Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27908

Evaluating Embedding Models: Investigating Factors Affecting Performance Through the Medical Domain

Eson Pu

Winchester College, Winchester, Hampshire, England
Pueson@outlook.com

Abstract. In this paper, we investigated the performance of embedding models on two different domains, using models of different embedding dimensions, and even varying dimension to fully understand the impacting factors for the models. We ran tests for seven state-of-the-art models across natural questions and medial QA datasets to investigate the models' out-of-domain performance, and how factors such as the number of training parameters and fine-tuning affect accuracy and speed. We also explored the potential impact of several other variables, such as how model architecture could be the only factor affecting processing speed and how training data could have a significant effect on out-of-domain accuracy. The results show that the accuracy of the models follows a decreasing exponential relationship with the number of embedding dimensions and that the number of training parameters has little effect on model performance. The models we found to have the best accuracy were the nq-distillbert for natural questions and MedEmbed-base for medical QA. The model with the lowest encode time and the best accuracy-to-speed ratio by far across both datasets, as well as the best out-of-domain performance, was the all-miniLM.

Keywords: Embedding models, Medical domain, Performance evaluation, out-of-domain performance, natural language processing

1. Introduction

In recent years, embedding models, machine learning models that map text to numeric vectors, have become central for a wide range of tasks, including semantic search, text classification, and question answering. The effectiveness of these applications is highly dependent on the quality of the embeddings these models produce. As such, rigorous evaluation is needed to help developers select the most suitable model for their specific needs. Prior work has provided valuable insights into model performance on various tasks. For example, one study [1] evaluated the performance of six models, both independent of and in the context of specific NLP tasks and another study [2] investigated the performance of embedding models in the context of knowledge base completion. Also, it was discovered that state-of-the-art models often fail to capture domain-specific semantic meaning[3] and that constructing domain-specific word embeddings is worthwhile even if the corpus size is small [4]. Whilst these papers are significant and include detailed investigation on model performance and out-of-domain performance, they did not include detailed evaluations on other factors such as embedding dimensions, fine-tuning or encoding speed. Furthermore, many

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD27908

papers were published before the release dates of the models we choose, and could potentially be outdated in the context of recent technological developments, meaning more up to date investigations may be required. In addition to that, this paper evaluates the performance of both state-of-the-art and domain-specific embedding models in a medical domain. We will be using the natural questions dataset from sentence-transformers [5]and the medical-o1-reasoning-SFT dataset from FreedomIntelligence [6]. We chose the natural questions dataset to act as a baseline as it is the most similar to the models' training data. The medical dataset will be used to evaluate the models' out-of-domain performance, assessing their ability to perform consistently even on text that is dissimilar to their training data. The med QA dataset will also be used to evaluate the importance of model fine-tuning using models fine-tuned for medical retrieval.

Table 1. Datasets we are using.

Dataset	Rows	Av query len	Av answer len
natural-questions	100231	9.1	100.3
medical reasoning	19704	46.1	94.0

Table 2. nq sample.

query	answer
when did jack and the beanstalk take place	Jack and the Beanstalk According to researchers ...
where does the name led zeppelin come from	"Led Zeppelin The band completed the Scandinavian ...
where is the heart located in the rib cage	Thorax The anatomy of the chest can also be described ...
who sings everybody in the club get tipsy	"Tipsy (song) ""Tipsy"" is a record by American ...
how many mg of thc to test positive	Cannabis drug testing The main metabolite excreted ...

Table 3. med QA sample.

query	answer
Given the symptoms of sudden weakness ...	The specific cardiac abnormality most likely to ...
A 33-year-old woman is ...	In this scenario, the most likely anatomical ...
In a patient with dermatomyositis...	In patients with dermatomyositis, ovarian cancer...
AA 14-year-old girl is brought to the physician...	The symptoms and clinical findings in the 14-year-old ...
What is the most likely diagnosis for ...	Based on the described symptoms and the unusual ...

