BIT.UA at BioASQ 13B: Revisiting Evaluation, DPRF-Enhanced Retrieval and Fine-Tuned LLMs

Notebook for the BioASQ Lab at CLEF 2025

Richard A. A. Jonker[1,*], Tiago Almeida[1], João R. Almeida[1] and Sérgio Matos[1]

[1]IEETA/DETI, LASI, University of Aveiro, Aveiro, Portugal

Abstract

Biomedical information retrieval and question answering are critical for navigating the vast and continually expanding body of biomedical literature. The BioASQ Task B Challenge provides a valuable benchmark for developing and evaluating systems capable of retrieving relevant documents and generating high-quality answers to biomedical questions. This paper describes our participation in the thirteenth edition of the BioASQ challenge, focusing on Task B, which is based off our participation in the twelfth edition of the challenge. For Phase A, we employed a hybrid two-stage retrieval pipeline combining BM25-based retrieval with transformer-based rerankers such as BioLinkBERT and PubMedBERT. We showcased the performance Dense Pseudo Relevance Feedback (DPRF) using the BGE-M3 model to enhance retrieval. In Phase B, we used a range of large language models (LLMs) for answer generation, including OpenBioLLM, LLaMA Nemotron, and a custom fine-tuned Gemma-3 27B model. We also unified our summarization and ensembling strategy from last year into a single generation step to improve efficiency and coherence. A key insight from this year's participation was the persistent misalignment between automatic evaluation metrics and human-judged answer quality, a discrepancy that influenced both our system design and our interpretation of results in previous years. For phase A our systems consistently achieved top rankings. We discuss these outcomes in light of evaluation challenges and outline promising directions for future work. All code is publicly available at https://github.com/bioinformatics-ua/BioASQ13B.

Keywords

Information Retrieval, Dense Retrieval, Semantic Search, Large Language Model, Answer Generation, Pseudo Relevance Feedback

1. Introduction

Biomedical information retrieval and question answering remain vital yet challenging tasks, propelled by the explosive growth of biomedical literature. Effective systems for retrieving relevant documents and accurately answering complex biomedical questions are essential for researchers and clinicians aiming to make informed decisions. The BioASQ Task B Challenge continues to provide a critical platform for benchmarking and advancing these technologies in the biomedical domain.

Our participation in the 13th BioASQ challenge [1] employed a hybrid two-stage retrieval pipeline—combining classical BM25-based retrieval with transformer-based neural rerankers—and a retrieval-augmented generation (RAG) framework leveraging large language models (LLMs), almost identical to our previous years submission [2]. We incorporated Dense Pseudo Relevance Feedback (DPRF) using semantic embeddings and used ensemble techniques such as reciprocal rank fusion to improve retrieval robustness. For answer generation, multiple LLMs, including fine-tuned versions of Gemma models, were combined with snippet-based context and novel answer selection methods.

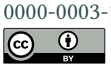
However, a key finding from our previous work was the significant misalignment between automatic evaluation metrics, such as ROUGE-based F1 scores, and human evaluation of answer quality. This is further emphasized by the fact that there is no true gold-standard which these metrics are run on, and the metrics get updated aswell (including MAP in phase A). This misalignment resulted in misleading

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

richard.jonker@ua.pt (R. A. A. Jonker); tiagomeloalmeida@ua.pt (T. Almeida); joao.rafael.almeida@ua.pt (J. R. Almeida); aleixomatos@ua.pt (S. Matos)

0000-0002-3806-6940 (R. A. A. Jonker); 0000-0002-4258-3350 (T. Almeida); 0000-0003-0729-2264 (J. R. Almeida); 0000-0003-1941-3983 (S. Matos)



conclusions and optimizing for certain metrics, which did not translate into better human-evaluated performance. For instance, runs achieving top ranks by automatic metrics were often rated poorly by human evaluators, and vice versa. This discrepancy exposed fundamental limitations in relying on automatic metrics alone to guide system development and evaluation.

These misleading conclusions motivated a fundamental rethink of our methodology for the 13th BioASQ challenge. We prioritized reproducibility and robustness, emphasizing consistent evaluation across large batches and reducing reliance on potentially unreliable automatic metrics, aided by the fact that automatic metrics were not released throughout the competition this year. Our approach shifted towards less reliance on individual model performance and using a large variety of models in ensemble. This year we also try to keep our submissions relatively consistent to truly gauge the performance.

A central theme of our work this year is evaluative uncertainty: recognizing and addressing the instability and unreliability of current automatic metrics, and exploring strategies that maintain performance across batches and evaluation regimes. We see this as a necessary evolution for BioASQ and similar challenges where generative quality is increasingly difficult to quantify automatically.

For phase A, we refined retrieval strategies by adjusting the number of reranked documents to balance performance and computational cost, as well as focusing more on DPRF. To address generation, historically our weakest component, we significantly revamped our pipeline. We transitioned to faster inference engines and expanded our model lineup to include LLaMA Nemotron, and OpenBioLLM, along with a two-stage fine-tuning of the larger Gemma-3 27B model. We also unified ensembling and summarization into a single generation step, improving answer coherence and efficiency.

