AQAMS and AQAMS2: Multi Agent Systems for Biomedical Question Answering

Notebook for the BioASQ Task 13b Lab at CLEF 2025

Johanna Angulo[1,*], Víctor Yeste[1]

[1]School of Science, Engineering and Design, Universidad Europea de Valencia, Paseo de la Alameda, 7, 46010 Valencia, Spain

Abstract

This paper presents AQAMS and AQAMS2, two multi-agent biomedical question answering systems developed for the BioASQ 13B challenge. Both systems employ a two-agent architecture comprising a Researcher Agent for evidence gathering and a Writer Agent for answer synthesis. AQAMS targets open-domain retrieval (Phases A/A+) using hybrid vector search and PubMed API integration, while AQAMS2 focuses on snippet-based analysis (Phase B) incorporating biomedical NER and UMLS concept mapping.

AQAMS achieved 87.1% average yes/no accuracy in Phase A+, with notable adaptation capability demonstrated by 173% relative improvement in factoid questions between batches. However, performance was constrained by incomplete vector database indexing (15% of PubMed corpus). AQAMS2 demonstrated substantially improved performance in Phase B, achieving up to 95.45% yes/no accuracy, 63.1% list F-measure, and 18.88% ROUGE-2 F1 for ideal answers, representing peak improvements of 51–183% across metrics compared to Phase A+. In comprehensive evaluation across all batches, AQAMS2 demonstrated competitive performance while indicating areas for improvement.

The comparative analysis reveals that multi-agent coordination performs more effectively with high-quality, focused context compared to broad retrieval scenarios. The architecture demonstrates measurable benefits for systematic integration of heterogeneous tools and adaptive processing capabilities. While the results validate the viability of the multi-agent approach for biomedical question answering, they indicate that further development is needed to achieve performance comparable to leading systems in this challenging domain.

Keywords

Biomedical Question Answering, GenAI, AI Agents, Large Language Models, Prompt Engineering, RAG, BioASQ 13B

1. Introduction

Recent advances in Natural Language Processing (NLP) and Information Retrieval (IR) have significantly enhanced the capability to answer biomedical questions at scale [1, 2]. The BioASQ challenge series provides a standardized benchmark for evaluating such systems on real-world biomedical queries, fostering the development of advanced approaches that serve researchers and clinicians. Task 13B of BioASQ (2025) continued this initiative by requiring participants to build systems capable of retrieving relevant information and generating accurate answers for questions posed by biomedical experts [3]. This task presents particular challenges due to the volume and complexity of biomedical literature, which contains specialized terminology and undergoes rapid evolution [4]. This paper describes two complementary systems—AQAMS and AQAMS2—developed for BioASQ 13B Task B, which leverage the synergy between information retrieval and large language models. Both systems adopt a multi-agent pipeline architecture where a Researcher Agent handles information gathering or analysis, followed by a Writer Agent that produces final answers. AQAMS was designed for the open-retrieval scenario of Phase A and A+ (where systems retrieve answers from the entire biomedical literature), while AQAMS2 was tailored for the closed-snippet scenario of Phase B (where relevant text snippets are provided by the organizers). By integrating vector-based semantic search, traditional API-based retrieval, prompt

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

*Corresponding author: Johanna Angulo (johanna.angulo@gmail.com)

johanna.angulo@gmail.com (J. Angulo); victor.yeste@universidadeuropea.es (V. Yeste)

https://www.linkedin.com/in/johannaangulo/ (J. Angulo); https://victoryeste.com/ (V. Yeste)

0009-0005-6965-0604 (J. Angulo); 0000-0002-3660-8347 (V. Yeste)

engineering techniques [5], and biomedical NER with UMLS concept mapping, our systems aim to deliver both precise "exact" answers and comprehensive "ideal" answers. We evaluate the performance of AQAMS and AQAMS2 on the official BioASQ 13B test batches and compare their effectiveness across different question types (yes/no, factoid, list, summary). We also analyze the impact of our design choices in the context of recent work in biomedical question answering, including hybrid retrieval pipelines and few-shot prompt templates, and discuss potential improvements. The remainder of this paper is organized as follows: Section 2 describes the BioASQ 13B task setup; Section 3 provides an overview of the techniques used; Section 4 details the methodology of our AQAMS and AQAMS2 systems; Section 5 presents the experimental results; Section 6 offers analysis and discussion; and Section 7 outlines future research directions.

2. Overview of the Task 13B

BioASQ is a series of international challenges that promotes advances in large-scale biomedical semantic indexing and question answering [6, 7, 8]. The BioASQ challenge serves as a long-running benchmark for biomedical question answering, with recent editions (2023–2024) demonstrating notable progress in the field [9, 8]. Competing systems address distinct challenges across multiple phases. Phase A focuses on information retrieval from PubMed articles to identify relevant snippets for specific questions. In Phase A+, participants must address multiple question types (yes/no, factoid, list, summary) by retrieving relevant literature from PubMed and providing both exact answers and paragraph-length ideal answers. Phase B requires participants to provide exact and ideal answers using provided text snippets. Overall performance has improved consistently year over year, and in BioASQ 2024, "most of the participating systems achieved competitive performance, suggesting the continuous advancement of the state-of-the-art in the field" [3]. Large language models have made a significant impact on biomedical question answering performance. One study demonstrated that GPT-4 and GPT-3.5 (ChatGPT) in a zero-shot setting, when provided with relevant snippets, could nearly match the performance of the best BioASQ 11b systems on factoid and list questions [10]. This finding underscores the rapid advancement achieved by general domain models even without domain-specific fine-tuning. In 2024 (BioASQ 12b), this trend continued with many teams integrating generative large language models with retrieval systems. Nentidis et al. reported strong results from systems that combine information retrieval with GPT-based answer generation [9]. The BioASQ challenge results from 2023–2025 highlight that ensemble approaches combining retrieval modules with large pre-trained models now dominate biomedical question answering, significantly outperforming earlier methodologies. The availability of the manually curated BioASQ dataset, containing thousands of question-answer pairs and gold standard snippets, continues to enable these advances [11].

3. Overview of Used Techniques

This section summarizes recent advances in the field and situates the techniques employed in our participating systems within the context of the current state of the art.

