COMP6211J Course Report

Lau Kwun Hang
khlaube@connect.ust.hk

Abstract

Retrieval-Augmented Generation (RAG) systems have exhibited transformative potential in knowledge-intensive applications by integrating information retrieval with natural language generation. Embedding model is the crucial component of RAG systems , which represent the semantic meaning of data in dense vector spaces and have advanced from static embeddings to sophisticated contextual representations driven by state-of-the-art model architectures. This report provides an extensive review of embedding model underpinning RAG pipelines, exploring different model architectures, training approaches, and evaluation benchmarks, with a focus on understanding their critical role in enabling effective retrieval and generation. Despite significant progress, embedding model still faces challenges in generalizing to complex tasks, such as multi-hop reasoning. To address these limitations, this report proposes two novel methodologies: (1) Synthetic Hypothetical Document Training, which utilizes large language models (LLMs) to generate synthetic documents designed to reinforce semantic relationship modeling, and (2) Task Abstraction for Instruction-Based Fine-Tuning, which incorporates high-level task-specific instructions during training to improve model generalization. The proposed strategies will be evaluated against state-of-the-art embedding baselines through comprehensive benchmarking and ablative analysis to assess their contributions. By enhancing embedding models' representational quality and reasoning capabilities, this research aims to advance the effectiveness of RAG systems in solving knowledge-intensive and reasoning-driven tasks.

1 Literature Review

Retrieval-Augmented Generation (RAG) systems have transformed numerous knowledge-intensive applications such as open-domain question answering, summarization, and dialogue systems by integrating information retrieval mechanisms with text generation models. Unlike stand-alone generative systems, RAG pipelines rely on retrieving relevant external knowledge to guide the generative process, making its outputs more accurate, contextually aware, and factually grounded.

The backbone of RAG systems is embedding models, which represent textual, visual, or multimodal data in dense vector spaces. In recent years, the transition from static embeddings (e.g., word2vec, GloVe) to contextual embeddings powered by transformer architectures (e.g., BERT, T5, and GPT) has revolutionized information retrieval and natural language generation. These contextual embeddings serve as the foundation for both retrieval and generation stages in RAG.

The remainder of this literature review is structured as follows. In Section 2, I provide a detailed categorization of embedding techniques used in retrieval stage in RAG pipelines, including encoder-only, encoder-decoder, decoder-only, and multi-modal embeddings, while highlighting their respective applications in retrieval, generation, and RAG systems. Section 3 explores strategies to enhance embedding models. Section 4 discusses existing benchmarks, such as the Massive Text Embedding Benchmark (MTEB), and evaluates their ability to measure the effectiveness of embedding models for RAG. Section 5 concludes by summarizing the critical role of embeddings in RAG systems and the opportunities for future advancements in the field.

Preprint. Under review.

2 Type of Embedding Models for Retrieval-Augmented Generation

Embedding models are foundational to RAG systems, encoding semantic representations for both retrieval and generation tasks. These models can be categorized into four major types based on their architecture and functionality: *encoder-only*, *encoder-decoder*, *decoder-only*, and *multi-modal embeddings*.

2.1 Encoder-Only Embeddings

Encoder-only models are built around transformer encoders that output fixed-length dense embeddings capable of capturing contextual and semantic relationships within text. These embeddings are essential for RAG systems, particularly in the retrieval step, as they enable efficient similarity computation in high-dimensional vector spaces.

BERT [10] is one of the most impactful encoder-only models. Its use of deep bidirectional transformers alongside the masked language modeling (MLM) objective enables it to capture rich contextual representations. BERT has been widely adapted for tasks such as semantic search and open-domain question answering, where fine-grained contextual embeddings significantly enhance retrieval efficacy.

