



Overview of BioASQ 2026: The Fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering

Anastasios Nentidis¹✉, Georgios Katsimpras¹, Anastasia Krithara¹, Martin Krallinger², Miguel Rodríguez-Ortega², Eduard Rodríguez-López², Natalia Loukachevitch³, Igor Rozhkov³, Elena Tutubalina^{4,5}, Dimitris Dimitriadis⁶, Vasiliki Patsiou⁷, Grigorios Tsoumakas^{6,8}, George Giannakoulas⁶, Alexandra Bekiaridou⁷, Athanasios Samaras⁶, 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,mirodrig8,eduard.rodriguez}@bsc.es

³ Lomonosov Moscow State University, Moscow, Russia
louk_nat@mail.ru, fulstocky@gmail.com
louk_nat@mail.ru

⁴ Artificial Intelligence Research Institute, Moscow, Russia
tutubalinaev@gmail.com

⁵ HSE University, Moscow, Russia

⁶ Aristotle University of Thessaloniki, Thessaloniki, Greece
dndimitri@csd.auth.gr, {g.giannakoulas,ath.samaras.as}@gmail.com
dndimitri@csd.auth.gr

⁷ Northwell Health, New Hyde Park, NY, USA
{spatsiou19,ampekariidou}@gmail.com

⁸ Archimedes, Athena Research Center, Marousi, Greece

⁹ University of Padua, Padua, Italy

Abstract. This paper presents an overview of the fourteenth edition of the BioASQ challenge, organized in the context of the Conference and Labs of the Evaluation Forum (CLEF) 2026. BioASQ is an international challenge series that supports progress in biomedical language processing tasks ranging from semantic indexing and information extraction to question answering and summarization. In 2026, BioASQ included six shared tasks: a) *Task 14b* on biomedical semantic question answering. b) *Task Synergy14* on question answering for developing biomedical topics. c) *Task MultiClinSum-2* on multilingual clinical summarization. d) *Task BioNNE-R* on extracting relations between nested named entities in Russian and English. e) *Task ELCardioCC* on clinical coding in cardiology. f) *Task GutBrainIE* on gut-brain interplay information extraction. Across these six tasks, 87 distinct teams participated, submitting

more than 1000 runs overall. As in previous editions, several submissions reached competitive performance, reflecting the continued progress of state-of-the-art methods across biomedical language processing tasks.

Keywords: Biomedical knowledge · Semantic Indexing · Question Answering

1 Introduction

BioASQ was launched in 2012 to advance large-scale semantic indexing and question answering (QA) in biomedicine, and it has gradually expanded to include additional biomedical language processing tasks [72]. It pursues this goal through annual shared tasks and benchmark datasets that capture the real information needs of biomedical experts. These tasks include updated versions of established tracks that remain timely, as well as novel tracks introduced to address emerging biomedical information needs. Through these shared tasks, research teams worldwide working on clinical coding, information extraction, information retrieval, QA, and summarization in biomedicine gain access to public datasets, a common evaluation infrastructure, and opportunities to exchange ideas through the BioASQ challenges and workshops.

This paper presents a compact overview of the shared tasks, the respective datasets, the participating systems, and their performance in the fourteenth BioASQ challenge, held as a CLEF 2026 lab. The rest of the paper is structured as follows. Section 2 first gives a general overview of the shared tasks in BioASQ for 2026, together with the respective datasets. Section 3 then offers a condensed description of participation in each shared task. More detailed descriptions of systems can be found in the corresponding extended task overviews and in the proceedings of the BioASQ lab. Next, Sect. 4 briefly reports the performance of the systems in each task, using state-of-the-art evaluation measures or manual assessment. Finally, Sect. 5 presents some overall conclusions.

2 Overview of the Tasks

The fourteenth edition of the BioASQ challenge consisted of six tasks [45]: (i) *Task 14b* on biomedical semantic question answering [52]. (ii) *Task Synergy 14* on question answering developing biomedical topics [52]. (iii) *Task MultiClinSum-2* on multilingual clinical summarization [61]. (iv) *Task BioNNE-R* on extracting relations between nested named entities in Russian and English [62]. (v) *Task ELCardioCC* on clinical coding in cardiology [14]. (vi) *Task GutBrainIE* on gut-brain interplay information extraction [40]. In this section, we describe this year's editions for each of these six tasks with a focus on differences from previous editions of the challenge [46–48, 54].

2.1 Task 14b

BioASQ *task 14b* is the fourteenth edition of the established BioASQ *task b* on biomedical QA [52]. In 2026, the task was organized in three phases: i) Phase A, where systems received biomedical questions in English and had to retrieve relevant PubMed documents and snippets. ii) Phase A+, where systems had to produce both “exact” and “ideal” answers. Depending on the question type, the exact answer is either *yes* or *no* (yes/no), a single entity name, such as a disease or gene (factoid), or a set of entity names (list). The ideal answer is a paragraph-length response for all question types. iii) Phase B, where BioASQ experts supplied selected relevant material for each question and systems had to generate answers using this additional evidence.

For testing, 280 new biomedical questions were annotated with gold documents, snippets, and answers, including both exact and ideal answers. The training set contained 5,729 biomedical questions from earlier *task b* editions, together with answers and supporting documents and snippets, providing a unique resource for developing biomedical QA systems [28]. Table 1 summarizes the training and test datasets for task 14b. The test data were divided into four independent bi-weekly batches of 80, 80, 60, and 60 questions, respectively, as shown in Table 1.

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

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	5729	1541	1130	1695	1363	10.74	13.97
Test 1	80	17	21	23	19	5.25	12.41
Test 2	80	21	13	20	26	3.64	7.98
Test 3	60	11	17	17	15	3.40	6.95
Test 4	60	16	20	11	13	3.78	8.35
Total	6009	1606	1201	1766	1436	10.43	13.75

2.2 Task Synergy 14

BioASQ *task Synergy* was first introduced in 2021 to encourage research on developing biomedical topics, such as COVID-19 [30,31]. The task is designed as an ongoing dialogue in which experts pose open-ended questions on developing topics, without knowing in advance whether a definitive answer can already be produced. Systems return relevant material, namely documents and snippets, and the experts assess it. They then provide feedback about relevance and indicate whether the available evidence is sufficient by marking questions as *ready to answer*. The process is repeated across four rounds, with each round considering

new material from the evolving document resource [53]¹. For questions marked as *ready to answer*, systems also provide exact and ideal answers, which the experts assess, providing feedback that can help systems improve their answers in later rounds. Additionally, the experts may mark a question as *closed* when they receive a fully satisfactory answer that is not expected to change, or when the question is no longer of interest.

A training dataset of 400 questions on developing topics, incrementally annotated with relevant material and answers, is available from earlier versions of *task Synergy* [49–51, 55, 56]. In *task Synergy 14*, 63 new questions on developing health topics were added, namely cortisol biology and Cushing syndrome, cancer treatment response, molecular and genetic mechanisms, infectious diseases and prophylaxis, diagnostic methods, and mental health topics. Additionally, 3 questions that remained open from earlier editions were added in round two to be enriched with newer evidence and revised answers [52]. Overall, 66 distinct questions were considered in the four Synergy 14 rounds, which contained 63, 66, 64, and 48 questions, respectively. The corresponding distinct yes/no, list, factoid, and summary counts were 15, 17, 14, and 20 in total.

2.3 Task MultiClinSum-2

Clinical content, including patient medical records, detailed clinical case reports, and electronic health records (EHRs), continues to grow rapidly across healthcare systems worldwide, covering numerous languages beyond English and reflecting the global scope of medical practice. The extensive length of many clinical reports presents a significant barrier for healthcare professionals who must efficiently extract essential clinical information from these documents. Recent advances in generative and encoder-based transformer models have demonstrated considerable potential for automated summarization, offering the capability to synthesize comprehensive clinical narratives into concise summaries that preserve relevant patient details and clinical findings while significantly reducing document length. Such developments highlight the need to rigorously assess and compare the performance of clinical summarization systems across multiple languages.

