SINAI at BioASQ@CLEF 2025: A Multi-Stage RAG Pipeline for Biomedical Semantic Question Answering

CLEF2025 BioASQ Task 13 Synergy Working Note

Sara Dueñas-Romero*, L. Alfonso Ureña-López and Eugenio Martínez-Cámara

SINAI - Research Group. Center for Advanced Studies in ICT (CEATIC). Universidad de Jaén

Abstract

This working note presents a multi-stage Retrieval-Augmented Generation (RAG) pipeline for the challenging BioASQ Task 13 Synergy, which focuses on biomedical semantic Question Answering with interactive expert feedback and no dedicated training data. Our system employs question analysis with Named Entity Recognition (NER) for query enhancement, dynamic FAISS indexing with S-PubMedBert for relevant document and snippet retrieval, and a biomedical fine-tuned Llama-based LLM with few-shot prompting for generating exact and 'ideal' answers. Evaluations in Round 4 highlighted our top performance in snippet retrieval. While document retrieval improved with updated PubMed data, exact answer performance varied by question type. For 'ideal' answers, manual expert evaluation favored our optimized system despite automatic metrics suggesting otherwise. Error analysis revealed areas for future improvement, including inference strategies and answer granularity. Overall, our results demonstrate the effectiveness of combining updated retrieval, entity-driven queries, and tuned LLM prompting for biomedical QA in this interactive setting.

Keywords

natural language processing, retrieval augmented generation, question answering, large language models.

1. Introduction

The Conference and Labs of the Evaluation Forum (CLEF) 2025 [1] continues its established tradition (since 2010) of fostering advancements in multilingual and multimodal information access through a peer-reviewed conference and a series of rigorous evaluation labs and workshops. Within this framework, the BioASQ Task Synergy [2] emerges as a significant and timely challenge within biomedical semantic Question Answering (QA), specifically targeting the retrieval and synthesis of information concerning evolving issues in biomedicine.

Distinct from conventional QA tasks that heavily rely on static training data, BioASQ Synergy 13 adopts a novel, interactive approach centered on incremental expert feedback provided across four rounds of evaluation. Participants are tasked with developing systems capable of answering a diverse set of English biomedical questions (yes/no, factoid, list, and summary) by retrieving relevant PubMed articles, extracting pertinent snippets, and generating both exact and summary 'ideal' answers. Notably, this iteration of Synergy does not provide a dedicated training set, making the expert feedback mechanism crucial for system refinement. The test dataset [3] comprises 74 questions with a distribution of 31% yes/no, 26% list, 24% summary, and 19% factoid questions.

To address this challenging task, we propose a system grounded in a Retrieval-Augmented Generation (RAG) architecture. Our approach begins with a preliminary question analysis that developed into an entity extraction methodology for query enhancement, which we utilize to formulate queries for the PubMed API to efficiently retrieve relevant document identifiers, thereby overcoming the data volume challenge.

Subsequently, we employ a reranking strategy with semantic searching for effective document and

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

*Corresponding author.

sduenas@ujaen.es (S. Dueñas-Romero); laurena@ujaen.es (L. A. Ureña-López); emcamara@ujaen.es (E. Martínez-Cámara)
0009-0006-3434-4815 (S. Dueñas-Romero); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-5279-8355 (E. Martínez-Cámara)

snippet retrieval from the title and abstract sections of the identified articles. These retrieved snippets serve as context within carefully constructed few-shot example prompts to guide a Llama-based large language model (LLM), fine-tuned with biomedical data, in generating the answers in the desired format. Our submitted systems achieved top rankings on the 'snippet retrieval' leaderboard, and demonstrated strong performance in answer extraction, the details of which will be further discussed in the results and error analysis sections.

This paper will detail our information retrieval and indexing approach, the prompt engineering strategy, and a comprehensive explanation of the system workflow, culminating in an analysis of the results and identified errors.

2. Related Work

The field of biomedical question answering (QA) has witnessed significant advancements in recent years, fueled by progress in both information retrieval methodologies and the capabilities of large language models (LLMs). While early BioASQ challenges primarily focused on static datasets and traditional retrieval pipelines, the Synergy series, including previous iterations; like in CLEF 2024 [4], introduces a dynamic and interactive evaluation paradigm driven by expert feedback.

Our work builds upon two crucial and interconnected strands of research: Retrieval-Augmented Generation (RAG) architectures for knowledge-intensive natural language processing (NLP) tasks, and the adaptation and fine-tuning of LLMs for enhanced performance within the biomedical domain.

2.1. Large Language Models in Biomedical QA

The recent surge in the capabilities of Large Language Models (LLMs) has profoundly impacted various NLP tasks, and biomedical QA is no exception. These models, with their ability to understand and generate human-like text, offer unprecedented opportunities for extracting and synthesizing information from complex biomedical literature. A key direction in this field involves fine-tuning general-purpose LLMs on extensive biomedical corpora, such as PubMed abstracts, clinical notes, and other specialized texts. This domain-specific adaptation significantly enhances the models' understanding of medical terminology, biological processes, and clinical contexts, leading to improved performance on biomedical QA benchmarks.

The Llama 3 "Herd" [5] exemplifies the advancements in open-source LLMs, showcasing a range of model sizes optimized for different computational constraints while achieving state-of-the-art results on both general and specialized evaluations. Similarly, the "Aloe" series of healthcare-focused LLMs [6], trained on rich clinical and research data, has demonstrated superior reasoning abilities in critical areas like disease understanding and drug interactions, outperforming general-purpose models. Furthermore, the development of models like MEDITRON-70B [7], which underwent large-scale pre-training specifically on medical data, underscores the benefits of domain-focused training for achieving high performance in biomedical NLP tasks. These specialized LLMs form a crucial component in modern biomedical QA systems, providing the generative power necessary to produce comprehensive and accurate answers.

2.2. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a highly effective paradigm for tackling knowledge-intensive NLP tasks, particularly in domains characterized by vast and rapidly evolving information, such as biomedicine. The core idea behind RAG is to augment the knowledge of a generative model by explicitly retrieving relevant information from an external knowledge source before generating the final output. This approach contrasts with relying solely on the parametric knowledge encoded within the LLM's weights, which can be limited or outdated, especially when dealing with emerging topics as in the BioASQ Synergy challenge.

Lewis et al. [8] first formalized this framework for open-domain QA, demonstrating significant improvements in answer accuracy and factuality by conditioning the language model on retrieved relevant passages. Subsequent research has explored and extended the RAG paradigm for various knowledge-intensive applications [9], highlighting its versatility and effectiveness in scenarios where access to up-to-date and specific information is crucial. In the biomedical domain, RAG offers a compelling solution for navigating the extensive PubMed literature, enabling systems to dynamically retrieve relevant abstracts and full-text articles in response to complex biomedical questions. This is particularly relevant for the BioASQ Synergy task, where the questions often pertain to developing issues, necessitating the integration of the latest research findings [10].

Prior work within the BioASQ challenges, including systems described in CLEF 2024 [4], has successfully employed RAG architectures, further validating its suitability and effectiveness for biomedical question answering. By combining the retrieval capabilities of information retrieval systems with the generative power of LLMs, RAG offers a promising avenue for building robust and accurate biomedical QA systems capable of addressing the dynamic nature of biomedical knowledge.

3. Proposed system

To effectively tackle the multifaceted challenges presented by the BioASQ Task Synergy, particularly the need for dynamic information retrieval and the generation of diverse answer formats based on evolving biomedical knowledge, we propose a novel multi-stage pipeline anchored in a Retrieval-Augmented Generation (RAG) architecture.

As visually depicted in Figure 1, our system strategically integrates three key modules: data preparation, relevant information retrieval, and response generation. The **data preparation** stage focuses on the initial processing of the incoming biomedical questions. The subsequent **IR module** is designed for the intelligent retrieval of relevant information from the vast PubMed repository.

Finally, the **response generation** module leverages this retrieved context, guided by carefully engineered prompts, to produce comprehensive answers tailored to the specific question type (summary, yes/no, factoid, list) and adhering to the task's requirements for article references, snippets, and ideal answers. Our approach leverages state-of-the-art Natural Language Processing (NLP) models and techniques within each stage to ensure both the accuracy and the relevance of the generated responses in this demanding interactive biomedical QA task.

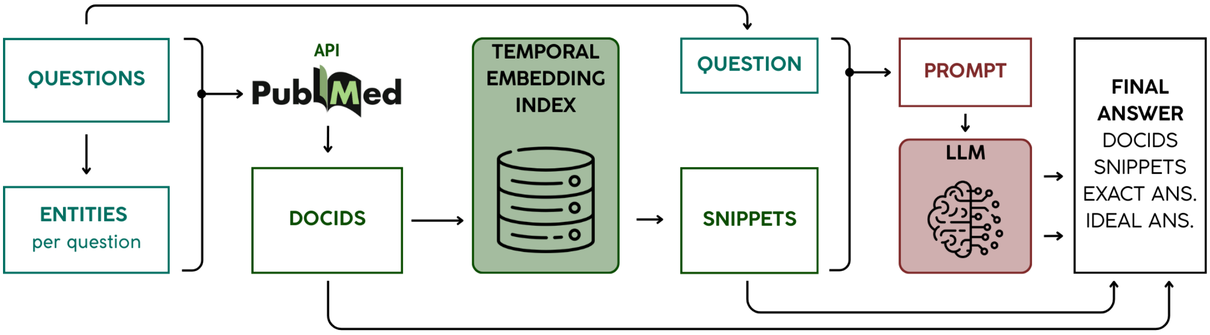


Figure 1: Complete workflow and architecture of the proposed system

3.1. Data Preparation

The initial phase focuses on acquiring and structuring the biomedical literature corpus required for the task.

