# NCU-IISR: BioASQ 14b Phase A/A+, and Phase B Challenge

Yi-Chieh Hung[1], Jen-Chieh Han[2], Hsi-Chuan Hung[3] and Richard Tzong-Han Tsai[1,2,3,*]

[1]International Graduate Program in Artificial Intelligence, National Central University, Taoyuan, Taiwan
[2]Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
[3]Department of Medical Research, Cathay General Hospital, Taipei, Taiwan

**Abstract**

This paper presents the systems and submissions developed for the BioASQ 14b challenge, extending prior research on leveraging large language models (LLMs) to improve biomedical question answering (QA). The proposed architecture integrates a hybrid information retrieval pipeline (Phase A/A+, systems IR1 to IR5) with a structured answer generation and ensemble fusion engine (Phase B, comprising systems Bio26NIA, Another, Dif-C, NewM, and EHM-9). In Phase A and A+, the retrieval pipeline utilizes BM25 over the PubMed 2025 corpus followed by either direct reranking or a hybrid dense retrieval plus reranking strategy to optimize candidate document and snippet selection, delivering retrieved evidence to GPT and Gemini series models for answer generation. In Phase B, building upon our previous pipeline, systems employ either direct inference or a two-stage Chain-of-Thought inference under strict formatting instructions. Across four test batches, the few-shot strategy progressively evolved: Batches 1 and 2 utilized a static approach with curated 13b examples to benchmark performance variations across newer and older models, including GPT-5.4, o3, GPT-4o, and Gemini 3.1 Pro; Batch 3 implemented a refined static approach using a condensed set of 10 examples from the 14b training data; and Batch 4 shifted to a dynamic few-shot framework using FAISS vector search for context anchoring. These configurations are integrated via a nine-system Ensemble Hybrid Merge (`EHM-9`) mechanism using GPT-5.5 as a judge to execute synonym collapse and multi-model voting. Systematic evaluation across multiple batches investigates the performance trade-offs introduced by upstream semantic filtering, indicating that the additional dense filtering stage does not consistently outperform traditional lexical baselines. In the downstream generation phase, the evaluation reveals that LLM response structures are sensitive to prompt alignment while validating the competitive performance of our dynamic anchoring configuration. These findings indicate that the proposed progressive few-shot adaptations and multi-model ensemble approach improve structural stability across factoid, list, and polar answer taxonomies, providing a grounded method for biomedical QA.

**Keywords**

Biomedical Question Answering, Retrieval-Augmented Generation (RAG), BM25, Dense Semantic Retrieval, Dynamic Few-Shot, Multi-Model Ensemble Fusion, Large Language Models (LLMs), BioASQ

## 1. Introduction

The BioASQ challenge [1] has long stood as a foundational benchmark for biomedical semantic indexing and question answering (QA), driving advancements that address the critical divide between vast biomedical literature and clinical practice. In its 14th edition, Task 14b Phase B mandates that systems synthesize both concise *exact* answers and expert-level, long-form *ideal* answers from retrieved evidence. This dual-target evaluation requires modern QA architectures to balance structural parsing precision with high-fidelity text generation.

As illustrated in Table 1, the BioASQ 14b corpus [2] is a highly specialized resource comprising 5,729 questions—including 340 newly curated instances that demand robust cross-source evidence integration. The dataset spans four primary taxonomies, each requiring distinct structural outputs (Table 2). While summary queries require only free-text ideal answers, exact answer extraction is mandatory for *yes/no*, *factoid*, and *list* queries, imposing strict format compliance constraints on the generation pipeline where minor structural deviations can lead to parsing failures.

Although Large Language Models (LLMs) have achieved unprecedented performance in general-purpose QA [3], their direct deployment within the biomedical domain remains challenging. Biomedical

*CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany*
*Corresponding author.
emilyjay313@gmail.com (Y. Hung); joyhan@cc.ncu.edu.tw (J. Han); cgh23991@cgh.org.tw (H. Hung); thtsai@csie.ncu.edu.tw (R. T. Tsai)

**Table 1**
Distribution of question types in the BioASQ Task 14b training dataset compared to Task 13b. The 14b dataset incorporates 340 new questions to enhance domain coverage.

| Question Type | 13b (Prior) | 14b (New) | Total (14b) | Percentage |
|---|---|---|---|---|
| Factoid | 1,600 | 135 | 1,735 | 30.3% |
| Yes/No | 1,459 | 82 | 1,541 | 26.9% |
| List | 1,047 | 83 | 1,130 | 19.7% |
| Summary | 1,283 | 40 | 1,323 | 23.1% |
| **Total** | **5,389** | **340** | **5,729** | **100%** |

knowledge is highly technical, rapidly evolving, and demands rigorous factual grounding to mitigate the risk of clinical misinformation. Unlike general-domain benchmarks, specialized biomedical datasets require expert curation, necessitating systematic benchmarking against evolving gold standards.

**Table 2**
Representative examples of the four question types in BioASQ Phase B, illustrating the required formats for *exact* and *ideal* answers.

| Type | Example |
|---|---|
| **Yes/No** | **Q:** Are transcription and splicing connected?<br>**Exact:** yes<br>**Ideal:** Yes. There is strong evidence that splicing and transcription are intimately coupled in metazoans, with genome wide surveys show that most splicing occurs during transcription. Chromatin structure, RNA polymerase dynamics, and recruitment of splicing factors through the transcriptional machinery are factors that explain a role for transcription in the regulation of splicing. |
| **List** | **Q:** Which are the different isoforms of the mammalian Notch receptor?<br>**Exact:** [["Notch-1"],["Notch-2"],["Notch-3"],["Notch-4"]]<br>**Ideal:** Notch signaling is an evolutionarily conserved mechanism, used to regulate cell fate decisions. Four Notch receptors have been identified in man: Notch-1, Notch-2, Notch-3 and Notch-4. |
| **Factoid** | **Q:** Orteronel was developed for treatment of which cancer?<br>**Exact:** castration-resistant prostate cancer<br>**Ideal:** Orteronel was developed for treatment of castration-resistant prostate cancer. |
| **Summary** | **Q:** Is Hirschsprung disease a mendelian or a multifactorial disorder?<br>**Ideal:** Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are involved in the development of Hirschsprung disease. The majority of these genes was shown to be related to Mendelian syndromic forms of Hirschsprung's disease, whereas the non-Mendelian inheritance of sporadic non-syndromic Hirschsprung disease proved to be complex; involvement of multiple loci was demonstrated in a multiplicative model. |

In this work, we present a system and submission developed for the BioASQ 14b Phase B task. Extending our laboratory's prior QA pipeline [4], our framework addresses the critical bottleneck of context-evidence alignment through three technical innovations:

1. **Dynamic RAG Architecture:** A dense FAISS-based retrieval mechanism [5] that replaces static templates with query-specific, type-aware contextual anchoring.
2. **Model-Specific Prompting:** Tailored adaptation layers that leverage the unique reasoning characteristics of diverse backbones, including *o3* and *GPT-5.4/5.5*.
3. **Ensemble Hybrid Merge (`EHM-9`):** A consensus-driven framework that normalizes multi-model outputs, enhancing structural stability in factoid extraction and summarization.

Our findings demonstrate that dynamic, query-specific few-shot anchoring consistently outperforms

static templates, and that multi-model consensus serves as a robust strategy for mitigating system-specific output hallucinations in biomedical QA tasks.

## 2. Related Work

**Retrieval-Augmented Generation (RAG).** RAG has established itself as a standard architectural paradigm for improving the factual grounding of generative models [6]. By decoupling knowledge retrieval from inference-time answer generation, RAG-based systems enable large language models (LLMs) to leverage external biomedical evidence [7], which helps mitigate parametric hallucinations—a critical constraint in domain-specific tasks. Within previous BioASQ iterations, benchmarking overviews indicate a progressive shift from direct snippet utilization toward hybrid retrieval-reranking workflows [1, 8]. This evolution underscores the operational dependency where high-precision upstream retrieval serves as the prerequisite for downstream generation stability, motivating the dual-phase (Phase A and Phase B) structure analyzed in this study.

