SPECTRA-MoE: Signal-Processing and High-Dimensional Reduction for Local Compression of Frontier Mixture-of-Experts Language Models

Andrew Kiruluta
UC Berkeley, School of Information
kiruluta@berkeley.edu

August 20, 2026

Abstract

Frontier large language models increasingly combine trillion-scale mixture-of-experts (MoE) parameter counts, multimodality, long-context mechanisms, and agentic post-training. Kimi K3 is now an open-weight, 2.8-trillion-parameter multimodal model with 104 billion activated parameters per token, 93 transformer layers, 69 Kimi Delta Attention (KDA) layers, 24 Gated Multi-Head Latent Attention (MLA) layers, 896 routed experts with 16 selected per token, two shared experts, and a 1,048,576-token context window [28, 29]. Its official Hugging Face repository is approximately 1.56 TB and contains 96 Safetensors weight shards, while the public configuration confirms MXFP4-packed routed linear weights together with higher-precision or excluded components [27, 29]. A 32 GB Apple-silicon computer can realistically reserve only about 18–22 GiB for resident weights, so the released artifact is more than seventy times larger than a 20 GiB resident-weight target.

This manuscript proposes *SPECTRA-MoE 2.1*, a signal-processing-driven framework for structural compression rather than scalar re-quantization alone. It combines activation-aligned transforms, random-matrix compressibility diagnostics, multirate filter-bank coding, shared expert representations, joint truncation–quantization allocation, vector quantization, sparse/low-rank repair, spectral distillation, and demand-paged refinements. Four hypotheses receive particular emphasis: (i) a causal expert graph whose edges are estimated by a tractable screen-and-audit counterfactual procedure; (ii) a router-margin-certified global rate–distortion objective; (iii) a successively refinable, router-addressable bitstream; and (iv) Fisher-shaped encoder-side error-feedback quantization.

The released K3 checkpoint allows the proposal to replace symbolic architecture assumptions with exact dimensions and to quantify the offline calibration and paging problem. We derive a streaming-sketch memory budget, reduce exhaustive expert-pair intervention to sparse candidate screening plus audited counterfactuals, formalize packet-granularity selection under measured SSD bandwidth and access latency, and provide a benchmark-specific pilot protocol on the fully open OLMoE-1B-7B model before any K3 parity claim. We additionally report implementation smoke tests from the public SPECTRA repository; these verify codec and progressive-reconstruction code paths on synthetic Kimi-style tensors but are explicitly not evidence of K3 quality retention. The central claim remains falsifiable: determine how much K3-class task behavior can be retained by a compact, multiscale, dynamically reconstructed model under a 32 GB unified-memory constraint. Repo here:https://github.com/andrew-jeremy/spectra-k3

Keywords: large language model compression; mixture of experts; spectral compression; graph signal processing; rate–distortion optimization; Apple silicon.

1 Introduction

The dominant approach to local large-language-model deployment is to quantize a dense checkpoint from 16-bit or 8-bit arithmetic to 4, 3, or 2 bits per parameter. This is highly effective for models whose uncompressed parameter count is already within one order of magnitude of the target memory budget. It is not, by itself, a solution for trillion-scale MoE systems. Kimi K2 and Kimi K2.5 each contain one trillion total parameters and activate 32 billion parameters per token [23, 25]; Kimi K3 scales the total count to 2.8 trillion parameters and activates 16 of 896 experts per token [26]. Although conditional computation reduces arithmetic per token, it does not automatically reduce checkpoint storage: all experts must either be resident, stored externally and paged, merged, pruned, reconstructed, or replaced by a smaller student.

The target hardware in this manuscript is a general-purpose Apple-silicon Mac with 32 GB of *unified memory*. The phrase "32 GB VRAM" is common but technically imprecise for this platform: the CPU and GPU share a common memory pool. PyTorch exposes the GPU through the Metal Performance Shaders (MPS) backend, whereas Apple's MLX framework is designed around unified memory and allows arrays to be used by CPU and GPU operations without explicit copies [2, 21]. Unified memory improves programmability and can reduce transfer overhead, but it does not remove the physical capacity constraint. The operating system, application, runtime, activations, temporary buffers, routing tables, and context state all compete with model weights.

The proposed research starts from a classical signal-processing observation: successful compression is rarely obtained by quantizing raw samples independently. Images, audio, video, radar signals, and communication waveforms are first transformed into a domain in which energy, predictability, or perceptual importance is concentrated. Coefficients are then allocated unequal rates, thresholded, vector quantized, entropy coded, and reconstructed with controlled distortion. LLM weight tensors and activations are also highly structured, but common quantization pipelines mostly treat them as blocks of scalar values after limited rotations or scaling. Recent methods such as QuIP# demonstrate that randomized Hadamard incoherence processing and lattice codebooks can substantially improve extreme low-bit quantization [38]; ASVD and SVD-LLM show that activation-aware coordinate changes improve low-rank compression [41, 46]; wavelet-based work has begun to explore multiscale representations of weights and activations [6, 39, 43]. These results motivate a more comprehensive transform-coding formulation.

The status of the target model changed after the first version of this manuscript. Moonshot AI released the full Kimi K3 weights and technical report in July 2026. The official model card reports 2.8T total parameters, 104B activated parameters, 93 layers, 69 KDA plus 24 Gated MLA attention layers, hidden size 7168, latent MoE dimension 3584, expert hidden dimension 3072, 896 routed experts, top-16 routing, two shared experts, a 401M-parameter MoonViT-V2 encoder, and a 1,048,576-token context [28, 29]. The public configuration also exposes the released quantization scheme and exact layer schedule [27]. Consequently, this revision no longer treats K3 as a hypothetical architecture. It treats the released checkpoint as the primary structural target while retaining smaller open MoEs as the tractable environment for end-to-end ablation and causal intervention.

1.1 Research hypothesis

The central hypothesis is that a frontier MoE checkpoint contains several distinct forms of redundancy that can be separated and coded at different rates:

1. **Within-matrix correlation:** rows, columns, heads, and intermediate channels are correlated under the model's activation distribution.

2. **Across-layer correlation:** neighboring layers often implement related transformations and may share bases, codebooks, or low-frequency components.

3. **Across-expert correlation:** experts in the same layer may share a large semantic subspace while differing through sparse, task-specific innovations.

4. **Routing sparsity:** only a small subset of experts is used for each token, permitting demand-driven reconstruction or paging.

5. **Task sparsity:** a local user needs a subset of the global production model's capabilities, languages, modalities, tools, and knowledge.

6. **Temporal and contextual redundancy:** KV states, recurrent attention states, and long contexts contain compressible low-rank and multiscale structure.

The corresponding engineering claim is conditional: a 32 GB local system may reproduce a declared slice of Kimi-class behavior if the system is allowed to (i) structurally reduce the model, (ii) distill the desired behavior, (iii) preserve a compact resident core, (iv) page sparse refinements from SSD, and (v) replace raw million-token context with hierarchical memory. Exact, lossless, all-domain equivalence is outside the feasible set.

1.2 Contributions of this manuscript

This revision makes the following contributions:

1. A released-checkpoint feasibility audit grounded in the public Kimi K3 configuration and model card rather than pre-release assumptions.

2. A quantitative separation between the approximately 1.56 TB released weight artifact and the 20 GiB resident-weight objective, showing that structural reduction and/or hierarchical residency are mandatory even after native MXFP4 quantization.

3. A notation-normalized signal-processing formulation in which every expert matrix is explicitly defined as $W_{\ell,e}^{(t)} \in \mathbb{R}^{m_{\ell,t} \times n_{\ell,t}}$, with layer, expert, tensor-type, subband, packet, and codec indices separated.

4. A *screen-and-audit causal expert graph*: observational sketches propose a sparse neighborhood, local expert replay scores candidates, and only a bounded uncertain/high-impact subset receives full downstream counterfactual evaluation.

5. A *router-margin-certified rate–distortion objective* that controls the probability of compression-induced changes to the teacher's top-k expert set.

6. A *successively refinable, router-addressable bitstream* whose packet size and grouping are selected by an explicit overfetch/latency/metadata objective rather than an unspecified paging heuristic.

7. A *Fisher-shaped error-feedback quantizer* that moves coefficient error toward lower-sensitivity transform directions before final scalar/vector coding.

8. An explicit calibration-compute and memory analysis showing that statistics collection is an *offline* streaming process and is not constrained to execute inside the 32 GB deployment envelope.

9. A benchmark-specific experimental gate on OLMoE-1B-7B-0924, followed by Kimi K2/K2.5 and shard-level K3 audits, with fixed calibration budgets, baselines, failure criteria, and reproducibility manifests.

10. Preliminary implementation smoke tests that validate bit packing, graph transforms, progressive reconstruction, routing certificates, packet integrity, and an expert-graph codec on synthetic tensors. These tests establish software-path correctness only and are not presented as model-quality evidence.

2 Target System and Problem Definition

2.1 Released Kimi K3 checkpoint

Kimi K3 is no longer a hypothetical target. Moonshot AI's released model card and configuration provide the concrete architecture summarized in Table 1 [27–29].

The public configuration identifies the text model as Kimi Linear, specifies the full KDA/MLA layer schedule, top-16 routing over 896 experts, and an MXFP4 packed quantization configuration with group size 32 for targeted linear layers while exempting several attention, shared-expert, output, and vision components [27]. This mixed representation explains why the physical checkpoint is larger than the idealized $2.8 \times 10^{12} \times 4/8 = 1.40$ TB calculation. The official Hugging Face repository reports an aggregate model size of approximately 1.56 TB and exposes 96 Safetensors weight shards [29]. In binary units this is approximately 1.42 TiB. We use the released repository size for systems planning and the official model card/configuration for architecture claims.

The release also changes the experimental hierarchy. K3 can now be inspected shard-by-shard and its exact tensor schema can be mapped, but a complete teacher forward pass still requires infrastructure far beyond a 32 GB Mac. Therefore, *offline compression* and *local deployment* are treated as separate resource regimes: distributed/high-memory hardware performs calibration, structural fitting, and distillation; the 32 GB Apple-silicon target runs only the exported resident core and pageable refinement stream.

2.2 Notation and dimensions

To remove ambiguity between layer, expert, matrix type, subband, and packet indices, Table 2 fixes the notation used throughout the paper.

Table 2: Core notation.

Symbol	Meaning
$\ell \in \{1, \dots, L\}$	transformer-layer index; $L = 93$ for released K3.
$e \in \{1, \dots, E_\ell\}$	routed-expert index in MoE layer ℓ; $E_\ell = 896$ for K3 MoE layers.
$t \in \mathcal{T}_W$	weight-matrix type (e.g., gate/up/down projection, attention projection).
$W_{\ell,e}^{(t)} \in \mathbb{R}^{m_{\ell,t} \times n_{\ell,t}}$	expert-specific matrix of type t; non-expert matrices omit e.

Symbol	Meaning
$P_{\ell,t} = m_{\ell,t} n_{\ell,t}$	scalar coefficients in one matrix of type t at layer ℓ.
j	multiresolution scale index.
$b \in \mathcal{B}_{\ell,t}$	subband identifier within a transformed matrix; $B_{\ell,t} = \lvert\mathcal{B}_{\ell,t}\rvert$.
p	serialized packet identifier in the progressive bitstream.
$q_i \in \mathcal{Q}_i$	codec choice for coefficient block i.
R_i	allocated coded rate in bits for block i.
M_w	resident weight/codebook budget on the deployment machine.
$z_\ell(x)$	router logits at layer ℓ for input/token state x.
$\mu_\ell(x)$	teacher top-k router margin, separating the kth and $(k+1)$th logits.
Ψ_ℓ^E	expert-graph analysis transform for MoE layer ℓ.

2.3 Hardware model

Let the physical unified memory be

$$M_{\text{phys}} = 32\,\text{GiB}.$$

A deployable runtime must satisfy

$$M_{\text{weights}} + M_{\text{scales}} + M_{\text{routing}} + M_{\text{activations}} + M_{\text{state}} + M_{\text{workspace}} + M_{\text{application}} \leq M_{\text{phys}}. \quad (1)$$

A conservative production budget is

$$M_{\text{weights}} \in [18, 22]\,\text{GiB},$$

with the remainder reserved for macOS, MLX or llama.cpp, token buffers, the vision encoder if used, KV or recurrent states, temporary dequantization tiles, and the host application. The manuscript uses 20 GiB as the default resident-weight budget.

Storage capacity is not the only constraint. Autoregressive decode on consumer hardware is frequently memory-bandwidth bound. A compressed representation that requires expensive full-matrix decompression each token may reduce disk footprint while reducing throughput. Accordingly, the optimization target must include:

- resident bytes;
- bytes read from unified memory per generated token;
- SSD bytes and page faults per token;

- transform/dequantization operations;
- p50 and p95 time-to-first-token and inter-token latency;
- energy per generated token;
- quality and safety retention.

