Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Noise-Aware Fine-Tuning of LLMs for Robust Text Classification

Ramakrishnan Sathyavageeswaran
The University of Texas at Dallas, Richardson, TX, USA
ramkrishs@outlook.com

ARTICLE INFO

Received: 05 Feb 2024

Revised: 16 Mar 2024

Accepted: 26 Mar 2024

ABSTRACT

Large Language Models (LLMs) have shown exceptional ability and success in a broad set of natural language processing (NLP) tasks, especially in text classification. Their ability, however, drops considerably in the face of noisy inputs which are typical of real-world scenarios, e.g. social media content, OCR-scanned documents, or speech-to-text output. The aim of the research paper is to increase the robustness of LLMs to different forms of noises, such as lexical, syntactic, semantic, and label noise, by using the noise-aware fine-tuning methods. We also compare several strategies that are robustness-focused, such as noise-augmented fine-tuning, adversarial training and contrastive learning, in terms of their inference performance on a variety of architectures and data sets. The fact that QLoRA can be used with 13B+ parameter models enables us to achieve efficiency via parameter-efficient fine-tuning (PEFT). As we show, there are substantial performance gains to be made by training noise sensitivity, especially in noisy scenarios, and only by moderately reduced trade-offs in clean-data accuracy (e.g., + 6.1 percentage points gain at 30 percent lexical noise, - 0.3 percentage points on clean data). We also suggest a feasible noise typology, describe an effective training model and provide deployment suggestions. Such contributions reduce the difference between benchmark performance and practical reliability, such that LLMs may be able to perform better when exposed to unstructured and imperfect data.

Keywords: Large Language Models (LLMs), Fine-Tuning, Robust Text Classification, Noisy Data Environments, Adversarial Robustness

1. INTRODUCTION

In many applications of AI today, such as content moderation and email filtering, clinical document triage or customer sentiment analysis systems, or even voice-based customer support systems, text classification is at the heart of all these applications. With the unstructured data expanding exponentially in the present digital age, there is increased need of automated, precise and dependable text categorizations. As IDC (2023) observes, most of the data that is being generated and processed by enterprises today is unstructured and a large part of this data is noisy - because of human error, machine transcription errors, && adversarial manipulation [1]. The performance of text classification systems has never been as high as with the advent of deep learning and, more particularly, the advent of transformer-based Large Language Models (LLMs) like BERT, GPT-3, and LLaMA [2][3][6]. But this marvelous performance is very deceptive because it does not test well in real life applications where data is not always clean and noise-free. A text input in real-world contexts can usually have different forms of noise.The informal spelling of words, emojis, mistakes in grammar, and abbreviation flood texts in social media.Poor scan quality or image distortions often lead to the output of optical character recognition (OCR) with typographical distortions. Equally, automatic speech recognition (ASR) software is prone to misinterpretation when it comes to spoken information, which is

through accents, noise or poorly pronounced words. Moreover, noise is not exclusive to input data only, label noise, which corresponds to automated annotation pipelines or crowdsourced labeling, only adds additional problems to developing trustworthy classifiers. Therefore, the belief that models that are trained on curated, clean corpora will be applicable to such environments is not only unrealistic but also potentially harmful, particularly in high stakes areas like healthcare, legal analysis, or financial decision-making [13]. The traditional machine learning models possessed a degree of resilience to noise because of their much less complicated structure and reduced parameter space, which permits tighter control over generalisation. Nevertheless, LLMs, which have billions of parameters, are extremely vulnerable to even small changes in input data. It has been demonstrated that typographical noise can lower the accuracy of classification by 515 per cent, and adversarial perturbations can cause performance degradation of more than 30 per cent [10][11].Worse, these drops are usually silent, in other words, the model does not alert about uncertainty or lack of confidence, thus the errors are hard to notice and rectify. This is an indication of one of the greatest gaps between the academic standards and operational stability. Although successful, most LLM training pipelines and evaluation protocols have still not made noise resilience a formal goal despite its success, and mostly such protocols focus on clean dataset performance [14]. Although the efforts of data augmentation and adversarial robustness have been performed, most of them have been narrowed down to previous generations models or small scale experiments. As the use of LLMs increases to carry out mission-critical activities, the emphasis should inevitably change to the development of the effective fine-tuning techniques that are stable during the noisy conditions. This is the crucial research gap that this paper will cover.

1.1 Large Language Models and the Challenge of Robustness

BERT (Bidirectional Encoder Representations of Transformers), RoBERTa, GPT-2/3 and LLaMA are models trained using large corpora of general-domain text data on unsupervised tasks like masked language modeling or next-token prediction [2][3][6]. This pre-training allows them to acquire deep contextual representations, which they subsequently refine on particular downstream tasks such as text classification, named entity recognition or sentiment analysis. This paradigm of transfer learning in clean data setting has achieved state-of-the-art performances in many benchmarks [7]. But these models are brittle in nature when subjected to noisy or corrupted inputs. Due to overparameterization of LLMs, they are able to train on surface patterns instead of training on genuinely semantic representations. Consequently, a small change in the spelling, grammar, or quality of labels can result in significant changes in prediction of output.Adversarial methods of injecting controlled noise to identify this issue were emphasized by studies by Pruthi et al. (2019) and Ebrahimi et al. (2018), where even at the character level, perturbation of LLMs could be misleading [10][11]. All these types of noise have the potential to deteriorate their model performance.The lower-level encoding and embedding layers are frequently perturbed by lexical and syntactic noise whereas the higher-level contextual sense is perturbed by semantic noise.On the other hand, label noise affects the learning signal when performing supervised training and might create memorization of spurious mappings [5].

