



ARTICLE

MIL-UDFT: An Unsupervised Domain Adaptation Framework for Text Embedding Models in Military Information Retrieval

Yongqi Wen¹, Ruopeng Yang^{2,*}, Changsheng Yin², Yihao Zhong¹, Yongqi Shi¹, Yu Tao¹,
Zhaoyang Gu¹ and Kaige Jiao¹

¹Military Intelligence, Department of Information and Communication Command, National University of Defense Technology, Changsha, China

²Military Intelligence, Department of Information and Communication Command, Information Support Force Engineering University, Wuhan, China

*Corresponding Author: Ruopeng Yang, Email: yangruopeng17@nudt.edu.cn

Received: 14 May 2026; Accepted: 14 August 2026; Published: 22 October 2026

ABSTRACT: Military information retrieval plays a pivotal role in scenarios such as intelligence analysis, equipment querying, and command decision support. However, three critical challenges—the scarcity of training data, the absence of public evaluation benchmarks, and strict data sensitivity constraints—severely hinder the performance of general-purpose embedding models in military contexts. To address these limitations, we propose MIL-UDFT, a fully local, unsupervised, data-driven fine-tuning framework. It constructs high-quality synthetic training data from publicly available military documents via an automated pipeline and adapts general-purpose embedding models to the military domain through contrastive learning-based fine-tuning. The pipeline comprises five stages: document parsing, LLM-driven corpus quality filtering, synthetic query generation, multi-dimensional query quality assessment, and positive-anchored threshold-based hard negative mining. Executed entirely within a local offline environment, it ultimately yields 163,835 training samples. Additionally, we introduce Mil-IR, the first Chinese military information retrieval evaluation benchmark, constructed based on the CMNEE dataset. Experiments across six embedding models, encompassing both encoder and decoder-only architectures, demonstrate consistent positive retrieval gains in the military domain. Notably, the BGE-base-zh-v1.5 model improves its average Mil-IR score across the three metrics (NDCG@10, Recall@10, and MRR@10) from 0.8283 to 0.9000 (an 8.66% relative increase), while experiencing only a marginal 2.06% performance drop on the C-MTEB retrieval subset. Furthermore, our findings indicate that domain-specific fine-tuning is more effective than merely scaling up model parameters, offering practical guidance for resource allocation in military AI systems.

KEYWORDS: Text embedding; synthetic data; fine-tuning; military information retrieval; domain adaptation

1 Introduction

Military information retrieval (MIR) is a critical task for identifying and extracting texts relevant to user queries from large-scale military document collections, and is essential for military applications such as intelligence analysis, equipment querying, and assisted command decision-making. However, military documents typically contain numerous specialized terms such as weapon models, tactical formations, and organizational structures, and exhibit complex document structures that demand a high level of domain understanding. Furthermore, confidentiality constraints on military data result in dispersed data storage and

complex knowledge systems, making it difficult for manually organized approaches to meet the retrieval needs of large-scale military information.

Retrieval-Augmented Generation (RAG) [1] provides an effective paradigm for domain-specific question answering by combining external knowledge retrieval with large language model (LLM) generation. In military scenarios, RAG anchors LLM responses to verified domain knowledge, mitigating the problems of hallucination and lagging knowledge updates in LLMs. However, the effectiveness of RAG systems depends heavily on the quality of the retrieval component [2]; if the retriever fails to accurately recall relevant documents, it not only fails to provide effective contextual knowledge to the LLM but may also introduce noise that misleads generation.

Contemporary retrieval systems primarily employ dense embedding models as retrievers, encoding queries and documents into dense vector representations and ranking them based on semantic similarity. In recent years, several text embedding models have been released and widely adopted, such as BGE [3] and Qwen3-Embedding [4]. These models achieve strong performance on general-domain retrieval tasks through large-scale pre-training and contrastive learning fine-tuning, providing a solid foundation for semantic retrieval.

However, when these general embedding models are directly applied to specialized domains, their performance degrades significantly [5]. This is because general embedding models are trained on Internet data, and due to the sensitive nature of the military domain, professional military terminology appears with extremely low frequency in general training corpora. Meanwhile, the narrative structure and writing style of military documents differ from those of the general training corpora (web text) used for general embedding models, making it difficult for them to accurately represent military texts.

To bridge the domain gap, existing work has explored domain adaptation methods, which can be broadly divided into two categories. One is traditional supervised fine-tuning methods that rely on annotated data (e.g., query-document similarity scores) [6], which achieve good performance but incur high annotation costs and require domain expert involvement. The other is weakly supervised methods that employ contrastive learning with query-positive-negative sample pairs [7], offering a more lightweight alternative.

In recent years, advances in synthetic data generation technology have attracted increasing attention to weakly supervised contrastive learning fine-tuning approaches. This technology enables the low-cost creation of large-scale query-text pair datasets. Synthetic data created based on large language models such as GPT-4 [8] and Qwen3 [9] has provided viable solutions to the persistent data scarcity problem, and has been validated in domains including accelerator physics [10], law [11], and hydrology [12], effectively improving the domain-specific text retrieval capability of embedding models.

However, existing methods still face challenges when applied to the military domain. First, available training data in the military domain is virtually non-existent, and there are no public evaluation benchmarks for military information retrieval. The sensitivity of military data also imposes an additional constraint that all processing must be performed locally. These unique challenges make existing domain adaptation methods difficult to directly transfer to military information retrieval scenarios.

To address these challenges, this paper introduces MIL-UDFT, an unsupervised, data-driven, and fully local fine-tuning framework for text embedding models. In contrast, prior studies in fields such as accelerator physics and law [10,11] typically generate synthetic training data through external cloud-hosted LLM services and apply quality control only at the generation stage. Against this backdrop, the methodological novelty of MIL-UDFT is twofold: First, a fully offline end-to-end pipeline—under the military data-sensitivity constraint, every stage from heterogeneous document parsing, corpus filtering, and query generation to hard-negative mining is executed locally; Second, a multi-stage quality-control mechanism—built upon

locally deployed open-source Qwen3-series LLMs, it applies multi-dimensional filtering and human validation to the corpus and synthetic queries to ensure the professionalism and reproducibility of the data. On this basis, we construct a high-quality military-domain corpus and a contrastive-learning fine-tuning dataset.

Additionally, to ensure the transparency and reproducibility of MIL-UDFT, we use publicly collected military-related documents for our experiments and construct a military information retrieval test set by adapting the CMNEE [13] military news dataset. The main contributions of this paper are summarized as follows:

1. We propose MIL-UDFT, an unsupervised data-driven fine-tuning framework for military information retrieval. We design an automated data pipeline comprising five stages: document parsing, LLM-driven corpus quality filtering, synthetic query generation, multi-dimensional query quality assessment, and positive-anchored threshold-based hard negative mining. All stages are executed in a local offline environment, satisfying military data sensitivity requirements.
2. We construct a Chinese military corpus ChMil (containing 90,926 high-quality text passages) and generate 163,835 synthetic query-document pair training samples based on this corpus, along with the first Chinese military information retrieval evaluation benchmark Mil-IR (containing 18,774 documents and 6179 queries), providing a complete experimental resource from training to evaluation for military domain retrieval research.
3. We conduct comprehensive experiments on six embedding models covering both encoder and decoder architectures. Our findings reveal that general capability rankings cannot predict domain adaptation potential, and that small-parameter models with domain fine-tuning can surpass the zero-shot performance of larger models, providing practical guidance for resource allocation in military retrieval scenarios.

The remainder of this paper is organized as follows: [Section 2](#) reviews related work, discussing prior methods and their limitations; [Section 3](#) presents the MIL-UDFT framework in detail, including military domain corpus filtering, synthetic training data generation, and model fine-tuning on domain data; [Section 4](#) describes the experimental setup, evaluation metrics, and embedding models used; [Section 5](#) presents and analyzes the experimental results; [Section 6](#) concludes the paper and suggests future research directions.

2 Related Work

This section is organized as follows: [Section 2.1](#) reviews the development of text embedding models and the need for domain adaptation; [Section 2.2](#) surveys research progress in embedding model fine-tuning; [Section 2.3](#) discusses data augmentation methods for embedding model fine-tuning.

2.1 Text Embedding Models

The development of text embedding models has evolved from sparse to dense representations and from static to dynamic paradigms, broadly encompassing three stages.

Early text retrieval primarily relied on statistically driven representations, such as the sparse term-weighted vectors of TF-IDF and BM25 and the dense low-rank vectors of Latent Semantic Analysis (LSA), which are obtained by applying singular value decomposition to the term-document matrix. While computationally efficient and straightforward, they are constrained by the curse of dimensionality and the semantic gap, which prevent them from capturing deep semantic relationships among terms.

Static word embedding methods, exemplified by Word2Vec and GloVe, significantly improved semantic representation quality by mapping words into low-dimensional dense vector spaces, where semantically similar terms are positioned in closer proximity. However, these static approaches assign a fixed vector to

each word, rendering them incapable of handling polysemy and limiting their effectiveness in sentence-level semantic matching tasks.

