Jurnal Pengembangan Rekayasa dan Teknologi
Vol.10, No.1, Mei 2026, pp. 24 – 32
p-ISSN: 1410-9840 & e-ISSN: 2580-8850
https://journals.usm.ac.id/index.php/jprt
DOI : https://doi.org/10.26623/jprt.v10i1.10528

Evaluasi Komparatif Model Embedding untuk Pencarian Semantik Dokumen Institusional Indonesia

Kasmawaru*[1], Nurhaedar[2], Rachmat[3], Muh. Rafli R[4]
Universitas Dipa Makassar[1,4], Politeknik Lembaga Pendidikan Dan Pengemabangan Profesi Indonesia Makassar [2], Universitas Pejuang Republik Indonesia[3]
kasmawaru@undipa.ac.id[1] , nurnurhaedar@gmail.com[2] , rachmat27udinus@gmail.com[3], raflir180478@gmail.com[4]

Informasi Artikel

Dikirim :20-03-2026
Direview :10-04-2026
Diterbitkan :30-05-2026

Kata Kunci

Pencarian Semantik, Model Embedding, Dokumen Institusional, Bahasa Indonesia, Retrieval

Abstrak

Penelitian ini menyajikan evaluasi komparatif terhadap berbagai model embedding dalam konteks pencarian semantik pada dokumen institusional berbahasa Indonesia. Permasalahan yang dikaji berkaitan dengan keterbatasan sistem pencarian tradisional berbasis kata kunci yang sering gagal mengambil dokumen relevan karena ketidaksesuaian kosakata antara kueri pengguna dan redaksi dokumen. Penelitian ini membandingkan Sentence-BERT, Multilingual E5, dan M3-Embedding pada skenario pengambilan dokumen Indonesia. Hasil evaluasi menunjukkan bahwa M3-Embedding (All) mengungguli model lain pada nDCG@10 dan Recall@100, sehingga pemilihan model perlu didasarkan pada evaluasi kontekstual dan spesifik domain.

1. PENDAHULUAN

Penelitian dalam Jurnal Pengembangan Rekayasa dan Teknologi sebelumnya juga menunjukkan bahwa sistem temu kembali informasi berbasis kata kunci dapat digunakan untuk mendukung pencarian judul tugas akhir, meskipun pendekatan tersebut masih berfokus pada pencocokan leksikal dan belum memanfaatkan representasi semantik berbasis embedding (Christioko & Daru, 2018).

Perkembangan berikutnya menghadirkan model embedding multibahasa yang lebih maju, seperti Multilingual E5 (Wang et al., 2024a) dan BGE-M3 atau M3-Embedding (Chen et al., 2024), yang dirancang untuk mendukung pengambilan lintas bahasa, pengambilan multi-tugas, dan teks dengan panjang yang bervariasi. Dalam literatur pengambilan informasi yang lebih luas, arsitektur dense retrieval seperti DPR, RocketQA, dan Contriever juga menunjukkan bahwa pembelajaran representasi dapat meningkatkan efektivitas pengambilan secara signifikan dibandingkan pendekatan yang sepenuhnya leksikal, terutama ketika kueri dan dokumen tidak memiliki bentuk permukaan yang sama (Karpukhin et al., 2020; Qu et al., 2021; Izacard et al., 2021). Namun, kemajuan tersebut tidak secara otomatis menyelesaikan persoalan pemilihan model. Semakin banyaknya alternatif justru memunculkan pertanyaan metodologis baru mengenai model mana yang paling sesuai dengan karakteristik dokumen institusional Indonesia.

Penelitian-penelitian sebelumnya telah menyediakan landasan evaluasi yang kuat melalui tolok ukur berskala besar seperti BEIR dan MTEB. Thakur et al. (2021) mengembangkan BEIR untuk mengevaluasi generalisasi zero-shot model pengambilan informasi pada beragam domain, sedangkan Muennighoff et al. (2023), melalui MTEB, menunjukkan bahwa kinerja model embedding sangat bergantung pada jenis tugas, dataset, dan bahasa yang digunakan.

