LABEL SUPERVISED LLAMA FINETUNING*

A PREPRINT

Zongxi Li[1†], **Xianming Li**[2†], **Yuzhang Liu**[3], **Haoran Xie**[4], **Jing Li**[2],
Fu-lee Wang[1], **Qing Li**[2], **Xiaoqin Zhong**[3]

[1] School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR
[2] Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR
[3] Shanghai 100me Internet Technology Co., Ltd., Shanghai, China
[4] Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR
[†] Corresponding authors: zoli@hkmu.edu.hk, xianming.li@connect.polyu.hk

ABSTRACT

The recent success of Large Language Models (LLMs) has gained significant attention in both academia and industry. Substantial efforts have been made to enhance the zero- and few-shot generalization capabilities of open-source LLMs through finetuning. Currently, the prevailing approach is instruction-tuning, which trains LLMs to complete real-world tasks by generating responses guided by natural language instructions. It is worth noticing that such an approach may underperform in sequence and token classification tasks. Unlike text generation tasks, classification tasks have a limited label space, where precise label prediction is more appreciated than generating diverse and human-like responses. Prior research has unveiled that instruction-tuned LLMs cannot outperform BERT, prompting us to explore the potential of leveraging latent representations from LLMs for supervised label prediction. In this paper, we introduce a label-supervised adaptation for LLMs, which aims to finetuning the model with discriminant labels. We evaluate this approach with Label Supervised LLaMA (LS-LLaMA), based on LLaMA-2-7B, a relatively small-scale LLM, and can be finetuned on a single GeForce RTX4090 GPU. We extract latent representations from the final LLaMA layer and project them into the label space to compute the cross-entropy loss. The model is finetuned by Low-Rank Adaptation (LoRA) to minimize this loss. Remarkably, without intricate prompt engineering or external knowledge, LS-LLaMA substantially outperforms LLMs ten times its size in scale and demonstrates consistent improvements compared to robust baselines like BERT-Large and RoBERTa-Large in text classification. Moreover, by removing the causal mask from decoders, LS-unLLaMA achieves the state-of-the-art performance in named entity recognition (NER). Our work will shed light on a novel approach to adapting LLMs for various downstream tasks.

1 Introduction

Large Language Models (LLMs), such as GPT-3 [Brown et al., 2020] and GPT-4 from OpenAI, LLaMA [Touvron et al., 2023a,b] from Meta, and PaLM [Chowdhery et al., 2022, Anil et al., 2023] from Google, have demonstrated impressive language understanding and human-like response generation abilities. The large-scale pretraining and increased parameter size lead to phenomenal *emergent abilities* [Wei et al., 2022a], endowing LLMs with strong generalization capacity for unseen tasks, even in zero- and few-shot settings. Such capabilities of LLMs, unseen in smaller models, have significantly revolutionized methodologies across various natural language processing (NLP) tasks, ranging from label-supervised parameter optimization to prompt-based response generation. A key driving factor behind these transformations is the decoder-only architecture of these LLMs, which incorporates causal masks to prevent forward information exposure. Consequently, until now, the latent representations of LLMs have primarily been employed for predicting the next token in autoregressive text generation.

*Preprint. Work in progress.

LLMs have been extensively evaluated in text classification and information extraction tasks, both in zero- and few-shot settings [Zhao et al., 2021, Wei et al., 2023, Li et al., 2023a], as well as with instruction tuning Wang et al. [2023], Lei et al. [2023]. Recent studies have revealed that zero-shot LLMs struggle to achieve satisfactory performance in these domains. For instance, GPT-3 (175B) achieves a classification accuracy of only 76% on SST-2 and 43.9% on AGNews [Zhao et al., 2021], while GPT-3.5-Turbo (154B) attains an F1 score of 18.22% for named entity recognition (NER) on OntoNotes [Wang et al., 2023]. Even the popular ChatGPT can only achieve F1 scores of 67.2% and 51.1% for NER on CoNNL2003 and OntoNotes, respectively Li et al. [2023a]. The above results fall short of state-of-the-art benchmarks. Although metric performance can be enhanced through careful prompt engineering and instruction-tuning techniques, these refined LLMs can hardly outperform discriminative encoder models such as BERT [Devlin et al., 2019] and RoBERTa [Liu et al., 2019].

