Degree Project in Machine Learning

Second cycle, 30 credits

Design Novel Effective Method for Large Language Model Compression

BiLD: Bi-directional Logits Difference Loss for Large Language Model Distillation

MINCHONG LI

Design Novel Effective Method for Large Language Model Compression

BiLD: Bi-directional Logits Difference Loss for Large Language Model Distillation

MINCHONG LI

Master's Programme, Machine Learning, 120 credits
Date: June 24, 2024

Supervisors: Amirhossein Layegh Kheirabadi, Xiaohui Song
Examiner: Amir Hossein Payberah
School of Electrical Engineering and Computer Science
Host company: OPPO
Swedish title: Utformning en Ny Effektiv Metod för Komprimering av Stor Språkmodell
Swedish subtitle: BiLD: Bi-directional Logits Difference Loss för Destillering av Stor Språkmodell

Abstract

In recent years, Large Language Models (LLMs) have shown exceptional capabilities across various Natural Language Processing (NLP) tasks. However, such impressive performance often comes with the trade-off of an increased parameter size, posing significant challenges for widespread deployment. Knowledge Distillation (KD) provides a solution by transferring knowledge from a large teacher model to a smaller student model. In this thesis, we explore the task-specific distillation of LLMs at the logit level. Our investigation reveals that the logits of fine-tuned LLMs exhibit a more extreme long-tail distribution than those from vision models. Moreover, existing logits distillation methods often struggle to effectively utilize the internal ranking information from the logits. To address this, we propose the **Bi**-directional **L**ogits **D**ifference (BiLD) loss. The BiLD loss filters out the long-tail "noise" by utilizing only top-k teacher and student logits, and leverages the internal logits ranking information by constructing logits differences. To evaluate BiLD loss, we conduct comprehensive experiments on 13 datasets using two types of LLMs. Our results show that the BiLD loss, with only the top-**8** logits, outperforms supervised fine-tuning (SFT), vanilla Kullback–Leibler (KL) loss, and five other distillation methods from both NLP and Computer Vision (CV) fields.

Keywords

Large Language Model, Model Compression, Knowledge Distillation

Sammanfattning

På senare år har stora språkmodeller (LLMs) visat exceptionella förmågor över olika NLP-uppgifter. Men sådan imponerande prestanda kommer ofta med en kompromiss i form av ökad parameterstorlek, vilket innebär betydande utmaningar för utbredd användning. Kunskapsdistillation (KD) erbjuder en lösning genom att överföra kunskap från en stor lärarmodell till en mindre studentmodell. I denna avhandling utforskar vi uppgiftsspecifik distillation av stora språkmodeller på logitnivå. Vår undersökning visar att logiterna från finjusterade LLMs uppvisar en mer extrem långsvansfördelning än de från visionsmodeller. Dessutom kämpar befintliga metoder för logitdistillation ofta med att effektivt utnyttja den interna rankningsinformationen från logiterna. För att åtgärda detta föreslår vi förlustfunktionen BiLD (Bi-directional Logits Difference). BiLD-förlusten filtrerar bort långsvansens ”brus” genom att endast använda de översta k lärar- och studentlogiterna, och utnyttjar den interna logitrankningsinformationen genom att konstruera logitskillnader. För att utvärdera BiLD-förlusten genomför vi omfattande experiment på 13 datamängder med två typer av LLMs. Våra resultat visar att BiLD-förlusten, med endast de översta 8 logiterna, överträffar både övervakad finjustering (SFT), vanilj-KL-förlust och fem andra distillationsmetoder från både NLP- och CV-fälten.

Nyckelord

Stor Språkmodell, Modellkompression, Kunskapsdestillation

Acknowledgments

This thesis was supported by OPPO Research Institute. I am profoundly grateful to my company supervisor, Xiaohui Song, for his unwavering support and invaluable advice throughout this thesis. His guidance and insight play an indispensable role in the completion of my degree project.

I would also like to thank my KTH supervisor, Amirhossein Layegh Kheirabadi, and my examiner, Amir Hossein Payberah, for their mentorship and for the constructive discussions we have. Their expertise and thoughtful perspectives have been instrumental in refining my work.

Beijing, China, June 2024
Minchong Li

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

CV Computer Vision

KD Knowledge Distillation
KL Kullback–Leibler

LLMs Large Language Models

NLP Natural Language Processing

Chapter 1

Introduction

The last few years have witnessed Large Language Models (LLMs) risen to prominence, demonstrating remarkable proficiency in natural language understanding and generation. However, these capabilities come at the cost of an ever-increasing number of parameters. Due to constraints on computational resources, the formidable size of LLMs hinders their democratization and widespread deployment. Knowledge Distillation (KD), as a classic model compression method[1], provides a solution for reducing model size while striving to maintain performance. KD transfers knowledge from a large teacher model to a smaller student model, thereby enhancing the student model's performance and making it a viable alternative for deployment.

As an important branch of KD, logits distillation has gained popularity due to its straightforward application. The goal of logits distillation is to minimize the Kullback–Leibler (KL) divergence between the teacher and student logits. A significant portion of research on logits distillation has focused on vision models [2, 3, 4, 5]. However, the application of these methods to distill LLMs has yet to be thoroughly explored due to potential differences in structure, data distribution, and output space between vision and language models.

For LLMs, research on logits distillation is still emerging, with methods such as reverse KL [6, 7, 8] and those based on optimal transport metrics [9]. However, in practical applications, the former suffers from the "mode-seeking" problem [10, 11], while the latter is computationally too complex for open-source large models with billions of parameters.

In this thesis, we investigate the characteristics of logits in LLMs. Compared to the limited output space of vision models, LLMs' output space comprises sequences of discrete tokens of potentially infinite length, making LLM logits significantly more complex. Furthermore, LLM logits exhibit a noticeable long-tail distribution, indicating a substantial portion of ”noise” beyond a small amount of ”key knowledge”. We also observe that in LLM text generation, common strategies like top-k sampling and top-p sampling are influenced by the internal ranking of logits when selecting output tokens. However, existing logits distillation methods often struggle to exploit this latent ranking information [5].

Inspired by these characteristics, we design a novel loss, the Bi-directional Logits Difference (BiLD) loss, for task-specific LLM distillation. BiLD loss emphasizes reducing long-tail ”noise” and explicitly utilizes the ranking information in logits. It computes KL divergence based on reconstructed ”logits differences,” which are obtained by calculating the internal pairwise differences of values from top-k teacher (student) logits and the corresponding student(teacher) logits. Our experiments show that BiLD loss, using only the top-8 logits, achieves state-of-the-art (SOTA) results across various Natural Language Processing (NLP) tasks.

To conclude, we make the following contributions:

- We investigate the characteristics of LLMs’ logits, discussing their intrinsic distribution and the significance of logits ranking.

- We propose the Bi-directional Logits Difference (BiLD) loss for logits distillation in LLMs. BiLD filters out inherent ”noise” in logits while leveraging logits ranking information to enhance performance. Our method can serve as an alternative to the vanilla KL loss in existing LLM distillation methods.

- To demonstrate the effectiveness of BiLD loss, we conduct comprehensive experiments on 13 NLP datasets using two open-source LLMs, BLOOM [12] and Qwen1.5 [13]. We evaluate various logits distillation methods from both Computer Vision (CV) and NLP domains. Experimental results show that our BiLD loss outperforms SFT, vanilla KL loss and five other methods using only the top-8 logits. Furthermore, our comparison of teacher and student logits shows that BiLD loss promotes better imitation of teacher behavior at the logit level.

1.1 Problem

Existing logits distillation methods are not specifically tailored to the unique characteristics of LLMs’ logits. Consequently, the research problem of this thesis is to develop a novel distillation method that better aligns with the characteristics of LLMs’ logits, thereby improving distillation performance.

The research questions of the thesis can be formulated as follows:

- What are the characteristics of LLMs' logits? How can these characteristics be leveraged to design effective distillation loss?
- Can we improve logits distillation performance by designing methods that better align with the characteristics of LLMs' logits?

1.2 Purpose

The purpose of this project is to explore logit-based knowledge distillation methods suitable for LLMs. While logits distillation methods for vision models are well-established, the unique characteristics of text data, such as output space size and internal logits distribution, prevent the methods from the CV field directly applicable to LLMs. Therefore, this thesis aims to design distillation methods specifically tailored to the characteristics of LLMs' logits, thereby advancing related research.

1.3 Goals

The goal of this project is to design a logit-based knowledge distillation method for LLMs. This has been divided into the following three sub-goals:

1. Show the characteristics of LLMs' logits through simple experiments.
2. Design a logit distillation method tailored to the characteristics of LLMs' logits.
3. Conduct comprehensive experiments on various datasets to validate the method's effectiveness and analyze the results.

The expected outcome of this thesis is a distillation algorithm that can be used directly for LLMs.

1.4 Research Methodology

This research is grounded in a pragmatist perspective, emphasizing practical outcomes and empirical evidence. We first qualitatively explored the characteristics of LLM logits, discovering that they exhibit a highly extreme long-tail distribution. Furthermore, existing distillation methods fail to effectively utilize the internal ranking information of LLMs. To address these issues, we designed the BiLD loss for LLM distillation. To thoroughly demonstrate the effectiveness of our approach, we considered various baseline methods from both the CV and NLP domains. In our result analysis, we primarily employed quantitative methods, calculating accuracy, EM score, F1 score, and more for the distillation experiments with different methods. Additionally, we proposed a novel metric, called overlap@k, to evaluate the performance of different methods at the logit level.

1.5 Delimitations

Apart from knowledge distillation, approaches like quantization and pruning are also utilized for compressing LLMs. However, this thesis focuses solely on investigating knowledge distillation methods for LLM compression. This decision stems from OPPO's concurrent exploration of diverse research directions. Following discussions with OPPO supervisor, I am tasked with delving into the avenue of knowledge distillation for model compression.

Simultaneously, this project conducts experiments exclusively on two LLMs, BLOOM and Qwen, encompassing both teacher training and student distillation. This choice arises from the significant computational overhead associated with training teachers on different models. Moreover, due to the substantial size of LLMs, effective methods are often unbiased towards specific model types.

1.6 Structure of the thesis

Chapter 2 presents relevant background information about LLMs and KD, as well as some related works about logits distillation and other distillation methods for LLMs. Chapter 3 introduces the formulation of traditional logits distillation, the characteristics of LLMs' logits, as well as our proposed BiLD loss. In Chapter 4, we introduce the datasets, baselines and our implementation details, following our main results and analysis for the results, as well as two analyses about the impact of temperature and k value in BiLD loss. Finally, Chapter 5 presents our conclusion, limitations and reflections.

Chapter 2

Background

This chapter aims to provide a comprehensive overview of the research area of this thesis and some related works. This chapter is divided into four parts: section 2.1 introduces the development and current process in the field of LLMs. section 2.2 introduces knowledge distillation, a model compression technique used in this thesis. Section 2.3 presents some works related to our thesis, mainly focusing on two aspects: logits distillation and distillation methods for LLMs.

2.1 Large Language Models

LLMs have become a cornerstone in the field of NLP, powering a wide range of applications from machine translation to conversational agents. These models are characterized by their massive scale, typically consisting of billions of parameters. The success of LLMs is attributed to their ability to capture complex patterns in vast amounts of text data, enabling them to generate human-like text and understand nuanced linguistic contexts.

