Evaluating Cross-Language Information Retrieval Models on Indonesian–Arabic Fiqh Texts: A Case Study

Muhammad Syifaurrohman [1] **Edi Winarko** [1]

Abstract

Cross-Language Information Retrieval (CLIR) for highly specialized domains, such as querying classical Arabic jurisprudence (Fiqh) using Indonesian, presents severe vocabulary mismatch and zero-resource training challenges. To resolve this lexical mismatch, we demonstrate that LLM-prompted domain-aware translation successfully captures strict legal terminology where standard machine translation fails. Concurrently, to address the absence of human relevance judgments, we employed the JH-POLO framework to generate synthetic in-domain triplets for fine-tuning a multilingual dense retriever. By synergizing these context-aware sparse signals with the semantic reasoning of the dense bi-encoder via Reciprocal Rank Fusion (RRF), we propose a highly effective hybrid architecture. Empirical evaluations on an expert-curated test collection reveal that while the lexical baseline dominates short queries, this late-fusion pipeline achieves the highest overall accuracy and acts as a robust safety net that consistently maximizes recall for complex, verbose queries. Code available at https://github.com/syifaurrr/MusIML26-Fiqh-CLIR

1. Introduction

Classical Arabic literature, traditionally known as *Kitab Kuning* (yellow books), serves as the primary lesson material and legal reference across Indonesian Islamic boarding schools (*pesantren*) (Rosidin et al., 2022). Although millions of students are educated within this system (Kementerian Agama Republik Indonesia, 2026), navigating these morphologically rich, unvocalized texts presents a steep linguistic barrier. To overcome this challenge, there is a critical need for Cross-Lingual Information Retrieval (CLIR) systems that empower users to access these classical Arabic manuscripts using their native Indonesian language.

Despite rapid advancements in the field of CLIR, current research predominantly focuses on high-resource language pairs (e.g., English-Arabic) and general-domain corpora such as news articles (Chen & Eickhoff, 2021). The intersection of CLIR for Indonesian-Arabic within the specialized domain of Fiqh remains largely unexplored. To address this critical literature gap, this paper introduces a comprehensive analysis of CLIR architectures designed to bridge Indonesian queries with classical Arabic texts, utilizing the seminal Fiqh manuscript, Fath al-Muin, as the primary corpus.

In this study, we evaluate and benchmark three distinct retrieval pipelines: (1) a translation-based sparse retrieval model (BM25), (2) a dense retrieval model utilizing a multilingual bi-encoder, and (3) a hybrid retrieval approach leveraging Reciprocal Rank Fusion (RRF). Our empirical results demonstrate that the hybrid RRF architecture successfully synthesizes the exact-match precision of lexical models with the semantic abstraction of dense retrievers. This approach not only overcomes the cross-lingual constraints but also establishes a highly robust retrieval baseline, significantly enhancing the accessibility of classical Islamic literature for non-native speakers.

2. Related Work

CLIR Paradigms and Preprocessing. CLIR predominantly employs translation-based sparse retrieval (Jaleel et al., 2004) or shared-embedding dense retrieval via multilingual Large Language Models (mLLMs) (Goworek et al., 2025). However, Arabic's highly derivational morphology presents contradictory preprocessing challenges. While sparse models require aggressive normalization and light stemming to mitigate vocabulary mismatch (Larkey et al., 2007), neural dense retrievers rely on sub-word tokenization and can degrade under aggressive stemming. This dichotomy necessitates decoupled preprocessing pipelines for hybrid systems.

Information Retrieval in Islamic Literature. Although

[1]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia. Correspondence to: Muhammad Syifaurrohman <muhammadsyifaurrohman@mail.ugm.ac.id>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Arabic CLIR pipelines perform well on general-domain datasets (Chen & Eickhoff, 2021; Abdallah et al., 2024), applications to classical Islamic jurisprudence (Fiqh) remain underexplored. Existing Fiqh literature predominantly focuses on monolingual retrieval using class-based term weighting (Fauzi et al., 2017) or adapted XLM-R models (Pavlova, 2025). Consequently, bridging the linguistic gap for Indonesian users remains an open challenge, necessitating hybrid retrieval architectures explicitly optimized for Indonesian-Arabic search in classical texts.

