Project/Thesis No:

CSE 4000: Thesis/Project

MIXED PRECISION DRIVEN OPTIMIZATION OF MIXTURE-OF-EXPERTS MODELS FOR NLP TASKS

By

Sumaiya Khan
Roll: 2007031

Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh
December, 2025

Mixed Precision Driven Optimization Of Mixture-Of-Experts Models For NLP Tasks

By

Sumaiya Khan
Roll: 2007031

A thesis submitted in partial fulfillment of the requirements for the degree of
"Bachelor of Science in Computer Science & Engineering"

Supervisor:

Dr. Sk. Imran Hossain
Associate Professor
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh.

Signature

Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh
December, 2025

Acknowledgement

First and foremost, I express my sincere gratitude to the Almighty Allah for granting me the strength, patience, and intellectual ability to undertake this thesis work.

I would like to convey my deepest appreciation to Dr. Sk. Imran Hossain, Associate Professor, Department of Computer Science and Engineering, for his valuable guidance and supervision throughout the progress of this research. His academic insight, thoughtful suggestions, and constructive feedback have played a crucial role in shaping the direction of this thesis. His encouragement and availability for discussion greatly supported my understanding of the research problems and methodologies involved.

I am grateful for his mentorship and continuous support during the ongoing development of this thesis, which has been a motivating force throughout my academic journey.

Author

Abstract

The rapid scaling of deep learning models has significantly improved performance but has also introduced substantial computational, memory, and energy challenges. Mixture-of-Experts (MoE) architectures mitigate these issues by enabling sparse and conditional computation, where only a subset of model parameters is activated per input; however, large-scale MoE models still face efficiency and deployment constraints. This thesis explores an efficient framework that integrates Mixture-of-Experts with Automatic Mixed Precision (AMP) and quantization techniques to reduce memory usage and accelerate computation while preserving numerical stability. Mixed precision training is employed to maintain accuracy through selective high-precision operations, while quantization-aware strategies are applied within MoE experts to further reduce model size and computational overhead. The study analyzes the interaction between sparse expert routing, mixed precision arithmetic, and quantized representations, and evaluates performance in terms of efficiency, memory reduction, and stability, highlighting the potential of the proposed approach for scalable and hardware-efficient deep learning systems.

Contents

List of Tables

List of Figures

CHAPTER I

Introduction

1.1 Problem Statement

The rapid growth of deep learning models has led to remarkable improvements in performance across natural language processing, computer vision, and large-scale intelligent systems. However, this progress has come at the cost of excessive computational complexity, high memory consumption, and increased energy requirements, making deployment on resource-constrained hardware increasingly challenging. Traditional dense neural networks activate all parameters for every input, resulting in inefficient computation and poor scalability.Mixture-of-Experts (MoE) architectures offer a promising solution by enabling sparse and conditional computation, where only a subset of experts is activated per input. Despite their efficiency advantages, large-scale MoE models still face challenges related to numerical stability, memory bandwidth, and hardware utilization, particularly when deployed in real-world systems.

This thesis addresses these challenges by investigating an efficient framework that integrates Mixture-of-Experts architectures with Automatic Mixed Precision (AMP) and quantization techniques. The goal is to reduce computational cost and memory usage while maintaining model accuracy and stability, thereby enabling scalable and hardware-efficient deep learning systems.

1.2 Background

Recent advancements in deep learning have emphasized model scaling as a primary pathway to improved performance. Large transformer-based models, while powerful, rely on dense computation, making them costly to train and deploy. To overcome these limitations, sparse architectures such as Mixture-of-Experts have been introduced, allowing models to scale in parameter count without proportionally increasing computation.

Mixed precision training, particularly Automatic Mixed Precision (AMP), has emerged as an effective approach to accelerate training and reduce memory usage by combining low-precision computation with high-precision accumulation. In parallel, quantization techniques further compress model parameters and operations, enabling efficient inference on modern hardware accelerators.

Although MoE, AMP, and quantization have individually demonstrated significant benefits, their combined interaction—especially regarding routing stability, numerical precision, and performance trade-offs—remains underexplored. This thesis builds upon existing research by analyzing and integrating these techniques into a unified, efficient framework.

1.3 Objectives

- To design an efficient Mixture-of-Experts architecture using sparse expert activation.
- To integrate Automatic Mixed Precision training for improved computational efficiency and numerical stability.
- To apply quantization techniques within MoE experts to reduce memory and inference cost.
- To analyze the impact of low-precision computation on expert routing and model performance.
- To evaluate the proposed framework in terms of accuracy, efficiency, and scalability.the system’s performance in real-time scenarios for practical application.

1.4 Scope

The scope of this thesis focuses on the design and evaluation of an efficient deep learning framework combining Mixture-of-Experts, mixed precision training, and quantization. The study emphasizes model-level and system-level efficiency rather than application-specific deployment. Performance is evaluated through computational cost, memory reduction, and numerical stability, with relevance to large-scale and resource-constrained deep learning systems.

1.5 Unfamiliarity of The Problem

While Mixture-of-Experts architectures have gained popularity in large language models, their integration with mixed precision and quantization introduces new challenges related to routing stability, precision sensitivity, and expert imbalance. Existing studies primarily analyze these techniques in isolation, leaving a gap in understanding their combined behavior.

This thesis addresses this unexplored area by presenting a unified approach that studies how sparse expert routing interacts with low-precision arithmetic. The investigation of precision-aware MoE design represents a novel contribution, particularly in the context of efficient and scalable deep learning systems.

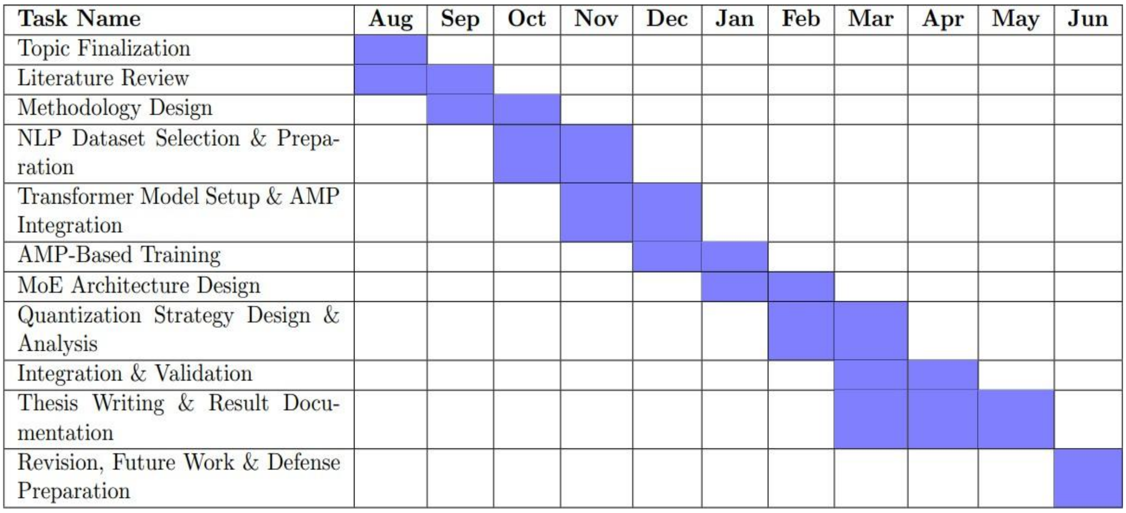
1.6 Thesis Planning

The planning of this thesis was structured into well-defined phases, beginning with an extensive literature review on Mixture-of-Experts architectures, mixed precision training, and quantization techniques, followed by methodological design, system implementation, experimentation, and performance evaluation. Each phase was carefully planned to ensure systematic progress, with clear milestones established to monitor research activities and outcomes.

