

Title. A Scalable Framework for Benchmarking Embedding Models for Semantic Medical Tasks

Author Block. Shelly Soffer¹, Benjamin S Glicksberg^{2,3}, Patricia Kovatch⁴, Orly Efros^{5,6}, Robert Freeman², Alexander W Charney², Girish N Nadkarni^{*2,3}, Eyal Klang^{*2,3}

* Indicates equal contribution

Author Affiliations.

¹ Institute of Hematology, Davidoff Cancer Center, Rabin Medical Center; Petah-Tikva, Israel.

² Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place; New York, NY 10029, United States.

³ Division of Data Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai; New York, NY 10019, United States.

⁴ Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States.

⁵ School of Medicine, Tel Aviv University; Tel Aviv, Israel.

⁶ National Hemophilia Center and Thrombosis Institute, Sheba Medical Center; Ramat Gan, Israel.

Corresponding Author.

Eyal Klang, MD

Division of Data Driven and Digital Medicine, Department of Medicine, Icahn School of
Medicine at Mount Sinai, New York, NY 10019, United States

Eyal.Klang@mountsinai.org

Word Count. 3285

Number of tables and figures. 8

ABSTRACT

Text embeddings convert textual information into numerical representations, enabling machines to perform semantic tasks like information retrieval. Despite its potential, the application of text embeddings in healthcare is underexplored in part due to a lack of benchmarking studies using biomedical data. This study provides a flexible framework for benchmarking embedding models to identify those most effective for healthcare-related semantic tasks. We selected thirty embedding models from the multilingual text embedding benchmarks (MTEB) Hugging Face resource, of various parameter sizes and architectures. Models were tested with real-world semantic retrieval medical tasks on (1) PubMed abstracts, (2) synthetic Electronic Health Records (EHRs) generated by the Llama-3-70b model, (3) real-world patient data from the Mount Sinai Health System, and the (4) MIMIC IV database. Tasks were split into 'Short Tasks', involving brief text pair interactions such as triage notes and chief complaints, and 'Long Tasks', which required processing extended documentation such as progress notes and history & physical notes. We assessed models by correlating their performance with data integrity levels, ranging from 0% (fully mismatched pairs) to 100% (perfectly matched pairs), using Spearman correlation. Additionally, we examined correlations between the average Spearman scores across tasks and two MTEB leaderboard benchmarks: the overall recorded average and the average Semantic Textual Similarity (STS) score. We evaluated 30 embedding models across seven clinical tasks (each involving 2,000 text pairs), across five levels of data integrity, totaling 2.1 million comparisons. Some models performed consistently well, while models based on Mistral-7b excelled in long-context tasks. 'NV-Embed-v1,' despite being top performer in short tasks, did not perform as

well in long tasks. Our average task performance score (ATPS) correlated better with the MTEB STS score (0.73) than with MTEB average score (0.67). The suggested framework is flexible, scalable and resistant to the risk of models' overfitting on published benchmarks. Adopting this method can improve embedding technologies in healthcare.

INTRODUCTION

Text embeddings are numerical representations of text that capture the semantic meaning of words, phrases, or entire documents in a continuous vector space¹.

Currently, most text embeddings are generated by dedicatedly trained large language models (LLMs).

Text embeddings enable key tasks like semantic search and Retrieval Augmented Generation (RAG), which have transformative potential in various fields, including healthcare.^{2,3} Despite their potential, such advanced natural language processing (NLP) methods remain largely untapped in the medical domain. Traditional approaches to handling medical text often fall short in capturing the nuanced and specialized language used in clinical settings. This gap emphasizes the necessity for robust text embedding models that can handle the complexity and diversity of medical texts.

The Multilingual Text Embeddings Benchmark (MTEB), a known general case embedding models benchmarking framework, is designed to evaluate text embedding models across multiple domains⁴. By providing standardized datasets and evaluation metrics, MTEB facilitates direct comparisons of model performance on tasks such as classification, clustering, and semantic textual similarity (STS). MTEB includes a broad array of datasets; however, its representation of the medical field remains limited, with datasets such as MedrxivClusteringP2P⁵, NFCorpus⁶, and BIOSSES⁷ being small in size. This hampers the effective comparison of the latest embedding models with healthcare-specific text. Moreover, another well-known risk of existing benchmarks for LLMs is the risk of models' overfitting on the published benchmark datasets.

This study aims to provide a flexible framework for evaluating the performance of leading text embedding models in capturing semantic similarity within medical texts. We used this framework to evaluate multiple embedding LLMs on multiple clinical and biomedical tasks.

METHODS

Overall Design

We aimed to identify and rank embedding models for semantic applications specific to the medical field. We selected thirty embedding models across different size groups from the MTEB leaderboard, available on GitHub⁸, chosen based on their relevance and demonstrated performance metrics. We designed a series of tasks simulating real-life medical semantic retrieval challenges to test models' performance in embedding medical terminology and contextual nuances. To evaluate the performance under varying data quality conditions, we designed an experiment where we deliberately introduced noise into the text pairs used. We then used Spearman rank correlation analysis to measure the relationship between the models' performance and the different levels of data integrity we created. This research was conducted with the approval of the Institutional Review Board (IRB) of the Mount Sinai Health System.

Model Selection Criteria

Our study evaluated models listed on the MTEB leaderboard. We selected open-source models available on Hugging Face with implementations via the sentence transformers library. Each model's inclusion required the presence of implementation code within its Hugging Face model card. We systematically chose five models from each of MTEB size groups: '<0.1 billion parameters', '0.1-0.25 billion parameters', '0.25-0.5 billion parameters', '0.5-1 billion parameters', '1-5 billion parameters', and '>5 billion parameters', prioritizing the best-performing models in each category. Among these, 'Bio_ClinicalBERT' served as a baseline reference⁹. This model utilizes contextual embeddings derived from Google's BERT architecture and is trained on PubMed and the MIMIC III dataset. Notably, 'Bio_ClinicalBERT' was not specifically trained for semantic embedding tasks, making it a standard benchmark against more specialized models. All included evaluated models and their characteristics are detailed in **Table 1**.

