

Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering

Anastasios Nentidis¹, Georgios Katsimpras¹, Anastasia Krithara¹, Martin Krallinger², Miguel Rodriguez-Ortega², Eduard Rodriguez², Natalia Loukachevitch³, Andrey Sakhovskiy^{5,6}, Elena Tutubalina^{4,5}, Dimitris Dimitriadis⁷, Grigorios Tsoumakas⁷, George Giannakoulas⁷, Alexandra Bekiaridou⁸, Athanasios Samaras⁷, Giorgio Maria Di Nunzio⁹, Nicola Ferro⁹, Stefano Marchesin⁹, Marco Martinelli⁹, Gianmaria Silvello⁹, and Georgios Paliouras¹

¹ National Center for Scientific Research “Demokritos”, Athens, Greece
{tasosnent, gkatsibras, akrithara, paliourg}@iit.demokritos.gr

² Barcelona Supercomputing Center, Barcelona, Spain
{martin.krallinger, miguel.rodriguez, eduard.rodriguez}@bsc.es

³ Moscow State University, Russia
louk_nat@mail.ru

⁴ Artificial Intelligence Research Institute, Russia

⁵ Kazan Federal University, Russia

⁶ SberAI & Skoltech, Russia
{andrey.sakhovskiy, tutubalinaev}@gmail.com

⁷ Aristotle University of Thessaloniki, Greece
{dndimitri,greg}@csd.auth.gr, {g.giannakoulas, th.samaras.as}@gmail.com

⁸ Northwell Health, USA
ampekiaridou@gmail.com

⁹ University of Padua, Italy
{name.surname}@unipd.it

Abstract. This is an overview of the thirteenth edition of the BioASQ challenge in the context of the Conference and Labs of the Evaluation Forum (CLEF) 2025. BioASQ is a series of international challenges promoting advances in large-scale biomedical semantic indexing and question answering. This year, BioASQ consisted of new editions of the two established tasks, b and Synergy, and four new tasks: a) *Task Multi-ClinSum* on multilingual clinical summarization. b) *Task BioNNE-L* on nested named entity linking in Russian and English. c) *Task ELCardioCC* on clinical coding in cardiology. d) *Task GutBrainIE* on gut-brain interplay information extraction. In this edition of BioASQ, 76 competing teams participated with more than 1000 distinct submissions in total for the six different shared tasks of the challenge. Similarly to previous editions, most of the participating systems achieved competitive performance, suggesting the continuous advancement of the state-of-the-art in the field.

Keywords: Biomedical knowledge · Semantic Indexing · Question Answering

1 Introduction

The BioASQ challenge has been focusing on the advancement of the state-of-the-art in large-scale biomedical semantic indexing and question answering (QA) for more than 10 years [42]. To this end, it organizes different shared tasks annually, developing respective benchmark datasets that represent the real information needs of experts in the biomedical domain. This allows the participating teams from around the world, who work on the development of systems for biomedical semantic indexing and question answering, to benefit from the publicly available datasets, evaluation infrastructure, and exchange of ideas in the context of the BioASQ challenge and workshop.

Here, we present the shared tasks and the datasets of the thirteenth BioASQ challenge in 2025, as well as an overview of the participating systems and their performance. The remainder of this paper is organized as follows. First, Section 2 presents a general description of the shared tasks, which took place from January to May 2025, and the corresponding datasets developed for the challenge. Then, Section 3 provides a brief overview of the participating systems for the different tasks. Detailed descriptions for some of the systems are available in the respective extended overviews of each task and the proceedings of the BioASQ lab. Subsequently, in Section 4, we present the performance of the systems for each task, based on state-of-the-art evaluation measures or manual assessment. Finally, in Section 5 we draw some conclusions regarding the 2024 edition of the BioASQ challenge.

2 Overview of the tasks

The thirteenth edition of the BioASQ challenge consisted of six tasks [28]: (i) *Task b* on biomedical semantic question answering. (ii) *Task Synergy* on question answering developing biomedical topics. (iii) *Task MultiClinSum* on multilingual clinical summarization. (iv) *Task BioNNE-L* on nested named entity linking in Russian and English. (v) *Task ELCardioCC* on clinical coding in cardiology. (vi) *Task GutBrainIE* on gut-brain interplay information extraction. In this section, we first describe this year’s editions of the two established tasks b (task 13b) and Synergy (Synergy 13) [31] with a focus on differences from previous editions of the challenge [29, 33]. Additionally, we also introduce the four new BioASQ tasks, MultiClinSum [21], BioNNE-L [8], ELCardioCC, and GutBrainIE.

2.1 Task 13b

BioASQ *task 13b* is the thirteenth edition of the established BioASQ *task b* on Biomedical QA. This year, it took place in three phases: i) Phase A: biomedical

questions in English were provided, and the systems had to retrieve relevant material (PubMed documents and snippets). ii) Phase A+, the systems had to provide ‘exact’ and ‘ideal’ answers. Depending on question type, the ‘exact’ answer can be a *yes* or *no* (yes/no), an entity name, such as a disease or gene (factoid), or a list of entity names (list). The ‘ideal’ answer is a paragraph-sized summary, regardless of question type. ii) Phase B: Some relevant material was provided for each question, selected by the BioASQ experts, and the systems had to provide new answers given this additional information.

About 340 new biomedical questions annotated with golden documents, snippets, and answers (‘exact’ and ‘ideal’), were developed for testing. In addition, a training set of 5,389 biomedical questions, accompanied by answers, and supporting evidence (documents and snippets), was available from previous versions of the tasks, as a unique resource for the development of question-answering systems [?]. Table 1 presents some statistics of both training and test datasets for task 13b. The test data for task 13b were split into four independent bi-weekly batches consisting of 85 questions each, as presented in Table 1.

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	5389	1459	1047	1600	1283	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	5729	1541	1130	1695	1363	9.33	12.23

2.2 Task Synergy 13

BioASQ *task Synergy* was originally introduced in 2020 with the aim of promoting research in developing biomedical topics, such as COVID-19 [16, ?]. The design of this task as an ongoing dialogue allows experts to pose open-ended questions for developing topics, for which they do not know in advance whether a definitive answer can be given, in order to obtain relevant material (documents and excerpts) retrieved from the systems. After assessing this material, they provide feedback to the systems on its relevance and on whether it is sufficient to answer their question, by marking respective questions as *ready to answer*. This process is repeated iteratively in rounds with new material considered in each round, based on updates to the original document resource [32]¹⁰. For *ready to answer* questions, they receive exact and ideal answers as well, assess them, and provide feedback that can be used by the systems to improve their responses to

¹⁰ As of 2023, this evolving document resource is PubMed [?]

these questions in the remaining rounds. The experts can also mark a question as *closed* if they receive a fully satisfactory answer that is not expected to change or if they are no longer interested in the question.