RoBERTa [27] extends BERT by refining pretraining procedures, such as dynamic masking, training with larger corpora, and removing the Next Sentence Prediction task. These improvements result in more robust embeddings, addressing inefficiencies in standard BERT while maintaining architectural simplicity. Dense Passage Retrieval (DPR) [17] builds on encoder-only models by introducing a dual-encoder architecture optimized for retrieval tasks. Its use of contrastive learning pairs questions and passages during training, producing embeddings tailored for retrieval scenarios.

Recent advancements such as BGE [8] and E5[50] explore broader applications and optimization techniques for encoder-only embeddings. By emphasizing multilingual capabilities and leveraging large-scale pretraining, these models further refine the embedding utility for RAG while maintaining efficiency and scalability.

2.2 Encoder-Decoder Embeddings

Encoder-decoder models simultaneously process input sequences and generate outputs, making them particularly well-suited for integrating retrieval and generation in RAG pipelines. These models encode input into intermediate embeddings that encode the semantic and contextual structure of text, which is then processed through the decoder for task-specific outputs.

T5 [39] exemplifies the strength of encoder-decoder architectures by framing all tasks within a unified text-to-text paradigm. By leveraging a large corpus, T5 enhances transfer learning capabilities, efficiently combining retrieved context with input queries for generative tasks. Its architecture integrates retrieval embeddings to produce task-relevant generations while maintaining flexibility across diverse NLP tasks.

2.3 Decoder-Only Embeddings

Decoder-only embeddings are derived from autoregressive language models, which produce outputs token by token. Recent methods demonstrate how decoder-only architectures, originally designed for generation, can also produce high-quality embeddings for retrieval tasks.
For instance, E5-mistral [51] leveraged powerful decoder-only LLM Mistral[15] to generate text embeddings by fine-tuning on synthetic datasets. The embeddings are derived by appending an [EOS] token to the input and extracting the last layer's [EOS] vector. This process reduces the dependency on labeled data while achieving competitive results. NV-Embed [21] incorporates innovations such as latent attention pooling and contrastive fine-tuning to enhance embedding performance. These models are notable for their capacity to handle long token limits, making them particularly valuable for tasks requiring extended context. However, their large model sizes and inference costs present challenges compared to encoder-only architectures.

2.4 Multi-Modal Embeddings

Multi-modal embeddings extend the capabilities of retrieval-augmented systems by processing and aligning diverse data modalities, such as text and images, within a shared embedding space. Early methods, including CLIP [38] and BLIP [25], leveraged paired image-text datasets and contrastive learning to project these modalities into unified vector spaces. These foundational models established paradigms for cross-modal retrieval while highlighting challenges in alignment and generalization. Recent approaches explore fusion mechanisms to integrate visual and linguistic features more effectively. For instance, UniIR[54] combines multi-modal embeddings by merging text and image features, providing a simpler yet effective solution for cross-modal alignment.

Recent advances aim to develop universal multi-modal embedding models with improved cross-modal alignment. VLM2Vec[16] repurposes pre-trained vision-language models Phi-3.5-V[] using contrastive learning, achieving robust generalization across diverse tasks. MM-GEM[29] integrates generation and embedding into a unified forward path, enabling fine-grained tasks such as region-specific retrieval and long-form multi-modal alignment. These developments underline ongoing progress toward versatile and efficient multi-modal embedding frameworks.

3 Embedding Models Training Approach

Improving the effectiveness of embedding models is a critical pillar for advancing RAG systems. Several state-of-the-art approaches aim to enhance embeddings through carefully designed training methodologies. This section highlights four key strategies: model pre-training, contrastive learning and task-specific fine-tuning.

3.1 Model Pre-training

Pre-training is a foundational step in building effective embedding models, enabling them to acquire general-purpose representations from large-scale, unlabeled data. Traditional approaches, such as masked language modeling (MLM), have been widely utilized in models like BERT [10]and RoBERTa [27], laying the groundwork for dense retrieval tasks. Recent advancements, such as RetroMAE [55], refine pre-training specifically for retrieval-oriented tasks through a masked auto-encoding paradigm. This framework employs asymmetric masking ratios alongside an encoder-decoder architecture to increase reconstruction difficulty, thereby improving the quality of the learned sentence embeddings.

