Calibrating Language Models for Neural Ranking under Noisy Supervision with Relaxed Labels

Arnab Sharma
Heinz Nixdorf Institute
Paderborn University
Paderborn, Germany
arnab.sharma@upb.de

Daniel Vollmers
Heinz Nixdorf Institute
Paderborn University
Paderborn, Germany
daniel.vollmers@upb.de

Axel-Cyrille Ngonga Ngomo
Heinz Nixdorf Institute
Paderborn University
Paderborn, Germany
axel.ngonga@upb.de

Abstract

In recent years, we have seen an increased usage of neural ranking models in the information retrieval domain. Although language model-based rankers have shown significant progress in performing ranking tasks, little to no work has addressed the issue of fine-tuning them in the presence of label noise in the training data. In a general learning setting, training models in the presence of noisy labeled data is studied extensively. To this end, *confidence calibration* approaches have shown significant promise; however, their usage in training neural ranking models is relatively less studied. In this work, we address this gap by adapting and analyzing regularization-based calibration approaches to reduce the effect of label noise in ranking tasks. Specifically, we study *label relaxation* in neural ranking models. We demonstrate the effectiveness of this approach by performing extensive evaluations comparing the label relaxation approach to standard loss functions. Additionally, we analyze the calibration error associated with the loss functions. After evaluating on five different noise levels, two different ranking models, and four diverse ranking datasets, the results suggest that label relaxation can improve the performance of the ranking models under noisy labels. Furthermore, we find that label relaxation reduces calibration error, although it suggests a better metric to be used for neural ranking models.

1 Introduction

The advancements of language models have enabled their rapid usage in various application domains. One of such prominent application areas is *neural ranking* wherein the task is to estimate the relevance of several candidate documents or entities based on their relevance to the given *query* (Reimers and Gurevych, 2019; Nogueira and Cho, 2019), which is typically a question presented in a natural language form. With the recent progress in NLP domains, models like BERT (Devlin et al., 2019) have achieved significant progress in capturing the semantic contextual information for a given query. Existing works focus on improving the ranking tasks considering several aspects of the learning framework (Sil et al., 2018; Yamada et al., 2020; Ganea and Hofmann, 2017; Fang et al., 2019; Zhang et al., 2020). However, to the best of our knowledge, only a few works considered approaches to develop *robust* ranking models when noisy labels are prevalent in the training data.

Label noise in the training data for ranking tasks can be caused due to several reasons. For instance, in a question answering dataset, noise can stem from distant supervision, weakly supervised data generation, bad annotations, among other reasons. Such noise can essentially lead to the generation of models with degraded generalization and unstable predictions (Liu and Tao, 2016; Natarajan et al., 2013; Patrini et al., 2017). This issue becomes particularly critical in ranking tasks, where the quality of predictions directly impacts the rank order of documents or entities, thereby affecting the overall effectiveness of the system. Furthermore, studying the risks associated with label noise in ranking models is also important, as improper handling of noise can lead to misleading rankings and reduced model reliability. One of the few works in the NLP domain by Zhu et al. (2022) studied the robustness of the BERT model and showed that in sentence classification tasks, weakly supervised noise can severely degrade the performance of the model. In classification and general learning settings, this problem has been tackled often by using several types of model calibration approaches by Zhu et al. (2021); Wei and Liu (2021); Ding et al. (2021); Cheng and Vasconcelos (2022); Ghosh et al. (2022); Moon et al. (2020); Ma and Blaschko (2021); Liu et al. (2022); Lienen and Hüllermeier (2021, 2024). These approaches typically work by ensuring that the *confidence* of the underlying model in predicting an

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 259–272

input instance should also reflect the true likelihood of the prediction. In other words, the model should not confidently predict wrong labels, and in contrast, when predicting the correct labels, it should exhibit sufficient confidence. The idea is to calibrate the overconfident models, which are vulnerable to memorizing incorrect labels (Guo et al., 2017). Label smoothing (Szegedy et al., 2016) is considered a standard approach, wherein the idea is to distribute a specific amount (decided based on a hyperparameter) of probability mass taken from the actual label to all the other labels. Although label smoothing can be quite effective, it still relies on precise probabilistic labels, which might degrade the generalization performance (Li et al., 2020). Therefore, Lienen and Hüllermeier (2021) proposed *label relaxation*, which considered a set of candidate distributions, instead of a single smoothed distribution. Label relaxation essentially replaces fixed (and possibly incorrect) label distributions with sets of plausible distributions, thereby allowing the learner to learn a bounded range of acceptable target labels.

In this paper, we tackle label noise in order to develop robust ranking models in the fine-tuning step. We consider two different directions, considering model calibration techniques. Firstly, we introduce label relaxation into the ranking paradigm as a principled approach to fine-tune models under noisy conditions. More specifically, considering the pairwise ranking loss, we integrate relaxation in several widely used neural ranking models. Then we compare the performance of two different calibration approaches, i.e., smoothing with relaxation, to gain some initial insights into which approach performs better. Secondly, by analyzing the calibration error, we aim to understand how well the models' confidence reflects their true performance under noisy conditions. Thus, we assess the associated risks of poor confidence calibration, which can lead to suboptimal ranking decisions. We model the label noise in the ranking tasks by considering a *proximity-aware* approach. Experimental results considering these two different calibration approaches, 5 different noise levels, 4 diverse datasets, and two ranking models suggest the potential of relaxation under label noise in fine-tuning ranking models. Our contributions can be summarized as,

- We introduce label relaxation to perform calibration for ranking models under the presence of label noise.
- We formally define the relaxation considering the pairwise ranking loss.
- We evaluate the performance of label relaxation considering 5 different noise levels.
- We give a comparative analysis comparing label relaxation to the standard calibration approach, label smoothing.
- We analyze the calibration error to understand the risks associated with two calibration approaches in the presence of label noise.
- We make the code publicly available [1].

