Overview of BioASQ Tasks 14b and Synergy14 in CLEF2026

Anastasios Nentidis[1], Georgios Katsimpras[1], Anastasia Krithara[1] and Georgios Paliouras[1]

[1]NCSR Demokritos, Athens, Greece

Abstract

This paper provides an overview of the Question Answering (QA) tasks in the fourteenth edition of the BioASQ challenge (BioASQ 14), on large-scale biomedical semantic indexing and QA, which is organised as part of the Conference and Labs of the Evaluation Forum (CLEF) 2026. Over more than a decade, BioASQ has acted as a central venue for advancing the state-of-the-art in semantic indexing, information extraction, information retrieval, and QA in biomedicine. This year, 48 teams participated in the biomedical QA tasks 14b, and Synergy14, developing more than 150 systems. In particular, 45 of them took part in task 14b, on biomedical semantic QA, and 7 on task Synergy14, on QA about developing biomedical topics where the questions may still miss a definite answer. The strong performance reached by several of the participating systems in the QA tasks of BioASQ 14 confirms the steady progress of state-of-the-art methods in the field, consistent with earlier editions of the tasks.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction

This paper gives an overview of the fourteenth BioASQ challenge (2026), with a particular focus on the shared tasks 14b and Synergy14. We outline the datasets that were used to train and evaluate the participating systems. Tasks 14b and Synergy14, which ran from March to May and from January to March 2026, respectively, are detailed in Section 2. Section 3 offers a short summary of the participation in these two tasks. A more detailed analysis of the approaches adopted by the participating systems will appear in the proceedings of the fourteenth BioASQ workshop [1]. The paper closes with a brief discussion of our main observations.

2. Overview of the Tasks

The fourteenth edition of the BioASQ challenge ws structured into six main tasks: (1) task 14b on biomedical QA, (2) task Synergy 14 on biomedical QA for open developing issues, (3) task MultiClinSum-2: the second edition of the task on multilingual summarization of clinical documents, (4) task BioNNE-R on extracting relations between nested named entities in English and Russian, (5) task ELCardioCC: the second edition of the task on automatic clinical coding in cardiology from Greek discharge letters, and (6) task GutBrainIE: the second edition of the task on extracting structured information from biomedical abstracts related to the gut-brain axis [2].

In this paper, we focus on the current versions of the first two well-established biomedical QA tasks in English, denoting them as task 14b and task Synergy14 in the context of the fourteenth BioASQ edition. Detailed descriptions of the MultiClinSum-2, BioNNE-R, ELCardioCC, and GutBrainIE tasks are available in [3], [4], [5], and [6], respectively. In addition, an extensive introduction to the BioASQ challenge and its original task design can be found in [7].

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

tasosnent@iit.demokritos.gr (A. Nentidis); gkatsibras@iit.demokritos.gr (G. Katsimpras); akrithara@iit.demokritos.gr (A. Krithara); paliourg@iit.demokritos.gr (G. Paliouras)

2.1. Biomedical semantic QA - Task 14b

Task 14b poses a complete question-answering challenge in the biomedical domain, asking participants to build systems that cover all stages of question answering. As in earlier editions, the task targets four question types: "yes/no", "factoid", "list", and "summary" questions [8].

For the fourteenth edition of the BioASQ Challenge, the participating teams were provided with an updated release of the BioASQ QA training dataset [9], comprising 5,729 questions annotated with relevant golden elements and answers carried over from previous versions of the task [10, 11, 12]. These questions were the main material used by the participants to develop their systems. For testing, 280 new biomedical questions were created and annotated with gold documents, snippets, and both exact and ideal answers. Table 1 gives a detailed view of both the training and the test sets for task 14b. A clear pattern in these statistics is that the average number of documents and snippets is considerably higher in the training data than in the test batches. Two main reasons account for this. First, a part of the training dataset consists of questions from the early years of BioASQ, where the experts annotated relevant documents and snippets exhaustively, aiming to retrieve as many relevant items as possible from the corpus. At present, only a sufficient set of relevant answers is required when preparing the first version of the data. Later, once the participants submit their responses, the experts review the submitted items and extend the ground-truth data with any additional relevant items that the participants identified. Second, the counts of relevant items for the test sets in Table 1 are preliminary, reported before the enrichment performed during the assessment process, which is still ongoing. The final evaluation of the participants will be carried out against the enriched relevant items, so that every submitted item that is relevant is treated as such.

Table 1
Statistics on the training and test datasets of Task 14b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	5,729	1,541	1,130	1,695	1,363	10.74	13.97
Test 1	80	17	21	23	19	5.25	12.41
Test 2	80	21	13	20	26	3.64	7.98
Test 3	60	11	17	17	15	3.40	6.95
Test 4	60	16	20	11	13	3.78	8.35
Total	6,009	1,606	1,201	1,766	1,436	10.43	13.75

Task 14b, like the previous version of the task (13b), was organised into three phases [13]. As in earlier versions, task 14b was split into four independent bi-weekly batches, and the three phases of each batch ran on two consecutive days [14]. The three phases of task 14b were: (phase A) retrieval of the required information, (phase A+) answering the question without golden material, and (phase B) answering the question with golden material. For each phase, participants had 24 hours to submit their systems' responses after receiving the corresponding test set. This year, the four test sets contained 80, 80, 60, and 60 questions, respectively. For every test set, the questions, written in English, and their type were first released for Phase A and Phase A+. In Phase A, the participants had to locate and submit relevant elements from designated resources, namely PubMed/MEDLINE articles and snippets extracted from those articles. The same questions were considered in phase A+, where the participating systems had to return *exact answers*, i.e. entity names or short phrases, and *ideal answers*, i.e. natural language summaries of the requested information. Finally, in phase B, manually selected relevant articles and snippets for these questions were also made available, and the participating systems were again asked to provide *exact answers* and *ideal answers*.

2.2. Synergy14 Task

First introduced in the ninth BioASQ edition [15], the Synergy task was designed to encourage collaboration between biomedical experts working on COVID-19 and the automated question-answering

systems participating in BioASQ. Its central goal is to establish a synergy in which experts evaluate the systems' responses and this feedback is then used to refine the systems responses in an iterative fashion [16]. The task remained focused on COVID-19 in the tenth BioASQ edition [17], and from the following editions onward it was broadened to cover more developing biomedical topics [18, 19, 20]. As depicted in Figure 1, the competing systems first provide their responses to open questions on emerging topics. These responses, together with relevant documents and snippets, are then assessed by the experts. The experts subsequently return feedback to the systems and address any new or pending questions.

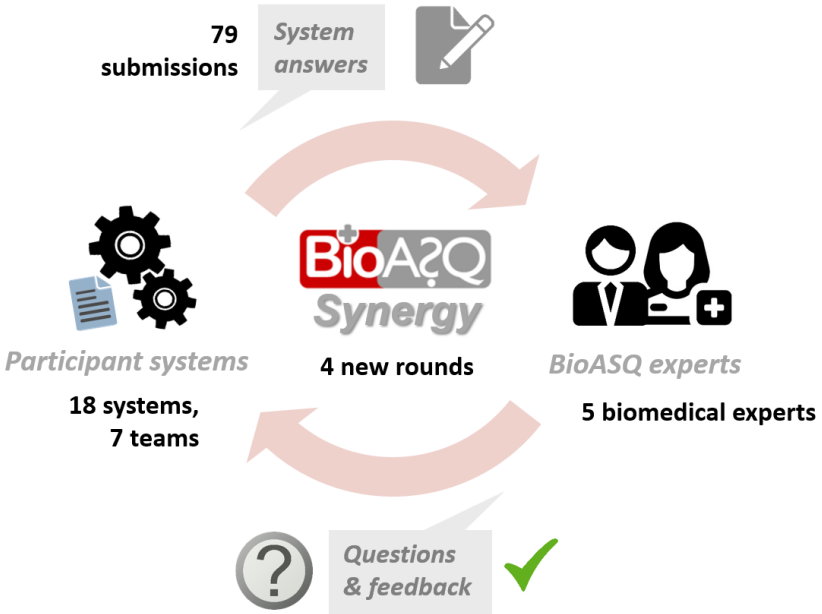

Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy task on question answering for open developing topics.

