Preserving Zero-shot Capability in Supervised Fine-tuning for Multi-label Text Classification

Si-An Chen*
Google
sianchen@google.com

Hsuan-Tien Lin
National Taiwan University
htlin@csie.ntu.edu.tw

Chih-Jen Lin
MBZUAI and
National Taiwan University
cjlin@csie.ntu.edu.tw

Abstract

Zero-shot multi-label text classification (ZMTC) requires models to predict multiple labels for a document, including labels unseen during training. Previous work assumes that models leveraging label descriptions ensures zero-shot capability. However, we find that supervised methods, despite achieving strong overall performance, lose their zero-shot capability during training, revealing a trade-off between overall and zero-shot performance. To address the issue, we propose OF-DE and OF-LAN, which preserve the zero-shot capabilities of powerful dual encoder and label-wise attention network architectures by freezing the label encoder. Additionally, we introduce a self-supervised auxiliary loss to further improve zero-shot performance. Experiments demonstrate that our approach significantly improves zero-shot performance of supervised methods while maintaining strong overall accuracy.

1 Introduction

Zero-shot multi-label text classification (ZMTC) is a crucial task in natural language processing (NLP) where models must assign multiple labels to a text document, including labels the model has never encountered during training. These labels, predefined in a label set but not associated with any instance from the training data, are termed zero-shot or unseen labels. To address ZMTC, each label is typically accompanied by a description, allowing the model to leverage semantic information for better predictions. ZMTC is applied across various domains, such as legislative document tagging (Loza Mencía and Fürnkranz, 2008; Chalkidis et al., 2019), medical code prediction (Mullenbach et al., 2018; Rios and Kavuluru, 2018), and product categorization (Lewis et al., 2004; McAuley and Leskovec, 2013). For instance, in medical code prediction, using code descriptions to predict rare, unseen diseases can significantly improve diagnostic coverage.

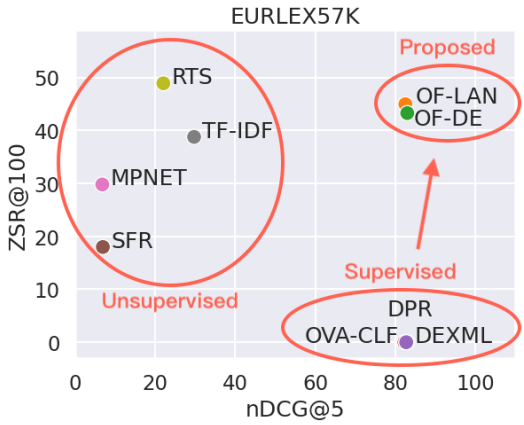


Figure 1: Overall performance (nDCG@5) and zero-shot performance (ZSR@100) of existing methods and our solutions on EURLEX57K dataset. Our method improves zero-shot performance while maintaining the superior overall performance of supervised approaches.

Existing approaches often assume that incorporating label descriptions inherently enables zero-shot capability, allowing models to predict unseen labels based on their semantics (Lee et al., 2018; Aggarwal et al., 2023). However, our analysis reveals that this assumption does not always hold. We categorize ZMTC methods into unsupervised and supervised approaches, based on whether document-label mappings are used during obtaining the model (detailed in Sec 2). As illustrated in Fig. 1, there is a trade-off between overall performance (which evaluates both seen and unseen labels) and zero-shot performance (which focuses solely on unseen labels). Supervised methods, while achieving strong overall performance, tend to lose zero-shot capability even when incorporating label descriptions. This phenomenon echoes population bias in recommendation systems (Abdollahpouri et al., 2019a,b), where models overfit

*The work was done while the author was a student at National Taiwan University

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5714–5727

to seen items, limiting their ability to recommend unseen items effectively. To the best of our knowledge, we are the first to concretely demonstrate this trade-off in multi-label classification.

In this work, we aim to address the trade-off by preserving the zero-shot capabilities for supervised fine-tuning approaches. We analyze the most common dual encoder architecture, where embeddings for documents and labels are generated by encoders and predictions are made based on the similarity between these embeddings. We observed that during supervised training, the embeddings of zero-shot labels generated by the label encoder often collapse into a small, isolated region, limiting the model's ability to generalize to zero-shot labels. This motivates us to propose a novel framework, One-sided Fine-tuned Dual Encoders (OF-DE), which freezes the label encoder to preserve its semantic richness while fine-tuning only the document encoder. This strategy retains zero-shot capabilities without sacrifice overall performance. Additionally, we introduce a self-supervised auxiliary loss that aligns document and label encoders through training on label descriptions, improving generalization to unseen labels.

Following the success of OF-DE, we propose an advance architecture, One-sided Fine-tuned Label-wise Attention Networks (OF-LAN), inspired by the state-of-the-art label-wise attention networks (LAN, Mullenbach et al., 2018). The original LAN applies label-wise attention to focus on relevant parts of the document for each label, leading to better overall performance. However, it relies on label-specific parameters learned from seen labels, which limits its effectiveness for zero-shot labels. To overcome this limitation, OF-LAN generates label-specific parameters from label descriptions, allowing it to leverage label information effectively. As a result, OF-LAN not only achieves state-of-the-art overall performance but also enhances the model's ability to predict unseen labels.

Our analysis demonstrates that our methods prevent the collapse of label embeddings, ensuring a more balanced distribution. This enhances the model's ability to generalize to unseen labels, effectively addressing the trade-off between overall and zero-shot performance.

In summary, our contributions are:

- We highlight the under-explored trade-off between overall and zero-shot performance in ZMTC, showing its connection to supervised and unsupervised approaches.
- We analyze the reasons behind the loss of zero-shot capability in models after supervised fine-tuning.
- We propose a one-sided fine-tuning framework and a self-supervised auxiliary loss that preserve zero-shot generalization without sacrificing overall performance.
- Extensive experiments demonstrate that our approach preserves strong zero-shot capabilities while achieving overall performance competitive with state-of-the-art methods.

Implementation for experiments are available at https://www.csie.ntu.edu.tw/~cjlin/papers/zero_shot_one_side_tuning/.

2 Related Work

We categorize existing work on ZMTC into two main approaches, unsupervised and supervised, based on whether they utilize ground truth document-label mappings, as illustrated in Fig. 2.