The rest of the paper is organized as follows. Section 2 describes our submissions from the previous year, highlighting discrepancies between the conclusions we initially drew and those informed by the final human evaluations. Section 3 outlines the changes made to our system this year. Section 4 presents both our internal validation results and batch-wise competition outcomes, discussing the performance of individual systems. Section 5 offers a critical analysis of our findings, including challenges related to evaluation metrics. Finally, Section 6 summarizes our conclusions and outlines directions for future work.

2. Previous Work

In our participation in BioASQ 12 Task B [2], our approach comprised a two-stage retrieval pipeline, neural reranking, and retrieval-augmented generation (RAG) techniques. For the document retrieval phase (Phase A), we implemented a hybrid strategy that began with first-stage retrieval using BM25 from the PISA framework [3, 4] to efficiently filter candidate documents from the vast PubMed corpus. This was followed by neural reranking using transformer-based models, specifically PubMedBERT [5] and BioLinkBERT [6]. Typically, we applied neural reranking to the top 1,000 documents retrieved via BM25. To further enhance retrieval performance, we incorporated Dense Pseudo Relevance Feedback (DPRF) through semantic search using BGE-M3 [7] embeddings. The goal was to identify documents semantically similar to the top 50-100 ranked by the neural reranker. The outputs from various models were then combined using reciprocal rank fusion (RRF), which proved beneficial for producing robust final rankings.

Our conclusions from this phase include:

1. Large models and small models performed similarly.
2. According to internal validation, higher data quality led to better performance than data quantity. This contradicted initial findings from preliminary results but was ultimately supported by the official evaluation. This discrepancy highlights a case of misaligned automatic evaluation.
3. DPRF yielded performance improvements, although these were minimal.

For answer generation (Phases A+ and B), we adopted a RAG framework. We provided the top five retrieved documents as context to several LLMs, including Llama 3 70B, Nous-Hermes2-Mixtral, and a fine-tuned Gemma 2B model. The Gemma 2B model was fine-tuned on BioASQ training data using

Low Rank Adaption (LoRA). To ensure the relevance and accuracy of generated answers, we developed answer selection mechanisms that used our neural reranker to evaluate candidate answers as pseudo-documents. We also experimented with snippet-based context and implemented answer truncation strategies to comply with the BioASQ 200-word limit. Finally, we used Mixtral to summarize model outputs, producing more concise answers that aligned better with BioASQ evaluation criteria. Our conclusions have evolved significantly since the release of the human evaluation results. A comparison of rankings from last year (Phases A+ and B) is shown in Table 1 and Table 2. The conclusions for Phases A+ and B are summarized as follows:

1. According to automatic metrics, the fine-tuned Gemma model achieved the second-best F1 score (system-3, phase A plus batch 2). However, in human evaluation, it was ranked only 9th—outperformed by one of our other systems, which ranked 17th in F1 and 1st in Recall. Notably, our run which ranked 3rd in human evaluation (phase A plus, system-3, batch 4) had an F1 rank of 20 and a recall rank of 23, illustrating a significant misalignment between ROUGE-based automatic metrics and human judgments. Further, we tried optimizing for high recall last year, which did not correlate with top performance in human evaluations.
2. Document sources did not play a significant role in performance (phase A plus). In the context of a competition, it appears more effective to keep the document source constant and vary other model or pipeline components.
3. Summarization techniques (phase A plus, system-3,4 batches 3 and 4, also in phase B) led to modest F1 gains at the expense of recall. However, these changes ultimately translated into substantial improvements in human evaluation scores.
4. In Phase B, the main difference was the use of snippets, which consistently yielded top recall scores (all bolded recall results relied on snippets). However, this did not always translate into high human evaluation scores.

Table 1
Rankings of our Phase Ap submissions in BioASQ 12B, showing the relative rank across different metrics (F1, Recall, Human Evaluation) for Phase A Plus.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	F1	Rec	HE	F1	Rec	HE	F1	Rec	HE	F1	Rec	HE
system-0	16	3	12	18	2	10	26	2	14	22	2	6
system-1	15	4	15	22	3	12	27	4	16	23	3	9
system-2	14	5	17	17	**1**	**7**	**16**	**1**	12	**16**	**1**	7
system-3	13	2	**11**	**2**	10	9	22	19	7	20	23	**3**
system-4	**12**	**1**	14	11	19	21	25	17	**6**	21	25	5

Table 2
Rankings of our Phase B submissions in BioASQ 12B, showing the relative rank across different metrics (F1, Recall, Human Evaluation) for Phase B.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	F1	Rec	HE	F1	Rec	HE	F1	Rec	HE	F1	Rec	HE
system-0	24	7	**21**	31	3	26	32	**4**	28	**30**	**4**	23
system-1	26	8	25	26	**1**	**20**	31	28	22	33	29	17
system-2	**22**	**1**	23	32	32	32	33	6	29	31	5	25
system-3	25	20	22	**22**	29	31	34	29	21	34	34	24
system-4	23	5	24	29	2	27	**29**	26	**15**	32	33	**16**

3. Methodology

Building on some of the conclusions from last year, along with new intuitions and assumptions, we began refining our methodology. For Phase A, our main objectives were to (1) successfully reproduce last year's submission and (2) place greater emphasis on DPRF, which we had previously underutilized. Due to these requirements, our first batch felt relatively weak upon submission, as we didn't have everything fully prepared in time and were only able to train five models. After submitting the first batch, we focused on validating models and adjusting the data used to train the reranker models, as well as setting up DPRF, which was not working in B1. Unlike last year, we approached this phase with a new perspective: rather than focusing on top-performing models based on validation, we assumed that most models would perform similarly and could all contribute meaningfully when used in an ensemble. As a result, we prioritized training diverse models while maintaining relatively strong individual performance, as demonstrated in the Validation Results section. From B2 to B4, the methodology remained largely consistent, with one key change introduced in B3: we reduced the number of documents reranked by the neural reranker from 1,000 to 100. A small validation test confirmed this had negligible impact on performance. The rationale was that we lacked the time to run inference for all models and apply DPRF across all 2025 models. Finally, in B4, we also experimented with incorporating our 2023 models into the ensemble.