2.4 Operational definition of comparable performance

Definition 1 (Task-conditioned comparability)**.** *Let $\mathcal{T} = \{T_j\}_{j=1}^{J}$ be a declared evaluation suite, $s_j(\mathcal{M})$ the score of model $\mathcal{M}$ on task T_j, and $w_j \geq 0$ normalized task weights. A compressed model $\widehat{\mathcal{M}}$ is* ρ-comparable *to teacher $\mathcal{M}$ on $\mathcal{T}$ if*

$$\text{Retention}(\widehat{\mathcal{M}}; \mathcal{M}, \mathcal{T}) = \sum_{j=1}^{J} w_j \frac{s_j(\widehat{\mathcal{M}}) - s_j(\mathcal{M}_{\text{chance}})}{s_j(\mathcal{M}) - s_j(\mathcal{M}_{\text{chance}})} \geq \rho, \tag{2}$$

subject to declared memory, latency, context, and safety constraints.

The recommended goals are:

1. **Targeted parity:** $\rho \geq 0.95$ on a user-specific suite such as coding, research synthesis, or scientific reasoning.
2. **Broad retention:** $\rho \geq 0.90$ across reasoning, knowledge, instruction following, code, long-context retrieval, and tool use.
3. **Safety stability:** no material increase in harmful compliance, privacy leakage, or calibration error under an independent safety suite.
4. **Local usability:** peak memory below 30 GiB, no memory-pressure termination, and interactive decode throughput.

These are research targets, not claimed results.

3 Feasibility Analysis and Memory Lower Bounds

3.1 Raw storage

For P parameters encoded with an average of b bits per parameter, ignoring scales, indices, codebooks, padding, and alignment,

$$M_{\text{raw}}(P, b) = \frac{Pb}{8} \text{ bytes.} \tag{3}$$

For $P = 2.8 \times 10^{12}$, the idealized K3 storage is shown in Table 3.

At a 20 GiB resident-weight budget, the maximum number of unstructured parameters is approximately 42.95B at four bits, 57.27B at three bits, 85.90B at two bits, or 108.73B at 1.58 bits. Thus, even ideal 1.58-bit scalar coding leaves K3 about 25.8 times too large. Four-bit coding leaves it about 65.2 times too large. The released checkpoint is larger still because it is not uniformly four-bit. Using the official approximately 1.56 TB repository size (about 1.42 TiB), the physical release is approximately 72.7 times the 20 GiB resident-weight target [29]. This ratio is the practical structural-plus-residency reduction that the complete system must bridge.

Proposition 1 (Structural reduction is necessary)**.** *Let P be the total checkpoint parameter count, $b_{\min}$ the minimum average code length per retained scalar coefficient, M_w the resident weight budget, and $r_s \in (0, 1]$ the fraction of scalar degrees of freedom retained after pruning, factorization, expert merging, or distillation. A necessary storage condition is*

$$r_s \leq \frac{8M_w}{Pb_{\min}}. \tag{4}$$

For K3, $P = 2.8 \times 10^{12}$, $M_w = 20$ GiB, and $b_{\min} = 2$, one requires $r_s \leq 0.0307$, i.e., at least a $32.6\times$ structural reduction before metadata.

Proof. The encoded retained coefficients require at least $r_s P b_{\min}/8$ bytes. Requiring this value not to exceed M_w yields Eq. (4). □

Remark 1. *Conditional activation does not invalidate Proposition 1. It reduces the number of experts evaluated per token, but a local checkpoint still needs a representation of inactive experts unless they are deleted, merged, generated from a shared code, or stored outside resident memory.*

3.2 Why ordinary quantization is not enough

GPTQ, AWQ, SmoothQuant, SpQR, QuIP#, and AQLM provide strong 4-bit to 2-bit compression regimes [11, 14, 18, 32, 38, 44]. These methods are essential components of the proposed stack, but they operate after the model's principal degrees of freedom have been chosen. Applying 2-bit AQLM directly to 2.8 trillion parameters still produces hundreds of gigabytes. K3 is not merely reported to use low precision: the released configuration explicitly declares MXFP4 packed quantization with group size 32 for targeted linear layers, while excluding attention, shared-expert, output-head, vision, and selected MLP components from that rule [27]. The easiest scalar precision redundancy has therefore already been partially exploited. A community MLX conversion at roughly 2.71 effective bits per logical parameter still occupies about 948 GB and requires two 512 GB Macs, illustrating the scale of the remaining structural problem; this is an engineering datapoint rather than a peer-reviewed result [4]. Further progress toward 32 GB must come from expert reduction/sharing, transform sparsity, task conditioning, distillation, and hierarchical residency.

3.3 Context-state feasibility

A one-million-token context cannot be treated as a free feature on a 32 GB machine. The state memory of a conventional attention model scales as

$$M_{\mathrm{KV}} \approx 2Ln_{\mathrm{kv}}d_hT\frac{b_{\mathrm{KV}}}{8}, \tag{5}$$

where L is the number of cached attention layers, n_{kv} the number of key/value heads, d_h the head dimension, T the context length, and the factor two represents keys and values. MLA and KDA reduce this cost through latent projections and recurrent state, and Kimi Linear reports a hybrid KDA/MLA design [24]. Nevertheless, the local system must impose a context-state budget. The proposal therefore distinguishes:

- **native active context**, e.g., 8K–32K tokens held in quantized model state;
- **compressed episodic memory**, represented by multiscale summaries, retrieval vectors, and selected verbatim spans;

- **external document memory**, indexed on SSD and retrieved on demand.

Comparable long-horizon task performance may be achievable without reproducing the teacher's raw one-million-token state representation.

4 Signal-Processing View of LLM Compression

4.1 Modern LLM compression as transform coding

Several successful compression methods can be interpreted through classical source coding. Activation-aware SVD is an input-weighted Karhunen–Loève transform; randomized Hadamard rotations are incoherence processing; lattice and trellis quantizers recover vector-quantization space-filling gain; and mixed-precision allocation is a transform-coding bit-allocation problem. This correspondence is useful only when it produces testable decisions. SPECTRA-MoE therefore measures transform coding gain, effective rank, route sensitivity, and expert substitution regret before selecting a codec rather than assuming every tensor is spectrally compressible.

4.2 Random-matrix and transform-gain diagnostics

For an $m \times n$ noise-like matrix with aspect ratio $q = m/n \leq 1$ and entry variance σ^2/n, the Marchenko–Pastur support for squared singular values is

$$\lambda_{\pm} = \sigma^2(1 \pm \sqrt{q})^2. \tag{6}$$

Singular values above the fitted upper bulk edge are candidate structured directions; the bulk is not automatically discarded, but it is a signal that low-rank coding may be expensive. The estimated outlier count becomes a data-driven initial rank.

For a candidate transform T, define the transform coding gain from coefficient-block variances $\{v_i\}_{i=1}^{N_b}$ as

$$G_{\mathrm{TC}}(T) = \frac{N_b^{-1} \sum_i v_i}{(\prod_i v_i)^{1/N_b}}. \tag{7}$$

A gain close to one argues against elaborate subband coding; a large gain supports unequal rate allocation and thresholding. These diagnostics are evaluated in activation-whitened coordinates and are reported per layer and per expert family.

4.3 Weights as high-dimensional signals

Consider a linear projection

$$\boldsymbol{y} = W\boldsymbol{x}, \qquad W \in \mathbb{R}^{m \times n}.$$

Raw matrix indices are not necessarily spatially meaningful. Applying a DCT or wavelet transform to an arbitrary row order may produce little energy compaction. The first problem is therefore to infer coordinates in which neighboring channels are statistically or functionally related. Let calibration activations be $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $Y = WX$. Define regularized covariances

$$C_x = \frac{1}{N} XX^\top + \epsilon I, \qquad C_y = \frac{1}{N} YY^\top + \epsilon I. \tag{8}$$

These matrices define activation geometry. They can be used to whiten inputs, build channel graphs, estimate Fisher sensitivity, or learn fast orthogonal transforms.

A two-sided transform representation is

$$W = U_y \widetilde{W} U_x^\top, \qquad \widetilde{W} = U_y^\top W U_x, \tag{9}$$

where U_x and U_y are orthogonal or approximately orthogonal. The goal is to choose transforms for which $\widetilde{W}$ is sparse, low-rank, block-smooth, codebook-friendly, or shared across related matrices.

4.4 Transform coding

Classical transform coding has four steps:

1. apply an invertible or approximately invertible transform;
2. partition coefficients into bands or blocks;
3. allocate rates according to energy and distortion sensitivity;
4. quantize and entropy-code coefficients.

The DCT is effective when signals are locally correlated [1]; wavelets provide localized multiresolution analysis [10, 33]; the Karhunen–Loève transform/PCA is optimal for mean-square energy compaction under known second-order statistics; graph Fourier transforms generalize frequency to irregular domains [37]. Compressed sensing establishes that sparse signals can be recovered from fewer measurements under suitable incoherence and restricted isometry assumptions [5, 12]. These ideas motivate model representations in which "frequency" means variation across activation-correlated channels, experts, layers, heads, or time.

4.5 Multirate filter banks

A critically sampled two-channel analysis filter bank maps a signal $w[n]$ into low-pass and high-pass subbands:

$$a_1[k] = \sum_n h[n - 2k] w[n], \tag{10}$$

$$d_1[k] = \sum_n g[n - 2k] w[n]. \tag{11}$$

Repeated decomposition of a_1 yields a multiscale tree. For a weight matrix, separable transforms can be applied across input and output dimensions, or a graph filter bank can be applied across channel graphs. Perfect reconstruction is possible before thresholding and quantization. Compression is obtained by coding low-frequency coefficients accurately and discarding or coarsely coding small detail coefficients.

For an MoE layer, multirate structure has an additional interpretation:

- a low-pass expert component captures behavior common to an expert cluster;
- intermediate bands encode domain specialization;
- high-pass innovations encode rare, precise distinctions;
- only bands required by the current routing context are reconstructed.

This creates a natural hierarchy between a resident base model and pageable refinements.

4.6 Rate–distortion formulation

Let z_i denote transform coefficients or coefficient blocks, R_i the number of bits allocated, and $D_i(R_i)$ the resulting task-weighted distortion. The classical constrained problem is

$$\min_{\{R_i\}} \sum_i D_i(R_i) \quad \text{subject to} \quad \sum_i R_i \le B, \qquad R_i \ge 0. \tag{12}$$

Unlike image compression, LLM distortion should not be measured only by coefficient mean-square error. A suitable local approximation weights perturbations by activation or Hessian information:

$$D_\ell(W_\ell, \widehat{W}_\ell) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{cal}}}\left[\|(W_\ell - \widehat{W}_\ell)\boldsymbol{x}\|_2^2\right] = \operatorname{tr}\left(\Delta W_\ell C_{x,\ell} \Delta W_\ell^\top\right), \tag{13}$$

where $\Delta W_\ell = W_\ell - \widehat{W}_\ell$. A richer criterion combines this local distortion with teacher logits, hidden-state spectra, routing distributions, and downstream task loss.

4.7 Reverse water-filling jointly selects rank and precision

For a Gaussian source with activation-weighted eigenvalues $\{\lambda_i\}$, reverse water-filling gives

$$D_i = \min(\theta, \lambda_i), \qquad R = \sum_i \frac{1}{2}\log_2 \frac{\lambda_i}{D_i}, \tag{14}$$

where θ is selected by the total rate. Directions with $\lambda_i \le \theta$ receive zero bits and are truncated; retained directions are quantized. Thus rank selection and precision selection are two branches of one allocation problem. For layer ℓ, an implementable relaxation is

$$\min_{\{r_\ell, b_\ell\}} \sum_\ell \alpha_\ell \left[\sum_{i>r_\ell} (\sigma_i^{(\ell)})^2 + c_q 2^{-2b_\ell} \sum_{i\le r_\ell} (\sigma_i^{(\ell)})^2\right] \quad \text{s.t.} \quad \sum_\ell b_\ell (m_\ell + n_\ell) r_\ell \le B_{\text{tot}}. \tag{15}$$

Contemporary work also jointly searches structural and precision choices; the claim here is narrower: the relaxation initializes a MoE-wide discrete allocation over channel bands, expert-graph bands, repair rank, resident bytes, and page traffic.

4.8 Vector quantization and residual incoherence

At the 2–3 bit operating points needed for the proposed system, scalar quantization leaves material space-filling loss. Lattice, trellis, product, and additive vector quantizers are therefore treated as first-class codecs. The residual is whitened or incoherence-processed *after* low-rank and shared expert components have been removed, because the remaining innovation is expected to be more isotropic and better matched to a codebook. This ordering is evaluated against the reverse ordering in ablation.

