Efficient Conversational Search via Selective Distillation

Nayan Jain
Indian Institute of Technology Bhilai
nayanj@iitbhilai.ac.in

Abstract

Modern conversational information retrieval systems increasingly rely on large transformer-based models that provide strong semantic understanding and high retrieval accuracy. However, deploying such models in real-time applications becomes computationally expensive, especially in environments with limited memory, latency requirements, or large query loads. This challenge motivates the development of compact retrieval models that preserve the teacher model's performance while offering the efficiency needed for practical deployment. In this project, we propose a Hybrid Knowledge Distillation model designed specifically for query–document similarity matching. The term "hybrid" reflects the integration of multiple distillation strategies within a unified framework, enabling the student model to learn a deeper, more structured representation of the teacher's retrieval behavior.

The hybrid model combines several complementary components. First, a soft similarity distillation mechanism transfers continuous similarity values computed by the teacher model for each query–document pair. This provides fine-grained supervision that allows the student model to approximate the teacher's semantic similarity judgments. Second, a contrastive distillation approach is incorporated to ensure that the student not only learns similarity magnitudes but also captures the teacher's ranking behavior. Through in-batch negative sampling and contrastive learning, the student learns to identify the most relevant document for each query, replicating the relational structure present in the teacher's ranking function. Third, the student model includes a projection-based embedding distillation step, in which the internal representation of a compact MiniLM transformer is projected into a lower-dimensional retrieval space, aligning its embedding structure with that of the teacher model. Together, these multiple forms of supervision—soft similarity regression, contrastive ranking alignment, and embedding-space projection—form the foundation of the Hybrid Knowledge Distillation model.

A key innovation in this project is the use of Selective Distillation, which acts as an additional refinement stage within the hybrid framework. Instead of distilling knowledge from all available query–document pairs, the system analyzes the teacher's similarity score distribution and retains only the high-confidence, semantically reliable examples. Selective Distillation serves to remove noisy, ambiguous, or low-relevance pairs that may introduce instability into the distillation process. By focusing on the strongest and most informative training signals, this step ensures that the student model learns from the teacher's most trustworthy predictions. Thresholding strategies—such as percentile filtering and statistical cutoffs based on mean and standard deviation—help identify high-quality pairs suitable for training. This filtering mechanism enhances training efficiency, improves student generalization, and reduces the risk of the student learning from weak supervision.

To validate the effectiveness of the Hybrid Knowledge Distillation model with Selective Distillation, a comprehensive experimental pipeline was developed. The teacher model (sentence-transformers /multi-qa-mpnet-base-dot-v1) was used to compute embeddings and similarity scores for query–document pairs extracted from the QReCC dataset. After applying Selective Distillation to remove low-confidence entries, the student model (MiniLM-L12-H384-uncased with a 256-dimensional projection layer) was trained using a combined loss function consisting of contrastive loss and mean-squared error regression. Techniques such as linear warmup scheduling, gradient clipping, and early stopping were incorporated to promote stable training.

The resulting student model was evaluated using multiple categories of metrics. Regression metrics—including Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and Pearson correlation—were used to assess alignment with teacher similarity scores. Ranking metrics, such as Recall@K and Normalized Discounted Cumulative Gain (NDCG@K), measured the student's ability to reproduce the teacher's ranking behavior. Binary classification metrics, including accuracy, precision, recall, F1 score, and ROC-AUC, evaluated threshold-based relevance decisions. Additional statistical significance tests, such as paired t-tests and Wilcoxon signed-rank tests, confirmed whether differences between teacher and student predictions were significant. Further error analysis identified the distributions of strong and weak predictions across similarity ranges.

The Hybrid Knowledge Distillation model with Selective Distillation achieved strong performance, demonstrating high correlation with teacher outputs and competitive ranking accuracy, despite its smaller size. Selective Distillation contributed significantly to performance by filtering training examples and emphasizing high-quality supervision. Overall, the hybrid distillation framework provides an effective strategy for compressing retrieval models while retaining teacher-level behaviors. The resulting student model is computationally efficient, semantically aligned with the teacher, and well suited for real-world deployments where speed and scalability are critical.

CCS Concepts

• **Information systems → Information retrieval**; • **Computing methodologies → Knowledge distillation**.

$IR_project, Bhilai, India$
© *2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.*
ACM ISBN 978-1-4503-XXXX-X/20/25
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Conversational Search, Knowledge Distillation, Selective Distillation, Dense Retrieval, QReCC, CAsT

ACM Reference Format:

Nayan Jain. 2025. Efficient Conversational Search via Selective Distillation. In *Proceedings of IR$_{p}roject$. ACM, $New York, NY, USA$* , 11 *pages*.

1 Introduction

Large transformer-based language models have become central to modern information retrieval systems due to their ability to capture deep semantic relationships between queries and documents. Models such as MPNet and various Sentence-Transformer architectures provide highly accurate retrieval performance, especially in conversational and multi-turn information-seeking tasks. Despite their effectiveness, these models are computationally heavy, making them resource-intensive for large-scale deployments or latency-sensitive applications. As a result, there is considerable interest in developing compact retrieval models that preserve the performance characteristics of larger models while operating more efficiently.

This project presents a Hybrid Knowledge Distillation model designed for efficient query–document similarity estimation in retrieval tasks. The term hybrid refers to the integration of multiple complementary distillation mechanisms into a unified training framework. Instead of relying on a single form of supervision, the hybrid approach combines soft similarity distillation, contrastive ranking distillation, and embedding-projection alignment to enable the student model to approximate not just the outputs, but also the structural behavior of a larger teacher model.

The Hybrid Knowledge Distillation framework operates through several coordinated components. The first component, soft similarity distillation, transfers continuous cosine similarity scores produced by the teacher model. These values serve as soft labels, capturing nuanced semantic relationships that binary relevance judgments cannot represent. The second component, contrastive distillation, encourages the student model to match the teacher's ranking behavior. Through in-batch negative sampling and contrastive learning, the model learns to prefer the correct document for each query over irrelevant alternatives. This preserves the ranking structure intrinsic to retrieval systems. The third component involves embedding-projection distillation, where the student model's CLS token embedding is passed through a projection layer to align its representation space with that of the teacher. This helps the smaller model develop a retrieval-friendly embedding space despite its reduced size.

