Panagiotis (Panos) Gkilis

Machine Learning Engineer — Production ML Systems · MLOps · Speech & Language AI

Thessaloniki, Greece · open to Remote / Athens / Oslo · bedvibe@bedvibe.studio · Greek, English, Norwegian (all fluent) · EU citizen, Norwegian-registered business (ENK)

GitHub Mormolykos · Hugging Face pan82 · LinkedIn · ORCID 0009-0007-3805-170X

Everything below is live, running and open to inspection — please look before you decide.

ai.bedvibe.studio/work — start here: everything I have built and published, on one page, with the links into all of it.

tts.bedvibe.studio/portfolio — 36 project write-ups you can open right now: architecture, screenshots, evaluation reports and the failures I found in my own work.

ai.bedvibe.studio/articles — engineering notebook: how each system was debugged, measured and gated, written up in full.

SUMMARY

Six years of end-to-end production ML: models trained from scratch, deployed behind live APIs, and operated for paying customers. Sole engineer of BedVibe Studios — a multilingual speech-AI platform built on a ~730M-parameter TTS model trained on consumer GPUs across 13 languages, running on FastAPI, PostgreSQL/Supabase, Docker and Linux/Nginx with Stripe-billed usage. Nine published research records with Zenodo DOIs — seven publications and two software releases. Four open-source Python libraries, all published on PyPI, plus a real-time telemetry engine in Rust.

HOW I BUILD

> I develop with AI coding agents — Claude Code and Codex, cross-checked against each other through MCP tooling — and treat their output as untrusted until it clears a gate: automated evaluation suites, ASR-validated model QA, and static checks. All four of my open-source libraries *are* those gates: **ttsproof** (TTS failure-mode QA), **trainproof** (training-run linter), **spkproof** (speaker-verification measurement checks) and **notchecked** (coverage accounting — what was checked, what could not be, and what was never in scope, as typed states rather than pass/fail). The method is why a LoRA collapse and a live RAG regression were caught before release, not after.

> **Reviewed and adopted by outside engineers:** an IBM Certified Expert IT Architect (21 years at IBM) reviewed the **notchecked** schema publicly and contributed three corrections, all accepted; the maintainer of SpeechSDK, a commercial API spanning 17+ speech providers, requested an integration, which shipped; **trainproof** is the schema's first production adopter.

CORE SKILLS

ML & modeling — transformer training from scratch (AR + NAR neural codec LM), LoRA fine-tuning, speaker embeddings (ECAPA-TDNN, WavLM, ResNet), SentencePiece tokenization, Whisper/WhisperX ASR & forced alignment, retrieval/RAG evaluation, LLM-output benchmarking

MLOps & deployment — FastAPI inference APIs, async job queues, release-gating automated evaluation suites (ASR-validated WER/CER), distributed tracing (OpenTelemetry + Langfuse joined on a shared trace id, Jaeger), LLM gateways with per-key quota, token metering, bounded admission and deadline-aware failover, monitoring & production debugging, systemd service management, nginx/TLS reverse proxying, firewall hardening, GPU inference, ONNX export, Docker, Linux, HTTPS/domain routing, CI/CD (GitHub Actions, PyPI trusted publishing)

AI orchestration & data — LangChain/LangGraph stateful validate-retry graphs, CrewAI (evaluated against LangGraph on eight fault scenarios and rejected on measurement), MCP servers *and* clients with schema-validated responses, RAG pipelines with pluggable vector backends (FAISS IVF/FlatIP, Qdrant HNSW, brute-force baseline) benchmarked on recall@k against an exact index, hybrid retrieval (dense + BM25), schema-enforced output (Pydantic), self-hosted n8n; speech-dataset curation (100+ hours), metadata schema design (1.55M records), custom binary training formats

Backend & languages — Python, SQL, JavaScript, Rust, HTML/CSS · PostgreSQL (schema design, RLS, plpgsql), Supabase, REST, JWT/auth, Stripe & idempotent webhooks, AWS SDK, Deno/TypeScript edge functions

PROFESSIONAL EXPERIENCE

Founder / Machine Learning Engineer · BedVibe Studios

2020 – Present · Greece / Remote · Norwegian-registered sole proprietorship (ENK)

- **Model training** — trained a ~730M-parameter multilingual neural-codec TTS model from scratch (24-layer AR + NAR transformer over Encodec/DAC tokens, Vocos vocoder): spanning 13 languages, 15 human speakers plus 2 synthetic voices, and 6 emotional states — three independent spans, not a crossed grid — on ~3,000–4,000 GPU-hours on consumer GPUs.
- **Production deployment** — serve it behind FastAPI (20+ REST endpoints) with async queues for long-form generation, 200–400 ms GPU inference per utterance, and Stripe-billed token accounting. Live at tts.bedvibe.studio.
- **Infrastructure** — operate a 10-service backend on a self-managed Linux server: nginx/TLS reverse proxy, each service isolated as a systemd unit, PostgreSQL 16, default-deny ufw firewall.
- **Evaluation & monitoring** — built an ASR-driven QA harness (WER/CER, silence/loop/replay detection, hold-out cases) gating every checkpoint; it caught a total LoRA fine-tune collapse — 60/60 hard failures vs 14/60 baseline — before release.
- **Data pipelines** — engineered a 108,000-sample multilingual corpus with a custom 16k SentencePiece tokenizer, an ECAPA-TDNN speaker-embedding pipeline (4,287 samples, zero failures), and a memory-mapped binary format that removed a 3–5× dataloading

bottleneck.

