Relevance Scores Calibration for Ranked List Truncation via TMP Adapter

Pavel Posokhov[1,2], Sergei Masliukhin[1,2], Stepan Skrylnikov[1], Danil Tirskikh[1,2], Olesia Makhnytkina[2],
[1] STC-Innovations Ltd, St. Petersburg, Russia
[2] ITMO University, St. Petersburg, Russia
{posokhov,maslyukhin,skrylnikov,tirskikh}@speechpro.com,
makhnytkina@itmo.ru

Abstract

The ranked list truncation task involves determining a truncation point to retrieve the relevant items from a ranked list. Despite current advancements, truncation methods struggle with limited capacity, unstable training and inconsistency of selected threshold. To address these problems we introduce TMP Adapter, a novel approach that builds upon the improved adapter model and incorporates the Threshold Margin Penalty (TMP) as an additive loss function to calibrate ranking model relevance scores for ranked list truncation. We evaluate TMP Adapter's performance on various retrieval datasets and observe that TMP Adapter is a promising advancement in the calibration methods, which offers both theoretical and practical benefits for ranked list truncation.

1 Introduction

Determining the appropriate truncation point is a fundamental problem in information retrieval and recommendation systems. An excessively long ranked list can overwhelm users with redundant or less relevant information. Conversely, an overly short list risks omitting highly relevant items that could enhance user satisfaction. Thus, optimizing the cutoff point is essential to balance relevance, diversity, and usability. The problem of determining the optimal cutoff point in a ranked list, also known as ranked list truncation or relevance filtering, has been approached using two primary methods: adaptive thresholding and global thresholding.

Adaptive thresholding focuses on predicting an optimal cutoff point for each individual list. BiCut (Lien et al., 2019) leverages a bidirectional LSTM to model sequential dependencies and predict truncation points. Choppy (Bahri et al., 2020) employs a Transformer architecture for the same task. AttnCut (Wu et al., 2021) further incorporates attention mechanisms and reward augmented maximum likelihood for direct optimization. LeCut (Ma et al., 2022) improves upon these by adding contextual features from the retrieval task to better model document semantics. In the realm of personalized recommendations, PerK (Kweon et al., 2024) estimates the expected user utility to determine the ideal list size. More recently, GenRT (Xu et al., 2024) combines reranking and truncation in a joint model using sequence generation.

Global thresholding aims to calibrate relevance scores, enabling the use of a universal threshold across queries. This approach often involves transforming raw retrieval scores into more interpretable values. TCM (Zhang et al., 2024) introduces a margin-based loss that facilitates a consistent distance threshold and, RCR (Bai et al., 2023), a regression-compatible ranking approach, ensures alignment between ranking and regression objectives. JRC (Sheng et al., 2023) consolidates optimization across all samples using a contextualized hybrid model. The Cosine Adapter (Rossi et al., 2024) maps cosine similarity scores to interpretable relevance scores and Surprise (Bahri et al., 2023) employs statistical methods to adjust a ranked list using. These methods contrast with adaptive thresholding by seeking a single, universally applicable cutoff.

Despite the promising progress, we discover that existing methods suffer from three main issues: (i) low capacity, especially for Large Language Models. (ii) unstable training, especially for low-data training. (iii) threshold inconsistency especially in case of distribution shift between the training and test. We address these issues by proposing improved Adapter architecture and training method with Threshold Margin Penalty inspired by TCM.

2 Methodology

2.1 Threshold Margin Penalty

We propose an additive penalty function with adaptive margin for contrastive loss functions. The goal

Findings of the Association for Computational Linguistics: ACL 2025, pages 7728–7734
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

of this function is to minimize the number of pair scores s located in the truncation threshold area to improve threshold consistency and global pair separation of positive S^+ and negative S^- scores. Threshold Margin Penalty is defined in Equation 1

$$TMP = w_{pos} * P_{pos} + w_{neg} * P_{neg} - w_m * R_m \quad (1)$$

Where P_{pos} and P_{neg} are penalties for positive and negative scores defined in equations 2 and 3.

$$P_{pos} = \frac{\sum_{s \in S^+} \max\left(0, m^+ - s\right)}{\sum_{s \in S^+} \begin{cases} 1; & s \leq m^+ \\ 0; & s > m^+ \end{cases}} \quad (2)$$

$$P_{neg} = \frac{\sum_{s \in S^-} \min\left(0, s - m^-\right)}{\sum_{s \in S^-} \begin{cases} 1; & s > m^- \\ 0; & s \leq m^- \end{cases}} \quad (3)$$