Building on the first edition of the task, we present MultiClinSum-2, the second edition of our shared task focusing on automatic summarization of clinical case reports, now expanded to 15 languages: English, French, Spanish, Portuguese, Italian, Russian, Catalan, Norwegian, Danish, Romanian, German, Greek, Dutch, Czech, and Swedish. The task leverages a comprehensive corpus of full clinical cases paired with reference summaries derived from biomedical literature. Table 2 reports the corpus statistics for each language sub-track.

2.4 Task BioNNE-R

In the BioNNE-R Shared Task [62], we address the challenge of extracting relations between nested named entities, i.e. entities that contain other enti-

¹ As of 2023, this evolving document resource is PubMed [47].

Table 2. Descriptive statistics per language sub-track of the MultiClinSum-2 dataset. For each language, the number of case-summary pairs and mean word and sentence counts are reported for both full clinical cases and their corresponding reference summaries.

Language	Nr. Pairs	Mean Word Count		Mean Sent. Count	
		Full case	Summary	Full case	Summary
Catalan (ca)	27699	604.80	118.56	27.49	5.59
Czech (cs)	27488	469.89	86.37	30.78	5.73
Danish (da)	27652	485.14	93.76	28.75	5.70
German (de)	27424	510.20	96.89	30.47	5.82
Greek (el)	26749	547.05	100.84	29.53	5.62
English (en)	27943	528.72	98.75	28.41	5.59
Spanish (es)	27743	591.95	116.16	27.57	5.59
French (fr)	27455	615.43	119.87	27.69	5.58
Italian (it)	27553	572.88	113.19	27.46	5.58
Norwegian (nb)	27609	485.48	95.27	28.28	5.65
Dutch (nl)	27587	511.66	100.19	27.11	5.58
Portuguese (pt)	27593	551.87	108.86	27.22	5.58
Romanian (ro)	27579	551.83	110.71	26.94	5.58
Russian (ru)	27086	481.95	88.55	29.66	5.58
Swedish (sv)	27594	459.82	88.72	27.53	5.59

ties within their boundaries, from biomedical texts. Most relation extraction datasets are annotated with flat named entities, which cannot contain other named entities. Extraction of relations between nested entities provides additional challenges for models: a relation can connect an uppermost entity and its internal entity, or intersect boundaries of an uppermost entity connecting its internal entity and some external entity. These issues can be problematic even for state-of-the-art relation extraction approaches [76]. One of the key features of BioNNE-R shared task is the focus on relations between nested entities that are (i) derived from the annotations in the NEREL-BIO corpus [36, 37] and (ii) supplemented by newly annotated data in both English and Russian. The annotation includes 8 entity types and 14 relation types. The competition was organized into three subtasks that fell under two evaluation tracks: 1. Two **monolingual tracks** 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 are summarized in Tables 3 and 4. All BioNNE-R materials can be found on the shared task’s GitHub² and Codabench pages³.

² <https://github.com/nerel-ds/NEREL-BIO/tree/master/BioNNE-R>.

³ <https://www.codabench.org/competitions/13132/>.

Table 3. BioNNE-R Shared Task general document statistics. Entities are annotated entity mentions; relations are gold relation annotations.

Split	Docs	Entities	Rels
[RU]			
train	716	37,879	23,773
dev	50	3,210	2,521
test	154	9,339	7,078
[EN]			
train	55	3,861	3,193
dev	50	3,478	2,967
test	152	10,316	8,350
[Totals]			
RU	920	50,428	33,372
EN	257	17,655	14,510
BI	1,177	68,083	47,882

Table 4. BioNNE-R Shared Task relations and their count in all documents.

Relation	Russian				English				Total
	train	dev	test	total	train	dev	test	total	
ABBREVIATION	862	67	244	1173	97	74	238	409	1582
AFFECTS	2493	288	774	3555	362	393	920	1675	5230
ALTERNATIVE_NAME	649	97	201	947	95	98	285	478	1425
APPLIED_TO	284	45	41	370	43	54	70	167	537
ASSOCIATED_WITH	3295	242	987	4524	325	233	977	1535	6059
FINDING_OF	1356	85	453	1894	121	83	529	733	2627
HAS_CAUSE	2710	407	790	3907	579	487	971	2037	5944
ORIGINS_FROM	566	67	135	768	120	79	214	413	1181
PART_OF	1483	208	502	2193	204	248	428	880	3073
PHYSIOLOGY_OF	1127	131	313	1571	150	178	425	753	2324
SUBCLASS_OF	7517	669	2187	10373	743	757	2475	3975	14348
TO_DETECT_OR_STUDY	771	84	239	1094	164	116	414	694	1788
TREATED_USING	282	46	61	389	101	88	182	371	760
USED_IN	378	85	151	614	89	79	222	390	1004
TOTAL	23773	2521	7078	33372	3193	2967	8350	14510	47882

2.5 Task ELCardioCC

Cardiovascular diseases affect a significant portion of the global population, accounting for 32% of global deaths according to WHO⁴. Automated clinical coding plays a crucial role in transforming unstructured real-world medical data gathered from patients into structured information, facilitating clinical research, healthcare analytics, and clinical decision support. This need is particularly important for Greek hospitals, where discharge letters are often stored as free-text clinical narratives, making large-scale analysis, interoperability, and

⁴ <https://www.who.int/health-topics/cardiovascular-diseases>.

efficient reuse of clinical knowledge challenging. The digitization and structuring of discharge letters through automated ICD-10 coding can therefore substantially improve electronic health record management and support downstream healthcare applications. However, existing research predominantly focuses on English clinical text, leaving other languages, such as Greek, underrepresented. To address this gap, the *ELCardioCC* shared task [13] focuses on the automatic assignment of cardiology-related ICD-10 codes to discharge letters collected from Greek hospitals.

In detail, participants in the *ELCardioCC* task of 2026 were asked to develop artificial intelligence systems capable of processing Greek cardiology discharge letters and predicting the appropriate ICD-10 diagnostic codes at the document level. To accomplish this goal, participants could employ a wide range of natural language processing and machine learning techniques, including Named Entity Recognition (NER), Entity Linking (EL), and multi-label classification methods. The provided corpus was specifically designed to support experimentation with such approaches, enabling systems to identify clinically relevant concepts and map them to standardized medical knowledge in support of ICD-10 prediction. The training dataset includes 2,500 discharge letters, while the test set comprises 500 letters. System performance was evaluated using the F1 score.

2.6 Task GutBrainIE

The scientific literature on the gut-brain axis, which investigates the interplay between the *gut microbiota* and neurological and psychiatric disorders, is rapidly expanding [2, 7, 10, 17]. PubMed publications on this topic more than doubled between 2020 and 2025, increasing from approximately 600 to more than 1,500 articles per year, making it increasingly challenging for clinicians and researchers to keep track of new findings. For this reason, we propose the second edition of the *GutBrainIE* task, which aims to foster the development of Information Extraction (IE) systems able to identify, normalize, and structure relevant biomedical knowledge from scientific publications related to the gut-brain axis and its implications in Parkinson's, Alzheimer's, and Multiple Sclerosis [40].