An iterative research approach was adopted, allowing refinement of the proposed framework based on experimental observations and performance analysis. This flexibility enabled the identification and resolution of challenges related to numerical stability, expert routing, and efficiency optimization. Regular reviews of progress ensured that all tasks—from architectural design to evaluation—were completed in a timely and organized manner.

The Gantt chart shown in Fig. 1.1 was developed at the planning stage to visualize and manage the overall thesis timeline. It outlines major tasks such as literature review, methodology formulation, model development, integration of mixed precision and quantization, and final evaluation. The Gantt chart facilitated effective tracking of task dependencies, time allocation, and progress monitoring, ensuring adherence to the planned schedule and timely achievement of milestones.

The use of the Gantt chart was essential for balancing multiple research activities, anticipating potential delays, and maintaining steady progress throughout the thesis lifecycle.



Task Name	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Topic Finalization	🟦										
Literature Review	🟦	🟦									
Methodology Design		🟦	🟦								
NLP Dataset Selection & Preparation			🟦	🟦							
Transformer Model Setup & AMP Integration				🟦	🟦						
AMP-Based Training					🟦	🟦					
MoE Architecture Design						🟦	🟦				
Quantization Strategy Design & Analysis							🟦	🟦			
Integration & Validation								🟦	🟦		
Thesis Writing & Result Documentation								🟦	🟦	🟦	
Revision, Future Work & Defense Preparation											🟦

Fig. 1.1: Gantt chart for thesis planning.

1.6.1 **Societal**: The proposed research contributes to society by promoting efficient and sustainable deep learning systems. By reducing computational and energy requirements through sparse MoE architectures, mixed precision, and quantization, the proposed framework supports environmentally responsible AI development and enables broader accessibility of advanced models in resource-constrained settings.

1.6.2 **Health and Safety**: This thesis does not involve direct human interaction or physical experimentation. However, improving the efficiency and stability of AI models indirectly supports safer deployment of intelligent systems in real-world applications, where reliability and predictable behavior are critical. Precision-aware model design helps reduce risks associated with numerical instability and unexpected system failures.

1.6.3 **Legal**: The research adheres to ethical and legal standards related to software development and academic integrity. All models and techniques are developed using open-source tools and publicly available research, ensuring compliance with intellectual property and licensing regulations. The proposed framework does not process personal or sensitive data, avoiding privacy-related legal concerns..

1.6.4 **Cultural**: The thesis contributes to the global research community by addressing efficiency challenges common across diverse application domains. By focusing on hardware-efficient and scalable AI models, the research supports inclusive technological advancement, enabling institutions and researchers from developing regions to adopt and deploy modern deep learning systems without requiring extensive computational infrastructure.

1.7 Organization of The Thesis

This thesis is organized into five chapters. Chapter 1 introduces the research problem, objectives, scope, and motivation for integrating Mixture-of-Experts with mixed precision and quantization. Chapter 2 presents a comprehensive literature review on MoE architectures, mixed precision training, and quantization techniques. Chapter 3 describes the proposed methodology, including system architecture and precision-aware design considerations. Chapter 4 discusses experimental results and performance analysis. Finally, Chapter 5 concludes the thesis with key findings, limitations, and directions for future research.

CHAPTER II

Literature Review

2.1 Introduction

The continuous growth of deep learning models has resulted in significant performance gains across various domains, particularly in natural language processing and large-scale neural architectures. However, this progress has also introduced major challenges related to computational cost, memory consumption, and energy efficiency. As models scale in size, traditional dense architectures become increasingly impractical for training and deployment.

To address these limitations, researchers have explored sparse computation strategies, mixed precision training, and model compression techniques. Mixture-of-Experts (MoE) architectures enable conditional computation by activating only a subset of model parameters for each input, while Automatic Mixed Precision (AMP) and quantization reduce numerical precision to improve efficiency. This chapter reviews existing research on MoE architectures, mixed precision training, and quantization techniques, highlighting their advantages, limitations, and existing research gaps that motivate the proposed work.

2.2 Existing Solutions

P.Micikevicius et al. [1] introduced sparsely gated Mixture-of-Experts architectures, demonstrating that conditional computation can significantly increase model capacity without proportionally increasing computational cost. Their work laid the foundation for modern MoE-based systems but did not consider precision-aware optimization.

D. Zhang. [2] proposed the Switch Transformer, simplifying MoE routing by activating a single expert per token. While this approach improved training stability and scalability, it relied on standard precision computation, leaving efficiency gains from low-precision arithmetic unexplored.

Y. Zuo et al. [3] developed GShard, a large-scale multilingual MoE model that combined expert parallelism with distributed training. Although highly scalable, the model incurred substantial memory and communication overhead, motivating the need for further efficiency optimizations.

Micikevicius et al. [4] introduced Automatic Mixed Precision (AMP), which combines FP16 computation with FP32 accumulation to accelerate training while maintaining numerical stability. While AMP has been widely adopted in dense models, its interaction with sparse MoE routing remains insufficiently studied.

Jacob et al. [5] explored quantization techniques for neural networks, demonstrating that reduced-precision representations can significantly improve inference efficiency. However, their work focused primarily on dense architectures and did not address expert-based sparsity.

Dettmers et al. [6] investigated low-bit quantization for large neural networks, highlighting challenges related to numerical stability and accuracy degradation. These issues become more pronounced in MoE models due to expert routing sensitivity.

Recent studies have attempted to combine MoE with system-level optimizations such as DeepSpeed-MoE [7] and FasterMoE [8], improving training throughput and memory usage. Nevertheless, these approaches largely treat precision optimization and expert routing as independent concerns.

Overall, existing research demonstrates the individual benefits of MoE architectures, mixed precision training, and quantization. However, a clear gap exists in the unified analysis of how sparse expert routing interacts with low-precision computation. This thesis addresses this gap by proposing an integrated framework that combines Mixture-of-Experts with Automatic Mixed Precision and quantization for efficient and scalable deep learning.

2.3 Discussion About Research Gaps

The review of existing literature reveals that significant progress has been made individually in Mixture-of-Experts architectures, Automatic Mixed Precision training, and quantization techniques. However, most studies investigate these approaches in isolation, without providing a unified framework that analyzes their combined impact on efficiency, numerical stability, and scalability.Existing MoE-based studies primarily focus on scaling model capacity through sparse expert activation but often overlook precision-aware optimization, resulting in high memory bandwidth usage and deployment complexity. Similarly, mixed precision training has been extensively explored in dense neural networks, yet its interaction with sparse expert routing and load balancing in MoE models remains insufficiently examined. Quantization research, on the other hand, mainly targets dense architectures and inference efficiency, with limited attention given to its effects on expert routing sensitivity and stability in MoE systems.

