Dense Retrievers Can Fail on Simple Queries: Revealing The Granularity Dilemma of Embeddings

Liyan Xu[1] **Zhenlin Su**[2] **Mo Yu**[1] **Jiangnan Li**[1]
Fandong Meng[1] **Jie Zhou**[1]
[1]Pattern Recognition Center, WeChat AI
[2]South China University of Technology
liyanlxu@tencent.com

Abstract

This work stems from an observed limitation of text encoders: embeddings may not be able to recognize fine-grained entities or events within encoded semantics, resulting in failed retrieval even in simple cases. To examine such behaviors, we first introduce a new evaluation dataset, CapRetrieval, in which passages are image captions and queries are phrases targeting entity or event concepts in diverse forms. Zero-shot evaluation suggests that encoders often struggle with these fine-grained matching, regardless of training sources or model size. Aiming for enhancement, we proceed to finetune encoders with our proposed data generation strategies, enabling a small 0.1B encoder to outperform the state-of-the-art 7B model. Within this process, we further uncover the *granularity dilemma*, a challenge for embeddings to capture fine-grained salience while aligning with overall semantics. Our dataset, code and models in this work are publicly released at https://github.com/lxucs/CapRetrieval.

1 Introduction

Dense Passage Retrieval (DPR) is a crucial component in searching, characterized by the classic dual-encoder paradigm (Reimers and Gurevych, 2019): queries and candidate passages are independently encoded to embeddings that capture their semantics, then top candidates are retrieved based on their embedding similarity. For large language models (LLMs) especially, DPR serves a pivotal role in Retrieval Augmented Generation (RAG) (Lewis et al., 2020). The training of such encoders has thereby become an active direction in representation learning (Ramesh Kashyap et al., 2024).

However, despite their growing capabilities in encoding complex queries and documents, embedding matching can still fail on simple cases, and yet unable to fully supersede the conventional lexical-based methods such as BM25 (Arabzadeh et al., 2021; Ren et al., 2023). Consequently, retrieval systems in practice usually take a hybrid approach for optimized performance (Chen et al., 2022; Kulkarni et al., 2023; Luo et al., 2023; Chen et al., 2024). In this work, we probe into such limitation of DPR encoders[1], and highlight that even for simple passages, state-of-the-art encoders still often **lack a fine-grained view of concepts** and the integration of world knowledge, resulting in failed retrieval on entities or events. To remedy this issue, we delve into data generation strategies towards enhanced encoder training, and we identify a *granularity dilemma* within this process, posing a challenge for general training data composition.

Illustrated by Table 8&9, with strong encoders of both BGE (Xiao et al., 2024) and GTE (Zhang et al., 2024) series, spanning from 0.1B to 7B models that rank top on the MTEB leaderboard[2] for Chinese, all their embeddings fail to rank the more relevant passages against obviously less relevant ones, for the simple query *fried chicken*, *purple flower*, and *watermelon* respectively, indicating that such phenomena occur universally, regardless of training sources, model sizes or languages.

To facilitate the analysis of such behaviors and potential improvements, we first construct a new evaluation dataset in both Chinese and English, tailored to a practical image search scenario, where the candidate passages are image captions, and queries are short phrases of entities or events reflected in captions. The dataset overall comprises seemingly straightforward queries and captions, dubbed **CapRetrieval**, and features two unique aspects. First, it naturally requires fine-grained semantic encoding of passages, inquiring concepts of entities or events reflected in the captions. Second, distinguished from prior caption-related datasets, e.g. Ecom/Video Retrieval (Long et al., 2022),

[1]We focus on embedding-based DPR in this work, and leave reranking or other hybrid paradigms as a separate scope.
[2]https://huggingface.co/spaces/mteb/leaderboard

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19295–19305
November 4-9, 2025

whose labels are collected from real-world user clicks and thus may contain many false negatives, CapRetrieval provides explicit annotations for each query-passage pair, enabling more reliable and in-depth analysis by near 1.3 million pair labels.

Upon our zero-shot evaluation on CapRetrieval using open-source encoders of various sizes, all of them are revealed certain flaws performing these fine-grained matching, underscoring the current limitations of embeddings. We then proceed to fine-tune the encoder with our proposed data generation strategies in Sec. 4, where we leverage LLMs for automatic keyword generation as training data to enforce the semantic encoding of precise salience.