Our main changes this year focused on generation, which had been our weakest component in previous years—last year, we only achieved a single second-place ranking in one of the batches for Phase A+. Phase A+/B is a rapidly evolving and challenging task to keep pace with, especially given our reliance on LLMs as the core of our RAG pipeline. While many of our conceptual approaches remained similar to last year, the implementation changed significantly. We switched our inference engine from Ollama[8] to LMDeploy [9], greatly increasing inference speed, and transitioned our model lineup to include Llama-Nemotron-70B[1] [10, 11] , and OpenBioLLM[2] [12]. Following this, we experimented with new prompting strategies and also restructured how we handle summarization and ensembling.

In last year's approach, we would ensemble multiple LLM outputs and then use one of our neural rerankers to select the best answer for a given question. Summarization, at the time, was treated independently and served only to compress model outputs into shorter texts. This year, we merged ensembling and summarization into a single step. Specifically, we designed a summarization prompt that takes multiple outputs from different models and combines them into one unified answer. This approach allows us to achieve both ensembling and summarization within a single generation pass.

The last major change was introduced in B4, where we revisited the use of fine-tuned models. This year, we chose to fine-tune a significantly stronger and larger model—Gemma-3 27B[3] [13]. The fine-tuning process consisted of two stages. First, we fine-tuned the model on a custom dataset [4] using LoRA [14] and Unsloth[5], serving as an initial knowledge injection phase. In the second stage, we fine-tuned the model specifically for RAG tasks using Prompt 5 and incorporating five document abstracts as context. A more minor change involved further experimentation with the number of documents used during generation. Regarding Phase A+, all of our runs relied on documents from our Phase A run 4 submission, assuming these were the strongest Phase A results. In Phase B, we continued to emphasize the use of snippets, further exploring their role and effectiveness in the generation pipeline.

The prompts used for generation generally follow similar styles, with Prompt 1 being the simplest and producing the weakest answers. Prompt 2 is slightly more complex, while Prompts 3 to 5 introduce higher complexity, eliciting reasoning steps and prompt 5 providing an example. In terms of summaries, we offered two prompts. These prompt changes occurred at the same stage as Phase A, with the primary difference being the inclusion of an example. The prompts can be seen in Appendix A.

[1] https://huggingface.co/ibnzterrell/Nvidia-Llama-3.1-Nemotron-70B-Instruct-HF-AWQ-INT4
[2] https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B
[3] https://huggingface.co/google/gemma-3-27b-it
[4] Dataset is currently under a double blind review
[5] https://docs.unsloth.ai/

4. Results

In this section, we describe the different models and configurations we evaluated as part of our participation in BioASQ 12 Task B. We outline the submissions made and report the official preliminary results provided by the organizers. The focus is on both retrieval and generation components, with performance assessed using standard automatic evaluation metrics. We also provide observations on how different approaches and design choices influenced the overall outcomes. As a reminder the results here present the official BioASQ metrics (MAP, ROUGE-2), as described in our previous work [2].

4.1. Validation Results

Validation was conducted on the first and second batches of the 2024 dataset. For reference, our best scores from these batches were 0.4142 for Batch 1 (B1) and 0.4412 for Batch 2 (B2), indicating that our current validation results are competitive with the large ensembles we used last year. This section presents validation results for Phase A (no validation was conducted for the generation task). The validation was performed prior to the release of Batch 2, and a summary of the findings is shown in Table 3. These results are based on an arbitrarily selected baseline model, with various hyperparameters altered to observe their impact. Overall, we observed a 2-point spread between the best- and worst-performing models, suggesting that most variations yielded relatively minor effects.

Table 3
Performance metrics with varying hyperparameters only

Model	Epoch	Sampler	Data	KN	Trainer	ExPOS	Warmup	B01	B02	Total
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	**0.41162**	0.41426	0.41294
BioLinkBERT-large	-	-	-	-	-	-	-	0.39634	0.42470	0.41052
BiomedBERT-base	-	-	-	-	-	-	-	0.40352	**0.42606**	**0.41479**
BiomedBERT-large	-	-	-	-	-	-	-	0.40030	0.41463	0.40747
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	**0.41162**	**0.41426**	**0.41294**
-	1	-	-	-	-	-	-	0.40898	0.39657	0.40277
-	10	-	-	-	-	-	-	0.39418	0.40456	0.39937
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	0.41162	**0.41426**	0.41294
-	-	-	-	-	-	-	True	**0.41909**	0.40796	**0.41353**
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	**0.41162**	**0.41426**	**0.41294**
-	-	-	quality	-	-	True	-	0.40818	0.40596	0.40707
-	-	-	quantity	-	-	False	-	0.39208	0.40497	0.39852
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	**0.41162**	**0.41426**	**0.41294**
-	-	-	-	-	pointwise	-	-	0.38568	0.40430	0.39499
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	0.41162	**0.41426**	0.41294
-	-	-	-	2	-	-	-	**0.42562**	0.41033	**0.41797**
BioLinkBERT-base	4	basic	quality	1	pairwise	False	False	**0.41162**	**0.41426**	**0.41294**
-	-	basicv2	-	-	-	-	-	0.39860	0.41370	0.40615

