q-MAP: sensitive QA systems with mapped question embeddings

BSRC "Alexander Fleming" at CLEF 2026

Dimitra Panou[1,2], Alexandros C. Dimopoulos[2,3], Manolis Koubarakis[1,4] and Martin Reczko[2,*]

[1]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
[2]Institute for Fundamental Biomedical Science, Biomedical Sciences Research Center "Alexander Fleming", Greece
[3]Department of Informatics & Telematics, School of Digital Technology, Harokopio University, Greece
[4]Archimedes, Athena Research Center, Greece

Abstract

In this work, we describe the Fleming systems submitted to the BioASQ14 challenge for both Task B and the Synergy task. Our work focuses on improving Phase A document retrieval through a multi-stage dense-sparse hybrid retrieval and re-ranking pipeline. Our approach combines dense semantic retrieval using KaLM embeddings with sparse BM25 retrieval over PubMed abstracts, followed by weighted Reciprocal Rank Fusion (wRRF) and cross-encoder re-ranking using Gemma2-LW and MxBai-Large models. To reduce the semantic gap between biomedical questions and relevant abstracts, we introduce a novel question-side mapper (q-MAP) that projects question embeddings closer to the embedding of correct answers. The dense and sparse retrieval outputs are merged using wRRF, followed by a subsequent wRRF step to combine the dual cross-encoder re-ranking models. A dual pipeline that fuses RRF scores from both mapped and unmapped queries yielded the best performance in the competition. For snippet identification, we evaluate both existing sliding-window semantic similarity methods and modified LLM-based extraction strategies. Exact and ideal answers are generated using our previous ensembles of open-source large language models.

Our systems achieved strong results in BioASQ14: in Task B Phase A document retrieval, we obtained 1st place in Batch 3 and 2nd place in Batch 2. In the Synergy 2026 task, we achieved two 1st places for exact answers in rounds 3 and 4, a 1st place for snippets in round 3, and top rankings for ideal answers under the preliminary automatic evaluation.

Keywords

Biomedical Question Answering, BioASQ, Rerankers, Multi-Stage Retrieval, Large Language Models, Retrieval-augmented generation

1. Introduction

Modern Information Retrieval (IR) and Retrieval-Augmented Generation (RAG) pipelines increasingly rely on neural dense retrieval to bridge the structural and semantic gaps between short user queries and long target documents. The architectural state of the art spans several core paradigms, balancing computational latency against token-level interaction granularity.

The highest accuracy bound is established by **Cross-Encoders**, which concatenate the query and document into a single Transformer pass [1]. This permits full, unconstrained cross-attention across all token pairs, though it introduces an exhaustive runtime complexity for candidate passages.

Conversely, **Asymmetric Bi-Encoders** (Two-Tower models) map queries and documents independently into a shared low-dimensional vector space $\mathbb{R}^d$ via separate pooling strategies [1]. This shifts the heavy computational burden offline, allowing sub-millisecond maximum inner product search at

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
panou@fleming.gr (D. Panou); dimopoulos@fleming.gr (A. C. Dimopoulos); koubarak@di.uoa.gr (M. Koubarakis); reczko@fleming.gr (M. Reczko)
https://dsit.di.uoa.gr/dimopoulos-cv/ (A. C. Dimopoulos); https://cgi.di.uoa.gr/~koubarak/ (M. Koubarakis); https://www.fleming.gr/research/ifbr/staff-scientists/reczko-lab (M. Reczko)
0000-0002-9824-4489 (D. Panou); 0000-0002-4602-2040 (A. C. Dimopoulos); 0000-0002-1954-8338 (M. Koubarakis); 0000-0002-0005-8718 (M. Reczko)

the expense of an extreme information bottleneck where a full document is compressed into a single vector. This paradigm is further split into symmetric single-weights networks and parameter-disjoint configurations like **Dense Passage Retrieval (DPR)** [2], which utilizes two completely distinct model weights to account for linguistic differences between questions and answers.