Empirical results show that our finetuned 0.1B encoder outperforms all baselines on CapRetrieval, surpassing 7B encoders, validating that the data generation strategies are effective strengthening a more subtle semantic matching. Nevertheless, we also identify a dilemma upon further analysis: with the introduced keywords added into training, the model improves on granular saliency and concepts, but may lose grasp on the overall critical semantics in a bigger picture, resulting in degradation on certain scenarios. We describe it as the *granularity dilemma* to handle both coarse and fine granularity (Sec. 4), showing that it still remains a challenge to obtain general embeddings of a full semantic view.

2 Dataset: CapRetrieval

Our dataset CapRetrieval is constructed for typical retrieval evaluation towards a practical image search scenario, composed of three following parts.

Passages A pool of image captions are prepared as candidate passages, derived from three steps. (1) A set of images is collected from personal phone albums for diverse coverage of different types, including daily photographs, web pictures, screenshots of various apps or articles. (2) Each image is transcribed into a short caption by prompting GPT-4o. (3) Each caption is manually reviewed and anonymized to ensure privacy compliance.

Queries User queries are short phrases to search relevant images, collected by two rounds. The first round of queries are brainstormed by researchers of this work, being usual entities or events that align with real-world usage scenarios. For the second round, a combination of queries in the first round is performed, then manually revised to form more complex queries. Among these queries, the easiest ones can be resolved by lexical matching, while others require fine-grained concept understanding and world knowledge to resolve.

	Records	Min	Max	Avg
Passages	3024	8	75	30.90
Queries	404	1	16	3.78

Table 1: Basic statistics of CapRetrieval dataset: number of records, and min/max/avg number of tokens per passage and query. The full annotations of all 1.3M query-passage pairs comprise 4,683 positive pairs.

Labels Following the convention of typical retrieval datasets in MTEB, the label is a relevance score of $[0, 1, 2]$ for a pair of query and passage, signaling no/weak/strong relevance. Unlike prior web-collected datasets with large-scale passage pools, our controlled setting enables exhaustive annotations for each query-passage pair, which minimizes false negatives for a more reliable evaluation.

Note that the annotation and the entire retrieval process only concern textual captions, without involving vision features. The accuracy of image captioning thus does not affect the evaluation.

- Strong Relevance (2): the query is certain or almost certain directly reflected in the caption according to commonsense. For example, "*the evening sky filled with dense clouds, with sunlight streaming through them*" is labeled 2 to the query "*sunset*".
- Weak Relevance (1): the query is likely but not certainly reflected in the caption, or the query is not directly reflected but indeed highly related. For example, "*a floor plan of a 2B2B apartment of approximately 90.89 square meters*" does not directly mention the action "*real estate purchase*", but is labeled 1 due to high relatedness by search intents.
- No Relevance (0): those who do not meet up with the strong or weak relevance.

Statistics Table 1 depicts the basic statistics of CapRetrieval. For analysis, we roughly categorize all queries in CapRetrieval into eight types, of which four types (*object*, *person*, *place*, *concept*) together can be regarded as *Singleton Entity*. Both *Singleton Entity* and *Singleton Event* are relatively straightforward phrases inquiring entities/events without imposing extra conditions, whereas other query types involve more complex constraints and semantics. More details and statistics on query types and labels are described in Appx. B.

CapRetrieval was originally constructed in Chinese. We further employ GPT-4o to translate all queries and passages into English to provide the corresponding English version, of which the translation quality is verified by human experts. The remaining sections in this paper reports on the original CapRetrieval without further specification.

3 Zero-Shot Evaluation

To examine the performance of off-the-shelf encoders on CapRetrieval, we evaluate five popular open-source encoder series for Chinese on Huggingface, offering various model sizes:

- BGE[3] (Xiao et al., 2024): 0.1B / 0.3B
- GTE[4] (Zhang et al., 2024): 0.1B / 1.5B / 7B
- E5[5] (Wang et al., 2024a,b): 0.1B / 0.3B / 7B
- Conan-v1[6] (Li et al., 2024): 0.3B
- Qwen3[7] (Zhang et al., 2025): 0.6B / 8B

Our experimental settings comply with the retrieval protocol in MTEB (Muennighoff et al., 2023), adopting nDCG@10 as the main evaluation metric. Additionally, since full labels of all pairs are annotated for CapRetrieval, we also provide nDCG@1/5 for more precise evaluation. For each model, we follow the recommended usage by its publishers' instructions. Performance of BM25 is also provided as a baseline for reference. More implementation details are described in Appx. C.