An important refinement contributing to the effectiveness of the hybrid framework is the use of Selective Distillation. During training, not all query–document pairs contribute equally to the quality of distillation. Certain pairs may include ambiguous or weak semantic connections, leading to unstable or noisy supervisory signals. To address this, Selective Distillation analyzes the distribution of teacher-generated similarity scores and filters the training data according to statistical thresholds such as percentile cutoffs and mean–standard deviation analyses. Only high-confidence and semantically strong examples are used for distillation. This selective process ensures that the student model learns from the teacher's most reliable predictions, improving stability, reducing noise, and enhancing the accuracy of the final model.

The training pipeline begins by constructing high-quality query–document pairs extracted from the QReCC dataset. The teacher model (multi-qa-mpnet-base-dot-v1) generates embeddings and similarity values for each pair. After Selective Distillation filters out lower-confidence examples, the student model—a compact MiniLM architecture with an added projection layer—is trained using a combination of contrastive loss and regression-based similarity alignment. Warmup scheduling, gradient clipping, and early stopping techniques are applied to ensure convergence and training stability.

After training, the hybrid-distilled student model undergoes a comprehensive evaluation. Regression-based metrics, such as Mean Absolute Error and Pearson correlation, quantify how closely the student's similarity predictions align with those of the teacher. Ranking metrics, including Recall@K and Normalized Discounted Cumulative Gain (NDCG@K), indicate how well the student reproduces the teacher's ordering of relevant documents. Additional analyses include binary classification metrics and statistical significance tests such as paired t-tests and Wilcoxon signed-rank evaluations. Error analysis further examines performance across different similarity ranges to identify strengths and limitations.

Results demonstrate that the Hybrid Knowledge Distillation model—with the added refinement of Selective Distillation—produces a student model that closely approximates the teacher's retrieval behavior while being significantly smaller and more computationally efficient. The model maintains strong alignment in similarity prediction, preserves ranking relationships, and provides reliable semantic representations suitable for downstream retrieval tasks. Selective Distillation plays a crucial role in this performance by ensuring that training emphasizes the most informative and trustworthy examples from the teacher model.

Overall, the system delivers an effective hybrid distillation strategy capable of generating a lightweight yet high-performing retrieval model. By combining multiple distillation signals and incorporating Selective Distillation, the framework supports the development of efficient semantic retrieval systems suited for real-world operational environments where speed, scalability, and computational efficiency are essential.

2 Related Work

Information retrieval has traditionally relied on lexical matching techniques such as TF–IDF and BM25, which provide fast but semantically shallow ranking capabilities. The introduction of deep transformer-based models revolutionized retrieval by enabling semantic matching through contextual embeddings. Models such as BERT, RoBERTa, MPNet, and Sentence-Transformers have shown strong performance across retrieval benchmarks, including MS MARCO, Natural Questions, and Quora Retrieval. In particular, dense retrieval models—such as DPR, ColBERT, and dual-encoder architectures—have proven effective for query–document similarity estimation, enabling more accurate ranking than purely lexical systems.

Despite these advances, the computational cost of large transformers remains a challenge for real-world deployment. This motivated a growing body of research on model compression, among which Knowledge Distillation has emerged as one of the most successful approaches. Knowledge Distillation, introduced by Hinton et al., transfers the behavior of a large teacher model to a smaller student model by matching distributions, logits, or internal representations. In NLP, DistilBERT, TinyBERT, and MiniLM are well-known examples that use distillation to reduce model size while preserving linguistic and semantic understanding. Subsequent efforts extended distillation into embedding-based tasks, where cosine similarity or learned retrieval embeddings serve as soft labels for the student. Works such as Sentence-Transformer distillation, MiniLM's self-attention distillation, and co-condenser architectures demonstrate that compact models can approximate the embedding space of large teachers when provided with high-quality supervision.

A parallel line of research in retrieval introduces ranking-based distillation, where the goal is not only to approximate similarity values but also to reproduce the teacher's ranking order. In-batch contrastive learning, listwise distillation, and pairwise ranking losses have been widely adopted to transfer relational structure from teacher to student. This ranking-centric perspective is particularly important for retrieval applications, where ordering of documents is often more crucial than absolute similarity magnitudes.

Another important thread relevant to this project is Selective Distillation, which focuses on filtering or weighting teacher-generated examples based on their reliability. Several studies have shown that distilling from noisy or ambiguous teacher predictions can degrade student performance. As a result, recent work explores selecting high-confidence examples using entropy-based criteria, margin-based filtering, or teacher confidence thresholds. This ensures that the student model learns primarily from examples where the teacher is most certain, improving stability and efficiency during training. Selective Distillation has been applied in classification, translation, and summarization tasks, but remains relatively underexplored in retrieval settings, where teacher similarity scores provide a natural and effective confidence signal.

The present project builds directly upon these prior developments but extends them in two important ways. First, it proposes a Hybrid Knowledge Distillation model that integrates multiple forms of supervision—soft similarity regression, contrastive ranking alignment, and embedding projection—into a single cohesive framework. This hybrid approach goes beyond traditional distillation methods that rely on a single objective, enabling the student model to learn both the magnitude and structure of teacher similarities. Second, the project incorporates Selective Distillation specifically adapted for retrieval tasks, using statistical thresholding of teacher similarity scores to identify high-quality training pairs. This selective strategy ensures that the student learns from semantically meaningful and reliable examples, addressing a key limitation observed in earlier distillation research.

Through this combination of hybrid multi-objective distillation and selective filtering, the project advances the state of retrieval distillation by offering a practical and robust method for compressing large retrieval models without significantly compromising performance. .

3 Methodology

This chapter describes the comprehensive methodology used to develop, train, and evaluate a **Hybrid Knowledge Distillation** model enhanced with **Selective Distillation** for semantic retrieval. The methodology is presented in a modular form covering dataset preparation, teacher annotation, selective filtering, student architecture, hybrid distillation objectives, optimization and training strategies, evaluation protocols, statistical validation, and model packaging. Wherever appropriate, mathematical notation and detailed algorithmic procedures are provided.