3.1. Multiagent systems

The developed system is based on multiagent architecture principles, defined as a computational paradigm where multiple autonomous entities (agents) collaborate to solve complex problems that exceed the capabilities of monolithic systems [12]. In the biomedical context, multiagent systems have demonstrated particular effectiveness in handling the inherent complexity of the medical domain, where different aspects of information processing require specialized expertise [13].

According to Ferber's taxonomy [14], the implemented Q&A system corresponds to a cooperative multiagent architecture where agents share common objectives and collaborate to maximize the global

utility of the system. This classification aligns with the coherence and precision requirements necessary in biomedical applications [15].

Following Russell and Norvig's taxonomy [16], the implemented agents can be classified as goal-based agents that maintain internal representations of the problem state and use domain-specific knowledge to make decisions.

3.2. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a pivotal approach for biomedical question answering in recent years. RAG systems integrate a document retrieval component with a text generation model, enabling large language models to ground their answers in external knowledge [17]. This is particularly important in biomedicine, where up-to-date factual accuracy is required beyond an LLM's static training data. Multiple studies from 2023–2024 demonstrate the effectiveness of RAG in this domain. For example, Merker et al. built a RAG-based pipeline that first retrieves PubMed abstracts using BM25 and neural re-ranking, then feeds top snippets into GPT-3.5 or GPT-4 to generate answers [18]. They observed that providing relevant snippets significantly improves answer accuracy, and in their experiments, answers based on snippet-context outperformed those generated from full abstracts.

Similarly, Jeong et al. introduced Self-BioRAG, a framework that augments a transformer model with biomedical knowledge via retrieval and even a self-reflection step [19]. Self-BioRAG selectively pulls in domain-specific content from curated corpora and knowledge bases and was shown to boost accuracy by approximately 7–8% over baseline LLMs on medical QA benchmarks. These improvements echo the general findings in open-domain QA that RAG can reduce hallucinations and inject up-to-date facts [17]. Indeed, integrating retrieval is seen as essential in biomedicine: Zhang et al. emphasize that RAG is "a pivotal innovation that improves the accuracy and relevance of LLM responses by integrating LLMs with a search engine and external sources of knowledge" [17]. In practice, many state-of-the-art biomedical QA systems now use RAG or related hybrid strategies. Traditional pipelines had separate information retrieval and answer stages [20], but modern RAG blurs this boundary, often iteratively retrieving and generating.

3.3. Use of PubMed E-Utilities API in Biomedical Question Answering

Given that PubMed/MEDLINE is the primary source of biomedical literature, many QA systems rely on NCBI's E-Utilities API to fetch relevant articles and snippets. The PubMed E-Utilities (RESTful web services) enable querying the vast biomedical literature programmatically, which is crucial for up-to-date QA. In the BioASQ challenges, for instance, participants are explicitly tasked with retrieving answers from designated resources including PubMed/MEDLINE articles [21]. Systems therefore use the E-Utilities (such as the esearch and efetch endpoints) to automate literature search and retrieval of article metadata, abstracts, and even full-text snippets when available. Merker et al. describe using the PubMed search API to retrieve up to 200 abstracts per query as an initial document set, before applying local re-rankers [18]. This highlights that even with advanced neural methods, the pipeline often begins with PubMed API calls to ensure comprehensive coverage of relevant papers. The NCBI E-Utilities have thus become a standard tool: many systems incorporate modules or use libraries to run keyword queries, possibly with MeSH term filtering. For example, the BioASQ dataset paper notes that curators formulated multiple PubMed queries with field tags and MeSH filters for each question, retrieving documents and snippets that systems can use [11]. The advantage of using the API is that it provides access to the entire up-to-date MEDLINE corpus (33+ million citations) in real time [22, 23].

3.4. Prompting Strategies and Few-Shot Prompting in Biomedical QA

With the advent of large language models, prompting strategies have become critical for biomedical QA performance. In particular, few-shot prompting – providing a handful of example Q&A pairs or instructions in the prompt – can substantially improve an LLM's ability to specialize to biomedical questions without fine-tuning. Research in 2023–2024 demonstrates that carefully engineered prompts

can bridge the gap between general-purpose LLMs (like GPT-3.5/GPT-4) and domain-specific needs [24, 25].

Few-shot prompting is especially useful in scenarios with limited training data. Ateia and Kruschwitz explored the few-shot performance of open-source versus commercial LLMs in biomedical tasks [21]. They observed that the performance gap between GPT-4 and smaller open models can be largely closed by providing around 10 exemplars (few-shot) – Mixtral (an 8×7B ensemble) became competitive with GPT-4 in a 10-shot setting [21]. This finding underlines that domain-specific examples help models follow instructions and leverage biomedical terminology more effectively. Another important strategy is chain-of-thought prompting, where the model is guided to reason step-by-step. Merker et al. employed a few-shot chain-of-thought prompt to extract answer-relevant snippets from abstracts [18]. By showing the model examples of reasoning through a passage to find key information, they improved snippet extraction performance.

Prompting can also specify format and context: e.g., instructing the model to answer with a certain style or to cite sources. In general, prompt design has become an art in biomedical QA, often involving trial-and-error to see which instructions yield the most factual and concise answers. Empirical results in 2023 indicate that even zero-shot, well-crafted prompts allow GPT-3.5/4 to rival specialized systems on many tasks [10]. However, adding a few examples (few-shot) tends to further boost reliability and reduce errors, especially for complex multi-step questions or when dealing with rare biomedical terms. In summary, effective prompting – including few-shot exemplars, step-by-step cues, and explicit instructions – is now recognized as a key factor in getting the best out of large generative models for biomedical question answering [26, 21]. It allows QA systems to leverage general LLMs while injecting domain knowledge and context through the prompt, often obviating the need for extensive model fine-tuning.

3.5. Role of Named Entity Recognition (NER) in Biomedical Semantic QA

Named Entity Recognition (NER) plays a foundational role in biomedical QA systems. Biomedical questions and texts are dense with technical entities (drug names, genes, diseases, etc.), and identifying these entities is crucial for understanding queries and retrieving precise answers [27]. Many pipelines therefore include a BioNER step to tag key terms in the question and candidate passages. This has several benefits: (1) it helps map words to standardized concepts (e.g., recognizing that "heart attack" corresponds to myocardial infarction), (2) it guides the retrieval module by focusing on important keywords, and (3) it can assist in answer extraction by ensuring the system outputs a properly identified entity. Early BioASQ challenges recognized this – the benchmark datasets even provide "concepts" (from UMLS, MeSH, etc.) associated with each question [11].

