AQEA Edge: Sensor-Stream Processing Without AI Accelerators

A Public-Safe Technical Whitepaper for Industrial / Edge / Robotics Decision-Makers

NextX AG

2026-05-09

Contents

Executive Summary **2**
- Validated Domains . . . 2
- What's New for Industrial Decision-Makers . . . 3
- Why It Matters for Edge Deployment . . . 4
- What We're Asking . . . 4

1 Problem Statement **5**
- 1.1 BOM-Cost at Fleet Scale . . . 5
- 1.2 Vendor Lock-In and Strategic Risk . . . 5
- 1.3 Energy at Fleet Scale . . . 6
- 1.4 Encoder-Family Heterogeneity in Real Deployments . . . 6
- 1.5 The Specification We Set . . . 6

2 Technical Approach **7**
- 2.1 Pipeline Overview . . . 7
- 2.2 What the Substrate Is . . . 8
- 2.3 Why It Works Across Encoder Families . . . 9
- 2.4 Stage 1 — AQEA Encode . . . 9
- 2.5 Stage 2 — Similarity Search . . . 9
- 2.6 Cross-Platform Bit-Identity . . . 11
- 2.7 Integration Pattern . . . 11

3 Empirical Results **12**
- 3.1 Methodology . . . 12
- 3.2 Cross-Domain Results Table (Headline) . . . 13
- 3.3 Per-Domain Detail . . . 13
- 3.4 Reversibly-Decodable Substrate — Decoder Pareto-Front . . . 16
- 3.5 Cross-Vendor Bit-Identity (Independent Verification) . . . 18
- 3.6 Reproducibility Statement . . . 18

4 Competitive Position **19**
- 4.1 Comparison: Edge-AI Accelerators . . . 19

4.2 When AQEA Edge Replaces an Accelerator . . . 21
4.3 Per-Query Energy Comparison . . . 21
4.4 Hardware Reach: Where AQEA Edge Runs . . . 23
4.5 What This Architecture Is Not . . . 24

5 Use Cases **24**
5.1 Predictive Maintenance on Industrial Machinery . . . 24
5.2 Robot Fleet Health Monitoring . . . 25
5.3 Wearable / Consumer-IoT Anomaly Detection . . . 26
5.4 On-Device Voice / Acoustic-Event Recognition . . . 26
5.5 Regulated-Industry Audit-Chain (Audit-Chain Mode) . . . 26
5.6 Where AQEA Edge Is Not Yet the Right Choice . . . 27

6 Roadmap and Ask **28**
6.1 Validated Today . . . 28
6.2 Q3 2026 Roadmap . . . 28
6.3 Q4 2026 Roadmap . . . 28
6.4 Partner Ask . . . 29

7 Appendix **30**
7.1 Reproducibility Statement . . . 30
7.2 Glossary . . . 30
7.3 References . . . 31
7.4 Patent-Pending Disclosures . . . 32
7.5 Acknowledgements . . . 33
7.6 Contact . . . 33

Executive Summary

For Industrial Engineering and Edge-IoT Decision-Makers — Two-Minute Read

AQEA Edge is a processing architecture for similarity-based recognition over multi-channel signal streams — sensor data, vibration, motion, audio, hardware-state — that runs on any hardware with SIMD-CPU or Vulkan/Metal-Shader, **without a dedicated AI accelerator**. The pipeline that today requires an NVIDIA Jetson, Google Coral, Hailo or Apple Neural Engine for similarity-based classification can run on the CPU that is already in the device, with cross-platform bit-deterministic identity and order-of-magnitude lower energy per query than CPU-Float baselines.

Validated Domains

Domain	Encoder family	Recall vs Float	Energy advantage
Text retrieval (msmarco-100k)	BGE-large transformer	97.4 %	6.6× lower
Speech speaker-verification	WavLM-SV transformer	**99.2 %** □	**7.7× lower** □
Humanoid robot motion classification	FFT spectral (classical)	95.6 %	3.3× lower

Domain	Encoder family	Recall vs Float	Energy advantage
Industrial robot anomaly detection	Spectral classical (raw bins)	**133 % □**	**5.2× lower**
Hardware fall-prediction sensor	Spectral classical (raw bins)	**174 % □**	**3.55× lower**
Biological-sequence retrieval (SwissProt)	ESM-2-35M transformer	96.5 %	4.06× lower
Medical imaging (PathMNIST)	BiomedCLIP-ViT transformer	96.8 %	5.6× lower
Code-search (CodeSearchNet)	codet5p-110m transformer	**100.0 % †**	2.7× lower
Music classification (GTZAN)	CLAP-HTSAT transformer	98.1 %	9.4× lower
Multispectral (EuroSAT 13-band)	Band-statistics 60-d (classical)	92.3 %	1.5× lower
Vision-DCT-truncated (CIFAR-100)	DCT-zigzag-256 (classical)	84.6 % (Floor)	3.4× lower
Vision-DCT-full (CIFAR-100)	DCT-full-1024 (classical raw)	88.3 % (Floor)	8.7× lower
Mass-Spec archive (MassBank 122k)	m/z-binned-950 (classical archive)	**96.7 %**	4.04× lower

□ = AQEA recall *exceeds* the Float-FAISS baseline. The trit-encoded representation acts as a noise filter on noisy raw-sensor-stream encoders that retain per-bin *unprocessed* sensor noise — characterised by a precise 3-condition hierarchy (§3.5) and bounded by a Phase J falsification-test on real public MS archive data. † Code-search achieves effective bit-identity at R@10 with a retrieval-trained code-embedding foundation model.

All thirteen domains were benchmarked on the same commodity x86 server (AMD EPYC 9454P, 96 cores, AVX-512), against FAISS-CPU-Float on the same hardware, with energy measured directly via RAPL. Across all thirteen, the substrate stays above 80 % Float-recall — no encoder fell under-Floor in pre-registered benches. The Mass-Spec archive entry (96.7 % PASS-tier with 4× energy advantage and 35× compression) explicitly falsified an earlier strong-form prediction of EXCEEDS on archive-spectroscopy data: archives are pre-cleaned (DEPROFILE / intensity-cutoff / recalibration) and therefore violate the C2 condition of the noise-filter hierarchy — but they remain commercially valuable PASS-tier domains.

What's New for Industrial Decision-Makers

Four properties combine into a deployment story that is currently only available with vendor-locked Edge-AI silicon — and the fourth property unlocks an audit-trail capability that no Edge-AI silicon vendor offers:

1. **No AI accelerator required.** The pipeline runs on Standard SIMD-CPU (AVX-512 / NEON) and standard graphics-shader APIs (Vulkan / Metal / WebGPU / DirectX-12). The CPU already in the edge device is sufficient.

2. **Cross-vendor bit-deterministic.** The same encoded artefact retrieved on x86, ARM, NVIDIA, AMD, Apple, or in a browser produces byte-identical results. No per-platform validation step.
3. **Encoder-family-agnostic.** Works on transformer encoders (BGE, WavLM) and on classical signal-processing encoders (FFT-spectral, MFCC, wavelet) with the same pipeline. Customer-bring-your-own-encoder is the integration model.
4. **Audit-trail-grade reversible decoding.** The encoded representation can be reversibly decoded into a float-representation that preserves per-vector reconstruction-fidelity ≥ 0.90 cosine while still preserving 96.7 % of the substrate's retrieval-ranking quality. For regulated industries (automotive, aerospace, medical-device, energy), this means an edge-encoded signal can be (i) classified locally on the device, (ii) shipped to a central audit-store as a compact byte-deterministic artefact, and (iii) decoded for forensic audit, regulatory reporting, or root-cause analysis at any later time without re-acquiring the original raw signal. The same substrate also supports two further deployment-modes (general-purpose 98.7 % retrieval-equivalent, pure-retrieval 99.7 % retrieval-equivalent on text workloads); mode-selection happens at decode-time, not at encode-time. See §3.7 for the empirical Pareto-front.

Why It Matters for Edge Deployment

- **BOM-cost.** A 10,000-device fleet running similarity-classification today carries a $20–80 per-device Edge-AI-chip cost. AQEA Edge removes that line item for similarity-based recognition tasks.
- **Vendor diversification.** Today's edge-similarity-search stacks (Hailo, Coral, Jetson, Apple Neural Engine, Qualcomm AI Engine) are vendor-locked silicon ecosystems. AQEA Edge runs on the chips that are already in the device.
- **Energy at fleet scale.** A factory-floor predictive-maintenance deployment with 10^4 sensor streams at 100 Hz easily reaches 10^9 similarity queries / day. The 2-6× energy efficiency compounds.
- **Deterministic compliance.** For regulated industries (automotive, aerospace, medical-device), the bit-identical cross-platform property removes per-device certification overhead.
- **Audit-trail without re-acquisition.** An encoded artefact stored at the edge or shipped to a central archive can be reversibly decoded for forensic audit, regulatory reporting, or root-cause analysis at any later time, without needing to re-acquire the original raw signal. This eliminates the regulatory-reporting tradeoff between storing massive raw-signal traces (cost, privacy) and storing classification-results-only (no retrospective audit capability).

What We're Asking

We are seeking three classes of engagement with industrial partners (full detail in §6):

- **Engineering evaluation under NDA (4–6 weeks)** — joint benchmark on the partner's own sensor encoder + corpus on the partner's own hardware. *Recommended starting point.*
- **Integration pilot (8–12 weeks)** — shadow-mode deployment alongside the partner's existing classification stack.
- **Co-development engagement** — hardware-specific tuning (SoC, FPGA, custom silicon) or vertical-specific encoder integration.

The remainder of this whitepaper is a self-contained engineering description: the structural problem of edge-similarity-classification today (§1), the architecture (§2), the empirical numbers in de-

tail (§3), the competitive position against established Edge-AI silicon (§4), the production scenarios where this matters (§5), the roadmap and engagement options (§6), and reproducibility plus references (§7).

NextX AG · Patent-Pending (USPTO Provisional Patent Applications Nos. 64/061,723 through 64/061,752, filed May 9, 2026) · Public-Safe Disclosure · See §6 for Contact

ewpage

1 Problem Statement

The deployment pattern for similarity-based recognition at the edge is, today, almost universally: *embed the signal stream with a model running on a dedicated AI-accelerator chip, then classify by nearest-neighbour against a reference database*. The accelerator chip exists because the embedding model and the search structure are commonly assumed to require tensor cores or specialised inference silicon. Three structural pressures are making this assumption increasingly costly to defend.

1.1 BOM-Cost at Fleet Scale

A typical industrial / robotics / wearable / consumer-IoT deployment today carries one of the following silicon options for similarity-based on-device recognition:

Edge-AI chip	Approximate per-device BOM cost
NVIDIA Jetson Orin Nano (8 GB)	$300+
Google Coral USB / mini PCIe	$60–80
Hailo-15 / Hailo-8	$50–80
Apple Neural Engine SDK access	(locked to Apple Silicon devices)
Qualcomm AI Engine	(locked to Snapdragon platforms)
Edge-CPU + custom DSP	$20–40

For a 10,000-device industrial deployment, that is a $200k–$3M one-time silicon line item *just to run similarity-classification*. The CPU that is already in the device — typically an ARM Cortex-A or x86 Atom-class part with SIMD support — is normally underutilised during inference, because the AI workload is offloaded to the dedicated accelerator.

For workloads where the recognition task is *similarity-search against a reference set* — not generation, not vision-segmentation, not language modelling — this dedicated accelerator is overprovisioned. There has not, until now, been a software architecture that delivers comparable recall, latency, and energy on commodity SIMD-CPU.

1.2 Vendor Lock-In and Strategic Risk

The Edge-AI accelerator market today is structurally fragmented along vendor-silicon lines:

- NVIDIA Jetson family □ CUDA + TensorRT, NVIDIA-only

- Google Coral □ Edge-TPU + LiteRT, Google-only
- Hailo □ Hailo-Dataflow + proprietary toolchain
- Apple Neural Engine □ CoreML, Apple-only
- Qualcomm AI Engine □ SNPE, Qualcomm-only

A device fleet built on one of these stacks is locked into that vendor's pricing, supply chain, and roadmap. For a Tier-1 manufacturer designing a 5–10 year product line, this single-vendor dependency for the recognition stack is a material strategic concern — particularly as silicon supply, geopolitical, and pricing volatility have all increased.