**Prompt Engineering and Reasoning.** Beyond evidence retrieval, prompt structure represents a primary mechanism for governing model response behavior [9]. Systematic prompting techniques, notably Chain-of-Thought (CoT) paradigms [10], have been shown to elicit step-by-step reasoning traces from foundation models. In specialized areas like medicine, explicit instructional constraints and contextual anchoring have been reported to improve compliance in exact-answer extraction [3, 7]. Based on these paradigms, our approach implements structured context management tailored to the empirical traits of backbones such as *o3* and *GPT-5.4*, establishing a framework to evaluate prompt alignment and structural consistency across factoid, list, and polar question taxonomies.

**Ensemble Learning in QA.** Model ensembling is frequently utilized to stabilize performance and reduce prediction variance across generative systems. For question answering tasks, integrating outputs from diverse model families can help mitigate individual architectural biases [1, 4]. Our framework builds upon this concept through the implementation of the `EHM-9` architecture, which utilizes *GPT-5.5* strictly as an administrative arbitrator. Rather than introducing additional prompt-level reasoning variables, this component is restricted to executing synonym collapse and semantic consensus voting, serving as a structural layer to address formatting non-compliance without artificially inflating model capabilities.

## 3. Methods

This section details our proposed architecture and its modular workflow, engineered for the BioASQ 14b challenge. Extending our laboratory's previous biomedical QA research [4], the pipeline is organized into two distinct yet integrated modules to ensure algorithmic clarity and experimental reproducibility.

First, Section 3.1 (Phase A/A+) details our hybrid information retrieval paradigm, which augments traditional lexical baselines with dense semantic filtering and multi-model reranker ensembles to minimize noise in candidate selection. Second, Section 3.2 (Phase B) delineates our guarded answer generation and ensemble fusion framework, which focuses on dynamic context construction, cross-generational model orchestration, and robust consensus-driven merging mechanisms. Synthesizing these stages yields a unified, end-to-end Retriever-Generator architecture that balances semantic recall with factual and structural grounding under strict clinical constraints.

### 3.1. Phase A and A Plus: Hybrid Dense Information Retrieval

Our submitted systems for BioASQ 14b Phase A and A+ are built upon our previous RAG-based framework proposed in [8]. The overall architecture remains largely unchanged, consisting of retrieval, reranking, and LLM-based answer generation components. Compared with the previous system, the

main modifications focus on the retrieval stage through the introduction of an optional dense filtering step and revised candidate selection strategies. The following sections describe the system workflow and the configurations of the submitted systems in detail.

### 3.1.1. System Overview

Our system follows the same RAG framework proposed in our previous work [8], which consists of three main components: a retrieval module, a reranker, and an LLM-based answer generation module. The overall architecture and Pipeline A/B remain unchanged. Figure 1 illustrates the extended workflow when the dense filtering stage is enabled.

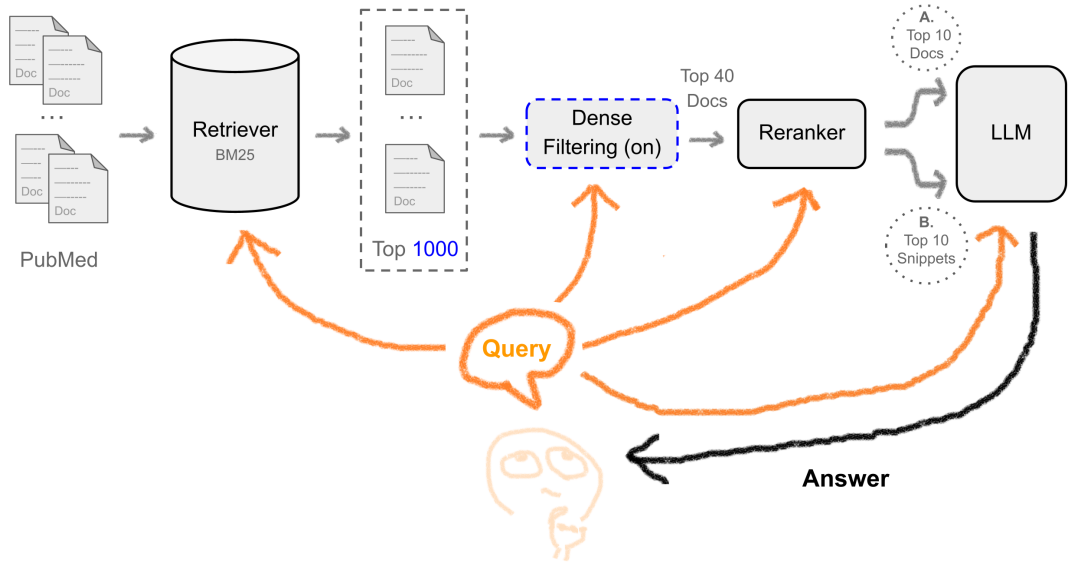


**Figure 1:** RAG workflow with the dense filtering stage enabled. An additional semantic filtering step is introduced between retrieval and reranking for candidate document selection.

Compared with the previous system, the main modification is introduced in the retrieval stage through an additional dense filtering step and adjusted candidate selection strategies. The dense filtering stage retains the top 40 semantically relevant documents for subsequent processing. This threshold was selected based on preliminary experiments on the BioASQ 13b Batch 1 documents retrieval task, where candidate sizes of 30, 40, and 50 documents were evaluated. Although the performance differences were relatively small, retaining the top 40 documents consistently achieved slightly better retrieval performance. Therefore, this setting was adopted for the current system.

For document retrieval, BM25 first retrieves the top 1000 candidate documents when dense filtering is enabled. A dense semantic encoder is then applied to generate vector representations for the input question and candidate documents. Similarity scores are computed between these representations, and the top 40 semantically relevant documents are selected for further processing. These selected documents are subsequently passed to the document reranker, which produces the final top 10 documents.

For snippet retrieval, instead of generating candidate snippets from the final top-ranked documents as in our previous system, the top 40 semantically selected documents are directly used as candidate sources. All sentences from these documents are extracted and reranked at the sentence level, and the top 10 snippets are selected as the final output.

When dense filtering is disabled, the system follows the original retrieval workflow by directly retrieving the top 100 candidate documents using BM25 before reranking.

Therefore, our system supports two retrieval configurations: the original BM25-based pipeline and an extended pipeline with an additional dense filtering stage.

The answer generation stage remains unchanged. Pipeline A utilizes the selected top-ranked documents for document-based generation, while Pipeline B uses the selected snippets for snippet-based

**Table 3**
Overview of submitted systems for BioASQ 14b Phase A.

| System | BM25 | Dense Filter | Reranker |
|---|---|---|---|
| IR1 | Top100 | – | BGE-v2-m3 |
| IR2 | Top1000 | Nomic-v1 | Jina-m0 |
| IR3 | Top1000 | BGE-M3 + Nomic-v1 + MPNet | Jina-m0 |
| IR4 | Top1000 | BGE-M3 + Nomic-v1 + MPNet | BGE-v2-m3 + Jina-m0 + MXBAI-v2 |
| IR5 | Top100 | – | BGE-v2-m3 + Jina-m0 + MXBAI-v2 |

**Table 4**
Overview of submitted systems for BioASQ 14b Phase A+. Retrieval configurations follow the corresponding Phase A systems with the same system identifiers.

| System | Pipeline | Generation Setup |
|---|---|---|
| IR1 | A | GPT-5-mini → GPT-5.4 (select) |
| IR2 | B | Chih workflow + Gemini-3.1-Pro-Preview |
| IR3 | B | Chih workflow + Gemini-3.1-Pro-Preview |
| IR4 | A | Gemini-3-Flash-Preview → Gemini-3.1-Pro-Preview (select) |
| IR5 | A | Gemini-3-Flash-Preview → Gemini-3.1-Pro-Preview (select) |

generation following the LLM generation framework proposed by Chih et al. [4]. As the framework evolved throughout the competition, Pipeline B was updated in later batches to incorporate the corresponding modifications.

### 3.1.2. Submitted Systems

Tables 3 and 4 summarize the submitted system configurations for Phase A and Phase A+, respectively. For readability, abbreviated model names are used in the submitted system tables, while full model names are provided when first introduced in the corresponding method descriptions.