The development of LLMs began with the introduction of the Transformer architecture [14], which revolutionized the field by providing a more efficient way to handle long-range dependencies in text compared to previous recurrent neural networks (RNNs) and convolutional neural networks (CNNs). The self-attention mechanism in Transformers allows for the parallelization of computations, making it feasible to train on large datasets and scale up the model size significantly.

One of the key milestones in the evolution of LLMs is the release of BERT (Bidirectional Encoder Representations from Transformers) [15]. The structure of BERT

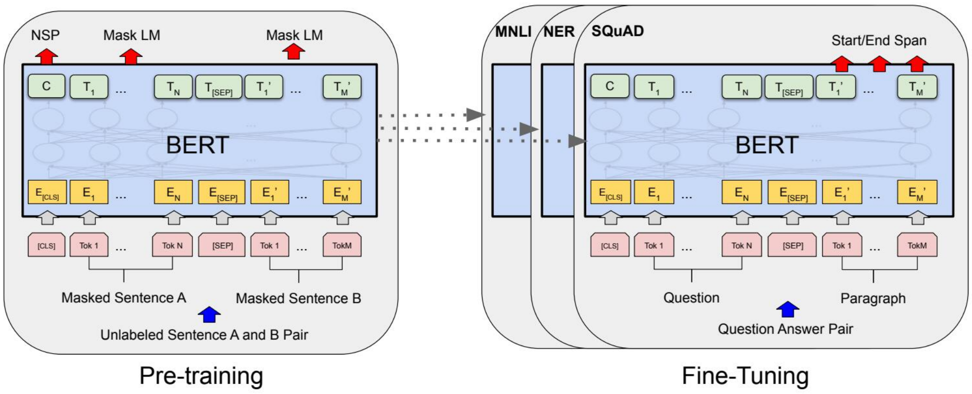

Figure 2.1: The structre of BERT model.

is shown in Figure 2.1. It introduces a novel pre-training approach based on masked language modeling and next sentence prediction, setting new benchmarks in various NLP tasks. Following BERT, models like GPT-2 [16] and GPT-3 [17] by OpenAI demonstrated the potential of autoregressive language models trained on vast and diverse datasets to generate coherent and contextually relevant text.

In recent years, the development of LLMs has advanced to even larger scales and higher performance levels. Models such as GPT-4 [18] and Gemini [19] have pushed the boundaries of model size and capability, boasting hundreds of billions of parameters and achieving remarkable results across numerous benchmarks and real-world applications. Concurrently, there has been a growing movement towards creating open-source, smaller-scale LLMs that are suitable for private deployment. Models like BLOOM [12] and Qwen[13], offer powerful language understanding and generation capabilities while being more accessible and easier to deploy in resource-constrained environments. This dual trajectory of scaling up for performance and scaling down for accessibility reflects the diverse needs of the NLP community and the broader technology landscape.

Despite the impressive capabilities, LLMs come with challenges such as high computational costs, substantial memory requirements, and the need for large-scale annotated data for fine-tuning. Moreover, deploying these models in resource-constrained environments remains a significant hurdle. To mitigate these issues, the thesis has focused on model compression techniques, specifically distillation methods.

2.2 Knowledge Distillation

KD is a technique used to transfer knowledge from a larger, more complex model (the teacher) to a smaller, more efficient model (the student). This method is pivotal in the field of machine learning, especially for deploying models in resource-constrained environments where computational power and memory are limited.

The concept of knowledge distillation was first introduced in [20] and later formalized by [1]. The main idea is to train the student model to mimic the output of the teacher model. Instead of training the student model directly on the ground truth labels, it is trained to reproduce the teacher model' s output probabilities (logits). These logits often contain richer information than the hard labels because they encode the relative probabilities of all classes, providing a more informative signal for training.

The distillation process involves two primary steps. 1) Train the Teacher Model. The teacher model, usually a deep and complex neural network, is trained on a large dataset to achieve high accuracy. 2) Train the Student Model. The student model, which is typically smaller and less complex, is trained to match the softened output (logits) of the teacher model. The soft targets are generated using a higher temperature in the softmax function, which smooths the output distribution of the teacher, providing more information about which classes the teacher found to be similar. The loss function used in KD is a combination of the traditional cross-entropy loss with the ground truth labels and the Kullback-Leibler divergence loss with the teacher' s softened output. This dual objective helps the student model learn both the exact labels and the generalization characteristics of the teacher model.

In the context of LLMs, knowledge distillation is particularly valuable. LLMs are extremely resource-intensive, making them impractical for many real-world applications. By using KD, smaller versions of these models can be created, which retain much of the performance of the original models but require significantly fewer resources.

2.3 Related Works

2.3.1 Logits Distillation

One representative approach of knowledge distillation is logits distillation, which transfers knowledge by minimizing the divergence of output logits [21]. For vision models, there has been substantial research on logits distillation. Approaches like DKD [2] and NKD [22] decouple the target and non-target components of logits, applying weighting or regularization. NormKD [4] dynamically customizes temperatures during the distillation process. However, the differences in structure, data, and output space between vision models and LLMs make it challenging to directly apply these methods to LLMs.

Recent research has introduced several logit distillation methods suited for LLMs. Reverse KL (RKL) [6, 8] has been used to mitigate the ”mode-averaging” problem; however, it can sometimes lead the student model towards ”mode-seeking” behavior. DistiLLM [23] proposes mixing the logits distributions of the teacher and the student, but this introduces additional hyperparameters, increasing its complexity in practical applications. SinKD [9] replaces KL divergence with Sinkhorn Distance, but its computational demands can pose challenges when applied to larger models.

Our work continues the paradigm of reducing the divergence of logits. However, unlike previous approaches, we calculate the divergence using logits differences instead of the logits themselves. Our method focuses the model on the ”key knowledge” in the teacher logits without introducing excessive hyperparameters that require extensive tuning.

2.3.2 Other Distillation Methods for LLMs

Previous works on distillation for LLMs extend beyond logits-based methods, primarily falling into two categories: white-box and black-box approaches [24]. White-box distillation [8, 25, 26] leverages the teacher's internal representations and hidden states to facilitate knowledge transfer. However, these methods often rely on structural similarities between the teacher and student models. In contrast, black-box distillation only permits the student to access the teacher's outputs. Current research in black-box distillation mainly focuses on learning from the teacher's output texts [27, 28, 29]. While BiLD can be classified as black-box distillation, it serves as an alternative to the vanilla KL divergence loss and can be easily integrated with white-box distillation methods.

Chapter 3

Methods

In this chapter we explain the methods used in thesis. We introduce the theory about KL divergence, teacher and student models' role in KD, as well as logits distillation (section 3.1, 3.2, 3.3). Then we analyse the characteristics of LLMs' logits through a toy experiment in section 3.4. Base on these two sections, we formally propose the BiLD loss in section 3.5.

3.1 Brief Review of KL Divergence

KL divergence, or relative entropy, is a metric used to compare two data distributions. It is a concept of information theory that contrasts the information contained in two probability distributions. The form of KL divergence can be represented as:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{3.1}$$

In this formula, P and Q are the probability distributions, and i represents each possible outcome. This expression calculates the KL divergence from distribution Q to distribution P.

Knowledge distillation in the context of large language models (LLMs) typically uses KL divergence as the loss, as it involves training a smaller "student" model to imitate the behavior of a larger "teacher" model. Instead of using the hard labels from the original dataset, the student model learns from the soft targets provided by the teacher model, which are probability distributions over the possible vocabularies. KL divergence is well-suited for comparing these probability distributions because it measures how one probability distribution diverges from a second, expected probability distribution. This helps the student model learn to produce a similar distribution to the teacher model.

3.2 Brief Review of the Teacher Model and Student Model in Knowledge Distillation

In the distillation of LLMs, the teacher model is a large, pre-trained LLM that serves as a source of knowledge. The text generated by the teacher can be represented as a sequence of tokens, and each token can be represented as a logit. The length of a logit corresponds to the size of the LLM's vocabulary. Each position in the logit represents the model's predicted score for each word in the vocabulary, indicating the likelihood that the current token is that specific word.

To elaborate, when a sentence is input into the teacher model, it generates a logit vector for each token in the sequence. This logit vector contains the model's prediction scores for all possible words in its vocabulary. By applying a softmax transformation to these scores, we obtain a probability distribution over the vocabulary, showing which word is most likely to be the current token.

For instance, consider a vocabulary consisting of ["wolf", "cat", "sheep"]. If the teacher model processes the phrase "the dangerous grey" and is going to generate the next token. Assume it generates a logit vector $[1.2, 0.9, -0.3]$ for the next token, applying the softmax function to this vector might yield a probability distribution of $[0.5092, 0.3772, 0.1136]$. This distribution indicates that the model predicts a 50.92% probability for "wolf", 37.72% for "cat", and 11.36% for "sheep".

In the distillation process, we aim for the student model to learn these crucial probability distributions rather than replicating the teacher's output tokens exactly. Through different distillation loss, we focus the student on the the logits from the teacher. This approach effectively utilizes the internal knowledge in logits, promoting student model's imitation of the teacher.

3.3 Brief Review of Logits Distillation

Logits distillation calculates the divergence between the teacher's and student's output logits as the optimization target. Consider a teacher model t and a student model s, both with a vocabulary size N. During the process of single token prediction, the teacher logits $\mathbf{z}^t$ and student logits $\mathbf{z}^s$ at a certain position can be represented as:

$$\mathbf{z}^t = \left[z_1^t, z_2^t, \cdots, z_N^t\right] \in \mathbb{R}^{1\times N},$$
$$\mathbf{z}^s = \left[z_1^s, z_2^s, \cdots, z_N^s\right] \in \mathbb{R}^{1\times N}. \tag{3.2}$$

Logits are the raw outputs of language models and cannot be directly used to calculate the loss. We process the logits into probabilities $\mathbf{p}^t$ and $\mathbf{p}^s$, where the element p_i from $\mathbf{p}^t$ or $\mathbf{p}^s$ represents the probability of the token at the i-th position being sampled as the output:

$$\mathbf{p}^t = \frac{\exp(\mathbf{z}^t/T)}{\sum_N^{i=1}\exp(z_i^t/T)} \in \mathbb{R}^{1\times N},$$
$$\mathbf{p}^s = \frac{\exp(\mathbf{z}^s/T)}{\sum_N^{i=1}\exp(z_i^s/T)} \in \mathbb{R}^{1\times N}, \tag{3.3}$$

where T is the temperature during normalization. The vanilla KL divergence loss is defined as:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left[\mathbf{p}^t \parallel \mathbf{p}^s\right]. \tag{3.4}$$

By aligning the student's logits distribution with that of the teacher using vanilla KL loss, the student can imitate the teacher's performance at the logit level, thereby facilitating knowledge transfer.