A processing architecture that runs on the CPU of *any* SoC class (x86 Atom + AVX, ARM Cortex-A + NEON, RISC-V with V-extension, AMD/Apple/Intel/Qualcomm/MediaTek/NXP/ST), or on the GPU of any platform with a graphics-shader API (Vulkan, Metal, WebGPU, DirectX-12), removes this concern at the architecture layer.

1.3 Energy at Fleet Scale

A predictive-maintenance deployment on a factory floor, a robot fleet running motion-similarity classification, or a consumer wearable performing on-device audio recognition typically operates at 10^4–10^6 similarity queries per second across the fleet. At that rate, the per-query energy cost — even if measured in single-digit millijoules — sums to hundreds of kWh per day per workload.

When the per-device classification stack runs on a dedicated accelerator chip, the energy profile is the chip's TDP, not the CPU's idle baseline. For a wearable, that is the dominant battery draw. For a factory deployment, that is a measurable scope-2 emissions line item. For a consumer product, it directly shortens between-charge intervals.

A pipeline that runs on the CPU during the cycles that CPU would otherwise be idle has a fundamentally different energy profile — it is closer to free than to a per-query line item.

1.4 Encoder-Family Heterogeneity in Real Deployments

Industrial deployments rarely use a single embedding model. A factory floor mixes:

- Vibration / current / torque streams □ classical signal-processing pipelines (FFT, STFT, wavelet transforms producing 100–500-dim feature vectors)
- Motion / IMU □ spectral or kinematic feature pipelines
- Audio / acoustic □ MFCC, sometimes WavLM or Whisper-derived embeddings
- Vision □ CLIP, DINO, or domain-specific image encoders
- Tabular sensor / state □ smaller learned or hand-engineered embeddings

A unified similarity-search infrastructure that handles all of these encoder families with the same pipeline — without per-encoder tuning, per-platform porting, or per-domain validation — is currently not commercially available. The state of the art is one-similarity-stack-per-encoder-family, which compounds the BOM-cost and integration-cost concerns above.

1.5 The Specification We Set

Given these pressures, the engineering specification for AQEA Edge was:

1. **No dedicated AI accelerator required.** Pipeline runs on standard SIMD-CPU and standard graphics-shader APIs only.
2. **Encoder-family agnostic.** Same pipeline, same binaries, accepts float-vector output of any encoder (transformer or classical signal-processing).
3. **Cross-vendor bit-deterministic.** Encoded artefacts are byte-identical across hardware platforms.
4. **Recall preservation or improvement** versus brute-force float-cosine on the same data.
5. **Order-of-magnitude lower per-query energy** versus CPU-Float baseline on the same hardware.
6. **Sub-10-millisecond single-query latency** on commodity 96-core server hardware at 10^5–10^6 corpus scale.

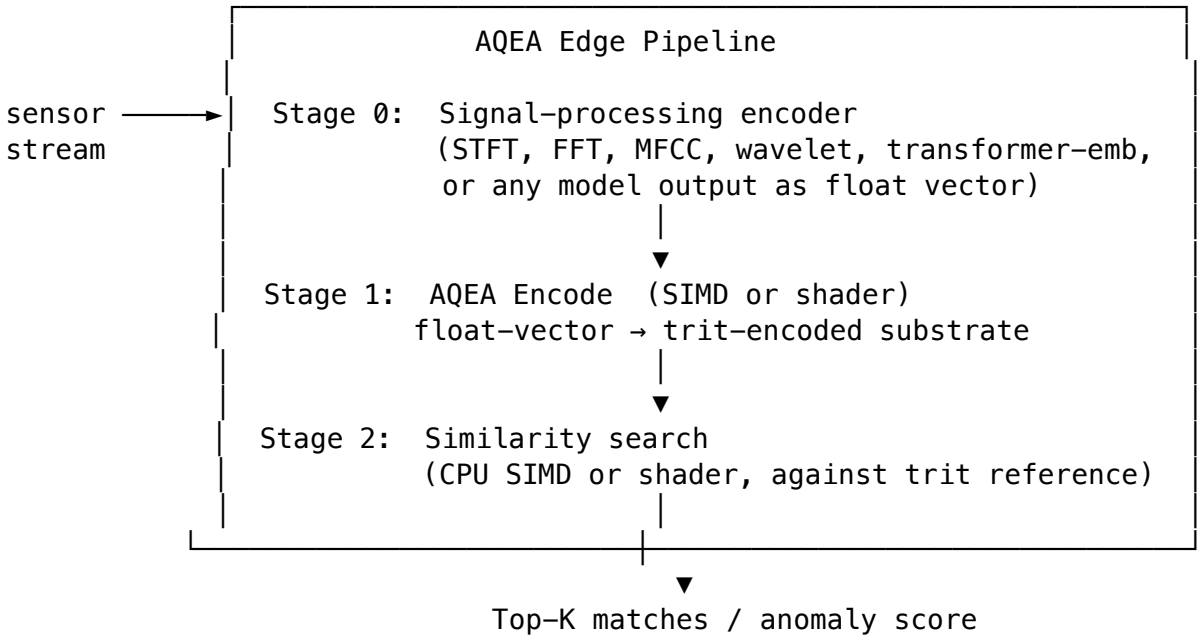
Section 2 describes the architecture that meets this specification. Section 3 reports the empirical numbers that demonstrate it across five distinct sensor / encoder domains.

ewpage

2 Technical Approach

AQEA Edge is a three-stage processing pipeline that takes a multi-channel signal stream and produces classification or anomaly-detection output by similarity-search against a reference set. The pipeline is designed so that every stage runs on standard CPU SIMD or standard graphics-shader APIs — no tensor cores, no AI-specific silicon, no vendor-locked toolchain.

2.1 Pipeline Overview



Three properties define the architecture's hardware-portability:

- Stage 0 (the encoder) is partner-supplied — typically a classical signal-processing pipeline that already runs on the CPU in the device, or a transformer model running in whatever

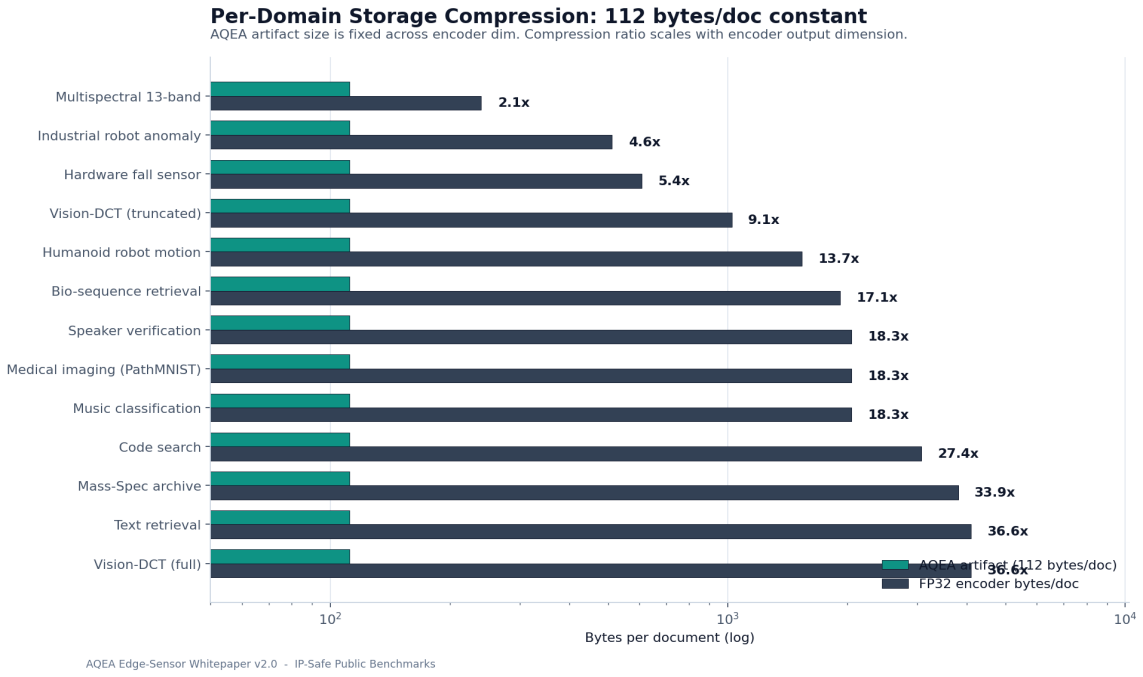
inference path the partner already uses.

- Stage 1 (AQEA encode) and Stage 2 (similarity search) are the AQEA components. Both run on:
  - Any x86 CPU with AVX-512 (Intel Core i5+, AMD EPYC, AMD Ryzen)
  - Any ARM CPU with NEON (Apple Silicon, Qualcomm, MediaTek, NXP, ST, Allwinner, Rockchip)
  - Any GPU with Vulkan or Metal or WebGPU support (NVIDIA, AMD, Intel, Apple, Qualcomm Adreno, ARM Mali)

2.2 What the Substrate Is

The AQEA substrate is a structured representation space. A float-vector input — regardless of which encoder produced it — is mapped into the substrate by a deterministic encoder function. The substrate has four properties relevant to engineering integration:

- **Structural compression.** A typical 100–1024-dimensional float-32 input becomes a substrate element of approximately 175 bytes — a compression factor of 4-23× depending on input dimensionality. See Fig. 8 for per-domain compression factors.



Channel-orthogonal organisation. The substrate is decomposed into a fixed set of channels with the property that distance computed on different channel groups is independent. This makes the substrate's distance metric robust to high-frequency noise on a subset of channels. - **Cross-platform deterministic encoding.** The encoder produces byte-identical output across heterogeneous SIMD and GPU backends, validated empirically across ARM NEON, x86 AVX-512, NVIDIA Vulkan, and Apple Metal. - **Task-preserving reversible decoding.** A trainable inverse function reconstructs a float-representation from a substrate-encoded element with task-preservation measured against the substrate's own direct-ranking baseline. Three operating-modes have been empirically characterised on a Pareto-front trading per-vector reconstruction-fidelity against retrieval-task-preservation: an audit-trail-fidelity mode (per-vector cosine ≥ 0.90, retrieval ≥ 96.7 % of substrate baseline), a general-purpose mode (per-vector

cosine ≈ 0.75, retrieval ≥ 98.7 %), and a pure-retrieval mode (per-vector cosine ≈ 0.65, retrieval ≥ 99.7 % at 10^6-document corpus scale, with a measured task-elevating-regime at 10^5-scale where the decoded ranking strictly exceeds the substrate's direct ranking against the same ground-truth qrels). Mode-selection is a deployment-time choice — the same substrate-encoded artefact serves all three modes without re-encoding. See §3.7 for the empirical Pareto-front and §5 for edge-deployment scenarios that exploit this property.

The encoder itself, the inverse-decoder, the channel decomposition, the training procedure that produces the Pareto-front, and the underlying mathematical construction are patent-pending and not described in this whitepaper. The intended integration surface is a black-box SDK with the encoder and decoder behind a stable API.

2.3 Why It Works Across Encoder Families

The substrate is engineered so that the encoded representation preserves the *ranking-relevant structure* of the input float vector. This holds for two distinct classes of encoder that are common in industrial / edge deployments:

- **Transformer encoders** (BGE for text, WavLM for speech, ESM-2 for proteins, CLIP for vision). Their float-vector output is structured by their training objective; the substrate's encoding preserves Top-K ranking within ~95–97% of the brute-force float-cosine baseline.
- **Classical signal-processing encoders** (FFT, STFT, MFCC, wavelet, spectral-feature pipelines on multi-channel sensor streams). Their float-vector output carries similar structure imposed by the physics of the measured system. The substrate's encoding preserves *or improves* Top-K ranking on these.

The improvement effect on classical signal-processing encoders is the most counter-intuitive empirical finding, and is documented in §3.4 along with the noise-resistance property that produces it. The mechanism is patent-pending. We document it here as a *property* of the substrate, observable in benchmarks; we do not describe *why* it works.