Kebutuhan terhadap pencarian semantik menjadi semakin penting pada pengelolaan dokumen institusional, seperti regulasi, surat keputusan, pedoman akademik, laporan kebijakan, dan dokumen administratif lainnya. Dokumen semacam ini umumnya memiliki struktur formal, istilah teknis, serta variasi redaksi yang tidak selalu identik dengan kata kunci yang digunakan oleh pengguna. Kondisi tersebut menyebabkan sistem pencarian berbasis pencocokan kata secara langsung berpotensi gagal menemukan dokumen relevan meskipun maknanya sesuai dengan kueri. Pendekatan berbasis embedding menawarkan solusi karena merepresentasikan teks ke dalam ruang vektor semantik, sehingga hubungan makna antara kueri dan dokumen dapat ditangkap secara lebih baik dibandingkan metode leksikal murni. Temuan pada pengembangan dense retrieval menunjukkan bahwa representasi semantik mampu meningkatkan efektivitas pencarian ketika terdapat kesenjangan kosakata antara kueri dan dokumen.

Meskipun berbagai model embedding telah dikembangkan, pemilihan model terbaik tidak dapat dilakukan hanya berdasarkan popularitas atau ukuran model. Hasil evaluasi pada benchmark lintas tugas menunjukkan bahwa tidak ada satu model embedding yang selalu unggul pada seluruh jenis tugas, bahasa, dan domain penggunaan. Kinerja model perlu dinilai berdasarkan konteks evaluasi yang spesifik agar rekomendasi model menjadi lebih tepat. Dalam konteks bahasa Indonesia, kebutuhan ini menjadi semakin relevan karena sumber daya evaluasi retrieval multibahasa mulai tersedia melalui dataset seperti MIRACL, yang menyediakan kueri dan anotasi relevansi untuk berbagai bahasa, termasuk bahasa Indonesia. Dataset tersebut dapat digunakan sebagai landasan awal untuk menguji kualitas model retrieval sebelum diterapkan pada dokumen institusional yang lebih khusus.

Berdasarkan kondisi tersebut, penelitian ini diarahkan untuk mengevaluasi secara komparatif performa beberapa model embedding yang relevan bagi pencarian semantik berbahasa Indonesia. Model yang dibandingkan mencakup pendekatan retrieval klasik, dense retrieval generasi awal, serta model embedding multibahasa modern seperti multilingual-e5-large dan M3-Embedding. Evaluasi ini penting karena multilingual-e5 dikembangkan melalui pelatihan skala besar pada pasangan teks multibahasa, sedangkan M3-Embedding dirancang untuk mendukung fungsi retrieval yang lebih luas, termasuk dense, sparse, dan multi-vector retrieval dalam lebih dari seratus bahasa. Dengan demikian, studi ini diharapkan dapat memberikan dasar empiris awal dalam menentukan model yang lebih sesuai untuk pengembangan sistem pencarian semantik dokumen institusional Indonesia.

2. METODOLOGI

Instrumen penelitian terdiri atas korpus dokumen institusional, kumpulan kueri, pedoman anotasi relevansi, dan perangkat komputasional yang digunakan untuk pengambilan serta evaluasi. Instrumen teknis mencakup model embedding yang dibandingkan, komponen pengindeksan dan pengambilan vektor, serta skrip berbasis Python untuk menghitung metrik pengambilan. Dalam eksperimen pengambilan informasi yang dapat direproduksi, kerangka kerja seperti Pyserini relevan karena mendukung

implementasi dan evaluasi alur pengambilan sparse, dense, dan hybrid secara transparan dan dapat direplikasi (Lin et al., 2021). Instrumen nonkomputasional adalah lembar anotasi relevansi yang digunakan untuk membangun dataset ground-truth. Validitas evaluasi bergantung pada keterwakilan korpus, kejelasan kueri, dan konsistensi label relevansi.

Penelitian ini menggunakan pendekatan deskriptif-evaluatif dengan desain komparatif. Fokus utama penelitian adalah membandingkan performa beberapa model pengambilan informasi dan model embedding pada konteks pencarian dokumen berbahasa Indonesia. Penelitian tidak bertujuan membangun model baru, melainkan menilai kecenderungan performa model yang telah tersedia melalui hasil benchmark yang dapat dipertanggungjawabkan secara ilmiah. Pendekatan ini relevan untuk memberikan dasar pemilihan model sebelum diterapkan pada sistem pencarian dokumen institusional yang lebih spesifik.

