

Artem Melnytskyi

Principal AI Engineer / AI Architect

Stuttgart, Germany | amelnytskyi@gmail.com | +49 152 57231124

GitHub: FastAnt | LinkedIn: artem-melnytskyi-635959118 | Atmos demo



PROFESSIONAL SUMMARY

AI engineer and architect with 13+ years in software engineering, embedded systems and automotive R&D. Designs and trains custom neural architectures for gesture and human activity recognition, audio and visual decisions. Combines generative and multimodal AI, efficient edge inference and agent harness engineering with hands-on prototyping and technical leadership.

PROFESSIONAL EXPERIENCE

Mercedes-Benz — AI Specialist, Entertainment R&D

Sindelfingen, Germany | Company tenure: May 2022–Present | Previously Lead Developer

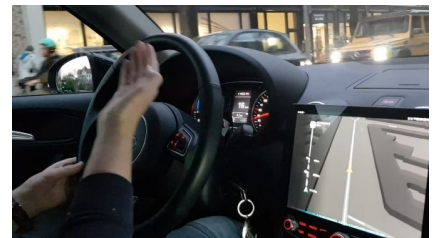
- Design and train CLM models within entertainment R&D, spanning model architecture and training.
- Develop and train neural codecs as part of entertainment R&D, combining model engineering with software prototyping.
- Build n8n orchestration workflows and design AI-agent interactions for software development.
- Define entertainment-domain architecture, design components and resolve integration issues on new hardware.
- Lead technology evaluation, architecture reviews, technical presentations and hackathons to guide implementation decisions.



Luxoft — AI Solutions Architect / Team Leader

Berlin / Kyiv / Odesa | Company tenure: Jul 2016–2022

- Developed and trained high-performance gesture-recognition models; built data-collection systems, integrated recognition into navigation prototypes and optimized inference for NVIDIA hardware.
- Developed an NLP-based search engine and its data-collection infrastructure; designed media frameworks and Flutter navigation proofs of concept.
- Led navigation HMI architecture and C++/Qt/QML development, including Android and QNX porting, integration planning and technical mentoring.



Earlier role: Software Engineer, Jul 2016–Jul 2017 — navigation HMI development in C++/Qt/QML and Android integration.

Bankomzvjazok JSC — Software Engineer

Kyiv, Ukraine | May 2013–Jul 2016

Developed embedded C/C++ payment software and fuel-station payment systems under memory, networking and cryptography constraints.



Earlier Research & Teaching

E. O. Paton Electric Welding Institute — Scientist. Sep 2015–Oct 2017, Kyiv. Applied machine learning to materials and structural diagnostics.

National Technical University of Ukraine — Physics Teacher. May 2010–Oct 2014, Kyiv. Delivered physics lectures and practical classes.

National Pedagogical Dragomanov University — Technician. May–Oct 2010, Kyiv. Supported automated student assessment.

Research and teaching dates overlap with engineering appointments.

SELECTED AI & SOFTWARE PROJECTS

Independent projects and research prototypes | Selected work, 2025–2026

Samsone + CLM — Audio Decision Models

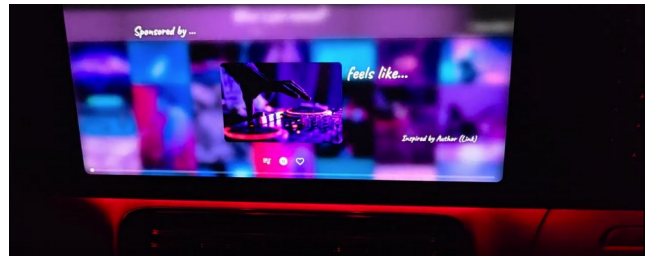
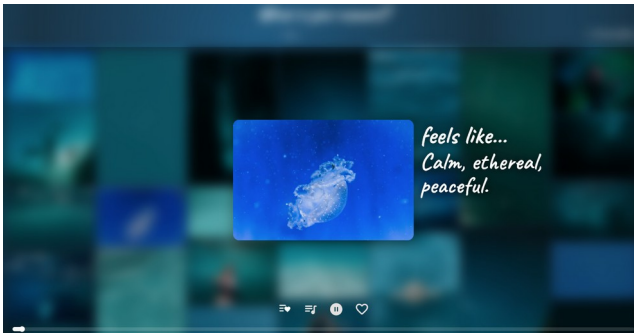
Built a non-generative audio question-answer ranking system around frozen Samsone-356M. Designed a shared 960 → 4096 adapter and fine-tuned pretrained CLM state/action heads to produce 512-dimensional representations. Prepared 404,473 musical training QA examples, including FMA genre data and replay; added calibrated scoring, a Python API and an audio-upload web demo. Added ExecuTorch exports for Kotlin Android integration. Python, PyTorch.

LFM + CLM — Visual Decision Models

Designed and trained CLM-style state/action projection heads over a frozen LiquidAI LFM2.5-VL-450M backbone. Implemented hidden-state extraction, cached features, text/vision specialist routing and candidate ranking without autoregressive answer generation. Built a comparison playground and experiments across head capacity, learning rates and source-balanced sampling. Python, PyTorch, Transformers.