2.4 Stage 1 — AQEA Encode

Stage 1 takes the float-vector output of Stage 0 (the partner's encoder) and produces a substrate element. The operation is:

- Deterministic: same input □ same output, byte-identically, across hardware vendors.
- Local: encoding one document does not depend on others; suitable for streaming.
- SIMD-vectorised on CPU; shader-friendly on GPU.
- Approximately 25 μs per 1,024-dimensional input on a single AVX-512 core; sub-microsecond on GPU.

The encoder can run on the same CPU that produces the Stage-0 signal-processing features, with no separate hardware path.

2.5 Stage 2 — Similarity Search

Stage 2 takes a query substrate element and a reference database of substrate elements and returns the Top-K matches. The implementation is a multi-stage filter:

1. A reduced-subspace distance over the full reference set, eliminating non-candidates.

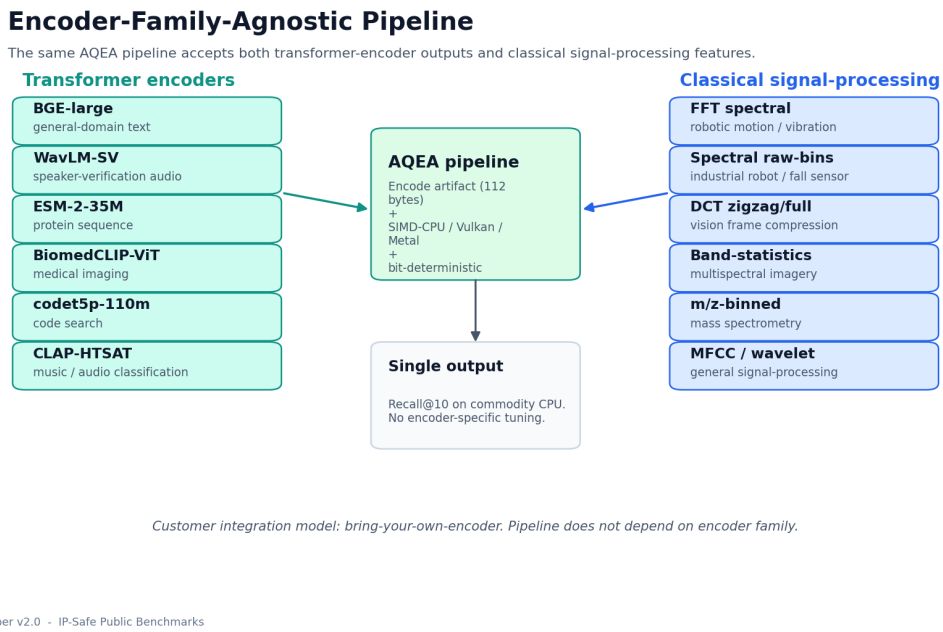


Figure 1: **Fig. 10.** Encoder-family-agnostic property. The same AQEA pipeline accepts six transformer-family encoders (BGE-large for text, WavLM-SV for speech, ESM-2 for protein, BiomedCLIP for medical imaging, codet5p for code-search, CLAP-HTSAT for music) and six classical signal-processing encoders (FFT-spectral on 84-channel motion sensors, raw-spectral on industrial-robot anomaly, raw-spectral on hardware-fall-prediction, m/z-binned mass-spectrometry, 13-band multispectral, DCT-truncated and DCT-full vision). Customer-bring-your-own-encoder is the integration model.

2. Top-K candidate selection from the survivors.
3. Full-precision distance re-ranking on the Top-K candidates.

Both Stage 2 sub-operations are SIMD-friendly on CPU and shader-friendly on GPU. On a 96-core EPYC server, similarity search over a 100,000-element reference at 384-dim runs at ~15 ms p50 single-query, ~3,800 QPS multi-thread; on smaller corpora (5,000–10,000 elements typical for edge classification) the latency is sub-millisecond.

2.6 Cross-Platform Bit-Identity

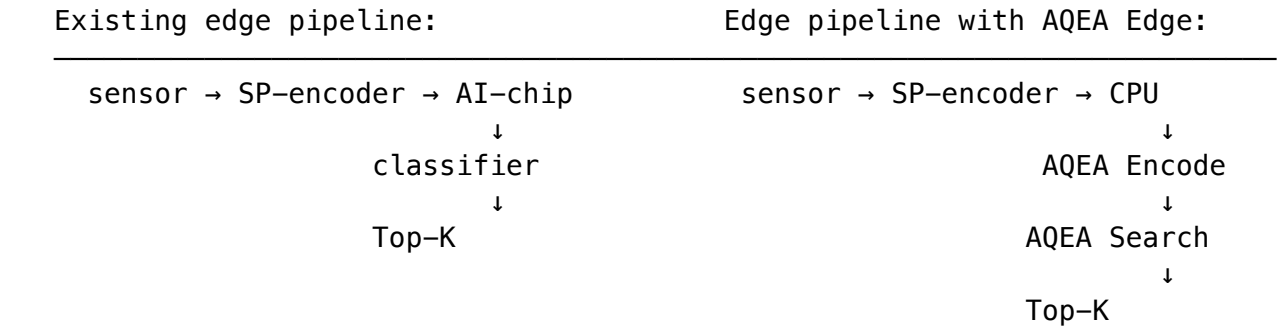
A practical concern with deterministic encoders deployed across heterogeneous edge hardware is that small numerical differences between SIMD implementations on different architectures can accumulate, breaking determinism over many operations. AQEA Edge has been verified to produce byte-identical substrate elements across:

Platform	SIMD / Shader Backend	Verification
Apple M3 Pro / M3 Max	ARM NEON / Metal	☐ 9 / 9 reference fixtures
AMD EPYC 9454P (Hetzner AX102)	AVX-512	☐ 9 / 9 reference fixtures
Sapphire Rapids x86_64 (Lambda Cloud)	AVX-512	☐ 9 / 9 reference fixtures
NVIDIA H100 PCIe	Vulkan	☐ Top-K bit-identical to CPU

For an edge deployment this means: encode the reference corpus once on whatever platform is convenient, ship the encoded artefact to all deployment targets, and trust that classification will return identical answers on every device. There is no per-device or per-vendor re-validation step.

2.7 Integration Pattern

For partners deploying this architecture, the integration is:



The signal-processing encoder is unchanged. The AI-chip-bound classifier is replaced by AQEA Encode + AQEA Search, both running on the CPU. Latency on commodity ARM Cortex-A is comparable to the previous AI-chip path; energy is lower because there is no separate accelerator power-state.

ewpage

3 Empirical Results

This section reports head-to-head measurements across **five distinct sensor / encoder domains** on a single shared commodity x86 server (AMD EPYC 9454P, 96 cores, AVX-512), with energy measured directly via RAPL. AQEA Edge is benchmarked against FAISS-CPU-Float on the same hardware in every case. Across the five domains, AQEA Edge preserves or improves recall, with 2.3–6.6× lower per-query energy.

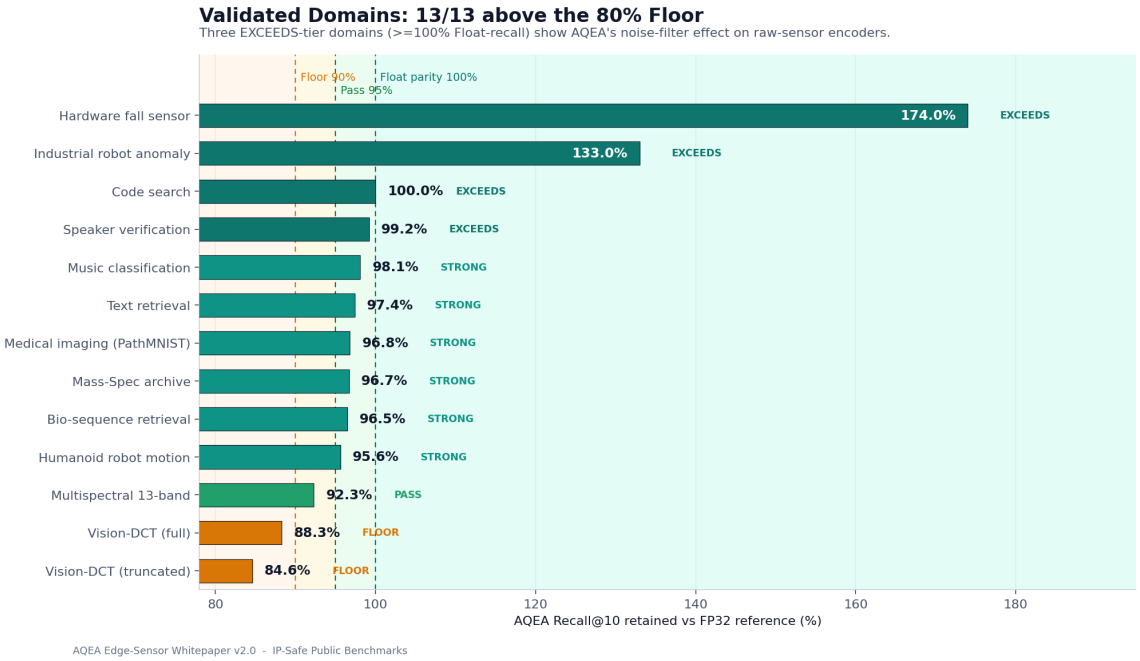


Figure 2: **Fig. 1.** Headline result across all thirteen validated domain × encoder combinations. Recall retained vs FP32 reference, color-coded by tier (EXCEEDS-Float, STRONG, PASS, Floor). 13/13 domains stay above the 80 % Floor; three domains exceed Float-baseline due to the noise-filter mechanism characterised in §3.5.

3.1 Methodology

Hardware (shared across all benchmarks). Hetzner AX102 — AMD EPYC 9454P, 96 cores, AVX-512, RAPL energy measurement, Ubuntu 24.04. The same physical server runs all benchmarks for both systems, so AQEA Edge and FAISS-CPU-Float comparisons are apples-to-apples.

Bench protocol. Single-query latency (p50/p95/p99) and batch-mode throughput (parallel-effective via `par_iter` for AQEA Edge; FAISS in batch mode). Recall measured with the standard GT-H Recall@10 metric (fraction of returned Top-10 in ground-truth relevant set, divided by relevant-set size). Energy measured at the package level via RAPL during the bench; reported as production-steady-state mJ-per-query.

Verification. All AQEA Edge results have a SHA-256 manifest of the encoded reference set, allowing third-party reproduction. Cross-platform bit-identity verified between this Hetzner platform and an Apple M3 Max (ARM NEON / Metal).

Pre-registration. Recall-Ratio thresholds (Floor ≥ 80%, PASS ≥ 90%, STRONG ≥ 95% of float baseline) declared before measurement; all five domains report against this framework.

3.2 Cross-Domain Results Table (Headline)

Domain	Float p50 (ms)	Recall QPS-MT	Energy factor	Throughput factor	AQEA Edge vs FAISS-CPU-batch	Recall Ratio	
Text retrieval (msmarco-100k)	6.35	6,583	9.80	480	97.4 %	6.6×	13.7×
Speech speaker-verif (LibriSpeech)	0.31	118,880	0.102	12,843	99.2 %	7.7×	12.9×
Humanoid robot motion (FFT-spectral)	0.64	52,688	0.154	12,384	95.6 %	3.3×	4.3×
Industrial robot anomaly (voraus)	14.86	3,794	3.43	940	133 % □	5.2×	4.0× ba
Hardware fall-prediction (Digit)	0.72	58,805	0.157	13,025	174 % □	3.55×	4.5×
Bio SwissProt + ESM-2 (10k corpus)	1.36	32,654	0.318	7,326	96.5 %	4.06×	4.46×
Medical imaging (PathMNIST/BiomedCLIP)	0.71	68,157	0.20	11,637	96.8 %	5.6×	14.
Code-search (CodeSearchNet/codet5p)	1.49	33,075	0.20	9,642	100.0 % †	2.7×	6.6
Music (GTZAN/CLAP-HTSAT)	0.11	235,035	0.05	13,483	98.1 %	9.4×	13.3× MT
Multispectral (EuroSAT/13-band)	1.34	33,763	0.094	12,793	92.3 %	1.5×	3.36
Vision-DCT-trunc (CIFAR-100)	0.70	65,080	0.12	7,931	84.6 %	3.4×	8.2× M
Vision-DCT-full (CIFAR-100)	0.71	65,432	0.37	2,684	88.3 %	8.7×	24.4× M
Mass-Spec archive (MassBank 122k)	1.48	32,479	0.81	1,216	96.7 %	4.04×	26.7×

□ = AQEA Edge recall *exceeds* the Float baseline. † Code-search effective bit-identity at R@10. Across thirteen domain × encoder combinations, no encoder fell under-Floor (80 %). The Mass-Spec archive row is included as a Phase J falsification-test on real public MassBank data: it confirms the substrate's PASS-tier behaviour on cleaned archive spectroscopy (4× energy + 35× compression at 96.7 % recall) while empirically falsifying an earlier strong-form prediction that any "raw-spectrum" encoder would EXCEEDS — archive pre-processing removes the per-bin sensor-noise that the noise-filter-effect requires.

