Overview of BioASQ Tasks 13b and Synergy13 in CLEF2025

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

[1]NCSR Demokritos, Athens, Greece

Abstract

This paper presents an overview of the Question Answering (QA) tasks in the thirteenth edition of the BioASQ challenge, on large-scale biomedical semantic indexing and QA, which is part of the Conference and Labs of the Evaluation Forum (CLEF) 2025. For more than a decade, BioASQ has been serving as a key platform for advancing the state-of-the-art in biomedical information retrieval, NLP, and QA. In this paper, we present a comprehensive overview of the biomedical QA tasks 13b and Synergy13 of the thirteenth BioASQ challenge (BioASQ 13). This year, 49 teams with more than 160 systems participated in these two tasks of the challenge, with 46 focusing on task 13b, on biomedical semantic QA, and 5 on task Synergy13, on QA for open questions on developing biomedical topics. The competitive performance achieved by several participating systems in the QA tasks of BioASQ 13 highlights the continuous advancement of state-of-the-art methods in the field, in alignment with previous editions of the tasks.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction

This paper presents an overview of the thirteenth BioASQ challenge (2025), with a specific focus on shared tasks 13b and Synergy13. We describe the corresponding datasets utilized for training and evaluating participating systems. Details of tasks 13b and Synergy13, which ran from March to May and January to March 2025, respectively, are provided in Section 2. Section 3 provides a brief overview of the participation in these two tasks. A comprehensive analysis of the methodologies employed by participating systems will be included in the BioASQ workshop proceedings in [1]. The paper concludes with a brief discussion of our key findings.

2. Overview of the Tasks

The thirteenth edition of the BioASQ challenge featured six primary tasks: (1) a biomedical question answering task (task b), (2) a task on biomedical question answering for open developing issues (task Synergy), both tasks considering documents in English, (3) a new task focused on the multilingual summarization of clinical documents (task MultiClinSum), (4) a new task focused on Biomedical Nested Named Entity Linking in English and Russian(task BIONNE-L), (5) a new task on developing named entity recognition, entity linking and multi-label classification-explainable AI systems using a specialized corpus of discharge letters from the cardiac department of a tertiary hospital (task ElCardioCC), and (6) a new task focused on extracting structured information from biomedical abstracts related to the gut microbiota and its connections with Parkinson's disease and mental health (task GutBrainIE) [1].

In this paper, we describe the current versions of the first two established tasks, referring to them as task 13b and task Synergy13 within the context of the thirteenth BioASQ edition. In-depth information on the new MultiClinSum, BIONNE-L, ElCardioCC, and GutBrainIE tasks can be found in [2], [3], [4], and [5], respectively. Furthermore, a detailed introduction to the BioASQ challenge and its initial task structure is provided in [6].

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

*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 13b

Task 13b introduces a comprehensive question-answering challenge in the biomedical field, requiring participants to develop systems that address all stages of question answering. As in previous editions, the task focuses on four question types: 'yes/no,' 'factoid,' 'list,' and 'summary' questions [7].

In the thirteenth edition of the BioASQ Challenge, participating teams received an updated version of the BioASQ QA training dataset [8], containing 5,389 questions that had been annotated with relevant golden elements and answers from previous task versions [9, 10]. These questions served as the basis for developing their systems. Table 1 provides a detailed overview of both the training and testing sets for task 13b. A notable observation from these statistics is the significantly larger average number of documents and snippets within the training data compared to the test batches. This can be attributed to two main factors. First, in the early years of BioASQ the annotation with relevant documents and snippets by the experts was exhaustive, in an attempt to identify as many relevant items as possible in the corpus. These questions are part of the training datasets affecting the average number of relevant items per question. Currently, only a sufficient number of relevant answers is required when the initial version of the data is developed. Still, when the participants submit their responses, the experts assess the submitted items and enrich the ground-truth data with potential additional relevant items detected by the participants. The numbers of relevant items for the test sets in Table 1 are preliminary, before the enrichment by the assessment process, which is still in progress. The final evaluation of the participants will be against these enriched relevant items, ensuring that all the submitted items that are relevant are indeed handled as such.

Table 1
Statistics on the training and test datasets of Task 13b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	5,389	1,459	1,047	1,600	1,283	9.74	12.78
Test 1	85	17	23	26	19	2.68	3.74
Test 2	85	17	19	27	22	2.71	3.06
Test 3	85	22	22	20	21	3.00	3.66
Test 4	85	26	19	22	18	3.15	3.92
Total	5,729	1,541	1,130	1,695	1,363	9.33	12.23

Task 13b, similar to the previous version of the task (12b), was structured into three phases [11]. As in older task versions, task 13b was divided into four independent bi-weekly batches, and the three phases for each batch ran for two consecutive days [12]. The three phases of task 13b included: (phase A) the retrieval of the required information, (phase A+) answering the question without golden feedback, and (phase B) answering the question with golden feedback, which ran for two consecutive days for each batch. Participants were given 24 hours to submit their system's responses after receiving the test set for each respective phase. This year, each test set consisted of 85 questions. For each test set, the respective questions, written in English, were released for phase A, requiring participants to identify and submit relevant elements from designated resources, including PubMed/MedLine articles and snippets extracted from these articles. Then, these questions were also released in phase A+ and the participating systems were asked to respond with *exact answers*, that is, entity names or short phrases, and *ideal answers*, that is, natural language summaries of the requested information. Finally, during phase B, manually selected relevant articles and snippets related to these questions were also made available, and participating systems were once again asked to provide *exact answers* and *ideal answers*.

2.2. Synergy13 Task

Introduced in the ninth BioASQ edition [13], the Synergy task aimed to foster collaboration between biomedical experts studying COVID-19 and automated question-answering systems participating in BioASQ. The core objective is to create a synergy where experts assess system responses, and this

feedback is used to iteratively improve the systems [14]. The task continued to focus on COVID-19 in the tenth edition of BioASQ [15], and it was extended to any developing biomedical topic in the subsequent editions of the challenge [16, 17, 18].

In the process depicted in Figure 1, competing systems provide their initial responses to open questions related to emerging problems. These responses, along with relevant documents and snippets, are evaluated by experts. Subsequently, the experts provide feedback to the systems and address any new or pending questions.

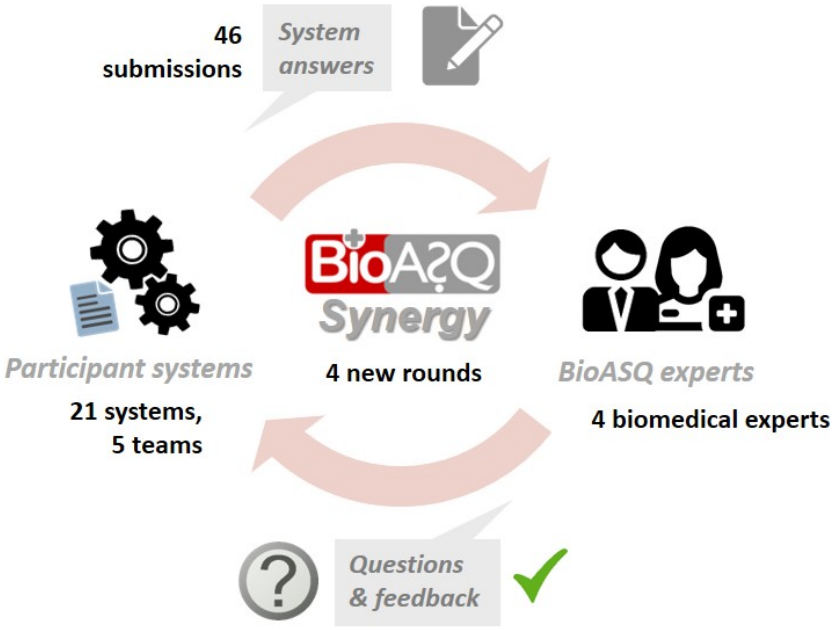

Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy13 task on question answering for open developing problems.

