UniTor at BioASQ 2025: Modular Biomedical QA with Synthetic Snippets and Multiple Task Answer Generation

Federico Borazio[1,*], Andriy Shcherbakov[2], Danilo Croce[1,*] and Roberto Basili[1]

[1]University of Rome Tor Vergata, Via del Politecnico 1, Rome, 00133, Italy
[2]Reveal srl, Via Kenia 21, Rome, 00144, Italy

Abstract

Recent advances in large language models (LLMs) have enabled notable progress in open-domain question answering, yet their application to specialized biomedical tasks remains limited by challenges in factual reliability, domain coverage, and evidence traceability. Retrieval-augmented generation (RAG) approaches have shown promise in addressing these issues by grounding model outputs in external documents. In this work, we introduce UniTor@BioASQ, a retrieval-augmented pipeline for biomedical question answering, developed for the BioASQ 2025 Challenge (Task 13b). The system combines multi-stage document retrieval with LLM-guided snippet extraction and unified answer generation across diverse biomedical question types. A key feature of our approach is the use of synthetic LLM-generated snippets as semantic anchors for document reranking, a strategy not previously explored at scale in this context. We systematically evaluate the impact of this component, as well as pseudo-relevance feedback, fine-tuned snippet extraction, and multi-task answer generation, through ablation studies on official BioASQ 2025 test batches. Results show that UniTor@BioASQ achieves robust performance: our system consistently ranks among the top five for both ideal and factoid answer generation, and among the top ten for snippet extraction, despite using compact, open-source models. Analysis highlights the contribution of snippet-based semantic reranking to retrieval effectiveness, the resilience of fine-tuned snippet extractors under noisy evidence, and the viability of a unified answer generator for heterogeneous biomedical QA tasks.

Keywords

Biomedical Question Answering, Retrieval-Augmented Generation, Multi-Stage Retrieval, Large Language Models, Synthetic Snippet Generation, Relevant Text Highlighting, Contextual Answer Generation

1. Introduction

Large Language Models (LLMs) have become foundational in modern natural language processing, demonstrating impressive generalization capabilities across a wide range of tasks, including text generation, summarization, and question answering [1, 2, 3]. Their success, however, often hinges on access to large, diverse datasets and a strong contextual understanding. In specialized domains such as biomedicine, these models face additional challenges, including the need for factual accuracy, domain-specific knowledge, and interpretability. In high-stakes applications like biomedical question answering, the risk of generating incorrect or misleading information, so-called "*hallucinations*", is a critical concern, raising the need for robust and trustworthy solutions [4].

Retrieval-Augmented Generation (RAG) frameworks, since [5], have emerged as a promising approach to mitigate these limitations by integrating document retrieval mechanisms with LLMs, allowing the model to ground its generation in external, factual sources. In the biomedical domain, RAG systems have shown considerable promise: for instance, BiomedRAG [6] demonstrated how incorporating retrieved documents into the model context can enhance factual grounding, while MedBioLM [7] highlighted the benefits of combining domain-specific fine-tuning with retrieval for improving biomedical QA. These systems typically follow a multi-stage pipeline, comprising document retrieval from large biomedical corpora (e.g., PubMed), document reranking based on relevance, snippet extraction for contextual relevance, and final answer generation conditioned on the selected evidence. However, each of these stages

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
borazio@ing.uniroma2.it (F. Borazio); shcherbakov@revealsrl.it (A. Shcherbakov); croce@info.unirma2.it (D. Croce); basili@info.unirma2.it (R. Basili)
0009-0000-0193-2131 (F. Borazio); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)

introduces its challenges: suboptimal retrieval can lead to irrelevant or noisy documents, aggressive filtering may discard useful information, and improper conditioning can result in poor answer quality.

In this paper, we present **UniTor@BioASQ** [1], our system developed for the BioASQ [8] 2025 Challenge (Task 13b) [9]. UniTor@BioASQ implements a fully modular RAG pipeline for biomedical question answering. BioASQ is an annual challenge on large-scale semantic indexing and question answering in the biomedical domain[2], organized as part of the Conference and Labs of the Evaluation Forum (CLEF)[3]. The 2025 edition (13th BioASQ Challenge) includes several tasks, with Task 13b specifically focusing on biomedical semantic question answering. Task 13b is structured into three distinct phases:

- **Phase A**: Systems are provided with English biomedical questions and are required to retrieve up to 10 relevant articles from the PubMed Annual Baseline Repository, along with up to 10 relevant text snippets extracted from these articles. The goal is to identify and extract evidence that may help answer the questions. In the intermediate phase called **Phase A+**, participants are required to generate answers directly by leveraging the retrieved evidence, articles, and snippets, as contextual support. Specifically, they must produce either exact answers (yes/no, factoid, or list) or ideal answers (short paragraph-style summaries) for each question, using the evidence gathered in Phase A. The questions cover different biomedical information needs and are categorized into four types:
  - Yes/no questions: require a binary answer (yes/no), such as "*Do CpG islands colocalise with transcription start sites?*"
  - Factoid questions: expect a short answer (up to 5 entities), such as "*Which virus is best known as the cause of infectious mononucleosis?*"
  - List questions: expect a list of entities (up to 100 entries), such as "*Which are the Raf kinase inhibitors?*"
  - Summary questions: require a short paragraph summarizing the most relevant information, such as "*What is the treatment of infectious mononucleosis?*"
- **Phase B**: In this final phase, systems are provided with the same set of questions as in Phase A, along with the corresponding gold standard articles and snippets selected by biomedical experts. Using this curated evidence, systems must generate exact answers (for yes/no, factoid, and list questions) and/or ideal answers (paragraph-sized summaries).

This multi-phase structure enables the evaluation of systems on various aspects of the biomedical QA pipeline: information retrieval (Phase A), context-aware answer generation (Phase A+), and answer generation using gold-standard evidence (Phase B). Task 13b was scheduled in 4 batches with two weeks in between and ran from March 26 to May 08.

In this context, UniTor@BioASQ is a modular system designed to tackle all the main BioASQ 13b tasks. UniTor@BioASQ integrates:

- **Document Retrieval**: A multi-stage retrieval workflow combining traditional full-text search (BM25/Solr), dense embedding-based similarity using pre-trained sentence encoders, and a supervised re-ranking model based on transformers, to select the most relevant PubMed abstracts for each question.
- **Snippet Extraction**: A task-adapted LLM, fine-tuned to identify and highlight the most relevant text spans within retrieved documents, given a biomedical question.
- **Answer Generation**: A dedicated LLM, trained to synthesize final answers of different types (yes/no, factoid, list, and ideal/summary) using both the selected documents and extracted snippets as contextual input.

[1] https://github.com/crux82/BioASQ2025-UNITOR
[2] https://www.bioasq.org/
[3] https://clef2025.clef-initiative.eu/

While these components reflect strategies widely used in RAG systems for BioASQ [10, 11], our work differs in that each module is explicitly designed, trained, and evaluated in isolation, allowing for a systematic investigation of its impact on overall system performance. This experimental setup enables us to address the following research questions:

Q1. **Document Reranking Enhancement**: Does generating plausible candidate answer snippets as an intermediate step, using a generative LLM, provide additional semantic guidance that improves unsupervised document reranking?
Q2. **Query Expansion**: How effective are pseudo-relevance feedback strategies that expand the original question with actual evidence extracted by the snippet module, in terms of retrieving additional relevant documents?
Q3. **Robust Snippet Extraction**: Can fine-tuning the snippet extraction LLM with a balanced set of positive, negative, and borderline training examples improve the robustness and precision of relevant evidence identification?
Q4. **Aggressive Filtering**: What are the risks of using the snippet extraction LLM as a hard filter to discard irrelevant documents, in terms of loss of recall and overall system performance?
Q5. **Multi-Task Answer Generation**: Does training the answer generation LLM in a multi-task fashion, leveraging both abstracts and highlighted snippets as input context, improve answer quality for different biomedical question types?

By evaluating UniTor@BioASQ on all official batches of BioASQ 2025, we provide a detailed empirical analysis of each pipeline component. Our results show that generating plausible candidate snippets for document reranking offers consistent, though modest, improvements in retrieval effectiveness. Query expansion via pseudo-relevance feedback yields mixed results, improving recall in some batches but not consistently. Fine-tuning the snippet extraction model with balanced positive and negative examples substantially enhances precision (with our systems placing among the top 10 systems in several batches out of around 50 systems), although aggressive snippet-based filtering can reduce overall recall. Finally, multi-task answer generation leveraging both abstracts and extracted snippets yields the most significant improvements for factoid and ideal answer types, with UniTor@BioASQ not only achieving consistently above-median performance but also here placing among the top 10 and top 5 systems in several batches and categories, for example, ranking 1th out of 73 for factoid questions and 1th out of 58 for ideal answers. These findings offer practical insights for designing robust and modular RAG pipelines in biomedical question answering.

In the remainder of the paper, we first discuss related works (Section 2), highlighting prior approaches to biomedical question answering and RAG pipelines. Section 3 details the methodology we employed to build a modular biomedical QA system grounded on PubMed, outlining each pipeline component and its integration. Section 4 presents the challenge official evaluation framework, including the metrics used and a comprehensive analysis of results from the preliminary BioASQ Task 13b 2025 automatic evaluation. Finally, Section 5 concludes the paper by summarizing our findings and outlining directions for future research.

2. Related Works

Large Language Models (LLMs) have demonstrated remarkable versatility across a wide range of natural language processing tasks, from text generation to information extraction and reasoning [12, 13, 14, 15]. In recent years, Retrieval-Augmented Generation (RAG) has emerged as a key approach for biomedical question answering (QA) [16]. By integrating external retrieval mechanisms into Large Language Models, RAG aims to address limitations of closed-book LLMs, such as hallucination and outdated knowledge, by providing retrieved documents as context for generation. For example, [17] introduced the MIRAGE benchmark, covering 7,663 questions from five biomedical QA datasets, including BioASQ, PubMedQA, and MedQA, and demonstrated that RAG pipelines can improve accuracy by up to 18% over

standalone LLMs like GPT-3.5 and Mixtral. MIRAGE highlights that combining multiple biomedical corpora and diverse retrievers yields better performance than using single-source retrieval alone.

Beyond traditional retrieval, recent work has also explored automatic topic discovery and structured prompt generation to enhance interpretability and classification performance. For example, [X] propose a linguistically guided pipeline that expands user-provided seed terms via Word2Vec to generate semantically rich Subject-Verb-Object (SVO) triples, which are clustered and transformed into natural language assertions using LLMs. This approach allows dynamic prompt creation for zero-shot classification tasks, enabling analysts to uncover and define latent topics from large document collections.

Many recent works have focused on BioASQ, the primary benchmark for biomedical QA. [16] proposed BioRAGent, an interactive RAG system built on both proprietary and open-source LLMs, which leverages LLMs for query rewriting, snippet extraction, and answer generation while providing transparent links to PubMed documents. Similarly, [18] describes their participation in BioASQ Task 12b, using a Llama-2-7B model adapted with LoRA within a hierarchical RAG pipeline. Their system builds BM25 indexes over PubMed Central, employs an ensemble retriever (combining sparse and dense retrievers), and feeds the top-ranked paragraphs to the LLM, achieving significant gains in both phase A+ and final answer generation. These studies demonstrate that efficient parameter fine-tuning (PEFT) of LLMs, combined with hybrid RAG architectures, substantially boosts biomedical QA performance.

Beyond BioASQ, PubMedQA (yes/no questions based on PubMed abstracts) and MedQA (multiple-choice questions from US medical licensing exams) have been used to evaluate RAG systems. [17] include PubMedQA and BioASQ-Y/N in their benchmark, testing ability of RAG to retrieve relevant context even in the absence of in-context examples. [19] specifically analyzes the impact of the "*lost-in-the-middle*" problem, where relevant information is not placed early enough in the context, on RAG systems over PubMedQA and BioASQ-Y/N. MIRAGE also incorporates MedQA-US and MedMCQA (additional medical exam datasets) to assess RAG performance on multiple-choice questions in clinical contexts [17]. Overall, these works confirm that RAG methods provide clear improvements over closed-book LLMs, though careful retrieval design remains critical.

Regarding retrieval strategies, biomedical QA systems employ a variety of approaches, from sparse lexical retrievers (e.g., BM25) to dense embedding models. Many systems use BM25 as a fast, effective lexical baseline [17, 18]. Popular dense retrievers include Contriever (contrastive pretraining over general corpora) and SPECTER (scientific document embedding model), both validated in general-domain settings [17]. For biomedical tasks, models like MedCPT [20], trained on PubMed search logs, demonstrate strong performance in encoding clinical queries and documents. Some approaches integrate multiple resources to maximize information coverage: [17], for instance, uses Reciprocal Rank Fusion (RRF) to combine BM25 and MedCPT results, yielding optimal retrieval performance. Hybrid retrieval strategies, retrieving with BM25 for efficiency and reranking with BERT-based models or domain-specific dense retrievers, are also gaining traction, balancing retrieval accuracy and latency [21].

Collectively, these studies confirm that RAG pipelines leveraging LLMs, whether general-purpose or domain-specific, outperform closed-book LLMs on biomedical QA tasks. Nonetheless, challenges remain in optimal retriever selection and combination, scalability to large biomedical corpora (e.g., full PubMed), and handling "*context-wrangling*" issues like lost-in-the-middle effects. The literature suggests that carefully designed RAG pipelines, integrating LLM contextualization with targeted retrieval from biomedical repositories, are key to achieving reliable performance in the medical domain.

3. Methodology

In this section, we provide a comprehensive description of the datasets, resources, and methodological choices underlying our participation in the BioASQ 2025 Challenge (Task 13b). We first present the corpora and preprocessing steps employed, then detail each module of the UniTor@BioASQ system. Each component is designed to address one or more of the research questions outlined beforehand, enabling a systematic investigation of how different retrieval, filtering, and generation strategies impact

overall biomedical question answering performance.