3.1.1. Data Download

The system utilizes publicly available PubMed data. Specifically, the process involves downloading the PubMed baseline corpus and subsequent update files up to the date specified for the relevant challenge round. These files are initially obtained in xml.gz format from an FTP server[1]. Following download, the XML data is parsed and converted into a more manageable Comma-Separated Values (CSV) format.

3.1.2. Data Format

To optimize storage and processing efficiency, the raw text data from the CSV files is transformed into the Apache Parquet format. This columnar storage format significantly reduces storage footprint and improves read performance. The corpus is further processed by segmenting the text (primarily titles and abstracts as per task constraints) into overlapping chunks or snippets, as shown in Figure 2. A maximum chunk size of 512 characters is enforced, with an overlap of 64 characters between consecutive snippets. This overlap helps preserve contextual information across snippet boundaries.

Due to the large volume of data, this chunking process is performed in batches. For each batch, text content is extracted from the CSV rows and segmented using a text splitter. Each resulting snippet is stored as a structured record containing the text itself, a unique iterative snippet identifier (ID), the source document's PubMed Identifier (PMID), the section of the document it originated from (e.g., TITLE, ABSTRACT), and the start and end character positions of the snippet within its section.

To handle the large scale of the data efficiently without loading entire datasets into memory, processing relies on Polars LazyFrames. This allows for defining computation graphs on the data that are executed only when results are materialized, minimizing memory usage. Each processed chunk is represented as a Document object, encapsulating its text content and associated metadata (ID, PMID, SECTION, START, END).

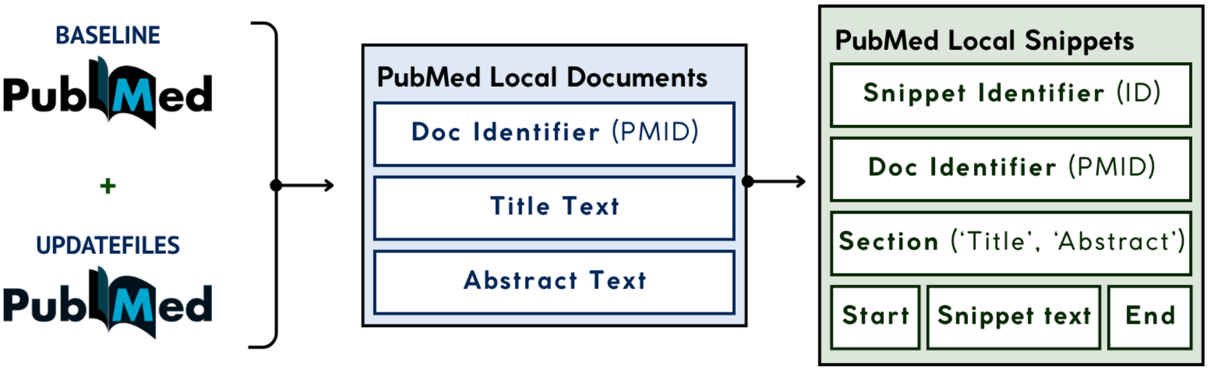


Figure 2: Processing of PubMed documents for local storage.

Finally, a mapping file is generated. This file acts as an index, storing the relationship between PMIDs and the Parquet files containing their corresponding pre-computed snippets. It records the start and end PMID for the documents primarily contained within each Parquet file, facilitating efficient lookup of relevant files based on a list of PMIDs. Notably, update files may contain documents with non-sequential PMIDs. This prepared, chunked, and indexed corpus forms the basis for subsequent retrieval steps.

3.2. Relevant Information Retrieval Module

This module is responsible for identifying and retrieving relevant textual evidence from the prepared corpus in response to a given biomedical question. It involves two main stages: retrieving relevant

[1] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/, https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/

document identifiers (PMIDs) and then extracting specific, semantically relevant text snippets from those documents.

3.2.1. Document Retrieval

This first step aims to identify a candidate set of PubMed articles likely to contain information relevant to the input question.

Question and Query Analysis. Initial experiments revealed that querying the PubMed API directly with the full natural language question often resulted in insufficient or irrelevant document retrieval. The presence of common words, stop words, and specific phrasing nuances can dilute the query's focus when using keyword-based or simple search mechanisms likely employed by the PubMed API. To address this, a query enhancement strategy based on Named Entity Recognition (NER) is employed.

The system utilizes the d4data/biomedical-ner-all model [11] , which is built upon distilbert-base-uncased and specifically fine-tuned on the Maccrobat dataset. This dataset consists of clinical case reports from PubMed Central, making the model particularly well-suited for identifying biomedical entities within the target domain. The model is configured to extract entities primarily of type 'chemical' and 'disease', as these are often the core concepts in biomedical questions.

As the employed DistilBERT-based NER model often outputs subword-tokenized entities, a crucial aggregation step is implemented. This custom logic reconstructs full entity names by combining consecutive subword tokens that share the same entity label (e.g., "trypa" and "##nosomiases" become "trypanosomiases").

Simultaneously, the entity type is maintained, and a representative confidence score is computed as the average of the subword token scores. This aggregation process is vital for transforming the fragmented NER output into coherent and precise query terms, thereby significantly improving the relevance of the retrieved documents from PubMed.

PubMed API for PMID retrieval. The PubMed API[2] is then queried using two distinct approaches for each input question: once with the original, full question text, and separately with the aggregated list of extracted 'chemical' and 'disease' entities. The primary goal of these queries is to retrieve the PubMed Identifiers (PMIDs) of potentially relevant articles. While the API might return additional information, such as full text excerpts, only the retrieved PMIDs are utilized in the subsequent stages of the pipeline, adhering to the task constraints of using only pre-processed title and abstract data.

The lists of PMIDs obtained from querying the original question and the extracted entities are then combined. Duplicate PMIDs are removed to generate a final, unique set of candidate PMIDs associated with the input question. This list forms the input for the snippet extraction phase.

3.2.2. Snippet extraction

Given the list of relevant PMIDs, the next goal is to identify and rank the specific text snippets (chunks) from the corresponding documents that are most semantically relevant to the input question. These snippets serve as the direct evidence for the final answer generation module.

The process begins by retrieving all pre-computed snippets associated with the identified PMIDs. This is handled by a function that takes the list of PMIDs and the path(s) to the relevant Parquet file(s) (located using the previously-generated distribution mapping) as input. It reads the specified files and filters the data to select only the rows whose PMID matches one of the input PMIDs, and converts these rows into Document objects containing the chunk text and metadata (PMID, SECTION, START, END).

Subsequently, a semantic search is performed to rank these retrieved snippets based on their relevance to the question. This relies on a sophisticated embedding model: pritamdeka/S-PubMedBert-MS-MARCO

[2] https://pmc.ncbi.nlm.nih.gov/tools/developers/

[12]. This model is particularly suitable for this task as it is derived from microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext and has been specifically fine-tuned on the MS-MARCO dataset for information retrieval tasks within the medical domain.

It maps sentences and paragraphs into a 768-dimensional dense vector space, employing a mean pooling strategy over token embeddings to generate the final sentence vector. Mean pooling is a common and effective technique for creating robust sentence embeddings from transformer outputs, ensuring contribution from all tokens.

For efficient semantic search, a FAISS (Facebook AI Similarity Search) index is constructed. Notably, a temporary, query-specific index is built in memory for each incoming question, rather than maintaining a single, persistent index for the entire corpus. This design choice implies that the number of candidate snippets retrieved per query (based on the initial PMID list) is computationally manageable, making the overhead of per-query index creation acceptable. It simplifies system architecture by avoiding the complexities of maintaining and updating a massive, persistent index of all PubMed snippets.

The index performs an exact search based on the maximum inner product between the query vector and the indexed vectors. When vectors are L2-normalized (as is standard practice for cosine similarity search with sentence transformers), maximizing the inner product is equivalent to maximizing cosine similarity. The choice of IndexFlatIP prioritizes retrieval accuracy, guaranteeing that the truly closest snippets within the candidate set are identified. While approximate nearest neighbor indices like IndexHNSW offer faster search times on very large datasets , the exact search provided by IndexFlatIP is preferred here to ensure the highest quality evidence is passed to the generation module, especially since the candidate pool has already been narrowed by the PMID retrieval step.

The **index construction** process involves:

1. Identifying the relevant Parquet files containing snippets for the query's PMIDs using the mapping CSV.
2. Extracting all corresponding Document objects.
3. Eliminating duplicate chunks to prevent redundancy in the index. Duplicates are identified using a unique key generated from a combination of the chunk's text content and its metadata (PMID, section, start, end position)
4. Embedding the text content of the unique snippets using the embedding model.
5. Adding these embeddings and their corresponding Document objects to the index.
6. Once the temporary index is built, the input question is embedded using the same model.
7. A similarity search is then performed, which retrieves the top 20 most similar snippets based on cosine similarity (achieved via normalized embeddings and the inner product search), along with their respective similarity scores.
8. Finally, the retrieved snippets are sorted in descending order based on their similarity scores.
9. The text content of these top-ranked snippets is concatenated into a single string, separated by newline characters, creating the summaries input for the generation module.
10. The function returns both this concatenated summaries string and the ranked list of the top Document objects including their scores.

