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

Yi-Ting Chiu, Zong-Han Bai

Independent research by an unaffiliated team of BBA students in QF, DS & CS, National Tsing Hua University

MOTIVATION

Traditional Chinese is underrepresented in CLIR, especially in finance where precise terminology matters. Dense models often miss key terms, returning semantically similar but legally incorrect results. This motivates revisiting translation-based retrieval—aligning queries and documents to improve term-level matching.

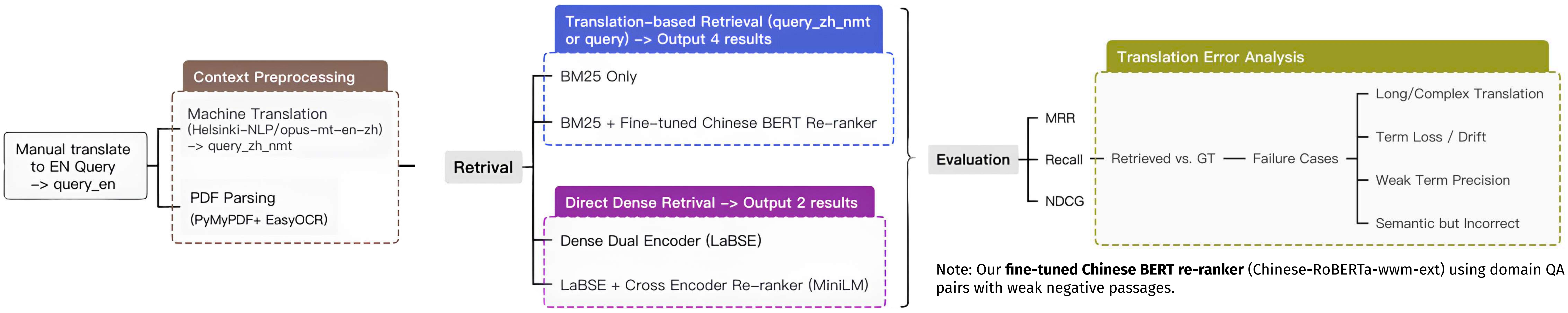
We evaluate its advantage over multilingual models on financial QA tasks. This final version integrates all 11 reviewer suggestions to inform robust CLIR in high-stakes, terminology-rich domains.

DATASET

We evaluate six retrieval pipelines using the **AI Cup 2024 dataset:**

- 1,600+ Traditional Chinese documents from financial reports, insurance contracts, and FAQs
- OCR fallback for scanned or image-heavy PDFs
- 150 **manually translated English** queries covering regulatory and operational intent
- All documents and query pairs are in the finance/legal domain, where retrieval failure carries real-world risk.

METHODOLOGY


Manual translate to EN Query -> query_en
Context Preprocessing
Machine Translation (Helsinki-NLP/opus-mt-en-zh) -> query_zh_nmt
PDF Parsing (PyMyPDF+ EasyOCR)
Retrival
Translation-based Retrieval (query_zh_nmt or query) -> Output 4 results
BM25 Only
BM25 + Fine-tuned Chinese BERT Re-ranker
Direct Dense Retrival -> Output 2 results
Dense Dual Encoder (LaBSE)
LaBSE + Cross Encoder Re-ranker (MiniLM)
Evaluation
MRR
Recall
NDCG
Translation Error Analysis
Retrieved vs. GT
Failure Cases
Long/Complex Translation
Term Loss / Drift
Weak Term Precision
Semantic but Incorrect

Note: Our **fine-tuned Chinese BERT re-ranker** (Chinese-RoBERTa-wwm-ext) using domain QA pairs with weak negative passages.