3.2 Contrastive Learning

Contrastive learning has emerged as a powerful paradigm for training embedding models, focusing on distinguishing between semantically similar and dissimilar inputs. SimCSE [12] integrates contrastive objectives to improve the alignment and uniformity of embeddings while addressing representation degeneration in their latent space. Multi-modal extensions, such as CLIP [38] and GMC [37], successfully adapt contrastive frameworks to align visual and textual modalities, broadening the applicability of RAG across diverse domains.

3.3 Task-Specific Fine-Tuning

Task-specific fine-tuning is essential for adapting embedding models to downstream applications by addressing inter-task conflicts and tailoring embeddings to task-specific requirements. A prominent approach is instruction-based fine-tuning, where task-specific instructions are appended to inputs during training to provide contextualization. For example, BGE-base-en utilizes verbal prompts (e.g., "search relevant passages for the query") to help the model effectively differentiate between diverse tasks[56]. Recent workss have proven effective in aligning embeddings to varying retrieval objectives [4, 44].

Negative sampling plays a complementary role in enhancing model discriminability. Techniques like SimLM and E5 employ hard negatives derived from mined samples or cross-encoder distillation[49, 50] , while methods such as ANCE use approximate nearest neighbor (ANN) indices to identify challenging negatives from the corpus[57]. By combining these strategies with instruction-based fine-tuning, as demonstrated by BGE-base-en, embedding models are better equipped to robustly distinguish between related and unrelated content, which is vital for dense retrieval tasks.

4 Evaluation Benchmarks and Tools Embedding Models

Comprehensive evaluation frameworks play a pivotal role in assessing the efficacy of embedding models within RAG systems. These frameworks enable systematic comparison across tasks, fostering deeper insights into model performance while identifying shortcomings and areas for improvement. A variety of benchmarks have been developed to evaluate text embeddings, multilingual capabilities, long-context retrieval, and multimodal tasks, emphasizing versatility and robustness.

The Massive Text Embedding Benchmark (MTEB)[35] is one of the most extensive benchmarking frameworks for text embeddings. Spanning eight tasks and 58 datasets across 112 languages, it evaluates embeddings on tasks such as retrieval, classification, and clustering. While MTEB primarily focuses on cross-task generalizability, frameworks like Benchmarking-IR (BEIR)[45] concentrate on zero-shot information retrieval (IR). BEIR encompasses 18 datasets drawn from heterogeneous text retrieval tasks, offering insights into embedding models' generalization across out-of-distribution scenarios. Both MTEB and BEIR highlight the ongoing challenge of developing uniformly high-performing text embeddings that balance generalizability and computational efficiency.

Multilingual capabilities in embedding models are another critical evaluation axis, with benchmarks like MKQA[28], MLDR[8], and Tatoeba [3] offering diverse perspectives. MKQA is designed for open-domain question answering, featuring 10,000 curated queries with answers aligned across 26 languages, including many low-resource ones. MLDR evaluates embeddings specifically for multilingual long-document retrieval using datasets derived from Wikipedia, mC4, and Wudao. Beyond this, the Tatoeba benchmark enables testing embeddings for linguistic similarity tasks across a broad spectrum of languages, further advancing multilingual evaluation. Beyond these benchmarks, MTEB also applicable for evaluating models on multilingual tasks across its 112 languages.

For long-context retrieval, datasets such as MLDR and NarrativeQA[18] are highly applicable. MLDR assesses the retrieval performance of embedding models over extended text sequences using multilingual long documents, offering high relevance for applications like legal and scientific texts. NarrativeQA, on the other hand, focuses on English documents by providing Wikipedia summaries, links to full-length narratives, and question-answer pairs that challenge models to comprehend longer sequences. Together, these benchmarks address the growing need for effective embeddings in scenarios requiring substantial context length, tightly aligned with the increasing prevalence of such demands in real-world tasks.