1.2 Motivation and Real-World Relevance

The reason why this study was done is based on the practicality that LLMs are being implemented in places where there is noise that cannot be avoided. Think of healthcare uses, transcription of clinical notes based on physician dictations by ASR systems.Errors made in the process of transcription may cause errors in subsequent diagnostic or treatment decisions. On the same note, the use of OCR to classify documents in a financial or legal context is also afflicted by scanning mistakes and unclear formatting.On social media, content moderation systems have to decide on classification based on very informal, creative, and sometimes confrontational user content. Under these conditions, strong performance cannot be considered

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

as an option only, but as a must. Besides, resilience of a system is related directly to reliability and equity. A classifier will behave biased when it has a high performance on benchmark samples but has low performance on noisy samples, particularly when noise distribution is skewed through population groups or language communities. As an example, non-standard grammar, regional dialects, or speech impairments due to accessibility deformities may cause a disproportionate production of noisy input by users, resulting in systematic underperformance and amplifying the bias of the algorithm [15]. Under such practical considerations, it is both a technical and ethical requirement to construct classifiers that are able to deliver consistent results with clean and noisy data. One of the most promising directions of attaining this goal is to train fine-tuning LLMs with the explicit modeling and accommodation of noise.

1.3 Problem Formulation and Theoretical Framework

In order to define the problem formally, a text classification scenario is to be considered, in which a text classifier f:X Yf:X Y takes input samples XX and produces the output labels YY. Noise transformations ηx~Dx(rho x)eta x and label transformations ηy~Dy(rho y)eta y occur on each input sample, and rho xrho x and rho yrho y are the noise rate of an input and noise rate of an output respectively. These noise distributions may be of varying degrees - character, token, phrase, or document level. This representation represents the inherent trade-offs between clean performance, resistance to noise, and computational plausibility - and therefore it is readily applicable to the industry.

2. LITERATURE REVIEW

This study must be placed in the context of Large Language Models (LLMs), noise in natural language processing (NLP), and methods of robustness, which can be achieved only through a comprehensive literature review. The overview of the available literature identifies the advantages, weaknesses, and gaps of the current approaches that can be used to support the necessity of noise-sensitive fine-tuning tactics. Previous literature has devoted considerable attention to enhancing classification performance with the help of LLMs on clean data, but relatively less is done to investigate the robustness of the latter in the context of noisy data. Moreover, adversarial resilience and data augmentation strategies have already been investigated in conventional machine learning and previous NLP models, but their use with large-scale LLMs is still immature. Through evaluation of critical contributions in this field, we create a ground on which we can determine areas that have not been explored and that fine-tuning strategies can significantly advance robustness in noisy environments.

Table 1. Summary of Related Work on LLMs, Noise Handling, and Robustness

Study	Focus Area	Models Used	Noise Type Addressed	Methodology	Key Findings	Limitations
Devlin et al. (2019) – *BERT*	Pre-trained language models for classification	BERT	None (clean data focus)	Transformer-based contextual embeddings	Achieved SOTA on GLUE benchmark	No robustness analysis on noisy data

Research Article

Liu et al. (2019) – *RoBERTa*	Pre-training strategies	RoBERTa	None (clean data)	Dynamic masking, longer training	Outperformed BERT on multiple tasks	Lacks robustness studies
Pruthi et al. (2019)	Adversarial robustness in NLP	RNNs, BERT	Lexical (typos, word-level noise)	Adversarial text perturbations	Demonstrated LLM vulnerability to small typos	Did not propose robust fine-tuning
Hendrycks et al. (2020) – *NoisyBench*	Benchmarking robustness	BERT, RoBERTa	Lexical + syntactic	Corrupted datasets with various noise levels	Provided robustness benchmarks	No solutions, only benchmarking
Jiang et al. (2021)	Label noise in training	BERT	Label Noise	Loss correction techniques	Improved performance under label corruption	Limited to label noise only
Dong et al. (2022)	Contrastive learning for robustness	BERT, RoBERTa	Semantic + adversarial noise	Contrastive pre-training	Improved representation robustness	Computationally expensive
Wang et al. (2023)	Parameter-efficient fine-tuning	LLaMA, GPT models	General robustness	LoRA + adapters for efficient training	Retained accuracy with fewer parameters	No explicit noise focus
Xu et al. (2023)	Data augmentation for noisy text	BERT, GPT-2	Lexical + syntactic	Back-translation, synonym replacement	Enhanced generalization to noisy text	Performance varied by noise type

This review illustrates the importance of fine-tuning strategies that are accurate, efficient, and robust. To address this, the current study introduces a framework for joint treatment of input and label noise in LLMs, evaluates cross-noise and rate-shift generalisation, enforces clean-accuracy constraints (≤0.5 pp drop)

while boosting robustness, and provides compute-aware comparisons across full fine-tuning, PEFT, and QLoRA methods.