Sumber data evaluasi berasal dari hasil benchmark MIRACL pada subset bahasa Indonesia. MIRACL merupakan dataset pengambilan informasi multibahasa yang dirancang untuk menilai kemampuan retrieval monolingual melalui korpus, kueri, dan penilaian relevansi yang dikembangkan secara sistematis. Penggunaan subset bahasa Indonesia dipilih karena penelitian ini berfokus pada kemampuan model dalam menangani retrieval pada teks berbahasa Indonesia. Meskipun korpus MIRACL berbasis Wikipedia, dataset tersebut tetap relevan sebagai benchmark eksternal awal untuk melihat perbandingan performa model retrieval sebelum dilakukan pengujian lanjutan pada dokumen institusional Indonesia.

Model yang dibandingkan dalam penelitian ini terdiri atas enam pendekatan, yaitu BM25, mDPR, mContriever, multilingual-e5-large, M3-Embedding (Dense), dan M3-Embedding (All). BM25 digunakan sebagai baseline retrieval leksikal, sedangkan mDPR dan mContriever merepresentasikan pendekatan dense retrieval generasi sebelumnya. multilingual-e5-large dan M3-Embedding dipilih karena keduanya merupakan model embedding multibahasa modern yang dirancang untuk mendukung pengambilan informasi lintas bahasa dan pencarian semantik. Pemilihan variasi model tersebut memungkinkan analisis yang lebih luas antara pendekatan leksikal, dense retrieval awal, dan embedding multibahasa mutakhir.

Analisis kinerja dilakukan dengan membandingkan dua metrik utama yang tersedia pada hasil benchmark, yaitu nDCG@10 dan Recall@100. Nilai nDCG@10 digunakan untuk menilai kemampuan model menempatkan dokumen relevan pada posisi peringkat teratas. Nilai Recall@100 digunakan untuk melihat kemampuan model mengambil sebanyak mungkin dokumen relevan pada seratus hasil teratas. Kedua metrik tersebut penting karena sistem pencarian yang baik tidak hanya harus menemukan dokumen relevan, tetapi juga menempatkannya pada urutan yang mudah dijangkau oleh pengguna. Perbandingan dilakukan secara deskriptif melalui pemeringkatan nilai antar-model dan analisis selisih performa pada masing-masing metrik.

Prosedur analisis dilakukan melalui tiga tahap. Tahap pertama adalah identifikasi model yang dievaluasi dalam benchmark MIRACL subset bahasa Indonesia. Tahap kedua adalah pengumpulan nilai performa masing-masing model berdasarkan metrik nDCG@10 dan Recall@100. Tahap ketiga adalah interpretasi komparatif untuk menentukan model dengan performa terbaik dan menjelaskan kecenderungan perbedaan antara baseline leksikal, dense retrieval, dan embedding multibahasa modern. Hasil dari tahapan ini kemudian disajikan melalui tabel dan grafik agar pola perbandingan performa model dapat dibaca secara lebih jelas.

Gambar 1. FlowChart Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

Dataset ini secara khusus dirancang untuk mendukung evaluasi pengambilan informasi lintas bahasa yang beragam menggunakan kueri berbahasa asli dan anotasi relevansi, sehingga menjadi benchmark eksternal yang kuat untuk analisis pengambilan informasi berbahasa Indonesia (Zhang et al., 2023). Meskipun MIRACL dibangun berdasarkan Wikipedia, bukan korpus institusional, dataset ini menyediakan benchmark eksternal yang andal untuk membandingkan model embedding berorientasi pengambilan dalam pengaturan pengambilan informasi berbahasa Indonesia.

1.Precision@k

Precision pada posisi k mengukur berapa banyak dokumen relevan yang ditemukan dalam hasil pencarian top-k. Rumusnya adalah:

$$\text{Precision@k} = \frac{|\text{Rel} \cap \text{Top-k}|}{\text{k}} \qquad (1)$$

Hasil pengukuran ini akan disajikan dalam bentuk tabel atau grafik yang menunjukkan nilai Precision pada berbagai nilai k (misalnya, k = 1, k = 5, dan k = 10). Misalnya, pada Tabel 1, sebagaimana ditunjukkan dalam artikel, Precision@k dapat dilaporkan untuk setiap model embedding yang diuji.