5 Related Work

5.1 Quantization and pruning

Post-training quantization methods exploit activation statistics, approximate second-order information, rotations, sparse outliers, and vector codebooks. GPTQ performs efficient layerwise second-order weight quantization [18]; AWQ protects activation-salient channels [32]; SmoothQuant

migrates activation difficulty into weights through equivalent scaling [44]; SpQR stores sensitive outliers separately [11]; QuIP# uses randomized Hadamard incoherence processing and lattice codebooks [38]; AQLM uses additive multi-codebook quantization in the 2–3 bit regime [14]. SparseGPT demonstrates one-shot pruning at large model scale [16]. Deep Compression established the broader prune–quantize–code pipeline in earlier neural networks [20].

These methods validate several signal-processing principles—decorrelation, outlier separation, vector quantization, and entropy-aware representation—but generally do not jointly model multiscale expert structure and dynamic subband residency.

5.2 Low-rank, tensor, and structured transforms

The Eckart–Young theorem identifies truncated SVD as the optimal rank-r approximation in Frobenius norm [13]. Randomized algorithms make approximate decompositions practical for large matrices [19]. ASVD and SVD-LLM adapt low-rank compression to activation distributions and truncation loss [41, 46]. Tensor-train, Tucker, Kronecker, and block-circulant decompositions offer additional parameter sharing; tensor train is especially useful when matrices can be tensorized into balanced modes [35]. Butterfly factorizations represent broad classes of fast transforms as products of sparse factors [9].

Low-rank methods can fail when LLM matrices have slowly decaying singular spectra. SPECTRA-MoE therefore treats low-rank approximation as one band or residual mechanism rather than a universal representation.

5.3 Wavelets and spectral methods

Learnable wavelet transforms have been used to compress neural linear layers [43]. WaveletGPT imposes multiscale structure on intermediate embeddings [39], while recent work explores Haar-enhanced extreme quantization [6]. Graph signal processing provides tools for defining frequencies over covariance or interaction graphs [37]. These lines support the hypothesis that LLM tensors can be compressed more effectively after architecture- and activation-aware spectral organization.

5.4 Mixture-of-experts reduction

MoE architectures trade total parameter storage for sparse compute. Existing compression methods prune experts, merge experts, factorize each expert, share bases, quantize expert weights, or offload cold experts. MoBE, for example, represents expert matrices through shared basis matrices and expert-specific factors [7]; QMoE demonstrates extreme low-bit coding of trillion-parameter MoEs [17]; and recent systems optimize mixed-precision offloading and asynchronous expert materialization [40, 45]. These works substantially narrow any claim that a shared expert dictionary or speculative prefetch is new by itself.

The proposed departure is the combination of (i) a causal substitution test for deciding which experts may share information, (ii) a graph wavelet-packet transform applied along the expert axis rather than only per-expert matrix factorization, and (iii) random-access enhancement packets jointly optimized with route stability and I/O. A recent causal audit reports that common observational expert-importance statistics need not predict interventional importance [15]; this directly motivates the causal expert graph in Section 6.7.

5.5 Closest contemporary work and narrowed novelty posture

Several elements in the reconciled design have close contemporary precedents. Joint pruning/architecture and mixed-precision search is an active line of work [30]; cross-layer predictive weight coding has been proposed using alignment, keyframes, and residuals [31]; routing-aware quantization now explicitly preserves top-k ordering and boundary margins [36]; scalable neural-network parameter streams predate this manuscript [42]; and speculative expert prefetch is central to MoE-SpeQ [40]. These techniques are included as baselines or enabling components, not claimed wholesale as original.

The primary novelty hypotheses retained in this revision are: causal expert-graph wavelet packets; a global chance-constrained routing certificate coupled to subband allocation; Fisher-shaped error feedback in activation/graph spectral coordinates; and a router-addressable successive-refinement format that permits independent random access to expert enhancement bands. Each has a distinct falsification test in Section 15.

5.6 Knowledge distillation

Knowledge distillation transfers soft targets and intermediate behavior from a teacher to a student [22]. For the present goal, distillation is not an optional polish stage; it is one of the mechanisms that converts task-conditioned teacher behavior into fewer degrees of freedom. The proposed method adds spectral hidden-state matching and router-distribution matching to conventional logit distillation.

6 The Proposed SPECTRA-MoE Framework

6.1 System overview

SPECTRA-MoE is a nine-stage pipeline:

1. collect multimodal, code, reasoning, long-context, tool-use, and safety calibration traces;
2. estimate per-layer activation covariance, Fisher/Hessian surrogates, and expert co-activation statistics;
3. learn channel permutations and fast activation-aligned transforms;
4. transform weights and group related experts into spectral atlases;
5. decompose coefficients into multirate subbands and low-rank/sparse residuals;
6. solve a global rate–distortion allocation problem;
7. quantize/code each band with an appropriate scalar, vector, additive, or sparse representation;
8. repair the compressed model through constrained distillation;
9. export a resident-core plus pageable-refinement runtime for MLX/Metal.

For matrix type t, a conceptual compressed linear operator is

$$\widehat{W}_{\ell}^{(t)}(\boldsymbol{x}) = U_{\ell,t}^{\mathrm{out}} \left[\widehat{A}_{\ell,t,0} + \sum_{b \in \mathcal{B}_{\ell,t}} g_{\ell,t,b}(\boldsymbol{x}) \widehat{D}_{\ell,t,b} + \widehat{R}_{\ell,t} \right] \left(U_{\ell,t}^{\mathrm{in}}\right)^{\top}, \tag{16}$$

where $\widehat{A}_{\ell,t,0}$ is the resident low-pass core, $\widehat{D}_{\ell,t,b}$ is subband b, $g_{\ell,t,b}(\boldsymbol{x}) \in \{0,1\}$ or $[0,1]$ gates that detail band, $\widehat{R}_{\ell,t}$ is a sparse or low-rank correction, and $U_{\ell,t}^{\text{in/out}}$ are fast transforms. The subband index b is distinct from the coded rate R_i and from the retained rank r_ℓ. In a fused implementation, the transforms and quantized multiply are not materialized as dense matrices.

6.2 Calibration trace design

Compression quality depends on calibration coverage. Let

$$\mathcal{D}_{\text{cal}} = \mathcal{D}_{\text{text}} \cup \mathcal{D}_{\text{code}} \cup \mathcal{D}_{\text{math}} \cup \mathcal{D}_{\text{vision}} \cup \mathcal{D}_{\text{tools}} \cup \mathcal{D}_{\text{long}} \cup \mathcal{D}_{\text{safety}}.$$

For each layer and expert, record:

- input/output moments and robust quantiles;
- top singular/eigen directions through streaming sketches;
- router probabilities, selected experts, entropy, and transition statistics;
- per-channel activation maxima and outlier rates;
- approximate diagonal Fisher information;
- teacher logits, hidden states at selected anchor layers, and recurrent/KV states;
- task tags and modality identifiers.

Full covariance matrices are too expensive at frontier dimensions. Use randomized sketches, block-diagonal approximations, Nyström approximations, or sparse k-nearest-neighbor channel graphs.

6.3 Calibration overhead and streaming-statistics budget

The 32 GB limit applies to *deployment*, not to the offline teacher-calibration job. Nevertheless, the calibration state itself must remain bounded so that the procedure scales to K2/K3. Let d_ℓ be the hidden width and let r_s be the rank of a streaming covariance sketch. A dense per-layer covariance requires

$$M_{\text{cov,dense}} = 4 \sum_{\ell=1}^{L} d_\ell^2 \quad \text{bytes (FP32)}, \tag{17}$$

whereas a rank-r_s sketch requires

$$M_{\text{cov,sketch}} = 4 \sum_{\ell=1}^{L} d_\ell r_s. \tag{18}$$

For K3, taking $L = 93$, $d_\ell = 7168$, and $r_s = 128$ gives approximately 17.8 GiB for dense covariances but only 325.5 MiB for the sketches. A diagonal Fisher vector at the same hidden width costs only about 2.5 MiB across all 93 layers. A sparse expert graph with 16 candidate neighbors per expert, storing an expert id and one FP32 score per edge, is on the order of 10 MiB, whereas raw per-token routing traces are never retained in full. Statistics are updated online and reduced across workers.

We therefore divide calibration into three passes with explicit budgets:

1. **Pass C0—moments/router telemetry:** one teacher forward stream; update moments, quantiles, rank-r_s sketches, top-k counts, sparse co-selection tables, and 64–128-dimensional expert-output sketches. No raw hidden-state archive is required.

2. **Pass C1—sensitivity:** compute diagonal Fisher or low-rank gradient sketches on a stratified 1–5% calibration subset. This is the only pass requiring backward information.

3. **Pass C2—causal audit:** replay cached layer inputs for a sparse set of candidate expert substitutions; perform full downstream counterfactual continuation only for uncertain or high-impact edges selected by the screen-and-audit procedure in Section 6.7.

To make time overhead reproducible, we report calibration cost in *teacher-forward equivalents*. If $F(N)$ is the cost of one teacher forward pass over N calibration tokens, Pass C0 costs $1.0F(N)$. Pass C1 is capped at fraction $f_F \leq 0.05$ and a forward-plus-backward step is budgeted at three forward equivalents, hence at most $0.15F(N)$. Local expert replay in C2 is capped at $0.10F(N)$ and true downstream audits at $0.25F(N)$, yielding a predeclared total budget

$$C_{\text{cal}} \leq 1.50F(N). \tag{19}$$

This converts directly to wall-clock time on the chosen teacher infrastructure: if a 5M-token K2/K3 forward takes T_F hours on a declared cluster, the calibration job is budgeted to complete within $1.5T_F$ hours excluding checkpoint staging. If confidence intervals remain too wide at that budget, the affected causal edges are marked unknown and the graph codec is disabled rather than increasing compute without bound.

For the 1T-parameter Kimi K2/K2.5 surrogate ($L = 61$, hidden width 7168), the same rank-128 scheme uses approximately 213.5 MiB for hidden-state covariance sketches versus about 11.7 GiB for dense FP32 covariances, approximately 1.7 MiB for diagonal hidden-state Fisher vectors, and only a few MiB for a sparse 16-neighbor expert graph. Thus calibration is a bounded offline streaming job; it is not performed inside the 32 GB deployment process.

The calibration corpus is processed from distributed object storage or NVMe in chunks. The final statistics bundle is designed to remain below 1 GiB per worker/rank by construction. The deployment Mac receives only the compressed artifact and packet index; it is never required to host the 1.56 TB teacher or run the full calibration pipeline.

6.4 Activation-aligned graph spectral transforms

For layer ℓ, define an input-channel graph

$$A_{\ell,ij} = \mathbf{1}\{j \in \mathcal{N}_k(i)\} \exp\left(-\frac{d_{\ell,ij}^2}{\tau_\ell^2}\right), \tag{20}$$

where $d_{\ell,ij}$ measures dissimilarity between channel activation traces, normalized covariance profiles, or Fisher-weighted responses. Let $L_\ell = D_\ell - A_\ell$ be the graph Laplacian. Its eigenvectors define a graph Fourier basis:

$$L_\ell = U_\ell \Lambda_\ell U_\ell^\top.$$

Smooth graph signals have energy concentrated at small eigenvalues. Direct storage of dense U_ℓ would erase compression gains, so SPECTRA-MoE approximates the transform by one of:

1. a shared DCT/Hadamard transform after spectral seriation/permutation;

2. a product of sparse Givens rotations;

3. a butterfly factorization with $O(d \log d)$ parameters and operations;

4. a low-rank correction to a fixed fast transform;

5. a polynomial graph filter bank that avoids explicit eigenvectors.

The transform-learning objective is

$$\min_{\theta,P} \sum_{\ell \in \mathcal{G}} \left[\|\mathcal{H}_\lambda(T_\theta^\top P^\top W_\ell P T_\theta)\|_0 + \alpha \, \mathcal{D}_{\text{act}}(W_\ell, \widehat{W}_\ell) \right] + \beta \, \Omega(T_\theta), \tag{21}$$

where P is a channel permutation, T_θ a fast approximately orthogonal transform, $\mathcal{H}_\lambda$ a thresholding operator, and Ω penalizes deviation from orthogonality or fast structure.

6.5 Multirate subband decomposition

After channel organization, each transformed matrix is partitioned by scale:

$$\widetilde{W}_\ell = A_{\ell,J} + \sum_{j=1}^{J} \sum_{o \in \mathcal{O}} D_{\ell,j,o}, \tag{22}$$

where $A_{\ell,J}$ is a coarse approximation and $D_{\ell,j,o}$ are detail bands at scale j and orientation o. For non-square matrices, use separable transforms with independent depths on input and output axes. For head-structured tensors, apply a three-axis transform across head, input-channel, and output-channel dimensions.

Each band is classified as:

- **resident critical:** high Fisher sensitivity or frequent use;
- **resident coarse:** low-frequency backbone at low bit width;
- **pageable refinement:** useful only for selected tasks or experts;
- **discarded:** negligible task-weighted distortion contribution;
- **repair-coded:** approximated through a low-rank or sparse residual.

