Hyper-Efficient On-Device Small Language Models for Structured Agentic Workflows

Pablo Leyva
Undergraduate Researcher – Applied Statistics and Data Analytics, NJIT
Prev AI/ML Intern at Apple
Advisor: Dr. Ding

Phase-2 URI Student Seed Grant Proposal

Abstract

Recent advances in language model architecture and quantization have enabled unprecedented efficiency in model deployment without sacrificing quality. This project proposes the design and evaluation of an on-device small language model (SLM) optimized for agentic business process execution, such as customer onboarding, document processing, and procedural task automation. The approach synthesizes five key innovations: (1) Apple Foundation Model (AFM)-style efficient backbones, (2) BitNet b1.58 ternary weight quantization to minimize memory and computation, (3) super weight identification and preservation to maintain critical model behavior, (4) Evolution-Strategy (ES)-based fine-tuning to achieve stability without reinforcement learning overhead, and (5) recursive reasoning with compact architectures inspired by TinyNet paradigms. Together, these methods enable specialized on-device models to handle deterministic workflows.

The research will design, implement, and benchmark such a model (\~3B parameters) to demonstrate that hyper-efficient SLMs can reliably execute long chained tasks on edge devices, bridging the gap between research LLMs and practical, economical deployment.

A. Problem Statement

Large Language Models (LLMs) have achieved impressive reasoning and language capabilities but introduced severe technical, environmental, and economic challenges. The surge in AI chatbots like ChatGPT has driven a global spike in data center electricity use—about **415 TWh in 2024**, projected to exceed **945 TWh by 2030**—as GPT-scale training consumes gigawatt-hours and each query far surpasses the energy of a standard web search. Expanding AI infrastructure now demands **100 MW or more** per facility—comparable to small power plants—placing heavy strain on regional grids, increasing fossil-fuel reliance, and escalating water use for cooling, intensifying calls for sustainable AI solutions to reduce carbon and infrastructure impact.

Widespread adoption of cloud-based chatbots has intensified privacy concerns, as tools like ChatGPT routinely collect user data—ranging from names, emails, and uploaded files to IP addresses and device identifiers—that may be retained for training unless chat history is disabled or enterprise privacy controls apply. Stanford researchers found that providers such as OpenAI, Google, and Microsoft aggregate chatbot inputs into broader data ecosystems, enabling long-term profiling and inference of sensitive details like health or financial status. Moreover, OpenAI’s policies permit data sharing with affiliates and contractors, while legal warrants can compel disclosure, as shown in the 2025 DHS case. For organizations, these practices pose compliance risks, as confidential data used during chatbot interactions could reenter training datasets, heightening exposure to breaches and regulatory violations. Together, these issues highlight the urgent need for privacy-preserving, on-device AI inference solutions.

Fine-tuning large foundational models for domain-specific use is still costly and time-intensive. Expenses vary by model size, dataset complexity, and tuning method—full fine-tuning can reach tens of thousands of dollars and take weeks, while LoRA or QLoRA lower costs but still require specialized compute and expertise. Additional time and money go into data preparation, evaluation, and maintenance, creating major economic and technical barriers that limit small and medium-sized enterprises from adopting or innovating with tailored AI models.

Together, these challenges—energy inefficiency, environmental strain, high operational cost, fine-tuning expense, and data vulnerability—expose critical weaknesses in today's centralized AI model paradigm. Addressing them requires new approaches that minimize resource use, reduce adaptation costs, enhance data privacy, and enable AI operation directly on edge devices. Addressing them requires new approaches that minimize resource use, reduce adaptation costs, enhance data privacy, and enable AI operation directly on edge devices.

B. Significance

The significance of this research lies in its potential to redefine how AI is built and deployed—transforming it from an energy-intensive, privacy-compromised, and costly system into a sustainable, secure, and accessible tool for innovation. This shift is critical as small and medium-sized businesses (SMBs) enter a new growth phase driven by LLM-based autonomous agents. The rise of affordable models, no-code platforms, and modular AI ecosystems is creating a multi-billion-dollar opportunity for workflow automation, where responsible adoption will determine the next wave of productivity and competitiveness.