To retain cross-attention performance without its associated inference costs, **Late Interaction Models**, such as ColBERT, maintain token-level multi-vector representations for both inputs [3]. Relevance scoring is deferred to the end of the pipeline using a fast, parallelizable MaxSim operator. While highly precise, this paradigm demands a massive storage footprint due to storing multi-vector sequences for the entire corpus.

Alternatively, methods can restructure the text space prior to embedding. **Hypothetical Document Embeddings (HyDE)** use a generative large language model (LLM) at query time to synthesize a pseudo-document, converting an asymmetric query-to-document task into a more reliable symmetric document-to-document search [4].

At the ingestion side, **Document Expansion** algorithms like Doc2Query utilize offline sequence-to-sequence networks to predict and append prospective queries directly to documents before indexing, eliminating query-time overhead while mitigating vocabulary mismatch [5]. A summary of these trade-offs is synthesized in Table 1.

Table 1
State-of-the-Art Retrieval Architectural Comparison

Approach	Granularity	Latency	Storage
Cross-Encoder	Full Attention	Extreme	Low
Bi-Encoder (Symmetric)	Shared Pool	Ultra-Low	Low
DPR (Asymmetric Two-Tower)	Disjoint Pool	Ultra-Low	Low
Late Interaction	Token Multi-Vec	Low	High
HyDE	Document-Gen	High	Low
Doc2Query	Pre-Expanded	Ultra-Low	High

While classical dense retrievers—including both symmetric single-weights models [1] and the in theory superior parameter-disjoint Asymmetric Two-Tower architectures like DPR [2]—and contemporary top-tier models on the Massive Text Embedding Benchmark (MTEB) [6] share an architectural foundation in decoupling query and document encoding, they represent entirely distinct computational generations.

Classical bi-encoders are typically constrained to small, encoder-only Transformer backbones (e.g., BERT-base, $\approx 110 \times 10^6$ parameters) and a rigid context boundary of 512 tokens. While the parameter-disjoint design of DPR allows the question tower and document tower to optimize for separate grammar spaces independently, the underlying representation engine remains bottlenecked. This structural limitation forces an aggressive information compression, requiring long documents to be heavily truncated or fragmented into isolated chunks.

Conversely, modern state-of-the-art (SOTA) MTEB embedders leverage massive auto-regressive decoder LLM trunks spanning from 1.5×10^9 to 27×10^9 parameters. These architectures natively support expanded context windows ranging from 8,192 to 131,072 tokens, allowing for the deep semantic compression of entire long-form documents or complex codebases into a unified vector space.

A core drawback of early dual-tower models is their static projection mechanism; queries and documents pass unconditionally through the networks, leaving the model to implicitly resolve asymmetric intent through text matching. Modern architectures overcome this via *Instruction-Tuning*, where a natural language task description prefix adjusts the model's internal attention weight distribution at query time. This aligns abstract search intents dynamically, transforming asymmetric query-to-passage tasks into optimized symmetric mappings without needing physically distinct, parameter-disjoint networks.

Furthermore, contemporary models replace rigid mean pooling with advanced latent attention mechanisms and *Matryoshka Representation Learning* (MRL) [7], used also in our pipeline. MRL explicitly

trains the network to pack the highest-density semantic information into the earliest vector dimensions. This enables dynamic truncation (e.g., downstream compression from 3072 to 256 dimensions) at query time, sacrificing less than 1% retrieval accuracy while yielding up to a 12× reduction in downstream vector database storage and distance computation costs, as summarized in Table 2.

