

Vistrit Pandey

Chicago, IL • +1 215 452 9375 • vistritpandey@gmail.com • vistrit.space • linkedin.com/in/vistrit

Software Engineer • Applied AI and ML Systems • Research and Product Engineering • Forward Deployed Engineering

Experience

Blinkfire Analytics

Jan 2026 to Present

Software Engineer II, AI R&D

Chicago, IL

- Built **reinforcement learning**, synthetic data generation, and model evaluation workflows over **12+ years of proprietary multimodal data** for LLM training, benchmarking, and fine tuning.
- Designed **local LLM inference** infrastructure using **CUDA, TensorRT, NVIDIA Triton, Ollama, Docker, Kubernetes, and Linux** to improve serving latency, throughput, deployment control, and cost.
- Developed **AI agents** and retrieval systems combining **RAG, semantic search, tool execution, and customer scoped context** to generate analytics, recommendations, reports, and structured files.
- Built **multimodal** systems using **Whisper, LLMs, and vision language models** for audio understanding, contextual classification, automated review, and reduced manual processing.
- Partnered with product, infrastructure, leadership, and **enterprise stakeholders** to ship production AI services, backend APIs, and full stack product workflows.

Blinkfire Analytics

Apr 2024 to Dec 2025

Software Engineer I, AI R&D

Chicago, IL

- Shipped AI and analytics products using **Python, Go, TypeScript, REST, gRPC, Redis, Next.js, Vue.js, Firebase, and Google Cloud**.
- Integrated **GPT, Gemini, Llama, Hugging Face, and open source models** into semantic search, automation, **content intelligence**, and real time inference workflows.
- Built model fine tuning, quantization, serving, and inference optimization workflows using **PyTorch, TensorFlow, CUDA, TensorRT, Triton, Docker, and Kubernetes**.
- Developed backend APIs, asynchronous processing systems, **annotation interfaces, and customer facing applications** across distributed service architectures.

Drexel University

Sept 2022 to Sept 2023

Course Assistant, Computer Science

Philadelphia, PA

- Mentored **500+ students** across algorithms, machine learning, computer vision, computational imaging, programming, and software engineering through labs, debugging, code reviews, and project evaluations.

Projects

Meethealth | Founder and Lead Software Engineer

meethealth.us

- Built an **AI operating system** for dietitians using **Next.js, Python, Google Cloud, CUDA, and local LLM inference**, supporting more than **200 organizations** across clinical planning and patient management workflows.
- Generated more than **one million synthetic clinical diet plans locally** to create a domain specific training and evaluation dataset, then trained a nutrition language model that achieved **89% clinical accuracy, outperforming all frontier models** evaluated on the same internal benchmark.

Meetext | Founder and Lead Software Engineer

meettext.us

- Built and **scaled an AI communication platform** using **Next.js, Python, Firebase, speech recognition, LLM APIs, and Stripe**, reaching **20,000+ users** while leading seven engineers and designers.

Research and Open Source

Grounded Introspection | MLX, Python, Model Interpretability

github.com/VistritPandey/grounded-introspection

- Developed a **white box framework testing** whether LLM verbal uncertainty **tracks internal next token entropy** using read only hidden state and logit access, bounded **autonomous generation, and permutation controls**.
- Published reproducible code, raw step level records**, and a working paper from a preliminary **Qwen3.5 0.8B and 9B pilot**, explicitly treating the results as underpowered.

MachBoost | MLX, Hugging Face, Multimodal Inference

github.com/VistritPandey/machboost

- Built a **local inference server and Python package** with resident model serving, OpenAI and Ollama compatible APIs, **context backed speculative decoding, multimodal caching, and architecture aware state reuse**.
- Designed paired benchmarks showing up to **1.96x text, 18.33x repeated image, and 1.70x first view visual prefill speedups** under documented workload and quality constraints.

Education

University of Illinois Urbana Champaign

Aug 2026 to May 2028

Master of Computer Science

Illinois, US

Drexel University

Sept 2019 to Sept 2023

Bachelor of Science in Computer Science, Minor in Information Systems

Pennsylvania, US

Technical Skills

Languages: Python, Go, TypeScript, JavaScript, SQL, Java, C++, Bash

AI and ML: PyTorch, TensorFlow, Hugging Face, MLX, reinforcement learning, fine tuning, model evaluation, synthetic data, quantization, RAG, semantic search, multimodal AI, Whisper, vision language models

Systems and Product: CUDA, TensorRT, NVIDIA Triton, Docker, Kubernetes, Linux, GCP, AWS, Redis, REST, gRPC, distributed systems, React, Next.js, Vue.js, Firebase