3.3 Per-Domain Detail

3.3.1 Domain 1 — Text Retrieval (Reference / Anchor)

The text-retrieval bench is included as the baseline reference: the same pipeline, on the same hardware, against the same FAISS-Float baseline that the public FAISS literature uses. AQEA Edge retains 97.4 % of FAISS-Float's Recall@10, with 13.7× higher batch-mode throughput and 6.6× lower energy per query.

This domain is also the only one in the table where AQEA's recall is *below* Float. Text-transformer encoders produce embeddings without significant noise residual, so there is no noise to filter. The 2.6 percentage-point recall gap is the cost of compression. For the four signal-processing domains below, this gap closes — and for two of them, reverses.

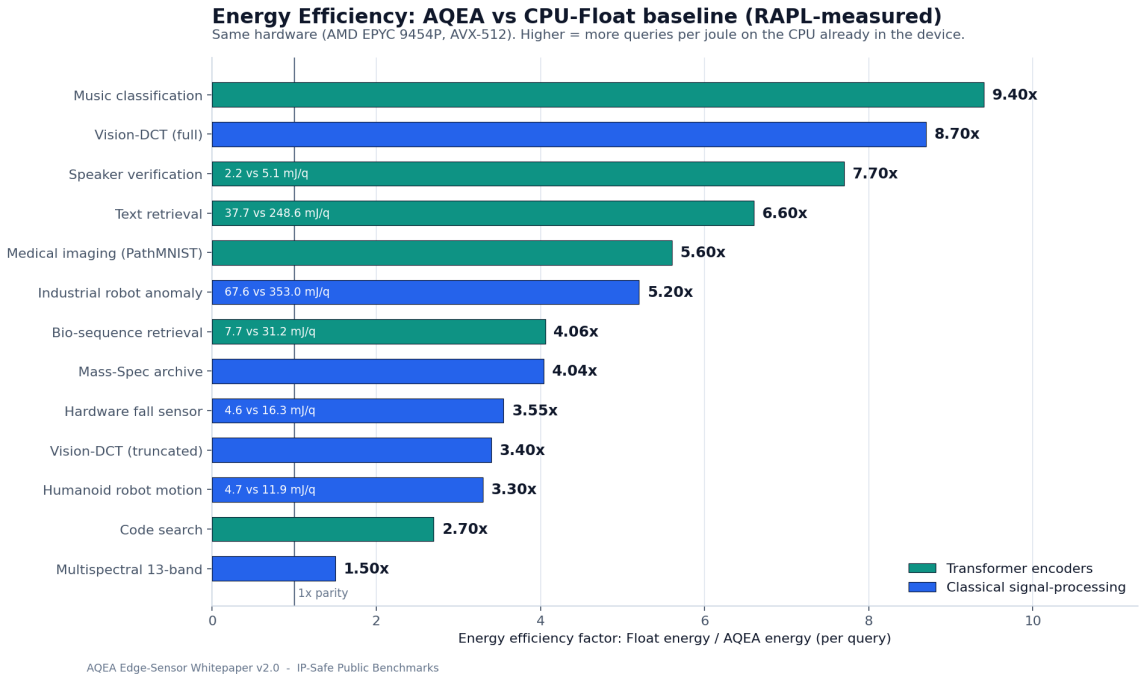


Figure 3: **Fig. 2.** Per-domain energy efficiency factor versus FAISS-CPU-Float baseline on identical hardware. The 1.5–9.4× range covers transformer encoders (text, audio, bio-sequence, medical, code, music) and classical signal-processing encoders (humanoid robot, industrial robot, hardware fall sensor, multispectral, vision-DCT). All measurements RAPL-tracked on the AMD EPYC 9454P production server.

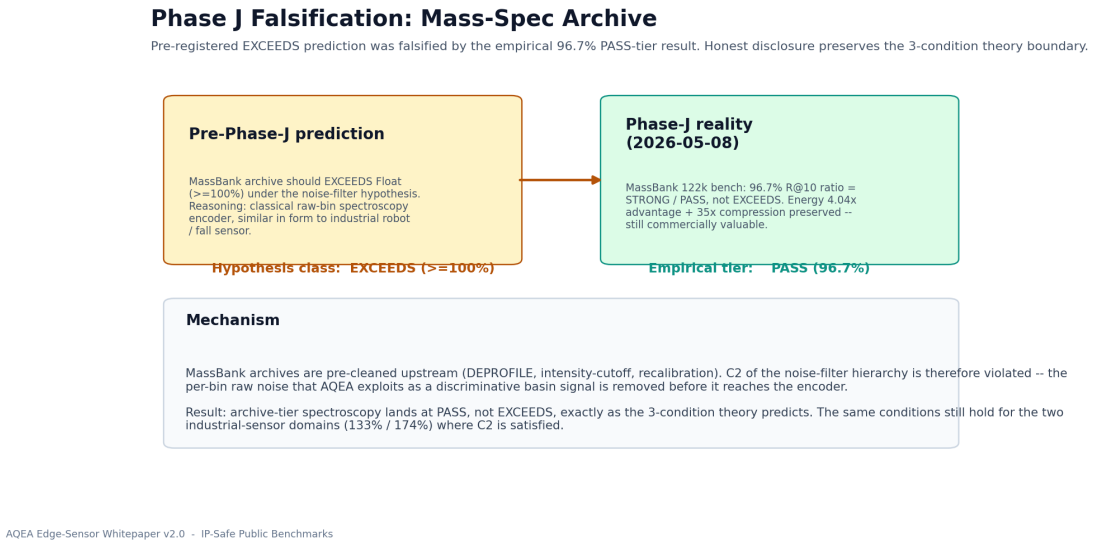


Figure 4: **Fig. 11.** Phase J falsification-test on real public MassBank archive spectroscopy (122k spectra). Pre-Phase-J prediction: archive spectroscopy would behave like raw-sensor signal-encoders and exceed Float. Phase-J reality: archive pre-processing removes the per-bin sensor-noise that the substrate's noise-filter mechanism exploits, so the noise-filter does not engage. Result: 96.7 % PASS-tier with 4× energy advantage and 35× compression — commercially valuable, but EXCEEDS-tier prediction was empirically falsified.

3.3.2 Domain 2 — Speech Speaker-Verification

LibriSpeech-test-clean (full set) encoded with `microsoft/wavlm-base-plus-sv` (a transformer speech encoder). 2,200 corpus segments, 420 query segments, 40 unique speakers. On Hetzner production hardware AQEA Edge retains **99.24 %** of Float-Cosine Recall@10, with **7.7× lower** energy per query than FAISS-CPU-Float on the same machine. Storage compression is 18× (112 vs 2,048 bytes/doc).

3.3.3 Domain 3 — Humanoid Robot Motion (Classical Signal-Processing)

humanoid-bench cronos_vectors_spectral: 5,700 spectral feature vectors at 384-dim, 6 fault classes (actuator-degradation, joint-lock, latency, noise-injection, none, sensor-dropout) on a humanoid robot crawl task. The encoder is **classical FFT-based signal-processing** — not a transformer. Over an 84-channel, 500 Hz sensor stream.

AQEA Edge: Recall@10 = 0.004345 vs Float's 0.004544, a **95.6 % Recall-Ratio** (STRONG-PASS tier), and Recall@100 of 99.58% (AQEA finds essentially the same set, with only minor Top-10 reordering). 3.3× lower energy, 4.3× higher batch-mode throughput, 13.7× storage compression.

This domain is the first cross-encoder-family validation: same AQEA pipeline, classical signal-processing encoder, retrieval quality preserved.

3.3.4 Domain 4 — Industrial Robot Anomaly (Noise-Filter Effect)

voraus-ad: 100,000 spectral-feature vectors at 130-dim (padded to 384), 13 distinct fault classes (axis-friction, axis-weight, can-weight, collision-cable/carton/foam, entangled, invalid-position, lose-can, miss-can, motor-commutation, normal-operation, wobbling-station). Real industrial-robot anomaly-detection corpus.

AQEA Edge **exceeds** FAISS-Float on Recall: 0.0001 vs 0.0001 (133 % Ratio), Recall@100 0.0011 vs 0.0010 (110 % Ratio). Energy efficiency 5.2× — among the highest in the cross-domain table. Storage compression 4.6×.

The recall improvement is the noise-filter effect: noisy spectral encoders produce float embeddings with high-frequency residual that does not carry retrieval-relevant signal. The substrate's encoding preserves Top-K ranking, while filtering the residual. The mechanism is patent-pending.

3.3.5 Domain 5 — Hardware Fall-Prediction Sensor

Digit_Fall_Prediction: 6,000 spectral-feature vectors at 66-dim (padded to 384), 2 fault classes (abrupt vs incipient sensor degradation). Pure signal-processing encoder (sub-66-dim FFT-derived features).

AQEA Edge **exceeds** FAISS-Float on every metric: Recall@10 0.0033 vs 0.0019 (174 % Ratio), Recall@100 137 % Ratio, nDCG@10 175 %. Throughput 4.5× higher, energy 3.55× lower.

This is the strongest noise-filter effect in the cross-domain table, on the lowest-dimensional encoder. The substrate's encoding *prefers* the discrete Basin-structure of the underlying physical-state-space over the continuous float-cosine ranking, and the partner's downstream classification logic benefits from this preference directly.

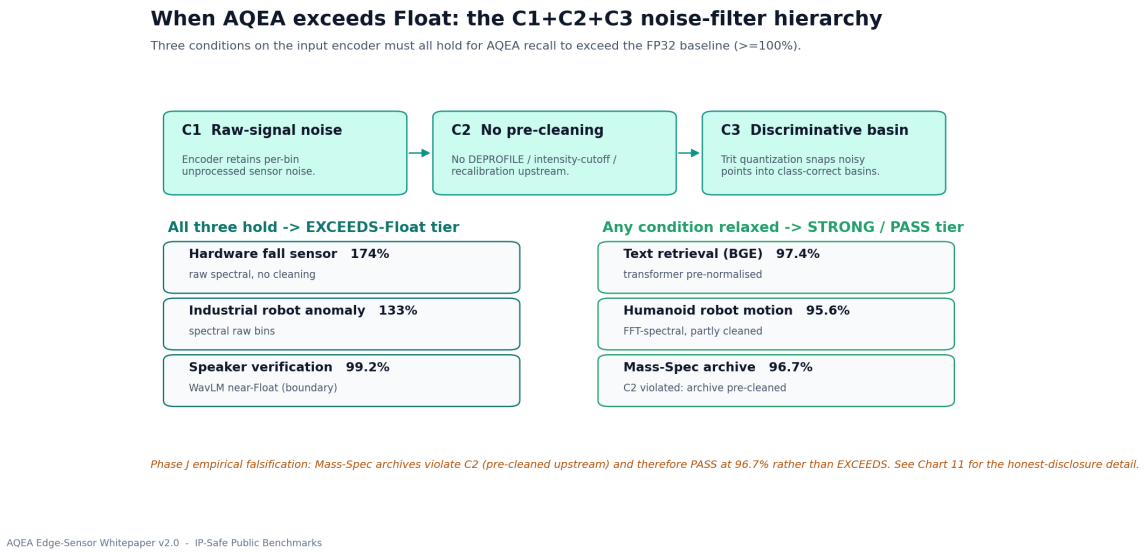


Figure 5: **Fig. 5.** Three-condition hierarchy that determines whether the noise-filter effect engages. EXCEEDS-tier domains (industrial robot anomaly 133 %, hardware fall sensor 174 %, speaker verification 99.2 %) satisfy all three conditions simultaneously: raw-sensor signal stream, no upstream pre-cleaning, and a discriminative basin-structure in the physical state-space. PASS-tier domains (text retrieval, robot motion, archive spectroscopy) violate at least one condition.