As in previous years, this edition of the Synergy task (Synergy14) was organised into four rounds, one every two weeks, and addressed emerging biomedical topics, drawing on relevant documents from the current PubMed version [14, 21, 13]. As in earlier versions, the questions were open-ended and their answers could be more volatile [22, 23].

In the Synergy task, the system responses and the expert feedback in each round concern the same questions, except for those already closed by the experts after receiving a complete and definitive answer. In Synergy14, the experts contributed 63 new questions on developing health topics, namely cortisol biology and Cushing syndrome, cancer treatment response, molecular and genetic mechanisms, infectious diseases and prophylaxis, diagnostic methods, and mental health. In the second round, 3 questions that had remained open from earlier editions were added to be enriched with more recent evidence and updated answers. Overall, 66 distinct questions were considered in this edition of the task. The experts assessed both the retrieved material (documents and snippets) and the responses submitted by the participating systems in all four rounds. Table 2 shows the details of the datasets used in task Synergy14 per round. Synergy14 addresses the same four question types as task 14b: yes/no, factoid, list, and summary, and the two answer types exact and ideal. Moreover, the evaluation of the systems relies on the same measures as task 14b.

3. Overview of participation

Across all six shared tasks of BioASQ 14, 87 distinct teams took part in the challenge. In the two QA tasks 14b and Synergy14 covered by this paper, more than 150 distinct systems participated, developed

Table 2
Statistics on the datasets of Task Synergy14. "Answer ready" stands for questions marked as having enough relevant material to be answered after the assessment of material submitted by the systems in the respective round.

Round	Size	Yes/No	List	Factoid	Summary	Answer ready
1	63	15	15	13	20	0
2	66	15	17	14	20	42
3	64	14	17	13	20	55
4	48	10	13	11	14	47

by a total of 48 teams, 36 of which were new to BioASQ. The large fraction of teams joining the BioASQ challenge for the first time, around 75%, reflects the sustained interest of the community in large-scale biomedical question answering. Specifically, 45 of these teams submitted to at least one phase of task 14b and 7 to task Synergy14, with 4 teams taking part in both tasks, as shown in Figure 2.

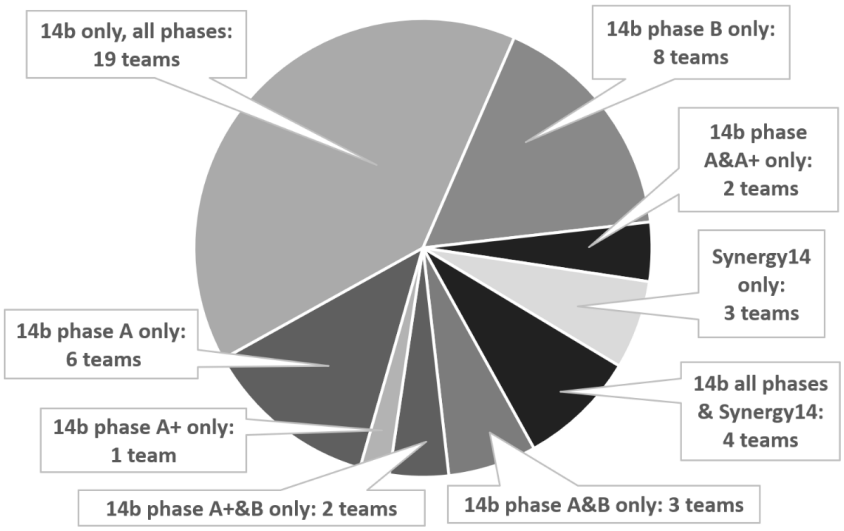

Figure 2: The distribution of participant teams in BioASQ task 14b phases and task Synergy14.

In line with previous years [23, 14], task b attracted more participants than Synergy, while the overall number of participating teams remained high this year, as illustrated in Figure 3.

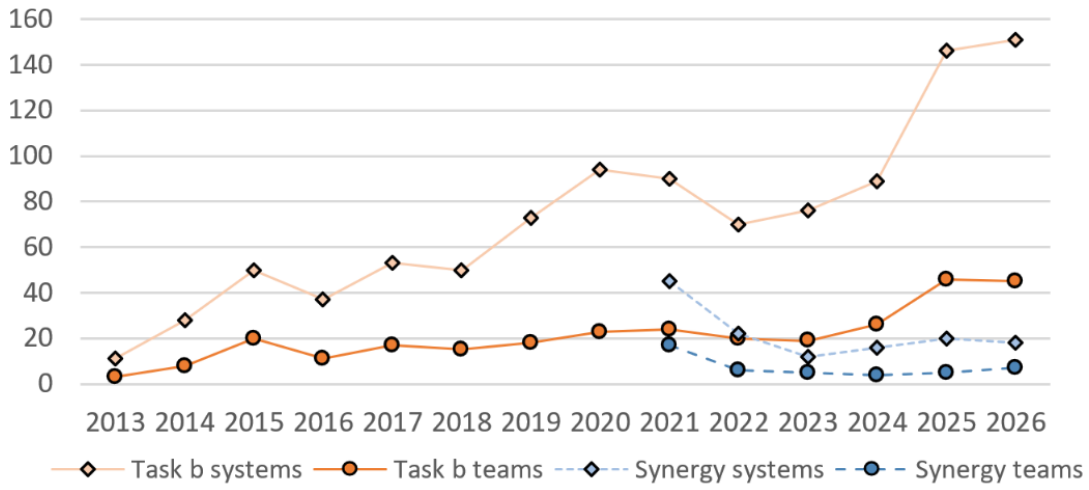

Figure 3: Participation evolution in task b and Synergy during the fourteen years of BioASQ.

3.1. Task 14b

This year, 45 teams took part in task 14b, producing 821 submissions from 151 distinct systems across the four batches and the three phases A, A+, and B. Specifically, 34, 28, and 36 teams competed in phases A, A+, and B, with 100, 81, and 112 distinct systems and 261, 241, and 319 submissions, respectively. Twenty-three of these teams were active in all three phases, with four of them participating in Synergy as well, as depicted in Figure 2.

An overview of the technologies used by the teams is presented in Table 3. As in previous years, the open-source system OAQA [24], which reached top performance in earlier editions of BioASQ [25, 26], served as a baseline for phase B *exact answers*. This system is built on the UIMA framework and relies on traditional NLP and Machine Learning approaches and tools, such as MetaMap and LingPipe[1].

Table 3
Systems and approaches for task 14b. Systems for which no information was available at the time of writing are omitted.

Systems	Phase	Ref.	Approach
IR_Y/IR_J	A,A+,B	[27]	BM25, RAG, reranking, query expansion, GPT, DeepSeek, BioMistral
Fleming	A,A+,B	[28]	BM25, KaLM, reranking, RRF, Gemma2-LW, MxBai
UR-IW	A,A+,B	[29]	OpenSearch, reranking, RRF, prompting
IR	A,A+,B	[30]	BM25, reranking, GPT, Gemini, chain-of-thought prompting
DSGT	A,A+,B	[31]	query expansion, reranking, RRF, OpenBioLLM, prompting
multistage	A,A+,B	[32]	BM25, splade, reranking, RRF, MedCPT, MxBai, MedGemma, Llama, Claude
dictycite	A,A+,B	[33]	BM25, RM3, RRF, reranking, llama
BioXR	A,A+,B	[34]	BM25, reranking, BioXR-GTE, prompting, ensemble
lean_rag	A,A+,B	[35]	BM25, RRF, reranking, prompting
MedQA	A,A+,B	[36]	BM25, BGE, query expansion, RRF, Claude, Gemini, GPT
RMC	A,A+,B	[37]	RAG, Nomic, BERT, llama
bioinfo	A,A+,B	[38]	BM25, Qdrant, HyDE, Context-1 agentic search, reranking, multi-agent quorum mechanism, Gemma
SATO	A,A+,B	-	BGE-M3, Qdrant, Distribution-Based Score Fusion, reranking, MedGemma
NUST	A,A+,B	-	BM25, MiniLM, llama, prompting
interactos.ai	A,A+,B	-	Qwen, LatentMAS, Agentic
health-nlp	A,B	[39]	MedCPT, reranking, Parametric-RAG, Qwen
MIT Lincoln Lab	A+,B	[40]	PubMedBERT, BM25, RRF, Gemma
University of Macedonia	A	[41]	BM25, BGE, reranking
University of Amsterdam	A	[42]	splade, DeBERTa, reranking
ASK	A+	[43]	multi-agent architecture, prompting, Claude, GPT
DMIS	B	-	prompting, chain-of-thought, multi-agent architecture
CSA-IISR	B	[44]	prompting, multi-model comparative setup

The IR_Y/IR_J systems from the Zhongyuan University of Technology utilized a RAG pipeline structured into four stages. They performed MeSH-based knowledge expansion using SciSpacy to extract biomedical entities and synonyms followed by a multi-stage PubMed retrieval with two-phase re-ranking process that combined BM25 with heuristic features. Their systems employed three LLM backends: GPT-4o-mini, DeepSeek, and a fine-tuned BioMistral model.