Results Table 2 shows the results of zero-shot evaluation. As most queries and captions in CapRetrieval do not possess complex semantics, all models are able to achieve decent scores as expected, with nDCG@10 above 76. The best performance is obtained by GTE-7B with 86.55 nDCG score. Several observations can be further made:

• **All encoders exhibit flaws** matching these fine-grained queries, even the capable 7B/8B LLM models. Though, they all outperform BM25 by at least 10%, underscoring that lexical matching alone is not able to resolve this task.

• **Model size is not the principal factor**. Within the same GTE series, the much smaller 0.1B model even outperforms the 1.5B model, and falls behind the 7B by 7%, despite the huge size difference.

Query Analysis Table 3 shows the decomposed performance on CapRetrieval by query types (described in B.1). *Singleton Entity* has the highest nDCG score as 82.05, while *Singleton Event* and *Simple Condition* have low scores as 73.2 and 73.8 respectively. The overall trend suggests that the encoder performs worse on more abstract queries, i.e. events or phrases with conditions. Entity-centric queries are relatively easier to resolve (80+ nDCG), though not by a large margin.

		nDCG@1	nDCG@5	**nDCG@10**
	BM25	74.40	69.30	66.54
0.1B	BGE	81.30	78.97	78.86
	GTE	82.49	80.48	79.67
	E5	80.11	77.31	76.33
0.3B	BGE	83.42	78.94	79.15
	E5	82.76	81.17	81.01
	Conan-v1	78.78	77.30	77.04
0.6B	Qwen3	85.41	81.14	81.04
> 1B	GTE-1.5B	81.70	77.20	77.35
	GTE-7B	**89.12**	**86.94**	**86.55**
	E5-7B	77.59	76.02	76.40
	Qwen3-8B	87.00	84.95	84.61
	Human	100.00	98.57	97.83

Table 2: Evaluation results of zero-shot experiments on CapRetrieval, with encoders of different model sizes. Human performance is evaluated on a 10% subset. Evaluation on the according English dataset (CapRetrievalEn) is separately presented in Table 6.

	nDCG@10	E>B	E<B	E=B
Singleton Entity	**82.05**	28%	**40%**	**32%**
Singleton Event	73.21	50%	25%	25%
Conjunction	80.60	38%	38%	25%
Simple Cond.	73.80	58%	20%	22%
Complex Cond.	77.30	**73%**	7%	20%

Table 3: Zero-shot performance of BGE 0.1B encoder per query type. The right part depicts the comparison with BM25: the ratio of queries when **E**mbeddings obtain higher/lower/similar (>/</=) scores than/to **B**M25.

Embedding vs. BM25 The right section of Table 3 presents a comparison between embedding and BM25. BM25 exhibits more polarized performance, where it outperforms embedding on entity-centric queries, but lags behind on more abstract queries. The limitation of BM25 is especially pronounced for *Complex Condition*, with a 66% gap, highlighting the **necessity of embedding-based retrieval**. Overall, it also reveals room for improvement in embeddings as follows.

False Negatives We identify three error types as the common shortcomings of current embeddings.

[3] bge-{base,large}-zh-v1.5
[4] gte-multilingual-base; gte-Qwen2-{1.5B,7B}-instruct
[5] multilingual-e5-{base,large}; e5-mistral-7B-instruct
[6] Conan-embedding-v1 (v2 has not been released yet)
[7] Qwen3-Embedding-{0.6B,8B}

• *Direct miss*: embeddings may miss entities or events reflected directly in passages, indicating that the current embedding **lacks a full view of semantics**, despite queries and passages are relatively short already. Errors can be further divided into two types within this scope.

i) *Literal Error*: embeddings fail to retrieve passages that contain the full or partial query terms verbatim - passages that BM25 can successfully recall. Though these cases can be remedied by adding lexical search in practice, we advocate that embeddings should encode concepts with full information view. Resolving these seemingly simple matches is still challenging for state-of-the-art encoders, which can be regarded as the *embedding* version of the "needle in a haystack" test for LLMs.

ii) *Semantic Error*: the query is reflected by paraphrasing or in a more abstract way in passages, where embeddings generally outperform BM25 but still have room for improvement.

