BIT.UA at BioASQ 12: From Retrieval to Answer Generation

Notebook for the BioASQ Lab at CLEF 2024

Tiago Almeida[1,*], Richard A. A. Jonker[1], João Reis[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 vital for accessing and processing the ever-increasing volume of biomedical data. Effective systems in this domain are essential for researchers, clinicians, and medical experts to make well-informed decisions. The BioASQ Task B Challenge fosters the development of advanced retrieval and question-answering systems by providing a platform for evaluating and comparing diverse approaches.

This paper presents our participation in the twelfth edition of the BioASQ challenge, focusing on Task B. For Phase A, we employed a two-stage retrieval pipeline with the BM25 model from PISA and transformer-based neural reranking models, including PubMedBERT and BioLinkBERT. Additionally, we enhanced BM25 results with semantically similar documents using the BGE-M3 model and augmented the BioASQ training data. Outputs from these models were combined using reciprocal rank fusion (RRF). In Phases A+ and B, we utilized instruction-based transformer models such as Llama 3, Nous-Hermes2-Mixtral, and a BioASQ fine-tuned version of Gemma 2B for conditioned zero-shot answer generation. Our systems in Phase A achieved competitive results, consistently scoring on top or near the top across all batches. In Phases A+ and B, our systems remained competitive, especially in terms of Recall.

Keywords

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

1. Introduction

Biomedical information retrieval and question answering are important and complex tasks, driven by the need to access and process vast amounts of biomedical data. As the volume of published biomedical literature grows exponentially, effective retrieval and accurate question-answering systems are crucial for researchers, clinicians, and medical experts to make informed decisions.

The BioASQ Task B Challenge aims to push the state of the art in retrieval and question-answering systems within the biomedical domain. By providing a platform for evaluating and comparing different approaches, BioASQ encourages the development of innovative algorithms and techniques that can handle the unique challenges posed by growing biomedical data. Participants are tasked with developing systems that can efficiently retrieve relevant documents and snippets, as well as generate accurate answers to biomedical questions.

In this paper, we outline the participation of the Biomedical Informatics and Technologies group of the University of Aveiro (BIT.UA) in the 12th BioASQ challenge [1]. Our team participated in Phase A, Phase B, and the newly added Phase A+. Phase A centered on information retrieval, primarily identifying the top documents responding to a biomedical query. Both Phase A+ and B can be characterized as Retrieval Augmented Generation (RAG) tasks, where the objective is to answer a query using a document that provides relevant context.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

tiagomeloalmeida@ua.pt (T. Almeida); richard.jonker@ua.pt (R. A. A. Jonker); joaoreis16@ua.pt (J. Reis); joao.rafael.almeida@ua.pt (J. R. Almeida); aleixomatos@ua.pt (S. Matos)

0000-0002-4258-3350 (T. Almeida); 0000-0002-3806-6940 (R. A. A. Jonker); 0009-0002-3579-0711 (J. Reis); 0000-0003-0729-2264 (J. R. Almeida); 0000-0003-1941-3983 (S. Matos)

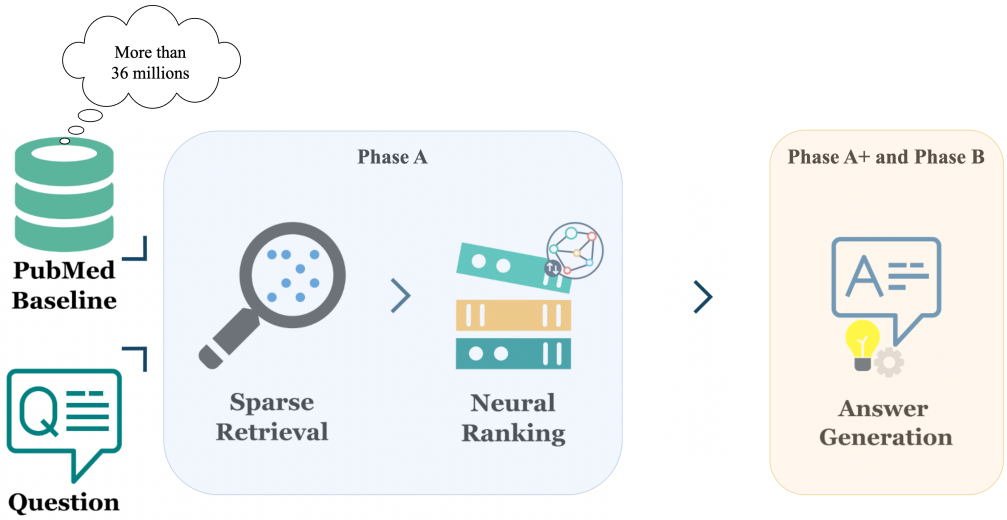

Figure 1: High-level overview of the entire system pipeline concerning the BioASQ tasks.

Figure 1 presents an overview of our solution as an end-to-end system for the BioASQ tasks. For Phase A, we followed the approach of Almeida et al. [2] with a two-stage retrieval pipeline. As a first stage, we used the traditional BM25 model from PISA [3, 4], followed by transformer-based neural reranking models, specifically PubMedBERT [5] and BioLinkBERT [6]. Additionally, we explored a semantic search approach with the BGE-M3 [7] model in two ways: enhancing BM25 results with semantically similar documents and augmenting the BioASQ training data with these documents. We used reciprocal rank fusion (RRF) to ensemble outputs from the various models.

For Phases A+ and B, we employed instruction-based transformer models such as Llama 2 [8] and 3, Nous-Hermes2-Mixtral [9], and a BioASQ fine-tuned version of Gemma 2B [10] for conditioned zero-shot answer generation. Specifically, we utilized the top-5 most relevant articles to generate an ideal answer and explored using relevant snippets in Phase B. Furthermore, we investigated methods to select the optimal answer from a pool of candidates generated by these models.

2. Related Work

Biomedical information retrieval and question answering have made significant strides in recent years, partly due to advancements in Deep Learning and Large Language Models [11], where evaluation platforms like BioASQ have provided a valuable testbed for benchmarking these advancements, continually pushing the state-of-the-art.

Looking at the information retrieval task, two-stage retrieval approaches continue to be the most adopted solutions [12, 13, 14, 15, 16, 17, 2, 18]. These methods combine the efficiency of sparse retrieval models, which reduce the candidate pool using weighted keyword matching, with the efficacy of neural reranking models, which refine the initial list by considering semantic understanding and contextual relevance. BM25 [19] is the most popular choice for sparse retrieval, while BERT-like models [20] are the preferred architecture for neural reranking.

Although less popular, there were some efforts on exploring semantic search approaches, also known as dense retrieval [21, 13, 18], as replacements for sparse retrieval methods. These efforts were mainly motivated by the ability of dense retrieval models to capture semantic similarities between queries and documents, overcoming the limitations of term-based models like BM25, which rely on exact term matches. However, Ma et al. [21] demonstrated that dense retrieval alone was not capable of surpassing BM25 in terms of retrieval performance. Despite its ability to capture semantic similarities, the dense retrieval model struggled with exact lexical matching, which is a strength of BM25. Consequently, combining dense retrieval with BM25 in a hybrid model was found to be more effective, leveraging the

strengths of both approaches to achieve superior results in the BioASQ challenge.

Regarding answer generation, with the advent of Large Language Models (LLMs), most solutions now follow a Retrieval-Augmented Generation (RAG) methodology. In this approach, relevant documents are first retrieved and then used to generate comprehensive and accurate answers. The main differences between the participants' approaches lie in the models they use and how these models are configured. Some systems directly use ChatGPT [22, 23, 24], while others employ open-source models [2, 25], each leveraging distinct configurations to optimize performance.