By employing a specialized biomedical embedding model and a temporary, question-specific FAISS index, the system efficiently identifies and ranks the most semantically relevant text snippets from the retrieved documents.

This focused approach ensures that the subsequent generation module is provided with highly pertinent evidence, thereby maximizing the potential for accurate and comprehensive answers across the diverse question types.

3.3. Generation Module

The Generation Module utilizes a Large Language Model (LLM), Llama3.1-Aloe-Beta-8B, to synthesize answers based on the retrieved snippets. `Llama3.1-Aloe-Beta-8B` is an open healthcare LLM that achieves state-of-the-art performance on several medical tasks. The Aloe Beta family, which includes 7B, 8B, 70B, and 72B model sizes, is trained using a consistent recipe on top of both Llama3.1 and Qwen2.5 model families. Aloe models are trained on 20 medical tasks, resulting in robust and versatile healthcare capabilities. Evaluations demonstrate that Aloe models rank among the best in their class.

Its goal is to produce responses that are not only accurate but also adhere to the specific format requirements of the input question type (summary, yes/no, factoid, list) and the evaluation criteria defined by the challenge. This involves carefully crafted prompting strategies, including few-shot learning and a two-stage generation process for certain question types.

3.3.1. Few-Shot Prompting

The core of the generation process relies on structured prompting designed for instruction-tuned LLMs. A role-based format (System, User) is employed to guide the model's behavior effectively.

The System Prompt sets the context and constraints for the LLM. It defines the desired persona (e.g., acting as a "biomedical expert"), establishes strict operational boundaries (e.g., instructing the model to base its answer solely on the provided context snippets), specifies the mandatory output format, and encourages a step-by-step reasoning process (Chain-of-Thought, CoT) [13].

The User Prompt provides the specific inputs for the current question, dynamically inserted into a predefined template. These inputs typically include the original question (question), the concatenated text of the top-ranked retrieved snippets ({summaries}), and few-shot examples ({examples}), extract from the challenge feedback on from Round 4, to guide the model's response style and format. For generating "exact" answers (detailed below), the user prompt also incorporates a summary of the reasoning ({answer} summary) derived from a preceding "ideal" answer generation step.

Few-Shot Learning is incorporated by including examples of high-quality question-answer pairs from the feedback of previous challenge rounds within the prompt ({examples}). These examples are curated using a dedicated function. This function processes feedback files, filters for questions marked as *answerReady*, extracts and normalizes the ideal and exact answers (e.g., handling list formats appropriately), and saves these curated examples.

For *ideal* answers, the function extracts all questions of the specified type that have a non-empty ideal answer from the feedback file, formats each as a question followed by its ideal answer, and presents them in a standardized format (e.g., `Question: "..." -> "answer": "..."`). For *exact* answers, the function similarly filters for questions of the target type with non-empty exact answers, and for each, it constructs a prompt line including the question, a summary of the ideal answer (if available), and the exact answer in a structured output format (e.g., `Question: "..." -> Summary of relevant information: "..." - "answer": "..."`). Including these examples helps the LLM adapt to the specific nuances, expected level of detail, and precise formatting requirements of the target task.

A key aspect of the generation strategy is a Two-Stage Generation Process, employed for all question types except "summary". This approach separates the synthesis of information from the extraction of the final concise answer:

Stage 1: Ideal Answer Generation

The first stage aims to generate the ideal answer. The prompt includes the question, the retrieved snippets (summaries), and few-shot examples. The objective is to produce a comprehensive, paragraph-sized summary, written in an expert tone, that directly answers the question using only the information present in the provided snippets.

This ideal answer includes a chain of thought explaining the reasoning process. For questions explicitly designated as "summary" type, the output of this stage constitutes the final response. This initial step leverages the LLM's strength in synthesis and summarization, forcing it to reconcile potentially disparate or subtly conflicting information from various snippets into a coherent narrative and logical reasoning path.

Listing 1: Ideal prompt for 'yesno' questions

```
{
    ‘‘role’’: ‘‘system’’,
    ‘‘content’’: ‘‘You are an expert in the biomedical field. Your task is to answer yes
        /no questions based solely on the knowledge provided in the fragments extracted
        from the title and abstract sections of biomedical articles. You must ONLY use
        the information in these fragments to infer your answer. Sometimes the context
        does not provide a direct answer; in such cases, you must understand and join
        all the information provided and infer a correct conclusion.Follow a brief chain
        -of-thought process to arrive at your conclusion and include this reasoning in
        your response. Provide your final output in JSON format with the keys ‘‘
        chain_of_thought’’ and ‘‘answer’’, where answer is a single paragraph that
        completely answers the question and summarizes the most relevant information
        from the context in the style of a biomedical expert.If no answer can be
        inferred from the context, set ‘‘answer’’ to exactly ‘‘No answer provided". Pay
        attention to these examples of questions + answer types:{examples}In conclusion,
         this must be the answer exact format you must generate:{{"chain_of_thought’’:
        Detailed reasoning process, ‘‘answer’’: Paragraph detailing a complete
        informative answer or ‘‘No answer provided"}}Do not use line breaks in your text
         nor use double quotes like ’’ directly in your answer text, only use single
        quotes like ’ or double quotes with an escape character symbol to not compromise
         the JSON format and structure.The answer should resemble one given by an expert
        ."
},
{
    ‘‘role’’: ‘‘user’’,
    ‘‘content’’: ‘‘These are the fragments of information extracted from the articles
        from where you must infer the answer to the given question: {summaries}The
        question you must answer in the JSON format is: {question}"
}
```

Stage 2: Exact Answer Generation

This stage is executed only if the question type is not "summary" (i.e., for "yesno", "factoid", "list").

- **Input Preparation**. The JSON output containing the ideal answer from Stage 1 is parsed, and a concise summary of its chain-of-thought is extracted.
- **Prompt Construction**. A new prompt is constructed specifically for generating the exact answer. This prompt includes the original question, the same set of retrieved snippets (summaries) used in Stage 1, relevant few-shot examples, and the summary of the reasoning derived from the ideal answer's chain of thought.
- **Generation Goal**. The LLM is tasked with generating a concise, factual answer in the specific format required by the question type (e.g., 'yes' or 'no' for "yesno" questions; one or more entity names for "factoid" or "list" questions). By providing the reasoning path from the ideal answer, this stage guides the LLM to produce an exact answer consistent with the synthesized understanding developed in Stage 1. This makes the task more constrained, effectively becoming an extraction or classification based on the LLM's own prior reasoning, reducing the likelihood of pulling incorrect facts directly from potentially noisy snippets.

The LLM inference is performed using a generate function. Key parameters are carefully tuned to promote factual and consistent outputs: a low temperature (0.11) minimizes randomness, a "repetition penalty" (1.15) discourages verbatim repetition, and "max new tokens" (1024) provides sufficient length for both the chain-of-thought and the answer components.

Listing 2: Exact prompt for 'yesno' questions

```
{
    ‘‘role’’: ‘‘system’’,
    ‘‘content’’: ‘‘You are an expert in the biomedical field. Your task is to answer yes
        /no questions based on both the knowledge provided in the fragments extracted
        from the title and abstract sections of biomedical articles and an additional
        short generated answer. Carefully analyze the fragments of text and use only
        that information and the generated summary to infer your answer.\nFollow a brief
         chain-of-thought process to arrive at your conclusion and include this
        reasoning in your response. Provide your final output in JSON format with the
        keys ‘‘chain_of_thought’’ and ‘‘answer’’, where ‘‘answer’’ must be exactly
        either ’yes’ or ’no’. If no answer can be inferred from the context, set ‘‘
        answer’’ to exactly ’no’. \nPay attention to these examples of questions +
        summary + answer types:\n{examples}\nIn conclusion, this must be the answer
        exact format you must generate:\n{{"chain_of_thought’’: Detailed reasoning
        process, ‘‘answer’’: ’yes’/’no’}}\nDo not use line breaks in your text nor use
        double quotes like ’’ directly in your answer text, only use single quotes like
        ’ or double quotes with an escape character symbol to not compromise the JSON
        format and structure.\n"
  },
  {
    ‘‘role’’: ‘‘user’’,
    ‘‘content’’: ‘‘These are the fragments of information extracted from the articles
        from where you must infer the answer to the given question: {summaries}\nTake
        into account this inferring process done from the context to infer a summary of
        the relevant information for the question: {answer}\nThe question you must
        answer in the JSON format is: {question}\nRemember to answer in this exact JSON
        format:\n{{"chain_of_thought’’: Detailed reasoning process, ‘‘answer’’: ’yes’/’
        no’}}\n"
  }
```

3.3.2. Response Post-processing

After the LLM generates the response, a final post-processing step is applied to ensure the output strictly conforms to the expected format, which is crucial for automated evaluation pipelines and downstream consumption. Minor variations in LLM output, such as missing characters or inconsistent casing, could otherwise lead to evaluation failures.

These post-processing steps act as essential guardrails, ensuring the system’s output is robustly formatted and machine-readable, maximizing compatibility with automated evaluation scripts.

4. Evaluation and results

In this section we discuss the performance of our three system variants across retrieval, snippet extraction, exact answer generation, and ideal answer quality, drawing on both automatic and manual evaluation metrics.