The models we chose were for their position on the MTEB leader-board as well as the tasks they were trained to perform.The all-MiniLM-L6-v2 model represents an efficient architecture that achieves competitive performance while maintaining computational efficiency through its reduced parameter count, making it particularly suitable for resource-constrained environments. The paraphrase-multilingual extends embedding capabilities across multiple languages and, despite

DOI: 10.54254/2755-2721/2025.LD27908

sharing the same embedding dimensionality as the all-MiniLM variant, incorporates the largest parameter set within our evaluation corpus to support its multilingual functionality. In contrast, the all-mpnet-base-v2 model prioritizes embedding quality through increased dimensionality, though this enhancement comes at the cost of reduced inference speed. The all-distilroberta-v1 model, derived through knowledge distillation from the RoBERTa architecture, demonstrates a balanced trade-off between computational efficiency and representational quality. For domain-specific applications, we include the nq-distilbert-base-v1 model, which has been specifically optimized for natural question processing tasks. Finally, we examine two medical domain models: MedEmbed-base-v0.1, which is fine-tuned from the bge-base-en-v1.5 foundation model for medical question-answering retrieval tasks, and its compact variant MedEmbed-small-v0.1, which sacrifices some representational capacity in favor of improved inference efficiency.

Table 4. Model Information.

Model	Embedding Dim	Model Size	MTEB ranking	release date
all-MiniLM-L6-v2	368	22.7M	117	Aug 2021
paraphrase-multilingual	386	118M	64	Jun 2021
all-mpnet-base-v2	786	109M	102	Aug 2021
all-distilroberta-v1	768	82.1M	n/a	Aug 2021
nq-distilbert-base-v1	786	66.4M	n/a	Feb 2021
MedEmbed-base-v0.1	786	109M	n/a	Oct 2024
MedEmbed-small-v0.1	368	33.4M	91	Oct 2024

2. Results

The results were obtained by evaluating each model on a sample from each dataset. Note that the original Natural Questions dataset contains over 100,000 training examples and the Medical dataset comprises of around 20,000 samples, which would require substantial computational resources for comprehensive evaluation. To ensure computational feasibility while maintaining statistical validity, we conducted preliminary experiments to verify that reducing the evaluation dataset size to 1,000 samples does not significantly impact the overall performance metrics, aside from yielding slightly more optimistic results. Based on these validation tests, we randomly sampled 1,000 examples from each dataset to accelerate the evaluation process without compromising the reliability of our comparative analysis.

2.1. Retrieval performance metrics

Recall@K serves as a fundamental metric for evaluating retrieval completeness, defined as the proportion of relevant items retrieved within the top-k results relative to the total number of relevant items in the corpus. This metric provides an assessment of the system's capacity to identify all pertinent documents, independent of their ranking positions within the retrieved set. Normalized Discounted Cumulative Gain (NDCG@K) and Mean Average Precision (MAP@K) constitute complementary ranking-aware metrics that prioritize the placement of relevant items at higher positions in the retrieved list. NDCG@K accommodates graded relevance judgments through continuous or multi-level relevance scores, enabling nuanced assessment of item pertinence and

DOI: 10.54254/2755-2721/2025.LD27908

making it particularly suitable for applications where relevance exists on a spectrum. Conversely, MAP@K operates under binary relevance assumptions, treating items as either relevant or non-relevant, which renders it more appropriate for scenarios requiring definitive relevance determinations and absolute precision requirements. Mean Reciprocal Rank (MRR) quantifies retrieval effectiveness by computing the reciprocal of the rank position of the first relevant item for each query, then averaging these reciprocal ranks across all test queries. Like NDCG@K and MAP@K, MRR emphasizes early retrieval performance, making it particularly sensitive to the system's ability to place at least one relevant item in high-ranking positions.

2.2. Computational efficiency metrics

To assess model computational performance, we employ several temporal metrics. Encoding time captures the cumulative duration required for models to generate embeddings for all queries and candidate documents in the evaluation set. Similarity computation time measures the milliseconds required by the ranking system to calculate pairwise similarities between query and document embeddings. Additionally, we report total processing time and throughput (measured in texts processed per second) to provide comprehensive efficiency characterization across different operational scales and deployment scenarios.

2.3. Model performance

To calculate the evaluation metrics of a model, a matrix of cosine similarities between each query and answer embeddings produced by the model is calculated where item [i, j] corresponds to the similarity between query i and answer j (Assuming for each query i, correct answer is also at index i). For each query, its list of similarity scores (calculated for each query between every individual answer) is sorted in descending order to produce the rankings where the index of the correct answer can then be found (correct answer found at a higher index represents higher model accuracy). The rank of the correct answer for every query can then be used to calculate all the relevant evaluation metrics. As for table columns, NDCG@10, MRR@10, Recall@10 and MAP are just the evaluation metrics discussed earlier, with MAP referring to MAP@K applied to the entire list. Dim refers to the dimensions of the model and will be relevant later when looking at reduced dimensions. Encode time is the sum of the time taken to encode every query and answer in each dataset. Texts per second (texts/s) is obtained from dividing total texts (queries and answers) by the encode time after adding on the relatively insignificant similarity time (sim time/ms) which is the time taken to calculate the cosine similarity of two word embeddings produced by a model. We evaluated every model on a subset of 1000 words from both the Natural Questions and Medical QA dataset. The results are listed in the two tables below.