These concepts essentially come from NER and linking performed by the curators, and they can be used by QA systems for query expansion or checking answer correctness. Traditional approaches to biomedical QA often featured an NER component [28]. By tagging entities in both questions and documents, they could better match relevant information. NER is also critical in identifying the type of answer expected. A question like "What enzyme digests fibrin?" implies the answer should be an enzyme name – if the system's NER can detect that, it can constrain the answer extraction to enzyme entities [29].

Modern transformer-based QA models implicitly learn some entity recognition, but integrating explicit NER has still proven useful in many cases. Jin et al. in their survey highlight that handling of synonyms and variant entity names is a big challenge in biomedical QA – which is essentially an NER+normalization problem [4].

3.6. UMLS APIs and Mapping NER to UMLS Concepts

The Unified Medical Language System (UMLS) is a comprehensive repository of biomedical vocabularies and concepts, and it has been widely used to enhance QA systems. By mapping recognized entities (via NER) to UMLS concepts, systems can normalize different terms to the same underlying idea and

even retrieve additional information (synonyms, definitions, related concepts) through UMLS services. Several approaches from 2023–2025 explicitly integrate UMLS knowledge. One line of work is to use UMLS for query expansion or synonym resolution [30]. Once entities are linked, the UMLS API (via the NIH UMLS Terminology Services) can be used to fetch concept details: e.g., hierarchical relations, or all synonyms in various terminologies. Incorporating this knowledge can improve recall – a QA system might automatically expand a query with UMLS synonyms or check an answer candidate against UMLS definitions. Recent research has also looked at injecting UMLS knowledge into language models. Park et al. introduced a method of infusing UMLS knowledge into a transformer via adapter-based fine-tuning [31]. By semantically partitioning the UMLS knowledge graph and training adapters, they imbued the model with domain knowledge that improved performance on biomedical QA datasets [31].

3.7. Importance of Evidence-Grounded Responses in Biomedical QA

Biomedical Question Answering aims to extract an answer to the given question from a biomedical context [32]. However, despite tremendous growth, BQA still needs to mature and faces many challenges, such as corpora scaling, annotation, lexical answer type prediction, and complex terminology [33]. In the biomedical domain, it is critical that QA systems provide evidence-grounded responses – that is, answers supported by authoritative sources (e.g., peer-reviewed articles). Unlike casual Q&A, an incorrect answer in healthcare could have serious consequences. Hence, recent biomedical QA research places great emphasis on grounding and justification [34].

This need for evidence-grounded responses in the biomedical field inspired us to extend our system beyond the BioASQ 13B Challenge and, on top of exact and ideal answers, provide a "supported answer" with references to snippets used to build the answer. We also implemented this in the user interface to provide a user-friendly visualization of the system responses, helping to trust the AI system or realize when the answer does not have the expected level of evidence support.

4. Methodology

The objective of our research is to develop a multi-agent, hybrid question answering system where modular resources and knowledge bases interact synergistically.

For that end, we have implemented a two-agent architecture for both tasks, with variations customized to the retrieval scenario of Phase A/A+ and the snippet-based scenario of Phase B. Figure 1 outlines the general architecture of our systems (AQAMS for Phase A/A+ and AQAMS2 for Phase B).

The Researcher Agent is responsible for interpreting the question and gathering relevant evidence, which includes querying databases or APIs in AQAMS, and performing named entity recognition (NER) and UMLS concept mapping in AQAMS2. The Writer Agent is a large language model (GPT-4) prompted to generate the final answers (exact and ideal, with an additional supported explanation in Phase B). We selected an OpenAI GPT model for the Writer Agent due to its strong few-shot learning capabilities and fluent language generation [35]. To ensure domain specificity and factual accuracy, the Writer Agent is always provided with contextual information—either retrieved snippets or extracted entities—through a structured prompt. Below, we describe each system in detail.

4.1. AQAMS: Phase A/A+ System

AQAMS (Automated Question Answering with Multi-agent System) was developed for the open-domain scenario of BioASQ 13B Phase A and A+. The system implements a two-stage pipeline for biomedical question answering. In the first stage (Phase A, document/snippet retrieval), AQAMS's Researcher Agent performs a hybrid retrieval process. The agent first queries a local Qdrant vector database that indexes PubMed article embeddings. We constructed this vector store using LlamaIndex embeddings of PubMed abstracts, enabling semantic search beyond keyword matching. However, due to time constraints, the vector index was only partially populated (approximately 15 per cent of the PubMed