2.1 Unsupervised Approaches

Unsupervised methods include traditional statistical approaches such as TF-IDF (Salton and Buckley, 1988) and BM25 (Robertson et al., 2009), which rely on term co-occurrence. These methods are computationally efficient but fail to capture deeper semantic relationships. Advanced approaches use pre-trained language models, such as SentenceBERT (Reimers and Gurevych, 2019), BGE (Xiao et al., 2024), GTE (Zhang et al., 2024), and SFR (Meng et al., 2024). These models are used without fine-tuning, as shown in Fig. 2(b). However, without learning the document-label relationships, these models usually lead to weak overall performance.

Self-supervised pre-training approaches, such as MACLR (Xiong et al., 2022) and RTS (Zhang et al., 2022), improve the models by fine-tuning with task-specific documents and label descriptions using self-supervised training, as shown in Fig. 2(c). Though these methods enhance data representations, the lack of using document-label mapping still limits their effectiveness in ZMTC.

Generative-based approaches, such as ICXML (Zhu and Zamani, 2024), prompt large language models to generate predictions. While these methods benefit from the extensive

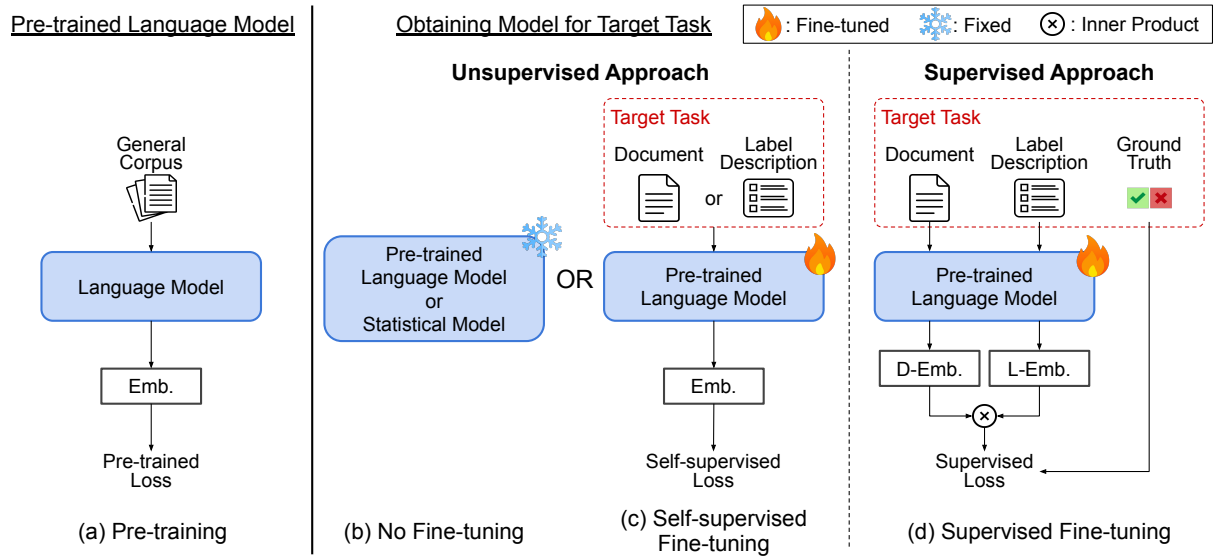


Figure 2: Overview of obtaining models for a target task. (a) **Pre-training**: Language models are pre-trained on a massive corpus to learn general-purpose text embeddings. Then two approaches can be taken: unsupervised (b, c) or supervised (d). (b) **No Fine-tuning**: Pre-trained or statistical models are directly applied to the target task without any fine-tuning. (c) **Self-supervised Fine-tuning**: Pre-trained models are fine-tuned using documents and label descriptions from the target task through self-supervised learning, without using ground truth document-label mappings. (d) **Supervised Fine-tuning**: Pre-trained models are fine-tuned using documents, label descriptions, and ground truth document-label mappings via supervised learning.

knowledge embedded in large models, they are beyond the scope of our study due to their high computational cost and the need for complex, task-specific prompt engineering.

2.2 Supervised Approaches

Supervised methods fine-tune models using ground truth document-label mappings, as shown in Fig. 2(d). These include symmetric dual encoder architectures like DEXML (Gupta et al., 2024), asymmetric dual encoders such as DPR (Karpukhin et al., 2020) and SemSup-XC (Aggarwal et al., 2023), and statistical models like 0-BiGRU-LWAN (Chalkidis et al., 2020) and ZestXML (Gupta et al., 2021). Despite utilizing label descriptions, these methods often lose their zero-shot capabilities after fine-tuning, as demonstrated in Fig. 1.

Classifier-based methods, originally not designed for zero-shot tasks, typically combine a document encoder with a linear classifier to predict logits for each label. Examples include OVA-classifier (Devlin et al., 2019), LAN (Chalkidis et al., 2020), AttentionXML (You et al., 2019), DeepXML (Dahiya et al., 2021), and Renee (Jain et al., 2023). Some approaches, such as ECLARE (Mittal et al., 2021b), DECAF (Mittal et al., 2021a), NGAME (Dahiya et al., 2023a), and DEXA (Dahiya et al., 2023b), combines classifiers with label encoders to achieve better performance. However, these methods rely on label-specific weights, which are only trained for labels seen during training and therefore unsuitable for ZMTC tasks.

IRENE (Yadav et al., 2024), a plug-and-play approach, learns classifier weights for zero-shot labels to enable the prediction of unseen labels. However, its performance is highly dependent on the underlying classifier, making it less comparable to our work.

3 Problem formulation

In ZMTC, the goal is to assign a subset of labels from a set $\mathcal{L}$ to each document instance in $\mathcal{X}$, where $\mathcal{L} = \mathcal{L}_{\text{seen}} \cup \mathcal{L}_{\text{unseen}}$, and $\mathcal{L}_{\text{seen}}$, $\mathcal{L}_{\text{unseen}}$ represent the sets of seen and unseen labels, respectively. Seen labels are those associated with at least one instance in the training set, whereas unseen labels have no associations with any training instances. Given a dataset of N document instances $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, and $|\mathcal{L}|$ label descriptions $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{|\mathcal{L}|}\}$, the task is to predict the ground truth document-label mapping $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ indicates that Z_j is relevant to X_i.