DOI: 10.54254/2755-2721/2025.LD27908

Table 5. Combined Model performance across Natural Questions (blue) and Med QA (red) (1000 subset).

Model	dim	NDCG@10	MRR@10	Recall@10	MAP	encode time(s)	sim time/ms
all-MiniLM	384	0.971/0.942	0.962/0.93	0.996/0.976	0.962/0.931	94.45/123.82	15/23
mpnet-base	768	0.982/0.946	0.976/0.935	0.999/0.979	0.976/0.935	747.38/1012.0	36/32
distilroberta	768	0.979/0.926	0.972/0.911	0.999/0.974	0.972/0.912	366.22/478.94	53/29
multilingual	384	0.892/0.855	0.873/0.838	0.952/0.908	0.875/0.841	128.89/194.69	30/16
nq-distilbert	768	0.988/0.777	0.984/0.755	1.000/0.843	0.984/0.760	350.22/482.51	30/28
MedEmbed-base	768	0.975/0.971	0.967/0.964	0.996/0.99	0.967/0.965	705.47/963.09	31/40
MedEmbed-small	384	0.962/0.966	0.952/0.958	0.990/0.991	0.952/0.959	200.61/278.67	21/18

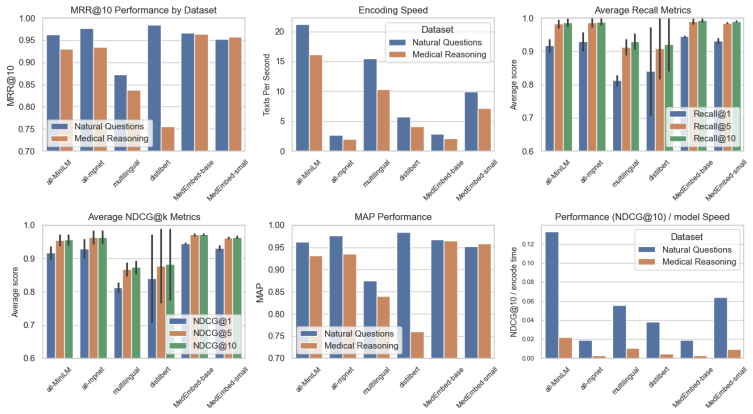

Figure 1. Model performance original dim (1k).

Surprisingly enough, the model that had the best performance on the natural questions dataset, nq-distillbert-base, actually had the worst performance by far on the medical QA dataset. This is most likely due to the fact that the model was trained explicitly on natural questions, giving it little generalisability across unseen data. Paraphrase-multilingual has the most training parameters of all the models but consistently performed poorly across both datasets. This could be attributed to the fact that as it is a multilingual model, most of its parameters are from foreign languages other than English as the model is designed to work with many different languages. This coupled with its relatively small embedding dimensions provides a possible explanation as to why the embeddings it produced are not as accurate as other, smaller models with the same embedding dimensions.

2.4. In- and out-of-domain performance

The models (excluding fine-tuned models) which had the most similar NDCG@10 score across the two datasets (in order) is all-MiniLM with a 0.029 difference, all-mpnet-base with a 0.036 difference, paraphrase-multilingual with 0.037, all-distillroberta with 0.053 and nq-distillbert-base

DOI: 10.54254/2755-2721/2025.LD27908

with a massive 0.211 difference. Although the MedEmbed models technically had the greatest generalisability (both models only had a 0.004 difference in NDCG@10), they were both models trained on natural questions and fine-tuned for medical QA; they might not perform as well on other unseen datasets. Fine-tuning affects the quality of embeddings very significantly. Whereas the MedEmbed-base and MedEmbed-small both had mediocre accuracy scores on natural questions (ranking 4th and 6th respectively), they had the best performance by far on the Med QA dataset (ranking first and second despite the MedEmbed-small model's few embedding dimensions and training parameters). Not only that, whereas the NDCG@10 of every single non-fine-tuned model decreased significantly from natural questions to medical QA, the performance of the MedEmbed models did not change much, with the NDCG@10 score of MedEmbed-base being reduced by 0.004 and increased by 0.004 for MedEmbed-small.