A training dataset of 366 questions on developing topics with incremental annotations with relevant material and answers is already available from previous versions of *task Synergy* [?, ?, ?, ?]. During the *task Synergy 13*, this set was extended with 47 new questions on developing health topics, such as infectious, rare, and genetic diseases, and women’s and reproductive health. Meanwhile, 27 questions from the previous version of the task remained open and were enriched with more recent evidence and updated answers. Overall, 74 questions were considered in the four rounds of *task Synergy 13*. The number of yesno, list, factoid, and summary questions was 23, 19, 14, and 18, respectively.

2.3 Task MultiClinSum

There is a rapid accumulation of different types of clinical content, including medical records and clinical case reports. These are generally written not only in English but in a variety of languages. Some of the clinical reports can be very long and thus it is challenging for domain experts to read and keep track of key clinical insights. Large Language Models (LLMs) have shown promising results for summarization approaches that can be useful to condense lengthy clinical cases into a shorter version of text, reducing the size of the initial text while preserving key clinical informational elements. Thus there is a pressing need to evaluate and benchmark how well clinical summarization works for case reports written in different languages.

We introduce the *MultiClinSum* task covering the automatic summarization of lengthy clinical case reports written in different languages, namely English, Spanish, French, and Portuguese. The *MultiClinSum* task will rely on a corpus of manually selected full clinical case reports and their corresponding clinical case report summaries derived from case report publications written in the previously mentioned languages. For evaluation purposes, automatically generated summaries will be compared against manually generated summaries generated by the original authors, exploring Rouge-2 scores and BERTScore [?] for evaluation assessment. As clinical case reports do share commonalities with medical discharge summaries, insights provided by the *MultiClinSum* results can be of practical relevance also for clinical records summarization scenarios.

As mentioned, the manually selected dataset consists of full text-summary pairs extracted from clinical case reports. The total number of pairs in English is 1,280, in Spanish 534, in Portuguese 108, and in French 54. In order to increase the size of the dataset, each of the pairs of one language has been translated into the other languages, resulting in a total of 1,976 pairs for each language. An additional large-scale dataset has also been created from the clinical cases published in the PMC-Patients subset. In this subset, patients’ clinical cases are presented in a condensed format, making them ideal for use as summaries of the case. With regard to the full text of the case, manually selected headers from the corresponding PubMed abstracts have been extracted for that purpose.

Table 2. Statistics for the datasets provided for MultiCardioNER. “Annot.” stands for “annotations”, while “Chars” stands for “characters”. Unique annotations refer to the number of distinct annotated strings after converting all annotations to lowercase. The number of tokens has been calculated using the following spaCy models: “es_core_news_sm”, “en_core_web_sm” and “it_core_news_sm”.

Dataset	Lang.	Entity	Docs	Tokens	Chars	Annot.	Unique Annot.
DisTEMIST	ES	Diseases	1,000	406,137	2,335,968	10,664	6,739
DrugTEMIST	ES	Drugs	1,000	406,137	2,335,968	2,778	925
	EN	Drugs	1,000	404,194	2,230,631	2,814	875
	IT	Drugs	1,000	421,251	2,393,002	2,808	893
CardioCCC	ES	Diseases	508	568,297	3,215,774	18,232	7,692
	ES	Drugs	508	568,297	3,215,774	4,227	755
	EN	Drugs	508	576,772	3,114,833	4,231	734
	IT	Drugs	508	595,332	3,345,466	4,385	752

2.4 Task BioNNE-L

In the BioNNE-L Shared Task, we address the medical entity linking task, also known as Medical Concept Normalization (MCN), which is to map given entities to the most relevant vocabular entries from an external source, e.g., concepts from the UMLS metathesaurus identified with concept unique identifiers (CUIs). Although the task has been widely explored in recent years, existing approaches usually treat each entity individually, medical entities often form a nested structure, where an entity can be a subpart of another entity. One of the key features of BioNNE-L is the focus on nested entities that are (i) derived from the MCN annotation of the NEREL-BIO corpus [24, 25] and (ii) supplemented by newly annotated data in both English and Russian. The annotated entity types are disorders (DISO), anatomical structures (ANAT), and chemicals (CHEM). The competition was organized into three subtasks that fell under two evaluation tracks: 1. **Monolingual track** that treated English and Russian data independently; 2. **Bilingual track** that required a single bilingual model for the combined Russian and English data. Data statistics for both tracks is summarized in Table 3.

All BioNNE-L materials can be found on the shared task’s GitHub¹¹ and Codalab pages¹². Annotated data and normalization vocabulary is also available at HuggingFace¹³.

2.5 Task ELCardioCC

Cardiovascular diseases affect a significant portion of the global population, accounting for 32% of global deaths according to WHO¹⁴. Automated clinical

¹¹ <https://github.com/nerel-ds/NEREL-BIO/tree/master/BioNNE-L.Shared.Task>

¹² <https://codalab.lisn.upsaclay.fr/competitions/21568>

¹³ <https://huggingface.co/datasets/andorei/BioNNE-L>

¹⁴ <https://www.who.int/health-topics/cardiovascular-diseases>

Table 3. BioNNE-L 2025 statistics for Disorder (**DISO**), Chemical (**CHEM**), and Anatomical Structure (**ANAT**) among Russian and English entities as well as vocabulary statistics.

Entity type	Refined NEREL-BIO				Novel data		Vocabulary	
	Train		Dev		Test			
	Ru	En	Ru	En	Ru	En	Ru	En
# documents	716	54	50	50	154	154	—	—
Number of entities								
DISO	11,168	1,200	925	1,029	2,811	3,068	91,867	1,825,048
CHEM	4,741	579	531	564	1,218	1,345	47037	1,732,096
ANAT	8,346	911	878	901	2,186	2,248	6899	345,043
	24,255	2,690	2,334	2,494	6,215	6,661	145,803	3,902,187

coding plays a crucial role in transforming unstructured real-world medical data gathered from patients into structured information, in order to facilitate clinical research and analysis. However, existing research predominantly focuses on English clinical text, leaving other languages, such as Greek, underrepresented. To this end, we propose a new *ELCardioCC* task, which concerns i) the assignment of cardiology-related ICD-10 codes to discharge letters from Greek hospitals, ii) the extraction of the specific mentions of ICD-10 codes from the discharge letters.

In detail, the participants in the *ELCardioCC* task were tasked with developing named entity recognition (NER), entity linking (EL) and multi-label classification - explainable AI (MLC-X) systems using a specialized corpus of discharge letters. These discharge letters, which were written in Greek contained valuable medical information about patients’ conditions, treatments, and outcomes. The corpus was meticulously annotated with the positions of mentions (such as chief complaint, diagnosis, prior medical history, drugs and cardiac echo) and their corresponding ICD-10 codes. The training dataset includes 1,000 discharge letters, while the test set comprises 500 letters. System performance was evaluated using the micro F1 score.

2.6 Task GutBrainIE

Recent scientific evidence suggests a connection between *brain-related diseases* and the *gut microbiota* that may play a critical role in mental health-related disorders or diseases like Parkinson’s, and Alzheimer’s [2, 5, 7, 12]. The scientific literature on this topic is rapidly expanding, making it increasingly challenging for clinicians and researchers to stay up to date. For example, in 2020, approximately 200 articles were published on the relationship between gut microbiota and mental health; by 2024, this number had more than doubled to over 450 publications. The *GutBrainIE* task aims to foster the development of Information Extraction (IE) systems that support experts by automatically extracting

and linking knowledge from biomedical abstracts, facilitating the understanding of gut-brain interplay and its role in mental health and neurological diseases.