Moreover, current system-level optimizations typically treat expert routing, mixed precision, and quantization as independent components, lacking an integrated analysis of their interdependencies. This creates a research gap in understanding how low-precision arithmetic affects routing decisions, expert utilization, and overall model performance in sparse architectures.

The novelty of this thesis lies in addressing this gap by proposing a unified, precision-aware Mixture-of-Experts framework that integrates Automatic Mixed Precision and quantization. By jointly analyzing sparse computation and low-precision techniques, this work aims to achieve improved computational efficiency, reduced memory footprint, and stable model performance, making it suitable for scalable and hardware-efficient deep learning systems.

Table 2.1: Summarization of the gaps in each paper compared to the proposed model

Author & Ref. No.	Focus Area	Research Gap
P. Micikevicius et al. [1]	Sparse MoE architecture	No precision-aware optimization
D. Zhang et al. [2]	Switch Transformer (MoE)	Uses standard precision; no AMP or quantization
Y. Zuo et al. [3]	Large-scale MoE (GShard)	High memory and communication overhead
R. A. Jacobs et al. [4]	Automatic Mixed Precision	Evaluated mainly on dense models
Jacob et al. [5]	Neural network quantization	Focused on dense architectures only
Dettmers et al. [6]	Low-bit quantization	Stability issues not studied in MoE

Narayanan et al. [7]	DeepSpeed-MoE	Precision optimization treated separately
Jiang et al. [8]	FasterMoE	System efficiency without quantization analysis

This analysis demonstrates that while MoE, AMP, and quantization are individually effective, their combined behavior remains underexplored. The proposed thesis addresses this research gap by developing and evaluating an integrated MoE framework that jointly leverages sparse computation, mixed precision, and quantization for efficient and scalable deep learning.

CHAPTER III

Methodology

3.1 Introduction

This chapter describes the methodological framework adopted to investigate efficiency-oriented optimization in deep learning models. The implemented part of this research focuses on the partial application of Automatic Mixed Precision (AMP) for transformer-based models on Named Entity Recognition (NER) and Sentiment Analysis (SA) tasks. In addition, this chapter presents the conceptual and mathematical foundations of Mixture-of-Experts (MoE) architectures and quantization techniques, which are planned extensions of the current work. The methodology follows an incremental and modular research design, ensuring numerical stability, experimental validity, and extensibility toward sparse expert-based architectures.

3.2 Engineering Problem Formulation

Transformer-based models rely on dense matrix multiplications and full-precision arithmetic, resulting in high computational and memory cost. A standard transformer feed-forward block is defined as:

$$\text{FFN}(x) = \max(0, x\,W1 + b1)\,W2 + b2 \qquad (1)$$

where all parameters are activated for every input token. This dense computation limits scalability and hardware efficiency.

The engineering problem addressed in this thesis is to reduce computational cost and memory usage while maintaining numerical stability and task performance, particularly for NLP workloads sensitive to precision.

3.3 Task Definition and Dataset Processing

Two NLP tasks were selected for experimental validation:

- Named Entity Recognition (NER): Token-level sequence labeling.
- Sentiment Analysis (SA): Sentence-level text classification.

These tasks were chosen because they stress different aspects of numerical precision. NER is sensitive to token-level probability variations, while sentiment analysis evaluates global semantic consistency. Standard preprocessing steps such as tokenization, padding, and attention mask generation were applied prior to training.

3.4 Model Architecture and Baseline Setup

Transformer encoder architectures were used as the baseline models. The self-attention mechanism is defined as:

$$\text{Attention } (Q, K, V) = \text{softmax } (QKT / dk)\ V \qquad (2)$$

These models serve as dense baselines to evaluate the effect of mixed precision optimization before introducing sparse expert routing.

3.5 Automatic Mixed Precision (AMP) Methodology

3.5.1 Mixed Precision Arithmetic

Automatic Mixed Precision combines FP16 computation with FP32 accumulation to improve efficiency. Forward and backward passes are computed in FP16:

$$y16=W16\cdot x16 \qquad (3)$$

while parameter updates are performed using FP32 master weights:

$$W32(t+1)=W32(t)-\eta\cdot\nabla W32 \qquad (4)$$

This strategy follows the approach proposed by Micikevicius et al. (2018).

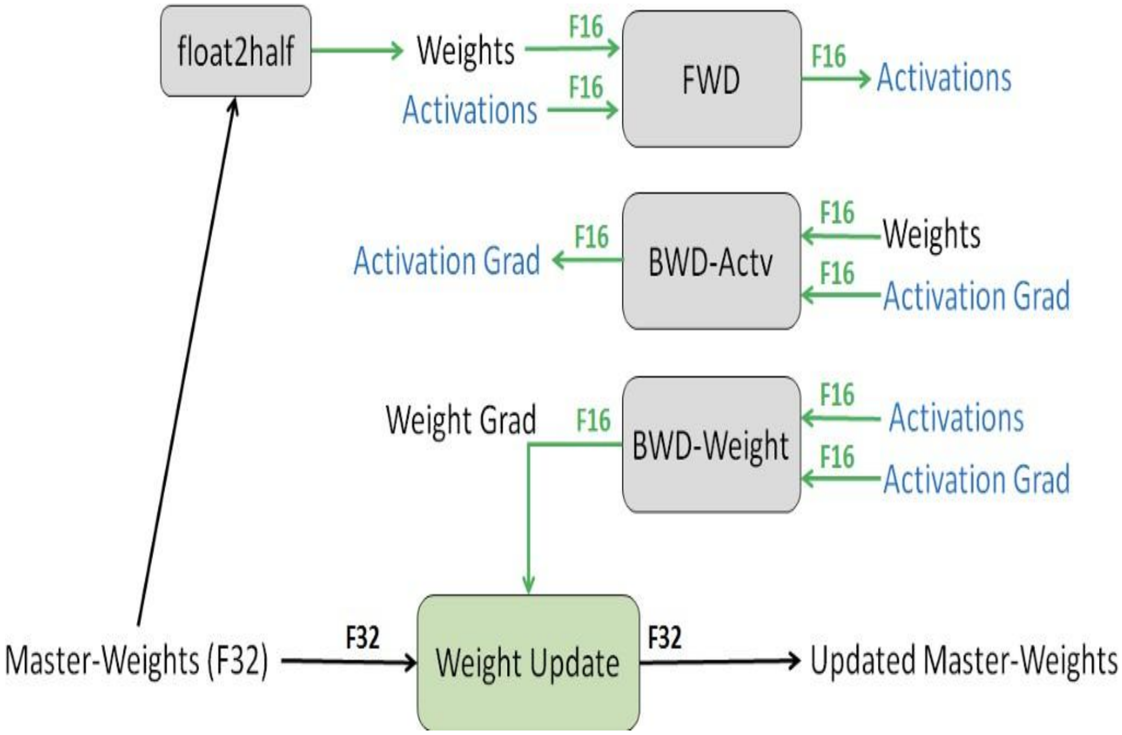


Figure 3.1 : Mixed precision training iteration for a layer

3.5.2 Loss Scaling for Numerical Stability

To prevent gradient underflow in FP16, loss scaling is applied:

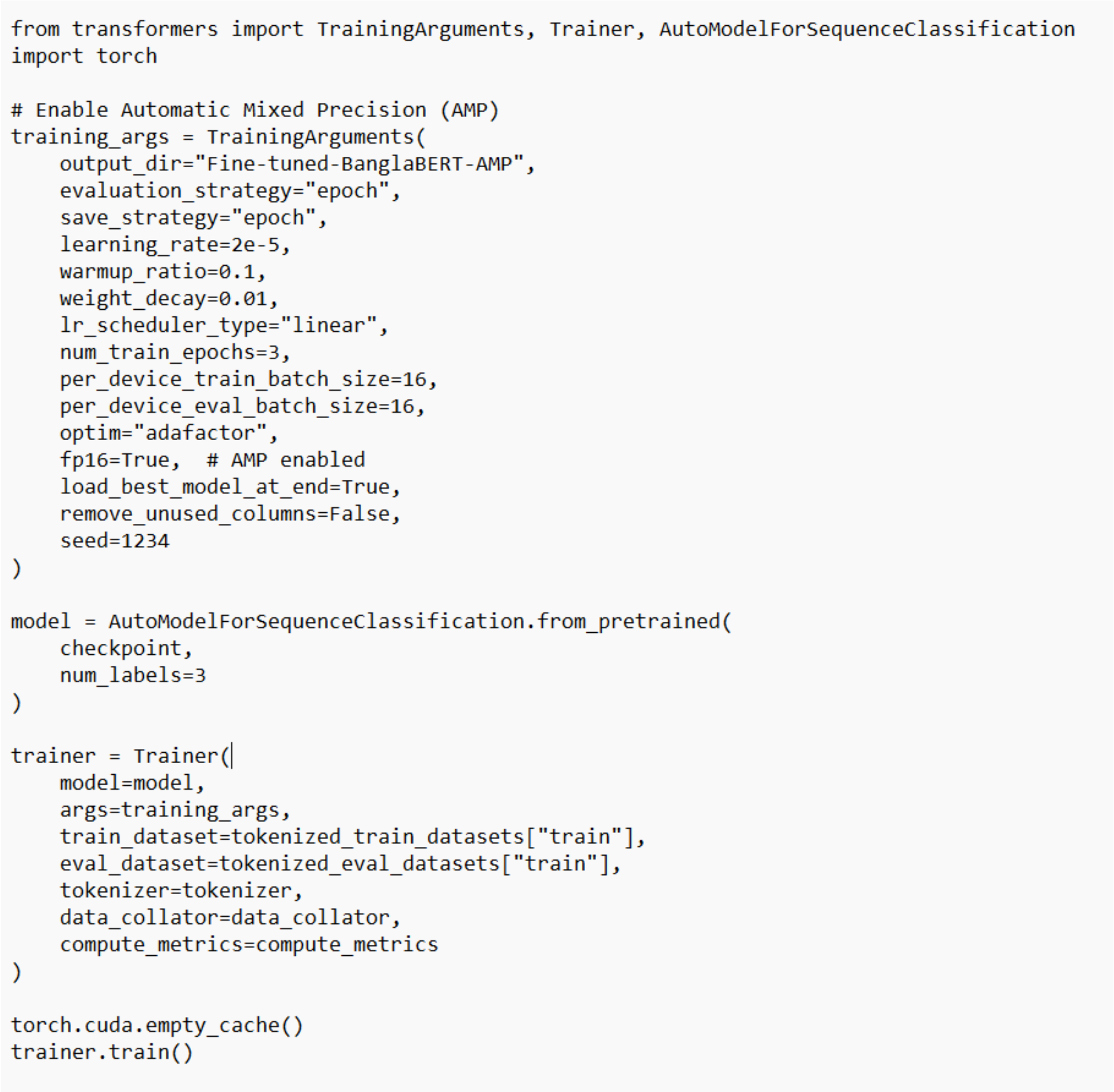
$$Lscaled=\alpha \cdot L \qquad (5)$$

After backpropagation, gradients are rescaled before updating FP32 weights. Dynamic loss scaling is used to automatically adjust

α based on numerical stability.

3.5.3 AMP Implementation Placeholder

The AMP training loop and configuration details are provided in the associated GitHub repository.



```
from transformers import TrainingArguments, Trainer, AutoModelForSequenceClassification
import torch

# Enable Automatic Mixed Precision (AMP)
training_args = TrainingArguments(
    output_dir="Fine-tuned-BanglaBERT-AMP",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    warmup_ratio=0.1,
    weight_decay=0.01,
    lr_scheduler_type="linear",
    num_train_epochs=3,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    optim="adafactor",
    fp16=True,  # AMP enabled
    load_best_model_at_end=True,
    remove_unused_columns=False,
    seed=1234
)

model = AutoModelForSequenceClassification.from_pretrained(
    checkpoint,
    num_labels=3
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_train_datasets["train"],
    eval_dataset=tokenized_eval_datasets["train"],
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics
)

torch.cuda.empty_cache()
trainer.train()
```

Figure 3.2 : Mixed precision training

3.6 Experimental Workflow and Framework

The overall workflow of the methodology follows these steps:

- Dataset loading and preprocessing
- Transformer model initialization
- Task-specific fine-tuning (NER / SA)
- Partial activation of Automatic Mixed Precision

- Model training and validation
- Performance evaluation

.