2 Related Work

Ranking models As mentioned beforehand, with the advancement of language models, we have seen significant progress in the domain of neural ranking (Reimers and Gurevych, 2019; Nogueira and Cho, 2019; Déjean et al., 2024; Zhang et al., 2022; Wu et al., 2020a). One of the first works was Sentence-BERT (Reimers and Gurevych, 2019), which adapted the BERT architecture into a Siamese network to produce sentence-level embeddings. Nogueira and Cho (2019) extended this idea by further showing that BERT-based models could be fine-tuned specifically for passage re-ranking. This has shown substantial improvements in retrieval performance. Subsequent work has continued to explore more scalable and generalizable ranking solutions. For example, Wang et al. (2022) introduced a family of embedding models trained with contrastive learning on massive collections of text pairs. *Cross encoders* (Déjean et al., 2024) are shown to outperform the previous approaches in re-ranking tasks at the cost of a high training time.

Calibrated Loss Calibration refers to the alignment between the model's predicted confidence and the actual likelihood of correctness. A perfectly calibrated model assigns a probability of 0.7 to a prediction if, on average, 70% of such predictions are correct. There exist two categories of approaches that perform model calibration, (i) *post-hoc* and (ii) *regularization-based*. In order to perform calibration, post hoc approaches adjust the output predictions (Cheng and Vasconcelos, 2022; Wei et al., 2022; Hebbalaguppe et al.,

[1] https://github.com/dice-group/RobustRanking/tree/label-relaxed-ranking

2022). However, this requires additional validation on held-out datasets. Furthermore, this approach assumes the training and test distributions to be the same, which often is not. Regularization-based approaches do not require any extra data and perform calibration during the training step while computing the loss (Cheng and Vasconcelos, 2022; Wei et al., 2022; Hebbalaguppe et al., 2022). Label smoothing is often used as a standard technique to soften the hard target labels by redistributing the probability mass to non-target labels (Szegedy et al., 2016; Müller et al., 2019). However, typical smoothing distributes the probability mass uniformly. There exist approaches that essentially follow more advanced approaches, such as bootstrapping techniques (Reed et al., 2015), wherein a self-supervised approach is used to distribute the probability mass. Self-distillation and model distillation approaches also follow a similar approach by replacing the hard labels with the soft ones from the *teacher* model (Yun et al., 2020; Zhang et al., 2019). Although in typical classification settings such approaches have been extensively studied, in the NLP domain, this is relatively less explored. Huang et al. (2024) introduced confidence-aware label smoothing for alignment tasks considering language models and have shown the potential of the calibration approaches. Kobyzev et al. (2023) also showed the potential of several calibrated approaches in fine-tuning language models.

Note that we consider the idea of label relaxation introduced by Lienen and Hüllermeier (2021) wherein a single fixed target distribution is replaced with a set of candidate probability distributions. Another work by Kim et al. (2021) proposed relaxed labels in metric learning, which relaxes binary pairwise relation labels by replacing them with continuous similarity weights from a source embedding space. Alike our work, Purpura et al. (2022) also study learning to rank from relevance judgment distributions. They use KL divergence to align model predictions with empirical distributions, thereby directly capturing inter-annotator disagreement. Note that we assume that such distributions are not consistently available across ranking datasets. Instead, we propose label relaxation, which defines a credal set of admissible label distributions. This approach allows us to model epistemic uncertainty and mitigate label noise without requiring multiple annotations per query. To the best of our knowledge, this is the first work that studies calibration in this context.

3 Calibration in Ranking Model

In our work, we consider two different ranking approaches, both of which fall under the category of bi-encoder models. More specifically, we use the pre-trained BERT (Wu et al., 2020a) and the E5 (Wang et al., 2022) models.

E5 model (Wang et al., 2022) encodes both the query and candidate entities using a language model, producing dense vector embeddings. These token-level embeddings are then averaged via a pooling layer to obtain fixed-size vectors. Finally, a scoring function computes a probability score $\hat{y} \in [0, 1]$, reflecting the likelihood that the candidate entity is the correct match for the query.

BERT model follows a similar approach to E5 model in performing ranking tasks. However, the only difference is that BERT is pre-trained to perform binary relevance classification tasks between a query and a document (Devlin et al., 2019), whereas in contrast, E5, is additionally pre-trained on several ranking tasks. Next, we formalize the ranking task and subsequently define the term label relaxation in this context.

Note that the document ranking step often consists of two different steps, namely retrieval and ranking. In this work, we only consider the ranking stage. In a typical supervised ranking setup, each training sample consists of a query q and a set of candidate documents defined as $\mathcal{D}_q = \{d_1, d_2, \ldots, d_K\}$, with only one document labeled as relevant (Wang et al., 2022; Zhang and Braun, 2024; Tran et al., 2024). Herein, we assume the set of candidate documents is already correctly retrieved by a retrieval model. Typically, within a training step, a batch of queries and corresponding documents are presented, wherein the size of the batch is determined by the training configuration. For a batch consisting of N queries and K candidates per query, we define the label matrix $Y \in \{0, 1\}^{N \times K}$ as follows.

$$Y_{i,j} = \begin{cases} 1 & \text{if } d_j \in \mathcal{R}_{q_i} \\ 0 & \text{otherwise.} \end{cases}$$