The *GutBrainIE* task comprises four subtasks of increasing difficulty: Named Entity Recognition (NER), which identifies and classifies entity mentions in PubMed abstracts about the gut-brain interplay focusing on mental health and the Parkinson’s disease; Binary Tag-based Relation Extraction (BT-RE), which detects whether pairs of entities are in relation without specifying the relation type; Ternary Tag-based Relation Extraction (TT-RE), which extends BT-RE by also assigning a relation label to each related pair; and Ternary Mention-based Relation Extraction (TM-RE), which further localizes the exact entity mentions involved in each relation and assigns the appropriate relation label.

The dataset includes over 1000 documents with annotated entity mentions and relations, organized into Training, Development, and Test sets. The train set is further divided into four quality tiers: expert-curated (Platinum), expert-annotated (Gold), student-annotated (Silver), and automatically generated (Bronze). Development and Test sets contain only expert annotations (Platinum+Gold).

Table 4. Dataset statistics for *GutBrainIE*.

Collection	# Docs	# Entities	Ents/Doc	# Rels	Rels/Doc
Train Platinum	111	3638	32.77	1455	13.11
Train Gold	208	5192	24.96	1994	9.59
Train Silver	499	15275	30.61	10616	21.27
Train Bronze	749	21357	28.51	8165	11.90
Development Set	40	1117	27.93	623	15.58
Test Set	40	1237	30.92	777	19.42

3 Overview of participation

Overall, 76 distinct teams participated in the thirteenth edition of the BioASQ challenge, submitting more than 1000 distinct runs for the six different shared tasks of the challenge.

3.1 Task 13b

This year, 46 teams participated in task 13b, submitting a total of 734 different submissions generated by 146 distinct systems across all four batches for the 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 in previous years, the open-source system OAQA [43], which achieved top performance in older editions of BioASQ, was used as a baseline for phase B *exact answers*.

The “*ITMO*” team participated in both phases of the task experimenting in its “pa” systems with differing solutions across the batches. In general, for

document retrieval the systems follow an approach with two stages. First, they identify initial candidate articles based on BM25, and then they re-rank them using variations of BERT [9], fine-tuned for the binary classification task with the BioASQ dataset and pseudo-negative documents. They extract snippets from the top documents and rerank them using biomedical Word2Vec based on cosine similarity with the question. A more detailed overview of the methods developed by the participating teams is available in the extended overview of *task 13b* and *Synergy 13*.

3.2 Task Synergy 13

In the thirteenth edition of BioASQ, five teams participated in the Synergy task (Synergy 13). 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 Synergy 13. In particular, the BSRC Alexander Fleming team participated with four systems. Similar to task b, their systems focused on LLMs with optimized prompts and majority voting. Also, the UR team from the Universität Regensburg competed with two systems. Their systems employed 2-shot and zero-shot learning with GPT-3.5-turbo and GPT-4. Furthermore, their systems utilized handcrafted examples for the 2-shot learning as well as incorporated query expansion methods. The Gatech team from the Georgia Institute of Technology participated in five systems. As with task b, their systems relied on pre-trained LLMs and prompt engineering. More detailed descriptions for some of the systems are available at the proceedings of the workshop.

A more detailed overview of the methods developed by the participating teams is available in the extended overview of *task 13b* and *Synergy 13*.

3.3 Task MultiClinSum

31 teams registered for the MultiCardioNER task, out of which 7 teams submitted at least one run of their predictions. Specifically, 6 teams participated in the CardioDis subtrack, while 5 teams participated in the MultiDrug subtrack (with one of those teams participating only in the Spanish part). Overall, a total of 70 runs were submitted, with each team being allowed up to 5 runs per subtrack and language.

Table ?? gives an overview of the methodologies used by the participants in each of the sub-tasks. Following the trend of previous similar shared tasks [20, 19, 18], all participants used some variant of Transformers-based models, with RoBERTa [23] models being the most popular. Other than that, ensembles were quite popular and provided good results (e.g. BIT.UA [15]), as were the use of custom datasets and dictionaries (e.g. Enigma [1]), data augmentation or window sliding (e.g. Data Science TUW [39]).

3.4 Task BioNNE-L

In total, we’ve received 23 Codalab registrations for the BioNNE-L task, with 7 teams submitting predictions during the evaluation phase. The systems submitted by the participants are summarized in Table 5.

Table 5. Overview of the approaches presented by participants for the BioNNE task. EN stands for the English-oriented and RU for the Russian-oriented tracks.

Team	Track	Approach
verbanexialab	EN	SapBERT w/ lexical and semantic reranking
LYX_DMIIP_FDU	Bilingual,EN,RU	BERGAMOT fine-tuning
BlancaPlanca	Bilingual,EN,RU	BERGAMOT w/ language-specific preprocessing
EeyoreLee	Bilingual,EN,RU	Two-step retrieval and ranking pipeline
dstepakov	Bilingual,RU	RoBERTa fine-tuning with contrastive learning
ICUE	Bilingual,EN,RU	BERT, BioSyn, LLM 0-shot reranking
NLPIMP	Bilingual	Russian LaBSE model pre-trained on medical data

Team **verbanexialab** leveraged a SapBERT [22], pre-trained on UMLS concepts, to obtain entity embeddings, followed by a multicomponent re-ranking. They combined embedding cosine similarity with Jaccard similarity for lexical overlap recognition and Levenshtein distance for character-level alignment.

Team **LYX_DMIIP_FDU** fine-tuned a BERGAMOT [37] model for each task via contrastive learning using the train- and dev-set entities to enrich the original vocabularies. The textual context of each entity was used as additional input to enhance the entity representation.

Team **BlancaPlanca** used BERGAMOT for zero-shot retrieval based on entity-concept cosine similarity. They apply language-specific lemmatization for Russian and speed up the inference by chunking the original vocabulary into type-specific parts of 100k entries each.

Team **dstepakov** performed the nearest-neighbor search based on the cosine similarity of RoBERTa embeddings [23], fine-tuned contrastively on anchor-positive-negative term triplets via the InfoNCE objective [35].

Team **ICUE** fine-tuned BioSyn [40] using the vocabularies reduced to less than 100k entries each. They fine-tune a separate BERT-based model [10] for English [3], Russian¹⁵, and multilingual [41] tracks, respectively.

Team **NLPIMP** performed the zero-shot ranking using a Russian LaBSE [?] model¹⁶ pre-trained contrastively on an in-house Russian medical corpus.

¹⁵ <https://huggingface.co/KoichiYasuoka/bert-base-russian-upos>

¹⁶ <https://huggingface.co/sergeyzh/LaBSE-ru-turbo>

3.5 Task ELCardioCC

Five teams participated in the *ELCardioCC* task, with each team permitted to submit up to five systems. In the Named Entity Recognition (NER) subtask, four teams took part, submitting a total of 14 systems in addition to our baseline model. The Entity Linking (EL) subtask also saw participation from four teams, who submitted 14 systems alongside the provided baseline. For the Multi-label Classification and Explainable AI (MLC-X) subtask, four teams contributed systems for the multi-label classification component, while one team submitted a system for the explainable AI component. We provided baseline models for both parts of the MLC-X subtask. In total, 13 systems were submitted for MLC-X.

