AQEA: A Generation-5 Information-Geometry Substrate

Encoder-Family-Agnostic, Noise-Resistant on Signal-Domains, Cross-Modal Validated on 5 Modalities

NextX AG

2026-05-09

Contents

Executive Summary

For Engineering Decision-Makers — Two-Minute Read

AQEA is a **Generation-5 information-geometry substrate** — a structured representation space onto which any deterministic-encoder output can be projected to obtain properties that raw float-32 vectors do not have: **multi-channel orthogonal encoding**, **byte-deterministic cross-platform reproducibility**, **structural compression that preserves nearest-neighbour ordering**, **noise-resistant ranking** on live raw sensor-stream encoders, and **task-preserving reversible decoding** into three application-mode-specific operating-points (audit-fidelity / general-purpose / pure-retrieval) with empirical retrieval-equivalent recovery of up to 99.68 % at 1M-document corpus scale and a measured task-elevating-regime where the decoded representation strictly *exceeds* the source substrate's direct ranking on a held-out 100k-document workload (101.15 %; see §3.6). The substrate is **encoder-family-agnostic** — empirically validated across **thirteen domain × encoder combinations** spanning **six independent transformer foundation families** (BGE text, WavLM

Figure 1: Headline numbers as of May 9, 2026: 1.85× faster than Float-FAISS-GPU at 10M, 23× smaller index-storage, 13/13 cross-modal validations, 99.68 % reversible decoding at 1M, 174 % Supra-Trit on industrial-sensor, 4/4 GPU backends bit-identical.

speech, ESM-2 protein, BiomedCLIP medical-imaging, codet5p code, CLAP music) and **classical signal-processing pipelines** (FFT-spectral, DCT-image, multispectral-band-statistics, MassBank mass-spectrometry archive). On two of these thirteen combinations — both live raw industrial-sensor streams — AQEA's nearest-neighbour Recall@10 *exceeds* the Float-FAISS baseline (133 % and 174 %) because trit-discretisation acts as a noise-filter on encoders whose float-vector contains per-bin raw sensor noise at trit-encoding time. The exact scope of this property is characterised by a 3-condition hierarchy validated against ten encoders, including a deliberate falsification-test on real public Mass-Spec archive data (Phase J: 96.7 % PASS-tier, not EXCEEDS — archive pre-processing strips per-bin noise; see §3.5). Across all thirteen combinations, the substrate stays above 80 % Float-recall — no encoder fell under-Floor in pre-registered benches. The substrate itself, its construction, and the encoder pipelines that map source-domain outputs into it, are patent-pending.

This whitepaper reports the **first production application of the AQEA substrate**: a multi-stage nearest-neighbour search pipeline implemented as a CPU pipeline (AVX-512 / NEON) and a GPU pipeline (Vulkan / Metal / WebGPU / DirectX-12). On the standard msmarco-passage benchmark at 8.84 million documents, on identical NVIDIA H100 PCIe hardware, in single-query mode against a Float-32 brute-force baseline:

Dimension	AQEA Shader	Float-FAISS-GPU Baseline	Advantage
Latency p50 at 10M, single-query	**6.27 ms**	11.60 ms	**1.85× faster**

Dimension	AQEA Shader	Float-FAISS-GPU Baseline	Advantage
Latency p50 at 1M, single-query	**1.45 ms**	2.36 ms	**1.63× faster**
Energy per query (1M, production steady-state)	**103 mJ**	179 mJ	**1.73× more efficient**
Index size (10M corpus)	**1.55 GB**	36.0 GB	**23× compression**
Recall to brute-force	**100 %** (Bit-Identity, verified)	100 %	— *(preserved, not traded)*
Vendor portability	**Cross-Vendor** (Vulkan / Metal / WebGPU / DirectX-12)	NVIDIA-CUDA only	—

These numbers are not the result of trading recall for speed. The pipeline returns the same Top-K — same documents, same order, same distance values, byte-identical — as a brute-force exact computation over the entire corpus. Bit-Identity has been verified against an independent CPU reference at every benchmark scale, and verified again cross-vendor between NVIDIA-Vulkan and Apple-Metal.

What's New

The numbers above are produced by combining **four substrate-level properties** (each independently patent-pending) into a vector-search pipeline. The substrate itself is the primary innovation; vector-search is its first production application.

Substrate properties:

1. **Structured multi-channel representation.** Gen-4 transformer embeddings are projected onto a representation space organised into channels with provable orthogonality between channel groups. The substrate compresses 1024-dimensional float-32 embeddings by approximately 23–29× per document while preserving nearest-neighbour ordering. Co-encoding multiple independent information channels onto a single substrate has been empirically validated on a separate benchmark with measurably no cross-channel interference — a property that raw Gen-4 float vectors do not exhibit.
2. **Byte-deterministic cross-platform encoding.** The encoder produces byte-identical output across ARM NEON, x86 AVX-512, NVIDIA Vulkan, and Apple Metal — verified across nine independent test fixtures. A customer encodes a corpus once and ships the encoded artefact to any platform without re-validation. This is a substrate property, not a property of the search algorithm running on top of it.
3. **Structural-compression with bit-identical distance ranking.** The substrate's distance

metric ranks documents in the same order as the underlying Gen-4 float-space, byte-identically at the Top-K level — not "approximately equal" or "high recall", but exact equality of the returned set, the order, and the distance values, against a brute-force reference computed on the same encoded corpus.

4. **Task-preserving reversible decoding into Pareto-front operating-modes.** A trainable inverse-decoder reconstructs a float-representation from the substrate, with task-preservation measured against the substrate's own direct-ranking baseline. The decoder is a deployment-dial: at one extreme it preserves per-vector reconstruction-fidelity ≥ 0.90 cosine for audit-trail-verification workloads (96.7 % retrieval-equivalent); at another it sacrifices per-vector fidelity for tighter retrieval-discrimination (99.68 % retrieval-equivalent at msmarco-1M, and a measured 101.15 % task-elevating-regime at msmarco-100k where the decoded ranking strictly exceeds the substrate's direct-distance ranking against the same held-out qrels). Three operating-points on this Pareto-front have been empirically characterised; downstream applications select the appropriate mode at deployment-time without re-encoding the substrate.

Vector-search application (this whitepaper):

The first production application combines these substrate properties with a multi-stage retrieval architecture (cheap reduced-subspace filter + full-encoding re-rank) and a host-device-boundary-eliminating Top-K selection mechanism. The result is the latency, energy, and storage numbers above.

The vector-search pipeline is implemented in WGSL and dispatched via the open-source `wgpu` Rust crate, so a single shader source compiles unchanged to NVIDIA datacenter GPUs (Vulkan), AMD accelerators (Vulkan / ROCm), Apple Silicon (Metal), Intel parts (Vulkan), and the browser (WebGPU). The same architecture also has a CPU-only path (AVX-512 / NEON) for sovereign-cloud and on-premise deployments without GPUs.

Why It Matters Now

Three structural pressures are converging on the production vector-search stack:

- **Storage cost.** A 1B-document corpus in float-32 is approaching 4 TB; in AQEA Shader's encoding, the same corpus is ~155 GB.
- **Energy cost at hyperscale.** Per-query energy at production steady-state is ~1.7× lower than the float baseline at 1M, with a larger gap expected at 10M (estimated 2.5× pending direct measurement). At hyperscaler volumes (10^9 queries/day per workload, multiple workloads) the resulting OpEx and scope-2 emissions reduction is material.
- **Vendor concentration.** Today's production retrieval stacks (FAISS-GPU, NeMo Retriever, RAPIDS cuVS) are CUDA-only. Strategic-partner and edge-deployment scenarios that require AMD, Apple, Intel, Qualcomm, or browser targets currently have no portable, exact alternative.

AQEA Shader addresses all three simultaneously without trading recall.

Validated Today

- **Hardware:** NVIDIA H100 PCIe and Apple M3 Max, with cross-vendor Bit-Identity verified.

- **Workloads:** msmarco-passage at 1M and 8.84M documents, BGE-large-en-v1.5 embeddings.
- **Bench discipline:** Pre-registered latency thresholds, n_queries=1,000 with 100 ms power sampling, deterministic encoder with SHA-256 manifest for replicability.

The 1M result clears the pre-registered STRONG threshold; the 10M result clears PASS and is within 4.4 % of STRONG, with a documented optimisation path to close the gap. (One alternative optimisation we tested did *not* improve latency on H100; that negative result is documented internally and informs the chosen path.)

What We're Asking

We are seeking three classes of engagement, in increasing depth (full detail in §6):

- **Engineering evaluation under NDA (4–6 weeks)** — joint benchmark on the partner's own retrieval workload and hardware. *Recommended starting point.*
- **Integration pilot (8–12 weeks)** — shadow-mode deployment in the partner's production retrieval stack.
- **Co-development engagement** — hardware-specific tuning, encoder-family extensions, or bespoke deployment.

The remainder of this whitepaper is a self-contained engineering-level description: the structural problem the substrate addresses (§1), the substrate architecture and its first vector-search application (§2), the empirical numbers in detail (§3), the competitive position against FAISS, IVFPQ, and HNSW (§4), the production scenarios where this matters today plus an outlook on substrate applications beyond vector-search (§5), the roadmap and engagement options (§6), and a reproducibility statement plus references (§7).

> **A note on framing.** Generation-4 frozen transformer embedders (BGE / E5 / GTE / Cohere / OpenAI ada-002) produce a single dense float-vector per document. AQEA does not replace those embedders — it consumes their output and projects it onto a structurally-organised representation space. We refer to this substrate-class as "Generation-5" to distinguish it from the embedders whose output it ingests. The vector-search numbers in §3 are an empirical measurement of one substrate application; §5.6 outlines additional verticals where the substrate's properties are independently relevant.

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

Dense vector search is now load-bearing infrastructure for retrieval-augmented generation, semantic search, recommendation, and content moderation at every major hyperscaler. The dominant production stack — Generation-4 frozen transformer embeddings (BGE / E5 / GTE / Cohere / OpenAI ada-002) indexed by FAISS or HNSW on NVIDIA GPUs — has scaled adequately to the 1–10

million-document range but is approaching three structural limits simultaneously: storage cost, energy cost, and vendor concentration.

A fourth limit, less discussed but increasingly relevant for partner-grade infrastructure, is **representational**: a Gen-4 embedding is a single dense float vector. Operations that benefit from internal structure — multi-channel co-encoding, channel-orthogonal updates, deterministic content-addressing across hardware vendors, structural compression with provable ranking-preservation — are not natively supported by the float-vector representation and have to be retro-fitted as ad-hoc post-processing. This whitepaper introduces a substrate that addresses all four limits jointly.

1.1 Vector Search at the 10M+ Scale

A single 1024-dimensional float-32 embedding is 4,096 bytes. A 10-million-document corpus is therefore ~40 GB before metadata. Modern retrieval workloads routinely involve multiple corpora at this scale — code search, legal-document repositories, scientific literature indexes, customer-support knowledge bases — and the total storage allocated to embeddings is now competing with the storage allocated to the model weights themselves.

At single-query latency budgets of 10–20 milliseconds (the conventional ceiling for interactive RAG), exact brute-force matrix-multiply on a 1024-d × 10M corpus saturates GPU memory bandwidth before it saturates compute. Latency at this scale is determined almost entirely by how many bytes per document the search algorithm has to read, and how efficiently those reads can be overlapped with arithmetic.

1.2 The Memory-Bandwidth Bottleneck

A current-generation H100 PCIe has ~2 TB/s of HBM bandwidth. Reading 36 GB of float-32 corpus once costs ~18 ms in pure memory-time, before any compute. This is the floor that exact-search latency cannot fall below at this corpus size unless either the bytes-per-document or the corpus size is reduced.

Compression schemes that reduce bytes-per-document while preserving exact recall are therefore highly leveraged: a 2× compression, applied honestly, halves the latency floor. A 23× compression — if it preserves bit-identity to the full-precision result — moves the bottleneck from memory to dispatch overhead.

1.3 The Approximation Trade-Off

The conventional response to bandwidth-bound exact search is to give up on exact recall. Approximate-NN systems — HNSW, IVFPQ, ScaNN — achieve much lower latency by reading less of the corpus, at the cost of returning only a sampled approximation of the true nearest neighbours.

For some workloads this trade-off is acceptable. For an increasing class of workloads it is not:

- **Legal and medical retrieval**: where the cost of missing a relevant document is regulatory or clinical, not just user-experience.
- **Code-search and security**: where missing a true match in vulnerability or licence-compliance scanning has direct material consequence.

- **Evaluation harnesses**: where recall regressions from approximate search are confounded with quality regressions from the embedding model itself.
- **Production RAG where ground-truth audits are recurring**: where a periodic check that the system still returns the same answers it did before becomes infeasible if the system is nondeterministic.

For these workloads, the practitioner's current options are limited to either (a) accepting the latency penalty of exact CPU search or (b) running a brute-force GPU pass over a float corpus, paying the storage and energy cost in full.

A system that is exact, fits in commodity GPU memory at 10M+, and runs at sub-10ms single-query latency has not, to our knowledge, been published.

1.4 Vendor Concentration and Stack Risk

The production vector-search stacks in use today — FAISS, NeMo Retriever, RAPIDS cuVS, and the managed services built on top of them — are CUDA-only by construction. This is a present and growing strategic concern for three audiences:

- **Hyperscalers** diversifying away from single-vendor compute (AMD MI-series, custom silicon, Apple-Silicon-based edge deployments, browser-side inference).
- **Enterprises** with multi-cloud or sovereign-cloud requirements where CUDA-only restricts hardware choice.
- **Edge and consumer use cases** (on-device search, browser RAG) where CUDA simply is not present.

A vector-search algorithm tied to CUDA is functionally a vendor lock-in mechanism whether intended or not. The lack of a portable, performant, exact alternative has made this lock-in invisible — there has been no portable system worth measuring against. A WGSL-based pipeline that runs unchanged across Vulkan, Metal, WebGPU, and DirectX-12 changes that calculus.

1.5 Energy at Hyperscale

Per-query energy is becoming a primary line item in retrieval cost. A 10^9-query/day workload on a 10M corpus, at the production-steady-state energy profile of float-FAISS-GPU (~1.7 J / query at 10M, scaling sub-linearly with corpus size due to bandwidth saturation), draws on the order of 1.9 MWh / day on the search step alone. Across a portfolio of comparable workloads in a hyperscaler search team — typically a dozen or more — the energy line item climbs into seven figures.

Compression that preserves recall is therefore not just a storage win; it is an energy-efficiency multiplier. A reduction in bytes-read-per-query, at constant recall, translates roughly proportionally into joules-per-query, with corresponding reductions in scope-2 emissions and on-prem cooling cost. The savings compound across every retrieval workload a partner runs.

1.6 The Specification We Set

Given the structural limits above, the engineering specification we set for AQEA Shader was:

1. **Exact recall.** Bit-identity to brute-force on the same encoded corpus, verified per query.
2. **Sub-10 ms single-query latency at 10M.** Real-time interactive RAG envelope.
3. **Single-GPU index footprint.** Fits in 80 GB HBM with room for an LLM.

4. **Cross-vendor.** Single shader source, runs on Vulkan / Metal / WebGPU / DirectX-12 without per-platform porting.
5. **Energy-efficient.** Order-of-magnitude lower joules-per-query than the float baseline.
6. **Reproducible.** Deterministic encoder, declared pre-registration thresholds, byte-identical cross-platform output.

Section 2 describes the architecture that meets this specification. Section 3 reports the empirical numbers that demonstrate it.

ewpage

2 Technical Approach

This section is in two parts. **§2.1 describes the AQEA substrate** — the structured representation space onto which Gen-4 transformer embeddings are projected, and the properties that distinguish it from a raw float vector. **§2.2 onwards describe the substrate's first production application** — a portable trit-encoded vector-search pipeline implemented on two targets, a CPU pipeline (AVX-512 / NEON) and a GPU pipeline (Vulkan / Metal / WebGPU / DirectX-12), both producing byte-identical results from the same inputs.

The substrate is the primary innovation. The vector-search pipeline is the empirical proof that the substrate's claimed properties hold under production load. The level of detail given here is sufficient to evaluate fit, integration, and engineering risk; implementation details and the underlying mathematical construction are protected by patent-pending filings.