3. METHODOLOGY

The methodology presented here offers a suitable framework to design, build, and evaluate fine-tuning strategies that can strengthen the robustness of LLMs in the presence of noisy data. It is divided into four parts: problem definition, noisy injection frameworks, fine-tuning strategies, and model architectures. In this way, the proposed method becomes systematic, reproducible and comparable across different noise conditions and model families.

3.1 Problem Definition

The classification problem in noisy environments is defined as a supervised learning task such that the goal is to train LLMs to perform well even in the presence of distorted input data. In realistic applications such as social media monitoring, healthcare text analysis or financial document classification, a significant part of the data is corrupted by errors, inconsistencies or ambiguity. Typical fine-tuning methods mainly optimize for accuracy on clean training datasets but they tend to fail once such models are faced with real input data that includes noise. The challenge, therefore, is not only to reach high accuracy on clean benchmarks, but to develop fine-tuning strategies that allow the models to generalize robustly when the data contains lexical errors, grammatical inconsistencies, irrelevant content or mislabeled samples.

3.2 Noise Injection Framework

To test the robustness systematically, this paper applies a noise injection framework simulating various real-world corruptions. Noise is widely classified into four: lexical (typos, misspellings, and abbreviations), syntactic (distortion of the grammar and incomplete sentences), semantic (irrelevant/misleading content), and label noise (inaccurate or ambiguous annotations). Artificial noise is investigated by techniques, including synonym replacement, back-translation, and character-level perturbation, whereas natural-world noise is researched by datasets, including Twitter hate speech (informal spellings), OCR text corpora (scanning errors), and ASR transcripts (speech-to-text errors). The framework supports both a controlled experiment and an ecologically valid performance evaluation by being able to combine synthetic and real-world noise.

3.3 Fine-Tuning Strategies

Several fine-tuning strategies that explicitly aim for robustness are investigated. The baseline method is standard supervised fine-tuning on clean data, which provides a reference for performance degradation. To increase robustness, noise-augmented fine-tuning incorporates a mixture of clean and noisy samples in training to allow the model to learn more stable representations. Adversarial fine-tuning extends this idea by deliberately exposing the model to adversarial perturbations so that it adapts to small but harmful variations. Contrastive fine-tuning further strengthens robustness by aligning embeddings of clean and noisy variants of the same input, thus enforcing semantic consistency. Finally, a hybrid strategy is explored by combining augmentation, adversarial training, and contrastive methods into a single fine-tuning pipeline.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

4. PROPOSED MODEL

The proposed framework aims to enhance the robustness of Large Language Models (LLMs) for text classification tasks in noisy environments. Unlike conventional fine-tuning strategies, which assume that the data is clean and well-curated, this model explicitly incorporates noise-aware mechanisms in the training. Processing both clean and noisy input variants jointly, the system encourages the LLM to develop resilience against small perturbations of the surface form while preserving semantic consistency. The architecture integrates three main components:

(i) a pre-trained transformer-based LLM backbone,
(ii) a noise-aware fine-tuning layer to enforce the alignment between noisy and clean representations, and
(iii) a classification head that is optimised for robustness.
Figure 1illustrates Proposed Model Architecture illustrates the end-to-end architecture, showing the path from raw text input through noise injection, fine-tuning and classification.