3.4 The Characteristics of LLMs' Logits

Compared to vision models, LLMs have an output space consisting of infinitely long sequences of tokens, making their logits more complex. We conduct a toy experiment to compare the logit characteristics of vision models and LLMs. We choose ResNet-101 [30] and Qwen-4B [13] for the toy case. We randomly select five images and five sets of instructions from our test data as inputs for the vision and language models (details about images and instructions are provided in Appendix C). We use kurtosis to measure the extremity of logits' long-tail distribution and calculate the proportion of top-k logit values. We report the experimental results in Table 3.1. The kurtosis of text logits is 2-3 orders of magnitude higher than that of image logits, suggesting that text logits are much "sharper" than image logits. Given that text logits are much longer than image logits, the proportion of top-k logit values also indicates that text logit values are more concentrated than those of image logits.

Input Image / Text	Model	Kurtosis	Top-k logits percentage (%)			
			k=8	k=64	k=512	k=1024
cat.jpg	ResNet-101	975	99.540%	99.642%	99.993%	\
dogs.jpg		782	93.977%	98.433%	99.882%	\
lioness.jpg		995	99.904%	99.973%	99.999%	\
mushroom.jpg		914	99.756%	99.968%	99.998%	\
hat.jpg		906	83.982%	93.643%	99.646%	\
Instruction 1	Qwen-4B	135404	99.991%	99.996%	99.997%	99.998%
Instruction 2		46163	99.998%	99.998%	99.998%	99.998%
Instruction 3		79604	99.982%	99.990%	99.993%	99.994%
Instruction 4		50719	99.528%	99.604%	99.634%	99.651%
Instruction 5		116329	94.778%	94.826%	94.977%	95.081%

Table 3.1: The kurtosis and top-k proportion of image logits and text logits.

Moreover, previous logits distillation methods have not fully utilized the internal rank information of logits [31, 5], even though this ranking information significantly affects LLMs' generation performance. When LLMs generate text, two sampling strategies, top-k sampling and top-p sampling, are commonly used to control the diversity of the generated content. Top-k sampling controls the maximum length of the candidate tokens list, while top-p sampling filters tokens according to cumulative probability. The ranking of logit values impacts the selection process in both strategies, as higher-ranked tokens are more likely to be selected as candidates. Therefore, maintaining rank consistency will better assist the student in imitating the teacher's generating patterns.

3.5 Bi-directional Logits Difference Loss

3.5.1 Overview

The Bi-directional Logits Difference (BiLD) loss is a novel optimization target for task-specific LLM distillation. It filters out the "noise" in the long-tail distribution of LLMs' logits and constructs bi-directional differences that reflect the internal ranking of logits. Our goal is not for the student logits to fully match the teacher's but for the student to effectively learn the key knowledge represented in the non-long-tail part. The detailed process of BiLD is shown in Figure 3.1.

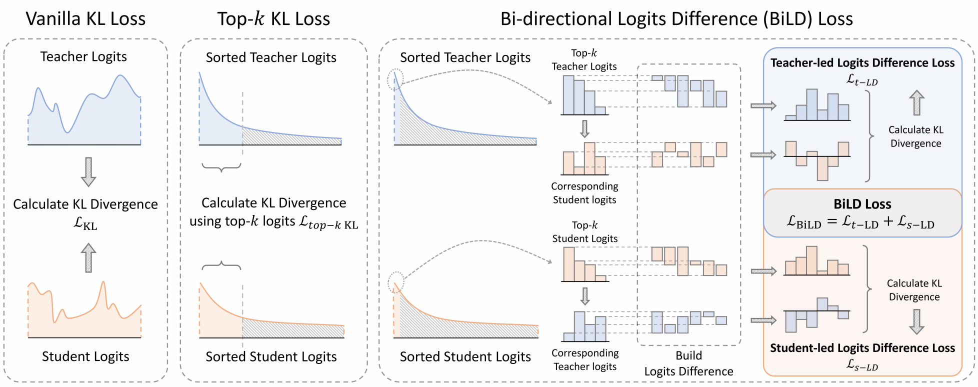

Figure 3.1: An illustration of vanilla KL, top-k KL and our BiLD loss. The vanilla KL loss directly calculates the KL divergence between teacher and student logits, whereas the top-k KL loss uses clipped logits instead of the full logits. In contrast to these methods, our BiLD loss computes KL divergence based on reconstructed "logits differences." The logits difference is derived by calculating the pairwise differences between logit values. We construct two groups of logits differences and compute the KL divergence within each group as a loss: the top-k teacher logits and their corresponding student logits are used to calculate the teacher-led logits difference (t-LD) loss, while the top-k student logits and their corresponding teacher logits are used to calculate the student-led logits difference ((s-LD)) loss. The BiLD loss is the sum of these two losses.

3.5.2 Formal Definition

The BiLD loss consists of two components: the teacher-led logits difference (t-LD) loss and the student-led logits difference (s-LD) loss. Given the similarity between the two components, we explain the process using the calculation of the t-LD loss. First, we select the top-k teacher logits and sort them in descending order to build the teacher-led logits $\mathbf{z}^t_{\text{led}}$:

$$\mathbf{z}^t_{\text{led}} = \left[z^t_{i_1}, z^t_{i_2}, \cdots, z^t_{i_k}\right] \in \mathbb{R}^{1\times k}, \tag{3.5}$$

where the elements of $\mathbf{z}^t_{\text{led}}$ satisfy $z^t_{i_1} \geq z^t_{i_2} \geq \cdots \geq z^t_{i_k}$. Then, we create the corresponding student logits $\mathbf{z}^s_{\text{cor}}$ by selecting the student logit values at the corresponding positions $[i_1, i_2, \cdots, i_k]$:

$$\mathbf{z}^s_{\text{cor}} = \left[z^s_{i_1}, z^s_{i_2}, \cdots, z^s_{i_k}\right] \in \mathbb{R}^{1\times k}. \tag{3.6}$$

Next, we build the logits differences $\mathbf{d}^t_{\text{led}}$ and $\mathbf{d}^s_{\text{cor}}$ by calculating the internal pairwise value differences of $\mathbf{z}^t_{\text{led}}$ and $\mathbf{z}^s_{\text{cor}}$ respectively:

$$\begin{aligned} \mathbf{d}_{\text{led}}^{t} &= \left[z_{i_m}^{t} - z_{i_n}^{t} \mid 1 \leq m < n \leq k\right], \\ \mathbf{d}_{\text{cor}}^{s} &= \left[z_{i_m}^{s} - z_{i_n}^{s} \mid 1 \leq m < n \leq k\right], \end{aligned} \tag{3.7}$$

where both $\mathbf{d}_{\text{led}}^{t}$ and $\mathbf{d}_{\text{cor}}^{s} \in \mathbb{R}^{1\times\frac{k(k-1)}{2}}$. Then we normalize $\mathbf{d}_{\text{led}}^{t}$ and $\mathbf{d}_{\text{cor}}^{s}$ into probabilities:

$$\begin{aligned} \mathbf{p}_{\text{led}}^{t} &= \frac{\exp(\mathbf{z}_{\text{led}}^{t}/T)}{\sum_{\frac{k(k-1)}{2}}^{i=1}\exp(z_{\text{led},i}^{t}/T)}, \\ \mathbf{p}_{\text{cor}}^{s} &= \frac{\exp(\mathbf{z}_{\text{cor}}^{s}/T)}{\sum_{\frac{k(k-1)}{2}}^{i=1}\exp(z_{\text{cor},i}^{s}/T)}. \end{aligned} \tag{3.8}$$

To obtain the teacher-led logits difference loss $\mathcal{L}_{t-\text{LD}}$, we calculate the KL divergence between $\mathbf{p}_{\text{led}}^{t}$ and $\mathbf{p}_{\text{cor}}^{s}$:

$$\mathcal{L}_{t-\text{LD}} = D_{\text{KL}}\left[\mathbf{p}_{\text{led}}^{t} \parallel \mathbf{p}_{\text{cor}}^{s}\right]. \tag{3.9}$$

The calculation of the s-LD loss is similar to that of the t-LD loss. The key difference is that the s-LD loss selects the top-k student logits $\mathbf{z}_{\text{led}}^{s}$ and extracts the corresponding teacher logits $\mathbf{z}_{\text{cor}}^{t}$. Based on these, we can sequentially calculate the logits differences $\mathbf{d}_{\text{led}}^{s}$ and $\mathbf{d}_{\text{cor}}^{t}$ as well as the probabilities $\mathbf{p}_{\text{led}}^{s}$ and $\mathbf{p}_{\text{cor}}^{t}$. The s-LD loss can be represented as:

$$\mathcal{L}_{s-\text{LD}} = D_{\text{KL}}\left[\mathbf{p}_{\text{cor}}^{t} \parallel \mathbf{p}_{\text{led}}^{s}\right]. \tag{3.10}$$

Finally, we obtain the BiLD loss:

$$\mathcal{L}_{\text{BiLD}} = \mathcal{L}_{t-\text{LD}} + \mathcal{L}_{s-\text{LD}}. \tag{3.11}$$

To aid comprehension, we outline the calculation process of the BiLD loss in Algorithm 1.

3.5.3 The Application of BiLD loss: An Example

For simplicity, we explain the BiLD loss process with $k = 3$ and $T = 1$. In a knowledge distillation scenario, we assume there is a knowledgeable teacher model and a smaller student model that aims to learn from the teacher. We input a text sequence, ”a dangerous grey.” At this time, the teacher and student will give their predictions of next token.

Algorithm 1 Calculation of BiLD Loss

Input: teacher logits $\mathbf{z}^t$, student logits $\mathbf{z}^s$, temperature T, hyperparameter k that controls the number of clipped logits
Output: the BiLD loss $\mathcal{L}_{\text{BiLD}}$
1: select top-k teacher logits $\mathbf{z}^t_{\text{led}}$ (Equation 3.5)
2: select corresponding student logits $\mathbf{z}^s_{\text{cor}}$ (Equation 3.6)
3: build the teacher and student logits differences $\mathbf{d}^t_{\text{led}}$ and $\mathbf{d}^s_{\text{cor}}$ (Equation 3.7)
4: normalize differences to probabilities $\mathbf{p}^t_{\text{led}}$ and $\mathbf{p}^s_{\text{cor}}$ (Equation 3.8)
5: calculate the teacher-led logits difference loss $\mathcal{L}_{t-\text{LD}}$ (Equation 3.9)
6: calculate $\mathcal{L}_{s-\text{LD}}$ (Equation 3.10), generally following steps 1-5
7: sum $\mathcal{L}_{t-\text{LD}}$ and $\mathcal{L}_{s-\text{LD}}$ to obtain $\mathcal{L}_{\text{BiLD}}$ (Equation 3.11)

Step 1 Select the top-k teacher logits and sort them in descending order. We need the teacher model to predict the next token in the form of logits. Typically, the length of the logits corresponds to the entire vocabulary. We select the top-3 logit values and arrange them in descending order. Suppose the top-3 predicted tokens by the teacher are [”wolf”, ”cat”, ”sheep”], with corresponding logit values of $\mathbf{z}^t_{\text{led}} = [1.2, 0.9, -0.3]$.
Step 2 Select the corresponding student logits. We select the student logits corresponding to the words [”wolf”, ”cat”, ”sheep”], with values of $\mathbf{z}^s_{\text{cor}} = [-0.5, 0.6, 0.4]$.
Step 3 Construct the logits differences for teacher top-k logits $\mathbf{z}^t_{\text{led}}$ and corresponding student logits $\mathbf{z}^s_{\text{cor}}$. The teacher top-$k$ logits difference is:

$$\mathbf{d}^t_{\text{led}} = [1.2 - 0.9, 1.2 - (-0.3), 0.9 - (-0.3)] = [0.3, 1.5, 1.2] \,.$$