3.1. Data and Resources

A robust biomedical QA system like UniTor@BioASQ critically depends on the availability of both high-quality domain knowledge and reliable supervised signals for system development and evaluation. Our methodology leverages two primary resources, each with a distinct and complementary role in the overall architecture.

PubMed/MEDLINE as the Evidence Source. The backbone of our system is the PubMed/MEDLINE corpus, which serves as the universal source of biomedical knowledge and evidence. For the 2025 BioASQ Challenge, we utilize the official PubMed baseline[4], a collection of approximately 38 million records, containing titles and abstracts from the biomedical literature. These records are the only information accessible at retrieval time, reflecting the real-world constraints and evaluation settings of BioASQ. Efficient large-scale retrieval is a fundamental challenge in this context. To enable fast and flexible access to the entire PubMed collection, we indexed all titles and abstracts using Apache Solr[5], a scalable and widely adopted search engine. Each document entry in the index is enriched with metadata and precomputed dense embeddings (Sentence-BERT[6] [22], PubMedBERT[7]), which play a dual role: enabling both traditional sparse (lexical) retrieval and dense (semantic) reranking. This hybrid design allows our system to balance the high recall of keyword-based methods with the semantic sensitivity of modern language models, ensuring that relevant documents are surfaced even in the presence of vocabulary mismatch or nuanced biomedical phrasing.

BioASQ Training Data as Supervised Signal. While PubMed provides the raw knowledge base, effective system development also requires curated supervision: high-quality question–answer pairs and expert judgments on what constitutes relevant evidence. For this, we rely on the BioASQ training dataset [23], which includes $5,389$ manually constructed biomedical questions from the twelve previous editions of the challenge. Each question is categorized (yes/no, factoid, list, or summary) and annotated with gold-standard relevant documents, evidence snippets, exact answers, and ideal answers. This dataset not only reflects the diversity and complexity of biomedical information needs but also provides ground-truth labels for training and evaluating each system module.

Aligning Data and Pipeline Objectives. The interplay between these two resources shapes every stage of our pipeline. The PubMed corpus is exclusively used for retrieval and evidence extraction, while the BioASQ training set is used to supervise and fine-tune our LLM-based models for tasks such as document reranking, snippet extraction, and answer generation.

To fully exploit BioASQ annotations, we construct a series of task-specific datasets, each aligned to a core subproblem:

- For document retrieval and reranking, we generate positive pairs from the gold-standard (question, relevant abstract) and hard negatives from top-ranked, but non-relevant, PubMed documents. This exposes models to realistic, challenging retrieval scenarios.
- For snippet extraction, we reformat abstracts to highlight gold evidence spans and supplement training with negative/borderline examples, teaching models both to extract relevant text and abstain when no answer is present.
- For answer generation, we aggregate all available context, questions, evidence snippets, abstracts, and train models to produce type-specific answers in the required BioASQ format.

[4] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
[5] https://solr.apache.org/docs/9_8_1/
[6] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
[7] https://huggingface.co/NeuML/pubmedbert-base-embeddings

Each derived dataset is split 80/10/10 (train/validation/test), ensuring rigorous model selection and avoiding overfitting. By combining the scale and breadth of PubMed with the precision and supervision of BioASQ, UniTor@BioASQ is able to support both robust retrieval and reliable, context-aware answer generation. This dual-resource approach is critical for meeting the high standards of factual accuracy, recall, and explainability required in biomedical QA, as well as for systematically evaluating the impact of each pipeline module in the downstream tasks.

3.2. UniTor: System Overview

A robust biomedical question answering (QA) system must seamlessly integrate multiple specialized modules into a coherent Retrieval-Augmented Generation (RAG) pipeline. The overarching objective is to guide large language models (LLMs) to produce answers that are not only fluent and contextually appropriate but also firmly grounded in factual evidence, thus minimizing the risk of hallucinations, a particularly pressing concern in biomedical applications.

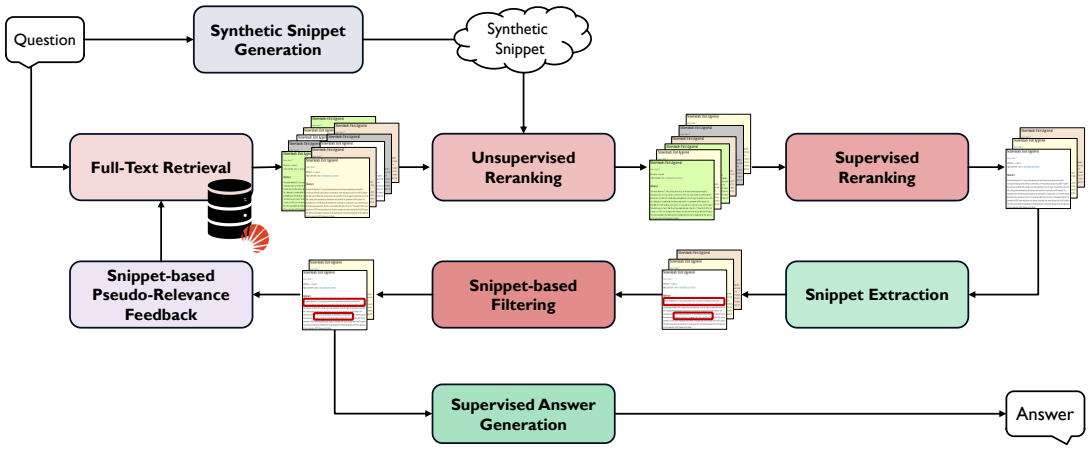

Figure 1: Overview of the UniTor@BioASQ pipeline for biomedical question answering. The system begins with the input question and follows a sequence of retrieval, reranking, snippet extraction, and answer generation stages. A synthetic snippet, generated via an LLM, is used for semantic guidance during unsupervised reranking, while extracted real snippets enable pseudo-relevance feedback for query expansion. Filtering, supervised reranking, and targeted snippet extraction refine the evidence passed to the final supervised answer generator. Feedback loops (e.g., snippet-based pseudo-relevance feedback) are clearly highlighted, illustrating the iterative and evidence-driven nature of the pipeline.
Note: For visual clarity, arrows from the "Question" to all pipeline modules are omitted; in practice, the question is provided as input to every component of the pipeline, not just those explicitly shown.

Figure 1 provides a high-level overview of UniTor@BioASQ modular pipeline, which is organized around three guiding principles:

- **Effective retrieval** of potentially relevant information from a vast biomedical corpus.
- **Precise identification** and filtering of the most pertinent evidence within retrieved documents.
- **Resilient answer generation** capable of coping with noisy or incomplete context.

To illustrate the workflow, consider the question: *"What are the common side effects of the drug Imatinib?"* The pipeline proceeds as follows:

1. **Full-text Retrieval.** The pipeline begins by submitting the user's biomedical question as a query to our Solr-based retrieval system, which indexes the entire PubMed corpus. Using the classical BM25 algorithm and inverted indexing, the system efficiently retrieves a broad pool of candidate documents that exhibit high lexical overlap with the query. To ensure comprehensive coverage, we typically collect the top $1,000$ abstracts per question at this stage. While this method is highly effective for recall and speed, it may overlook semantically relevant documents whose wording differs from the query due to the specialized and varied vocabulary of the biomedical domain.

2. **Synthetic Snippet Generation for Soft Reranking Guidance.** A fundamental limitation of standard retrieval systems is that, when given a biomedical question as a query, they are designed to retrieve documents that are lexically or semantically similar to the query itself, which is typically a question, rather than to the actual answer. In an ideal scenario, if we already possessed the answer (or a passage containing it), we could append it to the query, thereby enabling the retriever to prioritize documents that directly contain the sought information. However, since the answer is unknown by definition, this strategy is not directly applicable. To address this gap, we leverage recent advances in prompt-based query expansion [24, 25], introducing a novel intermediate step: synthetic answer snippet generation. Specifically, we fine-tune a decoder-only LLM on historical BioASQ data [26], training it to generate plausible answer snippets for a given question by learning from pairs of previous BioASQ questions and their corresponding gold-standard snippets. When presented with a new biomedical question, the model is thus able to hypothesize one or more candidate answer passages. For example, in response to *"What are the common side effects of the drug Imatinib?"*, the model may generate: *"Our research shows that Imatinib is known to cause side effects such as nausea, rash, and edema."* Importantly, these generated snippets are not used to directly reformulate the retrieval query, as this would risk introducing factual errors or hallucinations into the retrieval process. Instead, we encode the generated snippet into a dense vector and use it as "soft guidance" during the subsequent semantic reranking phase. By incorporating the snippet embedding into the document scoring function, the system is biased to prefer documents that semantically align not just with the original question, but also with the structure and content of a plausible answer. This strategy aims to bridge the gap between question-centric retrieval and answer-centric evidence gathering, exploiting the generative capabilities of LLMs to inject answer-like signals into the retrieval pipeline without contaminating the search with potentially hallucinated facts. This approach directly addresses our first research question (Q1): *Does generating plausible candidate answer snippets as an intermediate step, using a generative LLM, improve unsupervised document reranking?*
3. **Unsupervised and Supervised Reranking.** After the initial retrieval, which typically returns the top 1000 abstracts from the corpus, UniTor@BioASQ applies a two-step reranking pipeline that balances efficiency with precision. In the first step, *unsupervised (dense) reranking*, all candidate abstracts are encoded as dense vectors using pre-trained sentence encoders (e.g., Sentence-BERT, PubMedBERT). Since these document embeddings are precomputed offline for the entire corpus, the semantic similarity calculation between each abstract and both the original question and the generated synthetic snippet can be performed rapidly at inference time, typically as simple dot products in vector space. This allows the system to efficiently rescore all 1000 candidates, producing a semantically informed ranking that prioritizes documents likely to contain answer-relevant content, as hypothesized by the generated snippet (see Research Question Q1). The top 100 abstracts from this dense reranking stage are then forwarded to a *supervised reranker*, which is based on a fine-tuned model encoder transformer. This classifier is more computationally expensive, as it processes full question-document pairs to predict a scalar relevance (or entailment) score, trained using positive and hard-negative examples from BioASQ. Applying this more selective, high-precision model only to the top 100 documents ensures tractability while substantially increasing the precision of the final ranking. Ultimately, the supervised reranker identifies the 10 best abstracts to serve as evidence for downstream snippet extraction and answer generation. This two-step design enables UniTor@BioASQ to combine the scalability and broad coverage of fast unsupervised similarity with the deep discrimination of supervised relevance modeling, maximizing both recall and precision in evidence selection.
4. **Supervised Sequence Labeling for Snippet Extraction and Snipped-based Document Filtering.** The next stage addresses the identification and extraction of precise evidence from the top-ranked abstracts. We employ a decoder-only LLM, fine-tuned to perform sequence labeling: given a question and a candidate abstract, the model generates the same abstract with relevant text spans explicitly marked using special tags (e.g., `[BS]` ... `[ES]`), while non-relevant portions of the abstract are omitted from the output. For example, given the question *"What are the common side*

effects of the drug Imatinib?", and a candidate abstract, the model might output only the relevant snippet: `[BS]`Side effects include nausea, rash, edema, and muscle cramps`[ES]`. All surrounding, non-relevant content from the abstract is discarded at this stage, so that only the evidence most directly supporting the answer is retained for downstream processing. To make the snippet extraction model robust to overgeneration (Research Question Q3), we explicitly address a key issue in sequence labeling for QA. If the model is fine-tuned only on positive examples, that is, (question, document) pairs where a relevant answer snippet is always present, it tends to always extract a span, even when none should be found. This results in high recall but poor precision, as the system is likely to produce false positives when no relevant answer is present in a document. This problem has been previously documented in the context of extractive QA, most notably in the development of the SQuAD2.0 dataset [27], which demonstrated that models trained exclusively on answerable questions struggle to abstain when faced with unanswerable or irrelevant passages. To overcome this, we explicitly introduce negative and borderline examples during fine-tuning. Negative examples are abstracts that do not contain the answer to the question: some are sampled randomly from the corpus (ensuring they are topically unrelated), while others are selected as "borderline negatives." These borderline cases are abstracts that appear in the mid-rank positions (e.g., between the 50th and 100th position) of the retrieval list for a question but are not annotated as containing an answer in BioASQ. Such documents are particularly useful: although they may discuss related topics, they lack a true answer, making them realistic distractors.
By training the model on a mix of positive, negative, and borderline (difficult negative) instances, we explicitly teach it not only to identify relevant snippets but also to abstain from marking any span when no answer is present. In such cases, the expected output is a special "empty snippet" tag, e.g., `[BS][ES]`, indicating that the abstract contains no relevant evidence. This design improves the model calibration and greatly enhances its precision, making it less likely to hallucinate evidence in irrelevant documents, a property essential for high-stakes biomedical QA. After sequence labeling, the model outputs marked spans for each abstract. If no relevant span is detected, the document is filtered out from the pipeline (Research Question Q4). This aggressive filtering mechanism increases the signal-to-noise ratio of the evidence passed to downstream modules, but introduces a trade-off: It can increase precision, but risks a drop in recall if relevant information is mistakenly filtered out due to model error. Thus, this step explicitly tests how snippet-based filtering impacts the balance between evidence quality and retrieval coverage within the UniTor@BioASQ pipeline.

