NORTH SOUTH UNIVERSITY

SPRING 2026

DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEER

PROJECT PROPOSAL - GROUP: 03

Course Code: CSE 299
Course Title: Junior Design Course
Section: 04
Group Number: 03

Date of Submission: 09.02.2026
Submitted by Group Number: 03 [Three]
Group Members:

Name	ID
Md. Adnan Abdullah Sadi	2221345642
Makki Ammer Sakib	2231687642
Nahian Syed Ahanaf	2212705042

Course Instructor:

Dr. Mohammad Shifat-E-Rabbi [MSRb]
Assistant Professor
Department of Electrical & Computer Engineering
North South University

Evaluating Small Language Models as Resource-Efficient Alternatives to Large Language Models

Research Proposal

9th February 2026

1 Problem Statement

Recent Large Language Models (LLMs), including GPT-4–class systems and frontier proprietary models, require large-scale GPU infrastructure, high memory bandwidth, and sustained energy consumption, which limits accessibility for individual researchers, academic labs, and small organizations. [1]. Beyond computational cost, recent studies and industry reports highlight increasing privacy and data governance risks associated with cloud-hosted LLM deployment, particularly for sensitive or regulated data, creating vulnerabilities for exploitation and non-compliance with regulations such as GDPR and HIPAA [4]. This centralization of AI capabilities limits accessibility and raises sustainability concerns.

Small Language Models (SLMs) with billions rather than trillions of parameters offer a potential solution running on consumer-grade hardware while maintaining data privacy through local deployment. Recent advances in fine-tuning techniques (LoRA, QLoRA) and model optimization have significantly improved SLM performance [3]. However, it remains unclear whether current SLMs can deliver performance comparable to that of LLMs for practical applications. This research addresses the fundamental question: ***Can existing SLMs, enhanced through modern fine-tuning techniques, provide adequate performance for real-world tasks while operating on accessible hardware and ensuring complete data privacy?***

2 Background and Significance

Over the past decade, transformer-based architectures have enabled rapid scaling of language models from hundreds of millions to hundreds of billions of parameters, culminating in today's frontier LLMs.While such models achieve strong benchmark performance, recent analyses emphasize their high inference cost, environmental footprint, and deployment constraints, motivating research into more efficient alternatives [2].

Small Language Models (SLMs, 1B–13B parameters) offer a practical alternative. With techniques like LoRA and QLoRA, SLMs can run on consumer-grade hardware while maintaining competitive performance on selected benchmarks [3].

However, SLMs remain underexplored compared to LLMs, especially for complex reasoning and real-world deployment. Evaluating their performance and limitations is essential to guide resource-efficient, privacy-preserving adoption.

This research is significant for: (1) democratizing AI by enabling deployment on accessible hardware, (2) protecting privacy through on-device processing, (3) reducing environmental impact, (4) enabling offline applications, and (5) establishing fundamental understanding of sufficient model capacity for various tasks.

3 Research Objectives

Primary Objective: Evaluate whether current SLMs with modern fine-tuning can match LLM performance for common use cases reasoning while operating on consumer hardware with complete data privacy.

Specific Objectives:

1. Benchmark SLMs (1B-13B parameters) across standardized reasoing tasks
2. Compare fine-tuning methods: LoRA, QLoRA, and full fine-tuning Mathemetical and Logical reasoing
3. Quantify resource requirements: GPU memory, inference latency, energy consumption, and throughput on consumer hardware
4. Conduct head-to-head comparisons with commercial LLMs (GPT-4, Claude) to identify performance gaps
5. Develop evidence-based guidelines for selecting and deploying SLMs for specific applications

