

Tanish Gudise

tgudise@andrew.cmu.edu | linkedin.com/in/tanish-gudise/ | github.com/TanishGudise

EDUCATION

Carnegie Mellon University

Pittsburgh, PA

Bachelors of Science in Computer Science; Additional Major in Business Administration (Concentration in AI)

May 2027

- **Relevant Coursework:** Matrices & Linear Transformations, Business Computing, Principles of Imperative Computation, Introduction to Computer Sciences, Concepts of Mathematics, Functional Programming, Fundamentals of Programming, Parallel and Sequential Data Structures and Algorithms, Probability and Computing, Introduction to Machine Learning

TECHNICAL SKILLS

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

Systems/Distributed: distributed systems, real-time messaging, WebSockets, concurrency, gRPC, Redis, lock-free data structs

Tools & Platforms: Git, Linux, Docker, Kubernetes, AWS, GCP, Jenkins, PostgreSQL, MongoDB, Airflow, LangGraph, vLLM

EXPERIENCE

Airbnb

San Francisco, CA

Software Engineering Intern - AI Infrastructure

May 2026 – Aug 2026

- Built a **conversational AI agent** for Airbnb's internal prompt-engineering platform, applying **retrieval-augmented generation** and a **human-in-the-loop** confirm-required workflow, coordinating with the platform team on design.
- Instrumented **distributed tracing (OpenTelemetry)** across the agent's inference pipeline, applying **mesh-level proxy authorization** and span propagation to restore observability at scale for production debugging and monitoring workflows.
- Architected a **structured intake pipeline** and **mode-aware generation system** with real-time **diff rendering**, applying **few-shot grounding** over a curated production dataset to guide zero-shot prompt drafting for unseen input types.
- Developed an **in-context evaluation harness** integrated with the platform's **LLM execution runtime**, applying persistent result caching and a curated fallback dataset to validate agent-generated outputs reliably across seven categories.

Bliss

New York, NY

Software Engineering Intern

Jul 2025 – Jan 2026

- Developed and shipped an **MVP iOS mobile app** from the ground up using **Swift**, **REST APIs**, and **Firebase** for real-time data sync, implementing **MVVM architecture** with UI components across 15+ core consumer app features.
- Drove an 85% user task completion rate and a 25% increase in daily active users through **A/B tested UX** iterations.
- Connected app to a **distributed tech stack** via Chrome extension and **backend infrastructure**, achieving 99.5% uptime.

Lighthouse Invest

New York, NY

Software Engineering Intern

Jul 2024 – Dec 2024

- Designed a **PostgreSQL** schema and **connection pooling** layer using **PgBouncer** to handle concurrent trading account requests at scale during peak market hours, reducing query latency by 40% and improving throughput by 25%.
- Engineered a **REST API integration** with **Interactive Brokers** and **Twilio APIs** for automated account provisioning.
- Built **automated test coverage** and **CI/CD pipelines** with **GitHub Actions**, cutting deployment errors by 35%.

PROJECTS

Parameter Golf – OpenAI Model Craft Challenge | *PyTorch, CUDA, C, Distributed Training*

- Built a **16MB language model** on **8xH100s** stacking a **token-only n-gram tilt**, **asymmetric logit rescaling**, and **phased test-time training**; submission briefly topped the leaderboard at **1.0567 BPB** on **FineWeb** across three seeds.
- Ported a **C-based n-gram state machine** and the surrounding Python compute path from a merged precedent PR into the active codebase, then engineered an **in-eval-timer precompute** that fit a 167s hint table plus 355s phased TTT under the strict 600s validation budget across all three independent random seeds while staying under the 16MB artifact size cap.
- Ran over **60** carefully tracked ablations across architecture, **quantization** (**GPTQ** paired with **LQER** asymmetric rank-4), and test-time training hyperparameters; identified **PHASED_TTT_NUM_PHASES=1** as the single dominant compute-quality lever, cutting validation eval time by 49 seconds while simultaneously improving overall BPB on identical hardware.
- Synthesized **20+** merged competition PRs to map the full architecture lineage and stacked novel techniques on the most recent baseline while preserving full compatibility with the standard **CaseOps tokenizer** and BPB methodology.

StoryForge – AI Interactive Fiction Engine | *Python, React, TypeScript, LLM Agents*

- Architected **multi-agent AI storytelling platform** with specialized Narrator, Character, and Director agents, implementing orchestration protocols to generate branching narratives across **50+** dynamically generated story paths.
- Built **hierarchical memory system** using **ChromaDB** for **vector-based recall** and **Neo4j** for **causal knowledge graphs**, enabling persistent character memories and consequence tracking improving narrative consistency by **40%**.
- Developed full-stack **React/TypeScript** and **FastAPI** application featuring real-time scene generation, AI character portraits, and **embedding-based personality modeling** for consistent NPC behavior across **100+** interactions.
- Designed **RESTful API** layer with **JWT authentication**, story state persistence in **PostgreSQL**, and **WebSocket integration** for collaborative storytelling, supporting multiple concurrent users with sub-second latency and synchronization.