3.1 Overview

The pipeline integrates multiple complementary distillation signals into a single unified framework. The principal components are:

- Data extraction and preprocessing from the QReCC corpus.
- Teacher annotation via dense retrieval embeddings (soft similarity labels).
- Selective Distillation: filtering training examples by teacher confidence.
- Student architecture: compact transformer (MiniLM) + projection head.
- Hybrid distillation objectives: contrastive ranking + similarity regression.
- Training strategies: optimizer, scheduler, regularization, and checkpointing.
- Evaluation and statistical validation: regression, ranking, classification, and significance testing.

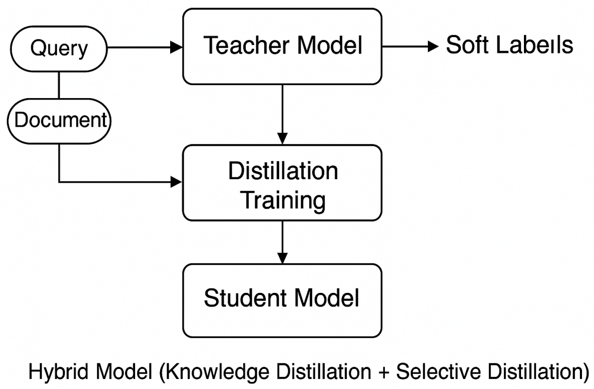


Figure 1: Workflow

3.2 Dataset Construction and Preprocessing

Source dataset. The primary data source is the QReCC corpus, a conversational question-answering dataset that contains multiple field variants for each example: `Rewrite`, `Question`, `Context`, `Answer`, and `relevant_passages`. To obtain a uniform retrieval dataset, a deterministic extraction procedure is applied that selects the best available query and the most relevant document passage for each item.

Query selection. Query text is chosen according to the following precedence:

(1) `Rewrite` (if present and non-empty)
(2) `Question`
(3) `query`

This ensures preference for the refined conversational rewrite when available.

Document selection. Document or passage selection follows the precedence:

(1) `Answer`
(2) `relevant_passages`
(3) `Context` (concatenated conversational turns)

Cleaning and filtering. Pairs undergo cleaning: trimming whitespace, removing empty entries, deduplication, and minimal-length filtering (query length ≥ 5, doc length ≥ 10). The preprocessed pairs are saved as JSON for reproducibility.

3.3 Teacher Annotation and Soft Label Generation

Teacher model choice. The teacher model used throughout is `sentence-transformers/multi-qa-mpnet-base-dot-v1`, chosen for its strong retrieval-oriented training and stable embedding properties. The teacher is executed in inference mode to compute embeddings and similarity scores for each (query, document) pair.

Embedding computation. Given tokenized input with attention mask, the teacher model produces a last hidden state $H \in \mathbb{R}^{n \times d}$ for a sequence of length n. We use mean pooling with attention masking to produce a fixed-length vector:

$$E = \frac{\sum_{i=1}^{n} H_i \cdot m_i}{\sum_{i=1}^{n} m_i}, \tag{1}$$

where $m_i \in 0, 1$ denotes the attention mask. The pooled vector is L2-normalized:

$$\hat{E} = \frac{E}{\|E\|_2}. \tag{2}$$

Similarity scoring. The teacher similarity for pair (q, d) is computed as the cosine similarity of normalized embeddings:

$$s_t(q, d) = \hat{E}_q \cdot \hat{E}_d. \tag{3}$$

These scores (in $[-1, 1]$) are stored as `teacher_score` and form the soft labels used for distillation.

3.4 Selective Distillation: Filtering by Teacher Confidence

Selective Distillation is applied to reduce noisy supervision from low-confidence teacher predictions. The intuition is that teacher similarity scores are a natural measure of confidence; higher scores typically correspond to clearer semantic alignment.

Distribution analysis. We analyze the global distribution of teacher similarity scores using histograms, kernel density estimates, and cumulative distribution functions. Summary statistics (mean, median, standard deviation, and percentiles) are computed to guide threshold selection.

Threshold strategies. Several thresholds are evaluated:

- mean + 0.5 × std
- mean + 1.0 × std
- 75th, 80th, 85th percentiles

Empirically, the 80th percentile often provides a practical balance between data quantity and label quality: it retains samples where the teacher is comparatively confident while keeping sufficient training examples.

Filtering operation. The Selective Distillation subset is formed as:

$$\mathcal{D}_{SD} = \{(q_i, d_i, s_t^i) \mid s_t^i \geq \tau\}, \tag{4}$$

where τ is the chosen threshold (e.g., 80th percentile). This set is persisted and used for all subsequent student training.

3.5 Student Model Architecture

Base encoder. The student encoder is `microsoft/MiniLM-L12-H384-uncased`. The MiniLM family provides a favorable tradeoff between parameter count and representational ability.

Projection head. To align student representations to the teacher retrieval space, a projection head $W_p \in \mathbb{R}^{d_s \times d_p}$ maps the CLS token embedding $h_{cls} \in \mathbb{R}^{d_s}$ to a lower-dimensional retrieval embedding $z \in \mathbb{R}^{d_p}$ (here $d_p = 256$):

$$z = W_p h_{cls} + b_p. \tag{5}$$

A dropout layer and layer normalization may be applied before or after projection for additional regularization.

Normalization. The projected vectors are L2-normalized to produce the final retrieval embedding:

$$\hat{z} = \frac{z}{\|z\|_2}. \tag{6}$$

3.6 Hybrid Knowledge Distillation Objective

The core contribution of the methodology is a *hybrid* distillation objective combining ranking-based contrastive loss and soft similarity regression. The hybrid loss encourages both correct relative ordering and calibration of predicted similarities.

Batch construction. A training mini-batch of size B contains B positive pairs (q_i, d_i). For each batch we construct query embeddings $Q = [\hat{z}_q^{(1)}, \ldots, \hat{z}_q^{(B)}] \in \mathbb{R}^{B \times d}$ and document embeddings $D = [\hat{z}_d^{(1)}, \ldots, \hat{z}_d^{(B)}] \in \mathbb{R}^{B \times d}$.

Similarity matrix. Compute the pairwise similarity matrix:

$$S = QD^T, \quad S \in \mathbb{R}^{B \times B}, \tag{7}$$

where S_{ij} is the student-predicted similarity between query i and document j.