EXPERIMENT

Retrieval Setup

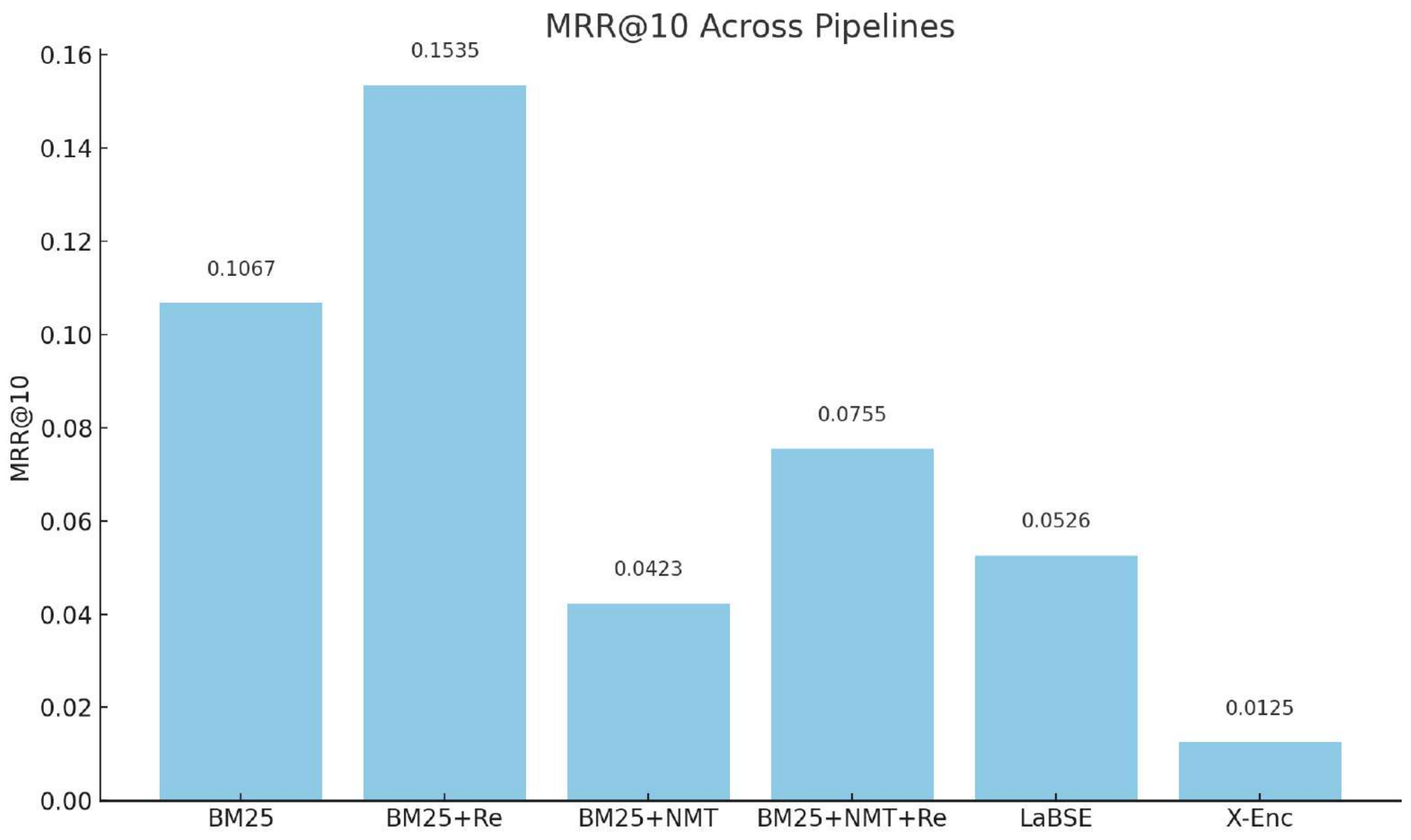
6 pipelines tested on 150 queries (ZH/NMT/EN).

BM25+Rerank yields best MRR@10.

Cross-encoder built on LaBSE top-100.

All rerankers = fine-tuned Chinese RoBERTa.

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



- BM25+Reranker gives best accuracy (MRR@10 = 0.1535) but is slow (~19s/query).

- LaBSE offers faster, balanced trade-off (0.27s/query).