The novelty is not the use of wavelets alone, but the interpretation of subbands as *runtime capability layers*. A local model can operate with its coarse bands and selectively increase precision or specialization by loading detail bands.

6.6 Shared expert spectral atlas

For a fixed expert matrix type t, let $W_{\ell,e}^{(t)} \in \mathbb{R}^{m_{\ell,t} \times n_{\ell,t}}$ denote expert e in MoE layer ℓ. After alignment, define a cluster $\mathcal{C}_{\ell,c}^{(t)}$ of experts with similar routing contexts and outputs. Represent

$$\widetilde{W}_{\ell,e}^{(t)} = B_{\ell,c(e)}^{(t)} + \sum_{r=1}^{R_{\ell,c,t}} a_{\ell,e,r}^{(t)} S_{\ell,c,r}^{(t)} + E_{\ell,e}^{(t)}, \tag{23}$$

where $B_{\ell,c}^{(t)}$ is a shared low-pass prototype, $S_{\ell,c,r}^{(t)}$ are shared spectral atoms, $a_{\ell,e,r}^{(t)}$ are expert-specific mixing coefficients, and $E_{\ell,e}^{(t)}$ is a sparse quantized innovation. The cluster objective is

$$\min_{\{B^{(t)},S^{(t)},a^{(t)},E^{(t)}\}} \sum_{e\in\mathcal{C}_{\ell,c}^{(t)}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\ell,e}} \left\| \left(\widetilde{W}_{\ell,e}^{(t)} - B_{\ell,c}^{(t)} - \sum_r a_{\ell,e,r}^{(t)} S_{\ell,c,r}^{(t)} - E_{\ell,e}^{(t)} \right) \widetilde{\boldsymbol{x}} \right\|_2^2 + \lambda_E \sum_{e\in\mathcal{C}_{\ell,c}^{(t)}} \|E_{\ell,e}^{(t)}\|_1 + \lambda_a \sum_{e\in\mathcal{C}_{\ell,c}^{(t)}} \|\boldsymbol{a}_{\ell,e}^{(t)}\|_0 + \lambda_S \sum_r \|S_{\ell,c,r}^{(t)}\|_*. \tag{24}$$

This representation interpolates among expert pruning (identical coefficients and zero innovation), expert merging (prototype only), and expert preservation (high-fidelity innovation) without changing matrix dimensions or silently mixing tensor types.

Expert clustering should not use raw weight distance alone. Recommended features include router co-selection, Jensen–Shannon divergence between routing-context distributions, centered kernel alignment between expert outputs, and activation-weighted reconstruction error.

6.7 Causal expert-graph wavelet packet coding

Weight distance, router frequency, and output correlation are observational proxies. Directly evaluating every ordered substitution among E_ℓ experts would require $E_\ell(E_\ell - 1)$ interventions per layer and is infeasible for $E_\ell = 896$. We therefore use a three-stage *screen-and-audit* estimator.

Stage 1: observational screening. For each expert, streaming router-context and output sketches define a proposal embedding $h_{\ell,e}$. Approximate nearest-neighbor search proposes at most $k_{\text{cf}} \ll E_\ell$ candidate substitutes. Observational similarity determines only which edges are tested; it is never interpreted as evidence of causal interchangeability.

Stage 2: local interventional replay. For a candidate edge $e \to f$, cache only layer-ℓ inputs on tokens for which e is selected and replay the two expert functions locally. Let $\Delta h_{\ell,e\to f}(x)$ be the change in the MoE output. A low-rank downstream influence sketch (g_ℓ, H_ℓ) produces the screening score

$$\widetilde{\Delta}_{\ell,e\to f}(x) = g_\ell(x)^\top \Delta h_{\ell,e\to f}(x) + \frac{1}{2} \Delta h_{\ell,e\to f}(x)^\top H_\ell \Delta h_{\ell,e\to f}(x). \tag{25}$$

This stage requires expert-local matrix multiplications rather than a complete continuation through the 93-layer teacher.

Stage 3: downstream audit and calibration. For each layer, a fixed budget B_{audit} of uncertain, high-regret, rare-route, and low-margin candidate pairs receives true downstream counterfactual evaluation:

$$\Delta_{\ell,e\to f}(x) = \mathcal{L}\Big(\mathcal{M}_T^{(\ell,e\to f)}; x\Big) - \mathcal{L}(\mathcal{M}_T; x). \tag{26}$$

The audited pairs calibrate the proxy in Eq. (25) and yield confidence intervals. Edges whose confidence intervals overlap a rejection threshold are retained as "unknown" rather than merged.

A symmetric causal distance is

$$d_{\ell,ef}^{\text{cf}} = \frac{1}{2} \mathbb{E}_{x\sim\mathcal{D}_{\ell,ef}} \left[(\Delta_{\ell,e\to f}(x))_+ + (\Delta_{\ell,f\to e}(x))_+ \right], \qquad A_{\ell,ef}^E = \exp\Big(-d_{\ell,ef}^{\text{cf}} / \tau_E\Big), \tag{27}$$

where audited values are used when available and calibrated conservative proxy bounds otherwise. The sampling distribution oversamples rare routes and small router margins.

For K3, exhaustive undirected all-pairs testing over 92 MoE layers would create about 36.9 million expert-pair families. With $k_{\mathrm{cf}} = 8$ for local screening and $B_{\mathrm{audit}} = 32$ true downstream pair families per MoE layer, only 2,944 pair families receive full downstream audits—approximately a 12,500-fold reduction in the expensive component. The cheaper local screening still covers only $92 \times 896 \times 8$ ordered candidates and can be batched by expert and layer.

Stack the corresponding aligned expert matrices of type t as

$$\mathscr{W}_{\ell}^{(t)} = \begin{bmatrix} \mathrm{vec}(\widetilde{W}_{\ell,1}^{(t)})^{\top} \\ \vdots \\ \mathrm{vec}(\widetilde{W}_{\ell,E_\ell}^{(t)})^{\top} \end{bmatrix} \in \mathbb{R}^{E_\ell \times P_{\ell,t}}. \tag{28}$$

The sparse causal affinity graph defines a graph-wavelet or lifting transform Ψ_{ℓ}^{E} along the expert axis:

$$\mathscr{C}_{\ell}^{(t)} = (\Psi_{\ell}^{E})^{\top} \mathscr{W}_{\ell}^{(t)}, \qquad \widehat{\mathscr{W}}_{\ell}^{(t)} = \widetilde{\Psi}_{\ell}^{E} \widehat{\mathscr{C}}_{\ell}^{(t)}. \tag{29}$$

Low graph frequencies encode causally substitutable behavior; high graph frequencies encode specialization boundaries. Dense eigenvectors are not stored: polynomial filters, graph lifting, or sparse Givens/butterfly factors implement the transform. The graph codec is disabled for any layer whose held-out causal smoothness, confidence interval width, or net byte saving fails a predeclared threshold.

6.8 Sensitivity-aware global bit allocation

For coefficient block i, let q_i identify a codec and bit-width, $m_i(q_i)$ its memory, $t_i(q_i)$ its runtime cost, and $d_i(q_i)$ its estimated task distortion. Solve

$$\min_{\{q_i\}} \sum_i d_i(q_i) + \lambda_t \sum_i t_i(q_i) + \lambda_p \sum_i p_i(q_i) \quad \text{s.t.} \quad \sum_i m_i(q_i) \le M_w, \tag{30}$$

where p_i penalizes page traffic or kernel incompatibility. Candidate codecs include:

- 8/6/5/4-bit scalar quantization for sensitive bands;
- 3/2-bit GPTQ-, QuIP#-, or AQLM-style coding;
- ternary coding for retrained low-sensitivity bands;
- product or residual vector quantization;
- sparse coordinate coding with high-precision outliers;
- low-rank factors with quantized singular vectors;
- zero bits for discarded bands.

A Lagrangian relaxation selects

$$q_i^{*}(\lambda) = \arg\min_{q \in \mathcal{Q}_i} \left[d_i(q) + \lambda m_i(q) + \lambda_t t_i(q) + \lambda_p p_i(q) \right], \tag{31}$$

and λ is adjusted by bisection until the memory target is met. This is the model-compression analogue of water-filling.

6.9 Router-margin-certified rate–distortion allocation

Let $z_\ell(x) \in \mathbb{R}^E$ be the teacher router logits and let $z_{\ell,(k)} \geq z_{\ell,(k+1)}$ denote the logits at the top-k boundary. Define the route margin

$$\mu_\ell(x) = z_{\ell,(k)}(x) - z_{\ell,(k+1)}(x). \tag{32}$$

For a candidate codec assignment q, let $\delta z_\ell(x; q)$ be the router-logit perturbation induced by all preceding compressed operators. In addition to expected task distortion, SPECTRA-MoE solves

$$\min_q D_{\text{task}}(q) + \lambda_M M(q) + \lambda_T T(q) + \lambda_I I(q) \quad \text{s.t.} \quad \Pr_{x \sim \mathcal{D}_{\text{cal}}} \left[\|\delta z_\ell(x; q)\|_\infty \geq \frac{\mu_\ell(x)}{2} \right] \leq \varepsilon_\ell, \quad \forall \ell \in \mathcal{E}. \tag{33}$$

The probability is estimated with held-out calibration traces and conservative upper confidence bounds. Tokens with small margins receive more weight, and router inputs, norms, and router weights may be protected at higher precision. Unlike a local router-alignment loss, the certificate is assigned to the complete upstream compression path and competes directly with memory and I/O in the global allocation.

6.10 Residual repair

Pure truncation can accumulate error across depth. For selected layers, fit a correction

$$R_\ell = A_\ell B_\ell^\top + S_\ell, \tag{34}$$

where $A_\ell B_\ell^\top$ is low-rank and S_ℓ is sparse. Fit the correction to teacher activations:

$$\min_{A_\ell, B_\ell, S_\ell} \mathbb{E}_{\boldsymbol{x}} \left\| W_\ell \boldsymbol{x} - (\widehat{W}_\ell + A_\ell B_\ell^\top + S_\ell)\boldsymbol{x} \right\|_2^2 + \lambda_r(\|A_\ell\|_F^2 + \|B_\ell\|_F^2) + \lambda_s \|S_\ell\|_1. \tag{35}$$

The correction budget is allocated to layers with the largest downstream amplification factors or validation loss gradients.

6.11 Attention and context-state compression

K3 combines KDA and MLA. SPECTRA-MoE applies different strategies to each:

1. **KDA state:** low-rank or diagonal-plus-low-rank factorization of recurrent state, block floating-point quantization, and periodic error-reset anchors.
2. **MLA latent cache:** per-head whitening/rotation followed by 4/3/2-bit quantization and outlier side channels.
3. **AttnRes memory:** compress cross-layer representations using shared bases and age-dependent precision.
4. **Context hierarchy:** maintain exact recent tokens, compressed mid-range states, semantic summaries, and external retrievable spans.

A generic state objective is

$$\min_{\widehat{\mathcal{S}}_{1:T}} \sum_{t=1}^{T} \left[\|z_t - \widehat{z}_t\|_{H_t}^2 + \gamma \, \mathrm{KL}(p_t \| \widehat{p}_t) \right] \quad \text{s.t.} \quad M(\widehat{\mathcal{S}}_{1:T}) \leq M_{\text{state}}, \tag{36}$$

where z_t is the teacher state and p_t the next-token distribution.

6.12 Fisher-shaped error-feedback quantization

Classical noise shaping does not reduce raw quantizer error; it redistributes error into frequencies that matter less to the final receiver. Let $c = [c_1, \dots, c_N]^\top$ be an ordered vector of transform or expert-graph coefficient blocks and let $H_c \succeq 0$ approximate task sensitivity in this coordinate system. During encoding, use a strictly lower-triangular feedback operator F:

$$u_i = c_i - \sum_{j<i} F_{ij} e_j, \tag{37}$$

$$\widehat{c}_i = Q_i(u_i), \tag{38}$$

$$e_i = \widehat{c}_i - u_i. \tag{39}$$

The stored coefficient error is

$$\widehat{c} - c = (I - F)e. \tag{40}$$

Choose a sparse, stable F to minimize

$$\min_F \operatorname{tr}\Big[H_c(I-F)\Sigma_e(I-F)^\top\Big] + \lambda_F \|F\|_1, \tag{41}$$

subject to a bounded feedback depth compatible with blockwise encoding. The decoder stores only $\widehat{c}$; no feedback state is required at inference. The idea is related to Sigma–Delta/frame quantization of neural networks [8], but the proposed experiment shapes graph/wavelet coefficient error with an activation/Fisher metric and evaluates interaction with low-rank removal and vector quantization.

