The Data Leakage Challenge in Vietnamese Legal Information Retrieval

Duc-Nhan Nguyen

May 26, 2024

Abstract

The integrity of evaluation datasets is paramount for objectively assessing Information Retrieval (IR) systems. This study investigates data leakage phenomena by comparing BM25 and four Dense Retrieval models (`aiteamvn`, `bkai`, `halong`, `hcmute`) across three dataset types: (1) "Extend,"a novel, manually-labeled high-integrity dataset; (2) "Zalo AI,"an internal dataset (potentially from Zalo AI Challenge 2021) suspected of test set leakage into training data; and (3) "Auto-labeled Subsets,"representing larger, automatically generated data. Using standard IR metrics, we found that `bkai` and `hcmute` exhibited anomalously high performance solely on the Zalo AI dataset, strongly suggesting data leakage. Conversely, BM25, `halong`, and `aiteamvn` showed consistent performance on "Extend"and "Auto-labeled Subsets,"establishing these as more reliable benchmarks. This research highlights the critical need for data integrity and proposes "Extend"and "Auto-labeled Subsets"for more realistic IR performance assessment, cautioning against results from potentially compromised datasets.

Keywords: Information Retrieval, BM25, Dense Retrieval, Data Leakage, Performance Evaluation, Manually Labeled Dataset, Auto-labeled Dataset, Zalo AI, Extend Dataset.

1 Introduction

Information Retrieval (IR) aims to retrieve relevant information from large data repositories. While Dense Retrieval models, leveraging vector embeddings, have shown significant advancements over traditional methods like BM25, their objective evaluation demands standardized datasets and rigorous protocols, critically including data integrity. Data leakage—where test set information inadvertently influences model training—can severely inflate performance metrics, misrepresenting a model's true capabilities on unseen data.

This study confronts this challenge by evaluating IR models across three distinct data paradigms:

1. **Extend Dataset**: A meticulously curated, manually labeled dataset, serving as a "gold standard"for clean evaluation.

2. **Zalo AI Dataset**: An internal dataset where suspicions of data leakage from its own test/index set into model training data have arisen, potentially skewing performance.

3. **Auto-labeled Subsets (test_0, test_1, test_2)**: Three automatically labeled data subsets, whose averaged results represent performance on larger-scale, potentially noisier, but more practically generated data.

By comparing a BM25 baseline against four Dense Retrieval models (`aiteamvn`, `bkai`, `halong`, and `hcmute`) on these datasets, this research aims to:

- Quantify performance variations across different data labeling methodologies and integrity levels.

- Investigate and provide evidence for data leakage phenomena, particularly concerning the Zalo AI dataset.

- Offer recommendations for robust model selection and reliable evaluation practices in IR, emphasizing the value of datasets like Extend and the averaged Auto-labeled Subsets for assessing real-world applicability.

Understanding these dynamics is crucial for advancing fair and accurate IR system benchmarking.

2 Methodology

This section outlines the datasets, retrieval models, and evaluation metrics employed in this study.

2.1 Datasets

Two datasets with distinct characteristics were used:

- **Extend Dataset**: Constructed with a strong emphasis on "cleanliness,"ensuring no significant overlap between potential training, testing, and index sets. This dataset serves as a reliable baseline for assessing true model performance.
- **Zalo AI Dataset**: An internal dataset previously used in challenges and internal research. Preliminary observations suggest potential data leakage, where data from various sources (e.g., training sets of publicly available models) might have inadvertently appeared in the Zalo AI test or index set.
- **Auto-labeled Subsets (test_0, test_1, test_2)**: Three subsets of a larger dataset, labeled automatically with specific constraints. These represent a scenario with potentially noisier labels compared to the manually curated Extend dataset and are used to evaluate model performance under such conditions by averaging results across them.

2.2 Retrieval Models

The following retrieval models were benchmarked in this study. For brevity, models are often referred to by their short names (`bkai`, `halong`, `hcmute`, `aiteamvn`) throughout the report. Further details and Hugging Face model card links are provided in Appendix A.

2.2.1 Baseline Model

- **BM25**: A classic lexical ranking model implementing the Okapi BM25 algorithm. It is a widely adopted and strong baseline in IR systems, serving as a non-neural comparison point.