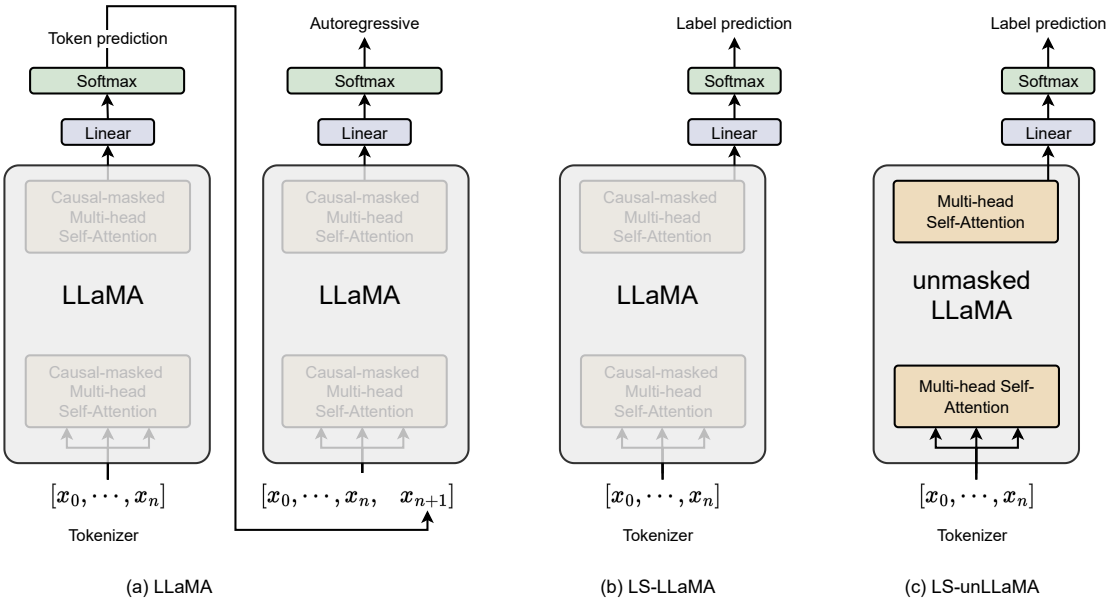
We consider that these observations align with expectations, as LLMs and BERT possess different architectures and strengths that lead to divergent performance on specific tasks like text classification and NER. In classification tasks, labels often consist of highly concentrated words or phrases, and NER tasks involve symbolic tags, causing difficulties for LLMs in understanding the semantic meanings of the labels effectively. Although conversational LLMs can be finetuned on labeled datasets for these tasks, their generation-focused architecture may not capture task-specific patterns as efficiently as label-supervised BERT models, which have consistently demonstrated superior performance on a wide range of label prediction tasks. Furthermore, label spaces are considerably more restricted compared to the entire vocabulary, making the autoregressive generation less effective and efficient in label prediction, particularly in NER, where outputs are structured sequences of NER tokens.

Motivated by the efficacy of finetuning BERT and RoBERTa for classification tasks, we are the first to explore the feasibility of finetuning LLMs with label supervision to achieve effective task-specific adaptation. In this study, we employ the open-sourced LLaMA-2-7B model. We directly extract latent vector representations from the final LLaMA decoder layer, which was originally designed for autoregressive next-token prediction. These representations are then mapped into the label space through feed-forward layers, yielding logits that are used for discriminant label classification. We calculate the cross-entropy loss and employ Low-Rank Adaptation (LoRA) [Hu et al., 2021] to fine-tune the LLaMA model. Our preliminary results on multiclass benchmarks have been remarkably promising. We have observed significant improvements over both zero-shot and instruction-tuned LLaMA-2-7B models, as well as consistent enhancements compared to finetuned BERT. Extensive experiments demonstrate the effectiveness and robustness of our proposed task-specific LLaMA adaptation. We name such a configuration **L**abel **S**upervised **LLaMA** (LS-LLaMA).

Before this work, open-sourced LLMs had yet been utilized for discriminant label classification as their latent representations were considered not suitable for language encoding, although Meta has provided the interface for sequence classification since LLaMA-1[2] [Touvron et al., 2023a]. Unlike the the encoder-based structure of Transformer [Vaswani et al., 2017] models, such as BERT and RoBERTa, conversational LLMs do not have an encoder structure Brown et al. [2020], Touvron et al. [2023a] and were trained primarily for language generation, focusing on predicting the next word in a sequence. The causal mask in the decoder blocks avoids forward information disclosure and also blocks bidirectional dependency extraction, leading to a myth that LLMs are not suitable for text encoding. Our research has has unveiled that the decoder output of LLMs contains a substantial amount of semantic meaning from the input text and can be effectively utilized as text representations for various classification tasks. However, the causal mask's presence causes fatal information loss at the token-level representation, leading to unsatisfactory results by LS-LLaMA in NER tasks. To overcome such a limitation, we propose an innovative solution by removing the causal mask from the decoders. After finetuning, the **L**abel **S**upervised **un**masked **LLaMA** (LS-unLLaMA) exhibits substantial improvements (up to 18%) over LS-LLaMA in NER benchmarks i.e., CoNNL2003 and OntoNotes, and even outperforms LS-LLaMA on multiple text classification tasks.

This research aims to introduce a label-supervised adaptation configuration for LLMs. We investigate the feasibility of employing latent representations from LLMs for discriminant label prediction in text classification tasks. Our experimental results demonstrate that LLMs' latent representations can effectively serve as a text encoding method. Furthermore, we identify a limitation when using LLM's representations for token-level applications, attributed to the causal masks in the decoder structure, which restrict bidirectional information flow. To overcome this limitation, we remove the causal masks from LLaMA, resulting in state-of-the-art performance in NER tasks.