6.13 Spectral distillation

Let teacher $\mathcal{M}_T$ and compressed student $\mathcal{M}_S$ produce logits $\boldsymbol{z}^T, \boldsymbol{z}^S$, selected hidden states H_ℓ^T, H_ℓ^S, and router distributions π_ℓ^T, π_ℓ^S. The repair loss is

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{KD}} T^2 \,\mathrm{KL}\big(\mathrm{softmax}(\boldsymbol{z}^T/T) \,\|\, \mathrm{softmax}(\boldsymbol{z}^S/T)\big) \\ & + \lambda_{\mathrm{hid}} \sum_{\ell\in\mathcal{A}} \big\|P_{\ell,k}H_\ell^T - P_{\ell,k}H_\ell^S\big\|_F^2 + \lambda_{\mathrm{spec}} \sum_{\ell\in\mathcal{A}} \big\|\Phi_\ell(H_\ell^T) - \Phi_\ell(H_\ell^S)\big\|_1 \\ & + \lambda_{\mathrm{route}} \sum_{\ell\in\mathcal{E}} \mathrm{KL}(\pi_\ell^T \| \pi_\ell^S) + \lambda_{\mathrm{safety}}\mathcal{L}_{\mathrm{safety}} + \lambda_{\mathrm{budget}}\mathcal{L}_{\mathrm{budget}}.\end{aligned} \tag{42}$$

Here $P_{\ell,k}$ aligns different hidden widths, and Φ_ℓ extracts graph/wavelet spectral energies. Spectral matching is less restrictive than elementwise matching and may preserve multiscale information even when the student uses fewer channels.

6.14 Task-conditioned capability profiles

A single 20 GiB checkpoint need not contain every capability at maximum fidelity. Define a resident base $\mathcal{M}_0$ and optional capability packs

$$\{\Delta\mathcal{M}_{\mathrm{code}}, \Delta\mathcal{M}_{\mathrm{math}}, \Delta\mathcal{M}_{\mathrm{vision}}, \Delta\mathcal{M}_{\mathrm{research}}, \Delta\mathcal{M}_{\mathrm{languages}}, \dots\}.$$

Each pack contains expert innovations, residual adapters, codebooks, retrieval indexes, and prompt/tool policies. At startup or query time, a capability classifier selects a bounded set of packs. This modularity converts global production capability into local task-conditioned comparability.

6.15 Router-addressable successive-refinement bitstream

The compressed artifact is encoded once as a base stream B_0 and enhancement packets $B_{j,a}$, where address a identifies a layer, expert neighborhood, modality, capability, or context-state band. A reconstruction at profile $\mathcal{A}$ is

$$\mathcal{M}^{(\mathcal{A})} = \mathrm{Decode}(B_0, \{B_{j,a} : a \in \mathcal{A},\ j \le J_a\}). \tag{43}$$

Enhancements use nested quantization grids or residual codebooks so that a coarse index is refined rather than replaced. Packets are independently memory-mappable and the manifest supports random access without decoding preceding experts. The runtime may truncate precision, omit a capability, or add graph-high-pass packets while preserving a valid base model.

This is an application of scalable source coding, not a claim that successive refinement itself is new. The proposed research question is whether MoE capability can be made approximately successively refinable under task distortion while retaining random expert access and bounded page amplification.

7 Mathematical Analysis

7.1 Transform-domain truncation error

Proposition 2 (Parseval-preserving coefficient error)**.** *Let $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ be orthogonal, $\widetilde{W} = U^\top W V$, and $\widehat{\widetilde{W}}$ any compressed coefficient matrix. Then*

$$\|W - U\widehat{\widetilde{W}}V^\top\|_F = \|\widetilde{W} - \widehat{\widetilde{W}}\|_F. \tag{44}$$

If $\widehat{\widetilde{W}}$ retains coefficient index set Ω, the squared error equals the energy of discarded coefficients:

$$\|W - \widehat{W}\|_F^2 = \sum_{(i,j)\notin\Omega} |\widetilde{W}_{ij}|^2. \tag{45}$$

Proof. The Frobenius norm is invariant under left and right multiplication by orthogonal matrices. The second statement follows by zeroing coefficients outside Ω. □

This proposition does not imply that small Frobenius error guarantees small task loss. It justifies transform energy as a first-stage proxy. Activation weighting in Eq. (13) is more relevant.

7.2 Activation-weighted whitening

Let $C_x = RR^\top$ be a Cholesky or symmetric square-root factorization. Then

$$\mathrm{tr}(\Delta W C_x \Delta W^\top) = \|\Delta W R\|_F^2. \tag{46}$$

Thus, compressing WR in Frobenius norm and applying R^{-1} at reconstruction aligns matrix error with the calibration activation distribution. This connects SPECTRA-MoE to activation-aware SVD and second-order quantization.

7.3 Network output perturbation

Consider a depth-L network

$$h_\ell = f_\ell(W_\ell h_{\ell-1}), \qquad \widehat{h}_\ell = f_\ell(\widehat{W}_\ell \widehat{h}_{\ell-1}),$$

where each f_ℓ is κ_ℓ-Lipschitz and $\|\widehat{W}_\ell\|_2 \le \omega_\ell$.

Theorem 1 (Layerwise error amplification)**.** *Assuming the same input $h_0 = \widehat{h}_0$,*

$$\|h_L - \widehat{h}_L\|_2 \le \sum_{\ell=1}^{L} \left[\kappa_\ell \|\Delta W_\ell\|_2 \|h_{\ell-1}\|_2 \prod_{j=\ell+1}^{L} \kappa_j \omega_j \right], \tag{47}$$

where $\Delta W_\ell = W_\ell - \widehat{W}_\ell$.

Proof. For each layer,

$$\begin{aligned} \|h_\ell - \widehat{h}_\ell\|_2 &\le \kappa_\ell \|W_\ell h_{\ell-1} - \widehat{W}_\ell \widehat{h}_{\ell-1}\|_2 \\ &\le \kappa_\ell \left(\|\Delta W_\ell h_{\ell-1}\|_2 + \|\widehat{W}_\ell (h_{\ell-1} - \widehat{h}_{\ell-1})\|_2 \right). \end{aligned}$$

Apply spectral-norm bounds and recursively expand. □

Theorem 1 motivates allocating more bits or repair rank to early or highly amplified layers, even when their local coefficient energy is modest. Residual architectures reduce practical amplification, but compression can disturb residual balance; AttnRes makes cross-layer sensitivity an explicit concern.

7.4 Top-k route stability certificate

Theorem 2 (Sufficient top-k stability condition)**.** *Let $z \in \mathbb{R}^E$ have a strict top-k margin $m = z_{(k)} - z_{(k+1)} > 0$. If a perturbation δz satisfies $\|\delta z\|_\infty < m/2$, then*

$$\mathrm{TopK}(z + \delta z, k) = \mathrm{TopK}(z, k).$$

Proof. Every selected logit decreases by less than $m/2$ and every unselected logit increases by less than $m/2$. Therefore the perturbed kth selected logit remains strictly larger than the perturbed $(k+1)$th logit, so no selected/unselected pair can cross the boundary. □

The theorem converts router margins into per-token error budgets. VSRAQ preserves ordering and margins through a calibration loss [36]; Eq. (33) instead uses Theorem 2 as a global chance constraint on accumulated upstream codec error.

7.5 Optimal rank and hybrid sparse–low-rank coding

For matrix $W = U\Sigma V^\top$, the best rank-r approximation is $W_r = U_r \Sigma_r V_r^\top$, with

$$\|W - W_r\|_F^2 = \sum_{i>r} \sigma_i^2.$$

If transformed coefficients are additionally sparse, use

$$W \approx U_y \, (L_r + S_k) \, U_x^\top, \tag{48}$$

where L_r is rank r and S_k has k nonzeros or nonzero blocks. The storage is approximately

$$M(r,k) \approx b_L r(m+n) + k(b_S + b_{\text{index}}) + M(U_x, U_y). \tag{49}$$

This representation is useful when a smooth common component is low-rank and sharp expert innovations are sparse.

7.6 High-rate bit allocation

Suppose block i has high-rate distortion model

$$D_i(R_i) = c_i 2^{-2R_i/d_i},$$

where d_i is block dimension and c_i captures sensitivity and variance. The Lagrangian

$$\mathcal{J} = \sum_i c_i 2^{-2R_i/d_i} + \lambda \left(\sum_i R_i - B \right)$$

gives

$$R_i^* = \frac{d_i}{2} \left[\log_2 \left(\frac{2 \ln 2 \, c_i}{\lambda d_i} \right) \right]_+ . \tag{50}$$

Hence high-variance, high-sensitivity blocks receive more bits, while blocks below a threshold receive zero bits. In practice, discrete codec choices replace the continuous approximation, but Eq. (50) provides initialization and intuition.

7.7 Shared expert storage

Suppose E experts each contain $P_{\ell,t}$ scalar parameters. Independent b-bit coding costs $EP_{\ell,t}b$. Under Eq. (23), let one prototype cost $P_{\ell,t}b_B$, R shared atoms cost $RP_{\ell,t}b_S$, each expert have R coefficients at b_a bits, and each innovation retain fraction ρ at b_E bits plus index overhead b_I. The total bits are

$$B_{\text{atlas}} = P_{\ell,t}b_B + RP_{\ell,t}b_S + ERb_a + E\rho P_{\ell,t}(b_E + b_I). \tag{51}$$

The compression factor relative to independent coding is

$$\Gamma_{\text{atlas}} = \frac{EP_{\ell,t}b}{P_{\ell,t}b_B + RP_{\ell,t}b_S + ERb_a + E\rho P_{\ell,t}(b_E + b_I)}. \tag{52}$$

Substantial gains require small R, sparse innovations, or shared atoms that replace many full experts. This equation gives a measurable criterion for whether expert similarity is sufficient.

7.8 Expert-graph packet storage and causal smoothness

Let $\mathscr{W} \in \mathbb{R}^{E_\ell \times P_{\ell,t}}$ be the expert signal and let Ψ be an orthogonal expert-graph transform. If packet index set Ω is retained, Parseval invariance gives

$$\|\mathscr{W} - \widehat{\mathscr{W}}\|_F^2 = \sum_{p \notin \Omega} \|\mathscr{C}_p\|_2^2 + \sum_{p \in \Omega} \|\mathscr{C}_p - \widehat{\mathscr{C}}_p\|_2^2. \tag{53}$$

More relevant than Frobenius error is causal graph smoothness

$$\mathcal{S}_{\text{cf}}(\mathscr{W}) = \operatorname{tr}(\mathscr{W}^\top L^E \mathscr{W}) = \frac{1}{2} \sum_{e,f} A^E_{ef} \|\widetilde{W}_e - \widetilde{W}_f\|_F^2. \tag{54}$$

Small $\mathcal{S}_{\text{cf}}$ is a necessary diagnostic for low expert-graph frequencies to dominate. The final acceptance test additionally requires low counterfactual substitution regret after reconstruction; weight smoothness alone is insufficient.

7.9 Monotonicity of an optional enhancement stream

Proposition 3 (Achievable distortion is non-increasing with optional packets)**.** *Let $\mathcal{B}_j$ be the set of decoders available after receiving base stream B_0 and the first j enhancement packets, with $\mathcal{B}_j \subseteq \mathcal{B}_{j+1}$ because a decoder may ignore packet $j+1$. For any distortion functional D,*

$$D^*_{j+1} = \inf_{g \in \mathcal{B}_{j+1}} D(g) \leq \inf_{g \in \mathcal{B}_j} D(g) = D^*_j.$$

This guarantees monotonicity of the best available reconstruction, not of a fixed greedy decoder. Every enhancement packet is therefore validated and may be rolled back when it harms a declared task or safety metric.

7.10 Paging constraint

Let p_e be the probability that expert/refinement e is needed, S_e its pageable bytes, q_e the cache hit probability, and B_{ssd} sustainable SSD bandwidth. Expected I/O time per token is bounded below by

$$t_{\text{io}} \geq \frac{\sum_e p_e(1-q_e)S_e}{B_{\text{ssd}}}. \tag{55}$$

Interactive decoding requires high cache hit rates or small refinements. Loading complete experts token-by-token is unlikely to be viable; the runtime must prefetch at the prompt, segment, or capability-pack level and use resident low-pass approximations during misses.

8 Compression Algorithms

9 Apple-Silicon Runtime Architecture

9.1 Framework choice

MLX is the preferred research runtime because it is designed for Apple-silicon unified memory, supports CPU/GPU execution, lazy computation, quantized models, and LLM inference/fine-tuning through MLX-LM [3, 21]. A second backend based on llama.cpp/GGUF is valuable for portability and highly optimized low-bit CPU/Metal kernels. The research format should be independent of either backend and converted into backend-specific packed layouts.