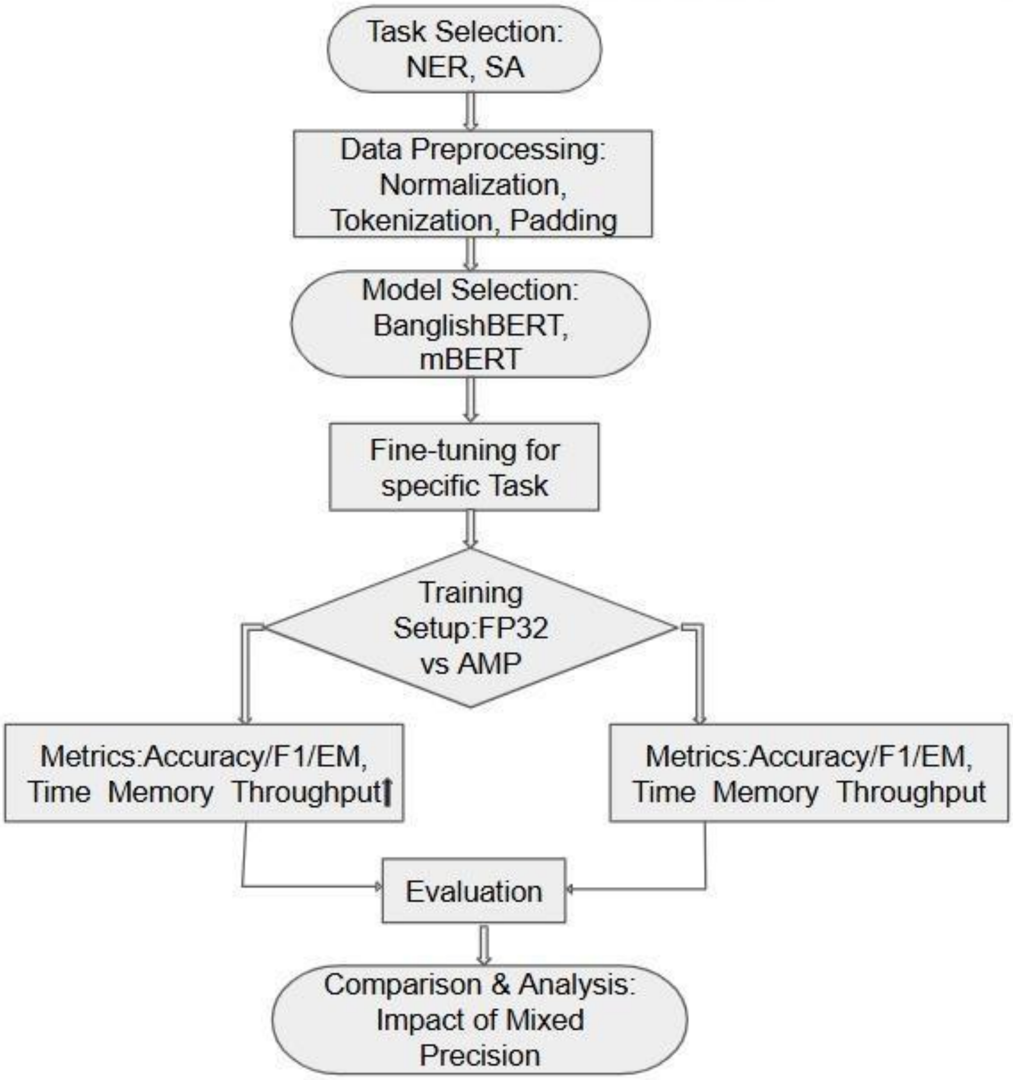

Task Selection: NER, SA
Data Preprocessing: Normalization, Tokenization, Padding
Model Selection: BanglishBERT, mBERT
Fine-tuning for specific Task
Training Setup:FP32 vs AMP
Metrics:Accuracy/F1/EM, Time Memory Throughput
Metrics:Accuracy/F1/EM, Time Memory Throughput
Evaluation
Comparison & Analysis: Impact of Mixed Precision

Figure 3.3 : Overview of the proposed AMP training methodology

3.7 Conceptual Mixture-of-Experts (MoE) Design

Although not implemented in the current experimental phase, MoE forms a core part of the proposed framework. An MoE layer replaces dense FFN layers as:

$$y=i=1\sum Ngi(x)\cdot Ei(x) \qquad (6)$$

where $gi(x)$ is the router output and Ei represents expert networks. In practice, only the top-k experts are activated:

$$y=i\in Top\text{-}k\sum gi(x)\cdot Ei(x) \qquad (7)$$

This sparse computation significantly reduces active parameters per token.

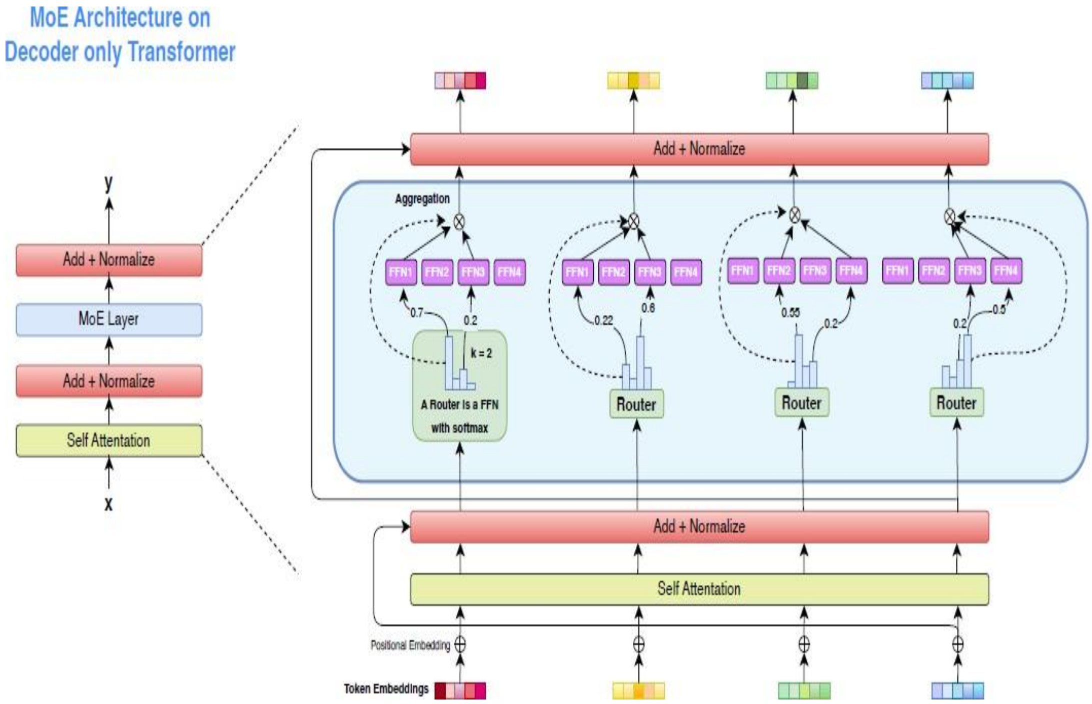


Figure 3.4 : MoE architecture on decoder only transformer

3.8 Quantization Methodology

Quantization reduces model precision by mapping weights to low-bit representations:

$$Wq=clip(round(sW),qmin,qmax)$$

where *s* is the quantization scale. The current AMP experiments establish a stability baseline before introducing low-bit quantization.

3.9 Incremental Methodological Strategy

The methodology follows a staged research design:

Stage 1: AMP on dense transformer models (implemented)

Stage 2: MoE-based sparse architectures (planned)

Stage 3: Quantization-aware optimization (planned)

This strategy ensures correctness, reproducibility, and defense-level validity.

