# Ensemble-Based Multi-Specialty Retrieval: Integrating Diverse Similarity Metrics for Enhanced Question Answering

**Chi-Hsiang Chao**
National Central University
Taipei, Taiwan
samuelchao921221@gmail.com

**Hsin-Fu Chang**
National Central University
Taipei, Taiwan
xuanchang@g.ncu.edu.tw

**Po-Yuan Teng**
National Central University
Taipei, Taiwan
davidteng00000@gmail.com

## Abstract

*x3We propose an innovative approach for **multi-specialty retrieval** in question answering systems by integrating diverse similarity measures through ensemble learning. Traditional machine reading comprehension methods struggle to accurately capture specialty-specific terminology and semantic nuances due to their reliance on generic models. Our framework addresses this challenge by leveraging multiple pre-trained embedding models tailored specifically for Chinese, English, and financial texts, combined with various similarity metrics, including cosine similarity, modified Euclidean similarity, and modified Manhattan similarity. The core novelty of our framework lies in effectively combining these domain-specific embeddings and diverse similarity metrics through both unsupervised and supervised ensemble strategies, enabling robust relevance prediction across heterogeneous contexts. Extensive experiments on domain-specific and challenging cross-specialty datasets demonstrate significant improvements in accuracy, F1-score, and precision compared to single-embedding and single metric baselines.*

## 1. Introduction

Recent advances in deep learning and large language models (LLMs) have significantly enhanced question-answering (QA) systems. Traditional Machine Reading Comprehension (MRC) methods often depend on large, parameter-intensive models, yet they frequently struggle to capture nuanced semantic information unique to specialized fields such as finance, healthcare, and multilingual content. Retrieval-Augmented Generation (RAG) has emerged as an effective alternative by supplementing generative models with externally retrieved information, thereby enhancing accuracy and comprehensiveness.

Despite these advancements, designing an efficient retrieval module for multi-specialty applications remains challenging. Each specialty possesses distinct terminologies, stylistic conventions, and data distributions, causing embedding models optimized for one specialty to suboptimized when applied to another. Each specialty-specific embedding model inherently exhibits unique strengths and weaknesses, making it difficult to select a universally optimal solution.

To overcome these limitations, we propose an ensemble-based approach that strategically integrates multiple domain-specific embedding models with diverse similarity metrics—cosine similarity, modified Euclidean similarity, and modified Manhattan similarity. By combining these embeddings and metrics through both unsupervised and supervised ensemble techniques, our approach effectively leverages their complementary strengths, enhancing retrieval robustness even under resource-constrained scenarios.

## 2. Dataset

### 2.1 Training Data

To capture specialty-specific characteristics across Chinese, English, and financial texts, we aggregated several publicly available QA datasets into a unified corpus. This unified approach allows our retrieval classification models to learn from diverse specialty-specific contexts.

| Source Dataset | Specialty | Selected QA Pairs |
|---|---|---|
| DuReader | Chinese | 6666 |
| ChineseSquad | Chinese | 6666 |
| WebQA | Chinese | 6668 |
| Finqa | Financial | 10000 |
| Sujet-Finance-QA-Vision-100k | Financial | 10000 |
| SearchQA | English | 6666 |
| Disfl_qa | English | 6666 |
| Duorc | English | 6668 |

### 2.2 Testing Data

The testing set consists of five datasets, each targeting different specialties and evaluation aspects.

| Source Dataset | Specialty | QA Pairs Quantity |
|---|---|---|
| DRCD | Chinese | 67906 |
| NQ | English | 15660 |
| Financial-QA-10K | Financial | 14000 |
| BiPaR | Chinese&English | 29336 |
| AICUP2024 | Chinese&Finance | 2570 |