Databases

We extracted medical data from the following databases

1. PubMed. Abstracts were extracted using the MeSH terms: "Artificial Intelligence", "Machine Learning", and "Deep Learning," spanning the last five years. We ensured abstracts and keywords were non-null. Paired keywords were directly extracted from PubMed. Additionally, using Llama-3-70b, we generated search queries from a collection of PubMed abstracts (The complete prompt with the JSON directive and formatting is in **Supplementary eFigure 1**).

2. LLM Synthetic Electronic Health Records (EHR) Notes. Synthetic notes were generated using the Llama-3-70b model to create a simulated dataset of EHRs. The notes were based on three term list (**Supplementary eFigure 2**) to ensure variability in the notes. Using Llama-3-70b, we also generated search queries paired with the synthetic notes (The complete prompt with the JSON directive and formatting is in **Supplementary eFigure 3**).

3. Mount Sinai Health System (MSHS) EHR. Clinical care data from actual patients from 2023, including triage notes, chief complaints, physician notes and admission H&P notes.

4. MIMIC IV Database. This open database of de-identified medical information served as a source for chest X-ray reports and discharge notes.¹⁰ Discharge notes were summarized using Llama-3-70b (The complete prompt with the JSON directive and formatting is in **Supplementary eFigure 4**).

Embedding Tasks Overview

Table 2 outlines the data sources and configurations for our embedding tasks. Each task in our study involved 2,000 pairs of source and destination text. These pairs were organized into two categories: 'Short Tasks' for brief text interactions and 'Long Tasks' for more extended text analyses. Below, we detail each task type.

Short Tasks

PubMed (Abstracts to Queries): To link abstracts with generated queries they inspired.

PubMed (Abstracts to Keywords): To match abstracts with extracted keywords.

LLM Synthetic EHR Notes: To connect synthetic notes with corresponding search queries.

MSHS EHR: To pair triage notes with corresponding chief complaints.

MIMIC IV Chest-XR Reports: To link observations noted in the 'Findings' section with interpretations from the 'Impression' section. To ensure unique findings and impressions, we first filtered out non-significant (“normal”) impressions. A complete list of terms used to categorize “normal” findings is presented in **Supplementary eTable 1**.

Long Tasks

MIMIC IV Discharge Notes: To match a random sample of discharge notes with their summaries generated using Llama-3-70b.

MSHS ED Physician Notes: To concatenate random sample of 2000 admitted patient cases from 2023, with corresponding admission H&P notes.

The selection criteria and process for data preparation, including the use of specific search parameters and data handling rules, including detailed prompts used for Llama-3-70b implementations are available in the **supplementary materials**.

Experimental Setup

We evaluated the performance of embedding models across varying levels of data integrity. The experimental design incorporated a gradient of data integrity levels to simulate different degree of data alteration:

- **0% Integrity:** Utilized pairs with unmatched source and destination (X source - Y destination).
- **25% Integrity:** Combined 25% of the original source (X) with 75% of an alternate source (Z), paired with the original destination (X destination).
- **50% Integrity:** Mixed 50% of the original source (X) with 50% of an alternate source (Z), paired with the original destination (X destination).
- **75% Integrity:** Merged 75% of the original source (X) with 25% of an alternate source (Z), paired with the original destination (X destination).
- **100% Integrity:** Paired the original source with its corresponding original destination (X source - X destination)

For short tasks, we used the first % of the original source string and completed it with the remaining (100%) from an alternate source. For long tasks, 25% integrity meant each one sentence from the original source (1/4) was followed by three sentences from an alternate source (3/4); 50% data integrity meant two sentences from each source (2/4, 2/4), and so on. We used different techniques for adding noise because short tasks may not have enough sentences for the long method, and in long tasks, short context window models are truncated at 512 tokens, which would primarily capture the 'X' part if we used the short task method. Using this approach allowed us to conduct 10,000 comparisons per task (5x2000).

We used Spearman rank correlation to analyze the relationship between the level of data integrity and the performance metrics (cosine similarity, Euclidean difference, dot

product). This methodology was designed to evaluate how effectively each model captures and maintains the key semantic features of the source material at different integrity levels. All computations and model evaluations were performed on a dedicated MSHS server equipped with H100 80GB GPUs. The analyses were conducted using Python version 3.9.18, with additional dependencies on several key libraries: PyTorch version 2.2.2+cu121, Transformers version 4.41.2, Sentence Transformers version 3.0.1, pandas version 2.1.4, and scikit-learn version 1.3.0.

Statistical Analysis Methods

To evaluate the performance of embedding models, we applied cosine similarity, Euclidean difference, and dot product metrics to measure the similarity or distance between source and destination vectors. For each integrity level, we assessed the models using these three metrics. We then calculated the Spearman rank correlation between the integrity levels and the performance metrics to determine the robustness and reliability of the models across varying data quality levels.

We also examined correlations between the average Spearman correlation for each model across tasks and two established MTEB leaderboard scores: the model's overall recorded average score and the average Semantic Textual Similarity (STS) score.

Additionally, STS scores were correlated to average MTEB scores to contextualize our findings.

To ensure that the natural variance of Spearman scores across tasks did not skew the overall assessment of each model, we ranked the models independently for each type

of task—short, long, and overall. We used the ‘Bio_ClinicalBERT’ model as a reference, assigning it a baseline rank of one for all tasks . Other models were then ranked according to how their Spearman scores compared to this reference (sorted order). We also compared model performance between clinical and PubMed tasks.