For Phase A, IR1 and IR5 follow the retrieval workflow used in our previous work. IR1 adopts the same configuration as last year and employs a single reranker, `BAAI/bge-reranker-v2-m3` [11], serving as the baseline system in this study. In contrast, IR5 utilizes an ensemble of three rerankers by additionally incorporating `jinaai/jina-reranker-m0` [12] and `mixedbread-ai/mxbai-rerank-base-v2` [13], where normalized scores from the three models are aggregated.

IR2–IR4 introduce the dense filtering stage. For dense retrieval, `nomic-ai/nomic-embed-text-v1` [14] was selected because it achieved the best performance in our previous experiments on the BioASQ 13b Batch 1 snippet retrieval task. Similarly, `jinaai/jina-reranker-m0`, adopted in IR2 and IR3, showed better performance than `BAAI/bge-reranker-v2-m3` on our previous evaluation data. Furthermore, IR3 and IR4 employ an ensemble of dense models. They additionally incorporate BAAI/bge-m3 [11] and sentence-transformers/multi-qa-mpnet-base-dot-v1 [15, 16]. These models were selected based on their superior performance among different model families in our preliminary experiments, and their scores were averaged after normalization.

For Phase A+, the generation configuration was determined according to the retrieval characteristics observed during system development while maintaining diversity among the submitted systems. Rather than concentrating all systems on a single generation strategy, we aimed to balance the configurations across the two pipelines. Systems showing stronger document retrieval performance were generally paired with Pipeline A, whereas systems with better snippet retrieval performance were more likely to be paired with Pipeline B.

To keep Table 4 concise, only the primary generation models are shown, while detailed generation workflows are described in the text. For Pipeline A, the selected top 10 documents are individually processed to generate candidate responses, followed by a second-stage selection process to determine

the final answer. Since Pipeline B adopts the generation framework proposed by Chih et al. [4], detailed descriptions and updates of its workflow are deferred to the subsequent Phase B section.

## 3.2. Phase B: Guarded Answer Generation and Ensemble Fusion

### 3.2.1. Overview

Phase B focuses on generating both *exact* and *ideal* answers for BioASQ queries [1], leveraging document snippets retrieved during Phase A. Extending our laboratory's prior QA pipeline [4], this phase incorporates advanced mechanisms for dynamic retrieval, guarded ensemble fusion, and adaptive format anchoring. The core of this phase lies in the system-level integration of these modules to attain state-of-the-art stability in biomedical QA. The pipeline comprises three sequential stages: (1) dynamic few-shot context construction, (2) multi-model answer generation, and (3) ensemble fusion with guarded post-processing.

To provide a transparent, immediate visual anchor for the system architecture, Figure 2 illustrates the comprehensive workflow of the Phase B answer generation pipeline.

As mapped in Figure 2, the execution schematic details the evolutionary trajectory of the prompt configurations across the four official BioASQ test batches. It tracks the intake of Phase A retrieved snippets through three distinct few-shot context construction strategies: *Static 13b*, *Condensed new14few_shot*, and *Dynamic FAISS*. Furthermore, the diagram visualizes the strict segregation of the parallel inference tracks—standard direct generation versus multi-step Chain-of-Thought (CoT) reasoning—specifically highlighting the integration of the `NewM` configuration in Batch 4. Finally, the workflow culminates in a post-generation decision node, delineating how outputs from Batches 1 through 3 underwent standard formatting compliance, whereas Batch 4 outputs were uniquely routed through the nine-system Ensemble Hybrid Merge (`EHM-9`) mechanism, adjudicated by a *GPT-5.5* meta-arbitrator.

### 3.2.2. Modular System Architecture

To ensure experimental reproducibility, the framework adopts a decoupled, modular architecture. A central orchestration service governs task distribution and batch management, while the inference tracks are strictly segregated: direct single-pass generation and multi-step reasoning (Chain-of-Thought) are executed through independent processing logic. Model interactions are abstracted into standardized API wrappers to support diverse LLM provider schemas (e.g., OpenAI and Google). Context management is similarly modularized, with auxiliary components—such as the dynamic FAISS retriever and historical prompt repositories—encapsulated within an isolated support subsystem. This design enables rapid structural iterations between experimental batches while preserving overall pipeline integrity.

### 3.2.3. Dataset

We utilized the BioASQ Task 14b Phase B training corpus [2], comprising 5,729 curated biomedical questions—a cumulative expansion that integrates 340 new instances into the historical benchmarks. The dataset spans four primary query taxonomies, distributed as follows: *factoid* (30.3%), *yes/no* (26.9%), *summary* (23.1%), and *list* (19.7%).

This distribution reflects the diverse complexity of clinical and scientific information needs in the biomedical domain. As shown in Table 1, the addition of 14b instances ensures that our few-shot learning strategies (Section 3.2.4) are conditioned on a representative sample of biomedical inquiry. Furthermore, while our prior work [4] was constrained by token limitations, advancements in LLM context windows allowed for the integration of all retrieved snippets per query, enhancing both factual grounding and answer comprehensiveness.

### 3.2.4. Few-Shot Prompt Construction

We evaluated three strategies to optimize context-driven downstream performance [9]:

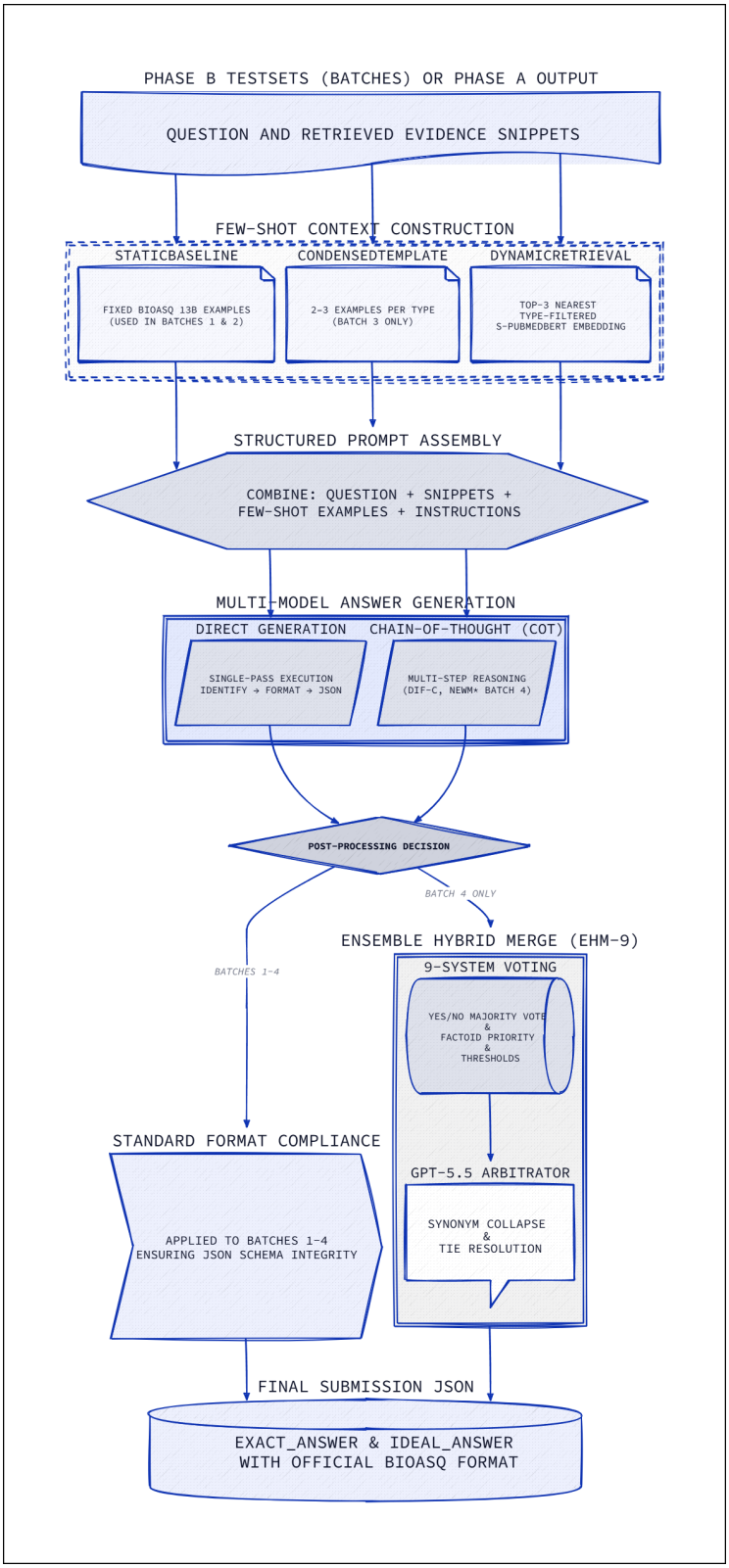


**Figure 2:** The comprehensive execution pipeline for BioASQ 14b Phase B. The flowchart delineates the integration of retrieved Phase A evidence, the batch-specific evolution of few-shot strategies, the parallel generation tracks (Direct vs. CoT), and the final ensemble fusion mechanism (EHM-9) deployed in Batch 4.