In terms of model architecture, the base variants generally performed better their large counterparts, with BioMedBERT-base achieving the best performance and BioMedBERT-large the worst. Among training durations, 4 epochs appeared optimal. The use of warm-up also provided slight performance improvements in some configurations. Looking at the data configurations, we found that using high-quality data led to better results than simply increasing data quantity. The ExPOS setting—which expands positive samples using semantically similar but unannotated documents—did not improve performance. While these documents may be relevant, their lack of annotation may have diluted their usefulness. However, we note that the true value of ExPOS might only be measurable through manual evaluation. Pointwise training continued to underperform relative to pairwise approaches. Increasing the number of negative samples (KN) to 2 showed a modest improvement. Experimentation with alternative sampling strategies, such as ‘basicv2‘ and ‘exponential‘, did not yield significant gains. In summary, Phase A validation primarily served to eliminate poorly performing configurations from

consideration. In total, we trained 24 models—12 based on BioLinkBERT and 12 on BioMedBERT—as detailed in Table 10 in Appendix B.

4.2. Official Results - Phase A

A summary of the systems submitted across each Phase A batch is presented in Table 4, with the performance results shown in Table 5. Each system varies in the combination of models used, including different training epochs, samplers, and ensemble strategies. Some techniques were developed between batches, which explains the evolving configurations across submissions, however we tried to keep most things stable between batches, especially since we did not get any intermediate results between batches.

Table 4
Summary of the systems submitted for phase A. Each system details the year and number of the models used in the ensemble as well as some extra information were applicable, for example DPRF being applied.

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	2025 + 2024 Pairwise(12)	2025 models (24)	2025 models (24)	2025 models (24)
system-1	Pairwise, BioLinkBERT(9)	2025 models + DPRF (20)	2025 models + DPRF (24)	2025 models + DPRF (24)
system-2	2025 models (5)	2025 models ExPos (16)	2024 models (46)	2023 (26) + 2024 (46) + 2025 (24)
system-3	2024 models (11)	2025 (24) + 2024 (46)	2025 (24) + 2024 (46)	2024 (46) + 2025 DPRF (24)
system-4	2025(4) + 2024 (11)	2025 + DPRF (4+20) + 2024 (46)	2025 DPRF (24) + 2024 (46)	2023 (26) + 2024 (46) + 2025 DPRF (24)

Table 5
Performance metrics for various systems across different batches for phase A. Bold values represent our best submission.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	MAP	Prec	Rank	MAP	Prec	Rank	MAP	Prec	Rank	MAP	Prec	Rank
system-0	41.41	10.47	4	41.93	10.94	5	30.79	8.88	4	17.37	5.76	2
system-1	40.82	10.71	5	**43.07**	11.65	3	**32.36**	9.41	1	**18.01**	6.00	1
system-2	42.14	10.00	2	41.93	10.94	5	31.48	8.53	2	16.25	5.65	6
system-3	41.75	10.47	3	42.26	10.71	4	30.79	8.65	4	16.19	5.76	8
system-4	**42.46**	10.47	1	41.66	10.82	8	30.84	8.59	3	16.24	5.65	7
Best Competitor	38.06	7.88	6	44.25	9.76	1	30.49	8.88	6	17.01	6.10	3
Median	27.16	7.18	24	30.25	9.96	20	20.11	7.53	22	6.27	3.79	39

In Batch 1, we submitted five systems, primarily composed of a small set of newly trained 2025 models (e.g., PubMedBERT and BioLinkBERT variants) and previously used 2024 systems. Systems 0 and 1 focused on pairwise models, with System-2 containing all the 2025 models and System-3 containing all the 2024 models. System-4—an ensemble of both 2025 and 2024 systems—achieved the best MAP (42.46), obtaining a slightly higher MAP than System-2 (42.14). System-1 had a marginally better precision than the others (10.71). Notably, all our systems ranked among the top 5, outperforming both the best external competitor (MAP: 38.06) and the median submission (MAP: 27.16).

In the second batch, we fixed a number of issues present in the first batch and also introduced DPRF. System-1, which applied DPRF to most of our 2025 models, achieved the best MAP (43.07) and precision (11.65) among our submissions, placing 3rd overall. Interestingly, System-0, which contained all of our newly trained 2025 models, obtained the same MAP as our 2025 models trained solely on ExPOS data. This suggests that the ExPOS data may, in fact, be beneficial. Systems 3 and 4 were our large ensembles combining 2025 and 2024 models. In this batch, we observed that including DPRF-enhanced models in the large ensemble actually resulted in worse performance compared to using only the base 2025 models, even though the DPRF models alone had a better performance than the base runs. This indicates that the ensemble strategy may need further refinement.