- **LLM / RAG evaluation** — benchmarked 5 retrieval strategies on a 240,767-word corpus against an 80-question gold set (3.3× context recall over naive baseline) and scored 9 external AI systems on separated evidence-grounding vs answer-coverage metrics.
- **Datasets & compliance** — recorded and curated 100+ hours of rights-cleared studio speech under signed talent contracts with GDPR-aware handling; commercial emotional-speech datasets sold to external customers.

Norwegian Language Teacher · globalcitizen

Sep 2013 – Jun 2019 · North Athens

Taught Norwegian to Greek- and English-speaking adults through B2; designed curricula and explained complex material across three languages.

Natural Sciences Teacher · Aristarhos Private High School

Sep 2014 – Jun 2017 · Peraia

Taught chemistry and natural sciences; prepared students for university-track examinations.

RESEARCH & PUBLICATIONS

Structure Substitutes for Scale in Cross-Document Temporal Reasoning — Zenodo, Aug 2026 · DOI 10.5281/zenodo.22169643. Thirty-eight questions ask whether one event precedes another when the two are narrated in different documents and share no character, place or causal link — the ordering exists only across the corpus, never inside it. Given the passages, every model scored 0/38 at four sizes, refusing 92–100% of the time: the control holds. Given the identical facts as a structured chronology, an 8B model scored 28/38, matching a 14B model given prose — structure substituting for roughly a doubling of parameters. The advantage is gone by 14B, and scrambling the supplied order collapses accuracy to 10.5%, ruling out memorisation. After publishing I audited the scorer behind my own headline: it proved strictly conservative (1,786 items re-scored, 30 gains, 0 losses) but had understated the prose conditions, so my claimed gap over the best prose representation halved. The equivalence held; the gap did not. Corrected in place, not retracted. ai.bedvibe.studio/structure-not-scale

Speaker Encoders Disagree About Who You Are When You Shout: A Content-Matched Benchmark of Fourteen Speaker Encoders Under Expressive Phonation — Zenodo, Aug 2026 · DOI 10.5281/zenodo.22158030. Fourteen encoders over one frozen list of 11,935 trials; eight speakers × the same 1,360 sentences × six phonation states, so content is held constant while delivery varies. Equal error rate spans 0.047–0.233 on identical audio — five-fold, from encoder choice alone — with 29 of 91 pairwise comparisons surviving Holm–Bonferroni. The encoder shipped inside a widely used open-source TTS system is the worst of the fourteen for every speaker and every condition (13/13 comparisons survive correction); the mechanism is impostor-score compression, not degraded genuine scores. Intervals from a vertex bootstrap over the speaker graph, whose correction cost two of my own already-written claims. Ten retractions published alongside the result, including an 11.7 dB effect withdrawn against a 12.09 dB instrument artifact budget. Score tables, analysis code and the full decision record public; every statistic regenerates line for line from a CSV. github.com/Mormolykos/speaker-encoders-disagree

Intra-Speaker Vocal Variation and Speaker-Embedding Displacement: A Matched-Content Replication Across Three Encoder Architectures — Zenodo, Aug 2026 · DOI 10.5281/zenodo.21921958. Four speakers, matched sentences, three encoders (ECAPA-TDNN, ResNet, WavLM); 30/30 speaker × condition × encoder cells negative. Diagnoses differential measurement error in F0 tracking (Fisher $p = 2.4\times10^{-11}$); pre-registered, with the adversarial prior-art audit and a withdrawn $p = 0.013$ result published in full.

The Loss Curve Is Not a Sufficient Statistic: Silent Objective Failures from Sentinel–Class Collisions in Neural Codec Language Models — Zenodo, Aug 2026 · DOI 10.5281/zenodo.21864659. A sentinel–class collision makes masked cross-entropy silently delete an entire class — usually the end-of-sequence token — while the loss curve still descends normally. Reproduced in 40 lines on CPU.

BedVibe-TTS: An Engineering Report on a Multilingual Neural Codec Language Model with Multi-Axis Conditioning — Zenodo, Apr 2026 · DOI 10.5281/zenodo.19781414. Architecture, data pipeline, training diagnostics and ablations, plus a reproducibility finding on attention-backend × precision interactions across GPU generations.