Compared with the previous edition, the 2026 edition of *GutBrainIE* introduces two main changes: the dataset has been substantially expanded, from around 1,500 to more than 6,000 annotated documents, and concept-level annotations have been added [39, 41]. Accordingly, we kept two subtasks from the previous edition: Named Entity Recognition (NER), which requires the identification and classification of entity mentions according to the GutBrainIE annotation schema, and Mention-level Relation Extraction (M-RE), which requires the identification and labeling of relations between specific entity mentions. We further introduced two new subtasks: Named Entity Recognition and Disambiguation (NERD), which extends NER by requiring each detected entity mention to be linked to a concept identifier from reference biomedical vocabularies, and Concept-level Relation Extraction (C-RE), which evaluates relations at the level of linked concepts rather than surface mentions.

The dataset is organized into Training, Development, and Test sets. The Training set is divided into three quality tiers: expert-curated (Gold), student-annotated (Silver), and automatically annotated (Bronze). The Development and Test sets contain only expert annotations (Table 5).

Table 5. Dataset statistics for *GutBrainIE*.

Collection	# Docs	# Entities	Ents/Doc	# Rels	Rels/Doc
Train Gold	639	20530	32.13	8556	13.39
Train Silver	1310	41409	31.61	21523	16.43
Train Bronze	2972	89987	30.28	29692	9.99
Development Set	80	2521	31.51	1261	15.76
Test Set	80	2850	35.62	1285	16.06

3 Overview of Participation

Overall, 87 distinct teams took part across the six shared tasks of the fourteenth edition of the BioASQ challenge. Participation by task amounted to 45 teams in task 14b, 7 in Synergy14, 10 in MultiClinSum-2, 7 in BioNNE-R, 6 in ELCardioCC, and 16 in GutBrainIE. Most teams focused on a single task, although four teams submitted to both task 14b and Synergy14. This section gives a concise overview of the approaches developed by the participating teams for each BioASQ task. More detailed descriptions will be provided in the extended overview of each task [14, 40, 52, 61, 62], and system-specific papers will appear in the proceedings of the fourteenth BioASQ workshop⁵.

3.1 Task 14b

Task 14b attracted 45 teams, which produced 821 submissions from 151 distinct systems across the four batches and the three phases A, A+, and B. Participation remained as high as in the previous edition, with a slightly larger number of submitted systems and runs, showing that the task continues to be timely and relevant [46]. Specifically, 34, 28, and 36 teams competed in Phases A, A+, and B, respectively, with 100, 81, and 112 distinct systems and 261, 241, and 319 submissions. Twenty-three teams participated in all three phases. As in previous years, the open-source system OAQA [77], which achieved top performance in older editions of BioASQ [29], was used as a baseline for phase B *exact answers*.

The participating systems explored a broad set of retrieval and generation methodologies. Beyond traditional retrieval methods like BM25, systems utilized dense and hybrid retrieval models, which were further refined using techniques

⁵ <https://www.bioasq.org/workshop2026/proceedings>.

such as re-ranking, query-expansion, ontology-guided enhancements and multi-query strategies. Most approaches relied on multi-stage Retrieval-Augmented Generation (RAG) pipelines that combined these methods with state-of-the-art Large Language Models (LLMs) (including GPT-4/5 variants, Gemini, Claude, Llama-3, BioMistral, OpenBioLLM, Qwen). Furthermore, these systems also integrated various advanced techniques such as few-shot prompting, chain-of-thought reasoning, ensemble voting, prompt optimization, and agentic workflows. [1, 3–5, 15, 24, 27, 32, 57, 59, 60, 65, 67, 70, 71, 73, 74, 81]

3.2 Task Synergy 14

In BioASQ 14, seven teams participated in task Synergy14. They submitted 79 runs from 18 distinct systems. Four of these teams also participated in task 14b, while the remaining three focused on Synergy14. The participating systems employed a variety of approaches including hybrid retrieval methods, which combined BM25 with cross-encoder models, task-specific LLMs with evidence-grounded prompting, and RAG architectures that utilized various prompting and optimization techniques [3, 18, 59, 67, 69]. More detailed descriptions for some of the systems will be available in the workshop proceedings.

3.3 Task MultiClinSum-2

In general, participation in the task was very satisfactory, with a broader engagement reflecting the expanded multilingual coverage of this second edition. A total of 9 teams submitted at least one run of their predictions, resulting in 75 runs across all sub-tracks, with each team allowed to submit up to 5 runs per sub-track. As expected, the highest participation was observed in the English sub-track, with 7 teams submitting 20 runs, followed by Spanish (5 teams, 14 runs) and Portuguese (5 teams, 9 runs). Catalan and Dutch also attracted considerable attention, with 4 and 3 teams respectively. Norwegian was the only language that received no submissions, although full coverage was ensured through the developed baseline systems. Table 6 presents the participation received per language sub-track.

3.4 Task BioNNE-R

In total, we have received 20 Codabench registrations for the BioNNE-R task, with 7 teams submitting predictions during evaluation. All seven participated in the English subtask, and five each in the Russian and bilingual subtasks.

Most teams framed nested relation extraction as a pair-classification problem: candidate entity pairs were enumerated, encoded with typed entity markers inserted at span boundaries, and classified by a fine-tuned transformer with an explicit no-relation class. Encoder choices reflected the bilingual setting, with biomedical English models such as BioBERT [20] and BioLinkBERT [78] used for the English subtask and multilingual models such as mDeBERTa-v3 [22],

Table 6. Participation overview of MultiClinSum-2 task per team and language sub-track. Numbers indicate runs submitted;—indicates no participation.

Team	EN	ES	FR	PT	IT	RU	CA	NO	DA	RO	DE	EL	NL	CS	SV	Total
ixa-sum	5	5	2	2	2	—	5	—	—	2	2	—	2	—	—	27
MediScribes	5	—	—	—	—	—	—	—	—	—	—	—	5	—	—	10
NLP4Health	3	3	3	2	—	—	—	—	—	—	—	—	—	—	—	11
DACHausa	1	1	1	1	1	1	1	—	1	1	1	1	1	1	1	14
InfoLab-FEUP	3	—	—	3	—	—	—	—	—	—	—	—	—	—	—	6
PMH-Aqeel	1	1	—	1	1	1	1	—	—	—	—	—	—	—	1	7
UMUTeam	—	4	—	—	—	—	—	—	—	—	—	—	—	—	—	4
Aleyna	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
RaRaMi	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1
Total runs	20	14	6	9	4	2	7	0	1	3	3	1	7	1	2	75

XLM-RoBERTa [9], and multilingual BERT [12] for the Russian and bilingual subtasks. As the inference-time candidate set is dominated by negative pairs, nearly all teams applied postprocessing to control precision: schema-based filtering of entity-type pairs unseen in training, per-class threshold calibration on the development set, and distance- or sentence-boundary-based candidate pruning. Several teams explored alternatives to the standard classifier: HSE NLP Team used an LLM pipeline to synthesize training examples for rare relation types, aakobiakova fine-tuned an 8B-parameter LLM with LoRA [23] and log-probability inference, and LODAC-NII complemented a typed-marker classifier with a document-level relational graph model. More detailed system descriptions are available in the extended task overview [62].

3.5 Task ELCardioCC

The ELCardioCC task engaged six teams, with a total of 19 systems submitted in addition to our baseline model.

The submitted systems reveal a strong trend toward hybrid and multi-stage architectures that combine Named Entity Recognition (NER), entity linking, and document-level ICD-10 classification. Several teams framed the task as a pipeline problem, first identifying clinically relevant spans and then mapping them to standardized diagnostic codes. Some approaches relied on explicit mention detection followed by concept normalization or classification, often leveraging Greek-specific transformer backbones such as BERT-based Greek models and multilingual variants including XLM-RoBERTa. These systems enriched semantic representations through hierarchical medical knowledge, entity linking strategies, translated terminologies, and external clinical resources. A common motivation behind these designs was to reduce the complexity of the original

multi-label document classification task by transforming it into finer-grained mention-level prediction problems, enabling more precise handling of clinical concepts and improving generalization across rare ICD-10 categories.