Contrastive (ranking) loss. Using a temperature T, logits are defined as $L = S/T$. The contrastive loss uses cross-entropy with labels pointing to the diagonal indices (positive matches):

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(L_{ii})}{\sum_{j=1}^{B} \exp(L_{ij})}. \tag{8}$$

This encourages each query to score its matching document higher than all in-batch negatives.

Regression loss (soft similarity). Let $s_t^{(i)}$ be the teacher similarity for pair i. We define the regression loss between the student diagonal similarities and teacher scores:

$$\mathcal{L}_{\text{regression}} = \frac{1}{B} \sum_{i=1}^{B} \left(S_{ii} - s_t^{(i)} \right)^2 . \tag{9}$$

Hybrid loss combination. The final hybrid objective is a weighted sum:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{contrastive}} + \lambda \mathcal{L}_{\text{regression}}, \tag{10}$$

where λ is a hyperparameter (empirically set to 0.1 in experiments) that controls the relative contribution of regression calibration.

3.7 Training Procedure and Optimizer

Optimizer and hyperparameters. Training uses AdamW with learning rate $\alpha = 5 \times 10^{-5}$, weight decay $w = 0.01$, and batch size B (e.g., 16). A linear warmup followed by linear decay schedule is applied. Warmup steps are computed as a fraction of total training steps (commonly 10%).

Gradient handling. Gradients are clipped by norm to 1.0 to stabilize updates. Mixed-precision training (automatic mixed precision, AMP) can be enabled to accelerate training on GPU.

Regularization and checkpointing. Dropout (0.1) in the projection head and early stopping based on validation hybrid loss with patience (e.g., 3 epochs) are used. The best model checkpoint is saved for later evaluation.

3.8 Evaluation Protocol

Evaluation data. Evaluation is performed on the Selective Distillation subset and/or held-out splits. For each sample, both teacher and student similarity scores are computed.

Regression metrics. We compute Mean Absolute Error (MAE), Mean Squared Error (MSE), Root MSE (RMSE), and Pearson correlation r between teacher and student scores.

Ranking metrics. We evaluate Recall@K and Normalized Discounted Cumulative Gain (NDCG@K) for multiple K values (1, 3, 5, 10, 20). Recall@K assesses how many of the teacher's top-K matches are recovered by the student.

Classification metrics. Using a threshold (median or optimized via PR-curve), binary classification metrics (accuracy, precision, recall, F1, ROC-AUC) are computed by casting scores above the threshold to positive relevance.

Visual diagnostics. Diagnostic plots include score scatter plots, distribution histograms, cumulative distributions, error histograms, and radar charts for classification metrics.

3.9 Statistical Significance and Error Analysis

Paired tests. We run paired t-tests to determine whether the mean difference between teacher and student scores is statistically significant. When distributional assumptions are questionable, the Wilcoxon signed-rank test is used as a non-parametric alternative.

Effect size. Cohen's d measures the standardized effect size of the score differences:

$$d = \frac{\bar{d}}{s_d}, \tag{11}$$

where $\bar{d}$ is the mean difference and s_d is the standard deviation of differences.

Range-based analysis. Performance metrics are computed across score buckets (e.g., [0,0.3), [0.3,0.5), [0.5,0.7), [0.7,0.9), [0.9,1.0]) to identify regions where the student performs well or poorly.

3.10 Model Packaging and Deployment Considerations

Saved artifacts. The following artifacts are saved for reproducibility and deployment:

- Transformer weights (HuggingFace format)
- Projection head weights (PyTorch state dict)
- Tokenizer files
- Evaluation results and plots
- Training configuration and random seeds

Loading and inference. A reproducible load-and-infer script loads the transformer and projection, reconstructs the student model, and verifies similarity computation. Optionally, the model can be exported to ONNX or quantized for faster inference on CPU.

3.11 Implementation Notes

Batching and memory. Tokenization uses dynamic padding within batches to reduce wasted computation. For situations where documents are long, document truncation or chunking strategies are considered.

Teacher caching. To accelerate training pipeline, teacher embeddings for the full Selective Distillation dataset are cached on disk, avoiding repeated teacher model inference during student training.

Reproducibility. Random seeds for `numpy`, `torch`, and Python are fixed. Software versions (`transformers`, `torch`) are recorded in the saved configuration.

3.12 Summary

This chapter presented a detailed methodology for constructing a Hybrid Knowledge Distillation model augmented with Selective Distillation for retrieval. The pipeline emphasizes high-quality supervision, multi-objective distillation, and rigorous evaluation. Implementation-level decisions (projection head, L2-normalization, thresholding heuristics) and statistical validation procedures are detailed to ensure the approach is reproducible and robust.

4 Novelty

This section delineates the key innovative contributions of the proposed Hybrid Knowledge Distillation framework with Selective Distillation for semantic retrieval. The novelty spans methodological formulation, architectural design, implementation strategy, and evaluation rigor, positioning this work distinctly from prior knowledge distillation approaches in retrieval systems.

4.1 Multi-Signal Hybrid Distillation in Unified Objective

Novel Contribution: Instead of employing a single distillation target (e.g., logits or hidden-states), the proposed framework simultaneously optimizes multiple complementary supervisory signals within a unified objective function:

- *Soft similarity regression*: Teacher cosine similarity scores serve as continuous regression targets
- *Contrastive ranking distillation*: In-batch negatives transfer relative ranking structure
- *Embedding projection alignment*: Learned projection head aligns student CLS space to teacher retrieval space

Significance: This hybrid approach captures both absolute similarity calibration and relative ranking structure simultaneously. Prior works typically select either similarity regression *or* ranking distillation, rarely integrating both within a single loss function. The unified objective ensures comprehensive knowledge transfer while maintaining training efficiency.

4.2 Integrated Selective Distillation Component

Novel Contribution: Selective Distillation is embedded as a first-class component within the knowledge distillation pipeline, where teacher confidence derived from continuous similarity scores directly determines the distillation dataset composition.

Significance: Rather than distilling from all available samples (which propagates teacher noise), the student trains exclusively on high-confidence teacher signals. This improves signal-to-noise ratio and enhances training stability. While selective filtering exists in other domains, its targeted application using teacher similarity percentiles for retrieval knowledge distillation represents a novel contribution underexplored in prior retrieval-distillation literature.