Table 2
Evolution of Bi-Encoder Architectures

Evaluation Dimension	Classical Bi-Encoder (2019 / DPR)	Modern SOTA MTEB Embedder
N parameters	$\approx 110 \times 10^6 - 220 \times 10^6$	$1.5 \times 10^9 - 27 \times 10^9$ (Decoder LLM)
Context Window Length	512 Tokens (Hard Cap)	8,192 – 131,072 Tokens
Average Retrieval NDCG@10	$\approx 45.0 - 55.0$ [8, 2]	**70.0 – 75.5+** [6]
Projection Mechanics	Static Vector (Single or Dual-Tower)	Dynamic Task-Instruction Prefixes
Dimensionality Handling	Rigid, Fixed Width (e.g., 768)	Flexible Matryoshka Representation

Our proposed Question-Side Embedding Mapper (q-MAP) occupies a distinct operational niche by combining a careful regression based exclusively on positive examples with an index-frozen transformation.

As neural retrieval frameworks have improved rapidly over recent years, the inherent human bias in the selection of "golden" documents by BioASQ [9] competition experts has become increasingly apparent. AI-generated document selections tend to be more comprehensive than expert selections, which are inherently limited by the volume of documentation a human can process. Consequently, an effective document retrieval system tailored for BioASQ must include specialized components optimized to mirror the preferences (or biases) of human experts, while simultaneously maintaining a high baseline sensitivity and precision. The components in our pipeline optimized to capture these human expert constraints are the Question-Side Embedding Mapper (q-MAP) and multiple weighted Reciprocal Rank Fusion (wRRF) [10] steps across all score fusions.

2. Methodology

Although modern embedding models leading MTEB yield superior retrieval metrics, their immense scale inflicts severe operational constraints. In streaming, production-grade QA systems, processing user queries through a multi-billion parameter autoregressive decoder model introduces non-negligible inference latencies and substantial hardware provisioning costs. Furthermore, migrating a vast, pre-existing static document index to a newly released MTEB model is often computationally and financially prohibitive.

To break this performance-infrastructure bottleneck, we introduce a novel cross-space mapping alignment framework that preserves the target document collection in a frozen representation space while substituting query-side LLM processing with a lightweight parametric layer.

2.1. Architectural Formulation

Let $\mathcal{D}$ represent a static document corpus indexed by a frozen, highly parameterized encoder E_D (e.g. from the MTEB leaderboard), yielding a fixed vector store where each document vector is defined as $\mathbf{e}_d = E_D(d) \in \mathbb{R}^n$, where n is the dimension of the embedding.

To eliminate the operational and infrastructure overhead of maintaining disparate underlying models, our framework enforces complete encoder symmetry on the input layer by utilizing the identical frozen SOTA MTEB network for query ingestion. An incoming runtime query q_i is passed through this encoder

to produce its baseline dense representation:

$$\mathbf{e}_{q_i} = E_D(q_i) \in \mathbb{R}^n \tag{1}$$

Although $\mathbf{e}_{q_i}$ and $\mathbf{e}_d$ reside within the same nominal vector space, they suffer from a geometric shift caused by the structural and linguistic asymmetry between short, interrogative queries and dense, declarative passages. To bridge this gap, we introduce a specialized neural mapping network $\mathcal{M}_\phi : \mathbb{R}^n \to \mathbb{R}^n$, parameterized by its weights ϕ. This network processes the query embedding exclusively, transforming it into an optimized target retrieval coordinate:

$$\mathbf{e}^*_{q_i} = \mathcal{M}_\phi\left(\mathbf{e}_{q_i}\right) \tag{2}$$

Unlike conventional dense retrieval layers trained via contrastive classification, the transformation network $\mathcal{M}_\phi$ is explicitly optimized as a coordinate-wise vector regression task. The training dataset $\mathcal{T} = \{(\mathbf{e}_{q_i}, \mathbf{e}_{d_i^+})\}_{i=1}^N$ is highly streamlined, pairing the baseline query embeddings $\mathbf{e}_{q_i} = E_D(q_i)$ exclusively with the pre-computed embeddings of their corresponding positive ("golden") target documents denoted with $\mathbf{e}_{d_i^+}$.