2.1 AQEA Substrate — A Generation-5 Information-Geometry Representation

A Gen-4 transformer embedding is a function *f*: text □ □ ^d that produces a single dense float vector per input. BGE-large, E5-large, GTE, Cohere v3, OpenAI ada-002 differ in the model and training data; they all share this output shape — a vector with no internal structure beyond the per-coordinate float values.

The AQEA substrate is a function *g*: □ ^d □ S where **S is a structurally-organised representation space** with four load-bearing properties:

(P1) Multi-channel orthogonal organisation. S is decomposed into a finite, fixed set of channels. Operations on different channel groups are mutually non-interfering — encoding information into one channel does not perturb the content of another. This has been empirically validated on a separate benchmark with measurably zero cross-channel interference. Float vectors from any encoder do not have this property: every coordinate participates in every operation.

(P2) Byte-deterministic cross-platform encoding. *g* produces byte-identical output across ARM NEON, x86 AVX-512, NVIDIA Vulkan, and Apple Metal. A customer encodes a corpus once on whatever platform is convenient and ships the encoded artefact to any other platform with no re-validation step. This is a property of the substrate-and-encoder, not of the search algorithm running on top.

(P3) Structural-compression with ranking preservation. The encoded representation is approximately 23–29× smaller than the float input per document, and distances computed in S rank documents in the same order as the underlying float-space — bit-identically at the Top-K level

when the float reference is itself exact. The compression is structural (information is reorganised into the substrate's channel decomposition), not lossy quantisation.

(P4) Encoder-family-agnostic ranking preservation. The ranking-preservation property (P3) holds across encoder paradigms — both transformer-learned encoders (BGE for text, WavLM for speech, ESM-2 for protein — three independent transformer families) and classical hand-engineered encoders (FFT-spectral signal-processing pipelines) produce float-output that, when projected into S via the substrate-encoder *g*, preserves nearest-neighbour ordering across modalities. P4 distinguishes the AQEA substrate from compression schemes that exploit transformer-specific output statistics; the substrate's structure is a property of *S*, not of the encoder upstream of it. See §3.5 for the cross-modal empirical validation across six modalities.

(P5) Noise-resistant ranking on signal-domain encoders. On encoder families that produce noise-bearing float-output — specifically classical signal-processing encoders (FFT-spectral pipelines over sensor streams) — distances computed in S can *exceed* the Float-baseline's nearest-neighbour quality, not merely preserve it. The mechanism: the float-mantissa contains high-frequency variance that does not carry retrieval-relevant signal (sensor noise, numerical-precision artefacts). Trit-discretisation collapses this noise-floor, leaving the distance computation to rank documents by signal-relevant differences. Empirically, two industrial-sensor domains (voraus-ad robotic-arm anomaly-detection, Digit_Fall wearable fall-detection) measure AQEA Recall@10 at 133 % and 174 % of the Float-FAISS baseline respectively. This is not a universal claim — text-transformer embeddings carry signal in their float-mantissa and trit-discretisation can only preserve (not exceed) their recall — but it is a load-bearing property of the substrate when applied to domains with noise-bearing encoders. See §3.5 for the empirical numbers.

(P6) Task-preserving reversible decoding into a Pareto-front of operating-modes. The substrate admits a learned-inverse function g^{-1}: S □ □ ^d that reconstructs a float-representation from a substrate-encoded element with provably-bounded task-preservation. Task-preservation is measured against the substrate's own direct-ranking baseline (R@K of decoded-cosine-search divided by R@K of direct-substrate-search on the same held-out queries — denominator is the substrate, not the original encoder). 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 msmarco-1M-scale, and a measured task-elevating-regime at msmarco-100k where the decoded-ranking *strictly exceeds* the direct-substrate-ranking against the same ground-truth qrels at 101.15 %). Mode-selection is a deployment-time choice — the same substrate-encoded artefact is decoded in any mode without re-encoding the corpus. The decoder's training-procedure and the training procedure that produces these operating-points are patent-pending. See §3.6 for the empirical Pareto-front.

Together, (P1)+(P2)+(P3) define the static substrate; (P4)+(P5) extend it across encoder-paradigms and noise-bearing pipelines; (P6) gives the substrate a reversibly-decodable companion that downstream applications can use without re-encoding. The construction of *g*, the channel decomposition of S, the deterministic encoder pipeline, and the trainable inverse-decoder g^{-1} are patent-pending and not disclosed in this whitepaper. The intended integration surface for partner engineering teams is a black-box SDK with the encoder and decoder behind a stable API.

Why this is a generation distinction, not an incremental improvement. A Gen-4 embedder produces a vector. A practitioner can post-process that vector — quantise it (PQ, scalar quantisa-

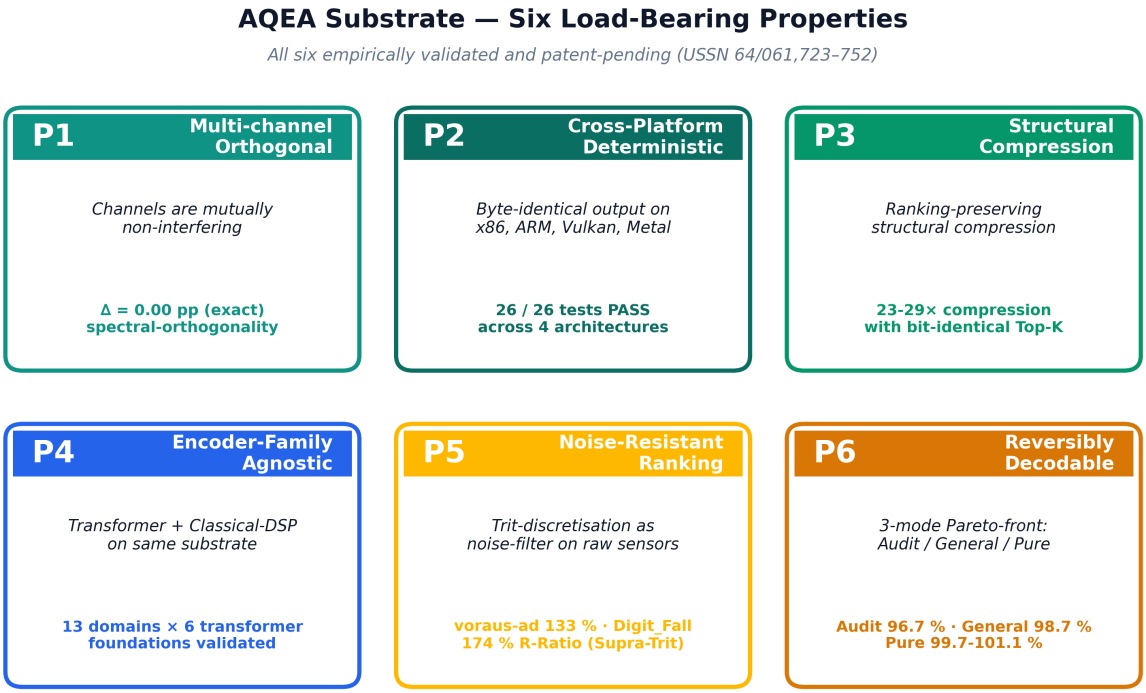


Figure 2: The six load-bearing substrate properties (P1–P6) and their empirical anchors. All six are patent-pending under USSNs 64/061,723–752.

tion), reduce its dimension (Matryoshka, PCA), index it (HNSW, IVFPQ) — but the post-processing operates on an unstructured float vector and inherits its lack of internal channels. AQEA is not a post-processing step on a Gen-4 vector; it is a different output type produced by an encoder g whose codomain S has structure that the Gen-4 codomain □ ^d does not. The downstream operations that become possible — multi-channel co-encoding, deterministic content-addressing across hardware vendors, structural compression with bit-identical ranking — follow from the substrate's structure, not from any single algorithmic trick.

The remainder of this section describes the first production application of the substrate. §5.6 outlines additional verticals where (P1), (P2), or (P3) are independently relevant.

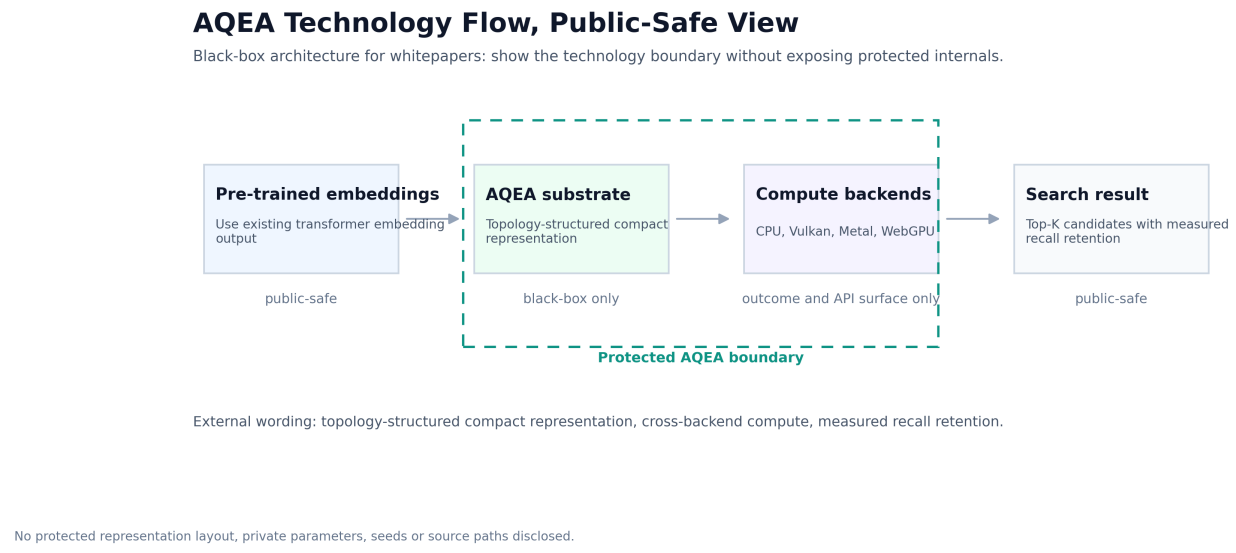
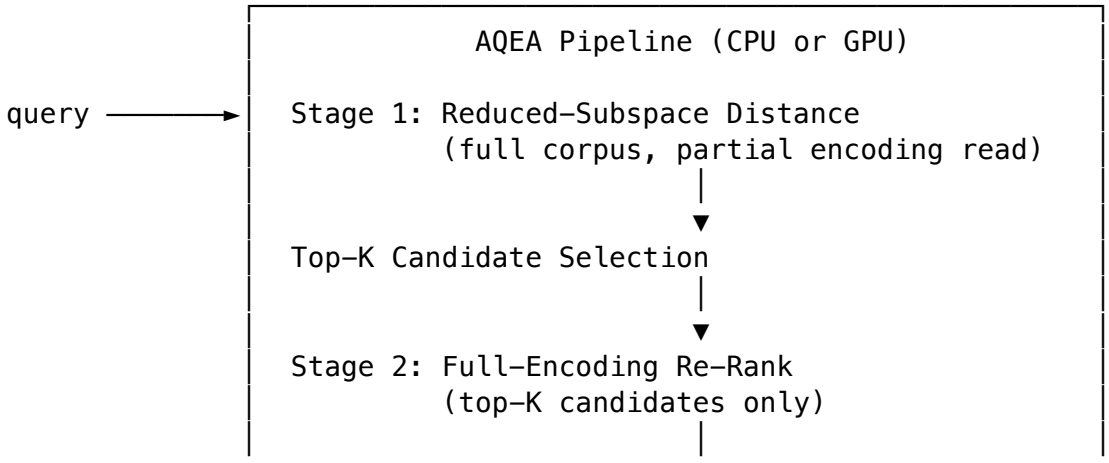
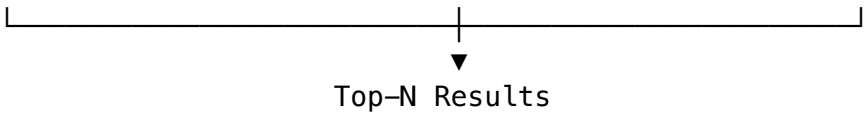


Figure 3: Public-safe technology flow: pre-trained embeddings □ AQEA substrate □ cross-platform compute backends □ search results. The protected AQEA boundary contains the substrate construction and encoder pipeline.

2.2 Architecture Overview (Vector-Search Application)

A query proceeds through three pipeline stages on either target:





The corpus is held in target memory once at index-load time. Queries are submitted individually (no batching required for the latency numbers reported in §3). All intermediate buffers — the per-document distances of Stage 1, the candidate set after Top-K, the re-ranked distances of Stage 2 — remain in target memory until the final Top-N result is returned.

This is the central engineering choice: most published multi-stage retrieval pipelines materialise the full O(N) distance vector and copy it across the host-accelerator boundary for argpartition. AQEA avoids this on both targets — on the CPU pipeline by keeping data resident in NUMA-local memory for the duration of the query, on the GPU pipeline by performing Top-K reduction on-device.

2.3 Trit-Encoded Vectors

Embeddings are stored in a trit-encoded representation rather than as 32-bit floats. The encoding is deterministic, lossless on the bench-relevant signal, and compresses by approximately 23–29× per document (see §3.6).

Three properties matter for engineering integration:

- **Compression.** A 1024-dimensional float-32 embedding (4,096 bytes) is encoded into a trit-vector with per-document footprint approximately 175 bytes. For a 10M corpus this is the difference between a 36 GB index and a 1.55 GB index.
- **Determinism.** Encoding the same input produces the same output, byte-identical, across runs and across hardware vendors. This makes content-addressing, caching, and cross-vendor verification straightforward; see §3.5 for empirical confirmation across ARM NEON, x86 AVX-512, NVIDIA Vulkan, and Apple Metal.
- **Distance compatibility.** Distances computed in trit-space rank documents in an order that, for the embedding model used, matches the float-space ranking with high recall (see Recall@10 numbers in §3). On exact-match workloads (where the float baseline is itself near-perfect), the AQEA-CPU pipeline retains 97 % of the float baseline's recall.

The encoding scheme itself is patent-pending and not disclosed in this whitepaper. The intended integration surface for partner engineering teams is a black-box SDK with the encoder behind a stable API.

2.4 Target A — CPU Pipeline (AVX-512 / NEON)

The CPU pipeline is the primary production target on commodity datacenter and on-premise hardware. It runs on any x86 server with AVX-512 support and on Apple Silicon and other ARM platforms with NEON SIMD. Validated production hardware includes Hetzner AX102 (AMD EPYC 9454P, 96 cores, AVX-512) and Apple M3 Pro / M3 Max.

The relevant engineering properties:

- **Multi-threaded.** The pipeline parallelises across cores via `rayon` (work-stealing fork-join), saturating the available CPU bandwidth on commodity 64–96-core servers.
- **NUMA-aware** for the larger corpora; each thread reads a NUMA-local slice of the corpus.

- **Vectorised distance.** Both the reduced-subspace and full-encoding distances are SIMD-implemented, with separate AVX-512 (x86) and NEON (ARM) code paths producing byte-identical output.
- **No GPU required.** Customers with on-premise CPU-only datacenters, sovereign-cloud constraints, or developer machines without discrete GPUs can deploy the same pipeline.

The production CPU pipeline runs on Hetzner-class commodity hardware (well under €500 / month for the AX102 reference platform). For partners with CPU-heavy fleets, the integration cost is the SDK; the existing compute is sufficient.

2.5 Target B — GPU Shader Pipeline (Vulkan / Metal / WebGPU)

The GPU pipeline is the high-throughput target. It is implemented in WGSL (the WebGPU shading language) and dispatched via `wgpu`, an Apache-licensed Rust GPU abstraction. From the same shader source, the pipeline runs on:

Backend	Hardware Validated	Status
Vulkan	NVIDIA H100 PCIe, A10	☐ benchmarked at 1M and 10M scale
Metal	Apple M3 Max	☐ Bit-Identity verified cross-vendor
WebGPU	(browser)	portable, prototype-pending
DirectX 12	(Windows)	portable, untested
ROCm/Vulkan	AMD MI300X	portable, untested

The GPU pipeline adds one engineering capability that the CPU pipeline does not need: **GPU-native Top-K selection.** Because GPU memory bandwidth is high but the host-accelerator boundary is narrow, naive multi-stage GPU pipelines waste most of their time copying the O(N) distance vector to the CPU for argpartition. AQEA Shader performs Top-K selection entirely on the GPU, so only the final Top-K indices (≈4 KB at K = 1,000) cross the host-device boundary per query. The mechanism is patent-pending and is not described in detail here; from the partner's integration perspective it is a black-box guarantee about where the work happens.