As in previous years, the Synergy task (Synergy13) comprised four rounds, with a two-week interval between each round, and focused on emerging biomedical topics, drawing from relevant documents in the current PubMed version [12, 11]. Consistent with earlier versions, the questions posed were open-ended, allowing for dynamic responses [19, 20].

In the Synergy task, during each round, the system responses and expert feedback address the same questions, unless those questions have already been closed by experts due to receiving a comprehensive and definite answer. Specifically, in Synergy13, a group of four biomedical experts contributed a total of 74 open biomedical questions. This set includes 47 new questions on developing health topics, such as infectious, rare, and genetic diseases, and women's and reproductive health, and 27 questions from the previous version of the task [11], which remained open and were enriched with more recent evidence and updated answers. The experts evaluated the retrieved material (including documents and snippets) and the responses submitted by participating systems in all four rounds. Table 2 shows the details of the datasets used in task Synergy13.

Table 2
Statistics on the datasets of Task Synergy13. "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	74	23	19	14	18	43
2	74	23	19	14	18	49
3	74	23	19	14	18	51
4	74	23	19	14	18	51

Synergy13, similar to task 13b, addresses four question types: yes/no, factoid, list, and summary, and two types of answers, exact and ideal. Moreover, the evaluation of systems relies on the same measures used in task 13b. Upon completing the Synergy13 task, relevant material was identified for answering roughly 80% of the questions. Additionally, around 47% of the questions had at least one ideal answer submitted by the systems, which was deemed satisfactory by the expert who posed the question.

3. Overview of participation

In this year's BioASQ challenge, over 160 distinct systems engaged in tasks 13b and Synergy13 with a total of 49 teams, 39 of which were new. The high percentage of teams joining the BioASQ challenge for the first time, which is almost 80%, indicates the enduring interest of the community in large-scale biomedical question answering. Specifically, 46 of these teams submitted on at least some phase of task 13b and 5 on task Synergy13, with several teams participating in more than one task and phases, as demonstrated in Figure 2.

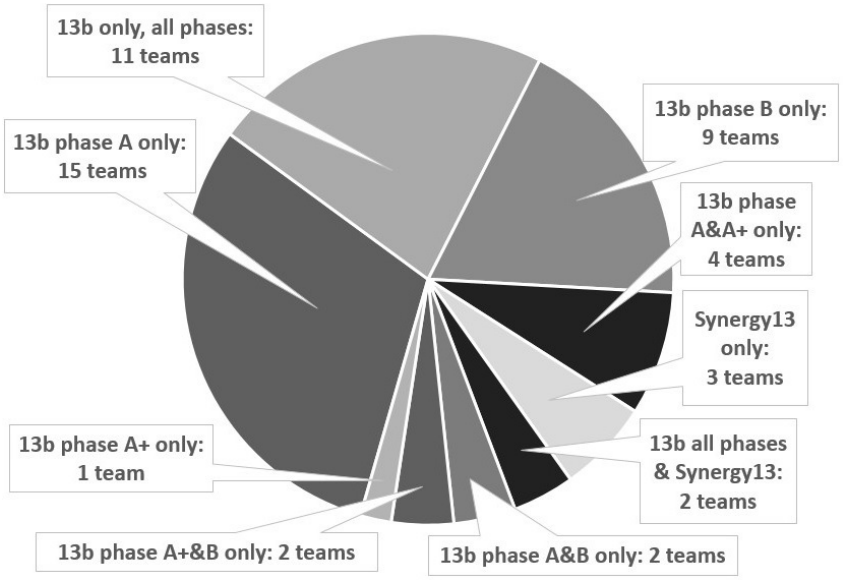

Figure 2: The distribution of participant teams in BioASQ task 13b phases and task Synergy13.

In line with previous years [20, 12], task b attracted more participants than Synergy, with a large increase in the total number of participating teams this year in comparison to last year, as illustrated in Figure 3. The increased participation in task 13b can be partly attributed to 15 student teams from a course on Advanced Information Retrieval at TU Wien, which incorporated the phase A of BioASQ task 13b as part of the course assignments, highlighting the educational potential of BioASQ. Specifically, these teams account for about 38% of all new teams in the QA tasks of BioASQ 13.

3.1. Task 13b

This year, 46 teams participated in task 13b, submitting a total of 146 distinct systems across all three phases A, A+, and B. Specifically, 34, 20, and 26 teams competed in phases A, A+, and B, with 95, 79, and 88 distinct systems, respectively. Eleven of these teams were involved in all three phases, as depicted in Figure 2.

An overview of the technologies utilized by the teams is outlined in Table 3. Additional details for specific systems can be found in the workshop proceedings. As in previous years, the open-source system OAQA [21], which achieved top performance in older editions of BioASQ [22, 23], was used as a baseline for phase B *exact answers*. This system is based on the UIMA framework and relies on

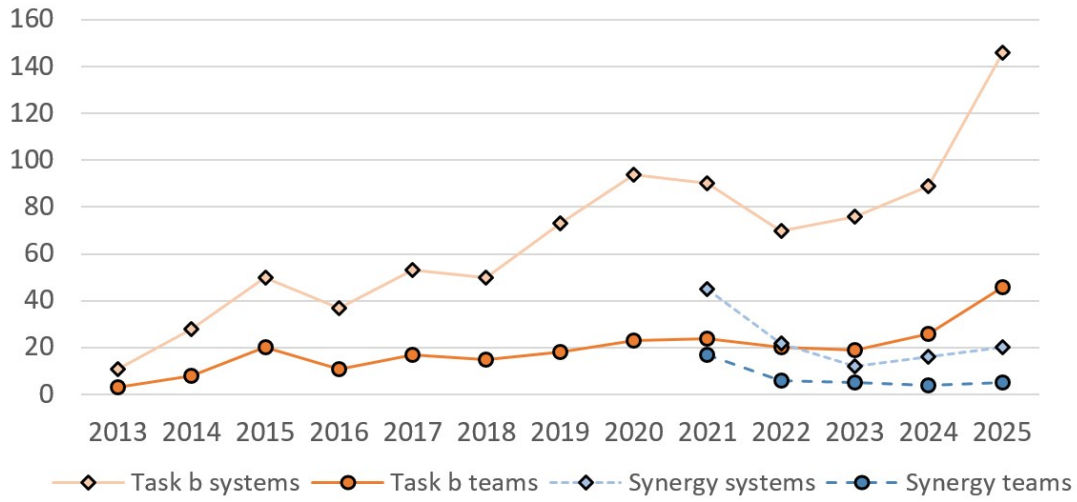

Figure 3: Participation evolution in task b and Synergy during the thirteen years of BioASQ.

traditional NLP and Machine Learning approaches and tools, such as MetaMap and LingPipe [1].