2.2.2 Dense Retrieval Models

These models leverage vector embeddings to encode queries and documents into dense vector spaces. Retrieval is performed by computing similarity (typically cosine similarity) between query and document embeddings. The specific pre-trained Vietnamese embedding models evaluated are:

- **`bkai` (BKAI Bi-Encoder)**: A Vietnamese bi-encoder model developed by BKAI Foundation Models. Available at: `https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder`
- **`halong` (Halong Embedding)**: An embedding model by Hieu H. Available at: `https://huggingface.co/hiieu/halong_embedding`
- **`hcmute` (DEk21 HCMUTE Embedding)**: An embedding model by Huy Dang. Available at: `https://huggingface.co/huyydangg/DEk21_hcmute_embedding`

- **aiteamvn (AITeamVN Embedding)**: A Vietnamese embedding model by AITeamVN. Available at: `https://huggingface.co/AITeamVN/Vietnamese_Embedding`

All the above Dense Retrieval models, along with BM25, were evaluated on the Extend dataset, the Zalo AI dataset, and the auto-labeled subsets (test_0, test_1, test_2).

2.3 Evaluation Metrics

Model performance was assessed using the following standard IR metrics. For brevity in tables, `hit_rate@k` is often denoted as `hit@k`, `MRR@5` as `mrr@5`, and `NDCG@5` as `ndcg@5`.

- **Hit Rate@k (Accuracy@k)**: The proportion of queries for which at least one relevant document is found within the top k retrieved results. This study primarily uses k=1, 3, 5.
- **Recall@5 (Recall@k)**: The proportion of relevant documents found within the top k results. We focus on Recall@5 as presented in the generated tables.
- **Mean Reciprocal Rank (MRR@5)**: The average of the reciprocal of the rank of the first relevant document, capped at the top k. We focus on MRR@5.
- **Normalized Discounted Cumulative Gain (NDCG@5)**: A measure of ranking quality that considers the relevance level of documents and their positions in the result list, capped at k. We focus on NDCG@5.

3 Experimental Setup

Each model was evaluated on the Extend dataset, the Zalo AI dataset, and the three auto-labeled test subsets (test_0, test_1, test_2). For the auto-labeled subsets, performance metrics were averaged across the three runs to provide a more stable measure. A total of five distinct evaluation scenarios were thus considered for each model: Extend, Zalo AI, and the average of test_0, test_1, test_2.

4 Results

This section presents the performance results of the evaluated models across the different datasets. The primary focus will be on the manually labeled "Extend Dataset"and the average performance across three "Auto-labeled Subsets"(test_0, test_1, test_2). Performance on the "Zalo AI Dataset"is also presented for completeness, though it is analyzed with caution due to suspected data leakage (see Section 5).

4.1 Comparative Performance on Key Datasets

Table 1 consolidates the performance metrics for each model on the "Extend Dataset"(manually labeled) and the average of the "Auto-labeled Subsets". This allows for a direct comparison of how models perform on high-quality, clean data versus potentially noisier, automatically generated data.

Table 1: Comparative Model Performance: Extend Dataset vs. Average Auto-labeled Subsets.

	Extend Dataset (Manually Labeled)						**Average Auto-labeled Subsets**					
Model	hit@1	hit@3	hit@5	recall@5	mrr@5	ndcg@5	hit@1	hit@3	hit@5	recall@5	mrr@5	ndcg@5
aiteamvn_Dense	**0.776**	**0.912**	**0.926**	**0.891**	**0.839**	**0.834**	**0.713**	**0.873**	**0.914**	**0.877**	**0.796**	**0.799**
bkai_Dense	0.360	0.561	0.630	0.594	0.464	0.482	0.318	0.486	0.558	0.522	0.409	0.424
bm25_BM25	0.508	0.716	0.785	0.745	0.615	0.635	0.472	0.654	0.722	0.686	0.568	0.583
halong_Dense	0.487	0.697	0.757	0.716	0.593	0.610	0.473	0.667	0.734	0.694	0.575	0.589
hcmute_Dense	0.501	0.661	0.745	0.698	0.593	0.599	0.388	0.574	0.648	0.611	0.486	0.503

4.2 Performance on Zalo AI Dataset (Suspected Leakage)

Table 2 presents the model performances on the Zalo AI dataset. As discussed in Section 5, these results should be interpreted with caution due to strong indications of data leakage.