The Fleming systems from the BSRC Alexander Fleming, employed a multi-stage hybrid retrieval and

[1] http://alias-i.com/lingpipe

re-ranking pipeline. Retrieval combined KaLM embeddings for dense semantic search and BM25 for sparse retrieval, merged via Reciprocal Rank Fusion (RRF). For re-ranking their systems used Gemma2-LW and MxBai-Large models. Also, their systems experimented with sliding-window semantic similarity and ensembling of open-source large language models.

The UR-IW systems from Universität Regensburg utilized a pipeline based on LLMs. For Phase A, it employed OpenSearch retrieval over a PubMed-style index, incorporating three complementary learned query strategies fused via RRF in later batches. For Phase A+ and Phase B the systems generated various answer types (yes/no, factoid, list, and ideal) by applying task-specific few-shot prompts.

The IR systems from National Central University Taiwan employed a RAG framework for Phase A/A+ and a specialized generation pipeline for Phase B. Phase A/A+ relied on hybrid retrieval which combined BM25, dense retrieval and reranking. Phase B focused on prompt-based generation using GPT and Gemini models (including GPT-5.4, GPT-5.5, and Gemini 3.1 Pro), applying both direct inference and Chain-of-Thought strategies. Their systems also employed ensemble mechanisms and multi-model voting.

The DSGT systems from Georgia Institute of Technology integrated query expansion, PubMed retrieval, neural semantic reranking, and RRF. Their systems also followed specialized strategies such as ontology-guided retrieval refinement, knowledge-graph-inspired expansion, and multi-step planner–retriever–critic workflows. For answer generation, the system utilized OpenBioLLM together with question-type-specific prompting.

The multistage $v1$ systems from Università degli Studi di Pavia employed BM25 for initial retrieval with SPLADE-v3 for reranking, while the $v2$ systems utilized a complex ensemble: BM25 initial retrieval, followed by MedCPT, mxbai-rerank-large-v2, and MedGemma-27B-IT listwise reranking. These multi-stage systems incorporated pseudo relevance feedback using S-PubMedBERT-MS-MARCO. For answer generation, their systems relied on Claude Llama-70B-IT.

The dictycite systems from the University of Ljubljana utilized a multi-staged architecture which combined BM25+RM3 retrieval with dense retrieval, merging results via RRF. For reranking their system relied on cross-encoder models. Then, the selected documents served as input to a Llama-3.3 model to perform final answer generation.

The BioXR systems from the University of Management and Technology Lahore used BM25 followed by a two-model cross-encoder reranking strategy (BioXR-BGE and BioXR-GTE) for Phase A. For Phase A+ and Phase B, their systems employed a few-shot LLM approach with type-specific strategies including multi-model ensembles and also incorporated adversarial consistency checks and key-term-guided reranking.

The MedQA systems from University of Technology Sydney utilized a hybrid retrieval (dense BGE and BM25 fused via RRF) and an agent-driven pipeline that expands queries across PubMed, Europe PMC, and iCite. Then, a BGE cross-encoder is used for passage scoring and re-retrieval is performed based on cost-pragmatic policies. For Phase B, their systems used Claude, cross-model ensemble (Claude, Gemini, and GPT-5.5), and LLM-as-judge strategies (including aggregation-based fusion methods).

The lean_rag systems from Universidade de Lisboa combined BM25 retrieval with type-specific query expansion and dense retrieval, also followed by RRF. For generation, the systems used a instruction-tuned LLM with a few-shot prompting strategy tailored to question types.

The RMC systems from Virginia Commonwealth University implemented a RAG pipeline utilizing Nomic-Embed-Text v1.5 embeddings for both full-abstract and sentence-level vectors. For answer generation, their systems employed a Llama2-based model (Synthia-13B-GPTQ), applying various prompting and answer processing techniques.

The bioninfo systems from the University of Aveiro used a hybrid retrieval pipeline integrating BM25 and Qdrant. This process was enhanced by HyDE-based query expansion, Context-1 agentic search strategy and retrieval negative sampling. For answer generation, the systems employed a multi-agent and LLM-as-a-judge framework, where multiple LLMs converged on a consensus.

The SATO systems from the Technical University of Kosice used a hybrid dense–sparse retrieval pipeline incorporating BGE-M3, Qdrant, and Distribution-Based Score Fusion, followed by a cross-encoder reranker. For answer generation, their systems relied on the MedGemma 27B language model.

The NUST team utilized a RAG pipeline. Their systems employed a hybrid BM25 and dense retrieval using a fine-tuned BAAI/bge-base-en-v1.5 bi-encoder, with results being fused and reranked by a fine-tuned MiniLM-L-12-v2 cross-encoder. For answer generation, their systems utilized LLaMA-3.2-3B-Instruct, incorporating various prompting strategies tailored to question types.

The interactos.ai systems utilized an architecture with five specialist agents (EntityResolver, EvidenceHunter, MechanismExplorer, ReasoningOracle, and AnswerSynthesizer) that coexist within a single Qwen model backbone. These agents are orchestrated by a LatentMAS pipeline, which routes each BioASQ question type to a specialized semantic sub-graph for optimized processing.

The health-nlp systems from the University of Tübingen employed distinct strategies for each phase. For Phase A, their systems used AutoBool followed by reranking with MedCPT. For Phase B, the team relied on Parametric Retrieval-Augmented Generation (PRAG) using the Qwen2.5-1.5B model. Their systems explored standard in-context learning, the original PRAG approach, and a modified PRAG variant incorporating scaled adapter merging.

The MIT Lincoln Lab systems utilized a hybrid approach that combined PubMedBERT and BM25 reranking, with final candidate lists fused via RRF. For answer generation the team experimented with two primary configurations varying the model backbone (Gemma 3 27B vs. Gemma 4 31B Dense).

The University of Macedonia systems utilized a two-stage retrieval pipeline. For document retrieval, their systems combined BM25 with a BGE cross-encoder for reranking. For snippet extraction, their systems also ranked results using a combination of BM25 scores and a heuristic scoring function.

For Phase A, the University of Amsterdam systems utilized SPLADE-v3 as the primary retriever and a DeBERTa-based model as the reranker. The team optimized the reranker using contrastive learning on the BioASQ training set, testing two fine-tuning strategies: one using all historical instances (pre-2025) and one using a recent subset (2020–2024).

The ASK systems from Janssen (Johnson & Johnson) utilized the Med.ai ASK framework across three configurations: a baseline LLM approach (Claude Sonnet 4.6), a ReAct-style agent that iteratively invoked tools (GPT-4.1, GPT-5.4, Claude Sonnet 4.6), and a hierarchical multi-agent system. This latter approach featured a supervisor that dynamically decomposed questions and dispatched specialized sub-agents equipped with domain-specific information and internal reasoning loops to generate answers.

The DMIS systems from Korea University employed a question-type-specific LLM framework, incorporating Chain-of-Thought, in-context learning, and a multi-agent architecture (extraction, generation, verification, and aggregation).

The CSA-IISR systems from ChingShin High School/Academy utilized a feedback-driven automated prompt optimization architecture across multiple LLMs.

3.2. Synergy14 Task

In the fourteenth edition of BioASQ, seven teams took part in the Synergy task (Synergy14). These teams submitted 79 runs from 18 distinct systems. Four of these teams also participated in task 14b, while the remaining three focused exclusively on task Synergy14. An overview of the systems and approaches used by the Synergy14 teams for which system information was available is given in Table 4.

Table 4
Systems and their approaches for task Synergy. Systems for which no description was available at the time of writing are omitted.