**Static 13b template.** Used in Batches 1 and 2, this baseline employed a fixed template from the BioASQ 13b training set, providing a uniform example matrix regardless of query content.

**Condensed type-aware template (`new14few_shot`).** Introduced in Batch 3, this configuration utilized a compact matrix with two to three examples per query type from the 14b pool. While this minimized the token footprint, the restricted context lacked sufficient "format anchoring" for *factoid* queries, resulting in structural anomalies (detailed in Section 4.2.2) that necessitated a shift to dynamic retrieval.

**Dynamic FAISS retrieval (`dynamic_few_shot`).** Implemented in Batch 4, this approach dynamically retrieves the top-three semantically aligned training examples at inference time to guide downstream generation. We constructed a dense similarity index over the BioASQ 14b corpus using the FAISS framework [5]. To generate high-fidelity representations, we map both target queries and historical candidate examples into a 768-dimensional vector space using the `S-PubMedBert-MS-MARCO` bi-encoder. This model leverages the PubMedBERT architecture [17], pre-trained on biomedical literature to capture specialized clinical semantics, and is further fine-tuned via the Sentence-BERT framework [15] on the MS-MARCO passage ranking dataset. During inference, we apply the question type as a strict categorical filter prior to vector similarity computation, ensuring that the retrieved few-shot context provides both high domain-relevance and robust structural format anchoring.

### 3.2.5. Multi-Model Answer Generation and Prompt Engineering

Each system pairs a specific few-shot strategy with a frontier large language model, including *GPT-4o*, *GPT-5.4*, *GPT-5.5* [18], *o3* [19], and *Gemini-3.1 Pro* [20]. The models process a structured prompt comprising the target query, retrieved evidence snippets, and selected few-shot examples. The foundational directive mandates a unified JSON-schema output, incorporating dynamic constraints based on the question taxonomy (e.g., restricting *factoid* queries to a maximum of five precise entities ordered by confidence).

Because different LLM architectures require distinct instructional guardrails to maximize their reasoning capabilities, we implemented model-specific prompt adaptations:

- **Reasoning Elicitation for *o3*:** To leverage the latent reasoning capabilities of the *o3* architecture, we introduced explicit step-by-step directives for *factoid* queries. The model was instructed to thoroughly evaluate the clinical evidence sequentially before finalizing entity names, thereby reducing premature entity commitment and improving ranking accuracy.
- **Strict Grounding for *GPT-5.4*:** To mitigate generative hallucinations and improve ROUGE scores in ideal answer synthesis, we designed a specific `<citation_rules>` block. This module enforces strict textual grounding by mandating that the model synthesize information exclusively from the provided snippets, ensuring that every major claim is implicitly traceable to the source texts.

For systems employing the Chain-of-Thought (CoT) [10] inference track (e.g., `Dif-C` and `NewM`), an explicit multi-stage reasoning protocol is prepended. The model must sequentially: (1) identify the question type, (2) logically justify the answer based on the provided snippets, and (3) format the final JSON output. These structured derivation strategies align with recent advances in biomedical QA [3, 7] by explicitly leveraging reasoning traces to mitigate domain-specific hallucinations.

To investigate the interaction between prompt-level instructions and model-level internal reasoning, this CoT track was evaluated on two distinct GPT architectures in Batch 4. In the standard configuration (`Dif-C`), the track was executed via *GPT-5.4* using traditional temperature scaling. In contrast, the standalone `NewM` system utilized *GPT-5.5*, replacing traditional temperature with an intrinsic `reasoning_effort` API parameter set to "high." This setup created a "Double CoT" scenario, deliberately overlaying our explicit step-by-step reasoning prompt onto the native internal reasoning pathways of *GPT-5.5*.

It is important to note that this specialized `reasoning_effort` configuration was restricted solely to the independent generation track of `NewM`. Within our ensemble framework (`EHM-9`), *GPT-5.5* functioned strictly as an administrative meta-arbitrator for multi-model voting. In this role, it operated without prompt-level reasoning interventions to avoid over-constraining the consensus logic.

#### 3.2.6. Ensemble Hybrid Merge (EHM-9)

The `EHM-9` system aggregates outputs from nine configurations—five *Static*, two *Condensed*, and two *Dynamic* systems—consolidated via a voting mechanism with *GPT-5.5* as the meta-arbitrator.

**Logic:** Yes/No queries follow a majority vote. Factoid queries prioritize candidates based on the hierarchy (dynamic $\succ$ condensed $\succ$ static). List queries require $\geq 2$ system consensus ($\geq 3$ for Gemini-exclusive outputs to mitigate over-retrieval). Summary answers are extracted directly from the top-performing configuration.

#### 3.2.7. Format Compliance and Post-Processing

A dedicated post-processing script enforces the required nested array structures (e.g., `[["entity"]]`) for *factoid* and *list* queries. This validation step resolves sporadic formatting anomalies—specifically the triple-nested array structures (e.g., `[[["entity"]]]`)—observed under highly compressed few-shot conditions, which otherwise would trigger parsing failures and null scores (as evidenced in Section 4.2.1).

## 4. Results

This section presents an empirical evaluation of our biomedical question-answering (QA) framework against the official BioASQ 14b benchmarks. To facilitate a rigorous and modular analysis, we partition our findings into two primary tracks corresponding to the retrieval and generation phases of the pipeline.

First, Section 4.1 (Phase A/A+) evaluates the performance of the upstream information retrieval modules, analyzing the stability and retrieval gains achieved by integrating dense semantic filtering layers and multi-model reranker ensembles with standard lexical baselines. Second, Section 4.2 (Phase B) provides an investigation into downstream prompt-based answer generation. This track systematically assesses the efficacy of dynamic few-shot contextual anchoring, cross-generational LLM variations, and consensus-driven ensemble fusion mechanisms (as detailed in Section 3.2) across both *exact* and *ideal* answer categories.

Building upon our laboratory's prior QA research [4], these complementary evaluations characterize the end-to-end operational dependencies and performance trajectories of our integrated architecture. By triangulating results from these distinct tracks, we demonstrate how our pipeline maintains semantic precision and structural stability under the constraints of the BioASQ 14b challenge.

### 4.1. Evaluation on Phase A / A+

This section presents the preliminary results of our submissions for BioASQ Task 14b. The final official results will be released in September following manual evaluation by BioASQ experts and the possible enrichment of the ground truth with additional correct answers. Therefore, rankings and final scores are not reported in this paper, and the current results are intended for reference only.

Since our system was developed concurrently with the competition timeline, several components and configurations were introduced progressively throughout the competition period. Consequently, the submitted systems varied across batches. We submitted four out of five configurations in Batch 3, while all five configurations became available from Batch 4 onward.