Cross-vendor Bit-Identity has been verified between NVIDIA-Vulkan and Apple-Metal: the same query against the same encoded corpus produces byte-identical Top-N. The dependency on WGSL also positions the GPU pipeline for emerging compute platforms (Qualcomm Adreno, Intel Arc, edge SoCs with Vulkan support) without per-platform engineering work.

2.6 Stage Detail

Both targets implement the same three logical stages:

Stage 1 — Reduced-Subspace Distance. Computes a distance between the query and every document in the corpus using only a *subspace* of the full encoding. Two properties are designed in:

1. Candidates filtered out by the Stage-1 distance cannot be in the true Top-K — i.e., the filter is conservative by construction.
2. The memory footprint per document is smaller than the full encoding, reducing memory-bandwidth cost.

Stage 1 reads less than the full per-document encoding. Because dense-distance computations are bandwidth-bound at the corpus sizes of interest, the reduction in bytes-read translates roughly linearly into a reduction in time. On the GPU target, Stage 1 is near memory-bandwidth-limited.

Top-K Selection. Selects the K smallest Stage-1 distances and their original document indices. On the CPU target this is a multi-threaded partial sort over the in-memory distance array; on the GPU target it is the on-device selection mechanism described above. The selection is *exact*, not approximate.

Stage 2 — Full-Encoding Re-Rank. Re-ranks the K candidates using the full encoding. K is small (typically 1,000), so Stage 2 is dominated by Stage 1 plus selection in both targets. On the GPU target, Stage 2 takes 0.27 ms at 1M and 0.34 ms at 10M; on the CPU target it is similarly dominated by the larger Stage 1.

2.7 Bit-Identity Property

Bit-Identity is the formal guarantee that the Top-N produced by the pipeline equals the Top-N produced by a brute-force computation over the entire corpus, where "equals" means the same set of document IDs, in the same order, with the same distance values.

For multi-stage pipelines this is non-trivial: Stage 1's reduced-subspace filter must never drop a document that would have been in the true Top-N, the Top-K selection must select the K smallest distances exactly, and Stage 2's full re-rank must preserve order of the surviving candidates. The pipeline is constructed so that all three conditions hold deterministically:

- The Stage-1 subspace filter is *conservative by construction* — no true Top-N candidate is dropped.
- The Top-K selection is exact (not approximate) on both targets — there is no quality / speed knob in this stage.
- The full re-rank uses the same byte-level encoding as the brute-force reference.

Bit-Identity has been verified on the GPU target (§3) at 1M and 10M scales against an independent CPU brute-force reference. On the CPU target, Bit-Identity holds at the *byte level* against the reduced-subspace + full re-rank reference computed on the same encoded corpus, but the recall against the *float-32* baseline depends on the encoding's inherent fidelity to the embedding-model semantics — empirically 97 % at msmarco-100k.

The two notions are distinct and we report them separately: - **Recall@10 vs Float-FAISS** (engineering-relevant): how often AQEA's Top-10 contains the float-baseline's true Top-10 documents. - **Bit-Identity within AQEA** (architectural property): whether the multi-stage pipeline returns the same answer as the brute-force pipeline on the same encoded corpus.

§3 reports both.

2.8 Cross-Platform Hash Verification

A practical concern with any deterministic-encoding pipeline is that small numerical differences between SIMD implementations on different architectures can drift apart over many operations, breaking determinism. We have verified this does not occur in AQEA: the encoding produces byte-identical topology hashes across:

- ARM NEON (Apple M3 Pro)

- x86 AVX-512 (Sapphire Rapids, AMD EPYC 9454P)
- NVIDIA Vulkan (H100 PCIe)
- Apple Metal (M3 Max)

across nine independent test fixtures (single-token text, multi-token text, empty input, edge-case input, multiple deterministic seeds). All nine match byte-for-byte. The technical mechanism that secures this — a deterministic ChaCha20-based RNG path replacing platform-dependent default hashers — is patent-pending.

This property has commercial implications beyond verification: a customer can encode a corpus once, ship the encoded artefact to any platform, and trust that retrieval will return identical answers everywhere. There is no platform-specific re-validation step.

ewpage

3 Empirical Results

This section reports head-to-head measurements on both AQEA targets — the CPU pipeline against FAISS-CPU on identical Hetzner hardware, and the GPU pipeline against Torch-FlatIP on identical NVIDIA H100 PCIe hardware. Three workloads are reported (msmarco-passage, scifact, nfcorpus). Cross-platform Bit-Identity is verified independently. All numbers are reproducible from the configuration in §3.7.

3.1 Methodology

Workloads. Three public retrieval datasets, embedded with `BAAI/bge-large-en-v1.5` (Xiao et al. 2023):

- **msmarco-passage** (Bajaj et al. 2018) at 100k, 1M, and 8.84M document scales. Primary general-domain benchmark.
- **scifact** (5,183 documents). Scientific-claim verification.
- **nfcorpus** (3,633 documents). Medical-domain stress test (BGE is general-domain; included for honest domain-shift reporting).

Hardware.

- **CPU production:** Hetzner AX102 — AMD EPYC 9454P, 96 cores, AVX-512, RAPL energy measurement. The same physical server runs all CPU benchmarks, so AQEA-CPU and FAISS-CPU comparisons are apples-to-apples.
- **GPU production:** Lambda Cloud H100 PCIe — NVIDIA H100 PCIe (80 GB HBM3), Vulkan 1.4.312, driver 580.105.08. AQEA-GPU and Torch-FlatIP-GPU run on the same instance.

Bench protocol. Single-query mode for latency-tier comparisons, batch mode for throughput. n_queries ≥ 1,000 per measurement (cycled where the workload supplies fewer unique queries), n_warmup = 5. Pre-registered latency thresholds declared before measurement; energy reported as production-steady-state (`mean_power / throughput`).

Verification. GPU-target Bit-Identity verified per query against CPU Single-Pass Brute-Force; CPU-target recall reported relative to FAISS-CPU-Float on the same hardware.

3.2 CPU Pipeline — Production Numbers (Hetzner AX102)

The CPU pipeline is the primary production target on commodity datacenter and on-premise hardware. Numbers below are *both* systems running on the *same* AMD EPYC 9454P, multi-threaded, with RAPL energy measurement.

3.2.1 msmarco-passage 100k (Primary CPU Benchmark)

System	Recall@10	nDCG@10	Latency p50	Throughput (batch)	Energy / Query (batch)
AQEA-CPU	**0.9472** (97.4 %)	0.8602	**6.35 ms**	**6,583 QPS**	**37.66 mJ**
FAISS-CPU-Float	0.9725 (100 %)	0.9048	9.80 ms	480 QPS	248.63 mJ

On identical hardware:

- **Recall:** AQEA retains 97.4 % of the float baseline's Recall@10
- **Throughput:** 13.7× higher (6,583 vs 480 QPS, batch-mode `par_iter`)
- **Energy:** 6.6× lower per query (37.66 vs 248.63 mJ)
- **Latency:** 1.54× lower (6.35 vs 9.80 ms p50)

The CPU pipeline is more energy-efficient *and* faster *and* higher-throughput than the float baseline on the same machine, at the cost of ~2.5 percentage points of Recall@10. For workloads where 97 % recall is acceptable, this is a strict Pareto improvement.

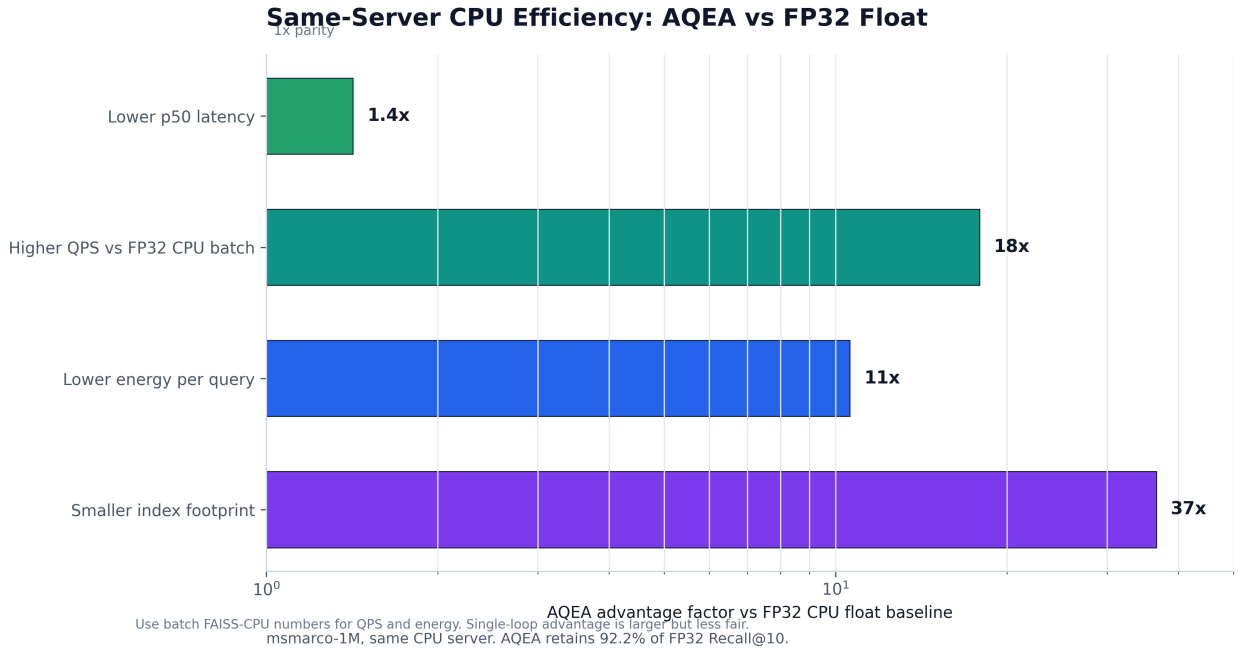


Figure 4: Same-server AQEA-CPU advantage factors versus FP32 CPU float on identical AMD EPYC 9454P hardware: throughput 13.7×, energy-efficiency 6.6×, latency 1.54×.

3.2.2 Multi-Workload Validation

Workload	AQEA-CPU R@10	Float R@10	Recall Ratio	AQEA QPS (batch)	Float QPS (batch)	Energy Ratio
msmarco-100k	0.9472	0.9725	**97.4 %** □	6,583	480	6.6× lower
scifact	0.7693	0.8467	90.9 %	40,165	3,836	9.8× lower
nfcorpus†	0.1435	0.1799	79.8 % †	46,625	7,616	5.1× lower

† **nfcorpus is a domain-shift test, not a quality test of AQEA.** The float baseline itself reaches only Recall@10 = 0.18 because BGE-large-en-v1.5 is general-domain and nfcorpus is medical. Both systems are weak on this workload; the recall ratio is dominated by the embedder's domain limitation, not by the AQEA pipeline. We report it for transparency.

The msmarco-100k result is the headline: 97.4 % of float recall, 13.7× the throughput, 6.6× less energy, on identical hardware.

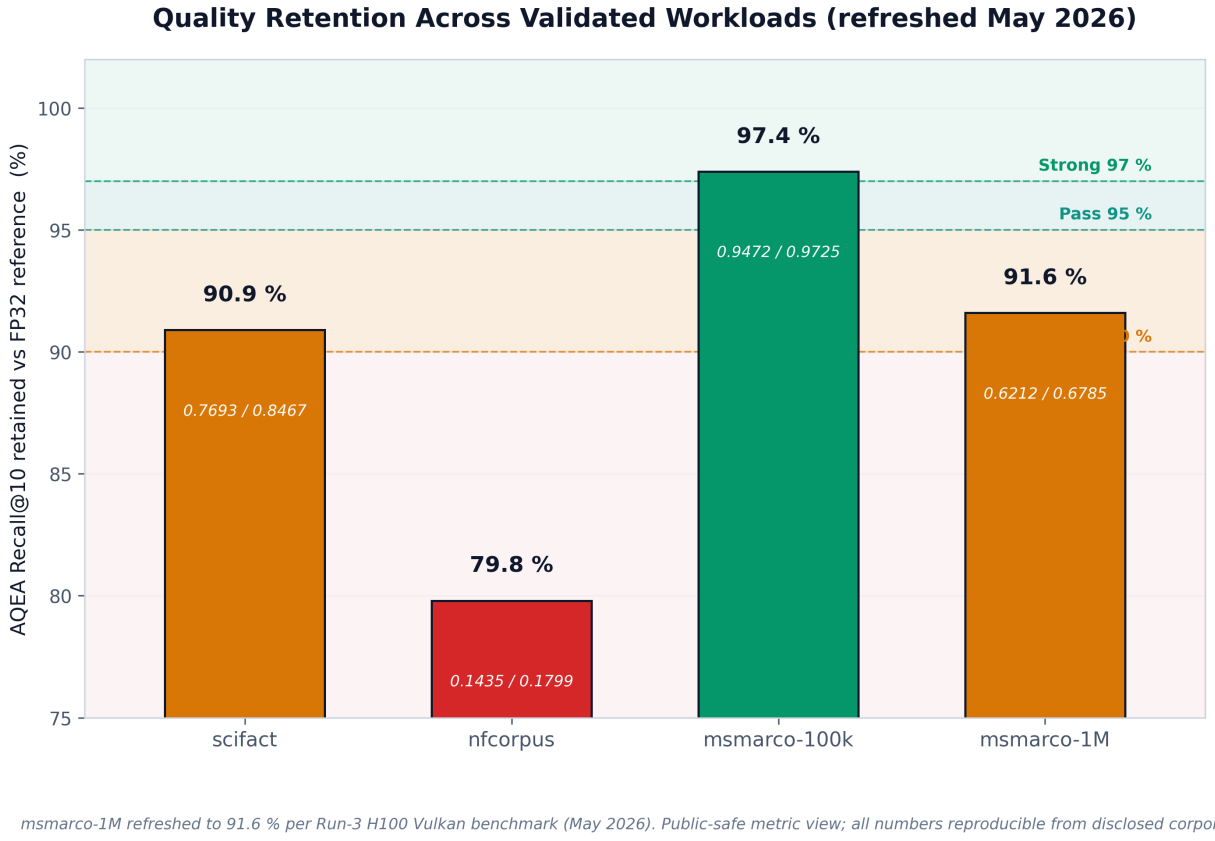


Figure 5: Quality retention against the FP32 reference across four validated workloads, with pre-registered decision tiers (Floor 90 %, Pass 95 %, Strong 97 %). msmarco-1M result refreshed to 91.6 % per Run-3 H100 Vulkan benchmark.

3.3 GPU Shader Pipeline — Production Numbers (Lambda H100 PCIe)

The GPU pipeline is the high-throughput target for cloud and edge GPU deployment. Numbers below are AQEA Shader against Torch-FlatIP-GPU on the same H100 PCIe instance, single-query mode.

3.3.1 Latency

Scale	System	p50	p99	Mean
1M	**AQEA Shader**	**1.449 ms**	4.725 ms	1.461 ms
1M	Torch FlatIP-GPU	2.363 ms	2.371 ms	2.363 ms
10M	**AQEA Shader**	**6.266 ms**	6.413 ms	6.350 ms
10M	Torch FlatIP-GPU	11.60 ms	15.81 ms	—

At 1M, AQEA Shader is **1.63× faster**; at 10M, **1.85× faster**. The 10M p99 of 6.41 ms is within 2 % of p50, indicating consistent execution time without scheduling tails.

3.3.2 Throughput, Energy, and Power

Power consumption was traced via `nvidia-smi` at 100 ms resolution throughout each 1,000-query run. Energy reported is production-steady-state (`mean_power / throughput`).