Note that in some cases, $\mathcal{L}_{\text{unseen}}$ is unknown dur-

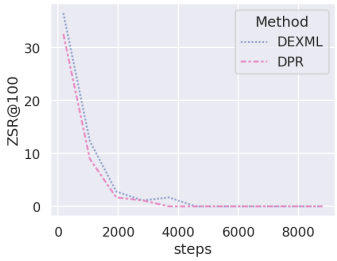
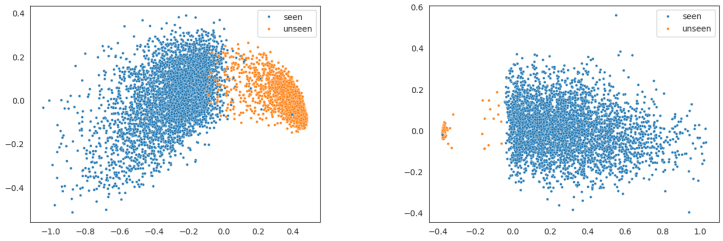


Figure 3: The progress of zero-shot performance of DEXML and DPR on EURLEX57K during training.



(a) DEXML (b) DPR

Figure 4: 2D visualization of label embeddings on the EURLEX dataset for DEXML and DPR, using PCA projection.

ing training. This work focuses on the scenario where $\mathcal{L}_{\text{unseen}}$ is known but also includes a discussion of the unknown case in Sec. 4.5 and Appendix D.

4 Proposed Method

We begin by analyzing why supervised fine-tuning approaches often lose their zero-shot capability. To address this issue, we propose the One-sided Fine-tuned Dual Encoder (OF-DE) and the more advanced One-sided Fine-tuned Label-wise Attention Network (OF-LAN). Additionally, we introduce a self-supervised auxiliary loss to further enhance zero-shot performance.

4.1 Motivation

Previous work in zero-shot learning often assumes that models incorporating label descriptions inherently possess zero-shot capability (Lee et al., 2018; Aggarwal et al., 2023). However, this assumption has not been thoroughly examined. In fact, we observe a significant challenge in supervised fine-tuning: the loss of zero-shot capability during training. As shown in Fig. 3, the zero-shot performance of supervised methods, such as DEXML (Gupta et al., 2024) and DPR (Karpukhin et al., 2020), consistently drops to zero as fine-tuning progresses. Despite improving performance on seen labels, these methods fail to retain their ability to predict unseen labels, which is crucial in ZMTC tasks.

Upon further analysis, Fig. 4 reveals the root cause of this issue. The embeddings of zero-shot label descriptions, which should ideally retain their semantic richness and diversity, tend to collapse into a small, isolated region during supervised fine-tuning. This suggests that fine-tuning on document-label mappings causes the model to overfit on seen labels. As a result, the collapse of these embeddings significantly hampers the model's ability to generalize to unseen labels, as they no longer effectively capture the distinctiveness of the zero-shot labels.

4.2 One-sided Fine-tuned Dual Encoder (OF-DE)

To address the issue discussed in Sec. 4.1, we propose a novel one-sided fine-tuning framework that freezes the label encoder in the dual encoder architecture, preserving the semantic richness and diversity of label embeddings, as shown in Fig. 5(b). In this framework, a document X_i is encoded using a pre-trained encoder $\mathcal{E}_{\text{doc}}(X_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the trainable parameters. The document embedding is computed as follows:

$$\begin{aligned} \boldsymbol{x}_i &= \mathcal{E}_{\text{doc}}(X_i; \boldsymbol{\theta}) \\ \boldsymbol{x}_i^{\text{pool}} &= \mathcal{P}(\boldsymbol{x}_i) \end{aligned} \tag{1}$$

where $\boldsymbol{x}_i$ is the sequence of token embeddings, $\mathcal{P}$ is a pooling operation that summarizes the sequence into a single embedding, and $\boldsymbol{x}_i^{\text{pool}}$ is the resulting document embedding.

For a label description Z_j, we use a fixed, distinct pre-trained model $\mathcal{E}_{\text{label}}(Z_j)$ to produce its label embedding:

$$\boldsymbol{z}_j^{\text{pool}} = \mathcal{P}(\mathcal{E}_{\text{label}}(Z_j)) \tag{2}$$

To address differences in embedding dimensions between the document and label encoders, we learn a linear projection layer to align them:

$$\boldsymbol{z}_j^{\text{proj}} = W_{\text{proj}} \boldsymbol{z}_j^{\text{pool}} + b_{\text{proj}}$$

where W_{proj} and b_{proj} are the trainable parameters of the projection layer.

The final prediction $\hat{y}_{ij}$ is computed as:

$$\hat{y}_{ij} = f(\boldsymbol{x}_i^{\text{pool}}, \boldsymbol{z}_j^{\text{proj}}) \tag{3}$$

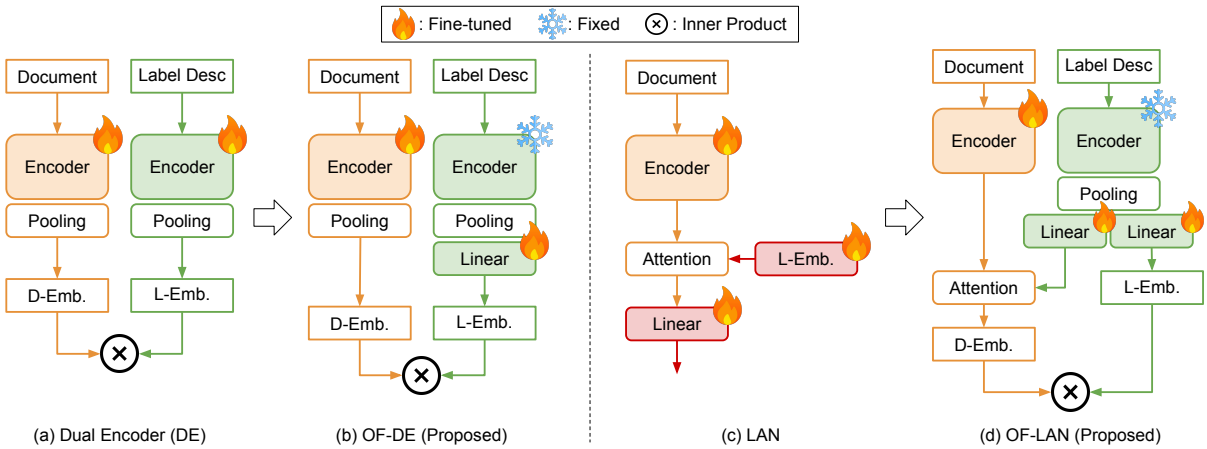


Figure 5: Architectures of existing supervised solutions (DE, LAN) and our proposed methods (OF-DE, OF-LAN). Red blocks indicate label-specific components that do not function for zero-shot labels.