System	Ref.	Approach
Fleming	[28]	BM25, KaLM, reranking, RRF, Gemma2-LW, MxBai
UR-IW	[29]	OpenSearch, reranking, RRF, prompting
RMC	[37]	RAG, Nomic, BERT, llama
SynapFlow	[45]	BM25, reranking, MedCPT, GPT
AgenticRAG	[46]	BM25, reranking, RAG

The BSRC Alexander Fleming systems used a similar architecture to those employed in task b.

Likewise, UR-IW systems from Universität Regensburg applied the same prompting approach as in task b, combining snippets with task-specific LLM prompts to generate answers. Also, the RMC team from Virginia Commonwealth University relied on their task b system design.

The SynapFlow systems from CMU utilized query normalization and named entity extraction, followed by a hybrid retriever followed by cross-encoder reranking and Maximal Marginal Relevance (MMR) for diversity. For answer generation, their systems employed GPT-4o-mini.

The AgenticRAG systems from DePaul University employed BM25 with phrase boosting and keyword matching for retrieval and a cross-encoder for reranking. A second LLM-based reranking phase is also performed to further refine the final results.

4. Results

4.1. Task 14b

This section presents the evaluation measures and the preliminary results for task 14b. These results are preliminary, since the final results will become available only after the manual assessment of the system responses by the BioASQ team of experts and the enrichment of the ground truth with potential additional relevant items, answer elements, and/or synonyms, which is still ongoing.

Phase A: The Mean Average Precision (MAP) was used to evaluate document retrieval. In particular, since BioASQ8 [47], the MAP computation uses a modified version of Average Precision (AP) that accounts both for the limit of 10 items allowed per question in each submission and for the actual number of golden items, which in practice is often smaller than 10 [48]. For snippets, where a single ground-truth snippet may overlap with several distinct submitted snippets, the interpretation of MAP is less straightforward. Therefore, since BioASQ9 [15], we adopt an F-measure based on character overlaps[2] [49]. Tables 5 and 6 report some indicative results in batch 1 for document and snippet retrieval, respectively. The full 14b results for phase A are available online[3].

Phases A+ and B: The official ranking of systems providing *ideal answers* is based on manual scores assigned by the BioASQ team of experts, who assess each *ideal answer* in the responses [8]. The final position of systems providing *exact answers* is determined by their average ranking across the three question types for which *exact answers* are required, namely "yes/no", "list", and "factoid". Summary questions, for which no *exact answers* are submitted, are excluded from this ranking. Specifically, the mean F1 measure is used for ranking list questions, the Mean Reciprocal Rank (MRR) for factoid questions, and the F1 measure, macro-averaged over the yes and no classes, for yes/no questions. Tables 7 and 8 present some indicative results on *exact answer* extraction. The full 14b results for both phase A+[4] and B[5] are available online.

The top performance of the participating systems in *exact answer* generation for each question type over the fourteen years of BioASQ is presented in Figure 4. The preliminary results for task 14b reveal that the participating systems continue to achieve relatively high scores in answering all question types. In batch 2, phase B, for example, presented in Table 8, a system manages to correctly answer all yes/no questions. In phase A+, on the other hand, where no ground-truth relevant material is available, correctly answering all yes/no questions is more demanding, although the top systems still reach very high yes/no scores, as shown in Table 7 for the same batch. The preliminary results of task 14b, phase B, indicate competitive performance for list and factoid questions, which, however, remain challenging, with performance fluctuating more across batches. On average, the top scores for list and factoid questions in the four batches of 14b (0.52 ± 0.074 MRR and 0.50 ± 0.11 F-measure) were lower than the corresponding preliminary results reported for BioASQ 13 (0.61 ± 0.07 MRR and 0.62 ± 0.02 F-measure) [13]. This may be related to the high proportion of questions contributed by new experts in BioASQ 14 (64%), who did not contribute to the training dataset. In fact, this preliminary performance

[2] http://participants-area.bioasq.org/Tasks/b/eval_meas_2022/
[3] https://participants-area.bioasq.org/results/14b/phaseA/
[4] https://participants-area.bioasq.org/results/14b/phaseAplus/
[5] https://participants-area.bioasq.org/results/14b/phaseB/

Table 5
Preliminary results for document retrieval in batch 1, phase A, task 14b. Only the top-50 systems are presented, based on MAP.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
dictycite-max-rew-sl	0.1188	**0.4090**	**0.1686**	**0.2835**	0.0048
dictycite-max	0.1138	0.3808	0.1594	0.2830	**0.0053**
dictycite-max-rew-si	0.1138	0.3934	0.1613	0.2763	0.0048
dictycite-snippet	0.1050	0.3636	0.1500	0.2577	0.0028
bioinfo-3	0.1050	0.3539	0.1488	0.2479	0.0029
Fleming-5	0.1063	0.3518	0.1490	0.2462	0.0045
Fleming-1	0.1063	0.3542	0.1490	0.2432	0.0047
Fleming-2	0.1063	0.3542	0.1490	0.2432	0.0047
Fleming-4	0.1050	0.3485	0.1476	0.2428	0.0043
Fleming-3	0.1050	0.3485	0.1476	0.2428	0.0043
bioinfo-4	0.1125	0.3698	0.1581	0.2398	0.0032
IR1	0.0950	0.3300	0.1358	0.2349	0.0027
dictycite-baseline	0.0938	0.3343	0.1349	0.2333	0.0023
bioinfo-2	0.1088	0.3576	0.1524	0.2220	0.0035
Biomedical QA system	0.0867	0.3152	0.1235	0.2176	0.0032
pancras_naive	0.1169	0.3000	0.1524	0.2078	0.0020
dmiip2024_1	0.0888	0.2851	0.1214	0.2023	0.0025
dmiip2024	0.0888	0.2864	0.1217	0.1986	0.0026
aida_2025_2	0.0776	0.2792	0.1119	0.1949	0.0011
dmiip2024_3	0.0938	0.2788	0.1276	0.1902	0.0039
dmiip2024_4	0.0900	0.2703	0.1225	0.1887	0.0038
dmiip2024_2	0.0900	0.2703	0.1225	0.1884	0.0038
pancras_crag	0.0934	0.2703	0.1248	0.1798	0.0014
porties-baseline	0.0700	0.2575	0.1002	0.1743	0.0004
health-nlp-1	0.0892	0.2769	0.1163	0.1696	0.0017
health-nlp-2	0.0892	0.2769	0.1163	0.1696	0.0017
lean_rag	0.0800	0.2483	0.1112	0.1691	0.0011
lean_rag_ft	0.0825	0.2620	0.1163	0.1650	0.0012
lean_rag_ft_sparse	0.0805	0.2547	0.1131	0.1630	0.0010
bioinfo-1	0.0863	0.2949	0.1212	0.1502	0.0015
semantic	0.0663	0.2335	0.0952	0.1497	0.0007
prompted	0.0663	0.2335	0.0952	0.1497	0.0007
main	0.0663	0.2335	0.0952	0.1497	0.0007
bioinfo-0	0.0850	0.2937	0.1200	0.1476	0.0014
UR-IW-1	0.0765	0.1942	0.1012	0.1473	0.0005
Biomedical QA s3	0.0732	0.2047	0.0913	0.1398	0.0003
UR-IW-2	0.1046	0.1966	0.1168	0.1340	0.0004
UR-IW-3	0.0938	0.1915	0.1029	0.1312	0.0005
UR-IW-5	0.0756	0.1658	0.0852	0.1263	0.0004
UR-IW-4	0.0610	0.1747	0.0828	0.1226	0.0004
IR_Y-4	0.0453	0.1685	0.0671	0.1077	0.0003
IR_Y-5	0.0453	0.1685	0.0671	0.1077	0.0003
IR_Y-3	0.0453	0.1685	0.0671	0.1077	0.0003
IR_Y-1	0.0453	0.1685	0.0671	0.1077	0.0003
IR_Y-2	0.0453	0.1685	0.0671	0.1077	0.0003
RMC_1	0.0484	0.1704	0.0690	0.1042	0.0003
RMC_2	0.1035	0.1270	0.1016	0.0998	0.0001
IR_J-2	0.0875	0.1292	0.0870	0.0954	0.0002
Organization name	0.0726	0.1525	0.0759	0.0902	0.0001
IR_J-1	**0.1286**	0.1293	0.0991	0.0874	0.0002
...	...	...	...	...	...