Another clear trend was the extensive use of ensemble learning, curriculum learning, and specialized optimization strategies to address class imbalance and improve robustness. Multiple teams combined heterogeneous models using weighted voting, routing mechanisms, or metaheuristic search procedures rather than relying on a single architecture. For example, some teams explored sophisticated ensemble frameworks integrating precision-oriented and recall-oriented subnetworks, threshold tuning, asymmetric loss functions, and declarative search strategies for model fusion. These methods highlight the growing emphasis on balancing performance across both frequent and long-tail diagnostic labels, a critical challenge in clinical coding tasks. At the same time, even simpler baselines based on Greek BERT models demonstrated the importance of strong language-specific pretrained transformers as foundational components. Overall, the approaches indicate a broader movement toward combining deep contextual language models with structured clinical knowledge, ensemble optimization, and task decomposition techniques to achieve accurate and scalable automated ICD-10 coding systems for cardiovascular discharge summaries.

3.6 Task GutBrainIE

The *GutBrainIE* task registered 18 participating teams. Among these, 17 teams participated in NER, 14 in NERD, 11 in M-RE, and 10 in C-RE.

For NER, most teams adopted token classification approaches based on biomedical pretrained transformer models, including PubMedBERT, BioLinkBERT, BiomedBERT, and BioMedELECTRA, often combined with CRF layers [8, 20, 78]. Several systems used GLiNER-style models fine-tuned on the GutBrainIE training data [79]. A common strategy was to leverage the quality-based stratification of the training set through filtering, re-weighting, and relabeling of noisier data. Many teams also employed ensemble methods, combining models trained with different architectures, hyperparameters, or annotation subsets.

For NERD, most participants extended their NER pipelines with a Named Entity Linking (NEL) component. Common strategies included dictionary-based matching, SapBERT-style biomedical embeddings, nearest-neighbor retrieval, and concept-definition encoders [19, 34, 75].

Across the RE subtasks, participants mainly relied on sequence classification approaches using biomedical pretrained language models finetuned to perform classification of entity pairs marked in-line. Common strategies included type-based constraints, negative sampling, and threshold tuning. Some teams also experimented with LLMs using prompt- and RAG-based approaches.

4 Results

4.1 Task 14b

This section reports the evaluation measures and preliminary results for task 14b. The evaluation in *task 14b* combines expert manual assessment of system responses with automatic scoring using established measures [38], as in earlier editions of the task [46, 48]. Table 7 summarizes the official measures by response and question type, based on the preliminary results. The final results will become available after the BioASQ experts complete manual assessment and enrich the ground truth with any additional relevant items, answer elements, and synonyms. The online results pages for Phase A⁶, Phase A+⁷, and Phase B⁸ will be updated when final results are available.

Table 8 reports the batch-level document and snippet retrieval performance of systems participating in Phase A of task 14b. Document retrieval was strongest in the first batch and lower in the later batches, with an average top score across batches of 0.142 ± 0.013 . Top snippet F1 remained in a relatively narrow range across the four batches with an average value of about 0.095 ± 0.008 . Both metrics are lower than respective preliminary results reported in the BioASQ 13, which were 0.323 ± 0.08 and 0.244 ± 0.021 , respectively [46]. This can be related to the high percentage of questions from new experts (64%) in BioASQ 14, who have not contributed to the training dataset.

Table 7. The evaluation measures for *task 14b* per response type and question type [38].

Resp. type (Phase)	Quest. type	Official measure
Documents (A)	All	Mean Average Precision (MAP)
Snippets (A)	All	F1 (based on character overlaps)
Exact ans. (A+ & B)	List	F1
	Yesno	macro F1 on “yes” & “no” classes
	Factoid	Mean Reciprocal Rank (MRR)
Ideal ans. (A+ & B)	All	Manual scores for precision, recall, repetition, readability

The top exact-answer performance of the participating systems is summarized in Table 9 for both Phase B and Phase A+. In Phase B, the leading systems reached perfect yes/no macro-F1 in all four batches, while factoid MRR ranged from 0.400 to 0.588 and list F1 from 0.361 to 0.654. In Phase A+, where systems were not given expert-selected relevant material, the best yes/no scores remained close to Phase B performance, whereas factoid and list scores were lower overall, highlighting that the retrieval of relevant material is a considerable limitation.

⁶ <https://participants-area.bioasq.org/results/14b/phaseA/>.

⁷ <https://participants-area.bioasq.org/results/14b/phaseAplus/>.

⁸ <https://participants-area.bioasq.org/results/14b/phaseB/>.

Table 8. The average and top scores of participant systems in Phase A, task 14b.

Batch	Documents		Snippets	
	Average MAP	Top MAP	Average F1	Top F1
1	0.163	0.283	0.041	0.091
2	0.138	0.239	0.040	0.098
3	0.130	0.211	0.033	0.084
4	0.137	0.231	0.044	0.106

Table 9. Top scores of participant systems in exact answers, task 14b.

Batch	Phase B			Phase A+		
	Factoid MRR	Yes/No macro-F1	List F1	Factoid MRR	Yes/No macro-F1	List F1
1	0.522	1.000	0.361	0.478	1.000	0.234
2	0.400	1.000	0.472	0.400	0.948	0.391
3	0.588	1.000	0.518	0.529	1.000	0.326
4	0.576	1.000	0.654	0.409	0.935	0.484

4.2 Task Synergy 14

In *task Synergy 14*, the same evaluation measures used for *task 14b* are applied, while the information retrieval component considers only newly retrieved material, following the residual collection evaluation approach [64]. 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.

Table 10 presents the best performance obtained by participating systems per round in *task Synergy 14*. In Round 1, we started with 63 new questions; no questions were answer-ready, and no answers were received in this round, only relevant material. The four-round dialogue between biomedical experts and QA systems in this task supported the progressive collection of enough relevant material to answer 61 of these 63 new questions (97%) and three additional questions from the previous version of the task that remained open. In particular, for 48 questions (73%), the systems managed to provide at least one ideal answer, which was considered ground-truth quality by the respective expert. Overall, the task aided the generation of *exact* and *ideal* answers to questions on developing topics, including the Cushing syndrome, cancer therapy, molecular genetics, infections and prophylaxis, diagnostics, and mental health.

4.3 Task MultiClinSum-2

In this second edition, a new detailed evaluation was designed with the objective of providing a more comprehensive and clinically-oriented summary assessment,

Table 10. The number of questions (Qs) and “Answer Ready” (AR) questions and the top system performance per round (R) in Task Synergy14. Retrieval of documents (Top MAP) and snippets (Top F1). Generation of exact factoid (Top MRR), list (Top F1), and yes/no (Top macro-F1) answers.

R	Qs	AR	Top MAP	Top F1 Snip.	Top MRR	Top F1 list	Top macro-F1
1	63	0	0.538	0.340	-	-	-
2	66	42	0.337	0.262	0.545	1.000	1.000
3	64	55	0.249	0.390	0.394	0.818	0.771
4	48	47	0.315	0.197	0.500	0.900	0.890

going beyond the lexical and semantic metrics used in the first edition. For the submissions evaluation, automatically generated summaries were compared against reference summaries using a combined evaluation framework that integrates automatic summarization metrics with an LLM-as-judge approach, capturing both surface-level lexical overlap and semantic similarity alongside deeper clinical quality dimensions. The evaluation metrics considered are described below.