[2] `https://huggingface.co/docs/transformers/v4.33.2/model_doc/llama#transformers.LlamaForSequenceClassification`



(a) LLaMA (b) LS-LLaMA (c) LS-unLLaMA

Figure 1: Comparison between conversational LLaMA and our proposed LS-LLaMA and LS-unLLaMA in sequence classification.

2 Methodology

This section presents the implementation of our proposed LS-LLaMA and LS-unLLaMA. Figure 1 depicts how they are different from the conventional autoregressive settings.

2.1 Label-supervised LLaMA

We tokenize the input sequence S with the default `AutoTokenizer` following the automatic operation of the transformers[3] library. The tokens T were fed into pretrained models through `LlamaForSequenceClassification` to extract the latent representation H from LLaMA for sequence classification,

$$\begin{aligned} T &= \text{Tokenizer}(S) \\ H_{seq} &= \text{LlamaForSeqClf}(T). \end{aligned} \tag{1}$$

The pooling operation is applied to the latent representation to obtain the vector representation h for sequence classification. `LlamaForSequenceClassification`'s default pooling operation takes out the last token vector from the final representation, as this is the only token that encodes all the historical information due to single-direction information flow caused by the causal masks. After passing through fully connected layers and a softmax layer, vector representation h is mapped to the label space. Cross-entropy loss is calculated based on the output logits and the ground-truth label.

LLaMA did not provide the interface for token classification. Therefore, we modify the LLaMA model to obtain all the token representations with `LlamaForTokenClassification`[4], $H_{tkn} = \text{LlamaForTokenClf}(T)$, for token classification tasks.

We apply LoRA [Hu et al., 2021] to finetune the LLaMA model to maximize the probability of the correct label.

[3] `https://github.com/huggingface/transformers`

[4] Code for LS-LLaMA is available on GitHub: `https://github.com/4AI/LS-LLaMA`

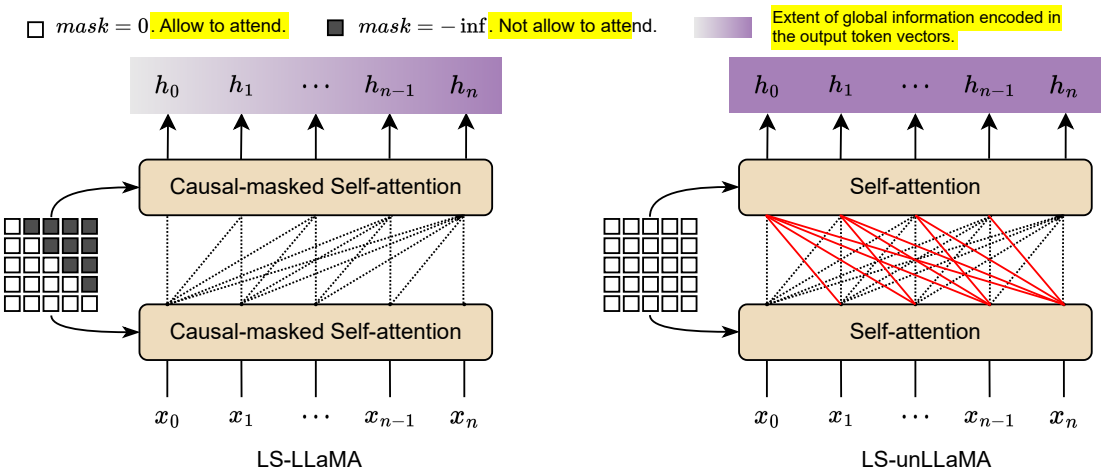


Figure 2: An illustration of the effects caused by different masking methods in LS-LLaMA and LS-unLLaMA.

2.2 Label-supervised unmasked LLaMA

The causal masks CM, as shown in Equation 2,

$$CM = \begin{bmatrix} 0 & -\inf & -\inf & \cdots & -\inf & -\inf & -\inf \\ 0 & 0 & -\inf & \cdots & -\inf & -\inf & -\inf \\ 0 & 0 & 0 & \cdots & -\inf & -\inf & -\inf \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & -\inf & -\inf \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\inf \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

in decoder blocks prevent information leaking, as the decoder is only allowed to attend to earlier positions in text generation. Bidirectional dependency extraction of the self-attention layer is reduced to single-direction, leading to critical information loss at the token level. Our empirical studies show that using token representations learned with causal masks significantly underperforms in token classification tasks. To address such an issue, we remove the causal masks from `LlamaForTokenClassification`[5] and extract the latent representations H' for token classification, as illustrated in Figure 2,

$$H'_{tkn} = \text{unmasked\_LlmaForTokenClf}(T). \tag{3}$$

The essential bidirectional information is expected to be replenished in token representations during finetuning as all the tokens can attend to each other.