Table 3
Systems and approaches for task 13b. Systems for which no information was available at the time of writing are omitted.

Systems	Phase	Ref.	Approach
UA	A,A+,B	[24]	BM25, PubMedBERT, BioLinkBERT, re-ranking, RAG, llama-3, gemma-3
UR	A,A+,B	[25]	ElasticSearch, gemini flash 2.0, o4-mini, o3-mini, deepSeek reasoner, query expansion, re-ranking, self-feedback
NCU	A,A+,B	[26, 27]	BM25, RAG, bge-reranker, llama-3, GPT-4o
BSRC	A,A+,B	[28]	BM25, RM3, query expansion, LLM ensemble,
PJs	A,A+,B	[29]	dense retrieval, reranking, MiniLM-L6, claude, Qwen2.5
JU_NLP	A,A+,B	[30]	minilm, T5, roberta
UniTor	A,A+,B	[31]	BM25, reranking, SentenceBERT, llama-3, phi-4
UniBol	A,A+,B	[32]	PubMedBERT, Qwen3
lasigeBioTM	A,A+,B	[33]	mistral, KG, NER
AQAMS	A,A+,B	[34]	hybrid retrieval, llama, openAI-gpt, NER
ZUT	A,A+,B	[35]	re-ranking, RAG, deepseek
FSU	A,A+	-	re-ranking, LLMs, IKraph
GT	A,B	[36]	bge-large, gpt, re-ranking
MQU	A+,B	[37]	zero-shot, gemini, claude, re-ranking
HCMC	A+,B	-	SciSpaCy, all-MiniLM-L12, GPT-4o-mini, query expansion
UTB	B	[38]	BERT
VCU	B	-	Synthia-13B, llama-3
Evidence	B	[39]	nomic-ai, LLM ensemble
DMIS	B	[40]	GPT-4, claude, LLM ensemble

The UA team from the University of Aveiro participated in all three phases of the task with five systems. In phase A, their systems followed a two-stage retrieval pipeline in line with previous years' submissions [41, 42]. To enhance the BM25 results they used the BGE-M3 model, and reciprocal rank fusion to combine model outputs. For phases A+ and B, their systems adapted RAG with prompts incorporating relevant abstracts using instruction-based transformer models such as llama-3 70B and Gemma-3 27B.

The UR team from the Universität Regensburg competed in all phases of the task with five systems.

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

Their systems employed few-shot prompting (typically 0-shot or 10-shot configurations) and an experimental two-step self-feedback mechanism. In this feedback approach, an LLM generated an initial answer, then provided critical feedback on its own output, and finally refined the answer based on this critique. This process was applied across yes/no, factoid, list, and ideal answer types. For the retrieval stages, their systems utilized LLM-driven query expansion, which also incorporated a similar self-feedback loop for query refinement based on initial search results, coupled with Elasticsearch for document retrieval, followed by LLM-based snippet extraction and re-ranking.

The NCU team from the National Central University participated with five systems. Their systems employed a basic Retrieval-Augmented Generation (RAG) framework consisting of a retriever, a reranker, and a LLM for language generation. The initial retriever used was a BM25, and the documents were further re-ranked using the bge-eranker-v2-m3 model to identify the most relevant articles and snippets. For answer generation, their systems used the meta-llama/Llama-3.1-8B-Instruct and GPT-4o modles. Furthermore, for the Phase A+ task, they extended the answer generation pipeline previously developed by Chih et al. [43]. For phase B, the systems utilized GPT-4o, o3-mini, and Gemini 2.0 Flash. Each system employed either direct or two-stage generation for ideal and exact answers.

Another team participating in all phases is the team from the BSRC Alexander Fleming Institute. For Phases A and A+, their systems focused on document retrieval using the BM25 algorithm enhanced with RM3, and re-ranking based on the relevance of associated text snippets, similar to the previous year participation [44]. In phases A+ and B, their systems explored three distinct prompting strategies for generating exact answers (snippet-based, abstract-based, and extended abstract-based). To further enhance answer quality, their systems build upon their earlier LLM ensemble strategy [44] and extend it to all exact answer types. For list and factoid questions, they use the union of answers across multiple LLMs to maximize recall.

The PJs team participated in all three phases, submitting five system configurations in total. For Phase A, their systems employed dense retrieval using the e5-base model and then re-ranking using different models (e.g. modernbert-embed-base, gte-large, granite-25m, gte ensemble, modernbert, and granite). For snippet selection, MiniLM-L6 was utilized. For Phase A+, their systems used Claude Sonnet 3 with prompt engineering on the top retrieved results, while for Phase B, the the same prompt design was utilized with different LLM modles including Qwen2.5-3B, Claude Sonnet 3.7, Claude Haiku 3.5, and an ensemble of Sonnet 3.7 and Sonnet 3.0 with post-processing to correct errors.

Also, the JU_NLP team from Jadavpur University participated in all phases. For phase A, their systems perform dense retrieval using a fine-tuned minilm model. Then, a RoBERTa-base-SQuAD2 model extracts the answer snippet. For phase B, a separately fine-tuned T5 model for generating ideal and exact answers is utilized. Finally, phase A+ combines these processes, first using the Phase A system to extract relevant articles and snippets, and then feeding those results into the Phase B system to produce ideal and exact answer.

The UniTor team from Università degli Studi di Roma Tor Vergata participated in all three phases. Their systems employed a multi-stage pipeline that combined sparse and dense retrieval techniques, supervised re-ranking, supervised LLM snippet extraction, and supervised LLM answer generation. All components were trained and evaluated using official BioASQ datasets and external biomedical resources such as PubMed.

The UniBol team from Universidad Tecnológica de Bolivar participated in all three phases. Their systems employed the PubMedBert model to perform dense retrieval and Qwen3 8b with prompt engineer to generate answers.

Another team that participated in all phases is the lasigeBioTM team. Their systems used Mistral as the baseline model for all phases. Furthermore, they incorporated external knowledge from various ontologies and knowledge bases like Human Disease Ontology and NCBI Gene, with BENT tool used for named-entity recognition and linking. Also, some systems employed a Decoding on Graphs methodology, utilizing Monarch KG and MARISA Trie for structured reasoning.

The AQAMS team from Universidad Europea took part in all phases. For phases A and A+ their systems used a two-stage pipeline. First, LlamaIndex embeddings were used to perform dense retrieval. Then, OpenAI GPT models with structured prompt templates were utilized to generate answers. For

phase B, their systems focused on biomedical NER using a fine-tuned model, optionally mapping terms to UMLS concepts, and generating answers, all deployed via a user-friendly Streamlit interface.

The ZUT team from Zhongyuan University of Technology participated in all phases. For phase A, their systems employed a query expansion-driven multi-stage retrieval and re-ranking framework that utilized different DeepSeek models. In phases A+ and B, their systems utilized different prompting strategies and also experimented with supervised fine-tuning of LLMs.

The FSU team from Florida State University participated in phases A and A+. Their systems followed a multi-stage retrieval and reasoning framework, leveraging an LLM for keyword extraction and answer generation, along with an internal search engine for document retrieval. A re-ranker model and an LLM-based scoring mechanism filter retrieved documents for quality. Additionally, their systems integrate structured biomedical knowledge from IKraph, an in-house knowledge graph built from PubMed abstracts using large-scale relation extraction and causal inference.