The advancement of pre-trained language models (PLMs) has driven a paradigm shift toward context-aware dynamic representations. Transformer-based encoder models like BERT learn deep bidirectional semantic representations through Masked Language Modeling (MLM). The pre-train-then-fine-tune paradigm has subsequently spawned a wide array of embedding models tailored for retrieval. Karpukhin et al. [14] introduced Dense Passage Retrieval (DPR), which integrates PLMs with contrastive learning to pioneer a new paradigm in dense retrieval. Following this, SimCSE [15] enhanced sentence embedding quality via unsupervised contrastive learning, while frameworks such as SBERT [16] and AugSBERT [6] further advanced retrieval alignment performance. General-purpose embedding models like BGE [3] have achieved state-of-the-art results on standard benchmarks. Concurrently, embedding models built upon decoder-only large language architectures, exemplified by Qwen3-Embedding [4], have demonstrated robust performance, thereby expanding the methodological paradigms for embedding models. Nevertheless, systematic empirical research by Tang and Yang [5] demonstrates that when a significant distribution shift exists between the training corpus of general-purpose models and the target domain, their retrieval performance degrades markedly in specialized fields. This finding strongly motivates the development of domain-adapted retrieval models.

2.2 *Embedding Model Fine-Tuning*

Recent embedding models, such as BGE [3], LLM2Vec [17], and Qwen3-Embedding [4], are predominantly built upon encoder- or decoder-only Transformer architectures. They leverage large-scale pre-training combined with contrastive learning-based fine-tuning to establish robust general-purpose embedding capabilities. However, direct deployment of these models in specialized domains often results in significant performance degradation [5].

Existing research on domain adaptation for embedding models has primarily explored two paradigms: continued pre-training and fine-tuning. In the continued pre-training paradigm, Gururangan et al. [18] demonstrated that further pre-training on large-scale target-domain corpora consistently enhances downstream performance within that domain. Similarly, Lee et al. [19] introduced BioBERT, which achieved substantial gains across multiple biomedical text mining tasks by continuing BERT's pre-training on biomedical literature. However, continued pre-training requires extensive domain-specific corpora and substantial computational resources, making it ill-suited for the military domain, which is constrained by limited resource availability and stringent data sensitivity requirements.

In contrast, supervised fine-tuning with contrastive learning using domain-relevant data has emerged as the more prevalent approach. This method achieves effective domain adaptation with only a limited amount of labeled or synthetic data, offering a significantly more favorable performance-to-cost ratio compared to continued pre-training. For instance, Szabó et al. [20] validated the effectiveness of LLM-based document embedding adaptation in engineering knowledge retrieval. A common thread across these studies [6,10,11] is the use of data augmentation to expand domain-specific fine-tuning samples, thereby mitigating the scarcity of annotated domain data. Collectively, these works demonstrate that contrastive learning-based fine-tuning on domain-specific corpora effectively improves the retrieval performance of embedding models in specialized fields. This body of research provides a valuable methodological reference for developing text embedding models tailored to military information retrieval.

2.3 Data Augmentation

In recent years, driven by the continuous advancement of Large Language Models (LLMs), LLM-based data augmentation methods have emerged as a highly effective approach to mitigating reliance on manually annotated data. These methods primarily follow two paradigms: LLM-based data annotation and LLM-based data generation.

2.3.1 Data Annotation

LLM-based data annotation methods leverage LLMs to automatically label existing domain corpora, mimicking human annotators in assigning judgment labels for tasks such as query-document relevance assessment and text classification. Tan et al. [21] provided a comprehensive survey on the application of LLMs in data annotation and synthesis, systematically covering three core aspects: annotation generation, quality assessment, and downstream utilization. Regarding specific methodologies, Dai et al. [22] proposed Promptagator, which exploits the few-shot generation capabilities of LLMs to synthesize queries for existing domain-specific documents using only a minimal set of task examples. This approach constructs training pairs to fine-tune dense retrieval models, outperforming ColBERT v2 (trained on large-scale annotated data) across 11 retrieval benchmarks. He et al. [23] introduced AnnoLLM, which employs a two-stage explain-then-annotate strategy. It first prompts an LLM to generate explanatory reasoning for a small seed of annotated examples, which subsequently guides the model in annotating unlabeled data for relevance. Experiments demonstrate that its annotation quality rivals or even surpasses that of crowdsourced annotators. However, annotation-based approaches inherently depend on existing domain corpora and a limited seed of labeled data. Given the highly classified, scarce, and fragmented nature of military corpora, this paradigm is ill-suited for direct application in military scenarios, necessitating alternative data construction strategies that operate without seed annotations.

2.3.2 Data Generation

Unlike data annotation, LLM-based data generation methods directly synthesize novel training samples using LLMs and can be further categorized into two sub-paradigms. The first operates without any real corpus as an anchor, where the entire training dataset is entirely generated by the LLM, effectively functioning as a form of knowledge distillation. For instance, DataGen proposed by Huang et al. [24] utilizes an LLM to uniformly generate synthetic datasets spanning multiple task formats. While highly flexible, it lacks guarantees regarding the factual accuracy of the generated content. The second paradigm relies on real corpora as anchors, where the LLM synthesizes queries based on existing documents to construct query-document pairs. InPars, introduced by Bonifacio et al. [25], exemplifies this corpus-anchored approach by prompting large language models to generate queries for existing documents. Along this line, Wang et al. [26] proposed the GPL method, which integrates query generation, BM25 negative sampling, and cross-encoder pseudo-labeling to form a complete unsupervised domain adaptation pipeline. Shen et al. [27] developed DRAGON for Retrieval-Augmented Generation (RAG) optimization scenarios, designing a multi-round generation and filtering iterative mechanism to enhance synthetic data quality. In comparison, while the former offers flexibility, it struggles to ensure factual grounding; the latter, anchored to authentic corpora, provides stronger factual reliability. Consequently, synthesizing training data anchored to real corpora remains the most viable technical route.

Nevertheless, implementing corpus-anchored data synthesis still faces practical challenges, including domain corpus scarcity, difficulties in data cleaning and curation, and the inherent difficulty in ensuring the quality of LLM-generated synthetic data. In the military domain, documents typically exist in unstructured

formats such as PDF, necessitating parsing tools like PaddleOCR [28] and MinerU [29] to extract clean text. The parsed raw corpora still contain noise, including advertisements, tables of contents, copyright notices, and off-topic content, requiring systematic quality filtering to ensure corpus integrity. To address similar challenges, Yu et al. [30] utilized an LLM to pre-filter sampled corpora and trained a BERT-based discriminator on the filtered data to extract educationally valuable content from multi-source web data, providing a reference pathway for automated quality filtering. Additionally, Shi et al. [31] proposed an unsupervised data-driven fine-tuning dataset generation framework tailored for vertical domains. By designing multi-dimensional quality evaluation metrics, they leveraged LLMs to filter synthetic data effectively. Drawing upon these methodological insights, this paper designs a comprehensive pipeline for military information retrieval, encompassing corpus collection and cleaning, real-corpus-anchored synthetic query generation, and a rigorous quality screening mechanism. The detailed methodology is presented in Section 3.

3 Methodology

3.1 Framework Overview

In this paper, we propose MIL-UDFT, an unsupervised data-driven fine-tuning framework for military information retrieval. The framework comprises three core stages: data pipeline, model fine-tuning, and model evaluation. The overall architecture is illustrated in Fig. 1.

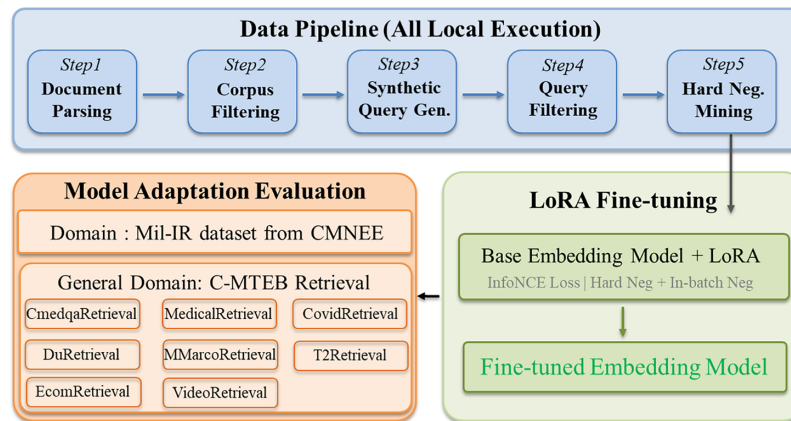


Figure 1: Overview of the MIL-UDFT framework. The framework consists of two core modules: the data pipeline (top) and model adaptation (bottom).

3.2 Data Pipeline of MIL-UDFT

This section details the data pipeline of the MIL-UDFT framework. Starting from publicly available military documents, the pipeline generates high-quality synthetic training data suitable for contrastive learning-based fine-tuning through five automated stages. All stages are executed entirely within a local offline environment, satisfying the environmental isolation requirements imposed by military data sensitivity.

3.2.1 Document Collection and Text Parsing

Due to the sensitive nature of authentic military data, all data utilized in this study were sourced from publicly accessible internet resources. We collected a total of 3220 Chinese military magazines, academic papers, and textbooks, primarily in PDF and Microsoft Word formats. These materials cover diverse topics including military equipment, tactical doctrines, and organizational structures, ensuring broad domain

coverage. Document parsing was performed using the MinerU tool [29], which extracts textual content while preserving paragraph structures and hierarchical chapter organization. Parsed documents were segmented into independent text paragraphs, and paragraphs containing fewer than 30 characters were filtered out to eliminate noise such as page numbers, advertisements, and copyright notices. After this preprocessing step, we obtained 100,486 text paragraphs with an average length of 326 characters, forming the initial military corpus.

