Akhil Theerthala

Bengaluru | akhiltvsn@gmail.com | 8179528501 | LinkedIn | Portfolio | GitHub | Hugging Face

Summary

Applied ML researcher and data scientist with 3+ years of experience building production NLP, VLM, and document-intelligence systems for financial domains. My research work spans data-centric language modeling, synthetic benchmark generation, financial reasoning, mechanistic interpretability, and evaluation pipelines, with accepted/submitted workshop publications and open-source Hugging Face artifacts reaching 56k+ downloads. Strong track record translating research prototypes into deployed systems, including latency reduction of 97.5%, accuracy gains of 27.6%, and domain-adapted multimodal models for financial document understanding.

Professional Experience

Senior Member Data Scientist | *Perfios Software Solutions* Apr 2025 – Present

- Awarded **Circle of Excellence** for research and contributions to Generative AI in financial document intelligence.
- Adapted **PaliGemma2** via LoRA on domain-specific financial data, achieving a **TEDS score of 0.85** (Tree-Edit-Distance Similarity) on internal document benchmarks.
- Developed reusable agentic workflow and evaluation frameworks for financial-service tasks including Lending-STP, RM Copilot, fact verification, market intelligence, and demand assessment; evaluated task accuracy, reliability, failure modes, and inference-time scaling strategies across internal baselines and local model deployments.
- Developed a reference-free visual quality estimation model using ViT regression to predict document legibility, filtering low-fidelity inputs at **92%** precision and reducing hallucination-prone downstream inference.

Member Data Scientist | *Perfios Software Solutions* Jun 2023 – Apr 2025

- Distilled and quantized a multimodal document classifier into a lightweight student model, reducing latency from **8s** to **200ms** while preserving F1 parity in production-scale evaluation.
- Improved generalized table detection accuracy by **27.6%**, lifting downstream TSR module performance, via semi-synthetic data curation and systematic evaluation of YOLOv8 variants.
- Integrated a semantic row-detection module (fine-tuned text encoders) into the production TSR pipeline with **<40 ms** added turnaround time, preserving existing throughput.

Research Experience

Research Volunteer | *FSIL & HCAI Labs, Georgia Tech* Aug 2025 – Present
(External Collaborator)

- Co-authored **FinForge**, a semi-synthetic benchmark generation pipeline for financial tasks — accepted at the **AAAI 2026 Workshop on Agentic AI in Financial Services**.
- Contributed to **Stable Steering in Activation Space for LLMs** project, submitted to the Mechanistic Interpretability Workshop at ICML 2026.

Open Source Contributor | *Hugging Science (AI for Food Allergies)* Oct 2025 – Nov 2025

- Curated high-quality public datasets for food allergy detection to support AI safety in health domains and engineered an interactive **Dataset Explorer**, enabling visualization and analysis of allergy data distributions.

Selected Projects & Open-Source Work

Reasoning Dataset Creation Challenge Winner – 1st Place (Global)

- Won 1st place globally out of 150+ teams in the Reasoning Dataset Creation Challenge hosted by Bespoke Labs, Hugging Face, and Together.ai.
- Built synthetic data pipelines using real-world personal finance queries to create instruction-tuning and reasoning datasets tailored to Indian financial contexts.
- Fine-tuned 8B/14B models and released datasets, LoRA adapters, and quantized weights on Hugging Face, reaching 56k+ downloads across direct and community GGUF adaptations.
- Demonstrated that high-quality domain-specific synthetic data can enable smaller 7B models to approach larger 14B/24B baselines on finance reasoning tasks.

Density vs. Diversity – Data Curation Strategy Validation for VLMs Hugging Face Article

- Investigated trade-offs between dense and diverse sampling strategies by curating 15k-sample synthetic datasets for Vision-Language Models (VLMs).
- Conducted comparative analysis of 4B and 8B models across in-domain and RealWorldQA benchmarks to assess reasoning and OOD generalization.

LazyInfer: A multi-stage LLM orchestration framework using YAML files. GitHub

- Built LazyInfer, a Python-based framework for configurable multi-stage LLM inference pipelines over JSONL datasets, enabling reusable and scalable QA/data generation workflows.
- Enabled production-style dataset processing with checkpoint recovery, YAML-driven pipeline configuration, structured output validation, and optional Hugging Face dataset publishing.

Causal Analysis of Social Media Signals on Crowdfunding Success GitHub

- Designed a multi-stage experimental framework to decouple the impact of social engagement metrics from fundamental content features on funding rates.
- Conducted stepwise ablation studies to quantify the marginal lift of social signals, demonstrating that content features drive success independently of engagement metrics

Publications

G. Matlin et al. (2026). **Stable Steering in Activation Space for LLMs**
Submission to Workshop on Mechanistic Interpretability, ICML 2026.

G. Matlin et al. (2026). **FinForge: A Semi-Synthetic Benchmark Generation Framework for Finance.**
AAAI 2026 Workshop on Agentic AI in Financial Services.

A. Theerthala (2025). **A Data-Centric Framework for Training Behaviour-Aware Personal Finance LMs.**
FinNLP, EMNLP 2025.

Education

MS, Data Science Apr 2026 – Expected 2028
International Institute of Information Technology, Hyderabad

B. Tech, Aerospace Engineering Aug 2019 – May 2023
Indian Institute of Technology, Kharagpur

Skills

Core ML: Python, SQL, PyTorch, Scikit-learn, NumPy, Pandas
LLMs & Fine-Tuning: Hugging Face Transformers, TRL, PEFT/LoRA, vLLM, instruction tuning, structured generation, RAG
Agentic Systems & Evaluation: Smolagents, Google Agent Development Kit, LangChain, LangGraph, tool-use agents, multi-agent workflows, inference-time scaling, task-specific evaluation, failure-mode analysis
Data-Centric ML: synthetic data generation, benchmark construction, dataset curation, ablation studies, model comparison, evaluation design and benchmarking
Multimodal & Document AI: vision-language models, document classification, table structure recognition, visual quality estimation, OCR-adjacent pipelines
Optimization & Engineering: distillation, quantization, async inference pipelines, Docker, Git, Linux

Selected Course Certifications

- **Agentic AI** (*Deeplearning.AI*)
- **The Reasoning Course** (*Hugging Face*)
- **Generative AI Nanodegree** (*Udacity*)
- **Machine Learning in Production** (*Coursera*)
- **Deep Learning Specialization** (*DeepLearning.AI*)