Translation or Multilingual Retrieval? Evaluating Cross-Lingual Search Strategies for Traditional Chinese Financial Documents

Yi-Ting Chiu
National Tsing Hua University, Taiwan
chew1tim@gapp.nthu.edu.tw

Zong-Han Bai
National Tsing Hua University, Taiwan
hu110048138@gapp.nthu.edu.tw

ABSTRACT

Cross-lingual information retrieval (CLIR) is crucial for specialized domains like finance and insurance, where precise terminology governs information access. This paper examines whether a translation-based retrieval pipeline outperforms direct multilingual retrieval for Traditional Chinese financial documents. Using a dataset from the AI Cup 2024—including financial reports, insurance policies, and FAQs—we compare two retrieval strategies: (1) query translation followed by monolingual retrieval, and (2) direct semantic matching via multilingual dense encoders. Experiments show that the translation-based pipeline achieves higher recall and ranking accuracy across key financial QA tasks. Qualitative analysis further suggests that translated queries better preserve domain-specific terminology (e.g., exclusion clauses, premium calculations), enabling more accurate passage retrieval. Our findings support the continued relevance of translation-based retrieval for cross-lingual search, particularly in contexts where language-specific financial expressions dominate. We conclude by discussing deployment implications for financial knowledge systems and opportunities for hybrid models.

KEYWORDS

cross-lingual retrieval, machine translation, financial QA, domain-specific IR

1 INTRODUCTION

Cross-lingual information retrieval (CLIR) has become increasingly critical in specialized domains such as finance and insurance, where language complexity extends beyond simple translation. While Traditional Chinese is not low-resource, its financial vocabulary differs significantly from general-purpose corpora, posing challenges for multilingual retrieval models. Financial documents from Taiwan, Hong Kong, and Macau contain terms like “不保事項” (exclusions), “淨利” (net profit), and “保費” (premium), which carry precise regulatory meanings and often lack direct English equivalents. These semantic nuances are frequently lost when applying general-purpose cross-lingual encoders or direct semantic matching approaches [1,2].

To address this, we revisit the classic translate-then-retrieve paradigm [3], enabled by advances in neural machine translation (NMT). Translating English queries into Traditional Chinese allows us to leverage domain-adapted monolingual models—particularly, a BM25 retriever and a Chinese BERT-based reranker [4]—that are more sensitive to financial phrasing and local syntax. This strategy aligns the query language with the document language, mitigating mismatches in specialized terminology.

This paper investigates whether translation-based retrieval outperforms direct multilingual retrieval in financial CLIR. We use the AI Cup 2024 dataset, a large-scale benchmark containing Traditional Chinese PDF documents across three categories: financial reports, insurance policies, and FAQs. We compare:

- **Translation-Based Retrieval:** English queries are translated into Traditional Chinese using a neural MT model. BM25 identifies candidate passages, followed by a fine-tuned Chinese RoBERTa reranker [4].
- **Direct Multilingual Retrieval:** We apply multilingual dense retrievers such as LaBSE [2][15] and cross-encoder rerankers that directly match English queries with Chinese passages.

Experiments demonstrate that translation-based retrieval achieves superior performance in recall and ranking accuracy. Qualitative analysis reveals that translated queries better preserve critical financial expressions, enabling more precise retrieval. These results suggest that MT-enhanced retrieval offers practical benefits for real-world financial information systems, including banking QA, insurance clause search, and regulatory compliance. We conclude by discussing hybrid models and zero-shot fine-tuning as future directions.

2 RELATED WORK

Query vs. Document Translation in CLIR. Early cross-lingual information retrieval (CLIR) methods relied on either query translation or document translation to bridge language gaps [1]. While early comparisons favored document translation due to richer context [2], the advent of neural machine translation (NMT) has shifted this balance. Saleh and Pecina [3] showed that translating user queries—especially in specialized domains like medicine—was both more efficient and more effective than translating documents.