Table 6
Preliminary results for snippet retrieval, batch 1, phase A, task 14b, ranked based on Mean F-measure.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
dmiip2024_3	0.0824	0.2188	**0.0914**	**0.1671**	0.0017
IR1	0.0689	**0.2418**	0.0877	0.1656	0.0010
dmiip2024_2	0.0767	0.2094	0.0852	0.1591	0.0014
dmiip2024_4	0.0766	0.2099	0.0851	0.1601	0.0015
UR-IW-2	0.0930	0.1416	0.0829	0.1074	0.0004
pancras_naive	0.0759	0.2000	0.0825	0.1402	0.0008
UR-IW-3	0.0761	0.1385	0.0791	0.1209	0.0003
dmiip2024_1	0.0704	0.2213	0.0791	0.1646	0.0013
Fleming-1	0.0652	0.2098	0.0769	0.1109	0.0015
Fleming-5	0.0639	0.2083	0.0759	0.1085	0.0012
Fleming-4	0.0621	0.2215	0.0757	0.1168	0.0020
Fleming-2	0.0640	0.2200	0.0757	0.1198	**0.0021**
dmiip2024	0.0682	0.2041	0.0755	0.1574	0.0011
Fleming-3	0.0608	0.2057	0.0735	0.1048	0.0011
UR-IW-1	0.0617	0.1326	0.0717	0.1193	0.0003
Biomedical QA system	0.0598	0.1790	0.0710	0.1475	0.0009
UR-IW-5	0.0720	0.1295	0.0702	0.1234	0.0003
lean_rag	0.0518	0.1369	0.0606	0.0403	0.0002
lean_rag_ft	0.0505	0.1537	0.0601	0.0403	0.0002
lean_rag_ft_sparse	0.0504	0.1537	0.0601	0.0404	0.0002
IR_J-1	**0.0953**	0.0836	0.0561	0.0565	0.0001
UR-IW-4	0.0481	0.1211	0.0554	0.0997	0.0003
aida_2025_2	0.0462	0.1541	0.0543	0.0769	0.0002
health-nlp-1	0.0595	0.0858	0.0542	0.0542	0.0002
health-nlp-2	0.0595	0.0858	0.0542	0.0542	0.0002
IR_J-2	0.0622	0.0982	0.0522	0.0633	0.0001
pancras_crag	0.0479	0.1094	0.0517	0.0619	0.0002
RMC_2	0.0360	0.1036	0.0436	0.1113	0.0001
Biomedical QA s3	0.0387	0.1200	0.0435	0.1078	0.0002
Organization name	0.0400	0.1033	0.0425	0.1089	0.0001
porties-baseline	0.0364	0.0871	0.0414	0.0609	0.0001
IR_Y-5	0.0161	0.1364	0.0254	0.0255	0.0001
RMC_1	0.0233	0.0764	0.0254	0.0357	0.0001
IR_Y-4	0.0161	0.1364	0.0254	0.0153	0.0001
IR_Y-1	0.0161	0.1364	0.0254	0.0255	0.0001
IR_Y-2	0.0161	0.1364	0.0254	0.0153	0.0001
IR_Y-3	0.0161	0.1364	0.0254	0.0255	0.0001
Biomedical QA s. v2	0.0179	0.0245	0.0170	0.0152	0.0000
ubuntu	0.0155	0.0290	0.0160	0.0179	0.0000
semantic	0.0072	0.0094	0.0069	0.0128	0.0000
prompted	0.0033	0.0186	0.0045	0.0064	0.0000
main	0.0034	0.0056	0.0040	0.0048	0.0000

is still higher than the respective preliminary performance in the collaborative batch of task 10b (0.33 factoid MRR, and 0.45 list F-measure), which consisted exclusively of questions by new experts in BioASQ 10 and which is depicted with black marks in Figure 4 [50, 51]. In phase A, the average top scores for document and snippet retrieval in BioASQ 14 (0.14 ± 0.013 MAP and 0.09 ± 0.008 F1) are also lower than the respective preliminary results of BioASQ 13 (0.19 ± 0.078 MAP and 0.12 ± 0.036 F1) [13].

Table 7
Results for batch 2 for *exact answers* in phase A+ of task 14b. Only the top-50 systems are presented, based on Yes/No macro-F1.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
mckpt2	**0.9481**	**0.9524**	0.2000	0.2000	0.2000	0.1605	0.1496	0.1537
UR-IW-5	**0.9481**	**0.9524**	0.1500	0.1500	0.1500	0.1833	0.2687	0.2028
MedQA-1	0.9443	**0.9524**	**0.3500**	0.4500	**0.4000**	0.3533	0.3777	0.3551
dictycite-max-rew-si	0.9443	**0.9524**	0.2500	0.3000	0.2750	0.3111	0.3777	0.3251
dictycite-max-rew-sl	0.9443	**0.9524**	0.2500	0.2500	0.2500	0.3671	0.3777	0.3629
MedQA-2	0.9443	**0.9524**	0.3000	**0.5000**	0.3850	0.3629	0.3777	0.3617
UR-IW-3	0.9443	**0.9524**	**0.3500**	0.3500	0.3500	0.2862	0.3361	0.2939
ASK-Skill1	0.9443	**0.9524**	0.3000	0.3000	0.3000	0.1692	0.2527	0.1875
dictycite-max	0.9443	**0.9524**	0.2500	0.2500	0.2500	0.3324	0.3905	0.3504
bioinfo-3	0.8929	0.9048	0.2000	0.3500	0.2450	0.2926	0.3841	0.3145
dictycite-baseline	0.8929	0.9048	0.3000	0.3000	0.3000	0.3821	0.3905	0.3762
Biomedical QA system	0.8929	0.9048	0.1000	0.2500	0.1625	0.2531	0.2725	0.2596
dmiip2024	0.8929	0.9048	**0.3500**	0.4000	0.3750	0.3231	0.2207	0.2371
Biomedical QA s3	0.8929	0.9048	0.1500	0.1500	0.1500	0.2299	0.2490	0.2366
bioinfo-1	0.8929	0.9048	0.2000	0.2500	0.2250	0.2391	0.2784	0.2479
Biomedical QA s. v2	0.8929	0.9048	0.0500	0.2500	0.1350	0.2568	0.2918	0.2693
UR-IW-4	0.8929	0.9048	0.3000	0.3000	0.3000	0.3406	0.2912	0.3017
ubuntu	0.8833	0.9048	0.1000	0.1500	0.1250	0.2429	0.2132	0.2168
UR-IW-1	0.8833	0.9048	**0.3500**	0.3500	0.3500	0.2322	**0.4558**	0.2883
UR-IW-2	0.8833	0.9048	**0.3500**	0.3500	0.3500	0.2608	0.3425	0.2722
ASK-Tool2	0.8833	0.9048	0.3000	0.4000	0.3417	0.1119	0.2890	0.1461
IR_J-1	0.8833	0.9048	0.2500	0.3000	0.2750	0.2314	0.1839	0.1822
ASK-Skill2	0.8833	0.9048	0.3000	0.3000	0.3000	0.1743	0.2987	0.1976
dmiip2024_4	0.8833	0.9048	**0.3500**	0.4000	0.3750	**0.4359**	0.3686	**0.3909**
Fleming-1	0.8833	0.9048	0.1500	0.3000	0.1933	0.3367	0.4141	0.3544
IR_J-3	0.8518	0.8571	0.1500	0.1500	0.1500	0.1722	0.1346	0.1233
IR_J-4	0.8518	0.8571	0.1500	0.2500	0.1917	0.2379	0.2447	0.2294
bioinfo-2	0.8444	0.8571	0.2000	0.2000	0.2000	0.2406	0.2249	0.2306
bioinfo-4	0.8444	0.8571	0.2500	0.3000	0.2750	0.2948	0.4034	0.3227
dmiip2024_3	0.8444	0.8571	0.3000	0.3000	0.3000	0.2974	0.1994	0.2294
bioinfo-0	0.8444	0.8571	0.2000	0.2000	0.2000	0.2401	0.1608	0.1778
dmiip2024_2	0.8444	0.8571	0.3000	0.3500	0.3167	0.3538	0.3521	0.3276
config-2	0.8444	0.8571	0.2500	0.3000	0.2667	0.2350	0.3885	0.2785
RMC_5	0.8444	0.8571	0.1000	0.1000	0.1000	0.3051	0.2490	0.2610
dictycite-snippet	0.8329	0.8571	0.2500	0.3000	0.2750	0.3833	0.3799	0.3584
lean_rag_ft_sparse	0.8329	0.8571	0.2500	0.2500	0.2500	0.3635	0.2639	0.2766
lean_rag_ft	0.8329	0.8571	0.3000	0.3000	0.3000	0.3661	0.2639	0.2772
lean_rag	0.8329	0.8571	0.2500	0.3000	0.2667	0.3269	0.2581	0.2721
MedQA-4	0.8329	0.8571	**0.3500**	0.4500	**0.4000**	0.3160	0.3521	0.3219
IR_Y-4	0.8329	0.8571	0.3000	0.3500	0.3250	0.0641	0.0662	0.0543
IR_Y-3	0.8329	0.8571	0.2500	0.3500	0.3000	0.0635	0.0662	0.0543
IR_Y-2	0.8329	0.8571	0.0000	0.3000	0.0775	0.0205	0.0449	0.0272
mckpt1	0.8152	0.8571	0.2500	0.3000	0.2667	0.3049	0.2993	0.2939
IR1	0.7981	0.8095	0.3000	0.3000	0.3000	0.4177	0.3120	0.3280
IR_J-2	0.7981	0.8095	0.2500	0.2500	0.2500	0.2277	0.1709	0.1776
ASK-Tool1	0.7981	0.8095	0.3000	0.3000	0.3000	0.1577	0.3955	0.2061
RMC_2	0.7981	0.8095	0.1000	0.1000	0.1000	0.1436	0.1064	0.1170
MedQA-3	0.7857	0.8095	0.2000	0.3000	0.2375	0.3002	0.3425	0.3082
dmiip2024_1	0.7857	0.8095	0.3000	0.3000	0.3000	0.4141	0.3072	0.3408
ASK-Baseline	0.7857	0.8095	0.3000	0.3500	0.3250	0.1654	0.4030	0.2026
...	...	...	...	...	...	...	...	...