3.6 Task GutBrainIE

The *GutBrainIE* task registered 17 teams submitting runs. Among these, 16 teams participated in NER, 12 in BT-RE, 13 in TT-RE, and 13 in TM-RE. Overall, a total of 391 runs were submitted: 101 for NER, 100 for BT-RE, and 95 for both TT-RE and TM-RE.

Most teams adopted supervised fine-tuning or transformer-based models pre-trained on biomedical text for the NER task. Standard backbones included PubMedBERT, BioBERT, BioLinkBERT, and ELECTRA [6, 13, 17, 44]. Specialized NER architectures, such as GLiNER [45], were also utilized and fine-tuned. Many groups trained multiple models with different random seeds to boost robustness and ensembled their outputs. All teams used platinum, gold, and silver collections for training. A few also used the noisier bronze set, employing cleaning or re-weighting approaches and integrating PubMed data augmentation.

Across the RE subtasks, participants primarily used biomedical pre-trained language models fine-tuned on entity-marker augmented inputs. Among these, the most widely employed include: SapBERT, PubMedBERT, BioBERT, RoBERTa, and ELECTRA [22, 13, 17, 23, 6]. Several teams reformulated RE as a seq2seq problem using REBEL-large [14], directly generating relation tuples or tagged spans in a single pass for all three subtasks. Few others leveraged model ensembling and trained with negative-pair subsampling to counter class imbalance and increase generalization capabilities. Finally, some groups experimented with few-shot or Retrieval-Augmented Generation (RAG), prompting large language models to extract relations with minimal fine-tuning (e.g., for the BT-RE task).

4 Results

4.1 Task 13b

This section presents the evaluation measures and preliminary results for Task 13b. The evaluation in *task 13b* is done manually by the experts that assess each system response and automatically by employing a variety of established evaluation measures [26] as in the previous versions of the task [30]. Table 6 provides a brief overview of the official measures per response and question

type. The results reported for *task 13b* are preliminary, as the final results will be available after the manual assessment of all system responses by the BioASQ team of experts and the enrichment of the ground truth with potential additional relevant items (i.e. documents and snippets), answer elements, and/or synonyms, which is still in progress. The online results pages for Phase A¹⁷, Phase A+¹⁸, and Phase B¹⁹ will be updated with the final results when available.

Table 6. The evaluation measures for *task 13b* per response type and question type [26].

Resp. type (Phase)	Quest. type	Official measure
Documents (A)	All	Mean Average Precision (MAP)
Snippets (A)	All	F-measure (based on character overlaps)
	List	F-measure
Exact ans. (A+ & B)	Yesno Factoid	macro F-measure on “yes” & “no” classes Mean Reciprocal Rank (MRR)
Ideal ans. (A+ & B)	All	Manual scores for precision, recall, repetition, readability

Table 7. Preliminary results for document retrieval in batch 1 of phase A of task 13b. Only the top 5 systems are presented, based on MAP.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
bioinfo-4	0.1039	0.3124	0.1485	0.2067	0.0016
bioinfo-3	0.1009	0.3047	0.1444	0.2024	0.0013
bioinfo-1	0.1156	0.3171	0.1581	0.2018	0.0015
bioinfo-2	0.1294	0.3369	0.1728	0.2006	0.0019
bioinfo-0	0.1151	0.3055	0.1570	0.1800	0.0009

Table 8. Preliminary results for snippet retrieval in batch 1 of phase A of task 13b. Only the top-6 systems are presented, based on F-measure.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
dmiip2024_2	0.0446	0.1490	0.0638	0.1149	0.0001
dmiip2024_3	0.0482	0.1645	0.0684	0.1062	0.0002
dmiip2024	0.0485	0.1645	0.0691	0.1061	0.0002
dmiip2024_1	0.0477	0.1568	0.0679	0.1039	0.0002
dmiip2024_1	0.0477	0.1568	0.0679	0.1039	0.0002

¹⁷ <https://participants-area.bioasq.org/results/13b/phaseA/>

¹⁸ <https://participants-area.bioasq.org/results/13b/phaseAplus/>

¹⁹ <https://participants-area.bioasq.org/results/13b/phaseB/>

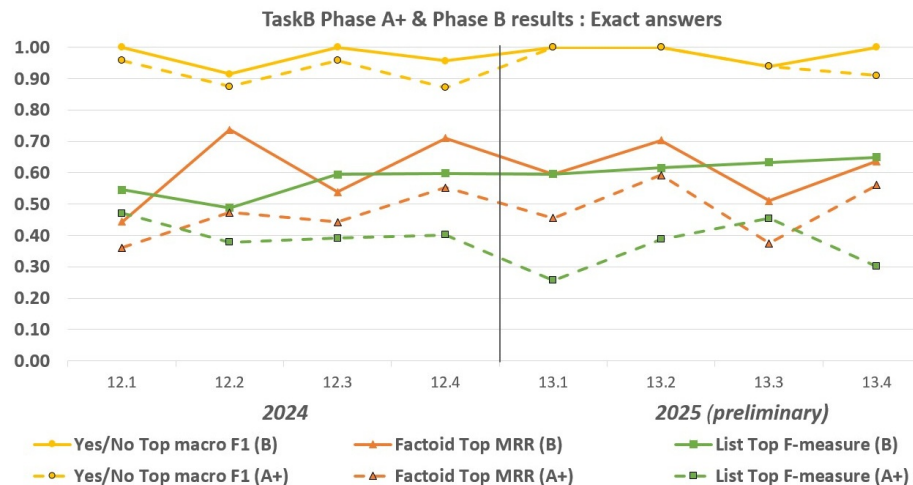


Fig. 1. The scores of the top systems in *exact answer* generation, for Phase A+ (dashed lines) and B (solid lines), across the test sets of *task 13b* and *task 12b* [34].

The top performance of the participating systems in *exact answer* generation for each type of question during the twelve years of BioASQ is presented in Figure 1. The preliminary results for task 12b, 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 3, for instance, presented in Table ??, several systems manage to correctly answer literally all yes/no questions. High performance, beyond 0.95 in macro F1 is also observed for yes/no questions in the remaining batches. More consistent performance is also observed in the preliminary results for list questions compared to the previous years, but there is still more room for improvement, as done for factoid questions where the performance across the batches fluctuates more.

4.2 Task Synergy 13

In *task Synergy 13* we use the same evaluation measures described for *task 13b*, considering only new material for the information retrieval part, an approach known as *residual collection evaluation*[38]. In addition, 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 indicated as “answer ready” were evaluated for *exact* and *ideal answers* per round.