Table 2: Performance on Zalo AI Dataset

Model	hit@1	hit@3	hit@5	recall@5	mrr@5	ndcg@5
aiteamvn_Dense	0.759	0.911	0.934	0.934	0.835	0.860
bkai_Dense	0.731	0.875	0.906	0.906	0.802	0.828
bm25_BM25	0.577	0.780	0.822	0.822	0.676	0.713
halong_Dense	0.578	0.734	0.781	0.781	0.660	0.690
hcmute_Dense	**0.873**	**0.966**	**0.978**	**0.978**	**0.920**	**0.935**

5 Discussion

The experimental results, particularly when visualized, reveal significant variations in model performance contingent on the evaluation dataset. These discrepancies strongly suggest the influence of dataset characteristics, primarily potential data leakage in the Zalo AI dataset and the differing nature of labeling methodologies (manual vs. automatic).

5.1 Model Performance Consistency and Dataset Integrity

A key observation from the scatter plots (Figure 1), particularly when comparing performance on the Extend dataset versus the Zalo AI dataset, is the commendable consistency exhibited by several models. Specifically, the traditional bm25 model, along with Dense Retrieval models halong and aiteamvn_Dense, demonstrate

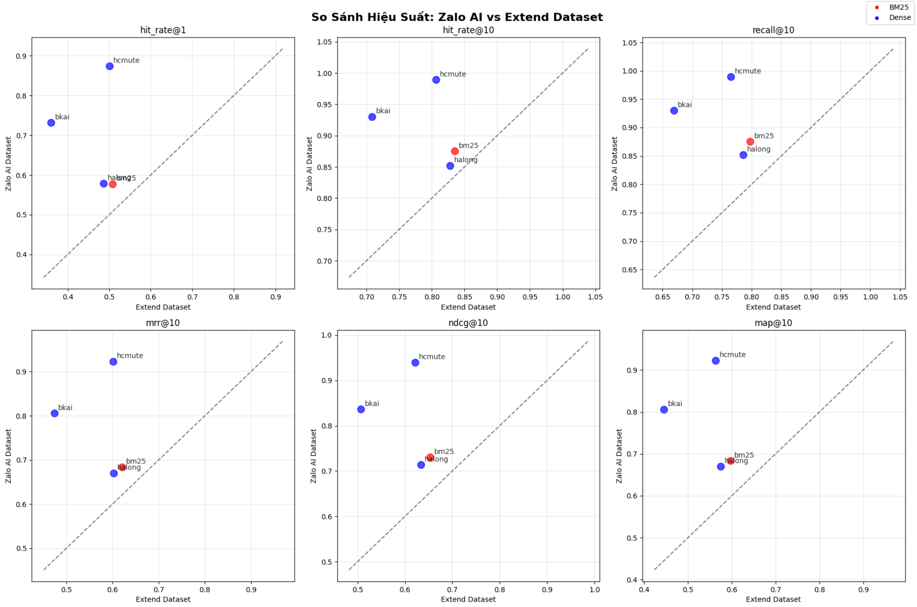

Figure 1: Performance Comparison: Zalo AI vs. Extend vs. Auto-Labeled Datasets. Scatter plots illustrating model performance shifts across different datasets for key metrics. Proximity to the diagonal when comparing Extend and Zalo AI suggests consistent performance, while significant deviation towards the Zalo AI axis by some models may indicate anomalous gains.

performance levels on Zalo AI that are largely proportional to their scores on the Extend dataset. Their data points tend to lie closer to the diagonal line in such comparisons, indicating that their performance scales predictably across these two environments. This consistent behavior across datasets with differing origins and potential characteristics suggests that these models are generalizing well and are less affected by dataset-specific anomalies.

Crucially, the stable performance of `bm25` (a lexical model inherently less prone to training data memorization), `halong` (reportedly trained on independent data), and `aiteamvn_Dense` across both Extend and Zalo AI datasets lends significant credibility to the Extend dataset as a robust and fair benchmark. Their predictable scaling implies that Extend effectively captures generalizable IR challenges, making it a suitable baseline for evaluating true model capabilities without the confounding factor of potential data leakage present in Zalo AI. The performance on the Auto-labeled Subsets (Table 1), while generally lower due to automatic labeling noise, further shows these models (`bm25`, `halong`, `aiteamvn_Dense`) maintain a reasonable level of performance, underscoring their robustness.