Aligning the query projections against a curated selection of target documents affords a powerful mechanism for post-hoc domain adaptation, steering the retrieval phase toward specific semantic target distributions without altering the underlying corpus index. Crucially, the objective of $\mathcal{M}_\phi$ is not to reconstruct the precise embeddings of the target documents d_i^+ and solve the entire retrieval task, but rather to minimize the intrinsic structural asymmetry between query and document representations and to induce a geometric transformation toward the desired semantic categories.

The optimization objective minimizes the cosine similarity loss for the mapped query and the true document representation:

$$\begin{aligned} \text{loss}(x_1, x_2) &= 1 - \frac{x_1 \cdot x_2}{\|x_1\|_2 \cdot \|x_2\|_2} \\ \mathcal{L}_{\text{cossim}} &= \frac{1}{N} \sum_{i=1}^{N} \text{loss}(\mathbf{e}^*_{q_i}, \mathbf{e}_{d_i^+}) \end{aligned} \tag{3}$$

where N denotes the size of the training set.

By freezing the document index and optimizing only the cross-space mapper $\mathcal{M}_\phi$ on query vectors, the system captures the extreme performance of top-tier MTEB models on long-form text documents while reducing query-time latency to a single pass through E_D and $\mathcal{M}_\phi$ with subsequent optimized vector search.

2.2. Implications of Symmetric Ingestion and Regression-Based Mapping

This architecture introduces several profound theoretical and practical implications for high-throughput QA systems:

1. **Resolution of Intra-Space Query-Passage Asymmetry:** By setting the query base encoder equal to the document encoder, the model no longer attempts to align entirely different semantic embedding topologies. Instead, $\mathcal{M}_\phi$ functions strictly as a geometric translation vector field within a singular, high-fidelity MTEB manifold. It acts as a specialized post-hoc corrector, learning the exact spatial transformation required to approximate a query's semantic coordinates towards its corresponding answer's structural neighborhood.
2. **Elimination of the Hard-Negative Mining Bottleneck:** Traditional contrastive learning depends heavily on complex hard-negative mining (e.g., BM25 or iterative dense filtering) to present non-trivial boundaries during training. By converting the task into a coordinate regression toward a fixed target point, the model computes an exact translation vector field, rendering offline negative-mining infrastructure entirely obsolete.

3. **Immunity to False Negative Label Noise:** Large-scale QA datasets frequently suffer from unlabeled positives (false negatives) within the corpus. Contrastive losses explicitly penalize models when a valid but unlabeled passage appears in the negative subset. Because our point-to-point regression pipeline only enforces a target-directed attraction without negative repulsion forces, it eliminates this source of gradient noise.
4. **Strict Ingestion-Side Invariance:** Unlike document expansion techniques (Doc2Query), our method keeps the pre-existing document database entirely frozen. Large-scale enterprise indexes can remain stored in their native, high-dimensional SOTA MTEB formats without requiring computationally expensive re-indexing, offline schema updates, or document-side projection modifications.
5. **Infrastructure Decoupling:** Maintaining a single unified encoder backbone (E_D) slashes cold-start memory footprints and production deployment costs, as only one large-scale LLM instance needs to be hosted in GPU memory or queried via downstream APIs. The runtime query-side translation is offloaded entirely to $\mathcal{M}_\phi$, which can be implemented as an ultra-lightweight multilayer perceptron (MLP) with residual connections, executing in sub-millisecond timeframes.