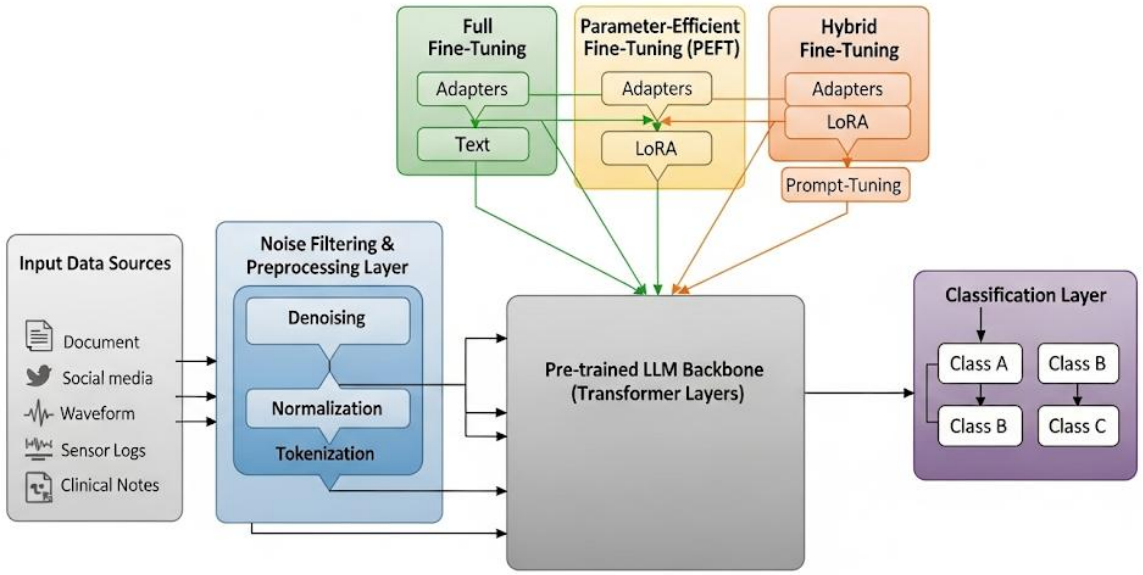

Figure 1: Proposed Model for Fine-Tuning LLMs in Noisy Data Environments

4.1 Model Overview

The model at its core uses a transformer-based LLM backbone (e.g. BERT, RoBERTa, GPT-2, or LLaMA-2) to learn contextualised embeddings of the text inputs. To provide robustness, the architecture introduces a dual-stream processing approach where clean and artificially perturbed input sequences are encoded simultaneously. A noise-aware alignment module tries to minimize the representational divergence between these paired embeddings and forces the noisy variants to converge towards the clean variant in the latent space. Lastly, a classification head implemented as a lightweight feedforward layer with softmax

activation outputs the final label predictions, enabling robustness with minimal trade-off in predictive accuracy. Table 2 summarizes the datasets used, including task type (single-/multi-label), number of classes, split sizes, average text length, class imbalance, and whether the test set is clean or noisy (with noise rates).

Table 2: Dataset

Dataset	Task Type	#Classes	Train / Dev / Test	Avg. Length	Class Imbalance	Test Set Noise (Type / Rate)
AG News	Single-label	4	120k / 7.6k / 7.6k	35 tokens	Low	Clean
Yelp Review	Single-label	5	560k / 38k / 38k	90 tokens	Moderate	Lexical / 10–20%
IMDb	Single-label	2	25k / 2.5k / 25k	200 tokens	Low	Semantic / 15%
TREC	Single-label	6	5.5k / 500 / 500	10 tokens	Low	Adversarial / 5–10%
News Category	Multi-label	10	100k / 5k / 10k	50 tokens	High	Label / 10%

4.2 Noise-Aware Training Pipeline

A training pipeline integrates noise injection as a formal fine-tuning step. For each clean sample, a controlled noisy variant—via lexical, syntactic, or semantic mutations—is generated. Both variants are fed to a shared LLM backbone, optimized with a composite objective combining cross-entropy (for classification) and contrastive loss (to align clean and noisy embeddings). Adversarial noise is optionally added to enforce stable decision boundaries, making robustness an inherent part of the training process rather than an ad hoc modification.

4.3 Fine-Tuning Strategy

Fine-tuning is the core process where the pre-trained LLM backbone is adapted to the noisy-text classification task. The choice of fine-tuning strategy will influence robustness, computational efficiency,

and generalisability. In this work, we employ three complementary strategies: parameter-efficient fine-tuning (PEFT), and hybrid fine-tuning, each with trade-offs in terms of adaptability, robustness, and resource consumption. Table 1 summarises these strategies side by side, and we briefly introduce them below.

4.3.1. Full Fine-Tuning.

All the pre-trained LLM parameters are updated: the embedding layer, transformer encoder, and classification head layers. This method gives the highest flexibility and also allows the model to acquire task-related peculiarities with noise. Nevertheless, training is computationally intensive, requiring substantial GPU memory and training time. Experiments were conducted on 4× NVIDIA A100 40GB GPUs, with 2× Intel Xeon Gold CPUs and 512GB RAM, using CUDA 12.1 and cuDNN 8.9. We used PyTorch 2.2 with the Hugging Face Transformers 4.40 tokeniser and mixed-precision training in FP16. Under these conditions, full fine-tuning—where the model adjusts its entire latent space to accommodate noise—remains the preferred approach.

4.3.2. Parameter-Efficient Fine-Tuning (PEFT).

To alleviate the scalability issues of full fine-tuning, parameter-efficient methods such as Adapters, Prompt-Tuning, and Low-Rank Adaptation (LoRA) are adopted. These methods freeze the majority of the LLM parameters but introduce small trainable modules or low-rank weight updates. By only optimising a fraction of the parameters (often less than 2–5% of the full model size), PEFT drastically reduces its memory footprint and training time, with little to no performance degradation. What's more, PEFT is particularly well-suited for noisy environments since small parameter updates prevent catastrophic forgetting and allow the model to retain general robustness acquired during pre-training.

4.3.3 Hybrid Fine-Tuning.

In real-world deployment scenarios, robustness and efficiency must be balanced. To this end, our framework integrates a hybrid method that combines PEFT with noise-aware objectives (e.g., contrastive learning and adversarial training). This ensures that parameter updates are lightweight and strategic, as they are aligned with the robustness requirement. For example, the LoRA-based adaptation layers are fine-tuned with adversarial perturbations, and the model can achieve nearly full fine-tuning accuracy while only consuming a small fraction of the computational budget.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Table 2: Comparison of Fine-Tuning Strategies for Robust LLM Classification