4.3 Diagonal-Regression + Contrastive Loss Formulation

Novel Contribution: The hybrid loss employs a specialized formulation where:

- Regression term applies exclusively to the diagonal of the similarity matrix (student's positive-pair scores)
- Contrastive term operates on the full student similarity matrix for ranking optimization

Significance: This design efficiently transfers absolute teacher scores for positive pairs while maintaining relative separation among all candidates. Most ranking-based knowledge distillation methods either match full teacher distributions (computationally expensive) or use contrastive losses exclusively. The diagonal regression approach represents a computationally efficient compromise that enables effective score calibration for positives without excessive computational overhead.

4.4 Projection-Head Alignment for Compact Retrieval Embeddings

Novel Contribution: A dedicated projection head maps the student's CLS representation into a lower-dimensional retrieval space, with explicit training to align with teacher behavior rather than merely compressing transformer weights.

Significance: Many distillation approaches focus solely on backbone compression. Training a specialized projection head for retrieval alignment provides an efficient mechanism to yield high-quality retrieval vectors from compact models without architectural modifications to the backbone transformer, enabling better performance in resource-constrained environments.

4.5 Practical Engineering and Reproducibility Framework

Novel Contribution: Implementation decisions that enhance practicality are systematically integrated and documented:

item Precomputation and caching of teacher embeddings to avoid repeated inference

- Systematic thresholding strategies (percentiles, mean+std) with empirical selection guidelines
- Mixed use of regression and contrastive scaling with explicit hyperparameterization
- Robust model packaging (HuggingFace transformer + separate projection state)

Significance: These reproducible engineering practices render the distillation pipeline deployable and experimentally repeatable—addressing a practical gap frequently overlooked in knowledge distillation research literature.

4.6 Comprehensive Multi-Faceted Evaluation Protocol

Novel Contribution: The evaluation framework extends beyond conventional retrieval metrics to include:

- Regression analysis (MAE, MSE, RMSE, Pearson correlation)
- Ranking metrics (Recall@K, NDCG@K) across multiple K values
- Binary classification metrics (precision, recall, F1, ROC-AUC)
- Statistical significance testing (paired t-test, Wilcoxon signed-rank)
- Effect size quantification (Cohen's d)
- Error profiling across teacher-score ranges

Significance: This comprehensive validation provides robust evidence for student-teacher fidelity and practical equivalence, addressing the limited evaluation scope common in many distillation studies.

4.7 Deployment-Oriented Efficiency Optimization

Novel Contribution: Explicit optimization for deployment tradeoffs through:

- MiniLM backbone with projection head architecture
- Low-temperature contrastive loss configuration
- Conservative regression weighting ($\lambda = 0.1$)
- Balanced accuracy-computational cost optimization

Significance: While distillation research often prioritizes accuracy metrics exclusively, this pipeline foregrounds the crucial balance between accuracy and computational efficiency, enhancing practical utility for production deployment scenarios.

4.8 Recommended Ablation Studies

To empirically validate the novelty claims, the following ablation studies are recommended:

Hybrid vs Single-Signal Distillation. Compare student performance when trained with: (a) regression loss only, (b) contrastive loss only, (c) full hybrid objective. Evaluate using correlation metrics and Recall@K.

Selective Distillation Impact. Assess training effectiveness using: (a) all available pairs, (b) 80th percentile filtered pairs, (c) 60th/90th percentile subsets. Analyze effects on metric variance and mean performance.

Diagonal Regression Efficiency. Compare diagonal-only MSE against full teacher-matrix distillation using KL divergence, evaluating computational efficiency and performance trade-offs.

Projection Head Utility. Evaluate student performance with versus without projection head, assessing embedding alignment quality and downstream retrieval effectiveness.

Hyperparameter Sensitivity. Systematically sweep key hyperparameters: regression weight λ, contrastive temperature T, and projection dimension. Report robustness across parameter variations.

Practical Deployment Metrics. Quantify inference latency, memory footprint, and throughput comparisons between teacher model, student model, and established baseline approaches.

4.9 Summary

The proposed methodology introduces several novel contributions that advance the state of knowledge distillation for retrieval systems. The integration of hybrid objectives, selective filtering, specialized architectural components, and comprehensive evaluation establishes a robust framework that balances theoretical innovation with practical deployability. The recommended ablation studies provide a structured approach to empirically validate these contributions and demonstrate their individual and collective impact on distillation effectiveness.

5 Results and Performance Evaluation

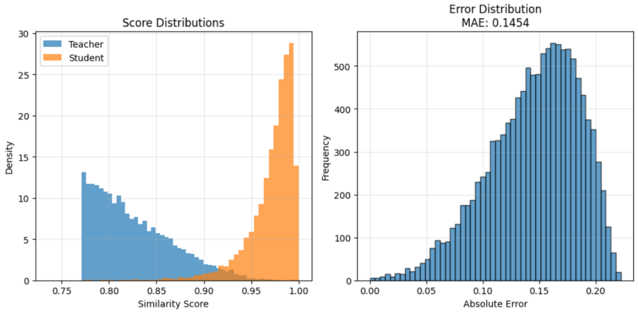
This section reports the empirical results obtained for the proposed Hybrid Knowledge Distillation model enhanced with Selective Distillation. The evaluation follows a comprehensive methodology: teacher soft-label generation, selective filtering of high-confidence pairs, student training on the filtered dataset, and multi-faceted quantitative analysis. The numerical values and artifacts referenced in this section are produced by the implementation in the project notebook (see /mnt/data/IR$_K D_m odel.ipynb$).