The environmental impact of AI is increasingly urgent: data centers already strain global power grids, and unchecked expansion risks industrial-scale energy use. Developing hyper-efficient Small Language Models (SLMs) directly addresses this issue, achieving high performance with drastically lower energy demands. The privacy challenge is equally pressing—most SMBs lack enterprise-grade infrastructure, while cloud chatbots collect sensitive inputs, metadata, and histories often retained or shared for training.

Finally, economic accessibility ensures equitable participation in this AI revolution. Fine-tuning large models remains costly and time-consuming, even with methods like LoRA that require specialized compute and expertise. By enabling lightweight, adapter-based customization, this research makes fine-tuning hundreds of times cheaper and faster, empowering SMBs to adopt tailored AI tools without prohibitive overhead.

In sum, sustainable on-device AI democratizes access to advanced intelligence, allowing smaller firms to compete with large enterprises while upholding privacy, affordability, and environmental responsibility—setting a precedent for inclusive, responsible technological progress.

C. Innovation

This research **bridges the cost–capability gap** between large cloud LLMs and the practical deployment needs of small businesses and edge devices. Fine-tuning foundational models for specific domains remains prohibitively expensive and time-intensive, limiting access for organizations without large-scale compute resources. This project advances innovation by developing **scalable, energy-efficient Small Language Models (SLMs)** that reduce adaptation cost and latency while maintaining competitive task performance in agentic business environments.

Technically, the project will implement a **modular SLM architecture (3–8 B parameters)** optimized for Apple Silicon via MLX, Apple's on-device machine-learning framework. The system will integrate **four innovations not previously unified in undergraduate research:**

1. **BitNet-style ternary quantization (b1.58)** for aggressive weight compression;
2. **Super-weight preservation**, a re-centering function that stabilizes accuracy under ternarization;
3. An **Apple Foundation Model-inspired memory layout** to enhance locality and inference throughput; and
4. **Evolution Strategies (ES)** for gradient-free fine-tuning, minimizing KL-divergence between base and adapted model distributions without reinforcement learning.

Evaluation Framework: Unlike traditional benchmarks that test static reasoning, this project adopts a **realistic, multi-step agentic evaluation paradigm** inspired by GTA: A Benchmark for General Tool Agents (2024). The resulting SLMs will be evaluated on business workflow simulations that emulate small-enterprise processes — such as invoice reconciliation, inventory management, expense tracking, and customer-response automation — each modeled as a sequential reasoning task involving planning, tool selection, and action execution.

Performance will be measured across both step-level and end-to-end metrics, including:

1. **Tool-selection accuracy** (correctly identifying and invoking tools for each subtask),
2. **Parameter correctness** (accuracy of arguments passed to each function),
3. **Chain-of-execution fidelity** (logical coherence across tool calls),
4. **Final outcome accuracy** (task success rate),
5. **Inference latency and throughput**, and
6. **Energy and adaptation cost per workflow.**

Building on insights from **Jolicoeur-Martineau (2025)**, the project will also explore **recursive reasoning modules** to enhance interpretability and compositional generalization in compact networks. Collectively, these methods establish a **sustainable, privacy-preserving, and explainable SLM framework** capable of executing real-world agentic business workflows on consumer-grade devices — a foundational step toward democratizing intelligent automation for small enterprises.

D. Research Literature Review

Existing literature highlights critical progress in efficient model architectures and quantization. Apple (2024) introduced AFM, demonstrating competitive performance with minimal compute. Ma et al. (2024) proposed BitNet b1.58, showing 1.58-bit precision viability for large models. Yu et al. (2025) identified the role of super weights in model integrity. Belcak et al. (2025) argued that small, specialized models represent the next wave of agentic AI. **Qiu et al. (2025)** extended fine-tuning methods through Evolution Strategies, offering scalable optimization without reinforcement learning. **Jolicoeur-Martineau (2025)** demonstrated that recursive reasoning with compact networks enhances reasoning efficiency and interpretability, reinforcing the feasibility of high-performing SLMs. This proposal builds directly upon these findings, integrating them into a unified, on-device deployment pipeline optimized for privacy, interpretability, and environmental efficiency.