9.2 Packed SPECTRA format

Each compressed tensor record contains:

- tensor identity, original shape, transform group, and checksum;
- input/output permutations or shared transform identifiers;
- subband tree and coefficient block descriptors;
- codec type, bit width, group size, scale/zero metadata;
- codebook identifiers and packed indices;
- sparse outlier coordinates and values;

Algorithm 1 Calibration and transform learning

Require: Teacher checkpoint $\mathcal{M}_T$, calibration corpus $\mathcal{D}_{\text{cal}}$, sketch rank r_s, channel graph degree k
Ensure: Layer statistics, fast transforms, expert similarity graph

```
Initialize streaming covariance/Fisher sketches and routing counters
for all calibration batches B ⊂ D_cal do
    Run teacher with hooks on selected layers and experts
    Update robust activation moments and randomized covariance sketches
    Update router frequency, co-selection, entropy, and transition statistics
    Accumulate teacher logits, anchor hidden states, and safety outputs
end for
for all compressible layer groups G do
    Construct sparse channel graphs from covariance/Fisher similarity
    Compute spectral ordering or approximate graph basis
    Fit fast transform T_θ (permutation + DCT/Hadamard/butterfly/Givens)
    Measure MP bulk edges, transform coding gain, and activation-weighted reconstruction curves
end for
for all MoE layers do
    Screen experts with routing/output features, then estimate counterfactual substitution regret
    Build sparse causal expert graph and select packet tree/cluster order by held-out distortion
end for
return statistics, transforms, clusters
```

- low-rank factor descriptors;
- causal expert-graph identifier, packet-tree node, capability address, and prefetch priority;
- base/enhancement dependency, nested-codebook level, and rollback checksum;
- router-margin calibration statistics and certified-risk target;
- reconstruction error statistics and calibration provenance.

The format should support memory mapping so that pageable bands are not copied into anonymous memory before use.

9.3 Packet granularity and router-addressable indexing

The enhancement packet cannot be arbitrarily small: tiny packets increase metadata, page faults, and IOPS; large packets overfetch unrelated experts and subbands. Let s_p be packet p's payload bytes, π_p its probability of demand in the current session profile, o_p the expected fraction of bytes read but unused, t_{seek} measured random-access latency, and B_{seq} measured sustained sequential SSD bandwidth. We select grouping G and packet sizes by minimizing

$$J_{\text{I/O}}(G) = \sum_{p \in G} \pi_p \left[\frac{s_p(1+o_p)}{B_{\text{seq}}} + t_{\text{seek}} \, \mathbf{1}_{\text{miss}}(p) \right] + \lambda_{\text{meta}} M_{\text{index}}(G) + \lambda_{\text{waste}} \sum_{p \in G} \pi_p s_p o_p. \tag{56}$$

The hardware terms are measured at installation time because internal and external Apple-silicon SSDs differ materially.

Algorithm 2 SPECTRA-MoE compression

Require: Teacher $\mathcal{M}_T$, statistics/transforms, memory target M_w, codec set $\mathcal{Q}$
Ensure: Resident core $\mathcal{M}_0$, pageable packs $\{\Delta\mathcal{M}_c\}$, manifest

for all layers and expert clusters **do**
    Transform weights into activation-aligned coordinates
    Fit shared prototype/atoms and construct the causal expert graph
    Apply channel filter banks and expert-graph wavelet packets
    Compute sparse specialization innovations
    Generate candidate representations for every band:
     prune, sparse, low-rank, scalar quantized, VQ/AQ, residual-coded
    Estimate task distortion, bytes, decode cost, and page cost of each candidate
end for
Initialize rank/precision by reverse water-filling; solve discrete allocation with router-margin chance constraints
Assemble resident low-pass bands and critical high-sensitivity blocks
Encode a base stream plus random-access nested enhancement packets
Fit low-rank/sparse repair and Fisher-shaped feedback quantizers under remaining budget
Run constrained spectral distillation
Re-estimate memory and latency; repeat allocation if constraints are violated
Export packed tensors, transforms, codebooks, routing metadata, checksums

The default search space uses 256 KiB, 512 KiB, 1 MiB, 2 MiB, 4 MiB, and 8 MiB payload targets. Co-demanding subbands are grouped using router-transition and capability-profile traces, then laid out contiguously. The manifest uses a two-level index

$$(\ell, \text{matrix type}) \to (\text{expert/cluster}, \text{subband}, \text{quality level}) \to \text{packet id},$$

with packet offsets, lengths, codec identifiers, checksums, and prefetch priorities. With 100,000 packets and a 64-byte fixed index record, the fixed index is only about 6.1 MiB; the dominant cost is therefore I/O overfetch, not metadata. A packetization is accepted only if warm-cache and cold-cache p95 latency both improve over a monolithic per-expert layout at equal model quality.

9.4 Kernel fusion

A naive implementation reconstructs $\widehat{W}$ and then performs matrix multiplication, incurring prohibitive memory traffic. Required fused kernels include:

1. transform–dequantize–matmul for fixed Hadamard/DCT/butterfly transforms;
2. codebook lookup–accumulate for product/additive quantization;
3. sparse innovation accumulation;
4. low-rank correction matmul;
5. fused router and expert prefetch prediction;
6. quantized KDA/MLA state update.

Metal threadgroup memory should hold tiles, scales, and codebook fragments. Transform stages should be selected to minimize global-memory passes.

Algorithm 3 Runtime expert/subband scheduling

Require: Prompt $x_{1:T}$, resident base stream, enhancement packets, draft/router predictor, cache capacity C

Classify modality, task, language, tools, and capability profile
Load selected capability packs asynchronously
Run prompt prefill with resident core and available refinements
for $t = T + 1, \ldots$ **do**
    Draft/router predictor forecasts expert neighborhoods and enhancement addresses
    Prefetch high-probability packets; evict by byte-cost- and miss-penalty-aware policy
    for all layers ℓ **do**
        Apply resident low-pass operator
        if required refinement is resident **then**
            Add decoded sparse/subband correction
        else
            Use coarse approximation and mark uncertainty
        end if
    end for
    Generate token and update compressed context state
    if uncertainty or verification risk exceeds threshold **then**
        Increase active bands, retrieve external memory, or run a verifier pass
    end if
end for

9.5 Memory tiers

The runtime uses four tiers:

1. **Tier 0—resident critical:** embeddings, norms, routers, attention core, shared low-pass expert atlas, output head.
2. **Tier 1—resident adaptive:** most likely capability packs and recently used detail bands.
3. **Tier 2—memory-mapped SSD:** compressed expert innovations and rare modality packs.
4. **Tier 3—external knowledge:** retrieval indexes and original documents, not model weights.

The system must avoid macOS swap amplification. Peak resident memory, not only nominal checkpoint size, is a release criterion.

9.6 Context manager

A bounded-memory context manager maintains:

$$\mathcal{C}_t = \mathcal{C}_t^{\text{recent}} \cup \mathcal{C}_t^{\text{anchors}} \cup \mathcal{C}_t^{\text{summary}} \cup \mathcal{C}_t^{\text{retrieval}}. \tag{57}$$

Recent tokens remain exact within the model's local window. Anchors preserve high-attention or high-surprisal events. Summaries are recursively compressed at increasing temporal scales. Retrieval points to external chunks. The teacher is used during training to label which facts, constraints, and reasoning state must survive compaction.

10 Experimental Methodology

10.1 Evidence status in this revision

The original review correctly identifies the absence of empirical evidence as the dominant weakness. This revision does *not* convert unrun experiments into claims. It separates evidence into three levels:

1. **Released-checkpoint evidence:** architecture, quantization metadata, layer schedule, and shard layout are taken from the public K3 model card/configuration and technical-report repository.

2. **Implementation smoke tests:** the accompanying SPECTRA repository has executable unit and end-to-end synthetic tests for transforms, packed quantization, expert-graph coding, routing certificates, progressive packets, and checksum-verified reconstruction.

3. **Model-quality experiments:** the conference-critical OLMoE/K2/K3 evaluations below are pre-registered protocols. Until executed, the manuscript does not claim task-quality superiority.

This distinction directly addresses the reviewer's concern that a research plan must not be presented as measured model performance.

10.2 Preliminary implementation smoke test

The public implementation at `https://github.com/andrew-jeremy/spectra-k3` was exercised on deterministic synthetic Kimi-style tensors. Seventeen automated tests covered 1–8 bit packing, grouped symmetric quantization, odd-shape Haar reconstruction, expert-graph transform inversion, shared-atlas factorization, calibration-bundle serialization, spectral/logit/router distillation losses, route-margin certification, global codec allocation, packet SHA-256 verification, quantized state anchors, router-based prefetch, distributed artifact assembly, and a complete synthetic checkpoint conversion.

The expert-graph smoke test uses a weight-distance proposal proxy rather than true counterfactual regret; it therefore tests the packet/transform code path but does not validate the causal hypothesis. The 17-test suite was rerun on August 20, 2026. On the same deterministic synthetic seed, Fisher-shaped feedback produced weighted MSE 9.49×10^{-5} versus 9.28×10^{-5} without shaping; therefore no improvement is claimed for that component, and the zero-feedback fallback in Section 11 remains active until a real-model matched-rate experiment demonstrates benefit.

10.3 Gate 1: fully open OLMoE pilot

The first publishable model experiment uses `allenai/OLMoE-1B-7B-0924-Instruct`. OLMoE is fully open and has 7B total parameters, 64 experts, eight experts selected per token, 16 layers, hidden size 2048, and expert intermediate size 1024 [34]. It fits comfortably on a 32 GB Mac in BF16, making full counterfactual substitution, route logging, and end-to-end evaluation feasible without a cluster.

Frozen calibration. Use exactly 262,144 text tokens, seed 20260820, sequence length 2048, drawn from a fixed manifest of non-evaluation text. Use 75% for transform/codec fitting, 12.5% for router-certificate calibration, and 12.5% as a held-out compression-validation split. Causal estimation uses $k_{\mathrm{cf}} \in \{4, 8\}$ candidate neighbors and at most 32 audited expert-pair families per layer.

Compression operating points. Report resident model budgets corresponding to 8, 6, 4, 3, 2, and 1 effective bits per original scalar where a baseline supports that regime, plus equal-byte budgets derived from 50%, 35%, 25%, and 15% of BF16 teacher bytes. Metadata, codebooks, sparse indices, and transforms count toward the budget.

Baselines. At minimum compare BF16, RTN/grouped quantization, GPTQ, AWQ, AQLM, per-expert activation-aware SVD plus quantized residual, routing-frequency expert pruning, output-similarity merging, a MoBE-style shared-basis factorization, VSRAQ-style routing-consistent quantization, and SPECTRA variants. QMoE is included on architectures supported by its reference implementation and is additionally evaluated on its native Switch-Transformer setting because it provides the strongest sub-1-bit trillion-MoE storage precedent [7, 17, 36]. RQ-MoE is a vector-codebook baseline for the quantizer component rather than an expert-weight compressor [47].

Frozen evaluation tasks. Use WikiText-2 perplexity; HellaSwag, ARC-Challenge, ARC-Easy, PIQA, WinoGrande, and MMLU accuracy through a pinned lm-evaluation-harness commit; 1,000 held-out prompts for teacher/student logit KL and route-set agreement; and system metrics on Apple silicon (peak unified memory, p50/p95 TTFT, median/p95 inter-token latency, tokens/s, bytes read from SSD/token, cache-hit ratio). For route stability, report top-k Jaccard, exact route-set match, margin-stratified flip probability, and expected regret conditioned on a flip.

Gate-1 acceptance criterion. The causal graph hypothesis proceeds only if it improves downstream retention over the strongest shared-basis/quantization baseline by at least 0.5 percentage points absolute at equal bytes on the predeclared aggregate, or reduces bytes by at least 10% at statistically indistinguishable retention. If neither condition holds, the causal graph path is disabled and the project falls back to the shared-basis/per-expert codec described in Section 11.

10.4 Gate 2: Kimi K2/K2.5 trillion-parameter surrogate

Kimi K2/K2.5 remains the first trillion-parameter end-to-end compression target because its ecosystem and existing MoBE results provide a strong comparison point. Compression is performed offline on distributed/high-memory hardware. The evaluation repeats the OLMoE metrics at budgets determined by storage and active-weight traffic. Required baselines include the native checkpoint, MXFP4/4-bit baselines where applicable, QMoE if adapted successfully, MoBE, routing-consistent quantization, expert pruning/merging, and SPECTRA with each primary hypothesis ablated. The exit criterion is a Pareto improvement over MoBE or the best feasible baseline in at least one task-conditioned profile without violating routing/safety constraints.