In multimodal evaluation, the Massive Multimodal Embedding Benchmark (MMEB)[16] provides a unified framework to assess models across a variety of modalities and tasks. Comprising 36 datasets, it spans classification, information retrieval, visual question answering, and visual grounding, offering scenarios that incorporate text, image, or both as input and output modalities. MMEB presents a realistic testbed for measuring embedding models' capacity to generalize across modalities, reflecting their applicability to diverse multimodal RAG systems.

These benchmarks collectively cover a broad range of capabilities required by embedding models, including multilingual processing, long-context retrieval, and multimodal understanding. By enabling rigorous and specialized evaluation, they provide the foundation for driving advancements in the design and optimization of embedding models in RAG systems.

5 Conclusion

This literature review highlights the pivotal role of embedding models in Retrieval-Augmented Generation (RAG) systems, focusing on encoder-only, encoder-decoder, decoder-only, and multi-modal embeddings. We explored training paradigms such as pretraining, contrastive learning, and task-specific fine-tuning, along with benchmarks like MTEB to assess embedding performance. While significant advances have been made, challenges remain in retrieval relevance, factual grounding, and multi-modal integration. Future work should address these limitations to enhance the effectiveness and scalability of embedding models in RAG pipelines.

6 Proposed Research Idea

Recent advancements in Retrieval Augmented Generation (RAG) heavily depend on embedding models for effective retrieval and reasoning, yet these models face critical challenges in real-world applications. Current embeddings often fail to generalize well to complex tasks such as multi-hop reasoning, which require understanding and connecting semantic relationships across multiple documents. To address these limitations, this research proposes two methods: (1) **Synthetic Hypothetical Document Training**, which leverages large language models (LLMs) to generate synthetic hypothetical documents that enrich embedding training and improve their capacity to capture semantic relationships, and (2) **Task Abstraction for Instruction-Based Finetuning**, which introduces task-specific, high-level instructions as auxiliary input during model training to align embeddings with task semantics and enhance their generalizability. Together, these approaches aim to advance the state of embedding models in RAG by improving both retrieval accuracy and semantic understanding while enabling better performance in multi-hop reasoning and task-specific scenarios.

7 Research Objectives

The objective of this research is to enhance the reasoning capacity of embedding models in RAG systems to improve semantic understanding, generalizability across complex tasks, and multi-hop reasoning. Current embedding models often struggle with the intricacies of reasoning over multiple documents, generalizing to diverse downstream tasks, and addressing challenges posed by complex tasks such as multi-hop reasoning, entity linking, and contextual integration. Furthermore, their reliance on extensive annotated data limits their scalability in real-world applications. This research aims to address these challenges by advancing embedding models to perform more robust semantic understanding and reasoning, enabling superior performance in retrieval accuracy, task adaptability, and diverse RAG scenarios.

8 Proposed Methodology

In this section, I describe the methodology by which I enhance the embeddings in RAG systems, leveraging large language models (LLMs) to generate synthetic query-document pairs and extract task abstractions to improve semantic reasoning and generalization.

8.1 Synthetic Hypothetical Document Training

A key challenge in training embeddings for RAG systems lies in ensuring that the representations capture fine-grained alignment between semantically-related inputs throughout the entire embedding space. To address this issue, I propose a training method called Synthetic Hypothetical Document Training. Inspired by techniques that use hypothetical document generation to aid retrieval stages [11, 53], my method extends this concept to the training of embedding models. Given a relevant query-document pair (q,d) , utilizing LLM to generate a concise summary S. Providing (q,S) to LLM and asking LLM to generate a synthetic hypothetical document (d_H). The embeddings are then optimized using a contrastive learning objective, which aligns the query, document, and hypothetical document within the representation space.