where f is a similarity function such as Euclidean similarity or cosine similarity. By keeping the label encoder unchanged, we may prevent the collapse of zero-shot label embeddings. In this situation, we can preserve the model's zero-shot capabilities while improving the overall performance.

This approach offers the additional advantage of reducing both training and inference costs in terms of computation time and memory. Since both the label descriptions and the label encoder $\mathcal{E}_{\text{label}}$ are fixed, the label embeddings $\boldsymbol{z}_j^{\text{pool}}$ only need to be encoded once. In contrast, traditional fine-tuning methods like DEXML and DPR require forward and backward operations on all label descriptions at each training step, making it computationally expensive, especially with a large number of labels. By eliminating this overhead, our method allows for the use of larger pre-trained models, which typically offer higher-quality embeddings and improved performance, while using fewer computational resources than previous methods.

4.3 One-sided Fine-tuned Label-wise Attention Network (OF-LAN)

4.3.1 Label-wise Attention Networks (LAN)

Label-wise Attention Networks (LAN) (Mullenbach et al., 2018) are designed to enhance multi-label text classification by allowing the model to focus on relevant parts of a document for each label. The architecture of LAN is shown in Fig. 5(c). For a sequence of token embedding $\boldsymbol{x}_i$ obtained in Eq. (1), the label-aware document embedding is computed using scaled dot-product attention:

$$\begin{aligned}\boldsymbol{x}_{ij}^{\text{attn}} &= \text{Attention}(\boldsymbol{z}_j^{\text{attn}}, \boldsymbol{x}_i, \boldsymbol{x}_i) \\ &= \text{softmax}\left(\frac{\boldsymbol{z}_j^{\text{attn}} \boldsymbol{x}_i^\top}{\sqrt{d}}\right) \boldsymbol{x}_i\end{aligned}$$

where d is the dimension of document embeddings, and $\boldsymbol{z}_j^{\text{attn}}$ is a label-specific trainable parameter for the j-th label. Then the prediction for the j-th label is made using a label-specific linear classifier:

$$\hat{y}_{ij} = \boldsymbol{w}_j^{\text{out}\top} \boldsymbol{x}_{ij}^{\text{attn}} + b_j^{\text{out}}$$

where $\boldsymbol{w}_j^{\text{out}}$ and b_j^{out} are label-specific trainable parameters.

The advantage of LAN lies in its ability to focus on relevant parts of the document for each label. This dynamic focus enhances the model's ability to make accurate predictions. However, a limitation of LAN is that it learns the label-specific parameters $\boldsymbol{z}_j^{\text{attn}}$, $\boldsymbol{w}_j^{\text{out}}$, and b_j^{out} from scratch during training, making it ineffective for zero-shot labels, as there are no training instances for these unseen labels.

4.3.2 One-sided Fine-tuned Label-wise Attention Network (OF-LAN)

To extend LAN to zero-shot settings, we propose the One-sided Fine-tuned Label-wise Attention Network (OF-LAN), illustrated in Fig. 5(d). Unlike the original LAN, where attention weights and linear classifiers are learned from scratch, OF-LAN leverages the fixed label embeddings $\boldsymbol{z}_j^{\text{pool}}$ obtained from Eq. (2) to replace these label-specific attention weights. Specifically, we use a linear transformation to generate the attention weights:

$$\boldsymbol{z}_j^{\text{attn}} = W_{\text{attn}} \boldsymbol{z}_j^{\text{pool}} + b_{\text{attn}}$$

where W_{attn} and b_{attn} are learnable parameters shared across all labels. Then the final prediction $\hat{y}_{ij}$ is computed similarly to OF-DE (Eq. (3)):

$$\hat{y}_{ij} = f(\boldsymbol{x}_{ij}^{\text{attn}}, \boldsymbol{z}_j^{\text{proj}})$$

where f is a similarity function and $\boldsymbol{z}_j^{\text{proj}}$ is the projected label embedding. This approach generates label-specific weights based on label descriptions, rather than learning them independently from scratch. As a result, OF-LAN retains the advantages of LAN in focusing on relevant document sections, while ensuring it can generalize to both seen and unseen labels.

4.4 Auxiliary Self-supervised Training on Label Descriptions

Motivated by self-supervised learning in recommendation systems (Yu et al., 2017; Yao et al., 2021), we introduce an auxiliary self-supervised task on label descriptions to further enhance generalization. The idea is to align the embeddings of label descriptions generated by both the document encoder and label encoder.

Specifically, for each label description Z_k, we treat it as a document and assign it a target label $y_{kj}^{\text{aux}} = \mathbb{I}(k = j)$, where k and j are label indices, and $\mathbb{I}(\cdot)$ is the indicator function. The prediction for the label description Z_k with respect to the j-th label is computed similarly to the prediction for documents. For example, in the case of OF-DE, the output is computed as:

$$\hat{y}_{kj}^{\text{aux}} = f\left(\mathcal{P}\left(\mathcal{E}_{\text{doc}}(Z_k; \boldsymbol{\theta})\right), \boldsymbol{z}_j^{\text{proj}}\right)$$

We incorporate this auxiliary task by optimizing the same objective function as the primary task. Let $\mathcal{L}(y_{ij}, \hat{y}_{ij})$ denote the original loss function, we extend it to include the auxiliary self-supervised task:

$$\sum_i \sum_j \mathcal{L}(y_{ij}, \hat{y}_{ij}) + \lambda \sum_k \sum_j \mathcal{L}(y_{kj}^{\text{aux}}, \hat{y}_{ij}^{\text{aux}}) \quad (4)$$

where λ is a hyperparameter that controls the weight of the auxiliary task.