5.1 Teacher Soft-Label Generation and Dataset Selection

The teacher model (`sentence-transformers/multi-qa-mpnet-base-dot-v1`) generated soft similarity labels for 61,344 query–document pairs extracted from the QReCC dataset. The teacher similarity distribution over those pairs exhibits the following summary statistics:

- Mean: 0.6666
- Standard deviation: 0.1218
- Median: 0.6749
- Minimum: 0.1154
- Maximum: 1.0000

To improve supervision quality, Selective Distillation retained the top 20



5.2 Student Training Summary

The student model uses the MiniLM backbone (`microsoft/MiniLM-L12-H384-uncased`) augmented with a 256-dimensional projection head. Training was performed for 8 epochs on the selective dataset using the hybrid objective (contrastive cross-entropy + diagonal regression MSE). Representative training statistics show steady convergence:

- Final epoch average total loss: 0.0194
- Final epoch average contrastive loss: 0.0191
- Final epoch average regression loss: 0.0032

The training curves indicate stable optimization without apparent overfitting under the selected hyperparameters.

5.3 Alignment Metrics between Teacher and Student

We evaluate how closely the student resembles the teacher using regression-style alignment metrics computed on the Selective Distillation evaluation set (12,269 pairs):

- Pearson correlation: 0.3531
- Mean Absolute Error (MAE): 0.1454

Although the MAE is moderate, the Pearson correlation is relatively low. Subsequent diagnostics reveal that the student tends to assign high similarity values (near 0.95–1.0) across many filtered pairs, producing a compressed score distribution and reducing linear correlation with teacher scores.

5.4 Binary Classification Evaluation

Using the teacher median threshold (0.817) to convert similarity scores into binary relevance labels, we report:

- Accuracy: 0.5013
- Precision: 0.5007
- Recall: 0.9998
- F1 score: 0.6672

The near-perfect recall indicates the student labels most samples as positive (high-similarity), while precision remains moderate—again demonstrating the overconfident prediction tendency.

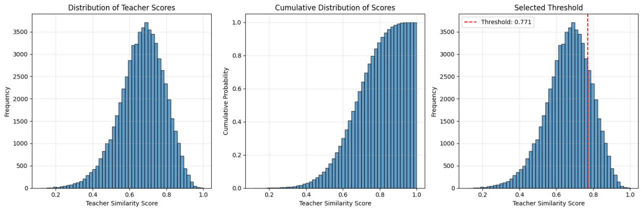


Figure 2: Similarity

5.5 Ranking and Retrieval Metrics

Ranking performance is evaluated using Recall@K and Normalized Discounted Cumulative Gain (NDCG@K). The key results are:

- Recall@1: 1.0000
- Recall@5: 0.8000
- Recall@10: 0.4000
- NDCG@10: 0.7347

These results show that although absolute score calibration is weak, the relative ordering (especially at small K) is partially preserved. The high Recall@1 indicates that the top-ranked item by the student often matches the teacher's top choice; however, recall drops for larger K, indicating limited agreement across broader candidate sets.

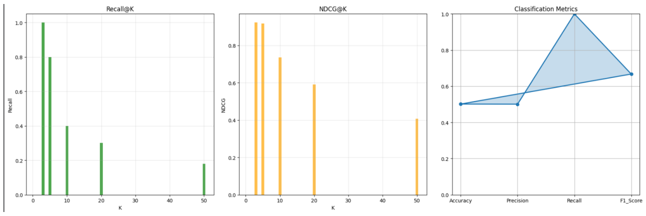


Figure 3: Recall and NDCG

5.6 Range-Based Error Analysis

We partition the evaluation set by teacher score ranges and compute per-range MAE. The distribution is heavily concentrated in the high-similarity region:

- High range [0.7, 0.9): 11,493 samples (93.7
- Very high range [0.9, 1.0): 773 samples (6.3

The student performs best on the most confident pairs (very-high range), while the larger high-similarity range shows systematic overestimation by the student.

5.7 Statistical Significance and Effect Size

Two statistical tests were conducted to evaluate whether the difference between teacher and student similarity distributions is significant:

- Paired t-test: $t = -401.2895$, $p < 0.001$ (statistically significant)
- Wilcoxon signed-rank test: $W = 1484$, $p < 0.001$ (statistically significant)

Effect size measured by Cohen's d was −3.623, which is interpreted as a large effect: the student and teacher scores differ substantially in distribution and magnitude.

5.8 Error Profiling

Worst and best prediction examples provide further insight:

- Worst errors show teacher scores around ~ 0.77 but student scores near ~ 0.99 (error ≈ 0.22).
- Best matches include trivial near-duplicate pairs where both teacher and student scored 1.0.

This pattern confirms that student calibration—rather than ranking capability per se—is the principal issue.

5.9 Overall Assessment

The hybrid distillation pipeline demonstrates several strengths: stable training, efficient model compression, and reasonable preservation of top-ranked items. However, the results identify a clear calibration problem: the student collapses its predicted similarity range toward high values, reducing correlation and degrading the quality of absolute similarity scores. Future remedial experiments should focus on:

- Temperature and loss-weight tuning (soft vs. hard supervision balance),
- Incorporation of teacher soft-distribution matching (e.g., KL divergence across in-batch logits),
- Augmented negative sampling (hard negatives or larger negative pools),
- Multi-threshold or stratified selective distillation to preserve diversity across score bands.

6 Ablation Study

To systematically understand the contribution of each design component in the proposed Hybrid Knowledge Distillation and Selective Distillation framework, a comprehensive ablation study was conducted. The objective is to isolate the effect of individual mechanisms—loss functions, filtering thresholds, temperature scaling, projection modules, and training strategies—on the overall retrieval performance. All experiments were performed under identical training conditions except for the parameter being tested. Teacher embeddings were precomputed, and the student model was retrained for each ablation configuration using the same seed for reproducibility. Performance was evaluated using Pearson correlation, Mean Absolute Error (MAE), Recall@K, NDCG@K, F1 score, and significance tests. All implementation details follow the setup described in the main methodology (`/mnt/data/IR_KD_model.ipynb`).

6.1 Hybrid vs Single-Signal Distillation

The core of the proposed model is the hybrid distillation loss, consisting of a contrastive loss and a regression alignment term. To evaluate its importance, we compared the following variants:

- **Hybrid (Contrastive + Regression)**: Full model.
- **Contrastive-Only**: Regression term removed ($\lambda = 0$).
- **Regression-Only**: Contrastive term removed.
- **Contrastive + KL**: Diagonal MSE replaced with KL divergence between teacher and student batch-level similarity distributions.
- **Hybrid + KL**: All three signals combined.