The GT team from Georgia Tech participated in both phases A and B. For phase A, their systems employed a custom index built with the bge-large-en embedding model and used a fine-tuned ms-marco-12 model for re-ranking, with some configurations also incorporating GPT-4o as an additional re-ranker. In phase B, the team utilized Mistral-7B-Instruct-v0.3 and GPT-4o-turbo, with a prompt-based approach using few-shot examples to generate answers.

The MQU team from Macquarie University participated in phases A+ and B. Their systems employed a zero-shot QA framework, using prompting multiple LLMs (Gemini variants and Claude) to generate answers based on snippets and full abstracts. A secondary synthesis step with confidence-based re-tries refines the candidate answers into a final response, improving precision and consistency across yes/no and factoid questions.

Also, the HCMC team from the University of Science participated in phases A+ and B. Their systems first distinguish queries as single-hop or multi-hop; multi-hop queries are decomposed into sub-queries. Relevant documents are retrieved using the PubMed API, with queries reformulated into Conjunctive Normal Form (CNF). Document sentences are encoded using SciSpaCy, and dense retrieval is performed using all-MiniLM-L12-v2 model. Finally, GPT-4o-mini is utilized to generate answers.

The VCU team from Virginia Commonwealth University participated with four different systems in phase B. Their systems are based on a zero-shot learning approach using generative LLMs, including Synthia-13B-GPTQ and llama-3. Their systems heavily relied on prompt engineering and answer processing.

The Evidence team from Evidence Prime participated in phase B. Their systems employed dense retrieval using the nomic-embed-text-v1 model. Various open-source LLMs and proprietary models were used, including GPT-4o, GPT-4.1, Claude 3.5, and Claude 3.7. For Yes/No and Factoid/List questions, an ensemble strategy was implemented, using voting or frequency analysis of outputs from multiple LLMs. Summary responses were generated using carefully designed, hand-crafted prompts to balance contextual similarity and appropriate length.

The DMIS team from the Korea University participated in phase B. Their systems employed multiple LLMs, including GPT-4o-mini, GPT-4, and Claude. They explored various prompting strategies, such as standard instruction, one-by-one snippet querying, randomized snippet order, and a no-snippet condition relying on prior knowledge. Final answers were derived through ensemble techniques, either by aggregating log probability scores or using majority voting.

Another team that participated only in phase B is the UTB team from the Universidad Tecnológica de Bolívar. Their systems were based on the well-known BERT model.

3.2. Synergy13 Task

In the thirteenth edition of BioASQ, five teams participated in the Synergy task (Synergy13). These teams submitted 46 runs from 21 distinct systems. Two of these teams participated in task 13b as well, while the remaining three focused exclusively on task Synergy13. An overview of systems and approaches employed is provided 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
BSRC	[28]	open source LLMs, sparse/dense/hybrid retrieval
SINAI	[45]	NER, RAG, llama

In particular, the BSRC Alexander Fleming team participated with four systems. Similar to task b, their systems focused on LLMs, specifically the DeepSeek-R1 model, with optimized prompts and majority voting. Also, the SINAI team from the Universidad de Jaén competed with five systems. Their systems built upon prior research by integrating a lightweight NER-based query module, dynamic indexing of PubMed abstracts, and few-shot prompt templates tailored to different question types. These components are utilized within a RAG framework based on a biomedical fine-tuned LLaMA model. The system employs a multi-stage pipeline—comprising data preparation, context extraction, and response generation—designed specifically for biomedical question answering across the required formats (summary, yes/no, factoid, list).

4. Results

4.1. Task 13b

This section presents the evaluation measures and preliminary results for the task 13b. These results are preliminary, as the final results will be available 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 in progress.

Phase A: The Mean Average Precision (MAP) was used for evaluation on document retrieval. In particular, since BioASQ8 [46], MAP calculation is based on a modified version of Average Precision (AP) that considers both the limit of 10 elements allowed per question in each submission and the actual number of golden elements that is often less than 10 in practice [47]. For snippets, where a single ground-truth snippet may overlap with several submitted ones, the interpretation of MAP is less straightforward. Hence, since BioASQ9 [13], we use the F-measure which is based on character overlaps[2] [48]. Tables 5 and 6 present some indicative results in batch 1 for document and snippet retrieval, respectively. The full 13b results for phase A are available online[3].

Phases A+ and B: The official ranking for systems providing *ideal answers* is based on manual scores assigned by the BioASQ team of experts that assesses each *ideal answer* in the responses [7]. The final position of systems providing *exact answers* is based on their average ranking in the three question types where *exact answers* are required, that is "yes/no", "list", and "factoid". Summary questions for which no *exact answers* are submitted are not considered in this ranking. In particular, the mean F1 measure is used for the ranking in list questions, the Mean Reciprocal Rank (MRR) is used for the ranking in factoid questions, and the F1 measure, macro-averaged over the classes of yes and no, is used for yes/no questions. Tables 7 and 8 present some indicative results on *exact answer* extraction. The full 13b 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 type of question during the thirteen years of BioASQ is presented in Figure 4. The preliminary results for task 13b, reveal that the participating systems keep achieving high scores in answering all types of questions, despite the addition of two new experts to the BioASQ team. In batch 2, phase B, for instance, presented in Table 8, several systems manage to correctly answer literally all yes/no questions, as well as in batches

[2] http://participants-area.bioasq.org/Tasks/b/eval_meas_2022/
[3] http://participants-area.bioasq.org/results/13b/phaseA/
[4] http://participants-area.bioasq.org/results/13b/phaseAplus/
[5] http://participants-area.bioasq.org/results/13b/phaseB/