- **ROUGE-1/ROUGE-2** [33]: Measurement of unigram and bigram overlap between the generated and reference summaries, providing surface-level assessment of lexical similarity.
- **ROUGE-Lsum** [33]: A sentence-level variant of the ROUGE-L metric that computes the longest common subsequence (LCS) score at the summary level, offering a measure of content coverage and structural similarity with respect to the reference summary.
- **BERTScore** [80]: A semantic similarity metric that leverages contextualized embeddings from pre-trained transformer models to capture meaning beyond surface-level lexical overlap.
- **LLM-as-judge**: An evaluation approach powered by EuroLLM-9B-Instruct model [43], aiming to assess generated summaries across four clinical dimensions: Faithfulness (factual consistency), Completeness (coverage of key clinical information), Fluency (grammatical correctness) and Consistency (coherence across language versions).

Participating systems presented a broad methodological diversity, with two main strategy identified. On one side, several teams applied fine-tuning approaches with varying degrees of complexity: some used parameter-efficient methods such as QLoRA [11] for supervised instruction tuning, others explored reinforcement learning through Group Relative Policy Optimization (GRPO) [66] in order to optimize generated summaries quality based on lexical and factual consistency metrics, and one team opted for a continual pretraining step on medical-domain models before applying task-specific instruction tuning. On the other hand, a second group of teams approached the task through inference-only systems, focusing on prompt engineering, dynamic example retrieval, or

even the addition of external structured knowledge such as medical ontologies or knowledge graphs.

The evaluation results for each language sub-track are summarized in Table 11, showing the top-2 performing teams per language ranked by BERTScore. Team ixa-sum, with their GRPO-based system, consistently ranked first across most sub-tracks on both lexical and semantic metrics, suggesting a clear advantage over other strategies. Other systems such as NLP4Health and MediScribes achieved notable performance on LLM-as-judge clinical dimensions, particularly *completeness* and *faithfulness*. In terms of lower-resource languages, team DACHausa was the main participating team, providing the only submission in Greek, Czech and Danish. Regarding LLM-as-judge evaluation, *consistency* was the weakest clinical dimension, suggesting the limited capability of submitted systems to generate equivalent summaries across different language versions.

Table 11. Top-2 participant teams per MultiClinSum-2 sub-track, ranked by BERTScore. Developed baseline system results are not included in the ranking (R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Com. = Completeness, Fl. = Fluency, Cons. = Consistency).

Lang.	Team	Run	R-1 (R-2)	R-L	BERT Score	Faith.	Com.	Fl.	Cons.
EN	ixa-sum	2	0.44 (0.22)	0.32	0.88	0.73	0.83	0.70	0.70
	NLP4Health	1	0.38 (0.16)	0.27	0.87	0.76	0.94	0.70	0.70
ES	ixa-sum	2	0.47 (0.23)	0.31	0.88	0.71	0.80	0.70	0.65
	NLP4Health	1	0.41 (0.18)	0.27	0.87	0.72	0.84	0.70	0.63
PT	NLP4Health	1	0.38 (0.16)	0.26	0.87	0.74	0.89	0.70	0.64
	ixa-sum	1	0.38 (0.15)	0.26	0.87	0.74	0.91	0.70	0.65
CA	ixa-sum	2	0.46 (0.23)	0.30	0.88	0.71	0.87	0.70	0.63
	DACHausa	1	0.37 (0.16)	0.25	0.84	0.66	0.69	0.69	0.34
NL	ixa-sum	2	0.35 (0.11)	0.24	0.86	0.73	0.95	0.71	0.66
	MediScribes	3	0.34 (0.11)	0.22	0.86	0.75	0.98	0.71	0.67
FR	ixa-sum	2	0.41 (0.17)	0.26	0.87	0.74	0.90	0.70	0.67
	NLP4Health	1	0.41 (0.18)	0.26	0.87	0.72	0.83	0.70	0.65
IT	ixa-sum	1	0.36 (0.14)	0.24	0.87	0.72	0.89	0.70	0.65
	PsychAI-Discovery-Lab	1	0.33 (0.12)	0.24	0.86	0.70	0.79	0.70	0.56
RO	ixa-sum	1	0.36 (0.15)	0.25	0.87	0.74	0.91	0.70	0.57
	RaRaMi	1	0.34 (0.15)	0.24	0.86	0.70	0.68	0.68	0.30
DE	ixa-sum	1	0.30 (0.09)	0.20	0.86	0.72	0.83	0.70	0.65
	DACHausa	1	0.27 (0.10)	0.19	0.82	0.63	0.46	0.69	0.34
RU	PsychAI-Discovery-Lab	1	0.29 (0.07)	0.28	0.86	0.70	0.58	0.70	0.39
	DACHausa	1	0.32 (0.07)	0.31	0.83	0.63	0.42	0.68	0.27
SV	PsychAI-Discovery-Lab	1	0.33 (0.12)	0.23	0.86	0.71	0.78	0.70	0.61
	DACHausa	1	0.29 (0.12)	0.21	0.83	0.64	0.58	0.69	0.36
DA	DACHausa	1	0.29 (0.12)	0.21	0.83	0.63	0.58	0.68	0.35
EL	DACHausa	1	0.31 (0.07)	0.30	0.85	0.65	0.58	0.69	0.28
CS	DACHausa	1	0.30 (0.10)	0.19	0.83	0.63	0.45	0.68	0.29

4.4 Task BioNNE-R

As an official baseline, we provide a relation classifier built on the OpenNRE framework [21] with a `bert-base-multilingual-cased` backbone [12]. Each candidate entity pair is encoded with special entity markers inserted around the head and tail spans, and classified into one of the 14 relation types or a `no_relation` class. For each candidate pair, the minimal sentence segment surrounding the two entities is extracted as model input. Training data is constructed by enumerating gold entity pairs and adding negative (`no_relation`) examples at a 3:1 negative-to-positive ratio, with candidate pairs constrained to entity-type combinations permitted by the annotation schema. Separate checkpoints were trained for the English, Russian, and bilingual tracks, the latter on the concatenation of the English and Russian training data.

Systems were evaluated against the gold relation annotations using precision, recall, and F1-score. Given the highly imbalanced relation-type distribution, macro-averaged F1 across relation types was adopted as the official ranking metric, computed independently for each subtask.

Table 12. Official BioNNE-R results: macro-F1 of each team’s best submission per subtask. Best result per subtask in bold; – indicates no submission.

Team	English	Russian	Bilingual
ELiRF-UPV	0.5060	0.4717	0.4540
LODAC-NII	0.5016	0.5283	0.5015
aakobiakova	0.4951	0.4403	0.4716
HSE NLP	0.4733	0.5057	0.4527
rabiaozdemir	0.4620	–	–
savvafq	0.4570	0.4829	0.4849
LSIUNED	0.4514	–	–
Baseline	0.2825	0.3070	0.3027

Table 12 reports macro-F1 for the best submission of each team. All participating systems substantially outperformed the baseline, yet scores cluster in a narrow 0.44–0.53 band across subtasks, indicating considerable headroom for nested relation extraction. ELiRF-UPV achieved the best English result (0.5060), while LODAC-NII led on both the Russian (0.5283) and bilingual (0.5015) subtasks, with only a 0.0044 margin separating the two top English systems. Despite the much larger Russian training set, Russian and English scores were comparable and among the five teams that entered both subtasks, the mean macro-F1 was 0.4858 for Russian against 0.4866 for English, although the single best Russian score (0.5283) did exceed the best English one. The document-level graph approach and the typed-marker pair-classification approaches performed comparably, with no approach type showing a clear advantage.

4.5 Task ELCardioCC

The results of the best systems regarding of F1 score of the participants for the ELCardioCC task are presented in Table 13.