3.2.2 LLM-Driven Corpus Quality Filtering

Rule-based methods relying on character-count thresholds can only filter enumerable noise patterns. The corpus may still contain low-quality content that is irrelevant to the military domain yet exhibits no obvious formatting anomalies, such as journal introductions, author biographies, and table-of-contents pages. To ensure the professionalism and information density of the training data, inspired by the work of Yu et al. [30], we designed a Large Language Model-based quality scoring mechanism to filter out such irrelevant texts. Specifically, we employed a locally deployed Qwen3-4B-Instruct model, leveraging the vLLM inference engine to score each paragraph. The prompt template used for scoring is illustrated in Table A1 of Appendix A. The distribution of corpus scores is shown in Fig. 2. In this study, we retained paragraphs with scores no lower than 4, ultimately obtaining 90,926 high-quality text segments, yielding a retention rate of 90.5%. We designate this curated corpus as **ChMil** (Chinese Military Corpus).

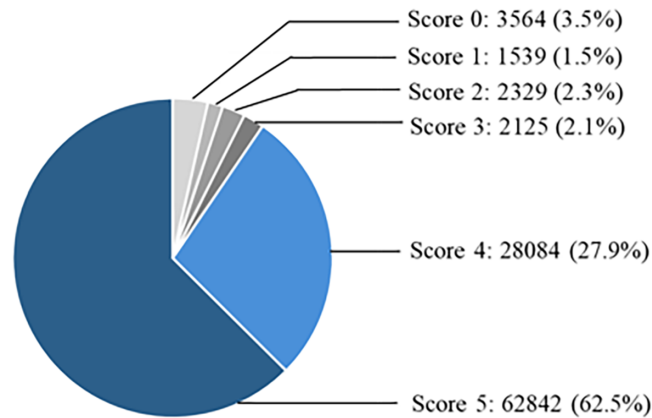


Figure 2: Corpus quality score distribution.

To intuitively demonstrate the effectiveness of the scoring mechanism, Table 1 presents representative examples from the low-score range (scores of 0–3). As shown, the mechanism successfully filters out: (1) noise texts entirely irrelevant to the military domain; (2) fragmented or low-information-density military-related texts; thereby ensuring that only text segments with substantial military reference value (scores of 4–5) are retained for downstream training.

Table 1: Corpus quality scoring examples (low-score range).

Score	Source	Text Excerpt	Filtering Rationale
0	World Military (magazine)	With the enthusiastic support of our readers, the press received a total of 1532 survey forms across all issues this year. Through careful statistical analysis and synthesis. . .	Reader survey/feedback content; entirely irrelevant to military domain knowledge

(Continued)

Table 1 (continued)

Score	Source	Text Excerpt	Filtering Rationale
1	Effects-Based Operations (monograph)	4.1 Extension of the OODA Loop ...100 4.2 Operations in the Cognitive Domain 101 4.3 From Cognitive Cycle to Effects-Based Actions ...115	Contains only section headings and page numbers; lacks substantive textual content
2	Military Digest (magazine)	If the evolution of UAV air combat is likened to a graceful youth. . . it has experienced a somewhat prolonged ‘gestation period’, a brief ‘infancy’, and a rapidly growing ‘childhood’ alongside the development of warfare. . .	Employs metaphorical narrative to describe UAV development; readable but lacks substantive technical information or actionable insights
3	Armament Knowledge (magazine)	The double-action trigger of the P99 is excessively heavy, while the single-action trigger travel is too long, requiring nearly full depression to fire. In terms of pricing, the P99 remains uncompetitive.	Provides specific equipment evaluation but lacks in-depth technical analysis or tactical doctrine context

3.2.3 Synthetic Query Generation

To construct high-quality paired data for contrastive training, we leverage a Large Language Model to generate synthetic queries from the ChMil corpus. Specifically, we adopt the same base model as Zhang et al. [4]—Qwen3-32B—and deploy it locally to ensure both the quality of synthesized queries and compliance with data sensitivity requirements. For each text paragraph, multiple synthetic queries are generated conditioned on its length and semantic content. The prompt template is provided in Table A2 of Appendix A, which instructs the model to formulate diverse military information retrieval queries from the perspective of commanders or staff officers, grounded in the paragraph content, thereby simulating authentic query demands in real-world military scenarios. This process ultimately yields 195,376 synthetic query-document pairs. The prompt for synthetic query generation along with representative output examples is illustrated in Fig. 3.

3.2.4 Synthetic Query Quality Assessment

To ensure the effectiveness of the synthesized queries, we evaluate each query along three dimensions:

1. **Relevance:** Measures the degree of association between the query and the military domain, filtering out queries that are grammatically correct but semantically deviate from military topics;
2. **Coverage:** Assesses whether the information requested by the query can be directly grounded in the content of the corresponding text paragraph, ensuring a clear semantic alignment between the query and the document;
3. **Independence:** Evaluates whether the query retains complete and unambiguous semantics even when detached from the original paragraph context, thereby preventing the generation of vague queries that rely on implicit contextual dependencies.

The assessment is performed using the same locally deployed Qwen3-4B-Instruct model. The prompt template designed for this evaluation is provided in Table A3 of Appendix A. Each query receives an independent score ranging from 1 to 5 on each of the three dimensions. The results of the synthetic query quality assessment are presented in Fig. 4. As observed, queries generated by the powerful Qwen3-32B model demonstrate excellent performance in terms of independence.

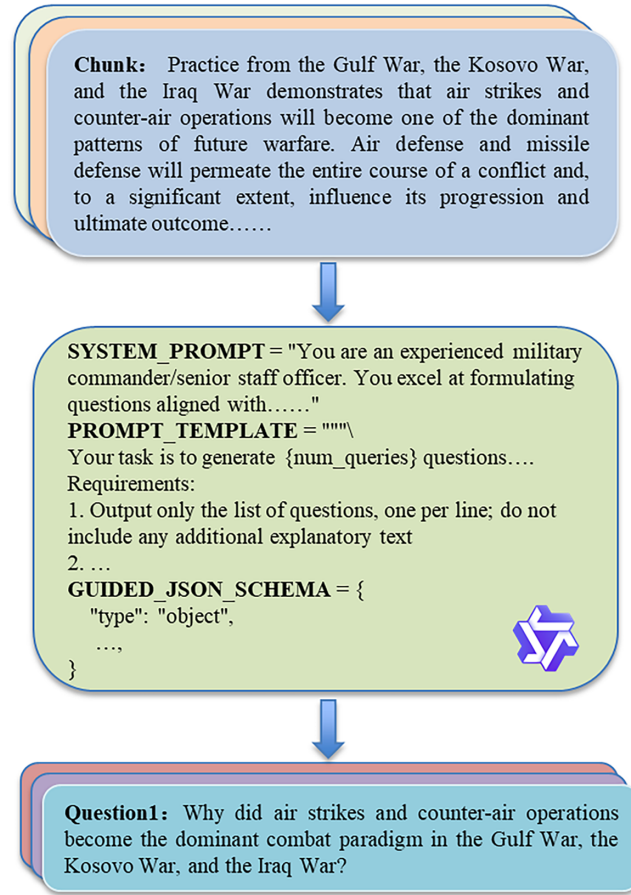


Figure 3: Example of synthetic query generation.

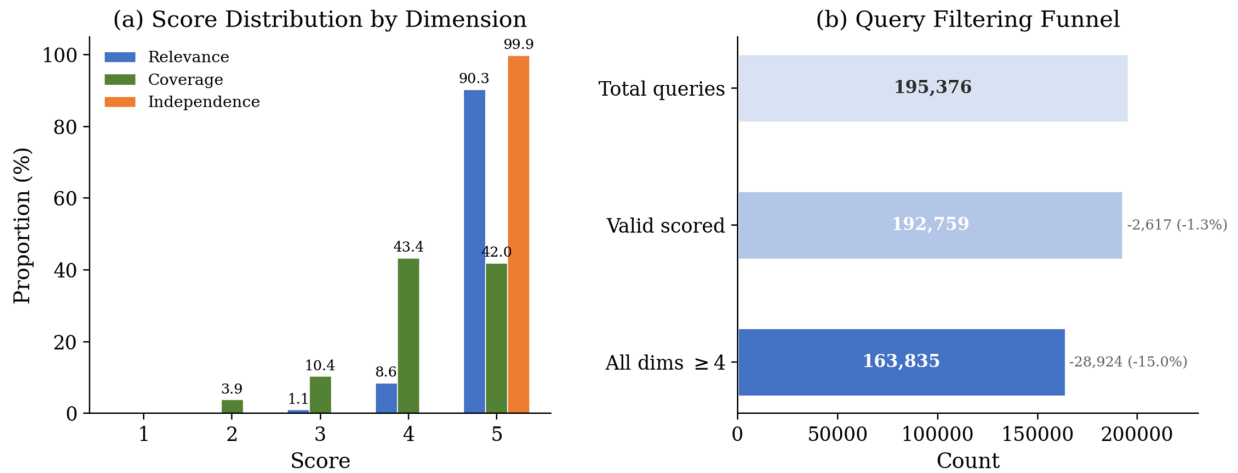


Figure 4: Synthetic query quality assessment results.