3. Methodology

This section of the paper provides a detailed description of the corpora and datasets utilized throughout the challenge, including the preprocessing steps undertaken. It then outlines the specific methodologies adopted for each task in which we participated.

3.1. Corpus and Dataset

The BioASQ dataset [26] encompasses questions from the last eleven editions, totalling 5,049 questions. These are classified into four categories: factoid (1,551), yes/no (1,357), summary (1,210), and list (967). Each question is accompanied by a list of relevant documents, snippets (taken form the relevant document), and an example of the ideal answer. For its corpus, the BioASQ challenge utilizes the PubMed/MEDLINE annual baseline. This year, the baseline corresponds to 2024 baseline and includes over 36 million documents.

As evidenced by Almeida et al. [2], the continual updates and removal of documents between each yearly baseline poses a challenge for maintaining consistency in document relevance across different editions of the BioASQ challenge. In other words, documents relevant to questions in earlier editions may not be present in the current edition's document collection. To mitigate this issue and ensure accurate training data, each question in the dataset is marked with the year it was featured. Furthermore, we download the PubMed/MEDLINE baselines from 2013 to 2024 to maintain a clear snapshot of the corpus as it existed when each question was initially posed. This approach enables us to align each question with the specific corpus version available at that time.

In terms of document preprocessing, we observed that some documents in the baselines lacked titles, abstracts, or both. To address this, we simply removed these incomplete documents from the collection.

Regarding the preprocessing of the training dataset, we primarily adopted two approaches. The first, inspired by Almeida et al. [2], focused on maximizing the quality of the dataset, while the second approach concentrated on maximizing the quantity of the dataset.

3.1.1. High-quality Dataset Preprocessing

Under the high-quality perspective, our goal is to ensure with confidence that every question in the dataset is valid, well-written, and correctly matched with relevant documents, even if this results in a reduction of training data.

To achieve this, we thoroughly reviewed the dataset and found instances of repeated or very similar questions paired with different relevant documents. Consequently, we decided to merge similar questions along with their corresponding sets of relevant articles. For efficient and automatic merging, we utilized the pre-trained SimCSE [27] model to calculate the similarity between questions. Questions with a cosine similarity score above 0.99 were automatically merged, while those with a similarity score between 0.90 and 0.99 underwent manual review, in total 43 questions were merged. Furthermore, considering the BioASQ guidelines, systems prior to the fourth edition of BioASQ could use full-text articles from PubMed Central (PMC) for judgment. This could lead to situations where later models, lacking access to full texts, might not have the necessary content to make accurate predictions, making the training data actually incorrect. Therefore, we decided to remove all question pairs from before the fourth edition of BioASQ. Additionally, the capabilities of earlier systems were arguably less sophisticated than those

available today, potentially leading to a less reliable gold standard compared to more recent editions. At the end of this process, the refined dataset comprised 3,795 questions (a reduction of 25%), totalling 28,910 positive question-document pairs.

3.1.2. High-quantity Dataset Preprocessing

On the other hand, in the high-quantity perspective, our aim is to maximize the number of annotated pairs, operating under the assumption that the sheer volume of data will outweigh any minor errors present in the dataset.

Thus, in this approach, we chose to retain all 5,049 questions in the dataset. Nevertheless, we still recognized the need to address the previously mentioned issue concerning the use of full-text abstracts in the early editions of the BioASQ challenge. Specifically, for questions from these early editions, we examined the list of relevant snippets, and any document was considered positive if its relevant snippet was derived from the article's abstract. At the conclusion of this process, the dataset included 5,049 questions, totalling 43,732 positive question-document pairs.

3.2. Phase A

In Phase A, we participated solely in the document retrieval subtask, and this section details the methods we employed. Inspired by the work of Almeida et al. [2], we developed a two-stage retrieval system. The first stage utilizes an efficient sparse retrieval method, followed by neural reranking model as the second stage. To effectively integrate the knowledge from different models, we adopted the reciprocal rank fusion [28] (RRF) method to ensemble the outputs from various models. Additionally, we explored methods to efficiently incorporate semantic search mechanisms into our pipeline, an aspect not extensively explored by previous teams in the challenge.

3.2.1. First stage: Sparse Retrieval

The objective of the first stage is to efficiently retrieve the best k candidate documents that potentially contain an answer to a given question, referred to throughout this document as the top-k documents. To achieve this, we utilized sparse retrievers, specifically the traditional BM25 [19] model from PISA [4, 3], a state-of-the-art text search engine written in C++ that supports advanced WAND-like search algorithms.

During preliminary experiments, we observed that setting k to 1000 guarantees a recall of 90.2%, providing the best balance between efficiency and effectiveness. The parameters for the BM25, specifically k_1 and b, were selected through a preliminary hyperparameter tuning process. Specifically, we conducted a grid search for k_1 values ranging from 0.1 to 1.2 with intervals of 0.1 and b values ranging from 0.1 to 1.0 with intervals of 0.1. The optimal parameters k_1 and b were found to be 0.4 and 0.3, respectively.

3.2.2. Second stage: Transformer-based Reranking

The second stage in our pipeline aims to thoroughly analyse and re-rank the top-1000 candidate documents previously retrieved in the first stage. To accomplish this, we employ a transformer-based cross-encoder architecture as our neural reranker model. This model encodes each question-document pair into a CLS representation, which a classifier then uses to compute the relevance score for each pair. We initialized our neural reranker models using the pretrained PubMedBERT [5] and BioLinkBERT [6] weights, and trained with pointwise (cross-entropy) loss and pairwise (hinge) loss using the Trainer API from HuggingFace.

While this section describes the methods adopted for the second stage, we will now briefly address other avenues that we explored but ultimately disregarded due to their lack of performance. One problem with our previous neural reranker model architecture is that its input size is limited to 512 tokens, which is too short for some question-document pairs, forcing us to truncate the document and potentially lose valuable information. To address this issue, we implemented a sentence-level neural

reranking model that processed any question-document pair regardless of size by splitting the document into individual sentences that fit within the model's size constraints. However, in all of our preliminary experiments, this sentence-level model did not surpass the performance of our other neural reranker models. Additionally, we explored a dynamic training regime where we used the neural reranker model currently during training to mine for negative documents. However, this approach also failed to yield any improvement in our preliminary results.

3.2.3. Adding Semantic Search

Semantic search, also known as dense retrieval, has become increasingly popular for performing information retrieval, particularly due to its capability to address the vocabulary mismatch problem [29]. Despite its potential, semantic search approaches often face performance challenges and struggle to compete, especially in contexts like the BioASQ challenge, where the exact match signal significantly enhances sparse retrieval performance. Furthermore, searching through over 36 million documents poses a substantial computational challenge.

Nonetheless, the value of semantic search is undeniable, especially in identifying documents that would otherwise be missed by sparse models. Inspired by MacAvaney et al. [30], our goal is to integrate semantic search as a complementary method to our sparse retrieval model, rather than as a hybrid method [31]. More precisely, we aim to identify documents semantically similar to the documents ranked higher by our neural re-ranking model, since, according to the Cluster Hypothesis [32], similar documents are likely to be relevant to the same question. After obtaining the semantically similar documents, they are then ranked by our neural reranker model to determine their placement in the final ranking order. We refer to this technique as Dense Pseudo Relevance Feedback (DPRF) and its integration with our two-stage retrieval pipeline can be seen in Figure 2. To efficiently gauge document similarity, we precomputed a similarity graph over the entire 2024 PubMed/MEDLINE database using the BGE-M3 model [7]. We only recorded connections between documents with a cosine similarity exceeding 0.85 to manage storage constraints, as recording all similarities would require petabytes of storage. Notably, by precomputing this graph for any document in the entire collection, we can instantaneously access a list of documents that have a similarity higher than 0.85.