Herein $\mathcal{R}_{q_i} \subseteq \mathcal{D}_{q_i}$ is the set of relevant documents for query q_i, typically of cardinality 1. Let us assume the ranking model as f, then it produces a score $f(q_i, d_j)$ for candidate d_j to query q_i. In the ranking tasks, the goal is to ensure that relevant documents are scored higher than non-relevant ones. To achieve this, the pairwise ranking loss is often

used herein. Let (q_i, d_{j^+}, d_{j^-}) denote a training triplet, where d_{j^+} is a document relevant to query q_i, and d_{j^-} is a non-relevant (or less relevant) document, i.e., $d_{j^-} \in \mathcal{D}_{q_i} \setminus \mathcal{R}_{q_i}$. Since we adopt an *in-batch negative sampling* strategy, we have a training batch containing N queries, each associated with K candidate documents (one relevant and $K-1$ non-relevant). For every query q_i, the model compares the relevant document d_{j^+} against all $K-1$ non-relevant candidates in the batch. The pairwise ranking loss scores a relevant document so that it exceeds that of a non-relevant one by $\gamma > 0$, a *defined margin*. This can be defined as,

$$\mathcal{L}_{\text{PR}} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq j^+}}^{K} \max\big\{0, \gamma - f(q_i, d_{j^+}) + f(q_i, d_{j^-})\big\} \tag{1}$$

Here, j^+ denotes the index of the relevant document among the K candidates for q_i. This loss penalizes cases where a non-relevant document scores too closely or higher than the relevant one.

Label Smoothing is a regularization technique that softens target labels to mitigate overconfidence (Szegedy et al., 2016; Müller et al., 2019). Rather than encoding the correct document as a one-hot vector, label smoothing redistributes a small fraction of the probability mass across all other candidates. Formally, the smoothed label distribution $\tilde{Y} \in [0,1]^{N \times K}$ can be defined as follows.

$$\tilde{Y}_{i,j} = \begin{cases} 1 - \varepsilon & \text{if } j = j^+ \\ \frac{\varepsilon}{K-1} & \text{otherwise} \end{cases}$$

As mentioned previously, we consider in-batch pairwise training; therefore, we use the smoothed score $\tilde{Y}_{i,j}$ in place of a hard label of 1 in the margin-based loss, resulting in the label-smoothed pairwise ranking loss as follows.

$$\mathcal{L}_{\text{LS}}^{pair} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq j^+}}^{K} \tilde{Y}_{i,j^+} \max\big(0, \gamma - f(q_i, d_{j^+}) + f(q_i, d_{j^-})\big) \tag{2}$$

Label Relaxation in Pairwise Ranking unlike label smoothing, which redistributes the probability mass of the target label uniformly, label relaxation replaces the target with a set of *plausible distributions* that reflect *epistemic uncertainty* (Lienen and Hüllermeier, 2021). This can help to reduce the uncertainty regarding the correct label. Label relaxation introduces a *relaxed set* of acceptable target distributions parameterized by $\alpha \in [0,1]$. We define Q^α as the set of all relevance probability distributions p satisfying $p(+) \geq 1 - \alpha$ and $p(-) \leq \alpha$. While Q^α is a set, in our implementation, we instantiate it via a canonical representative distribution p_r for loss computation as $Q^\alpha = \big\{p \in \Delta^2 : p(+) \geq 1 - \alpha, \quad p(-) \leq \alpha\big\}$.

This set essentially defines that the relevant document should be preferred with high probability; this does not pertain to a specific numeric value; rather, we allow the model to match any distribution within Q^α. The model can then generate any label that falls inside this plausible region, without penalizing it for deviations that are within the acceptable uncertainty bounds. The relaxation parameter $\alpha \in [0,1]$ controls the degree of permissible deviation from the one-hot target. We select α via validation set performance for each dataset.

Next, we apply the KL divergence on the predicted scores and the distribution Q^α. Note that, since $f(q_i, d_j)$ is a relevance score, for KL divergence we need to first convert it into a probability distribution over candidates using a softmax normalization, let us call this $\hat{p}_i(j)$. Afterwards, the label relaxation loss compares the predicted distribution $\hat{p}_i(j)$ with the distribution Q^α, denoted as p_r. Since we have two different distributions, instead of using any margin-based loss, it is computed using KL divergence as follows.

$$\mathcal{L}_{\text{LR}}^{\text{pair}} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq j^+}}^{K} \text{KL}(p_r \| \hat{p}_i(j)) \tag{3}$$

Herein p_r can be defined as follows.

$$p_r(y) = \begin{cases} 1 - \alpha & \text{if } y = + \\ \alpha & \text{if } y = - \end{cases}$$

Since the prediction is probabilistic, KL divergence penalizes differences in a way that reflects confidence mismatches, i.e., confident, however, when predictions are wrong, are more penalized than uncertain ones. Below we give an example to explain this clearly.

Example Consider (q_i, d_{j^+}, d_{j^-}) that has been judged by multiple annotators. Suppose 80%

of them preferred $d^+_{i,j}$, while 20% preferred d_{j^-}. This uncertainty would be difficult to capture with one-hot labels or a uniform label smoothing approach. Rather than forcing the model to match a hard label $[1, 0]$, or smoothing it arbitrarily to something like $[0.9, 0.1]$, label relaxation allows us to model more concisely. Let us choose the relaxation parameter as $\alpha = 0.2$, with which we can define the relaxed set as $Q^{0.2} = \left\{p \in \Delta^2 : p(+) \geq 0.8,\ p(-) \leq 0.2\right\}$.

Therefore, $p_r(+) = 0.8, \quad p_r(-) = 0.2$. This implies that if the model predicts any values between 0.8 and 1.0, the associated loss would be considered as 0. Otherwise, using the KL-divergence loss, the model is then trained to minimize the divergence. This formulation respects the ambiguity in the supervision and allows the model to output calibrated probabilities that reflect uncertainty, rather than overconfident or artificially smoothed predictions. As a result, label relaxation not only improves robustness to label noise but also enhances the model's ability to represent uncertainty, which is critical in real-world applications such as QA, recommendation, and information retrieval.