Table 5
Preliminary results for document retrieval in batch 1 of phase A of task 13b, ranked based on MAP.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
bioinfo-4	0.1047	0.5043	0.1605	**0.4246**	0.0104
bioinfo-2	0.1000	0.4916	0.1537	0.4214	0.0077
bioinfo-3	0.1047	0.5057	0.1605	0.4175	**0.0123**
bioinfo-0	0.1047	0.5037	0.1602	0.4141	0.0101
bioinfo-1	0.1071	**0.5106**	0.1636	0.4082	**0.0123**
Baseline top 20	0.0788	0.4720	0.1300	0.3806	0.0051
Baseline top 10	0.0788	0.4720	0.1300	0.3806	0.0051
dmiip2024_1	0.0718	0.4465	0.1192	0.3660	0.0034
Using LLM alone	0.0753	0.4598	0.1249	0.3622	0.0040
dmiip2024	0.0682	0.4353	0.1140	0.3590	0.0031
dmiip2024_4	0.0706	0.4327	0.1169	0.3528	0.0033
dmiip2024_3	0.0694	0.4278	0.1149	0.3469	0.0024
dmiip2024_2	0.0682	0.4249	0.1133	0.3428	0.0024
IR1	0.0718	0.4312	0.1173	0.3394	0.0036
lasigeBioTM	0.0624	0.3920	0.1029	0.3207	0.0023
IRIS_1	0.0682	0.4282	0.1129	0.3081	0.0033
UniTor_0	0.0565	0.3847	0.0948	0.3020	0.0026
UniTor_1	0.1042	0.3608	0.1486	0.3004	0.0016
Using KG for list q	0.1453	0.4225	0.1940	0.3004	0.0029
config-1	0.0553	0.3745	0.0934	0.2903	0.0021
Main pipeline	0.1512	0.3971	0.1908	0.2885	0.0020
UR-IW-5	**0.1677**	0.3471	**0.2038**	0.2865	0.0015
UniTor_2	0.0553	0.3725	0.0933	0.2784	0.0012
Fleming-1	0.0606	0.3863	0.1005	0.2716	0.0020
UniTor_3	0.0852	0.3471	0.1303	0.2712	0.0011
UR-IW-1	0.1415	0.3194	0.1776	0.2527	0.0010
llama	0.1339	0.2853	0.1642	0.2466	0.0006
config-3	0.0494	0.3539	0.0844	0.2342	0.0010
UR-IW-2	0.1376	0.2941	0.1699	0.2272	0.0007
mistral	0.1260	0.2876	0.1542	0.2271	0.0005
config-2	0.0506	0.3706	0.0869	0.2114	0.0011
UR-IW-3	0.1344	0.2547	0.1557	0.2064	0.0005
DS@GT BioASQ...	0.0584	0.2325	0.0843	0.1980	0.0003
UR-IW-4	0.0979	0.1892	0.1135	0.1739	0.0001
deepseek32b-full	0.1514	0.2298	0.1641	0.1712	0.0002
deepseek32b-me	0.1378	0.1690	0.1425	0.1428	0.0001
dense	0.0295	0.1857	0.0492	0.1324	0.0002
config-4	0.0619	0.4059	0.1042	0.1314	0.0017
config-5	0.0619	0.4059	0.1042	0.1302	0.0017
no expand...	0.1133	0.1480	0.1196	0.1268	0.0001
13b_phase_a	0.0430	0.1396	0.0591	0.1005	0.0001
deepseek-r1:8b	0.0381	0.1876	0.0611	0.0932	0.0001
deepseek-r1:14b	0.0251	0.1167	0.0395	0.0764	0.0001
deepseek-r1:32b	0.0381	0.1945	0.0598	0.0724	0.0001
GPT4O	0.0343	0.1363	0.0532	0.0665	0.0001
IRIS_3	0.0035	0.0186	0.0056	0.0082	0.0000
IRIS_2	0.0035	0.0235	0.0061	0.0034	0.0000

1 and 4. In phase A+, on the other hand, where no ground truth relevant material is available, correctly answering all yes/no questions is more challenging, though not infeasible. In batch 2, for instance, presented in Table 7, only one system manages to do so. The preliminary results of task 13b, Phase B,

Table 6
Preliminary results for snippet retrieval in batch 1 of phase A of task 13b, ranked based on Mean F-measure.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
UR-IW-5	**0.1189**	0.1928	**0.1202**	0.2768	0.0006
dmiip2024_1	0.0803	**0.3050**	0.1186	**0.4535**	**0.0014**
UniTor_0	0.0957	0.2120	0.1170	0.2770	0.0011
dmiip2024	0.0783	0.2891	0.1157	0.4533	0.0012
UR-IW-2	0.1136	0.1633	0.1110	0.2478	0.0005
dmiip2024_3	0.0741	0.2865	0.1105	0.4158	0.0010
UniTor_1	0.0833	0.2120	0.1081	0.2770	0.0011
dmiip2024_4	0.0716	0.2767	0.1077	0.4158	0.0009
dmiip2024_2	0.0713	0.2821	0.1076	0.4156	0.0009
UR-IW-1	0.0978	0.1594	0.1071	0.2762	0.0005
Main pipeline	0.0989	0.1630	0.0958	0.1566	0.0005
Using KG for list q	0.0897	0.1797	0.0939	0.1629	0.0007
Using LLM alone	0.0679	0.1959	0.0936	0.2051	0.0010
Baseline top 20	0.0671	0.2042	0.0933	0.2109	**0.0014**
Baseline top 10	0.0671	0.2042	0.0933	0.2109	**0.0014**
UR-IW-3	0.0863	0.1393	0.0912	0.2447	0.0003
UniTor_2	0.0624	0.1930	0.0874	0.2282	0.0007
llama	0.1159	0.1054	0.0870	0.1644	0.0002
UniTor_3	0.0569	0.1930	0.0817	0.2282	0.0007
UR-IW-4	0.0795	0.1035	0.0778	0.1844	0.0001
DS@GT BioASQ...	0.0562	0.1310	0.0761	0.1390	0.0001
mistral	0.1077	0.0717	0.0703	0.1170	0.0001
deepseek32b-full	0.0998	0.0575	0.0651	0.1131	0.0001
deepseek32b-me	0.0955	0.0517	0.0603	0.1085	0.0001
dense	0.0456	0.1117	0.0583	0.1466	0.0001
lasigeBioTM	0.0385	0.0560	0.0417	0.1426	0.0006
Fleming-1	0.0280	0.0924	0.0398	0.1009	0.0003
config-4	0.0284	0.0714	0.0367	0.0411	0.0001
config-5	0.0284	0.0714	0.0367	0.0396	0.0001
config-1	0.0229	0.0630	0.0310	0.0824	0.0001
google_serach_&_LLM	0.0428	0.0334	0.0289	0.0000	0.0000
config-2	0.0226	0.0510	0.0287	0.0561	0.0001
deepseek-r1:32b	0.0201	0.0355	0.0236	0.0304	0.0001
config-3	0.0177	0.0448	0.0235	0.0687	0.0001
no expand only...	0.0389	0.0176	0.0220	0.0478	0.0000
deepseek-r1:8b	0.0176	0.0386	0.0218	0.0446	0.0000
13b_phase_a	0.0209	0.0168	0.0173	0.0494	0.0000
deepseek-r1:14b	0.0122	0.0326	0.0173	0.0419	0.0000
GPT4O	0.0094	0.0166	0.0101	0.0162	0.0000

reveal improved and more consistent performance for list questions compared to the previous years, but room for improvement is still available, as is for factoid questions, where the performance across batches presents more fluctuations.