2.5. Impact of varying embedding dimensions

Embedding dimensions clearly has a significant effect on model performance. Whereas a vector of 384 dimensions could capture the semantic meaning of a sentence, the same cannot be done with just a single number. Across the natural questions dataset, every single model with 768 embedding dimensions outperformed those with 384 dimensions. This is to be expected, as more embedding dimensions allow for a more detailed representation of the semantic meaning of a sentence, thus allowing for more accurate rankings. The dimensions of the word embeddings of a model can be truncated to a desired dimension by passing the predicted vector embedding through a final pooling layer before being returned as the output. This allows the semantic meaning of text to be captured in less dimensions, but sacrifices data in the process. Thus, the quality of the embeddings with reduced dimensions is highly dependent on the quality of the original embeddings. As such, the rankings of the models with half of their original dimensions through NDCG@10 is the exact same as the rankings for models with original dimensions across both the natural questions dataset and the medical QA datasets. The only difference is that every single model performed slightly worse with their dimensions halved. Although that is to be expected, it is surprising how little change there was to their accuracy metrics (the all-mpnet-base had less than 0.0005 difference in NDCG@10). However, the differences are slightly more pronounced when the models' dimensions were reduced to 1/4 of their original dimensions.

Table 6: Model performance across Natural Questions and Med QA datasets (1000 subset, 1/2 dim).

Model	dim	NDCG@10	MRR@10	Recall@10	MAP	encode time(s)	sim time/ms
all-MiniLM	192	0.967/0.928	0.958/0.914	0.992/0.969	0.959/0.915	82.26/106.72	8/6
mpnet-base	384	0.982/0.944	0.976/0.933	0.999/0.979	0.976/0.933	647.53/846.4	12/16
distilroberta	384	0.978/0.918	0.971/0.901	0.998/0.968	0.971/0.902	316.91/407.26	12/14
multilingual	192	0.87/0.847	0.845/0.829	0.945/0.902	0.847/0.832	113.46/164.84	9/8
nq-distilbert	384	0.987/0.767	0.983/0.745	1.0/0.835	0.983/0.749	304.97/404.6	14/14
MedEmbed-base	384	0.974/0.97	0.966/0.964	0.997/0.988	0.966/0.965	592.58/796.43	14/16
MedEmbed-small	192	0.95/0.956	0.939/0.947	0.985/0.985	0.939/0.947	174.2/242.06	9/10

DOI: 10.54254/2755-2721/2025.LD27908

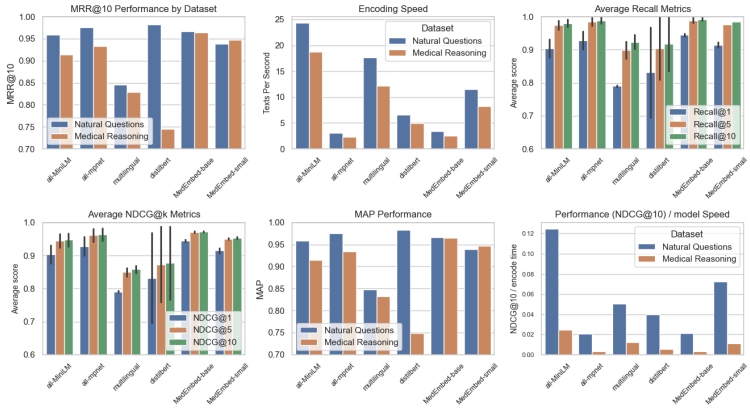

Figure 2. Model performance truncated dim (1000 subset, 1/2 dim).

Table 7. Model performance across Natural Questions and Medical QA (1000 subset, 1/4 dim).