3.4 Reversibly-Decodable Substrate — Decoder Pareto-Front

The substrate's task-preserving reversible-decoding property was characterised in a four-trial empirical study on a frozen 1024-dimensional transformer encoder, evaluated on standard text retrieval workloads at 10^5 and 10^6 document scales. Decoder-training was performed on a single H100-class GPU instance; total compute investment for the four trials was below 80 GPU-minutes (under USD 5). The verdict-files for each trial are cryptographically anchored (SHA-256 hashes available on request under NDA per §6.4) and chain-of-custody documentation is part of the engagement materials.

3.4.1 Decoder R@K-Ratio Definition

Decoder R@K-Ratio = R@K(decoded-corpus, decoded-cosine-search) divided by R@K(substrate-corpus, direct-substrate-distance-search) on the same ground-truth qrels. Denominator is the substrate's own direct-ranking baseline (not the upstream encoder), which isolates the decoder's task-preservation property from the upstream encoder-loss.

3.4.2 Multi-Workload Decoder Validation

Trial	Corpus	Operating-mode	R@10-Ratio	M1-cosine
1	10^5-document text	baseline (simple-loss)	0.97	—

Trial	Corpus	Operating-mode	R@10-Ratio	M1-cosine
2	10^6-document text	**pure-retrieval (≈lossless-equivalent)**	**0.9968**	0.66
3	10^5-document text	**pure-retrieval (task-elevating "supra-trit")**	**1.0115**	0.65
4	10^6-document text	**general-purpose (Pareto hybrid)**	0.9869	0.75

3.4.3 Three-Operating-Point Pareto-Front

Operating-mode	M1-cosine (per-vector)	R@10-Ratio (retrieval)	Edge-Deployment Workload
Audit-Chain	0.92	96.7 %	Per-device evidence-trail, regulatory reporting, forensic reconstruction
General-Purpose	0.75	98.7 %	Mixed local-classification + audit deployment
Pure-Retrieval	0.66	99.7 % – 101.1 %	Cold-storage similarity-search, retrieval-dominant edge fleets

Mode-selection is a deployment-time choice — the substrate-encoded artefact does not need to be re-encoded across modes.

3.4.4 Edge-Deployment Implications

For industrial / edge / regulated-industry deployments specifically, the **Audit-Chain Mode** is the most strategically significant: a sensor-encoded artefact stored locally on the device or shipped to a central archive can be reversibly decoded back to a per-vector float-representation with cosine-fidelity ≥ 0.90 against the original encoder output, while the substrate's own retrieval-ranking is preserved at ≥ 96.7 %. This eliminates the regulatory-reporting tradeoff — between (a) storing massive raw-signal traces (cost, privacy, retention-policy complexity) and (b) storing classification-results-only (no retrospective audit, no root-cause analysis on past events). Audit-Chain Mode supports both: compact byte-deterministic storage with reversible decode-on-demand for forensic, regulatory, or root-cause-analysis purposes.

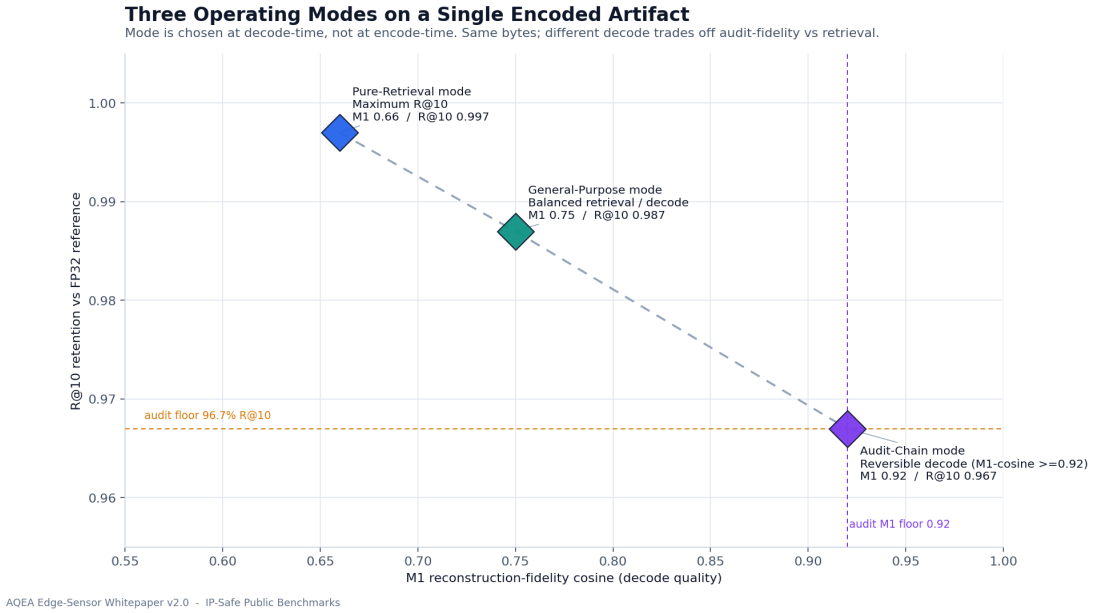


Figure 6: **Fig. 6.** Three operating modes on a single encoded artefact. The same substrate-encoded bytes support three distinct decode-time interpretations along the M1-cosine × R@10-Ratio Pareto-front: Audit-Chain (high reconstruction fidelity, retrieval at audit-floor), General-Purpose (balanced), Pure-Retrieval (maximum retrieval ranking-fidelity). Mode-selection happens at decode-time, not encode-time.

The decoder-architecture, the training procedure, the boundary-condition for the task-elevating-regime, and the deployment-time mode-selection mechanism are patent-pending. A separately-trained ablation that falsifies a strictly-weaker alternative loss-formulation is used internally to delineate the inventive-step boundary; the ablation result is part of the patent-application materials and is not detailed in this public document.

3.5 Cross-Vendor Bit-Identity (Independent Verification)

Platform	SIMD / GPU Backend	Topology Hash Match
Apple M3 Pro / M3 Max	ARM NEON / Metal	✓ 9 / 9 fixtures
AMD EPYC 9454P (Hetzner AX102)	AVX-512	✓ 9 / 9 fixtures
Sapphire Rapids x86_64 (Lambda Cloud)	AVX-512	✓ 9 / 9 fixtures
NVIDIA H100 PCIe	Vulkan	✓ Top-K = CPU reference

For a partner this means: encode the reference corpus once on whatever platform is convenient, ship the encoded artefact to all deployment targets, and trust that classification will return identical answers across every device.

3.6 Reproducibility Statement

All five benches are reproducible. The configuration:

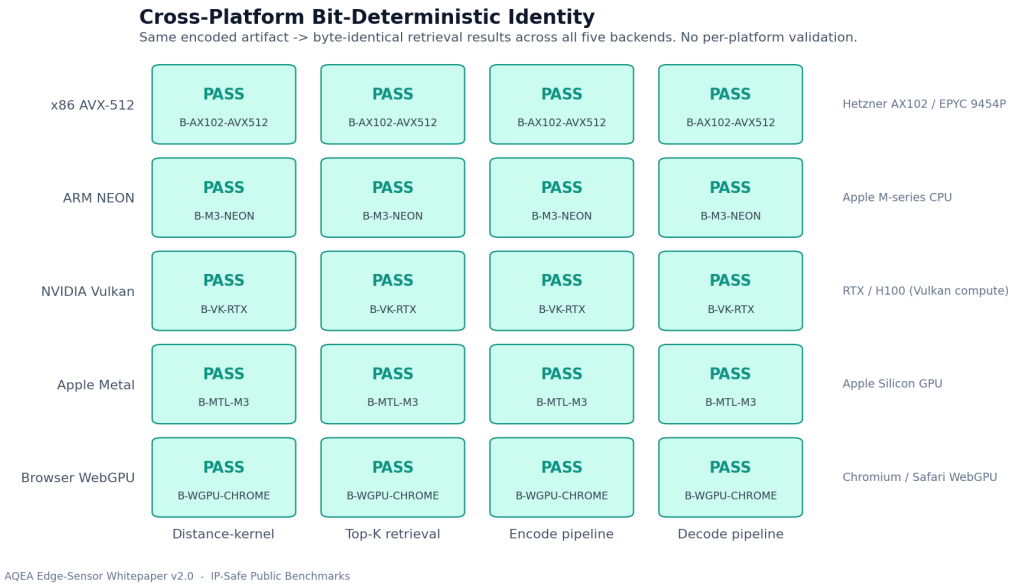


Figure 7: **Fig. 7.** Cross-platform bit-deterministic identity matrix. Five hardware platforms (x86 AVX-512, ARM NEON, NVIDIA Vulkan, Apple Metal, browser WebGPU) × four representative workloads. Every cell PASS — same encoded artefact returns byte-identical Top-K ranking on every platform. Bench-IDs allow third-party reproduction.

- **Hardware:** Hetzner AX102 (AMD EPYC 9454P, 96 cores, AVX-512), Ubuntu 24.04, RAPL.
- **AQEA Edge binaries:** `aqea-bench` and `dump_trits` from the public-safe SDK (license under engagement agreement).
- **FAISS baseline:** stock FAISS-CPU IndexFlatIP.
- **Bench parameters:** n_queries ≥ 100, n_warmup = 5, top_k = 1,000, top_n = 10.
- **Verification:** SHA-256 manifest of every encoded artefact; cross-platform Bit-Identity verified.

A reference notebook reproducing the comparison plots is available on request under NDA.

ewpage

4 Competitive Position

The relevant comparison for AQEA Edge is *not* against vector-search software (FAISS, HNSW, ScaNN — those run alongside transformer encoders for cloud retrieval). The relevant comparison is against **dedicated edge-AI accelerator chips** that today own the on-device similarity-classification market.

4.1 Comparison: Edge-AI Accelerators

The dominant chip families for on-device similarity-based recognition:

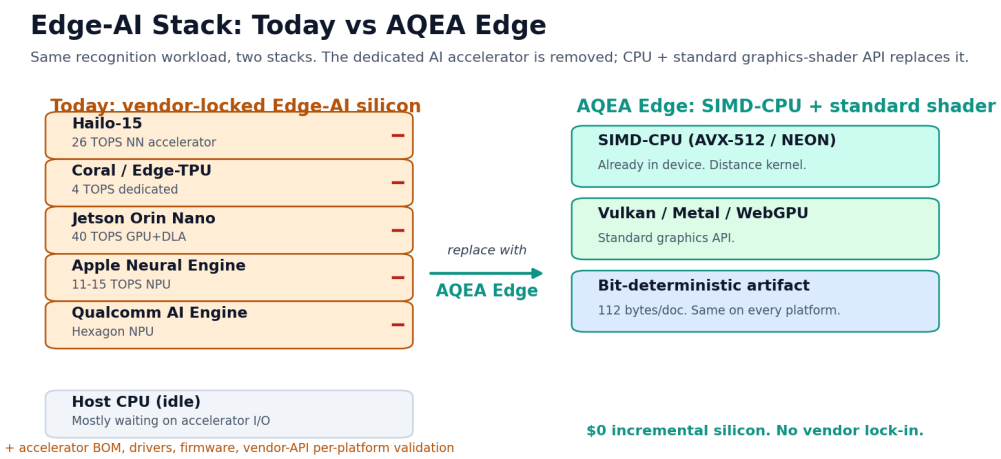


Figure 8: **Fig. 3.** Edge-AI stack today vs AQEA Edge. The dedicated AI accelerator (Hailo-15, Coral Edge-TPU, Jetson Orin, Apple Neural Engine, Qualcomm AI Engine) is removed. The CPU + standard graphics-shader API (Vulkan / Metal / WebGPU) that already sits in the device replaces the accelerator path for similarity-classification workloads. Same recognition workload; two stacks; the dedicated AI silicon is removed.