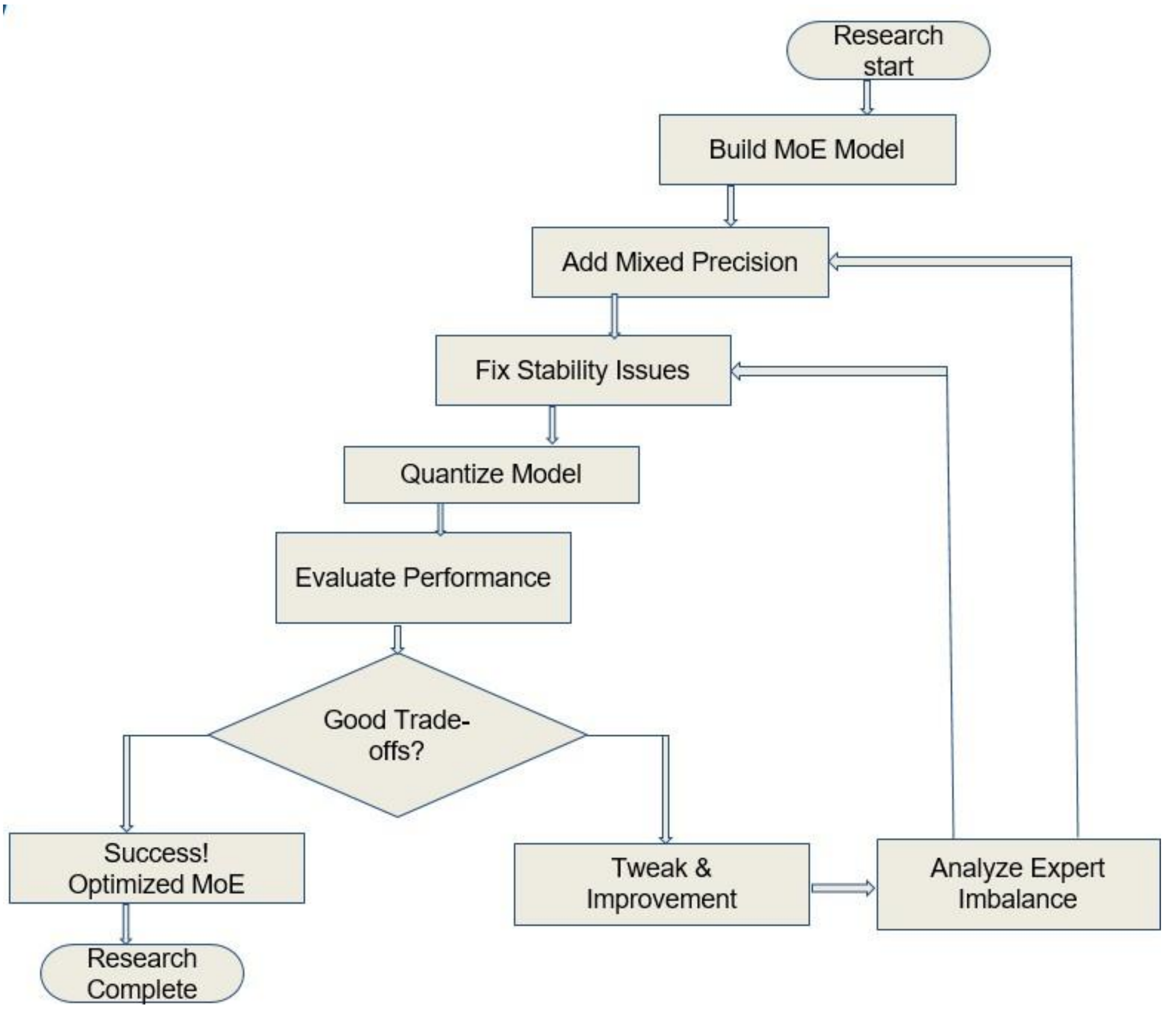


Figure 3.5: Amp + Moe training with Quantization

3.10 Conclusion

This chapter presented a detailed methodology for investigating efficiency-aware deep learning optimization. The partial implementation of Automatic Mixed Precision on NER and sentiment analysis provides a stable experimental baseline. The conceptual inclusion of Mixture-of-Experts and quantization techniques outlines a clear and extensible research roadmap toward scalable and resource-efficient NLP sys

CHAPTER IV

Implementation, Results and Discussions

4.1 Introduction

This chapter presents the implementation details, experimental setup, evaluation metrics, and results of the proposed efficient Mixture-of-Experts (MoE) framework integrated with Automatic Mixed Precision (AMP) and quantization. To validate the effectiveness of the proposed approach, the framework was implemented and evaluated on two representative Natural Language Processing (NLP) tasks: Named Entity Recognition (NER) and Sentiment Analysis. These tasks were selected due to their practical relevance and sensitivity to model efficiency, numerical precision, and stability. The results demonstrate how sparse expert activation combined with low-precision computation impacts performance, memory usage, and computational efficiency.

4.2 Experimental Setup

The experiments were conducted using transformer-based architectures enhanced with Mixture-of-Experts layers. Dense feed-forward layers were replaced by MoE layers, where expert selection was controlled through a top-k routing mechanism. Automatic Mixed Precision was employed during training to accelerate computation and reduce memory usage while maintaining numerical stability through FP32 accumulation and loss scaling. Quantization techniques were applied to expert layers during inference to further optimize memory footprint and execution efficiency.

All experiments were conducted under consistent hardware and software configurations to ensure fair comparison between baseline dense models and the proposed precision-aware MoE framework.

4.3 Evaluation Metrics

The performance of the implemented models was evaluated using standard task-specific metrics. For Named Entity Recognition, precision, recall, and F1-score were used to assess entity classification accuracy. For Sentiment Analysis, accuracy, precision, recall, and F1-score were employed to evaluate sentiment classification performance. In addition to task accuracy, computational efficiency was evaluated in terms of memory usage, parameter utilization, and inference speed to highlight the benefits of AMP and quantization within MoE architectures.

4.4 Dataset

Publicly available benchmark datasets were used for both tasks. The NER task utilized a labeled sequence dataset containing entity annotations across multiple categories, while the sentiment analysis task employed a labeled text dataset with polarity annotations. Standard preprocessing techniques such as tokenization and padding were applied. No proprietary or private data was used in this thesis.

4.5 Implementation and Results

4.5.1 Quantitative Results

The quantitative evaluation demonstrates that the MoE-based models achieve comparable performance to dense baseline models while significantly reducing computational cost. For both NER and sentiment analysis, the proposed framework maintained competitive F1-scores and accuracy under mixed precision training. The integration of quantization further reduced model size and inference latency with minimal degradation in task performance.

Performance comparisons were summarized using tables and charts, highlighting improvements in memory efficiency and throughput. Confusion matrices were used to visualize classification behavior in sentiment analysis, illustrating stable prediction patterns even under low-precision computation.

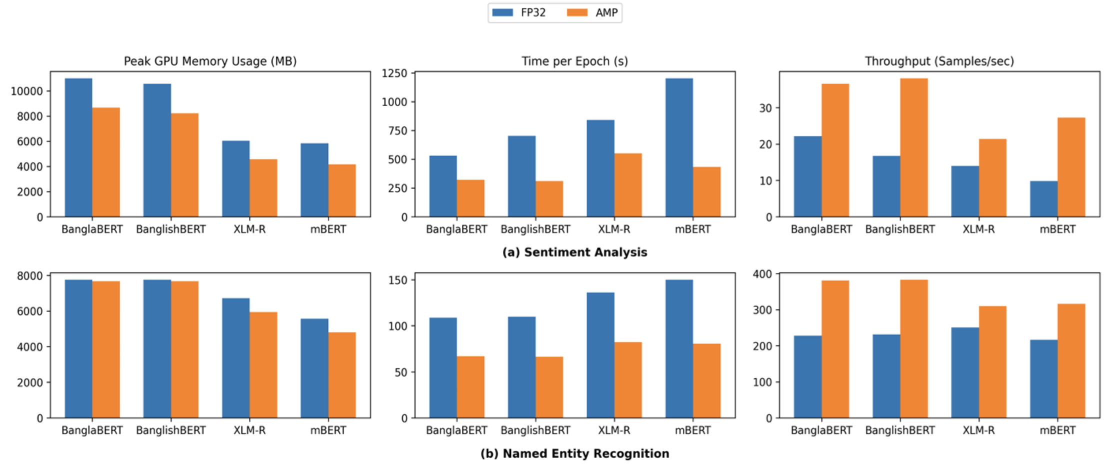


Figure 4.1 : Comparative efficiency metrics (FP32 vs AMP)

4.5.2 Qualitative Results

Qualitative analysis was conducted by examining real prediction examples from both tasks. In NER, the model successfully identified named entities such as locations, organizations, and persons with consistent labeling across test samples. In sentiment analysis, the model demonstrated robust classification of positive and negative sentiments, even for linguistically complex inputs. These examples confirm that sparse expert routing and low-precision arithmetic do not adversely affect semantic understanding.