We believe the essential global dependency is also helpful in the sequence classification, and hence, remove the causal masks in `LlamaForSequenceClassification`. With bidirectional self-attention resumed, we have more choices in pooling. We have tested three pooling methods, i.e., `max`, `average`, and `last`, and the experiments show that max-over-time pooling yields better performance than average pooling and last-token pooling in classification tasks without causal masks.

3 Experiment

3.1 Tasks and evaluation metrics

We have conducted extensive experiments on text classification and NER tasks against zero- and few-shot LLMs, instruction-following LLMs, and discriminant baselines to validate the effectiveness of the proposed label-supervised adaptation method based on LLaMA-2-7B.

[5]Note that this operation is different from the padding mask and requires modification on LLaMA's source code.

3.1.1 Multiclass text classification

Experiments were conducted on four English datasets from multiple domains, i.e., SST2 and SST5[6] (general sentiment analysis), AGNews (news categorization), and Twitter Financial News[7] (short as "Twitter-Fin", financial sentiment analysis). Table 1 contains the statistics and label classes of the datasets.

Moreover, we experimented on the German, English, Spanish, and Chinese subsets of Multilingual Amazon Reviews Corpus[8] [Keung et al., 2020] to test LS-LLaMA and LS-unLLaMA under the multilingual setting. The datasets contain Amazon product reviews for 31 product categories. The task is to predict the product category given the review content. For each language, the subset contains 200,000, 5,000, and 5,000 data samples in the training, development, and test sets respectively. We report the accuracy of text classification tasks.

Table 1: Dataset descriptions for multiclass text classification.

Dataset	Train	Test	Classes
SST2	67,300	872*	"Positive", "Negative"
SST5	8,540	2,210	"Very negative", "Negative", "Neutral", "Positive", "Very positive"
AGNews	120,000	7,600	"World", "Sports", "Business', "Sci/Tech"
Twitter Fin	9,938	2,486	"Bearish", "Bullish", "Neutral"

* Testing on the validation set for all baselines

3.1.2 Named entity recognition

For the NER task, we experiment on OntoNotes V5.0[9] and CoNLL2003 [Tjong Kim Sang and De Meulder, 2003], which are widely-adopted datasets for information retrieval. OntoNotes V5.0 has 59,924 training samples and 8,262 testing samples, and CoNLL2003 contains 14,987 training samples and 3,684 testing samples. We report the F1 scores on the NER tasks.

3.2 Implementation details

LLaMA-2-chat under the zero-shot setting was tested on the classification tasks. LLaMA-2-chat-7B and LLaMA-2-chat-13B were deployed on Nvidia GeForce RTX4090 and A100 GPUs, respectively.

The main experiments of label-supervised adaptation are based on LLaMA-2-7B. These experiments were run on a single Nvidia GeForce RTX4090 GPU. We adopt the standard parameter-efficient fine-tuning method LoRA to finetune the model. We configured the LoRA settings *lora_rank*, *lora_alpha*, and *lora_dropout* to 12, 32, and 0.1, respectively. We set the batch size to 8 and initial learning rate to $8e-5$ using grid search. For the sequence classification, we use the *last token* pooling on LS-LLaMA and the *max* pooling on LS-unLLaMA according to the ablation study on the pooling method.

We apply instruction-tuning based on LLaMA-2-7B on text classification tasks. The prompt is set as follows:

> `prompt_input` = "You are a classification model. Based on the given article, you need to predict the most relevant category label from `{all_labels}`. One article has only one label. \n ### Input article: `{input_text}` \n ### Output: "

The instruction-following model is finetuned for 50k steps on large datasets (training samples exceed 50,000) and 10k steps on small datasets (training samples less than 50,000) with a batch size of 16 and an initial learning rate of $2e-4$.

3.3 Experimental results

3.3.1 Multiclass text classification

We present the results of multiclass text classification experiments in Table 2, offering a comprehensive analysis of LLMs under different adaptation settings and discriminant models. Notably, LLMs such as LLaMA-2-7B/13B and GPT-3-175B exhibit suboptimal performance when subjected to zero- and few-shot settings. While instruction-tuning

[6] `https://huggingface.co/datasets/SetFit/sst5`
[7] `https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment`
[8] `https://huggingface.co/datasets/amazon_reviews_multi`
[9] `https://catalog.ldc.upenn.edu/LDC2013T19`

Table 2: Accuracy (%) in multiclass text classification.