**Table 5**
Results of our submissions for BioASQ 14b Phase A. Only MAP is reported. The dense rank for each MAP score is shown below the corresponding value. From Batch 3 onward, the highest score among the submitted systems for each retrieval setting is shown in bold.

| System | Batch 1 | | Batch 2 | | Batch 3 | | Batch 4 | |
|---|---|---|---|---|---|---|---|---|
| | Documents | Snippets | Documents | Snippets | Documents | Snippets | Documents | Snippets |
| Best | 0.2835 | 0.1671 | 0.2395 | 0.2108 | 0.2114 | 0.1440 | 0.2310 | 0.1750 |
| IR1 | 0.2349 (10/45) | 0.1656 (2/38) | 0.1881 (15/63) | 0.1494 (8/51) | 0.1404 (23/59) | 0.0469 (28/48) | 0.1955 (12/66) | 0.0817 (23/61) |
| IR2 | - | - | - | - | 0.1639 (15/59) | 0.0794 (17/48) | 0.1973 (9/66) | 0.0887 (18/61) |
| IR3 | - | - | - | - | **0.1689** (12/59) | **0.0892** (12/48) | 0.1917 (15/66) | 0.0876 (19/61) |
| IR4 | - | - | - | - | - | - | 0.1998 (7/66) | 0.0848 (21/61) |
| IR5 | - | - | - | - | 0.1677 (13/59) | 0.0666 (19/48) | **0.2025** (5/66) | **0.0970** (15/61) |
| Median | 0.1650 | 0.0542 | 0.1386 | 0.0543 | 0.1253 | 0.0359 | 0.1598 | 0.0452 |

#### 4.1.1. Phase A results

Table 5 summarizes the retrieval performance of our submitted systems across different batches, allowing comparisons between different retrieval and reranking configurations.

Compared with the previous year's findings, introducing the additional dense filtering stage did not consistently improve document retrieval performance. In several batches, systems using the original BM25-based retrieval pipeline achieved comparable or even better MAP scores than systems incorporating dense filtering. One possible explanation is that semantic filtering may discard lexically relevant documents retrieved by BM25 or reduce candidate diversity before reranking. These observations also suggest that the reranking stage may plays a more critical role in retrieval performance than the additional dense filtering stage.

Consistent with our previous observations, reranker ensembles did not demonstrate a clear advantage over single reranker configurations. Although the ensemble-based system slightly improved document retrieval in some batches, the improvement was not consistent across document and snippet retrieval tasks.

#### 4.1.2. Phase A+ results

Table 6 summarizes the performance of our submitted systems on exact and ideal answer generation, enabling comparisons across different retrieval pipelines and generation strategies.

For answer generation, systems using Pipeline B consistently achieved stronger performance on ideal answers. In both Batch 3 and Batch 4, IR3 obtained the highest dense rank among all submitted systems. This result suggests that integrating the snippet-based generation workflow remained beneficial for generating long-form answers.

During the competition period, Pipeline B adopted the LLM generation workflow from Chih et al. [4]. Since the Phase A+ schedule preceded the corresponding Phase B batches, the adopted workflow lagged behind the latest Phase B configurations. Specifically, the Batch 3 submissions used the same five-shot setup as the configurations employed in Phase B Batch 1 and Batch 2, while Batch 4 adopted the updated few-shot strategy introduced in the Phase B Batch 3 configuration.

**Table 6**
The results of the submissions for BioASQ 14b Phase A+. The table below presents the evaluation results for both exact and ideal answers. Only the primary evaluation metrics are reported here; for the complete set of scores, please refer to the official leaderboard. Each evaluation score is accompanied by its dense rank, shown directly below the value in the table. Bolded values indicate the best-performing system submitted by our team for question type from Batch 3 onward.

| Batch | Answer Type | Q Type | Evaluation | System | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Best | IR1 | IR2 | IR3 | IR4 | IR5 | Median |
| 1 | Exact | Yes/No | Macro F1 | 1.0000 | - | - | - | - | - | 0.8786 |
| | | Factoid | MRR | 0.4783 | - | - | - | - | - | 0.2609 |
| | | List | F-Measure | 0.2345 | - | - | - | - | - | 0.1527 |
| | Ideal | all | R-SU4 (F1) | 0.1821 | - | - | - | - | - | 0.1207 |
| 2 | Exact | Yes/No | Macro F1 | 0.9481 | 0.7981 (9/16) | - | - | - | - | 0.8444 |
| | | Factoid | MRR | 0.4000 | 0.3000 (8/26) | - | - | - | - | 0.2500 |
| | | List | F-Measure | 0.3909 | 0.3280 (10/57) | - | - | - | - | 0.2371 |
| | Ideal | all | R-SU4 (F1) | 0.1851 | 0.1205 (38/59) | - | - | - | - | 0.1349 |
| 3 | Exact | Yes/No | Macro F1 | 1.0000 | 0.7179 (6/13) | **1.0000** (1/13) | 0.8952 (2/13) | - | 0.8952 (2/13) | 0.8952 |
| | | Factoid | MRR | 0.5294 | 0.2353 (17/22) | 0.2647 (16/22) | 0.2647 (16/22) | - | **0.3529** (11/22) | 0.3235 |
| | | List | F-Measure | 0.3258 | 0.2290 (35/60) | 0.2604 (22/60) | **0.2724** (15/60) | - | 0.2486 (25/60) | 0.2358 |
| | Ideal | all | R-SU4 (F1) | 0.1537 | 0.0910 (52/60) | 0.1373 (12/60) | **0.1537** (1/60) | - | 0.1277 (23/60) | 0.1236 |
| 4 | Exact | Yes/No | Macro F1 | 0.9352 | **0.8730** (2/10) | 0.8667 (3/10) | 0.8667 (3/10) | 0.8667 (3/10) | 0.8667 (3/10) | 0.8667 |
| | | Factoid | MRR | 0.4091 | 0.0909 (12/13) | **0.3636** (2/13) | **0.3636** (2/13) | 0.2727 (5/13) | 0.2727 (5/13) | 0.2273 |
| | | List | F-Measure | 0.4840 | 0.2667 (37/61) | **0.3480** (19/61) | 0.3347 (24/61) | 0.2677 (36/61) | 0.2374 (45/61) | 0.3065 |
| | Ideal | all | R-SU4 (F1) | 0.1699 | 0.1169 (38/58) | 0.1655 (3/58) | **0.1699** (1/58) | 0.1422 (17/58) | 0.1566 (8/58) | 0.1319 |

## 4.2. Experimental Results and Discussion in Phase B

### 4.2.1. Overview of Evaluation Results

To provide a clear context for interpreting the empirical results, the evaluated system tracks were designed with distinct behavioral goals that progressively extend our baseline pipeline [4]. Specifically: (1) `Another` serves to evaluate frontier model performance (OpenAI *o3* and *Gemini-3.1 Pro*) under evolving contextual constraints; (2) `Bio26NIA` focuses on validating standard generation stability using various base GPT architectures; and (3) `Dif-C` and `NewM` explicitly implement Chain-of-Thought (CoT) inference tracks to enforce structured logical derivation and mitigate biomedical hallucinations. Crucially, to safeguard data structural integrity against generative anomalies under highly compressed prompts, these configurations introduce strict API-level JSON schema enforcement and rigid array-nesting post-validation (detailed in Sections 3.2.5 and 3.2.7) before scoring.

Tables 7 and 8 present the comprehensive evaluation results for exact and ideal answers across all four batches. For clarity, only official submission system names are listed; the underlying language models, generation strategies, and few-shot configurations are detailed in the batch-by-batch analysis below. Within each table, the highest score per metric is highlighted in bold. To complement these results, three focused comparison tables are provided in Section 4.2.3: a cross-batch List F-Measure progression (Table 9), a cross-batch ROUGE-SU4 F1 progression (Table 10), and a direct comparison of the two CoT systems in Batch 4 (Table 11).

Performance variation was observed across the evaluated systems, particularly in list-type questions and ideal-answer generation. In Batch 4, all systems demonstrated a marked improvement in List F-Measure, with gains of approximately +0.15 to +0.20 relative to Batch 3. This trajectory, alongside distinct patterns in factoid retrieval and summarisation quality, motivates the system-level analysis in Section 4.2.3.