To validate the consistency of the LLM-based evaluation results, we randomly sampled a subset of query-document pairs and invited military domain experts to conduct manual assessments. The comparative results, presented in Table 2, demonstrate that LLM-generated scores maintain high agreement with human evaluations across all three dimensions. This finding confirms that LLMs can serve as reliable tools for

automated query quality assessment. Consequently, we retain only those query-text pairs with scores no lower than 4 on all three dimensions, yielding a final dataset of 163,835 samples for subsequent contrastive learning-based fine-tuning.

Table 2: Sampling validation of query quality assessment.

Dimension	Query Example	LLM Score & Rationale	Expert Score & Rationale
Relevance	How can the Militia Military Training Manual adapt to national militia development?	5 pts. Core military training topic; high topical relevance.	5 pts. Focused question with accurate terminology; high practical value.
	How to determine orientation using solar position in spring/autumn?	3 pts. Indirect military relevance; applicable to wilderness survival (borderline scenario).	3 pts. Orientation has military use but also common non-military applications; limited domain value.
Coverage	How to identify friendly vs. hostile aircraft by markings?	5 pts. Text directly answers the query with specific identification methods.	5 pts. Source content fully addresses the question; strong coverage.
	How to balance personnel and weaponry roles in modern warfare?	2 pts. Text mentions both but lacks analysis of their balancing relationship.	2 pts. Query emphasizes integration; text focuses on lethality/protection; thematic mismatch.
Independence	What challenges does the U.S. Army’s Multi-Domain Transformation strategy face?	5 pts. Clear entities and topic; semantically complete without source context.	5 pts. Self-contained question with explicit subject; interpretable independently.
	Why did the Navy still endorse the proposal after revisions by naval and foreign affairs authorities?	3 pts. “The proposal” lacks explicit referent; ambiguous when detached from context.	2 pts. Pronoun reference unclear without context; compromises independent evaluability.

3.2.5 Hard Negative Mining

Contrastive learning learns discriminative representations by increasing the similarity between queries and positive examples while decreasing the similarity with negative examples. In this process, semantically similar but irrelevant “hard negatives” are particularly crucial for improving the model’s discriminative ability. Since the training data constructed in this paper only contains query-positive pairs, an additional negative set needs to be constructed. Specifically, for each training query, the base embedding model before fine-tuning is used to perform dense retrieval on the ChMil corpus, and the top-100 candidate documents from the retrieval results are taken as the candidate set. To avoid misclassifying semantically related but unlabeled documents as negatives (false negatives), we apply a positive-anchored threshold filtering strategy: the similarity score between the query and its positive document serves as an anchor, and documents from the candidate set whose similarity reaches 95% or above of the positive score are removed. This strategy draws on the approximate nearest neighbor negative idea from ANCE [32], avoiding the introduction of false negatives through dynamic thresholding. The 95% value serves as a robust default balancing false-negative avoidance against hard-negative retention: a lower threshold would over-prune informative hard negatives, whereas a higher one would raise the risk of false negatives. Fifteen documents are randomly sampled from the filtered candidates as hard negatives for each query. This count represents the upper bound afforded by our training resources; in contrastive learning, provided the negatives are hard (high-quality), a larger number yields

richer discriminative signals and stronger representation separation, so we adopt the maximum permitted by our compute budget, consistent with the default of the FlagEmbedding framework [3] and the common practice of mainstream embedding models such as BGE and Qwen3-Embedding.

3.3 Embedding Model Fine-Tuning

This section introduces the domain adaptation method for embedding models in the MIL-UDFT framework, including the parameter-efficient fine-tuning strategy based on LoRA (Low-Rank Adaptation) [33] (Section 3.3.1) and the contrastive learning training objective (Section 3.3.2).

3.3.1 Fine-Tuning Strategy

Considering the limited scale of synthetic training data, full fine-tuning carries the risk of overfitting and requires substantial computational resources. This paper adopts the LoRA [33] method for model domain adaptation, whose working principle is illustrated in Fig. 5.

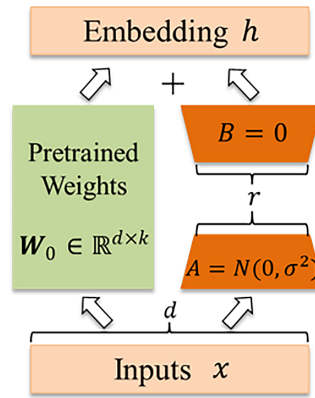


Figure 5: Working principle of LoRA.

Specifically, for an embedding model to be fine-tuned, suppose its weight parameter matrix is $W_0 \in \mathbb{R}^{d \times k}$. Its update process can be generally expressed as follows:

$$W_{\text{update}} = W_0 + \Delta W \quad (1)$$

$$\Delta W = BA, \quad B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k} \quad (2)$$

where W_0 is the original parameter matrix, ΔW is the parameter matrix to be updated, and W_{update} is the merged parameter. $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$ are parameter matrices obtained through low-rank decomposition. During fine-tuning, the original weights of the model remain frozen, and only the low-rank decomposition matrices injected via bypass are updated to adapt to the target domain. During forward propagation, for input x , the original intermediate state is:

$$h = W_0 x \quad (3)$$

Upon integrating LoRA, the intermediate representation is modified to:

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad (4)$$

After training completion, the adapted weights can be seamlessly merged with the original parameters via simple matrix addition, introducing zero additional computational overhead during inference. By updating only a minimal subset of parameters, this approach enables effective knowledge transfer even on small-to-medium-scale high-quality datasets, making it particularly well-suited for specialized domains with scarce data. Consequently, within the fine-tuning framework for retrieval tasks proposed in this work, adopting LoRA as the parameter-efficient fine-tuning method, with AdamW serving as the underlying optimizer, not only satisfies cost-efficient training requirements but also aligns optimally with the scale of our synthetic dataset.

3.3.2 Training Objective

We employ the InfoNCE [7] contrastive loss as the training objective. Its core principle is to learn discriminative embedding representations by maximizing the similarity between a query and its positive document while simultaneously minimizing the similarity with negative documents. For a query q , a positive document d^+ , and a set of negative documents $\{d_1^-, \dots, d_K^-\}$, the loss function is formulated as:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(q, d^+)/\tau)}{\exp(\text{sim}(q, d^+)/\tau) + \sum_{i=1}^K \exp(\text{sim}(q, d_i^-)/\tau)} \quad (5)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter. The negative samples comprise two components: explicitly mined hard negatives (detailed in Section 3.2.5) and in-batch negatives sampled during training. Together, they provide diverse negative signals that facilitate the learning of finer-grained semantic discriminative boundaries.

Regarding the in-batch negatives, we clarify that in-batch and cross-device negatives operate under distinct mechanisms, with the former formed within each mini-batch forward pass and the latter requiring the gathering of embeddings across multiple GPUs. All of our experiments were conducted on a single GPU, so cross-device negatives were not enabled. Constrained by GPU memory, the per-device batch size was set to the maximum afforded by each model, and gradient accumulation was applied to align the effective batch at 64, thereby stabilizing the gradient estimate. It should be noted that during accumulation each sub-batch is processed through an independent forward pass, so the number of in-batch negatives per forward pass still varies across models; nevertheless, as described in Section 3.2.5, all models are uniformly configured with 15 hard negatives that serve as the primary negative signal, which mitigates the impact of this variation. The experimental results corroborate this point. As shown by the results in Section 5.1, even Qwen3-0.6B, which has the fewest in-batch negatives, achieves steady Mil-IR improvements after fine-tuning, indicating that the differences in in-batch negative count did not limit fine-tuning effectiveness.

4 Experimental Setup

4.1 Base Embedding Models

This study incorporates six pre-trained embedding models, with their specific configurations detailed in Table 3. The model selection follows a structured rationale: First, the BGE series [3] demonstrates outstanding performance in Chinese text embedding and benefits from a mature ecosystem, making it the natural primary experimental subject. We evaluate three parameter counts: BGE-small (33M), BGE-base (110M), and BGE-large (326M). Pre-fine-tuning evaluations indicated that BGE-large performs nearly on par with BGE-base (see Section 5.1); consequently, BGE-base is designated as the primary fine-tuning target, while BGE-small is utilized to investigate performance variations across different model scales. Second, to examine the trade-off between generalization capability and domain-specific adaptation, we incorporate

Yuan-embedding-2.0-zh, which ranks 11th overall and 1st in retrieval on the C-MTEB benchmark [3]. Built upon a BERT architecture, this model employs mean pooling to generate sentence representations and appends a fully connected dense layer ($1024 \rightarrow 1792$ dimensions) post-pooling, yielding a higher embedding dimensionality than the other models. During fine-tuning, this dense layer is fully optimized without parameter freezing.

Table 3: Configuration of base embedding models used in experiments.