Model	dim	NDCG@10	MRR@10	Recall@10	MAP	encode time(s)	sim time/ms
all-MiniLM	96	0.95/0.91	0.94/0.89	0.99/0.96	0.94/0.89	80.45/106.71	7/5
mpnet-base	192	0.98/0.94	0.97/0.92	1.00/0.98	0.97/0.92	652.81/854.77	8/7
distilroberta	192	0.97/0.90	0.97/0.87	1.00/0.96	0.97/0.88	316.40/404.71	9/9
multilingual	96	0.83/0.82	0.80/0.80	0.92/0.88	0.81/0.81	114.44/164.41	4/4
nq-distilbert	192	0.99/0.75	0.98/0.72	1.00/0.82	0.98/0.73	291.04/396.34	9/9
MedEmbed-base	192	0.96/0.96	0.95/0.95	0.99/0.99	0.95/0.95	592.57/793.35	8/8
MedEmbed-small	96	0.92/0.94	0.90/0.93	0.97/0.97	0.90/0.93	172.11/235.96	5/4

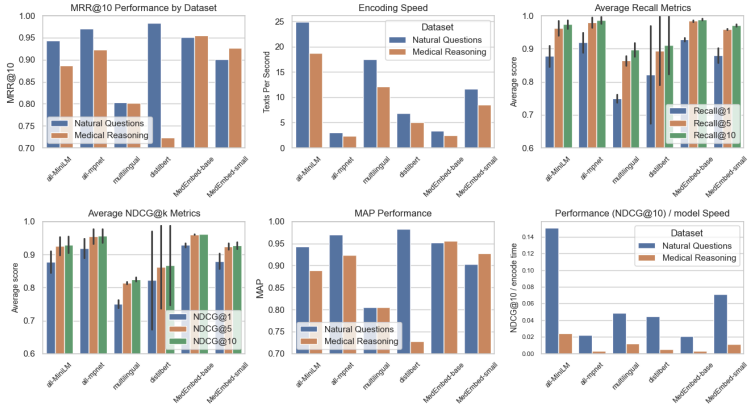

Figure 3. Model performance truncated dim (1000 subset, 1/4 dim).

We further investigated the relationship between reduced embedding dimensions and model accuracy by evaluating all-MiniLM across the two datasets with a range of embedding dimensions. We chose the all-MiniLM as it was the model with the highest out-of-domain performance and encoding speed whilst also having the best accuracy out of the 384 dim models, allowing us to use a

DOI: 10.54254/2755-2721/2025.LD27908

much larger subset of data whilst maintaining similar performance across the two datasets. The results obtained from a 10000 word subset from both datasets showed a decreasing exponential relationship.

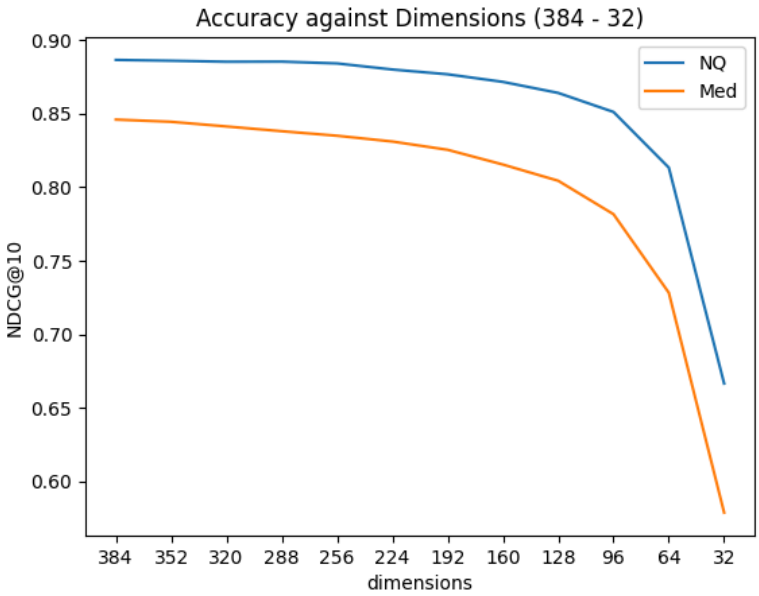

Figure 4. All-MiniLM performance (10,000 subset)/ dimensions (384 - 32).