5. **Snipped-based Pseudo-Relevance Feedback.** Building on the intuition developed in the earlier stage, where we leveraged LLM-generated hypothetical snippets to provide additional semantic guidance for reranking, we now exploit snippets directly extracted from actual retrieved documents. Unlike generated snippets, these are verbatim excerpts from biomedical literature, and thus can be considered trustworthy and free from hallucinations. The key idea is: if our pipeline can extract concise, high-quality evidence from real documents, why not use these validated snippets to further refine the retrieval process? To this end, after the sequence labeling model identifies the most relevant snippets, we concatenate the original question with a selection of these snippets to construct an expanded query. For example, the original query *"What are the common side effects of Imatinib?"* can be expanded as
"What are the common side effects of Imatinib? Side effects include nausea, rash, and edema." This augmented query is then fed back into the retrieval pipeline, starting again from the sparse retrieval stage, so the system can exploit both the explicit information needed and evidence already found in the corpus. Since these snippets are drawn from genuine PubMed abstracts, we are no longer concerned about introducing hallucinated or misleading content. As a result, this feedback loop enables the system to surface additional relevant documents that may have eluded the initial search due to differences in terminology, phrasing, or document structure. By incorporating real, high-confidence snippets as query expansion, UniTor@BioASQ can iteratively improve the breadth of retrieved evidence and increase the likelihood of finding overlooked but pertinent documents, thus further strengthening downstream answer generation.

6. **Supervised Answer Generation.** At the final stage of the UniTor@BioASQ pipeline, the curated evidence, in the form of top-ranked documents and their most relevant extracted snippets, is fed into a supervised, decoder-only LLM trained specifically for biomedical answer generation. This module is designed to address the full diversity of BioASQ question types, including yes/no, factoid, list, and summary (ideal answer) formats. To achieve robust performance, we fine-tune the answer generation model in a multi-task learning framework, exposing it to all question types and contexts during training. The model is presented with a prompt that includes the biomedical question, its type, and a set of retrieved abstracts and/or extracted snippets, each tagged to highlight their relevance. This setup enables the model to learn not only how to select the appropriate answer format, but also how to prioritize and synthesize information from heterogeneous and sometimes incomplete or noisy sources. For example, given the question *"What are the common side effects of Imatinib?"* and supporting snippets such as `[BS]`Side effects include nausea, rash, edema, and muscle cramps`[ES]`, the model is trained to produce a list-formatted answer: `[Nausea, Rash, Edema, Muscle Cramps]`. This multi-task training paradigm directly addresses the challenge in Research Question Q5: *Does training the answer generation LLM on both abstracts and extracted snippets, in a unified framework, improve answer quality across different biomedical question types?* By leveraging the complementary strengths of full abstract context (for coverage) and focused snippets (for precision), the system is able to maximize answer accuracy and adaptability, delivering concise factoid lists, confident yes/no answers, or informative summaries as appropriate. This design ensures that even when the evidence passed through the pipeline is partial, noisy, or sparse, the answer generator is capable of producing high-quality, context-aware outputs that are both faithful to the evidence and aligned with the requirements of biomedical question answering in the BioASQ framework.

In summary, UniTor@BioASQ brings together advanced retrieval, semantic reranking, targeted evidence extraction, and answer generation into a unified pipeline for biomedical question answering. The following sections describe each component in detail, outlining key design choices, training procedures, and empirical evaluation.

3.3. Phase A - Document Retrieval and Snippet Extraction

We participated in both the document retrieval and snippet extraction subtasks of the BioASQ Challenge Phase A. To address these tasks, we designed a multi-stage retrieval pipeline tailored to the complexities of biomedical question answering.

Our system consists of up to four sequential stages for document retrieval:

- Full-Text Retrieval: We employ BM25 (via Solr) to efficiently retrieve a high-recall set of candidate documents.
- Unsupervised Reranking: We reorder documents based on cosine similarity between dense embeddings of the query and documents.
- Supervised Reranking: A transformer-based model predicts entailment scores, refining document ranking with higher precision.
- Supervised Filtering (Optional): A decoder-only LLM filters out documents that lack relevant content based on titles and abstracts.

For the snippet extraction subtask, we use a supervised sequence labeling model based on a decoder-only LLM (e.g., LLaMA), trained to identify the most relevant text spans within documents. Additionally, the extracted snippets can optionally feed back into the retrieval process via a pseudo-relevance feedback loop, iteratively refining document and snippet selection.

3.3.1. Full-Text Retrieval

The initial stage of the UniTor@BioASQ pipeline is dedicated to efficiently retrieving a high-recall set of candidate documents that are likely to contain relevant evidence for a given biomedical question. For

this, we employ the BM25 ranking model as implemented in Solr, a well-established and competitive baseline in biomedical information retrieval [28, 29]. BM25 balances term frequency with document length normalization and has demonstrated robust performance across large-scale retrieval tasks, including BioASQ challenges [30].

To optimize the effectiveness of this step, we tuned the BM25 hyperparameters, K_1 (term frequency saturation) and b (document length normalization), using grid search over the BioASQ training data, targeting Mean Average Precision (MAP) on a validation split. Our results confirmed that the default values ($K_1 = 1.2, b = 0.75$) already achieve high recall (typically exceeding 85%), underlining their suitability for the biomedical domain.

Solr support for field-specific weighting allows us to balance the contribution of titles and abstracts during retrieval. While we hypothesized that titles, being concise, might provide stronger relevance signals, experiments showed that equal weighting for both title and abstract fields ($w_t = w_a = 1$) consistently yielded the best retrieval performance, suggesting that both fields provide complementary information.

Finally, we set $k = 1000$ as the number of top-ranked documents retrieved per query, based on empirical analysis showing this value consistently exceeds 85% recall while keeping downstream computational costs tractable. The output of the sparse retrieval stage is thus, for each query q, a set of k candidate documents $\{d_1, \ldots, d_k\}$, each scored as:

$$\text{FullTextScore}(q, d) = \text{BM25}(q, d; k_1, b, w_t, w_a)$$

where $K_1 = 1.2$, $b = 0.75$, $w_t = 1$, and $w_a = 1$. Here, d refers to either the title or abstract of the document, and w_t, w_a denote the respective weights for title and abstract fields in the index. This high-recall candidate pool provides the foundation for subsequent semantic reranking and answer extraction in the UniTor@BioASQ pipeline.

3.3.2. Question Reformulation via Synthetic Snippet Answer Generation

Query expansion with plausible answer snippets is a powerful technique for enhancing document retrieval, particularly in the biomedical domain where relevant information is often paraphrased or distributed across multiple abstracts [31, 25, 24]. However, naïvely concatenating generated snippets to the original query risks introducing hallucinated or misleading content, which could undermine retrieval precision and the trustworthiness of the pipeline.

Instead, UniTor@BioASQ adopts a more controlled strategy: generated snippets are *encoded* into dense vectors and used as "soft guidance" in the dense reranking phase, supplementing the original question embedding. This enriches the semantic signal for reranking, steering the model toward answer-bearing documents without altering the explicit query text or introducing unverified facts. To generate synthetic and plausible snippets that can guide the reranking phase, we train a neural model on the BioASQ dataset, leveraging the annotated abstracts and question-answer pairs to learn how to produce candidate snippets that are both semantically rich and relevant. These generated snippets aim to approximate the gold-standard answer snippets available in BioASQ, which serve as a reference for supervised learning.

For each question in the BioASQ training set, we have access to gold-standard snippets extracted from abstracts annotated as relevant. Using a pre-trained Sentence-BERT model, we measure the semantic similarity between the question and each associated snippet. Since Sentence-BERT is specifically trained to capture both syntactic and semantic relatedness, high-similarity snippets are likely to be directly answer-relevant, while lower-similarity snippets might capture more peripheral aspects. To ensure diverse supervision and to encourage the model to generate a spectrum of plausible, realistic answers, we sort all snippets for a question by similarity and partition them into three groups of three: the *top-3* (most similar), the *middle-3*, and the *bottom-3*. All available snippets are thus included, maximizing coverage. During training, each group forms an output "triple" for the LLM, which is prompted as shown in Figure 2. This grouping strategy, compared to randomly sampling a single snippet, exposes the model to various ways in which relevant information might be formulated, from very direct to

more nuanced or tangential expressions. Moreover, this "three-snippet grouping" forces the model to generalise across varying degrees of relevance, rather than memorising a single gold phrase. The snippet generator is trained, fine-tuning the LLM using Low-Rank Adaptation (LoRA, [32], on $9,222$ grouped question–triple instances, with batch size 32, for 3 epochs, learning rate 6×10^{-4}, LoRA rank/alpha $= 16$, and max sequence length 768 tokens. At inference, we feed only the question and let the model produce up to three candidate snippets, using greedy search (`do_sample=false`) to ensure maximum faithfulness and reproducibility. In practice, the model nearly always returns a single, concise snippet as its first output; if multiple are returned, we select only the first (most confident) one for downstream use. If an empty snippet is generated, the reranking step that relies on this additional signal is simply skipped. On development data, the 8-B parameter `unsloth/llama-3-8b-Instruct`[8] consistently yielded the highest MAP scores, narrowly outperforming the 14-B parameter `unsloth/Phi-4-mini-instruct`[9]; we therefore adopt the former as UniTor's default snippet generator.

During dense reranking (discussed below), the generated snippet embedding is combined with the question embedding to compute semantic similarity with each candidate document; further details and weighting strategies are described in the corresponding section. This "soft guidance" aims to bridge the gap between question-centric retrieval and answer-centric evidence gathering, addressing *Research Question Q1.* Quantitative ablation results supporting this choice are presented in the experimental section 4.2.

You are given a question related to biomedical research. Your task is to generate 3 relevant answer snippets that might realistically appear in a PubMed abstract or article discussing this topic. These snippets should be:

- Highly specific: Provide a precise and directly relevant response to the question.
- Concise: Limit the snippet to one or two sentences.
- Informative: Include well-defined biomedical details (e.g., specific proteins, genes, pathways, mechanisms, experimental methods, or effects).
- Contextually accurate: The snippet should resemble an excerpt from a scientific publication, ensuring clarity and rigor.
- Avoid vague or generic statements (e.g., T̈his paper explores...). Instead, focus on delivering a scientifically meaningful response with clear biological relevance.

Question: Which protein mediates gene loop formation in the yeast S. cerevisiae?

Snippets:
- Gene-loop formation is dependent on regulatory proteins localized at the 5' and 3' ends of genes, such as TFIIB.
- Gene looping, defined as the interaction of the promoter and the terminator regions of a gene during transcription, requires transcription factor IIB (TFIIB).
- Gene ...

Figure 2: Prompt template used for domain-adapted plausible snippet generation. Training triples are constructed by grouping question–snippet pairs based on semantic similarity, exposing the LLM to a range of plausible answer formulations for each question.

3.3.3. Dense/Unsupervised Reranking

The initial sparse retrieval step based on BM25 efficiently collects a high-recall pool of candidate documents, typically the top $1,000$ abstracts per question. However, as previously discussed, BM25 operates purely on lexical overlap and thus may miss relevant documents that use alternative terminology or paraphrase the information need. To address this limitation and move beyond surface-level matching, we introduce an unsupervised reranking step grounded in dense semantic representations. In this stage, each candidate document is scored based not only on its similarity to the original question but also on its alignment with the plausible answer snippet generated earlier by our LLM. This design leverages the insight (from Research Question Q1) that combining question-centric and answer-centric signals can significantly improve the selection of genuinely relevant evidence. If a document aligns closely with both the question and the generated snippet, which encodes the typical linguistic structure and

[8] https://huggingface.co/unsloth/llama-3-8b-Instruct
[9] https://huggingface.co/unsloth/Phi-4-mini-instruct

content of a true answer, it is likely to contain valuable information for answer synthesis. Formally, for each (question, document) pair (q, d), we separately encode:

- the question q,
- the plausible snippet s_q generated for q,
- the document's title t_d and abstract a_d,

As dense vectors using a pre-trained sentence encoder. The dense similarity score is computed by averaging the cosine similarities for title and abstract, and combining question-based and snippet-based signals as follows:

$$DenseScore(q, d) = 0.4 \times \frac{C(q, t_d) + C(q, a_d)}{2} + 0.6 \times \frac{C(s_q, t_d) + C(s_q, a_d)}{2}$$

where $C(x, y)$ denotes the cosine similarity between the corresponding embeddings. The $0.4/0.6$ weighting prioritizes the semantic match with the generated snippet, reflecting its role as a "soft proxy answer" (see Section 3.3.2), while retaining the user's original query as an anchor. These weights were selected via grid search over the BioASQ validation set, optimizing the Mean Average Precision (MAP); we observed that variations within ± 0.1 yielded similar results, indicating stability of this tuning. For the sentence encoder, we compared several options, including general-purpose and domain-specific models. Despite expectations that PubMedBERT[10] would excel in biomedical QA, our experiments showed that SentenceBERT[11] consistently achieved higher MAP, both on BioASQ validation and on out-of-domain biomedical queries. This aligns with recent evidence that SentenceBERT models provide robust generalization even in specialized domains [33]. All dense scoring in UniTor@BioASQ thus relies on SentenceBERT, with document vectors precomputed for scalability and cosine similarities computed efficiently in batch (using in-memory matrix operations)[12]. After reranking, we select the top 100 documents for subsequent supervised reranking and snippet extraction. This cutoff was chosen to balance coverage, maintaining over 80% recall of gold abstracts on BioASQ validation sets, with the need to control computational costs in downstream, more resource-intensive stages. In summary, this dense, answer-aware reranking step bridges sparse retrieval and fine-grained filtering, enriching the candidate pool with semantically relevant, answer-aligned documents. Experimental results (see Section 4) quantify the impact of this strategy, directly addressing Research Question Q1.

3.3.4. Supervised Reranking

While the unsupervised reranking step based on dense semantic similarity provides a valuable filtering mechanism, it cannot capture all the fine-grained, contextualized judgments needed to distinguish between genuinely relevant and superficially similar biomedical documents. To address this limitation, we introduce a supervised reranking stage, which leverages transformer-based models to further refine the candidate pool and select the top-10 documents for each question as required by the BioASQ challenge.