Domain Adaptation via Synthetic Data Generation. The primary barrier to deploying dense retrievers in Fiqh is domain mismatch, as models pre-trained on modern corpora fail to capture classical semantic nuances. To avoid the prohibitive cost of manual relevance judgments, recent advancements utilize Generative AI for synthetic data generation (Valentini et al., 2025; Yang et al., 2024). Methodologies like JH-POLO employ LLMs to autonomously generate domain-specific queries from the target corpus (Mayfield et al., 2023). By contrasting relevant passages with hard-negatives, JH-POLO produces high-fidelity triplets, effectively resolving data scarcity for fine-tuning dense models in zero-resource domains.

3. Methodology

3.1. Fiqh Corpus and Benchmark Construction

The retrieval corpus is constructed from the classical islamic jurisprudence text, *Fath al-Muin*, sourced from the OpenITI dataset (Nigst et al., 2020). To ensure structural integrity, the raw data underwent manual curation to resolve anomalies such as missing or misplaced passages, resulting in exactly 639 discrete passages. Each page of the book is treated as an individual document (passage).

To build a robust ground-truth benchmark, 153 test queries were formulated by three domain experts (Fiqh practitioners). Each expert selected 17 distinct passages and annotated each with three different types of queries: (1) short keyword queries, (2) natural language layperson questions, and (3) narrative case studies. All queries were formulated in Indonesian, while the target documents remained in classical Arabic. These queries and their corresponding relevance judgments serve as the ground truth for our evaluation.

Task Formulation and Evaluation. Given that exactly one ground-truth Fiqh passage was annotated per Indonesian query, we frame this as a *known-item retrieval* task. System performance is evaluated using Mean Reciprocal Rank (MRR) for exact top-ranking precision, and Success@k ($k \in \{10, 20, 50, 100\}$)—which is mathematically equivalent to Recall@k in single-target settings—to measure retrieval coverage.

3.2. Synthetic Data Generation

To address the scarcity of in-domain training data and the domain mismatch problem, we employed the JH-POLO methodology (Mayfield et al., 2023) to automatically generate synthetic training data using Large Language Models (LLMs). First, hard-negative passages were mined using BM25, filtered by specific criteria (e.g., a maximum 0.65 BM25 score ratio and a maximum 60% longest common substring) to ensure they were topically similar but semantically distinct from the positive passages.

Next, an LLM—prompted with the persona of a linguistic and Fiqh expert—was fed the positive and negative passage pairs to generate corresponding Indonesian queries. The resulting *(query, positive passage, negative passage)* triplets were subsequently validated using a multilingual cross-encoder. Triplets with a Softmax Margin score below a 0.15 threshold were filtered out to eliminate LLM hallucinations. From an initial generation of 3,091 triplets, 2,031 high-quality triplets were retained for fine-tuning the pre-trained mLLMs.

3.3. Dual-Preprocessing and Retrieval Pipelines

To accommodate the distinct characteristics of lexical and semantic matching, we implemented a dual-pipeline architecture. As a preliminary step, all Arabic documents were preprocessed by chunking, removing diacritics (harakat), and stripping elongation characters (tatweel).

For the sparse retriever, the Indonesian queries were translated into Arabic using two separate machine translation approaches: (1) Google Translate (NMT) and (2) a specifically prompted Gemini 3.1 Pro (LLM). Conversely, the dense retriever directly processed the original Indonesian queries using a shared multilingual embedding space. Finally, the outputs from both pipelines were fused to enhance overall retrieval performance.

3.3.1. Sparse Pipeline

In the sparse retrieval pipeline, both the translated Arabic queries and the Arabic documents underwent symmetrical lexical preprocessing. This included the normalization of characters (such as hamza, ligatures, and alef) and the application of a light-stemming algorithm to mitigate morphological variations without reducing words to their bare roots. We employed BM25 as the foundational lexical model, augmented with an RM3 pseudo-relevance feedback mechanism for query expansion to further bridge the vocabulary mismatch.