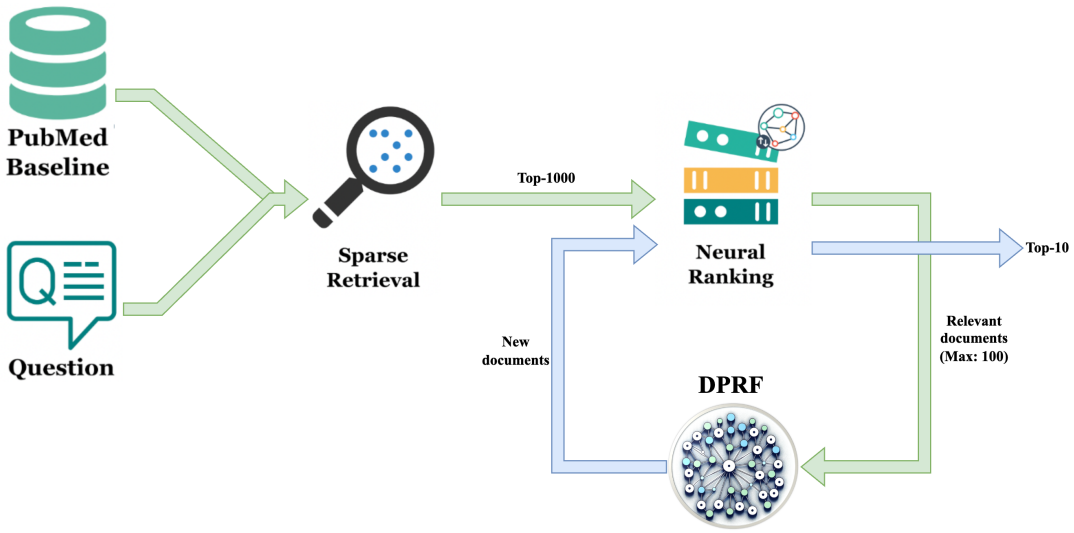

Figure 2: Overview of our two-stage retrieval pipeline using Dense Pseudo Relevance Feedback (DPRF).

Moreover, guided by the Cluster Hypothesis [32] and leveraging the new similarity graph, we propose augmenting our training data by increasing the number of positive documents. To achieve this, for each positive document from the dataset, we consider semantically similar documents with a similarity score above 0.95 as potentially relevant and include them in the training data.

3.3. Phase A+ and B

For phases A+ and B, we participated only in the ideal answer generation subtask. The primary difference between these phases is in the resources provided by the organizers. In phase B, participants are given a list of gold documents and snippets. In contrast, for phase A+, participants must utilize the documents retrieved during phase A. In terms of approach, similar to Almeida et al. [2], we focus primarily on zero-shot answer generation using large language models. Furthermore, given the vast array of available LLMs, we also explored a cost-effective method to determine which answer would be the best.

3.3.1. Answer Generation with LLMs

For answer generation, we primarily relied on the following large language models: Nous-Hermes2-Mixtral (referred to as Mixtral in the remaining of the paper), Llama2 70B, and Llama3 70B, which were released during the competition. Unlike Almeida et al. [2], our approach involved exploring multiple sources of information as context for generating an ideal answer. Specifically, we propose using several abstracts, and in the context of phase B, snippets as the context for these LLMs. Additionally, we encountered an issue with the length of the generated answers, as per competition criteria the maximum length allowed was of 200 words. To comply with this, we included a word limit constraint within the prompts, and preprocessing step of truncation. Below, we present the two main prompt variations used:

```
Act as a biomedical expert. You will receive several abstracts
(‘[abstract: Abstract]’) summarizing research findings and
methodologies. Along with this, a question will be
provided(‘[question]’).

Your role is to analyze the abstract and provide a
scientifically  accurate, concise answer to the question,
leveraging the information  from the abstracts.

[Abstract: CONTEXT]

[Question: QUESTION]

Answer in less than 150 words:
```

Listing 1: Prompt variation 2; detailed structure emphasizing biomedical expertise.

```
Context: CONTEXT

Question: QUESTION

Answer in less than 150 words:
```

Listing 2: Prompt variation 1; basic structure for generating concise answers.

In these prompts, the contexts were provided as a list of up to five abstracts. We observed a benefit in utilizing multiple documents, with five proving to be an optimal balance. This reinforces our idea of using multiple sources of information as context. Exclusively for phase B, we experimented with providing gold snippets instead of full abstracts, based on the rationale that feeding direct answers to the LLM would enable it to rewrite them in natural language. Additionally, we attempted to automatically extract exact answers with LLMs and used these as context for answer generation. However, this approach did not yield successful results, and we subsequently abandoned it.

Besides the zero-shot approaches, we also explored fine-tuning a large language model to produce answers more aligned with the expectations of the BioASQ challenge. For this purpose, we utilized

the unsloth library[1] to fine-tune the Gemma 2B model[2] [10]. Specifically, we used the 4-bit model and trained it using Low Rank Approximation (LoRA) [33]. For the training data, we employed the BioASQ dataset, where we structured prompts containing the question alongside multiple document abstracts with the goal of generating the gold-standard ideal answer. The prompt used to fine-tune the model is as follows:

```
Below is an instruction that describes a task, paired with an
input that provides further context. Write a response that
appropriately completes the request.

### Instruction:
Act as a biomedical expert. You will receive several abstracts
(‘[abstract: Abstract]’) summarizing research findings and
methodologies. Along with this, a question will be
provided(‘[question]’).

Your role is to analyze the abstract and provide a
scientifically  accurate, concise answer to the question,
leveraging the information  from the abstracts.

### Input:
[Abstract: CONTEXT]

[Question: QUESTION]

Answer in less than 200 words:

### Response
```

Listing 3: Prompt used to fine tune the Gemma model.

As previously mentioned, we needed to ensure that our answers contained fewer than 200 words. Despite adding an instruction and slightly tweaking the sampling generation hyperparameters, these adjustments could not reliably produce answers within the 200-word limit. Therefore, we propose a second generation step that we refer to as summarization, where the objective is to take the question and the original lengthy answer as inputs and make the answer more concise, aligning it with the BioASQ guidelines. To accomplish this, we utilized the Nous-Hermes2-Mixtral model with the following prompt:

```
Please provide a short and concise summary based on the
provided answer:

Question: QUESTION

Answer: ANSWER
```

Listing 4: Prompt used to summarise model outputs using Mixtral.

Overall, both task phase A+ and phase B followed the methods described above, however there are some slight variations which will be addressed in the submissions sections.

3.3.2. Answer Selection

Given the variety of prompts and models employed for answer generation, it is essential to establish a robust method for selecting the best possible answer. For this purpose, we propose using our neural re-ranking model from Phase A to select the ‘best’ answer. Intuitively, the model was trained to assess

[1] https://github.com/unslothai/unsloth
[2] https://huggingface.co/unsloth/gemma-2b-it-bnb-4bit

content relevance between a question and a document, making it suitably analogous for assessing the relevance between a question and an answer. In practical terms, for any given question and set of relevant documents, we would generate multiple answers using different combinations of prompts and LLMs. Then, we would feed each question-answer pair into our neural re-ranking model and select the answer with the highest score. This method also proves useful for detecting cases where the answers are hallucinatory or nonsensical.