Models	SST2	AGNews	Twitter-Fin	SST5
Zero- and few-shot				
LLaMA-2-7B (zero-shot)	76.26	37.39	23.40	39.05
LLaMA-2-13B (zero-shot)	69.90	59.40	38.74	37.01
GPT-3 175B (zero-shot)	54.3†	43.9†	–	–
GPT-3 175B (few-shot)	93.4†	84.3†	–	–
Instruction-tuning				
LLaMA-2-7B (instruction-tuning)	91.97	52.40	68.72	43.35
Discriminant baselines				
BERT-Base	92.78	94.51	88.19	55.07
BERT-Large	92.86	94.45	88.74	55.79
RoBERTa-Base	94.61	94.70	90.32	58.46
RoBERTa-Large	96.10	94.78	90.95	59.64
AGN [Li et al., 2021]	93.27	93.82	–	55.72
Ours				
LS-LLaMA-2-7B	96.67	95.38	**91.87**	**62.31**
LS-LLaMA-2-13B	96.90	95.66	91.20	62.17
LS-unLLaMA-2-7B	**97.36**	**95.68**	91.54	60.50
LS-unLLaMA-2-13B	92.77	95.44	87.94	52.99

† results reported in Zhao et al. [2021]

brings about notable enhancements on LLaMA-2-7B, it is apparent that the achieved results are not comparable to the discriminative capabilities demonstrated by BERT and RoBERTa. This observation highlights the challenges faced by conversational LLMs when predicting from a predefined label set.

Our proposed label-supervised adaptation method showcases substantial improvements over instruction fine-tuning. Notably, the magnitude of these enhancements is more outstanding in datasets where labels are more domain-specific and demand a deeper understanding of commonsense knowledge. For instance, in SST2, which features commonly-used binary sentiment labels ("Positive" and "Negative"), we observe an absolute improvement of 5.39% (equivalent to a relative improvement of 5.86%). In contrast, in SST5 with fine-grained sentiment labels like "Neutral", "Very positive", and "Very negative", our approach demonstrates an 18.96% absolute improvement (equivalent to a relative improvement of 43.74%). When applied to the Twitter-Fin dataset, which comprises domain-specific financial labels ("Bullish" and "Bearish"), our method produces a 23.15% absolute improvement (equivalent to a relative improvement of 33.69%). In the AGNews dataset, labeled with domain-specific news categories, we achieve a 43.28% absolute improvement (equivalent to an astonishing relative improvement of 82.60%). Remarkably, such improvements do not require sophisticated prompt engineering and external knowledge.

Furthermore, it is noteworthy that both LS-LLaMA and LS-unLLaMA consistently demonstrate significant improvements when compared to robust discriminative baselines such as BERT-Large and RoBERTa-Large. Our approach achieves improvements of 1.26%, 0.90%, 0.92%, and 2.67% when compared against RoBERTa-Large on SST2, AGNews, Twitter-Fin, and SST5, respectively. Interestingly, we also observe variations in performance between LS-LLaMA and LS-unLLaMA. Specifically, LS-unLLaMA outperforms LS-LLaMA on SST2 and AGNews, but exhibits inferior performance on Twitter-Fin and SST5. A potential explanation for these differences can be drawn from the dataset sizes, as indicated in Table 1. SST2 and AGNews are notably larger datasets compared to Twitter-Fin and SST5, a distinction that may provide insights. Since we removed causal masks from LS-unLLaMA, the model needs more training samples to reconstruct the parameters that were previously concealed during pretraining. This process is not strictly needed in LS-LLaMA as it can directly employ the last token for classification task. Therefore, we conclude that LS-LLaMA can quickly adapt on small-scale datasets, and LS-unLLaMA can achieve even better results with ample training samples.

3.3.2 Multilingual multiclass text classification

According to Touvron et al. [2023b], LLaMA-2's pretraining data primarily consists of English text, accounting for a substantial 89.70%, with non-English languages representing only a minority within the pretraining dataset—specifically, 0.17% German, 0.13% Chinese, and 0.13% Spanish. However, "[a] training corpus with a majority in English means

Table 3: Accuracy (%) in multiclass text classification on Multilingual Amazon Reviews Corpus.

Models	**DE**	**EN**	**ES**	**ZH**	**Avg.**
Zero-shot					
LLaMA-2-7B (zero-shot)	9.23	12.13	7.43	19.65	12.11
Discriminant baselines					
BERT-Base-Multilingual	53.54	53.42	46.22	67.35	55.13
RoBERTa-Base-Multilingual	52.61	52.56	44.32	66.98	54.12
RoBERTa-Large-Multilingual	52.76	53.64	44.76	66.90	54.52
Ours					
LS-LLaMA	56.80	58.82	49.28	68.72	58.41
LS-unLLaMA	**56.90**	**60.20**	**49.68**	**69.70**	**59.21**

that the model may not be suitable for use in other languages." Consequently, we conducted a series of tests to evaluate LLaMA's text classification capabilities under the zero-shot setting and with our proposed label-supervised adaptation method within multilingual environments. In particular, the adopted Amazon Reviews Corpus encompasses 31 product category labels, making the classification more difficult for both LLMs and discriminant baselines. Zero-shot LLaMA-2-7B presents low accuracy when asked to select one label from 31 candidates. With label-supervised finetuning, the model's performance gains improvements ranging from four to sixfold. In particular, compared with the multilingual version of BERT and RoBERTa, our approaches can still gain remarkable improvements of at least 3.36%, 6.56%, 3.46%, and 2.35% on the German, English, Spanish, and Chinese subsets, respectively, underlining the exceptional multilingual learning capacity of LLaMA. These results underscore the superior efficacy of the proposed label-supervised adaptation method when applied to a range of text classification tasks. Furthermore, LS-unLLaMA outperforms LS-LLaMA, suggesting the benefit brought by the removal of the causal mask.