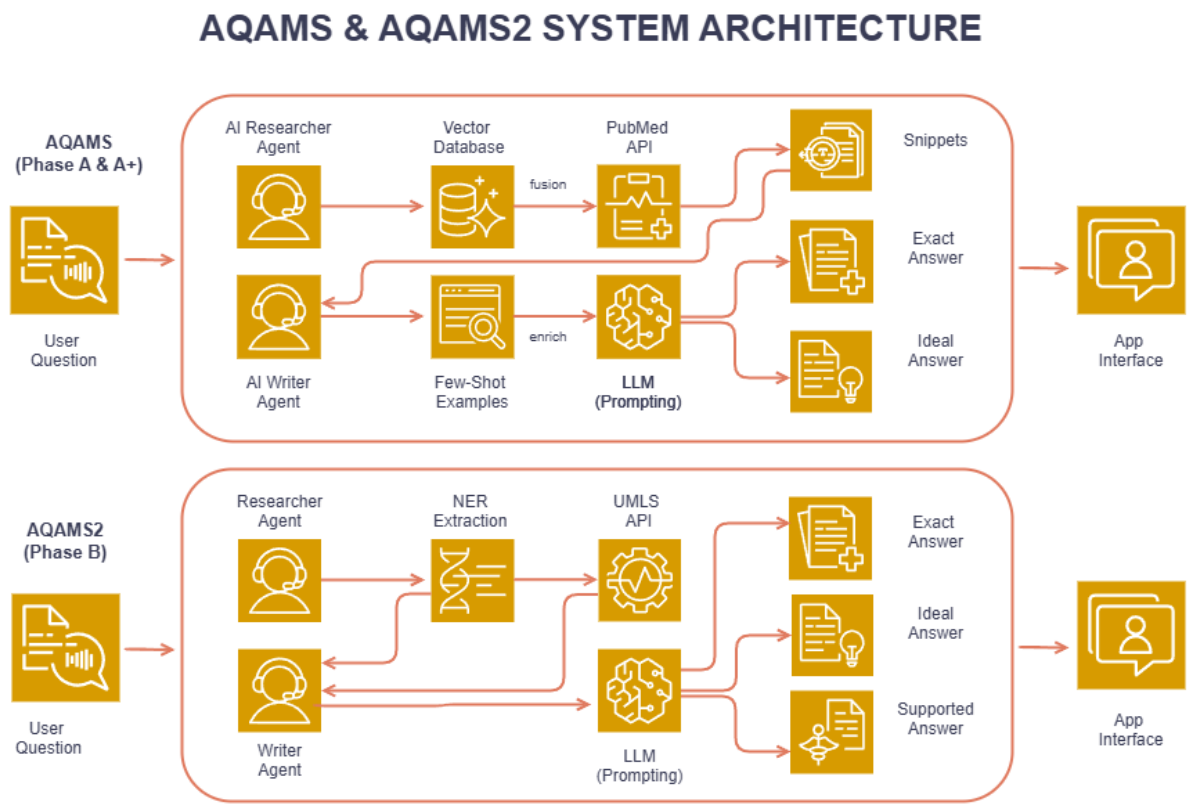

Figure 1: The architecture orchestrates the collaboration of distinct yet interconnected components mediated by two primary agents: the Researcher Agent and the Writer Agent. (Source: Elaborated by the authors).

corpus). This partial coverage limited the standalone effectiveness of vector search—if the relevant information was not in the indexed portion, the semantic query would fail.

Therefore, the Researcher Agent employs a backoff strategy: if the vector search yields insufficient results or if additional diversity is needed, it falls back to the NCBI PubMed E-Utilities API to fetch relevant articles via traditional keyword search. This hybrid approach combines the advantages of dense semantic retrieval and classic lexical retrieval, which has been shown to improve recall in biomedical question answering systems [9]. Specifically, semantic search helps retrieve conceptually relevant documents that lexical search might miss (for example, retrieving an article about "myocardial infarction" for a query on "heart attack"), while the PubMed API (a BM25-based engine) ensures up-to-date and high-precision results for exact keyword matches. Similar hybrid strategies combining BM25 with transformer-based dense retrieval have proven effective in past BioASQ challenges [9]. Figure 2 illustrates a screenshot of the AQAMS question-answering system application.

In our implementation, the Researcher Agent also had access to a small repository of few-shot example queries to aid prompt formulation. If a new question was similar to a previously solved question, the agent could adapt a known effective query pattern (e.g., adding specific MeSH terms or synonyms) based on those examples. This served as a form of prompt enhancement for retrieval, inspired by recent prompt-based retrieval augmentation techniques [36]. Once relevant documents and snippets are retrieved, the pipeline enters the second stage (Phase A+), where the Writer Agent generates the answers. The Writer Agent is built on a transformer-based large language model (specifically, OpenAI GPT-3.5 and GPT-4, accessed via API) with carefully designed prompt templates. Prompt engineering was crucial: we created structured prompts that provide the question, a selection of top-ranked snippets (with citations), and explicit instructions for the answer format. The agent was instructed to produce two outputs: an exact answer and an ideal answer. For exact answers, the prompt template varied based on question type.

We found that including a few demonstrations question-answer pairs in the prompt (few-shot

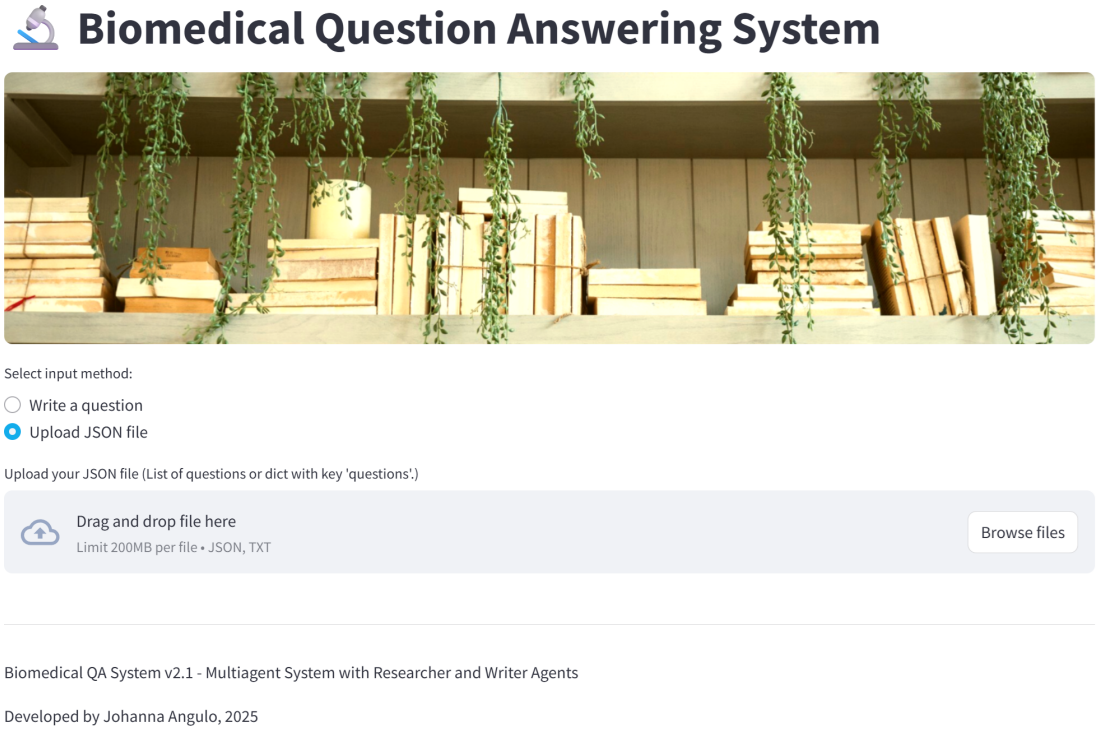

Figure 2: App User Interface of AQAMS System. & Biomedical Multiagent System. (Source: Elaborated by the authors).

prompting) further improved the consistency of the answers, consistent with findings that GPT models can follow examples to produce well-formatted outputs [35]. The ideal answer prompt was designed to elicit a comprehensive yet focused summary. It typically included a prefix such as "Using the information from the above snippets, write a detailed answer:" and instructed the model not to merely copy text but to synthesize and explain in a cohesive paragraph. The model was not allowed to cite external knowledge not found in the snippets, thereby minimizing unsupported statements. By providing relevant snippets as context, we essentially grounded the language model's generation in real biomedical evidence, an approach akin to retrieval-augmented generation that reduces hallucination [37]. The outputs from the Writer Agent (exact and ideal answers) were then returned and displayed to the user or evaluator. We deployed AQAMS via a Streamlit web interface for ease of use, supporting both single-question queries and batch processing.