It is not surprising that the accuracy score is lower for natural questions than for Medical QA, as the model consistently performed better on natural questions in previous evaluations. The rate of decrease of accuracy with respect to the number of dimensions is also higher for Med QA than it is for Natural questions as Med QA has longer queries so having fewer dimensions would have more impact on the accuracy of rankings. The relationship between vector dimensions and accuracy can be highlighted by how the larger models with halved embedding dimensions (768 truncated to 384) still perform better than models with original embeddings of the same size (384 dimensions). When evaluated on natural questions, the worst performing model with a halved embedding dimension of 384, MedEmbed-base, achieved a NDCG@10 score of 0.974 whereas the best performing model with an original embedding dimension of 384, all-MiniLM-L6-v2, achieved a lower NDCG@10 score of 0.971. This shows how the texts did not need a large (768 - 384) number of dimensions to be accurately captured and that reducing embedding dimensions only significantly affects accuracy when there are not enough dimensions to accurately capture all the data.

2.6. Impact of model parameter size

The size of the model does affect model accuracy to some extent but not as much as one might think, according to the results of our evaluation. The all-mpnet-base is a large model with 109M parameters and outperforms the smaller smaller models, all-MiniLM and all-distillroberta, across both natural questions and medical QA. However, this is where the correlation ends, as although all-distillroberta outperformed all-MiniLM across natural questions, all-MiniLM outperformed all-distillroberta on medical QA despite having half the embedding dimensions and around 4 times less training parameters (22.7M compared to 82.1M). Not only that, the MedEmbed-base performed reactively poorly on natural questions and was ranked 4th in terms of NDCG@10 despite having the same amount of training parameters as the all-mpnet-base which came second. However, the fact

DOI: 10.54254/2755-2721/2025.LD27908

that MedEmbed-base was fine-tuned for medical QA should be considered as that could have negatively impacted its performance on natural questions through the introduction of noise. The nq-distillbert-base was also able to achieve the best results on natural questions despite being the third smallest model with 66.4M parameters, although being trained explicitly for natural questions may have given it an inherent advantage over other models. Similarly for paraphrase-multiligual, although it is the largest model with 118M training parameters, it is a multilingual model designed to work with multiple language meaning only a fraction of its training data was in English, possibly being the main reason it ranked second to last and last in both natural questions and medical QA. Fine-tuning is shown to have a much greater effect than model size on model accuracy, as the MedEmbed-small ranked second in medical QA, outperforming every single other model despite being the second smallest model with 33.4M parameters (the smallest model actually ranked third, right after MedEmbed-small). Also, if nq-distillbert-base can be considered a fine-tuned model by how it is trained for natural questions, it also outperformed much larger models on the natural questions dataset.

2.7. Factors affecting model speed

When considering model performance, the speed at which models encode data should be considered along with model accuracy. In the case of a search engine where the rankings do not need to be extremely accurate, some accuracy is sacrificed in exchange for embeddings to be produced much more quickly, allowing relevant results to be displayed in less time. Embedding dimensions are shown to be correlated with model speed as in both the natural questions and medical QA datasets, the models with an embedding dimension of 384 had a smaller encode time than the ones with 768 dimensions. In addition, less time is taken to calculate the cosine similarity of vectors with fewer dimensions as the calculations are performed faster for each embedding. However, halving the embedding dimensions has the surprising effect of reducing the encode time. This is unexpected as when reducing the embedding dimension, the vector representation with the original embeddings is sent through a pooling layer to get the truncated embeddings. This is an extra added step at the end of the encoding process, meaning that the total encode time should be roughly the same, if not longer. One possible explanation for this could be that, as the tests were carried out on Google colab, varying server load and user volume could have affected resource allocation to the session, and thus the encoding speed would be more inconsistent. The fact that reducing dimensions should have no effect on encode time is also reflected by how the models with their dimensions reduced fourfold have almost identical encode times as the models with their dimensions reduced by a half. Not only that, the time taken to calculate the cosine similarity of word embeddings also varies significantly, although the inconsistencies in server processing speed may be amplified by the small timescale of the measurements. However, it still remains that models with smaller original embeddings have faster encoding times and similarity calculation times. This difference can be attributed to model architecture: models with fewer original embedding dimensions are most likely less complex models with fewer, simpler layers, requiring less computational resources, and thus taking less time to compute outputs. The size of the model does seem to have some effect on the encode speed. The fastest model, all-MiniLM also has the least training parameters whereas the slowest model, all-mpnet-base, has the most (English) training parameters. However, at the same time, paraphrase-multilingual, despite having the most training parameters of all the models, had the second lowest encode time, lower even than MedEmbed-small which has the same embedding dimension and one third the training parameters. Not only that, the nq-distillbert-base has a lower encode time than the all-distillroberta despite having fewer training parameters. Thus, model size might not have any