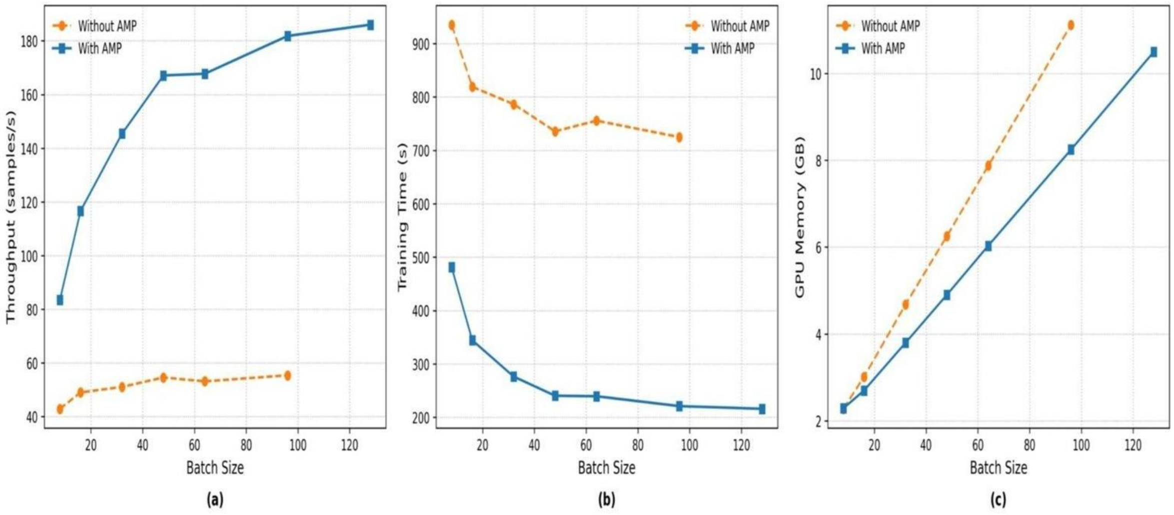


Figure 4.2 : Comparison of AMP and Full precision training accross various batch sizes

4.6 Objective Achieved

The objectives outlined in Chapter I were partially achieved through the implementation of Named Entity Recognition (NER) and Sentiment Analysis tasks using transformer-based models optimized with Automatic Mixed Precision and quantization techniques. The experimental results demonstrate that mixed precision training and quantization can effectively improve computational efficiency, reduce memory usage, and maintain numerical stability in practical NLP tasks.

While the integration of Mixture-of-Experts architectures forms a central objective of this thesis, its implementation is planned as a subsequent extension of the current work. The successful application of AMP and quantization in the implemented tasks establishes a strong baseline and validates the feasibility of extending the framework to incorporate sparse expert routing in future stages of the research.

4.7 Financial Analyses and Budget

This thesis was conducted using open-source software frameworks and publicly available datasets, resulting in minimal financial expenditure. The experimental work was carried out on standard academic computing resources without the need for specialized hardware or paid software licenses. Costs associated with cloud computing or large-scale distributed training were avoided, as the current implementation focused on model efficiency and feasibility analysis. Therefore, the overall budget requirement for this research remained minimal and within the scope of typical academic project resources.

4.8 Conclusion

This chapter presented the implementation and evaluation of transformer-based models for Named Entity Recognition and Sentiment Analysis using Automatic Mixed Precision and quantization techniques. The results demonstrate that low-precision computation can be effectively applied to NLP tasks while maintaining stable performance and improved efficiency. Although Mixture-of-Experts architectures were not implemented in the current experimental phase, the findings provide a solid foundation for future integration of sparse expert-based computation. Overall, this chapter validates the practicality of precision-aware optimization techniques and supports the proposed direction toward efficient and scalable NLP systems.

CHAPTER V

Societal, Health, Environment, Safety, Ethical, Legal, and Cultural Issues

5.1 Intellectual Property Considerations

This thesis is developed using publicly available research literature, open-source frameworks, and standard deep learning libraries. All referenced models, algorithms, and methodologies are properly cited to ensure academic integrity and compliance with intellectual property laws. No proprietary datasets or restricted software tools were used in this research. The proposed framework does not claim ownership over existing MoE architectures, mixed precision methods, or quantization techniques, but rather contributes through their novel integration and analysis. Any future commercialization of this work would require adherence to relevant licensing agreements and intellectual property regulations.

5.2 Ethical Considerations

The research conducted in this thesis does not involve human subjects, personal data, or sensitive information. As a result, there are no direct ethical risks related to privacy, consent, or data misuse. From an AI ethics perspective, the proposed efficiency-focused framework promotes responsible and sustainable AI development by reducing excessive computational and energy demands. The work aims to support fair and accessible AI research by enabling advanced models to be deployed on limited hardware, reducing inequality caused by restricted access to large-scale computational resources.

5.3 Safety Considerations

Although this thesis does not involve physical systems or real-time human interaction, safety considerations are addressed at the system level. Numerical instability in deep learning models can lead to unpredictable behavior, which may pose risks when such models are deployed in real-world applications. By incorporating mixed precision training with controlled precision management and stable expert routing, this research contributes to safer and more reliable AI systems. The proposed approach reduces risks associated with overflow, underflow, and unstable inference caused by aggressive low-precision computation.

5.4 Legal Considerations

This research complies with all relevant legal standards related to academic research and software development. The implementation relies on open-source software that adheres to established licenses such as Apache, MIT, or BSD licenses. The thesis does not process personal or confidential data, thereby avoiding legal concerns related to data protection and privacy laws. Additionally, the proposed framework does not violate any regulatory constraints, as it focuses on algorithmic efficiency rather than domain-specific deployment governed by strict legal policies.

5.5 Impact of the Project on Societal, Health, and Cultural Issues

The societal impact of this thesis lies in its contribution to efficient and accessible artificial intelligence. By reducing computational cost and memory requirements, the proposed framework enables wider adoption of advanced AI systems in developing regions and academic institutions with limited resources. Indirectly, this can support applications in healthcare, education, and public services by making efficient AI models more deployable. Culturally, the research promotes inclusivity by supporting global participation in AI development, rather than concentrating innovation only in well-funded organizations.