R_m - margin reward which encourages better separation given in equation 4.

$$R_m = m^+ - m^- \quad (4)$$

Since the optimal truncation point could change during training, we add tunable parameters m^+ and m^-, which are normalized using the sigmoid function that change positive and negative boundaries. This allows us to tune optimal margin placement and size during training. We also include penalty weights hyperparameters w_{pos}, w_{neg} and w_m empirically selected based on experimental results remaining close to main loss to save better convergence. w_{pos} and w_{neg} are codependent and guide the distributions bias. w_m determines margin size dynamics and should be proportional to the sum of w_{pos}, w_{neg}, increasing this parameter enhances scores separation but may lead to training instability.

2.2 TMP Adapter

We recognize the potential of the Cosine Adapter model; however, we also identify several limitations, including low consistency of the truncation threshold, insufficient generalization ability, and unstable training. In this study, we build upon the concept of the Cosine Adapter and address these issues by proposing the TMP Adapter, depicted in Figure 1. The adjusted score s is computed using a modified function presented in equation 5.

$$s = p_1 + s_{raw}^{p_3} * p_2^2 \quad (5)$$

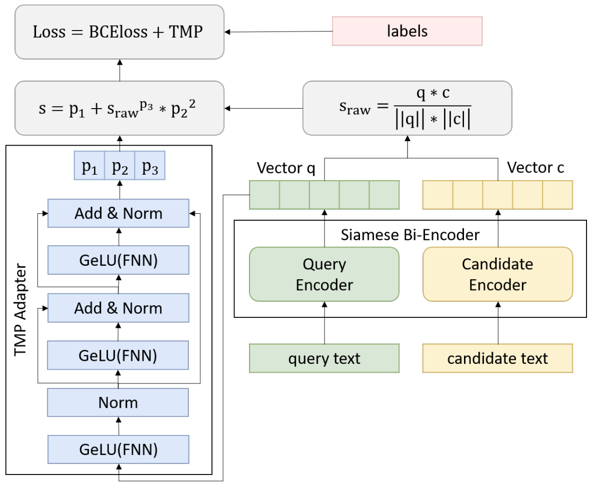

Figure 1: TMP Adapter architecture and training pipeline for Bi-Encoder scores calibration.

where s_{raw} is cosine similarity between query vector q and candidate vector c.

To enhance threshold consistency, measured by the deviation of the validation-set-optimized threshold from the optimal test-set threshold, we introduce the Threshold Margin Penalty. This method expands the optimal threshold region without encoder model tuning, similar to several previously mentioned methods. Additionally, we propose increasing the model's capacity and modifying its architecture by incorporating residual connections and GeLU activation functions to improve training stability (see Appendix A).

3 Experiments

3.1 Datasets

In this paper, we utilize three key information retrieval datasets from BeIR benchmark (Thakur et al., 2021). FiQA is a domain-specific dataset of financial questions and answers, designed for retrieval models evaluation. NFCorpus is a dataset of health-related documents with human-annotated relevance judgments, applicable for IR tasks in medicine. Robust04 is a widely used benchmark from the TREC Robust Track 2004, based on news articles with relevance assessments, designed to test the robustness of retrieval models across domains of varying difficulty. This setup provides diverse retrieval challenges from domain-specific to general information retrieval tasks. We selected an amount of varying datasets to evaluate threshold consistency and quality of ranked list truncation (which requires both training stability and model capacity) on different domains. Full datasets characteristics are available in Appendix B.

3.2 Metrics

In this paper we report Normalized Discounted Cumulative Gain at rank 10 (NDCG@10) as retrieval quality metric, as it accounts for both the relevance and position of retrieved documents. While NDCG@10 is the standard evaluation metrics of the retrieval task in MTEB benchmark (Muennighoff et al., 2023) and particularly relevant for encoder tuning experiments, it is not the primary metric to assess the proposed method. The TMP Adapter is implemented as score calibrator rather than reranker, leading to ranking metrics remaining unchanged. To comprehensively evaluate ranked list truncation we consider several key metrics. The maximum F1 score $(F1(M))$ represents the maximum F1 value for a given ranked search result list without reranking. We also report the oracle F1 score $(F1(O))$, obtained by optimizing the threshold on test subset. In contrast, the tuned F1 score $(F1(T))$ is derived by adjusting the threshold on the dev subset. For better interpretation we report $\frac{F1(T)}{F1(M)}$ that calculated as percentage of the maximum F1 score. To quantify the threshold consistency we compute the $\frac{F1(T)}{F1(O)}$ percentage ratio.

