

BENSON ZHENG

(570) 765-3735 | zheng.ben@northeastern.edu | [linkedin.com/in/zhengbenson](https://www.linkedin.com/in/zhengbenson) | github.com/SlothfulDreams | slfd.dev | US Citizen

EDUCATION

Northeastern University

Bachelor of Science in Computer Science | GPA: 3.7/4.0

Boston, MA

Graduation: May 2028

EXPERIENCE

IBM | Software Development Engineer Intern - IBM Bob

Aug 2026 - Present

PricewaterhouseCoopers (PwC)

Jun 2026 – Aug 2026

Software Engineering Intern

New York, NY

- Engineered an agentic AI harness with Claude Agent SDK and computer use (UI automation), orchestrating autonomous private equity fee calculations via MCP tool-calling, reducing manual calculation time by 85%
- Shipped an LLM-as-judge grading pipeline with Pydantic schema validation across 6 LLM graders, catching 95% of calculation errors through independent cross-check verification and self-correcting retry loops
- Hardened CI/CD in GitHub Actions by enforcing 5 blocking lint, SCA, and complexity checks and migrating 2 package managers to uv lockfile-pinned builds, eliminating environment drift across local and CI environments

Thinkstruct

Mar 2026 – Jun 2026

Software Engineering Intern

Cambridge, MA

- Developed an agentic support chatbot with Vercel AI SDK using SSE streaming, tool-calling RAG over internal docs, and prompt guardrails against hallucination and legal advice, reducing support tickets by 40%
- Constructed a multilingual patent machine translation (MT) pipeline over 19M+ BigQuery records, running BERTAlign alignment on EC2 GPU compute with automated quality gates to produce a validated parallel corpus
- Architected a self-correcting multi-agent LLM pipeline in Python that parses raw case filings and generates court-ready DOCX IPR petitions, using evaluator-optimizer loops to rewrite failing sections

Tetracorp

Jan 2026 – Jun 2026

Software Engineer

Remote

- Architected and led Bluffs, a real-money iMessage gaming platform, to 23,750+ users and a Top-40 peak in Strategy on the US App Store, building a React Native app and 4 SpriteKit/SceneKit games with deterministic Box2D physics
- Engineered real-money Convex backend across 16 tables: idempotent integer-cents ledger, per-wager HMAC-signed geofencing, billing-matched Sumsb KYC, and 4-signal device fingerprinting to shut out fraud and balance drift
- Launched an LLM prediction market on Anthropic Claude API that auto-resolves real-world wagers through web-search grounding and structured tool outputs, with retry-then-void fallback and signed receipts to block tampering

Coverix

Jan 2026 – Mar 2026

AI Engineering Intern

Cambridge, MA

- Built an LLM evaluation pipeline with chain-of-thought prompting and Pydantic structured outputs, running 200+ automated test scenarios to benchmark voice agent accuracy on tool use and lead qualification at 93%
- Integrated real-time transcript middleware in Express connecting Voicerun with interruption detection, leveraging Upstash Redis for sub-100ms latency across 6 production agents handling outbound calls via Twilio
- Designed a ReactFlow-based deployment system with drag-and-drop prompt chaining and dynamic context injection, reducing configuration time by 75% for enterprise clients

PROJECTS

Simulacra (Yale Hacks '26 K2 1st Place) | Python, FastAPI, LangGraph, Socket.IO | [GitHub](#)

Mar 2026

- Designed a multi-agent orchestration engine in FastAPI with LangGraph StateGraph, running concurrent cognitive loops via asyncio with shared memory streams and Pydantic-validated structured outputs
- Built per-agent memory retrieval with temporal decay scoring, threshold-triggered reflection, and Deffuant bounded-confidence opinion dynamics to model echo-chamber formation across 50+ agents streamed over Socket.IO

TerrarAI (HackPrinceton '25 xAI 1st Place) | C#, xAI, tModLoader | [GitHub](#)

Nov 2025

- Created an agentic NPC mod for Terraria using tModLoader with a ReAct-style reasoning loop via Grok-4-Fast, supporting 10+ tools across 50+ parallel agents, winning \$8K in xAI credits (800+ participants)
- Engineered a C# agentic framework with async inference bindings, schema-validated action registry, and automatic replanning on 90-second timeouts

TECHNICAL SKILLS

Languages: TypeScript, Java, Python, JavaScript, C/C++, C#, Lua, SQL, HTML/CSS, Racket/Scheme

Frameworks: React, Next.js, Node.js, Express, Flask, FastAPI, LangChain/LangGraph, GraphQL, SpriteKit, PyTorch

Developer Tools: Git, Linux, Vim/Neovim, VS Code, Cursor, Postman, Agile/Scrum, Pytest, Jest

Databases: PostgreSQL, MySQL, SQLite, Redis, MongoDB, Convex (real-time serverless backend)

Cloud/Infra: AWS (Lambda, S3, EC2), Vercel, Docker, GitHub Actions (CI/CD), Supabase, Upstash (serverless Redis)