Table 13. Performance of participating systems in the ELCardioCC 2026

Team	System	Recall	Precision	Micro-F1
stanimeros	ensemble_metaheuristic_p4_winner...	0.850979	0.882995	0.866691
Georgios.1	Baymax_submission_v1	0.865845	0.861783	0.863809
LSI_UNED	test_set_NER_greekuncased_greek...	0.818709	0.841595	0.829994
alikipliona	submission_A_reranker	0.818347	0.824626	0.821474
elcardiocc	baseline_bert_base_uncased_greek	0.659173	0.903579	0.762264
Pakakisd	max_micro_curriculum	0.809282	0.720000	0.762035
fernandogd97	Greek_ClinLinker-P_greek_bert_DA..	0.791516	0.734028	0.761689

The results show a relatively competitive top tier, with stanimeros achieving the best overall performance ($F1 = 0.8667$) using an ensemble-based approach that also yields the highest precision (0.8830), indicating strong balance between correct predictions and error control. Very closely behind, Georgios.1 reaches an F1 of 0.8638, suggesting that different high-quality modeling strategies converge in performance at the top end. A mid-performing cluster includes LSI_UNED and alikipliona, with F1 scores around 0.83082, reflecting solid but less optimized systems. Below them, several systems (elcardiocc, Pakakisd, and fernandogd97) converge around 0.76 F1, but with notably different trade-offs: for example, the baseline BERT system prioritizes precision (0.9036) at the cost of recall, whereas others such as max_micro_curriculum favor recall more strongly. Overall, ensemble and correction-heavy methods clearly dominate the leaderboard, while simpler or single-model approaches tend to trade off precision and recall more unevenly, leading to lower or less balanced F1 scores.

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 the same baseline architecture used in the previous edition, employing a fine-tuned NuNER model for NER and a fine-tuned ATLOP model for the RE subtasks [6, 82]. To support NERD and C-RE, we added a NEL module combining exact matching from the annotated training data with PubMedBERT-based semantic similarity over URI definitions for unmatched entities [20].

Tables 14, 15, 16 and 17 show each team’s top run beating the baseline for NER, NERD, M-RE, and C-RE, respectively. NER obtained the highest scores,

with a top micro- F_1 of 0.87, confirming the effectiveness of biomedical pretrained transformers for exact-span entity recognition. Performances drop substantially in NERD, where the best micro- F_1 is 0.69, showing that concept linking introduces significant challenges. A similar drop can be observed across RE subtasks: M-RE reaches 0.61 micro- F_1 , while C-RE drops to 0.35. These results show that, although concept-level evaluation is not impacted by surface-form variations, it is highly sensitive to errors accumulated across entity recognition and linking, leading to a performance drop of over 40% from M-RE to C-RE.

Table 14. GutBrainIE 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
TWIX [42]	merged6	0.9548	0.8058	0.8740
NightSun [35]	run21	<u>0.8756</u>	0.8110	<u>0.8420</u>
Graphwise [58]	18	0.8613	0.8169	0.8385
TEXA [68]	1	0.7980	<u>0.8302</u>	0.8138
unibuc-bionlp-bp	1	0.7816	0.8425	0.8109
SMTE [44]	R1	0.8109	0.7932	0.8019
GetGut@AAU [25]	1	0.8011	0.8017	0.8014
NLP-DE	1	0.7847	0.8173	0.8007
BASLINE	NuNerZero-Finetuned	0.7782	0.8221	0.7996

Table 15. GutBrainIE performance metrics of each team’s top run beating the baseline for NERD. The best result is in bold, the second-best is underlined (micro-averaged).

Team ID	Run name	Precision	Recall	F1
TWIX [42]	24merged6	<u>0.7527</u>	0.6353	0.6890
NightSun [35]	run16	0.6290	<u>0.6053</u>	<u>0.6169</u>
Graphwise [58]	9	0.6217	0.5897	0.6053
TUGW [26]	EXP4SUB1EL	0.7723	0.4437	0.5636
MindGut link	BiomedBERT	0.5738	0.5330	0.5527
GetGut@AAU [25]	1	0.5515	0.5519	0.5517
GBA	5	0.4385	0.5263	0.4784
TEXA [68]	1	0.4581	0.4766	0.4672
BASLINE	GLiNER-3stage	0.4281	0.4522	0.4398

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

Team ID	Run name	Precision	Recall	F1
TWIX [42]	mre10	0.8996	0.4651	0.6132
NightSun [35]	run18	0.4459	<u>0.4593</u>	<u>0.4525</u>
SMTE [44]	R2	0.4476	0.3997	0.4223
GetGut@AAU [25]	1	<u>0.4546</u>	0.3658	0.4054
GutHub [63]	1	0.4347	0.3754	0.4029
Graphwise [58]	BGPT5515GEXP8	0.3517	0.4497	0.3947
BASELINE	Atlop-finetuned	0.4444	0.3453	0.3886

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

Team ID	Run name	Precision	Recall	F1
TWIX [42]	11mre12	0.4903	0.2691	0.3475
NightSun [35]	run18	<u>0.2484</u>	<u>0.2798</u>	<u>0.2632</u>
Graphwise [58]	BGPT5515GEXP4	0.2092	0.2963	0.2452
GetGut@AAU [25]	1	0.2123	0.1926	0.2020
TEXA [68]	1	0.1556	0.1383	0.1464
BASELINE	Atlop-3stage	0.1403	0.1292	0.1345

5 Conclusions

This paper provides an overview of the fourteenth BioASQ challenge. This year, BioASQ consisted of six tasks: (i) *Task 14b* on biomedical semantic question answering. (ii) *Task Synergy14* on question answering for developing biomedical topics. (iii) *Task MultiClinSum-2* on multilingual clinical summarization. (iv) *Task BioNNE-R* on relation extraction between nested named entities in Russian and English. (v) *Task ELCardioCC* on clinical coding in cardiology. (vi) *Task GutBrainIE* on gut-brain interplay information extraction.

The results for task 14b indicate that the strongest participant systems achieved high scores, particularly for yes/no answer generation, even in Phase A+, where expert-selected relevant material was not provided, which is in alignment with task 13b. For list and factoid questions, performance remains less consistent, especially in Phase A+, leaving clear room for further improvement. For these question types, access to expert-selected relevant material appears to help systems produce higher-quality answers. These results also underline the importance of Phase A, which evaluates automatic retrieval of relevant material and shows variability across batches and a performance drop in task 14b, where more questions by new experts were included, compared to task 13b. Participants applied a diverse range of retrieval and generation strategies, including

traditional methods, LLM and agent-based frameworks, multi-stage and ensembling techniques, and domain-specific knowledge integration. The Synergy14 results, consistent with earlier editions, suggest that current systems can support biomedical scientists seeking specialized information on developing problems, while still leaving room for further advances.

The MultiClinSum-2 task focuses on the automatic summarization of clinical case reports across 15 languages, providing a comprehensive multilingual benchmark and evaluation framework to foster the development of automatic summarization systems capable of processing diverse clinical narratives within different linguistic contexts. The task received a total of 9 participating teams, with the highest participation concentrated in languages with greater resource availability such as English, Spanish or Portuguese. Participant systems showed rich and diverse methodological approaches, ranging from prompt engineering and agentic inference pipelines augmented with external biomedical knowledge such as UMLS, to parameter-efficient fine-tuning strategies such as QLoRA or reinforcement learning via GRPO. A key addition of this second edition is the introduction of an enhanced evaluation framework combining lexical and semantic summarization metrics with LLM-as-judge approaches, under which participating systems achieved notable results across all sub-tracks on both assessment dimensions.

The BioNNE-R task evaluated the extraction of relations between nested named entities in Russian and English biomedical texts. Seven teams participated, most framing the task as pair classification with typed entity markers and fine-tuned transformers, alongside one document-level graph model and LLM-based systems. All submissions outperformed the multilingual baseline, but the best macro-F1 scores stayed near 0.500.53, showing that relation extraction over nested entities remains an open challenge.