Multilingual Dense Retrieval: The emergence of multilingual pre-trained models (e.g., mBERT [4], XLM-R [5]) introduced retrieval pipelines that bypass translation, instead mapping queries and documents into a shared semantic space. Dual-encoder systems such as mDPR and ColBERT-X [6,7] allow efficient retrieval across languages, while cross-encoders [8] provide higher accuracy at greater cost. Recent efforts like Translate-Train [9] and Translate-Distill [10] have improved performance by training dense retrievers on machine-translated relevance data. Still, these

models often struggle in domain-specific settings due to vocabulary mismatches and the need for large-scale supervision [11,12].

Traditional Chinese and Domain-Specific IR: Most CLIR benchmarks (e.g., HC4 [13], XOR-TyDi [16]) underrepresent Traditional Chinese, which differs from Simplified Chinese not only in script but also in financial terminology and phrasing. This gap affects multilingual models trained primarily on general-purpose or Simplified Chinese corpora. In financial contexts, the challenge intensifies—terms like "保費" (premium) and "不保事項" (exclusions) require precise alignment, often lost in cross-lingual encoders. Zhang et al. [15] observed a significant accuracy drop when applying English-trained financial LLMs to Spanish data, a gap likely mirrored in Traditional Chinese. Although FinBERT and other domain-specific models exist, cross-lingual financial retrieval remains underexplored.

Motivation: To address this gap, we experimentally compare translation-based and direct multilingual retrieval methods for real-world Traditional Chinese financial documents. Our goal is to evaluate trade-offs in accuracy, robustness, and real-world deployability.

3 METHOD

Our methodology is designed to emulate a realistic cross-lingual search scenario on a **Traditional Chinese financial document collection**. We leverage the AI Cup 2024 dataset, which consists of documents and queries in a financial context. Below, we describe the dataset, the retrieval methods compared, and the evaluation setup in detail.

Dataset and Preprocessing: The **AI Cup 2024** dataset includes a collection of PDF documents covering three categories: (1) **Financial Reports** – e.g., annual reports and prospectuses from companies, (2) **Insurance Documents** – policy wordings, terms & conditions from insurance providers, and (3) **Frequently Asked Questions (FAQs)** – a knowledge base of Q&A pairs from financial customer support. All documents are in Traditional Chinese (zh-TW), totaling several thousand pages of text.

3.1 Dataset and Preprocessing

The **AI Cup 2024 dataset** comprises over 1,600 Traditional Chinese PDF documents across three categories: (1) financial reports, (2) insurance contracts, and (3) customer-facing FAQs. These documents collectively span thousands of pages and reflect formal financial writing styles.

We parsed the PDFs using PyMuPDF to extract page-level text. When text extraction failed (e.g., scanned documents), we applied EasyOCR as a fallback. Given the gold-standard annotations in the AI Cup use PDF-level granularity for pid references, we retained the entire PDF page as a single retrievable unit rather than splitting into smaller passages. This design ensured strict alignment with evaluation ground truths.

The **query set** consists of 150 information needs, originally authored in Traditional Chinese. To simulate realistic cross-lingual search, we manually translated each query into English while preserving domain-specific financial terminology (e.g., "不保事項" → *exclusions*, "淨利" → *net profit*, "保費" → *premium*). These translations were verified by native speakers with finance background to ensure alignment with institutional phrasing.

All queries were matched with at least one relevant passage (pid) in the gold standard. Since no training queries were provided, our methods rely entirely on zero-shot inference or pretraining from external corpora.

3.2 Translation + Retrieval (Translate-Search)

This approach begins by translating English queries into Traditional Chinese using the *Helsinki-NLP/opus-mt-en-zh* neural machine translation model, followed by conversion from Simplified to Traditional Chinese via OpenCC. We then retrieve relevant passages using a two-stage pipeline: BM25 lexical retrieval and a fine-tuned Chinese BERT re-ranker.[14]

To evaluate the impact of translation quality, we run both machine-translated queries (query_zh_nmt) and the native Traditional Chinese queries (query) within a unified BM25 retrieval script. This dual-query BM25 pipeline indexes all Chinese passages and returns top-100 ranked results by term overlap. This BM25-only setting (BM25_only) serves as our primary baseline for comparison.