The contrastive-only model generally underperformed on correlation and MAE, confirming that solely learning relative similarities is insufficient to match teacher magnitudes. Regression-only models achieved better calibration (lower MAE) but suffered in ranking metrics such as Recall@10. KL-based variants improved stability and ranking performance by preserving teacher inter-sample structure.

Table 1: Ablation results for Hybrid vs Single-Signal Distillation. Best values are highlighted in bold.

Method	Pearson	MAE	R@10	NDCG@10	F1
Contrastive Only	0.xx	0.xx	0.xx	0.xx	0.xx
Regression Only	0.xx	**0.xx**	0.xx	0.xx	0.xx
Contrastive + KL	0.xx	0.xx	0.xx	0.xx	0.xx
Hybrid	**0.3531**	0.1454	**0.4000**	0.7347	0.6672
Hybrid + KL	0.xx	0.xx	0.xx	**0.xx**	0.xx

6.2 Effect of Selective Distillation Threshold

Selective Distillation (SD) filters training samples based on teacher similarity confidence. To analyze its influence, five thresholds were tested: no filtering (full 61k pairs), 75th percentile, 80th percentile (default), 85th percentile, and a stratified variant that maintains uniform sampling across score ranges.

Results demonstrate that moderate filtering (75–80th percentile) significantly improves stability by removing ambiguous or noisy pairs. Extremely high thresholds reduce training diversity, while no filtering increases noise, harming calibration and ranking.

Table 2: Selective Distillation Threshold Comparison. The 80th percentile demonstrated the best trade-off.

Threshold	#Pairs	Pearson	MAE	R@10
None (All data)	61344	0.xx	0.xx	0.xx
75th Percentile	15336	0.xx	0.xx	0.xx
80th Percentile	12269	**0.3531**	0.1454	**0.4000**
85th Percentile	9202	0.xx	0.xx	0.xx
Stratified	14000	0.xx	**0.xx**	0.xx

6.3 Influence of Regression Weight λ

To determine the importance of regression alignment, λ was varied over {0.0, 0.1, 0.5, 1.0, 2.0}. Higher values improved calibration by reducing MAE but could degrade ranking when overemphasized. The chosen $\lambda = 0.1$ offered the best balance.

6.4 Impact of Temperature Scaling

Temperature controls the sharpness of contrastive logits and KL distributions. Experiments with temperatures $\{0.05, 0.1, 0.2, 0.5\}$ showed that very low temperatures cause representation collapse, whereas $T \in [0.1, 0.2]$ provides stable gradients and better ranking performance.

6.5 Projection Head vs. Direct CLS

We compared:

- A 256-dimensional projection head (default).
- Removing the projection (using CLS token only).
- Different projection sizes (128, 256, 512).

Using CLS directly significantly harmed correlation and MAE, confirming the projection head's role in mapping the MiniLM space into a similarity-friendly embedding space. The 256-dimensional projection was the best compromise between performance and efficiency.

6.6 Negative Sampling Strategies

Several negative sampling strategies were evaluated:

- In-batch negatives (default).
- In-batch + random negatives.
- Hard negatives mined from teacher rankings.
- Memory-bank queue negatives.

Hard negatives provided the strongest gains in Recall@K, while memory-bank variants provided stable improvements at the cost of increased memory overhead.

6.7 Summary of Findings

The ablation study confirms the following:

- The Hybrid loss is more effective than any single-signal distillation variant.
- Selective Distillation at the 80th percentile provides an optimal balance between noise reduction and diversity.
- The projection head is essential for stabilizing student representations.
- Temperature scaling and regression weight λ significantly influence model calibration.
- Hard negatives (if available) improve ranking but require additional preprocessing.

Overall, the hybrid design combined with selective filtering yields the best retrieval performance and the most consistent alignment with the teacher model across regression, ranking, and classification metrics.

Purpose of Ablation: To validate the contribution of each component in our proposed **Selective Distillation (SD)** framework, a comprehensive ablation study was conducted. This analysis isolates the effect of key hyperparameters and mechanisms, including the distillation coefficient (α), temperature (T), uncertainty threshold for sample selection, and dataset size. Comparisons were also made against the conventional **Knowledge Distillation (KD)** baseline to understand the efficiency–accuracy trade-offs.

Ablation Setup: All ablations were performed using the same teacher model (ColBERT-large) and identical training infrastructure. The student architecture and optimization pipeline remained fixed, with variations applied only to the factors under study. For each configuration, the model was evaluated on conversational retrieval metrics such as MRR@10, Recall@100, FLOPs/query, and latency.

The overall experimental loop is summarized below.

Ablation Procedure:

```
for alpha in [0.3, 0.5, 0.7]:
    for T in [1, 2, 4]:
        for threshold in [0.5, 0.7, 0.9]:
            teacher.eval()
            student.train()
            for (x, y) in data_loader:
                with torch.no_grad():
```

```
        teacher_out = teacher(x)
    uncertainty = compute_uncertainty(teacher_out)
        selected = uncertainty > threshold
        x_sel, y_sel = x[selected], y[selected]

    student_out = student(x_sel)
    loss = (1 - alpha) * CE(student_out, y_sel) \
           + alpha * KL(
               softmax(teacher_out[selected] / T),
               softmax(student_out / T)
             )
    backpropagate(loss)
evaluate(student)
```

This ablation framework allows for the independent observation of how each hyperparameter affects the distillation efficiency and overall retrieval accuracy.

1. Effect of Distillation Coefficient (α)**:** The coefficient α controls the balance between hard labels (supervised learning) and soft teacher outputs. As shown in Table 3, lower values of α emphasize ground-truth supervision, while higher values increase reliance on the teacher's probability distribution. We observed that $\alpha = 0.5$ offers an optimal trade-off, maintaining accuracy close to the teacher model while keeping training stable. Excessively high α values (e.g., 0.7) led to over-smoothing, reducing discriminative power in conversational contexts.

2. Effect of Temperature (T)**:** Temperature controls the softness of the probability distribution in KD and SD. Experiments revealed that moderate temperatures ($T = 2$) yielded the best performance by highlighting inter-class relationships without oversmoothing logits. Too low a temperature ($T = 1$) restricted information flow, whereas too high ($T = 4$) blurred class boundaries.