This self-supervised learning approach allows the document and label encoders to refine their weights using descriptions of zero-shot labels, potentially improving the model's generalization capabilities and its ability to predict these unseen labels.

4.5 Training With or Without Unseen Labels

In real-world scenarios, unseen labels may be unknown during training. Even when they are available, deciding whether to include them in training impacts model performance with a tradeoff between overall accuracy and zero-shot generalization. Since this is a complementary approach to our one-sided fine-tuning, we provide detailed experiments and analysis in Appendix D.

5 Experimental Setup

5.1 Data

We follow the evaluation setup of Chalkidis et al. (2020) and assess our methods on three large-scale ZMTC benchmarks: EURLEX57K (Chalkidis et al., 2019), MIMIC-III (Johnson et al., 2016), and AmazonCat-13K (McAuley and Leskovec, 2013). Detailed statistics for each dataset are shown in Table 1. More information about the datasets are provided in Appendix A.

5.2 Evaluation Metrics

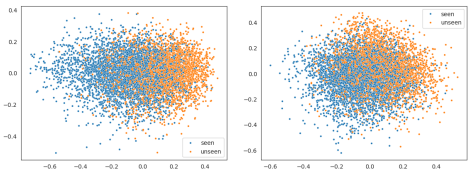
We evaluate both the overall performance and zero-shot performance of our models using the following metrics.

Overall performance. We use NDCG@K (Normalized Discounted Cumulative Gain) and PSP@K (Propensity-Scored Precision) to assess overall performance. NDCG@K is a common ranking metric that treats all labels equally, while PSP@K weights labels based on their frequency, giving more importance to tail labels, as introduced by Jain et al. (2016). For EURLEX57K and AmazonCat-13K, we set $K = 5$, and for MIMIC-III, we set $K = 15$, following the setup from Chalkidis et al. (2020).

Zero-shot performance. To measure zero-shot performance, we use ZSR@K (Zero-Shot Recall), following Yadav et al. (2024). This metric focuses on instances with zero-shot labels and measures how many correct zero-shot labels are predicted in the top-K. We set $K = 100$ since seen labels tend to rank higher than unseen labels. Further evaluations with different K values are available in Appendix E.

5.3 Baselines

We compare our methods against a range of both representative and state-of-the-art approaches.



(a) OF-DE (b) OF-LAN

Figure 7: 2D visualization of label embeddings on the EURLEX dataset for OF-DE, and OF-LAN, using PCA projection.

Figure 8: The progress of zero-shot performance of DEXML, DPR, OF-DE and OF-LAN across datasets during training. The x-axis represents the number of training steps, while the y-axis shows ZSR@100.

Dataset	$\lvert\mathcal{L}\rvert$	$\lvert\mathcal{L}_{\text{unseen}}\rvert$	N_{trn}	N_{val}	N_{tst}	L_{AVG}
EURLEX57K	4,271	163	45,000	6,000	6,000	5.07
MIMIC-III	8,921	443	47,723	1,631	3,372	15.45
AmazonCat-13K	13,330	579	948,992	237,247	306,782	5.04

Table 1: Dataset statistics. $|\mathcal{L}|$: number of labels, $|\mathcal{L}_{\text{unseen}}|$: number of unseen labels, N_{trn}: number of training instances, N_{val}: number of validation instances, N_{tst}: number of test instances, L_{AVG}: average number of labels per instance across the entire dataset.

Unsupervised approaches. We include TF-IDF (Salton and Buckley, 1988), a traditional statistical approach, and pre-trained models like MPNet (Song et al., 2020) and SFR (Meng et al., 2024). Additionally, we evaluate RTS (Zhang et al., 2022), a state-of-the-art self-supervised fine-tuning method designed to improve zero-shot performance.

Supervised approaches. We include OVA-CLF (Devlin et al., 2019), a one-vs-all classifier with a pre-trained encoder, and DPR (Karpukhin et al., 2020), which uses asymmetric dual encoders. We also compare against DEXML (Gupta et al., 2024), a state-of-the-art method that fine-tunes symmetric dual encoders with advanced loss functions. More details of these methods are provided in Appendix B.

5.4 Implementation Details

We follow the setup from Gupta et al. (2024) to use DistilBERT (Sanh, 2019) as the document encoder for all supervised methods, including ours, with cosine similarity as the similarity function. For the label encoder, our methods leverage the advantage of using larger models, whereas other methods typically use the same model for both the document and label encoders. Therefore, we utilize the 7B-parameter SFR-Embedding-2 (Meng et al., 2024) for the label encoder in our approach. The remaining implementation details are provided in Appendix C.

6 Experiment Results

6.1 Main Result

The results presented in Table 2 show our proposed methods, OF-DE and OF-LAN, demonstrate significant advantages over both unsupervised and supervised baselines. Compared to the unsupervised methods, both models achieve much higher overall performance, with OF-LAN obtaining the best NDCG and PSP scores on EURLEX57K and MIMIC-III. This highlights the effectiveness of fine-tuning to learn the task-specific relationship between documents and labels in MLTC tasks.

In terms of zero-shot performance, OF-DE and OF-LAN greatly outperform the supervised baselines, which struggle with unseen labels. Both models achieve substantial improvements in zero-shot recall, while the supervised baselines show zero or little ability to predict zero-shot labels. This demonstrates the strength of our one-sided fine-tuning approach in handling zero-shot scenarios without sacrificing overall accuracy.

6.2 Zero-shot Performance Analysis

We explain the improved zero-shot capabilities of our methods in two key aspects.

Label embedding distribution. Fig. 7 presents a 2D visualization of the label embeddings generated by our proposed methods. In contrast to the embeddings produced by supervised approaches (Fig. 4), where unseen label embeddings collapse into a small, isolated region, our methods maintain a more even distribution of unseen labels, allowing them to mix with seen labels. This suggests that our methods preserve better generalization for unseen labels, as the more uniform embedding distribution facilitates improved zero-shot label prediction.

