

# Shreyash Bhatkar

[shryshbhtkr@gmail.com](mailto:shryshbhtkr@gmail.com) | [linkedin.com/in/shr3yash](https://www.linkedin.com/in/shr3yash) | [github.com/Shr3yash](https://github.com/Shr3yash) | [shr3yash.github.io](https://shr3yash.github.io)  
Champaign, IL, USA | (+1) 217-298-6740

## EDUCATION

**University of Illinois Urbana-Champaign** | Master of Computer Science, GPA: **4.0/4.0** | Expected May 2027  
**University of Pune** | B.Tech Electronics and Computer Science, CGPA: **3.62/4.0**

## TECHNICAL SKILLS

**Programming:** Python, C++, Java, Go, SQL, Bash

**ML & Agent Systems:** PyTorch, FSDP2, vLLM, TRL/GRPO, Hugging Face Transformers, RAG, LangChain, FAISS, SWE-Bench Verified, LLM Evaluation

**Distributed/GPU Systems:** CUDA, NCCL, Distributed Training, LLM Inference, KV-Cache Optimization, GPU Profiling, Slurm/HPC, Linux

**Infrastructure:** AWS EC2/EBS, Docker, Kubernetes, Kafka, Redis, Git/GitHub, Prometheus, OpenTelemetry

## WORK EXPERIENCE

**SSAIL – Supercomputing Systems & AI Lab, UIUC** Jan 2026 – Present

*Graduate Researcher – Agentic RL & ML Systems | Advisor: Prof. Minjia Zhang*

*Champaign, IL*

- Built an agentic RL research stack for SWE-Bench Verified combining Qwen3.5-4B, TRL AsyncGRPO, vLLM inference, FSDP2 training, Docker-based environments, NCSA DeltaAI, and AWS Blackwell GPUs
- Diagnosed trainer-to-inference policy synchronization in AsyncGRPO at source level, testing inference DP topology and packed/non-packed NCCL weight transfer and narrowing a reproducible first-update hang beyond model loading, serving, and basic distributed training
- Implemented a hybrid HPC/cloud sandbox architecture enabling DeltaAI-hosted RL workflows to remotely execute and resolve official SWE-Bench Verified instances on persistent EC2 Docker environments

**Oracle Corporation**

Jul 2023 – Jan 2026

*Software Engineer | Client Engagements: Batelco Telecom & PPC Retail*

*India*

- Led a **4-person** team from prototype through production rollout of a RAG assistant using LangChain, Cohere embeddings, and Neo4j for **500+** engineers and **2K+** daily queries; confidence-thresholded retrieval reduced hallucinations **41%** and ticket-resolution time **33%**
- Architected **8** Java/Kotlin microservices and **20+** REST APIs processing **10M+** daily transactions; implemented saga compensation and Redis idempotency to maintain **99.95%** correctness under concurrent load
- Built Kafka event pipelines with automated retry and idempotent processing across **12+** integrations, increasing delivery throughput **35%** with zero duplicate-charge incidents after rollout

## OPEN SOURCE CONTRIBUTIONS

- vLLM:** Fixed silent lm\_head LoRA mis-application under speculative decoding in Model Runner V2 — adapters now map per logits row (bonus + drafts) instead of per request, so MTP batches neither drop the adapter nor leak onto neighbors (#59341 / #59161)
- verl:** Made FSDP merged-LoRA hybrid sleep use level 1 (weights stay resident; adapters reload on wake) so actor↔rollout sync does not force a full reload; Megatron stays level 2 (#8046 / #7904)
- SGLang:** Fixed Falcon-H1 tied LM-head crash under FP16/BF16 by stopping in-place `.float()` on shared `embed.tokens.weight` and upcasting only the matmul path (#41652 / #41463)
- vLLM:** Rejected chat turns the chat template silently drops (HTTP 400) instead of serving a corrupted prompt; Jinja2 AST role detection covers selectattr/rejectattr and mixed text+tool/image probes without  $O(S \times L)$  re-renders (#59108 / #59090)

## PROJECTS

**Adaptive LLM Inference Runtime** | Python, PyTorch, vLLM, CUDA, Llama

- Built an adaptive decoding runtime that routes between single- and multi-token generation using entropy, margin, energy, and agreement signals, reducing autoregressive decoding steps by **2.3×** and long-form latency by **28%**
- Batched inference and LLM-as-judge evaluation from **90s to 3.3s/sample**, delivering a **27×** evaluation speedup

**Semantic Search & Transcript Retrieval** | Python, SBERT, FAISS

- Trained SBERT embeddings with contrastive learning and FAISS hard-negative mining; achieved ROUGE-1 **0.34**, ROUGE-L **0.29**, and LLM-as-judge quality **4.2/5**

## AWARDS & CERTIFICATIONS

- Oracle Certified Generative AI Professional & Data Science Professional | Above and Beyond Award | Pat on the Back Award | Runner-Up, CGIU Memorial AI Hackathon