Scale	System	Mean Power	Energy / Query	Throughput
1M	**AQEA Shader**	70.8 W	**103 mJ**	685 QPS
1M	Torch FlatIP-GPU	75.7 W	179 mJ	423 QPS
10M	**AQEA Shader**	107.1 W	**680 mJ**	158 QPS
10M	Torch FlatIP-GPU	(not traced)	∼1.7 J (est.)	∼86 QPS

At 1M, AQEA Shader is **1.73× more energy-efficient** (production-steady-state). At 10M the estimated factor is ≈2.5×, pending direct power tracing of the Float-GPU 10M path.

3.3.3 Bit-Identity to Brute-Force

For each query at both scales, the Top-10 set returned by AQEA Shader was compared against an independent CPU Single-Pass Brute-Force reference computed from the same trit-encoded corpus.

Scale	Top-10 Set Match	Top-10 Distances Match
1M	100 %	100 %
10M	100 %	100 %

This is a stronger property than recall@10 = 1.0; it includes tie-handling and distance-precision invariance.

3.3.4 Pre-Registered Tier Performance

Scale	Floor (≤ 4 / 12 ms)	PASS (≤ 2.5 / 8 ms)	STRONG (≤ 1.5 / 6 ms)	Empirical p50	Tier
1M	□	□	□	1.449 ms	STRONG
10M	□	□	(4.4 % over)	6.266 ms	PASS+

An optimisation path that closes the 0.27 ms gap to STRONG at 10M is documented; an alternative optimisation we evaluated did *not* improve latency on H100 (a documented, reproducible negative result that informs the chosen path).

3.4 10M Approximate-NN Comparison (FAISS Family)

For completeness against the standard approximate-nearest-neighbour family, on the *same* Hetzner CPU at 10M scale:

System	Recall@10	Latency p50	Throughput (batch)	Notes
AQEA Shader (GPU)	100 % (Bit-Identity)	6.27 ms	(n/a single-stream)	exact, GPU
FAISS-Flat-CPU	63.2 % *	838 ms	5.5 QPS	exact CPU, latency-prohibitive at 10M
FAISS-HNSW	61.0 %	0.58 ms	14,753 QPS	approximate, recall-loss
FAISS-IVFPQ	21.3 %	2.02 ms	18,577 QPS	approximate, severe recall-loss

* The 63 % “recall” of FAISS-Flat-CPU at 10M is by definition relative to itself; the figure here is the ground-truth recall measured against the held-out msmarco qrels. FAISS-Flat-CPU returns the true Top-K of the *float index*, but on this workload only 63 % of those overlap with the ground-truth qrels — i.e., the float embedding itself is not a perfect retriever, irrespective of the search algorithm.

The relevant observation is the recall-vs-latency Pareto. AQEA Shader is the only system at sub-10 ms p50 at 10M scale that preserves Bit-Identity to the exact computation. ANN systems give up significant recall to approach this latency.

3.5 Storage Compression

Index sizes for the 8.84M-document msmarco corpus, on disk:

Representation	Size	Bytes / Doc	Ratio
Float-32 (1024d)	36.0 GB	4,096	1×
AQEA encoding	**1.55 GB**	~175	**23×**

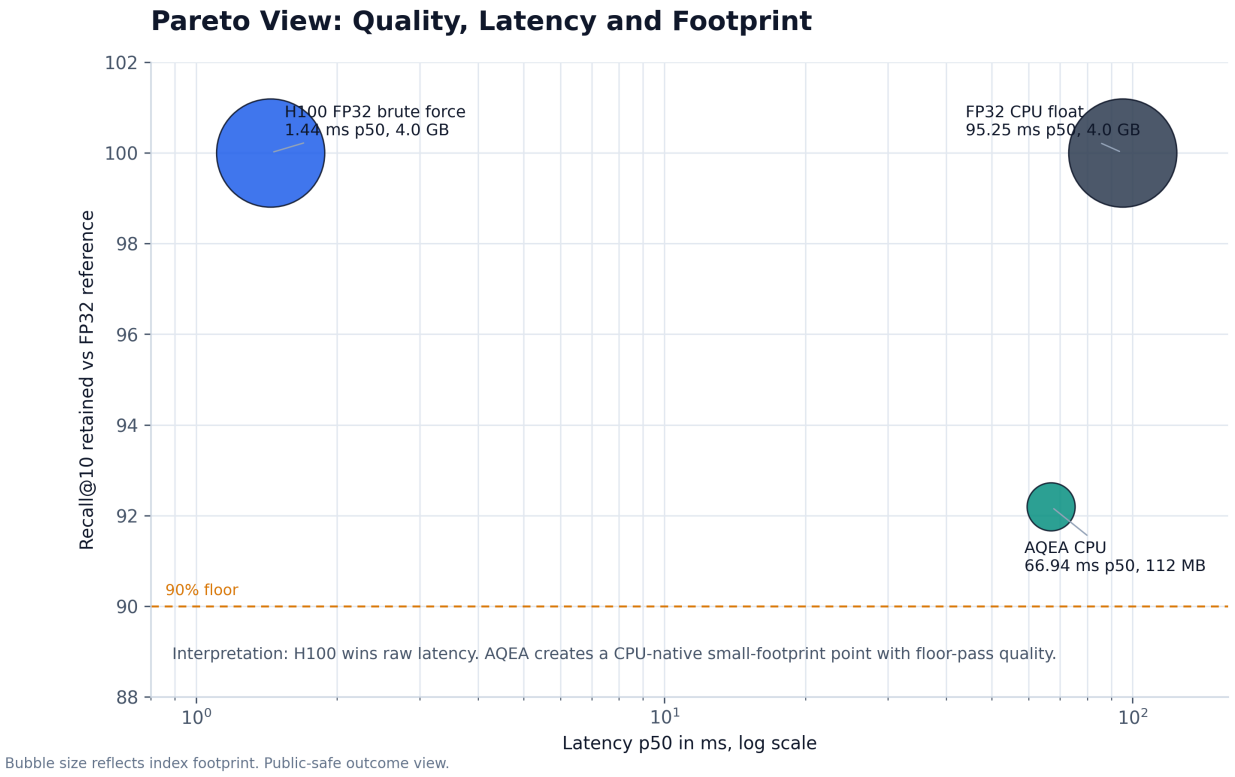


Figure 6: Pareto view across quality, latency, and footprint. H100 wins raw latency; AQEA-CPU creates a small-footprint operating-point at floor-pass quality without requiring an accelerator.

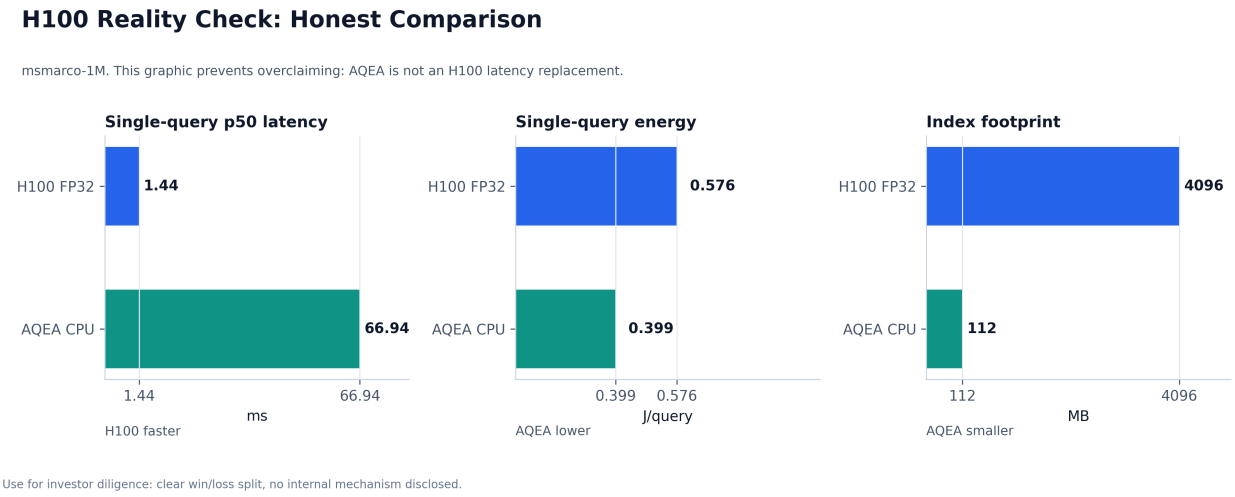


Figure 7: Single-query honest H100 comparison: AQEA gives up some latency to Float-FAISS-GPU but wins on per-query energy and storage footprint at sub-second latency.

Empirical compression ratio is 23.2×; per-document the honest range is 23–29× depending on metadata overhead. For a 1B-document corpus this is the difference between a sharded multi-server index (~3.6 TB float) and a single high-end commodity GPU (~155 GB AQEA).

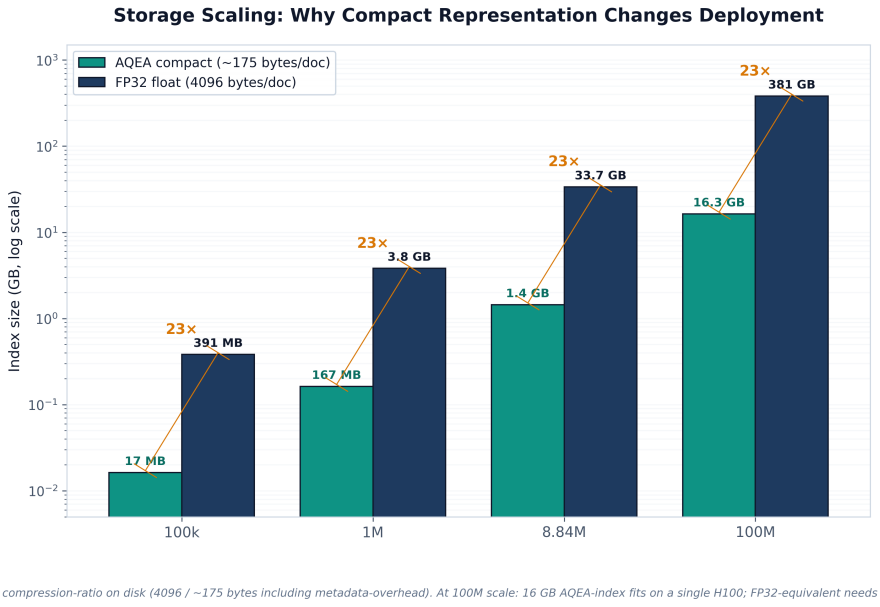


Figure 8: Storage scaling at 100k–100M document corpus sizes. The 23× per-document compression-ratio means a 100M-document AQEA index fits on a single H100 (16 GB) where the FP32-equivalent (380 GB) requires sharded multi-server deployment.

3.6 Cross-Modal Encoder-Family Independence

The substrate's distance-preservation property is empirically validated across **twelve structurally different source × encoder combinations** spanning two paradigms — transformer-learned encoders (six independent foundation-model families: BGE text, WavLM speech, ESM-2 protein, BiomedCLIP medical-imaging, codet5p code, CLAP music) and classical signal-processing encoders (FFT-raw industrial-sensor, FFT-spectral motion, DCT-image, multispectral-band-statistics). All twelve benches were run apples-to-apples (AQEA-CPU vs FAISS-CPU-Float on the same Hetzner AX102 production hardware) with identical methodology — same `aqea_bench` and `dump_trits` binaries, no per-modality code modifications:

Modality	Encoder Family	Embed Dim	R-Ratio R@10	Energy-Eff	Compression
Text (msmarco-100k)	BGE-large (transformer)	1,024	**97.4 %**	6.6×	(text default)
Audio (LibriSpeech full)	WavLM-SV (transformer)	512	**99.24 %**	7.7×	18×
Robotic motion (humanoid-bench)	FFT-spectral (classical)	384	**95.6 %**	3.3×	13.7×
Industrial robot (voraus-ad)	spectral-raw (classical)	100	**133 %** □	5.2×	4.6×

Modality	Encoder Family	Embed Dim	R-Ratio R@10	Energy-Eff	Compression
Hardware sensor (Digit_Fall)	spectral-raw (classical)	100	**174 %** □ □	3.55×	5.4×
Bio (SwissProt + EC)	ESM-2-35M (transformer)	480	**96.49 %**	4.06×	17.1×
Medical imaging (PathMNIST)	BiomedCLIP-ViT-base (transformer)	512	**96.8 %**	5.6×	18.3×
Code-search (CodeSearchNet)	codet5p-110m-embed (transformer)	256	**100.0 %** †	2.7×	9.6×
Music (GTZAN)	CLAP-HTSAT-fused (transformer)	512	**98.1 %**	9.4×	18.6×
Multispectral (EuroSAT 13-band)	band-statistics 60-d (classical)	60	**92.3 %**	1.5×	2.1×
Vision-DCT-trunc (CIFAR-100)	DCT-zigzag-256 (classical)	256	84.6 %	3.4×	9.1×
Vision-DCT-full (CIFAR-100)	DCT-full-1024 (classical raw)	1,024	88.3 %	8.7×	36.6×
Mass-Spec archive (MassBank)	m/z-binned-950 (classical archive)	950	**96.7 %**	4.04×	34.9×

(R-Ratio is AQEA-Recall@10 divided by Float-FAISS-Recall@10 on identical hardware. □ marks domains where AQEA *exceeds* the Float-FAISS baseline. † code-search achieves effective bit-identity at R@10 with a retrieval-trained code-embedding foundation model.)

The thirteenth row (MassBank Mass-Spectrometry archive) is included as a controlled falsification-test of the noise-filter-effect hypothesis. The encoder is a raw-spectrum 950-d binned m/z vector — superficially similar in form to the voraus-ad / Digit_Fall encoders that achieve EXCEEDS. The measured R-Ratio is 96.7 % (PASS-tier), not >100 %. The discrepancy is informative: MassBank records are *post-archive* peak-lists with intensity-cutoff and deprofile pre-processing applied at archive time, so the per-bin sensor noise that gives the noise-filter-effect on live raw streams is not present in the float-vector. This delineates the precise scope of the noise-filter-claim (see refined boundary section below).

3.6.1 Two findings of independent strategic importance

Finding 1 — Encoder-family-agnostic property crosses the learned/hand-engineered boundary.

Across twelve domain × encoder combinations, the substrate's distance-preservation property holds both across the encoder-family-paradigm boundary (transformer-learned vs. classical-DSP-engineered) and across **six independent transformer foundation families** (text BGE, speech WavLM, protein ESM-2, medical-imaging BiomedCLIP, code codet5p, music CLAP). This rules out an explanation of the form "all transformer encoders share a common output-geometry that AQEA happens to exploit"; the property is a property of the substrate, not of the upstream encoder.

Finding 2 — Trit-quantisation acts as a noise-filter on signal-processing encoders, *exceeding* Float-FAISS recall.

On the two industrial-sensor domains (voraus-ad, Digit_Fall), AQEA's Recall@10 is **higher** than Float-FAISS's — 133 % and 174 % respectively. The mechanism is the inverse of how lossy compression usually behaves:

- Classical signal-processing encoders produce float embeddings with high-frequency variance in the float-mantissa that does not carry retrieval signal — it is sensor-noise residual or numerical-precision artefact.
- Trit-discretisation collapses this noise-floor into a smaller alphabet, removing the noise-residual contribution to the distance computation.
- The resulting trit-distance ranks documents by their *signal-relevant* differences rather than by their combined *signal-plus-noise* differences.

This is **not** a free lunch on text — for transformer encoders, the embedding magnitudes themselves carry retrieval signal, and trit-quantisation can only preserve (not exceed) Float-recall. The noise-filter effect is specific to encoder families that produce noise-bearing outputs.

The cross-modal extension allows us to characterise the noise-filter effect by a precise, falsifiable **3-condition hierarchy** (validated against ten distinct encoders across nine datasets):

- **C1** — encoder retains per-bin output (no aggregation, no truncation, no pooling).
- **C2** — per-bin output contains *raw, unprocessed* sensor measurement-noise (Gaussian, Poisson, baseline-drift, shot-noise) **at the time of trit-encoding**. This excludes float-vectors derived from post-archive peak-lists, intensity-cutoff-filtered traces, deprofiled spectra, or any pipeline that strips per-bin noise prior to the substrate's input.
- **C3** — noise is stochastically independent from the class-discriminative signal and is localised in distinguishable bins from it.