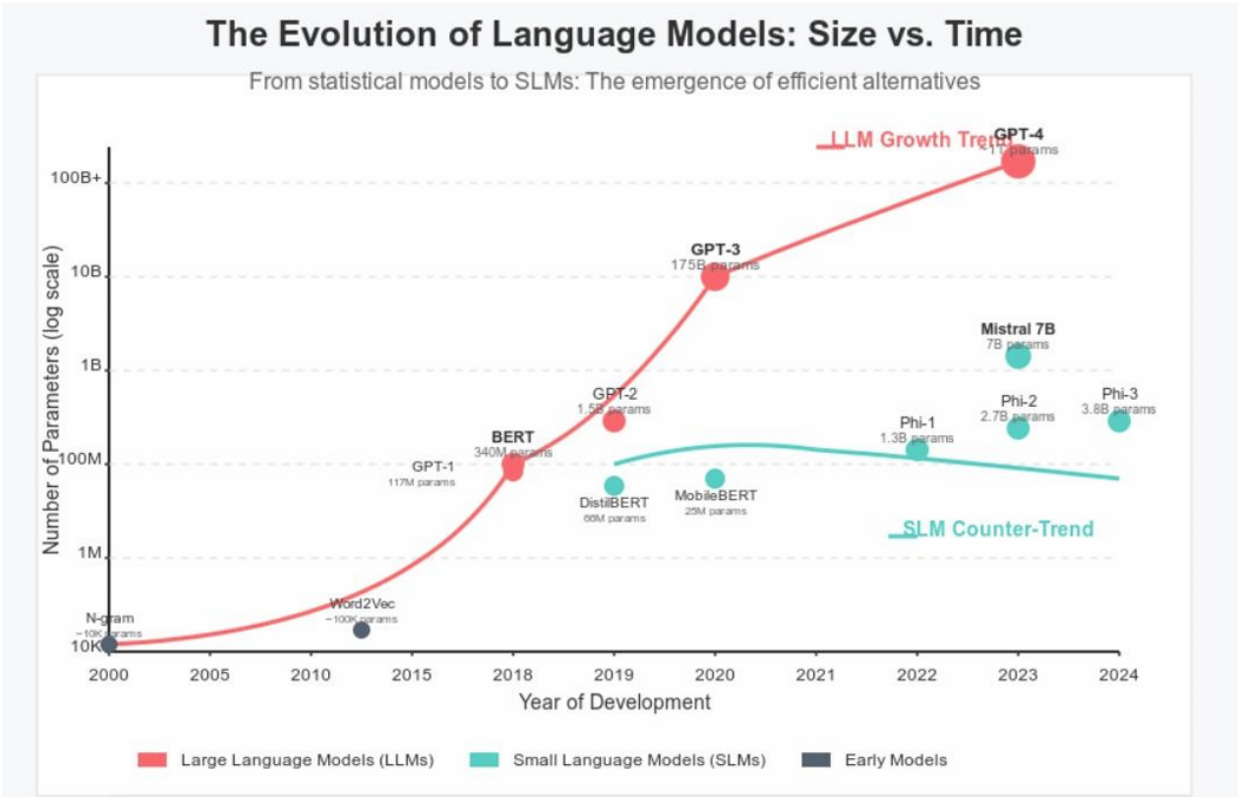

Figure 1: Evolution of language models from early statistical methods to Large Language Models (LLMs) and the emergence of Small Language Models (SLMs). Parameter counts are shown on a logarithmic scale and are based on publicly reported estimates.

Future Goals: Building upon our findings, we aim to develop optimized SLM architectures specifically designed for mathematical and logical reasoning tasks, potentially incorporating retrieval-augmented generation and chain-of-thought prompting techniques. Furthermore, we plan to explore federated fine-tuning approaches that enable collaborative model improvement across multiple institutions while maintaining strict data privacy, establishing a framework for privacy-preserving collective intelligence in specialized reasoning domains.

4 Current Statistics

Market Context: The global Natural Language Processing (NLP) market is projected to reach approximately $127 billion by 2028, reflecting rapid enterprise adoption of AI-driven language technologies. ChatGPT alone serves over 100 million monthly users, while enterprise LLM adoption increased by 270% between 2022 and 2024. However, operational costs remain a barrier, with GPT-4 API pricing ranging from $0.03 to $0.12 per 1K tokens, creating significant long-term expenses for large-scale deployments.

Computational Requirements: Large Language Models demand substantial computational infrastructure. GPT-4–scale systems require clusters of high-memory GPUs (e.g., 8× NVIDIA A100 80GB), whereas Small Language Models such as LLaMA 2 13B can operate on a single consumer GPU like the RTX 4090 (24GB). Training large models can emit up to 626,000 pounds of CO_2, and AI-related computing energy consumption is estimated to double every 3.4 months, raising sustainability concerns.

Privacy Concerns: Privacy risks further complicate LLM adoption. Approximately 38% of enterprises report concerns about sensitive data exposure when using cloud-based AI systems. The average cost of a data breach reached $4.45 million per incident. Consequently, organizations such as Samsung and several financial institutions have restricted or banned the use of public LLM platforms following internal data leakage incidents.

SLM Adoption: In response, interest in deployable, privacy-preserving models is rising. Open-