Model	Architecture	Language	Parameters	Embedding Dim.
BGE-small-zh-v1.5	Encoder-only	Chinese	33M	512
BGE-base-zh-v1.5	Encoder-only	Chinese	110M	768
BGE-large-zh-v1.5	Encoder-only	Chinese	326M	1024
BGE-M3	Encoder-only	Multilingual	600M	1024
Yuan-embedding-2.0-zh	Encoder-only	Chinese	~340M	1792
Qwen3-Embedding-0.6B	Decoder-only	Multilingual	600M	1024

It is worth noting that this fine-tuning scheme for Yuan-embedding-2.0-zh is dictated by its architectural specificity rather than being a free choice among multiple strategies. The model incorporates an independent dense projection layer ($1024 \rightarrow 1792$) outside the BERT encoder, whereas the low-rank adaptation of LoRA acts only on the linear sublayers within the encoder and does not apply to this projection layer; freezing it together with the other components would leave the model's output representation space unable to adapt to the target domain. We therefore had no choice but to fully optimize this projection layer while applying LoRA to the remaining components, yielding a fine-tuning method specific to this model. We acknowledge that this introduces a certain asymmetry between Yuan and the other models, but the experimental results (see Section 5) show that the strategy is effective and achieves the goal of domain adaptation.

Finally, to validate the cross-architecture applicability of the MIL-UDFT framework, we select Qwen3-Embedding-0.6B [4], which is based on a decoder-only architecture. Given that Qwen3-0.6B is a multilingual model and thus not directly comparable to the Chinese-specialized BGE-zh-v1.5, we also include BGE-M3 (another multilingual variant) as a controlled baseline to ensure a fair comparison.

4.2 Evaluation Benchmarks

We employ two benchmarks to assess the retrieval effectiveness of the models: the self-constructed military-domain benchmark Mil-IR, and the general-domain retrieval benchmark C-MTEB_Retrieval from C-MTEB [3].

Mil-IR is the first Chinese military information retrieval benchmark constructed in this work, built upon the CMNEE dataset [13]. Its document pool comprises 18,774 military-themed paragraphs covering topics such as military equipment, tactical operations, organizational structures, training exercises, and defense technologies. The query set consists of 6179 military-related queries synthesized using GLM-4.7-Flash. Each query is paired with one relevant document. The query generation process follows the same synthetic methodology as described for the training data in Section 3.2.3, but employs a distinct model instance to mitigate potential generation bias.

C-MTEB_Retrieval encompasses eight Chinese retrieval datasets and adopts the same zero-shot retrieval evaluation protocol as BEIR [34]. It provides a representative metric for assessing general-purpose Chinese retrieval capability.

Given that the primary objective of this work is to adapt embedding models for military-domain retrieval tasks, we simultaneously evaluate model performance on both the military-specific and general-purpose benchmarks. This dual-evaluation strategy enables a comprehensive assessment of:

1. The effectiveness of domain adaptation achieved through fine-tuning on our curated synthetic dataset;
2. The cross-domain generalization capability of the adapted models.

4.3 Evaluation Metrics

This paper employs three standard retrieval metrics at truncation position $K = 10$ to evaluate the models.

1. **Normalized Discounted Cumulative Gain (NDCG)** evaluates ranking quality with position-aware weighting, defined as $NDCG@k = \frac{DCG@k}{IDCG@k}$, where the Discounted Cumulative Gain is computed as:

$$DCG@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)} \quad (6)$$

Here, rel_i denotes the relevance grade of the result at rank i , and $IDCG@k$ represents the DCG value under the ideal ranking order.

2. **Recall** measures the proportion of relevant documents retrieved within the top- k results, defined as:

$$Recall@k = \frac{|R \cap T|}{|R|} \quad (7)$$

where R denotes the set of all relevant documents for a given query, and T represents the set of documents retrieved in the top- k ranked results.

3. **Mean Reciprocal Rank (MRR)** measures the ranking quality of the first relevant document, defined as:

$$MRR@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{rank_j} \quad (8)$$

where Q denotes the set of test queries, and $rank_j$ represents the rank position of the first relevant document for query j within the retrieved results. If no relevant document appears among the top- k results, the reciprocal rank is defined as $\frac{1}{rank_j} = 0$ for that query.

4. **Average Score (Avg@10)** is the arithmetic mean of the three metrics above, i.e., $Avg@10 = (NDCG@10 + Recall@10 + MRR@10)/3$, providing a single composite measure of overall retrieval quality at $k = 10$.

All evaluation metrics are computed based on inner-product similarity search using FAISS [35].

4.4 Experimental Environment and Hyperparameters

All experiments were conducted on a single NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of VRAM. The fine-tuning framework is built upon FlagEmbedding [3] (v1.4.0), implemented with PyTorch 2.10 and PEFT (v0.18.1) to enable LoRA-based adaptation. Synthetic data generation and quality assessment were performed using locally deployed Qwen3-32B and Qwen3-4B-Instruct models via the vLLM inference engine, respectively. Document parsing was carried out using MinerU [29]. All components operate entirely within a local environment without any external API calls. Table 4 summarizes the primary hyperparameter configurations for fine-tuning each model.

Table 4: Training hyperparameter settings.

Hyperparameter	BGE	Yuan-2.0-zh	Qwen3-0.6B
learning_rate	2×10^{-5}	2×10^{-5}	1×10^{-4}
temperature	0.02	0.02	0.02
batch_size	64	32	16
gradient_accumulation_steps	1	2	4
max_seq_length	512	512	512
epoch	2	2	1
lora_rank	32	32	32
lora_alpha	64	64	64
optimizer	AdamW	AdamW	AdamW
lr_scheduler_type	cosine	cosine	cosine
LoRA target modules	query, key, value, output.dense, intermediate.dense	query, key, value, output.dense, intermediate.dense*	q/k/v/o_proj, gate/up/down_proj

Note: *Yuan-embedding-2.0-zh additionally fully optimizes its dense projection layer ($1024 \rightarrow 1792$), which lies outside the BERT encoder and is not amenable to LoRA; see [Section 4.1](#).

It is worth noting that Qwen3-Embedding-0.6B adopts a higher learning rate (1×10^{-4}) and fewer training epochs (1) than BGE and Yuan, which is motivated by its decoder-only architecture. Decoder-only large language models are typically fine-tuned with LoRA at the 1×10^{-4} learning-rate scale (as in common recipes for the Qwen and LLaMA families), whereas BERT-based encoder models retain the lower 2×10^{-5} . In addition, as a 600M-parameter multilingual model, Qwen3-0.6B is more prone to catastrophic forgetting on domain data, so the number of epochs is limited to 1 to alleviate the degradation of general-domain capability. These values were determined based on architectural characteristics and established LoRA practice rather than a systematic hyperparameter search.

5 Experimental Results and Analysis

5.1 Main Experimental Results

[Table 5](#) presents the performance of the evaluated embedding models on the three evaluation metrics (NDCG@10, Recall@10, and MRR@10) across two benchmarks: Mil-IR and the C-MTEB retrieval subset, comparing results before and after fine-tuning.

Baseline Model Analysis. There is no strictly monotonic positive correlation between model parameter count and military domain retrieval performance. For instance, BGE-base-zh-v1.5 outperforms both BGE-large-zh-v1.5 and Yuan-embedding-2.0-zh in military retrieval. Despite its leading ranking on general benchmarks, Yuan-embedding-2.0-zh achieves only 0.7241 on Mil-IR, falling short of BGE-small-zh-v1.5 (0.7271), which possesses an order of magnitude fewer parameters. This reveals a pronounced domain-specific bias, likely rooted in differences in pre-training corpus composition: models trained predominantly on general or news-oriented corpora can excel on general benchmarks yet underrepresent military terminology, leaving all baselines with similarly weak military-domain priors. Regarding language scope, while Chinese-specific models surpass multilingual models in general-domain retrieval, they lack competitiveness in the military domain prior to fine-tuning; notably, the multilingual model BGE-M3 achieves the highest pre-fine-tuning score on Mil-IR. These findings indicate that general-purpose benchmark rankings cannot

directly predict a model’s potential for domain adaptation, because military retrieval depends less on model capacity than on the domain coverage of the pre-training corpus and tokenizer: military terms such as weapon designators and tactical or organizational vocabulary are largely out-of-distribution for general tokenizers, so models differ more in how their tokenization and pre-training data represent such entities than in raw parameter count.

Table 5: Performance of different models on Mil-IR and the C-MTEB retrieval subset. Bold values denote the best Avg@10 among the pre-trained baselines and among the fine-tuned models on each benchmark.