RESULTS

Data Overview

We assessed 30 embedding models across seven clinical embedding tasks, each involving 2,000 pairs of notes at different levels of integrity, from unchanged to fully mixed. This evaluation encompassed a total of 2.1 million comparisons, calculated as 30 models * 7 tasks * 5 levels of integrity * 2,000 pairs. The number of vectors was slightly higher, as the calculations also included destination vectors. This adjustment resulted in a total of 2.52 million vectors, computed as 30 models * 7 tasks * 6 (5 level of integrity and additional destination vector) * 2,000 pairs.

Table 3 presents the variations in word counts for different clinical embedding tasks, categorized as "Short Tasks" and "Long Tasks." These categories illustrate the range of textual demands, from concise interactions to more detailed clinical texts.

Metric Efficacy Across Models and Tasks

We evaluated three metrics—cosine similarity, Euclidean difference, and dot product—across multiple embedding models and tasks to determine their efficacy in capturing semantic similarity under different integrity levels. We used Spearman rank correlation to evaluate how the metrics correlate with varying levels of data integrity, which reflect model performance. Cosine similarity emerged as the most effective metric overall in maintaining semantic integrity across integrity levels. Model-specific performances for the best-performing metric are available in **Supplementary Excel Table 4**. While most

models showed minimal variation in metric efficacy between tasks, some differences were observed in specific cases, such as 'e5-base-4k' in the 'imaging' task, 'mmlw-e5-large' in the 'PubMed query' task, and 'Bio_ClinicalBERT' in the 'QA' task.

Spearman Rank Correlation Across Models and Tasks

Table 4 provides the Spearman rank correlation values for the best-performing metric of each model across the various tasks.

MTEB Correlation

We then compared the correlations between our average task performance score (ATPS) (across all clinical and biomedical tasks), the STS score from the MTEB suite, and the overall MTEB average score **Supplementary table 2**. The STS score, which assesses model performance on tasks requiring semantic understanding akin to our studies, correlates well with the overall MTEB average. Our ATPS correlated better with the MTEB STS score (0.73) than with the MTEB average score (0.67), reflecting the role of STS as a component of the overall MTEB metric. The correlation between the STS score and MTEB score is slightly higher than ours (0.70), however, the STS score is also a component of the MTEB score, thus affecting the correlation.

Overall Top-Ranking Models

Table 5 ranks the models based on the Spearman values.

In this evaluation, the overall top-ranking models, 'GIST-large-Embedding-v0' and 'b1ade-embed' , have achieved the highest rankings across all tasks.

The 'Bio_ClinicalBERT' model served as a reference and recorded the lowest scores . This outcome is not surprising, as the model was not trained for semantic embedding tasks. Unlike models specifically designed for semantic embedding tasks, Bio-Clinical-BERT is essentially BERT fine-tuned on some bio-medical data.

Interestingly, the large-scale models (>5b parameters) based on Mistral-7b did not reach the top overall ranks despite their capacity.

The relationship between the overall models' performance and their embedding times is visually represented in **Figure 1**.

Short-Tasks Top-Ranking Models

In the evaluation of short tasks (**Table 5**), 'b1ade-embed' and 'GIST-large-Embedding-v0', maintained their high performance with scores of 27.4 and 26.6, respectively. Additionally, the 'NV-Embed-v1', a large model and a top performer on the MTEB leaderboard, obtained the highest score, 28.4, outperforming other large models like 'SFR-Embedding-Mistral' and 'Linq-Embed-Mistral' . The relationship between the models' performance in short tasks and their embedding times is represented in **Figure 2**.

Long-Tasks Top-Ranking Models

In the evaluation of long tasks (**Table 5**), the largest models with extensive context windows demonstrated dominant performance. The top four performing models in the long context tasks were >5b parameters, Mistral-7b-based models (e5-mistral-7b-instruct , SFR-Embedding-2_R , SFR-Embedding-Mistral , Linq-Embed-Mistral . Contrarily, the 'NV-Embed-v1', despite its large size, did not perform as well in long tasks, ranking second to last among the models evaluated, although this model was top ranging across three of the short tasks. The relationship between the models' performance in long tasks and their embedding times is visually represented in **Figure 3**.

Comparison of Models Across Short Clinical vs. PubMed Tasks

In our analysis of model performance across "Short Clinical Tasks" and "Short PubMed Tasks," (**Table 5**) 'NV-Embed-v1' stands out by ranking highest in both categories. Other models exhibiting strong performance in both domains include 'b1ade-embed' . Similarly, 'Linq-Embed-Mistral' and 'instructor-xl' show consistency, ranking in the top tiers for both clinical and PubMed tasks.

Conversely, 'UAE-Large-V1' performs significantly better in clinical tasks than in PubMed, whereas 'gtr-t5-xxl' shows a reversed trend, highlighting their specialized strengths in respective areas.

Discussion

This study provides a framework to assess the performance of leading text embedding models for specific medical data. Additionally, it offers valuable insights into performance on specific tasks, including short and long texts, as well as biomedical and clinical data. Our findings can help identify the most effective models for those working with medical data, providing guidance for choosing models which may best suit their tasks. Our findings reveal that text embedding models, initially benchmarked on general-domain datasets, maintain high efficacy in the medical domain. The top-performing models demonstrated robust capabilities in handling medical language and contextual nuances, evidenced by their high Spearman rank correlations across various tasks.

Models like 'GIST-large-Embedding-v0'¹¹ and 'b1ade-embed'¹² excel in both short and long tasks, demonstrating superior embedding capabilities and impressive performance despite their smaller size, making them highly efficient choices.

Large-scale models, particularly those based on Mistral-7b, demonstrated superior performance in long-context tasks. This can be attributed to their longer context windows, which allow for better handling of detailed clinical texts, while shorter context window models are truncated after 512 tokens. Conversely, 'NV-Embed-v1'¹³, despite being a large 7b model and the top performer in short tasks, did not perform as well in long tasks, indicating a potential limitation in its training for handling extended contexts. This outcome suggests that sheer model size does not guarantee superior performance

across varied tasks. The adaptability and training specificity seem to play more critical roles in determining model effectiveness across such diverse testing scenarios.