DOI: 10.54254/2755-2721/2025.LD27908

effect on encoding speed. It might simply be the case that there is a correlation between model complexity and size and that the differences in model speed are only attributed to different model architecture. Fine-tuning and training data do not seem to have any major effect on model speed. The aforementioned inconsistencies in the relationship between model size and speed were highlighted through MedEmbed-small, nq-distillbert-base and paraphrase-multilingual, which are all models trained on specific datasets, fine-tuned, or trained in different languages. However, the relative encoding times of the models do not change even in the medical domain, possibly suggesting that the only reason for different encode speeds is differences in model architecture. This is supported by the fact that smaller embedding dimensions correspond to a shorter encode time, as original embedding dimensions are determined by the layers used in the model: Less output dimensions correlate to simpler model architecture.

3. Conclusion

In this paper, we compared the model performance in both NA and medical datasets, with the aim of understanding the effects of different factors in choosing the best model in both general and domain-specific applications. Although model size does seem to have some influence over model accuracy, the results were not consistent and other factors such as model architecture, training data, and fine-tuning seem to play a more significant role. Truncating embedding dimensions negatively impacts model accuracy with a decreasing exponential relationship. This indicates that model accuracy is only notably affected when there are not enough dimensions to capture sufficient semantic data. As for encode speed, the only significant factor that affects it seems to be the model architecture. The number of training parameters does seem to have some effect on encode speed but the results were inconsistent. The same can be said for truncating embedding dimensions: although the results showed slightly lower encode speeds, the differences are most likely due to technical inconsistencies with Google colab. However, it can be concluded that the data used to train the models and fine-tuning does not have any significant effect on encode speed. What does affect encode speed, however, is the text that is used during the evaluation. Unsurprisingly, it takes more time to encode longer texts, and as the answers and queries were much longer for medical QA than natural questions, it explains the differences in encode times across the two datasets. The model with the best performance-to-speed ratio by far was the all-miniLM (this is true for both the natural questions and medical QA datasets). The all-miniLM was able to achieve a surprisingly high accuracy and the best out-of-domain performance, displaying the lowest encode time out of all the models despite having the fewest training parameters and embedding dimensions. However, if accuracy is the priority, then models trained on specific datasets or fine-tuned models like nq-distillbert-base and MedEmbed-base would be more suitable (or all-mpnet-base as a more well rounded model, being the general model with the highest accuracy across both domains). Although NDCG@10 was the only accuracy measure considered in the evaluation, the results do not change even when looking at other metrics such as MRR@10 or MAP. This evaluation could be improved by further evaluating the models across even more domains to gauge out-of-domain performance more accurately, especially for the MedEmbed models (which were fine-tuned from general state-of-the-art models). In addition, we could have used larger subsets of data for more accurate results. The evaluation also did not cover all aspects of embedding models, as, for instance, we only covered traditional, encoder-only embedding models and did not investigate decoder-only models (like the text-embedding models from Openai and Gemini models from Google) which could potentially outperform their encoder-only counterparts. [1]Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019a). Evaluating word embedding models: Methods and experimental results. APSIPA

DOI: 10.54254/2755-2721/2025.LD27908

transactions on signal and information processing, 8:e19. [2]Wang, Y., Ruffinelli, D., Gemulla, R., Broscheit, S., and Meilicke, C. (2019b). On evaluating embedding models for knowledge base completion. [3]Tang, Y. and Yang, Y. (2025). Do we need domain-specific embedding models? an empirical investigation. [4]Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). [5]Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. [6]Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., and Wang, B. (2024). Huatuogpt-o1, towards medical complex reasoning with llms. arXiv preprint arXiv:2412.18925.

References

[1] Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019a). Evaluating word embedding models: Methods and experimental results. APSIPA transactions on signal and information processing, 8: e19.

[2] Wang, Y., Ruffinelli, D., Gemulla, R., Broscheit, S., and Meilicke, C. (2019b). On evaluating embed¬ ding models for knowledge base completion.

[3] Tang, Y. and Yang, Y. (2025). Do we need domain-specific embedding models? an empirical investi¬ gation.

[4] Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of domain-specific word em¬ beddings using knowledge resources. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

[5] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert net¬ works. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[6] Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., and Wang, B. (2024). Huatuogpt-o1, towards medical complex reasoning with llms. arXiv preprint arXiv: 2412.18925.