4.2. AQAMS2: Phase B System

AQAMS2 was developed for BioASQ 13B Phase B, where the challenge is to generate answers strictly from the given snippets without any external retrieval. Since all necessary information is provided in the input, AQAMS2 focuses on understanding and organizing the snippet content to produce correct answers. The system maintains the same overall two-agent architecture, but the roles of the agents are adjusted.

The Researcher Agent in AQAMS2 does not perform document retrieval; instead, it conducts biomedical named entity recognition (NER) and concept identification on the provided snippets. We integrated a fine-tuned biomedical NER model (based on a RoBERTa transformer) to tag entities such as diseases, drugs, and genes in the snippets. These recognized entities were then optionally mapped to standard concepts in the UMLS (Unified Medical Language System) ontology. For example, if "heart attack" is detected in a snippet, it would be mapped to the UMLS concept for "myocardial infarction (C0027051)".

This concept mapping (entity linking) is not part of the output but is used internally to cluster and reason about the information. Linking textual mentions to unique identifiers allows the system

Exact Answer Extraction Prompt

System prompt:

You are an AI specialized in extracting specific answers from text according to strict instructions. Only output the requested answer format.

User prompt:

```
Given the following context from biomedical text snippets, answer the question based *only* on the provided
text.
Question: Which genes are related to disease X?
Question Type: list

Instructions: Extract the list of items (e.g., genes, drugs, symptoms) requested in the question. Output
*only* the items, each on a new line.

Context:
Snippet (Relevance: 0.980):
GeneA, GeneB, and GeneC have been associated with disease X.
---
Snippet (Relevance: 0.880):
Symptoms include fever, fatigue, and weight loss.

Exact Answer:
```

Figure 3: Prompt example. This prompt is used for LLM-based extraction of exact, direct answers from retrieved biomedical texts. (Source: Elaborated by the authors).

to recognize when different snippets mention the same underlying concept through synonyms or abbreviations [10]. This alignment supports the Writer Agent in avoiding redundancy and ensuring consistency in the final answer. Prior work has shown that such ontology mappings enable unified reasoning over biomedical texts with varied terminology [38].

The Writer Agent in AQAMS2 takes as input the question and the set of provided snippets (augmented with any metadata from the Researcher, such as recognized entities or identified key facts). The prompt template is structured to encourage evidence-supported generation. In BioASQ Phase B, systems can output three fields: the exact answer, the ideal answer, and an optional "supported answer" which provides supporting evidence references. We utilized this by having the Writer Agent produce a supported answer that explicitly cites snippet indices. Specifically, after generating an exact answer (e.g., a specific drug name) and an ideal answer (a short paragraph explanation), the model is prompted to generate a series of sentences like: "[Answer] This conclusion is supported by [Snippet 2], which states that '...'", incorporating snippet identifiers. Figure 3 illustrates one of the prompts used.

The exact and ideal answer generation in AQAMS2 uses similar prompt techniques as in AQAMS, with adjustments for the available input. Because the input snippets already contain the relevant information, the Writer Agent's task is closer to summarization or extraction. For factoid or list questions, the model is prompted to identify the specific entities in the snippets that answer the question (benefiting from the NER results). For yes/no questions, the model looks for affirmative or negative language in snippets (e.g., "does not affect survival" indicates a "No") and any consensus among snippets.

The ideal answers in Phase B are typically shorter than in Phase A+, since there is a limited, focused context to draw from. We tuned the prompt to avoid verbosity and not introduce extraneous information beyond the snippets.

We employed diverse prompting strategies ranging from simple task-specific instruction prompts and chain-of-thought prompts to multi-type answer extraction prompts. Our approach utilizes a complex conditional prompt incorporating multiple advanced techniques: few-shot learning elements (type-specific examples and instructions), chain-of-thought (CoT) reasoning (structured decision process), template-based prompting (conditional logic based on question type), and constrained generation prompting (strict output format requirements).

This prompting technique demonstrates ensemble-like logic, as the conditional branching mimics