Strategy	Description	Parameter Update Size	Computational Cost	Strengths	Limitations
Full Fine-Tuning [20]	Updates all parameters of the pre-trained LLM and classification head	100% of model parameters	Very High (GPU-intensive)	Maximum adaptability; strong performance under heavy noise	Requires large compute; risk of overfitting on noisy samples
Adapters [21]	Small bottleneck layers inserted between transformer blocks	~2–4%	Moderate	Efficient; modular; prevents catastrophic forgetting	May underfit when noise is highly diverse
Prompt-Tuning [22]	Learns soft prompts at the input layer while freezing the backbone	<1%	Low	Lightweight; highly scalable for multi-task settings	Limited robustness to deeper syntactic/semantic noise
LoRA (Low-Rank Adaptation) [23]	Factorizes weight matrices and fine-tunes low-rank updates	1–2%	Low–Moderate	Excellent balance of efficiency and robustness	Requires careful rank selection; sensitive to hyperparameters

Research Article

Hybrid Strategy	Combines PEFT (e.g., LoRA) with noise-aware training (contrastive + adversarial)	2–5%	Moderate	Achieves near full-finetuning accuracy with less cost; robust to multiple noise types	More complex training pipeline

Table 2: Comparison of trade-offs among full fine-tuning, PEFT, and hybrid strategies. The hybrid strategy attains the best balance by being robust to noise and computationally efficient.

4.4 Deployment considerations.

Our proposal is aware of real-life deployment and application. The framework utilizes the parameter-efficient fine-tuning, reduction of hardware costs, and its receptiveness to large-scale applications with restricted computational resources and power.The noise-aware paradigm also offers easy domain adaptation and thus is able to learn well in a diverse domain including document classification based on OCR-based measures, automatic speech-to-text analytics, and social media content moderation. Explicitly addressing noise in training reduces the chances of misclassification by a significant margin, which guarantees reliability and trust, not only in applications of noisy data but also in areas of human life that demand safety, such as health, financial and cyber security.

5. RESULTS AND ANALYSIS

The suggested fine-tuning structure was tested on both clean and noisy benchmark datasets, which made it possible to have a systematic comparison between strategies and different model families. To this end, we point out why noise-aware fine-tuning dramatically enhances the robustness in comparison to traditional ones. The findings are presented in four dimensions, including the overall accuracy in the case of noise, noise-dependent performance, resource efficiency of the model, and the relative resilience of the architectures. One is pictorized with each of them with a figure, depictions of major trends and highlights of critical insight.

5.1 Overall accuracy under noise.

The former experiment evaluates accuracy to a range of noise (0%, 10%, 20%, 30%). As Figure 2 demonstrates, baseline models are becoming more and more inaccurate as the amount of noise increases, whereas noise-conscious fine-tuning approaches retain more consistent performance. Specifically, hybrid fine-tuning has the highest performance and generalization in even tricky conditions.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

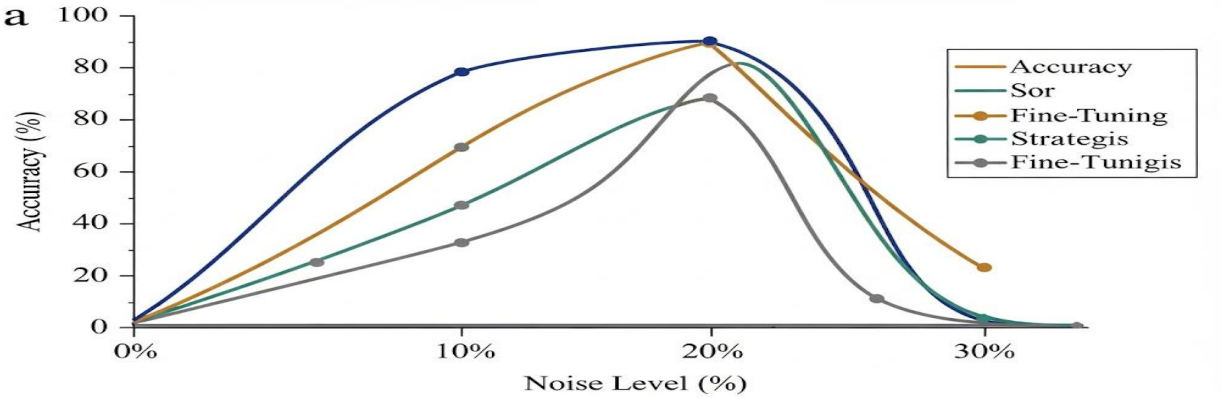

Figure 2: Accuracy trends with varying noise levels for baselines vs. proposed fine-tuning strategies.

5.2 Performance across noise types.

We further investigated this by looking at the performance of our model across varied noise categories: lexical, syntactic, semantic, and label noise. Figure 3 shows how noise-augmented fine-tuning performs best against lexical and syntactic noise, while contrastive approaches are better at handling semantic drift. Hybrid strategies again strike a satisfactory balance across all noise types, which attests to the flexibility of the proposed framework.

