YIRMEYAHU (JEREMY) MWANGELWA

AI Engineer

Location: Carnegie, PA
Email: jeremynizamedia@gmail.com
LinkedIn: linkedin.com/in/jeremy-mwangelwa
Githib: github.com/jeremycoder
Profile: yirmeyahu.me

Professional Summary

AI Engineer with 4+ years of professional software engineering experience. I build and deploy production AI systems — fine-tuned TTS models, 5B-parameter video diffusion models, NLP pipelines processing 444K+ words, and end-to-end content automation platforms. As founder of Zambian Tech, LLC, I architect and ship AI systems spanning voice synthesis, video generation, and large-scale text processing on cost-efficient serverless GPU and CPU infrastructure.

Technical Skills

AI / ML: PyTorch, model fine-tuning (TTS, diffusion), LLM integration (Claude API, OpenAI, Gemini, Perplexity), RAG, local & serverless GPU inference, diffusion models (Wan2.2), voice cloning

Infrastructure: Docker, CUDA 12.1, RunPod Serverless, AWS S3, AWS Lambda, PostgreSQL, Redis, Celery, FFmpeg

Backend: Python, Django, REST APIs, async task pipelines, PHP, Laravel, Node.js

Frontend: JavaScript, TypeScript, Vue, Nuxt, React, HTMX

Workflow: Event-driven architectures, human-in-the-loop systems, OAuth2 integrations, CI/CD, Agile/Scrum

---

Experience

Zambian Tech, LLC — Founder & Principal Engineer (2023–Present)

Solo engineer building and operating production AI systems. All architecture, infrastructure, and deployment decisions.

AssetFlow *(Flagship Platform)*

- Built an end-to-end AI production platform across 3 major versions, orchestrating content pipelines from creation through narration, rendering, and delivery
- Implemented multi-stage pipeline orchestration with automated handoffs, status tracking, and human-in-the-loop review gates for quality control
- Integrated AI-powered content generation: article writing, TTS narration, image creation, and video rendering in a single automated workflow
- Built event-driven asset ingestion from AWS S3 with version awareness and multi-user access control on shared asset libraries
- Built a RAG system using Gemini File Search grounding with corpus management, topic pools with round-robin scheduling, idempotent generation jobs, and citation extraction for grounded article generation
- Implemented YouTube publishing integration with OAuth2 for automated delivery
- *Stack: Python, Django, PostgreSQL, AWS S3/Lambda, Celery, Redis, OpenAI, Gemini API, FFmpeg*

Voice Synthesis & Cloning

- Fine-tuned the F5-TTS model on ~2 hours of custom voice recordings for production-quality voice cloning; deployed on serverless GPU with custom Docker container and CUDA 12.1
- Discovered and documented critical, previously undocumented reference audio requirements that eliminate common synthesis artifacts
- Deployed Qwen3-TTS (1.7B parameters) on serverless GPU as an additional voice cloning endpoint with flash attention and S3 storage
- Integrated Pocket TTS (100M-parameter model) for CPU-only inference at 6x real-time speed with ~200ms latency to first audio chunk on just 2 CPU cores
- *Stack: PyTorch, F5-TTS, Qwen3-TTS, Pocket TTS, CUDA 12.1, Docker, RunPod Serverless*

AI Video Generation

- Deployed the Wan2.2-TI2V-5B diffusion model (5 billion parameters) on serverless GPU for text-to-video and image-to-video generation
- Produces 720p video at 24fps with configurable duration, guidance scale, and seed control; supports batch processing with optional S3 storage
- *Stack: Wan2.2, PyTorch, Docker, RunPod Serverless, CUDA, S3*

NLP Translation Pipeline

- Built an AI-assisted word-by-word translation pipeline processing 30,286+ verses across 66 Biblical books (Hebrew OT and Greek NT) with full morphological analysis for every word
- Processed 444,785+ individual words with morphological tagging, etymology, Strong's cross-references, and interlinear formatting
- Built a structured lexicon database generating 475K+ SEO-friendly pages with etymological comparisons across 15 Niger-Congo Bantu languages
- Implemented a community contribution system with trust levels, voting, confidence scoring, and moderation workflows
- *Stack: Python, Django, PostgreSQL, Perplexity API, HTMX, DRF*

BookForge & SlideFlow

- Built an end-to-end non-fiction book creation pipeline: AI-driven market research, chapter-by-chapter generation with citations, and KDP-formatted DOCX export using Claude and Perplexity APIs
- Built a presentation generator that converts long-form articles into structured slide decks via an 11-type JSON schema with speaker notes, outputting to PowerPoint via python-pptx
- *Stack: Python, Django, Claude API, Perplexity API, python-docx, python-pptx*

Trader Interactive — Software Engineer II (Sep 2023–Feb 2026)

Sole engineer on the Media & Ad Technology stack for the majority of a 2.5-year tenure, owning the full engineering side serving 1.2 million active users.

- Redesigned the Google AdSense media advertising architecture, increasing annual revenue by $500,000 and improving ad performance by 75%
- Integrated a third-party API into the core system and Google Ad Manager, automating ad delivery processes and contributing to a $200,000 increase in annual revenue
- Developed new features and maintained legacy systems across the platform, ensuring seamless ad delivery and user experience at scale
- *Stack: PHP, JavaScript, TypeScript, Vue, Nuxt*

Rimsys — Associate Software Engineer (Apr 2021–Jul 2023)

Early engineer at a regulatory SaaS startup, contributing to core product from an early stage.

- Built and maintained responsive web applications using Vue.js and Nuxt.js for the core product frontend

- Implemented RESTful APIs and integrated backend services using PHP and Laravel, enabling data flow across system modules
- Adopted TypeScript across frontend codebases to improve code quality, maintainability, and type safety
- *Stack: Vue, Nuxt, TypeScript, PHP, Laravel*

Projects

NuxtIAM (2021–2024) — github.com/jeremycoder/nuxt-iam

- Built and open-sourced a Nuxt authentication and authorization framework in Node.js and TypeScript with cookie-based sessions, JWT, email verification, and password reset
- 97 GitHub stars, 27 forks

Education

PennWest University (formerly California University of Pennsylvania) Bachelor of Science in Computer Science, Magna Cum Laude — May 2021