One potential issue with this approach is the model's tendency to favour longer answers, which can be problematic given the strict word limit constraints of the BioASQ challenge. This bias arises because the model was trained with full abstracts, which are typically more comprehensive and detailed than the concise answers required here.

4. Results

In this section, we start by describing the evaluation metrics used throughout the paper. Then, we introduce some validation results that help us understand the performance of the proposed methods. Finally, we describe our submissions and the preliminary results.

4.1. Evaluation metrics

The BioASQ organizers employ well-known information retrieval metrics for evaluating Phase A systems and text generation metrics for Phases A+ and B. These metrics are computed by comparing the system's predictions against a "gold standard" dataset. At the time of writing, the organizers can only provide a preliminary "gold standard" that was created during the question construction phase. Subsequently, medical experts will carry out a manual judgment of all system predictions, constructing a more complete and final "gold standard". Note that this process can take several months. Therefore, the results presented in this paper are considered preliminary, as they are subject to change upon the release of the final "gold standard". It is also important to note that although the preliminary evaluation does not fully capture the true gold standard, it still offers valuable insights.

According to Malakasiotis et al. [34], the primary metrics for Phase A include Mean Average Precision (MAP) and Precision. MAP assesses the quality of ranked retrieval results by evaluating the precision at various recall levels and averaging across all queries. It is calculated as follows:

$$\text{MAP} = \frac{1}{n} \cdot \sum_{i=1}^{n} AP_i, \tag{1}$$

where AP_i is the average precision of the list returned for the question q_i and is defined as:

$$\text{AP} = \frac{\sum_{r=1}^{|L|} P(r) \cdot \text{rel}(r)}{min(|L_R|, 10)}, \tag{2}$$

where $|L|$ is the number of items, $|L_R|$ is the number of relevant items (typically 10), $P(r)$ is the precision when the returned list is treated as containing only its first r items, and $\text{rel}(r)$ equals 1 if the r-th item in the list is relevant (i.e., in the "gold standard") and 0 otherwise.

In Phases A+ and B, the evaluation focuses on Recall and F1 Score calculated from the ROUGE metric. ROUGE is a metric used to evaluate the quality of summaries by measuring the overlap of the generated summary with "gold standard". There are several versions of ROUGE. ROGUE-N, which uses n-grams to calculate the overlap between the artificial summary S and the reference summary $Refs$ is calculated as follows:

$$\text{ROUGE-N}(S|\text{Refs}) = \frac{\sum_{R \in \text{Refs}} \sum_{g_n \in R} C(g_n, S, R)}{\sum_{R \in \text{Refs}} \sum_{g_n \in R} C(g_n, R)}, \tag{3}$$

where g_n is a word n-gram, $C(g_n, S, R)$ is the number of occurrences of g_n in both S and reference R and $C(g_n, R)$ is the number of occurrences of g_n in R.

ROGUE-SU4, is a variation of ROGUE-N, where S represents skip bigrams as opposed to n-grams when computing the overlaps. The U also counts unigrams, and the 4 means a maximum distance between the words of any skip bigram is limited to 4. We utilize ROGUE-SU4 as the automatic metric in this work. The official metric for this task is human evaluation which includes readability, recall, precision, and repetition.

4.2. Validation Results

In order to deepen our understanding of the proposed methodology and determine how to organize the submissions, we conducted several validation experiments. Specifically, we explored which preprocessing strategy for the dataset—high quality versus high quantity—contributed to better performance. Additionally, we evaluated the capability of the Dense Pseudo Relevance Feedback (DPRF) method to identify documents that were missed by the BM25 retrieval method. Regarding the training of our neural reranking model, we aimed to understand which training methodology, pointwise or pairwise, worked best, the impact of adding semantic data augmentation, and which model initialization checkpoint was most effective. Regarding the validation data, we chose to use the questions and their respective "gold standard" from the first two batches of 2023 (BioASQ 11). The primary reason for selecting these two batches, rather than employing a standard train/validation split, is rooted in the progressive nature of the BioASQ challenge. Over time, participant systems tend to improve, leading to increasingly effective solutions [35]. Consequently, since the "gold standard" is developed based on medical expert judgments of the submissions from these systems, and because better-performing systems provide more accurate and relevant outputs for experts to review, the quality of the "gold standard" is likely to be higher with newer batches.

Regarding the experiments, we began by examining the performance trade-off between high-quality and high-quantity preprocessing methods. To this end, we trained a PubMedBERT-Base model using each resulting dataset and discovered that the high-quality preprocessing yielded better results. Specifically, the MAP from the high-quantity preprocessing was 45.36, while the high-quality preprocessing achieved a MAP of 46.78. In machine learning, data is often viewed on a spectrum, where achieving a balance between quality and quantity is necessary for optimal results. We believe that our high-quality preprocessing approach strikes the best balance here, showing the importance of data quality over quantity.

Now discussing our Dense Pseudo Relevance Feedback (DPRF) method, we initially had concerns that the additional overhead might not yield significant benefits and that the approach might fail to retrieve any new documents beyond those identified by sparse retrieval. To test this, we ran the sparse retrieval (BM25) across the entire training corpus, and then applied DPRF to the documents retrieved by BM25 to see if it could retrieve any of the positive documents that BM25 failed to retrieve. In this experiment, DPRF successfully retrieved an additional 2,326 documents that were not identified by the sparse retrieval. However, there were still 4,508 documents that remained unretrievable[3]. Although this approach does not recover every document, we believe it is promising and has the potential to enhance overall system performance.

One of the last approaches we investigated was the use of either pointwise or pairwise training. We observed that pairwise training slightly enhanced performance on the validation data, achieving a MAP of 45.98 compared to 45.44 with pointwise training. Furthermore, when we added semantic data augmentation, performance significantly improved for both approaches: 49.53 for pairwise and 49.16 for pointwise.

Lastly, we present the general performance of the various model checkpoints tested. We only show the validation results for pairwise models, however we reached similar conclusions for the pointwise models. The values for the models are as follows:

- BioLinkBERT-Large: 49.53
- BioLinkBERT-Base: 48.45

[3]Note that this can also include documents that are no longer available, as DPRF only uses the 2024 Baseline

- PubMedBERT-large: 48.26
- PubMedBERT-Base: 46.53

As expected, the large models outperformed the base counterparts, while the BioLinkBERT seems to have the overhand against the PubMedBERT model.

4.3. Phase A Results

A summary of the runs submitted for phase A can be seen in Table 1, with a more detailed description of the runs presented in Appendix A. It is important to mention that some of the methods previously described were only implemented in between batches. This is why we frequently changed the configuration of the runs we submitted during phase A. The results of the submissions can be seen in Table 2.