10.5 Gate 3: released K3 shard audit

Before full-teacher calibration, the 96 released K3 shards are scanned one at a time to avoid a multi-terabyte resident footprint. Select early, middle, and late KDA/MLA-cycle layers and compute, for every expert tensor type: singular-value decay, Marchenko–Pastur bulk diagnostics, transform coding gain, cross-expert shared-basis rank, innovation sparsity, and candidate graph smoothness. The audit exports only summary statistics and small basis/sketch artifacts. It is considered successful if at least one structural codec family yields a projected net byte reduction after transform/codebook/index overhead on a majority of sampled MoE layers. Otherwise full K3 compression is not attempted.

10.6 Gate 4: full K3 offline compression and 32 GB deployment test

Only after Gates 1–3 pass is the full teacher processed. The released K3 checkpoint is calibrated and compressed on distributed/high-memory infrastructure; the resulting resident core and pageable packets are copied to the 32 GB Mac. The Mac is evaluated only as the deployment runtime. The primary release target remains peak process-plus-GPU memory below 30 GiB with at least 95% task-conditioned retention on a predeclared profile, but failure to reach this target is a valid scientific outcome.

10.7 Benchmark and artifact manifest

Every experiment records model revision/hash, tokenizer/chat template, calibration manifest, random seed, codec per tensor, transform seed, rank, sparsity, effective bits including metadata, packetization, cache policy, benchmark commit, prompt template, sampling parameters, Apple hardware identifier, macOS/MLX versions, cold/warm system traces, and raw item-level results where licensing permits. The repository contains machine-readable YAML manifests matching this section.

10.8 Main result tables

11 Feasibility Dependencies and Contingency Plan

The pipeline is intentionally staged so that failure of one hypothesis does not invalidate the entire system.

Table 7: Dependency-aware contingency plan.

Failure / diagnostic	Trigger	Fallback
Causal graph too expensive	audit budget exceeded or confidence intervals remain wide	keep sparse observational graph only for packet co-demand; compress experts independently or with MoBE-style shared bases; make no causal claim.
Causal graph not smooth	graph coding gain or held-out regret fails Gate 1	disable expert-axis wavelets for that layer; use shared-basis + low-rank/sparse residual codec.
Activation transform has no gain	transform coding gain near one or kernel overhead erases byte savings	use identity/Hadamard whitening and direct VQ/AQLM/QMoE-style coding.
Router certificate vacuous	margins are too small for useful chance constraints	pin router path and upstream sensitive layers at higher precision; optimize empirical route consistency instead of claiming certification.

Failure / diagnostic	Trigger	Fallback
Fisher error feedback ineffective	no matched-rate reduction in held-out task distortion	set feedback operator to zero and retain ordinary residual VQ.
Progressive packets cause I/O stalls	p95 miss penalty exceeds latency budget	increase packet size, preload at prompt/session granularity, or ship static capability profiles.
32 GB target cannot meet quality gate	no Pareto point reaches declared retention within 30 GiB peak	publish the measured frontier and recommend the strongest natively fitting local model rather than claiming success.

12 Risk Analysis and Failure Modes

12.1 Insufficient spectral compressibility

Weight matrices may not become sparse under any transform cheap enough to store and execute. Activation graphs may be unstable across domains, or singular spectra may decay too slowly. The mitigation is to measure transform gain before designing kernels and to use model distillation when post-training representation is inadequate.

12.2 Expert diversity is functional, not redundant

Experts may appear similar under average calibration data while encoding rare but important capabilities. Aggressive merging can damage tail performance, safety, or multilingual behavior. Mitigations include rare-route oversampling, worst-case distortion objectives, expert leave-one-out tests, and preserving high-uncertainty innovations on SSD.

12.3 Dynamic paging destroys latency

Even a small average active set can produce random SSD access and latency spikes. The runtime should load capability packs at prompt boundaries, forecast expert demand several layers/tokens ahead, use large contiguous pages, and maintain a usable low-pass fallback. If the p95 latency target is not met, pageable refinement must be reduced or moved to session-level rather than token-level loading.

12.4 Compression changes routing

Small errors before an MoE router can change the top-k expert set discontinuously. This creates a feedback loop in which compressed activations visit experts not represented by the calibration distribution. Router margins should be measured, router inputs and weights assigned high precision, and distillation should explicitly match routing distributions and top-k sets.

12.5 Causal calibration cost is prohibitive

Counterfactual substitutions require additional forward evaluations and can be biased by the finite set of executed routes. The causal graph may be too expensive to estimate at K2/K3 scale. Mitigations include two-stage screening, low-rank Jacobian sketches, batched expert substitution, active pair selection, and rejecting causal graph coding when confidence intervals are too wide. Observational metrics may be used only as a proposal distribution, not as final evidence of substitutability.

12.6 Enhancement packets are not successively refinable in task space

An extra packet may improve average reconstruction error while damaging a particular task, router boundary, or safety behavior. Every packet must be independently validated, signed, and reversible. The system reports per-profile metrics rather than assuming that more bits always improve every capability.

12.7 Quantization-aware teacher is already compressed

K3's MXFP4 training reduces available precision redundancy. Further post-training quantization may cause larger degradation than on BF16 teachers. The research should prioritize structural redundancy and shared expert codes, not assume that moving from 4 to 2 bits will be free.

12.8 Multimodal and agentic parity is especially difficult

A local text-only student can score well on text benchmarks while failing to reproduce native vision, tools, long-horizon planning, or agent-swarm behavior. Claims must distinguish core language-model compression from the broader production system, which includes harnesses, tools, retrieval, execution environments, and post-training.

12.9 Teacher access and licensing

The K3 weights are now publicly released under the Kimi K3 License [29]. Offline compression therefore uses the released checkpoint rather than API-only teacher access, subject to the license and redistribution requirements. The 32 GB deployment artifact remains a derivative representation and must preserve required notices and comply with any applicable use restrictions.

13 Productionization Requirements

A production-level local release requires more than benchmark accuracy:

1. deterministic conversion with checksummed artifacts;
2. atomic capability-pack updates and rollback;
3. corruption detection for memory-mapped pages;
4. bounded allocator behavior under long sessions;
5. prompt-template and tokenizer fidelity;
6. tool-call schema compatibility;
7. streaming output, cancellation, and concurrency control;

8. telemetry for cache hit rate, page faults, memory pressure, and uncertainty;
9. safety-policy regression tests after every compression change;
10. reproducible model cards documenting excluded capabilities and contexts.

A local deployment should expose three profiles:

- **Fast:** resident low-pass core, shorter context, minimal refinements;
- **Balanced:** selected capability packs and 4/3-bit state;
- **Quality:** more bands, verifier passes, larger context memory, lower throughput.

This mirrors scalable signal decoders in which additional bits improve reconstruction quality.

14 Stage-Gated Research Roadmap

Table 8: Stage-gated work packages and hard exit criteria.

WP	Focus	Exit / failure criterion
WP0	Released K3 inventory	Pin checkpoint revision, parse 96-shard weight map/config, verify tensor schema and byte accounting.
WP1	OLMoE pilot	Run all frozen tasks/baselines; causal graph must meet the Gate-1 matched-byte or byte-saving criterion.
WP2	Calibration cost	Demonstrate bounded streaming-statistics state and record teacher-pass overhead; reject any statistic that requires unbounded activation archives.
WP3	Fast transforms/-codecs	Beat identity-coordinate quantization at matched bytes after transform metadata and kernel cost.
WP4	Expert representation	Beat per-expert and MoBE-style shared-basis codecs at a declared operating point, or disable causal wavelets.
WP5	Router protection	Held-out route-flip confidence bound must predict empirical flips; otherwise downgrade certification to empirical alignment.
WP6	Packetization	Select packet granularity from measured I/O curves; progressive layout must beat monolithic expert paging in cold/warm p95 latency.

WP	Focus	Exit / failure criterion
WP7	K2/K2.5 scale-up	Demonstrate a Pareto improvement over the strongest feasible trillion-scale baseline in a declared task profile.
WP8	K3 shard audit	Sampled layers must show net structural coding gain after all metadata/transform overhead; otherwise stop before full-teacher processing.
WP9	Full K3 offline compression	Produce checksummed resident-core + refinement artifact and complete quality/safety evaluation on distributed teacher infrastructure.
WP10	32 GB Apple deployment	Peak memory < 30 GiB; report retention, tok/s, TTFT, SSD I/O, and the opportunity-cost comparison. If the quality target fails, publish the measured Pareto frontier.

15 Novelty and Scientific Significance

The reconciled manuscript is materially stronger than a generic "wavelets plus quantization" proposal: it introduces random-matrix diagnostics, reverse-water-filling rank/bit allocation, a shared expert representation, and a hardware-aware multirate runtime. However, several of those ideas now have close prior or contemporaneous work. The defensible novelty should therefore be stated at the level of precise mechanisms and falsifiable departures, not the mere integration of known components.

The four primary hypotheses create a coherent signal-processing program. The causal graph defines the domain on which expert frequency is meaningful; graph wavelet packets create a multiresolution representation; Fisher-shaped feedback and vector quantization code the retained coefficients; router-margin constraints protect the discrete computation path; and the successive-refinement container maps the representation onto a memory hierarchy. This is more specific than an integration claim and can fail independently at each stage.

The scientific value extends beyond Kimi. The framework asks whether expert families are smooth under a causally meaningful graph, whether task-sensitive neural operators admit noise shaping, and whether model capability is approximately successively refinable. Negative results would establish useful limits: high causal graph variation would imply irreducible expert diversity; vacuous routing certificates would expose brittle computation paths; and non-monotone enhancement behavior would show that task capability is not organized like scalable media quality.

15.1 Claims deliberately not made

The manuscript does not claim that wavelets, graph Fourier transforms, vector quantization, reverse water-filling, shared expert bases, scalable coding, or speculative prefetch are individually new. It does not claim completed Kimi K3 compression or universal 95% retention. It proposes architecture- and systems-specific hypotheses whose novelty depends on evidence against the closest baselines in Table 9.

16 Limitations, Safety, and Responsible Claims

This proposal does not establish that Kimi K3 can be compressed to 32 GB with broad capability parity. The memory analysis shows that a large structural reduction is mandatory, and such reduction may require sacrificing tail capabilities or retraining a student. The released K3 checkpoint removes the earlier uncertainty about basic tensor dimensions, but it does not remove the central uncertainty: whether its expert families are structurally redundant enough for the proposed codec to retain useful capability at the required reduction ratio.

Compression can alter calibration, refusal behavior, privacy leakage, and robustness. Safety layers and router paths may be unusually sensitive to low-bit or spectral truncation. A compressed model should therefore undergo independent red-team evaluation and should not inherit the teacher's safety claims automatically.

Local deployment increases privacy and user control, but it also removes provider-side monitoring and can broaden access to powerful dual-use capabilities. Release decisions should consider model capability, licensing, safeguards, and the risk profile of capability packs.

Finally, benchmark similarity does not imply identical internal behavior. The goal is functional task retention under explicit constraints, not an assertion that a compact model is the same model.

17 Direct Answers to Reviewer Questions

Q1: Why not validate the causal expert graph on a smaller fully open MoE? Yes. The revision makes this the first mandatory empirical gate rather than an optional future step. OLMoE-1B-7B-0924-Instruct is selected because its 64-expert, top-8 architecture is fully open and fits on a 32 GB Mac, allowing true expert substitution, teacher-quality evaluation, and routing analysis in one reproducible environment [34]. K3-scale work does not proceed unless this pilot establishes a matched-budget benefit.

Q2: What is the calibration overhead on a trillion-scale teacher, and how can it fit in 32 GB? It is not required to fit in 32 GB because calibration is an offline teacher-side operation; the deployment machine only runs the exported compressed artifact. For K2/K2.5 ($L = 61$, hidden width 7168), rank-128 hidden-state covariance sketches require about 213.5 MiB versus about 11.7 GiB for dense FP32 covariances, while diagonal hidden Fisher vectors require about 1.7 MiB. At K3 dimensions the corresponding sketch is about 325.5 MiB versus 17.8 GiB dense, with about 2.5 MiB for diagonal hidden Fisher vectors. Time is capped in teacher-forward equivalents: one C0 forward pass, at most 0.15 forward-equivalents for the C1 sensitivity subset, 0.10 for local replay, and 0.25 for full downstream audits, for a total predeclared budget of at most $1.50F(N)$ over N calibration tokens. If a 5M-token teacher forward takes T_F hours on the declared cluster, calibration is budgeted at at most $1.5T_F$ hours excluding checkpoint staging. The paper gives the formulas and failure rule in Section 6.3.

Q3: How is enhancement-packet granularity chosen without creating an I/O bottleneck? Packetization is now an optimization problem, not a fixed heuristic. Candidate payload sizes from 256 KiB to 8 MiB are benchmarked against the actual target SSD. Router/capability co-demand determines grouping, and Eq. (56) jointly penalizes transfer time, miss latency, overfetch, and index metadata. A two-level layer/expert/subband index supports random access. The chosen layout must beat a monolithic per-expert paging baseline on both cold-cache and warm-cache p95 latency; otherwise the system falls back to larger session-level capability packs.