## 3. Experiment

### 3.1 Similarity Method

1. **Cosine Similarity**

$$\frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|+10^{-10}} \tag{1}$$

2. **Modified Euclidean Similarity**

$$e^{-0.1\|\mathbf{a}-\mathbf{b}\|} \tag{2}$$

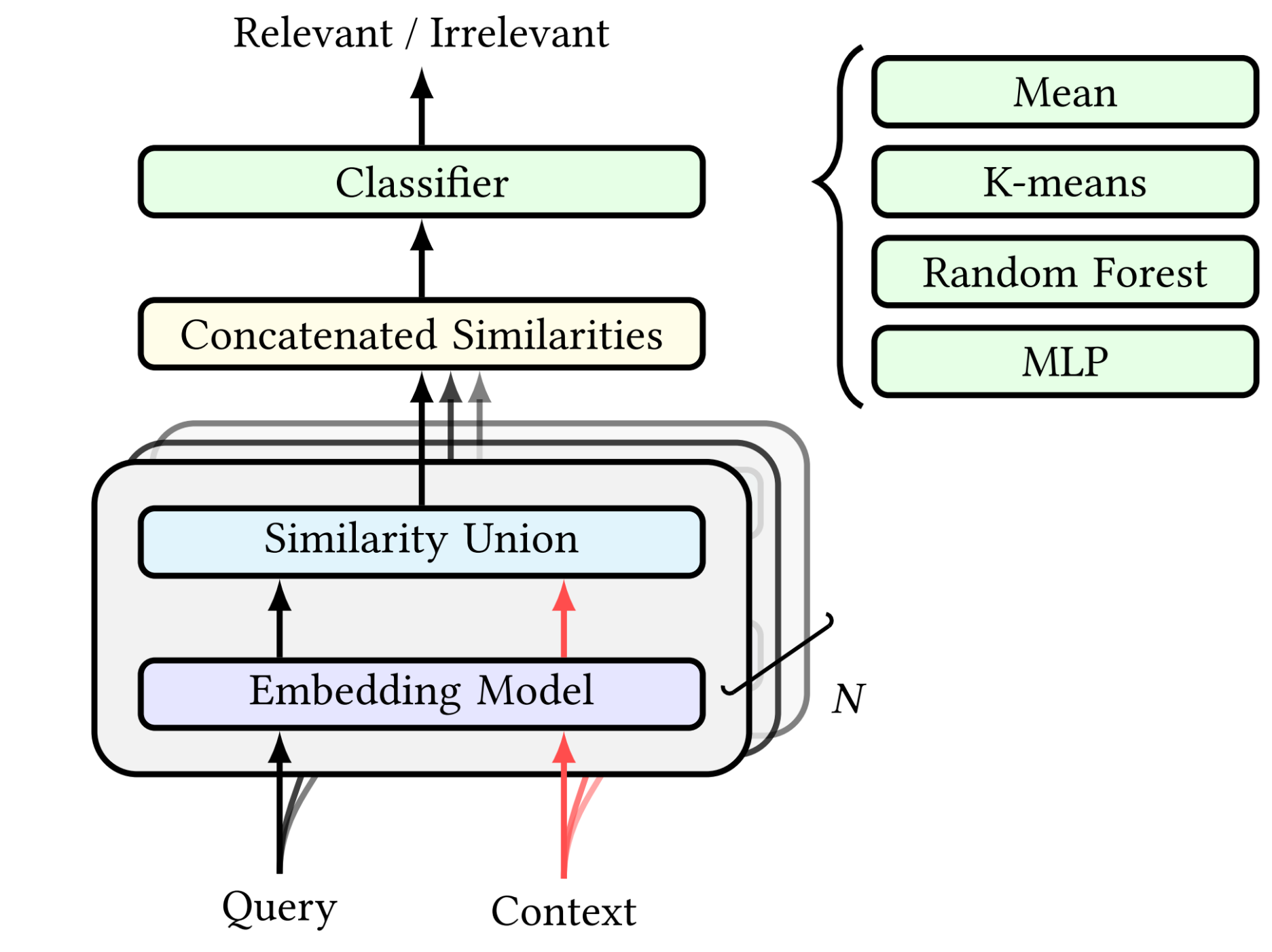
3. **Modified Manhattan Similarity**

$$e^{-0.1\sum|\mathbf{a}-\mathbf{b}|} \tag{3}$$

### 3.2 Embedding Models

1. Chinese Specialty : Yuan-embedding-1.0
2. English Specialty : all-MiniLM-L6-v2
3. Finance Specialty : finance-embeddings-investopedia