Table 1
Summary of the systems submitted for phase A. Each system is denoted with a structure specifying data source (Q for high-quality, T for high-quantity), model type (BL for BioLinkBERT, PM for PubMedBERT), model size (L for Large, B for Base). "2023" refers to models trained in 2023, and "All" refers to an RRF ensemble of all the previous runs. Furthermore, each system is an RRF ensemble, with the total number of runs indicated in parentheses. Note that in Batch 2, system-0 was trained with high-quantity data (T) instead of normal high-quality and DPRF is applied where specified. Batch 4 includes submissions with pairwise models and models trained over validation data. Generally, the final system in all batches uses an RRF ensemble of all runs, with Batch 3 switching to an ensemble of systems, and Batch 4 including 2023 model submissions

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	Q: BL(L3)	T: BL(L2) + PM(B2)	Q: BL(L5+B2) + PM(L2+B6)	Q: BL(L5+B2) + PM(L2+B6)
system-1	Q: BL(L3) + PM(L3)	System-0 + DPRF	System-0 + DPRF	System-0 + DPRF
system-2	Q: BL(B3) + PM(B3)	Q: BL(L1+B1) + PM(L1+B2)	Q: PM(B5)	Q: BL(L3+B5) + PM(B3) (Pairwise)
system-3	2023 (16)	2023 (16)	2023 (16)	Q: BL(L3+B5) (Best on validation)
system-4	All (28)	All (29)	All (4)	All+2023 (5)

Table 2
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	18.00	11.51	5	**22.17**	10.12	4	24.87	9.80	4	37.52	**12.47**	6
system-1	20.18	11.56	3	21.10	9.43	5	**24.93**	9.56	3	37.73	12.39	5
system-2	20.06	**12.94**	4	20.41	**10.85**	9	24.27	**9.82**	5	36.90	10.47	8
system-3	20.24	10.09	2	19.47	9.09	11	21.46	8.89	10	37.40	10.82	7
system-4	**20.67**	10.39	1	20.96	8.72	6	24.15	9.10	6	**39.03**	10.47	3
Best Competitor	16.12	7.06	6	22.93	9.53	1	25.49	8.59	1	39.30	10.00	1
Median	10.74	7.18	20	11.51	7.18	25	12.50	6.40	29	17.69	8.24	24

Regarding the results shown in Table 2, on the first batch, we trained only four models: PubMedBERT Base, PubMedBERT Large, BioLinkBERT Base, and BioLinkBERT Large. We utilized three checkpoints from these models for the submissions, combined with our models from last year (2023). These systems obtained the top five results for the first batch. From this batch, we observed that the ensemble of all runs performed best, and the large models and small models achieved similar performance (systems 1 and 2). However, relatively speaking, we believe that the systems submitted for this batch were weaker compared to other batches, as we did not have many models to create a robust ensemble. Consequently, our 2023 models outperformed our newly trained models in this first batch.

For the second batch, we focused on identifying the impact of training with high-quality vs high-quantity data. Here, system-0, which used the high-quantity training data, performed the best, contradicting our validation results. However, we must emphasise that these results are still preliminary, and

the final rankings will likely change. Erroneously, the models of system-0 were trained for 10 epochs, while the models trained with high-quality data were only trained for 5 epochs, which may cause the discrepancy in the results observed. Furthermore, we observed that within this batch, our newly trained models outperformed those from 2023.

Within the third batch, we applied the knowledge from our previous batches and trained several more models, varying the seeds and other model parameters. Notably, by using a larger number of models, 15 in case of system-0, it seems it contributed to achieve highly competitive results. Surprisingly, on this batch, the 2023 models were drastically outperformed by our newly trained models.

Compared to batch 3, batch 4 provided many similar runs, maintaining both systems 0 and 1, and adding a run with pairwise models, which were not able to outperform the pointwise models. Additionally, we included a submission containing our best performing models on validation. Our best model from this batch was an ensemble of all our submissions combined with the 2023 models, which outperformed our other models significantly.

Finally, we discuss the performance of the Dense Pseudo Relevance Feedback (DPRF) method on the submissions. As shown in the results table, DPRF outperformed its counterpart in batches 3 and 4 but failed to achieve better results in batch 2. We believe this is because batch 2 was when we first introduced the DPRF method, and at that time, we did not have a strong intuition on how to make it work properly. This understanding was explored and improved for batches 3 and 4. Additionally, it is important to note that this preliminary evaluation may not favour the DPRF method. We believe that the preliminary "gold standard" constructed by the organizers may be biased towards documents with more exact matches between the question and the documents.

Overall, in all batches, our systems outperformed the median in both metrics, obtaining the best MAP in batch 1, while being within less than one percentage point of the best submission, and achieving the best precision in batches 1, 3, and 4.

4.4. Phase A+ Results

A summary of the runs submitted for phase A+ is presented in Table 3, with a more detailed description of the runs available in Appendix B. A major challenge here was deciding which source of positive documents from Phase A to use, as it was crucial to maintain consistency in the input source for a fair comparison of the LLMs' performance when generating answers. The results of these submissions are shown in Table 4.

Table 3
Summary of the systems submitted for phase A+. The first number corresponds to the system number used as input source for the answer (e.g., 0 corresponds to system-0 from phase A, All corresponds to an ensemble using all sources), L refers to llama (either 2 or 3), M refers to the mixtral model, G refers to Gemma, and summ. refers to summarization using mixtral. The number in parentheses refers to how many runs are present in the ensemble, if no value is specified, 1 is assumed.

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	0 - L2 + M (4)	4 - L2 + M (3)	1 - L3 + M (4)	1 - L3 + M (4)
system-1	1 - L2 + M (4)	All - M (8)	2 - L3 + M (4)	2 - L3 + M (4)
system-2	4 - L2 + M (3)	All - L2 + M (8)	4 - L3 + M (4)	4 - L3 + M (4)
system-3	All - M (8)	All - Top 5 G (5)	4 - L3 + summ.	4 - L3 + summ.
system-4	All - L2 + M (8)	All - Top 1 G (5)	4 - L3 + summ.	4 - L3 + summ.

Also, it is important to note that between batches 2 and 3, the Llama 3 model was released as the successor to Llama 2. This release prompted us to switch from Llama 2 to Llama 3 for the remaining batches.

For the first batch submission, our primary goal was to understand the impact of the input source documents on our downstream task of answer generation. Upon reviewing the results, it appears that the source documents did not significantly impact performance, as almost all systems scored

Table 4
Performance metrics for various systems across different batches. The metrics reported is the ROGUE-SU4 recall and F1, where the rank is ordered by recall. Bold values represent our best submission.

System	Batch 1			Batch 2			Batch 3			Batch 4		
	REC	F1	Rank	REC	F1	Rank	REC	F1	Rank	REC	F1	Rank
system-0	33.21	6.79	4	28.85	8.15	3	**34.73**	11.21	1	32.19	9.83	2
system-1	**33.43**	7.46	1	28.98	8.06	2	32.83	10.44	4	29.84	9.13	3
system-2	31.39	**7.74**	5	**29.27**	7.78	1	34.04	10.75	2	**32.61**	10.96	1
system-3	33.36	7.30	2	22.92	**18.98**	10	24.45	12.26	14	23.11	**11.77**	18
system-4	33.24	6.93	3	17.06	13.81	20	27.11	**13.55**	9	22.19	11.03	19
Best Competitor	28.07	15.72	6	24.50	10.24	4	33.34	14.70	3	27.95	24.31	3
Median	25.29	11.52	11	19.59	11.04	13	24.53	15.84	14	23.90	13.08	14

closely. The only exception was system-2, which achieved a slightly lower ROUGE-SU4 (Recall) score but compensated with the best ROUGE-SU4 (F1) score. This outcome gives us more flexibility in the upcoming batches to focus primarily on the performance of the LLMs.

For Batch 2, we primarily focused on testing the performance of the fine-tuned Gemma models. We discovered that these models benefited from having access to more source documents. However, their recall metrics were significantly worse than our other models, although their F1 scores showed the best performance within the batch. Unfortunately, during the competition, we misinterpreted the Gemma results as underperforming, which led us to not use the fine-tuned approaches in the upcoming batches.