2.Recall@k

Recall pada posisi k mengukur berapa banyak dokumen relevan yang ditemukan dibandingkan dengan jumlah total dokumen relevan yang tersedia. Rumusnya adalah:

$$\text{Recall@k} = \frac{|\text{Rel} \cap \text{Top-k}|}{|\text{Rel}|} \qquad (2)$$

Hasil ini juga akan disajikan dalam bentuk tabel atau grafik yang menggambarkan nilai Recall pada berbagai posisi top-k untuk setiap model embedding yang diuji. Perbandingan ini memberikan gambaran mengenai seberapa lengkap hasil pencarian yang dihasilkan oleh setiap model embedding.

3.Mean Reciprocal Rank (MRR)

MRR mengukur rata-rata reciprocal rank dari dokumen relevan pertama yang ditemukan. Rumusnya adalah:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} i \ \frac{1}{\text{rank}_i} \qquad (3)$$

Nilai MRR disajikan untuk setiap model embedding dan digunakan untuk membandingkan kinerja model dalam menemukan dokumen relevan secara lebih cepat. Nilai MRR yang lebih tinggi menunjukkan bahwa model lebih baik dalam menampilkan dokumen relevan pada posisi peringkat yang lebih tinggi.

4.Normalized Discounted Cumulative Gain (nDCG@k)

nDCG mengukur relevansi kumulatif dari dokumen yang diambil dengan memberikan bobot lebih tinggi pada hasil pencarian relevan yang muncul pada posisi peringkat yang lebih tinggi. Rumusnya adalah:

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \qquad (4)$$

Hasil nDCG akan dibandingkan antar-model yang diuji dan digunakan untuk menunjukkan seberapa baik setiap model memprioritaskan dokumen relevan dalam hasil pencarian.

5.Perbandingan Model

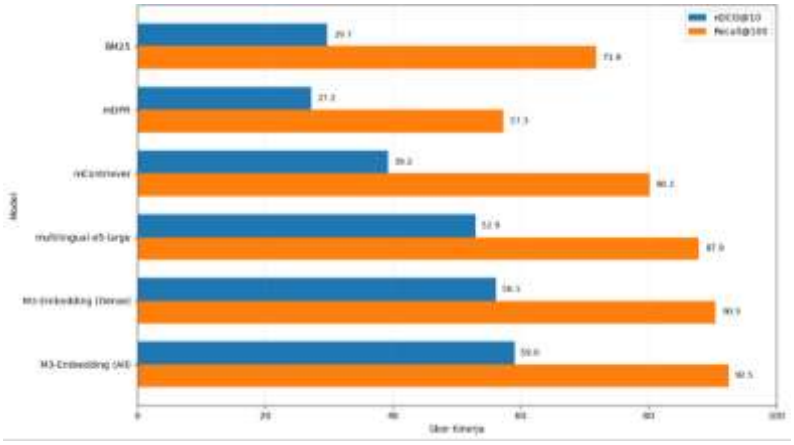
Perbandingan kinerja antara dua model dapat dilakukan dengan menghitung selisih antara metrik yang diukur, sebagai berikut:

$$\text{Selisih nDCG@k} = \text{nDCG@k(Model 1)} - \text{nDCG@k(Model 2)} \qquad (5)$$

Tabel 1 menyajikan kinerja pengambilan informasi yang telah dipublikasikan dari beberapa model baseline dan model embedding state-of-the-art pada irisan bahasa Indonesia MIRACL. Hasilnya menunjukkan bahwa M3-Embedding (All) mencapai kinerja keseluruhan terbaik, dengan nDCG@10 = 59.0 dan Recall@100 = 92.5. Model tersebut diikuti oleh M3-Embedding (Dense) dengan 56.1 dan 90.5, serta multilingual-e5-large dengan 52.9 dan 87.9. Sebaliknya, baseline leksikal BM25 memperoleh 29.7 dan 71.8, sedangkan mDPR menghasilkan kinerja terlemah di antara baseline yang dicantumkan dengan 27.2 dan 57.3. Hasil ini menunjukkan keunggulan substansial model embedding multibahasa terbaru dibandingkan pengambilan leksikal maupun baseline dense retrieval yang lebih lama untuk pengambilan informasi berbahasa Indonesia.