The comparison between clinical and PubMed tasks revealed that certain models have specialized strengths. For instance, 'UAE Large V1' performed better in clinical tasks, while 'sentence transformers gtr t5 xxl' excelled in PubMed tasks. This differential performance emphasizes the importance of choosing models based on the specific demands of the text type and context within medical data processing. Some models are versatile across domains, while others are tailored to specific types of medical text, providing insights into their application-specific efficacy.

We address limitations specific to healthcare in the MTEB framework. First, in the MTEB framework, the statistical measure employed assesses how well the model's outputs (similarity scores) align with actual human-assigned similarity scores⁴. In our study, we calculated the Spearman rank correlation between metrics and data integrity levels. This provides a measure of model performance stability across different data conditions and tasks. The consistency in performance across varying levels of data integrity demonstrates the models' ability to maintain semantic integrity. This is particularly crucial in medical applications, where data often varies in completeness and clarity.

Second, the MTEB framework does not predominantly focus on the medical domain. Within its leaderboard challenge, only a small number of biomedical datasets, such as BIOSSES⁷ (Biomedical Sentence Similarity Estimation System) are included. BIOSSES, consisting of 100 sentence pairs annotated with semantic similarity scores

by domain experts, may not capture the full diversity of biomedical literature. In contrast, our study employed a methodology that did not rely on human annotators, allowing us to utilize relatively large datasets varied in size and scope. Our method can provide an easy way to expand the types and sizes of the benchmarking datasets.

Third, the MTEB scores might reflect potential overfitting, as models' training data could inadvertently include the datasets used in MTEB, allowing for over-fitting on these specific tasks ¹⁴. In contrast, our study employed a diverse set of real-world new medical tasks and datasets, providing a more accurate measure of a model's performance in practical medical scenarios. Moreover, using our method, new benchmarking random datasets can be created dynamically. This approach ensures that the models are assessed based on their ability to handle a variety of medical data, rather than their performance on published benchmark tasks.

Another platform, "Papers with Code," tracks the "State-of-the-Art" (SOTA) in machine learning across a range of tasks, including specific leaderboards for medical data. However, for sentence embeddings in biomedical contexts, it also just utilizes the BIOSSES benchmark, with the above-mentioned limitations ¹⁵.

In some studies, researchers have independently developed models tailored to text embedding within medical contexts, such as BioSentVec ¹⁶⁻¹⁸. However, these individual efforts, while valuable, do not establish benchmarks against which to measure the performance of a broad array of existing models across varied medical tasks.

Our results provide a practical framework for medical professionals and researchers working with medical data and seeking to leverage NLP technologies effectively. By

identifying the optimal models for various medical tasks, we offer guidance on which models are best suited for medical applications. Additionally, several of the tasks and associated data evaluated in this study will now be available for public use, facilitating further model validation.

Our study has several limitations. Firstly, the scope of evaluated models was restricted to those available on the Hugging Face platform with implementations via the Sentence Transformers library, which may exclude potentially effective models do not present on this platform. Secondly, the generated synthetic data, particularly from the Llama-3-70b model, may not fully capture the complexity and variability of real patient data, however we did use real patient data from several databases. Additionally, our study primarily focused on English language texts, and the performance of these models on non-English medical texts remains to be explored. Furthermore, while data integrity levels were considered, the specific nature of the noise introduced may not fully capture the complexities and variations found in real-world medical data. Different types of noise (e.g., typographical errors vs. semantic errors) might affect model performance differently.

In conclusion, the suggested framework provides guidance for selecting embedding models tailored to various medical tasks. By leveraging task-specific models, we can enhance key applications such as semantic search and RAG, which, despite their potential, are still underutilized in healthcare.

References

1. Glicksberg BS, Miotto R, Johnson KW, et al. Automated disease cohort selection using word embeddings from Electronic Health Records. PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium; 2018: World Scientific; 2018. p. 145-56.
2. Glicksberg BS, Timsina P, Patel D, Sawant A. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. 2024.
3. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 2020; **33**: 9459-74.
4. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:221007316* 2022.
5. medrxiv-clustering-p2p. 2022. <https://huggingface.co/datasets/mteb/medrxiv-clustering-p2p>.
6. Boteva V, Gholipour D, Sokolov A, Riezler S. A full-text learning to rank dataset for medical information retrieval. *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016 Proceedings 38*; 2016: Springer; 2016. p. 716-22.
7. Soğancıoğlu G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 2017; **33**(14): i49-i58.
8. huggingface. MTEB leaderboard. 2024. <https://huggingface.co/spaces/mteb/leaderboard>.
9. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:190403323* 2019.
10. Johnson AE, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 2023; **10**(1): 1.
11. Solatorio AV. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:240216829* 2024.
12. w601sxs/blade-embed. 2024.
13. Lee C, Roy R, Xu M, et al. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint arXiv:240517428* 2024.
14. Mathieu Ciancone IK, Marion Schaeffer, Gabriel Sequeira, and Wissam Sibli. MTEB Leaderboard : User guide and best practices. 2024. <https://huggingface.co/blog/lyon-nlp-group/mteb-leaderboard-best-practices> (accessed March 13, 2024).
15. Sentence Embeddings For Biomedical Texts on BIOSSES. 2022. <https://paperswithcode.com/sota/sentence-embeddings-for-biomedical-texts-on>.
16. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. 2019 IEEE International Conference on Healthcare Informatics (ICHI); 2019: IEEE; 2019. p. 1-5.