Model	Mil-IR				C-MTEB_Retrieval			
	NDCG@10	Recall@10	MRR@10	Avg@10	NDCG@10	Recall@10	MRR@10	Avg@10
BGE-small-zh-v1.5	0.7271	0.8521	0.6875	0.7556	0.6171	0.7110	0.6198	0.6493
BGE-base-zh-v1.5	0.8032	0.9142	0.7675	0.8283	0.6940	0.7822	0.6909	0.7224
BGE-large-zh-v1.5	0.7984	0.9127	0.7617	0.8243	0.7006	0.7921	0.6957	0.7295
BGE-M3	0.8369	0.9353	0.8050	0.8591	0.6524	0.7435	0.6513	0.6824
Yuan-embedding-2.0-zh	0.7241	0.8598	0.6807	0.7549	0.8185	0.9053	0.8105	0.8448
Qwen3-Embedding-0.6B	0.8227	0.9232	0.7902	0.8454	0.6978	0.7857	0.6974	0.7270
BGE-small (finetuned)	0.7934	0.9060	0.7574	0.8189	0.6024	0.6745	0.6087	0.6285
BGE-base (finetuned)	0.8821	0.9617	0.8561	0.9000	0.6779	0.7662	0.6784	0.7075
BGE-large (finetuned)	0.8765	0.9597	0.8503	0.8942	0.6821	0.7754	0.6823	0.7123
BGE-M3 (finetuned)	0.8880	0.9637	0.8632	0.9050	0.6410	0.7304	0.6419	0.6711
Yuan-2.0-zh (finetuned)	0.8769	0.9580	0.8505	0.8951	0.7739	0.8647	0.7691	0.8026
Qwen3-0.6B (finetuned)	0.8882	0.9635	0.8635	0.9051	0.6150	0.7033	0.6219	0.6467

Fine-Tuned Model Analysis. All fine-tuned models exhibit consistent positive performance gains in the military domain, validating the broad applicability of the MIL-UDFT framework. Notably, the decoder-only Qwen3-0.6B (fine-tuned) demonstrates the strongest military retrieval performance, slightly surpassing the encoder-only BGE-M3 (fine-tuned) with a comparable parameter count. This indirectly suggests that decoder-only architectures may achieve higher training data utilization efficiency. However, its general-domain retrieval capability degrades more severely compared to BGE-M3 (fine-tuned) (see [Section 5.2](#)), indicating that encoder-only architectures exhibit superior robustness for text embedding tasks.

To assess the ecological validity of the synthetic Mil-IR benchmark, we additionally evaluated the six baseline models on a set of human-authored military queries. This query set comprises 300 questions, with passages appearing in the training set excluded to prevent data leakage, and retrieval was performed over a pool of 3236 military documents. The results show that the models’ average retrieval scores on these human-authored queries are strongly and significantly correlated with their Mil-IR scores, with a Spearman rank correlation coefficient of $\rho = 0.83$ ($p < 0.05$). In particular, the top-two models (BGE-M3 and Qwen3-Embedding-0.6B) and the lowest-ranked model (Yuan-embedding-2.0-zh) hold identical positions across both query sets, whereas only the closely matched BGE-small/base/large variants exhibit minor reordering, which falls within their mutual performance margin. This indicates that the synthetic benchmark preserves a realistic model-quality ordering, thereby supporting its ecological validity.

5.2 Domain Adaptation Analysis

[Fig. 6](#) illustrates the comparison of domain gains and general capability degradation across model configurations.

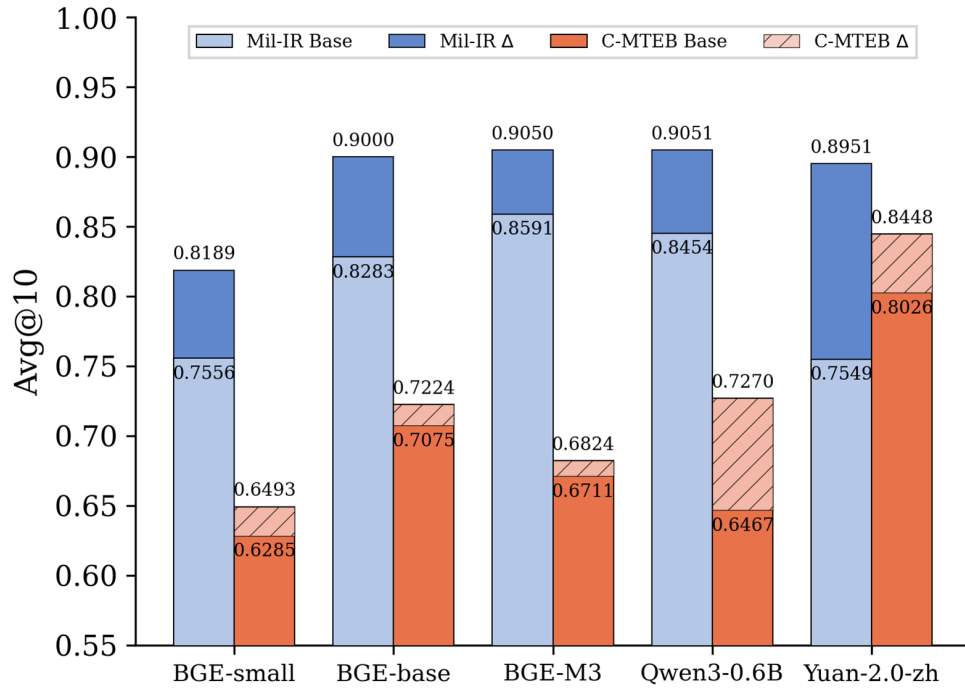


Figure 6: Comparison of domain gain (Mil-IR Δ NDCG@10) and general capability degradation (C-MTEB Δ NDCG@10) across model configurations.

Among the evaluated models, the most competitive performers are BGE-base (fine-tuned) and Yuan-2.0-zh (fine-tuned). Despite its relatively compact parameter count (110M), BGE-base-zh-v1.5 achieves exceptional military-domain retrieval performance while maintaining strong general-domain generalization capability. In contrast, Yuan-2.0-zh (fine-tuned) retains significantly superior general-domain retrieval performance compared to other models after fine-tuning; although its military-domain capability improves substantially, it remains slightly below that of BGE-base (fine-tuned).

To quantify the trade-off of domain adaptation, we report the precise degradation of each fine-tuned model on the general-domain C-MTEB_Retrieval (Avg@10): BGE-small drops 3.20%, BGE-base 2.06%, BGE-M3 1.66%, Yuan 5.00%, and Qwen3-0.6B 11.04%. It can be seen that the encoder-only models (the BGE series and Yuan) generally exhibit smaller general-domain degradation, whereas the decoder-only Qwen3-0.6B suffers the largest drop, corroborating the observation in [Section 5.1](#) that decoder-only architectures are less robust than encoder-only ones in the general domain.

It is worth noting that, although Yuan-2.0-zh (fine-tuned) adopts a model-specific hybrid fine-tuning strategy due to its architectural specificity, its results on Mil-IR and C-MTEB_Retrieval are comparable to those of the other models that use pure LoRA, indicating that this model-specific fine-tuning strategy is effective, achieves the goal of domain adaptation, and introduces no systematic bias into the cross-model comparison.

From a practical deployment perspective, Yuan-2.0-zh (fine-tuned) may be preferable when computational resources are abundant and strong out-of-domain generalization is required. However, given the prevalent resource-constrained nature of military scenarios, we recommend prioritizing BGE-base (fine-tuned) as the retrieval model for most military information retrieval applications.

5.3 Embedding Space Analysis

To elucidate the underlying mechanisms through which fine-tuning improves retrieval performance, this section analyzes the changes in the embedding space of BGE-base-zh-v1.5 before and after fine-tuning from two perspectives: similarity distribution and ranking behavior. Fig. 7 presents the analytical results, where panel (a) shows the kernel density estimation (KDE) curves of cosine similarity between queries and positive documents vs. random documents, and panel (b) displays a scatter plot illustrating the rank shifts of positive documents for all queries in the Mil-IR test set before and after fine-tuning.

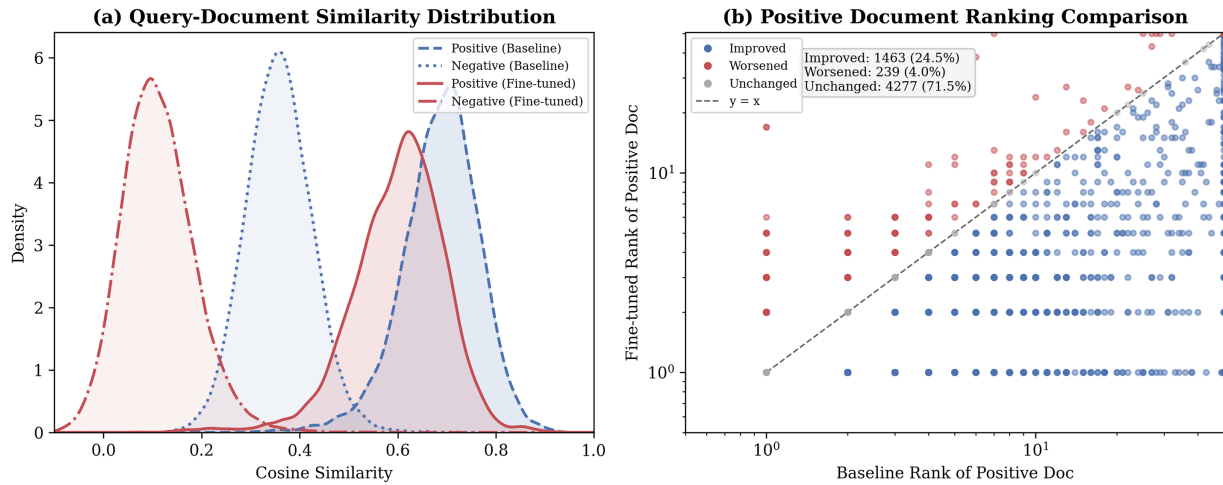


Figure 7: Embedding space analysis of BGE-base before and after fine-tuning.