2.3. Question-Side Embedding Mapper training

To train the query mapper used in this retrieval pipeline, we constructed a dataset containing 5488 unique question identifiers from the BioASQ competitions held between 2023 and 2025. The data were split approximately to a 70%/15%/15% split into training, validation, and test partitions with 3836, 820 and 832 question-document pairs, respectively. As the mapper $\mathcal{M}_\phi$ should only perform a small, bounded transformation of the question embeddings, we initially used a linear transformation that is initialized to the identity mapping. All mapped embeddings are normalized to unit length for direct use in dense retrieval. This linear layout prevents overfitting, acting as an implicit regularizer compared to unconstrained deep architectures. Shallow residual networks represent an alternative with similarly low parameter counts. We train the mapper using standard Adam optimization, a learning rate of 1×10^{-5}, and early stopping enforced via the validation set. After 52 epochs, the training loss was reduced from 0.227061 to 0.102816, and the validation loss dropped from 0.221634 to 0.216539. Moving beyond linear mappings, an initial application of gradient boosting over an ensemble of stabilized, low-rank weak learners to effectively minimize residual mapping errors showed promising results.

2.4. Document Retrieval Implementation

Our pipeline for document retrieval is visualized in figure 1.

Sparse Retrieval (BM25) Sparse retrieval was performed using BM25 implemented in Pyserini over a Lucene index built from the 2025 PubMed collection with the same parameters used in our previous systems [11, 12] (BM25: $k1 = 0.35$, $b = 0.37$, RM3: $terms = 17$, $docs = 14$, $qw = 0.7$). The top 4000 candidates for each question were retrieved from titles and abstract.

Dense Retrieval We constructed an SQLite-based retrieval database of PubMed abstracts published from 2010 onward. We utilized the `KaLM-embedding-multilingual-mini-instruct-v2.5` model [13] and extracted its 512-dimensional Matryoshka vector representations [7] to construct our dense index, downscaling from the model's native 768-dimensional output to optimize storage efficiency. The top 4000 candidates for each question were retrieved. In the systems named Fleming1 and Fleming2, unmapped question embeddings were used to search the dense index. The systems Fleming3 and Fleming4 use the question-side embedding mapper.

Dense/Sparse Fusion The top 4000 documents returned from dense and sparse retrieval were merged with Reciprocal Rank Fusion (RRF) [10] for batch 1 and the top 200 documents were selected. A weighted