The system names are listed below, we simplified the identifiers to improve the readability of the analysis:

- **Q&A based on RAG**: original identifier for [sinai-uja-v1]
- **Q&A based on RAG2**: original identifier for [sinai-uja-v2]
- **sinai_uja_RAG**: original identifier for [sinai-uja-v3]

The differences between the 3 system lie on slight alterations in the language of the prompt; the last system (sinai-uja-v3) being the most optimized for the task and the one shown before. Also, they differ on the pubmed files used; for system sinai-uja-v1 the pubmed files were not updated to the last permitted date. For the other 2 system the retrieval subset was the exact same and it is reflected in the retrieval related metrics below.

4.1. Results for the information retrieval module

Since sinai-uja-v2 and sinai-uja-v3 share the same updated corpus, their retrieval metrics are identical. These results confirm that simply refreshing the underlying PubMed data can drive meaningful improvements in RAG pipelines for emerging biomedical topics. Overall these results are quite poor all we will discuss the possible reasons for this in the error analysis section.

The document retrieval performance was evaluated using standard metrics such as Mean Precision, Recall, F-Measure, Mean Average Precision (MAP), and Geometric Mean Average Precision (GMAP). Precision measures the proportion of retrieved documents that are relevant, while Recall measures the proportion of relevant documents that are retrieved. F-Measure is the harmonic mean of Precision and Recall. MAP considers the ranking of retrieved documents, rewarding systems that place relevant documents higher in the list. GMAP is similar to MAP but uses a geometric mean, emphasizing performance on questions where systems struggled.

Table 1
Task 13 Synergy, Document Retrieval, Round 4

System	Mean Precision	Recall	F-Measure	MAP	GMAP
sinai-uja-v1	0.0353	0.0824	0.0423	0.044	0
sinai-uja-v2	**0.0417**	**0.1209**	**0.0533**	**0.0824**	**0.1**
sinai-uja-v3	**0.0417**	**0.1209**	**0.0533**	**0.0824**	**0.1**

As shown in Table 1, updating the PubMed snapshot for v2 (and v3) yields substantial gains over v1, which relied on an outdated index.

Precision increases from 0.0353 to 0.0417 (+18%), indicating that a higher fraction of the top-10 retrieved articles are relevant. *Recall* jumps from 0.0824 to 0.1209 (+47%), reflecting a slightly broader coverage of the gold-standard articles. *MAP* nearly doubles (0.044 → 0.0824) and *GMAP* rises from zero to 0.1, showing that retrieval freshness is critical for ranking quality.

On the other hand, snippet extraction performance, detailed in Table 2, utilizes similar metrics adapted to account for potential overlap between returned and golden snippets. The official measure for snippets in Phase A (which is analogous to the relevant material submission in Synergy) is Mean F-Measure. Interestingly, system **sinai-uja-v1** slightly outperformed **sinai-uja-v2** and **sinai-uja-v3** across all snippet extraction metrics.

This suggests that while the updated document set improved document retrieval, the older document set used by **sinai-uja-v1** might have contained snippets that were, on average, slightly more aligned or had better overlap with the golden snippets for the questions in this round, despite the overall lower document recall. The identical performance of **sinai-uja-v2** and **sinai-uja-v3** in snippet extraction is expected, given they used the same retrieval subset.

We must highlight how our RAG-based pipelines achieved the highest snippet extraction scores in Round 4, outperforming the next best competitor by nearly 0.20 points on the recall metric and 0.10 points across the rest. Our submissions led an otherwise tightly clustered field of nine teams, demonstrating that our dynamic indexing and overlap-aware matching deliver markedly superior snippet recall and precision—even in rapidly evolving biomedical domains.

Table 2
Task 13 Synergy, Snippet Extraction, Round 4

System	Mean Precision	Recall	F-Measure	MAP	GMAP
sinai-uja-v1	**0.2458**	**0.378**	**0.2518**	**0.3039**	**0.0143**
sinai-uja-v2	0.2451	0.3737	0.2505	0.3016	0.0142
sinai-uja-v3	0.2451	0.3737	0.2505	0.3016	0.0142

The results for the document and snippet retrieval modules present an intriguing dichotomy. While the document retrieval metrics (Mean Precision, Recall, F-Measure, MAP, GMAP) indicate a relatively modest performance across all three system variants, with the updated PubMed corpus demonstrably improving these metrics for v2 and v3, the snippet extraction results paint a significantly different picture.

Notably, all our RAG-based systems achieved the highest scores in snippet extraction in Round 4, outperforming competitors by a substantial margin, particularly in Recall and overall F-Measure. This suggests that our methodology for identifying and ranking relevant text excerpts within the retrieved documents is highly effective at pinpointing the most pertinent information, even if the initial document retrieval phase could be further optimized.

One possible explanation for this discrepancy is that while the initial set of retrieved documents might not have perfectly aligned with all the gold-standard documents, our semantic search and ranking strategy within those retrieved documents excels at identifying the key information-bearing snippets. The slight outperformance of v1 in snippet extraction, despite its outdated document corpus, could indicate that the specific content within its retrieved documents had a higher degree of overlap with the gold-standard snippets for the questions in this round.

The overall dominance of our systems in snippet retrieval underscores the strength of our dynamic indexing and semantic matching approach in extracting crucial evidence for downstream answer generation, even in the challenging context of evolving biomedical knowledge.

4.2. Results for the answer generation module

The answer generation module was evaluated based on the quality of both exact and ideal answers.

4.2.1. Evaluation of exact answers

For Yes/No questions, accuracy measures the proportion of correctly answered questions. F1-Yes and F1-No measure the F1 score for questions with "yes" and "no" as the golden answer, respectively. Macro F1 is the unweighted average of F1-Yes and F1-No, serving as the official measure.

Table 3 shows that both **sinai-uja-v1** and the most optimized system, **sinai-uja-v3**, achieved the highest Accuracy (0.9) and identical Macro F1 scores (0.899). System **sinai-uja-v2** performed slightly worse with an Accuracy of 0.8 and a Macro F1 of 0.7917. This suggests that the prompt language optimization in **sinai-uja-v3** was beneficial for Yes/No questions, bringing its performance in line with the initial **sinai-uja-v1** system.

Table 3
Task 13 Synergy, Q&A exact Yes/No, Round 4

System	Accuracy	F1 - Yes	F1 - No	Macro F1
sinai-uja-v1	**0.9**	**0.9091**	**0.8889**	**0.899**
sinai-uja-v2	0.8	0.8333	0.75	0.7917
sinai-uja-v3	**0.9**	**0.9091**	**0.8889**	**0.899**

Factoid questions require returning a list of up to 5 entity names. Strict Accuracy counts a question as correct only if the top returned entity matches the golden answer, while Lenient Accuracy considers a question correct if the golden answer is anywhere in the top 5 list. Mean Reciprocal Rank (MRR), the official measure, rewards systems that rank the correct answer higher.

Table 4 indicates that **sinai-uja-v3** achieved the best performance for Factoid questions with a Strict Accuracy, Lenient Accuracy, and MRR of 0.4. Systems **sinai-uja-v1** and **sinai-uja-v2** had identical and lower scores of 0.3 across all three metrics. This 33% relative boost in ranking the correct entity underscores the value of tailored few-shot examples and optimized prompt templates for factoid extraction in small-training scenarios.

Table 4
Task 13 Synergy, Q&A exact Factoid, Round 4

System	Strict accuracy	Lenient accuracy	MRR
sinai-uja-v1	0.3	0.3	0.3
sinai-uja-v2	0.3	0.3	0.3
sinai-uja-v3	**0.4**	**0.4**	**0.4**

For List questions, systems must return a list of entity names, and performance is evaluated using Mean Precision, Recall, and F-Measure against a golden list. Mean F-Measure is the official measure.

According to Table 5, **sinai-uja-v3** achieved the highest Mean Precision (0.2331) and F-Measure (0.2667), indicating better accuracy in the entities it returned. System **sinai-uja-v1**, however, achieved the highest Recall (0.5104), suggesting it was better at identifying more of the golden list entities, even if its precision was lower. System **sinai-uja-v2** performed in between the other two for Precision and F-Measure but had the lowest Recall. The prompt optimization in **sinai-uja-v3** appears to have improved the precision of the list of entities returned.

Table 5
Task 13 Synergy, Q&A exact List, Round 4

System	Mean Precision	Recall	F-Measure
sinai-uja-v1	0.2156	**0.5104**	0.2624
sinai-uja-v2	0.22	0.4479	0.2472
sinai-uja-v3	**0.2331**	0.4688	**0.2667**

The results for the list question answering reveal a trade-off between precision and recall across our system variants. While **sinai-uja-v3**, with its optimized prompt, demonstrates the highest precision and F-Measure, indicating a greater accuracy in the returned entities, **sinai-uja-v1** exhibits the highest recall. This suggests that despite potentially including more irrelevant entities (lower precision), **sinai-uja-v1** was more successful in identifying a larger proportion of the entities present in the gold-standard lists.

This higher recall in **sinai-uja-v1**, despite its less refined prompt and use of an older PubMed index, could be attributed to a broader or more exhaustive generation strategy. Perhaps the less constrained prompt in v1 encouraged the LLM to generate a wider range of potential entities, increasing the chances of overlap with the gold standard, even if it also led to the inclusion of more false positives. Conversely, the prompt optimization in **sinai-uja-v3**, while improving the accuracy of the generated list, might have inadvertently led to a more focused generation, potentially missing some relevant entities present in the gold standard.

4.2.2. Evaluation of ideal answers

Ideal answers, which are paragraph-sized summaries, were evaluated using both automatic and manual metrics.