**Table 7**
Exact answer evaluation results on BioASQ 14b Phase B. Evaluation metrics include Accuracy (Acc) and Macro F1 (maF1) for Yes/No questions; Strict Accuracy (SAcc), Lenient Accuracy (LAcc), and Mean Reciprocal Rank (MRR) for Factoid questions; and Mean Precision, Recall, and F-Measure (F1) for List questions. Dashes (—) indicate a formatting failure that produced no parseable score. Bold values denote the highest score within each batch.

| Batch | System | Yes/No | | Factoid | | | List | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | maF1 | SAcc | LAcc | MRR | Precision | Recall | F1 |
| Batch-1 | Another | **1.0000** | **1.0000** | 0.3478 | 0.3478 | 0.3478 | **0.2832** | 0.2345 | 0.2443 |
| | Bio26NIA | 0.9412 | 0.9377 | **0.4348** | **0.4783** | **0.4565** | 0.2822 | **0.3067** | **0.2866** |
| | Dif-C | 0.9412 | 0.9377 | 0.3913 | 0.4348 | 0.4058 | 0.2410 | 0.2216 | 0.2259 |
| Batch-2 | Bio26NIA | 0.9048 | 0.8929 | **0.3500** | **0.3500** | **0.3500** | **0.4830** | 0.4544 | **0.4605** |
| | Dif-C | 0.9048 | 0.8929 | **0.3500** | **0.3500** | **0.3500** | 0.4663 | **0.4644** | 0.4596 |
| | Another | **0.9524** | **0.9443** | 0.3000 | 0.3000 | 0.3000 | 0.4636 | 0.4640 | 0.4583 |
| Batch-3 | Another | **1.0000** | **1.0000** | **0.4118** | **0.4706** | **0.4412** | **0.4697** | **0.4823** | **0.4675** |
| | Dif-C | 0.9091 | 0.8952 | 0.3529 | 0.4118 | 0.3824 | 0.4092 | 0.4032 | 0.3971 |
| | Bio26NIA | **1.0000** | **1.0000** | — | — | — | 0.4119 | 0.4333 | 0.4150 |
| Batch-4 | Another | **0.8750** | **0.8667** | **0.4545** | **0.4545** | **0.4545** | **0.6353** | **0.6755** | **0.6324** |
| | EHM-9 | **0.8750** | **0.8667** | 0.3636 | 0.3636 | 0.3636 | 0.5908 | 0.6663 | 0.6030 |
| | Dif-C | **0.8750** | **0.8667** | 0.3636 | 0.3636 | 0.3636 | 0.5916 | 0.6275 | 0.5932 |
| | Bio26NIA | **0.8750** | **0.8667** | 0.3636 | 0.3636 | 0.3636 | 0.5798 | 0.6299 | 0.5884 |
| | NewM | **0.8750** | **0.8667** | 0.2727 | 0.3636 | 0.3182 | 0.5319 | 0.5610 | 0.5301 |

### 4.2.2. Batch-by-Batch Performance Analysis

**Batch 1 Performance Analysis** Batch 1 established a foundational performance baseline by applying a unified static few-shot template derived from the BioASQ 13b configuration across all three systems.

**Table 8**
Ideal answer evaluation results on BioASQ 14b Phase B, measured using ROUGE-2 Recall (R-2 Rec), ROUGE-2 F-Measure (R-2 F1), ROUGE-SU4 Recall (R-SU4 Rec), and ROUGE-SU4 F-Measure (R-SU4 F1). Bold values denote the highest score within each batch.

| Batch | System | R-2 (Rec) | R-2 (F1) | R-SU4 (Rec) | R-SU4 (F1) |
|---|---|---|---|---|---|
| Batch-1 | Dif-C | **0.2814** | **0.2487** | **0.2776** | **0.2381** |
| | Bio26NIA | 0.2642 | 0.2320 | 0.2734 | 0.2315 |
| | Another | 0.1617 | 0.1317 | 0.1716 | 0.1348 |
| Batch-2 | Dif-C | 0.2642 | **0.2351** | 0.2574 | 0.2255 |
| | Another | **0.2758** | 0.2334 | **0.2726** | **0.2258** |
| | Bio26NIA | 0.2742 | 0.2239 | 0.2685 | 0.2139 |
| Batch-3 | Another | **0.2489** | **0.2497** | 0.2447 | **0.2428** |
| | Bio26NIA | 0.2480 | 0.2436 | **0.2441** | 0.2371 |
| | Dif-C | 0.2425 | 0.2376 | 0.2336 | 0.2276 |
| Batch-4 | Another | **0.3562** | **0.2827** | **0.3422** | **0.2699** |
| | EHM-9 | 0.3205 | 0.2656 | 0.3100 | 0.2544 |
| | Dif-C | 0.2611 | 0.2455 | 0.2524 | 0.2337 |
| | Bio26NIA | 0.2876 | 0.2343 | 0.2781 | 0.2247 |
| | NewM | 0.2404 | 0.2312 | 0.2275 | 0.2196 |

`Another` employed the OpenAI *o3* model, `Bio26NIA` used *GPT-5.4*, and `Dif-C` applied a multi-stage Chain-of-Thought (CoT) generation strategy also built on *GPT-5.4*. As shown in Table 7, `Another` achieved perfect Yes/No Accuracy and Macro F1 (1.0000), whereas `Bio26NIA` led in all Factoid metrics (MRR: 0.4565) and List F-Measure (0.2866), reflecting a stronger capacity for entity-level extraction under the 13b static template. In ideal answer generation (Table 8), `Dif-C` obtained the highest ROUGE scores across all four metrics, suggesting that the CoT generation strategy produced more information-dense summaries within the same static prompting framework.

**Batch 2 Performance Analysis** Batch 2 evaluated the static prompting framework on updated model architectures. `Bio26NIA` and `Another` were transitioned to *GPT-4o* and *Gemini-3.1 Pro*, respectively, while `Dif-C` retained its *GPT-5.4 CoT* configuration. All three systems continued to use the static 13b few-shot template. `Another` achieved the highest Yes/No Accuracy (0.9524), while `Bio26NIA` and `Dif-C` tied across all Factoid metrics (MRR: 0.3500), with `Bio26NIA` narrowly leading in List F-Measure (0.4605 vs. 0.4596). For ideal answers, `Dif-C` led in ROUGE-2 F1 (0.2351) and `Another` in ROUGE-2 Recall (0.2758) and ROUGE-SU4 F1 (0.2258), indicating broadly comparable summarisation quality across all three systems.

**Batch 3 Performance Analysis** Batch 3 introduced a structural modification to examine the impact of extreme context compression on generation quality. `Another` and `Dif-C` retained their Batch 2 configurations (*Gemini-3.1 Pro* and *GPT-5.4 CoT* with the 13b static template). By contrast, `Bio26NIA` was reconfigured with a highly condensed, type-aware few-shot matrix (`new14few_shot`), which provided only two to three question-answer examples per question type, sampled from a newly curated 14b training pool.

`Another` achieved the top scores across all exact-answer metrics, including a Factoid MRR of 0.4412 and a List F-Measure of 0.4675. Conversely, the compressed `new14few_shot` configuration in `Bio26NIA` proved insufficient for robust structural format anchoring. While two to three examples are generally adequate for modern large language models to capture semantic intent, this minimal context lacked the specific format exemplars required to reliably enforce the required two-dimensional array structure in *factoid* queries. As a result, the model exhibited structural over-modification, erroneously producing unparseable three-dimensional array outputs (e.g., `[[["entity"]]]` in place of the required `[["entity"]]`). This formatting anomaly triggered parsing failures within the evaluation script, re-

sulting in the null *factoid* scores denoted by dashes in Table 7. These findings provided the empirical rationale for transitioning to the dynamically anchored framework in Batch 4.

**Batch 4 Performance Analysis** Batch 4 represented the most substantial architectural evolution in the experimental series, shifting from static few-shot sampling to both dynamic context retrieval and hybrid ensemble aggregation. `Another` integrated *Gemini-3.1 Pro* with a dynamic few-shot retrieval framework based on FAISS dense matching, retrieving the three most semantically similar training examples per question at inference time. `Bio26NIA` adopted the same dynamic retrieval mechanism with *GPT-5.4*. `Dif-C` served as a static baseline, retaining the *GPT-5.4 CoT* model with the 13b static template.

Two additional systems made their first appearance in this batch. `NewM` applied the *GPT-5.5* model under the same CoT inference strategy used in `Dif-C`, providing a direct generational comparison under an otherwise identical prompting approach. `EHM-9` implemented a nine-system Ensemble Hybrid Merge framework that aggregated outputs from all three prompting strategies (five systems using the 13b static template with *GPT-4o*, *o3*, *GPT-5.4*, *Gemini-3.1 Pro*, and *GPT-5.4 CoT*; two systems using the condensed `new14few_shot` template with *GPT-5.4* and *Gemini-3.1 Pro*; and two systems using the dynamic FAISS retrieval framework with *GPT-5.4* and *Gemini-3.1 Pro*). Aggregated outputs were resolved through multi-model voting, with *GPT-5.5* serving as the final arbitrator.