Some indicative results for the Synergy task are presented in Table ?. The full 12b results are available online²⁰. 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*

²⁰ http://participants-area.bioasq.org/results/synergy_v2024/

Table 9. "Answer Ready" questions and top system performance in Task Synergy 13 per round.

round	anser ready	Top MAP (docu- ments)	Top F score (snip- pets)	Top MRR (factoid)	Top F-measue (list)	Top ma F1 (yesno)
1						
2						
3						
4						

for several open questions for developing problems, such as Colorectal Cancer, pediatric sepsis, Duchenne Muscular Dystrophy, and COVID-19. In total, after the four rounds of Synergy 12, enough relevant material was identified to provide an answer to about 78% of the questions. In addition, about 51% of the questions had at least one *ideal answer*, submitted by the systems, which was considered of ground-truth quality by the respective expert.

4.3 Task MultiClinSum

All in all, the top scores for each subtrack were:

- **CardioDis.** The team BIT.UA attained the top position with an ensemble of RoBERTa-based models (roberta-es-clinical-trials-ner) that also uses multi-head CRF. Their runs integrated the provided datasets in different ways, with the highest scores being achieved by the models that use both the DisTEMIST and CardioCCC data. Their best run achieved an F1-score of 0.8199 and a recall of 0.8243. The team with the next best F1-score (0.8049) is Enigma, who uses a CLIN-X-ES model also fine-tuned on the DisTEMIST and CardioCCC data. Interestingly, the team PICUSLab achieves the best precision (0.8886) by a wide margin combining the predictions of multiple models trained on different parts of the data (including an augmented version of the CardioCCC corpus) and then using string matching techniques to enhance the final predictions.
- **MultiDrug.** In Spanish, the best F1-score is achieved by the ICUE team (0.9277), who also achieved the best recall (0.9412). Meanwhile, in English and Italian the winner team is Enigma, with an F1-score of 0.9223 and 0.8842, respectively.

The results for the CardioDis subtrack are shown in in Table 10, while the results for the MultiDrug subtrack are presented in Table ?? for Spanish, Table ?? for English and Table 11 for Italian. Due to space limitations, only the top-6 systems are presented, with the complete results being available in the MultiCardioNER overview paper [21].

In conclusion, the task’s results are quite good and varied, with scores ranging from 0.9277 (by the ICUE team in the Spanish MultiDrug subtrack) to 0.2201

Table 10. Results of the MultiCardioNER CardioDis subtrack. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
BIT.UA	run1-all-full	0.8155	0.8243	0.8199
BIT.UA	run0-top5-full	0.811	<u>0.8181</u>	<u>0.8145</u>
Enigma	3-system-CLIN-X-ES-pretrained	0.8016	0.8082	0.8049
Enigma	2-system-CLIN-X-ES-14	0.8052	0.8007	0.803
PICUSLab	aug_fus_sub2	0.7794	0.803	0.791
BIT.UA	run4-all	0.7981	0.7827	0.7903

Table 11. Results of the MultiCardioNER MultiDrug subtrack in Italian. Only the top-6 systems are presented. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
Enigma	1-system-XLMR	0.884	0.8844	0.8842
Enigma	3-system-Italian-Spanish-RoBERTa	0.8723	<u>0.8956</u>	<u>0.8838</u>
Enigma	2-system-XLMR-filtering	<u>0.9016</u>	0.8606	0.8806
Siemens	run1_IMR	0.8891	0.8689	0.8789
ICUE	run4_GPT_translation	0.9114	0.8461	0.8776
ICUE	run5_GPT_translation_all	0.9114	0.8461	0.8776

(by the DataScienceTUV team, who had some problems with the submission, also in the Spanish MultiDrug subtrack). Overall, the results for the MultiDrug subtrack are higher than those for the CardioDis subtrack, which was to be expected since drugs, as an entity type, are simpler and more straightforward than diseases.

As stated earlier, in terms of methodology, there’s a definite trend of using pre-trained Transformer-based systems (with a preference for RoBERTa models, perhaps due to their availability in Spanish), with most participants going beyond mere finetuning. Many of the presented runs incorporate new layers over their initial system results, be it an ensemble of multiple models and their predictions, multi-head CRFs, window sliding or using some kind of post-processing.

However, what seems to really make a difference in MultiCardioNER is the use of the cardiology-specific data (the CardioCCC dataset), which is one of the shared task’s main research points. All top-performing systems incorporate the released 258 documents from the CardioCCC corpus in some way. Meanwhile, participants that only use the DisTEMIST and DrugTEMIST corpora (which are made up of clinical case reports from varied clinical specialties) are able to achieve a really high precision but a much lower recall, thus achieving a not-so-high F1-score. This seems to indicate that, while these systems are able to retrieve many clinical entities correctly (i.e. high precision), they fail to recover those entities that are specific to the cardiology domain (i.e. low recall).

4.4 Task BioNNE-L

Following prior research on entity linking [25, 22, 36, 37], we address BioNNE-L as a retrieval task: given a mention, a model must retrieve the top- k concepts from the given UMLS vocabulary and employ two ranking-based evaluation metrics: (i) Accuracy@ k and (ii) Mean Reciprocal Rank (MRR). Accuracy@ k : Accuracy@ $k=1$ if the correct UMLS CUI is retrieved at rank $\leq k$, and Accuracy@ $k=0$ otherwise. $MRR = \frac{1}{|E|} \sum_{e \in E} \frac{1}{rank_e}$, where E is the set of entities, $|E|$ is the number of entities, $rank_e$ is the rank of entity e 's the first correctly retrieved concept among the top k retrieved concepts. As baseline, we adopt zero-shot ranking using BERGMOT with each entity type processed independently to reduce memory footprint caused by extensive vocabulary.

Table 12. Official evaluation results of the BioNNE-L task for the multilingual and monolingual tracks in terms of Accuracy@1 (**@1**), Accuracy@5 (**@5**), and MRR. The best results for each track and metric are highlighted in **bold**.

Team	Multilingual				English				Russian			
	#	@1	@5	MRR	#	@1	@5	MRR	#	@1	@5	MRR
verbanexialab	—	—	—	—	1	0.70	0.80	0.74	—	—	—	—
LYX_DMIIIP_FDU	1	0.68	0.84	0.75	2	0.66	0.84	0.74	2	0.71	0.84	0.76
BlancaPlanca	2	0.67	0.81	0.73	3	0.64	0.83	0.72	1	0.72	0.83	0.76
EeyoreLee	3	0.63	0.76	0.69	4	0.64	0.82	0.71	4	0.65	0.74	0.69
dstepakov	4	0.63	0.71	0.66	—	—	—	—	3	0.70	0.76	0.72
ICUE	5	0.58	0.76	0.66	6	0.51	0.79	0.62	5	0.62	0.72	0.67
baseline	6	0.53	0.70	0.60	5	0.57	0.78	0.66	6	0.52	0.59	0.55
NLPIMP	7	0.41	0.58	0.48	—	—	—	—	—	—	—	—

The official evaluation results, ordered by Accuracy@1 value, for BioNNE-L are summarized in Table 12. Most of the participants adopted various BERT-based [9] with the top-performance achieved by domain-specific models, such as BERGMOT, SapBERT. Specifically, Team LYX_DMIIIP_FDU ranked the first in the multilingual track and the second rank in the two monolingual track by fine-tuning BERGMOT. Top 1 results for the Russian and English data are achieved by multilingual BERGMOT (Team BlancaPlanca) and English SapBERT (Team verbanexialab) models, respectively.