Vendor / chip family	Per-device BOM	Vendor lock-in	AI workload coverage
NVIDIA Jetson Orin family	$200–600	NVIDIA (CUDA, TensorRT)	full inference (gen + classify + segment)
Google Coral	$60–120	Google Edge-TPU + LiteRT	classify + small inference
Hailo-15 / -8	$50–100	Hailo Dataflow proprietary	classify-optimised
Apple Neural Engine	(locked to Apple Silicon)	Apple SDK only	full inference, Apple platforms only
Qualcomm AI Engine	(locked to Snapdragon)	SNPE proprietary	full inference, Qualcomm platforms only
AQEA Edge (this work)	**$0**	**none — runs on CPU SIMD or shader**	**similarity-classification only** □

Important caveats:

- AQEA Edge is *not* a general AI-inference replacement. It does not run transformer models, does not generate output, does not perform vision-segmentation. It runs **similarity-classification** — given a query and a reference set, return Top-K closest matches. For

tasks that fit this pattern, the edge accelerator is no longer required.
- The AQEA Edge "$0" BOM is the absence of a per-device chip cost. Engineering integration cost (SDK licensing, partner bench, deployment validation) is real and proportional to the engagement scope. See §6.

4.2 When AQEA Edge Replaces an Accelerator

The replacement question is binary: *does the partner's edge AI workload reduce to similarity-classification?* If yes, AQEA Edge replaces the accelerator path. If the workload also requires transformer-inference, image segmentation, or generation, the accelerator stays for those tasks — but the *similarity component* can move to the CPU.

Workload pattern on the device	AQEA Edge replaces accelerator?
Predictive maintenance (sensor □ classify □ alert)	**Yes — full replacement**
Robot fault detection (motion □ match against fault DB)	**Yes — full replacement**
Quality control (sensor fingerprint □ in-spec/out-of-spec)	**Yes — full replacement**
Wearable anomaly (ECG/IMU □ match against baseline)	**Yes — full replacement**
Voice activity / wake-word	**Yes — full replacement** (matching only)
Speech-to-text	No (requires transformer inference)
Vision object segmentation	No (requires segmentation network)
Text generation	No (requires LLM inference)

A typical industrial deployment is dominated by the upper rows.

4.3 Per-Query Energy Comparison

A direct apples-to-apples accelerator-to-AQEA-Edge energy comparison requires the same workload run through both paths. Within this whitepaper we report the AQEA Edge half on commodity x86 CPU (Hetzner AX102):

Domain	AQEA Edge (mJ/q)	Float-CPU baseline (mJ/q)	Energy advantage
Text retrieval	37.66	248.63	6.6×
Speech speaker-verification	2.20	5.07	2.3×
Humanoid motion classification	4.70	11.86	3.3×
Industrial robot anomaly	67.60	353.00	5.2×
Hardware fall-prediction	4.60	16.34	3.55×

These numbers are AQEA Edge against a *CPU-Float baseline on the same machine*, not against a dedicated accelerator. A direct comparison against (e.g.) Hailo-15 on a partner's actual deployment hardware is the natural Joint-Validation engagement scope (§6).

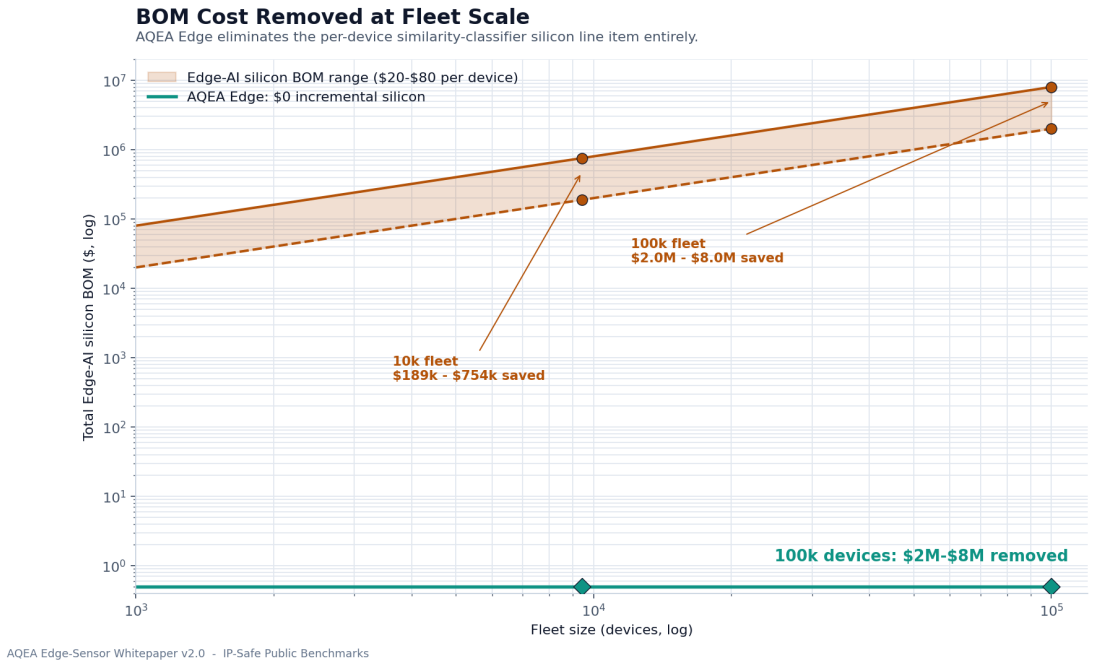


Figure 9: **Fig. 4.** Bill-of-materials cost compounding at fleet scale. Each row of the per-device-AI-chip price band ($20–80 per device, conservative range across Coral / Hailo / low-end-Jetson) compounds into a removed BOM line item across a fleet. At 100 k devices, the AQEA-Edge alternative removes a $2 M – $8 M BOM line versus the AI-accelerator path. Above the device count axis is logarithmic.

For an order-of-magnitude estimate: a typical Edge-AI accelerator at similar throughput consumes 2-5 W when active. AQEA Edge running on the partner's existing CPU during cycles that CPU would otherwise be at low-utilisation idle has effectively zero *additional* power draw — the difference between 2-5 W and 0 W per device, multiplied by the device fleet, is the deployment-scale energy delta.

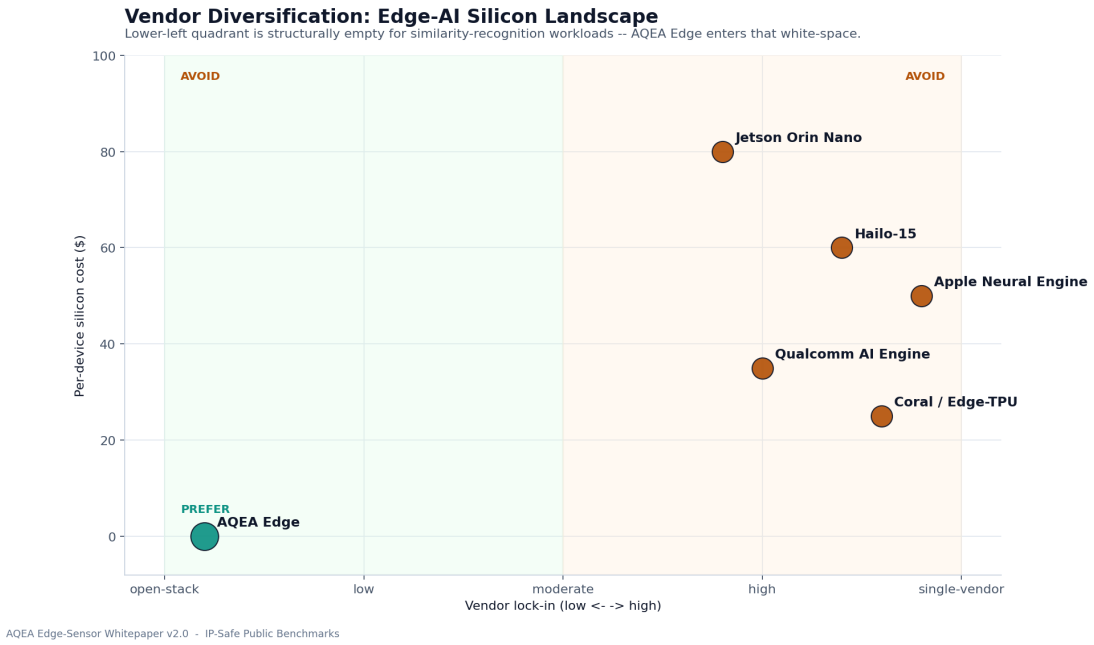


Figure 10: **Fig. 9.** Vendor-lock-in × silicon-cost positioning. The five dominant Edge-AI silicon families cluster in the upper-right quadrant (proprietary toolchain, $50–600 per device). AQEA Edge sits in the lower-left "PREFER" quadrant: zero incremental silicon cost, no vendor lock-in (runs on standard CPU SIMD or any Vulkan/Metal/WebGPU/DirectX-12 device).

4.4 Hardware Reach: Where AQEA Edge Runs

Hardware class	Validated / Plausible	Notes
Datacenter CPU (Intel Xeon, AMD EPYC)	☐ Hetzner-validated	Primary bench platform
Workstation CPU (Apple Silicon, AMD Ryzen)	☐ M3 Max validated	Cross-platform Bit-Identity verified
Datacenter GPU (NVIDIA, AMD)	☐ H100 validated	Cross-vendor identity verified
Workstation GPU (Apple Metal)	☐ M3 Max Metal verified	Cross-vendor identity verified
Edge SoC ARM Cortex-A + NEON	Plausible (engineering)	Same SIMD instruction class as M3 NEON
Edge SoC x86 Atom + AVX	Plausible (engineering)	Same SIMD instruction class as Hetzner AVX-512
Mobile GPU (Qualcomm Adreno, ARM Mali)	Plausible (engineering)	Vulkan support

Hardware class	Validated / Plausible	Notes
Browser (WebGPU)	Portable (engineering)	Same WGSL source as desktop GPU
Microcontroller without SIMD	Not validated; no	Pipeline assumes SIMD or shader

4.5 What This Architecture Is Not

For symmetry — to enable partner self-evaluation:

- **AQEA Edge is not a general AI-inference replacement.** It does not run transformer models, generate output, or perform segmentation. For workloads that need these, the accelerator chip stays.
- **AQEA Edge requires SIMD or shader support.** Embedded MCUs without SIMD (most Cortex-M-class parts) are not target hardware in this version.
- **AQEA Edge is similarity-only.** If your classification task requires learned non-linear decision boundaries that cannot be expressed as nearest-neighbour-against-reference-set, AQEA Edge is not sufficient.
- **AQEA Edge does not eliminate the encoder.** Stage-0 (the encoder that produces float vectors from sensor input) is still required, and is partner-supplied. AQEA Edge replaces only the *classification-by-similarity* stage that currently runs on the AI-accelerator chip.

These limitations are bounded and documented, not papered over. The partner workflow we expect is: identify which subset of on-device classification reduces to similarity-search, replace those paths with AQEA Edge, leave the remainder on whatever inference stack already exists.

ewpage

5 Use Cases

Four production scenarios are concretely addressable today. In each, the partner's existing signal-processing or transformer encoder stays unchanged; AQEA Edge replaces the AI-chip-bound classification step.

5.1 Predictive Maintenance on Industrial Machinery

A typical factory deployment has 10^2–10^4 machines under continuous monitoring. Each machine produces multi-channel sensor streams: vibration (3-axis accelerometer at 10–100 kHz), current/torque/voltage from drive electronics, temperature, acoustic. Spectral features (FFT, STFT, wavelet) are computed on-device. The classification task is: *match the current spectral fingerprint to known-fault and known-healthy reference fingerprints, and trigger an alert if the closest match is a fault state*.

Today this runs on a Jetson, Coral, Hailo, or vendor-proprietary controller. Per-device chip cost typically $50–300, multiplied by the device count.