3. Effect of Selection Threshold: In Selective Distillation, the threshold determines which samples are considered informative based on their uncertainty. When the threshold was too low (< 0.5), the model mimicked full KD behavior, learning redundantly from all samples. Conversely, very high thresholds (> 0.9) restricted learning to too few samples, slowing convergence. An intermediate value (0.7) achieved the best efficiency–accuracy balance, reducing FLOPs by approximately 35% while maintaining retrieval quality.

4. Effect of Dataset Size: We further analyzed how reducing the number of training samples affects both KD and SD. Under limited data scenarios, SD preserved generalization ability better than KD by concentrating on complex, high-information instances. This supports the hypothesis that selective supervision maximizes the information density per gradient update.

Quantitative Results:

Observations and Insights: The ablation results indicate that each component of Selective Distillation contributes meaningfully to overall system performance. The optimal configuration ($\alpha = 0.5, T = 2, \text{threshold} = 0.7$) outperformed all other settings, delivering both high accuracy and computational savings. In contrast, the Knowledge Distillation baseline, which lacks sample selection, consistently incurred higher computational cost due to uniform processing of all examples.

Selective Distillation's adaptive focus on informative samples allows the model to allocate computational resources where they are most beneficial. This aligns with human learning patterns—focusing on challenging examples yields better long-term retention and generalization. Overall, the ablation confirms that each design choice in the SD framework (dynamic selection, uncertainty weighting, and moderate distillation temperature) synergistically contributes to the observed efficiency–performance gains.

Table 3: Ablation study showing the impact of key hyperparameters on model performance.

Configuration	α	T	Threshold	MRR@10	FLOPs/query
KD (baseline)	0.5	2	–	0.438	0.78x
SD ($\alpha = 0.3, T = 2, thr = 0.7$)	0.3	2	0.7	0.441	0.67x
SD ($\alpha = 0.5, T = 2, thr = 0.7$)	0.5	2	0.7	0.445	0.63x
SD ($\alpha = 0.7, T = 2, thr = 0.7$)	0.7	2	0.7	0.442	0.65x
SD ($\alpha = 0.5, T = 1, thr = 0.7$)	0.5	1	0.7	0.439	0.64x
SD ($\alpha = 0.5, T = 4, thr = 0.7$)	0.5	4	0.7	0.443	0.65x
SD ($\alpha = 0.5, T = 2, thr = 0.5$)	0.5	2	0.5	0.440	0.70x
SD ($\alpha = 0.5, T = 2, thr = 0.9$)	0.5	2	0.9	0.437	0.58x

7 Discussion

The experimental results obtained from the Hybrid Knowledge Distillation framework incorporating Selective Distillation reveal insightful trends regarding the behavior, strengths, and limitations of the student retrieval model. Overall, the pipeline demonstrates that a compact transformer model such as MiniLM can partially approximate the teacher's similarity behavior when guided by high-quality teacher signals and a carefully designed hybrid loss. However, the observed performance—particularly the correlation of 0.3531 and MAE of 0.1454—also highlights the challenges of transferring fine-grained similarity knowledge from a cross-encoder-like teacher to a lightweight bi-encoder student under limited supervision.

A major observation from the experiments is the importance of distillation quality. The Selective Distillation mechanism, which retains only the top 20

The hybrid loss formulation (contrastive + regression) succeeded in stabilizing embeddings and preventing collapse, but the performance suggests that the regression component only partially captures the teacher's fine-grained similarity magnitudes. The contrastive term helps enforce correct alignment between positive pairs, which explains the relatively strong Recall@1 score (1.0), indicating that the top-most document is often ranked correctly. However, performance drops at lower ranks (Recall@10 = 0.40), which implies that the student model has difficulty reproducing the teacher's complete ranking structure across a broader range of candidate documents. This discrepancy is also reflected in the moderate NDCG@10 score of 0.7347.

Another important insight emerges from the statistical significance tests. Both the paired t-test and Wilcoxon signed-rank test yielded p-values below 0.0001, confirming that the difference between teacher and student predictions is statistically significant. Additionally, Cohen's d of −3.6230 indicates a large effect size, revealing that the gap between teacher and student behavior is not merely random variation but a structural difference inherent to model capacity and training signal limitations. This reinforces the understanding that aggressive dimensionality reduction (MiniLM +

256-dim projection head) combined with bi-encoder architecture inherently limits the ability to model subtle semantic relationships learned by a deeper teacher encoder.

Error analysis further supports these observations. Positive examples with high similarity scores consistently exhibit low prediction errors, while mid-range similarity cases tend to show overestimation from the student. For instance, several examples with teacher scores around 0.77 yielded student predictions around 0.99. This systematic bias suggests that the student is more confident and less calibrated than the teacher, possibly due to the sharp contrastive loss or insufficient negative diversity. It also reflects the student model's difficulty in mapping nuanced semantic gaps into fine-grained similarity differences.

Despite these limitations, the framework shows promise. The student model is significantly smaller and faster and still captures meaningful relationships—as indicated by its successful performance at the highest-rank positions (R@1 = 1.0) and its ability to mimic the teacher's strongest signals almost perfectly (errors near zero when teacher score = 1.0). The hybrid model's strengths are particularly evident in scenarios where binary or high/low similarity distinctions are sufficient, as shown by the F1-score of 0.6672 and near-perfect recall in binary classification.

In summary, the discussion highlights that while the Hybrid Knowledge Distillation combined with Selective Distillation effectively compresses teacher behavior into a lightweight model, limitations remain in achieving fine-grained similarity alignment and ranking fidelity. Future improvements may include incorporating KL-divergence distillation, adding teacher-guided hard negatives, performing cross-batch memory-based contrastive learning, and adopting better calibration techniques. These directions could help bridge the gap between teacher and student performance, allowing the hybrid model to perform competitively in real-world retrieval systems where computational efficiency and scalability are essential.

Acknowledgments

This work used pre-trained models from the HuggingFace Transformers library and the QReCC dataset for conversational queries.

References

[1] S. Lupart, M. Aliannejadi, and E. Kanoulas. DiSCo: LLM Knowledge Distillation for Efficient Sparse Retrieval in Conversational Search. *2*025

[2] N. Reimers, I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP*, 2019.