### 3.3 Model Architecture



## 4. Result

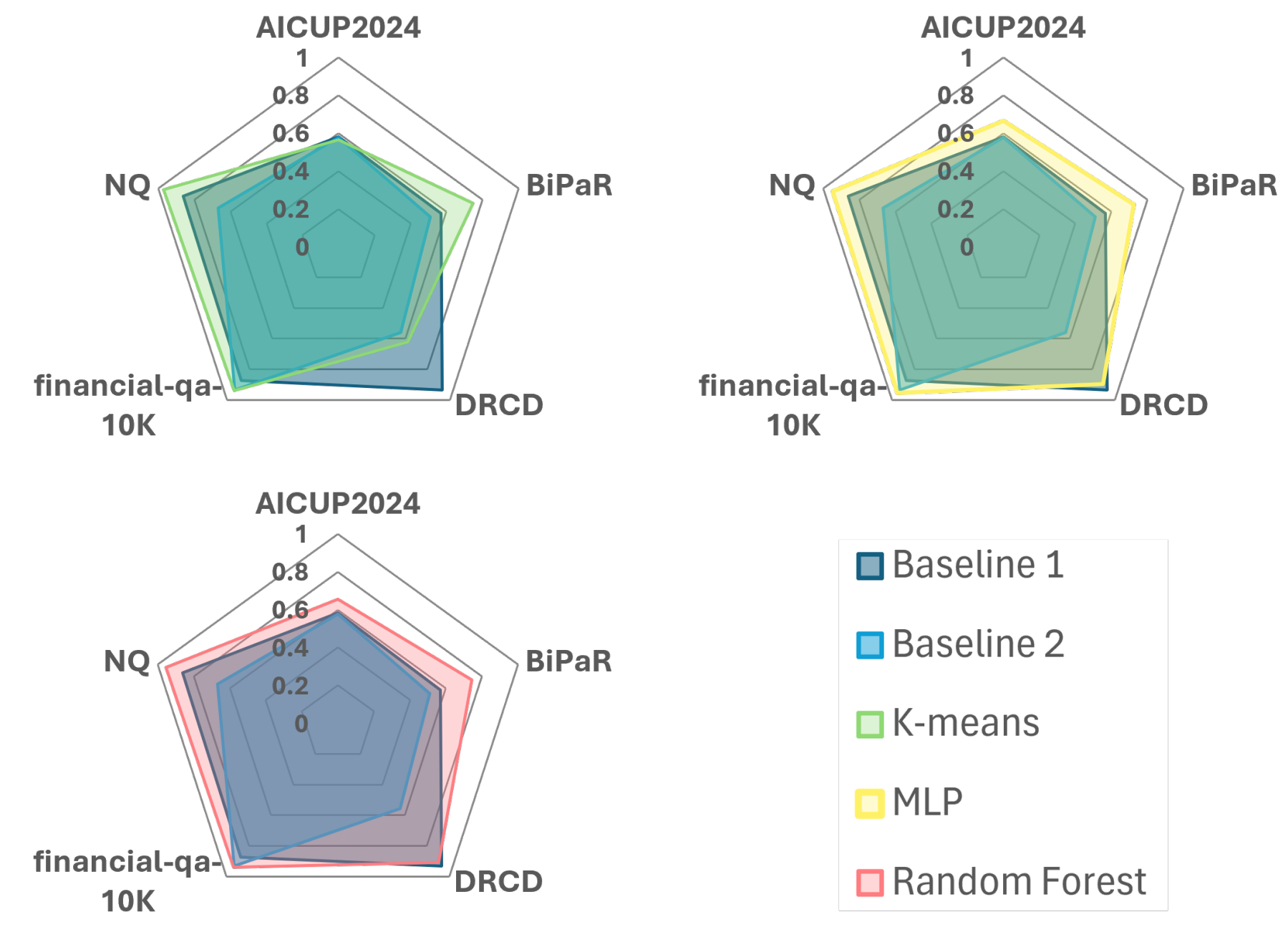


**Figure 1 :** Comparing Ensemble Models to Baselines Using Cosine Similarity

Figure 1 demonstrates that, regardless of the ensemble architecture—K-Means, Random Forest, or MLP—the accuracy measured by cosine similarity remains consistently high across all models.