4.5 Task ELCardioCC

The results of the participants for the three subtasks (i.e. NER, EL and MLC-X) are presented in tables 13, 14 and 15.

The team droidlyx consistently achieved the highest micro-F1 scores across all subtasks, leading NER (0.733), EL (0.678), and performing strongly in MLC-X (0.847). The ELCardioCC baseline remained highly competitive, particularly excelling in MLC-X with a micro-F1 of 0.827 and maintaining strong performance in NER and EL. Svassileva's systems were consistently near the top in

Team	System	Precision	Recall	Micro-F1
bhuang	5nm	0.6448	0.5205	0.5761
	1nm	0.4914	0.5331	0.5114
	2nm	0.6338	0.4000	0.4905
	3nm	0.5460	0.4387	0.4865
	4nm	0.6575	0.3601	0.4653
droidlyx	system1	0.7059	0.7618	0.7328
ELCardioCC_baseline	mbert_baseline	0.6959	0.7460	0.7201
pjmathematician	config1	0.2484	0.2586	0.2534
	config2	0.2892	0.2188	0.2491
	config3	0.3297	0.1732	0.2271
svassileva	greek-bert-exact-bge-m3	0.7012	0.7328	0.7167
	greek-bert-statistical-bge-m3	0.7012	0.7328	0.7167
	greek-bert-statistical-sap	0.7012	0.7328	0.7167
	xlmr-exact-bge-m3	0.7079	0.7222	0.7150
	xlmr-statistical-bge-m3	0.7079	0.7222	0.7150

Table 13. Performance of participating systems in the ELCardioCC Named Entity Recognition (NER) subtask. Results are reported using micro-averaged precision, recall, and F1 score.

NER and EL but did not participate in MLC-X, while bhuang showed promising results especially in classification, though with more variability.

4.6 Task GutBrainIE

Submitted runs were evaluated using micro- and macro-averaged precision, recall, and F1-score, with micro-F1 used as the reference measure for the leaderboards since it is better suited when classes are imbalanced.

We adopted a baseline employing a fine-tuned NuNER model for NER [4] and a fine-tuned ATLOP model for all RE subtasks [46]. The baseline has also been used to annotate the bronze collection automatically.

Tables 16-19 show each team’s top run beating the baseline for NER, TB-RE, TT-RE, and TM-RE, respectively. The performance gap between the subtasks is noticeable. While NER obtained a top micro-F1 of 0.84 utilizing pretrained biomedical transformers, RE subtasks were more challenging: both TB-RE and TT-RE peaked at approximately 0.65-0.69 micro-F1, and TM-RE reached only 0.46 micro-F1, demonstrating the added difficulty of simultaneously locating and labeling entities and identifying relations among these.

5 Conclusions

This paper provides an overview of the twelfth BioASQ challenge. This year, BioASQ consisted of four tasks: (1) Task 12b on biomedical semantic question answering in English and (2) Synergy 12 on question answering for developing problems, both already established from previous BioASQ versions, (3) the new

Team	System	Precision	Recall	Micro-F1
bhuang	5nm	0.5927	0.4852	0.5336
	1nm	0.4480	0.5000	0.4726
	3nm	0.4963	0.4211	0.4556
	2nm	0.5734	0.3677	0.4480
	4nm	0.5945	0.3374	0.4305
droidlyx	system1	0.6529	0.7046	0.6778
ELCardioCC_baseline	EL_baseline	0.6476	0.6942	0.6701
pjmathematician	config1	0.0616	0.0642	0.0629
	config2	0.0693	0.0525	0.0597
	config3	0.0212	0.0112	0.0146
svassileva	xlmr-statistical-bge-m3	0.6594	0.6728	0.6660
	xlmr-exact-bge-m3	0.6585	0.6719	0.6651
	greek-bert-exact-bge-m3	0.6548	0.6844	0.6693
	greek-bert-statistical-sap	0.6548	0.6844	0.6693
	greek-bert-statistical-bge-m3	0.6540	0.6835	0.6684

Table 14. Performance of participating systems in the ELCardioCC Entity Linking (EL) subtask. Results are reported using micro-averaged precision, recall, and F1 score.

Team	System	Subtask 3a (MLC)			Subtask 3b (X)		
		P	R	F1	P	R	F1
ELCardioCC_baseline	MLCX1_baseline	0.9339	0.7422	0.8271	-	-	-
	MLCX2_baseline	0.9531	0.5864	0.7261	0.6050	0.4442	0.5122
bhuang	1nm	0.6205	0.7676	0.6863	-	-	-
	2nm	0.5131	0.8355	0.6358	-	-	-
	3nm	0.5227	0.8237	0.6395	-	-	-
	4nm	0.4572	0.8576	0.5965	-	-	-
	5nm	0.6947	0.8250	0.7543	-	-	-
droidlyx	system1	0.8569	0.8377	0.8472	-	-	-
kbogas	w2l_cb	0.2115	0.3421	0.2614	-	-	-
	w2l_emb_la	0.2115	0.3421	0.2614	-	-	-
	w2l_la	0.2115	0.3421	0.2614	-	-	-
	w2l_tf_cb	0.2115	0.3421	0.2614	-	-	-
	w2l_tfidf_la	0.2115	0.3421	0.2614	-	-	-
pjmathematician	config4	0.6056	0.2257	0.3288	0.2326	0.0932	0.1331
	config5	0.5860	0.2656	0.3655	-	-	-

Table 15. Performance of participating systems in ELCardioCC Subtask 3a (Multi-label Classification) and Subtask 3b (Explainable AI). Metrics include Precision (P), Recall (R), and Micro-F1 score. A dash (-) indicates that the system did not participate in the corresponding subtask.

task MultiCardioNER on the automatic detection of disease and drug mentions on cardiology clinical case reports in Spanish, Italian, and English, and (4) the new task BIONNE on biomedical nested NER in English and Russian.

The preliminary results for task 12b reveal the high performance of the top participating systems, predominantly for yes/no answer generation, despite the

Table 16. Performance metrics of each team’s top run beating the baseline for NER. The best result is in bold, the second-best is underlined (micro-averaged).

Team ID	Run name	Precision	Recall	F1
GutUZH	AugEnsemble	0.8384	0.8432	0.8408
Gut-Instincts	5eedev	0.8286	0.8480	<u>0.8382</u>
NLPatVCU	ensemble1	0.8255	<u>0.8488</u>	0.8370
ICUE	ensemble5-th10	<u>0.8369</u>	0.8294	0.8331
LYX-DMIP-FDU	EnsembleBERT	0.8020	0.8513	0.8259
ata2425ds	transformer	0.7914	0.8432	0.8164
greenday	llmner	0.7957	0.8278	0.8114
Graphswise-1	NERWise	0.8066	0.7955	0.8010
BASELINE	NuNerZero-Finetuned	0.7639	0.8238	0.7927

Table 17. Performance metrics of each team’s top run beating the baseline for TB-RE. The best result is in bold, the second-best is underlined (micro-averaged).