The ELCardioCC task centered on extracting and classifying medical entities from Greek discharge letters, drawing participation from six teams who submitted a diverse set of systems. Overall, the results show a clear shift toward hybrid pipelines and ensemble-based systems that combine NER, entity linking, and ICD-10 classification to better capture clinical semantics. Strong performance is consistently achieved by models that integrate multiple strategies and leverage Greek-specific transformer backbones, especially when combined with optimization techniques to address class imbalance. While simpler BERT-based baselines remain competitive, they are generally outperformed by more complex ensemble and knowledge-enhanced approaches. This indicates that combining structured clinical knowledge with multi-model fusion is key to achieving robust and accurate ICD-10 coding in this task.

The GutBrainIE task, centered on information extraction for the gut-brain axis, challenged participants with Named Entity Recognition and increasingly fine-grained Relation Extraction subtasks. Teams achieving the strongest performance employed supervised deep-learning strategies, combining pretrained biomedical language models with ensemble strategies. Only a few participants experimented with prompt-based or generative approaches; however, these gen-

erally obtained lower scores, confirming the need to develop specialized models to effectively extract complex entities and relations in a specific biomedical domain.

Overall, BioASQ 14 brought together 87 distinct teams across established and newer tasks, with substantial participation in all six tracks. Several participating systems achieved competitive performance, and some improved over baselines or previous state-of-the-art results. In line with earlier editions, BioASQ continues to push research forward in semantic indexing, question answering, information extraction, and language processing in biomedicine, offering both long-running and newer tasks. The challenge first expanded beyond English biomedical literature with MESINESP [16], and it has continued to broaden its scope since then. In its fourteenth edition, BioASQ included MultiClinSum-2 [61], BioNNE-R [62], ELCardioCC [14], and GutBrainIE [40], alongside the established 14b and Synergy14 tasks. As a result, BioASQ 14 covered fifteen languages, multiple document types (biomedical articles, clinical case reports, and discharge letters), and specialized biomedical domains such as gut-brain interaction and cardiology.

Future directions for BioASQ include further expanding QA benchmark datasets through community-driven processes, broadening the network of biomedical experts involved, and enlarging the range of resources used across BioASQ tasks. This includes additional document types, more languages, and further specialized subdomains within biomedicine.

Acknowledgments. The fourteenth edition of BioASQ is sponsored by Ovid. The MEDLINE/PubMed data resources considered in this work were accessed courtesy of the U.S. National Library of Medicine. BioASQ is grateful to the CMU team for providing the *exact answer* baselines for task 14b. The work on the BioNNE-R task is an output of a research project (HSE-BR-2025-025) implemented as part of the Basic Research Program at HSE University. This research was funded by the Ministerio de Ciencia e Innovación (MICINN) under project BARITONE (TED2021-129974B-C22). This work is also supported by the European Union’s Horizon Europe Co-ordination & Support Action under Grant Agreement No 101080430 (AI4HF), as well as Grant Agreement No 101057849 (DataTool4Heartproject). ELCardioCC has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. The work on the Gut-BrainIE task was supported by the HEREDITARY Project, as part of the European Union’s Horizon Europe research and innovation programme (GA 101137074).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Antunes, D.: Retrieval-Bound Generation: Lean RAG pipelines for Biomedical QA. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
2. Appleton, J.: The gut-brain axis: influence of microbiota on mood and mental health. *Integrative Med. Clinician’s J.* **17**(4), 28 (2018)
3. Ateia, S., Kruschwitz, U.: Can LLMs refine their own keyword search strategies for biomedical RAG? In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)

4. Avdeiko, A., Topaloglu, P., Vaccari, T., Peikos, G.: MedCascade: A Multi-Stage Cascaded Retrieval and Generation System for Biomedical Question Answering. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
5. Bakker, J., Kamps, J.: Contrastive learning for efficient reranking in Task 14B of the CLEF 2026 BioASQ Track. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
6. Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., Bernard, E.: NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data (2024)
7. 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)
8. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations (2020)
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR abs/1911.02116 (2019). <http://arxiv.org/abs/1911.02116>
10. Cryan, J.F., O’Riordan, K.J., Sandhu, K., Peterson, V., Dinan, T.G.: The gut microbiome in neurological disorders. *Lancet Neurol.* **19**(2), 179–194 (2020)
11. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms (2023). <https://arxiv.org/abs/2305.14314>
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. ACL, Minneapolis, Minnesota (June 2019). <https://doi.org/10.18653/v1/N19-1423>
13. Dimitriadis, D., et al.: Overview of ElCardioCC task on clinical coding in cardiology at BioASQ 2025. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) CLEF 2025 Working Notes (2025)
14. Dimitriadis, D., et al.: Overview of elcardiocc task on clinical coding in cardiology at BioASQ 2026. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
15. Galat, D., Rizoiu, M.A.: Cost-pragmatic quality gating and selection–fusion multi-model combiners for BioASQ phase A+/B. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
16. Gasco, L., et al.: Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. In: Proceedings of the 9th BioASQ Workshop (2021)
17. 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)
18. Gibson, J., Nyberg, E.: SynapFlow at BioASQ@CLEF 2026: retrieval-augmented generation with diversity-enhanced hybrid retrieval. In: Sánchez Salido, E., Barrón-

- Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
19. Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, E., Garcia-Olano, D.: Learning dense representations for entity retrieval. In: Bansal, M., Villavicencio, A. (eds.) *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 528–537. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/K19-1049>, <https://aclanthology.org/K19-1049/>
 20. 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 Trans. Comput. Healthcare (HEALTH)* **3**(1), 1–23 (2021)
 21. Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., Sun, M.: Opennre: an open and extensible toolkit for neural relation extraction. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 169–174 (2019)
 22. He, P., Gao, J., Chen, W.: Debertav 3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing (2021)
 23. Hu, E.J., et al.: Lora: low-rank adaptation of large language models, vol. 1, p. 3 (2022)
 24. Hung, Y.C., Han, J.C., Hung, H.C., Tsai, R.T.H.: LAB Note of BioASQ 14b Phase A/A+, and Phase B Challenge. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 25. Jørgensen, S., Garavito Molina, E., Rodriguez, J.M., Dell’Aglia, D.: GetGut: a multi-stage pipeline for gut-brain axis information extraction. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 26. Kantz, B., Waldert, P., Lengauer, S., Schreck, T.: CHASTE: Constrained Holistic Annotator System with onTological Entities. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 27. Kouti, A.M., Refanidis, I.: Sparse retrieval with cross-encoder reranking and heuristic snippet scoring for biomedical question answering. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 28. Krithara, A., Nentidis, A., Bougiatiotis, K., Paliouras, G.: BioASQ-QA: a manually curated corpus for Biomedical Question Answering. *Sci. Data* **10**(1), 170 (2023)
 29. Krithara, A., Nentidis, A., Paliouras, G., Kakadiaris, I.: Results of the 4th edition of BioASQ Challenge. In: *Proceedings of the Fourth BioASQ Workshop* (2016). <https://www.aclweb.org/anthology/W16-3101.pdf>
 30. 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)
 31. Krithara, A., et al.: BioASQ Synergy: a dialogue between question-answering systems and biomedical experts for promoting COVID-19 research. *J. Am. Med. Inform. Assoc.*, ocae232 (08 2024). <https://doi.org/10.1093/jamia/ocae232>