Fig. 7a reveals the core mechanism underlying the performance gains from fine-tuning: contrastive learning-based fine-tuning enhances retrieval effectiveness not by pulling queries and positive documents closer in absolute similarity, but by pushing irrelevant documents farther away from queries to achieve stronger discriminative capability. Specifically, after fine-tuning, the average cosine similarity between queries and positive documents slightly decreases from 0.6868 to 0.6016 (-0.085), while the similarity between queries and random documents drops substantially from 0.3576 to 0.1121 (-0.246). Consequently, the similarity margin between positive and negative examples expands from 0.329 to 0.490, representing a 48.7% relative increase. This observation aligns with the theoretical underpinning of contrastive learning: models learn more discriminative representations by enlarging the similarity gap between positive and negative pairs, rather than merely minimizing the distance within positive pairs.

Fig. 7b corroborates the above analysis from a ranking perspective. Among the 6179 test queries in Mil-IR, 5979 were included in the rank-shift analysis; the remaining 200 were excluded because their relevant documents fell outside the top-10 results of the baseline model, leaving no baseline rank for comparison. Of the 5979 analyzed queries, the rank of the positive document improves for 1463 (24.5%), degrades for only 239 (4.0%), and remains unchanged for the remaining 4277—where the positive document was already ranked first by the baseline model. The ratio of improvements to degradations is approximately 6:1, indicating that fine-tuning yields widespread and consistent gains in retrieval quality.

The preceding analysis focuses on the encoder-only BGE-base. To examine the effect of architecture, we extend it to the decoder-only Qwen3-Embedding-0.6B (Fig. 8), and find that the two architectures exhibit different representation-adjustment patterns. After fine-tuning, the similarity distributions of both positive and negative examples in BGE-base tend to flatten—the distributions broaden and their peaks

lower, indicating more dispersed similarities among documents; in contrast, the distributions of Qwen3-Embedding-0.6B show the opposite trend, becoming more concentrated, with the distributions narrowing and their peaks rising, indicating more clustered similarities. That is, the encoder-only BGE-base tends to spread documents' similarities apart to enhance discrimination, whereas the decoder-only Qwen3-Embedding-0.6B compresses the similarities into a narrower range. This difference is consistent with the observation in Section 5.2 that Qwen3-Embedding-0.6B suffers a larger general-domain degradation: compared with the relatively dispersed adaptation of the encoder-only model, the compressive adaptation of the decoder-only model entails a more global representation shift, thereby perturbing general-domain capability to a greater extent.

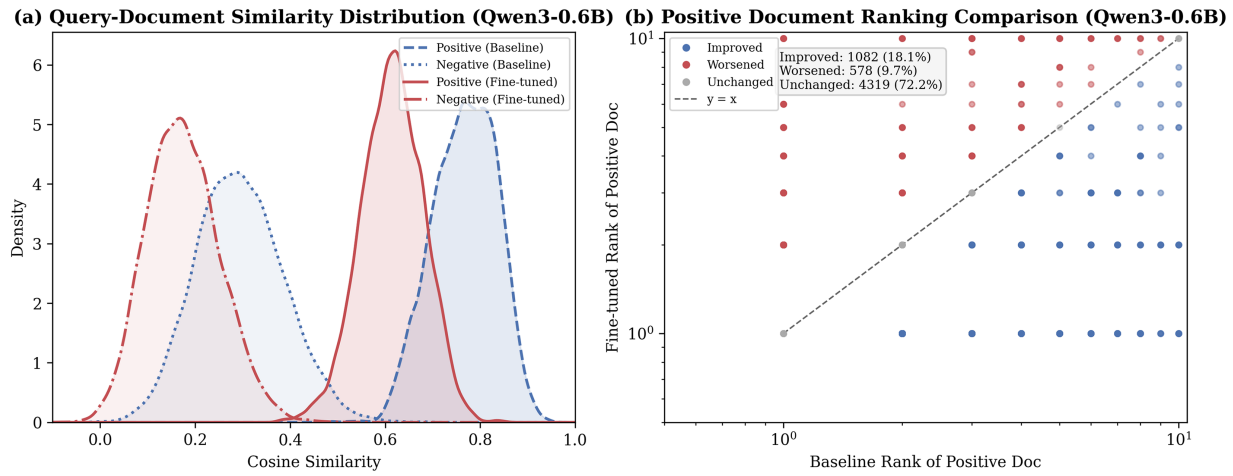


Figure 8: Embedding space analysis of Qwen3-Embedding-0.6B before and after fine-tuning.

5.4 Efficiency Analysis

Table 6 reports the training cost and inference latency of each model. Training time increases with model size, from 1.5 h for BGE-small-zh-v1.5 to 12.8 h for Qwen3-Embedding-0.6B. Encoder-only models exhibit generally low inference latency, while the decoder-only Qwen3-Embedding-0.6B incurs the highest latency (19.73 ms per query)—approximately 2.7–2.9 times that of the larger encoder-only models (BGE-large-zh-v1.5, BGE-M3, and Yuan-embedding-2.0-zh) and 5–10 times that of the smaller BGE variants. Considering training cost, inference efficiency, and retrieval performance jointly, BGE-base-zh-v1.5 offers the best trade-off—with moderate training and inference overhead and leading military-domain performance after fine-tuning—further supporting its selection as the recommended model.

Table 6: Efficiency evaluation of base embedding models.

Model	Params	Emb Dim	Train Time (h)	Latency (ms/q)
BGE-small-zh-v1.5	33M	512	1.5	1.89
BGE-base-zh-v1.5	110M	768	5.1	3.94
BGE-large-zh-v1.5	326M	1024	10.2	6.93
BGE-M3	600M	1024	11.2	7.29
Yuan-embedding-2.0-zh	~340M	1792	14.4	7.35
Qwen3-Embedding-0.6B	600M	1024	12.8	19.73

6 Conclusions

This paper proposes MIL-UDFT, a synthetic data-driven fine-tuning framework that requires no manual annotation, adapting general text embedding models to the military information retrieval domain. Through an LLM-driven automated data pipeline, the framework constructs a Chinese military corpus ChMil containing 90,926 high-quality text passages and 163,835 synthetic training samples from publicly available military documents, and establishes the first Chinese military information retrieval evaluation benchmark Mil-IR. All data processing stages are completed in a local offline environment, satisfying the inherent data sensitivity requirements of military applications. Experiments on six embedding models covering both encoder and decoder architectures demonstrate that contrastive learning fine-tuning based on synthetic data can effectively bridge the domain gap in military retrieval. Furthermore, the experiments reveal findings such as general capability rankings not predicting domain adaptation potential and small-parameter models with domain fine-tuning surpassing the zero-shot performance of larger models, providing practical guidance for resource allocation in military retrieval scenarios.

This study has the following limitations: first, synthetic data generation and quality filtering introduce additional computational overhead; second, the evaluation benchmark Mil-IR is based on a single relevant document setting and does not cover multi-relevant document scenarios; third, the current work is limited to Chinese monolingual text and does not address multilingual and multimodal retrieval scenarios. Fourth, controlled ablation experiments isolating the contribution of individual pipeline stages and the sensitivity of key design choices, such as the quality threshold and the number of hard negatives, have not yet been conducted; we defer such an analysis to future work as a valuable next step. Despite these limitations, the MIL-UDFT framework provides a scalable solution for data-scarce and data-sensitive domains, and its technical approach can be transferred to other low-resource specialized domains. Future work will explore the application of this framework in multilingual military information retrieval and multimodal military document understanding, and further optimize the quality and efficiency of synthetic data.

Acknowledgement: The authors would like to express their gratitude to three scholars from the Information Support Force Engineering University for their assistance in the expert validation of synthetic query quality and the construction of the human-authored evaluation query set.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yongqi Wen, Ruopeng Yang, Changsheng Yin; data collection: Yihao Zhong, Yongqi Shi; analysis and interpretation of results: Zhaoyang Gu, Kaige Jiao; draft manuscript preparation: Yongqi Wen and Yu Tao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available at <https://www.modelscope.cn/datasets/Gary0824/ChMil> and https://www.modelscope.cn/datasets/Gary0824/Military_Retrieval.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Prompt Templates

The prompt templates used in the data pipeline are presented in this appendix.

Table A1: Corpus quality scoring prompt template.

PROMPT_TEMPLATE = "" Please evaluate the provided article excerpt using the following 5 point scoring system to assess the text’s professionalism, information value, and readability: 0 points: Provides no military value whatsoever, consisting entirely of irrelevant information (e.g., advertisements, copyright notices, table of contents, publication information, etc.). 1 point: Contains a small amount of militaryrelated information, but the content is fragmented and incoherent, lacking substantive analysis. 2 points: Touches on military topics but remains superficial, possibly mixing military content with nonmilitary material, providing only a shallow overview of potentially useful subjects, or presented in an incoherent writing style. 3 points: Suitable for military use; introduces relevant concepts or practical information useful to military personnel. The text is coherent but may be incomplete or lack indepth analysis. 4 points: Highly relevant to military affairs, beneficial for enhancing military personnel’s relevant knowledge and capabilities without being overly abstruse. The logic is relatively clear and provides good reference value for understanding related military topics. 5 points: Possesses extremely high military value; content is highly professional, with indepth analysis, comprehensive and accurate information, containing no nonmilitary or impractical content. Offers great reference value and readability. Article excerpt: {text} Please systematically assign a score according to the listed criteria and output in JSON format: { “reasoning”: “Brief evaluation (no more than 100 characters, explaining the reason for the score)”, “score”: integer score (0 to 5) }

Table A2: Synthetic query generation prompt template. The prompts are English translations of the Chinese originals used in the experiments; the length constraint in Requirement 5 refers to Chinese characters, as all synthetic queries are generated in Chinese.