Method	EURLEX57K			MIMIC-III			AmazonCat-13K		
	NDCG@5	PSP@5	ZSR@100	NDCG@15	PSP@15	ZSR@100	NDCG@5	PSP@5	ZSR@100
Unsupervised									
TF-IDF	29.71	23.36	38.76	8.62	7.95	**26**	5.89	7.11	26.83
MPNet	6.75	6.77	29.78	5.2	4.29	12.42	13.64	20.66	76.15
SFR	6.89	5.45	17.98	3.83	3.23	9.33	13.05	21.16	**79.13**
RTS	22.03	18.6	**48.88**	10.33	8.88	21.54	12.63	19.38	75.23
Supervised									
OVA-CLF	82.33	65.95	0	52.4	31.57	0	**87**	74.64	0
DPR	82.99	67.18	0	56.19	34.69	0	86.92	75.23	4.59
DEXML	82.76	67.49	0	55.88	34.6	0	86.57	**75.51**	22.02
Proposed Method									
OF-DE	82.54	66.33	44.94	55.12	33.06	9.43	86.74	73.29	58.94
OF-LAN	**83.01**	**67.85**	43.26	**57.18**	**37.89**	16.25	86.35	74.25	64.45

Table 2: Performance comparison of baselines and proposed methods on three large-scale ZMTC datasets. The best results are highlighted in **bold**. Our proposed methods (OF-DE and OF-LAN) demonstrate competitive overall performance and superior zero-shot performance compared to state-of-the-art supervised methods.

Method	EURLEX57K			MIMIC-III			AmazonCat-13K		
	NDCG@5	PSP@5	ZSR@100	NDCG@15	PSP@15	ZSR@100	NDCG@5	PSP@5	ZSR@100
OF-DE ($\lambda = 0$)	82.56	66.01	32.58	55.45	33.26	7.55	86.8	72.61	56.42
OF-DE ($\lambda = 1$)	82.54	66.33	**44.94**	55.12	33.06	9.43	86.74	73.29	58.94
OF-LAN ($\lambda = 0$)	83	67.44	38.2	56.6	37.38	12.79	**87.02**	**74.87**	30.96
OF-LAN ($\lambda = 1$)	**83.01**	**67.85**	43.26	**57.18**	**37.89**	**16.25**	86.35	74.25	**64.45**

Table 3: Comparison of OF-DE and OF-LAN, with and without self-supervised loss.

Method	Label Encoder	NDCG@5	PSP@5	ZSR@100
OF-DE	BGE-small (33M)	80.9	63.73	42.13
	BGE-base (109M)	81.68	64.94	**48.88**
	BGE-large (335M)	81.9	65.3	44.94
	SFR (7B)	**82.54**	**66.33**	44.94
OF-LAN	BGE-small (33M)	82.22	66.96	39.89
	BGE-base (109M)	82.3	67.36	35.39
	BGE-large (335M)	82.62	67.51	41.01
	SFR (7B)	**83.01**	**67.85**	**43.26**

Table 4: Performance comparison of OF-DE and OF-LAN on EURLEX57K using different scales of label encoders. The numbers following the label encoder names indicate the number of parameters in each model.

Progress of zero-shot performance. Fig. 8 shows the progress of zero-shot recall (ZSR@100) across different datasets. The ZSR of our methods, OF-DE and OF-LAN, steadily improves throughout training, indicating that the models become increasingly capable of predicting unseen labels. In contrast, DEXML and DPR show a gradual decline in ZSR, eventually dropping to zero. This highlights the effectiveness of OF-DE and OF-LAN in maintaining or enhancing zero-shot generalization while improving overall performance, by preserving the quality and distribution of label embeddings during training.

6.3 Ablation Study

Impact of self-supervised training on label descriptions. Table 3 demonstrates the effect of incorporating the self-supervised auxiliary task. The results indicate that both OF-DE and OF-LAN exhibit reasonable zero-shot capabilities even without the auxiliary learning on label descriptions. However, adding the self-supervised auxiliary loss significantly boosts zero-shot performance (ZSR) across all datasets, with minimal impact on overall performance (NDCG and PSP). These findings suggest that the self-supervised auxiliary task is an effective and safe addition for enhancing zero-shot performance without sacrificing overall accuracy.

Impact of label encoder's scales. In both OF-DE and OF-LAN, label encoders $\mathcal{E}_{\text{label}}$ play important roles to ensure the label embedding quality, leading to performance. Therefore we analyze the impact of encoder's scales. We evaluate BGE (Xiao et al., 2024) models at three scales—small (33M), base (109M), and large (335M)—and a larger SFR model (7B). Table 4 shows that increasing the size of the label encoder generally improves overall

performance (NDCG@5, PSP@5) for both OF-DE and OF-LAN. Larger models like SFR achieve the best results in these metrics, highlighting the benefits of more comprehensive label representations. However, the impact on zero-shot performance (ZSR@100) is less straightforward. For OF-DE, the highest ZSR is achieved with the BGE-base model, outperforming larger models like BGE-large and SFR. This indicates that while larger encoders improve overall accuracy, zero-shot performance gains may not scale linearly with model size.

7 Conclusion

In this work, we addressed the challenge of preserving zero-shot capabilities in supervisedly fine-tuned models while maintaining competitive overall performance for ZMTC. We identified the issue of label embedding collapse during supervised training and proposed OF-DE to mitigate this problem by freezing the label encoder. Additionally, we introduced an advanced OF-LAN and a self-supervised auxiliary loss, both of which further enhance zero-shot performance. Experiments demonstrated that our approach significantly improves zero-shot performance while sustaining strong overall accuracy, offering a balanced solution to the trade-off.

8 Limitation

Our methods are applicable to extreme multi-label learning; however, due to resource constraints, we focus on moderate-scale benchmarks, which we believe are sufficient to demonstrate the effectiveness of our approach. Additionally, given the vast number of existing works in the field, we selected only the most representative methods and those critical for showing our approach's effectiveness.

Acknowledgements

This work was supported in part by the National Taiwan University Center for Data Intelligence via NTU-113L900901 and the Ministry of Science and Technology in Taiwan via 113-2628-E-002-003 and 110-2221-E-002-115-MY3.

References

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019a. Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference*.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019b. The unfairness of popularity bias in recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems*.