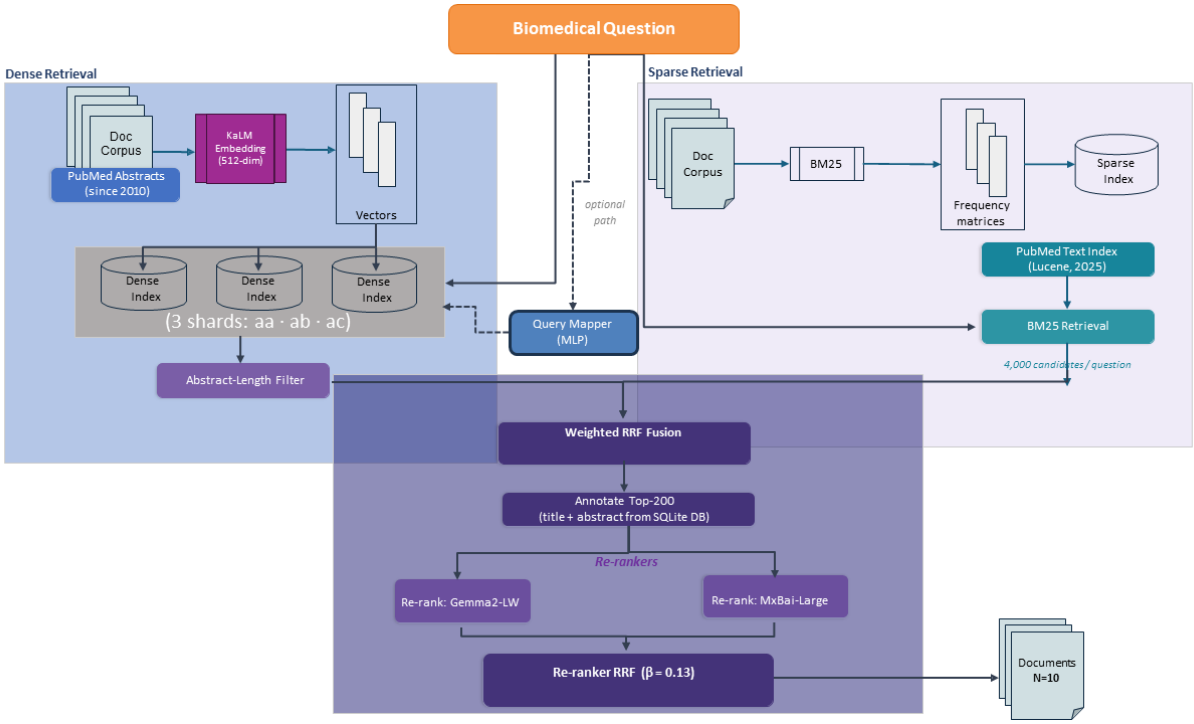

Figure 1: Document retrieval pipeline. In the left Dense Retrieval, the PubMed corpus and the Biomedical Question is embedded using KaLM and relevant abstracts are retrieved using cosine similarity. With Query Mapping, the embedding is first mapped with the Query Mapper and then cosine similarity search is applied. In the right Sparse Retrieval using BM25 an index is constructed from the corpus and abstracts relevant for the question are retrieved. The dense and sparse retrieval results are merged based on their scores using weigthed RRF and the top 200 abstracts are re-ranked using the question text, title and abstract and the two rerankers Gemma2-L and MxBai-Large. The reranker scores are finally merged with weighted RRF to return the final top10 documents.

RRF was optimized with the results of batch 1 and used in subsequent batches. The optimal weight (beta) for batch 1 is 0.04, emphasizing the documents from the dense retrieval.

Reranking The top 200 documents returned from the sparse and dense fusion were scored with the rerankers MxBai-Large and Gemma2-LW. In batch 1, RRF was used to merge the scores of the rerankers. A weighted RRF was optimized with the results of batch 1 and used in subsequent batches. The optimal weight (beta) is 0.13, emphasizing the documents scored with Gemma2-LW. The top 10 documents from this fusion constitute our final submissions.

2.5. Performance of Variants with Question-Side Embedding Mapping in BioASQ14

Table 3 summarizes the document retrieval ranks of our system variants across the BioASQ14 competition batches in the published ranking of all submitted systems that is based on the mean average precision (MAP). The “unmapped” configuration represents our baseline dense-sparse hybrid retrieval pipeline backed by double cross-encoder re-rankers. The “q-MAP” designation denotes an identical pipeline where query embeddings undergo parametric translation before dense retrieval begins. Lastly, the “q-MAP/unmapped fusion” introduces an RRF layer that combines the outputs of both the baseline and mapped streams.

The fused variant achieved 1st place in Test Batch 3 and consistently outperformed both standalone configurations across all evaluated batches. This indicates that a simple RRF blend using raw and mapped query embeddings effectively pushes relevant documents to the top of the results, successfully mitigating localized projection errors better than either contributing method could in isolation.

The drop in the ranks for the last batch can be attributed to a higher proportion of golden documents published before 2010 there that are currently not covered by the dense retrieval (the last golden

document the first 10% percentile of the PMID identifiers is published in 2007 for batches 1 to 3, while it is published in 2002 for batch 4).

Table 3
Document retrieval ranks in BioASQ14 for q-MAP variants.

System (name)	Batch 1 Rank	Batch 2 Rank	Batch 3 Rank	Batch 4 Rank	Total Rank
q-MAP/unmapped fusion (Fleming5)	6	5	**1**	24	**36**
q-MAP (Fleming3, Fleming4)	9	6	4	34	53
unmapped (Fleming1, Fleming2)	7	7	2	37	53

2.6. Phase A+ and Phase B

The generation of exact and ideal answers is performed as described in our previous system [12]. For exact answers, the identified optimal farms of LLMs were applied. For factoid type questions, the optimal ensemble is DeepSeek-R1-Distill-Llama-70B, LLaMA 3.3, Qwen3-14B, Qwen3-32B, Reflection and Smaug. For list type questions, the ensemble is DeepSeek-R1-Distill-Llama-70B, LLaMA 3.3 and Qwen3-14B. For Yes/No type questions, the ensemble is Aya, Qwen3-32B and Smaug. All ideal answers were genereated using LLaMA 3.3 and prompts containing all predicted or golden snippets for each question.