For Batches 3 and 4, we maintained the same submissions but shifted our focus to the summarization technique. According to the metrics, the systems employing summarization achieved higher ROUGE-SU4 (F1) scores at the expense of ROUGE-SU4 (Recall). We attribute this to the summarization producing shorter sentences, which naturally would decrease the recall while potentially increasing the precision in case of correct answers.

Discussing the results within the competition across all batches, our models consistently achieved the best ROUGE-SU4 (Recall) scores. However, these models generally underperformed in terms of the ROUGE-SU4 (F1) metric, likely due to our bias towards generating longer answers. The only exception to this trend were the fine-tuned Gemma models, which produced higher F1 scores but significantly lower recall, which should be further investigated in future works.

4.5. Phase B Results

A summary of the runs submitted for phase B can be seen in Table 5, with a more detailed description of the runs presented in the Appendix C. The results of the submissions can be seen in table 6. In this phase, we conducted experiments similar to those in Phase A+, with the primary difference being the source of input. Unlike in Phase A+, where inputs were derived from our retrieval models from Phase A, for Phase B, we used the "gold standard" abstracts and snippets provided by the organizers.

Table 5
Summary of the systems submitted for phase B. The first value corresponds to the source of answer generation either abstract (Abs.), snippet (Snipp.), or exact answer (EA). L refers to llama (either 2 or 3), M refers to the mixtral model, G refers to Gemma (either using Top 1 abstract, Top 5 or both), and summ. refers to summarization using mixtral. The number in parentheses refers to how many runs are present in the ensemble, if no value is specified, 1 is assumed.

System	Batch 1	Batch 2	Batch 3	Batch 4
system-0	Abs - L2 (2)	Abs - L2+M (4)	Snipp. - L3+M (4)	Snipp. - L3+M (4)
system-1	Abs - L2+M (4)	Snipp. - L2+M (4)	summ(system-0)	summ(system-0)
system-2	Snipp. - L2+M (4)	Abs - Top 1 G	Abs - L3+M (4)	Abs - L3+M (4)
system-3	EA - L2+M (2)	Abs - Top 5&1 G (2)	summ(system-2)	summ(system-2)
system-4	All (16)	All (10)	Snipp.+Abs - L3	Snipp.+Abs - L3

Table 6
Performance metrics for our submissions for phase B. The metrics reported is the ROGUE-SU4 recall and F1, where the rank is ordered by recall. Bold values represent our best submission.

System	Batch 1			Batch 2			Batch 3			Batch4		
	REC	F1	Rank	REC	F1	Rank	REC	F1	Rank	REC	F1	Rank
system-0	38.63	8.96	4	35.96	11.04	9	**49.72**	17.12	6	**45.13**	**16.59**	4
system-1	38.71	8.45	3	**39.84**	13.23	4	32.22	18.07	24	29.47	15.02	29
system-2	**41.19**	**9.47**	1	14.01	11.89	35	48.27	16.38	7	42.33	14.53	6
system-3	28.97	8.84	21	19.67	**16.34**	31	32.07	16.53	26	26.73	14.04	31
system-4	36.67	8.65	8	35.22	11.34	12	33.99	**19.58**	22	28.28	15.33	30
Best Competitor	39.24	13.69	2	44.62	40.35	1	52.25	36.43	1	49.05	21.44	1
Median	31.91	19.11	16	27.34	17.31	19	35.27	23.30	20	34.90	21.60	21

Similar to Phase A+, in the first batch, we primarily focused on understanding the impact of the input source on the models, examining abstracts, snippets, and exact answers. The performance of the snippets was particularly strong compared to the entire abstracts. Additionally, when comparing system-0 and system-1, we observed that incorporating the Mixtral models into the ensemble proved beneficial. Furthermore, as mentioned, we experimented with using exact answers as our source. This method involved using an LLM for zero-shot extraction of exact answers. However, this approach was not very successful, leading us to discard it in subsequent batches and not include it in this manuscript.

In the second batch, we primarily tested the performance of the fine-tuned Gemma model. Similar to Phase A+, this model exhibited lower ROGUE-SU4 (Recall) but higher ROGUE-SU4 (F1) scores compared to the Llama and Mixtral models. However, given that the ROGUE-SU4 (F1) score was significantly lower than that of the best-performing model, we opted not to continue investigating this model in Phase B. Once again, the snippet submissions outperformed the abstract submissions, showing their effectiveness.

In batches 3 and 4, we maintained the same submission strategy, with an increased emphasis on summarization techniques and more submissions utilizing snippets. We observed that these summarization techniques tend to improve ROUGE-SU4 (F1) scores at the expense of ROUGE-SU4 (Recall). Ideally, these techniques should undergo human evaluation before drawing definitive conclusions. Additionally, we found that snippets slightly outperformed abstracts alone; however, the smaller context fed into the models led to reduced inference time. Our final submission in these batches did not utilize an ensemble of all models but instead employed the newer Llama3 model with snippets, which resulted in surprisingly low performance. This outcome suggests that using both Llama and Mixtral models in an ensemble indeed enhances performance.

In all batches, our best models were obtained by using snippets, obtaining a top 1 recall in Batch 1. Similarly to phase A+ the Gemma model obtained a higher ROUGE-SU4 (F1), however in comparison to the competition, the ROUGE-SU4 (F1) was still below the median of the competition, which is why we focus more on our models which were performing well on recall. Comparing the results of phase A+ to phase B, we can see a significant improvement, indicating that the gold standard documents plays an impact in the generation.

5. Error Analysis on Answer Generation

In this section, we aim to provide more insights into the real performance of the language models used in phase A+ and phase B. For that, we start by showing, in Figure 3, examples of answer from our strategies and models previously described in the methodology section:

Figure 3: Output from various models for the question: “Are FOLFIRINOX plus Bevacizumab the preferred therapy for left-sided RAS wild type advanced colorectal cancer with aggressive clinical presentation?”.

Llama 2 - prompt 1

No, FOLFOXIRI-bevacizumab is not the preferred therapy for left-sided RAS wild-type advanced colorectal cancer with an aggressive clinical presentation. The study showed that FOLFOXIRI-bevacizumab achieved numerically superior survival outcomes compared to FOLFOX-panitumumab, but the difference was not statistically significant. Additionally, FOLFOXIRI-bevacizumab was associated with a higher incidence of chemotherapy-related adverse events, particularly grade 3 and 4 neutropenia. Therefore, doublet chemotherapy plus anti-EGFRs remains the preferred treatment for patients with left-sided RAS/BRAF wild-type mCRC, as it provides similar outcomes with less toxicity.

Llama 2 - prompt 2

No, FOLFIRINOX plus Bevacizumab is not the preferred therapy for left-sided RAS wild-type advanced colorectal cancer with an aggressive clinical presentation. The study showed that FOLFOXIRI-bevacizumab and FOLFOX-panitumumab have similar efficacy in terms of progression-free survival and overall survival, but FOLFOXIRI-bevacizumab had a numerically superior survival outcome. However, the study also showed that FOLFOXIRI-bevacizumab had more frequent chemotherapy-related adverse events, specifically grade 3 and 4 neutropenia. Therefore, the choice between the two regimens should be based on patient preference and potential impact on quality of life.

Mixtral summarization of Llama 2 - prompt 1

No, FOLFIRINOX plus Bevacizumab is not the recommended therapy for aggressive left-sided RAS wild type advanced colorectal cancer due to similar survival rates but higher incidence of adverse events compared to doublet chemotherapy and anti-EGFR treatment.