The student corresponding logits difference is:

$$\mathbf{d}^s_{\text{cor}} = [-0.5 - 0.6, -0.5 - 0.4, 0.6 - 0.4] = [-1.1, -0.9, 0.2] \,.$$

Step 4 Apply softmax to $\mathbf{d}^t_{\text{led}}$ and $\mathbf{d}^s_{\text{cor}}$ to get $\mathbf{p}^t_{\text{led}} = [0.1475, 0.4897, 0.3628]$ and $\mathbf{p}^s_{\text{cor}} = [0.1698, 0.2073, 0.6229]$.
Step 5 Calculate the KL divergence between $\mathbf{p}^t_{\text{led}}$ and $\mathbf{p}^s_{\text{cor}}$, which represents the teacher-led logits difference loss. $\mathcal{L}_{t-\text{LD}} = D_{\text{KL}}\left[\mathbf{p}^t_{\text{led}} \parallel \mathbf{p}^s_{\text{cor}}\right] = 0.0089$
Step 6 Select the top-k student logits and sort them in descending order. This time we use the student's prediction of the next token in the logits form. Assume the top-3 predicted tokens by the student are [”rock”, ”cat”, ”toy”], with corresponding logit values of $\mathbf{z}^s_{\text{led}} = [0.8, 0.6, -0.2]$.
Step 7 Select the teacher logits corresponding to the words [”rock”, ”cat”, ”toy”], with values of $\mathbf{z}^t_{\text{cor}} = [-0.4, 0.9, -0.6]$.
Step 8 Construct the logits differences for student top-k logits $\mathbf{z}^s_{\text{led}}$ and corresponding

teacher logits $\mathbf{z}^t_{\text{cor}}$. The student top-$k$ logits difference is:

$$\mathbf{d}^s_{\text{led}} = [0.8 - 0.6, 0.8 - (-0.2), 0.6 - (-0.2)] = [0.2, 1.0, 0.8] \,.$$

The teacher corresponding logits difference is:

$$\mathbf{d}^t_{\text{cor}} = [-0.4 - 0.9, -0.4 - (-0.6), 0.9 - (-0.6)] = [-1.3, 0.2, 1.5] \,.$$

Step 9 Apply softmax. We can get $\mathbf{p}^s_{\text{led}} = [0.1981, 0.4409, 0.3610]$ and $\mathbf{p}^t_{\text{cor}} = [0.0456, 0.2044, 0.7500]$.
Step 10 Calculate the KL divergence between $\mathbf{p}^t_{\text{cor}}$ and $\mathbf{p}^s_{\text{led}}$, which represents the student-led logits difference loss. $\mathcal{L}_{s-\text{LD}} = D_{\text{KL}} \left[\mathbf{p}^t_{\text{cor}} \parallel \mathbf{p}^s_{\text{led}}\right] = 0.0142$
Step 11 The BiLD loss $\mathcal{L}_{\text{BiLD}} = \mathcal{L}_{t-\text{LD}} + \mathcal{L}_{s-\text{LD}} = 0.0231$

3.5.4 Explanation about the Utilization of Logits Ranking

The calculation of the logits difference (Equation 3.7) ensures that the student model learns the ranking information embedded in the teacher logits. We demonstrate this by taking the calculation of the t-LD loss as an example. Since $\mathbf{z}^t_{\text{led}}$ satisfies $z^t_{i_1} \geq z^t_{i_2} \geq \cdots \geq z^t_{i_k}$, it is guaranteed that every element in the teacher-led logits difference $\mathbf{d}^t_{\text{led}}$ is non-negative. For the corresponding student logits difference $\mathbf{d}^s_{\text{cor}}$, consider an element $d^s = z^s_{i_m} - z^s_{i_n}$. If $z^s_{i_m} < z^s_{i_n}$, then $d^s < 0$. In this case, the order $z^s_{i_m} < z^s_{i_n}$ is inconsistent with the order in the teacher logits $z^t_{i_m} > z^t_{i_n}$. Therefore, the sign of the elements in the corresponding logits difference $\mathbf{d}^s_{\text{cor}}$ reflects whether the ranking of the teacher and student logits value pairs is consistent. When calculating $\mathcal{L}_{s-\text{LD}}$, this acts as a constraint, promoting the student logits to align their ranking order with the teacher logits.

Chapter 4

Experiments, Results and Analysis

In this chapter, we introduce our main experimental process, including an analysis of the datasets used and their characteristics (Section 4.1), the selection of baselines for comparison with BiLD loss (Section 4.2), and implementation details (Section 4.3). We provide a detailed description of the main results and their analysis in Section 4.4. A more detailed analysis of the experimental results can be found in Sections 4.5 and 4.6. In Section 4.7 and 4.8, we conduct ablation experiments on the impact of temperature and the k value in BiLD loss.

4.1 Datasets

We evaluate our BiLD loss on 13 NLP datasets: (1) 8 datasets from the SuperGLUE benchmark [32], including BoolQ [33], CB [34], COPA [35], MultiRC [36], ReCoRD [37], RTE [38], WiC [39] and WSC [40]; (2) 5 extra datasets used in previous works about model compression [41, 42], including: Arc-C, Arc-E [43], HellaSwag [44], PIQA [45] and WinoGrande [46]. We observe that these datasets vary significantly in size (the visualization of dataset sizes is presented in Appendix A). Using small datasets alone for SFT and distillation would result in severe overfitting. To prevent unreliable experimental results, we use these datasets collectively for SFT and distillation and evaluate each separately.

4.2 Baselines

We compare BiLD loss with seven baselines: (1) supervised fine-tuning (SFT), where all parameters are adjusted during adaptation to downstream tasks; (2) vanilla KL loss; (3) vanilla KL loss with only top-k logits (short as top-k KL), to demonstrate the performance improvements from noise filtering; (4) three logits distillation methods for vision models, including DKD [2], NKD [22], and NormKD [4]; (5) Reverse KL loss (RKL) used in MiniLLM [8], which has been proven to enhance distillation performance on LLMs.

4.3 Implementation Details

We conduct experiments using the BLOOM and Qwen1.5 (abbreviated as Qwen) models, chosen for their availability in various sizes. Specifically, We employ BLOOM-7B and Qwen-4B as teacher models. For student models, we select BLOOM-3B and BLOOM-1B from the BLOOM series, and 1.8B and 0.5B versions from Qwen.

We perform three epochs of SFT on each teacher model and eight epochs of distillation for each student. Both SFT and distillation processes are conducted with a batch size of 64 and a micro batch size of 2, using the full dataset. We employ a cosine scheduler with an initial learning rate of $1e-5$ for SFT and $2e-5$ for distillation. The warm-up steps are set to 64. During SFT, we utilize the cross entropy loss.

For the different distillation methods we tested, all parameters, except for temperature, are set to their default values. Due to the computational complexity of some distillation methods, we use the vanilla KL loss for the instruction part to expedite the distillation process, and apply different distillation losses to the output part. The temperature T for all loss functions is set to 3. For the top-k KL loss, we set k=1024, and for our proposed BiLD loss, we set k=8.

All our experiments are carried out on 8 NVIDIA A100 GPUs. To reduce memory usage, we employ DeepSpeed during both SFT and distillation processes, along with gradient checkpointing and BFLOAT16 mode [47]. We have not explored the minimum memory requirements. However, in practice, except for the DKD [2], NKD [22], and NormKD [4], experiments involving other methods can be conducted with half of the computational resources. During the evaluation, we employ vLLM [48] for faster inference. The evaluation can be performed with a single NVIDIA A100 GPU. More implementation details can be found in our open-source repository.