Automatic evaluation of ideal answers (Table 6) was performed using ROUGE metrics, specifically ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4), which measure the overlap of bigrams and skip-bigrams (with a maximum skip distance of 4), respectively, between the generated answer and a set of reference texts (including golden answers and relevant snippets).

R-2 and R-SU4 are reported as Recall (Rec) and F1 scores. System **sinai-uja-v1** surprisingly achieved the highest scores across all automatic ROUGE metrics (R-2 Rec 0.2093, R-2 F1 0.2047, R-SU4 Rec 0.2169, R-SU4 F1 0.2126). Systems **sinai-uja-v2** and **sinai-uja-v3** performed slightly lower, with **sinai-uja-v3** showing a minor edge in R-SU4 F1 over **sinai-uja-v2**.

This suggests that the ideal answers generated by **sinai-uja-v1** had a higher n-gram and skip-bigram overlap with the reference texts according to these automatic measures.

Table 6
Task 13 Synergy, Q&A ideal, Round 4

System	R-2 Rec	R-2 F1	R-SU4 Rec	R-SU4 F1
sinai-uja-v1	**0.2093**	**0.2047**	**0.2169**	**0.2126**
sinai-uja-v2	0.2008	0.1954	0.2035	0.1988
sinai-uja-v3	0.2005	0.2005	0.2008	0.2023

Manual evaluation by biomedical experts (Table 7) provides a human perspective on the quality of ideal answers, assessing Readability, Information Recall, Information Precision, and Information Repetition on a 1-5 scale.

Crucially, these expert judgments paint a different picture, in contrast to the automatic evaluation, the most optimized system **sinai-uja-v3**, received the highest average scores from the experts across all manual criteria (Readability 4.24, Recall 4.27, Precision 3.96, Repetition 4.27).

System **sinai-uja-v2** had slightly lower scores than **sinai-uja-v3** but was generally rated higher than **sinai-uja-v1**, except for Information Precision where **sinai-uja-v1** scored slightly higher.

This discrepancy between automatic and manual evaluation highlights that while ROUGE metrics measure surface-level overlap, expert judgment captures the nuances of information accuracy, completeness, coherence, and fluency that are critical for high-quality biomedical summaries. The prompt language optimization in **sinai-uja-v3** seems to have significantly improved the perceived quality of the ideal answers by biomedical experts.

Table 7
Task 13 Synergy, Q&A ideal for manual metrics, Round 4

System	Readability	Recall	Precision	Repetition
sinai-uja-v1	4.15	4.22	3.91	4.16
sinai-uja-v2	4.22	4.22	3.85	4.16
sinai-uja-v3	**4.24**	**4.27**	**3.96**	**4.27**

In summary, the updated PubMed dataset used by **sinai-uja-v2** and **sinai-uja-v3** led to improved document retrieval performance compared to **sinai-uja-v1**. For exact answers, the prompt language optimization in **sinai-uja-v3** resulted in better performance for Factoid and List questions and matched the performance of **sinai-uja-v1** for Yes/No questions, outperforming **sinai-uja-v2**. While automatic metrics favored **sinai-uja-v1** for ideal answers, manual evaluation by experts clearly indicated that the ideal answers generated by the optimized **sinai-uja-v3** system were of higher quality.

Overall, **sinai_uja_RAG** (v3) achieves the best end-to-end performance on Synergy Task 13, demonstrating that the combination of up-to-date retrieval, entity-driven query formation, and carefully tuned few-shot prompts yields the most effective biomedical QA system in this interactive, feedback-driven framework.

5. Error analysis

While the quantitative results presented in Section 4 demonstrate the overall strengths and weaknesses of our three system variants, a deeper look at the specific types of errors made by each component is essential to guide future improvements. In this section, we systematically compare the outputs of our best-performing pipeline (**sinai_uja_RAG**, v3) and its predecessors against the "golden" reference answers and expert feedback from the most recent Synergy round (Round 4). In total, we will analyse 55 questions, the ones that either had a golden-marked ideal answer or exact answer.

By examining discrepancies at each stage—document retrieval, snippet selection, ideal answer generation, and exact answer extraction; we aim to uncover common failure modes (e.g., missing key entities, over-reliance on noisy contexts, prompting ambiguities) and quantify their impact on end-to-end performance. We begin by aligning each system's retrieved PMIDs and snippet overlaps with the gold-standard set, then analyze patterns in incorrect "yes/no" and factoid outputs, and finally inspect qualitative feedback on ideal answers to identify recurring issues in coherence and factual completeness.

This error analysis not only elucidates where and why our models falter but also provides actionable insights for refining our multi-stage biomedical QA pipeline.

5.1. Answer error analysis by question type

5.1.1. Yes/no questions

Across all three versions, our system answered 10 "yes/no" questions in the exact-answer task. System v3 achieved perfect accuracy on 8 of these, but failed on two cases—an overall 80% exact accuracy for "yes/no" type. Both errors share a common pattern, they correspond to false positives on "no" questions:

- *63adca9ec6c7d4d31b00001d*: "Is Cinpanemab effective for Early Parkinson?s Disease?"
- *6593debd06a2ea257c00001f*: "Are there new European Union formal efforts to increase the number of clinical trials aimed at improving the mental health of children?"

In both instances, the provided snippets did not explicitly affirm the queried claim. Instead, they described related phenomena (e.g., clinical trials for hormonal contraception; mental-health studies in children) without supporting the specific assertion.

If we take a look at the snippets retrieved and the chain-of-thought generated by the model, it is clear that the model's tendency to infer a positive conclusion from loosely related or partially overlapping contexts appears to be the root cause.

The first misstep occurred with the question, "Are there new European Union formal efforts to increase the number of clinical trials aimed at improving the mental health of children?". Although the available snippets described a variety of important studies—a resilience-building RCT for young carers, a school-based mental-health program (PROMEHS), and epilepsy-focused trials involving children with comorbid mental-health disorders—nowhere did they reference an EU-wide mandate or policy explicitly intended to scale up pediatric clinical trials.

Listing 3: Snippets from correctly retrieved documents for question '6593debd06a2ea257c00001f'

```
{
  ’33669796’ = ‘‘It is estimated that 4-8% of youth in Europe carry out substantial care
       for a family member or significant other. To prevent adverse psychosocial
      outcomes in young carers (YCs), primary prevention resilience building
      interventions have been recommended. We describe the study protocol of an
      international randomized controlled trial (RCT) of an innovative group
      intervention designed to promote the mental health and well-being of adolescent
      YCs (AYCs) aged 15-17. The RCT will be conducted in six European’’,
  ’36360389’: ‘‘The challenges of today’s society demand high levels of socio-emotional
      skills in children and adolescents; therefore, mental health is an important issue
       to be addressed and promoted in schools. The present study aims to investigate
      the effectiveness of a school mental health program (Promoting Mental Health at
      Schools; PROMEHS) designed to promote socio-emotional learning and prevent
      psychosocial difficulties in children and adolescents. The study was conducted on
      a sample of 1392 students"
}
```

Instead, our v3 pipeline treated these investigator-driven research projects as if they were coordinated EU initiatives, and confidently answered “yes.” In reality, these are valuable individual studies, but they fall short of constituting a formal, Europe-level effort to increase trial volume.

A similar error pattern arose with Cinpanemab’s assessment in early Parkinson’s disease. The model’s chain of thought highlighted the antibody’s binding to aggregated alpha-synuclein and the dose-dependent biomarker changes observed in the SPARK trial. These pharmacodynamic results are encouraging indicators of target engagement, but they do not equate to clinical efficacy. There were no reported improvements in patient motor scores or disease progression metrics in the snippets provided. Nevertheless, the model conflated biomarker success with treatment effectiveness, again answering “yes” when the correct response was “no”.

Listing 4: Chain-of-thought generated for question ’63adca9ec6c7d4d31b00001d’

```
{
    ’chain-of-thought’: To determine if Cinpanemab is effective for early Parkinson’s
        disease, I analyzed the provided article fragments focusing on the specific
        details about Cinpanemab’s mechanism of action, its ability to bind to alpha-
        synuclein, and any clinical trial data mentioned. Cinpanemab is described as a
        human-derived monoclonal antibody that targets aggregated alpha-synuclein, a key
         component in the pathogenesis of Parkinson’s disease. The study highlights its
        ability to form stable complexes with alpha-synuclein in cerebrospinal fluid,
        indicating engagement with the target in the central nervous system. Furthermore
        , the evaluation of biomarker results from the Phase 2 SPARK clinical trial
        shows a significant correlation between Cinpanemab dosage and binding to alpha-
        synuclein, suggesting that the drug behaves as intended within the CNS. Although
         the overall assessment of Cinpanemab’s efficacy in treating early Parkinson’s
        disease requires further investigation beyond these findings, the available data
         indicate promise for its potential as a disease-modifying therapy.
}
```

Both errors underscore an overly permissive inference strategy: in the absence of an explicit statement, the model infers the most optimistic interpretation. To address this, our future prompts should require the model to pinpoint a direct quote or clearly labelled policy statement before granting a “yes” .

5.1.2. Factoid questions

Our systems encountered varied success on the factoid questions, where the goal is to extract a concise, factual answer. To quantitatively assess performance for specific questions, we employed BLEU [14] and BertScore [15] metrics, comparing the generated answers against the gold standards. This qualitative analysis of specific errors provides further insights into the systems’ limitations.