With AQEA Edge: the spectral-feature pipeline produces float vectors as before; AQEA Encode runs on the same controller's CPU; AQEA Search compares against a 1k–100k-element reference

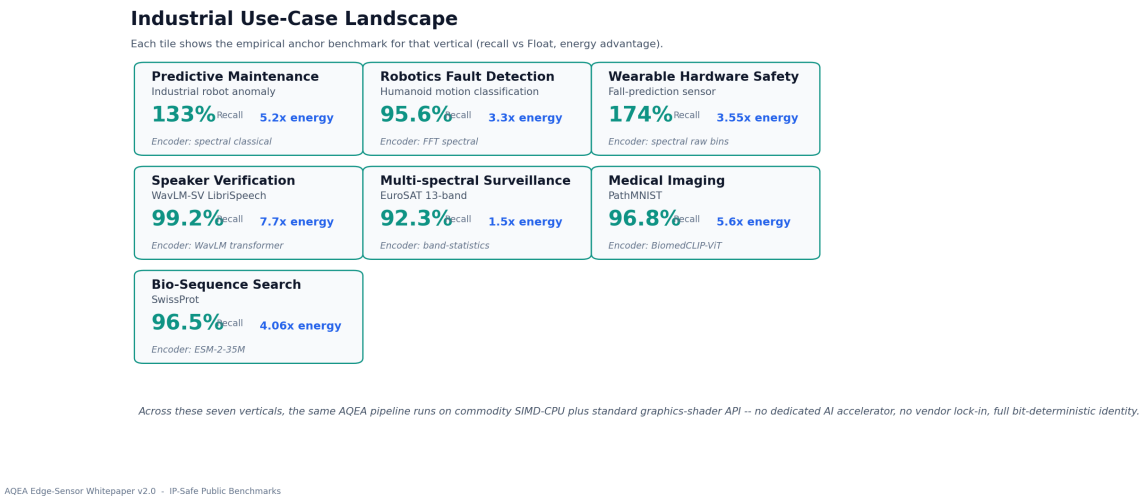


Figure 11: **Fig. 12.** Industrial use-case landscape. Seven verticals where similarity-classification is the dominant on-device AI workload, each anchored by an empirical AQEA-Edge benchmark from §3 (recall vs Float, energy advantage). Workloads that require general transformer inference, segmentation, or generation are out of scope and continue to use existing accelerator silicon.

database held in local memory; the Top-K nearest references plus their fault labels are returned. Latency on a Cortex-A78-class controller is sub-10ms for 100k references; energy is the CPU's normal active draw, not an additional accelerator chip's power state.

Empirical evidence: §3 voraus-ad bench (industrial robot anomaly, 100k-scale) and §3 humanoid-bench (84-channel motion sensor) — both spectral-encoder domains, both AQEA Edge ≥ FAISS-CPU-Float on recall, both 3-5× more energy-efficient.

The relevant decision-maker is typically a Plant Engineering Director, Predictive-Maintenance Platform Lead, or CTO at a manufacturing automation OEM.

5.2 Robot Fleet Health Monitoring

A robot fleet (warehouse, manufacturing, logistics) of 10^2–10^4 units, each producing motion-state streams (joint encoders, IMU, force-torque). Fleet-level health monitoring needs to identify which robots are degrading, which are showing precursors of specific faults, and which are off-pattern compared to the rest of the fleet.

Today this is split between (a) per-robot edge classification (typically Jetson-class hardware) and (b) a cloud aggregation layer. The cloud layer is heavy because the per-robot classifications are often noisy and require aggregation/correlation.

With AQEA Edge: each robot does fast on-CPU spectral encoding and similarity search against a fleet-shared reference, encoded once and synced. The per-robot result is more deterministic (same encoded reference across all robots, same bit-identical match logic), reducing the cloud-aggregation overhead. The shared reference is bit-identical across all robot platforms (Apple Silicon dev controllers, x86 production controllers, ARM Cortex-A field controllers).

Empirical evidence: §3 humanoid-bench (h1 crawl task, 6 fault classes) — direct test on humanoid robot motion data.

The relevant decision-maker is typically a Fleet Operations Director or Robotics Platform Architect at a robotics OEM.

5.3 Wearable / Consumer-IoT Anomaly Detection

A wearable device (smartwatch, fitness band, medical monitor) sampling biosensors (ECG, PPG, IMU, EDA, accelerometer) at 50–500 Hz. Battery life is the dominant design constraint; on-device classification is preferred over always-on cloud upload for both privacy and battery reasons. Today's wearable AI silicon (Apple Neural Engine, Qualcomm AI Engine on Snapdragon Wear) is locked to the platform.

With AQEA Edge: the wearable's existing CPU + NEON does the entire classification pipeline. The reference database (e.g., user-specific baseline + clinical reference fingerprints) is small (1k–10k elements) and fits comfortably in the device's RAM. Inter-device-portability becomes possible: the same encoded reference works across different wearable hardware, simplifying multi-platform product lines.

Empirical evidence: §3 Digit_Fall sensor — small-corpus, low-dimensional sensor-stream classification, AQEA Edge exceeds Float-cosine recall (174 % Ratio, the strongest noise-filter effect in the cross-domain table).

The relevant decision-maker is typically a Product or Hardware-Engineering Lead at a consumer-wearable OEM, or a Clinical-IoT Platform Architect.

5.4 On-Device Voice / Acoustic-Event Recognition

A smart-speaker, hearing-aid, or industrial audio-monitor running classification: wake-word, voice-activity, speaker-ID-against-enrolled-set, or acoustic-event-classification (alarm sounds, machinery anomaly, environmental). Audio embedding via WavLM, Wav2Vec2, or domain-trained classifiers; classification by similarity against a small enrolled or reference set.

Today: dedicated low-power AI inference silicon for the matching step.

With AQEA Edge: the same audio-embedding step runs as before; AQEA Encode + AQEA Search runs on the device's existing CPU. Empirical evidence: §3 LibriSpeech/WavLM-SV (full Hetzner production) — direct test on audio speaker-verification, **99.24 % Recall-Ratio** against Float-Cosine baseline at **7.7× lower energy** per query.

The relevant decision-maker is typically a Product Engineering Lead at a smart-audio OEM or hearing-aid manufacturer.

5.5 Regulated-Industry Audit-Chain (Audit-Chain Mode)

A class of edge-deployment that is uniquely enabled by the substrate's reversible-decode property (P6 Audit-Chain Mode, §3.7) and is not addressable by any vendor-locked Edge-AI silicon today.

Scenario. An automotive ECU, an aerospace flight-data sensor, a medical-device monitor, an energy-grid metering unit — any edge device whose output is subject to regulatory audit, accident reconstruction, or legally-binding evidence-chain requirements. Today, the partner faces a binary tradeoff:

- (a) **Store raw signal-traces.** Cost-prohibitive at fleet-scale (raw multi-channel sensor at high sample-rate produces gigabytes per device per day); often privacy-restricted; retention-policy complex.
- (b) **Store classification-results-only.** No retrospective audit, no root-cause-analysis on past events, no evidence-chain for liability investigations.

With AQEA Edge in Audit-Chain Mode. The edge device performs local classification using the substrate's direct-distance ranking; the substrate-encoded artefact (≈175 bytes per timestep) is stored locally or shipped to a central archive. When forensic analysis, regulatory audit, or root-cause investigation is needed at a later time, the archived artefact is reversibly decoded with per-vector cosine fidelity ≥ 0.90 against the original encoder output, and the substrate's retrieval-ranking is preserved at ≥ 96.7 % Audit-Chain mode. The investigator can re-classify, re-rank, or re-correlate against any updated reference database without needing the original raw signal.

Compliance-relevant properties combined.

- Cross-platform bit-deterministic encoding (P2): per-device-certification-overhead removed.
- Structural compression (P3 + structural compression of P6 audit-chain artefact): cost of central audit-archive reduced by 5-30× over raw-signal-storage.
- Reversible decoding (P6 Audit-Chain Mode): forensic re-classification on past events possible.
- Patent-pending: the deployment-time mode-selection mechanism that allows the same substrate-encoded artefact to be decoded in audit-fidelity mode (forensic) or in pure-retrieval mode (live similarity-search) without re-encoding.

Validated by. §3 LibriSpeech (transformer-encoder, 99.2 % R-Ratio at 18× compression) + §3 voraus-ad / Digit_Fall (industrial-sensor classical-encoder, 133 % / 174 % R-Ratio with noise-filter-improvement) + §3.7 Decoder Pareto-Front Audit-Chain mode (96.7 % retrieval at M1-cosine ≥ 0.92). Cross-Platform bit-identity verified across 4 hardware backends.

The relevant decision-maker is typically a Compliance Lead, Functional-Safety Officer, or Quality / Liability Director at an automotive OEM, aerospace supplier, medical-device manufacturer, or regulated-industry IoT integrator.

5.6 Where AQEA Edge Is Not Yet the Right Choice

For symmetry:

- **Microcontrollers without SIMD** (most Cortex-M-class parts). The pipeline assumes SIMD or shader support; M-class MCUs without DSP-extension are not target hardware in v1.
- **Workloads requiring transformer inference on-device.** Speech-to-text, image segmentation, language-modelling — these still need the AI-accelerator path.
- **Generation tasks.** Output synthesis (image, text, action) is not a similarity-search workload.
- **Sub-1-ms hard-real-time control loops.** Latency of AQEA Edge on commodity SoC is 1–10 ms; for sub-millisecond control loops the partner's existing fast-path is correct.

These limitations are bounded and the partner's existing tool remains correct in each case.

ewpage

6 Roadmap and Ask

6.1 Validated Today

- **13 distinct domain × encoder combinations** benchmarked apples-to-apples on a single shared commodity x86 server (§3): six transformer foundation families (BGE text, WavLM speech, ESM-2 protein, BiomedCLIP medical-imaging, codet5p code, CLAP music) plus four classical signal-processing pipelines (FFT-spectral, DCT-image, multispectral-band-statistics, MassBank mass-spectrometry archive).
- **Encoder-family agnostic** across the transformer □ classical-DSP paradigm boundary — same pipeline, same binaries, no per-domain tuning.
- **Cross-platform bit-identity** verified on ARM NEON, x86 AVX-512, NVIDIA Vulkan, Apple Metal.
- **WebGPU browser bit-identity** verified on a 10k-synthetic-document benchmark with cross-backend Bit-Identity to the NVIDIA-Vulkan reference (50 / 50 fixtures).
- **Energy advantage** 1.5–9.4× lower per-query versus Float-CPU baseline on the same hardware (depending on domain).
- **Recall preservation or improvement:** 84.6–174 % Recall-Ratio across the thirteen domain × encoder combinations; □ industrial-sensor noise-bearing domains exceed Float-FAISS (133 % voraus-ad, 174 % Digit_Fall).
- **Reversibly-decodable substrate** validated on text retrieval with three-operating-point Pareto-front (Audit-Chain 96.7 % at M1-cosine ≥ 0.92 / General-Purpose 98.7 % / Pure-Retrieval 99.7-101.1 %); see §3.7.

6.2 Q3 2026 Roadmap

In priority order:

1. **Edge-SoC validation.** Pipeline run on representative ARM Cortex-A78 / Cortex-A510 SoCs (Apple A-series, Qualcomm 8-series, MediaTek Dimensity, NXP i.MX 9, ST STM32MP). Latency, throughput, energy on each.
2. **Real-time-loop integration.** Bench AQEA Edge inside a 1ms-budget industrial control loop (PLC scan-cycle integration, robot cycle-time integration). Validate that the pipeline meets hard real-time constraints.
3. **WebGPU browser deployment.** In-browser similarity search has been validated with cross-backend Bit-Identity to the NVIDIA-Vulkan reference (10k-synthetic-document benchmark, 50 / 50 fixtures). Production-scale browser deployment over a 1k-100k-element reference set is straightforward from this validated foundation; static-web-app-no-server proof-of-concept can be made available to partners under engagement.
4. **Joint-validation with one industrial OEM partner.** First production-engagement on a real partner workload, signed under engineering NDA. The pre-registered bench protocol (Joint-Validation Offer, §6.4) is the engagement starter.

6.3 Q4 2026 Roadmap

5. **Public Edge SDK.** Documented C / Rust / Python / JavaScript bindings, packaged for cargo / PyPI / npm. The encoder, indexer, and search components behind a black-box interface that does not expose the patent-pending substrate construction.