Table 7
Results for batch 2 for *exact answers* in phase A+ of task 13b, ranked based on Yes/No F1.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
UR-IW-5	**1.0000**	**1.0000**	0.2963	0.3333	0.3086	0.1796	0.3432	0.2144
Fleming-1	0.9377	0.9412	0.2222	0.3704	0.2790	0.1934	0.3238	0.2242
dmiip2024_2	0.9377	0.9412	0.4444	0.4444	0.4444	0.1987	0.4024	0.2559
phaseB-5	0.9377	0.9412	0.3333	0.3333	0.3333	0.2395	0.2984	0.2574
phaseB-4	0.9377	0.9412	0.3333	0.3333	0.3333	0.2395	0.2984	0.2574
Baseline top 20	0.9328	0.9412	0.4444	0.4815	0.4630	**0.3785**	0.4357	**0.3880**
Main pipeline	0.9328	0.9412	0.3333	0.4074	0.3642	0.1814	0.4753	0.2238
Baseline top 10	0.9328	0.9412	0.4074	0.4815	0.4383	0.3609	0.4378	0.3748
Using KG for...	0.9328	0.9412	0.3333	0.4074	0.3642	0.1937	0.5016	0.2412
UR-IW-1	0.9328	0.9412	0.4074	0.4815	0.4383	0.2307	0.3536	0.2682
Fleming-2	0.9328	0.9412	0.2222	0.3704	0.2790	0.1934	0.3238	0.2242
deepseek-r1:14b	0.9328	0.9412	0.1481	0.1481	0.1481	0.1649	0.1750	0.1666
UR-IW-3	0.8786	0.8824	0.3704	0.4444	0.3920	0.1696	0.3213	0.2118
UR-IW-2	0.8712	0.8824	0.4815	0.5185	0.5000	0.3449	0.4626	0.3805
dmiip2024_3	0.8712	0.8824	0.4074	0.4444	0.4259	0.3614	0.4018	0.3711
mistral	0.8712	0.8824	0.4074	0.5185	0.4475	0.1986	0.3514	0.2330
deepseek32b-f	0.8712	0.8824	0.2963	0.2963	0.2963	0.2458	0.3160	0.2675
deepseek-r1:8b	0.8712	0.8824	0.1111	0.1111	0.1111	0.3358	0.3805	0.3422
GPT4O	0.8712	0.8824	0.1111	0.1111	0.1111	0.3358	0.3805	0.3422
bious4	0.8583	0.8824	0.4444	0.4444	0.4444	0.2128	0.2759	0.2345
deepseek32b-...	0.8583	0.8824	0.2963	0.2963	0.2963	0.3038	0.3072	0.2955
NN_Persona_1	0.8132	0.8235	0.2222	0.2222	0.2222	0.1631	0.3269	0.1939
NN_Persona_3	0.8132	0.8235	0.2963	0.4074	0.3519	0.1892	0.2919	0.2173
UniTor_0	0.8132	0.8235	0.4444	0.4444	0.4444	0.2765	**0.5081**	0.3296
UniTor_1	0.8132	0.8235	0.4444	0.4444	0.4444	0.2765	**0.5081**	0.3296
gpt 01 mini	0.8132	0.8235	0.0741	0.1111	0.0926	0.1781	0.2299	0.1867
bious1	0.8132	0.8235	0.3704	0.4074	0.3889	0.1868	0.2105	0.1956
dmiip2024_4	0.7984	0.8235	**0.5926**	**0.5926**	**0.5926**	0.2900	0.4419	0.3333
deepseek32b-...	0.7733	0.8235	0.2963	0.2963	0.2963	0.2199	0.3116	0.2390
NN_Persona_2	0.7571	0.7647	0.3333	0.3704	0.3519	0.2033	0.3182	0.2303
NN_Baseline	0.7571	0.7647	0.2963	0.2963	0.2963	0.1964	0.3080	0.2259
dmiip2024	0.7571	0.7647	0.4815	0.5556	0.5185	0.2478	0.3686	0.2792
dmiip2024_1	0.7571	0.7647	0.4815	0.4815	0.4815	0.2598	0.3598	0.2837
deepseek-r1:32b	0.7571	0.7647	0.1111	0.1111	0.1111	0.3270	0.4234	0.3563
bious3	0.7424	0.7647	0.3333	0.4444	0.3889	0.1386	0.1463	0.1355
bious5	0.7424	0.7647	0.3333	0.3704	0.3519	0.1792	0.2560	0.1933
UR-IW-4	0.7018	0.7059	0.5185	0.5556	0.5370	0.2859	0.3652	0.3023
IR2	0.7018	0.7059	0.4444	0.4444	0.4444	0.1938	0.2155	0.1974
bious2	0.7018	0.7059	0.2963	0.3333	0.3148	0.1635	0.2189	0.1829
UniTor_2	0.6886	0.7059	0.3704	0.3704	0.3704	0.2648	0.4433	0.3017
UniTor_3	0.6886	0.7059	0.3704	0.3704	0.3704	0.2648	0.4433	0.3017
Only uses GPT...	0.5882	0.5882	0.1852	0.1852	0.1852	0.1997	0.2261	0.2093
Using LLM...	0.5278	0.5294	0.2593	0.2593	0.2593	0.2013	0.3050	0.2288
lasigeBioTM	0.4632	0.4706	0.2593	0.2593	0.2593	0.3684	0.1058	0.1598
bioinfo-1	0.3929	0.6471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bioinfo-2	0.3929	0.6471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bioinfo-3	0.3929	0.6471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bioinfo-4	0.3929	0.6471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bioinfo-0	0.3929	0.6471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 8
Results for batch 2 for *exact answers* in phase B of task 13b, ranked based on Yes/No F1. Only the top-50 systems and the BioASQ Baseline are presented.