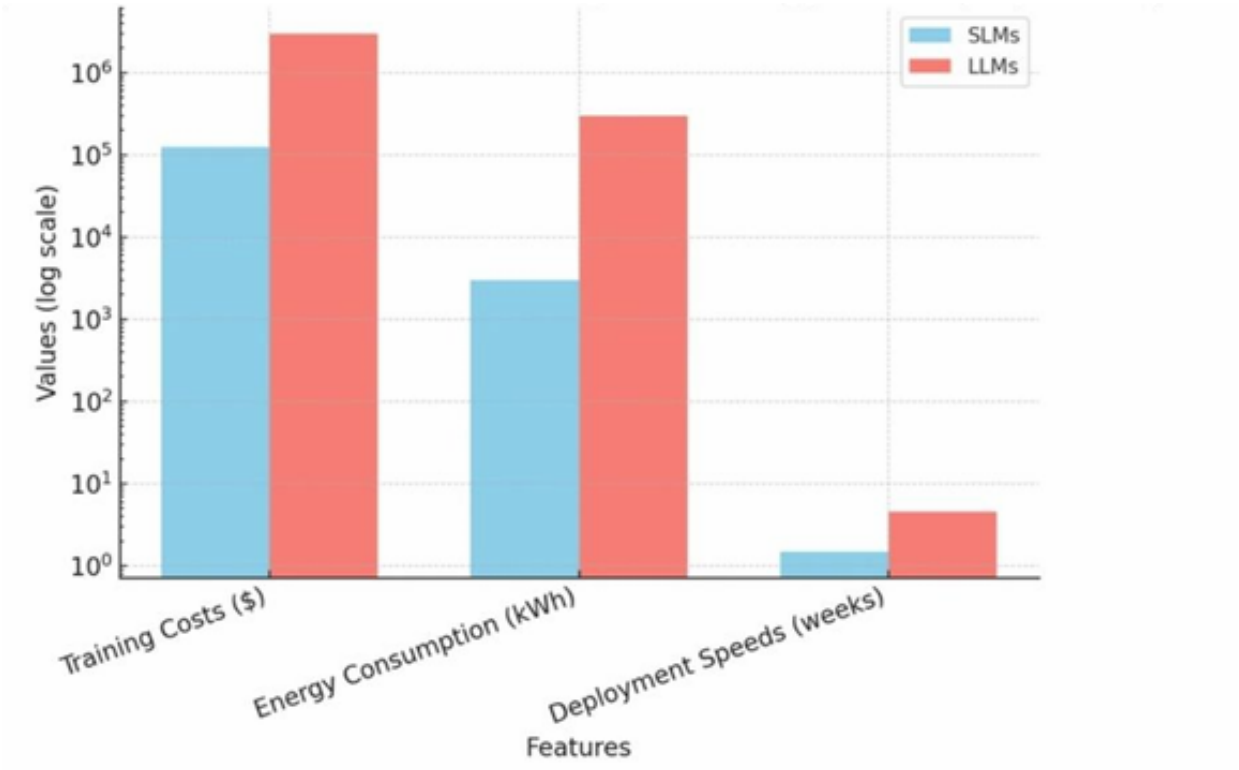

Figure 2: Comparison of Small Language Models (SLMs) and Large Language Models (LLMs) in terms of training cost, energy consumption, and deployment time. Values are presented on a logarithmic scale, highlighting the substantial resource efficiency advantages of SLMs.

source SLM downloads increased by 540% in 2023 alone. The Edge AI market is projected to reach $59.6 billion by 2030, with 45% of enterprises actively exploring on-premise or local model deployment strategies to ensure data sovereignty and regulatory compliance.

5 Literature Review

Recent large language models (LLMs) have demonstrated strong reasoning performance on benchmarks such as MMLU, GSM8K, MATH, and HumanEval [6]. However, these improvements are typically achieved using models with hundreds of billions of parameters, leading to high computational costs and significant deployment barriers, particularly for edge and on-device applications [6].

Small language models (SLMs), typically under 2B parameters, have emerged as efficient alternatives. The InfiR framework [6] introduces a 1B-parameter model trained on curated datasets including Infinity-Instruct and ScaleQuest-Math, achieving 69.45

Reflection-based optimization further enhances compact models. ReflectEvo [5] constructs the ReflectEvo-460k dataset to enable iterative self-reflection and self-correction, improving aggregated reasoning performance across multiple benchmarks by boosting Llama-3 from 52.4

Despite these advances, systematic evaluations that directly compare optimized SLMs with large-scale LLMs across diverse real-world reasoning tasks remain limited, motivating further investigation into parameter-efficient reasoning models.

6 Tools and Resources

Computational Resources: All experiments will be conducted using open-source frameworks, primarily **PyTorch** and **Hugging Face** Transformers with PEFT for efficient fine-tuning. Training and evaluation will be performed on freely available cloud platforms such as **Google Colab** and **Kaggle** to reflect realistic resource constraints. The study will utilize exclusively public and open-access datasets for instruction following, reasoning, and domain-specific evaluation. Experiments will focus on open-weight Small Language Models available through the Hugging Face ecosystem, while proprietary large language models will be used solely as performance reference baselines.

References

[1] Zhou, Y., et al. (2024). A survey on efficient large and small language models. *arXiv preprint arXiv:2403.01234*.

[2] Chen, L., et al. (2024). Towards resource-efficient foundation models. *ACM Computing Surveys*.

[3] Deng, Y., Zhang, A., Wang, N., Gurses, S., Yang, Z., and Yin, P. (2025). *CLoQ: Enhancing Fine-Tuning of Quantized LLMs via Calibrated LoRA Initialization*. arXiv preprint arXiv:2501.18475. :contentReference[oaicite:5]index=5

[4] Xu, R., et al. (2025). Small language models: Capabilities, limitations, and opportunities. *arXiv preprint*.

[5] Li, J., Dong, X., Liu, Y., Yang, Z., Wang, Q., Wang, X., Zhu, S., Jia, Z., Zheng, Z. (2025). *ReflectEvo: Improving Meta Introspection of Small LLMs by Learning Self-Reflection*. arXiv preprint arXiv:2505.16475.

[6] Xie, C., Cai, S., Wang, W., Li, P., Sang, Z., Yang, K., Zhang, Y., Li, Z., Zhu, G., Liu, Z., et al. (2025). *InfiR: Crafting Effective Small Language Models and Multimodal Small Language Models in Reasoning*. arXiv preprint arXiv:2502.11573.