E. Unmet Needs

There is a growing unmet need for **Small Language Models (SLMs)** that empower small-to-medium-sized business owners (SMBs) and independent developers to integrate intelligent, private, and affordable AI solutions into their operations. Current AI systems—large, cloud-hosted LLMs—are prohibitively expensive to fine-tune, demand high energy consumption, and risk exposing sensitive user data through centralized infrastructure. SLMs address these challenges by providing **on-device, privacy-preserving, and energy-efficient intelligence** that can run locally without external servers. They offer a path to sustainable, domain-specialized automation for smaller teams who lack access to massive compute resources. By combining **compute efficiency**, **inference efficiency**, **privacy**, and **LoRA-based modular adaptability** within the Apple MLX hardware ecosystem, this research directly targets these practical and market-driven gaps, ensuring that AI becomes an accessible and trustworthy tool for real-world business workflows.

F. Phase 2-Specific Goals

1. **Develop and benchmark a 2–4B parameter SLM** optimized for Apple MLX with int4 quantization and dynamic inference scaling.
2. **Design and deploy modular LoRA adapters** fine-tuned for five SMB sectors: Restaurants, HVAC/Service, Retail, Legal, and Accounting.
3. **Integrate agentic workflow automation** — connecting the SLM to email, scheduling, CRM, and document pipelines for small-business operations.
4. **Establish an evaluation framework** measuring performance on latency, accuracy, and energy consumption across Apple Silicon devices.

Examples of LoRA-based Agentic Use Cases

Industry	Agentic Workflow Example	Function
Restaurants	Summarize Yelp reviews and generate responses	Customer engagement
HVAC / Services	Generate service quotes and schedule follow-ups	Workflow automation
Retail	Write product listings and manage customer replies	Marketing & communication
Legal	Summarize NDAs and flag key clauses	Document intelligence
Accounting	Parse invoices and detect anomalies	Finance optimization

G. Methodology / Research Design

Transition:

During Phase-1, the focus is on constructing the base SLM pipeline integrating AFM-style architecture, BitNet b1.58 ternary quantization, and super-weight preservation. In the process, quantization will be assessed to push the bounds of the number of parameters and on-device memory used.

Phase-2 builds directly upon this foundation by transitioning from **infrastructure development** to **domain specialization and deployment optimization**. The emphasis shifts from proving quantization viability to demonstrating *applied performance* across diverse business environments and hardware configurations. This phase aims to validate both technical scalability and societal value—showing that small, transparent models can meet the real operational needs of independent developers and small enterprises.

Process:

- **Model Selection & Distillation**: Start with a 3B parameter open model (e.g., LLaMA 3B or AFM) and distill from a larger model on procedural task datasets.
- **BitNet b1.58 Quantization**: Apply absmean ternary quantization to all weights and measure perplexity before/after.
- **Super Weight Detection**: Identify key super weights for fp16 preservation.
- **Evolution Strategy Fine-Tuning**: Implement gradient-free optimization for LoRA adapters using Qiu et al. (2025) methods.
- **Adapter Integration**: Train LoRA adapters for specific workflows.
- **Agentic Reliability Harness**: Implement self-auditing checkpoints and benchmark task completion rates.
- **Deployment & Benchmarking**: Test on Mac and edge devices for latency, throughput, memory, and energy efficiency.

H. Expected Outcomes

Group A — Core SLM Capabilities

- **A1:** Build and fine-tune the SLM using small-business communication datasets.
- **A2:** Integrate Qdrant/MLX vector memory for local retrieval.
- **A3:** Implement compute-adaptive inference via MLX runtime quantization.

Group B — Agentic Workflow Integrations

- Build micro-agents for:
- Restaurant customer engagement (Yelp review summarization and response generation)
- HVAC service workflow automation (quote estimation and follow-up scheduling)
- Retail marketing & communication (product listing creation and customer reply management)
- Evaluate latency and memory footprint on MacBook M4 hardware.