4.4 Main Results

Model	Method	Arc-C (Acc.)	Arc-E (Acc.)	boolQ (Acc.)	CB (Acc.)	COPA (Acc.)	HellaSwag (Acc.)	MultiRC (F1a/EM)	PIQA (Acc.)	ReCoRD (F1/Acc.)	RTE (Acc.)	WiC (Acc.)	WinoGrande (Acc.)	WSC (Acc.)	Avg.
BLOOM-7B	Teacher	50.84	68.95	85.26	89.29	81.00	76.08	81.36/40.82	74.92	79.87/78.50	83.03	72.41	71.51	65.38	72.15
BLOOM-3B	SFT	44.15	61.75	84.04	87.50	67.00	57.00	77.09/36.20	70.84	76.05/74.59	78.34	69.75	69.69	64.42	66.56
	Vanilla KL	49.50	68.07	84.50	87.50	76.00	72.60	78.89/36.52	74.27	79.81/78.32	81.59	71.94	70.96	**74.04**	71.21
	RKL	**50.50**	68.42	84.62	87.50	**80.00**	72.20	78.95/36.41	74.48	79.63/78.13	82.31	72.57	71.35	68.27	71.29
	DKD	49.50	**69.82**	85.26	91.07	**80.00**	71.54	77.84/35.68	73.01	79.09/77.65	79.42	**73.20**	70.96	66.35	71.04
	NKD	50.17	67.19	84.01	**92.86**	79.00	72.68	79.69/37.67	73.50	78.50/77.09	81.23	71.32	**72.06**	66.35	71.16
	NormKD	48.16	67.54	**85.35**	89.29	79.00	70.57	77.19/35.57	71.82	78.44/76.98	80.87	72.88	70.48	68.27	70.52
	Top-k KL	47.49	68.25	84.19	87.50	77.00	**72.75**	79.39/37.67	**74.59**	79.40/78.01	**82.67**	72.10	70.80	64.42	70.57
	BiLD (ours)	49.83	67.54	84.86	91.07	**80.00**	72.10	**79.49/37.78**	73.61	**79.96/78.57**	**82.67**	72.88	71.98	71.15	**71.85**
BLOOM-1B	SFT	34.78	53.86	80.76	87.50	64.00	37.39	73.18/30.12	65.72	72.04/70.59	73.65	67.71	67.40	64.42	61.38
	Vanilla KL	45.48	64.39	83.67	87.50	73.00	65.31	77.66/33.37	70.95	77.11/75.67	77.62	68.03	68.43	68.27	67.82
	RKL	45.48	**65.44**	83.43	85.71	74.00	65.70	76.63/32.95	70.78	**77.51/76.10**	79.42	70.69	68.27	64.42	67.88
	DKD	42.47	64.56	**84.10**	85.71	72.00	63.72	75.49/31.79	69.48	75.78/74.46	79.78	**71.79**	68.98	**69.23**	67.55
	NKD	43.14	60.88	82.75	89.29	68.00	63.53	76.94/34.84	70.73	75.31/73.87	77.62	69.44	**69.30**	61.54	66.53
	NormKD	42.81	61.05	83.82	83.93	69.00	62.80	74.13/30.75	67.74	74.49/72.95	77.62	69.91	67.80	65.38	65.81
	Top-k KL	**49.50**	62.11	83.06	89.29	74.00	**65.72**	78.30/34.73	71.22	77.28/75.89	77.98	70.22	**69.30**	60.58	67.97
	BiLD (ours)	44.48	62.98	83.39	**91.07**	**77.00**	64.84	**78.37/35.78**	**72.20**	77.23/75.93	**80.14**	70.53	**69.30**	68.27	**68.92**
Qwen-4B	Teacher	68.23	81.40	87.43	96.43	89.00	86.30	85.85/51.63	82.10	82.59/81.10	87.73	72.73	80.82	74.04	79.92
Qwen-1.8B	SFT	52.17	73.86	83.88	91.07	86.00	72.58	79.95/39.66	75.90	77.37/76.05	84.12	71.79	72.06	61.54	72.36
	Vanilla KL	**55.52**	74.74	85.60	96.43	86.00	77.74	79.46/36.52	76.66	79.24/36.52	85.56	69.59	75.14	64.42	73.98
	RKL	50.84	76.14	85.14	94.64	87.00	77.85	79.52/39.14	76.39	79.49/77.98	84.48	71.47	76.64	69.23	74.38
	DKD	51.84	**77.02**	85.75	**98.21**	85.00	76.90	80.56/39.77	74.54	77.91/76.18	84.48	71.16	76.56	67.31	74.21
	NKD	51.84	73.33	84.53	92.86	**88.00**	77.49	81.98/42.18	76.61	79.03/77.58	84.12	70.85	74.98	66.35	73.90
	NormKD	52.84	76.49	85.26	96.43	85.00	77.24	80.81/40.50	74.76	78.22/76.48	**85.92**	70.53	**76.87**	**70.19**	74.50
	Top-k KL	53.85	76.14	**85.93**	96.43	82.00	**77.99**	81.81/41.03	76.71	**80.08/78.71**	83.39	71.32	75.85	67.31	74.36
	BiLD (ours)	54.85	73.16	84.53	96.43	**88.00**	77.56	**81.49/42.92**	**77.97**	79.87/78.56	85.56	**72.10**	76.01	68.27	**75.07**
Qwen-0.5B	SFT	37.46	62.11	80.40	87.50	77.00	46.71	74.24/28.54	68.44	71.19/69.79	77.26	66.30	69.38	59.62	63.88
	Vanilla KL	43.14	63.68	81.74	85.71	78.00	66.73	75.97/29.07	71.87	72.55/70.91	79.78	70.53	71.35	60.58	67.16
	RKL	**46.49**	64.39	81.53	87.50	**79.00**	67.06	75.37/29.38	71.16	71.46/69.55	**82.31**	69.91	70.80	58.65	67.52
	DKD	40.80	62.98	82.66	82.14	77.00	61.03	72.35/26.55	66.87	65.68/63.20	81.59	70.06	70.64	61.54	65.16
	NKD	41.14	63.86	82.42	94.64	78.00	68.30	79.33/36.20	73.01	74.81/73.35	**82.31**	67.40	72.22	71.15	69.54
	NormKD	41.14	61.40	82.72	83.93	77.00	62.31	74.13/29.07	68.55	67.17/64.79	**82.31**	**71.16**	71.43	62.50	66.02
	Top-k KL	43.14	65.79	82.39	94.64	77.00	68.58	78.83/35.89	71.82	74.30/72.95	**82.31**	69.28	**73.24**	62.50	69.19
	BiLD (ours)	41.81	**67.54**	**83.43**	**96.43**	78.00	**68.99**	**79.72/37.78**	**73.34**	**75.22/73.94**	81.59	69.75	72.22	**74.04**	**70.68**

Table 4.1: The overall performance of various distillation methods and SFT baselines, with best results shown in **bold**. When choosing the best results and calculating the Average Accuracy (Avg.), we use EM score for the MultiRC dataset and Accuracy for the ReCoRD dataset. The instruction templates for each dataset are listed in Appendix D.

We report the experimental results on all 13 datasets in Table 4.1. Across all four sets of distillation, the BiLD loss achieves the highest average accuracy, outperforming SFT, vanilla KL, and the other five methods we tested. In the distillation from Qwen-4B to 0.5B, the BiLD loss showed a significant improvement in average accuracy, surpassing the vanilla KL loss by 3.52%. This improvement is also observed in the distillation from Qwen-4B to 1.8B and from BLOOM-7B to 1B, with improvements of 1.09% and 1.10% over the vanilla KL loss, respectively. A notable case is the distillation from BLOOM-7B to 1B, where the student using vanilla KL loss can easily match the teacher's performance. In this scenario, our BiLD loss still maintained a consistent advantage, showing an average increase of 0.64% over the vanilla KL loss. In contrast, other methods achieve only marginal performance improvements or even experience declines. The robust performance of the BiLD loss across various distillation scenarios underscores its superiority and effectiveness.

4.5 Analysis of the Effectiveness of Clipping Logits

The experimental results in Table 4.1 indicate that in three sets of distillation, simply clipping the full logits to the top-k logits improves the performance of the KL loss. This suggests that filtering out the noise in the logits' long tail distribution can be a practical and straightforward approach to enhancing distillation performance. Our statistics show that the top-1024 logits cover over 99% of the probability in both Qwen-4B and BLOOM-7B teachers. For computational simplicity, we set k=1024 for the top-k KL loss to verify that excluding the long-tail distribution of logits can improve distillation results.

4.6 Analysis of Performance at the Logit Level

To demonstrate the performance of different distillation methods at the logit level, we introduce a new metric, top-k overlap (overlap@k). Consider an instruction I represented as a sequence of tokens. We denote the output tokens generated by the teacher with I as A^t, and the concatenated sequence of tokens as $C^t = I \oplus A^t$. The logits sequence for the teacher's output part can be represented as $\mathbf{Z}^t = [\mathbf{z}_1^t, \mathbf{z}_2^t, \cdots, \mathbf{z}_M^t]$, where M is the length of A^t. The element $\mathbf{z}_i^t$ within $\mathbf{Z}^t$ is the logits at the i-th position of the teacher's output part. By feeding the whole C^t into the student, we denote the student logits sequence corresponding to the positions of A^t as $\mathbf{Z}^s = [\mathbf{z}_1^s, \mathbf{z}_2^s, \cdots, \mathbf{z}_M^s]$. Consequently, we define the top-k overlap as:

$$\text{overlap@}k = \frac{1}{M}\sum_{i=1}^{M}\frac{\text{topk}(\mathbf{z}_i^t) \cap \text{topk}(\mathbf{z}_i^s)}{k}, \tag{4.1}$$

where $\text{topk}(\cdot)$ is a function to select tokens corresponding to the top-k logit values. The metric overlap@k measures the average degree of overlap for the top-k logits corresponding tokens at the same positions in $\mathbf{Z}^t$ and $\mathbf{Z}^s$. Specifically, overlap@1 evaluates if the token corresponding to the highest logit values of both the teacher's and the student's outputs match at each position. Overlap@1 can measure the efficacy of LLMs in greedy search mode, where LLMs generate text based on the token with the highest probability. For $k > 1$, overlap@k calculates the ratio of overlapping

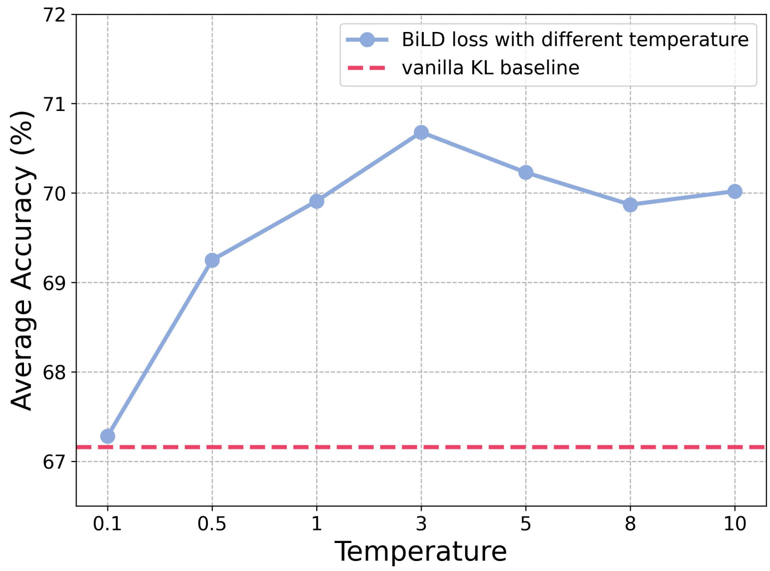

Figure 4.1: Impact of model temperature.

tokens corresponding to the top-k logits from both student and teacher at each position, reflecting how well the student imitates the important parts of teacher logits. From another perspective, overlap@1 measures the performance of models in scenarios where there is only one correct answer, while overlap@$k(k > 1)$ reflects the degree of similarity between the student and teacher responses in open-ended scenarios.

According to the results in Table 4.2, our proposed BiLD loss notably enhances overlap@8 while maintaining a competitive overlap@1. Compared to other methods, students trained with BiLD loss better imitate the teacher's primary behaviors at the logit level, indicating that BiLD loss helps student logits align with the important part of teacher logits.

4.7 Impact of Temperature

To understand the impact of temperature during the distillation of BiLD loss, we vary the temperature parameter $T \in \{0.1, 0.5, 1, 3, 5, 8, 10\}$ while keeping other hyperparameters and model architectures constant. The experimental results, as depicted in Figure 4.1, indicate that lower temperatures significantly degrade the performance of BiLD loss. We choose T=3, which yields the best performance, for our distillation experiments.

Model	Method	Overlap@1	Overlap@8
BLOOM-3B	SFT	74.89	44.61
	Vanilla KL	**82.51**	54.64
	RKL	82.31	54.64
	DKD	74.00	52.39
	NKD	82.11	53.25
	NormKD	48.80	36.95
	Top-k KL	81.67	55.73
	BiLD	81.72	**56.57**
BLOOM-1B	SFT	74.40	40.71
	Vanilla KL	80.82	51.91
	RKL	80.71	51.58
	DKD	75.44	48.83
	NKD	79.59	50.01
	NormKD	73.56	42.70
	Top-k KL	80.20	50.87
	BiLD	**81.21**	**52.86**
Qwen-1.8B	SFT	93.30	53.28
	Vanilla KL	94.35	68.02
	RKL	94.31	67.93
	DKD	94.09	67.01
	NKD	94.02	65.01
	NormKD	94.26	68.32
	Top-k KL	**94.43**	67.55
	BiLD	94.39	**70.97**
Qwen-0.5B	SFT	91.67	47.29
	Vanilla KL	92.72	61.81
	RKL	92.54	61.65
	DKD	91.50	56.62
	NKD	92.88	59.11
	NormKD	91.76	58.16
	Top-k KL	93.11	64.00
	BiLD	**93.23**	**68.58**

Table 4.2: The top-1 and top-8 overlap of different distillation methods on 4 distillation settings.