Our approach employs an encoder-based transformer model, ModernBERT[13][34], designed to process pairs of text as input, specifically, the biomedical question and either the title or abstract of a candidate document. Inspired by advances in natural language inference (NLI), where BERT-like architectures have excelled at modeling semantic entailment between sentence pairs[35, 36], we cast the reranking task as a binary classification problem during training: the model is trained to distinguish pairs where the document contains a correct answer (positive) from those where it does not (negative). However, during inference, we use the softmax probability of the model for the positive (entailment) class as a

[10] https://huggingface.co/NeuML/pubmedbert-base-embeddings

[11] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[12] A technical note: if the snippet generator outputs an empty snippet (i.e., no plausible answer could be hypothesized), we default to computing similarity solely between the question and the document (setting the 0.6 weight to zero), thus ensuring graceful fallback.

[13] https://huggingface.co/answerdotai/ModernBERT-base

continuous relevance score for each question-document pair, enabling fine-grained ranking based on the estimated degree of answer entailment. This continuous score serves as a fine-grained relevance signal for ranking, allowing us to order candidate documents not just by binary relevance, but by their estimated degree of answer entailment. This hybrid training/inference strategy, classification for supervision, probability scoring for prediction, ensures that the final document ranking is both discriminative and sensitive to subtle differences in evidence quality, which is particularly important in biomedical QA, where relevant information may be nuanced or implicit.

A key challenge in supervised reranking is the construction of high-quality training data, particularly the availability of both positive and negative examples for each question. Our dataset is built as follows:

- **Positive Instances:** For each question in the BioASQ training set, all abstracts annotated as relevant are paired with the question and labeled as positive.
- **Negative and Borderline Instances:** To expose the model to hard negatives (documents that are topically close but ultimately not answer-bearing), we:
    1. Use Solr BM25 to retrieve the top $k = 100$ documents for each question.
    2. Exclude all gold-standard abstracts from this set to prevent label leakage.
    3. Remove any document published in 2024 or later (to avoid future data leakage).
    4. Randomly sample 40 abstracts per question from the remaining pool, focusing on ranks 20 through 100 in the BM25 list. This ensures a mix of moderately and weakly relevant negatives, many of which are challenging for the model.

This construction yields a balanced set of positive pairs (q, d^+) and negative pairs (q, d^-), totaling $214,595$ training instances, and is crucial for teaching the model to distinguish subtle cues of true relevance. The supervised reranker is trained using standard binary cross-entropy loss, with positive and negative pairs as supervision. During inference, for each (question, document) pair, we take the softmax output associated with the positive (entailment) class as the document's final supervised relevance score. Because this step is computationally more expensive than prior stages, we apply it only to the 100 documents previously shortlisted by the unsupervised reranker. This design ensures that computational resources are focused where they are most likely to improve the final answer set. The supervised scores are then combined with the previous retrieval scores to determine the final document ranking. Specifically, for each (question, document) pair (q, d), we compute the final score as the product of the three relevance signals[14]:

$$Final\ Score(q, d) = FullText(q, d) \times Dense(q, d) \times Supervised(q, d)$$

Since the entailment probability output by the supervised model is normalized between 0 and 1, it can be directly combined multiplicatively with the BM25 and dense similarity scores to obtain the final ranking. The top-10 documents for each question, ranked by this final score, constitute the high-quality evidence set for downstream snippet extraction and answer generation modules. This approach ensures that only those documents most likely to directly address the question, based on both learned semantic entailment and classic IR features, are passed forward, thereby maximizing the precision and utility of evidence at later stages.

3.3.5. Snippet Extraction

Following document retrieval and reranking, the next key stage in the UniTor@BioASQ pipeline is snippet extraction, which fulfills the BioASQ requirement to identify concise, evidence-bearing text spans from the top-10 ranked abstracts. The resulting snippets provide focused evidence for subsequent answer synthesis and feedback steps. The objective is to isolate short, self-contained passages, typically

[14]We also explored alternative ranking strategies, such as using only the entailment score, or other aggregation methods, but empirical evaluation on a BioASQ development set (using MAP as the criterion) consistently favored this multiplicative approach.

one or a few sentences, that provide the essential facts or findings needed to answer the biomedical question. These snippets serve as an interpretable bridge between large-scale document retrieval and final answer synthesis.

We cast snippet extraction as a sequence labeling task implemented via text generation, leveraging decoder-only LLMs such as LLaMA-3 and Phi-4 [37]. Given a question and a candidate abstract, the model is prompted to reproduce the abstract while explicitly marking relevant spans with special tokens: [BS] (begin snippet) and [ES] (end snippet). This prompt-based design enables the use of autoregressive LLMs without the need for a separate token classifier and allows for robust, instruction-following behavior even in complex, multi-sentence biomedical contexts. Unlike general-purpose language models, an effective biomedical snippet extractor must demonstrate:

- **Domain adaptation**: Correct interpretation of biomedical terminology, reasoning about experimental evidence, and understanding nuanced clinical/research contexts.
- **Task adaptation**: Precise marking of text spans directly responsive to the question, and the ability to abstain from overgeneration (avoiding false positives).

To ensure both high recall and high precision, we again (as for the Supervised Re-ranking) curate a training set from BioASQ and additional PubMed abstracts as follows:

- **Positive instances**: For each question, every gold-standard relevant abstract forms a training pair. The ground-truth snippets are annotated, and the model learns to mark the corresponding spans.
- **Negative and borderline instances**: To combat overgeneration (as documented in SQuAD2.0 [27]), we include hard negative examples. For each question, among the top k BM25 candidates (excluding golds and 2024+ docs), we randomly select one abstract ranked 11–30 that is not annotated as relevant. These "borderline negatives" are semantically close but lack a true answer. The model is expected to output only [BS][ES] (empty snippet) in such cases, learning to abstain.

This results in a balanced and challenging dataset, crucial for calibration and for reducing false positives.

The input to the LLM consists of a pair: the biomedical question and the candidate abstract. As illustrated by the prompt in Figure 3, the model is instructed to carefully read both and return only the portions of the abstract that directly address the question, explicitly marking relevant text spans with [BS] and [ES] tags. If no answer is present, the expected output is simply [BS][ES]. The prompt emphasizes precision, discourages hallucination, and makes clear that the model should abstain from extracting snippets when the document is not informative. This explicit instruction and example-based format (see Figure 3) guides the model to focus on factual, evidence-based extraction. We trained our snippet extraction model on a set of $43,008$ annotated question–abstract pairs, using a batch size of 32 and running for two full epochs. The maximum sequence length for each example was set to $1,024$ tokens for the abstract, plus an additional 256 tokens to accommodate the question. For optimization, we used the Adam algorithm with a learning rate of 6×10^{-4}, and applied parameter-efficient fine-tuning using LoRA (with rank and alpha both set to 16). The model was trained to minimize cross-entropy loss, focusing specifically on the generated output tokens. At inference time, we generated snippets by prompting the model with each (question, abstract) pair and allowing it to generate up to 256 new tokens. To ensure maximum consistency and reliability, we used greedy decoding (i.e., always selecting the highest-probability token at each step). We experimented with two decoder-only language models, both available in resource-efficient versions through the Unsloth library: Llama-3[15] and Phi-4[16]. Based on validation results (evaluated using Mean Average Precision), Llama-3 consistently outperformed Phi-4 on the snippet extraction task, so we adopted Llama-3 as the default model for our pipeline.

[15] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-bnb-4bit
[16] https://huggingface.co/unsloth/phi-4-unsloth-bnb-4bit

You are an expert biomedical researcher skilled in extracting relevant information from scientific literature. Your task is to identify and extract key snippets from a given PubMed abstract or title that provide useful information to answer a specific biomedical question.

Instructions:
- Understand the question: Carefully analyze the biomedical question to grasp its key concepts, entities, and relationships.
- Analyze the document: Read the provided title or abstract carefully, identifying sentences or phrases that contain relevant information.
- Extract the snippet: If a portion of the text is relevant, extract it exactly as it appears in the original text and enclose it within the tags [BS] and [ES]. Example: [BS] extracted snippet [ES].
- Handle irrelevant cases: If the document does not contain any relevant information, return only [BS] [ES] with no content inside.
- Be precise: Ensure that extracted snippets are complete, self-contained, and directly relevant, without modifying or adding words. Always enclose every extracted snippet within [BS] and [ES] to ensure clarity and consistency.

Question: Which protein mediates gene loop formation in the yeast S. cerevisiae?

Abstract/Title:
Gene looping, defined as the physical interaction between the promoter and terminator regions of a RNA polymerase II-transcribed gene, is widespread in yeast and mammalian cells. Gene looping has been shown to play important roles in transcription. Gene-loop formation is dependent on regulatory proteins localized at the 5' and 3' ends of genes, such as TFIIB. However, whether other factors contribute to gene looping remains to be elucidated. Here, we investigated the contribution of intrinsic DNA and chromatin structures to gene looping. We found that Saccharomyces cerevisiae looped genes show high DNA bendability around middle and 3/4 regions in open reading frames (ORFs). This bendability pattern is conserved between yeast species, whereas the position of bendability peak varies substantially among species. Looped genes in human cells also show high DNA bendability. Nucleosome positioning around looped ORF middle regions is unstable. We also present evidence indicating that this unstable nucleosome positioning is involved in gene looping. These results suggest a mechanism by which DNA bendability and unstable nucleosome positioning could assist in the formation of gene loops.

Snippets:
[BS]Gene-loop formation is dependent on regulatory proteins localized at the 5' and 3' ends of genes, such as TFIIB.[ES] [BS]Gene-loop formation is dependent on regulatory proteins localized at the 5' and 3' ends of genes, such as TFIIB.[ES]

Figure 3: Example of the prompt used to train the LLM for snippet extraction. The model is instructed to identify and explicitly tag text spans in the abstract that answer the biomedical question, using the special markers [BS] and [ES]. The prompt also clarifies that if the document does not contain relevant information, the model should abstain and output only [BS][ES].

3.3.6. Snippet-based Filtering

After snippet extraction, UniTor@BioASQ applies an additional snippet-based filtering stage to further improve the precision and quality of evidence passed to the final answer generation module. This step directly addresses Research Question Q4: "*What is the impact of aggressive snippet-based filtering on system recall and precision?*".

The motivation for this component is straightforward: even after multi-stage retrieval and reranking, some of the top-10 candidate abstracts may not actually contain any answer-relevant content, or their relevance may be borderline or highly speculative. Allowing such documents to pass unfiltered into downstream answer generation can reduce answer quality and trustworthiness.

To mitigate this, we use the output of the supervised snippet extraction model as a hard filter on the candidate document set. Specifically, for each of the top-30 abstracts identified by the supervised reranker, we examine whether the snippet extraction model has marked any non-empty evidence span (i.e., a text segment wrapped in [BS]...[ES]). Only the first 10 documents containing at least one relevant snippet are retained; documents for which the model outputs only [BS][ES] (indicating no answer found) are filtered out. If fewer than 10 such documents are available, all available positives are used. This approach offers a simple, interpretable, and robust filter for document-level evidence. Compared to greedy selection of the top-10 ranked documents (irrespective of their snippet content), this strategy prioritizes documents for which the model can point to explicit, self-contained answer evidence. In effect, the snippet extraction model acts as a high-precision gatekeeper, further reducing noise and the risk of passing irrelevant context to the answer generator.

However, this "*aggressive*" filtering comes with a natural trade-off: while it increases precision by

removing uninformative documents, it may reduce recall if the snippet extractor occasionally fails to recognize relevant evidence in borderline or difficult cases. For this reason, we empirically compare both strategies, greedy top-10 and snippet-based filtering, in our experiments (see Section 4), analyzing the impact on downstream answer accuracy and robustness.

3.3.7. Pseudo Relevance Feedback via Extracted Snippets

In the final stage of our pipeline, we introduce a Pseudo Relevance Feedback (PRF) mechanism aimed at enhancing overall system recall by reformulating the query based on high-confidence evidence extracted during the retrieval process. While earlier stages employ LLM-generated plausible snippets to influence unsupervised reranking, enriching document scores with contextual hints, these generated snippets, though useful, are not guaranteed to be hallucination-free, as they are outputs of a generative model. Consequently, we deliberately avoid injecting them into the initial sparse retrieval step to prevent speculative or noisy content from distorting early retrieval.

By contrast, the PRF step leverages a more trustworthy signal: the top-ranked snippets extracted from PubMed abstracts by our snippet extraction model. These have passed through the full retrieval and reranking pipeline and thus reflect high-confidence, evidence-grounded content. Their reliability makes them an excellent basis for expanding the original query without risk of hallucination.

The mechanism proceeds as follows. First, we gather snippets from the top-10 documents previously selected by the supervised reranker. Snippets are sorted according to two criteria: (i) the rank of the source document, and (ii) their order of appearance within each document. From this sorted list, we select the first three snippets available, which may originate from one or multiple documents, depending on their distribution. These snippets are concatenated into a single string, which is then appended to the original question to form an augmented query.

Importantly, this snippet-based feedback operates strictly as a post-processing augmentation of the original question and does not propagate back to the synthetic snippet generation module. As illustrated in Figure 1, the pipeline preserves a clear separation: the initial LLM-generated synthetic snippet, used for semantic guidance in dense reranking, remains unaffected by any noisy or redundant evidence that could emerge from iterative query expansion. This design ensures that the generative LLM is never exposed to feedback signals potentially corrupted by extraction errors, thereby preserving the integrity and robustness of the answer hypothesis process.

The augmented query is then resubmitted to the sparse retrieval engine (BM25), enabling the system to recover documents that may have been missed initially due to lexical mismatch or implicit associations. These additional documents are processed through the full reranking pipeline, including dense and supervised stages, just like the original set.

This approach is conceptually aligned with pseudo-relevance feedback strategies in traditional information retrieval [38, 39], but tailored to the biomedical QA setting. Instead of heuristically selected terms, we expand the query using high-quality, domain-specific text segments. This principled, evidence-based expansion has been shown to improve recall and downstream answer quality.