Table 8
Results for batch 2 for *exact answers* in phase B of task 14b. Only the top-50 systems are presented based on Yes/No macro-F1.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
dictycite-baseline	**1.0000**	**1.0000**	0.3000	0.3000	0.3000	0.4465	0.4280	0.4301
Another	0.9443	0.9524	0.3000	0.3000	0.3000	0.4636	0.4640	0.4583
Biomedical QA s. v2	0.9443	0.9524	0.2500	0.2500	0.2500	0.1037	0.4775	0.1630
UR-IW-4	0.9443	0.9524	0.3500	0.3500	0.3500	0.3626	0.3987	0.3673
MedQA-1	0.9443	0.9524	0.3500	0.4500	**0.4000**	0.4492	0.4500	0.4433
dmiip2024_1	0.9443	0.9524	0.3000	0.3500	0.3250	0.3731	0.3152	0.3279
EP-2	0.9443	0.9524	0.3500	0.3500	0.3500	0.4603	0.3713	0.4002
EP-3	0.9443	0.9524	0.3500	0.3500	0.3500	0.4603	0.3713	0.4002
dictycite-max-rew-sl	0.9443	0.9524	0.2500	0.4000	0.3167	0.4499	0.4912	0.4620
EP-1	0.9443	0.9524	0.3500	0.3500	0.3500	0.4744	0.3521	0.3872
ku_dmis	0.9443	0.9524	0.2500	**0.5000**	0.3475	0.4672	0.4474	0.4538
EP-4	0.9443	0.9524	0.3000	0.3500	0.3250	0.4667	0.3521	0.3846
EP-5	0.9443	0.9524	0.3500	0.4000	0.3750	0.4846	0.4013	0.4188
UR-IW-5	0.8990	0.9048	0.2000	0.2500	0.2250	0.2707	0.4622	0.3183
bm25 + splade	0.8990	0.9048	0.3000	0.3000	0.3000	0.4309	0.4405	0.4268
IR_J-5	0.8990	0.9048	0.3000	0.3500	0.3250	0.4244	0.4517	0.4342
mckpt1	0.8990	0.9048	0.3500	0.3500	0.3500	0.2564	0.2909	0.2621
13b-1	0.8990	0.9048	0.3000	0.3000	0.3000	0.4250	0.3404	0.3630
13b_phase_a	0.8990	0.9048	0.3000	0.3500	0.3250	0.4250	0.3404	0.3630
RMC_2	0.8990	0.9048	0.2500	0.2500	0.2500	0.3692	0.3516	0.3438
Biomedical QA s.4	0.8929	0.9048	0.2000	0.2000	0.2000	0.4133	0.5036	0.4437
Biomedical QA s3	0.8929	0.9048	0.2000	0.2000	0.2000	0.4135	0.5228	0.4502
IR_Y-5	0.8929	0.9048	0.1500	0.3500	0.2417	0.1629	0.2095	0.1773
Bio26NIA	0.8929	0.9048	0.3500	0.3500	0.3500	0.4830	0.4544	0.4605
dmiip2024_2	0.8929	0.9048	0.3500	0.3500	0.3500	0.4615	0.4590	0.4259
ku_dmis_4	0.8929	0.9048	0.2500	**0.5000**	0.3475	0.4660	0.4672	0.4628
pancras_crag	0.8929	0.9048	0.2000	0.3500	0.2600	0.4464	0.4983	0.4586
Biomedical QA system	0.8929	0.9048	0.2000	0.2000	0.2000	0.4030	0.4993	0.4354
IR_Y-4	0.8929	0.9048	0.2500	0.3500	0.3000	0.1513	0.1839	0.1567
ku_dmis_5	0.8929	0.9048	0.2500	**0.5000**	0.3475	0.4660	0.4672	0.4628
pancras_naive	0.8929	0.9048	0.2000	0.3500	0.2600	0.4464	0.4983	0.4586
ku_dmis_3	0.8929	0.9048	0.2500	**0.5000**	0.3475	0.4444	0.3885	0.4068
ku_dmis_2	0.8929	0.9048	0.2500	**0.5000**	0.3475	0.4434	0.2918	0.3296
IR_Y-2	0.8929	0.9048	0.1500	0.3500	0.2200	0.1857	0.2453	0.1986
IR_Y-3	0.8929	0.9048	0.2000	0.3500	0.2750	0.1481	0.2031	0.1634
IR_J-4	0.8929	0.9048	0.3000	0.4000	0.3500	0.3185	0.4299	0.3590
IR_Y-1	0.8929	0.9048	0.1000	0.2500	0.1600	0.1792	0.2267	0.1919
bioinfo-1	0.8929	0.9048	0.3000	0.3500	0.3250	0.2266	0.2286	0.2270
bioinfo-3	0.8929	0.9048	0.3000	0.3000	0.3000	0.4137	0.4780	0.4366
MedQA-2	0.8929	0.9048	0.3000	0.4500	0.3750	0.4055	0.4372	0.4111
CSA-IISR 2nd	0.8929	0.9048	0.3500	0.4000	0.3750	0.4512	0.4998	0.4708
CSA-IISR 1st	0.8929	0.9048	0.3500	0.4000	0.3750	0.4516	0.4832	0.4632
bioinfo-2	0.8929	0.9048	0.3000	0.3000	0.3000	0.3657	0.3799	0.3709
MedQA-5	0.8929	0.9048	0.3500	0.4500	0.3917	0.4453	0.4576	0.4490
MedQA-4	0.8929	0.9048	0.3500	0.4500	0.3917	0.4453	0.4576	0.4490
Fleming-1	0.8929	0.9048	0.2500	0.3500	0.3000	0.3601	0.4570	0.3921
lean_rag	0.8929	0.9048	0.3000	0.3000	0.3000	0.4369	0.4185	0.4188
MedQA-3	0.8929	0.9048	0.3500	0.4500	0.3917	0.4073	0.4832	0.4349
CSA-IISR 3rd	0.8929	0.9048	0.3500	0.4000	0.3750	0.4564	0.4998	**0.4720**
CSA-IISR 5st	0.8929	0.9048	0.3500	0.4500	0.3917	0.4516	0.4832	0.4632
...	...	...	...	...	...	...	...	...