Calibration Error in Ranking is typically measured using expected calibration error (ECE) to evaluate the calibration of the model's probability outputs (Naeini et al., 2015; Guo et al., 2017). Calibration in this context refers to the agreement between predicted probabilities and the actual likelihood of correctness. More specifically, considering document ranking tasks, the goal is to ensure that the probability assigned to a document reflects its actual relevance to the query. A well-calibrated ranking model would assign a probability close to 1 to relevant documents and a probability close to 0 to non-relevant ones. To define it more formally, let us assume $\mathcal{R}_{q_i}$ be the set of relevant documents for query q_i, and $\mathcal{D}_{q_i} = \{d_1, d_2, \ldots, d_K\}$ the full set of candidate documents, ECE for neural ranking models is,

$$\text{ECE} = \sum_{i=1}^{N} \sum_{j=1}^{K} |\hat{p}_i(j) - \mathbb{I}(d_j \in \mathcal{R}_{q_i})| \cdot \mathbb{I}(\hat{y}_{i,j} \in \mathcal{B}). \tag{4}$$

Where $\mathbb{I}(d_j \in \mathcal{R}_{q_i})$ is the indicator function that is 1 if d_j is relevant to q_i, and 0 otherwise, and $\mathbb{I}(\hat{y}_{i,j} \in \mathcal{B})$ is an indicator function that checks whether $\hat{p}_i(j)$ falls into a bin $\mathcal{B}$ of predicted probability values. Specifically, $|\hat{p}_i(j) - \mathbb{I}(d_j \in \mathcal{R}_{q_i})|$ represents the absolute error between the predicted probability and the ground truth label. The summation is carried out over all candidate documents within each bin.

4 Evaluation

Datasets and Models For evaluation, we used four datasets, namely, (i) AIDA (Hoffart et al., 2011), (ii) Mintaka (Sen et al., 2022), (iii) LC-QuAD 2.0 (Dubey et al., 2019), and (iv) MS MARCO (Craswell et al., 2021). Datasets (i)–(iii) pertain to entity ranking tasks, and the MS MARCO dataset corresponds to document ranking tasks. The AIDA dataset contains news articles and entities that are linked to Wikipedia. Mintaka is generated through crowd workers, wherein the entities in question-and-answer pairs are linked to the Wikidata knowledge graph. LC-QuAD 2.0 (or in short, LC-QuAD) is also generated through crowd workers, but, contains SPARQL queries. Finally, the MS MARCO dataset is frequently used for diverse tasks to perform question answering, passage ranking, and document ranking. Both LC-QuAD and MS MARCO are question-answering datasets. The ranking models are taken from their original implementation given in `Hugging Face`, BERT [2], E5 [3]. Thereafter, using in-batch negative sampling (Wu et al., 2020b), we fine-tuned them on the datasets described above. Each of the models is trained using the default learning rates and the parameters considering 10 epochs. Finally, the evaluation is performed by using the model's embeddings indexed by using Faiss indexing API (Douze et al., 2024). Note that, for MS MARCO, while computing the hard negatives, we randomly selected 10,000 negative documents from the whole corpus at a time. Finally, for fine-tuning the models, we used a server with 128 GB of RAM and an NVIDIA RTX H100 GPU with 80 GB of RAM.

Label Noise In this work, we consider *semantic-aware* label noise, wherein instead of flipping labels randomly, our approach considers a more realistic scenario. More specifically, for a given ratio of noise addition, we intentionally introduce an error by replacing the correct (relevant) document with a non-relevant one that is semantically very similar to the former. This is done by first randomly choosing a subset of queries from the training batch.

[2] https://huggingface.co/docs/transformers/model_doc/bert

[3] https://huggingface.co/intfloat/e5-base-v2

Table 1: MRR and Recall results evaluated on four ranking datasets using the E5 model, grouped by noise ratio with comparisons across loss functions (PR, LS, LR). The results reported below are the best results obtained considering specific smoothing rates and relaxation parameters.

NR	LF	MRR ↑				Recall↑			
		Msmarco	**Lcquad**	**Mintaka**	**Aida**	**Msmarco**	**Lcquad**	**Mintaka**	**Aida**
0	PR	0.8823	0.9191	0.2419	0.2881	0.9678	0.8739	0.3290	0.1501
	LS	0.8819	0.8823	0.2433	**0.3007**	0.9666	0.8803	0.3701	**0.1692**
	LR	**0.9164**	**0.9194**	**0.3244**	0.2856	**0.9718**	**0.8818**	**0.4470**	0.1557
1	PR	0.8782	**0.9095**	0.2323	0.2782	0.9637	0.8713	0.3147	0.1492
	LS	0.8771	0.8915	0.2418	**0.2914**	0.9617	0.8701	0.3382	**0.1676**
	LR	**0.9165**	0.9090	**0.3385**	0.2835	**0.9713**	**0.8739**	**0.4640**	0.1534
2	PR	0.8757	**0.8922**	0.2119	0.2678	0.9603	0.8576	0.2856	0.1404
	LS	0.8819	0.8808	0.2247	**0.2812**	0.9597	0.8550	0.3003	**0.1498**
	LR	**0.9160**	0.8910	**0.3189**	0.2672	**0.9711**	**0.8593**	**0.4357**	0.1423
4	PR	0.8541	**0.8537**	0.1957	0.2489	0.9451	**0.8180**	0.2658	0.1241
	LS	0.8516	0.8332	0.2020	**0.2719**	0.9441	0.8000	0.2753	**0.1493**
	LR	**0.9128**	0.8452	**0.2818**	0.2530	**0.9703**	0.8058	**0.3748**	0.1302
5	PR	0.6805	0.8169	0.1813	0.2500	0.9129	0.7854	0.2433	0.1293
	LS	0.6907	0.8180	0.1877	**0.2688**	0.9091	0.7718	0.2612	**0.1403**
	LR	**0.9110**	**0.8187**	**0.2673**	0.2412	**0.9331**	**0.7857**	**0.3603**	0.1206

Then, for each selected query, the correct answer is changed and replaced with another candidate that is closest in meaning, based on a similarity score between the original relevant document and all the other candidates. Therefore, we simulate noisy supervision by replacing the correct document with a semantically similar but non-relevant one for a subset of queries. We vary the noise proportion across five levels: 0% (no noise) to 5% of the training labels, following a progressive corruption scheme. Concretely, at 2% noise, 2% of the queries in the training set have their relevant document replaced. Note that although we do not perform human verification to find out the plausibility of the noisy labels, we still ensure their semantic plausibility by selecting replacements based on embedding similarity [4]. Additionally, some datasets, for instance, Mintaka, originate from multiple human annotators, which in principle could provide empirical relevance distributions. However, the versions we use in our evaluation only provide single canonical labels. For consistency across benchmarks, we therefore did not compare against models trained on empirical annotation distributions.