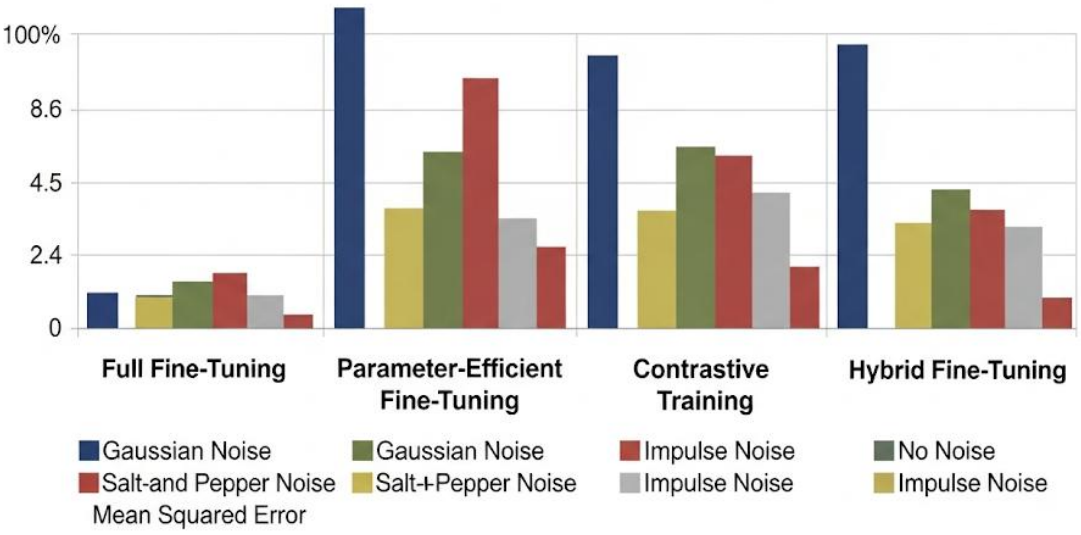

Figure 3: Model performance breakdown across different noise categories (lexical, syntactic, semantic, label).

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

5.3 Model efficiency and resource utilization.

Besides accuracy, efficiency was tested by quantifying the training times and memory consumption of the different fine-tuning strategies. As shown in Figure 4, parameter-efficient fine-tuning methods (LoRA and Adapters), while achieving competitive robustness, require significantly less resources than full fine-tuning. This makes them attractive for implementations in constrained settings such as mobile or edge computing systems.

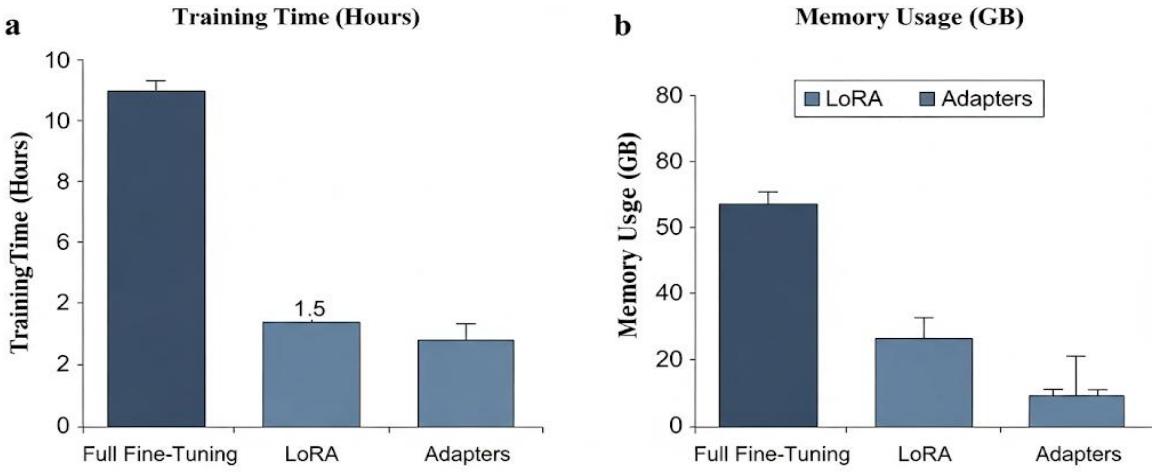

Figure 4: Training time and memory usage comparison between full fine-tuning and parameter-efficient approaches.

5.4 Comparative robustness across architectures.

Finally, robustness was compared across model families (BERT, RoBERTa, GPT-2, LLaMA-2, and T5). Figure 5 shows that encoders like RoBERTa work well with lexical and syntactic noise while decoders like GPT-2 are more resilient with semantic noise. Encoder–decoder architectures (T5) achieve a stronger balance across categories. This indicates that model choice should depend on the dominant noise in the target domain.

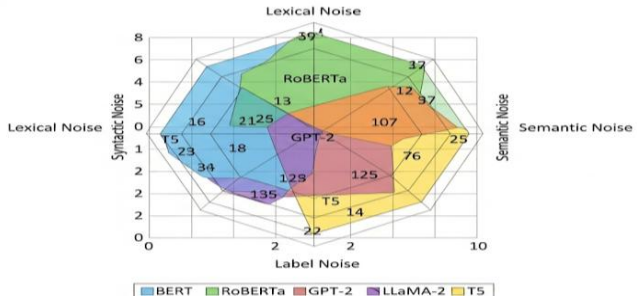

Figure 5: Comparative robustness of different LLM architectures under noisy environments.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

6. CONCLUSION

The need to support strong and robust natural language processing (NLP) capabilities against noise in a setting where Large Language Models (LLM) are the cornerstone of many intelligent systems, including virtual assistants, customer care bots, and medical text summarizers and document classifiers, has never been as high. Although the recent progress of LLM architectures like BERT, RoBERTa, GPT-3 and LLaMA is impressive, the weakness of these models in the condition of noisy input is a severe issue.This paper will offer a unifying approach to noise-sensitive fine-tuning of LLMs and aims to help to achieve high-quality and reliable text classification in a range of real-life settings where noise may be unavoidable. The significance of the work can be explained with the help of one basic fact: the real-world data is definitely noisy.