3.3.2. Dense Pipeline

The dense pipeline employs a bi-encoder architecture to map Indonesian queries and Arabic documents into a shared vec-

tor space. Unlike the sparse setup, text normalization was limited to removing diacritics and tatweels; light-stemming was strictly omitted to preserve the morphological integrity essential for the transformer's subword tokenization. We utilized mmBERT-base (Marone et al., 2025) as our representative dense model, fine-tuning it on the synthetic JH-POLO triplets via Multiple Negatives Ranking Loss (MNRL) to force separation between relevant and non-relevant passages. Finally, document embeddings were indexed using FAISS (IndexFlatIP) (Douze et al., 2024) for efficient inner-product retrieval.

3.3.3. Hybrid Fusion

To synergize the exact lexical matching of the sparse pipeline with the semantic reasoning of the dense model, we implemented Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). RRF naturally bypasses the need for complex score normalization between unbounded BM25 scores and dense cosine similarities by aggregating unsupervised rank positions. For a document d across a set of retrieval models M, the fusion score is calculated inline as $RRF(d) = \sum_{m \in M} 1/(k + rank_m(d))$. This late-fusion strategy effectively mitigates individual pipeline vulnerabilities and maximizes overall retrieval accuracy.

4. Experimental Results

Implementation Details. All computational experiments were conducted in a Kaggle environment equipped with dual NVIDIA Tesla T4 GPUs. For the sparse pipeline, BM25 hyperparameters were optimized to $b = 0.25$ and $k_1 = 1.0$, while RM3 pseudo-relevance feedback extracted 10 terms from the top 5 documents with a 0.8 interpolation weight. For the dense retriever, mmBERT was fine-tuned using Multiple Negatives Ranking Loss (MNRL) with a batch size of 8. We conducted a k-fold cross-validation grid search, utilizing the Spearman rank correlation of cosine similarity as an intrinsic heuristic to mitigate full-corpus evaluation costs. This established an optimal configuration of a 5×10^{-5} learning rate trained over 6 epochs. To ensure morphologically rich Fiqh passages were fully captured without truncation, the maximum sequence length was fixed at 512 tokens. Finally, the late-fusion RRF pipeline employed a standard stabilization constant of $k = 60$.

4.1. Main Retrieval Performance

Table 1 presents the overall performance of the evaluated retrieval pipelines. The empirical results reveal three critical findings regarding cross-lingual Fiqh retrieval.

First, standard machine translation systems struggle with the specialized vocabulary of classical Islamic jurisprudence. The lexical baseline utilizing Google Translate achieved an MRR of only 0.1853. This underperformance stems from the MT model's tendency to map queries into modern Arabic, failing to contextualize specific Fiqh terminology. Interestingly, applying RM3 pseudo-relevance feedback to these poor translations caused *query drift*—where expansion terms are extracted from irrelevant top-ranked documents—which exacerbated the failure and slightly degraded all retrieval metrics. In contrast, the LLM-prompted translation strategy (Gemini QT) successfully captured domain-specific lexical nuances, effectively doubling the initial precision. Because the initial retrieval yielded highly relevant documents, RM3 successfully extracted accurate expansion terms, steadily improving retrieval coverage across all Success@k depths and peaking at an MRR of 0.3929. This highlights that domain-aware translation is not only crucial for initial matching but serves as a strict prerequisite for effective query expansion.

Second, the results validate the efficacy of the JH-POLO synthetic data generation methodology. The off-the-shelf multilingual bi-encoder (mmBERT baseline) completely failed to map the Indonesian queries to the unvocalized Arabic texts, yielding an MRR of only 0.0182. However, fine-tuning the model using the synthetically generated in-domain triplets successfully bridged this severe domain gap, drastically improving the dense retriever's MRR to 0.3159. This demonstrates that LLM-generated synthetic data is a highly viable substitute for expensive human relevance judgments in zero-resource domains.

Finally, the proposed hybrid pipeline utilizing Reciprocal Rank Fusion (RRF) establishes a new state-of-the-art for this benchmark. By fusing the context-aware lexical signals of the Gemini-translated BM25+RM3 with the semantic abstraction of the fine-tuned mmBERT, the RRF model achieved the highest overall MRR of 0.4146. More importantly, the hybrid architecture acts as an absolute safety net, maximizing retrieval coverage and pushing the Success@100 metric to an impressive 98.04% (meaning the single correct ground-truth passage was successfully retrieved within the top 100 out of the 639 available chunks for almost all queries).