4.1 Results & Discussion

Tables 1 and 2 show the results in terms of MRRs and recall@10 of applying two different calibrated loss functions considering E5 and BERT models. NR depicts different noise ratios, and CL denotes different loss functions. We report results considering five different noise ratios. Noise ratio herein indicates the proportion of training queries for which the relevant document is replaced with a semantically similar but incorrect one. Note that in these tables, we show the results considering pairwise ranking loss. However, we also conducted experiments using cross-entropy loss. Since the results show the same trend, we omit them in the paper.

Considering Table 1, the results suggest that label relaxation can significantly improve the performance of the E5 model when fine-tuned on the Mintaka and MS MARCO datasets under noisy labels. However, considering the AIDA dataset, we find that, in fact, smoothing performs better, and in the LC-QuAD dataset, none of the calibration approaches lead to significant performance improvement. This is because the nature of the dataset determines the effectiveness of a calibration strategy. For AIDA, the relatively structured entity annotations and consistent alignment with the knowledge graph render the soft regularization of smoothing more effective than the plausibility distribution of labels used by label relaxation. In LC-QuAD, the queries are short and ambiguous, and the candidate space is limited, which might impact the calibration approaches. This might further reduce the impact of either calibration approach. These findings

[4]This is further mentioned in the Section 5

Table 2: MRR and Recall results evaluated on four ranking datasets using the BERT model, grouped by noise ratio with comparisons across loss functions (PR, LS, LR). The results reported below are the best results obtained considering specific smoothing rates and relaxation parameters.

NR	LF	MRR↑				Recall↑			
		Msmarco	**Lcquad**	**Mintaka**	**Aida**	**Msmarco**	**Lcquad**	**Mintaka**	**Aida**
0	PR	0.8331	**0.9410**	0.4223	0.3755	0.9254	**0.9271**	0.5111	**0.2231**
	LS	0.8310	0.9338	**0.4261**	**0.3761**	0.9051	0.9171	0.5331	0.2205
	LR	**0.8500**	0.9371	0.4117	0.3551	**0.9381**	0.9113	**0.5457**	0.2210
1	PR	0.7891	**0.9388**	**0.4235**	0.3421	0.8987	**0.9199**	0.5035	**0.2198**
	LS	0.8178	0.9381	0.4165	0.3383	0.8810	0.9090	0.5234	0.2171
	LR	**0.8438**	0.9358	0.4097	**0.3518**	**0.9341**	0.9049	**0.5434**	0.2000
2	PR	0.7517	**0.9108**	0.4058	**0.3353**	0.5900	**0.8989**	0.5021	**0.2065**
	LS	0.7234	0.9088	**0.4241**	0.3211	0.8571	0.8836	**0.5312**	0.2054
	LR	**0.8402**	0.9015	0.3793	0.3301	**0.9301**	0.8844	0.5083	0.1845
4	PR	0.6985	0.8441	**0.3963**	**0.3381**	0.4895	0.8110	0.4938	**0.2150**
	LS	0.7510	0.8419	0.3759	0.3230	0.8220	**0.8190**	**0.5114**	0.2065
	LR	**0.8400**	**0.8509**	0.3299	0.3104	**0.9301**	0.8176	0.4372	0.1718
5	PR	0.6885	0.8001	0.3543	0.3211	0.4074	0.7719	0.4255	0.2031
	LS	0.7491	**0.8199**	**0.3741**	**0.3230**	0.8113	0.7881	**0.5013**	**0.2063**
	LR	**0.7819**	0.8192	0.3019	0.2944	**0.9110**	**0.7898**	0.4009	0.1777

highlight that while label relaxation offers strong robustness under certain noise settings, its efficacy is still dataset-dependent and should be carefully selected based on the underlying characteristics of the data and task.

In Table 2, we see the results of the BERT model, wherein it can be observed that the label relaxation does not show significant performance improvement for Lc-QuAD, Mintaka, and Aida datasets. In those datasets, label smoothing performs slightly better. However, it also does not significantly improve the results in comparison to pairwise loss. These findings are consistent with the study by Zhu et al. (2022) wherein they reported that label smoothing does not improve the performance of the BERT model under label noise generated in the weakly supervised step.

E5 model, despite using the same underlying BERT model, is extensively weakly-supervised trained on the ranking dataset (Wang et al., 2022) that makes it inherently more robust to noisy supervision and better calibrated in its embedding space. This encourages the model to learn smoother decision boundaries and more stable representations. As a result, when fine-tuned with label relaxation, E5 is able to leverage its calibrated embedding space to better align the relaxed supervision with meaningful semantic gradients. In contrast, the standard BERT model lacks such domain-specific pre-training and starts from a relatively uncalibrated representation space for the ranking task, making it more sensitive to label noise and less responsive to relaxation-based regularization. However, we see that for the largest dataset, MS MARCO, label relaxation outperforms the other calibration approaches for the BERT model. This observation suggests that for very large datasets, the relaxed set can be helpful even when the model is not pre-trained on ranking datasets. Herein, the availability of training instances allows the model to benefit from the soft supervision, avoiding overfitting to incorrect labels. In contrast, the standard BERT model lacks such domain-specific pre-training and starts from a relatively uncalibrated representation space for the ranking task, making it more sensitive to label noise and less responsive to relaxation-based regularization. With these results, we highlight the following important findings.