4.8 Impact of the k Value in BiLD Loss

The hyperparameter k controls the length of clipped logits in BiLD loss. We experiment with $k \in \{1, 2, 4, 8, 12, 16, 32\}$ and evaluate the distillation results using average accuracy as well as top-1, top-8, and top-32 overlap, as defined in Equation 4.1. We report the results in Figure 4.2 and Table 4.3. Smaller k values ($k \in \{1, 2\}$) lead to

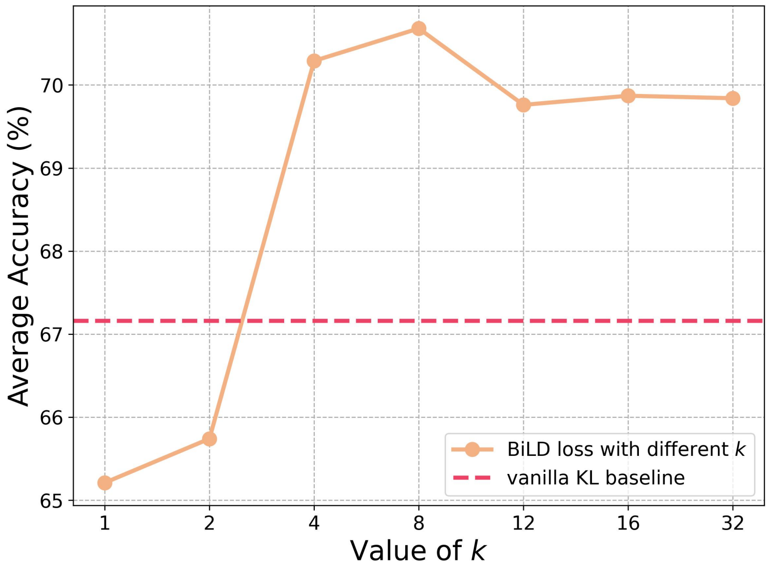

Figure 4.2: Impact of k values in BiLD loss.

overly short logits, resulting in poor performance. As k increases, both average accuracy and overlap@1 rise and then stabilize, while significant improvements can be seen in overlap@8 and overlap@32. However, higher k values lead to increased computational costs. Considering the trade-off between computation time and performance, we select k=8 for BiLD loss in our experiments.

top-k	Overlap@1	Overlap@8	Overlap@32
k=1	91.93	49.57	38.91
k=2	91.97	49.60	38.93
k=4	93.21	63.64	47.05
k=8	93.23	68.58	52.98
k=12	93.16	69.46	56.00
k=16	93.17	69.56	57.75
k=32	93.12	69.29	60.77

Table 4.3: Top-1, top-8 and top-32 overlap.

Chapter 5

Discussion and Conclusion

5.1 Conclusion

In this work, we propose the Bi-directional Logits Difference (BiLD) loss, a novel optimization objective for distilling LLMs. The BiLD loss enhances distillation performance by filtering out long-tail noise and leveraging internal ranking information from LLMs' logits. It achieves superior distillation performance using only the top-8 logits compared to vanilla KL loss using full logits and other distillation methods. Our extensive experiments across diverse datasets and model architectures confirm the effectiveness of the BiLD loss, demonstrating its ability to more efficiently capture key knowledge from the teacher model.

5.2 Limitations

Our approach falls within the realm of logits distillation, necessitating access to teacher logits. However, powerful LLMs such as GPT-4 [18] and Gemini [19] currently provide only output text or incomplete logits access, making our method unable to utilize these highly capable LLMs as teachers. Additionally, our Bi-directional Logits Difference (BiLD) loss requires shared vocabularies between the teacher and student models to ensure vector space alignment.

Another challenge lies in the computational complexity of our BiLD loss, particularly during the construction of logits differences using top-k logits. Although we demonstrate that using only the top-8 logits achieves better results than the vanilla KL loss, increasing the number of clipped logits leads to a rapid escalation in our method's time overhead, which becomes a practical concern.

Furthermore, our approach directly clips the long-tail part of logits during distillation. While this approach improves performance, it unavoidably results in the loss of knowledge contained within the long-tail distribution. Investigating methods to better utilize the knowledge hidden in the long-tail distribution represents a promising avenue for future research.

5.3 Reflections

From a sustainability perspective, the method we propose falls within the domain of knowledge distillation. It enables small student models to compete with larger teacher models. Substituting distilled smaller models for larger ones during inference not only significantly reduces computational power consumption and carbon emissions without noticeable performance degradation but also holds obvious significance for sustainable development.

References

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. [Pages 1 and 7.]

[2] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962. [Pages 1, 7, 18, and 34.]

[3] P. Yang, M.-K. Xie, C.-C. Zong, L. Feng, G. Niu, M. Sugiyama, and S.-J. Huang, "Multi-label knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 271–17 280. [Page 1.]

[4] Z. Chi, T. Zheng, H. Li, Z. Yang, B. Wu, B. Lin, and D. Cai, "Normkd: Normalized logits for knowledge distillation," *arXiv preprint arXiv:2308.00520*, 2023. [Pages 1, 8, and 18.]

[5] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [Pages 1, 2, and 12.]

[6] L. Tu, R. Y. Pang, S. Wiseman, and K. Gimpel, "Engine: Energy-based inference networks for non-autoregressive machine translation," *arXiv preprint arXiv:2005.00850*, 2020. [Pages 1 and 8.]

[7] H. Lee, Y. Park, H. Seo, and M. Kang, "Self-knowledge distillation via dropout," *Computer Vision and Image Understanding*, vol. 233, p. 103720, 2023. [Page 1.]

[8] Y. Gu, L. Dong, F. Wei, and M. Huang, "Minillm: Knowledge distillation of large language models," in *The Twelfth International Conference on Learning Representations*, 2023. [Pages 1, 8, and 18.]

[9] X. Cui, Y. Qin, Y. Gao, E. Zhang, Z. Xu, T. Wu, K. Li, X. Sun, W. Zhou, and H. Li, "Sinkhorn distance minimization for knowledge distillation," *arXiv preprint arXiv:2402.17110*, 2024. [Pages 1 and 8.]

[10] A. Chan, H. Silva, S. Lim, T. Kozuno, A. R. Mahmood, and M. White, "Greedification operators for policy optimization: Investigating forward and reverse kl divergences," *Journal of Machine Learning Research*, vol. 23, no. 253, pp. 1–79, 2022. [Page 1.]

[11] C. T. Li and F. Farnia, "Mode-seeking divergences: theory and applications to gans," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 8321–8350. [Page 1.]

[12] BigScience Workshop, "Bloom (revision 4ab0472)," 2022. [Online]. Available: https://huggingface.co/bigscience/bloom [Pages 2 and 6.]

[13] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023. [Pages 2, 6, and 11.]

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. [Page 5.]

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Page 5.]

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. [Page 6.]

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. [Page 6.]

[18] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Pages 6 and 25.]

[19] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. [Pages 6 and 25.]

[20] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541. [Page 7.]

[21] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 276–24 285. [Page 7.]

[22] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li, "From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 185–17 194. [Pages 7, 18, and 34.]

[23] J. Ko, S. Kim, T. Chen, and S.-Y. Yun, "Distillm: Towards streamlined distillation for large language models," *arXiv preprint arXiv:2402.03898*, 2024. [Page 8.]

[24] Z. Yuan, Y. Shang, Y. Zhou, Z. Dong, C. Xue, B. Wu, Z. Li, Q. Gu, Y. J. Lee, Y. Yan *et al.*, "Llm inference unveiled: Survey and roofline model insights," *arXiv preprint arXiv:2402.16363*, 2024. [Page 8.]

[25] C. Liang, S. Zuo, Q. Zhang, P. He, W. Chen, and T. Zhao, "Less is more: Task-aware layer-wise distillation for language model compression," in *International Conference on Machine Learning*. PMLR, 2023, pp. 20 852–20 867. [Page 8.]

[26] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist, and O. Bachem, "On-policy distillation of language models: Learning from self-generated mistakes," in *The Twelfth International Conference on Learning Representations*, 2024. [Page 8.]

[27] G. Sahu, O. Vechtomova, D. Bahdanau, and I. H. Laradji, "Promptmix: A class boundary augmentation method for large language model distillation," *arXiv preprint arXiv:2310.14192*, 2023. [Page 8.]

[28] Y. Fu, H. Peng, L. Ou, A. Sabharwal, and T. Khot, "Specializing smaller language models towards multi-step reasoning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 421–10 430. [Page 8.]

[29] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, and Y. Choi, "Symbolic chain-of-thought distillation: Small models can also" think" step-by-step," *arXiv preprint arXiv:2306.14050*, 2023. [Page 8.]

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [Page 11.]

[31] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022. [Page 12.]

[32] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019. [Page 17.]

[33] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," in *NAACL*, 2019. [Page 17.]

[34] M.-C. De Marneffe, M. Simons, and J. Tonhauser, "The commitmentbank: Investigating projection in naturally occurring discourse," in *proceedings of Sinn und Bedeutung*, vol. 23, no. 2, 2019, pp. 107–124. [Page 17.]

[35] M. Roemmele, C. A. Bejan, and A. S. Gordon, "Choice of plausible alternatives: An evaluation of commonsense causal reasoning," in *2011 AAAI Spring Symposium Series*, 2011. [Page 17.]

[36] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 252–262. [Page 17.]

[37] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, and B. Van Durme, "Record: Bridging the gap between human and machine commonsense reading comprehension," *arXiv preprint arXiv:1810.12885*, 2018. [Page 17.]

[38] D. Giampiccolo, B. Magnini, I. Dagan, and W. B. Dolan, "The third pascal recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, 2007, pp. 1–9. [Page 17.]

[39] M. T. Pilehvar and J. Camacho-Collados, "Wic: the word-in-context dataset for evaluating context-sensitive meaning representations," *arXiv preprint arXiv:1808.09121*, 2018. [Page 17.]

[40] H. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012. [Page 17.]

[41] X. Ma, G. Fang, and X. Wang, "Llm-pruner: On the structural pruning of large language models," *Advances in neural information processing systems*, vol. 36, pp. 21 702–21 720, 2023. [Page 17.]

[42] V. Egiazarian, A. Panferov, D. Kuznedelev, E. Frantar, A. Babenko, and D. Alistarh, "Extreme compression of large language models via additive quantization," *arXiv preprint arXiv:2401.06118*, 2024. [Page 17.]

[43] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018. [Page 17.]

[44] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" *arXiv preprint arXiv:1905.07830*, 2019. [Page 17.]

[45] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439. [Page 17.]

[46] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," *Communications of the ACM*, vol. 64, no. 9, pp. 99–106, 2021. [Page 17.]

[47] D. Kalamkar, D. Mudigere, N. Mellempudi, D. Das, K. Banerjee, S. Avancha, D. T. Vooturi, N. Jammalamadaka, J. Huang, H. Yuen *et al.*, "A study of bfloat16 for deep learning training," *arXiv preprint arXiv:1905.12322*, 2019. [Page 18.]

[48] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. [Page 18.]