4.2. Query Typology Analysis

To understand the behavioral differences between the lexical and semantic models, we evaluated the retrieval performance across three distinct user query typologies, as detailed in Table 2.

For short, **Keyword queries (Type 1; see Table 3 for examples)**, the translation-based lexical model (BM25+RM3) demonstrated superiority, achieving an MRR of 0.5389. In this scenario, integrating the dense retriever via RRF degraded the overall MRR to 0.4665. This decline occurs because short keyword queries lack the contextual depth

Table 1. Overall retrieval performance on the Fiqh benchmark corpus. Best scores are highlighted in **bold**.

Model Architecture	MRR	S@10 (%)	S@20 (%)	S@50 (%)	S@100 (%)
Sparse Retrievers					
BM25 Baseline (Gemini QT)	0.3911	62.75	71.24	79.74	88.24
BM25 + RM3 (Gemini QT)	0.3929	64.71	73.86	83.01	89.54
BM25 Baseline (Google Translate QT)	0.1853	30.27	38.56	57.52	70.59
BM25 + RM3 (Google Translate QT)	0.1850	30.07	39.87	56.21	69.93
Dense Retrievers					
mmBERT-base (Baseline)	0.0182	4.58	7.84	18.30	30.07
mmBERT-base (JH-POLO Fine-Tuned)	0.3159	47.06	62.75	75.82	88.24
Hybrid Pipeline (Proposed)					
RRF (BM25+RM3 + mmBERT FT)	**0.4146**	**67.97**	**84.31**	**94.12**	**98.04**

Table 2. Performance across query typologies (1: Keyword, 2: Natural Question, 3: Case Study). The sparse models utilize Gemini-prompted translation.

Query Type	Model	MRR	S@10 (%)
Type 1	BM25+RM3 (Gemini)	**0.5389**	**82.35**
	mmBERT FT	0.2740	35.29
	RRF Hybrid	0.4665	76.47
Type 2	BM25+RM3 (Gemini)	**0.4160**	62.75
	mmBERT FT	0.3628	50.98
	RRF Hybrid	0.3693	**64.71**
Type 3	BM25+RM3 (Gemini)	0.2238	49.02
	mmBERT FT	0.3108	54.90
	RRF Hybrid	**0.4080**	**62.75**

required by the dense bi-encoder. Without sufficient surrounding context, the transformer model fails to capture deep semantic relationships, causing it to retrieve noisy candidates that dilute the highly accurate exact-match signals of the lexical baseline.

For intermediate, **Natural Question queries (Type 2)**, the lexical model maintained the highest top-1 precision (MRR 0.4160). However, the RRF pipeline began to demonstrate its recall-enhancing capabilities, surpassing all standalone models with a Success@10 of 64.71%. This indicates that as query verbosity gradually increases, semantic fusion starts to rescue relevant passages that lexical matching alone might rank outside the top results.

Conversely, RRF exhibited synergistic behavior on complex, **Case Study queries (Type 3)**. The lexical model severely underperformed in this category (MRR 0.2238) due to profound vocabulary mismatch and query drift caused by pseudo-relevance feedback on overly verbose inputs. However, fusing the semantic abstraction of mmBERT with the lexical safety net yielded a substantial performance spike, propelling the hybrid MRR to 0.4080—significantly outperforming both standalone architectures.

5. Conclusion

This paper presented a comprehensive case study on Cross-Language Information Retrieval (CLIR) for classical Arabic jurisprudence (Fiqh), bridging the severe domain gap for Indonesian queries. Our empirical evaluations demonstrate that standard machine translation baselines fail to capture the specialized vocabulary of classical Fiqh literature. By substituting standard MT with an LLM-prompted translation strategy, we significantly elevated the lexical retrieval baseline. Furthermore, we validated that the JH-POLO synthetic data generation pipeline successfully adapts multilingual dense retrievers to highly specialized, zero-resource Arabic domains without requiring human-annotated training data. Finally, we established that a hybrid architecture utilizing Reciprocal Rank Fusion (RRF) acts as a crucial synergistic mechanism. While lexical and semantic models struggle individually with different query typologies, their fusion consistently maximizes recall and provides a highly robust safety net for complex, verbose queries, establishing a strong foundation for future research in low-resource religious text retrieval.