17. Chang D, Balažević I, Allen C, Chawla D, Brandt C, Taylor RA. Benchmark and best practices for biomedical knowledge graph embeddings. Proceedings of the conference Association for Computational Linguistics Meeting; 2020: NIH Public Access; 2020. p. 167.
18. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data* 2019; **6**(1): 52.
19. Xiao S, Liu Z, Zhang P, Muennighof N. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:230907597* 2023.
20. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:230803281* 2023.
21. all-MiniLM-L12-v2 2021. <https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>.
22. sentence-transformers/all-MiniLM-L6-v2. 2022.
23. sentence-transformers/all-mpnet-base-v2. 2021.
24. Zhu D, Wang L, Yang N, et al. LongEmbed: Extending Embedding Models for Long Context Retrieval. *arXiv preprint arXiv:240412096* 2024.
25. Li X, Li J. Angle-optimized text embeddings. *arXiv preprint arXiv:230912871* 2023.
26. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:240205672* 2024.
27. bge-m3-custom-fr. 2024. <https://huggingface.co/manu/bge-m3-custom-fr/tree/main>.
28. Dadas S, Perełkiewicz M, Poświata R. PIRB: A Comprehensive Benchmark of Polish Dense and Hybrid Text Retrieval Methods. *arXiv preprint arXiv:240213350* 2024.
29. Su H, Shi W, Kasai J, et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:221209741* 2022.
30. Ni J, Abrego GH, Constant N, et al. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:210808877* 2021.
31. Ni J, Qu C, Lu J, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:211207899* 2021.
32. Muennighoff N. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:220208904* 2022.
33. Meng R, Liu Y, Joty SR, Xiong C, Zhou Y, Yavuz S. Sfembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog* 2024; **3**.
34. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Improving text embeddings with large language models. *arXiv preprint arXiv:240100368* 2023.
35. Junseong Kim SL, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy-yong Sohn, Chanyeol Choi. Linq-Embed-Mistral:Elevating Text Retrieval with Improved GPT Data Through Task-Specific Control and Quality Refinement. 2024. <https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral>.

36. Rui Meng* YL, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, Semih Yavuz. SFR-Embedding-2: Advanced Text Embedding with Multi-stage Training. 2024.
https://huggingface.co/Salesforce/SFR-Embedding-2_R.
37. OrdalieTech/Solon-embeddings-large-0.1. 2023.

Table 1: Model Details According to Model Size

Model size	Model	Memory Usage (GB, fp32)	Window Size (Tokens)	Model Size (Million Parameters)
<0.1GB	GIST-small-Embedding-v0 ¹¹	0.12	512	33
	bge-small-en-v1.5 ¹⁹	0.12	512	33
	gte-small ²⁰	0.12	512	33
	all-MiniLM-L12-v2 ²¹	0.12	512	33
	all-MiniLM-L6-v2 ²²	0.09	512	23
0.1-0.25GB	GIST-Embedding-v0 ¹¹	0.41	512	109
	all-mpnet-base-v2 ²³	0.41	512	110
	gte-base ²⁰	0.41	512	109
	e5-base-4k ²⁴	0.42	4096	112
	Bio_ClinicalBERT ⁹	0.41	512	110

0.25-0.5GB	mxbai-embed-large-v1 ²⁵	1.25	512	335
	UAE-Large-V1 ²⁵	1.25	512	335
	GIST-large-Embedding-v0 ¹¹	1.25	512	335
	bge-large-en-v1.5 ¹⁹	1.25	512	335
	b1ade-embed ¹²	1.25	512	335
0.5-1GB	multilingual-e5-large-instruct ²⁶	2.09	514	560
	multilingual-e5-large ²⁶	2.09	514	560
	Solon-embeddings-large-0.1	2.09	512	560
	bge-m3-custom-fr ²⁷	2.12	8192	568
	mmlw-e5-large ²⁸	2.09	514	560
1-5GB	instructor-xl ²⁹	4.62	512	1241
	sentence-t5-xl ³⁰	4.62	512	1241
	sentence-t5-xxl ³⁰	18.12	512	4865

	gtr-t5-xxl ³¹	18.12	512	4865
	SGPT-2.7B-weightedmean-msmarco-specb-bitfit ³²	10.0	2048	2685
>5GB	SFR-Embedding-Mistral ³³	26.49	32768	7111
	e5-mistral-7b-instruct ³⁴	26.49	32768	7111
	Linq-Embed-Mistral ³⁵	26.49	32768	7111
	NV-Embed-v1 ¹³	29.25	32768	7851
	SFR-Embedding-2_R ³⁶	26.49	32768	7111

Table 2: Overview of embedding tasks.

Abbreviations: LLM: Large Language Model, EHR: Electronic Health Record, MSHS: Mount Sinai Health System, CXR: Chest X-ray, ICU: Intensive Care Unit, ED: Emergency Department, H&P: History and Physical

Length	Data Source	Source	Destination
Short Tasks	PubMed	Abstracts	Search queries
	PubMed	Abstracts	Keywords
	LLM Synthetic EHR Notes	Synthetic notes	Search queries
	MSHS EHR	Triage Notes	Chief Complaints
	MIMIC IV CXR Reports	Imaging Findings	Imaging Impressions
Long Tasks	MIMIC IV EHR Notes	ICU discharge notes	LLM Generated Summaries
	MSHS EHR	ED Physician Notes	H&P Notes

Table 3: Lengths of tasks inputs

Abbreviations: EHR: Electronic Health Record, ICU: Intensive Care Unit, ED: Emergency Department, H&P: History and Physical

Length	Data Source - Destination	Source Words (Mean ± Std Dev)	Destination Words (Mean ± Std Dev)
Short Tasks	PubMed Abstract - Search queries	216.0 ± 65.7	7.5 ± 2.5
	PubMed Abstract - Keywords	215.6 ± 64.4	10.7 ± 3.7
	Synthetic EHR Notes - Search queries	158.4 ± 8.3	30.2 ± 35.8
	Triage Note - Chief Complaint	29.5 ± 14.0	2.5 ± 1.3
	Imaging Findings - Impression	82.5 ± 27.3	17.8 ± 13.8
Long Tasks	ICU discharge - Summary	1476.4 ± 272.2	76.4 ± 14.5
	ED notes - H&P notes	1294.4 ± 239.8	1504.4 ± 279.1

Table 4: Spearman Correlation Rankings and Average Task Performance Score

(ATPS). This table presents the Spearman correlation rankings for each of the 30 embedding models across the seven clinical embedding tasks. Additionally, the table includes the ATPSs, which provide an average performance indicator for each model across all tasks.