• *Taxonomy knowledge*: certain scenarios require taxonomy knowledge involved, e.g. when inquiring a hypernym term such as "*household appliances*" or "*seafood*", or to recognize the matching between "*Cantonese-style roasted meats*" and "*char siu*".

• *Commonsense reasoning*: some matches require commonsense reasoning to resolve, for instance, the encoder needs to know the color of *lavender* to correctly handle the query "*purple flower*", or to realize the mention of *sitting* in a passage is highly relevant to the query "*chair*".

False Positives We further summarize two error types that appear common in false positives.

• *Over-generalization*: the passage contains relevant elements or shared tokens as the query, but does not reflect the actual query itself. For instance, the query "*shoe*" retrieves "*a labeled cardboard box with an anti-trample symbol*" before more relevant captions that actually mention shoes; "*subway*" ranks captions regarding "*high-speed train ticket*" before the ones mentioning subway stations.

• *Ignoring subjects or conditions*: the passage only addresses partial semantics but fails to accommodate full concepts, e.g. "*purple flower*" retrieves captions about *purple butterflies*; "*shopping cart screenshot*" retrieves a screenshot but of a ride-hailing app. As this kind of errors are quite common for queries with conditions, it suggests that **embeddings may not actually encode concepts**, but in a way towards superficial matching.

		CapRetr	EcomRetr	VideoRetr
	BGE	78.86	**64.55**	**69.91**
OOD	SM	84.74	63.26	68.69
	KW	87.23	60.49	63.82
	SM+KW	86.46	60.91	64.89
ID	SM	84.61	62.45	67.47
	KW	88.57	58.26	61.58
	SM+KW	**91.83**	60.24	65.16

Table 4: nDCG@10 on Cap/Ecom/Video-Retrieval. The same BGE encoder is continuously trained by using SM or KW or both as training queries, on the according OOD or ID corpus respectively (see Sec. 4.1 for acronyms).

4 Encoder Training

The results of zero-shot evaluation on CapRetrieval calls for more expressive embeddings that capture fine-grained concepts and world knowledge integration. We proceed with further examination by finetuning encoders with training data strategies.

4.1 Training Data Generation

Training pairs in existing large-scale resources such as mMARCO (Bonifacio et al., 2022) and DuReader (Qiu et al., 2022) mainly comprise user search queries and clicks. Consequently, for a passage, the queries associated with a given passage in the training set are often coarse-grained, such that they do not address the full semantic content. Motivated towards fine-grained semantic matching, we propose automatic query generation for enhanced training, with distinct granularity as follows.

- Overall summaries (**SM**): we ask LLMs to generate summaries and long questions regarding a passage, focusing on the overall saliency.
- Salient keywords (**KW**): given a passage, we ask LLMs to generate all salient keywords and hypernyms, as well as short phrases that may be inquired by users, focusing on precise saliency.

For passages, we prepare two different settings:

- Out-of-domain (**OOD**): we sample 20,000 passages from existing resources such as DuReader, mostly consisting of web articles and titles.
- In-domain (**ID**): we collect more image captions as the training passage pool. To mitigate memorization, we filter out all captions of ROUGE-L > 0.6 w.r.t. any test captions in CapRetrieval.

Table 5 illustrates training queries on an in-domain passage by the data generation strategies.

Experimental Settings We finetune BGE 0.1B as the backbone encoder and use CLS token for

Passage: 图片显示了上海市电力公司的月度账单，包括2021年5月至9月的用电费用和支付状态。
(*The image shows the monthly bill from the Shanghai Electric Power Company, including electricity charges and payment status from May to September 2021.*)

KW (keywords/phrases)	**SM** (summaries/queries)
Bill	Shanghai Electric Power Company bill inquiry
Electricity fee	Electricity usage details from May to September 2021
Utility bill	Monthly electricity bill details
Payment status	Electricity payment status records
Monthly electricity charges	Historical payment information from Shanghai Electric Power Company
Utility payment screenshot	Shanghai Electric Power Company electricity bill from May to September 2021
2021 electricity fee	...

Table 5: Example of training queries on an in-domain passage, generated by our data generation strategies (Sec. 4.1). KW strengthens the full view on precise keywords and concepts, while SM focuses on overall semantic saliency.

embeddings. The training follows the typical InfoNCE contrastive loss (Chen et al., 2020) with in-batch negatives. Training set statistics and more implementation details are provided in Appx. D.