In Batch 3, we retained most of the pipeline but reduced the number of reranked documents from 1000 to 100 to enable DPRF on all 2025 models. We also updated System-2 to use all our 2024 submissions. Once again, DPRF was the best-performing approach, with the 2024 models surprisingly coming in

second. This result is somewhat concerning, as we would expect the newer 2025 models to outperform the 2024 models, even with a smaller ensemble size. For the large ensembles, System-4 had a slight increase in MAP than System-3, which is a step in the right direction.

In the final batch, we experimented with including the 2023 runs in a large ensemble, as we were curious to see how well older models would perform. In general, we observed that the 2025 models with DPRF again achieved the best rank. However, the 2025 models, when included in larger ensembles, remained consistently behind the DPRF-only systems. Among the ensemble configurations, none outperformed Systems 0 and 1. Nevertheless, we note that the ensemble including the 2023 models did perform slightly better than others.

In general, the main conclusions we can draw from these automatic results (pending updates to the gold standard) are as follows: DPRF appears conclusively better this year and represents a promising step forward. Everything else remains relatively inconclusive. The weakness of the ensembles may stem from changes in how we performed ensembling: instead of conducting a single RRF, we grouped models into sub-ensembles (e.g., creating an ensemble of two sub-ensembles). The intuition behind this was to better balance the weights of the model sets—ensuring that the fewer, but newer, 2025 models could carry equal weight to the more numerous 2024 models. Overall, we consistently achieved top-tier performance, with only Batch 2 not securing a 1st place preliminary rank.

Across all batches, our systems consistently outperformed the median and often rivaled or exceeded the best competitors. In Batch 1, we obtained the best MAP and precision. In Batches 3 and 4, our systems ranked 1st overall. These results validate the evolution of our system design—starting from small ensembles in early batches to robust DPRF-integrated systems in later ones. Despite some inconsistencies between validation and submission results, likely due to training configuration discrepancies, our systems demonstrated strong performance across diverse configurations and evolving evaluation standards.

4.3. Official Results - Phase A plus

A summary of the systems submitted across each Phase A plus batch is presented in Table 6, with the performance results shown in Table 7. Each system varies in terms of prompting strategies and models. Some techniques were developed between batches, which explains the evolving configurations across submissions, however we tried to keep most things stable between batches, especially since we did not get any intermediate results between batches.

Table 6
Summary of the systems submitted for Phase A+. In general, **N** refers to Nemotron and **O** refers to OpenBioLLM. There are three types of submissions: (1) **Individual models** (e.g., systems 0–1), such as `N_5_3`, which refers to the Nemotron model using 5 abstracts and prompt 3. (2) **Summary models**, where we describe the idea behind the aggregation. For example, Summ. all @ 3 summarizes all models using prompt 3, while Summ. N P2 summarizes all Nemotron outputs generated with prompt 2. (3) **Custom models**, which is our double fine-tuned Gemma 3 27B model, with varying number of training epochs.

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	N_5_3	O_10_4	N_10_5	N_5_5
system-1	O_5_3	Summ. N P1 (3)	O_10_5	custom_E1
system-2	Summ. all @ 3 (6)	Summ. N P2 (3)	Summ. N P5 (3)	Summ. N P5 (3)
system-3	Summ. all @ 5 (6)	Summ. N P4 (3)	Summ. (different order) (3)	custom_E3
system-4	Summ. of 2, 3 (2)	Summ. of 1,2,3 (3)	Summ. All N (9)	Summ. All N (9)

In Batch 1, we began by establishing baseline systems. Systems 0 and 1 represented our strongest individual model/prompt configurations based on manual inspection: specifically, Nemotron and OpenBioLLM using Prompt 3 and 5 abstracts, respectively. We then tested our summarization strategy—believed to outperform individual runs based on qualitative assessment—on both configurations with 3 abstracts (System-3) and 5 abstracts (System-4). According to automatic evaluation metrics, System-0 achieved

Table 7
Performance metrics for various systems across different batches for phase A Plus. Bold values represent our best submission.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	F1	Rec	Rank	F1	Rec	Rank	F1	Rec	Rank	F1	Rec	Rank
system-0	**12.98**	15.97	14	**20.37**	25.15	8	13.18	18.34	24	8.80	11.83	33
system-1	12.43	30.62	19	11.63	18.40	31	**14.28**	22.63	19	**16.91**	17.61	1
system-2	11.14	24.43	26	11.86	21.38	30	10.45	18.58	35	6.97	14.72	47
system-3	10.53	23.34	30	11.33	21.61	34	10.76	19.30	32	13.78	17.04	10
system-4	9.23	21.20	38	10.63	21.85	40	9.17	18.83	39	7.20	15.32	46
Best Competitor	20.88	23.59	1	22.18	22.53	1	20.90	25.11	1	15.98	19.25	2
Median	10.67	20.57	28	12.30	20.43	25	11.60	18.32	27	8.75	14.98	33

the highest F1 score (12.98), with System-1 slightly behind (12.43). However, it is worth noting that System-1 obtained nearly double the recall of System-0. In contrast, our summarization-based systems performed worse than the individual runs, which was surprising given our initial expectations.

In Batch 2, we expanded our experiments with different prompt configurations. System-0 again consisted of a single model (OpenBioLLM), this time with 10 abstracts and using Prompt 4. This configuration performed reasonably well, achieving a rank of 8. Systems 1 through 3 explored various prompts using Nemotron, each paired with summarization. System-4 was an ensemble summary combining the outputs from the previous three. Despite our belief in the quality of these summaries, their performance did not surpass that of the single-model system.