Tabel 1. Kinerja pengambilan berbasis benchmark pada MIRACL

Model	**nDCG@10**	**Recall@100**	**Peringkat berdasarkan nDCG@10**
BM25	29.7	71.8	5
mDPR	27.2	57.3	6
mContriever	39.2	80.2	4
multilingual-e5-large	52.9	87.9	3
M3-Embedding (Dense)	56.1	90.5	2
M3-Embedding (All)	59.0	92.5	1



Gambar 2. Grafik Batang Perbandingan Kinerja Model Embedding pada MIRACL Subset Bahasa Indonesia

Berdasarkan Gambar 1, M3-Embedding (All) menunjukkan kinerja terbaik dibandingkan seluruh model lain. Model ini memperoleh nilai nDCG@10 sebesar 59,0 dan Recall@100 sebesar 92,5. Nilai tersebut menunjukkan bahwa M3-Embedding (All) paling efektif dalam menempatkan dokumen relevan pada posisi peringkat atas sekaligus mengambil dokumen relevan secara lebih lengkap.

Model M3-Embedding (Dense) berada pada posisi kedua dengan nilai nDCG@10 sebesar 56,1 dan Recall@100 sebesar 90,5. Posisi berikutnya ditempati oleh multilingual-e5-large dengan nilai 52,9 dan 87,9. Ketiga model tersebut menunjukkan performa yang jauh lebih tinggi dibandingkan BM25, mDPR, dan mContriever.

Sebaliknya, mDPR memiliki performa terendah dengan nilai nDCG@10 sebesar 27,2 dan Recall@100 sebesar 57,3. Hasil ini menegaskan bahwa model embedding multibahasa modern memiliki kemampuan yang lebih baik dalam mendukung pencarian semantik berbahasa Indonesia dibandingkan pendekatan leksikal maupun model dense retrieval yang lebih lama.

Temuan ini konsisten dengan studi pengambilan informasi sebelumnya yang menunjukkan bahwa model dense retrieval dan late-interaction retrieval dapat mengungguli pencocokan leksikal ketika kemiripan semantik lebih penting daripada tumpang tindih istilah secara tepat (Karpukhin et al., 2020; Santhanam et al., 2022; Jha et al., 2024). Hal ini khususnya relevan untuk pengaturan ketika kueri pengguna berbeda secara leksikal dari redaksi dokumen, karena model berbasis embedding dapat lebih baik menghubungkan ekspresi yang berhubungan secara semantik meskipun bentuk permukaannya tidak tumpang tindih secara langsung.

Dari perspektif praktis, hasil ini menjadikan M3-Embedding sebagai kandidat terkuat untuk pengambilan semantik tahap pertama, sementara multilingual-e5-large tetap menjadi alternatif yang kuat dan lebih sederhana dengan kinerja kompetitif. Selain itu, studi terbaru mengenai peningkatan embedding menunjukkan bahwa kinerja pengambilan informasi dapat terus memperoleh manfaat dari supervisi berskala lebih besar dan integrasi language model yang lebih kuat, yang membantu menjelaskan mengapa keluarga embedding multibahasa yang lebih baru semakin mengungguli baseline dense yang lebih lama (Wang et al., 2024b).

4. KESIMPULAN

Penelitian ini juga memberikan implikasi praktis bagi pengembangan sistem temu kembali informasi pada dokumen institusional Indonesia. Pemilihan model retrieval sebaiknya tidak hanya mempertimbangkan kemudahan implementasi, tetapi juga kemampuan model dalam memahami kesamaan makna secara kontekstual. M3-Embedding (All) dapat dijadikan kandidat utama untuk pengembangan sistem pencarian semantik yang menuntut ketepatan dan kelengkapan hasil, sedangkan multilingual-e5-large dapat menjadi alternatif yang tetap kompetitif. Namun, hasil penelitian ini masih bersifat evaluasi awal berbasis benchmark eksternal. Validasi lanjutan tetap diperlukan melalui pengujian langsung pada korpus dokumen institusional Indonesia agar rekomendasi model memiliki dasar penerapan yang lebih kuat dan kontekstual.