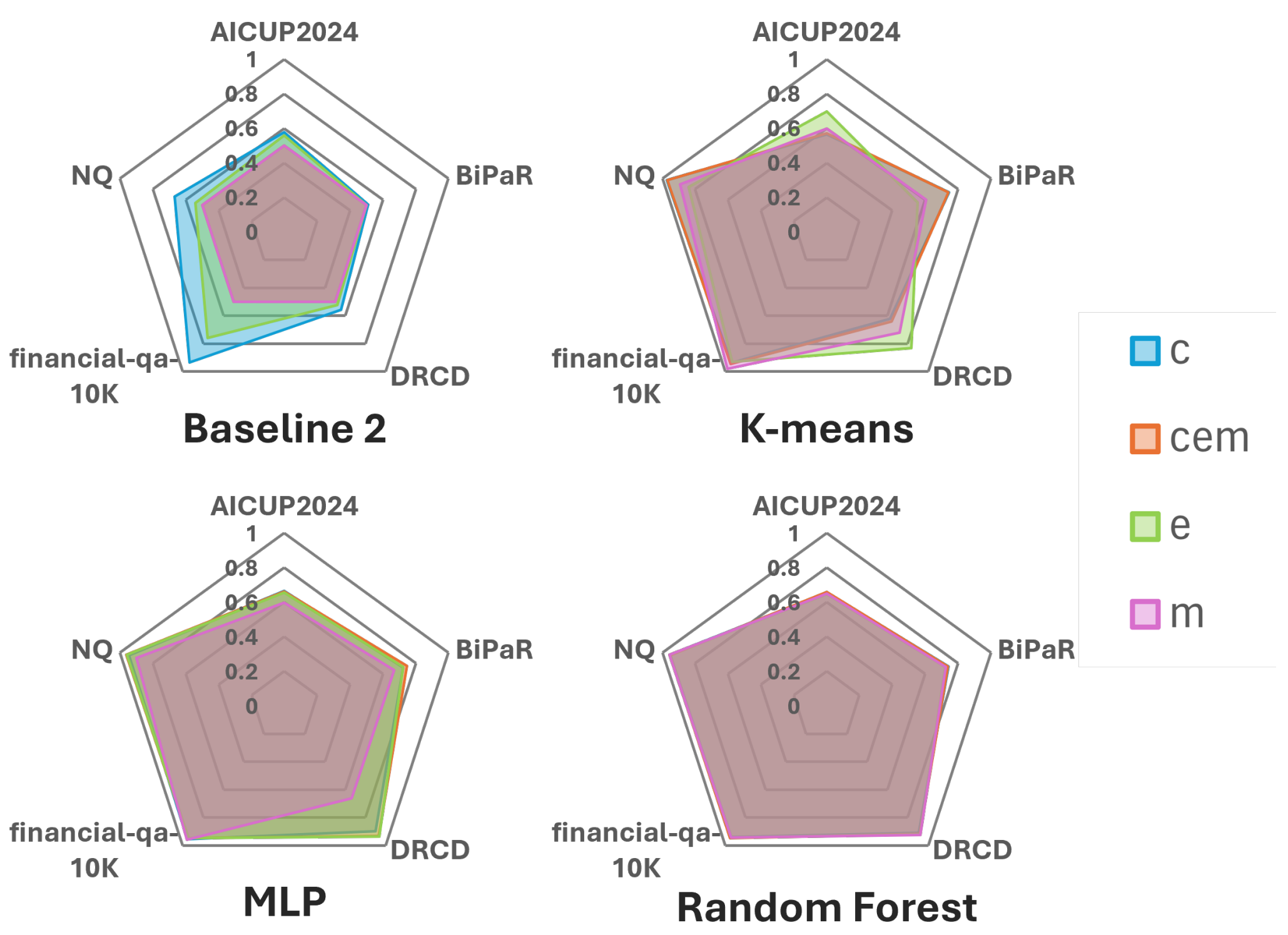


**Figure 2 :** Comparing Similarities Across Datasets in Various Embedding Models

Figure 2 shows that cosine similarity is the most broadly effective metric across all the ensemble models we evaluated. In contrast, Modified Euclidean similarity shines specifically within the K-means clustering framework, while Modified Manhattan similarity delivers its strongest performance when paired with the Random Forest model.

| Model | Dataset | Chinese(c) | English(c) | Finance(c) |
|---|---|---|---|---|
| Baseline 1 | DRCD | 0.9333 | 0.5094 | 0.5630 |
| | financial-qa-10K | 0.6159 | 0.8493 | 0.8739 |
| | NQ | 0.8626 | 0.6126 | 0.5430 |
| | BiPaR | 0.5683 | 0.5062 | 0.5026 |
| | AICUP2024 | 0.5821 | 0.5128 | 0.5198 |

| Model | Dataset | c | e | m | c,e,m |
|---|---|---|---|---|---|
| K-means | DRCD | 0.6227 | **0.8324** | 0.7213 | 0.6419 |
| | financial-qa-10K | 0.9393 | 0.9303 | <u>**0.9784**</u> | 0.9444 |
| | NQ | <u>**0.9732**</u> | 0.8397 | 0.8908 | 0.9711 |