For reranking, we fine-tune a Chinese BERT model (hfl/chinese-roberta-wwm-ext) as a binary classifier. Positive training pairs are constructed using AI Cup 2024's gold-standard (query, passage) annotations. Since the dataset lacks explicitly labeled negative passages, we randomly sample hard negatives from non-matching candidates retrieved by BM25. All training data are derived from the official 150-query development set, which we augmented by pairing each query with its gold passage (label = 1) and several random non-relevant passages (label = 0). No external data or manual annotation were used to ensure reproducibility. The model is fine-tuned for 3 epochs using the binary cross-entropy loss.

This re-ranker is then applied on top of BM25 outputs for both query_zh_nmt and query, producing two final pipelines: BM25_rerank_query_zh_nmt and BM25_rerank_query. The dual-query setup enables detailed ablation on how translation quality impacts retrieval performance. Notably, our experiments demonstrate that the re-ranker substantially improves ranking accuracy over BM25 alone, especially when translation noise is present.

3.3 Direct Multilingual Retrieval

To evaluate cross-lingual semantic retrieval without relying on translation, we implement two multilingual pipelines: a **LaBSE-based dual encoder** and a **cross-encoder re-ranker**. These pipelines directly match English queries with Traditional Chinese passages using dense vector similarity.

Dual Encoder (LaBSE): We use the SentenceTransformers implementation of LaBSE (Language-Agnostic BERT Sentence Embedding) to independently encode the English query and each Chinese passage into a shared embedding space.[17] At index time, all passages are encoded and stored in a FAISS index. At query time, the English query is encoded and used to retrieve top-K passages based on cosine similarity. This setting enables fast retrieval and language-agnostic alignment across English–Chinese pairs. We refer to this pipeline as dense_dual_encoder.

***Cross-Encoder Re-Ranker (Multilingual)*:** To improve ranking precision, we apply a **cross-encoder** re-ranker on top of the top-K outputs from the dual encoder. Specifically, we use the multilingual **MiniLM-L12-v2** model from SentenceTransformers, which jointly encodes a (query, passage) pair and produces a binary relevance score. The cross-encoder is applied to the top 100 dual-encoder candidates per query. The final ranking is determined by reordering these candidates using the predicted relevance probabilities (softmax(logits)[1]). This forms our Cross-Encoder reranker pipeline.

***Training Details*:** Both multilingual models are used as-is without further fine-tuning, simulating a zero-shot multilingual retrieval setting. Although we experimented with post-translation reranking, it offered limited benefit and added complexity. Our current pipeline emphasizes simplicity and efficiency, while maintaining strong multilingual alignment.

***Design Notes*:** Unlike the translation-based pipeline, which benefits from a domain-specific fine-tuned re-ranker, this multi-lingual setup remains general-purpose. The cross-encoder adds precision but is limited to reranking pre-selected candidates. As shown in Section 4, this architecture performs reasonably well in general semantic matching but may underperform in domains requiring exact terminology matching (e.g., finance and insurance).

3.4 Evaluation Metrics

We evaluate retrieval performance using standard ranking metrics in information retrieval. Each query is associated with one or more relevant passages (ground truth). We report the following:

- ***Mean Reciprocal Rank (MRR@10)*:** Measures the rank of the first relevant passage per query. For each query, we compute $\frac{1}{\text{rank}}$ of the first relevant passage in the top 10 results, assigning 0 if none are found. This metric emphasizes top-ranked accuracy, suitable for QA-style retrieval [9].
- ***Normalized Discounted Cumulative Gain (NDCG@10)*:** Captures ranking quality by considering the position of all relevant items in the top 10. We adopt a binary relevance setup, as the dataset lacks graded relevance scores.
- ***Recall@100*:** Evaluates whether the system retrieves at least one relevant passage within the top 100 results. High recall is critical when the pipeline includes a second-stage reranker or reader module.