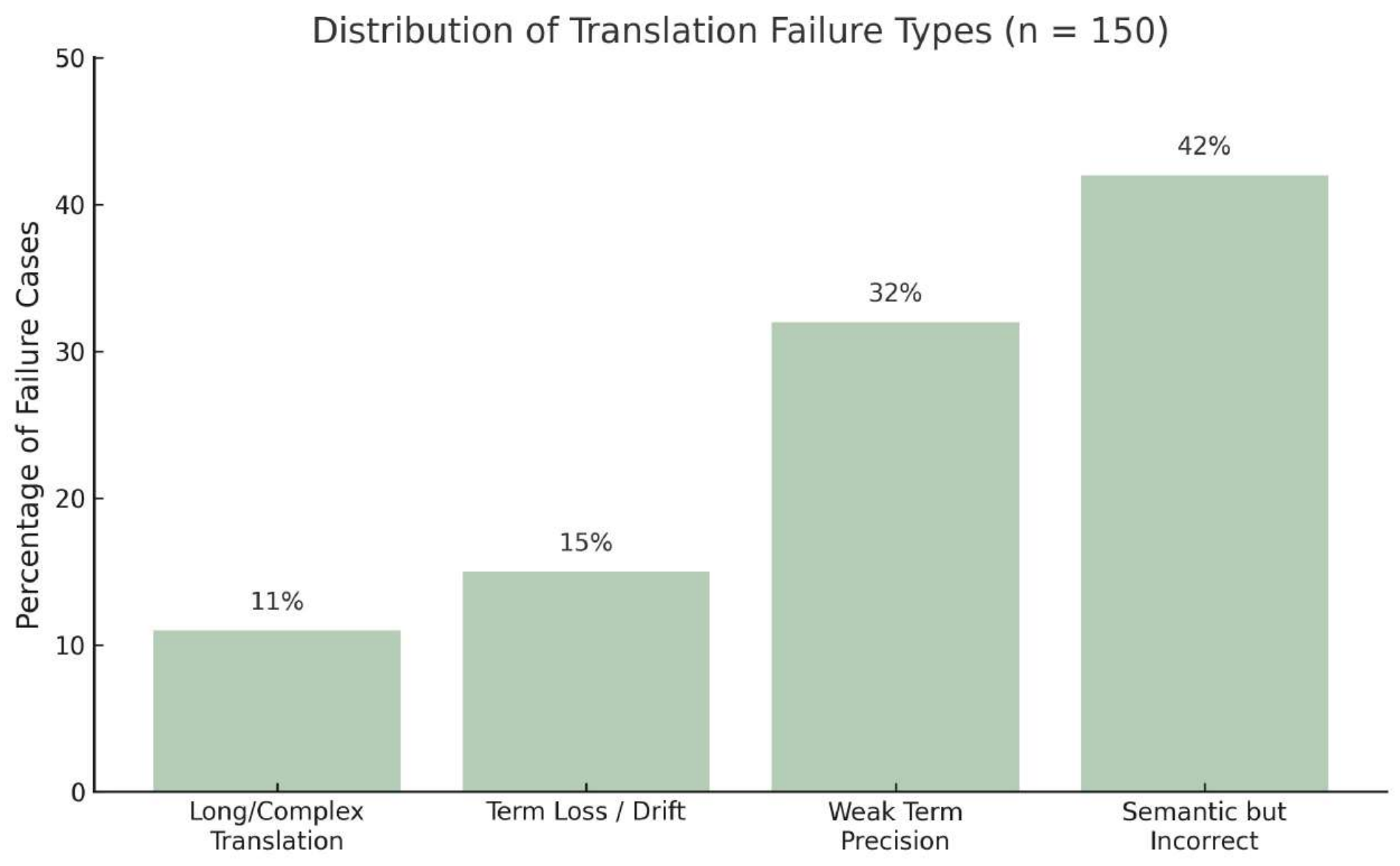
- Reranking boosts MRR 2× on both native and translated queries.

- Translation excels at regulatory keywords; dense models help with verbose concepts.

TRANSLATION ERROR ANALYSIS

We manually compared retrieved passages to gold answers and categorized translation-based failures into four types:
(1) Long or complex translations dilute key info,
(2) Term loss or drift from mistranslated terminology,
(3) Weak term precision leads to partial but insufficient matches
(4) Semantically plausible but incorrect results.

Most translation failures stem from term imprecision—preserved meaning, but distorted terminology hinders correct retrieval.



CONCLUSION

Translation-based retrieval, combining query translation, BM25, and a domain-adapted reranker, outperforms multilingual dense models on terminology-heavy financial documents.

These results highlight the importance of precise terminology alignment in specialized domains such as finance and insurance.

Multilingual encoders provide broad semantic coverage, but often underperform without fine-tuning.

Moving forward, we aim to design hybrid pipelines that combine multilingual embeddings for coverage with translated reranking for precision. We also plan to fine-tune cross-encoders, and lightweight on financial QA pairs and extend evaluation to compliance-related domains.

National Tsing Hua University
101, Section 2, Kuang-Fu Road, Hsinchu 300044, Taiwan R.O.C.
National Tsing Hua University, Taiwan

chew1tim@gapp.nthu.edu.tw
hu110048138@gapp.nthu.edu.tw

GitHub (Code & Models)
Open-source at:
github.com/Eric0801/NTCIR-18-CLIR-pipeline
Fork, replicate, build on top.