System	Yes/No		Factoid			List		
	F1	Acc.	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
UR-IW-5	**1.0000**	**1.0000**	0.5185	0.5556	0.5370	0.3916	0.6048	0.4463
Fleming-1	**1.0000**	**1.0000**	0.3704	0.6296	0.4704	0.5263	0.5516	0.5210
Fleming-2	**1.0000**	**1.0000**	0.3704	0.6296	0.4704	0.4810	0.6696	0.5356
Mistral7BIns10...	**1.0000**	**1.0000**	0.1481	0.1481	0.1481	0.2370	0.2526	0.2309
GPT4turbo	**1.0000**	**1.0000**	0.5185	0.5926	0.5556	0.5216	0.5657	0.5233
dmiip2024	**1.0000**	**1.0000**	0.5926	0.5926	0.5926	0.5719	0.6561	0.6010
dmiip2024_1	**1.0000**	**1.0000**	0.5926	0.5926	0.5926	0.5615	0.6145	0.5741
dmiip2024_3	**1.0000**	**1.0000**	0.6296	0.6667	0.6481	0.5728	0.5838	0.5683
dmiip2024_4	**1.0000**	**1.0000**	0.5556	0.6296	0.5926	0.6360	0.6315	**0.6152**
dmiip2024_2	**1.0000**	**1.0000**	0.4815	0.4815	0.4815	0.4640	**0.7132**	0.5365
Fleming-3	**1.0000**	**1.0000**	0.3704	0.7407	0.5148	0.5263	0.5516	0.5210
deepseek-r1:8b	**1.0000**	**1.0000**	0.4444	0.4444	0.4444	0.4912	0.5393	0.5000
phaseB-4	**1.0000**	**1.0000**	0.5185	0.5185	0.5185	0.5410	0.6296	0.5408
phaseB-5	**1.0000**	**1.0000**	0.4074	0.4074	0.4074	0.5705	0.5748	0.5412
simple trunc...	**1.0000**	**1.0000**	0.5185	0.6296	0.5679	0.5429	0.6074	0.5510
config-1	**1.0000**	**1.0000**	0.5185	0.5185	0.5185	0.5140	0.5026	0.5030
config-4	**1.0000**	**1.0000**	0.5926	0.6296	0.6111	0.4784	0.5385	0.4851
config-5	**1.0000**	**1.0000**	0.5926	0.5926	0.5926	0.5412	0.5801	0.5444
llama	**1.0000**	**1.0000**	0.4444	0.5185	0.4815	0.5004	0.5021	0.4832
2025-DMIS...	**1.0000**	**1.0000**	0.5556	0.5556	0.5556	0.5624	0.5701	0.5522
EP-4	**1.0000**	**1.0000**	0.4815	0.4815	0.4815	0.5412	0.5915	0.5362
IISR first submit	0.9377	0.9412	0.5185	0.5556	0.5370	0.4961	0.5209	0.4954
IISR 2nd submit	0.9377	0.9412	0.4444	0.4815	0.4630	0.4575	0.4522	0.4441
IISR 3rd submit	0.9377	0.9412	0.5185	0.5185	0.5185	0.5263	0.5626	0.5281
IISR 4th submit	0.9377	0.9412	0.4444	0.4444	0.4444	0.5905	0.5904	0.5800
bious1	0.9377	0.9412	0.3704	0.4074	0.3889	0.3393	0.4146	0.3582
GPT4O	0.9377	0.9412	0.3704	0.3704	0.3704	0.4967	0.4779	0.4666
deepseek32b...	0.9377	0.9412	0.6667	0.6667	0.6667	0.4718	0.5073	0.4723
deepseek32b-f	0.9377	0.9412	0.5926	0.5926	0.5926	0.4744	0.5336	0.4846
mistral	0.9377	0.9412	0.4815	0.6296	0.5556	0.5174	0.5801	0.5264
dense	0.9377	0.9412	0.5556	0.5556	0.5556	0.5192	0.5516	0.5250
2025-DMIS...	0.9377	0.9412	0.5185	0.5556	0.5370	0.5624	0.5701	0.5522
2025-DMIS...	0.9377	0.9412	0.5556	**0.7778**	0.6481	0.5599	0.5789	0.5545
EP-5	0.9377	0.9412	0.5185	0.5185	0.5185	0.6044	0.5415	0.5538
UR-IW-4	0.9328	0.9412	0.5185	0.5926	0.5556	0.4610	0.6425	0.5188
UniTor_0	0.9328	0.9412	0.6667	0.6667	0.6667	0.4070	0.5757	0.4487
UniTor_1	0.9328	0.9412	0.6667	0.6667	0.6667	0.4070	0.5757	0.4487
UniTor_2	0.9328	0.9412	**0.7037**	0.7037	**0.7037**	0.3462	0.4717	0.3807
UniTor_3	0.9328	0.9412	**0.7037**	0.7037	**0.7037**	0.3462	0.4717	0.3807
bious3	0.9328	0.9412	0.4444	0.5185	0.4815	0.4392	0.5170	0.4669
bious4	0.9328	0.9412	0.3704	0.4074	0.3889	0.3713	0.4337	0.3954
bious5	0.9328	0.9412	0.4444	0.4444	0.4444	0.3809	0.5026	0.4195
gpt 01 mini	0.9328	0.9412	0.1111	0.1111	0.1111	0.2791	0.2671	0.2493
config-2	0.9328	0.9412	0.5926	0.5926	0.5926	0.4754	0.5363	0.4943
config-3	0.9328	0.9412	0.5926	0.5926	0.5926	0.4754	0.5363	0.4943
2025-DMIS...	0.9328	0.9412	0.5185	0.5926	0.5556	0.5594	0.5723	0.5513
UR-IW-3	0.8786	0.8824	0.5185	0.5556	0.5309	0.4312	0.6771	0.5010
deepseek32b...	0.8786	0.8824	0.5556	0.5556	0.5556	0.5079	0.5770	0.5182
2025-DMIS...	0.8786	0.8824	0.5926	**0.7778**	0.6667	0.5477	0.6184	0.5670
kmeans	0.8786	0.8824	0.4815	0.5926	0.5370	0.4711	0.5008	0.4637
BioASQ_Baseline	0.4632	0.4706	0.1852	0.4444	0.2772	0.2693	0.3828	0.2528

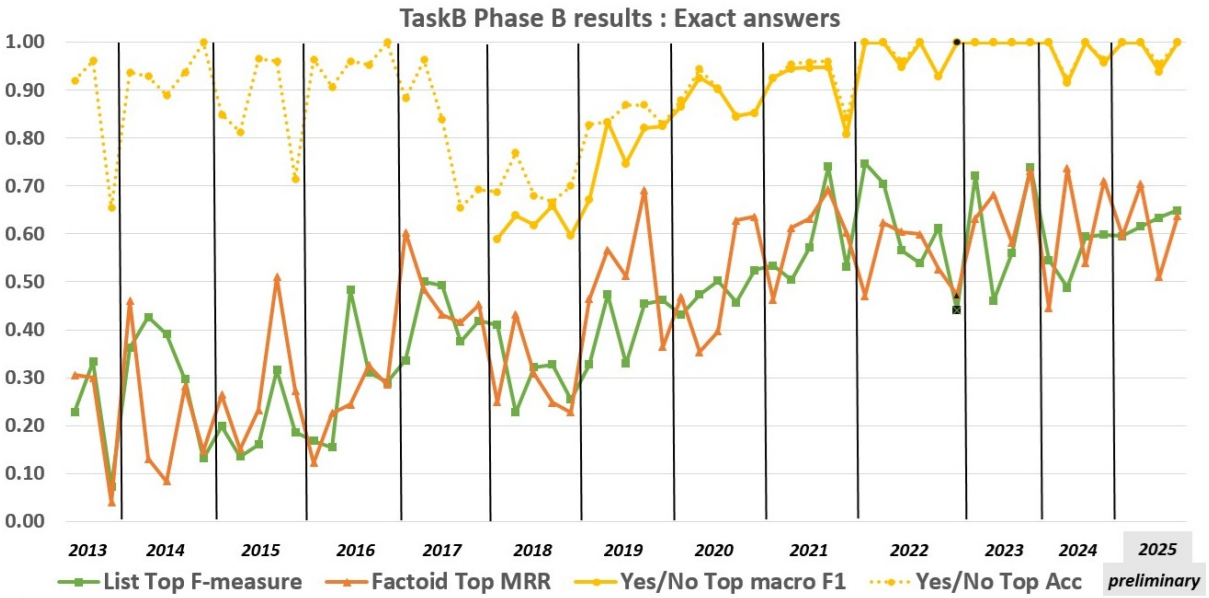

Figure 4: The evaluation scores of the best-performing systems in task B, Phase B, for *exact answers*, across the twelve years of BioASQ. Since BioASQ6 [49], accuracy (Acc) was replaced by macro F1 as the official measure for Yes/No questions. The black dot in BioASQ10 [50] indicates an additional batch consisting of questions from new experts exclusively [48].

4.2. Task Synergy13

In task Synergy13, no relevant material was initially available for new questions. For old questions, however, feedback from previous rounds was provided per question, that is the documents and snippets submitted by the participants with manual annotations of their relevance. Hence, the documents and snippets of the feedback, that have already been assessed and released, were not considered valid for submission in the subsequent rounds. As in task 13b, the evaluation measures for document and snippet retrieval are MAP and F-measure respectively.

In addition, due to the developing nature of the topic, no answer is available for all of the open questions in each round. Therefore only the questions indicated as "answer ready" were evaluated for *exact* and *ideal answers* in each round. Regarding the *ideal answers*, the systems were ranked according to manual scores assigned to them by the BioASQ experts during the assessment of systems responses as in phase B of task B [7]. As regards evaluation for the *exact answers*, similarly to task 13b, the mean F1 measure, the Mean Reciprocal Rank (MRR), and the macro F1 measure are used for the ranking in list, factoid, and yes/no questions respectively. Any *exact* or *ideal answer* that was assessed as 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 Synergy13 results are available online[6]. Overall, the collaboration between participating biomedical experts and question-answering systems allowed the progressive identification of relevant material and extraction of *exact* and *ideal answers* for several open questions for developing problems, such as infectious, rare, and genetic diseases, and women's and reproductive health. In total, after the four rounds of Synergy13, enough relevant material was identified to provide an answer to about 80% of the questions. In addition, about 47% of the questions had at least one *ideal answer*, submitted by the systems, which was considered of ground-truth quality by the respective expert.