Atmos — Immersive Music Player

Developed a synesthetic music-player prototype combining visual mood exploration, audio playback, playlists and favourites. Published a Flutter web demo and demonstrated the interface on an automotive display. Dart, Flutter, just_audio. [Live demo](#)



Neural Audio Codecs on Android

Ported SpectroStream to a standalone C++/LiteRT decoder: token-container parsing, RVQ reconstruction, context-aware inference and iSTFT. Added a streaming C API, Flutter FFI, persistent decoder workers and a startup-comparison demo. Implemented FP32/FP16 export and a separate SAGE W8A8 decoder bridge in Kotlin; validated SpectroStream CPU decoding on Android. C++, Python, Flutter, Kotlin, LiteRT/XNNPACK.



DeepReplay — Context Orchestration for Coding Agents

Built a DeepSeek Harness plugin for dynamic project-context selection and architecture-aware navigation. Implemented relevance scoring via remote Liquid and local llama.cpp backends, Handbook ingestion, context budgets, pins/exclusions, prompt snapshots and an interactive architecture view. Preserved canonical session history. TypeScript, Node.js, React, Cordis.

Harness Handbook — Language Support & Agent Integration

Extended Ruhan Wang's Harness Handbook with Kotlin parsing and call resolution, build-configuration analysis, configurable LLM endpoints, incremental updates with rollback and token accounting. Used the generated codebase maps to build the Handbook integration in a new DeepSeek Harness plugin (DeepReplay). Python, tree-sitter.

Android Multimodal Runtimes

Built a Kotlin/C++/JNI LFM2.5-VL runtime with camera/EGL preprocessing, native CLM heads and action caching; validated CPU inference on Galaxy S25. Separately integrated Samsone hidden-state extraction and general/music CLM scorers through ExecuTorch.

SELECTED RESEARCH & EXPERIMENTS

Audio grounding and training methodology. Built artist/source/PCM-disjoint splits and validation-based checkpoint selection; tested correct, random and silent audio to expose language-only shortcuts. FMA Large genre ranking reached 42.41% balanced accuracy across 11 classes on 569 reserved-test recordings (3,414 QA), versus 8.19% for random audio. Generalisation to unseen question wording remained weaker; genre performance is an experimental result.

Decision-head capacity and mobile inference. Compared head depth/width, sampling and learning-rate schedules. Selected visual heads reached 62.71% A-OKVQA validation accuracy (n=1,145), versus 36.94% without images; larger heads did not improve this evaluation. Separately benchmarked an Android text-ranking runtime: p50 27.9 ms over 1,000 requests with 100 cached actions on Galaxy S25 CPU/INT4. Mobile held-out ranking quality remains unverified.

Neural compression: fidelity versus computation. Compared chunk sizes, contextual decoding and FP32/FP16 exports. Context-aware C++ decoding improved boundary SNR against the Python reference from 4.91 to 61.14 dB on a 162-second fixture. Android CPU benchmarks recorded RTF 0.79 for the fast f25 mode with boundary artefacts, versus 2.59 for context-aware f75, both at six threads. The study separated decoder throughput from waveform fidelity.

TECHNICAL EXPERTISE

Main programming stack. C++, Python, Dart/Flutter, Kotlin, Qt/QML; TypeScript for agent tooling.

AI & model engineering. Design and training of custom neural network architectures; high-performance gesture and human activity recognition; transfer learning, fine-tuning, decision heads, representation learning, dataset preparation and ablation studies; PyTorch, TensorFlow, Keras, scikit-learn.

Visual & multimodal AI. Qwen and FLUX model families, Stable Diffusion; visual understanding, generative pipelines and custom LoRA training; Transformers, OpenCV.

Music & audio generation. YuE/YuE2, Stable Audio 2/3, Magenta RealTime 1/2, ACE-Step and MusicGen; custom model development, hybrid architectures, model adaptation and evaluation; torchaudio.

Neural compression & efficient AI. Neural audio codecs, RVQ latent representations, neural decoders, CLM/System-1 ranking, hidden-state inference, candidate caching; bitrate/quality/compute trade-offs.

Agentic systems & workflows. Agent architecture analysis and modification: DeepSeek Harness, Prime Agent, Kilo Code. Harness and plugin creation; RAG, GraphRAG, dynamic context, agent memory, tool orchestration, asynchronous APIs and experiment pipelines; n8n, llama.cpp.

Edge AI & optimisation. Android model integration, quantization diagnostics, model conversion, memory/performance profiling; LiteRT/TFLite, ExecuTorch, XNNPACK, ONNX, TensorRT, C++/JNI/FFI, NVIDIA hardware.

Architecture & platforms. Software/system architecture, distributed systems, embedded automotive infotainment; Linux, Android, QNX, gRPC, REST, IPC, Docker and CI/CD.

Languages. English — fluent; German — B1 (certified); Ukrainian — native; Russian.

EDUCATION

National Technical University of Ukraine “Kyiv Polytechnic Institute”

Master’s Degree in Physics — Faculty of Physics and Mathematics, 2008–2013.

Specialist Degree in Economics of Enterprise, 2010–2012.

Postgraduate research. Investigated neural-network models for predicting degradation in metal structures at the E. O. Paton Electric Welding Institute.

CERTIFICATIONS & PROFESSIONAL ACTIVITIES

Speaker: **Computer Visioners 2019 (AI Booster)** — co-presented “Gesture Recognition System for Automotive Industry”.

Additional talks: Embedded Fest; “Friendly Co-pilot as a Practical AI Application”.

Publication: **Application of Machine Learning Technologies in Diagnostics** — co-author, 2016.

Selected certificates: **Neural Networks and Computer Vision** (Samsung Research / Stepik, 2019); **Neural Networks and Text Processing** (Samsung Research / Stepik, 2020).