These synthetic documents are designed to provide nuanced semantic relationships that complement the original query-document pairs, adding diversity and depth to the embedding space. By introducing synthetic hypothetical documents into the training pipeline, this approach fosters a richer and more consistent embedding space with better semantic alignment, ultimately enhancing the performance of embeddings in downstream RAG tasks.

8.2 Task Abstraction for Instruction-Based Fine-Tuning

Embedding models often struggle to generalize effectively across diverse tasks that require intricate reasoning or domain-specific semantic comprehension. To address this limitation, I propose Task Abstraction for Instruction-Based Fine-Tuning, a methodology inspired by the prompt-based reasoning capabilities of large language models (LLMs) [59]. This approach introduces high-level task-specific abstractions as auxiliary inputs during training, providing contextual guidance to embeddings while

optimizing their representation space. For each query (q,d) , a task specific instruction is attached to the query side:

$$q' = \text{Instruct: } \{\text{task\_definition}\} \ \backslash n \ \text{Query: } \{q\} \tag{1}$$

where "{task_definition}" is a verbal prompt, which specifies the nature of the task. On top of that, I utilize LLMs to extract structured task abstractions which employ few-shot examples in prompt to produce a high-level concept. The task abstraction is apply the following template the instruction query as to form as a new query during the embedding model's training phase:

$$q'_{abs} = \text{Instruct: } \{\text{task\_definition}\} \ \backslash n \text{Abstraction: } \{\text{task\_abstract}\} \ \backslash n \text{Query: } \{q\} \tag{2}$$

where "{task_definition}" provides the general concept of the query, e.g. a complex query as "Who is the individual associated with the cryptocurrency industry facing a criminal trial on fraud and conspiracy charges, as reported by both The Verge and TechCrunch, and is accused by prosecutors of committing fraud for personal gain" and "task_definition" would be "Identify the person in the cryptocurrency industry accused of fraud and conspiracy". The resulting embeddings not only capture the semantic information in the data but also align more closely with the overarching task objective, enriching their representational capacity and bolstering their ability to handle novel or complex downstream tasks. I hope this approach enhances the reasoning and semantic generalization of embedding models within RAG systems, facilitating more robust performance across a broad spectrum of tasks.

9 Experiment Setup

9.1 Model Architecture

I adopt Low-Rank Adaptation (LoRA)[14], a parameter-efficient fine-tuning (PEFT) method, to effectively finetune the our model. I employ Mistral-7B [15], a decoder-only LLM, as the foundational architecture. The embeddings is obtained by appending the [EOS] token to end of sequence and taking the [EOS] vector. I adopt Low-Rank Adaptation (LoRA) , a parameter-efficient fine-tuning (PEFT) method, to effectively finetune the our model. I adopt the Focal-InfoNCE loss [13] $\mathbb{L}$ over the in-batch negatives and hard negatives. Within a mini-batch of N, given relevant query-document pair (q^+, d^+), the loss function is:

$$\min \ \ \mathbb{L} = -log \frac{e^{\sum_{i=1}^{N}(\phi(q^+,d^+))^2/\tau}}{\Pi_{i=1}^{N}\left[\sum_{n_i \in \mathbb{N}} e^{\phi(q^+,n_i)(\phi(q^+,n_i)+m)/\tau} + e^{(\phi(q^+,d^+))^2/\tau}\right]} \tag{3}$$

where τ is a temperature hyperparameter, m is a hardness-aware hyperparameter that offers flexibility in adjusting the re-weighting strategy, $\mathbb{N}$ denotes the set of all negatives and $\phi(q, d)$ denotes the cosine similarity score between embeddings of a query q and document d.