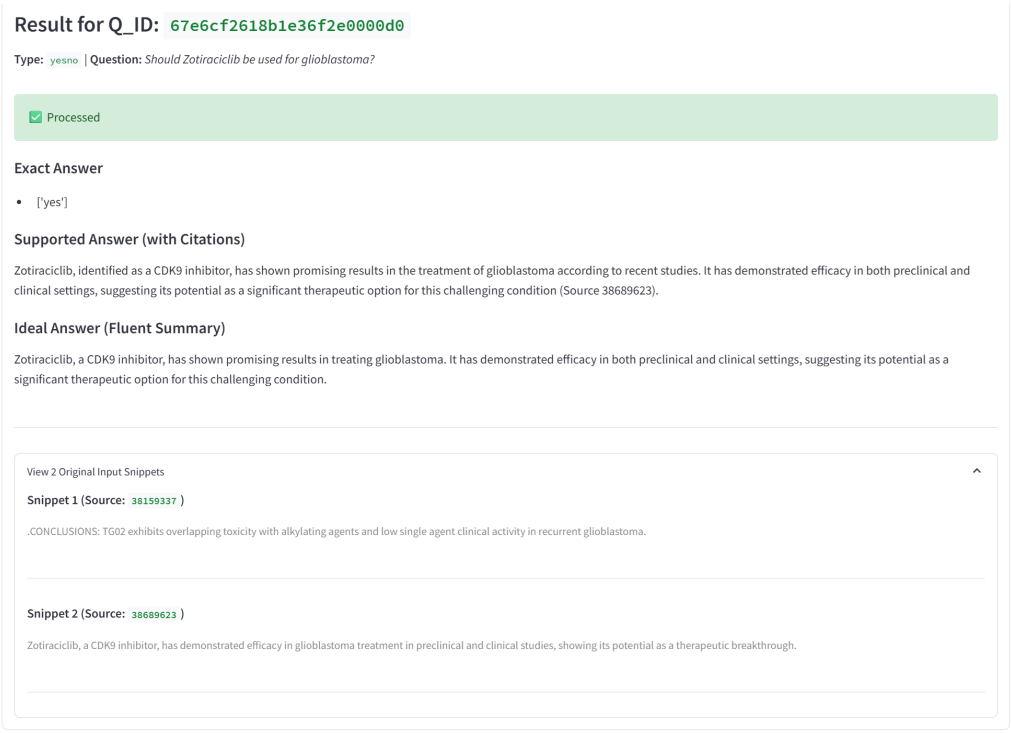

Figure 4: User interface of the AQAMS2 system application, displaying the Exact and Ideal answers, as well as the Supported Answer. (Source: Elaborated by the authors).

ensemble approaches by employing different strategies for each question type. This methodology represents what Hu et al. describe as a task-specific prompt framework that includes baseline prompts with task description and format specification, annotation guideline-based prompts, and error analysis-based instructions [39]. The conditional structure for different question types (factoid, list, yes/no) implements what Alhamzawi et al. classify as "heuristic prompts," utilizing domain knowledge and logical reasoning to guide the model's output.

AQAMS2 was also deployed with a Streamlit interface for consistency, allowing batch processing of Phase B questions. The interface displays the input snippets and the system's outputs (exact, ideal, supported answers). This mirrors a real-world scenario where a user might ask a question and want to know not just the response, but also which snippets support it. This transparency is important in biomedical question answering, as users need to trust and verify the source of an answer. Figure 4 illustrates the output of the AQAMS2 system.

5. AQAMS and AQAMS2 Results

Both AQAMS and AQAMS2 were evaluated on the official BioASQ 13B test sets. We report performance on all batches of the phases in which our systems participated: Batches 3 and 4 of Phases A and A+ for AQAMS, and Batches 3 and 4 of Phase B for AQAMS2. Tables 1–5 summarize the evaluation metrics as provided by the BioASQ organizers' evaluation tool.

5.1. Phase A Results (AQAMS)

In Phase A, the results were suboptimal due to incomplete vector database indexing (15% of PubMed corpus) and limitations of the PubMed API. These challenges significantly impacted downstream snippet extraction effectiveness, as the pool of candidate abstracts was inadequate for comprehensive retrieval.

Table 1
Phase A: Snippets

Batch	System	Mean prec. 1	Recall	F-Measure	MAP	GMAP
Batch 3	AQAMS	0.0254	0.1097	0.0392	0.0000	0.0000
Batch 4	AQAMS	0.0177	0.1318	0.0292	0.0000	0.0000

Table 2
Phase A+ Exact Answers (Batches 3-4)

Batch	System	Yes/No Acc.	Yes/No Macro F1	Factoid S.Acc.	Factoid MRR	List F-M
Batch 3	AQAMS	0.8182	0.7708	0.1500	0.1750	0.3478
Batch 4	AQAMS	0.9231	0.9023	0.4091	0.4091	0.2778

Table 3
Phase A+ Ideal Answers (Batches 3-4)

Batch	System	R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
Batch 3	AQAMS	0.2135	0.0690	0.2516	0.0778
Batch 4	AQAMS	0.1840	0.0616	0.2273	0.0763

Table 4
Phase B Exact Answers (Batches 3-4)

Batch	System	Yes/No Acc.	Yes/No Macro F1	Factoid S.Acc.	Factoid MRR	List F-M
Batch 3	AQAMS2	0.9545	0.9394	0.3000	0.3250	0.6310
Batch 4	AQAMS2	0.9231	0.8917	0.5455	0.5455	0.5277

5.2. Phase A+ Results (AQAMS)

AQAMS demonstrated heterogeneous performance across question types, with rankings varying from upper to middle tiers depending on task complexity. Yes/No questions showed substantial improvement (81.8% $\rightarrow$ 92.3% accuracy), achieving competitive performance in Batch 4 (2nd-5th position). This supports the multi-agent architecture's effectiveness in binary classification tasks. Factoid questions exhibited notable adaptation with 173% relative improvement between batches (15.0% $\rightarrow$ 40.9% strict accuracy), progressing from lower-middle to upper-middle tier performance. However, list questions maintained consistent middle-tier performance, while ideal answers showed performance decline, indicating synthesis limitations in the Writer Agent.

5.3. Phase B Results (AQAMS2)

AQAMS2 demonstrated substantially improved performance compared to Phase A+, with notable enhancements across all evaluation metrics. Yes/No questions achieved the maximum accuracy of 95.45% in Batch 3, tying for first place with approximately 30% of participating systems, while achieving 4th place overall among teams when averaged across all question types. List questions showed strong performance in Batch 3 (63.1% F-measure, 3rd-5th position), providing evidence consistent with the effectiveness of NER integration within the Researcher Agent, though formal ablation studies would be needed to establish causality. Factoid questions exhibited considerable adaptation capability (30.0% $\rightarrow$ 54.5%, +82% relative improvement). Ideal answers showed consistent upper-tier performance with 183% improvement over Phase A+ (18.4% vs 6.5% ROUGE-2 F1).

Table 5
Phase B Ideal Answers (Batches 3-4)

Batch	System	R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
Batch 3	AQAMS2	0.3567	0.1888	0.3643	0.1795
Batch 4	AQAMS2	0.3077	0.1787	0.3274	0.1837

6. Analysis and Discussion

6.1. Prompt Engineering and Few-Shot Integration

Our Writer Agent employs a hierarchical prompt structure that adapts to question types and integrates contextually relevant few-shot examples. The system-level prompt establishes the biomedical expert persona and output format rules, while question-specific templates provide targeted instructions.

Example Phase A+ Prompt Template (Yes/No Questions):

```
Question: {question}
Type: yesno
Relevant PubMed Publications (Snippets):
{ranked_snippets_with_pmids}

Instructions: Based strictly on the provided evidence, answer 'yes'
   or 'no'.
If evidence is ambiguous or insufficient, default to 'no'.
Respond with JSON: {"exact_answer": "yes|no", "ideal_answer": "..."}
```

Few-Shot Integration Mechanism: Our few-shot integration operates through dynamic example insertion during chat completion generation. The system constructs message sequences by: (1) placing the system prompt first, (2) inserting retrieved few-shot examples as alternating user-assistant message pairs, and (3) appending the current question as the final user message.