[6] http://participants-area.bioasq.org/results/synergy_v2025/

Table 9
Results for document retrieval of the first round of the Synergy13 task.

System	Mean precision	Mean Recall	Mean F-Measure	MAP	GMAP
dmiip2024	0.3275	0.4936	0.2899	**0.4060**	0.0436
dmiip2024_1	**0.3397**	**0.4944**	**0.2906**	0.4051	0.0497
dmiip2024_3	0.2410	0.4860	0.2743	0.3969	**0.0645**
dmiip2024_2	0.3217	0.4921	0.2848	0.3960	0.0423
dmiip2024_4	0.3051	0.4687	0.2703	0.3844	0.0309
Fleming-2	0.2238	0.3375	0.2121	0.2828	0.0076
Fleming-1	0.2426	0.3939	0.2403	0.2172	0.0142
SCIRE1 Results	0.3227	0.2072	0.2249	0.1928	0.0011

5. Conclusions

In this paper, we introduced the thirteenth version of the BioASQ challenge, focusing on the question answering tasks b and Synergy. These tasks have been well-established through previous versions of the challenge and remain timely and relevant, as indicated by the increased participation.

The preliminary results of task 13b reveal the strong performance of 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. System performance on list and factoid questions was more variable, especially in Phase A+, highlighting the presence of room for improvement. These results suggest that access to ground truth relevant material significantly enhances response quality for more complex question types. This emphasizes the critical role of Phase A, which involves the automatic retrieval of relevant documents and snippets for answering a biomedical question. Performance in Phase A showed greater variability across batches, potentially influenced by the domain expertise of the experts who authored the questions. A diverse set of retrieval and generation strategies was employed, including traditional IR methods, large language model (LLM)-based approaches, and systems enriched with domain-specific biomedical knowledge. Finally, the results of Synergy13 highlight that state-of-the-art QA systems can be useful in aiding biomedical researchers to address their specialized information needs, in alignment with the results of previous versions of the task, despite the persisting challenges and room for improvement.

Overall, several participating systems achieved competitive performance on the QA tasks of BioASQ 13, and some of them managed to improve over the state-of-the-art performance from previous years. Therefore, twelve years after its initial introduction, BioASQ keeps pushing the research frontier in biomedical question answering, offering two QA tasks and several response types that cover a range of biomedical information needs.

Acknowledgments

The thirteenth edition of BioASQ is sponsored by Ovid, Atypon Systems Inc, and Elsevier. 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 13b.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and ChatGPT to perform the following tasks: grammar and spelling checks, Paraphrasing, and rewording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, 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.

[2] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results., in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[3] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-L Task on Biomedical Nested Entity Linking in CLEF 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[4] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Kipouros, D. Papadopoulos, A. Bekiaridou, K. Barmpagiannos, A. Vasilopoulou, A. Barmpagiannos, A. Samaras, G. Giannakoulas, G. Tsoumakas, Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[5] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[6] 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. doi:10.1186/s12859-015-0564-6.

[7] 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.

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

[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, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), ????

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Z. Y. Y. Z. E. Nyberg, Learning to Answer Biomedical Questions: OAQA at BioASQ 4B, ACL 2016 (2016) 23.

[22] 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).

[23] 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.

[24] R. A. A. Jonker, T. Almeida, J. Almeida, S. Matos, BIT.UA at BioASQ 13B: Revisiting Evaluation, DPRF-Enhanced Retrieval and Fine-Tuned LLMs, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[25] S. Ateia, U. Kruschwitz, Can Language Models Critique Themselves? Investigating Self-Feedback for Retrieval Augmented Generation at BioASQ 2025 , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[26] C. Bing-Chen, J.-C. Han, H.-C. Hung, R. T.-H. Tsai, NCU-IISR: Biomedical Question Answering via

Gemini and GPT APIs in the BioASQ 13b Phase B Challenge , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[27] J.-C. Han, B.-C. Chih, H.-C. Hung, R. T.-H. Tsai, A Retrieval-Augmented Generation Approach for BioASQ 13b Phase A and A+ , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[28] D. Panou, A. Dimopoulos, M. Koubarakis, M. Reczko, Harnessing Collective Intelligence of LLMs for Robust Biomedical QA: A Multi-Model Approach , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[29] P. Vachharajani, Exploring Retrieval-Reranking and LLM-Based Answer Generation for Biomedical QA , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[30] H. P. Gupta, R. Banerjee, LLMs for Biomedical NER , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[31] F. Borazio, D. Croce, R. Basili, UniTor at BioASQ 2025: Modular Biomedical QA with Synthetic Snippets and Multiple Task Answer Generation , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[32] A. Morillo, E. Puertas, J. C. M. Santos, J. S. Castaneda, C. A. Palomino, VerbanexAI at BioASQ 13B: PubMed API and LLM-Driven Hybrid Retrieval for Biomedical Question Answering , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[33] P. R. C. Lopes, S. I. R. Conceição, M. Fernandes, F. M. Couto, lasigeBioTM: A lean biomedical QA system empowered by structured knowledge , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[34] J. Angulo, V. Yeste, AQAMS and AQAMS2: Multi Agent Systems for Biomedical Question Answering , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[35] J. Tang, H. Yang, K. Xiong, H. Li, P. Quaresma, H. Yu, W. Zhang, M. Song, Y. Jiang, Applying DeepSeek to BioASQ Task 13B: Using Supervised Fine-Tuning and Few-Shot Learning , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[36] S. Verma, F. Jiang, X. Xue, Beyond Retrieval: Ensembling Cross-Encoders and GPT Rerankers with LLMs for Biomedical QA , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[37] D. Galat, D. Molla-Aliod, LLM Ensemble for RAG: Role of context length in zero-shot Question Answering for BioASQ Challenge , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[38] E. Quiñones, E. A. P. D. Castillo, Evaluation of a System for Generating Exact and Ideal Responses , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[39] D. Stachura, J. Konieczna, A. Nowak, Are Smaller Open-Weight LLMs Closing the Gap to Proprietary Models for Biomedical Question Answering? , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[40] H. Kim, H. Lee, Y. Cho, J. Park, J. Park, S. Park, Y. T. Chok, S. Baek, D. Lee, J. Kang, Prompting Matters: Snippet-Aware Strategies for Biomedical QA with LLMs in BioASQ 13b , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[41] T. Almeida, R. A. A. Jonker, R. Poudel, J. M. Silva, S. Matos, Bit. ua at bioasq 11b: Two-stage ir with synthetic training and zero-shot answer generation., in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[42] T. Almeida, R. A. Jonker, J. Reis, J. R. Almeida, S. Matos, Bit. ua at bioasq 12: From retrieval to answer generation, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[43] B.-C. Chih, J.-C. Han, R. Tzong-Han Tsai, NCU-IISR: Enhancing Biomedical Question Answering with GPT-4 and Retrieval Augmented Generation in BioASQ 12b Phase B, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[44] D. Panou, A. Dimopoulos, M. Reczko, Farming Open LLMs for Biomedical Question Answering, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes,

2024.

[45] S. D. Romero, L. A. Ureña-López, E. Martínez-Cámara, SINAI at CLEF 2025: A Multi-Stage RAG Pipeline for Biomedical Semantic Question Answering , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[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: 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.