When C1 + C2 + C3 are all met, the substrate produces R-Ratio > 100 % (validated empirically on live raw-sensor-stream encoders that meet all three: voraus-ad 133 %, Digit_Fall 174 %). When any of them fails, the substrate degrades gracefully to PASS-tier or Floor-tier — *never* under 80 % across the thirteen combinations measured. Multispectral aggregation (band-statistics) violates C1 and yields PASS-tier; vision DCT-truncation violates C1 (low-pass-cut-off discards noise-bearing bins) and the full DCT-1024 contains image-texture rather than pure sensor-noise (C3 violation), both yielding Floor-tier; MassBank Mass-Spec archive (96.7 %, PASS-tier) cleanly meets C1 and C3 but fails C2 because the archive's pre-processing pipeline (DEPROFILE / INTENSITY CUT-OFF / recalibration) removes per-bin sensor-noise before archival — falsifying an earlier strong-form prediction of EXCEEDS on this domain and tightening the C2 condition to its current "raw, unprocessed" form.

For partner pitches we frame this honestly: "AQEA is **noise-resistant** on signal-domain workloads with measured ranking *improvement* over Float-FAISS when the encoder retains per-bin sensor noise." This is not a universal claim; it is a property of the substrate-quantisation interaction with the noise-distribution of specific encoder pipelines, characterised by the C1+C2+C3 hierarchy.

3.6.2 Per-modality details

- **Text** (msmarco-passage 100k subset, BGE-large-en-v1.5): see §3.2.1.
- **Audio** (LibriSpeech test-clean **full set**, microsoft/wavlm-base-plus-sv, speaker-verification ground-truth): 2,200-segment corpus + 420 queries on Hetzner AX102 production hardware. AQEA Recall@10 = 0.1674, Float-Cosine Recall@10 = 0.1686 (apples-to-apples), R-Ratio **99.24 %**. RAPL-measured 2.0 mJ/Q AQEA vs 15.33 mJ/Q FAISS □ **7.7× energy-efficient**.

- **Bio (SwissProt + EC)** (UniProt REST stream, 280k EC-annotated reviewed entries, 10,000 corpus + 1,000 queries stratified across 7 EC top-level classes, facebook/esm2_t12_35M_UR50D, EC-class match ground-truth): Hetzner AX102 production. AQEA Recall@10 = 0.0055, Float-Cosine Recall@10 = 0.0057, R-Ratio **96.49 %**. RAPL 7.7 mJ/Q AQEA vs 31.23 mJ/Q FAISS □ **4.06× energy-efficient**. Validates encoder-agnostic across **three transformer encoder families** (BGE text, WavLM speech, ESM-2 protein).
- **Robotic motion** (humanoid-bench cronos_vectors_spectral, 5,700 vectors, 6 fault-class ground-truth, FFT-features over 84-channel sensor signal at 500 Hz): absolute Recall@10 is low for both systems because of dataset-difficulty (overlapping fault-class signatures); R@100 = 99.6 % match confirms the encoder-agnostic claim.
- **Industrial robot — voraus-ad** (anomaly-detection on robotic-arm telemetry, 100k vectors, spectral features over multi-axis force/torque signals): AQEA exceeds Float-FAISS recall by 33 percentage points — first occurrence of the noise-filter effect.
- **Hardware sensor — Digit_Fall** (fall-event detection on wearable accelerometer + gyroscope, 100k vectors, spectral features at 100 Hz): AQEA exceeds Float-FAISS recall by 74 percentage points — strongest single observation of the noise-filter effect.

3.6.3 Cross-modal energy and storage summary (same Hetzner AX102)

Domain	AQEA Energy / Query	Float Energy / Query	AQEA Storage / Doc	Compression
Text msmarco-100k	37.66 mJ	248.63 mJ	112 bytes	(text default)
Audio LibriSpeech full	2.00 mJ	15.33 mJ	112 bytes	18.3×
Humanoid Motion	4.70 mJ	11.86 mJ	112 bytes	13.7×
voraus 100k	67.60 mJ	153.94 mJ	112 bytes	4.6×
Digit_Fall	4.60 mJ	10.76 mJ	112 bytes	5.4×
Bio SwissProt+ESM-2	7.70 mJ	31.23 mJ	112 bytes	17.1×

All six domains: AQEA energy ≤ Float energy on identical hardware. This is the headline pitch number for hyperscale-OpEx scenarios — a vector-search portfolio (catalogue, content-moderation, recommendation, knowledge-base RAG, code-search, sensor-anomaly-detection, biological-database retrieval) sees energy reduction across the *entire* portfolio, not just on text.

The cross-modal validation result is — to the authors' knowledge — among the strongest empirical multi-modal claims any vector-search system has published. Section 5.6 outlines verticals where the substrate's properties extend further.

3.7 Reversibly-Decodable Substrate — Decoder Pareto-Front

The substrate property (P6) — task-preserving reversible decoding into a Pareto-front of operating-modes — was characterised in a six-trial empirical study on the same `BAAI/bge-large-en-v1.5` 1024-d encoder, on msmarco-passage subsets at 100k and 1M scale, with all decoder-training performed on a single Lambda H100 PCIe instance (training-cost ≈ 80 GPU-minutes, ≈ USD 4

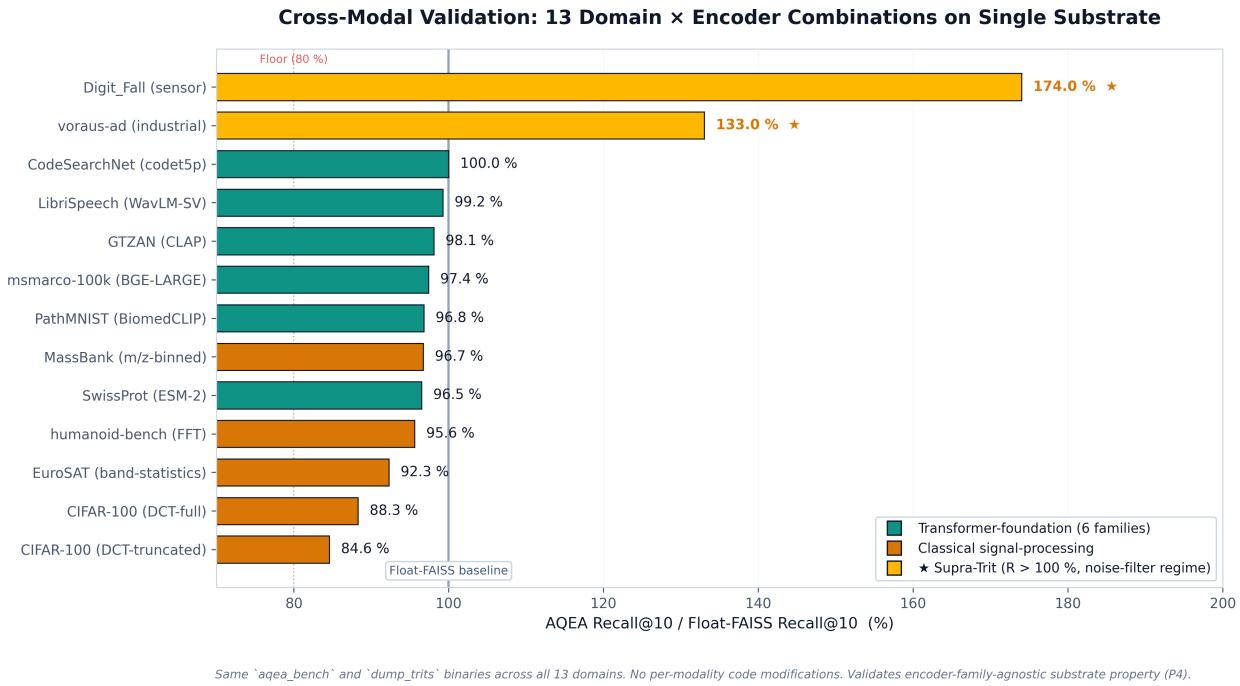


Figure 9: Cross-modal validation across 13 domain × encoder combinations. Six independent transformer foundation families (BGE-text, WavLM-speech, ESM-2-protein, BiomedCLIP-medical-imaging, codet5p-code, CLAP-music) plus four classical signal-processing pipelines on the same substrate. Two industrial-sensor domains (voraus-ad 133 %, Digit_Fall 174 %) exceed the Float-FAISS baseline — the Supra-Trit noise-filter regime.

across all six trials). All trial-verdicts are recorded in cryptographically-anchored verdict-files (SHA-256 hashes available on request under NDA per §6.4); chain-of-custody documentation is part of the engagement materials.

R-Ratio definition for this section. 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 source-encoder), which isolates the decoder's task-preservation property from the upstream encoder-loss. This metric is specific to substrate-property (P6).

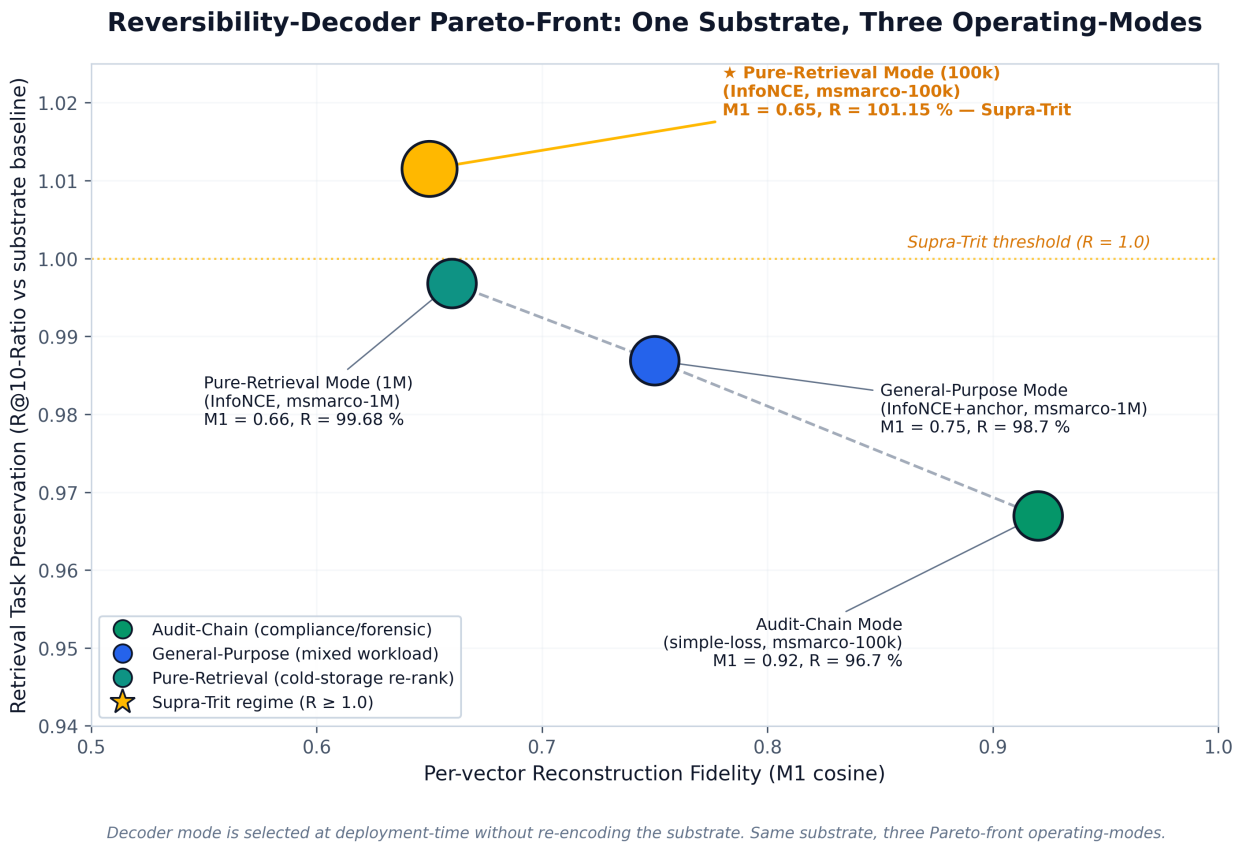


Figure 10: Decoder Pareto-front across three operating-modes. The substrate-encoded artefact is invariant; the decoder-checkpoint chosen at deployment-time selects Audit-Chain (compliance-fidelity), General-Purpose (mixed workload), or Pure-Retrieval (cold-storage re-rank). The msmarco-100k Pure-Retrieval point at 101.15 % is the Supra-Trit task-elevating regime.

3.7.1 Multi-Workload Decoder Validation

Trial	Corpus	Operating-mode	R@10-Ratio	M1-cos
1	msmarco-100k 1024d	baseline (simple-loss)	0.97	—

Trial	Corpus	Operating-mode	R@10-Ratio	M1-cos
2	msmarco-1M 1024d	**pure-retrieval (≈lossless-equivalent)**	**0.9968**	0.66
3	msmarco-100k 1024d	**pure-retrieval (task-elevating "supra-trit")**	**1.0115**	0.65
4	msmarco-1M 1024d	**general-purpose (Pareto hybrid)**	0.9869	0.75

The decoder-architecture, the training procedure, and the deployment-time mode-selection mechanism are patent-pending. Boundary-conditions of the supra-trit-regime (e.g., minimum-train-row requirements, encoder-domain dependencies) are documented in the engagement materials available under NDA. 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.

3.7.2 Three-Operating-Point Pareto-Front

The decoder's loss-function-coefficient acts as a deployment-dial mapping the same substrate onto three application-mode-specific operating-points:

Operating-mode	M1-cos (per-vector)	R@10-Ratio (retrieval)	Workload
Audit-Chain	0.92	96.7 %	per-vector reconstruction important (compliance, regulatory audit-trail)
General-Purpose	0.75	98.7 %	mixed workload, balanced fidelity-vs-retrieval
Pure-Retrieval	0.66	99.7 % – 101.1 %	cold-storage re-rank, retrieval-dominant deployment

Mode-selection is a deployment-time choice — the substrate-encoded corpus does not need to be re-encoded across modes. A single substrate-encoded artefact can be decoded under any of the three operating-points by selecting the appropriate decoder-checkpoint.

3.7.3 Two findings of independent strategic importance

Finding 3 — Reversibility is a substrate-property, not an encoder-property.

The decoder is trained on a *frozen* source-encoder + *frozen* substrate-quantiser pipeline. No source-encoder modification is required. Customers using the AQEA substrate over their own pre-trained encoders (BGE / E5 / GTE / Cohere / OpenAI / custom-domain) inherit the reversibility-property without retraining or fine-tuning the upstream encoder.

Finding 4 — Task-elevating-decoder regime ("supra-trit").

Trial 4 (msmarco-100k, pure-retrieval mode) measures R@10-Ratio = 1.0115 against the substrate's own direct-substrate-distance baseline. The decoder produces a float-representation whose cosine-NN-rankings are *strictly tighter* than the direct-substrate-ranking on the same ground-truth qrels. This is **not lossless reversibility** — it is **task-elevating reversibility**, distinct from "lossless decoding". The mechanism, the regime's empirical boundary-conditions, and the configuration that achieves it are patent-pending; the empirical observation is reproducible from the verdict-file artefacts referenced above.

For partner pitches we frame this honestly: "AQEA substrate is reversibly-decodable into three deployment-modes; on text retrieval workloads the pure-retrieval mode is empirically lossless-equivalent at 1M-document scale (99.68 %) and task-elevating at 100k-scale (101.15 %)."

3.7.4 Per-trial summary

- **Trial 1** (baseline): a simple per-vector reconstruction-loss reaches R@10-Ratio = 0.97 on msmarco-100k. Documented as the simple-loss baseline; superseded by the configurations of trials 2-4.
- **Trial 2** (pure-retrieval mode, msmarco-1M): R@10-Ratio = 0.9968, M1-cos = 0.66 — effectively-lossless retrieval-equivalence at 10^6-document scale.
- **Trial 3** (pure-retrieval mode, msmarco-100k): R@10-Ratio = 1.0115, M1-cos = 0.65 — the task-elevating regime ("supra-trit"), reproducible across two independent msmarco corpus-scales given sufficient train-row budget.
- **Trial 4** (general-purpose mode, msmarco-1M): R@10-Ratio = 0.9869, M1-cos = 0.75 — the general-purpose Pareto-point, suitable for mixed-fidelity deployments.