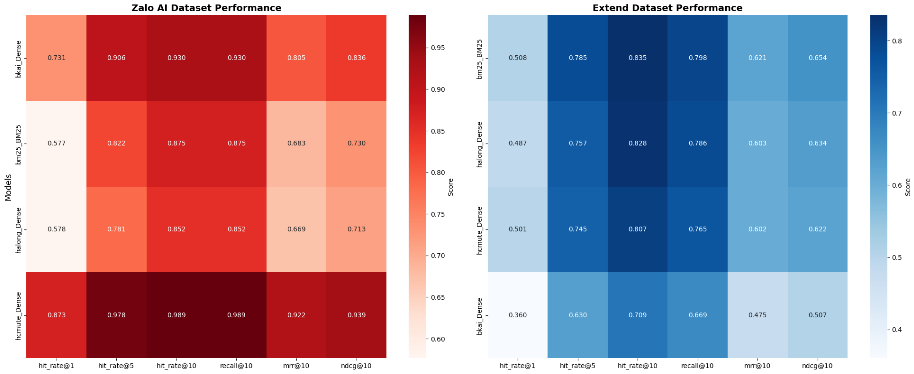

Figure 2: Performance Heatmaps across Datasets. Heatmaps for Zalo AI, Extend, and Average Auto-labeled datasets, visualizing relative model strengths on each.

5.2 Anomalous Performance and Suspected Data Leakage: The Case of `bkai` and `hcmute`

In stark contrast to the models discussed above, `bkai_Dense` and `hcmute_Dense` exhibit exceptionally and disproportionately high performance on the Zalo AI dataset (Table 2 and visualized in Figures 1 and 2). When visualized in the scatter plots comparing Zalo AI to Extend (Figure 1), these two models deviate significantly from the diagonal, showing massive performance jumps on Zalo AI that are not mirrored on the Extend dataset or the Auto-labeled Subsets. For instance, `bkai`'s `NDCG@5` surged from a modest 0.482 on Extend to 0.828 on Zalo

AI, and `hcmute`'s `NDCG@5` leaped from 0.599 on Extend to an outstanding 0.935 on Zalo AI.

This anomalous behavior strongly suggests that `bkai` and `hcmute` have benefited from data leakage. The lack of transparency regarding the training data for `hcmute` ("in-house dataset") and the explicit mention of Zalo 2021 Challenge data in `bkai`'s README file further substantiate these suspicions. It is highly probable that these models were trained on data that included parts or all of the Zalo AI test set, allowing them to "memorize"correct answers rather than learn generalizable retrieval strategies. Their significantly lower performance on both the "clean"Extend dataset and the Auto-labeled Subsets (Table 1) indicates that their inflated Zalo AI scores do not reflect their true capabilities on unseen, independently curated data. This underscores the critical problem of data leakage in skewing evaluation results and the importance of using verified, independent datasets like Extend for genuine performance assessment.

5.3 Model-Specific Performance Profiles Summary

Based on the collective evidence (Tables 1, 2, and Figures 1, 2):

- **`bm25`, `halong`, and `aiteamvn_Dense`**: These models display consistent and reliable performance across the Extend and Auto-labeled datasets, with moderate and expected scaling on the Zalo AI dataset. They serve as strong benchmarks for general IR capabilities. `halong` particularly shows good robustness to varying label quality.

- **`bkai` and `hcmute`**: Their Zalo AI performance is exceptionally high but appears to be an artifact of data leakage. Their scores on Extend and Auto-labeled Subsets are significantly lower, making their Zalo AI results unreliable indicators of general IR competence.

6 Conclusion

This comprehensive study underscores the critical impact of dataset integrity and labeling methodology on the perceived performance of information retrieval models. The findings provide compelling evidence that potential data leakage in the Zalo AI dataset significantly inflates the performance metrics of certain Dense Retrieval models, notably `bkai` and `hcmute`. Conversely, the manually labeled Extend dataset and the averaged results from Auto-labeled Subsets (test_0, test_1, test_2) offer more reliable benchmarks for assessing true model capabilities and robustness.

6.1 Key Conclusions

1. **Data Leakage Skews Zalo AI Performance:** The performance of `bkai` and `hcmute` on the Zalo AI dataset is anomalously high compared to their results on both the Extend and Auto-labeled datasets. This discrepancy,

corroborated by available information on their training data (or lack thereof for `hcmute`), strongly indicates data leakage, rendering their Zalo AI scores unreliable for gauging general IR competence.

2. **Extend and Auto-labeled Datasets as Reliable Benchmarks:** The Extend dataset, due to its manual and careful curation, serves as a robust baseline for "clean"performance evaluation. The averaged performance on Auto-labeled Subsets, while generally lower due to inherent noise, provides valuable insights into model robustness and performance on larger-scale, automatically generated data.