Impact Statement

This research aims to democratize access to Fiqh literature, facilitating easier navigation of Classical Arabic texts for the largest Muslim population globally in Indonesia. By lowering the linguistic barrier through robust cross-lingual retrieval, this work empowers scholars, students, and laypeople to engage with primary sources. To mitigate theological misinterpretation from retrieval errors, this system functions strictly as an assistive tool, not a *fatwa* generator. All legal conclusions require expert verification.

References

Abdallah, A., Kasem, M. S., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., and Jatowt, A. Arabicaqa: A comprehensive dataset for arabic question answering. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. URL https://api.semanticscholar.org/CorpusID:268691723.

Chen, Z. and Eickhoff, C. The cross-lingual arabic information retrieval (claire) system. *ArXiv*, abs/2107.13751, 2021. URL https://api.semanticscholar.org/CorpusID:236493337.

Cormack, G. V., Clarke, C. L. A., and Büttcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009. URL https://api.semanticscholar.org/CorpusID:12408211.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazar'e, P.-E., Lomeli, M., Hosseini, L., and J'egou, H. The faiss library. *ArXiv*, abs/2401.08281, 2024. URL https://api.semanticscholar.org/CorpusID:267028372.

Fauzi, M. A., Arifin, A. Z., and Yuniarti, A. Arabic book retrieval using class and book index based term weighting. *International Journal of Electrical and Computer Engineering*, 7:3705–3710, 2017. URL https://api.semanticscholar.org/CorpusID:30084732.

Goworek, R., Macmillan-Scott, O., and Özyigit, E. B. Bridging language gaps: Advances in cross-lingual information retrieval with multilingual llms. *ArXiv*, abs/2510.00908, 2025. URL https://api.semanticscholar.org/CorpusID:281705770.

Jaleel, N. A., Allan, J., Croft, W. B., Diaz, F., Larkey, L. S., Li, X., Smucker, M. D., and Wade, C. Umass at TREC 2004: Novelty and HARD. In Voorhees, E. M. and Buckland, L. P. (eds.), *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004. URL http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf.

Kementerian Agama Republik Indonesia. Statistik pesantren — satu data kemenag, 2026. URL https://satudata.kemenag.go.id/statistik?modul=pesantren.

Larkey, L. S., Ballesteros, L., and Connell, M. E. Light stemming for arabic information retrieval. 2007. URL https://api.semanticscholar.org/CorpusID:17788645.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., and Durme, B. V. mmbert: A modern multilingual encoder with annealed language learning, 2025. URL https://arxiv.org/abs/2509.06888.

Mayfield, J., Yang, E., Lawrie, D. J., Barham, S., Weller, O., Mason, M., Nair, S., and Miller, S. Synthetic cross-language information retrieval training data. *ArXiv*, abs/2305.00331, 2023. URL https://api.semanticscholar.org/CorpusID:258426878.

Nigst, L., Romanov, M., Savant, S. B., Seydi, M., and Verkinderen, P. Openiti: a machine-readable corpus of islamicate texts, October 2020. URL https://doi.org/10.5281/zenodo.4075046.

Pavlova, V. Multi-stage training of bilingual islamic LLM for neural passage retrieval. In Yagi, S., Yagi, S., Sawalha, M., Shawar, B. A., AlShdaifat, A. T., Abbas, N., and Organizers (eds.), *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pp. 42–52, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.clrel-1.4/.

Rosidin, R., Andriani, F., Kawakip, A. N., and Fauzi, M. M. The development history of the yellow book (kitab kuning) as islamic textbooks in indonesia based on the philology perspective. In *Proceedings of the International Symposium on Religious Literature and Heritage (ISLAGE 2021)*, pp. 233–242. Atlantis Press, 2022. ISBN 978-94-6239-538-1. doi: 10.2991/assehr.k.220206.030. URL https://doi.org/10.2991/assehr.k.220206.030.