An Automated Failure-Mode QA Framework for Neural Text-to-Speech Systems — Zenodo, Jun 2026 · DOI 10.5281/zenodo.20757553. 130 edge cases × 3 voices (390 samples), zero structural audio defects; blinded human validation of the 42 ASR-uncertain cases (controls 15/15) split 45/55 real failures vs ASR false-negatives, confirming quarantine-for-review.

Diagnosing Hierarchical Retrieval Failure in Long-Document RAG — Zenodo, Jun 2026 · DOI 10.5281/zenodo.20692451. Stage-level ablation isolating first-stage chapter-selection error compounding (27/80) and neighbour-expansion dilution (32/80); oracle-chapter routing recovered recall to 0.7844.

LongBook Verifier — Evidence-Grounded Evaluation of Long-Document Retrieval and AI-Generated Answers — Zenodo, 2026 · DOI 10.5281/zenodo.20513116. Retrieval baselines, model-output scoring and a 240,767-word stress-test benchmark, released as a research package.

A 2D Linear-Elasticity Finite Element Solver Written from First Principles, Validated Against the Kirsch Stress Concentration Factor — Zenodo, Aug 2026 · DOI 10.5281/zenodo.21892064. Hand-written Delaunay mesher, isoparametric elements and CG solver; no FEA library, no scipy. Recovers the analytic stress concentration factor as 3.00002 against an exact 3 the code never receives. Ten gates pre-registered; nine pass, the failure published.

SELECTED PRODUCTION SYSTEMS — LIVE AND PUBLICLY VERIFIABLE

BedVibe TTS Platform — multilingual emotional TTS with auth, token wallet, subscriptions and an async generation queue. Paying customers. tts.bedvibe.studio

ttsproof, trainproof, spkproof & notchecked (open source) — TTS failure-mode QA built from the published framework, benchmarking closed-source engines via a SpeechSDK integration; a training-reliability linter (92 rule IDs, 342 tests, 5 log formats, standard library only) that flagged a real run whose best checkpoint was step 99 while every saved checkpoint came from step 1600+; deterministic F0-contamination and trial-design checks for speaker-verification studies (9 rule IDs, exact tests, zero dependencies); and

coverage accounting that reports what a validator could *not* check as eight typed terminal states instead of silence. github.com/Mormolykos

Grounded Portfolio AI Agent (RAG) — retrieval-augmented Q&A over my work, hardened against 32 adversarial cases (prompt injection, credential invention), shipped with a public evaluation report; the eval suite re-runs on every corpus change and has caught a live model regression. Three vector backends behind one interface — brute force 0.86 ms, FAISS 0.035 ms, Qdrant gRPC 1.3 ms — with the simplest one shipped on measurement, and the whole answer path traced span by span (request p50 12.41 → 2.91 ms once tracing showed a model client being rebuilt on every answer). …/portfolio/agent-evals

basalt — LLM model gateway (open source) — one OpenAI-compatible endpoint over several backends: per-key quota, token metering, bounded admission, deadline-aware retries and failover. Removing backpressure let 32 callers all receive HTTP 200 while 31 arrived after they had given up — a 100% success rate and one usable answer; bounded admission raised that to six. github.com/Mormolykos/basalt

slate — MCP retrieval server & client (open source) — exposes a production RAG corpus to any agent over the Model Context Protocol, with typed responses instead of JSON strings so a drifting tool schema fails loudly rather than returning a confident wrong answer. Measured protocol overhead: 1.85 ms on top of 2.04 ms of retrieval. github.com/Mormolykos/slate

aether — real-time aerospace telemetry (Rust, open source) — live ADS-B ingestion with per-axis Kalman tracking and closed-form conjunction screening, exposed over a C ABI. A controlled A/B took rejection of valid observations from 16.1% to 1.9%. github.com/Mormolykos/aether

GEO Observatory — AI-crawler observability: append-only event ledger, deterministic deduplication, rule-based bot classification, public dashboard, automated daily collection on a hardened systemd deployment. geo.bedvibe.studio

AI Companion (desktop, Unity/C#) — supervises a 5-process local AI fleet (Rust VAD, Whisper, llama.cpp, FFmpeg) with CPU-specific model builds, MAC-bound licensing and Stripe; local inference to cloud TTS.

Nine further live systems — BookProof, Audio Drama Studio, Audiobook Studio Editor, Talking Avatars, The Mirelands browser game, a public API tester, a Rust/Axum Amazon Polly service, an MCP mail assistant and a forced-alignment microservice — with screenshots and technical write-ups at tts.bedvibe.studio/portfolio.

EDUCATION & LANGUAGES

Hellenic Open University — university-level coursework in Natural Sciences: 222 ECTS completed, 11 of 12 required modules (mathematics, physics, chemistry, biology), and six laboratory courses completed where three are required. Final module scheduled September 2026.

Greek (native) · English (fluent, professional technical writing) · Norwegian (fluent) · advanced technical documentation in Greek and English.