9.2 Training Data

To evaluate the proposed methodologies, I will employ a set of public datasets by adapting the setup from NV-Embed[21] to demonstrate the versatility of the proposed approaches in various embedding tasks, encompassing both retrieval and non-retrieval benchmarks. For retrieval tasks, I will utilize datasets including MS MARCO [5], HotpotQA [58], Natural Questions [19], PAQ [23], StackExchange [43], SQuAD [40], ArguAna [48], BioASQ [47], FiQA [32], and FEVER [46]. As datasets without its hard negatives, these will be obtained by utilizing $mE5_{base}$ [52] to mine the top 100 hard negatives. For non-retrieval tasks, I will utilize the training splits of datasets from three sub-tasks in the MTEB benchmark: classification, clustering, and semantic textual similarity (STS))—sourced from the MTEB Hugging Face datasets [34]. The included datasets will be: AmazonReviews-Classification [33], Amazon-Counterfactual-Classification [36], Banking77-Classification [6], Emotion-Classification [42], IMDB-Classification [30], MTOPIntent-Classification [24], ToxicConversations-Classification [1], TweetSentimentExtraction-Classification [31], clustering datasets such as TwentyNewsgroups [20], raw_arxiv, raw_biorxiv, and raw_medrxiv, and semantic similarity datasets such as STS12 [2], STS22 [9], and the STS-Benchmark [7]. To ensure meaningful evaluation, common content between

raw_arxiv, raw_biorxiv, raw_medrxiv, and TwentyNewsgroups-Clustering datasets will be filtered against the MTEB evaluation set to prevent overlap. I will only utilize the training splits from each dataset to maintain consistent comparison and benchmarking throughout this study.
As introduced in Section 8, for each query-document pair (q, d) in datasets associated with retrieval tasks, training pairs will be generated based on the proposed methodology. These include (q', d), (q', d_H), and ($q'abs$, d), where q' represents an instruction-based query, d_H denotes to synthetic hypothetical document of d, and $q'abs$ incorporates task abstraction. GPT-4 will be utilized to generate the hypothetical documents and task abstractions.

9.3 Evaluation

To comprehensively evaluate the proposed methodologies, the models will be assessed on the Massive Text Embedding Benchmark (MTEB), which encompasses 15 retrieval datasets, 4 reranking datasets, 12 classification datasets, 11 clustering datasets, 3 pair classification datasets, 10 semantic textual similarity datasets, and 1 summarization dataset, providing a holistic view of performance across diverse use cases. Additionally, the models' capabilities will be tested on the BEIR benchmark, a heterogeneous benchmark containing diverse information retrieval tasks. I will compare our model with recent frontier embedding models, including e5-mistral-7b-instruct[51], Google Gecko[22], SFR-Embedding[41], NV-Embed[21], gte-Qwen2-7B-instruct[26], to benchmark both retrieval accuracy and downstream task performance. The evaluation aims to validate the effectiveness of synthetic hypothetical document training and task abstraction methods in advancing retrieval-augmented generation embedding models.

9.4 Ablation Studies

To rigorously assess the contributions of each proposed component, I will conduct a series of ablation studies.

First, I will isolate the impact of the Synthetic Hypothetical Document Training by conducting experiments with and without augmenting training data with synthetic hypothetical documents generated by LLMs, measuring their influence on embedding model.

Second, to examine the impact of *Task Abstraction for Instruction-Based Fine-Tuning*, I will train models with and without task abstraction instructions and comparing their task generalization capabilities.This will reveal how task abstraction enhances embedding generalization and semantic alignment.

These ablation studies aim to provide deeper insights into the effectiveness and limitations of the proposed methods, validating their individual and combined impacts on enhancing embedding models.

10 Conclusion

In this research plan, I propose Synthetic Hypothetical Document Training and Task Abstraction for Instruction-Based Fine-Tuning to enhance the performance of embedding models in RAG tasks. Synthetic Hypothetical Document Training uses large language models to generate data that enriches embeddings with multi-hop reasoning capabilities, while Task Abstraction introduces task-specific instructions to improve generalization and task adaptation. These methodologies aim to advance retrieval accuracy, semantic representation, and reasoning in embedding models, addressing key limitations in real-world applications.