Mixtral summarization of Llama 2 - prompt 2

Short Summary: FOLFOXIRI-bevacizumab is not considered superior to doublet chemotherapy plus anti-EGFRs for left-sided RAS/BRAF wild-type metastatic colorectal cancer with aggressive clinical presentation due to numerically better survival outcomes but no statistical significance and higher toxicity. Therefore, the preferred treatment remains doublet therapy plus anti-EGFRs for such cases.

Gemma - Top 5

No, FOLFIRINOX plus Bevacizumab is not the preferred therapy for left-sided RAS wild type advanced colorectal cancer with aggressive clinical presentation.

Gemma - Top 1

Yes, FOLFIRINOX plus Bevacizumab is the preferred therapy for left-sided RAS wild type advanced colorectal cancer with aggressive clinical presentation.

Mixtral - prompt 1

The text does not provide a direct answer to this question, as it focuses on comparing FOLFOXIRI-bevacizumab and FOLFOX-panitumumab in left-sided RAS/BRAF wild-type mCRC. However, it does mention that initial therapy with FOLFOXIRI-bevacizumab is superior to doublets plus bevacizumab independently from primary tumor sidedness and RAS/BRAF status. FOLFIRINOX is a similar regimen to FOLFOXIRI, and both include all three drugs (5-fluorouracil, leucovorin, and irinotecan) but differ in the administration of oxaliplatin (given in FOLFOXIRI and not in FOLFIRINOX).
Considering this information, it is plausible to consider FOLFIRINO" *(truncated)*

Mixtral - prompt 2

Based on the abstract provided, the combination of FOLFOXIRI-bevacizumab (which is similar to FOLFIRINOX) and panitumumab were compared in patients with left-sided RAS/BRAF wild-type metastatic colorectal cancer (mCRC). The study found no significant difference in progression-free survival, overall survival, overall response rate, or disease control rate between the two treatment groups. However, there was a numerically superior survival outcome for FOLFOXIRI-bevacizumab, and it achieved a higher resection rate of metastases.
Regarding toxicity, more chemotherapy-related adverse events were observed in the FOLFOXIRI-bevacizumab group, particularly grade 3 and 4 neutropenia. The abstract does not specifically mention aggressive clinical presentation as a criterion for treatment selection. Nevertheless, both treatments appear to *(truncated)*

Looking first at an overview of all the models, we notice that Gemma provides the shortest outputs, offering very brief answers to questions, while sometimes lacking important information. After Gemma, we have the summarization technique which also provides short answers, while still maintaining some relevant information. Following this, it can be seen that both Mixtral and Llama provide similar length outputs, with Mixtral occasionally producing shorter responses.

Looking more closely at each of the systems, starting with Gemma, we can see an important difference between using only one document and using five documents as sources. The answer to the question completely changed, indicating that more information is often needed to answer the questions. We also observed that the Gemma models often output insufficient information, typically lacking a detailed answer justification.

Next, examining the Llama outputs, we observe very detailed answers, clearly explaining the responses to the questions. We also see that the prompt variation had limited impact, showing the robustness of the model. Looking now at the summaries created by Mixtral, we can see that it may offer some of the most valuable outputs. However, it may create some additional artifacts, such as “Short Summary”, though the approach appears promising.

Finally, examining the performance of the Mixtral models, we can see that they lack some consistency and are prone to several problems. In both cases, the output generation was not actually complete; multiple paragraphs were generated, and a direct answer to the question was not directly stated. This indicates that Mixtral may not be the best solution here, unlike what was demonstrated within previous results.

6. Conclusion

In this paper, we detail our team’s participation in the twelfth edition of the BioASQ challenge. Overall, our team’s performance was competitive, achieving top results in each of the tasks we participated in. Specifically, within phase A, we highlight our use of DPRF, which yielded promising results, as well as the addition of documents from semantic search to the training data. For the tasks in phases A+ and B, we presented a fine-tuned Gemma model and a summarization technique, which may pave the way for interesting future work. More specifically, within Phase B, we emphasize that the use of snippets significantly increases the performance of the models compared to those using the full abstract.

Acknowledgments

This work was funded by the Foundation for Science and Technology (FCT) in the context of the project doi.org/10.54499/UIDB/00127/2020. Tiago Almeida is funded by the grant doi.org/10.54499/2020.05784.BD. Richard A. A. Jonker is funded by the grant PRT/BD/154792/2023. This work was funded by FCT I.P. under the project Advanced Computing Project 2023.10766.CPCA.A0, platform Vision at University of Évora.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] T. Almeida, R. A. A. Jonker, R. Poudel, J. M. Silva, S. Matos, BIT.UA at BioASQ 11B: Two-Stage IR with Synthetic Training and Zero-Shot Answer Generation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 37–59. URL: https://ceur-ws.org/Vol-3497/paper-004.pdf.

[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] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[9] Teknium, theemozilla, karan4d, huemin_art, Nous hermes 2 mistral 7b dpo, 2024.

URL: https://huggingface.co/NousResearch/Nous-Hermes-2-Mistral-7B-DPO.

[10] G. DeepMind, Gemma: Open weights llm from google deepmind, https://github.com/google-deepmind/gemma, 2024. Accessed: 2024-05-29.

[11] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, S. Yu, Biomedical question answering: A survey of approaches and challenges, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3490238. doi:10.1145/3490238.

[12] T. Almeida, S. Matos, BIT.UA at bioasq 8: Lightweight neural document ranking with zero-shot snippet retrieval, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_161.pdf.

[13] J. Lu, J. Ma, K. B. Hall, Zero-shot hybrid retrieval and reranking models for biomedical literature, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 281–290. URL: https://ceur-ws.org/Vol-3180/paper-19.pdf.

[14] T. Almeida, S. Matos, Universal passage weighting mecanism (UPWM) in bioasq 9b, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 196–212. URL: https://ceur-ws.org/Vol-2936/paper-13.pdf.

[15] D. Pappas, R. McDonald, G.-I. Brokos, I. Androutsopoulos, Aueb at bioasq 7: Document and snippet retrieval, in: P. Cellier, K. Driessens (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2020, pp. 607–623.

[16] T. Almeida, A. Pinho, R. Pereira, S. Matos, Deep learning solutions based on fixed contextualized embeddings from pubmedbert on bioasq 10b and traditional IR in synergy, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 204–221. URL: https://ceur-ws.org/Vol-3180/paper-12.pdf.

[17] M. Lesavourey, G. Hubert, Bioasq 11b: Integrating domain specific vocabulary to bert-based model for biomedical document ranking, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 145–151. URL: https://ceur-ws.org/Vol-3497/paper-012.pdf.

[18] A. Shin, Q. Jin, Z. Lu, Multi-stage literature retrieval system trained by pubmed search logs for biomedical question answering, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 178–189. URL: https://ceur-ws.org/Vol-3497/paper-016.pdf.

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

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

[21] J. Ma, I. Korotkov, K. B. Hall, R. T. McDonald, Hybrid first-stage retrieval models for biomedical literature, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF

2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_92.pdf.

[22] C. Hsueh, Y. Zhang, Y. Lu, J. Han, W. Meesawad, R. T. Tsai, NCU-IISR: prompt engineering on GPT-4 to stove biological problems in bioasq 11b phase B, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 114–121. URL: https://ceur-ws.org/Vol-3497/paper-009.pdf.

[23] S. Ateia, U. Kruschwitz, Is chatgpt a biomedical expert?, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 73–90. URL: https://ceur-ws.org/Vol-3497/paper-006.pdf.