3.3 Baselines

We employ multiple baseline methods to ensure a comprehensive and reliable evaluation. First of all, we consider the AttnCut approach [1] and Cosine Adapter [2]. In addition we report two naive baselines: Greedy(k) - truncation based on global rank threshold; Greedy(s) - truncation based on global scores threshold.

To assess the effectiveness of ranked list truncation methods under current conditions, we identify state-of-the-art retrieval models and compare them to the approaches introduced in AttnCut (BM25) and polynomial Cosine Adapter (SimLM) (GLUE, 2022).

To address the use of the proposed method on different sized models we incorporate the small retrieval model Spice[3], which holds the highest ranking among small models having 33.4M parameters in the retrieval task of the MTEB leaderboard (as of January 30, 2025).

We also include NV-Embed-v2 (Lee et al., 2025) having 7B parameters, which is ranked first on the MTEB retrieval leaderboard (as of January 31, 2025) to benchmark our results against state-of-the-art Large Language Models. By incorporating these diverse baselines, we aim to provide a robust comparative analysis, highlighting the advantages and limitations of the proposed method in various retrieval scenarios and its compatibility both with small and large models. We maintain the original performance of baseline models without additional tuning, as we do not rerank the retrieved list. Proposed method serves exclusively as a calibrator for optimal threshold selection. All of the baselines are presented in Table 1. To further comparison of ranked list truncation methods we select two models with the best $F1(M)$ scores.

4 Results

4.1 Threshold Results

The relative results of the suggested TMP Adapter (for training details see Appendix C) and other truncation methods baselines are listed in Table 2. Absolute values are reported in Appendix D. Threshold consistency results of the TMP Adapter show an F1(T/O) increase in 4.25%pt over raw scores (Greedy(s)) and 2.24%pt over the best baseline model (Cosine Adapter). We attribute the use of TMP the primary factor leading to this increase in model's consistency.

TMP Adapter shows stable improvements in ranked list truncation metrics over all datasets in contrast to the Cosine Adapter, which indicates more stable training due to architecture's modifications.

All of these factors combined lead to ranked list truncation metrics improvement, allowing the TMP Adapter to achieve F1(T/M) increase both in raw scores (Greedy(s)) 9.08%pt, and an 5.75%pt improvement over the best baseline (Cosine Adapter), which confirms the effectiveness of proposed score calibration method.

4.2 Discussion

Experimental results indicate that the optimal threshold is changing during the training process. This dynamics can be observed visually analyzing F1-score curves obtained at model's validations at different training epochs (Appendix E). Notably, the peak F1-scores are achieved across wide range of thresholds, varying from 0 to 1. Therefore, margin penalty with fixed boundaries will prevent this behavior and reduce optimization efficiency due

[1] https://github.com/Woody5962/Ranked-List-Truncation
[2] https://github.com/juexinlin/dense_retrieval_relevance_filter
[3] https://huggingface.co/iamgroot42/spice

Model	FiQA		NFCorpus		Robust04	
	NDCG@10	F1(M)	NDCG@10	F1(M)	NDCG@10	F1(M)
NV-Embed-v2	**0.652**	**0.643**	0.449	0.381	0.405	**0.469**
Spice	0.63	0.623	**0.544**	**0.478**	**0.407**	0.345
MiniLM	0.188	0.212	0.231	0.2	0.178	0.143
BM25	0.253	0.391	0.342	0.344	0.343	0.228

Table 1: The evaluation of different ranking models on four datasets with ranking and truncation metrics.