3. **Consistent Performers Offer Generalizable Insights:** The traditional BM25 model, along with Dense Retrieval models `halong` and `aiteamvn`, exhibit more consistent and predictable performance across the Extend, Auto-labeled, and Zalo AI datasets (when accounting for expected generalization rather than leakage-driven inflation). This consistency highlights their better generalization capabilities and the validity of Extend and Auto-labeled Subsets for evaluation. `halong` particularly demonstrates good stability across varying data qualities.

4. **Data Integrity is Paramount:** Data leakage is a serious procedural flaw in IR evaluation. It can lead to misleading conclusions about model superiority, flawed model selection for deployment, and hinder genuine progress in the field.

6.1.1 For Model Evaluation and Selection

- **Mandate Cross-Dataset Evaluation:** Models should be evaluated on multiple, diverse, and verifiably independent datasets to assess true generalization and robustness. Performance on a single, potentially compromised dataset is insufficient.

- **Investigate Anomalous Gains:** Extremely high performance gains on one specific dataset compared to others, especially for models with opaque training data, should be a red flag for potential data leakage and warrant deeper investigation.

- **Model Selection for Production Based on Reliable Data:**

  - **`halong` and `aiteamvn`**: Recommended as strong candidates for production due to their consistent performance on Extend and Auto-labeled datasets. `halong`'s stability and embedding-slimming are notable.
  - **`bm25`**: Remains a crucial, fast, and data-leak-unaffected baseline, particularly valuable for new domains or as a sanity check.
  - **`hcmute` and `bkai`**: Their current Zalo AI scores should be disregarded for production decisions. These models require re-evaluation on clean, independent datasets after a thorough audit of their training data to remove any Zalo AI contamination. Transparency regarding `hcmute`'s "in-house"dataset is essential.

6.1.2 For Advancing Evaluation Standards

- **Promote Transparency in Model Training:** Encourage researchers and developers to clearly document the data sources and pre-processing steps used for training their models, especially when submitting to leaderboards or challenges.
- **Develop Leakage Detection Tools:** Investigate and develop automated or semi-automated methods for detecting potential data leakage, such as embedding similarity checks between training and test sets, or n-gram overlap analyses. Flagging high similarity (e.g., average cosine similarity $>$ 0.85-0.9 between supposed train/test queries and relevant documents) could be an initial step.
- **Community-Driven Clean Benchmarks:** Foster the creation and maintenance of more standardized, transparent, and verifiably clean evaluation datasets within the IR research community, particularly for specific languages or domains.

6.2 Future Research Directions

Future work should focus on quantifying the precise impact of different types of data leakage on various IR model architectures. Developing robust, automated techniques for sanitizing datasets or identifying leakage in pre-trained models remains a significant challenge. Furthermore, exploring evaluation frameworks that are inherently more resilient to subtle forms of data leakage would be a valuable contribution to the field.

A Retrieval Model Details

This appendix provides links to the Hugging Face model cards for the Dense Retrieval models evaluated in this study. Users are encouraged to consult these model cards for more detailed information regarding model architecture, training data (if disclosed), and intended use cases.

- **BKAI Bi-Encoder (`bkai`)**:
  - Hugging Face Link: `https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder`
  - Brief Note: As discussed, this model's README indicated training on Zalo 2021 Challenge data.
- **Halong Embedding (`halong`)**:
  - Hugging Face Link: `https://huggingface.co/hiieu/halong_embedding`
  - Brief Note: Reported not to use Zalo dataset for training; utilizes Matryoshka embeddings.
- **DEk21 HCMUTE Embedding (`hcmute`)**:
  - Hugging Face Link: `https://huggingface.co/huyydangg/DEk21_hcmute_embedding`
  - Brief Note: Training data source ("in-house dataset") not fully disclosed, leading to suspicions regarding Zalo AI data.
- **AITeamVN Embedding (`aiteamvn`)**:
  - Hugging Face Link: `https://huggingface.co/AITeamVN/Vietnamese_Embedding`
  - Brief Note: General purpose Vietnamese embedding model.

B Dataset Details

This appendix provides further details on the datasets employed in this study, including their statistical overview and an illustrative sample. The selection of these datasets was crucial for investigating the impact of data integrity and labeling methodologies on model performance.