Team ID	Run name	Precision	Recall	F1
Gut-Instincts	6219eedev3re	0.6304	0.7532	0.6864
ONTUG	ElectraCLEANR	0.7121	0.6104	<u>0.6573</u>
Graphswise-1	AtlopOnto	0.7418	0.5844	0.6538
ataupd2425-pam	BiomedNLP-FULL_DATASET	0.5671	<u>0.7316</u>	0.6389
BIU-ONLP	RobertaLarge	<u>0.7453</u>	0.5195	0.6122
BASELINE	Atlop-Finetuned	0.7584	0.4892	0.5947

Table 18. Performance metrics of each team’s top run beating the baseline for TT-RE. The best result is in bold, the second-best is underlined (micro-averaged).

Team ID	Run name	Precision	Recall	F1
Gut-Instincts	6229eedev3re	0.6280	<u>0.7572</u>	0.6866
ataupd2425-pam	BiomedNLP-FULL_DATASET	0.5853	0.7202	<u>0.6458</u>
ONTUG	ElectraCLEANR	0.7059	0.5926	0.6443
Graphswise-1	AtlopOnto	0.7326	0.5638	0.6372
ICUE	biolinkbertl_pp	0.4974	0.7860	0.6093
BIU-ONLP	RobertaLarge	<u>0.7362</u>	0.4938	0.5911
BASELINE	Atlop-Finetuned	0.7533	0.4650	0.5751

extension of the expert team with two new experts. However, room for improvement is still available, particularly for factoid and list questions, where the performance is less consistent. The results of the new Phase A+ also reveal that state-of-the-art QA approaches can achieve high performance, even without access to manually selected relevant material. Still, providing such material leads to answers of improved quality. This edition of the Synergy task as well, revealed that state-of-the-art systems, despite still having room for improvement, can be a useful tool for biomedical experts who need specialized information for addressing open questions in the context of several developing problems.

Table 19. Performance metrics of each team’s top run beating the baseline for TM-RE. The best result is in bold, the second-best is underlined (micro-averaged).

Team ID	Run name	Precision	Recall	F1
Gut-Instincts	6239eedev3re	0.4215	0.5147	0.4635
Graphswise-1	AtlopOnto	<u>0.4686</u>	0.3097	<u>0.3729</u>
ICUE	biolinkbertl_pp	0.2858	<u>0.5054</u>	0.3651
LYX-DMIP-FDU	BioLinkBERT	0.3682	0.3257	0.3457
ONTUG	ElectraCLEANR	0.3529	0.3231	0.3373
BASELINE	Atlop-Finetuned	0.4986	0.2453	0.3288

The new task MultiCardioNER presented two new challenging subtasks about annotations clinical case reports in Spanish, English, Italian with disease and drug mentions. Building on the work laid out in previous shared tasks like Dis-TEMIST [27], this task introduces the nuance of creating clinical Named Entity Recognition systems specifically for the cardiology domain. In addition, it expands the range of the task beyond Spanish by introducing a subtrack that also involves English and Italian text. In order to do this, two new datasets are released: the DrugTEMIST corpus, that includes drug mentions in Spanish, English and Italian in a group of clinical case reports of varied medical specialties, and the CardioCCC corpus, a collection of cardiology clinical case reports with disease and drug annotations. The results highlight the importance of having data specific to the language and specialty the systems are going to be applied in, even within domains that are already quite specific like the clinical one.

The ever-increasing focus of participating systems on deep neural approaches and Large Language Models, already apparent in previous editions of the challenge, is also observed this year. Most of the proposed approaches built on state-of-the-art neural architectures (BERT, PubMedBERT, BioBERT, BART etc.) adapted to the biomedical domain and specifically to the tasks of BioASQ. This year, in particular, several teams investigated approaches based on Generative Pre-trained Transformer (GPT) models and Retrieval Augmented Generation (RAG) for the BioASQ tasks.

The BioNNE task centered on extracting the eight most common biomedical entities in Russian and English from PubMed abstracts while accommodating potential nested structures. The top-performing approach employed a bi-encoder framework that leverages contrastive learning to map text spans and entity types into a common vector representation space. The performance of pre-trained LLMs without fine-tuning exhibited significantly lower results, underscoring the necessity for specialized training data.

Overall, several systems managed competitive performance on the challenging tasks offered in BioASQ, as in previous versions of the challenge, and the top performing of them were able to improve over the state-of-the-art performance from previous years. BioASQ keeps pushing the research frontier in biomedical semantic indexing and question answering for eleven years now, offering both well-established and new tasks. Aligned with the direction of extending beyond

the English language and biomedical literature, which started with the task MESINESP [11] and continued consistently ever since, this year BioASQ was further extended with two new tasks, MultiCardioNER [21] and BioNNE [8]. In addition, this year we introduced a new phase in the QA task 12b (phase A+) allowing the assessment of systems that produce answers directly, without access to manually selected relevant material. The future plans for the challenge include a further extension of the benchmark data for question answering through a community-driven process, extending the community of biomedical experts involved in the Synergy task, as well as extending the resources considered in the BioASQ tasks, both in terms of documents types, languages, and more focused sub-domains of biomedicine.

6 Acknowledgments

The thirteenth edition of BioASQ is sponsored by Ovid, Atypion 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. The MultiCardioNER track was funded by Spanish and European projects such as DataTools4Heart (Grant Agreement No. 101057849), AI4HF (Grant Agreement No. 101080430), BARITONE (Proyectos de Transición Ecológica y Transición Digital 2021. Expediente N^o TED2021-129974B-C21) and AI4ProfHealth (PID2020-119266RA-I00). The work on the BioNNE-L task was supported by the Russian Science Foundation [grant number 23-11-00358]. The work on the GutBrainIE task was supported by the HEREDITARY Project, as part of the European Union’s Horizon Europe research and innovation programme (GA 101137074).