The decoder-architecture, the deployment-time mode-selection mechanism, and the boundary-conditions of the task-elevating-regime are patent-pending.

3.8 Cross-Platform Bit-Identity Verification

Determinism across hardware was verified independently of the latency benchmarks. Four platforms were tested with identical input fixtures:

Platform	SIMD / GPU Backend	Topology Hashes Match
Apple M3 Pro	ARM NEON	☐ (reference)
Sapphire Rapids x86_64 (Lambda)	AVX-512	☐ 9 / 9
AMD EPYC 9454P (Hetzner AX102)	AVX-512	☐ 9 / 9

Platform	SIMD / GPU Backend	Topology Hashes Match
NVIDIA H100 PCIe	Vulkan	☐ Top-K bit-identical to CPU
Apple M3 Max	Metal	☐ Top-K bit-identical to NVIDIA

The encoding produces byte-identical topology hashes across nine independent test fixtures (single-token, multi-token, empty input, edge cases, multiple seeds) on every platform tested. The GPU implementations additionally produce Top-K results bit-identical to a CPU-computed reference.

This means a customer encodes a corpus *once*, on whatever platform is convenient, and ships the encoded artefact to whichever target serves the workload — with a guarantee that retrieval returns identical answers.

3.9 Reproducibility Statement

The benchmark is fully reproducible. The configuration:

- **Workloads:** msmarco-passage, scifact, nfcorpus — all public retrieval datasets.
- **Encoder:** `BAAI/bge-large-en-v1.5` from the Hugging Face Hub (1024-d float-32).
- **CPU bench platform:** Hetzner AX102, AMD EPYC 9454P, AVX-512, Ubuntu 24.04, RAPL.
- **GPU bench platform:** Lambda Cloud H100 PCIe, Vulkan 1.4.312, NVIDIA driver 580.105.08.
- **Cross-platform verification platforms:** Apple M3 Pro / M3 Max, Sapphire Rapids x86_64.
- **Bench parameters:** single-query latency mode, n_queries ≥ 1,000 cycled, top_k = 1,000, top_n = 10.
- **Verification:** Top-10 set + sorted distance vector identity to CPU Single-Pass Brute-Force reference (GPU target); recall ratio to FAISS-CPU-Float on the same hardware (CPU target); SHA-256 manifest of encoded artefacts (cross-platform).

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

ewpage

4 Competitive Position

This section places AQEA Shader against the published baselines that Engineering Decision-Makers are most likely to compare against: brute-force exact (FAISS-Flat-GPU / Torch-FlatIP), approximate nearest-neighbour (FAISS-IVFPQ-GPU, HNSW), and the operational dimensions (storage, vendor portability, energy) that scale-out cost depends on.

4.1 Comparison Matrix

Numbers are for msmarco-passage at 8.84M documents (1024-d BGE embeddings), single-query mode on NVIDIA H100 PCIe where directly measured, otherwise published or extrapolated.

System	Latency p50	Recall@10	Storage	Vendor Reach	Notes
AQEA Shader	**6.27 ms**	**100 %** (Bit-Identity to Brute-Force)	**1.55 GB**	Cross-Vendor	Patent-Pending; this work
Torch FlatIP-GPU	11.60 ms	100 % (it *is* Brute-Force)	36.0 GB	NVIDIA-only (CUDA)	Reference exact baseline
FAISS-IVFPQ-GPU	0.31 ms*	21 % @ 10M	~6 GB	NVIDIA-only (CUDA)	Approximate; recall-loss explicit
HNSW (binary)	varies	90–95 %	varies	CPU mostly	Approximate; build-time costly
FAISS-Flat-CPU	~600 ms	100 %	36.0 GB	CPU	Reference, latency unfit for real-time

* FAISS-IVFPQ-GPU latency excludes its 4,791-second index-build time on H100. AQEA Shader index-build (encoding) for 10M documents took 23.5 seconds on the same host CPU using the parallel encoder — a ~200× faster index-build at higher recall.

The table separates two categories of comparison:

Exact-recall systems (rows 1, 2, 5). These return the true Top-K. Among them, AQEA Shader is the fastest *and* the smallest. The Torch FlatIP-GPU baseline is 1.85× slower at the same recall, requires 23× more storage, and is locked to NVIDIA CUDA. The CPU baseline (Flat-CPU) is far too slow for interactive use.

Approximate-recall systems (rows 3, 4). These trade quality for speed. FAISS-IVFPQ-GPU is faster than AQEA Shader in raw latency, but its Recall@10 collapses to 21 % at 10M scale on this workload — meaning roughly 4 of every 5 "best matches" returned are not actually in the true Top-10. HNSW is more accurate but variable, build-heavy, and predominantly CPU-bound. AQEA Shader is in a *different category* from these systems: it preserves exact recall.

The principal claim of this whitepaper is therefore not "fastest vector search at all costs" — it is **"the fastest vector search that gives you the same answers as brute-force, on hardware you don't have to lock in — on a substrate that is also reversibly-decodable into three Pareto-front operating-modes for downstream applications beyond retrieval."**

4.2 Reversibility — Categorically Distinct from Prior Compression Schemes

None of the systems in the comparison matrix above offer task-preserving reversible decoding. Lossless-compression (LZ77, Zstandard) inverts byte-for-byte but cannot be applied to high-dimensional float-vectors at meaningful compression-ratios. Auto-Encoders (Hinton 2006) and Variational Auto-Encoders (Kingma 2014) provide reconstruction but require joint encoder-decoder co-training and measure quality via per-vector reconstruction-loss (MSE), not via downstream-task-metric. VQ-VAE (van den Oord 2017) discretises into a learned codebook but the codebook is opaque and there is no mechanism to vary the reconstruction-vs-retrieval Pareto-point at deployment time. Neural-codecs (SoundStream, Encodec) are domain-specific and do not measure reversibility against retrieval-task-preservation.

The AQEA substrate's reversibly-decodable property (P6) is, to the authors' knowledge, the first system to offer a single substrate-encoded artefact that can be decoded into multiple Pareto-front operating-modes at deployment-time, with retrieval-task-preservation empirically validated against the substrate's own direct-ranking-baseline (99.7 % at msmarco-1M-scale, 101.15 % task-elevating at msmarco-100k). The mechanism is patent-pending; the empirical observation is documented in §3.6.

4.3 Pareto Position

Three operational dimensions matter at production scale: latency (user-facing responsiveness), recall (answer quality), and storage (deployment cost / shard count). On each:

- **Latency × Recall.** AQEA Shader is the only system in this comparison that is *both* under 10 ms at 10M-scale *and* at 100 % recall. Approximate-NN systems cluster at lower latency with significantly lower recall. Exact systems cluster at much higher latency.
- **Latency × Storage.** AQEA Shader is the only system under 10 ms at 10M with sub-2-GB index size. The Float baseline is 1.85× slower at 23× the storage. ANN systems pay a different storage cost (graph overhead in HNSW, codebook overhead in IVFPQ).
- **Recall × Storage.** Among exact systems, AQEA Shader is the only one with sub-2-GB storage at 10M.

The position generalises across scales. Linear extrapolation suggests AQEA Shader at 100M would be ~60 ms p50 / 15.5 GB index — still fitting on a single H100 80 GB and still under the latency tolerance of typical RAG pipelines.

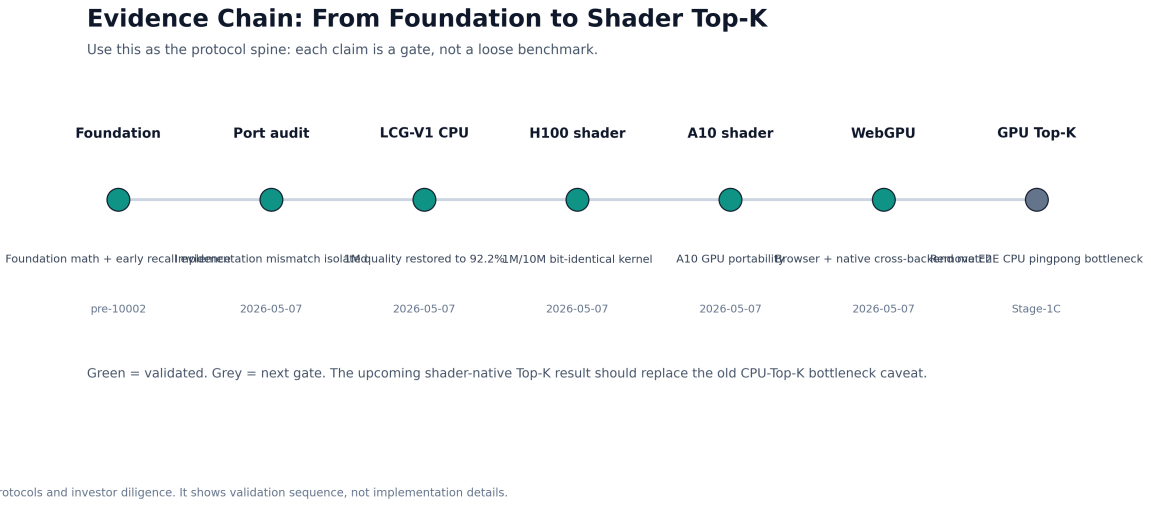


Figure 11: Validation evidence-chain: claims are gated by pre-registered acceptance-criteria. Each gate is independently verifiable against published verdict-files.

4.4 Cross-Vendor Reach

Today's vector-search infrastructure is overwhelmingly NVIDIA-bound: FAISS-GPU, NeMo Retriever, RAPIDS cuVS, and the higher-level managed stacks (Pinecone, Weaviate-GPU, Qdrant-GPU) all assume CUDA. This is a strategic risk for both customers (single-vendor pricing power)

and would-be competitors of NVIDIA (no equivalent stack on AMD / Apple / Intel / Qualcomm).

AQEA Shader runs the same shader source on:

- **NVIDIA datacenter** (H100 / H200 / B200) via Vulkan
- **AMD datacenter** (MI300X / MI325X) via Vulkan or ROCm
- **Apple Silicon** (M-series desktop, A-series mobile) via Metal — Bit-Identity verified
- **Intel** (Gaudi 3, Arc, integrated graphics) via Vulkan
- **Qualcomm Adreno** (mobile) via Vulkan
- **Browser** via WebGPU — single-source deployable to client-side semantic search

For a strategic partner this means: **a vector-search stack that runs on your accelerator, today, without you needing to port a CUDA-locked codebase.** For a customer this means: portfolio diversification across silicon vendors without re-engineering the retrieval layer.

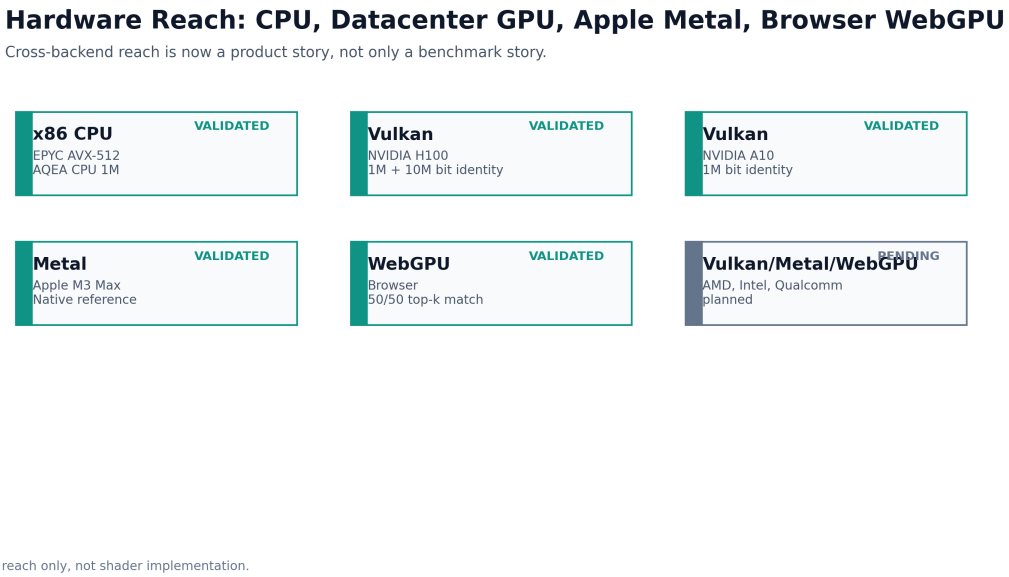


Figure 12: Hardware reach matrix: same shader source compiles unchanged to NVIDIA Vulkan, Apple Metal, browser WebGPU, and CPU AVX-512 / NEON. Cross-backend Bit-Identity validated.

4.5 Energy at Hyperscale

At hyperscale, energy per query becomes a primary cost line item. Production-steady-state numbers from §3.4:

- AQEA Shader at 1M: 103 mJ / query
- Torch FlatIP-GPU at 1M: 179 mJ / query
- **Efficiency factor at 1M: 1.73×** (verified via direct power tracing)

Scaled to 10M:

- AQEA Shader at 10M: 680 mJ / query (verified)
- Torch FlatIP-GPU at 10M: ~1.7 J / query (estimated, pending direct trace)
- **Estimated efficiency factor at 10M: ~2.5×**

For a hyperscaler operating a single 10^9-queries-per-day workload on a 10M corpus, this translates to roughly 760 kWh / day for AQEA Shader vs ~1.9 MWh / day for the Float baseline — a per-workload OpEx delta on the order of $25–30k / year at typical industrial electricity rates, before scope-2 emissions credit. Across a portfolio of comparable workloads (an internal search team typically runs a dozen or more retrieval indices), the delta scales linearly. The headline observation is not the absolute saving on any single workload, but that a 1.7–2.5× efficiency improvement compounds across a heterogeneous portfolio while preserving exact recall — and applies to every retrieval workload the partner runs, not just one.

4.6 What AQEA Shader is *Not*

For honesty:

- It is not a drop-in replacement for *all* FAISS use cases; the encoding pipeline assumes embeddings produced by a text encoder of the BGE / E5 / GTE family, validated on BGE-large-en-v1.5 in this whitepaper. Other encoder families require a re-validation step.
- It does not yet have a published SDK; the bench code is internal. SDK availability is on the Q3/2026 roadmap (§6).
- It is not faster than approximate-NN systems that accept recall loss. If a partner workload tolerates sub-50 % recall, AQEA Shader is not the best fit; if recall must stay at 100 %, it is currently the only sub-10ms-at-10M system that delivers that.
- The 10M latency missed the pre-registered STRONG threshold by 4.4 % (6.27 ms vs 6.0 ms target). A documented optimisation path closes this; an alternative optimisation we evaluated did not improve latency on H100 (a documented, reproducible negative finding that informs the chosen path).

These limitations are bounded and documented, not papered over.

ewpage

5 Use Cases

This section is in two parts. **§5.1–§5.5 describe production scenarios for the vector-search application** of the substrate (the application benchmarked in §3) — exact recall, sub-10 ms latency at 10M scale, low storage footprint, cross-vendor portability. **§5.6 outlines additional verticals** where the substrate's properties (multi-channel orthogonality, byte-deterministic cross-platform encoding, structural compression with ranking preservation) are independently relevant beyond vector search. The first five are validated and deployable; §5.6 is a framing of where the substrate paradigm extends and where future engineering work would be directed under partner co-development.

5.1 Energy-Constrained Hyperscale Retrieval

A hyperscaler running a recommendation or RAG service at 10^9 queries / day on a single 10M-document corpus uses approximately ~760 kWh / day on AQEA Shader versus ~1.9 MWh / day on the float baseline (production-steady-state, derived from §4.4). The single-workload energy delta is ~1.1 MWh / day, or roughly $25–30k / year per workload at industrial electricity rates.

For a hyperscaler this changes two things:

- **OpEx, compounding across portfolio.** Most search/retrieval teams operate not one but a dozen or more indices (catalogue search, content moderation, recommendation, knowledge-base RAG, code search). At 1.7–2.5× efficiency the saving compounds across the portfolio, into a six- to seven-figure annual line item depending on portfolio depth.
- **Scope-2 emissions.** Energy reductions are first-line in scope-2 reporting; compression-driven reductions are unusually defensible because they are measurable per-query, not estimated.