For evaluation, we also include EcomRetrieval and VideoRetrieval (Long et al., 2022) from MTEB, of which the passages are product/video titles, featuring similar lengths as image captions. As our experiments are conducted to examine the effect of training query granularity, we do not aim for other datasets nor general SOTA performance.

4.2 The Granularity Dilemma

Table 4 shows the results of the six training settings. For CapRetrieval, the encoder trained with both SM and KW on in-domain passages achieves state-of-the-art performance, surpassing the original BGE significantly by 13%, and outperforms the best baseline GTE-7B by 5+%. For Ecom and Video, encoders trained with SM are shown comparable with BGE, indicating that summary-based queries share similar characteristics to existing large-scale training sets. For CapRetrieval, models trained on in-domain corpus outperform those on OOD, while OOD models generalize better on Ecom and Video.

Coarse vs. Fine: we roughly regard summaries and questions as coarse-grained queries that grasp important text semantics, resembling the existing training paradigm of most open-source encoders. Keywords/phrases on the other hand, are deemed fine-grained to strengthen the full semantic view of precise entities or concepts. However, Table 4 suggests that while keywords drive substantial enhancement towards the entity and event retrieval, as shown by the clear improvement on CapRetrieval in both OOD and ID settings, they appear contributing little to Ecom and Video.

Upon further analysis, we identified that though training with keywords could facilitate fine-grained matching, it may overlook the overall saliency. For example, the query in VideoRetrieval "荒野独居第2季中文版" (*Alone Season 2 Chinese Edition*) should retrieve the TV show *Alone* with the specified requirement; whereas the encoder trained with KW may overemphasize terms on *Season 2* or *Chinese*, but fails to correctly prioritize the supposedly most critical concept *Alone*, showing a **misalignment of semantic importance**. We refer to this observation as the *granularity dilemma*.

We hypothesize the main reason is that, the semantic importance of those precise keywords in a passage is relative, e.g. *Season 2* is arguably less significant when accompanied with *Alone*. The current KW setting strengthens precise matching locally, but lacks training signals of fine-grained importance among keywords. This issue is not as severe for current open-source encoders, as their training queries comprise large-scale real-world queries that reflect user intents, which can implicitly curate importance through user clicking. However, for LLM-generated queries, it requires more engineering efforts to combat this issue. We reckon that further analysis on training dynamics and data composition are needed to resolve this dilemma.

5 Conclusion

We focus on the embedding granularity that stems from the observation, where text encoders can fail to recognize entities or events of even simple cases. A new evaluation set is introduced to probe the limitation of embeddings, and zero-shot experiments reveal the room for improvement on these fine-grained matching. We further investigate data generation strategies for encoder training, and identify the *granularity dilemma* that calls for future efforts towards more expressive embeddings.

Limitations

As this work focuses around the granularity problem of embeddings, and examines both the zero-shot evaluation and training strategies, there can be limitations regarding the following two aspects.

First, the conducted analysis and training involve single-embedding encoders but exclude other paradigms, such as ColBERT (Khattab and Zaharia, 2020; Santhanam et al., 2022) or multi-views (Zhang et al., 2022). Addressing the dynamics beyond single embeddings can be important for an enlarged scope on this topic.

Second, it is still an open question on how to fully resolve the *granularity dilemma*. As discussed in Sec. 4.2, we do provide our hypothesis on the keyword relative importance, and we leave the further investigation on training data composition outside the scope of this work.

Ethical Considerations

For the dataset introduced in this work, we have manually reviewed each case to ensure compliance with privacy and ethical standards, in accordance with ACL ethics guidelines. All passages and queries do not contain sensitive or biased content related to diversity or political viewpoints. Personally identifiable information was anonymized, except in cases involving public figures where such information is part of the public domain.

No external annotators were recruited or employed during the dataset creation process; all annotation and verification were conducted internally by in-house researchers of this work. The data preparation and annotation process is approved and audited by the in-house research department. All passages are derived from image transcriptions generated from GPT-4o. As such, there are no visible risks, copyright or legal concerns in using this dataset.