Batch 3 introduced yet another prompt variation, and we reintroduced Nemotron for direct comparison. System-0 featured the Nemotron run, while System-1 included the OpenBioLLM configuration, which again achieved the best score among our submissions in this batch—though with a lower overall rank than in earlier batches. An additional experiment in this batch assessed the effect of answer ordering within summaries. F1 scores showed only marginal differences between different orderings, suggesting the models were relatively robust to such variations.

Batch 4 marked a major shift with the introduction of our fine-tuned Gemma 3 model. We submitted two systems using this model: System-1 used the checkpoint after the first epoch, while System-3 used the checkpoint after the third. System-1 achieved the highest F1 score in the entire competition, while System-3 placed tenth. Based on these automatic metrics, it appears that fine-tuned models represent a significant advancement. However, we caution against overinterpreting these results without human evaluation. Last year, similar trends in automatic scores were ultimately contradicted by human assessment, which favored an LLM-generated response over our best-performing fine-tuned model.

Overall, this year's automatic results often contradicted our qualitative evaluations and intuitions. All submitted runs were qualitatively selected for answer quality and alignment with BioASQ standards. From this review, Nemotron appeared to produce significantly better responses than OpenBioLLM—yet this was not reflected in the automatic metrics. Similarly, our summarization-based systems were believed to be more aligned with BioASQ's ideal answers, but they consistently underperformed in F1 evaluations, even in ensemble settings. These findings highlight the limitations of automatic metrics in capturing the nuanced quality of biomedical question answering and reinforce the importance of incorporating human judgments into final system evaluations.

4.4. Official Results - Phase B

A summary of the systems submitted across each Phase B batch is presented in Table 8, with the performance results shown in Table 9. Each system varies in terms of prompting strategies and models. Some techniques were developed between batches, which explains the evolving configurations across submissions, however we tried to keep most things stable between batches, especially since we did not get any intermediate results between batches.

In Batch 1, we began by establishing baseline systems. System-0 represented our strongest individual

Table 8
Summary of the systems submitted for phase B. We keep the same naming schema as Table 6, with the addition of snippets.

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	N_snipp_3	N_snipp_5	N_snipp_5	N_snipp_5
system-1	Summ. N abstract (9)	Summ. Snipp. All (6)	Summ. Snipp. All (6)	Summ. Snipp. All (6)
system-2	Summ. N all (12)	Summ. Prompt 4 (8)	Summ. Prompt 4 (8)	custom_E1
system-3	Summ. O All (12)	Summ. Prompt 5 (8)	Summ. Prompt 5 (8)	custom_E3
system-4	Summ. 2+3 (2)	Summ. N Prompt 4+5 (4)	Summ. N Prompt 5 (4)	Summ. N Prompt 5 (4)

Table 9
Performance metrics for various systems across different batches for phase B Plus. Bold values represent our best submission.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	F1	Rec	Rank	F1	Rec	Rank	F1	Rec	Rank	F1	Rec	Rank
system-0	**18.64**	21.54	37	**21.29**	22.23	42	**18.07**	19.75	41	16.53	17.44	47
system-1	13.40	25.96	53	19.46	29.67	45	17.37	28.28	42	15.39	23.99	50
system-2	14.30	27.60	51	17.91	29.46	46	13.98	24.33	47	31.17	34.92	7
system-3	14.43	30.41	48	17.47	28.71	51	15.10	24.47	45	**32.02**	33.96	6
system-4	13.28	28.09	55	16.07	26.64	52	13.35	19.93	49	12.05	19.47	56
Best Competitor	41.22	43.50	1	44.17	46.68	1	35.20	34.29	1	36.04	37.45	1
Median	19.61	32.25	33	23.68	31.75	33	20.98	28.48	30	20.23	24.39	37

model/prompt configuration based on manual inspection: Nemotron using Prompt 3 and snippets. We then evaluated our summarization strategy across various settings: all Nemotron models using abstracts (9 answers, System-1), all outputs from Nemotron (12 answers, System-2), all outputs from OpenBioLLM (12 answers, System-3), and a final summary combining Systems 2 and 3. According to automatic evaluation metrics, System-0 achieved the highest F1 score (18.64), with all other systems performing worse.

In Batches 2 and 3, we kept the submissions largely consistent. System-0 again featured Nemotron with snippets, now using an updated prompt (Prompt 5). The remaining systems included summaries of all snippet outputs (System-1 and System-6), Nemotron with Prompt 4 (8 answers, System-2), Nemotron with Prompt 5 (8 answers, System-3), and a summary combining Systems 2 and 3. Once again, System-0 outperformed the others, although the performance gap narrowed. It remains unclear whether the updated prompt reduced System-0's relative effectiveness or improved the summaries' quality. In Batch 4, we introduced our custom fine-tuned model (Systems 2 and 3), which performed relatively well according to automatic metrics, achieving 6th and 7th place overall.

Overall, we opt not to provide a detailed analysis of individual systems, as we expect many conclusions to shift once human evaluation results are available. Additionally, the rank differences among our summarization systems are relatively small, suggesting limited significance in their comparative performance at this stage.