Dataset size & diversity. Large datasets such as MS MARCO and Mintaka work effectively with label relaxation since this distributes probability mass over semantically plausible candidates without overfitting to noisy labels.

Candidate space structure. Highly structured datasets like AIDA favor a calibrated loss function. Herein, smoothing gains top performance since it enforces small-entropy distributions.

Query ambiguity. In LC-QuAD, where ambiguity and candidate space constraints dominate, calibration does not yield notable performance gain.

Table 3: Loss functions to use when training on MS MARCO, Mintaka, LC-QUaD, and AIDA datasets.

Dataset	E5	BERT	Notes
MS MARCO	Label Relaxation	Label Relaxation	Largest dataset; soft supervision avoids overfitting
LC-QuAD	Pairwise loss	Pairwise loss	Small candidate space; calibration has little effect
Mintaka	Label Relaxation	Label Relaxation	Large, diverse queries; E5 benefits from calibrated embeddings
AIDA	Label Smoothing	Label Smoothing	Structured entity annotations; smoothing aligns better

Model pre-training. This is probably the most important finding of all. As mentioned previously, E5's extensive weakly-supervised pre-training on ranking data produces smoother embedding manifolds, allowing label relaxation to align gradients with semantically similar negatives. In contrast, standard BERT lacks such calibration and is more sensitive to label noise.

Based on the above observation, we provide a *practical guidance* as to when to use a specific type of loss function. This is summarized in Table 3.

Calibration Error Analysis Based on our proposed expected calibration error, defined in Equation 4, we evaluated the calibration of the E5 and BERT models, considering pairwise loss, label smoothing, and label relaxation. The results are reported in Table 4. Note that since LC-QuAD does not yield notable performance improvement using calibrated loss functions, we do not consider it.

We see that typically label relaxation leads to the lowest calibration errors for most of the datasets. However, the differences between the ECE values of the calibrated and non-calibrated loss functions are not remarkably high. In fact, considering the MS MARCO dataset, we find that ECE is lower for non-calibrated loss in high noise ratios compared to calibrated loss functions, even when the performance drops significantly with non-calibrated loss functions. This shows some known shortcomings of ECE, for instance, its histogram binning can mask differences, specifically considering high estimator bias and variance depending on bin count and scheme. Additionally, since ECE aggregates class- and score-conditional structure, work on ranking scale calibration similarly reports that off-the-shelf ECE can be misleading without class balancing or rank-aware structure (Widmann et al., 2019; Futami and Fujisawa, 2024; Yan et al., 2022). These works reported results on vision-based rankers. In this work, we find the same drawback in document ranking models as well.

As an alternative to calibration error, we analyze the behavior of the ranking models BERT and E5 under label noise by plotting the training and validation performance side-by-side and observe the differences in Figure 1 and 2 (Appendix A), respectively (). Therein, we see that as the label noise increases, the gap increases notably. This behavior is consistent with memorization under label noise. Specifically, the models eventually fit corrupted labels, inflating training metrics while harming generalization. This highlights that the performance of ranking under noise is not captured by ECE metric. The widening recall gap, as *memorization* error (Zhang et al., 2021; Han et al., 2025), is therefore a practical metric herein to guide calibration or early stopping.

5 Conclusion & Future Directions

In this work, we have studied label relaxation considering the neural ranking models in performing document ranking tasks. To this end, we first formally define the label relaxation in the context of the ranking task. Afterwards, we integrate it into bi-encoder ranking models. Additionally, to find out whether label relaxation can mitigate the impact of label noise in fine-tuning neural ranking models, we conducted extensive evaluations considering 2 different bi-encoder models, 4 different ranking datasets, and 5 different noise levels. We also compare our results to the popular label smoothing calibration approach. The results of our evaluation suggest that label relaxation can indeed be helpful in fine-tuning ranking models when label noise is prevalent in the ranking datasets. However, our findings also suggest that label relaxation is effective on the E5 model, which is extensively

Table 4: Calibration Error results evaluated on four ranking datasets using the E5 and Bi-Encoder models, grouped by noise ratio (NR) and loss functions (LF) with the best ϵ and α as smoothing and relaxation parameters. Lower is better; bold indicates the best value for a dataset at a given noise ratio.

NR	LF	E5↓			BERT↓		
		Msmarco	**Mintaka**	**Aida**	**Msmarco**	**Mintaka**	**Aida**
0	PR	0.0424	0.2914	**0.2098**	0.0090	0.3277	0.3016
	LS (0.1)	0.0317	0.2917	0.2101	0.0091	0.3199	0.2800
	LR (0.1)	**0.0269**	**0.2811**	0.2197	**0.0076**	**0.3185**	**0.2770**
1	PR	0.1538	0.2879	**0.2123**	0.0094	0.3391	0.3123
	LS (0.1)	0.1398	0.2883	0.2165	**0.0091**	0.3109	0.3001
	LR (0.2)	**0.1221**	**0.2846**	0.2193	0.1000	**0.3019**	**0.2877**
2	PR	**0.2024**	0.2726	0.2066	**0.1094**	0.3293	0.2893
	LS (0.2)	0.2119	0.2713	0.2081	0.1913	0.3150	0.2891
	LR (0.2)	0.2175	**0.2661**	**0.2041**	0.1911	**0.2854**	**0.2713**
4	PR	**0.2969**	0.2598	0.2033	**0.1111**	0.3373	0.3049
	LS (0.2)	0.3018	0.2561	0.2049	0.1101	0.3171	0.3098
	LR (0.3)	0.3161	**0.2476**	**0.1908**	0.2000	**0.3104**	**0.2811**
5	PR	**0.3150**	0.3109	0.2025	**0.2082**	0.3322	0.2943
	LS (0.3)	0.3310	0.2601	0.2019	0.2910	0.3091	0.2920
	LR (0.3)	0.3293	**0.2417**	**0.1902**	0.2989	**0.3047**	**0.2918**

weakly supervised pre-trained on the ranking tasks. On the other hand, if a pre-training on the ranking tasks is not performed, the results do not improve. Additionally, we find that the ECE might not be suitable to measure the calibration of the ranking models under noise and memorization errors could be helpful to get better insights.