References

Negar Arabzadeh, Xinyi Yan, and Charles L. A. Clarke. 2021. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2862–2866, New York, NY, USA. Association for Computing Machinery.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 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*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian, and Ophir Frieder. 2023. Lexically-accelerated dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, SIGIR '23, page 152–162, New York, NY, USA. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024. Conan-embedding: General text embedding with more and better negative samples. *Preprint*, arXiv:2408.15710.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3046–3056, New York, NY, USA. Association for Computing Machinery.

Man Luo, Shashank Jain, Anchit Gupta, Arash Einolghozati, Barlas Oguz, Debojeet Chatterjee, Xilun Chen, Chitta Baral, and Peyman Heidari. 2023. A study on the efficiency and generalization of light hybrid retrievers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1617–1626, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2017. MS MARCO: A human-generated MAchine reading COmprehension dataset.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pages 2366–2377.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, QiaoQiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. A thorough examination on zero-shot dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15783–15796, Singapore. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 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*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and ChristopherD. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Juyuan Wang, Rongchen Zhao, Wei Wei, Yufeng Wang, Mo Yu, Jie Zhou, Jin Xu, and Liyan Xu. 2025. Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. *Preprint*, arXiv:2508.10419.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by

weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 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)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Wu, Jiangnan Li, Yuqing Li, Lemao Liu, Liyan Xu, Jiwei Li, Dit-Yan Yeung, Jie Zhou, and Mo Yu. 2025. Sitemb-v1.5: Improved context-aware dense retrieval for semantic association and long story comprehension. *Preprint*, arXiv:2508.01959.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5822–5838, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000, Dublin, Ireland. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Related Work

Dense retrieval serves a critical role and has received growing attention in recent developments of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Sarthi et al., 2024; Xu et al., 2024; Edge et al., 2025; Qian et al., 2025; Wang et al., 2025). As a fundamental direction in representation learning, early works such as S-BERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021) and Contriever (Izacard et al., 2022) establish the effective training paradigm for embedding-based text representation, using contrastive learning on unsupervised or weakly supervised text pairs. Current state-of-the-art encoders usually adopt a multi-stage training that consists of both unsupervised and supervised finetuning stages (Neelakantan et al., 2022; Wang et al., 2024a; Li et al., 2023; Xiao et al., 2024). Among the popular supervised training resources, most of the datasets are collected through real-world user behaviors, such as MSMARCO (Nguyen et al., 2017) and DuReader (He et al., 2018; Qiu et al., 2022). Recently, synthetic data generation by LLMs is also reported positive gains in encoder training (Li et al., 2024; Wang et al., 2024b; Yang et al., 2025).

Beyond the conventional single-embedding encoders, other paradigms have been proposed for retrieval, such as ColBERT with token-level embeddings (Khattab and Zaharia, 2020; Santhanam et al., 2022), hybrid encoders with lexical features (Chen et al., 2022; Kulkarni et al., 2023; Luo et al., 2023) and sparse features (Chen et al., 2024). Recent efforts have also explored connecting global dependencies within embeddings (Wu et al., 2025).

As far as our knowledge, we are the first to focus on the in-depth analysis on the embedding granularity problem, with a newly introduced dataset and controlled experiments on the encoder evaluation and training data strategies.

B Dataset

The dataset is publicly released under the Apache 2.0 License.

B.1 Query Types

- *Singleton Person*: person-related queries, e.g. 男性 (*male*), 学生 (*student*).
- *Singleton Place*: place-related queries, e.g. 健身房 (*gym*), 沙漠 (*desert*).
- *Singleton Object*: other concrete entities, e.g. 食物 (*food*), 聊天记录 (*chat history*).

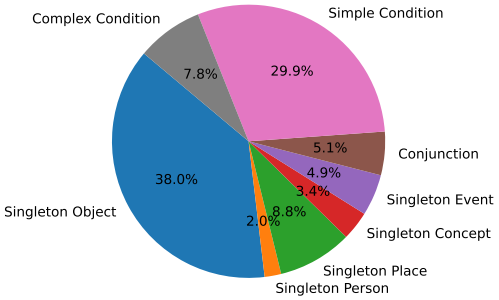

Figure 1: Query types in CapRetrieval (details in B.1).