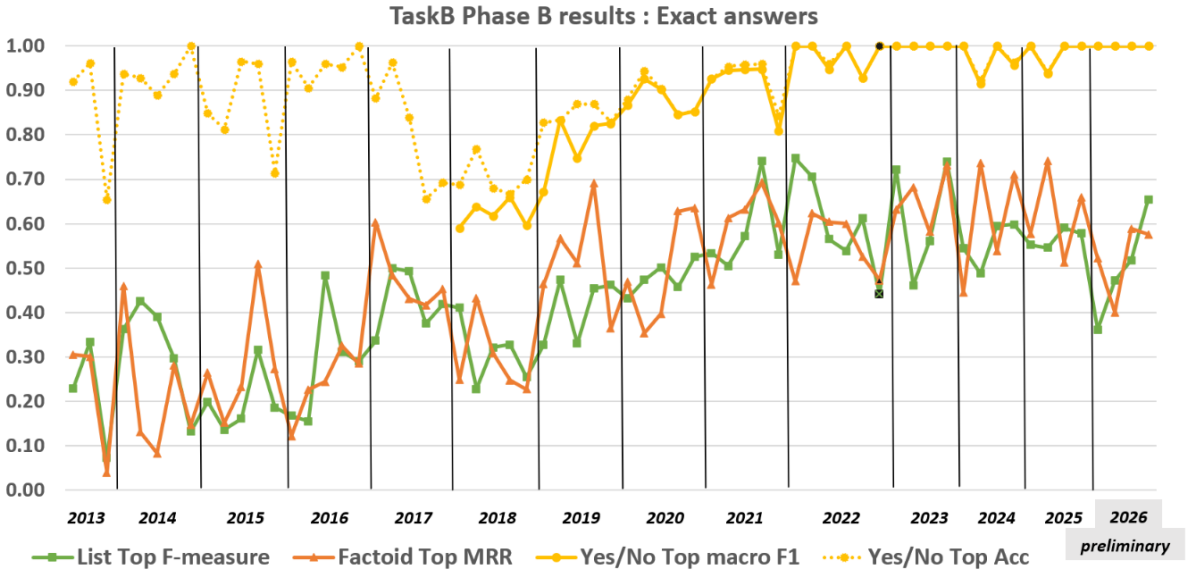

Figure 4: The evaluation scores of the best-performing systems in task B, Phase B, for *exact answers*, across the years of BioASQ. Since BioASQ6 [52], accuracy (Acc) was replaced by macro F1 as the official measure for Yes/No questions. The black dot in BioASQ10 [53] indicates an additional batch consisting of questions from new experts exclusively [49].

4.2. Task Synergy14

In task Synergy14, no relevant material was initially available for the new questions. For the questions carried over from previous rounds, however, per-question feedback was provided, i.e. the documents and snippets submitted by the participants together with manual annotations of their relevance. Consequently, the documents and snippets of the feedback, which had already been assessed and released, were not considered valid for submission in the following rounds. As in task 14b, document and snippet retrieval are evaluated with MAP and F-measure, respectively, while the information retrieval component considers only newly retrieved material, following the residual collection evaluation approach [54].

Moreover, due to the developing nature of the topics, no answer is available for all of the open questions in each round. Therefore, only the questions marked as "answer ready" were evaluated for *exact* and *ideal answers* in each round. For the *ideal answers*, the systems were ranked according to manual scores assigned to them by the BioASQ experts during the assessment of the systems' responses, as in phase B of task b [8]. As regards the *exact answers*, and similarly to task 14b, the mean F1 measure, the Mean Reciprocal Rank (MRR), and the macro F1 measure are used to rank list, factoid, and yes/no questions, respectively. Any *exact* or *ideal answer* judged to be of ground-truth quality by the experts was included in the feedback and provided to the participants before the next round.

Some indicative results for the Synergy task are presented in Table 9. The full Synergy14 results are available online[6]. Overall, the collaboration between the participating biomedical experts and the question-answering systems enabled the gradual identification of relevant material and the extraction of *exact* and *ideal answers* for several open questions on developing problems. In total, after the four rounds of Synergy14, enough relevant material was identified to answer 61 of the 63 new questions (about 97%), along with three additional questions carried over from the previous version of the task. In addition, for 48 questions (about 73%), the systems provided at least one *ideal answer* that was considered of ground-truth quality by the respective expert.

[6] https://participants-area.bioasq.org/results/synergy_v2026/

Table 9
Results for document retrieval of the first round of the Synergy14 task, ranked based on MAP.

System	Mean precision	Mean Recall	Mean F-Measure	MAP	GMAP
AgenticAI and RAG	**0.5295**	0.4281	**0.4235**	**0.5378**	0.3293
dmiip2024_3	0.4836	0.4456	0.4012	0.5023	**0.4191**
dmiip2024_2	0.4852	**0.4459**	0.4021	0.4995	0.4179
dmiip2024_4	0.4787	0.4420	0.3973	0.4892	0.4059
dmiip2024	0.4475	0.3885	0.3603	0.4304	0.2085
dmiip2024_1	0.4459	0.3868	0.3588	0.4293	0.2029
Fleming-1	0.4525	0.3388	0.3503	0.4224	0.1201
Fleming-3	0.4311	0.3154	0.3305	0.3913	0.0741
Fleming-2	0.4049	0.3052	0.3128	0.3878	0.1123
RMC_1	0.4230	0.3118	0.3276	0.3696	0.1141
RMC_2	0.4230	0.3118	0.3276	0.3696	0.1141

5. Conclusions

In this paper, we presented the fourteenth version of the BioASQ challenge, focusing on the question answering tasks b and Synergy. These tasks are well-established through the previous versions of the challenge and remain timely and relevant, as indicated by the sustained level of participation.

The preliminary results of task 14b reveal the strong performance of the top participating systems, particularly in generating yes/no answers, even under the constraints of phase A+, where no ground-truth relevant documents and snippets were provided. Performance on list and factoid questions was more variable, especially in phase A+, indicating that there is still room for improvement. These observations suggest that access to ground-truth relevant material substantially improves response quality for the more complex question types. This underlines the central role of phase A, which concerns the automatic retrieval of the relevant documents and snippets needed to answer a biomedical question. Performance in phase A showed greater variability across batches and was lower than in BioASQ 13, possibly influenced by the larger share of questions authored by new experts. A diverse set of retrieval and generation strategies was employed, including traditional IR methods, dense retrieval and re-ranking, retrieval-augmented generation, large language model (LLM)-based approaches, and systems enriched with domain-specific biomedical knowledge. Finally, the results of Synergy14 highlight that state-of-the-art QA systems can support biomedical researchers in addressing their specialized information needs on developing problems, in line with the results of previous versions of the task, despite the remaining challenges and room for improvement.

Overall, several participating systems achieved competitive performance on the QA tasks of BioASQ 14, and some of them improved over the state-of-the-art performance of previous years. Therefore, thirteen years after its initial introduction, BioASQ keeps pushing the research frontier in biomedical question answering, offering two QA tasks and several response types that span a range of biomedical information needs. The results suggest that these tasks remain far from solved, with considerable room for improvement still remaining, particularly in the retrieval of relevant material and in answering list and factoid questions, keeping biomedical question answering an open research problem.

Acknowledgments

The fourteenth edition of BioASQ is sponsored by Ovid. The MEDLINE/PubMed data resources considered in this work were accessed courtesy of the U.S. National Library of Medicine. BioASQ is grateful to the CMU team for providing the *exact answer* baselines for task 14b.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly, ChatGPT, and Claude to perform the following tasks: grammar, spelling, and consistency checks, paraphrasing, and rewording. After using these tools, the authors reviewed and edited the content as needed and took 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, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodriguez-Ortega, E. Rodriguez-Lopez, N. Loukachevitch, I. Rozhkov, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, D. Dimitriadis, A. Bekiaridou, A. Samaras, V. Patsiou, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, BioASQ at CLEF2026: The Fourteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 315–324. doi:10.1007/978-3-032-21321-1_42.

[3] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, A. Mash, M. Krallinger, Overview of MultiClinSum-2 at CLEF 2026: Data, Evaluation, and Results for Multilingual Clinical Case Report Summarization Across 15 Languages, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2025 Working Notes, 2026.

[4] I. Rozhkov, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[5] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Bekiaridou, A. Samaras, G. Giannakoulas, G. Tsoumakas, Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[6] M. Martinelli, V. Bonato, G. M. Di Nunzio, G. Faggioli, N. Ferro, O. Irrera, B. Kantz, S. Marchesin, L. Menotti, S. Merlo, R. Michelotto, G. Tussardi, F. Vezzani, P. Waldert, G. Silvello, Overview of GutBrainIE@CLEF 2026: Gut-Brain Interplay Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2025 Working Notes, 2026.