We believe label relaxation has a lot of potential to build well-calibrated models. We can envisage works that would explore adaptive label relaxation approaches that adjust relaxation based on model confidence or noise estimates, and investigate their effects on model calibration. Furthermore, extending relaxation to cross-encoder and LLM-based rankers, and studying new measures to compute calibration error in the context of document ranking, could also be a potentially interesting direction.

Limitations

The following are some of the key limitations of our study, which we acknowledge.

Model limitation Our experiments are restricted to bi-encoder architectures, i.e., BERT and E5. Although this choice allowed us to systematically analyze calibration under controlled conditions, the findings may not directly transfer to more complex architectures such as cross-encoders or large language model (LLM)-based rankers. However, such an extension would require additional challenges and opportunities.

Dataset limitation We considered experimenting with four datasets, AIDA, Mintaka, LC-QuAD, and MS MARCO. These vary in size, structure, and annotation quality; therefore, providing a diverse evaluation setting. However, our conclusions remain dataset-dependent, as seen from the differences in effectiveness across AIDA, LC-QuAD, and MS MARCO. Future work should examine a broader range of datasets, including multilingual and domain-specific ranking tasks, to assess the generalizability of these models.

Label noise modeling The label noise is modeled using a semantic-aware perturbation strategy, replacing relevant documents with semantically similar but incorrect ones. This provides a more realistic scenario compared to random flipping; however, real-world noise can be more diverse such as adversarial noise, annotation inconsistencies, or systematic bias. Our approach does not capture these variations; hence, the robustness of label relaxation under such conditions remains unexplored.

Calibration scope Our evaluation focused on *text-based calibration*, therefore, the probability estimates are aligned with annotation labels. We did not compare our approach against ranking-based calibration methods which adjust scores based on relative order or rank-aware confidence measures. These approaches are often stronger baselines in retrieval and relevance tasks, however, they fall outside the scope of this paper due to space limitations.

Hence, as part of future work, we would provide a more comprehensive view of calibration strategies for ranking.

Relaxation parameter The choice of relaxation parameter α is tuned using validation performance, which may not always be feasible in practice, especially when noisy labels affect the validation set itself. Adaptive or noise-aware strategies to determine relaxation parameters could further improve robustness and practicality.

Ethics Statement

This work relies exclusively on publicly available benchmark datasets (AIDA, Mintaka, LC-QuAD, and MS MARCO). No personally identifiable or sensitive information was collected or processed. All experiments were conducted in accordance with the terms of use of the respective datasets. Therefore, we believe our contributions pose no ethical risks beyond the general concerns of bias and fairness inherent in natural language processing and information retrieval research.

Acknowledgment

This work is supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL under the grant no NW21-059D, the project WHALE (LFN 1-04) funded under the Lamarr Fellow Network Programme by the MKW NRW, the European Union's Horizon Europe research and innovation programme under grant agreement No 101070305, and by the German Federal Ministry of Research, Technology and Space (BMFTR) within the project KI-OWL under the grant no 01IS24057B.

References

Jiacheng Cheng and Nuno Vasconcelos. 2022. Calibrating deep neural networks by pairwise constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13699–13708. IEEE.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. A thorough comparison of cross-encoders and llms for reranking SPLADE. *CoRR*, abs/2403.10407.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. 2021. Local temperature scaling for probability calibration. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6869–6879. IEEE.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II*, page 69–78, Berlin, Heidelberg. Springer-Verlag.

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 438–447. ACM.

Futoshi Futami and Masahiro Fujisawa. 2024. Information-theoretic generalization analysis for expected calibration error. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2619–2629. Association for Computational Linguistics.

Arindam Ghosh, Thomas Schaaf, and Matthew R. Gormley. 2022. Adafocal: Calibration-aware adaptive focal loss. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney,*

NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Andi Han, Wei Huang, Zhanpeng Zhou, Gang Niu, Wuyang Chen, Junchi Yan, Akiko Takeda, and Taiji Suzuki. 2025. On the role of label noise in the feature learning process. *CoRR*, abs/2505.18909.

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. 2022. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16060–16069. IEEE.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Baihe Huang, Hiteshi Sharma, and Yi Mao. 2024. Enhancing language model alignment: A confidence-based approach to label smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21341–21352.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. 2021. Embedding transfer with label relaxation for improved metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3967–3976. Computer Vision Foundation / IEEE.

Ivan Kobyzev, Aref Jafari, Mehdi Rezagholizadeh, Tianda Li, Alan Do-Omri, Peng Lu, Pascal Poupart, and Ali Ghodsi. 2023. Do we need label regularization to fine-tune pre-trained language models? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 166–177.

Weizhi Li, Gautam Dasarathy, and Visar Berisha. 2020. Regularization via structural label smoothing. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 1453–1463. PMLR.

Julian Lienen and Eyke Hüllermeier. 2021. From label smoothing to label relaxation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8583–8591. AAAI Press.

Julian Lienen and Eyke Hüllermeier. 2024. Mitigating label noise through data ambiguation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 13799–13807. AAAI Press.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022. The devil is in the margin: Margin-based label smoothing for network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 80–88. IEEE.

Tongliang Liu and Dacheng Tao. 2016. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):447–461.