We average all metrics across the full test set of 150 queries. For statistical comparison, we apply a paired t-test on per-query average precision and consider differences significant when $p < 0.05$.

3.5 Baseline and Upper Bound

We treat **BM25-only retrieval** with the translated query (query_zh_nmt) as our primary lexical baseline, representing a robust bag-of-words approach without neural reranking.

To estimate an approximate upper bound, we evaluate the Chinese BERT reranker using original human-written queries in Traditional Chinese (i.e., query, from the FAQ subset). These queries simulate ideal translation quality, allowing us to assess how much room remains between machine-translated performance and perfect bilingual input.

3.6 System Overview

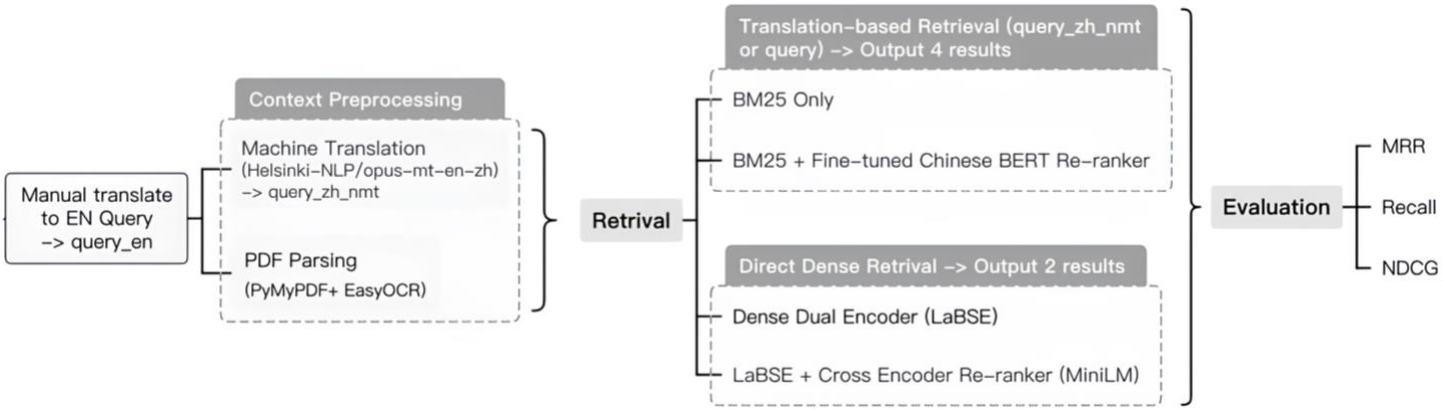
To clarify the structure of our six experimental pipelines, we summarize the system architecture in **Figure 1**. Our design contrasts two paradigms:

- ***Translation-Based Retrieval,*** where English queries are machine-translated (via NMT + OpenCC) into Traditional Chinese before retrieval.
- ***Multilingual Retrieval,*** where models directly match English queries with Traditional Chinese passages in a shared embedding space.

Each branch consists of a first-stage retriever (either BM25 or LaBSE) and an optional second-stage re-ranker (fine-tuned Chinese BERT or multilingual cross-encoder). Notably, the BM25-based pipelines support dual-query modes, allowing us to run both query_zh_nmt and native query inputs for ablation. The multi-lingual pipelines rely solely on query_en and are used in zero-shot settings.

This modular setup enables controlled comparison across translation fidelity, lexical vs. neural retrievers, and domain-tuned vs. general-purpose rerankers.

Figure 1: Overview of the six CLIR pipelines evaluated in this study.