Notably, our feedback loop remains modular and lightweight: it operates on already computed outputs (question and extracted snippets), requires no additional supervision, and does not alter previously trained components. It complements earlier pipeline stages by systematically exploiting the most reliable information available, grounded, extracted biomedical evidence, to recover additional relevant literature.

3.4. Phase A+ and Phase B: Retrieval Augmented Generation

For Phases A+ and B of the BioASQ challenge, we participated in the Answer Generation subtask, covering all types of questions: yes/no, factoid, list, and summary. The primary distinction between these phases lies in the resources provided by the organizers: in Phase B, participants are given the gold-standard documents and snippets, while in Phase A+, participants must rely on the documents retrieved during Phase A. Our focus was on developing an effective strategy to train a large language

model to be sensitive and robust to the specific biomedical context, enabling it to produce precise and relevant answers.

3.4.1. Answer Generation

The final stage of our pipeline focuses on the generation of accurate answers to biomedical questions, leveraging the context provided by the retrieved and processed documents. This component closes the loop of our RAG system, moving from relevant documents and extracted snippets to producing explicit, final answers in the format required by the BioASQ challenge.

To address the diverse nature of biomedical questions, including yes/no, factoid, list, and summary (ideal) answer formats, we employ a supervised decoder-only language model, such as LLaMA-3 or Phi-4, trained in a multitask learning framework. This approach allows the model to simultaneously learn to identify the appropriate answer format and to generate precise, contextually grounded responses.

The multitask paradigm explicitly addresses our fifth research question (Q5): *Does training the answer generation LLM in a multi-task fashion, leveraging both abstracts and highlighted snippets as input context, improve answer quality for different biomedical question types?* Indeed, cross-task knowledge transfer is central to our method [1, 40, 41]. For example, the model might leverage the ability learned from factoid questions to extract named entities, thereby enhancing its capability to generate concise and precise yes/no answers or identify salient information for ideal summaries.

To train our multitask model, we construct a dataset from BioASQ gold-standard annotations, following a structured process. Each training instance includes:

- The biomedical question.
- Explicit indication of the question type (yes/no, factoid, list, summary).
- A controlled context comprising either three or five gold-standard relevant abstracts, truncated to meet prompt length constraints and focused explicitly around the gold-standard snippets tagged with `[BS]` and `[ES]` markers.

This approach encourages the model to focus solely on informative passages, significantly reducing the risk of hallucinations or irrelevant outputs.

The output labels for each instance are tailored specifically to the answer type:

- Yes/no: concise answers (*yes*, *no*).
- Factoid: up to five named entities.
- List: comprehensive enumeration of relevant items.
- Summary: a concise and informative ideal-answer paragraph.

This training setup is illustrated in Figure 4, which shows the structured prompt format used to supervise the model. Each instance contains: (i) an explicit task instruction that defines the expected answer type (e.g., factoid), (ii) the biomedical question, and (iii) a set of relevant abstracts where relevant spans are pre-marked with `[BS]` and `[ES]`.

Unlike the snippet extraction stage, where the model task was to re-generate the input abstract while tagging spans, here we provide the full abstract as-is, with ground-truth snippets highlighted but not removed. This difference is essential and carefully motivated:

- **In snippet extraction**, the model was trained to copy the input and only modify it by placing snippet tags. However, because the loss is averaged over the entire generated sequence, the model can trivially achieve low loss by copying 99% of the text correctly while still failing to tag the relevant span, a failure mode that cannot be detected purely through perplexity or token-wise loss.
- **In answer generation**, such fragility would be detrimental. If a snippet alone is given as input and it happens to miss a key detail, the model will have no chance of recovering that missing evidence. Instead, we feed the full abstract and highlight informative regions using tags. This gives the model complete access to context while softly guiding its attention toward the most relevant spans. The model is thus better equipped to perform grounded, faithful answer generation.

This strategy aligns well with our goal of robustness, particularly for factoid and list questions, where answers may be expressed in multiple ways across different parts of the document. By training the model on full abstracts with localized highlights, we help it learn both *where to look* and *how to extract* precise answers, mitigating the risk of hallucination or omission.

A practical caveat of this approach is that, at training time, the original snippet extractor provides only the extracted spans, not the full tagged document. To overcome this, we automatically re-align each extracted snippet to its exact location in the source abstract. Notably, across the entire training corpus, we did not observe any case where a snippet generated by the extractor failed to match exactly a contiguous span in the original abstract, confirming the stability and reproducibility of our tagging procedure. To ensure robustness, training instances include variations in the number of provided abstracts (either 3 or 5). This design choice simulates real-world variability in evidence availability, enabling the model to generalize effectively across different contextual scenarios and evidential completeness.

We trained the answer generation model on $7,459$ examples, using a batch size of 32 and running for two full training epochs. Fine-tuning was done efficiently using the LoRA technique, which keeps training lightweight by updating only a small number of additional parameters (with rank and alpha both set to 32). The model was optimized using the Adam algorithm with a learning rate of 6×10^{-4}, and the training loss was computed only on the tokens the model was expected to generate, ensuring it focused on producing the correct answer.

You are a biomedical expert. Your task is to extract the most relevant factoid-based answer from the provided PubMed abstracts. Relevant information is marked with [BS] and [ES].
Rules:
- Use only the information provided, with a special focus on the relevant information.
- The answer must be a list of up to 5 short expressions (e.g., entity names, numbers).
- If the answer has more than one expression, be careful, the expressions must be ordered by decreasing confidence.
- Often the expressions represent the same concept but written differently.
- Do NOT provide explanations or extra text.
- Maintain this format strictly:[BE] expression1 [EE] [BE] expression2 [EE] ... [BE] expressionN [EE],

Question: Which protein mediates gene loop formation in the yeast S. cerevisiae?

PubMed resources:

Abstract truncated here ... gene, is widespread in yeast and mammalian cells. Gene looping has been shown to play important roles in transcription. [BS]Gene-loop formation is dependent on regulatory proteins localized at the 5' and 3' ends of genes, such as TFIIB.[ES] However, whether other factors contribute to gene looping remains ... Abstract truncated here.

[BS]Gene looping, defined as the interaction of the promoter and the terminator regions of a gene during transcription, requires transcription factor IIB (TFIIB).[ES] We have earlier demonstrated association of TFIIB with the ... Abstract truncated here.

Answer:
[BE]TFIIB[EE]

Figure 4: Example of the prompt format used to train the answer generation model. The instruction specifies the expected answer type (here, a factoid), followed by the input: a biomedical question, its type, and a set of PubMed abstracts. Within the abstracts, relevant evidence is explicitly marked using [BS] and [ES] tags. The model is trained to generate a precise answer, focusing on the tagged spans while having full access to the broader context. This setup ensures the model can rely on grounded information while avoiding hallucinated or unsupported content.

During inference, we used a deterministic decoding strategy (greedy search) to make the output consistent and avoid randomness, which helps ensure factual accuracy and reproducibility.

We tested two models: LLaMA-3[17] and Phi-4[18], both in efficient low-resource versions from the Unsloth library. Their performance was evaluated across all BioASQ question types using dedicated metrics: accuracy for yes/no questions, mean reciprocal rank (MRR) for factoids, F1 for lists, and ROUGE-2-F1 for ideal answers. The two models performed similarly overall, so we chose to alternate

[17] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-bnb-4bit
[18] https://huggingface.co/unsloth/phi-4-unsloth-bnb-4bit

them across submissions: LLaMA-3 was used for batches 1 and 2, and Phi-4 for batches 3 and 4. This setup also allowed us to directly compare their behavior across different challenge phases.

This multitask and multi-contextual training paradigm provides the UniTor@BioASQ pipeline with resilience and adaptability, enabling consistent, high-quality answer generation across diverse biomedical question-answering scenarios, directly addressing and empirically investigating Research Question Q5.

3.5. Hardware Complexity and Computational Footprint

All training and inference for our system was carried out on a single workstation equipped with four NVIDIA A100 GPUs (80GB each). This allowed us to efficiently handle parallel training runs and large-scale batch processing, as required by the BioASQ challenge.

A crucial aspect of our pipeline is that every component relies exclusively on standard, openly available models, such as LLaMA-3 8B, Phi-4 14B, and BERT-based variants. No proprietary or commercial cloud services (such as GPT-4 or Gemini) are used: all steps can be replicated on local hardware without external dependencies.

The end-to-end process, however, does require a non-trivial sequence of model invocations for each query. To clarify, below we summarize the models involved and their specific roles:

- **Synthetic Snippet Generation:** A LLaMA-3 (8B) model is used to generate a "*synthetic snippet*" based on the biomedical question. This serves as a semantic anchor for document retrieval.
- **Semantic Retrieval:** Sentence-BERT encodes both the query and candidate documents into dense embeddings, enabling efficient retrieval of relevant abstracts.
- **Supervised Re-Ranking:** A lightweight BERT-based classifier is applied approximately 100 times (once per candidate abstract) to re-rank retrieved documents by relevance.
- **Snippet Extraction:** The top 10 abstracts are then processed by a LLaMA-based model, which tags the relevant text spans (`[BS]` and `[ES]` markers) to highlight key evidence.
- **Answer Generation:** Finally, the answer generation module, again a supervised LLaMA-3 or Phi-4 model, takes as input the question and the extracted evidence to produce the final answer in the required BioASQ format.

While this modular pipeline requires running multiple neural models per query, it offers strong flexibility and state-of-the-art performance, especially when using multi-task training to streamline model weights across tasks.

It is important to note that all models employed are "*small*" by today's standards (ranging from 100M parameters for BERT to 8B or 14B for LLaMA and Phi-4). This makes our system far less demanding than any single large foundation model (e.g., LLaMA-2/3 70B, GPT-4). Thus, our pipeline is fully reproducible and feasible on modern academic or enterprise hardware, making it both practical and accessible for further research or deployment.

4. Results Analysis

In this section, we present and critically discuss the results obtained by our system in the BioASQ13b 2025 challenge. Our analysis is structured around the main research questions defined in Section 1, with the aim of evaluating the effectiveness and robustness of our pipeline across the key stages of biomedical question answering: document retrieval, snippet extraction, and answer generation.

We begin by briefly outlining the official evaluation metrics used in BioASQ, which provide the quantitative basis for our comparison with other participants. We then report and interpret the experimental results for each subtask and challenge phase, examining not only aggregate scores but also how different design choices affect system performance. Throughout, we highlight strengths, limitations, and lessons learned, with a focus on how our findings address the objectives and open questions motivating this work.

4.1. Evaluation Metrics

The evaluation of our system follows the official BioASQ framework, which adopts standard metrics tailored to each stage of the question answering pipeline:

- **Document Retrieval (Phase A):** Performance is assessed using Mean Average Precision (MAP), F1-score, with both precision and recall reported in detail.
- **Snippet Extraction (Phase A):** The same retrieval metrics are applied, but with adjustments to account for partial overlaps between system and reference spans.
- **Answer Generation (Phase A+, B):** Metrics are specific to question type:
  - *Yes/No:* Accuracy.
  - *Factoid:* Strict accuracy (correct answer at rank 1) and mean reciprocal rank (MRR) to reflect the importance of answer ordering.
  - *List:* Mean precision, recall, and F1-score, comparing predicted entities (normalized for synonyms) against gold standard sets.
- **Ideal Answer Generation:** Both automatic and manual criteria are used. ROUGE-2 and ROUGE-SU4 measure lexical overlap with gold summaries, while manual assessment by experts covers answer recall, precision, redundancy, and readability.

These metrics are consistently reported in our results tables for each module and task. All models and configurations were tuned on the previous year's BioASQ training and validation data, using these metrics for system selection and optimization. This ensures rigorous benchmarking and comparability with prior work in biomedical QA.

4.2. Phase A Results: Evaluation of Document and Snippet Retrieval Pipelines

Phase A of the BioASQ challenge focuses on retrieving relevant documents and extracting supporting snippets for each biomedical question. In this section, we evaluate UniTor@BioASQ, our complete evidence retrieval system (see Figure 1), along with several ablated variants. We report results for both document retrieval and snippet extraction, using the standard BioASQ evaluation metrics introduced above.

All module parameters were previously optimized on last year's official BioASQ training data, as detailed in Section 3; thus, in the following experiments, we do not re-tune hyperparameters. Instead, we systematically compare the full pipeline to versions with selected components (such as pseudo-relevance feedback or document filtering) removed. This allows us to directly assess the impact of each functionality on retrieval and extraction performance, and to address the research questions posed in Section 1.

Document Retrieval Stage. Table 1 reports the document retrieval results for UniTor@BioASQ and its ablated variants across the four official BioASQ batches, with Mean Average Precision (MAP) as the main metric, and F1-score and leaderboard rank provided for completeness.

We evaluated five system configurations:

- $\mathbf{UT_C}$: The complete UniTor@BioASQ pipeline, integrating all retrieval modules: BM25-based query formulation, unsupervised reranking (guided by both the original question and a synthetic snippet), supervised reranking, document filtering, and pseudo-relevance feedback (see Section 3.2).
- $\mathbf{UT_{noDF}}$: As above, but without document filtering (see Section 3.3.6).
- $\mathbf{UT_{noPRF}}$: As above, but without pseudo-relevance feedback (see Section 3.3.7).
- $\mathbf{UT_{noD\&P}}$: Lacking both document filtering and pseudo-relevance feedback.
- $\mathbf{UT_B}$: Baseline using only BM25, unsupervised reranking based on the question, and supervised reranking. It lacks unsupervised reranking with snippet guidance, filtering, and PRF (included for comparative analysis; not officially submitted).

All system hyperparameters were fixed as tuned on the previous year's BioASQ development data (see 3). The results reflect only the effect of enabling or disabling retrieval modules.