| | BiPaR | <u>**0.7457**</u> | 0.5545 | 0.6056 | 0.7411 |
| | AICUP2024 | 0.5642 | <u>**0.6969**</u> | 0.6000 | 0.5708 |
| RandomForest | DRCD | 0.9074 | 0.9216 | **0.9228** | 0.9160 |
| | financial-qa-10K | 0.9385 | 0.9419 | 0.9416 | **0.9483** |
| | NQ | 0.9550 | **0.9565** | 0.9554 | 0.9520 |
| | BiPaR | **0.7429** | 0.7240 | 0.7258 | 0.7389 |
| | AICUP2024 | 0.6556 | 0.6467 | 0.6490 | **0.6595** |
| MLP | DRCD | 0.8962 | <u>**0.9353**</u> | 0.6606 | 0.9301 |
| | financial-qa-10K | 0.9537 | 0.9456 | **0.9590** | 0.9466 |
| | NQ | 0.9463 | 0.9582 | 0.8997 | **0.9611** |
| | BiPaR | 0.7229 | 0.7238 | 0.6688 | **0.7470** |
| | AICUP2024 | **0.6650** | 0.6572 | 0.5977 | 0.6626 |

**Figure 3 :** Comparing Similarities Across Datasets in Various Embedding Models

In Figure 3, values displayed in boldface identify the highest-performing metric within each specific dataset-model pairing, whereas underlined values highlight the best overall performance achieved on a given dataset across all models.