Benefiting from localised semantic context at inference time, `Another` outperformed all other systems in Batch 4, achieving the highest Factoid MRR (0.4545), List F-Measure (0.6324), and ideal answer ROUGE-SU4 F1 (0.2699). `EHM-9` ranked second in both List F-Measure (0.6030) and ROUGE-SU4 F1 (0.2544), demonstrating that the multi-model consensus approach effectively filtered model-specific output inconsistencies while maintaining competitive aggregate performance.

### 4.2.3. Cross-Batch Comparative Analysis

To complement the batch-level analyses, this section examines long-term performance trends at the system level across all four evaluation cycles, as detailed in the progression matrices below. Critically, an overarching factor driving the score fluctuations across the experimental timeline is cross-batch query variability, which encompasses the shifting taxonomy distribution and latent clinical difficulty of the biomedical questions unique to each testset.

Within this framework, the baseline implementations evaluated in Batch 1—which paired earlier model configurations with a fixed, inherited static few-shot template—were intentionally deployed as an exploratory, foundational baseline, directly accounting for their lower initial metrics. As our system architectures evolved from Batch 2 through Batch 4 to actively mitigate these cross-batch difficulty variations, advanced features were systematically introduced, most notably the category-specific filtering and dynamic dense context retrieval via the *S-PubMedBert-MS-MARCO* bi-encoder utilized in the `Another` track. This progressive structural adaptation explains why the system-wide performance trajectories eventually established a stable, highly competitive plateau with incremental gains in the final rounds, rather than exhibiting unbounded exponential growth, thereby validating the stabilizing resilience of our dynamic context anchoring layers against dataset variability.

**Table 9**
List F-Measure progression across all four batches. Dashes indicate systems that did not participate in that batch. The Δ column reports the Batch 3-to-Batch 4 gain for systems active in both batches.

| System | B1 | B2 | B3 | B4 | Δ (B3→B4) |
|---|---|---|---|---|---|
| Another | 0.2443 | 0.4583 | 0.4675 | **0.6324** | +0.1649 |
| Bio26NIA | 0.2866 | 0.4605 | 0.4150 | 0.5884 | +0.1734 |
| Dif-C | 0.2259 | 0.4596 | 0.3971 | 0.5932 | +0.1961 |
| EHM-9 | — | — | — | 0.6030 | — |
| NewM | — | — | — | 0.5301 | — |

**Table 10**
ROUGE-SU4 F-Measure progression across all four batches, reflecting ideal answer generation quality. Dashes indicate systems that did not participate in that batch.

| System | B1 | B2 | B3 | B4 | Trend |
|---|---|---|---|---|---|
| Another | 0.1348 | 0.2258 | 0.2428 | **0.2699** | ↗ |
| Dif-C | 0.2381 | 0.2255 | 0.2276 | 0.2337 | → |
| Bio26NIA | 0.2315 | 0.2139 | 0.2371 | 0.2247 | ~ |
| EHM-9 | — | — | — | 0.2544 | — |
| NewM | — | — | — | 0.2196 | — |

**Table 11**
Direct performance comparison between `Dif-C` (*GPT-5.4* with CoT) and `NewM` (*GPT-5.5* with CoT) in Batch 4, under otherwise identical prompting conditions. Δ denotes the difference (NewM − Dif-C); negative values indicate regression.

| Metric | Dif-C (GPT-5.4) | NewM (GPT-5.5) | Δ |
|---|---|---|---|
| Yes/No Acc | 0.8750 | 0.8750 | 0.0000 |
| Factoid MRR | 0.3636 | 0.3182 | −0.0454 |
| List F1 | 0.5932 | 0.5301 | −0.0631 |
| ROUGE-SU4 F1 | 0.2337 | 0.2196 | −0.0141 |

**Another (Gemini-3.1 Pro with dynamic retrieval).** As shown in Table 9, `Another` achieved the highest List F-Measure in three of the four batches and demonstrated a consistent upward trajectory throughout the evaluation series. The most pronounced improvement was observed in ideal answer generation: ROUGE-SU4 F1 nearly doubled from 0.1348 in Batch 1 to 0.2699 in Batch 4 (Table 10), a gain no other system matched in magnitude. This progressive improvement suggests that dynamic few-shot retrieval compounds across batches as the retrieval pool accumulates more semantically aligned examples, yielding increasingly targeted contextual anchoring for the Gemini-3.1 Pro model.

**EHM-9 (nine-system Ensemble Hybrid Merge).** The multi-model consensus strategy employed in `EHM-9` produced stable and competitive results. Ranking second in both List F-Measure (0.6030) and ROUGE-SU4 F1 (0.2544) in Batch 4, it outperformed all static and CoT-only baselines on these metrics, demonstrating that cross-strategy aggregation effectively mitigates model-specific formatting errors and output outliers. However, the ensemble did not produce corresponding gains in Factoid MRR (0.3636), where it was matched by `Dif-C` and `Bio26NIA` and substantially outperformed by `Another` (0.4545). This disparity is attributable to the conservative factoid merging strategy: by selecting the most frequently agreed-upon candidate across nine systems, the voting mechanism can inadvertently suppress correct but minority-supported answers, thereby reducing ranking effectiveness on single-entity factoid tasks.

**Dif-C (GPT-5.4 with CoT).** `Dif-C` exhibited the most stable ROUGE-SU4 trajectory across all four batches, consistently scoring within the narrow range of 0.226 to 0.238 (Table 10). This consistency reflects the reliability of CoT prompting in producing structurally well-formed ideal answers under a static template. Nevertheless, while other systems — most notably `Another` — continued to improve progressively, `Dif-C`'s performance plateaued after Batch 1. These results suggest that stability without corresponding adaptability may be insufficient to sustain competitive performance across later evaluation rounds.

**NewM (GPT-5.5 with CoT).** Table 11 presents a direct performance comparison between `Dif-C` (*GPT-5.4* with CoT) and `NewM` (*GPT-5.5* with CoT) under identical prompting conditions. `NewM` exhibited a consistent performance regression across Factoid, List, and ROUGE metrics.

As detailed in Section 3.2.5, NewM operated under a "Double CoT" condition, where our explicit multi-stage prompt was superimposed onto the native high-effort reasoning pathway of *GPT-5.5*. This regression suggests that while explicit CoT structures effectively ground standard models, they may over-constrain or interfere with the internal reasoning mechanisms of newer architectures. Consequently, these findings indicate that architectural upgrades do not inherently guarantee task-level improvements without model-specific prompt alignment.

**Yes/No questions.** Across all batches, Yes/No question answering showed the smallest degree of system differentiation. In Batch 4, all five systems converged to an identical accuracy of 0.875, and both Another and Bio26NIA had already achieved perfect scores (1.000) in Batch 3. These results indicate that Yes/No question answering has become a largely saturated sub-task within the current evaluation setting, where most competitive systems are capable of achieving near-ceiling performance.

#### 4.2.4. Comparison with Challenge-Wide Performance Benchmarks

To contextualise our results within the broader BioASQ 14b competition, we compared peak system scores against the highest scores achieved by all participating teams across the four batches. The challenge-wide top scores were 1.0000 for Yes/No Accuracy and Macro F1, 0.5882 for Factoid MRR, and 0.6543 for List F-Measure. For ideal answer evaluation, the reference peaks were 0.4028 for ROUGE-2 Recall, 0.2607 for ROUGE-2 F1, 0.4155 for ROUGE-SU4 Recall, and 0.2547 for ROUGE-SU4 F1.

Our systems matched the challenge-wide perfect score (1.0000) for Yes/No questions and performed competitively in List questions, with a peak F-Measure of 0.6324 from the dynamic retrieval configuration, approaching the global top score of 0.6543. Most notably, in ideal answer generation, Another's Batch 4 dynamic configuration achieved a ROUGE-2 F1 of 0.2827 and a ROUGE-SU4 F1 of 0.2699, both of which exceed the corresponding challenge-wide reference scores of 0.2607 and 0.2547, respectively. These results demonstrate that the dynamic contextual retrieval and hybrid ensemble strategies explored in this study are highly competitive, with the ideal answer generation component achieving state-of-the-art performance relative to all participating teams in the challenge.