One notable error occurred for the question, “What is the burden of human trypanosomiases in European hospitals?” (id: *6593de2f06a2ea257c00001e*). The gold answer specified “Spain has the highest burden of Chagas disease in Europe” as shown in Listing 5 while our systems (v1, v2, and v3) returned “Imported cases” or “African trypanosomiasis,” resulting in a BLEU score of 0. The BertScore also indicated low similarity. This suggests a failure in correctly identifying the specific type and geographical distribution of trypanosomiasis requested. The retrieved snippets likely discussed trypanosomiases in a broader context, and the model failed to pinpoint the European-specific burden and the prominence of Chagas disease in Spain.

Listing 5: Factoid exact answer for question ’6593de2f06a2ea257c00001e’

```
{
    ‘‘question’’: ‘‘What is the burden of human trypanosomiases in European hospitals?’’
        ,
    ‘‘golden’’: ‘‘Spain has the highest burden of Chagas disease in Europe.’’,
    ‘‘results’’: [
        {"v1’’: ‘‘Imported cases’’, ‘‘bleu’’: 0, ‘‘bertscore’’: (0.4, 0.2, 0.27)},
        {"v2'': ``African trypanosomiasis'' ``bleu'': 0, ``bertscore'': (0.37, 0.2,
           0.26)},
        {"v3'': ``African trypanosomiasis'' ``bleu'': 0, ``bertscore'': (0.37, 0.2,
           0.26)},
    ]
}
```

Conversely, for the question “What is the percentage of women that have successfully undergone fertility treatment in the European Union?” (id: *6593d3ab06a2ea257c00001a*), all our systems correctly identified “24%,” which closely matches the gold answer “24.” While the BLEU score was 0 due to the presence of the percentage symbol, the high BertScore indicated strong semantic similarity. This highlights a limitation of exact-match metrics like BLEU in cases where the generated answer is semantically correct but differs slightly in formatting.

Another interesting case involved the question “What biological process is associated with Vitamin K?". The gold answers comprised a list of several processes, including “Blood coagulation,” “Bone metabolism,” and “protection against oxidative stress.” Our systems successfully identified some of these (e.g., “anti-inflammatory activity,” “protection against oxidative stress,” “blood coagulation,” “bone metabolism”).

However, they also generated non-gold answers like “brain development” (v1, v2) and “activation of PXR signaling pathway” (v3), leading to a lower overall performance score. This indicates that while the systems could retrieve some relevant information, they also included plausible but ultimately incorrect biological processes, highlighting a potential issue with over-generation or the inclusion of less directly supported information.

Listing 6: Factoid exact answer for question ’6772c765592fa48873000009’

```
{
    ``question'': ``What biological process is associated with Vitamin K?'',
    ``golden'': ['Blood coagulation', 'Bone metabolism', 'Bone Health Maintenance', '
        calcium metabolism', 'Inhibition of arterial calcification', 'Nervous System
        Function', 'anti-inflammatory activity', 'protection against oxidative stress'],
}
```

On the other hand, we observed **strong performances** on questions like “What is the best non-invasive method to diagnose endometriosis?” (id: *6777b471592fa4887300000c*) where all systems correctly answered “transvaginal ultrasonography,” achieving a high BLEU score. However, the systems failed to identify the other gold answer, “MRI,” suggesting a limitation in retrieving the full spectrum of correct answers.

Similarly, for “What disease is associated with chalk-stick fracture?” (id: *63adcb54c6c7d4d31b00001f*), all systems correctly answered “ankylosing spondylitis” resulting in a perfect BLEU score, despite a seemingly low BertScore which might be less reliable for short, exact answers. Lastly, for “What disease can be treated with vorasidenib?” (id: *677e8319592fa4887300001e*), systems v2 and v3 perfectly matched the gold answer “IDH-mutant gliomas,” while v1 returned “IDH-mutant glioma,” resulting in a slightly lower BLEU score. This minor variation likely stems from the prompt differences and the slightly different document set used by v1.

Overall, the factoid question analysis reveals that while our RAG-based system can often extract correct and precise answers, it still struggles with nuanced questions requiring specific geographical or contextual information, and can sometimes over-generate plausible but incorrect answers. Future work will focus on refining the prompting strategies and retrieval mechanisms to improve the accuracy and completeness of factoid answer extraction.

5.1.3. List questions

The evaluation of our systems on list questions, where a set of entities is expected as the answer, reveals several instances of poor performance as indicated by BLEU scores of 0. Examining these cases provides insights into the challenges our RAG-based approach faces in generating comprehensive and accurate lists.

For the question “Which are the most common psychiatric events associated with the consumption of cannabis?” (id: *63ac44c2c6c7d4d31b000011*), all three system variants returned a consistent list: [’panic attack psychotic states dependence abuse cognitive disorders amotival syndrome anxiety disorder suicidal idea and attempt Hallucinogenic effects Stimulant effects Aggressive behavior Addictive behaviors Bipolar Disorder’].

As shown in Listing 7, when compared against individual gold standard entities like ’psychosis’, ’paranoia’, and ’mood disorders’, the BLEU score is 0 across all systems. This suggests that while the generated list contains several relevant terms, it might be too broad or include terms not considered the most common by the gold standard. The low BertScore further supports a lack of close semantic overlap with the specific gold entities.

Listing 7: List exact answer for question ’63ac44c2c6c7d4d31b000011’

```
{
    ‘‘question’’: ‘‘Which are the most common psychiatric events associated with the
        consumption of cannabis?’’,
    ‘‘golden’’: ["psychosis’’, ‘‘paranoia’’, ‘‘mood disorders"],
    ‘‘results’’: [
       {"v1’’: ["panic attack psychotic states dependence abuse cognitive disorders
           amotival syndrome anxiety disorder suicidal idea and attempt Hallucinogenic
           effects Stimulant effects Aggressive behavior Addictive behaviors Bipolar
           Disorder"], ‘‘bleu’’: 0, ‘‘bertscore’’: (0.3134, 0.1506, 0.2034)},
       {"v2’’: ["panic attack psychotic states dependence abuse cognitive disorders
           amotival syndrome anxiety disorder suicidal idea and attempt Hallucinogenic
           effects Stimulant effects Aggressive behavior Addictive behaviors Bipolar
           Disorder"], ‘‘bleu’’: 0, ‘‘bertscore’’: (0.3134, 0.1506, 0.2034)},
       {"v3’’: ["panic attack psychotic states dependence abuse cognitive disorders
           amotival syndrome anxiety disorder suicidal idea and attempt Hallucinogenic
           effects Stimulant effects Aggressive behavior Addictive behaviors Bipolar
           Disorder"], ‘‘bleu’’: 0, ‘‘bertscore’’: (0.3134, 0.1506, 0.2034)}
    ]
}
```

A similar pattern of a lengthy generated list failing to precisely match specific gold entities is observed for the question “Which tools are used to predict mortality in paediatric sepsis?” (id: *658447ca06a2ea257c000001*). All systems returned an extensive list of potential predictors. However, when evaluated against individual gold entities like ’platelet count’, ’SIRS’, ’NEWS’, ’NGAL’, and ’PT/INR’, the BLEU score is consistently 0. This suggests that while the systems correctly identified various relevant factors, they might have included additional less critical or less established predictors, leading to a mismatch with the specific entities prioritized in the gold standard.

Continuing with the ongoing trend, the question “Most common physical signs of self-harm in pre-teenagers” (id: *6777b5e1592fa4887300000d*) resulted in the systems listing ’cutting’, ’scratching’, and ’burning’. Compared to broader terms like ’skin tearing’, ’wound-healing hindrance’, and ’striking objects’, the BLEU score is 0, indicating that the experts where aiming for more general answers rather than the specific ones generated.

These instances of low BLEU scores on list questions highlight a recurring challenge: our systems often generate lists of specific entities that, while potentially relevant, do not perfectly align with the level of abstraction or the exact set of entities prioritized in the gold standard.

This could be due to limitations in the retrieval process, the prompt’s influence on the specificity of the generated list, or the inherent difficulty in precisely matching the granularity and scope of expert-curated lists. Future efforts should focus on refining the system’s ability to generate lists that balance specificity with broader categories and ensure a closer match to the expected entities.

5.1.4. Summary questions and all ideal answers

For a more holistic error analysis, we also examined the ideal answers generated by our best-performing system (v3) across all question types (factoid, yes/no, list, and summary). These paragraph-length answers aim to synthesize the key information from the retrieved snippets.

We evaluated these generated ideal answers using the same metrics as before, Bleu and BertScore.

Ideal answers for yes/no, factoid and list questions

Analyzing the ideal answers generated for non-summary question types reveals both strengths and weaknesses in the system’s summarization capabilities.

The factoid question seen before with id ’*6593d3ab06a2ea257c00001a*’ asked for the percentage of women with successful fertility treatment in the EU. Our system’s ideal answer, “The percentage of women that have successfully undergone fertility treatment in the European Union is 14%,” achieved a low BLEU score of 0.346. This suggests a factual inaccuracy compared to the expected summary, which likely highlighted the reported 24% success rate. The low BertScore results (0.2490, 0.2071, 0.2142) further indicates a lack of semantic similarity. This error likely stems from the system extracting or synthesizing an incorrect numerical value from the retrieved documents.

