Jatin Avinash Salve

Machine Learning Engineer · Computer Vision & Visual Search · Generative AI · Responsible AI

+1 (352)-757-9671 | jatin.salve@ufl.edu | linkedin.com/in/jatin-salve | jatins-dev.github.io

SUMMARY

M.S. Computer Science (ML Track) student at the University of Florida (GPA 3.89/4.00) with hands-on production ML experience spanning computer vision, visual retrieval, generative AI, and responsible AI. Published at ACL 2024; second paper submitted to ACL 2026. Built and shipped full-stack ML systems serving 1M+ daily requests. Available Sep 21 – Dec 11, 2026 for Pinterest's Visual Search internship.

EDUCATION

University of Florida — M.S. Computer Science, Machine Learning Track Aug 2025 – May 2027

GPA: 3.89 / 4.00

Coursework: Machine Learning · Deep Learning · NLP · Large Language Models · Computer Vision · Distributed ML · Statistical Modeling · Agentic AI Systems

TECHNICAL SKILLS

- **Languages:** Python, Java, C++, CUDA, SQL, Go, TypeScript
- **ML / Deep Learning:** PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, XGBoost, MLflow; supervised & unsupervised learning, clustering, anomaly detection, time-series modeling
- **Computer Vision & Visual Search:** Image feature extraction, dense retrieval, transformer-based visual encoders, multimodal embeddings, ranking & reranking models, FAISS / ANN indexing
- **Generative AI & LLMs:** LLM fine-tuning (LoRA, QLoRA, RLHF), vLLM, RAG pipelines, multimodal generation, prompt engineering, evaluation pipelines, responsible AI / faithfulness verification
- **Agentic AI & Orchestration:** Multi-agent systems, LangGraph, tool-using agents, autonomous task decomposition, agent evaluation harnesses, structured outputs
- **ML Systems & Infra:** Ray Serve, FastAPI, Docker, Kubernetes, AWS (ECS, S3, Lambda), distributed training, CI/CD, A/B testing, GPU optimisation (batching, parallelism)

EXPERIENCE

Research Assistant — Computer Vision, Multimodal AI & Responsible AI Dec 2025 – Present

University of Florida · Advisor: Prof. Yonghui Wu Gainesville, FL

- Designed and evaluated multimodal retrieval pipelines combining visual and textual signals; improved NDCG@10 by 21% throughdense transformer embeddings, FAISS vector search, and learned reranking — directly applicable to Pinterest's Visual Search stack.
- Built claim-level verification and context-grounding strategies for generative AI outputs, reducing hallucination rate by 14%;developed responsible-AI evaluation harnesses benchmarking faithfulness, accuracy, and safety across model variants.
- Developed supervised and unsupervised ML models (classification, segmentation, clustering) combining NLP and structuredsignals; designed experimentation pipelines measuring quality and reliability trade-offs.
- Engineered two-agent LangGraph orchestration system with autonomous task decomposition and tool-use interface; reducedaverage response latency by 32% in production-ready Ray Serve deployment.

Machine Learning Engineer — Personalization, Ranking & Production ML Jul 2023 – Jul 2025

ICICI Bank Pvt Limited Mumbai, India

- Built ML-powered recommendation and ranking systems serving 1M+ daily requests; applied transformer-based ranking on GPUinfrastructure, achieving a 9% CTR lift and 12% improvement in recommendation relevance over production baselines.
- Developed user understanding and segmentation models using supervised and unsupervised learning (sequence modeling,clustering, anomaly detection) to capture preference and behavioral signals for real-time personalization.

- Built end-to-end Python training, serving, and experimentation workflows (FastAPI, Docker, AWS ECS); enabled rapid modelprototyping, iterative A/B-tested deployment, and reproducible evaluation at scale.
- Integrated AI tooling with large-scale data platforms for automated policy personalization — full-stack production ML experiencefrom data engineering and training through inference and product.

Research Intern — Neural Retrieval & Multimodal Representation Apr 2023 – Feb 2024

IIT Patna — AI-ML-NLP Lab Remote, India

- Built dense retrieval and reranking pipelines for content matching over 500K+ documents; improved NDCG@10 by 17% over BM25baseline using transformer embeddings.
- Developed unsupervised clustering and anomaly detection models over behavioral event streams; optimised GPU-based inferencepipelines, reducing end-to-end query latency by 40%.

Research Intern — Sequence Modeling & Behavioral Prediction May 2023 – Aug 2023

Polytechnique Montréal · MITACS Globalink Fellow (~200 awardees/year) Montréal, Canada

- Trained neural sequence models over 10M+ data points for temporal pattern learning; accelerated training 3.8× via parallelcomputing and optimised C++ pipelines.
- Implemented GPU-accelerated training loops and bottleneck profiling for production-grade behavioral prediction systems.

PROJECTS

Visual & Multimodal Retrieval System with Responsible-AI Evaluation Aug 2025 – Present

Python · PyTorch · FAISS · Hugging Face · Ray Serve

- Designed a multimodal retrieval pipeline combining vision and text encoders with dense FAISS indexing over 1M+ vectors; achieved8–12% gains in Recall@100 and 15–20% reduction in p95 latency via GPU-optimised Ray Serve batching.
- Integrated claim-level faithfulness verification and responsible-AI evaluation harnesses; benchmarked retrieval quality, latency, andfailure modes across model variants.

User Behavior Segmentation & Predictive Profiling Pipeline Jan 2026 – Present

Python · scikit-learn · PyTorch · XGBoost · Hugging Face

- Applied supervised and unsupervised learning (clustering, anomaly detection, sequence classification) to model user behavioralpatterns; benchmarked XGBoost, transformer, and LSTM architectures across latency and accuracy trade-offs.
- Built end-to-end profiling pipelines with temporal feature engineering and drift-detection harnesses — applicable to Pinterest's userunderstanding and recommender systems.

PUBLICATIONS

Structure-Grounded Medical QA: RDF Retrieval and Claim-Level Verification for Faithful Answering

ACL 2026 Workshop on Structured Understanding and Reasoning for Generative LLMs (SURGeLLM) — Submitted · 2026 | Stocks Natalias, Jatin S., Hunter, Kunchepu, Herrera, Youm, Gilda, Dorr

Contribution: Proposed claim-level verification as the core faithfulness mechanism; designed the structured retrieval pipeline; led evaluation design.

From Sights to Insights: Towards Summarization of Multimodal Clinical Documents

ACL 2024 (Long Paper, Main Conference) · Aug 2024 | Ghosh, Tomar, Tiwari, Saha, Jatin S., Sinha

Contribution: Designed the vision cross-attention fusion module; ran ablation studies isolating the contribution of multimodal grounding.

Anisotropic Noise Injection for Improving Utility in Differentially Private SGD

Under Review — University of Florida | Jatin S., Nahar, Mali

Contribution: Originated the core idea of shaping noise along gradient covariance eigenvectors; led theoretical derivation and empirical evaluation.

Availability: September 21 – December 11, 2026 (full-time, 40 hrs/week) · Open to San Francisco, Palo Alto, Seattle, New York, or Remote