The `generate_chat_completion` function receives few-shot examples as a parameter and automatically integrates them into the conversation flow before processing the current query. For example, if a factoid question about drug mechanisms is processed, the system retrieves similar question-answer pairs from our Qdrant vector store and inserts them as demonstration examples, enabling the model to learn the expected response format and reasoning patterns for that specific question type.

This approach provides contextual learning without requiring model fine-tuning, allowing the LLM to adapt its behavior based on relevant historical examples.

6.2. Architectural Validation and Context Quality Effects

The comparative performance between phases provides empirical support for our multi-agent architecture. The 183% improvement in ideal answers from Phase A+ to Phase B indicates that agentic coordination performs more effectively with high-quality, focused context compared to broad retrieval approaches. This observation aligns with our theoretical framework wherein traditional NLP pipelines implement static processing sequences, while our agentic framework enables contextual decision-making where specialized agents adapt strategies based on question characteristics.

6.3. NER Integration Success as Tool vs. Agent

Phase B's improved list performance (85% enhancement: 31.3% $\rightarrow$ 57.9% F-measure) supports our architectural decision to integrate biomedical NER as a specialized tool within the Researcher Agent rather than implementing a separate agent. This design adheres to the principle that agents function as autonomous decision-making entities, while tools serve as functional components activated by agents based on contextual requirements. NER as a specialized tool could be integrated into the Phase A system

as well to potentially improve performance. We intentionally did not implement this enhancement during the challenge in order to compare performance with and without NER integration, providing a controlled evaluation of the tool's impact on system effectiveness.

6.4. Cross-Phase Architectural Insights

Retrieval Dependency: Phase A performance indicates that agentic architectures exhibit critical dependency on knowledge base completeness. The incomplete vector indexing (15% of PubMed) created a retrieval bottleneck that constrained downstream performance.

Context Synthesis Performance: The 3× improvement in ROUGE-2 F1 (0.1888 vs. 0.0616) indicates that agent coordination performs effectively when synthesizing multiple information sources within curated contexts.

Adaptation Capability: The substantial factoid improvements in both phases (173% in A+, 82% in B) demonstrate the system's adaptability and learning capacity—a characteristic advantage of distributed decision-making architectures.

6.5. Competitive Positioning and Limitations

Our systems achieved competitive rankings across most evaluation metrics. AQAMS maintained consistent performance in the upper tier for yes/no questions and middle-tier performance in other categories. AQAMS2 demonstrated balanced performance across all question types, ranking within the upper quartile with particular effectiveness in structured extraction tasks.

Performance Limitations: The declining ideal answer performance in Phase A+ indicates challenges in answer synthesis when processing potentially irrelevant or incomplete retrieval results. Additionally, the Phase B list performance decline between batches (63.1% → 52.8%) suggests potential overfitting or batch-specific adaptation challenges.

6.6. Key Findings

Our multi-agent approach demonstrates competitive performance that suggests potential benefits of coordinated agent interaction for biomedical question answering tasks. The comparative performance between Phase A+ and Phase B indicates that the architecture performs more effectively with high-quality, focused context compared to open-domain retrieval scenarios, though the observed differences could be attributed to multiple factors including task complexity and data availability. AQAMS2 achieved a competitive position overall among participating teams with notable variation in performance (4th position in Batch 3, declining in Batch 4). While this ranking demonstrates the viability of the multi-agent approach, the results indicate that more development is needed to achieve performance comparable to leading systems.

7. Future Work

There are several lines of work to explore for improving both AQAMS and AQAMS2 in future editions:

7.1. Complete and Enhanced Indexing

The priority for AQAMS is to finish populating the PubMed vector index and possibly update it with the latest literature. A full corpus semantic index would improve recall significantly. In addition, integrating a neural re-ranker could help sort the retrieved snippets by relevance before passing to the Writer Agent.

7.2. Better Entity Handling and Answer Post-processing

Our use of NER and UMLS was a step towards deeper understanding. We plan to build on this by implementing an entity-based post-processor. For example, after the Writer Agent produces an answer, we could verify if all UMLS-linked concepts in the snippets that seem crucial to the question are reflected in the answer.

7.3. Long-term Agentic Vision

Based on our results analysis, we identify key architectural improvements that extend beyond traditional NLP pipeline enhancements:

Multi-Source Validation: Developing cross-validation mechanisms between Qdrant and PubMed results to ensure retrieval robustness and reduce dependency on single knowledge sources.

Context Quality Assessment: Implementing agents capable of evaluating context quality and dynamically adjusting retrieval strategies, addressing the critical finding that agentic coordination excels with high-quality, focused context rather than broad retrieval.

Beyond these immediate improvements, we plan to continue advancing our agentic systems approach to question answering, with the long-term goal of developing increasingly autonomous systems that can independently reason, validate, and refine their responses with minimal human intervention.

The empirical evidence from our Phase B results—demonstrating 183% improvement in ideal answers through focused context coordination—provides methodological support for this research direction. We believe these developments are important for building reliable question answering systems that can assist researchers and clinicians in navigating the expanding biomedical literature.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] K. A. Hambarde, H. Proenca, Information retrieval: recent advances and beyond, IEEE Access 11 (2023) 76581–76604.

[2] X. Cai, C. Wang, Q. Long, Y. Zhou, M. Xiao, Dataset distillation for domain LLM training, arXiv preprint arXiv:2501.15108 (2025).

[3] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, Bioasq at clef2024: The twelfth edition of the large-scale biomedical semantic indexing and question answering challenge, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 490–497.

[4] 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 Computing Surveys 55 (2023) 35:1 – 35:36. doi:10.1145/3490238.

[5] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, Y. Wang, An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study, JMIR Med Inform 12 (2024) e55318. URL: https://medinform.jmir.org/2024/1/e55318. doi:10.2196/55318.

[6] 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, et al., Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[7] 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.

[8] G. Tsatsaronis, et al., An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics (2015). URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6. doi:10.1186/s12859-015-0564-6.