References

[1] C. Adams, D. Borkan, J. Sorensen, L. Dixon, L. Vasserman, and N. Thain. Jigsaw unintended bias in toxicity classification, 2019.

[2] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, and D. Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[3] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, Nov. 2019.

[4] A. Asai, T. Schick, P. Lewis, X. Chen, G. Izacard, S. Riedel, H. Hajishirzi, and W. tau Yih. Task-aware retrieval with instructions, 2022.

[5] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[6] I. Casanueva, T. Temcinas, D. Gerz, M. Henderson, and I. Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, mar 2020. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

[7] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

[8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[9] X. Chen, A. Zeynali, C. Camargo, F. Flöck, D. Gaffney, P. Grabowicz, S. Hale, D. Jurgens, and M. Samory. SemEval-2022 task 8: Multilingual news article similarity. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States, July 2022. Association for Computational Linguistics.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[11] L. Gao, X. Ma, J. Lin, and J. Callan. Precise zero-shot dense retrieval without relevance labels, 2022.

[12] T. Gao, X. Yao, and D. Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

[13] P. Hou and X. Li. Improving contrastive learning of sentence embeddings with focal-infonce, 2023.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[16] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2024.

[17] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[18] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[19] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[20] K. Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995.

[21] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2024.

[22] J. Lee, Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, K. Hui, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. H. Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, and I. Naim. Gecko: Versatile text embeddings distilled from large language models, 2024.

[23] P. Lewis, Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, and S. Riedel. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115, 2021.

[24] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online, Apr. 2021. Association for Computational Linguistics.

[25] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[26] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[28] S. Longpre, Y. Lu, and J. Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021.

[29] F. Ma, H. Xue, G. Wang, Y. Zhou, F. Rao, S. Yan, Y. Zhang, S. Wu, M. Z. Shou, and X. Sun. Multi-modal generative embedding model, 2024.

[30] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[31] W. C. Maggie, Phil Culliton. Tweet sentiment extraction, 2020.

[32] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942, 2018.

[33] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA, 2013. Association for Computing Machinery.

[34] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[35] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark, 2023.

[36] J. O'Neill, P. Rozenshtein, R. Kiryo, M. Kubota, and D. Bollegala. I wish i would have loved this one, but i didn't–a multilingual dataset for counterfactual detection in product reviews. *arXiv preprint arXiv:2104.06893*, 2021.

[37] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic. Geometric multimodal contrastive representation learning, 2022.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[40] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[41] S. R. J. C. X. Y. Z. S. Y. Rui Meng, Ye Liu. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024.

[42] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen. CARER: Contextualized affect representations for emotion recognition. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[43] Stack-Exchange-Community. Stack exchange data dump, 2023.

[44] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W. tau Yih, N. A. Smith, L. Zettlemoyer, and T. Yu. One embedder, any task: Instruction-finetuned text embeddings, 2023.

[45] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021.

[46] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[47] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28, 2015.

[48] H. Wachsmuth, S. Syed, and B. Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, 2018.

[49] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Simlm: Pre-training with representation bottleneck for dense passage retrieval, 2023.

[50] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024.

[51] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Improving text embeddings with large language models, 2024.

[52] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[53] L. Wang, N. Yang, and F. Wei. Query2doc: Query expansion with large language models, 2023.

[54] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen. Uniir: Training and benchmarking universal multimodal information retrievers, 2023.

[55] S. Xiao, Z. Liu, Y. Shao, and Z. Cao. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder, 2022.

[56] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie. C-pack: Packed resources for general chinese embeddings, 2024.

[57] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval, 2020.

[58] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[59] H. S. Zheng, S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, and D. Zhou. Take a step back: Evoking reasoning via abstraction in large language models, 2024.