Group C — LoRA Adapter Expansion

- Develop 5 industry adapters (<50 MB each).
- Benchmark cross-adapter accuracy, context retention, and energy draw.

Group D — Evaluation Metrics

Metric	Target	Tool
Latency	↓ ≥ 35%	MLX Profiler
Accuracy (NER/Summarization)	±3% baseline	Custom test suite
Privacy compliance	100% local	Secure enclave validation

Expected Deliverables

- **SLM v2 prototype (4-bit MLX-optimized)**
- **Five LoRA adapters** for SMB industries
- **Open SDK + deployment dashboard**
- **Technical whitepaper** quantifying compute/inference gains

---

I. Resources and Budget Justification (Limit: \$3,000)

Item	Description / Purpose	Estimated Cost (USD)
Apple MacBook Pro 16″ (M4, 24 GB RAM)	Primary research device for SLM optimization, LoRA adapter training, and benchmarking across Apple Silicon hardware using the **MLX framework**. Enables local testing of compute-sharing between devices (Mac, iPhone, iPad).	\$2,700
External SSD (2 TB)	For model checkpoints, adapter weights, and dataset storage during iterative fine-tuning.	\$200
Miscellaneous Research Supplies	Includes cables, thermal monitors, and energy-usage tracking tools to measure inference efficiency.	\$100
Total		**\$3,000**

Justification:
The proposed MacBook Pro M4 serves as the cornerstone for this phase's experimental environment. The **MLX ecosystem** offers low-level access to Apple's on-device tensor cores, allowing the project to explore **computational sharing** across multiple Apple Silicon devices for distributed, energy-efficient inference. This hardware investment directly supports the research objectives of maximizing compute efficiency, privacy, and sustainability. The external SSD ensures reproducible experimentation and secure, local storage of sensitive model data. Together, these resources enable completion of the Phase-2 milestones and set the foundation for further multi-device SLM research.

References

1. Apple. (2024). *Apple Intelligence Foundation Language Model*. arxiv.org/2407.21075
2. Yu, M. et al. (2025). *The Super Weight in Large Language Models*. arxiv.org/2411.07191
3. Ma, S. et al. (2024). *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits*. arxiv.org/2402.17764
4. Belcak, P. et al. (2025). *Small Language Models are the Future of Agentic AI*. arxiv.org/2506.02153
5. Qiu, X. et al. (2025). *Evolution Strategies at Scale: LLM Fine-Tuning Beyond Reinforcement Learning*. arxiv.org/2509.24372
6. Jolicoeur-Martineau, A. (2025). *Less is More: Recursive Reasoning with Tiny Networks*. arxiv.org/2510.04871
7. International Energy Agency. (2024). *Global Data Centre Electricity Use*. iee.psu.edu
8. Goldman Sachs. (2025). *AI to Drive 165% Increase in Data Center Power Demand by 2030*. goldmansachs.com
9. Deloitte. (2025). *Data Center Infrastructure and Artificial Intelligence*. deloitte.com
10. MIT News. (2025). *Explained: Generative AI's Environmental Impact*. mit.edu
11. Nature. (2025). *AI Energy Usage and Climate Footprint*. nature.com
12. Wired. (2025). *New Research on AI Electricity and Energy*. wired.com
13. Carbon Brief. (2025). *AI and Data Centre Energy Use in Context*. carbonbrief.org
14. Private Internet Access. (2025). *ChatGPT Privacy Concerns*. privateinternetaccess.com
15. Just AI News. (2025). *Does ChatGPT Store Your Data?* justainews.com
16. Stanford University. (2025). *AI Chatbot Privacy Concerns and Risks Research*. stanford.edu
17. Forbes. (2025). *OpenAI Ordered to Unmask Writer of Prompts*. forbes.com
18. TechInformed. (2025). *Study: Data Risks Using Leading AI Chatbots*. techinformed.com