Secara lebih khusus, hasil evaluasi menunjukkan bahwa model M3-Embedding (All) memiliki kemampuan paling unggul dalam dua aspek penting pencarian informasi, yaitu menempatkan dokumen relevan pada peringkat atas dan mengambil dokumen relevan secara lebih luas. Capaian nilai nDCG@10 sebesar 59,0 dan Recall@100 sebesar 92,5 memperlihatkan bahwa model ini lebih efektif dibandingkan BM25, mDPR, mContriever, multilingual-e5-large, dan M3-Embedding (Dense). Temuan tersebut menegaskan bahwa model embedding multibahasa modern lebih sesuai untuk mendukung pencarian semantik berbahasa Indonesia, terutama ketika terdapat perbedaan istilah antara kueri pengguna dan isi dokumen.

Penelitian ini menyimpulkan bahwa pemilihan model embedding berpengaruh nyata terhadap kualitas pencarian semantik dokumen berbahasa Indonesia. Model embedding multibahasa modern menunjukkan kinerja yang lebih baik dibandingkan baseline leksikal dan dense retrieval lama, terutama dalam metrik nDCG@10 dan Recall@100. Berdasarkan bukti benchmark MIRACL, M3-Embedding (All) menjadi pilihan paling menjanjikan, sedangkan multilingual-e5-large dapat dipertimbangkan sebagai alternatif yang kompetitif. Namun, karena hasil benchmark masih berasal dari korpus Wikipedia, pengujian lanjutan pada korpus dokumen institusional Indonesia tetap diperlukan agar kesesuaian model dapat dibuktikan secara lebih kontekstual.

DAFTAR PUSTAKA

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 2318–2335). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.137

Christioko, B. V., & Daru, A. F. (2018). Sistem temu kembali informasi untuk pencarian judul tugas akhir berbasis kata kunci. Jurnal Pengembangan Rekayasa dan Teknologi, 14(2), 41-49. https://doi.org/10.26623/jprt.v14i2.1226

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning [Preprint]. arXiv. https://arxiv.org/abs/2112.09118 (arXiv)

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446. https://doi.org/10.1145/582415.582418 (ACM Digital Library)

Jha, R., Wang, B., Günther, M., Mastrapas, G., Sturua, S., Mohr, I., Koukounas, A., Akram, M. K., Wang, N., & Xiao, H. (2024). Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever. In Proceedings of the Fourth Workshop on Multilingual

Representation Learning (MRL 2024) (pp. 159–166). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.mrl-1.11 (ACL Anthology)

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550 (ACL Anthology)

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., & Nogueira, R. (2021). Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2356–2362). ACM. https://doi.org/10.1145/3404835.3463238 (Cheriton School of Computer Science)

Luo, K., Liu, Z., Xiao, S., Zhou, T., Chen, Y., Zhao, J., & Liu, K. (2024). Landmark Embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3268–3281). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.180 (ACL Anthology)

Ma, X., Wang, L., Yang, N., Wei, F., & Lin, J. (2023). Fine-tuning LLaMA for multi-stage text retrieval [Preprint]. arXiv. https://arxiv.org/abs/2310.08319 (arXiv)

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2014–2037). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.148

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., & Wang, H. (2021). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5835–5847). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.466 (ACL Anthology)

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019 (ACM Digital Library)

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3715–3734). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.272 (ACL Anthology)

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models [Preprint]. arXiv. https://arxiv.org/abs/2104.08663

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024a). Multilingual E5 text embeddings: A technical report [Preprint]. arXiv. https://arxiv.org/abs/2402.05672

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024b). Improving text embeddings with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 11897–11916). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.642

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. Transactions of the Association for Computational Linguistics, 11, 1114–1131. https://doi.org/10.1162/tacl_a_00595 (ACL Anthology)