Pranjal Aggarwal, Ameet Deshpande, and Karthik R Narasimhan. 2023. Semsup-xc: semantic supervision for zero and few-shot extreme classification. In *Proceedings of the International Conference on Machine Learning*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kunal Dahiya, Nilesh Gupta, Deepak Saini, Akshay Soni, Yajun Wang, Kushal Dave, Jian Jiao, Gururaj K, Prasenjit Dey, Amit Singh, et al. 2023a. Ngame: Negative mining-aware mini-batching for extreme classification. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*.

Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. 2021. Deepxml: A deep extreme multi-label learning framework applied to short text documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Kunal Dahiya, Sachin Yadav, Sushant Sondhi, Deepak Saini, Sonu Mehta, Jian Jiao, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2023b. Deep encoders with auxiliary parameters for extreme classification. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Nilesh Gupta, Sakina Bohra, Yashoteja Prabhu, Saurabh Purohit, and Manik Varma. 2021. Generalized zero-shot extreme multi-label learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Nilesh Gupta, Fnu Devvrit, Ankit Singh Rawat, Srinadh Bhojanapalli, Prateek Jain, and Inderjit S Dhillon. 2024. Dual-encoders for extreme multi-label classification. In *Proceedings of the International Conference on Learning Representations*.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Vidit Jain, Jatin Prakash, Deepak Saini, Jian Jiao, Ramachandran Ramjee, and Manik Varma. 2023. Renee: End-to-end training of extreme classification models. *Proceedings of Machine Learning and Systems*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*.

Rui Meng, , Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021a. Decaf: Deep extreme classification with label features. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021b. Eclare: Extreme classification with label graph correlations. In *Proceedings of the Web Conference 2021*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *Proceedings of the International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

V Sanh. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

World Health Organization WHO. 2004. *International Statistical Classification of Diseases and related health problems: Alphabetical index*. World Health Organization.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit S. Dhillon. 2022. Extreme zero-shot learning for extreme text classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sachin Yadav, Deepak Saini, Anirudh Buvanesh, Bhawna Paliwal, Kunal Dahiya, Siddarth Asokan, Yashoteja Prabhu, Jian Jiao, and Manik Varma. 2024. Extreme meta-classification for large-scale zero-shot retrieval. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. 2021. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*.

Hsiang-Fu Yu, Hsin-Yuan Huang, Inderjit Dhillon, and Chih-Jen Lin. 2017. A unified algorithm for one-cass structured matrix factorization with side information. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tianyi Zhang, Zhaozhuo Xu, Tharun Medini, and Anshumali Shrivastava. 2022. Structural contrastive representation learning for zero-shot multi-label text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Yaxin Zhu and Hamed Zamani. 2024. ICXML: An in-context learning framework for zero-shot extreme multi-label classification. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024*.

A Dataset

EURLEX57K (Chalkidis et al., 2019) dataset consists of legislative documents annotated with one or more labels from the EUROVOC [1] thesaurus. The dataset includes $4,271$ unique concepts, of which 163 are designated as zero-shot labels that appear only in the test set.

MIMIC-III (Johnson et al., 2016) consists of around 52k de-identified clinical discharge records, where each record is assigned to multiple ICD-9 (WHO, 2004) codes. The ICD-9 system includes $8,771$ labels, of which 443 are unseen in the training set.

AmazonCat-13K (McAuley and Leskovec, 2013) contains product descriptions from Amazon, with labels representing $13,330$ product categories. These labels are organized in an 8-level hierarchy, where all ancestors of a positive label are also assigned as positive. There are 579 zero-shot labels that appear exclusively in the test set.

B Baseline

Unsupervised methods include three no fine-tuning approaches and one self-supervised approach.

- **TF-IDF** (Salton and Buckley, 1988): A classical retrieval method using scaled term frequency vectors to represent documents and labels.
- **MPNet** (Song et al., 2020): A BERT-based pre-trained model which has been fine-tuned for retrieval tasks from Sentence-BERT (Reimers and Gurevych, 2019) family.
- **SFR** (Meng et al., 2024): A large-scale text embedding model based on Mistral-7B (Jiang et al., 2023) fine-tuned by Salesforce Research, ranked 4th on the Massive Text Embedding Benchmark (MTEB, Muennighoff et al., 2022).
- **RTS** (Zhang et al., 2022): A state-of-the-art self-supervised pre-training method specifically designed for ZMTC. It enhances embedding quality by leveraging document structures. It randomly splits documents into chunks, and creates title-document, document-document, and label-label pairs for contrastive learning.

Supervised methods include the one-versus-all classifier and dual encoder architectures.

- **OVA-CLF** (Devlin et al., 2019): A one-versus-all classifier that uses a pre-trained encoder followed by a linear layer for each label.
- **DPR** (Karpukhin et al., 2020): A retrieval-based method that employs asymmetric dual encoders to independently encode documents and labels, making it effective for retrieval tasks.
- **DEXML** (Gupta et al., 2024): A state-of-the-art extreme multi-label classification model that fine-tunes symmetric dual encoders using advanced loss functions for better performance in large-scale tasks.

C Implementation Detail

Training is carried out over 50 epochs for the EURLEX57K and MIMIC-III datasets, and 30 epochs for AmazonCat-13K. For the self-supervised auxiliary task, we set $\lambda = 1$ for Eq. (4) and sample mini-batches of label descriptions with the same size as those used for the documents. Following the practices recommended by Mosbach et al. (2021), we fine-tune the encoder with the AdamW optimizer (Loshchilov and Hutter, 2019), using a weight decay of 0.01, a dropout rate of 0.1, a warm-up rate of 0.1, and linear learning rate decay. We tune hyperparameters with learning rates from $\{2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$. The hyperparameters are selected based on NDCG@K, evaluated on the original validation set of EURLEX57K and an 80/20 split for AmazonCat-13K. The validation set for MIMIC-III is processed following the process in Mullenbach et al. (2018). The batch size is set to 32 per GPU, totaling 256 across 8 NVIDIA V100 GPUs.