5. Discussion and Future Work

This year, we would like to reflect more broadly on the competition itself, its evaluation process, and critically assess our own submissions. One key issue we observed—also highlighted by our experience in past years—is the importance of having access to the correct, finalized evaluation results. The discrepancy between intermediate results and the true gold standard has, in the past, led us to favor systems that later proved suboptimal. While we do not intend to critique the competition's structure, we do not draw any strong conclusions since the final gold standard will affect the results. This issue regarding misalignment is a known issue in existing literature where various authors discuss

misalignment between automatic metrics and human annotation [15, 16, 17].

This year, the lack of intermediate results between batches inadvertently had a positive effect: it encouraged us to focus more on consistency and robustness across batches rather than optimizing for short-term performance. Additionally, we would like to emphasize the importance of evaluating cross-batch consistency. Some systems perform well in specific batches but not in others, indicating that the data distribution can shift between batches. This variability underlines the importance of building systems that generalize well and highlights a valuable aspect of the BioASQ challenge.

In Phase A, we achieved strong and consistent performance, comparable to last year, which we find encouraging. Notably, our DPRF-based submissions achieved top results in all three batches where they were used, outperforming larger ensemble systems that we initially assumed would perform better. This is a promising direction for future work—pending confirmation from human evaluation. We also plan to further explore optimal ensemble strategies, as well as alternative embedding methods beyond BGE-M3 for DPRF. We would also like to further investigate different types of ensembling beyond RRF including reward modeling and performance aware selection. Additionally, we aim to experiment with prompt adaptation based on question similarity, potentially yielding more tailored and relevant answers.

Regarding Phase B (generation), we refrain from in-depth analysis at this stage due to the absence of human evaluation, which plays a crucial role in interpreting results. We prefer to avoid drawing premature conclusions that may not align with qualitative assessments once they become available.

6. Conclusion

In Phase A, our DPRF-based systems consistently performed better than our larger ensembles, indicating that carefully designed retrieval and reranking strategies can rival, and even exceed, the performance of larger ensembles, providing a good direction for future work. In Phase B, while our LLM Prompts showed promising output quality, automatic evaluation results did not always align with our qualitative assessments—highlighting the ongoing challenge of evaluation in generative tasks. On the other hand our fine-tuned model obtained very good automatic metrics. However, we cannot conclude whether this is due to its alignment with the style of BioASQ or because it genuinely outperforms the other submissions. We remain cautious in interpreting the results before final gold standards and human evaluations are available. Ultimately, BioASQ continues to be a valuable platform for rigorous benchmarking and reflective system development, and we look forward to further contributing in future editions.

Acknowledgments

This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT) under research unit UIDB/00127 – IEETA. Richard A. A. Jonker is funded by the FCT doctoral grant PRT/BD/154792/2023, with DOI identifier https://doi.org/10.54499/PRT/BD/154792/2023.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT as a writing assistant.

References

[1] 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, F. N. Maria Di Nunzio, Giorgio, 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: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S.

de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] T. Almeida, R. A. A. Jonker, J. Reis, J. R. Almeida, S. Matos, BIT.UA at BioASQ 12: From Retrieval to Answer Generation, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, Grenoble, France, 2024, pp. 47–67. URL: https://ceur-ws.org/Vol-3740/paper-05.pdf, notebook for the BioASQ Lab at CLEF 2024.

[3] A. Mallia, M. Siedlaczek, J. Mackenzie, T. Suel, PISA: performant indexes and search for academia, in: Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019., 2019, pp. 50–56. URL: http://ceur-ws.org/Vol-2409/docker08.pdf.

[4] S. MacAvaney, C. Macdonald, A python interface to PISA!, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3339–3344. URL: https://doi.org/10.1145/3477495.3531656. doi:10.1145/3477495.3531656.

[5] 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 Trans. Comput. Healthcare 3 (2021). URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.

[6] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining language models with document links, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8003–8016. URL: https://aclanthology.org/2022.acl-long.551. doi:10.18653/v1/2022.acl-long.551.

[7] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[8] O. Team, Ollama: Run large language models locally, https://ollama.com, 2025.

[9] LMDeploy Contributors, Lmdeploy: A toolkit for compressing, deploying, and serving llm, https://github.com/InternLM/lmdeploy, 2023.

[10] B. Adler, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, A. Brundyn, J. Casper, B. Catanzaro, S. Clay, J. Cohen, et al., Nemotron-4 340b technical report, arXiv preprint arXiv:2406.11704 (2024).

[11] Z. Wang, A. Bukharin, O. Delalleau, D. Egert, G. Shen, J. Zeng, O. Kuchaiev, Y. Dong, Helpsteer2-preference: Complementing ratings with preferences, 2024. URL: https://arxiv.org/abs/2410.01257. arXiv:2410.01257.

[12] M. S. Ankit Pal, Openbiollms: Advancing open-source large language models for healthcare and life sciences, https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

[13] G. Team, Gemma 3 (2025). URL: https://goo.gle/Gemma3Report.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

[15] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, S. Reddy, Evaluating correctness and faithfulness of instruction-following models for question answering, Transactions of the Association for Computational Linguistics 12 (2024) 681–699. URL: https://doi.org/10.1162/tacl_a_00667. doi:10.1162/tacl_a_00667.

[16] A. Chen, G. Stanovsky, S. Singh, M. Gardner, Evaluating question answering evaluation, in: Proceedings of the 2nd workshop on machine reading for question answering, 2019, pp. 119–124.