Valentini, F., Kozlowski, D., and Lariviere, V. CLIRudit: Cross-lingual information retrieval of scientific documents. In Adelani, D. I., Arnett, C., Ataman, D., Chang, T. A., Gonen, H., Raja, R., Schmidt, F., Stap, D., and Wang, J. (eds.), *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 226–242, Suzhuo, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-345-6. doi: 10.18653/v1/2025.mrl-main.16. URL https://aclanthology.org/2025.mrl-main.16/.

Yang, E., Lawrie, D. J., McNamee, P., and Mayfield, J. Extending translate-train for colbert-x to african language clir. *ArXiv*, abs/2404.08134, 2024. URL https://api.semanticscholar.org/CorpusID:269137335.

A. LLM Prompt Templates

A.1. Domain-Aware Query Translation Prompt (Gemini)

The exact system prompt utilized to translate the Indonesian user queries into classical Arabic is provided below. The prompt was executed in Indonesian to natively process the inputs. To ensure rigorous evaluation, the prompt enforces three strict constraints: (1) preserving the exact length and verbosity of the original query without summarization; (2) strictly mapping modern concepts to the classical Fiqh lexicon of *Fath al-Muin*; and (3) handling local Indonesian named entities via direct Arabic transliteration. *(Note: To ensure compatibility with standard LaTeX compilers, the original Arabic script in the prompt below has been converted using the Buckwalter transliteration scheme).*

```
Kamu adalah seorang Pakar Linguistik Arab Klasik dan Ahli Fiqih Mazhab Syafi'i,
khususnya spesialis leksikon dan terminologi yang digunakan oleh Syekh Zainuddin
Al-Malibari dalam kitab "Fathul Mu'in". Tugas utamamu adalah menerjemahkan kueri
pencarian (queries) dari Bahasa Indonesia ke Bahasa Arab (Arab Turats/Klasik).
Terjemahan ini akan diumpankan ke dalam mesin pencari AI (Dense Retriever) untuk
mencari teks yang relevan di dalam korpus kitab Fathul Mu'in. Saya akan
memberikan data dalam format CSV yang berisi dua kolom: [qid] dan [query].

ATURAN MUTLAK TERJEMAHAN:
1. PERTAHANKAN STRUKTUR DAN PANJANG ASLI (NO SUMMARIZATION)
Jangan pernah meringkas, memotong, atau merapikan kueri asli. Jika kueri aslinya
berupa kata kunci pendek, terjemahkan pendek. Jika kueri aslinya berupa cerita
panjang yang berbelit-belit (mengandung nama orang, latar tempat, kejadian),
terjemahkan seluruh cerita tersebut kata per kata secara utuh ke dalam bahasa Arab.
Mesin pencari kami sedang diuji ketangguhannya dalam membaca teks panjang (noise).

2. GUNAKAN DIKSI FIQIH KLASIK (LEKSIKON FATHUL MU'IN)
Ini adalah poin paling krusial. Ganti istilah umum/modern dengan istilah fiqih
klasik yang tepat:
- "Bunga bank/Rentenir/Pinjaman berbunga" -> [Buckwalter: AlrbA]
- "Kotoran hewan/Tahi" -> [Buckwalter: njAsp / rwv]
- "Bersetubuh/Berhubungan intim" -> [Buckwalter: wT' / jmAE]
- "Pakaian/Baju" -> [Buckwalter: vyAb / vwb]
- "Menggadaikan" -> [Buckwalter: rhn]
- "Menyewa/Sewa-menyewa" -> [Buckwalter: <jArp]
- "Mewakilkan/Menyuruh orang lain" -> [Buckwalter: twkyl]
- "Ganti rugi" -> [Buckwalter: DmAn]
- "Mabuk/Teler" -> [Buckwalter: skr]

3. PENANGANAN NAMA DAN LOKALITAS
Jika ada nama orang Indonesia (Budi, Joko, Tono) atau tempat, cukup lakukan
transliterasi hurufnya ke dalam abjad Arab (contoh: [Buckwalter: twnw, jwkw, bwdy])
tanpa perlu mengubahnya menjadi nama Arab.

FORMAT OUTPUT: Keluarkan hasil terjemahanmu SATU KALI SAJA dalam format tabel CSV
(dipisahkan tanda koma) dengan kolom: qid,query_indo,query_gemini
PENTING: Jangan memberikan penjelasan, basa-basi, atau komentar apa pun di luar blok
CSV tersebut agar outputmu bisa langsung saya simpan menjadi file. Pastikan tidak
ada tanda kutip ganda yang merusak struktur CSV.
```