- *Singleton Concept*: non-concrete concepts, e.g. 音乐 (*music*), 股票 (*stocks*).
- *Singleton Event*: event/action-related queries, e.g. 婚礼 (*wedding*), 演唱会 (*concert*).
- *Conjunction*: conjuncted entities, e.g. 烧烤加啤酒 (*BBQ and beer*), 樱花和传统建筑 (*cherry blossoms and traditional building*).
- *Simple Condition*: entities/events with simple conditions, e.g. 演唱会相关群聊 (*group chat regarding concerts*), 睡觉的婴儿 (*a sleeping baby*).
- *Complex Condition*: entities/events with more complex conditions, e.g. 一个人在田地里收割白菜 (*a person harvesting cabbages in the field*), 在沙发上的白猫 (*a white cat next to a sofa*).

The distribution of query types is shown in Fig. 1.

B.2 Label Distribution

As mentioned in Sec. 2, full labels are annotated for each query-passage pair in CapRetrieval, resulting in a total number of 1.3 million pair labels, comprising 4,683 positive pairs. The distribution of positive passages per query is provided in Fig. 2. Queries with the most number of relevant captions are general queries such as "男性" (*male*), "女性" (*female*), "食物" (*food*).

Distinguished from typical retrieval datasets, we allow queries with no positive passages in CapRetrieval. There are 27 such queries out of the total 404 queries, and they are excluded for ranking-based metrics, i.e. nDCG scores in Table 2 does not consider these queries. However, they can be helpful examining encoder performance in a classification setting or adversarial analysis, which is supported by CapRetrieval, since all query-passage pairs are annotated.

B.3 Annotation Procedure

Two in-house researchers of this work participate in the annotation process; no external annotators

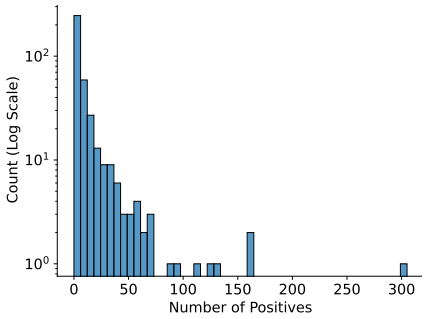

Figure 2: Histogram of the number of relevant captions (labels of $[1, 2]$) per query.

are recruited. The whole procedure is conducted by a double annotation workflow.

1. For each caption, each annotator independently assigns its labels for all the queries, such that each query-passage pair receives annotations from two annotators.
2. All label conflicts are collected and discussed by two annotators. The final label is either determined after reaching an agreement, or selected by the highest relevance received.
3. A round of correction is conducted by examining the retrieval results of preliminary experiments. Labels may be corrected after undergoing the same discussions by annotators.

The first two steps of the annotation process took around 96 working hours in total, and the third step took an additional 30 hours. The final annotator agreement across relevant pairs is 95.7%.

C Zero-Shot Evaluation

For embedding-based retrieval, all experiments follows the same setting: all text is firstly converted to lower case; all embeddings are normalized such that cosine similarity is the metric for retrieval. For the GTE-1.5B/7B, E5-7B and Qwen3 models that take a query instruction, we use the following instruction "*Given an image search query, retrieve relevant image captions*" in obtaining the query embedding for CapRetrieval.

Table 6 additionally provides zero-shot evaluation scores on the English version of CapRetrieval (dubbed CapRetrievalEn). GTE-7B achieves the best performance in both Chinese and English.

For BM25, we use *Jieba*[8] for Chinese word segmentation and NLTK[9] for English tokenization.

[8] https://github.com/fxsjy/jieba
[9] https://www.nltk.org

		nDCG@1	nDCG@5	**nDCG@10**
	BM25	72.68	70.30	69.56
0.1B	BGE	70.16	66.81	67.26
	GTE	78.12	76.60	75.77
	E5	77.72	75.52	74.53
0.3B	BGE	63.53	61.89	61.94
	E5	80.77	77.90	77.40
0.6B	Qwen3	80.64	78.36	77.80
> 1B	GTE-1.5B	73.21	72.46	72.04
	GTE-7B	**86.07**	**83.70**	**83.38**
	E5-7B	78.91	77.25	77.07
	Qwen3-8B	80.11	78.83	78.38

Table 6: Evaluation results of zero-shot experiments on CapRetrievalEn (the according English version). The same multilingual encoder models are evaluated as in Table 2, except for BGE that is now switched to its dedicated English version.

rank-bm25[10] is used for the Python BM25 implementation.