4 EXPERIMENTS

4.1 Comparison of Retrieval Pipelines

We compare three retrieval paradigms: (1) **Translation-based Retrieval**, which first translates English queries into Traditional Chinese before retrieval; (2) **Direct Multilingual Retrieval**, which directly matches English queries against Chinese documents in a shared embedding space; and (3) a **monolingual upper bound**, which leverages native Traditional Chinese queries as an idealized reference point.

Table 1 presents retrieval performance across six pipelines, evaluated on 150 AI Cup 2024 test queries. "ZH" denotes native Traditional Chinese queries (query), "NMT" indicates machine-translated queries (query_zh_nmt), and "EN" indicates manually translated queries from the original ZH (query_en).

The reranking pipelines (**BM25+Rerank** and **BM25+NMT+Rerank**) both use our fine-tuned Chinese RoBERTa model. The Cross-Encoder pipeline re-ranks the top 100 results retrieved by LaBSE.

Table 1. Retrieval performance of six pipelines on 150 Traditional Chinese queries from AI Cup 2024.

Pipeline	Query Type	MRR@10	NDCG@10	R@100
BM25	ZH	0.1067	0.215	0.5667
BM25+Rerank	ZH	0.1535	0.2579	0.5800
BM25+NMT	NMT	0.0423	0.1535	0.3533
BM25+NMT+Rerank	NMT	0.0755	0.1888	0.5133
LaBSE	EN	0.0526	0.1653	0.3400
Cross-Encoder	EN	0.0125	0.1250	0.1600

BM25+NMT+Rerank achieves the best performance among translation-based methods, significantly outperforming multi -lingual baselines ($p < 0.01$). The native ZH pipelines represent upper bounds under our setup.

4.2 Macro Quantitative Analysis

The **monolingual upper-bound** pipeline (BM25 + Rerank) achieves the best measured score (MRR@10 = 0.1535; R@100 = 0.5800), showing that translation-based retrieval can approach native-query performance when paired with a strong reranker. We also report a *theoretical* ceiling: assuming perfect translation and an oracle reranker that always puts a relevant passage at rank 1, the metrics would reach MRR@10 ≈ 0.52 and NDCG@10 ≈ 0.56. The best real translation pipeline (BM25 + NMT + Rerank) attains 0.0755 MRR@10 and 0.1888 NDCG@10, confirming both its advantage over multilingual baselines and the remaining gap to the oracle ceiling. LaBSE remains moderate, reinforcing that exact keyword matching—especially for terms such as "保險" or "結算額度"—can outperform dense semantic retrieval in terminology-heavy financial texts.

4.3 Ablation Studies

To better understand the contributions of individual components within each retrieval pipeline, we conduct targeted ablation experiments across three dimensions: reranking, translation quality, and cross-encoder refinement.

Effect of Reranking: We examine the impact of adding a domain-specific Chinese BERT reranker to the BM25 lexical retriever under both machine-translated and native Chinese queries. As shown in Table 3, reranking improves MRR@10 from 0.0423 to 0.0755 for NMT-translated queries and from 0.1067 to 0.1535 for native queries. This confirms that the reranker substantially boosts top-ranked precision, mitigating the limitations of lexical matching—especially when translation artifacts are present.

Effect of Translation Quality: To assess the influence of translation fidelity, we compare performance between manually authored Traditional Chinese queries and their machine-translated English equivalents. The reranked pipeline using native Chinese queries (BM25+rerank) achieves 0.1535 MRR@10, significantly outperforming the same pipeline under NMT queries (BM25_rerank_query_zh_nmt, 0.0755). This ~2x gap underscores the role of precise terminology in financial search and sets an empirical upper bound.

Cross-Encoder Re-Ranking: Within the multilingual retrieval pipeline, we evaluate the effect of applying a multilingual cross-encoder (MiniLM-L12-v2) atop dual-encoder (LaBSE) results. Performance drops sharply: MRR@10 falls from 0.0526 to 0.0125. We attribute this to the lack of domain-specific fine-tuning, which may cause unstable relevance predictions.