A.2. JH-POLO Synthetic Data Generation Prompt

The JH-POLO framework utilizes the following prompt template to construct robust training triplets. Executed in Indonesian, the prompt instructs the LLM to generate queries that perfectly match the positive document (*pos_text*) while remaining

strictly unanswerable by a topically similar hard negative document (*neg_text*). To ensure the dense retriever learns to handle diverse inputs, the prompt explicitly forces the generation of exactly ten queries per pair, strictly categorized into the three evaluated verbosity levels: Keyword (Type 1), Natural Question (Type 2), and Case Study (Type 3). The structured JSON output format is mandated for automated pipeline integration.

```
Anda adalah Asisten Peneliti Pakar Fiqih Syafi'i dan Ahli Information Retrieval (IR).
Tugas utama Anda adalah membantu membangkitkan data latih kueri sintetis (Synthetic
Query Generation) menggunakan metodologi JH-POLO.

Saya akan memberikan sekumpulan data yang berisi pasangan teks bahasa Arab dari kitab
Fathul Mu'in. Setiap baris data memiliki:
- pos_id & pos_text (Teks A / Dokumen Positif / Hard Positive)
- neg_id & neg_text (Teks B / Dokumen Negatif / Hard Negative)
Teks A dan Teks B memiliki kemiripan topik secara leksikal, namun mengandung detail
hukum, syarat, atau konteks Fiqih yang berbeda.

TUGAS ANDA:
Untuk setiap pasangan teks yang saya berikan, buatlah 10 kueri pencarian berbahasa
Indonesia dengan mematuhi aturan ketat berikut:

ATURAN JH-POLO (Keketatan Logika):
- Kueri HARUS merepresentasikan "Information Need" yang jawabannya ada secara spesifik
  dan akurat di dalam Teks A (pos_text).
- Kueri HARUS TIDAK BISA dijawab, atau akan menghasilkan jawaban yang salah/menyesatkan,
  jika dicarikan ke Teks B (neg_text).
- Fokuslah mencari "selisih makna" atau perbedaan detail terkecil (seperti pengecualian,
  syarat spesifik, atau perbedaan objek) antara Teks A dan Teks B untuk merumuskan kueri.

ATURAN TIPE KUERI (Komposisi 10 Kueri):
Setiap pasangan teks wajib memiliki tepat 10 kueri yang terbagi menjadi 3 tipe:

1. 3 Kueri TIPE 1 (Kata Kunci Pendek / Gaya Santri):
   Berupa frasa pendek (2-5 kata), langsung ke tujuan, dan menggunakan istilah teknis
   Fiqih (Arab yang di-Indonesiakan seperti "syarat sah wudhu", "hukum fasakh").
   Dilarang menggunakan kalimat tanya.

2. 4 Kueri TIPE 2 (Kalimat Tanya / Gaya Awam):
   Berupa pertanyaan langsung yang lengkap dan mengalir natural (menggunakan 5W1H).
   Gunakan bahasa Indonesia sehari-hari. Hindari istilah teknis Fiqih jika ada padanan
   bahasa Indonesianya yang lebih umum (misal: gunakan "air bekas mandi" alih-alih
   "air musta'mal").

3. 3 Kueri TIPE 3 (Studi Kasus / Deskriptif Panjang):
   Berupa skenario cerita atau studi kasus deskriptif yang dialami masyarakat
   sehari-hari (gunakan nama fiktif seperti Budi, Siti, dll., dan deskripsikan
   situasinya). Skenario ini harus menanyakan solusi hukum yang jawabannya hanya
   ada di pos_text dan bukan di neg_text.

FORMAT OUTPUT (JSON):
[
  {
    "pos_id": "ID_A",
    "neg_id": "ID_B",
    "tipe_1": ["...", "...", "..."],
```

```
    "tipe_2": ["...", "...", "...", "..."],
    "tipe_3": ["...", "...", "..."]
  }
]
```

B. Qualitative Query Examples

Table 3 presents a representative sample of queries from the expert-curated test collection, illustrating the three distinct query typologies (Keyword, Natural Question, and Case Study). To ensure precise orthographic representation and full compatibility with standard document compilers, all Arabic outputs are presented using the standard Buckwalter transliteration scheme.