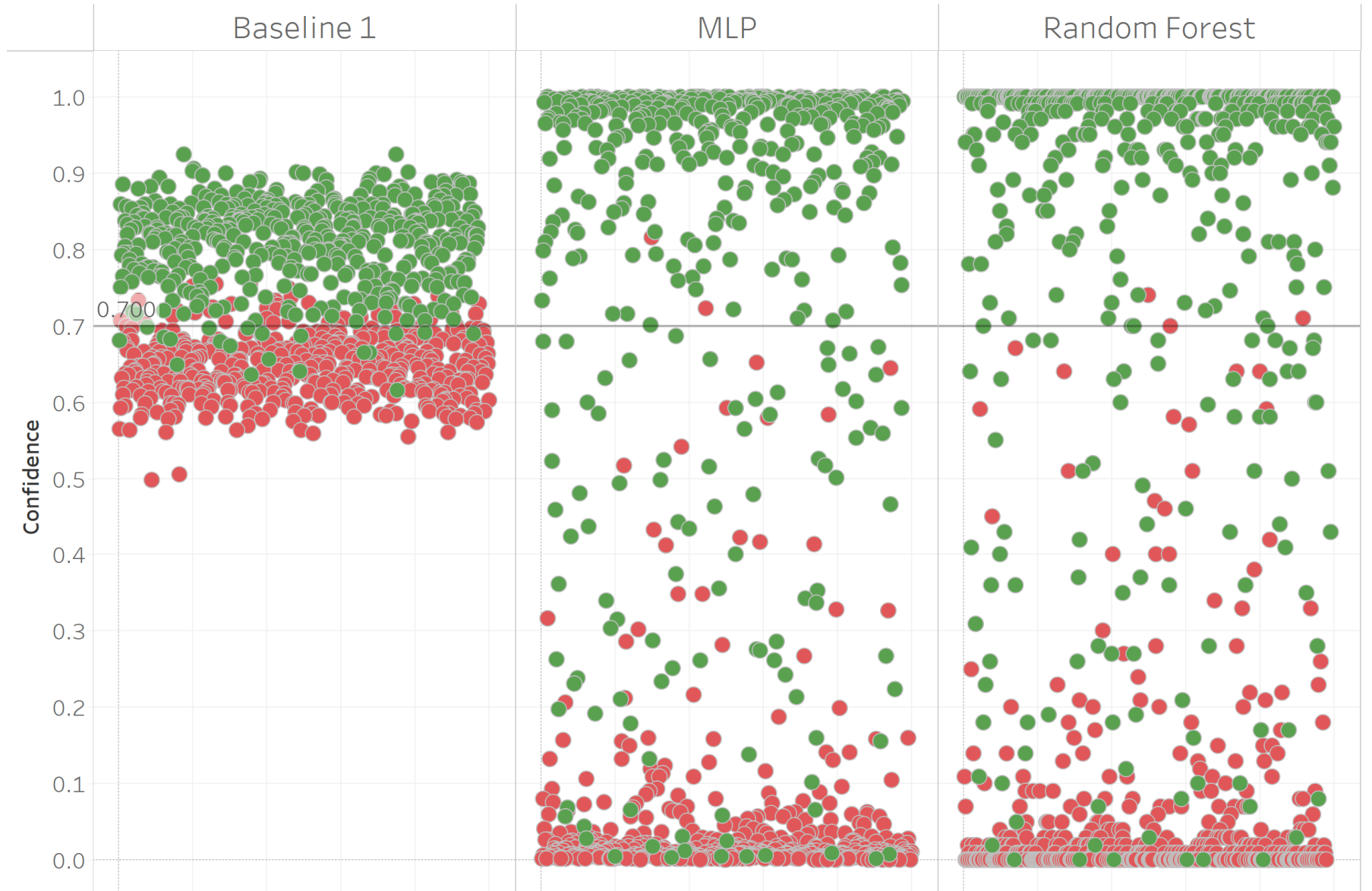


**Figure 4 :** Comparing Similarities Across Datasets in Various Embedding Models

Figure 4 illustrates that green points represent true label and red points denote false; unlike the baseline, in which both classes cluster around the 0.7 similarity cutoff, our method yields a far cleaner separation, concentrating true instances well above the threshold while relegating false ones to lower-confidence regions.

## 5. Acknowledgments

We would like to express our gratitude to E.SUN Bank for providing the AICUP2024 dataset. Their support and the accessibility of the data were instrumental in making this research possible. We greatly appreciate their contribution and assistance throughout this project.

## 6. Reference

1. Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, June 2024. arXiv:2402.03216 [cs].
2. Abdelmaseeh Felfel and Paul Missault. Deep Domain Specialisation for single-model multi-domain learning to rank, 2024.
3. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. 2023. Publisher: arXiv Version Number: 5.
4. Federico Tessari, Kunpeng Yao, and Neville Hogan. Surpassing Cosine Similarity for Multidimensional Comparisons: Dimension Insensitive Euclidean Metric (DIEM), December 2024. arXiv:2407.08623 [cs].

**NTCIR18 AICUP Poster**, National Institute of Informatics, 2025