Main findings:

- **Stable performance across batches**: The full UniTor@BioASQ system (UT_C) consistently ranks in the top half of submissions and surpasses the median (M) system in three out of four batches, with MAP scores ranging from 6.95 to 29.86. This indicates a high degree of robustness across variable test conditions.
- **Batch-specific variability**: The sharp performance drop observed in Batch 4 for all systems, including the best (BS), suggests increased difficulty or lower coverage in that dataset. This highlights the impact of batch-specific properties on retrieval outcomes.
- **Gap to state-of-the-art**: A persistent MAP gap (10–15 points) separates UT_C from the best system in each batch. However, without public details on BS methodology, no conclusive analysis of this discrepancy is possible.
- **Effectiveness of retrieval modules**: No ablated variant universally outperforms others, indicating non-trivial trade-offs. For instance, $UT_{noD\&P}$ attains the highest MAP in Batch 1, while UT_{noDF} performs best in Batches 2 and 4. These results show that removing document filtering typically increases MAP (due to improved recall) but decreases F1 (due to lower precision). The baseline UT_B is consistently outperformed by systems incorporating additional components, underscoring the incremental value of each module.

In summary, the modular architecture of UniTor@BioASQ enables interpretable performance improvements and competitive robustness relative to the challenge median. However, the relative benefits of each module vary by batch, confirming the need for adaptable, context-sensitive design. We further examine these dynamics in relation to the research questions below.

Table 1
Document Retrieval Stage: Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the MAP and F1, where the rank is ordered by MAP. Bold values represent our best submission. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3			Batch 4		
	MAP	F1	Rank	**MAP**	F1	Rank	**MAP**	F1	Rank	**MAP**	F1	Rank
UT_C	27.12	13.03	25/52	29.86	14.33	21/41	20.00	11.84	23/47	6.95	5.83	33/79
UT_{noDF}	27.84	9.33	23/52	**30.65**	9.64	20/41	20.29	8.21	21/47	**7.24**	4.63	30/79
UT_{noPRF}	30.04	14.86	18/52	29.21	14.57	24/41	20.11	11.49	22/47	6.26	5.79	40/79
$UT_{noD\&P}$	**30.20**	9.48	17/52	29.52	9.65	23/41	**20.72**	7.94	20/47	6.67	4.20	37/79
UT_B	29.27			26.60			19.74			6.54		
BS	42.46	16.05	1/52	44.25	15.46	1/41	32.36	14.45	1/47	18.01	9.27	1/79
M	25.27	11.40		29.86	15.16		18.34	10.64		6.26	5.22	

To address Q1 (Document Reranking Enhancement), we compare $UT_{noD\&P}$ with the baseline UT_B. The former incorporates a key component absent in the baseline: an unsupervised reranking model enhanced with a synthetic snippet generated via a large language model. Across all batches, $UT_{noD\&P}$ consistently outperforms UT_B in terms of MAP (e.g., 30.20 vs. 29.27 in Batch 1; 29.52 vs. 26.60 in Batch 2), with typical gains of 1–3 points. While this comparison does isolate the contribution of the synthetic snippet alone, the consistent improvements suggest that combining semantic guidance from snippet generation with supervised reranking yields measurable gains in document retrieval performance. This supports the hypothesis underlying Q1.

To investigate Q2 (Query Expansion via Snippet-based Pseudo-Relevance Feedback), we compare UT_C, which includes snippet-based PRF, with UT_{noPRF}, where this component is removed. The results show marginal and inconsistent effects: UT_{noPRF} slightly outperforms UT_C in Batch 1 (30.04 vs. 27.12) and Batch 3 (20.11 vs. 20.00), while UT_C is marginally better in Batch 2 (29.86 vs. 29.21) and Batch 4

(6.95 vs. 6.26). These mixed outcomes suggest that although PRF may retrieve some additional relevant documents, its overall contribution to retrieval effectiveness is limited and dataset-dependent. To further assess the role of PRF, we compare UT_{noDF} and $UT_{noD\&P}$, both lacking document filtering but differing in the presence of PRF. $UT_{noD\&P}$, which excludes PRF, achieves slightly better MAP in Batch 1 (30.20 vs. 27.84) and Batch 3 (20.72 vs. 20.29), while UT_{noDF} performs better in Batch 2 (30.65 vs. 29.52) and Batch 4 (7.24 vs. 6.67). Although none of these differences are large, $UT_{noD\&P}$ shows slightly lower variance across batches, suggesting that snippet-based reranking may offer a more stable and reliable signal for improving document retrieval. In sum, these findings offer partial support for Q2: snippet-based PRF can yield marginal improvements in specific cases, but reranking based on generated snippet guidance appears to be a more consistent strategy across different question sets.

To investigate Q4 (Impact of Document Filtering), we assess whether applying the snippet extraction model as a hard filter improves downstream precision at the expense of recall. We begin by comparing UT_{C}, which includes document filtering, with UT_{noDF}, where this component is disabled. Across all batches, UT_{noDF} achieves higher or comparable MAP (e.g., 30.65 vs. 29.86 in Batch 2; 7.24 vs. 6.95 in Batch 4), suggesting that disabling the filter broadens the candidate pool and improves recall. However, this gain comes at the cost of reduced F1 scores in every batch (e.g., 9.33 vs. 13.03 in Batch 1), reflecting lower precision in snippet selection. A similar trend is observed when comparing UT_{noPRF} (filtering enabled) and $UT_{noD\&P}$ (filtering disabled), both of which exclude pseudo-relevance feedback. Again, $UT_{noD\&P}$ yields higher MAP in Batch 1 (30.20 vs. 30.04) and Batch 3 (20.72 vs. 20.11), but at the cost of lower F1 (e.g., 9.48 vs. 14.86 in Batch 1). This reinforces the observation that document filtering helps suppress noise and improves snippet-level precision, albeit with a risk of discarding relevant content and hurting recall. In summary, using the LLM-based snippet extractor as a hard document filter introduces a measurable trade-off: it improves precision in downstream tasks but can constrain overall retrieval effectiveness. These results highlight the importance of carefully tuning the filter's threshold or considering softer filtering strategies to balance precision and recall, depending on the needs of the end-to-end QA pipeline.

Snippet Extraction Stage. As we also participated in the snippet extraction task, we conducted an in-depth analysis of the performance of our snippet extraction pipeline within the biomedical question answering setting. Table 2 reports the results of UniTor@BioASQ systems in this task, with MAP as the primary evaluation metric, and F1-score and leaderboard rank included for additional context. Since snippet extraction operates downstream of document retrieval, its effectiveness is naturally conditioned by the quality and relevance of the retrieved documents.

We evaluate the same four UniTor@BioASQ configurations introduced in Section 3.2, allowing us to assess how different retrieval strategies impact the snippet-level evidence selection.

Main findings:

- **Retrieval quality strongly influences snippet performance**: There is a clear positive correlation between document MAP and snippet MAP. For instance, UT_{noDF} and UT_{noPRF}, which achieved strong retrieval results, also score highest in snippet MAP in Batches 2 and 3, respectively. This supports the expected relationship that broader recall at the document level increases the chance of recovering high-quality snippets.
- **High leaderboard placement despite retrieval variability**: UniTor@BioASQ systems consistently outperform the median across all batches. In Batch 1, $UT_{noD\&P}$ and UT_{noPRF} achieve a rank of 6^{th} out of 51 participants, highlighting the competitive strength of our snippet module. Even in harder test sets (e.g., Batch 4), UT_{C} and UT_{noDF} maintain a strong rank (13^{th} out of 79), despite relatively low MAP values.
- **Robustness of the snippet extractor (Q3)**: Despite fluctuations in retrieval performance, the snippet model performs stably across variants. For example, $UT_{noD\&P}$, which lacks both document filtering and PRF, achieves the highest snippet MAP in Batch 1 (27.70), surpassing more complex configurations. This robustness supports **Q3**, suggesting that the model can effectively identify relevant evidence even from noisy or incomplete input. We attribute this to our training strategy:

the snippet extractor was fine-tuned on a dataset that includes not only positive and negative examples, but also carefully selected borderline cases. These examples were designed to be ambiguous or weakly relevant, helping the model learn fine-grained distinctions. This led to improved generalization and resilience, particularly in scenarios where retrieved documents are imperfect.

- **Trade-offs between MAP and F1 persist**: As with document retrieval, configurations that omit document filtering (e.g., UT_{noDF}) often show slightly higher MAP but lower F1, reflecting reduced snippet precision. In Batch 1, UT_{noDF} matches UT_{C} in MAP (22.82) but slightly improves F1 (8.74 vs. 8.17), a marginal shift likely due to retrieval noise.

In conclusion, snippet extraction quality is closely linked to retrieval effectiveness, but our results strongly support Q3: the model remains effective even when retrieval is noisy, thanks in part to its exposure to borderline training examples. This robustness enables consistently competitive performance—e.g., reaching $6^{th}/51$ in Batch 1, and confirms the value of designing the extractor to operate reliably under imperfect upstream conditions.

Table 2
Snippet Extraction Stage: Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the MAP and F1, where the rank is ordered by MAP. Bold values represent our best submission. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3			Batch 4		
	MAP	F1	Rank	**MAP**	F1	Rank	**MAP**	F1	Rank	**MAP**	F1	Rank
UT_{C}	22.82	8.17	12/51	33.96	9.24	6/41	18.49	6.02	13/47	8.25	5.51	13/79
UT_{noDF}	22.82	8.74	12/51	**33.96**	10.32	6/41	18.49	6.56	13/47	**8.25**	5.76	13/79
UT_{noPRF}	27.70	10.81	6/51	33.15	9.77	8/41	**19.74**	6.41	10/47	5.59	3.46	16/79
$UT_{noD\&P}$	**27.70**	11.70	6/51	33.15	10.71	8/41	19.57	6.53	12/47	5.59	3.46	16/79
BS	45.35	11.86	1/51	55.22	14.21	1/41	43.22	10.98	1/47	16.34	5.60	1/79
M	10.85	6.69		20.12	9.24		9.68	5.44		2.39	1.98	

4.3. Phase A+ Results: Multi-Task Answer Generation from Retrieved Evidence

This phase evaluates the **answer generation** capabilities of our biomedical QA pipeline, based on evidence previously retrieved from PubMed. All four submitted systems share the same initial information base, documents, and snippets processed by the retrieval stages described in Section 3.2. This controlled setup allows us to isolate the effect of answer generation strategies and directly address Research Question **Q5** (*Multi-Task Answer Generation*): can a single generative model effectively handle the full spectrum of biomedical question types, yes/no, factoid, list, and ideal, within a unified system?

To explore this question systematically, we varied two key experimental factors:

- **Snippet selection policy**: Contexts provided to the answer generator were constructed prioritizing snippet diversity. Specifically, snippets were chosen from distinct documents whenever possible to minimize redundancy and promote comprehensive coverage of relevant evidence.
- **Generative model architecture**: We fine-tuned two open-source large language models, LLaMA-3 and Phi-4, on the BioASQ training data. Although both models achieved similar performance on the development set, we initially favored LLaMA-3 due to its smaller size (8B vs. 14B parameters). However, after noticing slightly weaker results for factoid and ideal questions in early evaluation batches, we switched to Phi-4 for later batches, as it demonstrated slightly better performance on these tasks during tuning.

Additionally, we assessed the sensitivity of answer generation to context size by varying the maximum number of input snippets (3, 4, or 5).

The final system configurations submitted for evaluation were:

- **UT_C^5**: Up to 5 snippets retrieved by the complete pipeline (UT_C).
- **UT_{noDF}^4**: Up to 4 snippets from retrieval without document filtering (UT_{noDF}).
- **UT_{noPRF}^4**: Up to 4 snippets from retrieval without pseudo-relevance feedback (UT_{noPRF}).
- **$UT_{noD\&P}^3$**: Up to 3 snippets from retrieval without document filtering and pseudo-relevance feedback ($UT_{noD\&P}$).

Note that the primary differences among the four system configurations concern the number and diversity of snippets used to construct the input context and the generative model chosen. In the early batches (1-2), we employed LLaMA-3 due to its lighter computational footprint, while in later batches (3-4), we switched to Phi-4 to enhance performance on factoid and ideal questions, as indicated by tuning results. This variation in model architecture is reflected in the results tables, where an asterisk (*) denotes the use of Phi-4 in Batches 3 and 4.

The following analysis breaks down system performance across the four question types, drawing insights from automatic evaluation scores and rankings across all four batches.

- **Yes/No questions.** As shown in Table 3, performance was close to the median across all batches, with minor fluctuations across system variants. While the top-1 accuracy was not reached, several configurations achieved competitive rankings (e.g., 11^{th} in Batch 4). These results suggest stable handling of binary classification, with minimal sensitivity to model architecture or snippet count. Interestingly, performance across both LLaMA-3 and Phi-4 remained consistent, reinforcing the idea that yes/no questions may be less demanding in terms of context reasoning or model capacity compared to other subtasks. This further validates the feasibility of using a single lightweight model for multi-task biomedical QA, as explored in Q5. It is also worth noting that most systems converged toward high accuracy in Batch 1, indicating strong alignment between retrieved evidence and question intent when relevant binary cues are present. However, the drop in Batch 2 suggests that robustness under domain shift or question framing variability remains a challenge, which future work could address through more targeted training or specialized prompting.

Table 3
Phase A+ Answer Generation (Yes/No): Performance of UniTor@BioASQ systems across 4 batches. The metric reported is the Accuracy. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1		Batch 2		Batch 3*		Batch 4*	
	Acc	Rank	**Acc**	Rank	**Acc**	Rank	**Acc**	Rank
UT_C^5	94.12	14/56	70.59	37/49	81.82	18/58	88.46	11/67
UT_{noDF}^4	88.24	35/56	70.59	37/49	**81.82**	18/58	**88.46**	11/67
UT_{noPRF}^4	94.12	14/56	82.35	22/59	77.27	32/58	84.62	27/67
$UT_{noD\&P}^3$	**94.12**	14/56	**82.35**	22/49	77.27	32/58	84.62	27/67
BS	100.0	1/56	100.0	1/49	95.45	1/58	92.31	1/67
M	94.12		82.35		81.82		84.62	

- **Factoid questions.**Results in Table 4 show that UniTor@BioASQ systems consistently ranked in the top 10. Notably, $UT_{noD\&P}^3$ achieved 6^{th} place in both Batch 2 and 3, confirming the effectiveness of smaller, diverse contexts for pinpointing exact answers. Improvements observed in later batches support our switch to Phi-4, which appears better calibrated for short-form extraction under factoid constraints.