Listing 8: Ideal answer for question ’6772c615592fa48873000007’

```
{
    ‘‘question’’: ‘‘Is morphomics associated with the study of body image or body
        dysmorphia?’’,
    ‘‘golden’’: [
        ’Morphomics is not associated with the study of body image or body dysmorphia.
            Instead, it is focused on analyzing body composition, specifically through
            medical imaging techniques like CT scans, to assess and predict health
            outcomes related to body fat and muscle content.’,
        ’No, morphomics is not associated with the study of body image or body
            dysmorphia. It involves analyzing 3D medical imaging data to assess
            anatomical features, often for predicting health outcomes or assessing risk
            in medical situations.’,
        ’Morphomics involves the study and analysis of body composition through medical
            imaging techniques, primarily focusing on biological markers in muscle and
            adipose tissues. It does not concern itself with psychological aspects such
            as body image or body dysmorphia.’
    ],
    ‘‘results’’: [
       {
           ‘‘v3’’:
               ‘‘While morphomics is occasionally referenced in discussions about body
                   image or body dysmorphic disorders, its main focus lies in the
                   objective measurement and analysis of body composition through
                   medical imaging, making it less directly associated with the study of
                    body image or body dysmorphia compared to other fields. Its primary
                   goal is to quantify and analyze anatomical features, not to address
                   subjective perceptions of body image.’’,
           ‘‘bleu’’: 0.5502,
           ‘‘bertscore’’: (0.2712, 0.1739, 0.2076)
       }
    ]
}
```

"Is morphomics associated with the study of body image or body dysmorphia?” (id: *6772c615592fa48873000007*) was another yes/no question posed, the correct answer with respect with the feedback is “No". While the generated ideal answer correctly concludes that morphomics is not primarily associated with body image, the initial clause as seen in Listing 8 introduces unnecessary nuance and potential confusion that contradicts the direct “no” implied by the gold snippets. This suggests a tendency for the model to provide overly verbose or hedging answers even when the source information is direct.

On the other hand, question “Are there clinical trials on hormonal male birth control methods?” (id: *6593d46c06a2ea257c00001b*) was precisely the one marked by the BioASQ team as one of the golden answers in the feedback file used for this analysis. As shown in Listing 9, the system accurately extracted and synthesized the key details regarding the existence and focus of clinical trials on hormonal male contraception, demonstrating a strong ability to summarize factual information for yes/no questions.

Listing 9: Ideal answer for question ’6593d46c06a2ea257c00001b’

```
{
    ‘‘question’’: ‘‘Are there clinical trials on hormonal male birth control methods?’’,
    ‘‘golden’’: [
        ‘‘Yes, there are clinical trials on hormonal male birth control methods. These
            trials have evaluated combinations such as testosterone undecanoate and
            etonogestrel, as well as monthly depot medroxyprogesterone acetate with
            testosterone esters, showing effectiveness in suppressing spermatogenesis
            and promising potential as reversible contraception methods for men.’’,
        ‘‘Yes, there are clinical trials on hormonal male birth control methods. These
            trials focus on evaluating the effectiveness and acceptability of various
            hormonal substances for male fertility regulation, including different
            combinations of androgens and gestagens, as well as cyproterone acetate.’’,
        ‘‘Yes, there are clinical trials on hormonal male birth control methods. As of
            June 2023, two hormonal male contraceptive methods are undergoing phase II
            clinical trials for safety and efficacy. These methods involve the use of a
            long-acting combination of an androgen plus a progestin, which have been
            shown to be efficacious, reversible, and well-tolerated in clinical trials
            with healthy men. Additionally, a randomized, placebo-controlled clinical
            trial by the pharmaceutical industry has demonstrated the effectiveness of a
             combination of testosterone undecanoate and etonogestrel in suppressing
            spermatogenesis in volunteers. Overall, these developments suggest that the
            first marketed male hormonal contraceptive method is likely to be a long-
            acting hormonal method."
    ],
    ‘‘results’’: [
        {
            ‘‘v3’’:
                ‘‘Yes, there are clinical trials on hormonal male birth control methods.
                    These trials have evaluated combinations such as testosterone
                    undecanoate and etonogestrel, as well as monthly depot
                    medroxyprogesterone acetate with testosterone esters, showing
                    effectiveness in suppressing spermatogenesis and promising potential
                    as reversible contraception methods for men.’’,
            ‘‘bleu’’: 1,
            ‘‘bertscore’’: (0.2755, 0.1810, 0.2185)
        }
    ]
}
```

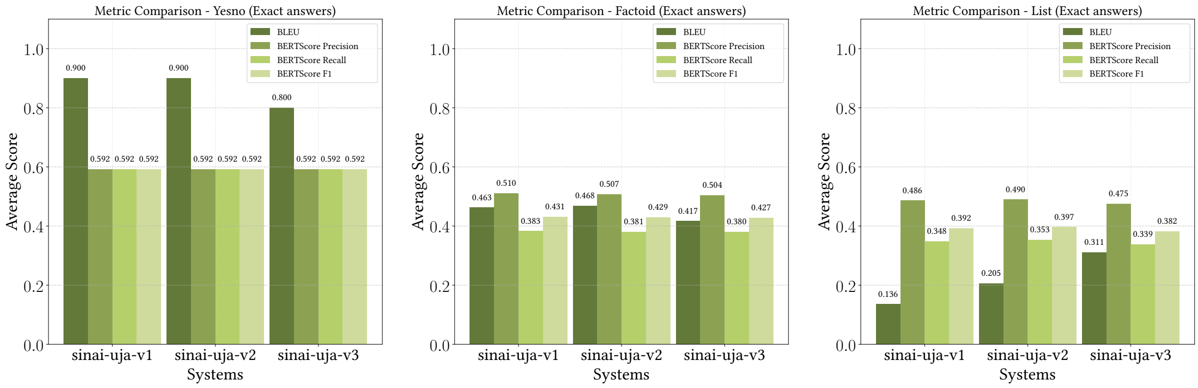
Answers for summary questions

In the realm of summary questions, we observed both successes and shortcomings in our system’s ability to generate concise and accurate summaries based on the retrieved information. For the question concerning Gene Set Enrichment Analysis (GSEA) (id: *677ecc12592fa48873000028*), the generated summary included details about statistical significance and database integration, which were not the primary focus of the expert-provided concise explanation of GSEA’s core function. This suggests a tendency to include potentially extraneous information. Similarly, for the question about Over-Representation Analysis (ORA) (id: *677ed8b8592fa4887300002b*), the system's summary, while capturing the main idea, also incorporated a discussion of ORA's limitations, which deviated from the purely descriptive nature of the first expert snippet and only partially aligned with the second.

On the other hand, for the summary question regarding gender-affirming care in minors "Should gender-affirming surgery be performed in people under 18 years of age?" (id: *6593d2e006a2ea257c000019*), the generated ideal answer effectively captured the cautious yet potentially beneficial stance, reflecting the nuances and emphasis on individualization found in the expert feedback. Similarly, for the question about the adverse effects of rare earth elements (REEs) (id: *6772e791592fa4887300000b*), the system provided a comprehensive summary listing various health risks associated with REEs, aligning well with the diverse information presented in the expert snippets. These examples of high BLEU scores indicate the system's capability to synthesize coherent and relevant summaries when the source information is well-aligned and the task is to consolidate multiple related facts or perspectives.

Error overview

The overall performance of our systems, as depicted in Figures 3a to 7 that showcase the avegare performance of our system across each question and answers types; reveals varying degrees of success across different question and answer types. For exact answers, system v1 and v2 achieved a perfect BLEU score of 0.9 for Yes/No questions, slightly outperforming v3 (0.8), while their BertScore F1 remained consistently high (around 0.59) across all variants (Figure 3a). In Factoid exact answers (Figure 3b), v1 and v2 showed marginally higher BLEU scores (0.46-0.468) compared to v3 (0.416), with BertScore F1 scores remaining very close across all systems (around 0.42-0.43). A notable improvement was observed in List exact answers (Figure 3c), where v3 achieved a significantly higher BLEU score (0.31) compared to v1 (0.135) and v2 (0.205), indicating that the prompt optimizations in v3 were particularly effective for generating more accurate lists, even if the BertScore F1 remained relatively consistent (around 0.38-0.39) across the variants.



(a) Exact Yes/No answers.

(b) Exact Factoid answers.

(c) Exact List answers.

Figure 3: Evaluation for exact answers across system variants.

Turning to ideal answers, the BLEU scores are generally higher than for exact answers, reflecting the greater flexibility in phrasing for longer generated texts. For Yes/No ideal answers (Figure 4), v2 showed the strongest performance with a BLEU score of 0.88, surpassing both v1 (0.758) and v3 (0.782), while BertScore F1 scores were consistently around 0.23. In Factoid ideal answers (Figure 5), v1 led with a BLEU score of 0.777, followed by v3 (0.701) and v2 (0.651). For List ideal answers (Figure 6), v3 and v2 performed similarly and slightly better than v1 in terms of BLEU (0.801 and 0.800 respectively vs. 0.767). Finally, for Summary ideal answers (Figure 7), v1 achieved the highest BLEU score (0.716), with v3 close behind (0.702) and v2 performing slightly lower (0.64). Across all ideal answer types, the BertScore F1 remained relatively low (around 0.17-0.23), suggesting that while the generated ideal answers share

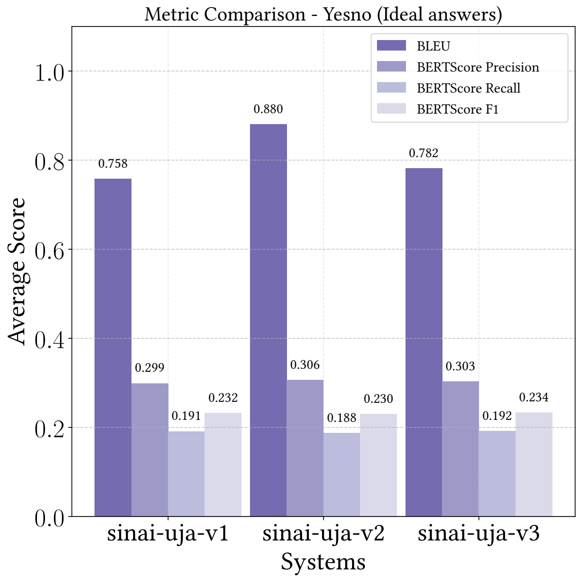


Figure 4: Evaluation for ideal Yes/No answers.