Abbreviations: ED: Emergency Department, H&P: History & Physical Notes, CXR: Chest X-ray, EHR: Electronic Health Record

Model size	Model	Discharge Notes to Summaries	ED to H&P	Triage Notes to Chief Complaints	CXR Findings to Impressions	PubMed Abstracts to Keywords	PubMed Abstracts to Queries	Synthetic EHR to Search Queries	Average Task Performance Score (ATPS)
<0.1GB	GIST-small-Embedding-v0 ¹¹	0.71	0.54	0.4	0.4	0.72	0.78	0.84	0.63
	bge-small-en-v1.5 ¹⁹	0.67	0.37	0.32	0.38	0.69	0.76	0.81	0.57
	gte-small ²⁰	0.69	0.5	0.42	0.41	0.69	0.75	0.75	0.6

	all-MiniLM-L12-v2 ²¹	0.62	0.26	0.38	0.32	0.58	0.66	0.55	0.48
	all-MiniLM-L6-v2 ²²	0.7	0.4	0.4	0.36	0.67	0.77	0.79	0.58
0.1- 0.25GB	GIST-Embedding-v0 ¹¹	0.72	0.52	0.45	0.4	0.71	0.77	0.85	0.63
	all-mpnet-base-v2 ²³	0.68	0.44	0.39	0.28	0.67	0.78	0.56	0.54
	gte-base ²⁰	0.67	0.51	0.44	0.41	0.69	0.74	0.77	0.6
	e5-base-4k ²⁴	0.67	0.55	0.26	0.31	0.66	0.68	0.77	0.56
	Bio_ClinicalBERT ⁹	0.21	0.14	0.17	0.17	0.22	0.21	0.37	0.21
0.25- 0.5GB	mxbai-embed-large-v1 ²⁵	0.74	0.49	0.44	0.38	0.73	0.79	0.86	0.63
	UAE-Large-V1 ²⁵	0.73	0.48	0.44	0.39	0.72	0.78	0.86	0.63
	GIST-large-Embedding-v0 ¹¹	0.75	0.57	0.48	0.4	0.74	0.8	0.86	0.66
	bge-large-en-v1.5 ¹⁹	0.72	0.46	0.41	0.4	0.72	0.78	0.83	0.62
	b1ade-embed ¹²	0.75	0.53	0.47	0.42	0.73	0.8	0.86	0.65
0.5- 1GB	multilingual-e5-large-instruct ²⁶	0.75	0.55	0.32	0.3	0.64	0.66	0.83	0.58

	multilingual-e5-large ²⁶	0.7	0.46	0.27	0.35	0.64	0.67	0.69	0.54
	Solon-embeddings-large-0.1 ³⁷	0.73	0.54	0.39	0.35	0.68	0.73	0.82	0.61
	bge-m3-custom-fr ²⁷	0.76	0.51	0.31	0.47	0.66	0.72	0.73	0.59
	mmlw-e5-large ²⁸	0.71	0.46	0.39	0.34	0.71	0.75	0.83	0.6
1-5GB	instructor-xl ²⁹	0.73	0.56	0.42	0.43	0.72	0.77	0.79	0.63
	sentence-t5-xl ³⁰	0.64	0.44	0.39	0.35	0.58	0.65	0.52	0.51
	sentence-t5-xxl ³⁰	0.7	0.45	0.38	0.3	0.57	0.64	0.53	0.51
	gtr-t5-xxl ³¹	0.66	0.51	0.45	0.49	0.74	0.77	0.78	0.63
	SGPT-2.7B-weightedmean-msmarco-specb-bitfit ³²	0.7	0.27	0.36	0.31	0.71	0.76	0.83	0.56
>5GB	SFR-Embedding-Mistral ³³	0.86	0.8	0.41	0.42	0.69	0.71	0.79	0.67
	e5-mistral-7b-instruct ³⁴	0.83	0.76	0.33	0.36	0.61	0.59	0.75	0.6
	Linq-Embed-Mistral ³⁵	0.86	0.77	0.41	0.43	0.7	0.77	0.75	0.67

	NV-Embed-v1 ¹³	0.35	0.21	0.48	0.5	0.77	0.79	0.84	0.56
	SFR-Embedding-2_R ³⁶	0.85	0.76	0.45	0.45	0.71	0.72	0.69	0.66

Table 5: Model Rankings Across Task Types. This table outlines the rankings of the 30 embedding models across three categories of tasks: short, long, and overall. To mitigate the impact of natural variance in Spearman correlation scores between tasks, we assigned independent rankings for each task type. The 'Bio_ClinicalBERT' model was used as a benchmark, with a baseline rank of one in all task categories. The rankings of other models were determined based on how their Spearman scores compared to this reference model.