[7] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138.

[8] G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, P. Gallinari, Evaluation Framework Specifications, Project deliverable D4.1, UPMC, 2013.

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

[10] 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, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 227–250. URL: https://link.springer.com/chapter/10.1007/978-3-031-42448-9_19.

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

[12] 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, G. Maria Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco 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.

[13] 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.

[14] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023, Working Notes of CLEF (2023). URL: https://ceur-ws.org/Vol-3497/paper-003.pdf.

[15] BioASQ at CLEF2021: Large-Scale Biomedical Semantic Indexing and Question Answering, in: Advances in Information Retrieval, Springer International Publishing, Springer International Publishing, Cham, 2021.

[16] A. Krithara, A. Nentidis, E. Vandorou, G. Katsimpras, Y. Almirantis, M. Arnal, A. Bunevicius, E. Farré-Maduell, M. Kassiss, V. Konstantakos, S. Matis-Mitchell, D. Polychronopoulos, J. Rodriguez-Pascual, E. G. Samaras, M. Samiotaki, D. Sanoudou, A. Vozi, G. Paliouras, BioASQ Synergy: a dialogue between question-answering systems and biomedical experts for promoting COVID-19 research, Journal of the American Medical Informatics Association (2024) ocae232. URL: https://doi.org/10.1093/jamia/ocae232. doi:10.1093/jamia/ocae232.

[17] A. Nentidis, A. Krithara, G. Paliouras, L. Gasco, M. Krallinger, BioASQ at CLEF2022: The Tenth Edition of the Large-scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer, Springer, 2022. URL: https://link.springer.com/chapter/10.1007/978-3-030-99739-7_53.

[18] A. Nentidis, A. Krithara, G. Paliouras, E. Farre-Maduell, S. Lima-Lopez, M. Krallinger, BioASQ at CLEF2023: The Eleventh Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 577–584.

[19] 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: European Conference on Information Retrieval, Springer, 2024, pp. 490–497.

[20] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, L. Menotti, G. Silvello, G. Paliouras, BioASQ at CLEF2025: The Thirteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp.

407–415.
[21] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024, booktitle = Working Notes of CLEF, 2024. URL: https://ceur-ws.org/Vol-3740/paper-01.pdf.
[22] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, G. Paliouras, Overview of bioasq tasks 9a, 9b and synergy in clef2021, in: Proceedings of the 9th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 2021. URL: http://ceur-ws.org/Vol-2936/paper-10.pdf.
[23] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022, in: Proceedings of the 10th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 2022. URL: https://ceur-ws.org/Vol-3180/paper-10.pdf.
[24] Z. Y. Y. Z. E. Nyberg, Learning to Answer Biomedical Questions: OAQA at BioASQ 4B, ACL 2016 (2016) 23.
[25] G. Balikas, A. Kosmopoulos, A. Krithara, G. Paliouras, I. Kakadiaris, Results of the BioASQ tasks of the question answering lab at CLEF 2015, CEUR Workshop Proceedings 1391 (2015).
[26] A. Krithara, A. Nentidis, G. Paliouras, I. Kakadiaris, Results of the 4th edition of BioASQ Challenge, in: Proceedings of the Fourth BioASQ workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, 2016, pp. 1–7. URL: http://aclweb.org/anthology/W16-3101. doi:10.18653/v1/W16-3101.
[27] L. Sun, H. Yang, J. Tang, P. Quaresma, Y. Jia, Z. Li, MeSH for BioASQ Task 14B: A Lightweight Knowledge-Enhanced RAG Pipeline, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[28] D. Panou, A. C. Dimopoulos, M. Koubarakis, M. Reczko, q-MAP: sensitive QA systems with mapped question embeddings, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[29] S. Ateia, U. Kruschwitz, Can LLMs Refine Their Own Keyword Search Strategies for Biomedical RAG?, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[30] Y.-C. Hung, J.-C. Han, H.-C. Hung, R. T.-H. Tsai, NCU-IISR: BioASQ 14b Phase A/A+, and Phase B Challenge, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[31] X. Zhao, M. Lee, A. Balaji, Retrieval-Augmented Biomedical Question Answering with Weak-Question Recovery and Neural Reranking for BioASQ Task 14b, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[32] A. Avdeiko, P. Topaloglu, T. Vaccari, G. Peikos, MedCascade: A Multi-Stage Cascaded Retrieval and Generation System for Biomedical Question Answering, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[33] Y. Wang, A Multistage Evidence Retrieval System for BioASQ Task 14b: Hybrid Retrieval, Reranking, and Snippet Selection, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[34] A. Shahzadi, M. Wasim, M. Tayyab Waqar, A Multi-Stage Retrieval and Ensemble-Based Answer Generation Pipeline for Biomedical QA in BioASQ Task 14b, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[35] D. Antunes, Retrieval-Bound Generation: Lean RAG pipelines for Biomedical QA, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[36] D. Galat, M.-A. Rizoiu, Cost-Pragmatic Quality Gating and Selection–Fusion Multi-Model Combiners for BioASQ Phase A+/B, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[37] D. Sims, K. Williams, S. Henry, RMC at BioASQ 2026: A local RAG approach for end-to-end Biomedical QA, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[38] A. Ribeiro, R. Garrido, A. Christiansen, R. A. A. Jonker, S. Matos, BIT. UA at BioASQ 14B: Modular Retrieval with pg_textsearch and Qdrant, and Agent-Based Answer Generation, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[39] S. Volz, B. Scheuvens, S. Ghosh, H. Scells, Health-NLP at the BioASQ 2026 Task B on Biomedical Semantic QA and GutBrainIE on Gut-Brain interplay Information Extraction Tasks, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[40] R. Leekha, D. Gwon, L. Shing, Agentic RAG with open-weight LLMs for Biomedical QA Task 14b, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[41] A. M. Kouti, I. Refanidis, Sparse Retrieval with Cross-Encoder Reranking and Heuristic Snippet Scoring for Biomedical Question Answering, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[42] J. Bakker, J. Kamps, Contrastive Learning for Efficient Reranking in Task 14B of the CLEF 2026 BioASQ Track, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[43] N. Nguyen, A. Elnagger, A. Kashyap, T. Schultz, Few but Deep vs. Many but Shallow: Comparing Licensed and Free Tool Sets for Agentic Biomedical Question Answering, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[44] R. T.-H. Tsai, Y.-S. Chiu, C.-Y. Hou, C.-H. Hsu, BioASQ Phase B: Exploring Diverse Optimization Techniques and Testing Frameworks for Enhanced Bio-Answering Performance, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[45] J. Gibson, E. Nyberg, SynapFlow at BioASQ@CLEF 2026: Retrieval-Augmented Generation with Diversity-Enhanced Hybrid Retrieval, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[46] R. Sukprasan, O. Alexandru Iulian, F. Jacob, R. Daniela, LLM-Guided RAG: A Biomedical Retrieval System for BioASQ Synergy 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[47] M. Krallinger, A. Krithara, A. Nentidis, G. Paliouras, M. Villegas, BioASQ at CLEF2020: large-scale biomedical semantic indexing and question answering, in: European Conference on Information Retrieval, Springer, 2020, pp. 550–556.
[48] A. Nentidis, A. Krithara, K. Bougiatiotis, M. Krallinger, C. Rodriguez-Penagos, M. Villegas, G. Paliouras, Overview of BioASQ 2020: The eighth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), Thessaloniki, Greece, September 22–25, 2020, Proceedings, volume 12260, Springer, 2020.
[49] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2021: The Ninth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 239–263.
[50] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, 2022. doi:10.1007/978-3-031-13643-6_22.
[51] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, G. Paliouras, Overview of BioASQ Tasks

10a, 10b and Synergy10 in CLEF2022, in: CEUR Workshop Proceedings, volume 3180, 2022, pp. 171–178.

[52] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, I. Kakadiaris, Results of the sixth edition of the BioASQ Challenge, in: Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 1, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–10. URL: http://aclweb.org/anthology/W18-5301. doi:10.18653/v1/W18-5301.

[53] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13390 LNCS, 2022, pp. 337–361. doi:10.1007/978-3-031-13643-6_22. arXiv:2210.06852.

[54] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science 41 (1990) 288–297. doi:10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8>3.0.CO;2-H.