D Effect of Removing Unseen Labels from Training

Excluding unseen labels from training is an alternative strategy to mitigate overfitting to seen labels. Table 5 compares models trained with and without zero-shot labels. Across all methods, excluding zero-shot labels improves ZSR@100, indicating better generalization to unseen labels. This effect

[1] https://eur-lex.europa.eu/browse/eurovoc.html

is similar to the improvements observed with one-sided fine-tuning. However, it comes at the cost of slightly lower NDCG@K and PSP@K, suggesting reduced ranking performance for seen labels.

This trade-off depends on application needs. If prioritizing rare or emerging labels, training without zero-shot labels can be beneficial. Conversely, if ranking across all labels is crucial, including zero-shot labels in training helps. The results show that this strategy is complementary to our one-sided fine-tuning.

E Results with Additional Metrics

We show evaluations of baselines and the proposed methods with additional metrics in Table 6.

Method	EURLEX57K			MIMIC-III			AmazonCat-13K		
	NDCG@5	PSP@5	ZSR@100	NDCG@15	PSP@15	ZSR@100	NDCG@5	PSP@5	ZSR@100
DPR	82.99	67.18	0	56.19	34.69	0	86.92	75.23	4.59
DPR w/o $\mathcal{L}_{\text{unseen}}$	82.28	66.53	30.9	56.47	34.82	2.52	86.77	75.07	54.13
DEXML	82.76	67.49	0	55.88	34.6	0	86.57	75.51	22.02
DEXML w/o $\mathcal{L}_{\text{unseen}}$	81.98	66.59	38.2	55.25	34.48	8.81	86.6	75.56	55.5
OF-DE ($\lambda = 0$)	82.56	66.01	32.58	55.45	33.26	7.55	86.8	72.61	56.42
OF-DE ($\lambda = 0$) w/o $\mathcal{L}_{\text{unseen}}$	82.41	66.11	38.2	54.27	32.28	10.59	86.71	72.81	58.72
OF-LAN ($\lambda = 0$)	83	67.44	38.2	56.6	37.38	12.79	87.02	74.87	30.96
OF-LAN ($\lambda = 0$) w/o $\mathcal{L}_{\text{unseen}}$	82.64	67.18	50	55.85	36.21	13.94	86.29	73.67	61.24
OF-DE ($\lambda = 1$)	82.54	66.33	44.94	55.12	33.06	9.43	86.74	73.29	58.94
OF-DE ($\lambda = 1$) w/o $\mathcal{L}_{\text{unseen}}$	82.27	66.01	48.88	54.08	32.17	11.01	86.75	73.33	59.63
OF-LAN ($\lambda = 1$)	83.01	67.85	43.26	57.18	37.89	16.25	86.35	74.25	64.45
OF-LAN ($\lambda = 1$) w/o $\mathcal{L}_{\text{unseen}}$	82.69	67.42	50.56	56.46	36.75	19.5	86.34	74.29	65.83

Table 5: Results of baselines and our proposed methods with and without considering unseen labels during training.

EURLEX57K									
Method	NDCG@1	NDCG@3	NDCG@5	PSP@1	PSP@3	PSP@5	ZSR@10	ZSR@50	ZSR@100
TF-IDF	44.22	32.73	29.71	29.28	24.37	23.36	15.17	30.34	38.76
MPNET	7.92	6.73	6.75	6.11	6.22	6.77	10.11	19.1	29.78
SFR	9.85	6.97	6.89	6.52	5.28	5.45	4.49	13.48	17.98
RTS	29.88	23.27	22.03	20.01	18.56	18.6	**21.91**	**41.01**	**48.88**
OVA-CLF	89.17	84.89	82.33	47.93	60.32	65.95	0	0	0
DPR	**90.2**	**85.76**	82.99	**51.78**	62.23	67.18	0	0	0
DEXML	89.18	85.47	82.76	51.17	**62.62**	67.49	0	0	0
OF-DE	89.53	85.51	82.54	49.3	61.17	66.33	4.49	33.71	44.94
OF-LAN	89.2	85.69	**83.01**	50.93	62.6	**67.85**	10.11	31.46	43.26
MIMIC									
Method	NDCG@3	NDCG@5	NDCG@15	PSP@3	PSP@5	PSP@15	ZSR@10	ZSR@50	ZSR@100
TF-IDF	11.33	10.13	8.62	7.13	7.2	7.95	**8.07**	**19.08**	**26**
MPNET	6.8	6.29	5.2	3.45	3.77	4.29	1.89	7.86	12.42
SFR	4.61	4.23	3.83	2.46	2.54	3.23	2.52	7.86	9.33
RTS	13.45	12.36	10.33	7.43	8.05	8.88	9.43	15.25	21.54
OVA-CLF	65.97	62.8	52.4	24.87	27.82	31.57	0	0	0
DPR	**70.99**	**67.55**	56.19	27.9	30.83	34.69	0	0	0
DEXML	70.29	67.09	55.88	27.61	30.72	34.6	0	0	0
OF-DE	70.27	66.9	55.12	26.4	29.55	33.06	0.63	4.72	9.43
OF-LAN	67.59	65.85	**57.18**	**28.24**	**31.93**	**37.89**	0	10.69	16.25
AmazonCat-13K									
Method	NDCG@1	NDCG@3	NDCG@5	PSP@1	PSP@3	PSP@5	ZSR@10	ZSR@50	ZSR@100
TF-IDF	8.63	6.43	5.89	7.9	7.08	7.11	15.37	24.54	26.83
MPNET	18.93	14.6	13.64	23.87	20.75	20.66	55.5	71.56	76.15
SFR	19.9	14.29	13.05	27.59	22.09	21.16	**61.93**	**73.85**	**79.13**
RTS	18	13.67	12.63	22.75	19.56	19.38	55.28	70.64	75.23
OVA-CLF	88.53	**87.69**	**87**	54.85	67.5	74.64	0	0	0
DPR	87.81	87.47	86.92	55.49	68.64	75.23	0	1.83	4.59
DEXML	87.14	87.02	86.57	**57.67**	**69.63**	**75.51**	3.21	15.6	22.02
OF-DE	**88.58**	87.63	86.74	48.87	64.29	73.29	15.37	47.71	58.94
OF-LAN	87.77	86.92	86.35	51.72	65.76	74.25	18.81	55.28	64.45

Table 6: Results of all baselines and our proposed methods on EURLEX57K, MIMIC-III, and AmazonCat-13K, evaluated with multiple performance metrics.