## 5. Conclusion

In this work, we presented an integrated question-answering (QA) system developed for the BioASQ 14b challenge, extending our laboratory's prior research on language model-driven pipelines [4]. Our architecture coordinates a hybrid information retrieval workflow (Phase A) with a structured answer generation and ensemble fusion engine (Phase B). In Phase A, we investigated the integration of dense semantic filtering layers and reranker ensembles with standard lexical baselines to evaluate candidate snippet selection. In Phase B, we examined dynamic contextual anchoring, cross-generational large language model characteristics, and multi-model ensemble fusion mechanisms to manage factual grounding and formatting compliance. Evaluating this unified pipeline across four sequential test batches provided an empirical analysis of the operational interplay between retrieval-stage configurations and downstream generation performance.

### 5.1. Retrieval and Pipeline Integration (Phase A and A+)

Based on the preliminary results, introducing an additional dense filtering stage did not consistently improve retrieval performance across different batches. While semantic filtering helped refine candidate selection, the original BM25-based configurations remained competitive and in several cases achieved comparable or better results. A possible explanation is that dense filtering may remove lexically relevant documents retrieved by BM25 before reranking, limiting its effectiveness in some cases.

Consistent with our previous findings, reranker ensembles did not demonstrate a clear advantage over single-reranker settings. Although ensemble approaches occasionally improved document retrieval, the gains were not stable across retrieval tasks.

For answer generation, systems using Pipeline B consistently achieved stronger performance on ideal answers. In particular, integrating the snippet-based generation workflow remained beneficial for long-form answer generation and achieved the highest rankings in multiple batches.

Overall, this year's results suggest that retrieval effectiveness is influenced not only by the addition of more complex components, but also by the interaction between retrieval strategies, candidate selection, and downstream generation workflows. Future work will further investigate more effective semantic filtering strategies and stronger retrieval-generation integration methods.

## 5.2. Prompting Strategies and Ensemble Robustness (Phase B)

Our empirical analysis in Phase B yields several insights regarding robust, prompt-based biomedical question answering.

### 5.2.1. Core Findings and System Observations

First, experimental results indicate that the structural configuration of few-shot context exercises a significant impact on performance. While static templates provided a stable baseline, they simultaneously established a performance ceiling. Conversely, extreme context compression—such as the `new14few_shot` template—reduced the token footprint but omitted essential formatting exemplars, inducing structural anomalies and subsequent parsing failures. Utilizing a dynamic FAISS-based retrieval framework to inject semantically aligned training examples at inference time mitigated these formatting failures. Automated evaluation metrics validated that this dynamic configuration enabled our top-performing track (`Another`) to achieve a ROUGE-SU4 F1 score of 0.2699 and a ROUGE-2 F1 score of 0.2827 in ideal answer generation, suggesting the utility of localized contextual anchoring for long-form summarization.

Second, the nine-system Ensemble Hybrid Merge framework (`EHM-9`) demonstrated that cross-strategy aggregation provides stability in neutralizing model-specific outliers. By synthesizing outputs across disparate prompting configurations and model families, `EHM-9` maintained stable performance in both List F-Measure (0.6030) and ROUGE-SU4 F1 (0.2544). However, the conservative consensus mechanism, while effective at satisfying strict target formatting boundaries, can suppress valid, minority-supported candidate entities in factoid tasks, indicating that different question taxonomies require more nuanced arbitration logic.

Third, cross-generational model upgrades do not inherently guarantee downstream task improvements under identical prompting. A head-to-head comparison between `Dif-C` (*GPT-5.4*) and `NewM` (*GPT-5.5*) revealed that explicit Chain-of-Thought (CoT) structures can be tightly coupled with the response distributions of specific base architectures. Consequently, migrating across model generations without proactive prompt realignment and task-specific customization can lead to performance regressions in structurally constrained biomedical tasks.

Finally, the results highlight a distinct trade-off between ideal and exact answer optimization. While polar yes/no tasks have reached near-ceiling performance, optimizing for information-dense ideal summaries occasionally introduces structural variances that impact exact answer accuracy, illustrating the intrinsic challenge of balancing generative text synthesis with rigid structural boundaries.

### 5.2.2. Limitations and Future Research Directions

Regarding the overall system design, the tight competition schedule precluded extensive component-level ablation studies, which we identify as a priority for future work to independently quantify the contribution of each pipeline module. Furthermore, the reliance on proprietary LLM architectures imposes limitations on fine-grained parameter-level analysis and domain-specific weight fine-tuning.

Furthermore, reliance on proprietary APIs introduces vulnerabilities regarding experimental reproducibility due to unannounced backend model updates. Navigating opaque, provider-specific safety filters also presents a significant challenge in biomedical domains, as these guardrails occasionally

misclassify rigorous clinical queries as restricted content, necessitating custom programmatic overrides to ensure continuous pipeline stability.

Regarding the Phase B answer generation framework, despite competitive results achieved via dynamic prompting and ensemble generation, the current architecture lacks tailored, model-specific customization across distinct LLM families. Because individual architectures exhibit varying sensitivities to instruction formatting and reasoning traces, a single prompt configuration cannot consistently yield optimal performance across all backbones. While automated prompt optimization frameworks have demonstrated potential for iterative refinement, development constraints during the competition prevented their integration.

Future research will focus on model-specific prompt alignment, entity-aware prompting, and resilient post-processing filters to further suppress structural hallucinations. Additionally, we plan to integrate localized biomedical knowledge graphs with our retrieval index. By anchoring raw textual snippet workflows within structured clinical ontologies, we aim to bridge the gap between standard document synthesis and expert-level semantic reasoning, ultimately advancing the utility and stability of high-fidelity biomedical QA systems.

## Acknowledgments

This research was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant No. 112-2221-E-008-062-MY3. We also thank the Intelligent Information Service Research Lab (IISR) members for their support.

## Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Gemini in order to assist in drafting, text translation, perform grammar and style correction, and polish the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

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

[2] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, Bioasq-qa: A manually curated corpus for biomedical question answering, bioRxiv (2022). doi:10.1101/2022.12.14.520213, preprint.

[3] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, et al., Can generalist foundation models outcompete special-purpose tuning? case study in medicine, arXiv preprint arXiv:2311.16452 (2023).

[4] B.-C. Chih, J.-C. Han, H.-C. Hung, R. T.-H. Tsai, Ncu-iisr: Biomedical question answering via gemini and gpt apis in the bioasq 13b phase b challenge, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 206–213.

[5] J. Johnson, M. Douze, H. Jegou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2021) 535–547.

[6] 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: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[7] Q. Chen, Y. Hu, X. Peng, Q. . Xie, M. B. Singer, et al., Benchmarking large language models for biomedical natural language processing applications and recommendations, Nature Communications 16 (2025) 3280.

[8] J.-C. Han, B.-C. Chih, H.-C. Hung, R. T.-H. Tsai, Ncu-iisr: A retrieval-augmented generation approach for bioasq 13b phase a and a+, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 292–301.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837.

[11] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: Findings of the association for computational linguistics: ACL 2024, 2024, pp. 2318–2335.

[12] Jina AI, jina-reranker-m0: Multilingual multimodal document reranker, https://jina.ai/news/jina-reranker-m0-multilingual-multimodal-document-reranker/, 2025. Official blog post, accessed: 2026-05-21.

[13] X. Li, A. Shakir, R. Huang, T.-f. A. Lee, J. Lipp, B. Clavié, J. Li, Prorank: Prompt warmup via reinforcement learning for small language models reranking, 2025. arXiv:2506.03487.

[14] Z. Nussbaum, J. X. Morris, A. Mulyar, B. Duderstadt, Nomic embed: Training a reproducible long context text embedder, Transactions on Machine Learning Research (2025). URL: https://openreview.net/forum?id=IPmzyQSiQE, reproducibility Certification.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[16] Sentence Transformers, multi-qa-mpnet-base-dot-v1, https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1, 2021. Accessed: 2026-05-21.

[17] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare 2 (2021) 1–23.

[18] OpenAI, Gpt-4o system card, https://arxiv.org/abs/2410.21276, 2024. ArXiv:2410.21276v1.

[19] OpenAI, OpenAI o3 and o4-mini System Card, Technical Report, OpenAI, 2025. URL: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[20] Gemini Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, et al., Gemini: A family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).