Table 4: F1 score (%) in NER.

Models	**CoNNL2003**	**OntoNotes V5**
Zero- and few-shot		
LLaMA-2-7B (zero-shot)	1.35	1.20
ChatGPT	67.20[†]	51.10[†]
GPT-3.5-Turbo	–	18.22[‡]
Instruction-tuning		
InstructUIE [Wang et al., 2023]	92.94	90.19
Discriminant baselines		
BERT-Base	92.40[§]	88.88[⋆]
BERT-Large	92.80[§]	89.27
RoBERTa-Base	92.13	91.55
RoBERTa-Large	92.59	91.72
RAN [Li et al., 2023b]	92.68	89.38
Ours		
LS-LLaMA-2-7B	74.76	77.41
LS-LLaMA-2-13B	74.12	77.73
LS-unLLaMA-2-7B	**93.19**	**92.10**
LS-unLLaMA-2-13B	91.46	91.07

[†] results reported in Li et al. [2023a]
[‡] results reported in Wang et al. [2023]
[§] results reported in Devlin et al. [2019]
[⋆] results reported in Li et al. [2023b]

3.3.3 Named entity recognition

Results on NER tasks were presented in Table 4. The highlight of the results lies in the performance margin between LS-LLaMA and LS-unLLaMA. In text classification tasks, these two variants can produce comparable results. However,

Table 5: Comparison between different pooling strategies

	SST2		SST5	
	LS-LLaMA w/ CM	LS-unLLaMA w/o CM	LS-LLaMA w/ CM	LS-unLLaMA w/o CM
`max`	95.53	**97.36**	58.79	**60.50**
`average`	55.92	95.96	53.80	59.41
`last`	**96.67**	95.76	**62.31**	46.74

CM stands for causal mask.

LS-unLLaMA achieves 18.43% and 14.69% higher F1 scores than LS-LLaMA on CoNNL2003 and OntoNotes, respectively.

The results pertaining to NER tasks are outlined in Table 4. LLaMA-2-7B under zero-shot setting can only produce single-digit result of F1 score, showing that LLaMA does not have suitable knowledge for addressing NER tasks. Our approach presents higher performance than BERT-Large and RoBERTa-Large. What deserves special attention is the distinctive performance margin observed between LS-LLaMA and LS-unLLaMA. In the realm of text classification tasks, these two variants tend to yield comparable results. However, in NER tasks, LS-unLLaMA exhibits much more reliable performance. Specifically, LS-unLLaMA achieves F1 scores that are 18.43% and 14.69% higher than those achieved by LS-LLaMA on the CoNLL2003 and OntoNotes datasets, respectively. This performance gap indicates the pivotal of global dependencies, encoded within token representations and facilitated by bidirectional self-attention mechanisms, in effectively addressing NER challenges. Such features were absent in LS-LLaMA and other LLMs.

3.4 Ablation study

3.4.1 Pooling method

We conducted a comprehensive study of various pooling methods, including `max`, `average`, and `last`, on the SST2 and SST5 datasets to determine their impact on model performance. The findings, as shown in Table 5, highlight distinct trends. Notably, when employing causal masks, LS-LLaMA achieves the best performance when adopting the `last` pooling method. Conversely, LS-unLLaMA without causal masks demonstrates optimal capabilities when utilizing the `max` pooling strategy.

The reason of such a difference attributes to the presence or absence of causal masks. In LS-LLaMA, causal masks enforces a single-direction information flow, allowing only the last token to attend to all the previos tokens. Consequently, the semantic meaning of the entire sentence aggregates at the last token. In contrast, the tokens within LS-unLLaMA, free from the constraints of causal masks, possess the capacity to attend to every other token in a bidirectional manner. The expanded connectivity facilitates the utilization of max-over-time pooling, which facilitates extracting global dependencies within the sentence.

3.4.2 Training process

We notice that the proposed approaches can outperform BERT on a range of benchmark datasets, hence it is interesting to observe how the LLaMA is finetuned in the same way as BERT. Therefore, we study the training process in 10 epochs of LS-LLaMA, LS-unLLaMA, and BERT on SST5 and ConLL2003 datasets. We recorded the training loss, evaluation loss, and evaluation accuracy / F1 score every 100 training steps and depict the data in Figures 3, 4, and 5. On SST5, in which all these three models performs well, LS-LLaMA and BERT converge faster than LS-unLLaMA, and LS-unLLaMA's testing loss and accuracy tend to be unstable after 5,000 training steps. As discussed in Section 3.3.1, LS-unLLaMA underperforms on small-scale datasets like SST5, as it struggles to reconstruct the self-attention weights that were not well trained during pretraining because of the causal masks. Small datasets cannot provide sufficient label supervision and leads to severer overfitting problem. In contrast, the training process of LS-LLaMA is more stable, and it suffers less overfitting than BERT.