Table 4
Phase A+ Answer Generation (Factoid): Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the Strict Accuracy and Mean Reciprocal Rank, where the rank is ordered by Strict Accuracy. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3*			Batch 4*		
	S-Acc	MRR	Rank	**S-Acc**	MRR	Rank	**S-Acc**	MRR	Rank	**S-Acc**	MRR	Rank
UT_C[5]	30.77	32.69	11/56	37.04	37.04	16/49	20.00	21.67	18/58	40.91	40.91	17/67
UT_{noDF}[4]	30.77	32.69	11/56	37.04	37.04	16/49	**25.00**	26.67	6/58	40.91	40.91	17/67
UT_{noPRF}[4]	30.77	30.77	11/56	44.44	44.44	6/49	15.00	15.00	25/58	45.45	45.45	7/67
$UT_{noD\&P}$[3]	**34.62**	34.62	7/56	**44.44**	44.44	6/49	25.00	25.00	6/58	**45.45**	45.45	7/67
BS	42.31	44.23	1/56	59.26	59.26	1/49	35.00	37.50	1/58	54.55	56.06	1/67
M	23.08	28.85		33.33	35.19		15.00	20.00		36.36	37.88	

- **List questions.** Table 5 shows that list QA was the most challenging subtask. Despite this, UT_{noPRF}[4] and $UT_{noD\&P}$[3] reached top-10 positions in Batches 1 and 2. The overall lower performance is likely due to the cap on input snippets (max 3–5), which restricts the system's ability to aggregate long-tail entities. While effective for precision, this limitation penalizes breadth, a key factor in list generation.

Table 5
Phase A+ Answer Generation (List): Performance of UniTor@BioASQ systems across 4 batches. The metric reported is the F1. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1		Batch 2		Batch 3*		Batch 4*	
	F1	Rank	**F1**	Rank	**F1**	Rank	**F1**	Rank
UT_C[5]	12.90	35/56	30.17	12/49	21.29	41/58	**25.32**	17/67
UT_{noDF}[4]	13.07	32/56	30.17	12/49	21.76	40/58	23.43	31/67
UT_{noPRF}[4]	**21.40**	10/56	32.96	9/49	26.21	35/58	24.98	19/67
$UT_{noD\&P}$[3]	17.20	21/56	**32.96**	9/49	**27.52**	32/58	25.03	18/67
BS	25.67	1/56	38.80	1/49	45.41	1/58	30.14	1/67
M	14.32		23.30		29.00		22.70	

- **Ideal questions.** Table 6 highlights the strongest performance of our systems. $UT_{noD\&P}$[3] achieved 1st place in Batch 1 and consistently ranked in the top 5 in all batches. These results indicate that the combination of targeted snippet selection and fine-tuned LLMs produces coherent, informative, and non-redundant long-form answers. The switch to Phi-4 in later batches appears to have improved stability and depth in this subtask, supporting its use in complex generation scenarios.

Overall, despite varying retrieval quality, the answer generation stage still shows notable robustness, especially when guided by focused snippet curation. These outcomes provide strong support for **Q5**: careful design of the input context and strategic model selection can substantially impact answer quality across tasks. Notably, our results suggest that smaller, fine-tuned models, when paired with well-structured evidence, can rival larger architectures in real-world biomedical QA.

A key goal of Q5 was to assess whether a single, shared model instance could effectively handle the full spectrum of biomedical QA subtasks, from binary classification to factoid identification, list aggregation, and long-form ideal generation. Our results confirm that this is feasible: all UniTor@BioASQ systems used the same model instance for all question types, without any task-specific specialization. Remarkably, this multi-task configuration yielded top-tier performance in several settings, including a 1st place ranking for ideal answers (Batch 1 and 3, see Table 6) and consistent top-10 placements for factoid QA.

These findings support the hypothesis that a single answer generator, when appropriately fine-tuned and guided by curated context, can perform competitively across heterogeneous biomedical QA tasks.

This is particularly compelling given the modest model sizes employed (LLaMA 8B and Phi 14B), suggesting that robust multi-task QA is achievable without resorting to extremely large-scale models, thus offering a practical path toward efficient and unified biomedical question answering systems.

Table 6
Phase A+ Answer Generation (Ideal Answer): Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the ROGUE-SU4 Recall and F1, where the rank is ordered by F1. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3*			Batch 4*		
	R-F1	R-Rec	Rank	**R-F1**	R-Rec	Rank	**R-F1**	R-Rec	Rank	**R-F1**	R-Rec	Rank
$UT_C{}^5$	20.06	22.47	2/56	20.67	23.77	6/49	19.48	18.41	7/58	**15.94**	15.82	4/67
$UT_{noDF}{}^4$	18.38	22.37	4/56	**21.17**	24.40	3/49	**20.85**	19.47	1/58	15.76	15.44	5/67
$UT_{noPRF}{}^4$	20.26	23.39	2/56	20.91	23.95	5/49	20.26	19.48	3/58	13.92	13.27	13/67
$UT_{noD\&P}{}^3$	**21.33**	24.67	1/56	21.00	24.16	4/49	19.44	18.60	8/58	14.12	13.35	11/67
BS	17.54	21.41	5/56	21.77	22.41	1/49	20.58	25.91	2/58	17.26	18.01	1/67
M	11.21	22.42		12.63	22.18		11.88	19.48		9.86	16.06	

4.4. Phase B Results: Answer Generation Leveraging Gold Snippet Context

A notable difference between Phase A+ and Phase B of the BioASQ 2025 Challenge lies in the availability of a *gold context* provided by the organizers in Phase B. Specifically, while in Phase A+ the answer generation relies solely on documents and snippets retrieved by the participant systems, in Phase B the official gold snippets can be leveraged as contextual evidence to support answer formulation. This distinction allows for a more controlled investigation of the impact of context selection strategies on answer quality, as the ground-truth relevant snippets serve as a reliable and unbiased source of information.

In this phase, we focus on evaluating various context selection policies utilizing the gold snippet lists available in Phase B. To this end, we investigate the following system configurations:

- **UT^3**: Context constructed from the first 3 snippets in the gold list.
- **UT^5**: Context constructed from the first 5 snippets in the gold list.
- **$UT_R{}^3$**: Context constructed from the top 3 snippets in the gold list, re-ranked according to similarity with the question using SentenceBERT (SBERT).
- **$UT_R{}^5$**: Context constructed from the top 5 snippets in the gold list, re-ranked according to similarity with the question using SBERT.

The rationale for applying a reranking step to the gold snippet list stems from the observation that the official order of gold snippets does not necessarily reflect their relative relevance to the given question. Therefore, for the $UT_R{}^3$ and $UT_R{}^5$ systems, we compute the cosine similarity between the embedding of the query and each gold snippet, using SentenceBERT as the embedding model, and select the highest-scoring snippets to construct the input context. This semantic reranking aims to further enhance answer quality by identifying the most pertinent pieces of evidence within the gold set.

In line with our experimental approach in Phase A+, for Batches 3 and 4 of Phase B, we adopted Phi-4 14 B as the generative model in place of LLaMA-3 8B. This choice allows us to further assess whether the increased model capacity and architectural improvements of Phi-4 result in better handling of complex biomedical terminology and nuanced relationships, potentially translating into improved answer quality when provided with optimal gold evidence.

We now report and discuss the results obtained by these systems across the four standard BioASQ question types, with particular attention to how different context selection policies interact with model architecture in shaping final answer quality.

By comparing systems utilizing the original ordering of gold snippets (UT^3, UT^5) against those employing semantic reranking ($UT_R{}^3$, $UT_R{}^5$), we can systematically assess the influence of context prioritization on the performance of biomedical question answering. This experimental setup enables a direct investigation into how different context selection and ranking strategies affect answer quality when reliable, expert-annotated evidence is available. The results of this analysis provide valuable insights regarding the role of evidence ordering and semantic relevance in supporting answer generation and may inform the design of future context construction pipelines for biomedical QA systems.

The following analysis breaks down system performance across the four question types, drawing insights from automatic evaluation scores and rankings across all four batches.

- **Yes/No Questions.** As expected, we observe a general increase in accuracy across all batches when models are provided with the gold snippet context in Phase B, compared to when answers are generated from retrieved evidence in Phase A+. For instance, considering our best submission in Batch 4, accuracy improves from 88.46% (Table 3) to 96.15% (Table 7) when switching from retrieved to gold evidence. When comparing different context selection strategies, we find that, for yes/no questions, the choice of input snippets, whether based on the original gold ordering or semantic reranking, and whether using three or five snippets, has negligible impact on performance. This result is intuitive, as binary questions often require only a single, clearly relevant snippet to provide sufficient information for the correct answer. Accordingly, accuracy is virtually identical across all tested context selection policies (UT^3, UT^5, $UT_R{}^3$, $UT_R{}^5$), with only minimal variation observed between them.

Table 7
Phase B Answer Generation (Yes/No): Performance of UniTor@BioASQ systems across 4 batches. The metric reported is the Accuracy. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1		Batch 2		Batch 3*		Batch 4*	
	Acc	Rank	**Acc**	Rank	**Acc**	Rank	**Acc**	Rank
UT^3	94.12	23/72	94.12	22/72	86.36	42/66	92.31	21/79
UT^5	94.12	23/72	94.12	22/72	86.36	42/66	96.15	3/79
$UT_R{}^3$	94.12	23/72	94.12	22/72	86.36	42/66	96.15	3/79
$UT_R{}^5$	**94.12**	23/72	**94.12**	22/72	**86.36**	42/66	**96.15**	3/79
BS	100.0	1/72	100.0	1/72	95.45	1/66	100.0	1/79
M	94.12		94.12		88.01		92.31	

- **Factoid Questions.** Analyzing the results for factoid questions, several observations can be made. As with yes/no questions, systems generally benefit from the gold context: in every batch, the Phase B systems (Table 8) outperform their Phase A+ counterparts (Table 4) by at least 10 points in terms of Strict Accuracy. Regarding the snippet selection mechanism, it is evident that choosing snippets semantically closest to the question is the most effective strategy, leading to clear benefits, particularly in the first two batches. Notably, in Batch 2, the system $\mathbf{UT}_R{}^5$ (based on LLaMA-3 and leveraging the top 5 most similar snippets to the question as context) achieved first place among 72 participating systems according to both Strict Accuracy and Mean Reciprocal Rank metrics. This result highlights how appropriate context selection, combined with careful fine-tuning of the LLM, enables the model to effectively suppress background noise and accurately identify the precise entity that answers the factoid question.

Table 8
Phase B Answer Generation (Factoid): Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the Strict Accuracy and Mean Reciprocal Rank, where the rank is ordered by Strict Accuracy. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3*			Batch 4*		
	S-Acc	MRR	Rank	**S-Acc**	MRR	Rank	**S-Acc**	MRR	Rank	**S-Acc**	MRR	Rank
UT^3	42.31	44.23	21/72	66.67	66.67	2/72	40.00	40.00	3/66	54.55	56.82	6/79
UT^5	38.46	40.38	37/72	66.67	66.67	2/72	**40.00**	42.50	3/66	54.55	56.82	6/79
$UT_R{}^3$	46.15	48.08	12/72	70.37	70.37	1/72	35.00	35.00	14/66	54.55	56.82	6/79
$UT_R{}^5$	**46.15**	48.08	12/72	**70.37**	70.37	1/72	40.00	40.00	3/66	**54.55**	56.82	6/79
BS	53.85	59.62	1/72	66.67	66.67	3/72	45.00	50.43	1/66	63.64	63.64	1/79
M	40.39	44.23		50.00	51.92		25.38	29.11		45.45	47.73	

- **List Questions.** Turning now to the results for list-type questions, where the system must enumerate a set of relevant biomedical entities for a given query, we observe a clear boost in F1 score in Phase B (Table 9) compared to Phase A+ (Table 5). For instance, our best-performing systems in Phase A+ achieved F1 scores of 21.40, 32.96, 27.52, and 25.32 across the batches, whereas in Phase B these values improve substantially to 46.76, 44.87, 58.58, and 49.69, respectively. Unlike factoid questions, there does not appear to be a significant effect associated with the specific context selection strategy adopted, whether using the original gold snippet order or applying semantic reranking. This outcome is reasonably explained by the nature of list questions, where answers typically require entities that are distributed across multiple documents or pieces of evidence. This characteristic is reflected in our results: the upper bound of 5 snippets provided as context acts as a limiting factor, as the system is unable to benefit from additional evidence that might be necessary to generate a more exhaustive list. As such, gains from more advanced snippet selection are inherently capped by the context window, and performance remains primarily constrained by the snippet budget rather than by the specific ordering or reranking of the provided evidence.