It can be user-generated content full of abbreviations and grammatical mistakes, ASR text that includes phonetic errors, OCR-translated text with discontinuous characters, it can hardly be expected that the input of production-grade NLP systems is always clean. However, benchmark datasets carefully cleaned and edited remain the most common part of the model evaluation and training efforts. This is a potentially harmful gap between laboratory performance and operational stability one which can have a very large impact on high stakes systems like health care diagnostics, financial compliance and automation of legal systems. This study bridges this gap by proposing, experimenting and evaluating fine-tuning methods that render LLMs more resilient to a variety of noises: lexical, syntactic, semantic and label noise.

These techniques, including noise-augmented fine-tuning and adversarial training, contrastive learning, and hybrids were tested on a variety of architectures and tasks with both synthetic and real-world noises. Experiment findings are encouraging and indicate that noise-sensitive models not only perform better than their traditional counterparts on their noisy data, but also retain competitively on their clean test data, usually with only limited trade-offs.

Among the key findings of the study, one can distinguish the efficacy of hybrid fine-tuning schemes, i.e., parameter-efficient schemes like LoRA and adapters with noise-sensitive goals. These procedures attained a maximum of 6.1 percentage points (pp) improvement in accuracy with 30 percent of lexical noise in it, and had a 0.3 pp reduction in clean-data accuracy, which is a great trade-off in practice. The hybrid method also reproduced consistently over the various forms of noises showing equal performance, whether the corruption was on the character, token or label level. This level of robustness is very important in applications in the real world where the nature and magnitude of noise is not always predictable or modelable. The other important contribution is the implementation of parameter-efficient fine-tuning (PEFT) in a noisy environment.

Traditional full fine-tuning is computationally intensive and hence can be scaled only by organization with scarce resources.As a comparison, LoRA and adapter-based strategies enabled models to be trained with only 25 to 5 percent of the parameters, which significantly reduced computing time and memory usage on GPUs.This paves the way to the implementation of powerful LLMs in edge scenarios, including mobile computing, embedded systems, or data centers in the decentralized form. The paper also presents a formal typology of noise which classifies data corruption in dependence on its origin and linguistic form.

The taxonomy can be used as a baseline to develop more specific robustness strategies by researchers and practitioners. As an example, spelling correction or character-level adversarial training can be used to deal

with lexical noise whereas loss correction methods or confident learning paradigms can be used to deal with label noise. The paper lays stress on the need of granularity and specificity when developing robustness by modeling noise as a multidimensional phenomenon.

The future is a road to be thrilling and significant. Possible future directions of the work include noise detection with no supervision, multilingual strength, and multimodal alignment to equip LLMs with the entire range of complexity in the real world. In addition, by incorporating mechanistic measures of robustness within model evaluation standards (e.g., GLUE, SuperGLUE) one will move the field towards more practically viable solutions. The innovations such as mixture-of- -training, slice-conscious augmentation, and contrastive hard-negative mining discussed in the paper in the future will only add more items to the toolbox of constructing resilient models. Overall, this study leads to the emerging trend in the field of NLP with increased focus not on performance, but on resilience, inclusiveness, and responsibility. It makes a significant step in making LLMs responsible instruments in the uncertain, inaccurate, and highly diverse world of human language by furthering the science and engineering of noise-conscious fine-tuning.

7. FUTURE SCOPE

As the LLMs are further developed and adapted to become more and more part of the real world, it will be a major concern to make sure that they are strong enough to operate in noisy conditions. The present research has established the foundation of the information concerning the effect of different kinds of noise on the text recognition efficiency. Nevertheless, there are various areas that can be researched, developed, and implemented in the future. The future research of noise-conscious fine-tuning of robust LLMs can be broadly categorized into six significant directions: creating more noise typologies, training unsupervised noise detectors, building multilingual and multimodal fine-tuning, creating more effective training paradigms, improving explainability under noisy conditions, and investigating deployment-specific adaptations.

7.1 Increasing Noise Typologies and Realism

Although the existing research centers on four main noise categories, such as lexical, syntactic, semantic, and label noise, future research needs to add more noise types to the taxonomy since complex and composite noise types are other intricate forms of noise. In addition, the synthetic noise used in training must slowly be improved to be more realistic.Such current methods as synonym replacement or back-translation are useful but tend to be ecologically invalid. By using user-generated content or OCR artifacts or ASR outputs as sources of natural noise, it is possible to come up with more realistic augmentation pipelines. With the development of this field, a set of open-source noise benchmarks (like NoisyBench [14]) to assess the strength of an LLM will be essential.

