ROUGE-SU4 scores reflect the known mismatch between automatic metrics and human judgment in biomedical QA. Ideal answer Results should be interpreted as indicative pending human evaluation.

mismatch has also been observed in BioASQ 2025 participant systems [8, 5]. Therefore, the bottleneck for ideal answer generation lies in snippet retrieval coverage and precision rather than the generation strategy.

5. Conclusion

This study introduced a modular RAG-based framework for biomedical QA, evaluated on the BioASQ Task 14b benchmark. The proposed system targets two limitations commonly observed in existing systems: single-stage reranking and consensus-only answer generation. Using two separate cross-encoder rerankers, GTE-ModernBERT and BGE-M3, on BM25 top-1000 candidates revealed that reranker architecture strongly affects retrieval effectiveness, with GTE-ModernBERT consistently better across both batches. Moreover, the two-stage hybrid snippet extraction approach improved downstream evidence quality. The central contribution of this work is the stance-diverse ensemble generation framework. For yes/no questions, using seven weighted votes from evidence-first, contrarian, and meta-reasoning positions in place of simple majority voting, together with adversarial consistency testing for narrow outcomes, proved to be quite effective. In Phase A+, this resulted in a perfect Macro F1. Question-specific approaches for factoid and list questions also showed that tailored generation methods perform better than generic single-call strategies. In particular, the most significant improvements were observed for Phase B list questions, where access to gold evidence increased the F-measure from 0.28 to 0.56. This finding suggests that recall of evidence remains the main limitation in biomedical entity enumeration tasks. Although the system performs competitively across most question types, summary-type answer generation was still a weak area, with ROUGE-SU4 scores remaining below the competition's best. This indicates the ongoing misalignment between automatic lexical metrics and true answer quality, a challenge widely discussed in previous BioASQ editions. Human evaluation could therefore produce different rankings from those reported using automatic metrics. Results are reported on two test batches without statistical significance testing; So the observed improvements and gaps may come from differences in the question sets rather than true generalized gains. Future work will prioritize improvements to both retrieval and generation strategies. A hybrid sparse-dense retrieval framework will be investigated to improve both document ranking and snippet convergence, addressing the primary bottleneck observed in summary question answering. In addition, fine-tuning the snippet extraction module on BioASQ- specific annotations is expected to improve the precision. In parallel, integrating higher-capacity LLMs into the ensemble framework may improve generation consistency and overall answer quality.

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT for language refinement to improve the clarity and readability and used Gemini to generate a Context Builder icon incorporated into the methodology diagram. After using these tools, the authors carefully reviewed, revised, and edited the content and take full responsibility for the accuracy, originality, and integrity of the final version of the manuscript.

References

[1] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, Bioasq-qa: A manually curated corpus for biomedical question answering, Scientific data 10 (2023).

[2] National Center for Biotechnology Information (NCBI), PubMed annual baseline repository, U.S. National Library of Medicine, 2026. URL: https://ftp.ncbi.nlm.nih.gov/pubmed/baseline, accessed: 2026.

[3] A. Hornback, H. Sathu, K. Kim, Y. Wang, Y. Zhu, M. Isgut, P. Avula, A. Khimani, M. D. Wang, Large language models in healthcare and biomedical informatics: A comprehensive review, Innovation and Emerging Technologies 13 (2026).

[4] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, et al., Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 173–198.

[5] D. Galat, D. Molla-Aliod, Llm ensemble for rag: Role of context length in zero-shot question answering for bioasq challenge, arXiv preprint arXiv:2509.08596 (2025).

[6] J. Tang, H. Yang, K. Xiong, H. Li, P. Quaresma, H. Yu, W. Zhang, M. Song, Y. Jiang, Applying deepseek to bioasq task 13b: Using supervised fine-tuning and few-shot learning, in: CLEF, 2025.

[7] D. Stachura, J. Konieczna, A. Nowak, Are smaller open-weight llms closing the gap to proprietary models for biomedical question answering?, arXiv preprint arXiv:2509.18843 (2025).

[8] D. Panou, A. C. Dimopoulos, M. Koubarakis, M. Reczko, Harnessing collective intelligence of llms for robust biomedical qa: a multi-model approach, arXiv preprint arXiv:2508.01480 (2025).

[9] A. Nentidis, et al., Overview of BioASQ 2026: The fourteenth BioASQ Challenge, Lecture Notes in Computer Science, Springer, 2026.

[10] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 14b and Synergy14 in CLEF2026, in: CLEF 2026 Working Notes, 2026.

[11] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of bioasq tasks 13b and synergy13 in clef2025, in: Proceedings of CLEF 2025, 2025.

[12] H. Yang, S. Li, T. Gonçalves, Enhancing biomedical question answering with large language models, Information 15 (2024). URL: https://www.mdpi.com/2078-2489/15/8/494. doi:10.3390/info15080494.

[13] H. Bolat, B. Şen, Document retrieval system for biomedical question answering, Applied Sciences 14 (2024).

[14] J.-C. Hana, B.-C. Chiha, H.-C. Hungb, R. T.-H. Tsaia, Ncu-iisr: A retrieval-augmented generation approach for bioasq 13b phase a and a (2025).

[15] R. A. Jonker, T. Almeida, J. Almeida, S. Matos, Bit. ua at bioasq 13b: Revisiting evaluation, dprf-enhanced retrieval and fine-tuned llms, in: CLEF, 2025.

[16] H. Kim, H. Lee, Y. Cho, J. Park, J. Park, S. Park, Y. T. Chok, S. Baek, D. Lee, J. Kang, Prompting matters: snippet-aware strategies for biomedical qa with llms in bioasq 13b, in: CLEF, 2025.

[17] F. Borazio, D. Croce, R. Basili, Unitor at bioasq 2025: modular biomedical qa with synthetic snippets and multiple task answer generation, in: CLEF, 2025.

[18] S. Ateia, U. Kruschwitz, Can language models critique themselves? investigating self-feedback for retrieval augmented generation at bioasq 2025, arXiv preprint arXiv:2508.05366 (2025).
[19] J. Zhu, J. Li, S. Zhao, Y. Deng, Y. Miao, J. Xu, Adapting llms for biomedical natural language processing: a comprehensive benchmark study on fine-tuning methods, The Journal of Supercomputing 82 (2026).
[20] J. Angulo, V. Yeste, Aqams and aqams2: Multi agent systems for biomedical question answering, in: CLEF, 2025.
[21] R. Abdel-Salam, M. Adewunmi, M. A. Abayomi, Caresai at biocreative ix track 1–llm for biomedical qa, arXiv preprint arXiv:2509.00806 (2025).
[22] C. Qian, X. Shi, S. Yao, Y. Liu, F. Zhou, Z. Zhang, J. Akram, A. Braytee, A. Anaissi, Optimized biomedical question-answering services with llm and multi-bert integration, in: 2024 IEEE International Conference on Data Mining Workshops (ICDMW), 2024. doi:10.1109/ICDMW65004.2024.00028.