The table highlights the qualitative translation differences between the standard machine translation baseline (Google Translate) and the LLM-prompted domain-aware translation (Gemini).

Table 3. Examples of query typologies. Arabic outputs are presented in Buckwalter transliteration. Gemini captures precise Fiqh terminology, whereas Google Translate defaults to modern conversational Arabic.

Typology	Original Indonesian Query	Google Translate	Gemini
Type 1	Kesunnahan membaca basmalah dalam wudhu	mn Alsnp qrA'p Albsmlp >vnA' AlwDw'	snyp qrA'p Albsmlp fy AlwDw'
	nikahnya perempuan dzimmi	zwAj Amr>p mn >hl Al*mp	nkAH Al*myp
Type 2	Bagaimana hukum membaca basmalah di tengah-tengah wudhu?	mA hw Hkm qrA'p Albsmlp fy mntSf AlwDw'?	mA Hkm qrA'p Albsmlp fy >vnA' AlwDw'?
	Apa saja syarat perempuan yang boleh dinikahi	mA hy Al$rwT Alty tsmH llmr>p bAlzwAj?	mA hy $rwT Almr>p Alty yjwz nkAHhA?
Type 3	Fahmi merupakan santri yang selalu melakukan kesunnahan ketika berwudhu, mulai dari membaca basmalah, berkumur, memasukkan air ke dalam hidung dan kesunnahan-kesunnahan lainnya. Suatu ketika, Fahmi lupa melakukan membaca basmalah di awal wudhu, dan ia teringat setelah berada di tengah-tengah wudhu. Apakah Fahmi tetap disunnahkan membaca basmalah?	fhmy TAlbN yHrS dA}mFA ElY >dA' Alsnn >vnA' AlwDw', bd'FA mn Albsmlp, vm AlmDmDp, vm Astn$Aq AlmA', wgyr *lk mn Alsnn. fy <HdY AlmrAt, nsy fhmy Albsmlp fy bdAyp AlwDw', vm t*krhA fy mntSfh. fhl yuEd~ mn Alsnn >n yqr> fhmy Albsmlp?	fhmy TAlb Elm yfEl dA}mA snn AlwDw', mn qrA'p Albsmlp, wAlmDmDp, wAlAstn$Aq wgyrhA mn Alsnn. w*At mrp, nsy fhmy qrA'p Albsmlp fy >wl AlwDw', wt*krhA whw fy >vnA' AlwDw'. fhl ysn lfhmy qrA'p Albsmlp?
	Marak terjadi di zaman sekarang pernikahan beda agama. Bagaimana islam memandang hal tersebut?	>SbHt AlzyjAt byn Al>dyAn Almxtlfp $A}Ep b$kl mtzAyd h*h Al>yAm. kyf ynZr Al<slAm <lY h*A Al>mr?	kvr fy h*A AlzmAn AlnkAH bAxtlAf Aldyn. fkyf ynZr Al<slAm <lY *lk?

C. Evaluation Scope and Limitations

We acknowledge certain evaluation scopes and limitations in this work. First, regarding the training of the dense retriever, target passages from the evaluation set were not explicitly excluded from the synthetic training corpus. Because this framework targets a specific historical text in a closed-book retrieval scenario, the model's learning is evaluated based on its ability to accurately match the indexed passages against completely unseen, human-generated queries.

Second, while the expert-curated benchmark provides high-quality, domain-specific relevance judgments, the modest size of the test collection (153 queries) and the single known-item evaluation setup preclude broad statistical significance testing. The empirical findings primarily serve as a foundational case study. Future work will aim to extend this hybrid retrieval framework to larger, multi-book classical Arabic benchmarks to further validate the consistency and generalizability of the proposed models.