Method	FiQA		NFCorpus		Robust04	
	F1(T/M)	F1(T/O)	F1(T/M)	F1(T/O)	F1(T/M)	F1(T/O)
			Spice			
Greedy(s)	56.34	98.60	53.97	92.81	64.35	95.28
Greedy(k)	58.91	89.95	54.18	95.57	63.48	95.22
AttnCut	**64.52**	–	55.65	–	60.87	–
Cosine Adapter	59.23	99.73	60.25	97.30	56.81	93.33
TMP Adapter	64.04	**99.75**	**65.27**	**99.36**	**66.67**	**97.87**
			NV-Embed-v2			
Greedy(s)	52.41	89.63	58.53	96.96	68.44	96.69
Greedy(k)	69.21	**100**	56.17	**100**	68.23	97.86
AttnCut	67.19	–	55.91	–	68.66	–
Cosine Adapter	67.19	98.18	62.99	97.96	67.59	95.48
TMP Adapter	**71.54**	99.14	**66.40**	99.61	**74.63**	**99.72**

Table 2: The results of ranked list truncation on three datasets and two encoder model for baselines and our approach. Metric $F1(T/M)$ shows percentage ratio $F1(T)$ to $F1(M)$ and reveal the calibration quality. Metric $F1(T/O)$ shows percentage ratio $F1(T)$ to $F1(O)$ and reveal the threshold consistency. Dashes in the table indicate the absence of oracle value for AttnCut method, making it impossible to compute threshold consistency.

to counteracting the main pairwise loss function. Consequently, the optimal threshold margin cannot be reliably determined using a fixed grid search approach but must be dynamic. These results are supported by heatmap shown in Figure 2.

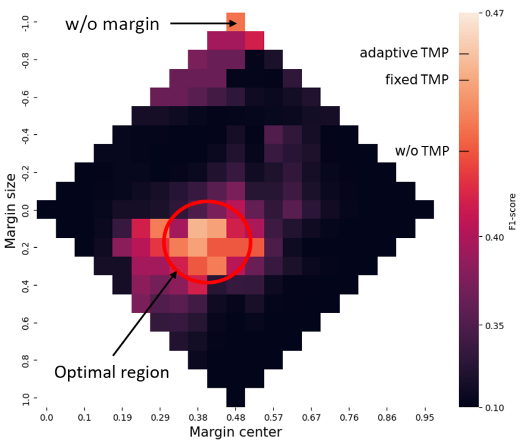

Figure 2: Performance of TMP Adapter with various fixed margin center and margin size parameters computed on FiQA dataset for NV-Embed-v2 model.

5 Conclusion

In this paper, we introduce Threshold Margin Penalty Adapter, a novel approach designed to calibrate ranking model relevance scores for ranked list truncation. Proposed TMP Adapter extends the improved adapter model by integrating the Threshold Margin Penalty as an additive loss function. This innovation enhances the model's ability to maintain threshold consistency and improves the separation between positive and negative pairs, which is critical for effective ranking list truncation. We evaluate TMP Adapter's performance on four datasets and observe a consistent and stable improvement in the F1-score, highlighting the model's effectiveness for score separation. Additionally, we observe a significant enhancement in threshold consistency, which underscores the model's in-domain robustness to maintain reliable decision boundaries. These findings show that TMP Adapter is a promising advancement in calibration methods, offering both theoretical and practical benefits for ranked list truncation.

Limitations

While the proposed method improves in-domain threshold consistency and training stability, it has limitations. First of all, it struggles with out-of-domain generalization, performing poorly outside its training domain. This restricts its applicability in diverse real world applications. Furthermore, requiring a sufficient number of training pairs for effective score calibration, similar to the Cosine Adapter, makes this approach challenging in training with small amount of data, despite enhancing training stability. These limitations highlight the need for further research into domain adaptation, data-efficient calibration, and computational optimization to enhance its real-world applicability.

References

Dara Bahri, Yi Tay, Che Zheng, Donald Metzler, and Andrew Tomkins. 2020. Choppy: Cut transformer for ranked list truncation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1513–1516, New York, NY, USA. Association for Computing Machinery.

Dara Bahri, Che Zheng, Yi Tay, Donald Metzler, and Andrew Tomkins. 2023. Surprise: Result list truncation via extreme value theory. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2404–2408, New York, NY, USA. Association for Computing Machinery.

Aijun Bai, Rolf Jagerman, Zhen Qin, Le Yan, Pratyush Kar, Bing-Rong Lin, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2023. Regression compatible listwise objectives for calibrated ranking with binary relevance. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4502–4508, New York, NY, USA. Association for Computing Machinery.

PLM MS-MARCO GLUE. 2022. Simlm: Pre-training with representation bottleneck for dense passage retrieval. *ELECTRA*, 31:89–4.

Wonbin Kweon, SeongKu Kang, Sanghwan Jang, and Hwanjo Yu. 2024. Top-personalized-k recommendation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3388–3399, New York, NY, USA. Association for Computing Machinery.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Yen-Chieh Lien, Daniel Cohen, and W. Bruce Croft. 2019. An assumption-free approach to the dynamic truncation of ranked lists. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 79–82, New York, NY, USA. Association for Computing Machinery.