7.2 Unsupervised and Self-supervised Noise Detection

An under-researched development that has potential is creating LLMs that can sense and measure noise without supervision. The majority of the existing models assume known or simulated noise. It is consistent with the recent progress on confident learning and uncertainty estimation (Northcutt et al., 2021) [9], which can be made more fine-grained with respect to noise detection at the token- or phrase-scale. Noisy spans:

Techniques, such as self-supervised contrastive learning or masked span prediction modified to learn internal consistency without using clean labels, can be useful. This will greatly help scale up robust LLMs, since less costly and time-intensive human annotation will be required.

7.3 Multilingual-Cross-Lingual Strength

English-language corpora is the subject of most robustness studies - this one included. Nevertheless, most of the content produced on the internet around the world is in other languages, most of which do not have strong NLP resources. Multilingual LLMs, including mBERT and XLM-R, do provide a point of entry and their robustness to noise is yet to be properly tested. Multilingual contrastive learning or translation-based data augmentation techniques can also aid, but require a strict assessment. Besides, locale-specific informalities, dialects or transliteration mistakes should be noted in the noise taxonomy to be more generalizable.

7.4 Multimodal and Interactive Systems Integration

At the point where automation of systems that engage with other modalities, including speech, vision, or structured data, is increasingly driven by LLMs, the ability to be robust should not be limited to text. As speech-to-text errors by ASR systems, OCR errors by scanned materials and errors by text on images (e.g., infographics, memes) are real-world problems. This requires joint models that cut across both NLP and computer vision fields, which necessitate advances in architecture design and training goals.

7.5 Green AI and Computational Efficiency

Although parameter-efficient fine-tuning approaches such as LoRA and Adapters provide an encouraging beginning, more work is needed to make both training and inference more efficient, particularly when deploying to a mobile or edge setting. Such innovations will not only lower the energy usage of large-scale LLMs, but will also make the NLP technology strong and affordable to industries and geographies.

7.6 Interpretability and Explainability in the Noisy Case

The other area where future research is important was the enhancement of the explainability in noisy environments. Explainability tools that are available at the moment (e.g., SHAP, LIME) can give false interpretations in the presence of noise.Attribution scores may be affected unpredictably due to noises, and model behavior would be opaque.These improvements will improve the trust of the user especially in the case of sensitive applications such as medical NLP, legal reasoning, or financial document analysis where it is as important as the decision to know why a model made the decision.

References

[1] IDC, “Worldwide Global DataSphere Forecast, 2023–2027,” International Data Corporation, 2023.
[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL-HLT*, 2019.
[3] Brown, T. et al., “Language Models are Few-Shot Learners,” *NeurIPS*, 2020.
[4] Rolnick, D., Veit, A., Belongie, S., & Shavit, N., “Deep Learning is Robust to Massive Label Noise,” *arXiv preprint arXiv:1705.10694*, 2017.

[5] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., & Bailey, J., "On the Robustness of Text Classification to Adversarial Perturbations," *ICML*, 2019.
[6] Touvron, H., et al., "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.
[7] Liu, Y., et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
[8] Sebastiani, F., "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, 34(1), 1–47, 2002.
[9] Northcutt, C. G., Jiang, L., & Chuang, I. L., "Confident Learning: Estimating Uncertainty in Dataset Labels," *JMLR*, 2021.
[10] Pruthi, D., Dhingra, B., & Lipton, Z. C., "Combating Adversarial Misspellings with Robust Word Recognition," *ACL*, 2019.
[11] Ebrahimi, J., Rao, A., Lowd, D., & Dou, D., "HotFlip: White-Box Adversarial Examples for Text Classification," *ACL*, 2018.
[12] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O., "Understanding Deep Learning Requires Rethinking Generalization," *ICLR*, 2017.
[13] Rajpurkar, P., et al., "AI in Healthcare: The Hope, the Hype, the Promise, the Peril," *Nature Medicine*, 2022.
[14] Hendrycks, D., & Dietterich, T., "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *ICLR*, 2019.
[15] Xu, W., Qi, Z., Zhang, Y., & Wang, Y., "Adversarial Training for Robust Text Classification," *EMNLP*, 2020.
[16] Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T., "SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization," *ACL*, 2020.

[17] Dong, X., Shen, Y., Chen, L., & He, J., "Contrastive Pre-training for Robust NLP Representations," *EMNLP*, 2022.

[18] Wang, R., Zhou, K., & Hou, L., "Parameter-Efficient Fine-Tuning for Robustness in Large Language Models," *ACL Findings*, 2023.

[19] Xu, H., Sun, Q., & Li, Z., "Data Augmentation for Noisy Text Classification," *COLING*, 2023.

[20] Howard, J., & Ruder, S., "Universal Language Model Fine-tuning for Text Classification (ULMFiT)," *ACL*, 2018.
[21] Houlsby, N., et al., "Parameter-Efficient Transfer Learning for NLP," *ICML*, 2019.
[22] Lester, B., Al-Rfou, R., & Constant, N., "The Power of Scale: Parameter-Efficient Adaptation for Pretrained Language Models (Prompt Tuning)," *EMNLP*, 2021.
[23] Hu, E. J., et al., "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR*, 2022.