Abbreviations: ED: Emergency Department, H&P: History & Physical Notes, CXR: Chest X-ray, EHR: Electronic Health Record

Model	Size	Discharge Notes to Summaries	ED to H&P	Triage Notes to Chief Complaints	CXR Findings to Impressions	PubMed Abstracts to Keywords	PubMed Abstracts to Queries	Synthetic EHR to Search queries	Short Tasks	Long Tasks	Overall Rank
GIST-large-Embedding-v0 ¹¹	0.25-0.5b	25	26	29	17	28	29	30	26.6	25.5	26.3
b1ade-embed ¹²	0.25-0.5b	24	20	28	23	27	30	29	27.4	22.0	25.9

mxbai-embed-large-v1 ²⁵	0.2-5.0.5b	22	14	23	14	26	27	28	23.6	18.0	22.0
instructor-xl ²⁹	1-5b	19	25	21	25	24	19	17	21.2	22.0	21.4
UAE-Large-V1 ²⁵	0.2-5.0.5b	21	13	24	16	23	25	27	23.0	17.0	21.3
GIST-small-Embedding-v0 ¹¹	<0.1b	16	22	16	19	25	26	25	22.2	19.0	21.3
Linq-Embed-Mistral ³⁵	>5b	30	29	17	26	17	20	9	17.8	29.5	21.1
NV-Embed-v1 ¹³	>5b	2	2	30	30	30	28	24	28.4	2.0	20.9
GIST-Embedding-v0 ¹¹	0.1-0.2-5b	18	19	26	18	19	18	26	21.4	18.5	20.6
SFR-Embedding-2_R ³⁶	>5b	28	27	27	27	18	10	7	17.8	27.5	20.6
gtr-t5-xxl ³¹	1-	5	17	25	29	29	21	14	23.6	11.0	20.0

	5b										
SFR- Embedding- Mistral ³³	>5b	29	30	19	24	13	9	16	16.2	29.5	20.0
bge-large- en-v1.5 ¹⁹	0.2 5- 0.5 b	17	10	18	20	22	23	20	20.6	13.5	18.6
gte-base ²⁰	0.1- 0.2 5b	6	18	22	22	14	13	13	16.8	12.0	15.4
gte-small ²⁰	<0. 1b	10	15	20	21	15	15	11	16.4	12.5	15.3
Solon- embeddings- large-0.1	0.5- 1b	20	21	13	10	12	12	19	13.2	20.5	15.3
mmlw-e5- large ²⁸	0.5- 1b	15	11	11	8	21	14	22	15.2	13.0	14.6
bge-m3- custom-fr ²⁷	0.5- 1b	26	16	4	28	9	11	8	12.0	21.0	14.6
e5-mistral- 7b-instruct ³⁴	>5b	27	28	7	13	5	2	10	7.4	27.5	13.1
all-MiniLM- L6-v2 ²²	<0. 1b	11	6	15	12	11	22	15	15.0	8.5	13.1

multilingual-e5-large-instruct ²⁶	0.5-1b	23	24	5	4	6	6	23	8.8	23.5	13.0
SGPT-2.7B-weightedmean-msmarco-specb-bitfit ³²	1-5b	12	4	8	5	20	16	21	14.0	8.0	12.3
bge-small-en-v1.5 ¹⁹	<0.1b	7	5	6	15	16	17	18	14.4	6.0	12.0
all-mpnet-base-v2 ²⁶	0.1-0.25b	9	7	12	2	10	24	5	10.6	8.0	9.9
e5-base-4k ²⁴	0.1-0.25b	8	23	2	6	8	8	12	7.2	15.5	9.6
multilingual-e5-large ²⁶	0.5-1b	13	12	3	11	7	7	6	6.8	12.5	8.4
sentence-t5-xl ³⁰	1-5b	4	8	14	9	4	4	2	6.6	6.0	6.4
sentence-t5-xxl ³⁰	1-5b	14	9	10	3	2	3	3	4.2	11.5	6.3
all-MiniLM-L12-v2 ²¹	<0.1b	3	3	9	7	3	5	4	5.6	3.0	4.9

Bio_Clinical BERT ⁹	0.1-										
	0.2	1	1	1	1	1	1	1	1.0	1.0	1.0
	5b										

Figure 1: Bubble chart illustrating the relationship between the embedding models' overall rank and their embedding time in nanoseconds (ns). Each bubble represents a different embedding model, with the size of the bubble corresponding to the model's size. 'GIST-large-Embedding-v0' and 'b1ade-embed', have achieved the highest rankings across all tasks with low embedding times.