SYSTEM_PROMPT = “You are an experienced military commander/senior staff officer. Skilled at raising questions that conform to practical combat thinking and have command decision-making value, based on the given text content.”

PROMPT_TEMPLATE = “”\

Your task is to generate {num_queries} questions from the given text. The text below, delimited by “—”, is the knowledge source for your questions. Generate questions based on the following text content.

—

{text}

—

Requirements:

1. Output only the list of questions, one per line. Do not output any extra explanatory text.
2. Questions should be independent, meaning they remain complete and semantically clear even when detached from the provided text.
3. Questions can be broad or specific in granularity, but must be supported by the provided text content; do not raise questions beyond the scope of the text.
4. The generated questions must be related to the military domain.
5. Each question should be between 15 and 40 Chinese characters in length.

Please start generating questions:“”

GUIDED_JSON_SCHEMA = {

“type”: “object”,

“properties”: {

“queries”: {

“type”: “array”,

“items”: {“type”: “string”},

}

},

“required”: [“queries”],

}

Table A3: Query quality assessment prompt template.

PROMPT_TEMPLATE = "" You are a quality assessment expert in the field of military information retrieval. Your task is to evaluate the quality of synthesized retrieval queries by scoring them on three independent dimensions (1–5) and providing brief justifications (each no more than 100 characters). Scoring criteria: [Relevance] The degree to which the query is related to the military domain. - 5: Clearly involves core military topics such as military equipment, strategy and tactics, military operations, etc. - 4: Closely related peripheral topics (e.g., defense industry, civilmilitary integration). - 3: General security topics (e.g., international relations, geopolitics), but military orientation is not clear enough. - 2: Only implies military relevance, requiring context to infer. - 1: Unrelated to the military domain. [Coverage] The extent to which the information requested by the query can be directly supported by the given text passage. - 5: The text passage completely and directly answers the query. - 4: The text passage covers most of the query, with only a few details needing inference. - 3: The text passage provides some relevant information, but is incomplete. - 2: The text passage is only indirectly related to the query topic and cannot directly answer it. - 1: The text passage does not contain the information requested by the query at all. [Independence] Whether the query is semantically complete and clearly targeted when detached from the original text. - 5: The query contains clear entities, events, or concepts and is understandable independently. - 4: The query is basically complete; readers can understand the intent, but a small amount of context is missing. - 3: The query is ambiguous and requires the original context to determine its specific referent. - 2: The query heavily relies on pronouns or omissions in the original text and is difficult to understand independently. - 1: The query is completely incomprehensible without the original text. "" USER_TEMPLATE = "" Please evaluate the quality of the following synthesized query. {query},{document} Please output the evaluation result strictly in JSON format, with an integer score from 1 to 5 for each dimension and a justification of no more than 100 characters.""
--

References

1. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*. 2020;33:9459–74.
2. Huang Y, Huang JX. A survey on retrieval-augmented text generation for large language models. *ACM Comput Surv*. 2026;58(12):1–38.
3. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D, Nie JY. C-Pack: packed resources for general Chinese embeddings. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2024 Jul 14–18; Washington, DC, USA; 2024. p. 641–9.

4. Zhang Y, Li M, Long D, Zhang X, Lin H, Yang B, et al. Qwen3 embedding: advancing text embedding and reranking through foundation models. arXiv:2506.05176. 2025.
5. Tang Y, Yang Y. FinMTEB: finance massive text embedding benchmark. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025 Nov 4–9; Suzhou, China; 2025. p. 3620–38.
6. Thakur N, Reimers N, Daxenberger J, Gurevych I. Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 June 6–11; Online. p. 296–310. doi:10.18653/v1/2021.naacl-main.28.
7. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748. 2018.
8. OpenAI. GPT-4 technical report. arXiv:2303.08774. 2024.
9. Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, et al. Qwen3 technical report. arXiv:2505.09388. 2025.
10. Hellert T, Montenegro J, Venturini M, Pollastro A. Domain-specific text embedding model for accelerator physics. *Phys Rev Accelerators Beams*. 2025;28(4):044601. doi:10.1103/physrevaccelbeams.28.044601.
11. Phung NH, Nguyen CT, Nguyen MT, Nguyen TH, Le HL, Nguyen TP. A fine-tuning framework based on question, context, and answer relationships for enhancing legal information retrieval. *Eng Appl Artif Intell*. 2025;159:111570. doi:10.1016/j.engappai.2025.111570.
12. Sajja R, Sermet Y, Demir I. Domain-specific embedding models for hydrology and environmental sciences: enhancing semantic retrieval and question answering. *Water Sci. Technol*. 2025;92(9):1328–42. doi:10.2166/wst.2025.156.
13. Zhu M, Xu Z, Zeng K, Xiao K, Wang M, Ke W, et al. CMNEE: a large-scale document-level event extraction dataset based on open-source Chinese military news. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024 May 20–25; Torino, Italy; 2024. p. 3367–79.
14. Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 6769–81. doi:10.18653/v1/2020.emnlp-main.550.
15. Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Punta Cana, Dominican Republic. p. 6894–910.
16. Reimers N, Gurevych I. 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); 2019 Nov 3–7; Hong Kong, China. p. 3980–90.
17. BehnamGhader P, Adlakha V, Mosbach M, Bahdanau D, Chapados N, Reddy S. LLM2Vec: large language models are secretly powerful text encoders. In: First Conference on Language Modeling (COLM 2024); 2024 Oct 7–9; Philadelphia, PA, USA.
18. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 July 5–10; Online. p. 8342–60.
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40.
20. Szabó BN, Gyöngyössi NM, Dalos A. Applied domain adaptation of LLM-based document embeddings for engineering knowledge retrieval. In: Proceedings of the Intelligent Robotics FAIR 2025; 2025 Jun 23–24; Budapest, Hungary. p. 29–37.
21. Tan Z, Li D, Wang S, Beigi A, Jiang B, Bhattacharjee A, et al. Large language models for data annotation and synthesis: a survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 930–57.
22. Dai Z, Zhao VY, Ma J, Luan Y, Ni J, Lu J, et al. Promptagator: few-shot dense retrieval from 8 examples. In: The Eleventh International Conference on Learning Representations (ICLR 2023); 2023 May 1–5; Kigali, Rwanda.

23. He X, Lin Z, Gong Y, Jin AL, Zhang H, Lin C, et al. AnnoLLM: making large language models to be better crowd-sourced annotators. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Industry Track; 2024 Jun 16–21; Mexico City, Mexico. p. 165–90.
24. Huang Y, Wu S, Gao C, Chen D, Zhang Q, Wan Y, et al. DataGen: unified synthetic dataset generation via large language models. In: The Thirteenth International Conference on Learning Representations (ICLR 2025); 2025 Apr 24–28; Singapore.
25. Bonifacio L, Abonizio H, Fadaee M, Nogueira R. InPars: unsupervised dataset generation for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022 July 11–15; Madrid, Spain. p. 2387–92.
26. Wang K, Thakur N, Reimers N, Gurevych I. GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2022 Jul 10–15; Seattle, WA, USA. p. 2345–60.
27. Shen H, Yan H, Xing Z, Liu M, Li Y, Chen Z, et al. DRAGON: domain-specific robust automatic data generation for RAG optimization. In: Findings of the Association for Computational Linguistics: EACL 2026; 2026 Mar 24–29; Rabat, Morocco. p. 1065–78.
28. Cui C, Sun T, Liang S, Gao T, Zhang Z, Liu J, et al. PaddleOCR-VL-1.5: towards a multi-task 0.9B VLM for robust in-the-wild document parsing. arXiv:2601.21957. 2026.
29. Wang B, Xu C, Zhao X, Ouyang L, Wu F, Zhao Z, et al. MinerU: an open-source solution for precise document content extraction. arXiv:2409.18839. 2024.
30. Yu Y, Dai Z, Wang Z, Wang W, Chen R, Pei J. OpenCSG Chinese corpus: a series of high-quality Chinese datasets for LLM training. arXiv:2501.08197. 2025.
31. Shi R, Xuan S, Ye Y, Shi W, Song Y, Teng J. Unsupervised data-driven fine-tuning dataset generation framework for vertical domains. *Comput Eng Appl.* 2026;62(9):122–32. (In Chinese).
32. Xiong L, Xiong C, Li Y, Tang KF, Liu J, Bennett P, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: The Ninth International Conference on Learning Representations (ICLR 2021); 2021 May 3–7; Online.
33. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations (ICLR 2022); 2022 April 25–29; Online.
34. Thakur N, Reimers N, Rücklé A, Srivastava A, Gurevych I. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021); 2021 Dec 6–14; Online.
35. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. *IEEE Trans Big Data.* 2021;7(3):535–47. doi:10.1109/tbdata.2019.2921572.