[17] R. Rahnamoun, M. Shamsfard, Multi-layered evaluation using a fusion of metrics and llms as judges in open-domain question answering, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 6088–6104.

A. Prompts

Prompt 1

Context: {context}
Question: {question}

Answer in less than 150 words, present a final json containing your answer {{"answer": answer}}

Prompt 2

Act as a biomedical expert. You will receive several {d_type} summarizing research findings and methodologies. Along with this, a question will be provided ('[question]').
Your role is to analyze the {d_type} and provide a scientifically accurate, concise answer to the question, leveraging the information from the {d_type}.

Answer in less than 150 words, present a final json containing your answer {{"answer": answer}}.

Question: {question}

{context}

Prompt 3

Act as a biomedical expert. You will receive several {d_type} summarizing research findings and methodologies. Along with this, a question will be provided. Your role is to analyze the {d_type} and provide a scientifically accurate, concise answer to the question, leveraging the information from the {d_type}.

First read and understand the relevant information present in the several {d_type}, extracting all relevant facts.

After thinking about the information presented, present a final json containing your answer {"answer": answer}. Please show all your reasoning first.

Answer in around 50–150 words, without using any markdown.

Question: {question}

{context}

Prompt 4

Act as a biomedical expert. You will receive several {d_type} summarizing research findings and methodologies. Along with this, a question will be provided. Your role is to analyze the {d_type} and provide a scientifically accurate, concise answer to the question, leveraging the information from the {d_type}.

First read and understand the relevant information present in the several {d_type}, extracting all relevant facts, explaining all your reasoning.

After thinking about the information presented, present a final json containing your answer {"answer": answer}.

Answer in around 50–150 words, use a concise format only with plain text (no lists or markdown).

Question: {question}

{context}

Prompt 5

Act as a biomedical expert. You will receive several {d_type} summarizing research findings and methodologies. Along with this, a question will be provided. Your role is to analyze the {d_type} and provide a scientifically accurate, concise answer to the question, leveraging the information from the {d_type}.

First read and understand the relevant information present in the several {d_type}, extracting all relevant facts, explaining all your reasoning.

After thinking about the information presented, present a final json containing your answer {"answer": answer}.

Answer in around 50–150 words, use a concise format only with plain text (no lists or markdown).

For example:

Question: "What is the use of P85-Ab?"

Insert your thinking here.

{"answer": "P85-Ab is a promising novel biomarker for nasopharyngeal carcinoma screening."}

Question: {question}

{context}

Summary Prompt 1

Act as a biomedical expert. You will receive multiple answers to a given question. Your task is to analyze these responses, extract all relevant information, and synthesize a concise yet comprehensive final answer (50–150 words, at least one complete sentence).

1. Carefully read and understand the key facts and insights from the provided answers.
2. Thoughtfully evaluate the information to form a well-reasoned conclusion.
3. Present your reasoning step-by-step before delivering the final response.

Finally, output a JSON object in the following format:
{"answer": "<your concise answer>"}

Guidelines:
- Ensure the answer is informative, clear, and medically accurate.
- Do not use Markdown formatting.
- Keep the response within the word limit.

Question: {question}

Answers: {answers}

Summary Prompt 2

Act as a biomedical expert. You will receive multiple answers to a given question. Your task is to analyze these responses, extract all relevant information, and synthesize a concise yet comprehensive final answer (50–150 words, at least one complete sentence).

1. Carefully read and understand the key facts and insights from the provided answers.
2. Thoughtfully evaluate the information to form a well-reasoned conclusion.
3. Present your reasoning step-by-step before delivering the final response.

Finally, output a JSON object in the following format:
{{"answer": "<your concise answer>"}}

Guidelines:
- Ensure the answer is informative, clear, and medically accurate.
- Do not use Markdown formatting.
- Keep the response within the word limit.

For example:

Question: "What is the use of P85-Ab?"
Insert your thinking here.
{{"answer": "P85-Ab is a promising novel biomarker for nasopharyngeal carcinoma screening."}}

Question: {question}

Answers: {answers}

B. Phase A trained models

Table 10
Model configurations trained for Phase A

Model	Size	Seed	Epochs	Sampler	Trainer	ExPOS	Warmup
BioLinkBERT	base	42	4	basic	pairwise	False	False
-	-	100	2	-	-	-	-
-	-	100	2	-	-	True	-
-	-	-	1	-	-	-	-
-	-	-	2	-	-	-	-
-	-	-	2	basicv2	-	-	True
-	-	-	4	-	-	True	-
-	-	-	2	exponential	-	True	-
-	large	100	2	-	-	True	-
-	large	100	4	-	pointwise	True	-
-	large	-	1	-	-	True	-
-	large	-	3	-	pointwise	True	-
BiomedBERT	base	42	4	basic	pairwise	False	False
-	-	100	2	-	-	True	-
-	-	-	1	-	-	-	-
-	-	-	2	-	-	True	-
-	-	-	2	basicv2	-	-	True
-	-	-	2	exponential	-	True	-
-	-	-	2	-	pointwise	True	-
-	large	100	2	-	-	True	-
-	large	100	4	-	pointwise	True	-
-	large	-	3	-	pointwise	True	-
-	large	-	2	-	-	True	-