In NER task, the training process of LS-LLaMA is more unstable than those of LS-unLLaMA and BERT, as the token representations of LS-LLaMA intrinsically lacks global dependency due to the causal masks. We can also observe that the testing loss of LS-unLLaMA bounces back later than BERT, suggesting that LS-unLLaMA is more resistant to overfitting issue than BERT with a large training set.

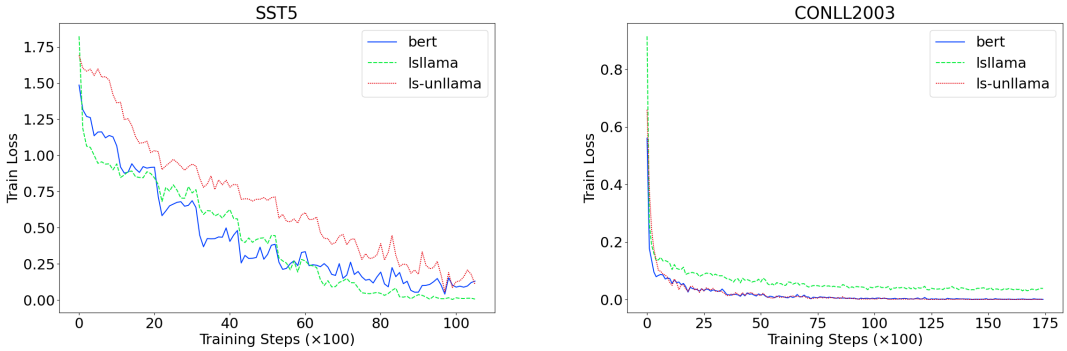


Figure 3: Training loss of BERT, LS-LLaMA, and LS-unLLaMA in 10 training epochs.

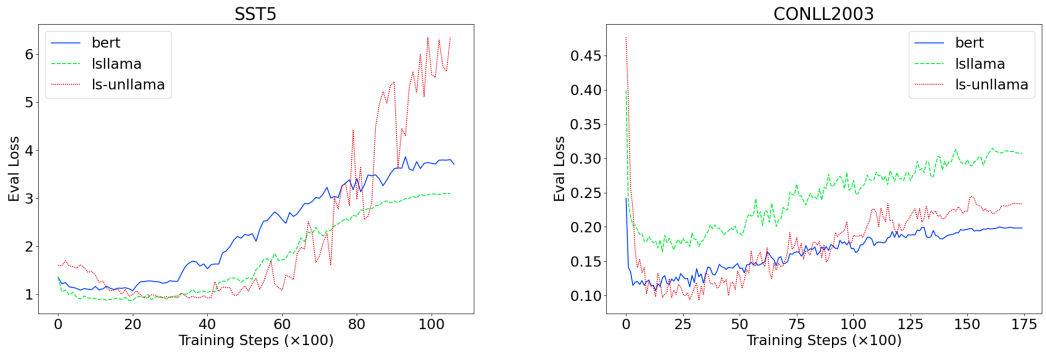


Figure 4: Testing loss of BERT, LS-LLaMA, and LS-unLLaMA in 10 training epochs.

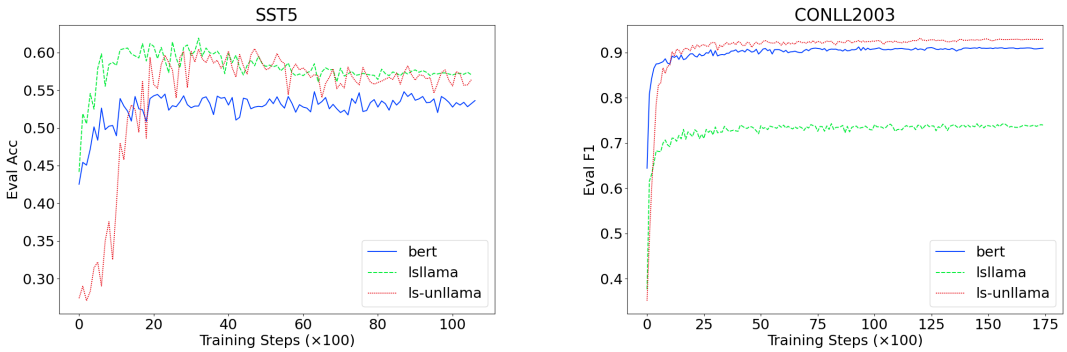


Figure 5: Testing accuracy / F1 score of BERT, LS-LLaMA, and LS-unLLaMA in 10 training epochs.

3.4.3 Model size

In light of the outstanding performance achieved by the smaller-scale 7B LLaMA-2, one may be curious about whether a larger LLM like 13B LLaMA-2 would yield even more impressive results. Therefore, we experimented the proposed label-supervised adaptation approach based on LLaMA-2-13B on sequence and token classification tasks. The results were listed in Tables 2 and 4.