Yixiao Ma, Qingyao Ai, Yueyue Wu, Yunqiu Shao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Incorporating retrieval information into the truncation of ranking lists for better legal search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 438–448, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicholas Rossi, Juexin Lin, Feng Liu, Zhen Yang, Tony Lee, Alessandro Magnani, and Ciya Liao. 2024. Relevance filtering for embedding-based retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4828–4835, New York, NY, USA. Association for Computing Machinery.

Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Joint optimization of ranking and calibration with contextualized hybrid model. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4813–4822, New York, NY, USA. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2021. Learning to truncate ranked lists for information retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4453–4461.

Shicheng Xu, Liang Pang, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. List-aware reranking-truncation joint model for search and retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1330–1340, New York, NY, USA. Association for Computing Machinery.

Q Zhang, L Xu, Q Tang, J Fang, YN Wu, J Tighe, and Y Xing. 2024. Threshold-consistent margin loss for open-world deep metric learning. International Conference on Learning Representations (ICLR).

A Architecture Modification

To determine the optimal architecture with sufficient capacity, we conduct an ablation study. The results for various adapter architectures are presented in Table 3. Training values are reported on the dataset split used for model training, while test metrics evaluate the model's performance on a previously unseen dataset split.

#Layers	Residual Connection	Activation Function	F1(T) Train	F1(T) Test
3	False	ReLU	0.499	0.432
4	False	ReLU	0.511	0.430
4	True	ReLU	0.530	0.446
4	True	GeLU	0.535	0.450
5	False	ReLU	0.502	0.402
5	True	ReLU	0.525	0.420
5	True	GeLU	0.528	0.422

Table 3: Evaluation of various adapter architectures with NV-Embed-v2 model modification on FiQA dataset. The table includes number of additional fully-connected layers, the use of residual connections, activation function between layers, train and test metrics.

B Dataset Description

We use question answering and information retrieval datasets, commonly used to evaluate truncation methods and included both in BEIR and MTEB benchmarks. Their characteristics are shown in Table 4.

Dataset	FiQA	NFCorpus	Robust04
Domain	Finance	Medicine	News
#Docs	57.6K	3.6K	528K
#Queries	6.6K	3.2K	250
#Positives	3	43	70
#Train Set	5.5K	2.6K	150
#Val Set	500	324	50
#Test Set	648	323	50
#Labels	2	4	3
Doc Length	136	221	605

Table 4: Overview of Datasets used in research including their domains, sizes, query counts, label distributions, and document lengths in words. Used datasets significantly vary in domains and scope, with Robust04 having the most number of relevant documents, while FiQA having the most queries.

C TMP Adapter Training Setup

We train the TMP Adapter without tuning the encoder models, utilizing a modified Cosine Adapter pipeline and the proposed TMP Adapter model trained with the parameters specified in Table 5.

Spice			
Dataset	FiQA	NFCorpus	Robust04
#Epochs	40	25	25
Batch Size	128	128	32
Optimizer	AdamW	AdamW	AdamW
Adapter lr	0.001	0.002	0.0005
Margin lr	0.008	0.005	0.01
w_{pos}	0.109	0.198	0.212
w_{neg}	0.1	0.1	0.104
w_m	0.25	0.19	0.25
NV-Embed-v2			
#Epochs	50	25	20
Batch Size	128	128	32
Optimizer	AdamW	AdamW	AdamW
Adapter lr	0.001	0.001	0.0005
Margin lr	0.01	0.005	0.01
w_{pos}	0.102	0.202	0.209
w_{neg}	0.1	0.093	0.106
w_m	0.2	0.18	0.26

Table 5: TMP Adapter Training parameters for Spice and NV-Embed-v2 models used in research.

D Absolute F1 Values

In addition to the relative results of the TMP Adapter described in the paper, we report absolute values of tuned F1 and oracle F1 metrics for more comprehensive and complete description in Table 6.

E Threshold Shifting

To provide a clearer demonstration of the threshold shifting during training, that is observed for all adapter models, we report a curve of the validation F1 metric values, recorded every five epochs in Figure 3.