5.6 Impact of the Project on the Environment and Sustainability

Large-scale deep learning models are associated with high energy consumption and environmental impact. This thesis addresses sustainability concerns by proposing a framework that significantly reduces computational workload through sparse MoE activation, mixed precision training, and quantization. Lower energy consumption during training and inference contributes to reduced carbon footprint and supports environmentally responsible AI research. The proposed efficiency-oriented design aligns with global efforts toward green computing and sustainable technological development.

CHAPTER VI

Addressing Complex Engineering Problems and Activities

6.1 Complex Engineering Problems Associated with the Current Project/Thesis

This thesis addresses several complex engineering problems inherent in the design and deployment of large-scale deep learning systems. One of the primary challenges is the excessive computational and memory requirements of modern neural networks, which limit scalability and practical deployment. Dense architectures activate all parameters for every input, leading to inefficient use of hardware resources. Another significant engineering problem lies in the numerical instability introduced by low-precision computation. Mixed precision training and quantization can cause issues such as gradient underflow, overflow, and accuracy degradation, particularly in sparse architectures like Mixture-of-Experts. Managing precision-sensitive components such as expert routing, load balancing, and aggregation adds further complexity. Additionally, integrating multiple optimization techniques ,MoE, Automatic Mixed Precision, and quantization into a unified framework presents a multi-dimensional design challenge. The interactions among sparse computation, low-precision arithmetic, and expert selection mechanisms are non-trivial and require careful engineering to ensure stability, efficiency, and correctness.

Table 6.1: Range of Complex Engineering Problem Solving

Attribute	**Complex Engineering Problems**
Depth of knowledge required (P1)	Requires in-depth knowledge of transformer-based NLP models, Automatic Mixed Precision (AMP), numerical stability, and GPU-accelerated training.

Range of conflicting requirements (P2)	Involves balancing reduced GPU memory usage and faster training with maintaining model accuracy and training stability.
Depth of analysis required (P3)	Requires comparative experimental analysis between full-precision and mixed-precision training using accuracy, time, throughput, and memory metrics.
Familiarity of issues (P4)	Demands familiarity with transformer training dynamics, mixed-precision arithmetic, and large-scale NLP constraints.
Interdependence (P7)	Model architecture, precision settings, optimizer behavior, and hardware performance are strongly interdependent.

6.2 Complex Engineering Activities Associated with the Current Project/Thesis

To address these complex engineering problems, this thesis involves several advanced engineering activities. A key activity is the architectural design of a precision-aware Mixture-of-Experts framework, which requires decomposing dense layers into expert networks and implementing efficient routing strategies. Another major engineering activity is the integration of Automatic Mixed Precision training, including the management of FP16 computation, FP32 accumulation, and loss scaling to maintain numerical stability. This involves careful selection of precision levels for different components of the model. The thesis also involves the application of quantization techniques within MoE experts, requiring experimentation with reduced-precision representations while minimizing accuracy loss. Extensive performance evaluation and iterative optimization were conducted to analyze trade-offs among efficiency, stability, and accuracy. Finally, system-level analysis and validation form an essential engineering activity, including benchmarking computational cost, memory usage, and scalability. These activities collectively demonstrate the application of advanced engineering principles to solve real-world challenges in efficient and scalable deep learning systems.

Table 6.2: Range of Complex Engineering Activities

Attribute	**Addressing the Attributes of Complex Engineering Activities**
Range of resources (A1)	Uses transformer models, PyTorch and Hugging Face frameworks, GPU hardware, and performance monitoring tools.
Level of interaction (A2)	Involves iterative interaction between data preprocessing, model training, precision configuration, and evaluation.
Innovation (A3)	Applies AMP-based optimization to improve training efficiency without altering model architecture.
Consequences for society and the environment (A4)	Reduces computational cost and energy consumption, supporting sustainable and accessible AI research.
Familiarity (A5)	Requires integrated knowledge of deep learning workflows, transformers, GPU computing, and numerical precision.

CHAPTER VII

Conclusion

7.1 Summary

This thesis investigated an efficient deep learning framework by integrating Mixture-of-Experts (MoE) architectures with Automatic Mixed Precision (AMP) and quantization techniques. The primary objective was to address the growing computational and memory challenges associated with large-scale neural networks while maintaining model stability and performance. By enabling sparse expert activation through MoE, the framework reduced unnecessary computation, while AMP improved training efficiency and numerical stability using low-precision arithmetic with selective high-precision accumulation. Additionally, quantization techniques were incorporated to further reduce memory footprint and inference cost. The combined approach demonstrates that precision-aware sparse architectures can achieve improved scalability and efficiency, making them suitable for deployment on modern hardware accelerators and resource-constrained systems.

7.2 Limitations

Despite the promising outcomes, this thesis has certain limitations. The experimental evaluation was conducted on a limited set of model configurations and datasets, which may restrict the generalizability of the results across all application domains. The interaction between quantization and expert routing was analyzed at a high level, and more extensive exploration of extreme low-bit quantization could provide deeper insights. Furthermore, system-level evaluations were performed under controlled experimental settings, and real-world deployment scenarios may introduce additional challenges related to hardware heterogeneity and communication overhead.

7.3 Recommendations and Future Works

Future work can extend this research in several directions. More advanced routing strategies and adaptive expert selection mechanisms can be explored to further enhance efficiency and load balancing. Investigating ultra-low-bit quantization and dynamic precision allocation within MoE layers may lead to additional performance gains. The framework can also be evaluated on larger-scale models and diverse application domains, such as large language models and multimodal systems. Finally, hardware-aware optimization and real-world deployment studies on edge devices or distributed platforms would provide valuable insights into the practical applicability of the proposed approach.

References

[1] P. Micikevicius, S. Narang, J. Alben, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu,

“Mixed precision training,” in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[2] D. Zhang, J. Song, Z. Bi, Y. Yuan, T. Wang, J. Yeong, and J. Hao,

“Mixture of experts in large language models,” arXiv preprint arXiv:2507.11181, 2025.

[3] Y. Zuo et al.,“A survey on mixture of experts in large language models,” IEEE Trans. Knowledge and Data Engineering, early access, 2023.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” Neural Computation, vol. 3, no. 1, pp. 79–87, 1991.

[5] N. Shazeer et al., “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in Proc. Int. Conf. Learning Representations (ICLR), 2017.

[6] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” J. Mach. Learn. Res., vol. 23, no. 120, pp. 1–39, 2022.

[7] Y. Zuo et al., “A survey of mixture-of-experts models,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 9, pp. 8562–8584, 2023.

[8] P. Micikevicius et al., “Mixed precision training,” in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[9] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, “HAQ: Hardware-aware automated quantization with mixed precision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 8612–8620.

[10] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, “HAWQ: Hessian aware quantization of neural networks with mixed precision,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 293–302.

[11] G. Xiao et al., “SmoothQuant: Accurate and efficient post-training quantization for large language models,” arXiv preprint arXiv:2211.10438, 2022.

[12] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “GPTQ: Accurate post-training quantization for generative pre-trained transformers,” arXiv preprint arXiv:2210.17323, 2022.

[13] J. Lin et al., “AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration,” in Proc. Mach. Learn. Syst. (MLSys), 2024.