In summary, the application of our label-supervised adaptation approach to 13B LLaMA-2, both with and without the causal mask, did not yield substantial improvements on either of the tasks. In the realm of sequence classification tasks, LS-LLaMA-2-13B does exhibit better performance compared to BERT and RoBERTa, as well as its 7B counterpart, on larger datasets like SST2 and AGNews. However, it falls short of the results achieved by LS-unLLaMA-2-7B. On smaller datasets such as Twitter-Fin and SST5, both 13B LS-LLaMA-2 and LS-unLLaMA fail to surpass their 7B counterparts. Moreover, LS-unLLaMA-2-13B also suffers from significant performance degradation, possibly caused by the insufficient volume of training set.

The challenge is also evident in token classification tasks. Finetuning the 13B LLaMA model, employing cross-entropy loss and LoRA, leads to deteriorated performance than finetuning the 7B LLaMA and BERT in the same manner. These observations point to the challenge of fine-tuning the 13B LLaMA-2, particularly when the training samples are limited. The scarcity of training samples may cause severe overfitting problems on the limited training set.

In essence, our experiments show that the performance of LLaMA-2 under label-supervised adaptation does not exhibit the expected linear scalability with an increase in model size. The availability of sufficient training samples becomes a critical determinant in applying label-supervised adaption for larger LLMs.

4 Related Work

Continuous efforts have been devoted to improving the problem-solving abilities of LLMs with their superior text generation capacity. One of the main research directions focuses on prompt engineering, which aims to generate higher-quality responses by harnessing LLMs' existing knowledge. Without modifying the parameters, one can give instructions with a few input-output exemplars or carefully crafted in-context prompts to help the model better understand the task and elicit a profound inferencing and reasoning process. Various prompting techniques have been proposed, such as contextual calibration [Zhao et al., 2021], prompt programming [Reynolds and McDonell, 2021], chat-based prompt [Wei et al., 2023], chain-of-thought [Wei et al., 2022b], and tree-of-thought [Yao et al., 2023]. These methods are especially preferred for extra-large LLMs like ChatGPT and PaLM-540B, on which tuning the parameters may be less feasible, they have been effective in a range of tasks including information extraction [Li et al., 2023a], semantic textual similarity Li and Li [2023], and reasoning [Kojima et al., 2022].

Despite the success of zero- and few-shot settings, LLMs frequently struggle in domains that require specific knowledge or precise response generation. To generalize LLMs on more downstream tasks, researchers have also investigated various instruction-tuning methods [Brown et al., 2020, Wei et al., 2021, Wang et al., 2022a, Peng et al., 2023, Phang et al., 2023, Zadouri et al., 2023] to replenish domain knowledge and enhance LLMs performance. Instruction-tuning approaches tune the pretrained parameters with instructional data, which contain instructional commands and human-annotated expected outcomes [Sanh et al., 2021, Wang et al., 2022b]. Peng et al. [2023] finetuned LLaMA using 52K English and Chinese instruction-following instances generated using GPT-4. Phang et al. [2023] proposed HyperTuning that uses a hypermodel to generate task-specific parameters for a fixed downstream model for model adaptation. Instruction-following LLMs present substantial improvements in zero-shot performance on unseen tasks.

So far, no attempts have been made to finetune an LLM with discriminant labels. Our study verifies the feasibility of the proposed label-supervised adaptation approach on sequence and token classification tasks.

5 Conclusion

In conclusion, this paper embarks on a comprehensive exploration of label-supervised adaptation for enhancing LLMs' performance in both sequence and token classification, surpassing existing approaches like prompt engineering and instruction-tuning. Our two proposed variants, LS-LLaMA and LS-unLLaMA, have demonstrated remarkable and consistent improvements over robust benchmarks such as BERT and RoBERTa, across a range of text classification and under multilingual settings. With causal masks removed in the decoders, LS-unLLaMA yields state-of-the-art performance on token classification tasks like NER. This study depicts the potential of LLMs as robust text encoders, with latent representations can be applied in a broader spectrum of applications when explicit label supervision is provided.

The implications of our findings may extend beyond the specific tasks explored in this study. The proposed label-supervised adaptation offers an accessible and highly effective configuration that can serve as a novel interface for various downstream tasks, such as domain-specific text classification and token classification, by finetuning a small-scale LLMs such as LLaMA-2-7B. This method could potentially reshape the landscape of LLM applications.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*, 2023a.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

Xianming Li, Zongxi Li, Haoran Xie, and Qing Li. Merging statistical feature via adaptive gate for improved text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13288–13296, 2021.

Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang, and Qing Li. Recurrent attention networks for long-text modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3006–3019, Toronto, Canada, July 2023b. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neweural Information Processing Systems*, 35:24824–24837, 2022b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022a.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. Hypertuning: Toward adapting large language models without back-propagation. In *International Conference on Machine Learning*, pages 27854–27875. PMLR, 2023.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2021.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022b.