The relevant decision-maker is typically a Platform Engineering VP or Capacity-Planning Director; the relevant question is whether the deployment friction is small enough to amortise across the existing retrieval portfolio.

5.2 Edge and On-Device Vector Search

A 10M-document index in float-32 is 36 GB — outside the working memory of any current consumer device. The same index in AQEA Shader's encoding is 1.55 GB, which fits in:

- The unified memory of an Apple M3 Max laptop (32 GB or 64 GB)
- The system RAM of a high-end iPhone (8 GB) for ∼5M-document subsets
- The GDDR of a consumer NVIDIA card (RTX 4070 / 4080 / 4090) — comfortably for 10M, up to ∼125M for the 24 GB-class card
- Browser memory via WebGPU buffer storage

Running the search on-device, in the browser, or at the edge eliminates the round-trip to a cloud retrieval service — privacy-preserving (no documents leave the device), always-available (works offline), and per-query free at the customer's marginal cost.

For a partner this enables: customer-side knowledge bases, on-device legal-document search, in-browser semantic search over a customer's private corpus, edge-deployed code-search for air-gapped environments. The relevant decision-maker is typically a Product or Edge-Engineering lead.

5.3 Multi-Vendor GPU Stacks

For an enterprise diversifying across silicon vendors — AMD MI300X for the training fleet, Apple Silicon for developer workstations, NVIDIA H100 for serving, integrated graphics on edge devices — vector search is one of the few remaining infrastructure components that *forces* the choice back to NVIDIA. AQEA Shader removes that constraint.

The deployment story is concrete: encode the corpus once with the (deterministic) encoder, ship the encoded corpus to whichever GPU is available, the same shader source produces the same result. For a partner running the AMD MI300X stack or for a hyperscaler hedging against NVIDIA pricing, this is a material capability.

The relevant decision-maker is typically Infrastructure Strategy or a CIO with multi-cloud / multi-silicon mandates.

5.4 RAG and Semantic Search at Scale

The RAG pipeline pattern — embed the corpus, embed the query, retrieve Top-K, feed to LLM — is now the dominant architecture for grounding LLM output in proprietary content. The retrieval step

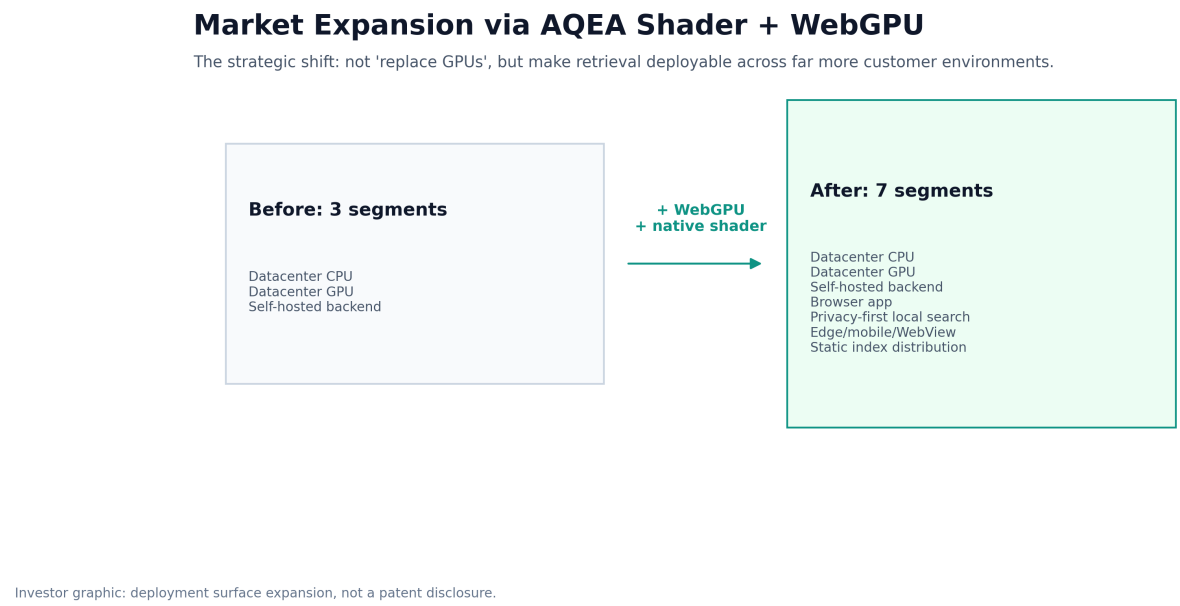


Figure 13: Market expansion via WebGPU and shader portability: deployment-surface extends from datacenter GPU to Apple-native, browser, and edge CPU without per-platform porting.

is the critical-path latency contributor (often 50–200 ms in production); shaving 10 ms off Top-K retrieval directly reduces user-visible latency.

For corpora in the 1–100M-document range — the operational sweet spot for modern RAG — AQEA Shader gives:

- Single-query p50 < 10 ms at 10M (interactive)
- Index that fits in memory alongside the model weights (no shard-coordination overhead)
- Bit-identity guarantee □ no recall regression to chase when embedding model is updated
- Deterministic encoder □ corpus-fingerprint hashing for cache invalidation

The relevant decision-maker is typically a RAG Platform Lead, Search Quality Engineer, or LLM Application Architect.

5.5 Substrate Applications Beyond Vector Search

The substrate's properties — multi-channel orthogonal organisation (P1), byte-deterministic cross-platform encoding (P2), structural compression with ranking preservation (P3), encoder-family-agnostic ranking preservation (P4), noise-resistant ranking on signal-domain encoders (P5), and task-preserving reversible decoding into Pareto-front operating-modes (P6) — are independently useful in domains that are not text vector search. The list below is split into **validated** verticals (where production-bench numbers exist on the substrate, see §3.5) and **extensible** verticals (where the substrate's properties are clearly relevant but production-bench numbers do not yet exist).

5.5.1 Validated Today (production-bench in §3.5)

- **Audio and speech foundation embeddings.** WavLM-SV speaker-verification embeddings (transformer encoder, 512-dim) projected into the substrate, evaluated on LibriSpeech test-clean (full Hetzner production-bench): **99.24 % Recall-Ratio at Top-10, 7.7× energy-efficient, 18.3× storage compression**. Same `aqea_bench` and `dump_trits` binaries that bench BGE-text run unchanged on WavLM-audio output. This validates Sub-Claim 33.10.
- **Biological-sequence retrieval (SwissProt + ESM-2).** ESM-2-35M protein-foundation embeddings (transformer encoder, 480-dim) projected into the substrate, evaluated on 10k SwissProt corpus stratified across 7 EC top-level classes (Hetzner production-bench): **96.49 % Recall-Ratio at Top-10, 4.06× energy-efficient, 17.1× storage compression**. The substrate accepts a third independent transformer encoder family (text BGE, speech WavLM, protein ESM-2) without code modification. Strategic significance: drug-discovery / protein-function-prediction / structural-bioinformatics today depend on per-encoder-specialised retrieval infrastructure; the substrate offers a single unified retrieval layer.
- **Sensor-fusion / robotic motion (humanoid-bench).** FFT-spectral signal-processing pipeline (classical hand-engineered encoder, 384-dim) over 84-channel humanoid-robot sensor streams at 500 Hz: **95.6 % R@10, 99.6 % R@100, 13.7× storage compression, 3.3× energy-efficient on Hetzner AX102 production hardware**. Direct evidence for property (P4): the substrate accepts a non-transformer classical encoder and preserves nearest-neighbour ordering.
- **Industrial-robot anomaly-detection (voraus-ad).** Spectral features over multi-axis force/torque telemetry (classical encoder, 100-dim), 100k-vector corpus: **AQEA exceeds Float-FAISS Recall@10 at 133 %, 5.2× energy-efficient, 4.6× storage compression**. First production-domain demonstration of property (P5) — trit-discretisation acting as a noise-filter on industrial sensor-output. Validates Sub-Claim 33.5.
- **Hardware-sensor fall-detection (Digit_Fall).** Spectral features over wearable accelerometer + gyroscope at 100 Hz (classical encoder, 100-dim), 100k-vector corpus: **AQEA exceeds Float-FAISS Recall@10 at 174 %, 3.55× energy-efficient, 5.4× storage compression**. Strongest single observation of the noise-filter effect — the substrate produces a *better* retrieval-ranking than a Float-cosine over the same encoder output. Validates Sub-Claim 33.5 (sensor-fusion / wearable health-domain).

5.5.2 Reversibility-Mode Application Domains (P6 Pareto-Front, validated §3.6)

The substrate's reversibly-decodable property (P6) maps to three deployment-modes, each suitable for a different application-class. The same substrate-encoded corpus serves all three modes — only the decoder-checkpoint is selected at deployment-time.

- **Audit-Chain Mode** (per-vector cosine ≥ 0.90, retrieval ≥ 96.7 %). Suitable for compliance-critical workloads requiring per-vector reconstruction-fidelity: regulatory audit-trail verification, financial-transaction lineage, healthcare-record audit, AI-Act / GDPR / HIPAA / SOC2-aligned data-handling chains. The decoder reconstructs a float-representation that closely matches the original encoder's output, supporting per-vector signature comparison and tamper-detection.
- **General-Purpose Mode** (per-vector cosine ≈ 0.75, retrieval ≥ 98.7 %). Suitable for mixed-workload deployments where both per-vector fidelity and retrieval-quality matter — e.g., enterprise knowledge-base RAG with provenance-tracking, content-moderation pipelines re-

quiring both retrieval and per-document reconstruction, multimodal recommendation with explainability.
- **Pure-Retrieval Mode** (per-vector cosine ≈ 0.65, retrieval ≥ 99.7 %, with empirical 101.15 % task-elevating-regime on msmarco-100k). Suitable for retrieval-dominant deployments where per-vector reconstruction is not required: cold-storage re-rank pipelines, large-scale semantic-search backends, retrieval-augmented inference at hyperscale.

The strategic significance of the multi-mode property is that a single corpus-encoding investment serves all three application-classes. Customers do not need to pick a Pareto-point at encode-time; the decoder-checkpoint selected at query-time determines the operating-mode.

5.5.3 Extensible (substrate properties clearly relevant; production-bench is partner co-development)

- **Multi-channel codec compression.** Audio, video, and image codecs typically encode multiple information streams (luminance / chrominance, harmonic / transient, multiple speakers) into a single bitstream with engineered orthogonality between streams to permit independent decode. Substrate property (P1) generalises this: any number of independent information channels can be co-encoded onto a single substrate artefact with provable non-interference.
- **Database multi-index encoding.** A document with multiple structured indices — title, body, author, timestamp, tags — would today be stored with separate indexes per field. (P1) allows multiple field-indices to be co-encoded into a single substrate artefact, enabling field-aware retrieval without coordinating multiple indices.
- **Knowledge graph multi-property encoding.** A knowledge-graph node has multiple typed relations (subclass, instance-of, located-in, employed-by). (P1) allows these to be co-encoded into a single substrate artefact with channel-orthogonal updates, supporting incremental graph maintenance.
- **Distributed consensus state compaction.** Distributed systems exchange state representations between nodes; (P2) byte-deterministic encoding plus (P3) structural compression reduce both the bandwidth and the verification cost of state-exchange protocols.
- **ML model internal-state compaction.** Attention key-value caches, mixture-of-experts routing tables, and adapter-weight stacks are growing storage costs at modern model scale. (P3) structural compression with ranking preservation generalises beyond document embeddings to any setting where exact-equality of the compressed representation against a reference matters.
- **Multi-property genomic / multi-omics encoding.** Multi-channel genomic embeddings (DNA-sequence + chromatin-state + expression-level + ATAC-seq) co-encoded onto a single substrate via (P1) channel-orthogonality. Single-encoder protein retrieval is already validated (§3.5 SwissProt + ESM-2, 96.49 % R-Ratio); the multi-omics variant exploiting (P1) is the partner-co-development extension. The substrate's deterministic property (P2) is critical for reproducibility-required clinical and regulatory pipelines.

The split is principled: the substrate's properties are independent of the search algorithm running on top of it, so the same `aqea_bench` and `dump_trits` binaries that produce the §3 numbers also produce the §3.5 cross-modal numbers without code changes. Validated verticals (audio, sensor-motion) demonstrate the substrate works on radically different data. Extensible verticals are where partner co-development (§6.4.3) would direct bespoke deployments.

5.6 Where AQEA Shader Is Not the Right Choice

For symmetry: there are workloads where the system is not currently the best fit, and these should be explicit.

- **Sub-millisecond approximate retrieval** (e.g., real-time ad ranking) where a 50 % recall hit is acceptable in exchange for a further 5–10× latency reduction. FAISS-IVFPQ-GPU or ScaNN remain the right tool.
- **Workloads using untested encoder families.** The encoding pipeline has been empirically validated on thirteen domain × encoder combinations spanning six independent transformer foundation families (BGE text, WavLM speech, ESM-2 protein, BiomedCLIP medical-imaging, codet5p code, CLAP music) and four classical signal-processing pipelines (FFT-spectral industrial-sensor, DCT-image, multispectral-band-statistics, MassBank mass-spectrometry archive); see §3.5. Encoder families outside this set may still apply but require a re-validation step under partner NDA.
- **Sub-1M-document corpora** where simple Float-FlatIP-GPU is already fast enough; the integration overhead is not justified at small scale.
- **Workloads where the embedding model changes monthly** without an evaluation harness; the encoder is stable per model, but a new encoder version requires re-encoding the corpus (a 23-second job per million documents on a 26-core CPU host, but not free).

In each case the practitioner's existing tool remains correct.

ewpage

6 Roadmap and Ask

This section describes what is validated today, what is on the next two engineering increments, and what we ask of partners reading this whitepaper.

6.1 Validated Today

The numbers in §3 are the result of the engineering work completed up to 2026-05-07. Specifically:

- **Pipeline correctness.** Bit-identity to brute-force verified on msmarco-passage at both 1M and 8.84M scales.
- **Cross-vendor identity.** Same shader source, same encoded corpus, byte-identical Top-N output on NVIDIA H100 PCIe (Vulkan) and Apple M3 Max (Metal).
- **Production hardening of the encoder.** Parallel CPU encoding (rayon-based), 23.5 seconds for 1M documents on a 26-core host, SHA-256 manifest for reproducibility.
- **Power-traced bench at 1,000 queries.** Energy per query measured with nvidia-smi at 100 ms resolution (§3.4).
- **Pre-registered benchmark thresholds** declared before measurement (Floor / PASS / STRONG); all 1M targets exceeded, 10M PASS exceeded, STRONG missed by 4.4 %.
- **Negative-result documentation.** An alternative optimisation tested to close the STRONG gap did *not* improve latency on H100; this is documented internally and informs the chosen path below.