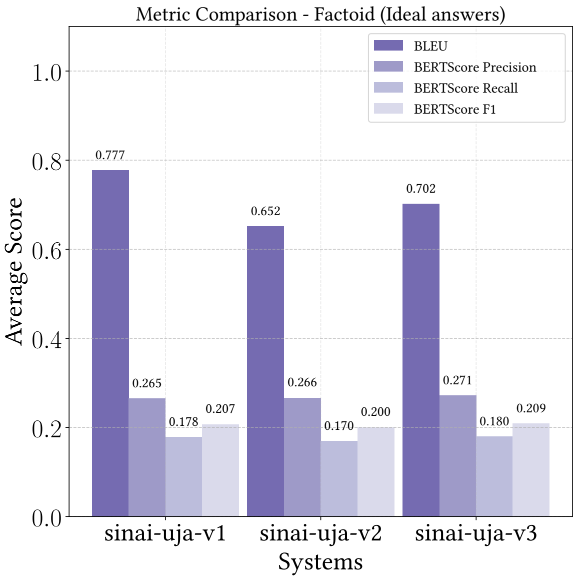


Figure 5: Evaluation for ideal Factoid answers.

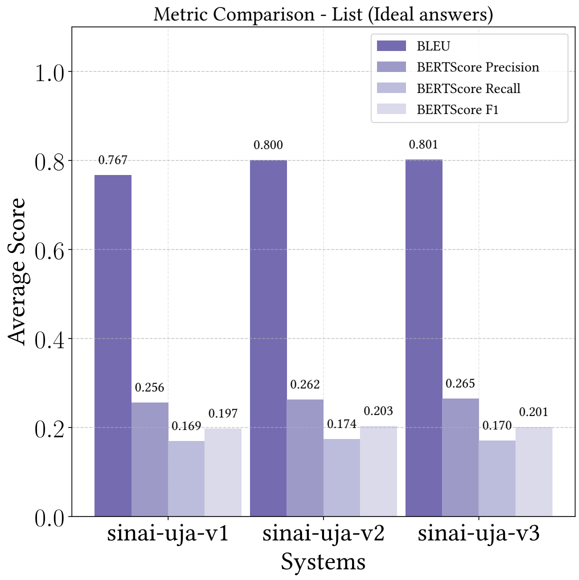


Figure 6: Evaluation for ideal List answers.

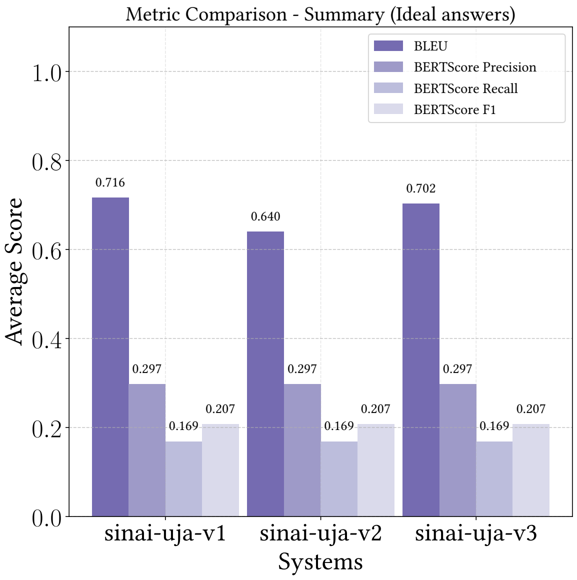


Figure 7: Evaluation for Summary answers.

n-gram overlap with the gold standards, there might be room for improvement in semantic precision or conciseness as perceived by BertScore.

Overall, these metrics highlight that no single system variant consistently outperformed the others across all metrics and question types, indicating that the subtle changes in prompts and data snapshots had specific, rather than universal, impacts on performance.

6. Conclusion

In this paper, we presented a comprehensive Retrieval-Augmented Generation (RAG) system designed to address the dynamic and challenging biomedical semantic Question Answering tasks within the BioASQ Task Synergy 13. Our multi-stage pipeline integrated a preliminary question analysis with an entity extraction methodology for query enhancement, a robust semantic search for document and snippet retrieval, and a fine-tuned Llama-based Large Language Model for answer generation across various formats. This work underscored the critical role of interactive expert feedback in refining systems for evolving biomedical information.

Our evaluation revealed several key insights into the performance of RAG architectures in this demanding domain. While our initial document retrieval performance, though improved by updating the PubMed corpus, remained modest overall, our systems demonstrated exceptional capabilities in snippet extraction. Notably, our RAG-based pipelines achieved the highest snippet extraction scores in Round 4, significantly outperforming other competitors. This highlights the strength of our dynamic indexing and semantic matching strategy in identifying the most relevant textual evidence from a given set of documents, even if the initial document pool was not perfectly optimized.

The analysis of exact answers showed varied performance across question types. For Yes/No questions, our systems achieved high accuracy, though with a tendency for overly permissive inference, leading to false positives when explicit confirmation was absent. In Factoid questions, we observed a mixed bag, with successes in extracting precise facts but also challenges in handling nuanced geographical or contextual information, and a propensity for over-generation of plausible but incorrect answers. For List questions, our prompt optimizations in system v3 led to improved precision and F-Measure, though system v1 surprisingly maintained the highest recall, suggesting a trade-off between specificity and coverage.

Regarding ideal answers, our systems generally produced coherent and informative summaries, especially for Yes/No and some List questions, achieving high BLEU scores. However, the error analysis identified areas for improvement, including issues with numerical accuracy, the introduction of extraneous details, and occasional over-synthesis of information not strictly central to the expert's ideal summary. The consistent, albeit low, BertScore F1 across ideal answers suggests room for enhancing semantic precision and conciseness.

In conclusion, our participation in BioASQ Task Synergy 13 provided invaluable insights into the practical challenges and opportunities of applying RAG to real-world, dynamic biomedical QA. The results confirm the effectiveness of our snippet retrieval approach and the potential of LLMs in synthesizing complex information. Future work will focus on refining our query enhancement strategies to improve initial document retrieval, developing more robust prompt engineering techniques to balance precision and recall in exact answer generation, and enhancing the LLM's ability to produce more concise and precisely targeted ideal answers by mitigating over-generation and improving numerical accuracy. These efforts will further strengthen our RAG pipeline, moving towards more robust and reliable biomedical question answering systems.

Acknowledgments

This work was partly supported by the grants FedDAP (PID2020-116118GA-I00), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) and CONSENSO (PID2021-122263OB- C21) funded by MCIN/AEI/10.13039/501100011033, "ERDF A way of making Europe" and "European Union NextGenerationEU/PRTR". This work was also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and, occasionally, Perplexity to assist with LaTeX syntax as well as grammar and spelling checks in English. After using these tools, the author(s) carefully reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, 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: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

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

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

[4] B.-W. Huang, Generative large language models augmented hybrid retrieval system for biomedical question answering, CLEF Working Notes (2024).

[5] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. G. et al., The llama 3 herd of models, 2024. arXiv:2407.21783.

[6] A. K. Gururajan, E. Lopez-Cuena, J. Bayarri-Planas, A. Tormos, D. Hinjos, P. Bernabeu-Perez, A. Arias-Duart, P. A. Martin-Torres, L. Urcelay-Ganzabal, M. Gonzalez-Mallo, S. Alvarez-Napagao, E. Ayguadé-Parra, U. C. D. Garcia-Gasulla, Aloe: A family of fine-tuned open healthcare llms, 2024. arXiv:2405.01886.

[7] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, A. Bosselut, Meditron-70b: Scaling medical pretraining for large language models, 2023. URL: https://arxiv.org/abs/2311.16079. arXiv:2311.16079.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[9] R. C. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. Rasmussen, C. Matuszek, B. S. Alexandrov, Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization, 2024. URL: https://arxiv.org/abs/2410.02721. arXiv:2410.02721.

[10] M. Li, H. Kilicoglu, H. Xu, R. Zhang, Biomedrag: A retrieval augmented large language model for biomedicine, Journal of Biomedical Informatics 162 (2025) 104769. URL: https://www.sciencedirect.com/science/article/pii/S1532046424001874. doi:https://doi.org/10.1016/j.jbi.2024.104769.

[11] S. Raza, D. J. Reji, F. Shajan, S. R. Bashir, Large-scale application of named entity recognition to biomedicine and epidemiology, PLOS Digital Health 1 (2022) e0000152.

[12] P. Deka, A. Jurek-Loughrey, P. Deepak, Improved methods to aid unsupervised evidence-based fact checking for online health news, Journal of Data Intelligence 3 (2022) 474–504.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[14] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, Bleu: a method for automatic evaluation of machine translation, 2002. doi:10.3115/1073083.1073135.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.