D Encoder Training

		Passages	Tokens	Queries	Length
OOD	SM	20K	2.8M	7.0	13.7
	KW	20K	2.8M	13.9	6.0
ID	SM	40K	1.5M	7.0	11.6
	KW	40K	1.5M	12.2	4.0

Table 7: Statistics of the training set settings described in Sec. 4.1: number of passages, total tokens of passages, averaged number of generated queries per passage, averaged number of tokens per query.

Table 7 shows the statistics of the training set settings described in Sec. 4.1, and 5% queries are randomly sampled as the holdout set. We continuously train *bge-base-zh-v1.5* on a single Nvidia GPU with the typical InfoNCE contrastive loss, learning rate 5×10^{-6}, weight decay 0.1, temperature 0.01. The number of epochs and batch size is adjusted to derive around 4K training steps.

The evaluation results in Table 4 demonstrate the effectiveness of our proposed data generation strategies on strengthening the fine-grained embedding matching, surpassing the baseline using either in-domain corpus or out-of-domain passages. Through cross-examination, we further raise the *granularity dilemma* discussed in Sec. 4.2.

[10] https://pypi.org/project/rank-bm25

Query	Passages	Label	Similarity
炸鸡 (*fried chicken*)	一桌丰盛的餐点包括烤肉串、炸薯条和春卷。 (*A table full of delicious dishes includes grilled meat skewers, French fries, and spring rolls.*)	0	0.48
	图片展示了麦当劳麦辣鸡翅（2块）20次券的电子优惠券，售价185.3元，单份低至7.9元。 (*The image shows a digital coupon for 20 servings of McDonald's Spicy Chicken Wings (2 pieces), priced at 185.3 RMB, bringing the cost as low as 7.9 RMB per serving.*)	2	0.38
紫色的花 (*purple flower*)	一辆紫色轿车停在路边，车顶和车窗上装饰有花束，车前挡风玻璃上有红色标签。 (*A purple sedan is parked by the roadside, decorated with flower bouquets on the roof and windows, and a red tag on the front windshield.*)	0	0.57
	图片中有四只紫色的蝴蝶，背景为浅紫色。 (*The image features four purple butterflies against a light purple background.*)	0	0.57
	一辆白色轿车停在树下，背景是紫色花田和远处的山脉。 (*A white sedan is parked under a tree, with a purple flower field and distant mountains in the background.*)	2	0.48
	图片展示了一片薰衣草田，背景是蓝天白云，文字内容为“只要你欢乐，我就幸福浓浓。朋友，愿你欢乐无忧。早上好”。 (*The image shows a lavender field with a background of blue sky and white clouds. The text reads: "As long as you're happy, my heart is full of joy. My friend, may you be cheerful and carefree. Good morning."*)	2	0.37

Table 8: Examples on dense retrieval selected from the zero-shot experiments on our new evaluation set CapRetrieval. Passages in Red are labeled irrelevant to queries (label 0), and passages in Green are relevant (label 2). For both queries, encoders retrieve irrelevant passages before the more relevant ones, despite all queries and passages are straightforward to comprehend. The cosine similarity is provided rightmost using the popular open-source encoder for Chinese *bge-large-zh-v1.5*. However, it should be noted that all encoders from both the popular BGE and GTE encoder series fail on the above examples, spanning from 0.1B to even 7B models. Overall, encoders can exhibit flaws on fine-grained embedding matching even on simple cases, regardless of training sources and model sizes.

Query	Passages	Label	Similarity
西瓜 (*watermelon*)	图片中有一个装满水果的篮子，旁边有生菜、猕猴桃和小番茄。	0	0.50
	一辆装满西瓜的三轮车停在商店门口。	2	0.47
watermelon	In the picture, there is a basket full of fruit, with lettuce, kiwis, and cherry tomatoes next to it.	0	0.60
	A tricycle loaded with watermelons is parked in front of the store.	2	0.55

Table 9: A more extreme example to illustrate the embedding granularity problem. For the query 西瓜 (*watermelon*) and passages in both Chinese and English accordingly, both the popular BGE large encoders *bge-large-zh/en-v1.5* fail to retrieve the obviously more relevant passage (label 2) before the irrelevant one (label 0). Though this case can be simply resolved by lexical matching, it demonstrates that the embedding granularity problem exists across languages. In this work, we primarily focus on the regarding evaluation in Chinese.