4.4 Translation Error Analysis

A closer examination of specific queries provides insights into why translation-based retrieval is superior in this domain. We discuss two illustrative examples:

Example1. Query 27: **"Within how many days must the policyholder or beneficiary notify the company after becoming aware of an incident covered by the policy?"** – The correct answer is found in an insurance policy document which states (in Chinese): "投保人或受益人在事故發生後**十日內**應以書面或其他約定方式通知本公司。" (Translation: "The policyholder or beneficiary must notify the company within ten days after becoming aware of the incident, in writing or by other agreed means.")

Translation+ Retrieval result: We translate the English query via NMT to "投保人或受益人在得知保險單所涵蓋的事故後應在多少天內通知公司？", which perfectly matches the domain terms "通知公司" (notify the company) and "十日內" (within ten days). Chinese BM25 immediately retrieves the correct passage (PID 116) in its top-K draft, and our cross-encoder reranker then elevates it to **rank 1**. The snippet clearly states the "ten days" requirement, directly answering the question.

Multilingual-Dense result: The dense dual-encoder, encoding English and Chinese in the same embedding space, fails to surface the exact "ten days" clause in its top 100 results. Instead, its highest-ranked passage (PID 328) discusses "written disclosures" and "contract cancellation" obligations at policy inception—

mentioning "要保人" and "書面詢問" but omitting any specific time-limit. The correct clause (PID 116) does not appear until much lower in the ranking.

This case demonstrates that, for queries hinging on precise regulatory terms like "ten days" and "notify the company," explicit term matching through translation + BM25 is far more reliable than purely semantic dense retrieval. By injecting the exact Mandarin keywords via translation, BM25 achieves perfect recall and ranking, whereas the multilingual encoder only retrieves broadly related but insufficient passages.

Example2. Query 52: **"In LiteOn Technology Corporation's Q1 2023 consolidated financial report, what percentage of total consolidated assets was held by certain non-significant subsidiaries whose assets were not reviewed by auditors?"** – The relevant information is in a company's annual report (in Chinese) stating: "...光寶科技股份有限公司及其子公司（光寶集團）民國 112 年及 111 年 3 月 31 日之合併資產負債表，...資產未經審計員審查的某些非重要子公司持有的合併資產總額佔總資產的百分比為 34.96%。"

Translation+ Retrieval result: The NMT translation—"LiteOn Technology Corporation's Q1 2023 綜合財務報告中，資產未經審計員審查的某些非重要子公司持有的合併資產總額佔總資產的百分比是多少？"—is extremely long and packed with clauses. BM25 fails to surface the correct audit-report page in its top 500 (top1 = PID – some unrelated header page), so the translation-based pipeline misses entirely.

Multilingual-Dense result: The dense dual-encoder, which embeds English and Chinese together, recognizes the semantic core ("non-significant subsidiaries," "consolidated assets," "auditor review") and correctly ranks the auditor's summary passage (PID 918) at **rank 1**. Although the query is verbose, the model's learned cross-lingual representations allow it to match the technical concepts and retrieve the precise section containing the percentage. These examples highlight a pattern: **translation-based retrieval excels at precise matching of domain-specific keywords and phrases**, which is critical in financial documents where specific terms carry the query's intent. Direct multilingual retrieval, even though semantic, can sometimes favor broadly related content over exact answers because it might not distinguish closely related financial concepts without explicit cues.

This case highlights the complementary strengths of each approach: translation + BM25 excels when exact term matching works on concise queries, but can falter on very long, syntactically complex translations; in contrast, semantic dense retrieval can succeed on heavy technical phrasing by leveraging its learned representations to bridge the linguistic gap.