[9] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioASQ 2023: The eleventh bioASQ challenge on large-scale biomedical semantic indexing and question answering, Conference and Labs of the Evaluation Forum (2023) 227 – 250. doi:10.1007/978-3-031-42448-9_19.

[10] Q. Jin, R. Leaman, Z. Lu, Is ChatGPT a biomedical expert? – exploring the zero-shot performance of current GPT models in biomedical tasks, arXiv preprint arXiv:2306.16108 (2023). URL: https://arxiv.org/abs/2306.16108.

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

[12] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd Edition | Wiley, 2009.

[13] D. Isern, D. Sánchez, A. Moreno, Agents applied in health care: A review, Int J Med Inform 79 (2010) 145–166. doi:10.1016/j.ijmedinf.2010.01.003.

[14] J. Ferber, Multi-Agent Systems : An Introduction to Distributed Artificial Intelligence, Harlow : Addison-Wesley, 1998.

[15] A. Croatti, M. Gabellini, S. Montagna, A. Ricci, On the Integration of Agents and Digital Twins in Healthcare, J Med Syst 44 (2020) 161. doi:10.1007/s10916-020-01623-5.

[16] R. Stuart, N. Peter, Artificial Intelligence: A Modern Approach, Hoboken, NJ, 2021.

[17] H. Zhang, et al., Leveraging long context in retrieval augmented language models for medical question answering, npj Digital Medicine 8 (2025) 12. URL: https://pubmed.ncbi.nlm.nih.gov/40316710/. doi:10.1038/s41746-025-01651-w.

[18] A. Merker, et al., Mibi at BioASQ 2024: Retrieval-augmented generation for answering biomedical questions, in: CEUR Workshop Proceedings, volume 3740, 2024, p. paper 16. URL: https://ceur-ws.

org/Vol-3740/paper-16.pdf.

[19] M. Jeong, J. Sohn, M. Sung, J. Kang, Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, Bioinformatics 40 (2024) i119 – i129. doi:10.1093/bioinformatics/btae238.

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

[21] S. Ateia, U. Kruschwitz, Can open-source LLMs compete with commercial models? exploring the few-shot performance of current GPT models in biomedical tasks, in: CEUR Workshop Proceedings, volume 3740, 2024, p. paper 7. URL: https://ceur-ws.org/Vol-3740/paper-07.pdf.

[22] R. Bramley, et al., Notes on the data quality of bibliographic records from the MEDLINE database, Database: The Journal of Biological Databases and Curation 2023 (2023) baad070. URL: https://academic.oup.com/database/article/doi/10.1093/database/baad070/7231777. doi:10.1093/database/baad070.

[23] E. Sayers, A general introduction to the e-utilities, in: E. Sayers (Ed.), Entrez Programming Utilities Help [Internet], updated 2022 nov 17 ed., National Center for Biotechnology Information (US), 2009, pp. 1–15. URL: https://www.ncbi.nlm.nih.gov/books/NBK25497/, available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/.

[24] S. Vatsal, et al., A survey of prompt engineering methods in large language models for different NLP tasks, Department of Computer Science New York University (2024). URL: https://arxiv.org/html/2407.12994v1.

[25] G. Marvin, N. Hellen, D. Jjingo, J. Nakatumba-Nabende, Prompt engineering in large language models, in: I. J. Jacob, S. Piramuthu, P. Falkowski-Gilski (Eds.), Data Intelligence and Cognitive Informatics, Springer Nature Singapore, Singapore, 2024, pp. 387–402.

[26] X. Hu, Y. Chen, L. Zhang, R. Wang, Q. Liu, Task-specific prompt frameworks for biomedical text mining: A systematic review, Journal of Biomedical Informatics 141 (2024) 104494. URL: https://pubmed.ncbi.nlm.nih.gov/38281112/.

[27] B. Alshaikhdeeb, et al., Biomedical named entity recognition: A review, 2016. URL: https://www.researchgate.net/publication/311917426_Biomedical_Named_Entity_Recognition_A_Review.

[28] N. Perera, et al., Named entity recognition and relation detection for biomedical information extraction, 2020. URL: https://pubmed.ncbi.nlm.nih.gov/32984300/.

[29] D. Molla, et al., Named entity recognition for question answering, 2006. URL: https://aclanthology.org/U06-1009.pdf.

[30] M. Monajatipoor, LLMs in biomedical: A study on named entity recognition, arXiv preprint arXiv:2404.07376 (2024). URL: https://arxiv.org/abs/2404.07376, preprint.

[31] H. Park, J. Son, J. Min, J. Choi, Selective UMLS knowledge infusion for biomedical question answering, Scientific Reports 13 (2023). URL: https://www.researchgate.net/publication/373519004_Selective_UMLS_knowledge_infusion_for_biomedical_question_answering. doi:10.1038/s41598-023-41423-8.

[32] Y. Du, J. Yan, Y. Lu, Y. Zhao, X. Jin, Improving biomedical question answering by data augmentation and model weighting, IEEE/ACM Transactions on Computational Biology and Bioinformatics PP (2022) 1–1. doi:10.1109/TCBB.2022.3171388.

[33] 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 Computing Surveys 55 (2022). URL: http://dx.doi.org/10.1145/3490238. doi:10.1145/3490238.

[34] W. Zhao, Z. Deng, S. Yadav, P. S. Yu, Heterogeneous knowledge grounding for medical question answering with retrieval augmented large language model, in: Companion Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1590–1594. URL: https://doi.org/10.1145/3589335.3651941. doi:10.1145/3589335.3651941.

[35] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. J. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are

few-shot learners, Neural Information Processing Systems 33 (2020) 1877 – 1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[36] Y. Zhang, D. Merrill, K. Logan, G. Ananiadou, Optimizing biomedical information retrieval with a keyword frequency-driven prompt enhancement strategy, BMC Bioinformatics 25 (2024). URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05123-4.

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

[38] E. Jimenez-Ruiz, B. C. Grau, I. Horrocks, R. Berlanga, Logic-based assessment of the compatibility of UMLS ontology sources, Journal of Biomedical Semantics 2 (2011) S2. URL: https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-2-S1-S2. doi:10.1186/2041-1480-2-S1-S2.

[39] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, H. Xu, Improving large language models for clinical named entity recognition via prompt engineering, Journal of the American Medical Informatics Association 31 (2024) 1812–1820. doi:10.1093/jamia/ocad259. arXiv:2303.16416, published online January 27, 2024.