Appendix A

Details about Datasets

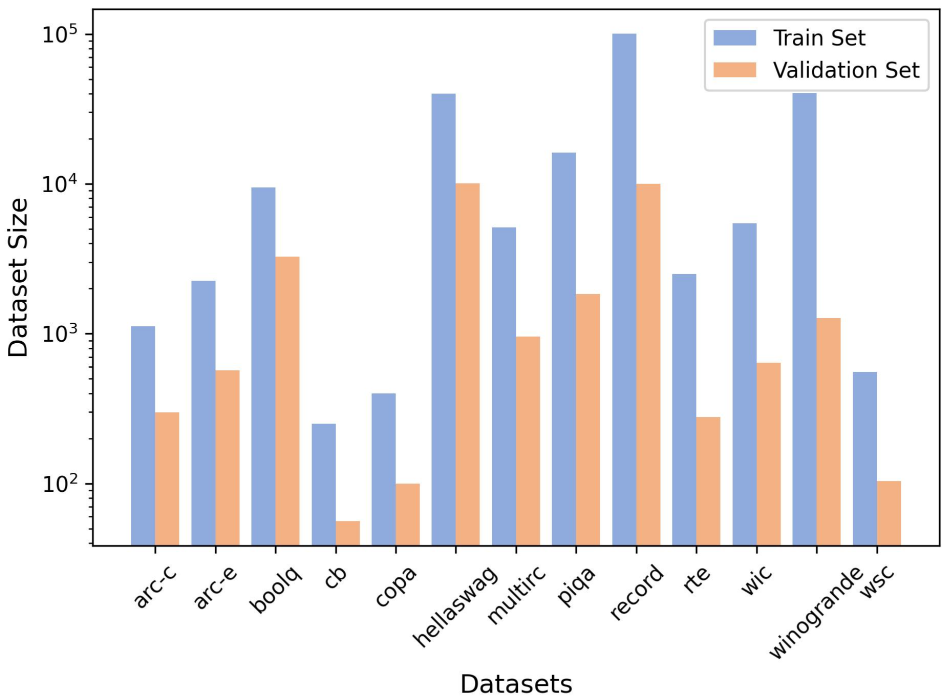

Figure A.1: A visualization of the dataset sizes. There are significant size differences among the datasets, with the smallest datasets (CB, COPA, WSC) differing by three orders of magnitude from the largest dataset (ReCoRD).

Appendix B

Calculation Efficiency of BiLD

In Figure B.1, we visualize the distillation speed of various methods during the distillation from Qwen-4B to 0.5B. Compared to the vanilla KL loss, our BiLD loss achieves better distillation performance with an acceptable increase in training time. Among all methods, DKD [2] and NKD [22], which are designed for vision models, have the slowest computation speeds due to the calculation of numerous intermediate variables. In contrast, the computation speeds of RKL, NormKD, and top-k KL are comparable to the vanilla KL loss.

In the code implementation, the BiLD loss consists of two main steps: selecting the top-k logit values and calculating the internal pairwise differences. Our analysis reveals that the latter step is where the significant time expenditure occurs. The time complexity for computing the internal pairwise differences is $\mathcal{O}(n^2)$, and it frequently necessitates extracting values from the tensor. This has become the time bottleneck for the BiLD loss.

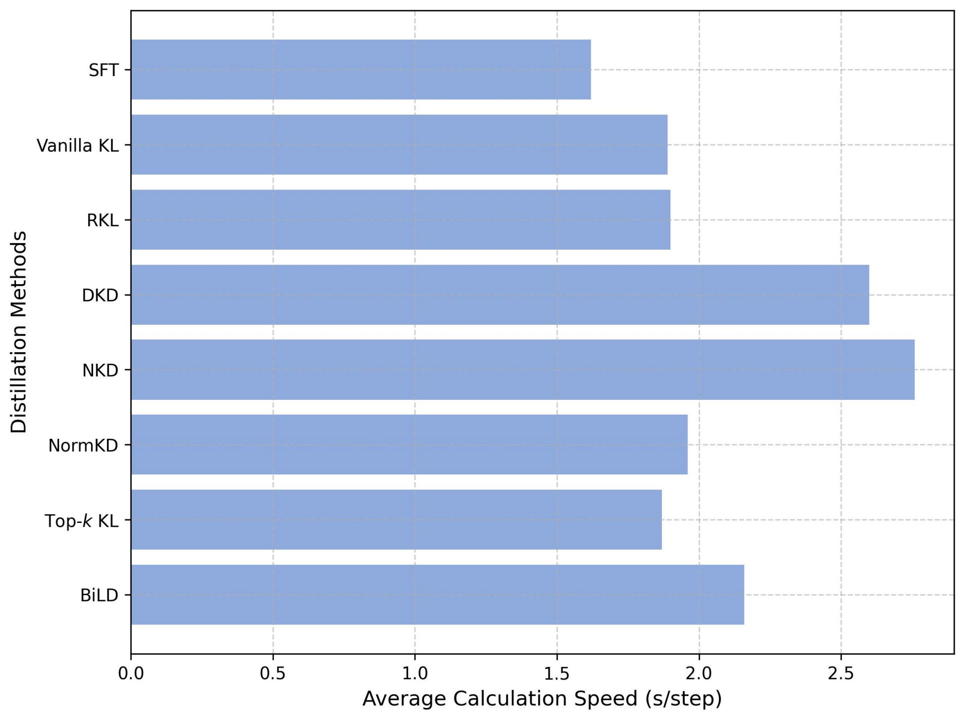

Figure B.1: The average calculation speed of different distillation methods.

Appendix C

Toy Experiment to Compare Vision Model and LLMs’ Logits

The five images we used in the toy experiments are shown in Figure C.1, and the five sets of instructions are in Table C.1.

((a)) cat.jpg

((b)) dogs.jpg

((c)) lioness.jpg

((d)) mushroom.jpg

((e)) hat.jpg

Figure C.1: Five images used in the toy experiment.

	Instructions Content
Instruction 1	Question: A mass of air is at an elevation of 1000 meters in the low pressure center of a Northern Hemisphere storm. Which of the following best describes the motion of air particles in this air mass due to storm conditions and the rotation of Earth as the air mass moves outward? Choices: ['Air particles move up and to the left.', 'Air particles move up and to the right.', 'Air particles move down and to the left.', 'Air particles move down and to the right.'] Answer:
Instruction 2	Premise: A: No, I don't either. B: Uh, I mean it's, you know it, A: I don't think it's going to change very much Hypothesis: it's going to change very much Question: Determine whether the premise entails the hypothesis or not. Choices: ['entailment', 'neutral', 'contradiction'] Answer:
Instruction 3	Goal: Keep laptop from overheating. Choose the most sensible solution to achieve the goal. Choices: ['Use on top of egg carton.', 'Use on top of egg shells.'] Answer:
Instruction 4	Choose the most sensible text to replace the "_" in the following sentence: Kyle asked Brett for some tips on healthy eating because _ has recently lost weight. Choices: ['Kyle', 'Brett'] Answer:
Instruction 5	Meanwhile, in the forest, the elephants are calling and hunting high and low for Arthur and Celeste , and their mothers are very worried. Fortunately, in flying over the town, an old marabou bird has seen them and come back quickly to tell the news. Question: In the above text, does 'their' refer to 'their mothers'? Choices:['true', 'false'] Answer:

Table C.1: Five instructions used in the toy experiment.

Appendix D

Templates

The template of each dataset can be seen in Table D.1.

Dataset	Template
Arc-C	Question: A scientist is measuring the amount of movement along a fault. Which tool is best used for making this measurement? Choices: ['barometer', 'stopwatch', 'meter stick', 'magnifying lens'] Answer:
Arc-E	Question: Which color shirt will reflect the most light on a hot, sunny day? Choices: ['black', 'blue', 'red', 'white'] Answer:

BoolQ	Turn on red –Right turns on red are permitted in many regions of North America. While Western states have allowed it for more than 50 years; eastern states amended their traffic laws to allow it in the 1970s as a fuel-saving measure in response to motor fuel shortages in 1973. The Energy Policy and Conservation Act of 1975 required in §362(c)(5) that in order for a state to receive federal assistance in developing mandated conservation programs, they must permit right turns on red lights. All 50 states, the District of Columbia, Guam, and Puerto Rico have allowed right turns on red since 1980, except where prohibited by a sign or where right turns are controlled by dedicated traffic lights. (On January 1, 1980, Massachusetts became the last US state to allow right turns on red.) The few exceptions include New York City, where right turns on red are prohibited, unless a sign indicates otherwise. Question: is it legal to turn right on red in california? Choices: ['true', 'false'] Answer:
CB	Premise: B: And I've worked in the hospital for fifteen years and I've taken care of a few AIDS patients. A: Uh-huh. B: Uh, when they asked us did we want to, uh, keep it the same or, uh, spend more, spend less, uh, I think right now what they're spending is adequate. Uh, for my personal opinion. Uh, because I think it's something that's going to take them a while to come up with a, uh, vaccine for. A: Yeah. Uh-huh. Uh-huh. B: I don't think it's going to be that easy to come up with Hypothesis: it is going to be that easy to come up with Question: Determine whether the premise entails the hypothesis or not. Choices: ['entailment', 'neutral', 'contradiction'] Answer:
COPA	Premise: The woman betrayed her friend. Question: What could be the possible effect of the premise? Choices: ['Her friend sent her a greeting card.', 'Her friend cut off contact with her.'] Your answer:

HellaSwag	Please choose the most appropriate text to complete the passage below: Passage: [header] How to clean a plastic retainer [title] Rinse the retainer with warm or cold water. [step] The water will prep your retainer for the cleaning process. [title] Apply a mild soap onto a toothbrush. Choices: ['[step] Rinse the retainer under the faucet bowl with warm water. Suds will accumulate on the toothbrush.', '[step] Rinse the retainer slowly from top to bottom and then wipe it on the toothbrush. Soap can effectively clean a plastic retainer but it can potentially cause irritation.', '[step] If you are using an old toothbrush, you may brush the bristles for pleasure. Fill a bucket, then fill it with a cup of liquid soap.', '[step] You can use liquid castile soap or a mild dishwashing detergent. Additionally, use a soft-bristled toothbrush.'] Answer:

MultiRC	Passage: One of the most dramatic changes in priorities proposed by the City Council would shift $25.6 million from funding for court-appointed lawyers to the Legal Aid Society. In a document released yesterday to justify its reordered priorities, the Council contended that Legal Aid can achieve greater economies of scale than lawyers appointed pursuant to Article 18-B of the County Law. The Council document also noted that ïnexplicablÿ 18-B lawyers are handling 50 percent of the indigent criminal cases in New York City, even though their mandate is to handle only multi-defendant cases where the Legal Aid Society had a conflict. In past years, the City Council had consistently added $5.6 million to the $54.7 million proposed for the Legal Aid Society by former Mayor Giuliani, bringing the total to just a shade over $60 million. But this year for the first time, the Council is proposing shifting more than $20 million in funds earmarked by the Mayor for 18-B lawyers to the Legal Aid Society, which would increase its total funding to $80.4 million. That would reflect a jump in its current finding of about one-third. Meantime, the City Council proposed slashing the Mayor's allocation of $62.8 million for 18-B lawyers by 66 percent, to $21.4 million. Question: By increasing current funding to the Legal Aid society by $25.6 million, how much is the Council increasing their funding? Choices: ['$60 million', '$62.8 million', 'One third', '$54.7 million', '$80.4 million'] Note: 1. there can be multiple correct answers. 2. each line contains one answer. 3. If no correct answer, reply 'none'. Your answer:
PIQA	Goal: how do you flood a room? Choose the most sensible solution to achieve the goal. Choices: ['fill it with objects.', 'fill it with water.'] Answer:

ReCoRD	A father has admitted killing his 13-year-old son by giving him a morphine tablet when the boy complained that he was feeling ill. Kevin Morton gave his son Kye Backhouse an extremely strong painkiller, a court heard - a mistake which he says he will 'have to try and live with it for the rest of my life'. He could now face jail after pleading guilty to manslaughter over the teenager's death at Preston Crown Court. Tragedy: Kevin Morton, right, has admitted killing his son Kye Backhouse, left, by giving him morphine 'Happy-go-lucky' Kye was found dead at his home in Barrow-in-Furness, Cumbria in October last year.@highlight Morton gave Kye Backhouse a strong painkiller when he was ill@highlight teenager subsequently died and his father has admitted manslaughter@highlight, 49, faces jail when he is sentenced next month Question: Death: @placeholder, 23, complained of feeling unwell before his father gave him the strong painkiller What is the @placeholder? Answer:
RTE	Premise: Euro Disney is one of the most popular theme parks of USA. Hypothesis: Euro-Disney is an Entertainment Park. Question: Determine whether the premise entails the hypothesis or not. Choices: ['entailment', 'not_entailment'] Answer:
WiC	Sentence1: An early movie simply showed a long kiss by two actors of the contemporary stage. Sentence2: We went out of town together by stage about ten or twelve miles. Question: Does 'stage' have the same meaning in both sentences? Choices: ['true', 'false'] Answer:
WinoGrande	Choose the most sensible text to replace the '_' in the following sentence: Natalie was less religous than Patricia, therefore _ attended church services more often on Sundays. Choices: ['Natalie', 'Patricia'] Answer:

WSC	The mothers of Arthur and Celeste have come to the town to fetch them. They are very happy to have them back, but they scold them just the same because they ran away. Question: In the above text, does 'them' refer to 'mothers'? Choices:['true', 'false'] Answer:

Table D.1: The template of each dataset.

TRITA – EECS-EX 2024:0000
Stockholm, Sweden 2024

www.kth.se

€€€€ For DIVA €€€€

{
”Author1”: { ”Last name”: ”Li”,
”First name”: ”Minchong”,
”Local User Id”: ”u1dlm95i”,
”E-mail”: ”mincli@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
}
},
”Cycle”: ”2”,
”Course code”: ”DA233X”,
”Credits”: ”30.0”,
”Degree1”: {”Educational program”: ”Master’s Programme, Machine Learning, 120 credits”
,”programcode”: ”TMAIM”
,”Degree”: ”Master’s Programme, Machine Learning, 120 credits”
,”subjectArea”: ”Machine Learning”
},
”Title”: {
”Main title”: ”Design Novel Effective Method for Large Language Model Compression”,
”Subtitle”: ”BiLD: Bi-directional Logits Difference Loss for Large Language Model Distillation”,
”Language”: ”eng” },
”Alternative title”: {
”Main title”: ”Utformning en Ny Effektiv Metod för Komprimering av Stor Språkmodell”,
”Subtitle”: ”BiLD: Bi-directional Logits Difference Loss för Destillering av Stor Språkmodell”,
”Language”: ”swe”
},
”Supervisor1”: { ”Last name”: ”Kheirabadi”,
”First name”: ”Amirhossein Layegh”,
”Local User Id”: ”u14l95xy”,
”E-mail”: ”amlk@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Supervisor2”: { ”Last name”: ”Song”,
”First name”: ”Xiaohui”,
”E-mail”: ”songxiaohui@oppo.com”,
”Other organisation”: ”OPPO Research Institute, Department of Audio and Language”
},
”Examiner1”: { ”Last name”: ”Payberah”,
”First name”: ”Amir Hossein”,
”Local User Id”: ”u1a73o9d”,
”E-mail”: ”payberah@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Cooperation”: { ”Partner_name”: ”OPPO”},
”National Subject Categories”: ”102, 10208”,
”Other information”: {”Year”: ”2024”, ”Number of pages”: ”1,43”},
”Copyrightleft”: ”copyright”,
”Series”: { ”Title of series”: ”TRITA – EECS-EX” , ”No. in series”: ”2024:0000” },
”Opponents”: { ”Name”: ”A. B. Normal & A. X. E. Normalè”},
”Presentation”: { ”Date”: ”2022-03-15 13:00”
,”Language”:”eng”
,”Room”: ”via Zoom https://kth-se.zoom.us/j/ddddddddddd”
,”Address”: ”Isafjordsgatan 22 (Kistagången 16)”
,”City”: ”Stockholm” },
”Number of lang instances”: ”2”,
”Abstract[eng]”: €€€€

```
In recent years, \gls{LLMs} have shown exceptional capabilities across various \gls{NLP} tasks. However, such
impressive performance often comes with the trade-off of an increased parameter size, posing significant
challenges for widespread deployment. \Gls{KD} provides a solution by transferring knowledge from a large
teacher model to a smaller student model. In this thesis, we explore the task-specific distillation of
\gls{LLMs} at the logit level. Our investigation reveals that the logits of fine-tuned \gls{LLMs} exhibit a
more extreme long-tail distribution than those from vision models. Moreover, existing logits distillation
methods often struggle to effectively utilize the internal ranking information from the logits. To address
```

this, we propose the \textbf{Bi}-directional \textbf{L}ogits \textbf{D}ifference (BiLD) loss. The BiLD loss filters out the long-tail "noise" by utilizing only top-k teacher and student logits, and leverages the internal logits ranking information by constructing logits differences. To evaluate BiLD loss, we conduct comprehensive experiments on 13 datasets using two types of \gls{LLMs}. Our results show that the BiLD loss, with only the top-\textbf{8} logits, outperforms supervised fine-tuning (SFT), vanilla \gls{KL} loss, and five other distillation methods from both \gls{NLP} and \gls{CV} fields.

€€€€,
"Keywords[eng]": €€€€
Large Language Model, Model Compression, Knowledge Distillation €€€€,
"Abstract[swe]": €€€€

% \generalExpl{Enter your Swedish abstract or summary here!}
\sweExpl{Alla avhandlingar vid KTH \textbf{måste ha} ett abstrakt på både \textit{engelska} och \textit{svenska}.\\
% Om du skriver din avhandling på svenska ska detta göras först (och placera det som det första abstraktet) - och du bör revidera det vid behov.}

% \engExpl{If you are writing your thesis in English, you can leave this until the draft version that goes to your opponent for the written opposition. In this way, you can provide the English and Swedish abstract/summary information that can be used in the announcement for your oral presentation.\\If you are writing your thesis in English, then this section can be a summary targeted at a more general reader. However, if you are writing your thesis in Swedish, then the reverse is true – your abstract should be for your target audience, while an English summary can be written targeted at a more general audience.\\This means that the English abstract and Swedish sammnfattning or Swedish abstract and English summary need not be literal translations of each other.}

% \warningExpl{Do not use the \textbackslash glspl\{\} command in an abstract that is not in English, as my programs do not know how to generate plurals in other languages. Instead, you will need to spell these terms out or give the proper plural form. In fact, it is a good idea not to use the glossary commands at all in an abstract/summary in a language other than the language used in the \texttt{acronyms.tex file} - since the glossary package does \textbf{not} support use of more than one language.}

% \engExpl{The abstract in the language used for the thesis should be the first abstract, while the Summary/Sammanfattning in the other language can follow}

På senare år har stora språkmodeller (LLMs) visat exceptionella förmågor över olika NLP-uppgifter. Men sådan imponerande prestanda kommer ofta med en kompromiss i form av ökad parameterstorlek, vilket innebär betydande utmaningar för utbredd användning. Kunskapsdistillation (KD) erbjuder en lösning genom att överföra kunskap från en stor lärarmodell till en mindre studentmodell. I denna avhandling utforskar vi uppgiftsspecifik distillation av stora språkmodeller på logitnivå. Vår undersökning visar att logiterna från finjusterade LLMs uppvisar en mer extrem långsvansfördelning än de från visionsmodeller. Dessutom kämpar befintliga metoder för logitdistillation ofta med att effektivt utnyttja den interna rankningsinformationen från logiterna. För att åtgärda detta föreslår vi förlustfunktionen BiLD (Bi-directional Logits Difference). BiLD-förlusten filtrerar bort långsvansens "brus" genom att endast använda de översta k lärar- och studentlogiterna, och utnyttjar den interna logitrankningsinformationen genom att konstruera logitskillnader. För att utvärdera BiLD-förlusten genomför vi omfattande experiment på 13 datamängder med två typer av LLMs. Våra resultat visar att BiLD-förlusten, med endast de översta 8 logiterna, överträffar både övervakad finjustering (SFT), vanilj-KL-förlust och fem andra distillationsmetoder från både NLP- och CV-fälten.

€€€€,
"Keywords[swe]": €€€€
Stor Språkmodell, Modellkompression, Kunskapsdestillation €€€€,
}

acronyms.tex

```
%%% Local Variables:
%%% mode: latex
%%% TeX-master: t
%%% End:
% The following command is used with glossaries-extra
\setabbreviationstyle[acronym]{long-short}
% The form of the entries in this file is \newacronym{label}{acronym}{phrase}
%                                      or \newacronym[options]{label}{acronym}{phrase}
% see "User Manual for glossaries.sty" for the  details about the options, one example is shown below
% note the specification of the long form plural in the line below
\newacronym[longplural={Debugging Information Entities}]{DIE}{DIE}{Debugging Information Entity}
%
% The following example also uses options
\newacronym[shortplural={OSes}, firstplural={operating systems (OSes)}]{OS}{OS}{operating system}

% note the use of a non-breaking dash in long text for the following acronym
\newacronym{IQL}{IQL}{Independent Q-Learning}

% example of putting in a trademark on first expansion
\newacronym[first={NVIDIA OpenSHMEM Library (NVSHMEM\texttrademark)}]{NVSHMEM}{NVSHMEM}{NVIDIA OpenSHMEM Library}

\newacronym{BiLD}{BiLD}{Bi-directional Logits Difference}
\newacronym{CV}{CV}{Computer Vision}
\newacronym{KD}{KD}{Knowledge Distillation}
\newacronym{KTH}{KTH}{KTH Royal Institute of Technology}
\newacronym{KL}{KL}{Kullback - Leibler}
\newacronym{LLMs}{LLMs}{Large Language Models}
\newacronym{NLP}{NLP}{Natural Language Processing}
\newacronym{RKL}{RKL Divergence}{Reverse Kullback - Leibler Divergence}
\newacronym{SFT}{SFT}{Supervised Fine-tuning}
\newacronym{SOTA}{SOTA}{State-of-the-art}


\newacronym{VM}{VM}{Wireless Fidelity}
\newacronym{LAN}{LAN}{Wireless Fidelity}
% note the use of a non-breaking dash in the following acronym
\newacronym{WiFi}{Wi-Fi}{Wireless Fidelity}

\newacronym{WLAN}{WLAN}{Wireless Local Area Network}
\newacronym{UN}{UN}{United Nations}
\newacronym{SDG}{SDG}{Sustainable Development Goal}
```