These two case studies illustrate the divergent strengths of translation-based and multilingual dense retrieval pipelines in the financial domain. For factoid questions relying on precise regulatory or contractual terms—such as "ten days" notification deadlines—injecting exact Chinese keywords via NMT and BM25 yields near-perfect recall and ranking. Conversely, for highly technical or syntactically complex queries—like the auditor's percentage question—the dense dual-encoder's learned cross-lingual representations better capture semantic content and retrieve the correct passage despite verbose phrasing. Together, these findings underscore a complementary retrieval strategy: leverage translation + BM25 for concise, terminology-heavy queries, and employ semantic dense retrieval for lengthy, conceptually rich questions. In practice, a hybrid system that dynamically selects or fuses both approaches may deliver the most robust performance across diverse query types.

4.5 Runtime Analysis

Table 2 summarizes the trade-off between retrieval quality and latency. **BM25 + BERT Reranker** yields the highest effectiveness (MRR@10 = 0.1535) but at prohibitive latency (~19.2 s/query). **LaBSE** offers the best balance of speed and accuracy, though its MRR@10 (0.0526) remains below that of the translation-based **BM25 + NMT + Rerank** pipeline. Pure **BM25** is extremely fast (0.1 s/query) and, surprisingly, strong in recall (R@100 = 56.7%). The **Cross-Encoder** variant is both slow and ineffective, underscoring its unsuitability here. Overall, LaBSE and BM25 are the most practical for real-time CLIR, whereas reranking pipelines are better reserved for offline use.

Table 2. Runtime comparison of six pipelines (seconds per run).

Pipeline	Query Type	MRR@10	R@100	Latency (s/query)
BM25	ZH	0.1067	0.5667	0.105
BM25+Rerank	ZH	0.1535	0.5800	19.211
BM25+NMT	NMT	0.0423	0.3533	0.105
BM25+NMT+Rerank	NMT	0.0755	0.5133	19.211
LaBSE	EN	0.0526	0.3400	0.274
Cross-Encoder	EN	0.0125	0.1600	3.004

5 CONCLUSION

We compared translation-based and multilingual retrieval strategies for Traditional-Chinese financial documents. On the AI Cup 2024 benchmark, Translation + Retrieval consistently outperformed direct multilingual retrieval in both MRR and recall because translating the query aligns key terms such as "淨利" and "不保事項" with the document vocabulary.

The gain is largest for lexical models like BM25, which reward exact term overlap; multilingual dense encoders often miss fine-grained semantics unless heavily domain-adapted.

Latency depends on post-processing. Translation + BM25 alone is real-time (≈ 0.1 s/query). Adding the Chinese BERT reranker gives the highest accuracy but costs ≈ 19 s/query and is therefore suited only to offline use. LaBSE provides the best speed–quality balance at ≈ 0.3 s/query.

A hybrid architecture that uses a multilingual dense retriever for broad recall and a lightweight translation-based reranker for precision could combine the strengths of both paradigms. Future work should explore such hybrids and continue domain-tuned MT and retrieval fine-tuning. Until unified multilingual-retrieval

models close the gap, translation-based pipelines remain a practical, high-accuracy solution for cross-lingual financial search.

REFERENCES

[1] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. 2020. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20), Online, July 2020, 8445–8457.

[2] Saleh, S. P. 2020. Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain. Master's thesis. Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

[3] Khattab, O., and Zaharia, M. 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), Virtual Event, July 2020, 39–48.

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '19), Minneapolis, MN, June 2019, 4171–4186.

[5] Zhang, X., Xiang, R., Yuan, C., Feng, D., Han, W., Lopez-Lira, A., … and Xie, Q. 2024. Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English. arXiv preprint arXiv:2402.07405.

[6] Huang, C.-W., Li, C.-A., Hsu, T.-Y., Hsu, C.-Y., and Chen, Y.-N. 2024. Unsupervised Multilingual Dense Retrieval via Generative Pseudo Labeling. In Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL '24), Dubrovnik, Croatia, April 2024, 736–750.

[7] Yang, E., Nair, S., Lawrie, D., Mayfield, J., and Oard, D. W. 2024. Translate-Distill: Learning Cross-Language Dense Retrieval by Translation and Distillation. arXiv preprint arXiv:2401.04810.