References

1. Aksenova, A., Datseris, A., Vassileva, S., Boytcheva, S.: Transformer-Based Disease and Drug Named Entity Recognition in Multilingual Clinical Texts: MultiCardioNER challenge. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
2. Appleton, J.: The gut-brain axis: influence of microbiota on mood and mental health. *Integrative Medicine: A Clinician’s Journal* **17**(4), 28 (2018)
3. Beltagy, I., Lo, K., Cohan, A.: Scibert: Pretrained language model for scientific text. In: EMNLP (2019)
4. Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., Bernard, E.: Nuner: Entity recognition encoder pre-training via llm-annotated data (2024)
5. Carabotti, M., Scirocco, A., Maselli, M.A., Severi, C.: The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* **28**(2), 203 (2015)
6. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)

7. Cryan, J.F., O’Riordan, K.J., Sandhu, K., Peterson, V., Dinan, T.G.: The gut microbiome in neurological disorders. *The Lancet Neurology* **19**(2), 179–194 (2020)
8. Davydova, V., Loukachevitch, N., Tutubalina, E.: Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024. In: *CLEF Working Notes* (2024)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**(Mlm), 4171–4186 (oct 2018), <http://arxiv.org/abs/1810.04805>
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018), <http://arxiv.org/abs/1810.04805>
11. Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., , Murasaki, R.T., Primo-Peña, E., Bojo-Canales, C., Paliouras, G., Krallinger, M.: Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. (2021)
12. Ghaisas, S., Maher, J., Kanthasamy, A.: Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. *Pharmacology & therapeutics* **158**, 52–62 (2016)
13. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**(1), 1–23 (2021)
14. Huguet Cabot, P.L., Navigli, R.: REBEL: Relation extraction by end-to-end language generation. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. pp. 2370–2381. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), <https://aclanthology.org/2021.findings-emnlp.204>
15. Jonker, R., Almeida, T., Matos, S.: BIT.UA at MultiCardioNER: Adapting a Multi-head CRF for Cardiology. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) *CLEF Working Notes* (2024)
16. Krithara, A., Nentidis, A., Paliouras, G., Krallinger, M., Miranda, A.: BioASQ at CLEF2021: large-scale biomedical semantic indexing and question answering. In: *Advances in Information Retrieval: ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II* 43. pp. 624–630. Springer (2021)
17. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746* (2019)
18. Lima-López, S., Farré-Maduell, E., Brivá-Escalada, V., Gascó, L., Krallinger, M.: MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. *Procesamiento del Lenguaje Natural* **71** (2023)
19. Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., Rodríguez-Miret, J., Krallinger, M.: Overview of SympTEMIST at BioCreative VIII: Corpus, Guidelines and Evaluation of Systems for the Detection and Normalization of Symptoms, Signs and Findings from Text. In: *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models* (2023)

20. Lima-López, S., Farré-Maduell, E., Gascó, L., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In: Working Notes of CLEF 2023 (2023)
21. Lima-López, S., Farré-Maduell, E., Rodríguez-Miret, J., Rodríguez-Ortega, M., Lilli, L., Lenkowicz, J., Ceroni, G., Kossoff, J., Shah, A., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
22. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4228–4238. Association for Computational Linguistics, Online (Jun 2021). <https://doi.org/10.18653/v1/2021.naacl-main.334>, <https://aclanthology.org/2021.naacl-main.334>
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
24. Loukachevitch, N., Manandhar, S., Baral, E., Rozhkov, I., Braslavski, P., Ivanov, V., Batura, T., Tutubalina, E.: NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities. *Bioinformatics* (04 2023). <https://doi.org/10.1093/bioinformatics/btad161>, <https://doi.org/10.1093/bioinformatics/btad161>, btad161
25. Loukachevitch, N., Sakhovskiy, A., Tutubalina, E.: Biomedical concept normalization over nested entities with partial UMLS terminology in Russian. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2383–2389. ELRA and ICCL, Torino, Italia (May 2024), <https://aclanthology.org/2024.lrec-main.213/>
26. Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Nentidis, A.: Evaluation measures for task b. Tech. rep., Tech. rep. BioASQ (2022), http://participants-area.bioasq.org/Tasks/b/eval_meas.2022
27. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., , Krallinger, M.: Overview of DISTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. (2022)
28. Nentidis, A., Katsimpras, G., Krithara, A., Krallinger, M., Ortega, M.R., Loukachevitch, N., Sakhovskiy, A., Tutubalina, E., Tsoumakas, G., Giannakoulas, G., Bekiaridou, A., Samaras, A., Di Nunzio, G.M., Ferro, N., Marchesin, S., Menotti, L., Silvello, G., Paliouras, G.: Bioasq at clef2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge. In: Advances in Information Retrieval. pp. 407–415. Springer Nature Switzerland, Cham (2025)
29. Nentidis, A., Katsimpras, G., Krithara, A., Lima-López, S., Farré-Maduell, E., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Experimental IR Meets Multilinguality, Multimodality, and Interac-

- tion. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)
30. Nentidis, A., Katsimpras, G., Krithara, A., Lima-López, S., Farré-Maduell, E., Krallinger, M., Loukachevitch, N., Davydova, V., Tutubalina, E., Paliouras, G.: Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) (2024)
 31. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
 32. Nentidis, A., Katsimpras, G., Vondrou, E., Krithara, A., Gasco, L., Krallinger, M., Paliouras, G.: 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. pp. 239–263. Springer (2021)
 33. Nentidis, A., Katsimpras, G., Vondrou, E., Krithara, A., Miranda-Escalada, A., Gasco, L., Krallinger, M., Paliouras, G.: 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 (2022). https://doi.org/10.1007/978-3-031-13643-6_22
 34. Nentidis, A., Krithara, A., Paliouras, G., Krallinger, M., Sanchez, L.G., Lima, S., Farre, E., Loukachevitch, N., Davydova, V., Tutubalina, E.: BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge. In: ECIR2024. pp. 490–497. Springer (2024)
 35. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018), <https://api.semanticscholar.org/CorpusID:49670925>
 36. Sakhovskiy, A., Semenova, N., Kadurin, A., Tutubalina, E.: Graph-enriched biomedical entity representation transformer. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 109–120. Springer Nature Switzerland, Cham (2023)
 37. Sakhovskiy, A., Semenova, N., Kadurin, A., Tutubalina, E.: Biomedical entity representation with graph-augmented multi-objective transformer. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 4626–4643. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). <https://doi.org/10.18653/v1/2024.findings-naacl.288>, <https://aclanthology.org/2024.findings-naacl.288/>
 38. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science **41**(4), 288–297 (jun 1990). [https://doi.org/10.1002/\(SICI\)1097-4571\(199006\)41:4<288::AID-ASIS;3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASIS;3.0.CO;2-H)
 39. Styll, P., Campillos-Llanos, L., Kusa, W., Hanbury, A.: Cross-Linguistic Disease and Drug Detection in Cardiology Clinical Texts: Methods and Outcomes. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.) CLEF Working Notes (2024)
 40. Sung, M., Jeon, H., Lee, J., Kang, J.: Biomedical entity representations with synonym marginalization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3641–3650. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.335>, <https://aclanthology.org/2020.acl-main.335/>

41. Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., Navigli, R.: WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2521–2533. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.findings-emnlp.215>, <https://aclanthology.org/2021.findings-emnlp.215/>
42. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* **16**, 138 (2015)
43. Yang, Z., Zhou, Y., Eric, N.: Learning to answer biomedical questions: Oaqa at bioasq 4b. *ACL 2016* p. 23 (2016)
44. Yasunaga, M., Leskovec, J., Liang, P.: Linkbert: Pretraining language models with document links. In: Association for Computational Linguistics (ACL) (2022)
45. Zaratiana, U., Tomeh, N., Holat, P., Charnois, T.: GLiNER: Generalist model for named entity recognition using bidirectional transformer. In: Duh, K., Gomez, H., Bethard, S. (eds.) *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 5364–5376. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). <https://doi.org/10.18653/v1/2024.naacl-long.300>, <https://aclanthology.org/2024.naacl-long.300/>
46. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2021)