Table 9
Phase B Answer Generation (List): Performance of UniTor@BioASQ systems across 4 batches. The metric reported is the F1. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1		Batch 2		Batch 3*		Batch 4*	
	F1	Rank	**F1**	Rank	**F1**	Rank	**F1**	Rank
UT^3	**46.76**	40/72	**44.87**	45/72	54.55	32/66	45.08	45/79
UT^5	51.09	27/72	44.87	45/72	**58.58**	18/66	**49.69**	35/79
$UT_R{}^3$	44.47	46/72	38.07	53/72	52.44	33/66	34.50	59/79
$UT_R{}^5$	44.83	45/72	38.07	53/72	56.01	24/66	41.33	49/79
BS	59.59	1/72	61.52	1/72	63.37	1/66	64.92	1/79
M	47.23		48.77		47.24		47.17	

- **Ideal Questions.** For ideal answer generation, we again observe a strong positive impact from using expert-annotated gold snippets as context, with a substantial boost in ROGUE-SU4 scores, exceeding 10% improvement in each batch compared to Phase A+ results. In this subtask, the specific snippet selection policy (original gold ordering vs. semantic reranking, three vs. five snippets) does not yield any consistent measurable gains, as evident from Table 10. All variants perform comparably, suggesting that, when the context is already curated and relevant, further reordering has little effect on long-form answer generation quality. Across all batches, our systems consistently achieve top-5 positions in ROGUE-SU4 F1, maintaining the trend observed

in Phase A+. Notably, in Batch 2 and Batch 3, the $UT_R{}^5$ and $UT_R{}^3$ systems, based on LLaMA-3 and Phi-4, respectively, achieve second place out of all submissions. These results highlight both the robustness of our answer generation pipeline and the key role of gold evidence in supporting the production of comprehensive, high-quality biomedical summaries.

Table 10
Phase B Answer Generation (Ideal Answer): Performance of UniTor@BioASQ systems across 4 batches. The metrics reported are the ROGUE-SU4 Recall and F1, where the rank is ordered by F1. Bold values represent our best submission. Star (*) indicates the batch in which the Phi-4 model is used, as a base model, instead of LLaMA-3. BS represents the Best competitor System in the challenge leaderboard, while M indicates the median.

Sys	Batch 1			Batch 2			Batch 3*			Batch 4*		
	R-F1	R-Rec	Rank	**R-F1**	R-Rec	Rank	**R-F1**	R-Rec	Rank	**R-F1**	R-Rec	Rank
UT^3	36.39	39.75	5/72	38.22	40.07	3/72	32.28	31.12	7/66	29.11	27.56	10/79
UT^5	**37.60**	41.72	3/72	38.22	40.07	3/72	33.59	32.99	4/66	**30.10**	29.64	8/79
$UT_R{}^3$	35.26	39.92	7/72	38.57	40.43	2/72	**34.01**	32.77	2/66	28.34	27.41	12/79
$UT_R{}^5$	36.37	41.74	6/72	**38.57**	40.43	2/72	33.86	33.39	3/66	29.83	29.34	9/79
BS	40.08	42.95	1/72	42.87	46.05	1/72	34.39	33.83	1/66	35.15	37.25	1/79
M	18.36	31.03		21.11	30.06		19.46	28.27		19.41	24.49	

4.5. Key Insights

The results of our participation in the BioASQ13b 2025 challenge provide a comprehensive answer to the research questions driving this work, while highlighting several broader insights into modular biomedical QA system design:

- **Document Retrieval and Evidence Selection (Q1, Q2, Q4):** Our ablation studies reveal that combining semantic reranking (via LLM-generated synthetic snippets, Q1) and snippet-based pseudo-relevance feedback (Q2) produces moderate but reliable gains in retrieval effectiveness. However, aggressive document filtering (Q4) increases snippet-level precision at the expense of recall, indicating that softer or more adaptive filtering strategies may offer better trade-offs in future systems.
- **Robustness in Snippet Extraction (Q3):** The snippet extraction module consistently demonstrated high robustness and competitive MAP and F1 scores, even with noisy upstream retrieval. We attribute this to a fine-tuning regime that included not only positive and negative samples but also carefully selected *borderline* cases. This enabled the extractor to learn fine-grained distinctions and reliably filter relevant evidence, maintaining strong leaderboard placement regardless of retrieval variability.
- **Unified and Effective Answer Generation (Q5):** Our answer generator, based on a single compact LLM, performed competitively across all biomedical question types (yes/no, factoid, list, ideal) in a unified multi-task setting. When provided with retrieved evidence (Phase A+), the system delivered robust results, especially for factoid and ideal answers, though list QA remained bottlenecked by the snippet budget and retrieval coverage. With gold-standard context (Phase B), answer quality improved substantially, leading to multiple top-5 and even $1^{st}/2^{nd}$ place rankings, particularly in factoid and ideal subtasks. Notably, semantic reranking of snippets was crucial for factoid questions, but less so for yes/no and ideal answers, suggesting diminishing returns from reordering when context is already expertly curated.

Take-home messages:

- Semantic reranking and feedback (Q1, Q2) add value, but retrieval and aggregation remain the primary bottlenecks.

- Robust, evidence-centric QA requires snippet extractors trained on ambiguous/borderline cases (Q3).
- The precision–recall trade-off in document filtering (Q4) should be dynamically calibrated to task needs.
- Multi-task answer generation is feasible with compact LLMs, especially when paired with high-quality, diverse evidence (Q5).
- Closing the gap between retrieved and gold-standard evidence is essential for practical, scalable biomedical QA.

Overall, these findings motivate further research into adaptive retrieval, advanced context construction, and robust answer generation. By systematically grounding our pipeline design and evaluation in the research questions, we provide both empirical evidence and practical guidance for future unified biomedical QA systems, bringing the field closer to reliable, real-world deployment at scale.

5. Conclusions

This paper presented **UniTor@BioASQ**, a biomedical QA system designed to address several open research questions on robust evidence retrieval and unified answer generation in BioASQ-style tasks. Our approach was guided by the need to (i) understand the limits and benefits of semantic reranking using synthetic snippets, (ii) assess the impact of different evidence aggregation strategies, and (iii) test whether a single generative model can effectively handle the full spectrum of biomedical question types.

To the best of our knowledge, we are the first to systematically exploit LLM-generated synthetic snippets as semantic anchors in document reranking for biomedical QA. This hypothesis-driven reranking proved efficient, yielding consistent gains of 1–3 MAP points over standard baselines and improving recall without introducing prohibitive noise.

Across the 2025 BioASQ13b challenge, UniTor@BioASQ achieved its best results in snippet extraction and ideal answer generation. Our snippet module was especially robust: in every batch, at least one configuration ranked in the top five, even when upstream retrieval was imperfect. We also reached the top five for ideal answer generation in all batches, and attained **first place** in two out of four batches. In the factoid subtask, our best systems placed in the top five in two batches and ranked in the top ten overall. These results confirm that, with careful evidence selection, even compact open-source LLMs can deliver state-of-the-art generation quality on complex biomedical questions.

Performance on yes/no and list questions was generally close to the challenge median, with main bottlenecks traced to evidence recall (for lists) and question framing (for yes/no). When provided with gold-standard snippets (Phase B), answer quality improved significantly for factoid and ideal answers, validating the centrality of evidence quality. Interestingly, semantic reranking of the gold snippets led to further measurable improvements for factoid QA, but had less effect on yes/no and ideal answers.

In summary, UniTor@BioASQ results suggest that (i) LLM-generated snippet hypotheses are an effective tool for semantic reranking in retrieval, (ii) robust snippet extraction can reliably filter noisy evidence, and (iii) a single answer generator can compete across diverse biomedical QA tasks. The main limitation remains evidence recall in retrieval, especially for list questions. Future work will focus on adaptive retrieval strategies and larger, more diverse context aggregation to further close the gap to the best systems.

Acknowledgments

We acknowledge financial support from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and support from Project ECS 0000024 Rome Technopole - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

Declaration on Generative AI

Parts of the writing and editing process for this manuscript were supported by the use of generative AI tools, specifically OpenAI's ChatGPT. These tools were employed to enhance clarity and correct spelling. All AI-assisted content was carefully reviewed and validated by the authors to ensure accuracy, originality, and compliance with ethical and scientific standards. The authors bear full responsibility for the final content.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, accessed: 2024-11-15.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[4] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: http://dx.doi.org/10.1145/3703155. doi:10.1145/3703155.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[6] M. Li, H. Kilicoglu, H. Xu, R. Zhang, Biomedrag: A retrieval augmented large language model for biomedicine, 2024. URL: https://arxiv.org/abs/2405.00465. arXiv:2405.00465.

[7] S. Kim, Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation, 2025. URL: https://arxiv.org/abs/2502.03004. arXiv:2502.03004.

[8] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. Maria Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[9] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 13b and Synergy13 in CLEF2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[10] M. Lesavourey, G. Hubert, Enhancing Biomedical Document Ranking with Domain Knowledge Incorporation in a Multi-Stage Retrieval Approach., in: 12th BioASQ Workshop at CLEF 2024, volume 3740, Grenoble, France, 2024. URL: https://hal.science/hal-04744454.

[11] J. H. Merker, A. Bondarenko, M. Hagen, A. Viehweger, Mibi at bioasq 2024: retrieval-augmented generation for answering biomedical questions, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, volume 3740, 2024, pp. 176–187.

[12] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou,

D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. URL: https://arxiv.org/abs/2206.07682. arXiv:2206.07682.

[13] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL: https://arxiv.org/abs/2303.12712. arXiv:2303.12712.

[14] F. Borazio, C. D. Hromei, E. Passone, D. Croce, R. Basili, Mm-iglu-it: Multi-modal interactive grounded language understanding in italian, in: A. Artale, G. Cortellessa, M. Montali (Eds.), AIxIA 2024 – Advances in Artificial Intelligence, Springer Nature Switzerland, Cham, 2025, pp. 64–78.

[15] F. Borazio, D. Croce, G. Gambosi, R. Basili, D. Margiotta, A. Scaiella, M. Del Manso, D. Petrone, A. Cannone, A. M. Urdiales, C. Sacco, P. Pezzotti, F. Riccardo, D. Mipatrini, F. Ferraro, S. Pilati, Semi-automatic topic discovery and classification for epidemic intelligence via large language models, in: H. Afli, H. Bouamor, C. B. Casagran, S. Ghannay (Eds.), Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 68–84. URL: https://aclanthology.org/2024.politicalnlp-1.8/.

[16] S. Ateia, U. Kruschwitz, Bioragent: A retrieval-augmented generation system for showcasing generative query expansion and domain-specific search for scientific q&a, 2024. URL: https://arxiv.org/abs/2412.12358. arXiv:2412.12358.

[17] G. Xiong, Q. Jin, Z. Lu, A. Zhang, Benchmarking retrieval-augmented generation for medicine, 2024. URL: https://arxiv.org/abs/2402.13178. arXiv:2402.13178.

[18] Y. Gao, L. Zong, Y. Li, Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation, CLEF Working Notes (2024).

[19] G. Zhang, Z. Xu, Q. Jin, F. Chen, Y. Fang, Y. Liu, J. F. Rousseau, Z. Xu, Z. Lu, C. Weng, Y. Peng, Leveraging long context in retrieval augmented language models for medical question answering, NPJ Digital Medicine 8 (2025) 239. URL: https://doi.org/10.1038/s41746-025-01651-w. doi:10.1038/s41746-025-01651-w.

[20] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, Bioinformatics 39 (2023). URL: http://dx.doi.org/10.1093/bioinformatics/btad651. doi:10.1093/bioinformatics/btad651.

[21] L. Stuhlmann, M. A. Saxer, J. Fürst, Efficient and reproducible biomedical question answering using retrieval augmented generation, 2025. URL: https://arxiv.org/abs/2505.07917. arXiv:2505.07917.

[22] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[23] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[24] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query expansion by prompting large language models, 2023. URL: https://arxiv.org/abs/2305.03653. arXiv:2305.03653.

[25] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9414–9423. URL: https://aclanthology.org/2023.emnlp-main.585/. doi:10.18653/v1/2023.emnlp-main.585.

[26] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, L. Menotti, G. Silvello, G. Paliouras, Bioasq at clef2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: Advances in Information Retrieval, Springer Nature Switzerland, Springer Nature Switzerland, Cham, 2025. URL: https://link.springer.com/chapter/10.1007/978-3-031-88720-8_61. doi:10.1007/978-3-031-88720-8_61.

[27] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://aclanthology.org/P18-2124/. doi:10.18653/

v1/P18-2124.

[28] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[29] Z. Lu, Pubmed and beyond: a survey of web tools for searching biomedical literature, Database 2011 (2011) baq036. URL: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baq036/1252270/baq036.pdf. doi:10.1093/database/baq036.

[30] T. Almeida, R. A. A. Jonker, J. A. Reis, J. R. Almeida, S. Matos, Bit.ua at bioasq 12: From retrieval to answer generation, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271772518.

[31] A. Rivas, E. Iglesias, L. Borrajo, Study of query expansion techniques and their application in the biomedical information retrieval, TheScientificWorldJournal 2014 (2014) 132158. URL: https://doi.org/10.1155/2014/132158. doi:10.1155/2014/132158.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, volume abs/2106.09685, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[33] J.-B. Excoffier, T. Roehr, A. Figueroa, J.-M. Papaioannou, K. Bressem, M. Ortala, Generalist embedding models are better at short-context clinical semantic search than specialized embedding models, 2024. URL: https://arxiv.org/abs/2401.01943. arXiv:2401.01943.

[34] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[36] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, J. Kang, Transferability of natural language inference to biomedical question answering, 2021. URL: https://arxiv.org/abs/2007.00217. arXiv:2007.00217.

[37] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, E. Chen, Large language models for generative information extraction: A survey, 2024. URL: https://arxiv.org/abs/2312.17617. arXiv:2312.17617.

[38] J. Xu, W. B. Croft, Quary expansion using local and global document analysis, SIGIR Forum 51 (2017) 168–175. URL: https://doi.org/10.1145/3130348.3130364. doi:10.1145/3130348.3130364.

[39] L. Pizzato, D. Molla Aliod, Pseudo-relevance feedback using named entities for question answering, 2006, pp. 83–90. doi:10.13140/2.1.2044.4166.

[40] F. Borazio, D. Croce, R. Basili, Adapting llms for domain-specific retrieval: A case study in nuclear safety, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 116–122.

[41] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.