Xingchen Ma and Matthew B. Blaschko. 2021. Meta-cal: Well-controlled post-hoc calibration by ranking. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7235–7245. PMLR.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7034–7044. PMLR.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1196–1204.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2233–2241. IEEE Computer Society.

Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2022. Learning to rank from relevance judgments distributions. *Journal of the Association for Information Science and Technology*, 73(9):1236–1252.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

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 and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5464–5472. AAAI Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Hung-Nghiep Tran, Akiko Aizawa, and Atsuhiro Takasu. 2024. An encoding–searching separation perspective on bi-encoder neural search. *Preprint*, arXiv:2408.01094.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. 2022. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23589–23614. PMLR.

Jiaheng Wei and Yang Liu. 2021. When optimizing f-divergence is robust with label noise. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

David Widmann, Fredrik Lindsten, and Dave Zachariah. 2019. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12236–12246.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020a. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020b. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6442–6454. Association for Computational Linguistics.

Le Yan, Zhen Qin, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2022. Scale calibration of deep ranking models. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4300–4309. ACM.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13873–13882. Computer Vision Foundation / IEEE.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

Leixin Zhang and Daniel Braun. 2024. Twente-BMS-NLP at PerspectiveArg 2024: Combining bi-encoder and cross-encoder for argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 164–168, Bangkok, Thailand. Association for Computational Linguistics.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3712–3721. IEEE.

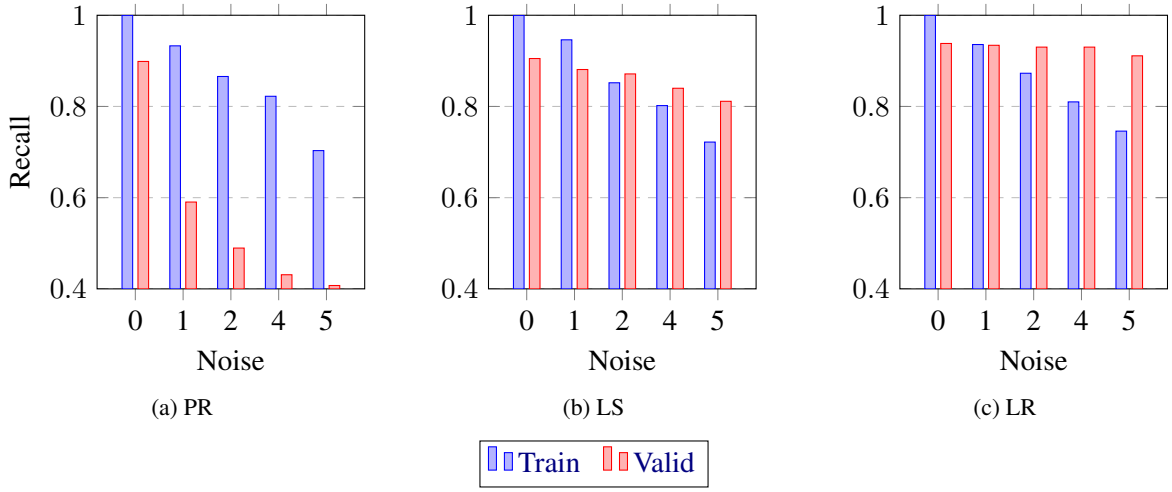
Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. Entqa: Entity linking as question answering. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhenyu Zhang, Xiaobo Sind, Tingwen Liu, Zheng Fang, and Quangang Li. 2020. Joint entity linking and relation extraction with neural networks for knowledge base population. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? A study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP, Insights@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 62–67. Association for Computational Linguistics.

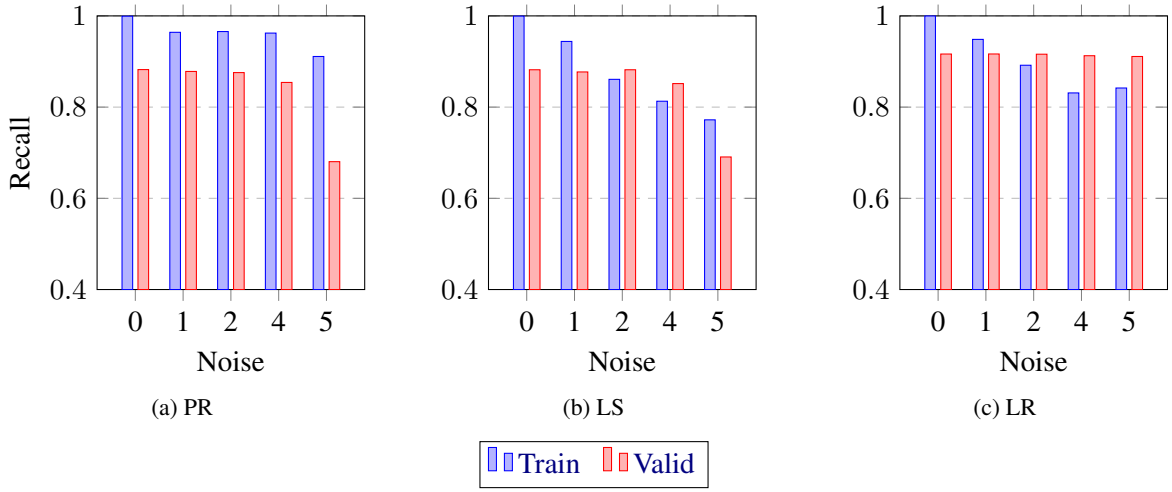
Zhaowei Zhu, Tongliang Liu, and Yang Liu. 2021. A second-order approach to learning with instance-dependent label noise. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10113–10123. Computer Vision Foundation / IEEE.

A Additional Results



(a) PR (b) LS (c) LR

Figure 1: Train vs. validation recall across noise levels for PR, LS, and LR for BERT.



(a) PR (b) LS (c) LR

Figure 2: Train vs. validation recall across noise levels for PR, LS, and LR for E5.