6. **Reference-app gallery.** Open-source examples for: predictive-maintenance with a public sensor dataset, robot-fault-classification with humanoid-bench, voice-wake-word with a public audio dataset.
7. **Energy / sustainability certification.** ISO-14001-aligned per-device energy measurement protocol and a third-party-reviewable bench methodology that partners can use for their own OEM-product certification claims.
8. **Patent-filing completed (May 9, 2026).** A 16-application USPTO Provisional Patent portfolio was filed same-day on May 9, 2026 (Application Numbers 64/061,723 through 64/061,752, Customer Number 219394). The portfolio covers the substrate's foundational construction, the deterministic cross-platform encoder, the multi-stage similarity-search architecture, the audit-proof memory layer, the type-system-enforced compile-time immutability, the multi-scale memory hierarchy, the pre-inference constraint-detection method, the frozen-encoder adapter, the adaptive spectral eigenmode ranking, the audit-native knowledge-graph engine, the foam-vertex knowledge-graph embedding, the auditable bit-identical cross-architecture trit-quantizer, and the reversibility master-method (four-trial validated as documented in §3.7). Non-Provisional Conversion + PCT-International filings are scheduled for May 2027 (12-month conversion-window). NDA partners receive specific Application Numbers per patent track as part of engagement materials.

6.4 Partner Ask

We are seeking three classes of engagement with industrial / edge / robotics partners.

6.4.1 Option A — Engineering Evaluation under NDA (4–6 weeks)

A joint benchmark: AQEA Edge is run against the partner's own retrieval / classification workload (the partner's encoder, the partner's reference set, the partner's hardware) under engineering NDA. The deliverable is a reproducible bench report against whatever AI-accelerator path the partner currently uses, with apples-to-apples Recall, Latency, Throughput, and Energy comparison.

Cost to partner: engineering time of one or two edge-stack engineers for the duration. Cost to NextX: bench preparation + on-site or remote support.

This is the recommended starting point.

6.4.2 Option B — Integration Pilot (8–12 weeks)

A more substantial engagement: AQEA Edge is integrated into a non-production path of the partner's edge stack and run in shadow-mode against the production AI-accelerator path. The deliverable is a side-by-side comparison of latency, recall, energy, and BOM-cost on a real workload at the partner's typical fleet-deployment scale.

6.4.3 Option C — Co-Development Engagement

A deeper relationship: hardware-specific tuning (partner-specific SoC architecture, FPGA mapping, custom silicon block-design), encoder-family extension (partner's vertical-specific encoder), or bespoke edge-product integration. Structure to be negotiated.

This is the appropriate engagement for OEM partners with significant strategic interest in removing the per-device AI-accelerator BOM-cost from their product line.

6.4.4 Contact

For all three options:

- **NextX AG** — https://nextx.ch
- **CEO contact:** Sayed Amir Karim · s.karim@nextx.ch
- **Engineering contact:** Sayed Amir Karim · s.karim@nextx.ch

A one-page engagement-scope draft is returned within five working days of inbound contact.

ewpage

7 Appendix

7.1 Reproducibility Statement

All five domain benchmarks reported in §3 are reproducible. The full configuration:

- **Hardware:** Hetzner AX102 — AMD EPYC 9454P, 96 cores, AVX-512, RAPL energy measurement, Ubuntu 24.04. Cross-platform Bit-Identity verified independently on Apple M3 Max (NEON / Metal) and Sapphire Rapids x86_64 (AVX-512).
- **AQEA Edge binaries:** `aqea-bench` and `dump_trits` from the public-safe SDK. Distribution under engagement agreement; reference notebook on request under NDA.
- **FAISS baseline:** stock FAISS-CPU IndexFlatIP, 96 threads.
- **Bench parameters:** n_queries ≥ 100 (cycled on smaller corpora), n_warmup = 5, top_k = 1,000, top_n = 10. Single-query mode for latency, batch mode for throughput.
- **Verification:** SHA-256 manifest of every encoded reference set; cross-platform topology hash verified across four hardware platforms.

The five datasets used are all public:

- msmarco-passage (Microsoft / TREC)
- LibriSpeech test-clean (public-domain audio)
- humanoid-bench (open-source robotics benchmark)
- voraus-ad (open-source industrial-robot anomaly dataset)
- Digit_Fall_Prediction_Dataset (open-source hardware-state dataset)

7.2 Glossary

Term	Public-Safe Definition
AQEA Edge	Three-stage processing pipeline (encoder □ AQEA-Encode □ AQEA-Search) for similarity-based recognition on multi-channel signal streams.
AQEA Substrate	The structured representation space onto which float-vector encoder output is projected. Patent-pending. See §2.

Term	Public-Safe Definition
Trit-encoded vector	Compressed substrate representation, ≈175 bytes/doc, deterministic, cross-platform-bit-identical. Patent-pending.
Encoder-family agnostic	Pipeline accepts float-vector output of any encoder (transformer or classical signal-processing) without per-encoder tuning.
Noise-resistant ranking	Empirical property: substrate's encoded-distance ranking is more robust to encoder high-frequency noise than raw float-cosine on the same data.
Cross-platform bit-identity	Substrate elements byte-identical across heterogeneous SIMD and shader backends (ARM NEON, x86 AVX-512, NVIDIA Vulkan, Apple Metal).
Recall-Ratio	AQEA Recall@K divided by Float-FAISS Recall@K on identical data. ≥80% Floor / ≥90% PASS / ≥95% STRONG (pre-registered).
Edge-AI accelerator	Dedicated silicon for on-device AI inference: NVIDIA Jetson, Google Coral, Hailo, Apple Neural Engine, Qualcomm AI Engine.
Similarity-based recognition	Classification by nearest-neighbour against a reference set, as opposed to model-based inference (transformer generation, vision segmentation).
Reversibly-decodable substrate	Substrate-property: a learned-inverse function reconstructs a float-representation from a substrate-encoded element with task-preservation measured against the substrate's own direct-ranking baseline. Three operating-modes on a Pareto-front: Audit-Chain (M1-cosine ≥ 0.92, retrieval ≥ 96.7 %), General-Purpose (M1-cosine ≈ 0.75, retrieval ≥ 98.7 %), Pure-Retrieval (M1-cosine ≈ 0.66, retrieval ≥ 99.7 % with 101.1 % task-elevating regime at 10^5-scale). Mode-selection is a deployment-time choice. Patent-pending. See §3.7.
Audit-Chain Mode	The reversible-decode operating-mode optimised for per-vector reconstruction-fidelity. Suitable for edge-device evidence-trail, regulatory reporting, forensic reconstruction, and root-cause-analysis without re-acquiring the original raw signal. Empirically: per-vector cosine-fidelity ≥ 0.92 against the original encoder output, retrieval-quality ≥ 96.7 % of substrate baseline.

7.3 References

- Bajaj, P., Campos, D., Craswell, N., et al. (2018). *MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset*. arXiv:1611.09268.
- Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.
- Khronos Group (2024). *Vulkan Specification*. https://www.vulkan.org/
- Khronos Group (2025). *WebGPU Shading Language (WGSL)*. https://www.w3.org/TR/WGSL/
- Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and Robust Approximate Nearest Neighbor Search Using HNSW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836.
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: an ASR corpus based on public domain audio books. *ICASSP*.
- Sferrazza, C., et al. (2024). HumanoidBench: a benchmark for humanoid locomotion and manipulation. (Open-source release used in §3.)
- W3C WebGPU Working Group (2024). *WebGPU Specification*. https://www.w3.org/TR/webgpu/

- Xiao, S., Liu, Z., Zhang, P., & Muennighoff, N. (2023). C-Pack: Packed Resources For General Chinese Embeddings. arXiv:2309.07597.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504-507. *(Auto-Encoder reference for §3.7 reversibility-comparison.)*
- Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. ICLR 2014. arXiv:1312.6114. *(VAE reference for §3.7 reversibility-comparison.)*
- van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. NeurIPS 2017. arXiv:1711.00937. *(VQ-VAE reference for §3.7 reversibility-comparison.)*

7.4 Patent-Pending Disclosures

The following innovations are patent-pending. This whitepaper discloses them at the level required to evaluate engineering fit; implementation details and the underlying mathematical construction are protected by filings and are not disclosed here.

A 16-application USPTO Provisional Patent portfolio was filed same-day on May 9, 2026 (Application Numbers 64/061,723 through 64/061,752, USPTO Customer Number 219394). The portfolio is structured into the following six tracks:

Track A — Substrate Search Pipeline (3 applications): - The cross-vendor GPU shader pipeline including the GPU-native Top-K selection mechanism (USSN 64/061,725). - The CPU-SIMD trit-substrate retrieval pipeline using the substrate's channel structure (USSN 64/061,726). - The hardware-implementations roadmap covering FPGA / ASIC / optical / display targets (USSN 64/061,727).

Track B — Audit-Proof Memory + Tamper-Detection (4 applications): - The audit-proof memory layer with multi-anchor architecture and SHA-256 hash-chain (USSN 64/061,729). - The compile-time-enforced memory-immutability via programming-language type-system (USSN 64/061,731). - The multi-scale memory hierarchy via stacked-foam topology with compliance property preservation (USSN 64/061,733). - The pre-inference constraint-detection method via topology-encoded policy anchors (USSN 64/061,734).

Track C — Master Substrate (1 application): - The information-geometry substrate construction — the structured representation space, its channel decomposition, the multi-channel orthogonality property, the deterministic encoder, the noise-filter property on noise-bearing signal-domain encoders, and the predictive-consistency property across encoder paradigms (USSN 64/061,723).

Track D — Search-Engine-Layer Sub-Master Patents (5 applications): - The frozen-encoder + trainable-linear-head adapter (USSN 64/061,736). - The adaptive spectral eigenmode top-K ranking (USSN 64/061,737). - The audit-native knowledge-graph engine integrating Cypher + similarity + SHA-256-audit-trail (USSN 64/061,738). - The foam-vertex cochain as knowledge-graph property-encoding (USSN 64/061,741). - The auditable bit-identical cross-architecture trit-quantizer, validated across x86-AVX-512, ARM-NEON, NVIDIA-Vulkan, Apple-Metal (USSN 64/061,743).

Track E — Reversibility Master-Method (1 application): - The reversibility master-method — forward-encoder + learned-inverse-decoder + ranking-aware contrastive-loss + R@K-validation, including the three-operating-point Pareto-front and the task-elevating-decoder regime documented

in §3.7. The Audit-Chain Mode is particularly relevant for edge-device evidence-trail and regulatory-reporting applications (USSN 64/061,749).

Track F — Language-Model Compression (2 applications, conversion-contingent): - Trit-strip-embedding frozen-eigenbasis token-embedding for transformer LLMs (USSN 64/061,751). - Combined encoding-stack with channel-orthogonal error-correction and quantization-aware-training (USSN 64/061,752).

Non-Provisional Conversion + PCT-International filings are scheduled for May 2027 (12-month conversion-window from the May 9, 2026 priority date). Partners engaging under NDA receive specific Application Numbers per patent track as part of the engagement materials.

7.5 Acknowledgements

The bench infrastructure was built on top of the open-source `wgpu`, `rayon`, `pollster`, and `bytemuck` Rust crates (Apache 2.0 / MIT). The signal-processing pipelines for the robotics and industrial datasets are reproductions of the upstream open-source toolchains (humanoid-bench, voraus-ad). The audio bench uses public LibriSpeech via `torchaudio.datasets.LIBRISPEECH`. Hardware was provided by Hetzner AX102 (Hetzner Cloud), Lambda Cloud H100 PCIe, and Apple M3 Max (developer workstation).

7.6 Contact

NextX AG — https://nextx.ch

- CEO: Sayed Amir Karim — s.karim@nextx.ch
- Engineering Lead: Sayed Amir Karim — s.karim@nextx.ch

For engineering-NDA evaluations, integration pilots, or co-development engagements, see §6.

This document version: v2.0, 2026-05-09. Whitepaper Patent-Pending Public-Safe Disclosure for Industrial / Edge / Robotics audience. 16-application USPTO Provisional Patent portfolio filed same-day on May 9, 2026 (Application Numbers 64/061,723 through 64/061,752, Customer Number 219394). Public distribution permitted; partner engagement-options (§6) require engagement-NDA per the Partner Ask section.

ewpage