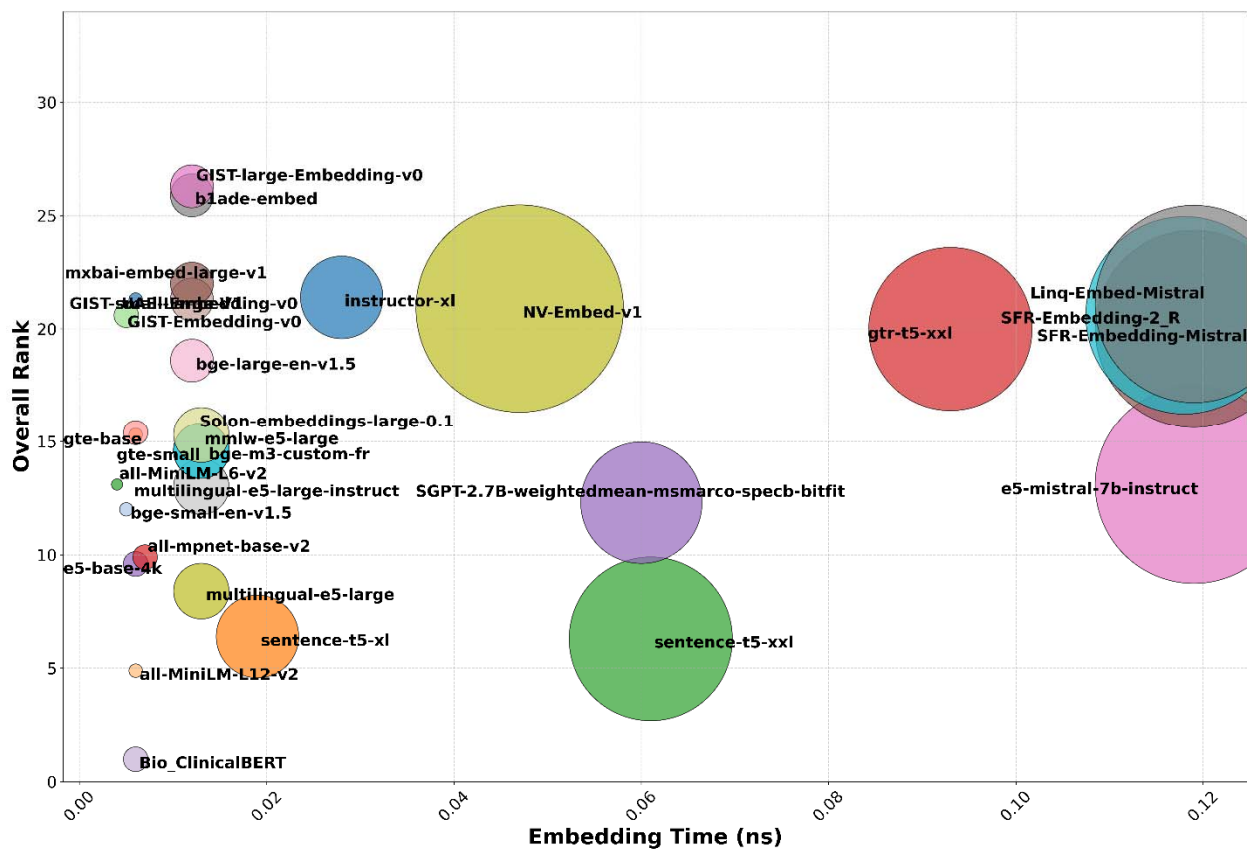


Figure 2: Bubble chart illustrating the relationship between the embedding models' rankings on short tasks and their embedding time in nanoseconds (ns). Each bubble represents a different embedding model, with the size of the bubble corresponding to the model's size. 'NV-Embed-v1', a large model, obtained the highest score. Models like GIST-large-Embedding-v0 and b1ade-embed show competitive rankings with relatively lower embedding times.

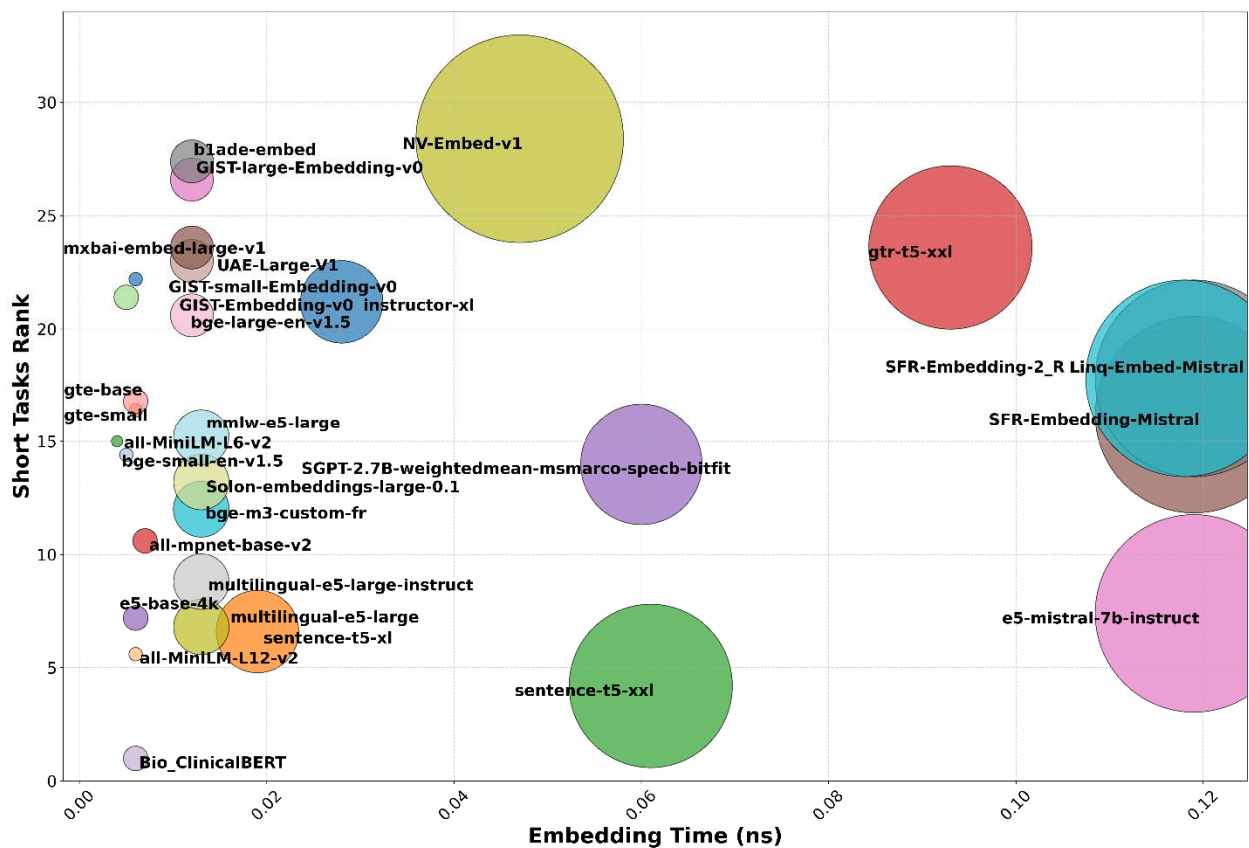


Figure 3: Bubble chart illustrating the relationship between the embedding models' rankings on long tasks and their embedding time in nanoseconds (ns). Each bubble represents a different embedding model, with the size of the bubble corresponding to the model's size. Models such as *SFR-Embedding-Mistral* and *Linq-Embed-Mistral* demonstrate strong performance with top rankings on long tasks.