[24] H. Kim, H. Hwang, C. Lee, M. Seo, W. Yoon, J. Kang, Exploring approaches to answer biomedical questions: From pre-processing to GPT-4, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 132–144. URL: https://ceur-ws.org/Vol-3497/paper-011.pdf.

[25] D. Galat, M. Rizoiu, Enhancing biomedical text summarization and question-answering: On the utility of domain-specific pre-training, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 102–113. URL: https://ceur-ws.org/Vol-3497/paper-008.pdf.

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

[27] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. URL: https://aclanthology.org/2021.emnlp-main.552. doi:10.18653/v1/2021.emnlp-main.552.

[28] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[29] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The vocabulary problem in human-system communication, Commun. ACM 30 (1987) 964–971. URL: https://doi.org/10.1145/32206.32212. doi:10.1145/32206.32212.

[30] S. MacAvaney, N. Tonellotto, C. Macdonald, Adaptive re-ranking with a corpus graph, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1491–1500. URL: https://doi.org/10.1145/3511808.3557231. doi:10.1145/3511808.3557231.

[31] Y. Luan, J. Eisenstein, K. Toutanova, M. Collins, Sparse, Dense, and Attentional Representations for Text Retrieval, Transactions of the Association for Computational Linguistics 9 (2021) 329–345. URL: https://doi.org/10.1162/tacl_a_00369. doi:10.1162/tacl_a_00369. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00369/1924040/tacl_a

[32] O. Kurland, The cluster hypothesis in information retrieval, in: M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, K. Hofmann (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2014, pp. 823–826.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

[34] P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, A. Nentidis, Evaluation measures for task b, In *BioASQ-EvalMeasures-taskB (Version 1.1)*. Intelligent Information Management, Targeted

Competition Framework, ICT-2011.4.4(d), Project FP7-318652 / BioASQ, 2020. Retrieved from http://www.bioasq.org.
[35] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of bioasq tasks 11b and synergy11 in CLEF2023, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 19–26. URL: https://ceur-ws.org/Vol-3497/paper-003.pdf.

A. Phase A runs

A.1. Batch 1

All models in this batch were trained in a pointwise fashion.

- **system-0**: Ensemble of 3 BioLinkBERT-Large checkpoints.
- **system-1**: Ensemble of 3 BioLinkBERT-Large checkpoints and 3 PubMedBERT-Large checkpoints (6 total).
- **system-2**: Ensemble of 3 BioLinkBERT-Base checkpoints and 3 PubMedBERT-Base checkpoints (6 total).
- **system-3**: Our 2023 submission excluding some models (16 total)
- **system-4**: Ensemble of all (28 total)

A.2. Batch 2

All newly trained models were trained for 10 epochs in a pointwise fashion. we were testing our new training data source.

- **system-0**: Ensemble of 2 BioLinkBERT large models and two PubMedBERT Base models with varying seeds, trained with "new data" (4 total).
- **system-1**:DPRF applied to each system-0 submission.
- **system-2**: Ensemble of models trained on the previous data, PubMedBERT-Base, Large and BioLinkBERT-Base, Large. two PubMedBERT-Base modesl totaling 5 models.
- **system-3**: Our 2023 submission excluding some models (16 total)
- **system-4**: Ensemble of all (29 total)

A.3. Batch 3

All models utilize old data from this batch onwards.

- **system-0**: Ensemble of 15 models: BiolinkBERT-Large x5, BiolinkBERT-Base x2, PubMedBERT-Large x2, PubMedBERT-Base x6.
- **system-1**: DPRF applied to each system-0 submission.
- **system-2**: Ensemble of 5 PubMedBERT-Base models.
- **system-3**: Our 2023 submission excluding some models (16 total)
- **system-4**: Ensemble of each submission file (the 4 files above)

A.4. Batch 4

- **system-0**: Same as batch 3 system-0.
- **system-1**: DPRF applied to each system-0 submission.
- **system-2**: Ensemble of 11 pairwise models: BiolinkBERT-Large x3, BiolinkBERT-Base x5, PubMedBERT-Base x3.
- **system-3**: Our top 8 models over validation: BiolinkBERT-Large x3, BiolinkBERT-Base x5, 7 of these models were trained in a pairwise manner.
- **system-4**: Ensemble of all the runs, combined with 2023 runs (5 total)

B. Phase A+ runs

B.1. Batch 1

- **system-0**: This run utilized an ensemble of the two models and two prompts from the first submission (system-0) of Phase A.
- **system-1**: This run utilized an ensemble of the two models and two prompts from the second submission (system-1) of Phase A.
- **system-2**: This run utilized an ensemble of the two models and two prompts from the fifth submission (system-4) of Phase A.
- **system-3**: This run utilized an ensemble of all Mixtral models and two prompts.
- **system-4**: This run utilized an ensemble of all runs (18 total).

B.2. Batch 2

- **system-0**: This run utilized an ensemble of the two models and two prompts from the fifth submission (system-4) of Phase A.
- **system-1**: This run utilized an ensemble of all Mixtral models and two prompts (8 total).
- **system-2**: This run used an ensemble of all runs from either Mixtral or Llama models (20 total).
- **system-3**: This run was an ensemble of Gemma with the top 5 sources per question from each source from Phase A.
- **system-4**: This run was an ensemble of Gemma with the best sources per question from each source from Phase A.

B.3. Batch 3+4

- **system-0**: This run utilized an ensemble of the two models and two prompts from the second submission (system-1) of Phase A.
- **system-1**: This run utilized an ensemble of the two models and two prompts from the third submission (system-2) of Phase A.
- **system-2**: This run utilized an ensemble of the two models and two prompts from the fifth submission (system-4) of Phase A.
- **system-3 and system-4**: These runs utilized an ensemble of the two prompts from the fifth submission (system-4) of Phase A, using only Llama, with Mixtral summarization.

C. Phase B runs

C.1. Batch 1

- **system-0**: This run utilized an ensemble of 2 prompts using the llama2 model, with abstracts as sources.
- **system-1**: This run utilized an ensemble of 2 prompts using both the llama2 + mixtral model model, with abstracts as sources.
- **system-2**: This run utilized an ensemble of 2 prompts using both the llama2 + mixtral model model, with snippets as sources.
- **system-3**: Using exact answers to generate an answer.
- **system-4**: Ensemble of all submissions (16 total)

C.2. Batch 2

- **system-0**: This run utilized an ensemble of 2 prompts using both the llama2 + mixtral model model, with abstracts as sources.

- **system-1**: This run utilized an ensemble of 2 prompts using both the llama2 + mixtral model model, with snippets as sources.
- **system-2**: Utilised the fine-tuned gemma model, with a single abstract.
- **system-3**: This run utilized an ensemble of two gemma submissions, one with a single abstract and one with multiple abstracts.
- **system-4**: Ensemble of all submissions (10 total)

C.3. Batch 3 + 4

- **system-0**: This run utilized an ensemble of 2 prompts using both the llama3 + mixtral model model, with snippets as sources.
- **system-1**: This run utilized an ensemble of 2 prompts using both the llama3 + mixtral model model, with snippets as sources, with a mixtral summary applied.
- **system-2**: This run utilized an ensemble of 2 prompts using both the llama3 + mixtral model model, with abstracts as sources.
- **system-3**: This run utilized an ensemble of 2 prompts using both the llama3 + mixtral model model, with abstracts as sources, with a mixtral summary applied.
- **system-4**: This run utilized an ensemble of 2 prompts using only the llama3 model, with abstracts and snippets as sources,