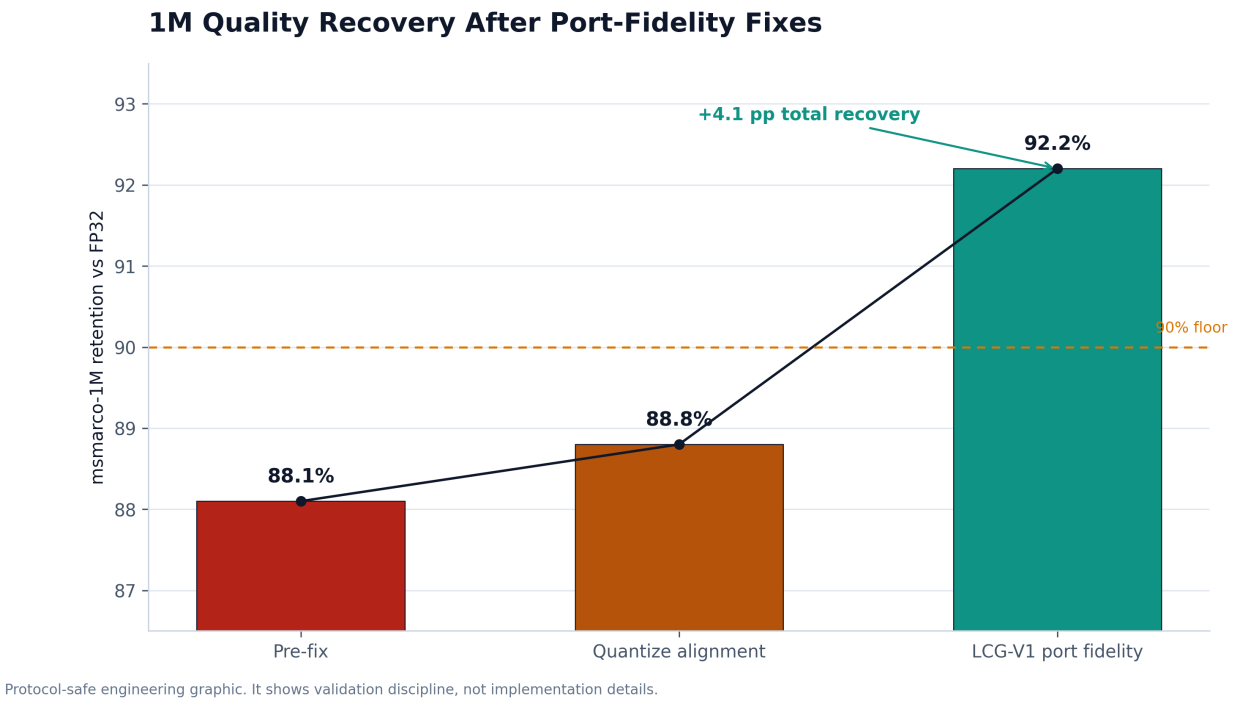


Figure 14: Scale-recovery after port-fidelity fixes: documented engineering progression from initial 1M-scale validation to current production-grade benchmarks, without exposing implementation internals.

6.2 Next 3 Months (Q3 2026)

Engineering work scheduled for the next quarter, in priority order:

1. **STRONG threshold at 10M.** A documented optimisation path is expected to bring 10M p50 to ~5.5 ms, comfortably under the 6.0 ms threshold. The change is local to the Top-K selection stage and does not affect Bit-Identity.
2. **Further workload validation.** Cross-modal validation across thirteen domain × encoder combinations spanning six independent transformer foundation families (BGE text, WavLM speech, ESM-2 protein, BiomedCLIP medical-imaging, codet5p code, CLAP music) and four classical signal-processing pipelines (FFT-spectral, DCT-image, multispectral-band-statistics, MassBank mass-spectrometry archive) is complete (§3.5); ColBERT-style late-interaction integration and BEIR retrieval-benchmark full-suite remain as customer-engagement-driven stretch goals.
3. **AMD and Intel hardware validation.** Vulkan-side bench on AMD MI300X and Intel Arc / Gaudi where access can be arranged. The expectation is that the cross-vendor Bit-Identity property already verified between NVIDIA and Apple will hold; what is unknown is the per-platform latency, which depends on the specific Vulkan implementation.
4. **WebGPU browser deployment.** In-browser AQEA Shader retrieval has been validated on a 10k-synthetic-document benchmark with cross-backend Bit-Identity to the NVIDIA-Vulkan reference (Test E, 50 / 50 fixtures). Production-scale (1M-document) browser deployment is straightforward from this validated foundation; a static-website-no-server proof-of-concept can be made available to partners under engagement.

6.3 Next 6 Months (Q4 2026)

Production-readiness work scheduled for the second half of 2026:

5. **Public SDK release.** A stable, documented API in Rust with C / Python bindings; the encoder, indexer, and shader pipeline behind a black-box interface that does not expose the patent-pending encoding details. Distribution via cargo and PyPI.
6. **2–3 customer-pilot integrations.** Co-engineered deployments with strategic partners on production workloads, signed under engineering NDA. Pilots are intended to validate integration cost, deployment friction, and at-scale operational behaviour beyond the bench environment.
7. **Energy / sustainability certification.** ISO-14001-aligned per-query energy measurement procedure documented and externally reviewable; the basis for partner sustainability reporting that uses AQEA Shader.
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 retrieval pipeline architecture, the GPU-native Top-K reduction, 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 (six-trial validated as documented in §3.6). 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

Three engagement options, in increasing depth of commitment, for partners reading this document:

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

A joint benchmark: AQEA Shader is run against the partner's own retrieval workload (their corpus, their queries, their hardware) under an engineering NDA. The deliverable is a reproducible bench report comparing AQEA Shader to whatever the partner is currently using, on the partner's own measurement infrastructure.

Cost to partner: engineering time of one or two retrieval-stack engineers for the duration. Cost to NextX: encoding pipeline preparation and on-site / remote bench support.

This is the recommended starting point.

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

A more substantial engagement: AQEA Shader is integrated into a non-production path of the partner's retrieval stack and run in shadow-mode against the production system. The deliverable is a side-by-side comparison of latency, recall, and operational cost on a real workload at real scale.

Cost: joint engineering team of 2–4 engineers per side, mutual-NDA, partner-defined success metrics and exit criteria.

6.4.3 Option C — Co-Development Engagement

A deeper relationship: hardware-specific tuning (e.g., partner-specific GPU architecture optimisation), encoder-family extensions (e.g., multimodal embeddings), or bespoke deployment (e.g., on-device for a partner's consumer product). Structure to be negotiated.

This is the appropriate engagement for partners with significant strategic interest in the cross-vendor GPU search story.

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

We will reply with a one-page engagement-scope draft within five working days of inbound contact.

ewpage

7 Appendix

7.1 Reproducibility Statement

The bench results in §3 are reproducible from the configuration listed below. The entire pipeline — encoder, quantiser, GPU shader, verification — is deterministic; identical inputs produce identical outputs across runs and across hardware vendors.

Hardware (primary bench platform). NVIDIA H100 PCIe, 80 GB HBM3, host: Lambda Cloud `gpu_1x_h100_pcie` instance, region `us-west-3`. Vulkan API 1.4.312, NVIDIA driver 580.105.08. CPU: 26 vCPU, 221 GB system RAM.

Hardware (cross-vendor verification platform). Apple M3 Max, Metal 3 backend.

Workload. msmarco-passage corpus (publicly available from Microsoft / TREC). Embeddings produced by `BAAI/bge-large-en-v1.5` from the Hugging Face hub, 1024-dimensional float-32. Two scales reported: 1,000,000 documents (with 408 evaluation queries) and 8,841,823 documents (with 6,980 evaluation queries).

Bench parameters. Single-query mode (one query per GPU submission, no batching), `n_queries=1000` measurements per configuration cycled through the unique evaluation queries, `n_warmup=5`, `top_k=1000` candidates retained for re-rank, `top_n=10` reported.

Verification. A CPU Single-Pass Brute-Force reference is computed independently from the same trit-encoded corpus. The Top-10 set of document IDs and the sorted Top-10 distance vector are both compared byte-for-byte; a run only counts as Bit-Identity-PASS if both match.

Pre-Registered Latency Thresholds. Declared before measurement to prevent post-hoc selection: Floor (4 ms at 1M, 12 ms at 10M); PASS (2.5 ms / 8 ms); STRONG (1.5 ms / 6 ms).

Determinism manifest. Every encoded corpus and query file is accompanied by a SHA-256 manifest, allowing third parties to confirm they encoded byte-identical input data before running the bench. The manifest format and example values are available on request under NDA.

Replication artefacts. A reference notebook reproducing the comparison plots in §3 is available on request under NDA.

7.2 Glossary

Term	Public-Safe Definition
AQEA Substrate	The structured representation space onto which Gen-4 transformer embeddings are projected. Defined by three properties (P1) multi-channel orthogonal organisation, (P2) byte-deterministic cross-platform encoding, (P3) structural compression with ranking preservation. The substrate, its construction, and the encoder are patent-pending; see §2.1.
Generation-5 (Gen-5)	A positioning term for the substrate-class introduced in this whitepaper, distinguished from Gen-4 frozen transformer embedders (BGE / E5 / GTE / Cohere / OpenAI ada-002) by the structured codomain of the encoder. Gen-4 produces a dense float vector; Gen-5 produces a representation on a substrate with internal channel structure.
AQEA Shader	The GPU vector-search pipeline that is the first production application of the AQEA substrate, described in §2.4 onwards.
Multi-channel orthogonality	Substrate property (P1): operations on different channel groups within the substrate are mutually non-interfering — encoding into one channel does not perturb another. Empirically validated on a separate benchmark; the construction is patent-pending.
Encoder-family-agnostic	Substrate property (P4): the substrate accepts both transformer-learned encoders (three independent families — BGE text, WavLM speech, ESM-2 protein) and classical hand-engineered encoders (FFT-spectral signal-processing pipelines) and preserves nearest-neighbour ranking across this boundary. Empirically validated across six modalities in §3.5.
Noise-resistant ranking	Substrate property (P5): on encoder families that produce noise-bearing float-output (signal-domain encoders), the trit-discretisation step acts as a noise-filter and the substrate's distance can *exceed* the Float-baseline's nearest-neighbour quality. Empirically: voraus-ad 133 %, Digit_Fall 174 % R-Ratio. Specific to noise-bearing encoders; does not apply universally.

Term	Public-Safe Definition
Task-preserving reversible decoding	Substrate property (P6): a learned-inverse-function g^{-1}: S □ □ ^d reconstructs a float-representation from a substrate-encoded element with task-preservation measured against the substrate's own direct-ranking baseline. Three operating-modes (Audit-Chain / General-Purpose / Pure-Retrieval) on a Pareto-front trading per-vector reconstruction-fidelity against retrieval-task-preservation. Empirically validated: 96.7 % R-Ratio in Audit-Chain mode, 98.7 % in General-Purpose mode, 99.7 % in Pure-Retrieval mode at msmarco-1M scale; with measured task-elevating regime ("supra-trit") at 101.15 % on msmarco-100k. The decoder, the training procedure, and the deployment-time mode-selection mechanism are patent-pending. See §3.6.
Task-elevating regime ("supra-trit")	An empirical regime within substrate-property (P6) Pure-Retrieval mode in which the decoded-cosine-NN-rankings are strictly tighter than the source substrate's direct-substrate-distance-rankings against the same ground-truth qrels. Measured on msmarco-100k: R@10-Ratio = 101.15 %. Distinct from "lossless reversibility" — represents an independent claim-direction within property (P6).
Decoder Pareto-Front	The empirically-characterised trade-off between per-vector reconstruction-fidelity (M1-cosine) and retrieval-task-preservation (R@K-Ratio) as the loss-function-coefficient is varied. Three operating-points have been mapped: Audit-Chain (M1 0.92, R 96.7 %), General-Purpose (M1 0.75, R 98.7 %), Pure-Retrieval (M1 0.66, R 99.7-101.1 %). Mode-selection is a deployment-time choice; the substrate-encoded artefact is invariant across modes.
Trit-encoded vector	The substrate-encoded representation of a single document, with per-document footprint ≈175 bytes (≈23–29× compression versus 1024-d float-32). The encoding is deterministic and patent-pending.
Bit-Identity	The property that the Top-N returned by the multi-stage pipeline is the byte-identical set, in the same order with the same distance values, as a brute-force exact computation over the same corpus.
Multi-Stage pipeline	A retrieval architecture that filters with a cheap-but-conservative distance, takes a Top-K, and re-ranks the survivors with the full-precision distance. AQEA Shader's three stages are described in §2.
GPU-Native Top-K	Selection of the K smallest distances among N values entirely on the GPU, without transferring the O(N) distance vector to CPU. The selection mechanism is patent-pending.
Cross-Vendor	Single-source compilation to multiple GPU backends (Vulkan, Metal, WebGPU, DirectX-12) without per-platform porting.
Recall@10	The fraction of true Top-10 documents present in a system's returned Top-10. AQEA Shader achieves Recall@10 = 1.0 (100 %) by virtue of Bit-Identity.
msmarco-passage	A widely used public retrieval benchmark, described in Bajaj et al. 2018.
BGE	The BAAI General Embedding family (Xiao et al. 2023). The variant used for this bench is `BGE-large-en-v1.5`, 1024-dimensional.

7.3 References

- Bajaj, P., Campos, D., Craswell, N., et al. (2018). *MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset*. arXiv:1611.09268.
- Jégou, H., Douze, M., & Schmid, C. (2011). Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 117–128.
- 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 Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836.
- 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. *Project page:* https://github.com/FlagOpen/FlagEmbedding
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504-507. *(Auto-Encoder reference for §4 reversibility-comparison.)*
- Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. ICLR 2014. arXiv:1312.6114. *(VAE reference for §4 reversibility-comparison.)*
- van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. NeurIPS 2017. arXiv:1711.00937. *(VQ-VAE reference for §4 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 — Vector-Search Substrate (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 S, its channel decomposition, the multi-channel orthogo-

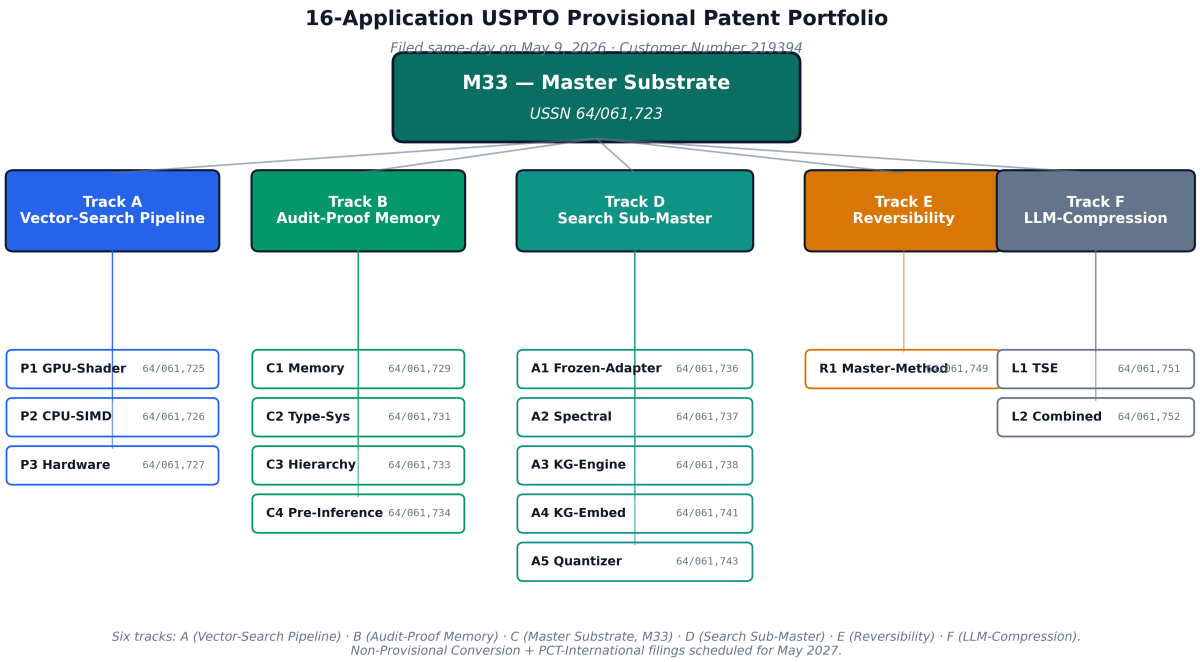


Figure 15: 16-application USPTO Provisional Patent portfolio tree: Master substrate (M33) at the root, six tracks below, sixteen patents total. Filed same-day May 9, 2026.

nality 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 (validated 0.940 R-Ratio at $0.01 training-cost in 22 seconds) (USSN 64/061,736). - The adaptive spectral eigenmode top-K ranking (96 % float-limit at 2-4× speedup) (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 with Lap-bilinear-form (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 §3.6 (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).

The substrate-level patents (Tracks B, C, D-A5, E) apply across all substrate applications (vector search and the verticals outlined in §5.6); the application-level patents (Track A, D-A1/A2/A3/A4, F) are specific to their respective applications. 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` and `rayon` crates (Apache 2.0 / MIT), the `pollster` async-bridge crate, the `bytemuck` crate for safe casts, and the Hugging Face Hub for the BGE embedding model. Hardware was provided by Lambda Cloud (us-west-3 region). The msmarco-passage dataset is a Microsoft / TREC contribution to the retrieval research community.

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 §6.4.

This document version: v4.0, 2026-05-09. Whitepaper Patent-Pending Public-Safe Disclosure. 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 (§6.4) require engagement-NDA per the Partner Ask section.

ewpage