B.1 Dataset Statistics

Table 3 presents a statistical overview of the datasets used, detailing the size of the corpus, the number of queries, and the relevance judgments (qrels). For the "Auto-generated dataset,"statistics for each of its three subsets (test_0, test_1, test_2) are provided.

Table 3: Statistical Overview of Datasets Used in the Report.

Dataset	Subset	Final Corpus Size	Number of Final Queries	Number of Valid Qrels	Avg. Qrels Per Query
Extend Dataset	test	10376	419	499	1.19
Zalo AI Dataset	test	10539	640	640	1.00
Auto-generated dataset	test_0	22603	3897	4574	1.17
Auto-generated dataset	test_1	22556	3897	4542	1.17
Auto-generated dataset	test_2	22548	3898	4560	1.17

B.2 Illustrative Data Sample

To provide a concrete understanding of the data format, particularly for query-document pairs, an example typical of the legal domain explored in parts of our data is shown below. This sample illustrates a query, a relevant passage retrieved from a document, and associated metadata.

Query Example:

```
Khi nào Luật Nhà ở 2023 có hiệu lực?
```

Relevant Document Snippet Example:

> Căn cứ quy định tại Điều 197 Luật Nhà ở 2023 như sau:
>
> **Hiệu lực thi hành**
>
> *1. Luật này có hiệu lực thi hành từ ngày 01 tháng 01 năm 2025.*
>
> *2. Luật Nhà ở số 65/2014/QH13 đã được sửa đổi, bổ sung một số điều theo Luật số 40/2019/QH14, Luật số 61/2020/QH14, Luật số 62/2020/QH14, Luật số 64/2020/QH14 và Luật số 03/2022/QH15 hết hiệu lực kể từ ngày Luật này có hiệu lực thi hành, trừ trường hợp quy định tại điểm b khoản 1, các điểm a, c, đ, e và g khoản 2, khoản 3, các điểm a, b, c, d, đ và e khoản 5 Điều 198 của Luật này.*

3. Nhà ở thuộc sở hữu nhà nước được quy định tại văn bản quy phạm pháp luật về nhà ở ban hành trước ngày Luật này có hiệu lực thi hành là nhà ở thuộc tài sản công.

Luật Nhà ở 2023 có hiệu lực thi hành từ ngày 01/01/2025.

Associated Metadata Example (JSON-like format):

```
[{’article’: ’Điều 197’, ’law’: ’Luật Nhà ở 2023’}]
```

B.3 Extend Dataset

- **Origin and Nature:** A proprietary dataset, manually curated by the research group. Over 500 initial samples were reviewed, and relevant legal articles/clauses were meticulously hand-labeled to ensure high-quality query-document pairs for "clean"evaluation.
- **Key Characteristics:** Designed for high data integrity, avoiding overlap with known public or challenge datasets to genuinely test model generalization. Statistics are in Table 3.
- **Role in Study:** Serves as the "gold standard"benchmark, free from suspected data leakage.
- **Availability:** Internal to the Information Retrieval Research Group; data generation process and dataset not publicly released with this study.

B.4 Zalo AI Dataset

- **Origin and Nature:** An internal dataset associated with Zalo AI, likely derived from or similar to data used in past events such as the Zalo AI Challenge 2021. Specific test queries and an associated index from this source were utilized.
- **Key Characteristics:** Suspected of data leakage, as some models show disproportionately high performance, suggesting prior exposure. The exact data generation or curation process for this dataset is not publicly detailed by its originators. Statistics are in Table 3.
- **Role in Study:** Used to investigate and highlight data leakage phenomena.
- **Availability:** Internal to Zalo; used here under specific research conditions. Not publicly released by the research group with this study.

B.5 Auto-labeled Subsets (test_0, test_1, test_2)

- **Origin and Nature:** Three distinct subsets generated from a larger internal document collection.
- **Labeling Method:** A hybrid approach was used. The Qwen2.5-7B model was employed to initially extract potentially cited legal articles. These extractions were then refined and validated through a rule-based system and predefined constraints to identify relevant query-document pairs.
- **Key Characteristics:** Represents larger-scale, automatically (and thus potentially noisier) labeled data compared to Extend. Statistics for each subset are in Table 3.
- **Role in Study:** Evaluates model robustness to label noise and performance in more practically generated data conditions.
- **Availability:** Internal to the Information Retrieval Research Group; data generation process and dataset not publicly released with this study.