THEO HYTOPOULOS

Seattle, WA | theo.hyto@gmail.com | hytopoulos.github.io | linkedin.com/in/hytopoulos

AI infrastructure and MLOps engineer with proven agentic coding experience. Founding engineer who built and runs the AWS inference stack behind a realtime voice-AI product and cut per-session inference cost from $9/hr to $1/hr. M.S. in Computer Science; first author on LLM fine-tuning research. US work authorized, no sponsorship needed.

SKILLS

Infrastructure and MLOps: Amazon Web Services (AWS) EC2, Auto Scaling Groups, spot GPU instances; Terraform; Docker; CI/CD; ONNX CPU inference; PostgreSQL; Supabase; Railway; Grafana; Sentry

ML and LLMs: PyTorch; Hugging Face PEFT; QLoRA fine-tuning; Group Relative Policy Optimization (GRPO); Retrieval-Augmented Generation (RAG); embeddings and vector retrieval; Model Context Protocol (MCP); context engineering

Languages: Python, TypeScript, C/C++, Java, SQL, Bash. Spoken: English (native), Spanish (intermediate)

WORK EXPERIENCE

MUSE AI LLC — Remote

Co-Founder and Chief Architect — Oct 2025 – Present

Pre-launch iOS voice-AI app that interviews people to record oral histories (in beta).

- **Built and operate the production inference stack on AWS:** self-hosted speech-synthesis, enhancement and IPA transcription services on EC2 Auto Scaling Groups, provisioned end to end with Terraform.
- **Cut voice inference cost ~55%** with a decision model that wakes the realtime voice model only when needed, using cheap transcription and pre-recorded backchannels while it sleeps.
- **Designed a provider-neutral realtime voice-agent runtime** (OpenAI Realtime, Gemini Live, Qwen Audio); a mode state machine swaps prompts and tool sets per interview phase across 17 server-side tools with vector-memory retrieval.
- **Made open-ended voice sessions reliable:** durable context checkpoints, context compression that carries conversations across model-session rotations and reconnects, and per-session usage metering for cost tracking.
- **Shipped the React Native app (iOS/Android):** full-duplex voice streaming, session resume across redeploys, in-app purchases, and TestFlight/Play release pipelines.
- Turned the CEO's product requirements into the technical architecture and roadmap; **managed and trained the first engineering hire, and am handing off ownership of the platform to him.**

WESTERN WASHINGTON UNIVERSITY — Bellingham, WA

Graduate Research Assistant — 2023 – 2025

- Led the first baseline study and first open-source dataset for automatically identifying the human-error root causes of software defects, from dataset construction through modeling: 599 labeled program versions from 155 participants.
- **Fine-tuned 1.3B–7B-parameter code LLMs (DeepSeek-Coder, Code Llama, CodeSage) with QLoRA** for multi-label classification on one H100 GPU; best model hit 37.5% precision vs. 23% for untrained human analysts (63% relative gain), 0.693 macro AUC-ROC.

PROJECTS

Reinforcement-learning decompiler: trained a Qwen2.5-Coder-7B LoRA adapter with GRPO (verl, Modal) to decompile assembly into compilable C++, with a verifiable reward that recompiles each output and scores it against the target assembly. Raised compilable outputs from 10 to 32 of 39 validation samples and mean reward from 0.06 to 0.39 (GPT-4.1: 0.33).

PUBLICATIONS

T. Hytopoulos, T. Nguyen, K. Schutte, M. Enertson, S. Wehrwein, M. Vieira, F. Huang. "Identifying Human Error Root Causes of Software Defects Using LLMs: A Baseline Study." *IEEE Transactions on Dependable and Secure Computing*, submitted 2026, under review.

T. Hytopoulos, M. Chan, K. Roth, R. Wasson, F. Huang. "An Approach to Cognitive Root Cause Analysis of Software Vulnerabilities." *International Conference on Product-Focused Software Process Improvement (PROFES)*, Springer, 2024, pp. 11–26.

ADDITIONAL EXPERIENCE

Midjourney, QA Tester (remote), Dec 2022 – Dec 2024: reviewed image generations, informally tested internal models, and moderated the community Discord.

Western Washington University, Teaching Assistant (2022 – 2025) and CS Tutor (2022 – 2024): led programming labs of 30 students; courses included data structures, object-oriented design, and computer organization.

Leviathan Games, Software Engineer Intern, Dec 2021 – May 2022: wrote C to port NFL Blitz (1997) from MIPS to ARM Linux arcade hardware.

EDUCATION

Western Washington University: **M.S. Computer Science**, June 2025; **B.S. Computer Science** (CS Honors Program), June 2024