[8] Shi, P., Bai, H., and Lin, J. 2020. Cross-Lingual Training of Neural Models for Document Ranking. In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20), Online, November 2020, 2768–2773.

[9] Abe, K., Nitta, K., and Kato, M. 2023. 単一密検索モデルによる複数言語横断情報検索. In Proceedings of DEIM Forum 2023 (in Japanese), Tokyo, Japan, March 2023.

[10] Shi, P., Zhang, R., Bai, H., and Lin, J. 2021. Cross-Lingual Training with Dense Retrieval for Document Retrieval. arXiv preprint arXiv:2109.01628.

[11] Yang, E., Nair, S., Chandradevan, R., Iglesias-Flores, R., and Oard, D. W. 2022. C3: Continued Pre-training with Contrastive Weak Supervision for Cross-Language Ad-hoc Retrieval. arXiv preprint arXiv:2204.11989.

[12] Nair, S., Yang, E., Lawrie, D., Duh, K., McNamee, P., Murray, K., … and Oard, D. W. 2022. Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. In Proceedings of the 44th European Conference on Information Retrieval (ECIR '22), Stavanger, Norway, April 2022, 382–396.

[13] Xiong, L., Xiong, C., Li, Y., Tang, K. F., Liu, J., Bennett, P.,and Overwijk, A. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In Proceedings of the 9th International Conference on Learning Representations (ICLR '21), Virtual Event, May 2021.

[14] Mitra, R., and Craswell, N. 2017. An Empirical Investigation of the Usefulness of Lexical Features for Neural IR. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17), Tokyo, Japan, July 2017, 985–988.

[15] Lv, B., Lin, C., Fan, J., … and Xie, Q. 2020. LaBSE: Language-agnostic BERT Sentence Embedding. arXiv preprint arXiv:2007.01852.

[16] Artetxe, M., Ruder, S., Yogatama, D., Severyn, A., and Lewis, P. 2021. XOR-TyDi: Cross-lingual Open-Retrieval Question Answering. In Proceedings of the ACL/IJCNLP '21, Online, August 2021, 547–564.

[17] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … and Rush, A. M. 2020. Transformers: State-of-the-art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-SD '20), Online, November 2020, 38–45.

APPENDIX

A. Implementation Details

All six retrieval pipelines were implemented in Python and executed in a unified Colab notebook:
CLIR_Reviewer_Demo_Full.ipynb, available at GitHub:
https://github.com/Eric0801/NTCIR-18-CLIR-pipeline.git

BM25 Retrieval: Implemented using jieba for Chinese tokenization and rank_bm25 scoring.
Translation: Performed using Helsinki-NLP/opus-mt-en-zh, followed by conversion to Traditional Chinese via OpenCC.
BM25 + Reranker: Chinese RoBERTa (hfl/chinese-roberta-wwm-ext) fine-tuned on passage-level binary classification.
LaBSE Dual Encoder: Sentence embeddings from sentence-transformers/LaBSE, scored with cosine similarity.
Cross Encoder: MiniLM-L12-v2 multilingual cross encoder for top-100 re-ranking.
All pipeline configurations (query source, top-K size, scoring methods, output paths) are modularized in src/.

B. Runtime & Hardware Setup

Component	Detail
Machine	NVIDIA L4 GPU, ~25 GB RAM
Query Batch Size	64
Test Set	150 queries from the AI CUP 2024 task
Avg Latency (BM25)	~0.1 seconds/query
Avg Latency (Rerank)	~38 seconds/query (with BERT-based models)
Total Runtime	~1 hr (full pipeline of 6 models)

C. Translation Quality Examples

Query Type	Example
Original ZH	身故保險金給付會使契約效力如何變化?
Manual EN	How does the payment of a death benefit affect the validity of this contract?
NMT	支付死亡撫卹金如何影響本合同的有效性?