32. Leekha, R., Gwon, D., Shing, L.: Agentic RAG with open-weight LLMs for Biomedical QA Task 14b. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
33. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop, pp. 74–81., Barcelona, Spain (2004) Text Summarization Branches Out
34. 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://www.aclweb.org/anthology/2021.naacl-main.334>
35. Liu, Y., Niu, Z., Li, C.: NightSun at CLEF2026 GutBrainIE: a first-place system for NER, entity linking, and relation extraction. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
36. Loukachevitch, N., et al.: NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities. *Bioinformatics* (04 2023). <https://doi.org/10.1093/bioinformatics/btad161>, btad161
37. 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/>
38. 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
39. Martinelli, M., et al.: Overview of GutBrainIE@CLEF 2025: gut-brain interplay information extraction. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) CLEF 2025 Working Notes (2025)
40. Martinelli, M., et al.: Overview of GutBrainIE@CLEF 2026: gut-brain interplay information extraction. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
41. Martinelli, M., et al.: A domain-specific curated benchmark for entity and document-level relation extraction. In: Demberg, V., Inui, K., Marquez, L. (eds.) Findings of the Association for Computational Linguistics: EACL 2026, pp. 5693–5711. Association for Computational Linguistics, Rabat, Morocco (Mar 2026). <https://doi.org/10.18653/v1/2026.findings-eacl.301>, <https://aclanthology.org/2026.findings-eacl.301/>
42. Martinelli, M., Menotti, L.: TWIX: a Two-stage approach for end-to-end named entity recognition and relation extraction. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
43. Martins, P.H., et al.: Eurollm: Multilingual language models for europe (2024). <https://arxiv.org/abs/2409.16235>
44. Maule, S., Di Nunzio, G.M.: SMTE – GutBrainIE@CLEF 2026: biomedical named entity recognition and mention-level relation extraction for the gut-brain axis. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)

45. Nentidis, A., et al.: BioASQ at CLEF2026: the fourteenth edition of the large-scale biomedical semantic indexing and question answering challenge. In: *Advances in Information Retrieval*, pp. 315–324. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-21321-1_42
46. Nentidis, A., et al.: Overview of BioASQ 2025: the thirteenth BioASQ Challenge on large-scale biomedical semantic indexing and question answering. In: Carrillo-de Albornoz, J., et al. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 173–198. Springer, Cham (2026)
47. Nentidis, A.: Overview of BioASQ 2023: the eleventh BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 227–250. Springer, Cham (2023)
48. Nentidis, A., et al.: 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)*
49. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023. In: *CEUR Workshop Proceedings (2023)*
50. 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 2024 Working Notes (2024)*
51. Nentidis, A., et al.: Overview of BioASQ Tasks 13b and Synergy13 in CLEF2025. In: Spina, D. (ed.) *CLEF 2025 Working Notes (2025)*
52. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ tasks 14b and synergy14 in CLEF2026. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) *CLEF 2026 Working Notes (2026)*
53. Nentidis, A., et al.: 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)
54. Nentidis, A., et al.: Overview of BioASQ 2022: The Tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer (2022). https://doi.org/10.1007/978-3-031-13643-6_22
55. Nentidis, A., Katsimpras, G., Vandroou, E., Krithara, A., Paliouras, G.: 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. CEUR Workshop Proceedings (2021)*. <http://ceur-ws.org/Vol-2936/paper-10.pdf>
56. Nentidis, A., Katsimpras, G., Vandroou, E., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022. In: *CEUR Workshop Proceedings*, vol. 3180, pp. 171–178 (2022)
57. Nguyen, N., Elnagger, A., Kashyap, A., Schultz, T.: Few but Deep vs. many but shallow: comparing licensed and free tool sets for agentic biomedical question answering. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) *CLEF 2026 Working Notes (2026)*
58. Nikolova, I., et al.: You Don't Need a Frontier LLM: Domain-Specific Encoders can compete in Gut–Brain Information Extraction. In: Sánchez Salido, E., Barrón-

- Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
59. Panou, D., Dimopoulos, A.C., Koubarakis, M., Reczko, M.: q-MAP: sensitive QA systems with mapped question embeddings. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 60. Ribeiro, A., Garrido, R., Christiansen, A., Jonker, R.A.A., Matos, S.: BIT. UA at BioASQ 14B: Modular Retrieval with pg_textsearch and Qdrant, and Agent-Based Answer Generation. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 61. Rodríguez-Ortega, M., Rodríguez-López, E., Lima-López, S., Mash, A., Krallinger, M.: Overview of MultiClinSum-2 at CLEF 2026: Data, Evaluation, and Results for Multilingual Clinical Case Report Summarization Across 15 Languages. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 62. Rozhkov, I., Loukachevitch, N., Tutubalina, E.: Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 63. Rydter, A.C., Christensen, S.B., Austys, D., Rodriguez, J.M., Dell’Aglia, D.: GutHub: tackling the GutBrainIE task through entity recognition, disambiguation, and relation extraction. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 64. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* **41**(4), 288–297 (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)
 65. Shahzadi, A., Wasim, M., Tayyab Waqar, M.: A multi-stage retrieval and ensemble-based answer generation pipeline for biomedical QA in BioASQ Task 14b. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 66. Shao, Z., et al.: Deepseekmath: pushing the limits of mathematical reasoning in open language models (2024). <https://arxiv.org/abs/2402.03300>
 67. Sims, D., Williams, K., Henry, S.: RMC at BioASQ 2026: a local RAG approach for end-to-end Biomedical QA. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 68. Smirnov, D., Khashechian, B., Di Nunzio, G.M.: TEXA at GutBrainIE 2026: a multi-stage approach for entity recognition, linking, and relation extraction in gut-brain biomedical abstracts. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 69. Sukprasan, R., Alexandru Iulian, O., Jacob, F., Daniela, R.: LLM-Guided RAG: a biomedical retrieval system for BioASQ synergy 2026. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
 70. Sun, L., Yang, H., Tang, J., Quaresma, P., Jia, Y., Li, Z.: MeSH for BioASQ Task 14B: A Lightweight Knowledge-Enhanced RAG Pipeline. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)

71. Tsai, R.T.H., Chiu, Y.S., Hou, C.Y., Hsu, C.H.: BioASQ Phase B: exploring diverse optimization techniques and testing frameworks for enhanced bio-answering performance. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
72. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.* **16**, 138 (2015)
73. Volz, S., Scheuvens, B., Ghosh, S., Scells, H.: Health-NLP at the BioASQ 2026 Task B on Biomedical Semantic QA and GutBrainIE on gut-brain interplay information extraction tasks. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
74. Wang, Y.: A multistage evidence retrieval system for BioASQ Task 14b: hybrid retrieval, reranking, and snippet selection. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
75. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable zero-shot entity linking with dense entity retrieval. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 6397–6407. Association for Computational Linguistics, Online (Nov 2020). <https://doi.org/10.18653/v1/2020.emnlp-main.519>, <https://aclanthology.org/2020.emnlp-main.519/>
76. Yandutov, A., Loukachevitch, N.: Approaches to relation extraction for nested named entities. *Lobachevskii J. Math.* **44**(1), 249–258 (2023)
77. Yang, Z., Zhou, Y., Nyberg, E.: Learning to answer biomedical questions: OAQA at BioASQ 4B. In: Kakadiaris, I.A., Paliouras, G., Krithara, A. (eds.) *Proceedings of the Fourth BioASQ workshop*. pp. 23–37. ACL, Berlin, Germany (Aug 2016). <https://doi.org/10.18653/v1/W16-3104>
78. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: Pretraining Language Models with Document Links. In: *Association for Computational Linguistics (ACL)* (2022)
79. 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. ACL, Mexico City, Mexico (Jun 2024). <https://doi.org/10.18653/v1/2024.naacl-long.300>
80. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. In: *International Conference on Learning Representations (ICLR)* (2020). <https://arxiv.org/abs/1904.09675>
81. Zhao, X., Lee, M., Balaji, A.: Retrieval-augmented biomedical question answering with weak-question recovery and neural reranking for BioASQ Task 14b. In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes (2026)
82. 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)