3. Conclusion

In this work, we presented the architectural framework and competitive outcomes of the Fleming systems submitted to the BioASQ14 challenge. By engineering a multi-stage dense-sparse hybrid retrieval pipeline combined with an ensemble of advanced cross-encoders (Gemma2-LW and MxBai-Large), our configuration established top-tier performance benchmarks across both Task B and the Synergy task [14].

The primary architectural innovation introduced was the Question-Side Embedding Mapper (q-MAP), a lightweight coordinate regression network designed to close the query-document semantic gap. By shifting from traditional, noise-sensitive contrastive objectives to a streamlined, positive-only cosine alignment task, q-MAP successfully modeled the specific selection habits and systemic biases of human medical experts. Crucially, it achieved this domain-specific adaptation while keeping the underlying multi-million document PubMed index completely frozen and invariant, eliminating the prohibitive infrastructural overhead of corpus re-indexing.

Empirical evaluations from the BioASQ14 competition confirm that fusing the retrieval outputs of mapped and unmapped query representations via Reciprocal Rank Fusion yields optimal results, culminating in a 1st place finish in Test Batch 3. Future work will explore expanding the mapping architecture.

Acknowledgments

We thank the anonymous reviewers for their useful comments. We express our gratitude to the BioAsq challenge organizers for organizing the event and offering continuous support. The GPU computations were executed on two servers acquired as part of project ID 16624, titled "Creation - Expansion - Upgrading of the Infrastructures of research centers supervised by the General Secretariat for Research and Innovation (GSRI)" with the code MIS 5161770. This project received funding under the "National Recovery and Resilience Plan Greece 2.0".

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini, Copilot in order to: Grammar and spelling check, paraphrase and reword. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of EMNLP-IJCNLP, 2019, pp. 3982–3992.

[2] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. tau Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[3] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[4] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of ACL (Volume 1: Long Papers), 2023, pp. 99–111.

[5] R. Nogueira, J. Lin, From doc2query to doctttttquery, arXiv preprint arXiv:1912.08797 (2019).

[6] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive text embedding benchmark, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023, pp. 2012–2037. doi:10.18653/v1/2023.eacl-main.148.

[7] V. Khrulkov, I. Oseledets, Matryoshka representation learning, in: Advances in Neural Information Processing Systems, volume 36, 2023.

[8] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 5274–5290. doi:10.18653/v1/2021.emnlp-main.429.

[9] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[10] G. V. Cormack, M. R. Grossman, Combining document rankings using reciprocal rank fusion, in: Proceedings of the 2017 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17), ACM, 2017, pp. 308–311. doi:10.1145/3121050.3121091.

[11] D. Panou, M. Reczko, Semi-supervised training for biomedical question answering, in: Proceedings of the CLEF 2023 Working Notes, CEUR-WS, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-013.pdf.

[12] D. Panou, A. C. Dimopoulos, M. Koubarakis, M. Reczko, Harnessing collective intelligence of llms for robust biomedical qa: a multi-model approach, arXiv preprint arXiv:2508.01480 (2025).

[13] X. Zhao, X. Hu, Z. Shan, S. Huang, Y. Zhou, X. Zhang, Z. Sun, Z. Liu, D. Li, X. Wei, Y. Pan, Y. Xiang, M. Zhang, H. Wang, J. Yu, B. Hu, M. Zhang, Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model, 2025. URL: https://arxiv.org/abs/2506.20923. arXiv:2506.20923.

[14] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 14b and Synergy14

in CLEF2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2025 Working Notes, 2026.