Method	FiQA		NFCorpus		Robust04	
	F1(T)	F1(O)	F1(T)	F1(O)	F1(T)	F1(O)
Spice						
Greedy(s)	0.351	0.356	0.258	0.278	0.222	0.233
Greedy(k)	0.367	0.408	0.259	0.271	0.219	0.230
AttnCut	0.402	–	0.266	–	0.210	–
Cosine Adapter	0.369	0.370	0.288	0.296	0.196	0.210
TMP Adapter	0.399	0.400	0.312	0.314	0.230	0.235
NV-Embed-v2						
Greedy(s)	0.337	0.376	0.223	0.230	0.321	0.332
Greedy(k)	0.445	0.445	0.214	0.214	0.320	0.327
AttnCut	0.432	–	0.213	–	0.322	–
Cosine Adapter	0.432	0.440	0.240	0.245	0.317	0.332
TMP Adapter	0.460	0.464	0.253	0.254	0.350	0.351

Table 6: The results of ranked list truncation on three datasets and two encoder model for baselines and our approach in absolute values.

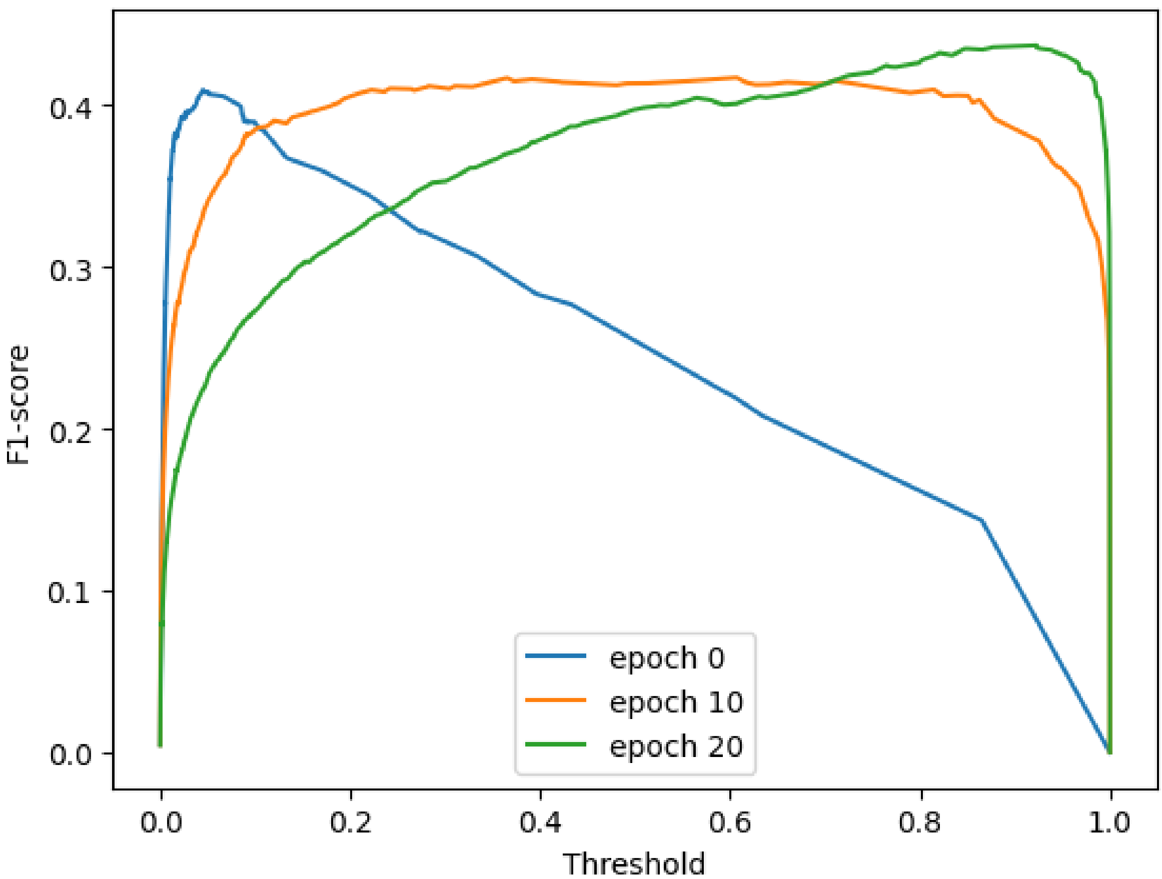

Figure 3: Validation F1 curve Cosine Adapter on FiQA dataset for NV-Embed-v2 model.