18 Conclusion

Running a Kimi K3-class model on a 32 GB Mac is not principally a scalar quantization problem. K3's 2.8 trillion parameters require approximately 1.27 TiB even at an idealized four bits per parameter, and the model is already trained with low-precision weights. A viable local system must remove or share most degrees of freedom, distill the behavior that matters, compress context state, and reconstruct specialization dynamically.

SPECTRA-MoE proposes a rigorous path based on established signal-processing concepts: decorrelating transforms, graph spectra, multirate filter banks, subband coding, vector quantization, sparse innovations, and rate–distortion allocation. These are combined with modern LLM techniques—activation-aware compression, low-rank repair, expert routing, knowledge distillation, and hardware-aware kernels—to create a resident semantic backbone with pageable refinements. The program is deliberately falsifiable: it specifies memory lower bounds, quality definitions, baselines, ablations, and production metrics. The immediate empirical gate is the fully open OLMoE pilot, followed by K2/K2.5 scaling, a released-K3 shard audit, and finally full K3 offline compression only if the preceding gates succeed. Success would demonstrate that a substantial portion of frontier-model capability can be represented as a compact, multiscale signal rather than a monolithic checkpoint; failure would establish useful limits on the compressibility of modern MoE intelligence.

A Detailed Memory Budget

A candidate 32 GB deployment budget is: In practice, macOS and other applications may require more headroom. A safer release target is below 28–30 GiB peak process-plus-GPU allocation.

B Candidate Hyperparameter Ranges

C Reproducibility Checklist

Every published experiment should include:

- teacher checkpoint hash, license, tokenizer, and chat template;
- calibration dataset manifest and exclusion rules;
- transform seeds, graph construction, and approximation tolerance;
- per-tensor codec, bit width, scale format, sparsity, and rank;
- resident/disk byte accounting including metadata;
- MLX, macOS, compiler, and hardware versions;
- exact benchmark commits and prompt templates;
- cold/warm system traces and memory-pressure logs;
- raw item-level evaluation outputs where licensing permits;
- safety evaluation and known capability regressions.

D Suggested Repository Layout

```
spectra-moe/
  pyproject.toml
  README.md
  configs/
    hardware/macos_32gb.yaml
    models/kimi_k3_release.yaml
    compression/spectra_default.yaml
    experiments/olmoe_pilot.yaml
    experiments/k3_shard_audit.yaml
    runtime/packet_granularity.yaml
  spectra/
    calibration/
    transforms/
    graph_spectral/
    filterbanks/
    expert_atlas/
    quantization/
    rate_distortion/
    distillation/
    runtime_mlx/
    runtime_gguf/
    context/
    evaluation/
  metal/
    transform_qmatmul.metal
    aq_lookup_matmul.metal
    sparse_residual.metal
    quantized_state_update.metal
  scripts/
    collect_calibration.py
    analyze_spectra.py
    compress_checkpoint.py
    distill_repair.py
    export_spectra.py
    benchmark_local.py
  tests/
  results/
  manifests/
```

References

[1] Nasir Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

[2] Apple Inc. Metal. https://developer.apple.com/metal/, 2026. Accessed July 20, 2026.

[3] Apple Machine Learning Research. Mlx-lm: Run large language models with mlx, 2026. URL https://github.com/ml-explore/mlx-lm.

[4] avlp12. Kimi-k3-alis-mlx-dynamic-2.71bpw. Hugging Face community model artifact, 2026. URL https://huggingface.co/avlp12/Kimi-K3-Alis-MLX-Dynamic-2.71bpw. Community engineering artifact; not peer reviewed; accessed August 20, 2026.

[5] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[6] Ningning Chen, Weicai Ye, and Ying Jiang. Hbllm: Wavelet-enhanced high-fidelity 1-bit quantization for llms. *arXiv preprint arXiv:2512.00862*, 2025.

[7] Xiaodong Chen, Mingming Ha, Zhenzhong Lan, Jing Zhang, and Jianguo Li. Mobe: Mixture-of-basis-experts for compressing moe-based llms. In *International Conference on Learning Representations*, 2026.

[8] Wojciech Czaja and Sanghoon Na. Frame quantization of neural networks. *arXiv preprint arXiv:2404.08131*, 2024.

[9] Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *International Conference on Machine Learning*, 2019.

[10] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.

[11] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

[12] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[13] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[14] Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *International Conference on Machine Learning*, 2024.

[15] Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models. *arXiv preprint arXiv:2606.10703*, 2026.

[16] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, 2023.

[17] Elias Frantar and Dan Alistarh. Qmoe: Sub-1-bit compression of trillion-parameter models. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

[18] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

[19] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[20] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

[21] Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. Mlx: Efficient and flexible machine learning on apple silicon, 2023. URL https://github.com/ml-explore/mlx.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[23] Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[24] Kimi Team. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

[25] Kimi Team. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[26] Kimi Team. Kimi k3: Open frontier intelligence. https://www.kimi.com/blog/kimi-k3, 2026. Accessed August 20, 2026.

[27] Kimi Team. Kimi k3 released configuration. https://huggingface.co/moonshotai/Kimi-K3/blob/main/config.json, 2026. Accessed August 20, 2026.

[28] Kimi Team. Kimi k3: Open frontier intelligence. Technical report, Moonshot AI, 2026. URL https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf. Technical report, accessed August 20, 2026.

[29] Kimi Team. Kimi k3: Open frontier intelligence – model card and open weights. Hugging Face model repository, https://huggingface.co/moonshotai/Kimi-K3, 2026. Accessed August 20, 2026.

[30] Hoang-Loc La, Truong-Thanh Le, Amir Taherkordi, and Phuong Hoai Ha. Joint structural pruning and mixed-precision quantization for llm compression. *arXiv preprint arXiv:2606.07819*, 2026.

[31] Ismail Lamaakal. Motion-compensated weight compression. *arXiv preprint arXiv:2605.24754*, 2026.

[32] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024.

[33] Stéphane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[34] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh

Hajishirzi. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

[35] Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5): 2295–2317, 2011.

[36] Hancheol Park, Geonho Lee, Tairen Piao, and Tae-Ho Kim. Value-and-structure alignment for routing-consistent quantization of mixture-of-experts models. *arXiv preprint arXiv:2606.05688*, 2026.

[37] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 30(3): 83–98, 2013.

[38] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning*, 2024.

[39] Prateek Verma. Wavelet gpt: Wavelet inspired large language models. *arXiv preprint arXiv:2409.12924*, 2024.

[40] Wenfeng Wang, Jiacheng Liu, Xiaofeng Hou, Xinfeng Xia, Peng Tang, Mingxuan Zhang, Chao Li, and Minyi Guo. Moe-speq: Speculative quantized decoding with proactive expert prefetching and offloading for mixture-of-experts. *arXiv preprint arXiv:2511.14102*, 2025.

[41] Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *arXiv preprint arXiv:2403.07378*, 2024.

[42] Xing Wang and Jie Liang. Scalable compression of deep neural networks. *arXiv preprint arXiv:1608.07365*, 2016.

[43] Moritz Wolter, Shaohui Lin, and Angela Yao. Neural network compression via learnable wavelet transforms. *arXiv preprint arXiv:2004.09569*, 2020.

[44] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

[45] Yuchen Yang, Yaru Zhao, Pu Yang, Shaowei Wang, and Zhi-Hua Zhou. Zipmoe: Efficient on-device moe serving via lossless compression and cache-affinity scheduling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[46] Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

[47] Zhengjia Zhong, Shuyan Ke, Zaizhou Lin, Jiaqi Song, Hongyi Lan, and Hui Li. Rq-moe: Residual quantization via mixture of experts for efficient input-dependent vector compression. *arXiv preprint arXiv:2605.14359*, 2026.

Table 1: Released Kimi K3 architecture used in this revision.

Quantity	Released value
Total parameters	2.8T
Activated parameters per token	104B
Transformer layers	93
Dense layers	1
Attention composition	69 KDA + 24 Gated MLA
Attention hidden dimension	7168
Attention heads	96
Latent MoE dimension	3584
Per-expert hidden dimension	3072
Routed experts per MoE layer	896
Selected routed experts per token	16
Shared experts	2
Vocabulary	160K
Maximum context	1,048,576 tokens
Vision encoder	MoonViT-V2, 401M parameters
Training/deployment precision	MXFP4 weights / MXFP8 activations after QAT
Checkpoint layout	96 Safetensors weight shards

Table 3: Idealized storage of 2.8 trillion parameters. Real formats require additional scales, codebooks, sparse indices, alignment, and runtime buffers.

Representation	Bits/parameter	Decimal TB	Binary TiB
FP16/BF16	16.00	5.600	5.093
INT8/FP8	8.00	2.800	2.547
MXFP4/INT4 idealized	4.00	1.400	1.273
3-bit	3.00	1.050	0.955
2-bit	2.00	0.700	0.637
Ternary entropy floor proxy	1.58	0.553	0.503

Table 4: Measured software smoke-test results. These validate implementation paths only; they are not Kimi K3 or open-MoE quality results.

Test	Base / source	Enhanced / artifact
Synthetic checkpoint bytes	296,960	∼79.1 KB artifact
Attention projection relative MSE	0.761	0.052
Eight-expert source bytes	131,072	16,504
Expert-graph relative MSE	0.150	0.018
Route-certificate violations	0/128 synthetic perturbations	
Automated tests	17 passed	

Table 5: Required OLMoE pilot result table. Blank cells are intentionally unclaimed until the experiment is run.

Method	GiB	eff. bits	PPL	Retention	Route match / systems
BF16 teacher		16		1.000	
AWQ/GPTQ 4-bit		4			
AQLM 2-bit		2			
MoBE-style shared basis					
VSRAQ-style quantization					
SPECTRA without causal graph					
SPECTRA full					

Table 6: Required 32 GB K3 deployment table. Blank cells remain blank until the full experiment exists.

Method	Res. GiB	Disk GiB	Ret.	tok/s	p95 TTFT	SSD GB/tok
Native released checkpoint	–	∼1454	1.000	–	–	–
Community low-bit reference	–	∼948	–	–	–	–
SPECTRA static						
SPECTRA progressive						

Table 9: Novelty matrix. "Primary" denotes a proposed research claim; "enabler" denotes an incorporated technique that is not claimed as original.

Concept	Closest precedent	Proposed departure and falsification test
Causal expert-graph wavelet packets (Primary)	Shared expert bases/-factorizations such as MoBE; causal audits of expert importance	Build the expert graph from counterfactual substitution regret and apply a multiscale transform along the expert axis. Reject if it does not beat shared-basis/prototype coding at matched bytes and route loss.
Router-margin-certified rate allocation (Primary)	Router-alignment and margin-preserving quantization such as VSRAQ	Impose held-out chance constraints on accumulated upstream logit perturbation inside a global memory/latency/I/O allocation. Reject if certificates are vacuous or do not predict route flips.
Fisher-shaped error feedback (Primary)	Sigma–Delta/frame quantization and activation-aware quantization	Shape graph/wavelet coefficient error with an activation/Fisher metric before low-bit VQ. Reject if matched-rate task distortion does not improve.
Router-addressable successive-refinement stream (Primary, application-level)	Scalable neural-network parameter compression and progressive media coding	Encode a complete base model plus independent random-access expert/capability enhancements with rollback. Reject if metadata erases coding gain, or if profiles are not monotone after validation.
Joint truncation and precision (Qualified)	Reverse water-filling; contemporary joint pruning/architecture and quantization search	Use one MoE-wide allocation spanning channel bands, expert-graph bands, repair rank, residency, and I/O. The novelty is the constrained scope, not reverse water-filling itself.
Speculative expert prefetch (Enabler)	MoE-SpeQ and related offloading systems	Reuse as a systems baseline and integrate with enhancement addresses; no standalone novelty claim.
Cross-layer prediction (Enabler)	Motion-compensated/inter-layer weight codecs	Optional candidate codec for compatible layers; not a primary contribution.

Table 10: Illustrative memory allocation. Values are design targets, not measured results.

Component	Budget
Compressed resident weights and codebooks	20.0 GiB
Runtime, transforms, routing, metadata	1.5 GiB
Activations and temporary matmul workspace	2.5 GiB
KV/KDA/MLA/context state	3.0 GiB
Vision encoder or optional modality workspace	1.0 GiB
Application and safety/verifier buffers	1.0 GiB
Operating-system/headroom reserve	3.0 GiB
Total	32.0 GiB

Table 11: Initial search ranges.

Parameter	Candidate values
Calibration tokens	1M, 5M, 20M, task-stratified
Graph degree k	8, 16, 32, 64
Transform family	permutation+DCT, Hadamard, Givens, butterfly, graph polynomial
Wavelet family	Haar, CDF 5/3, CDF 9/7, learned lifting
Decomposition levels	1–5, shape dependent
Expert cluster count	chosen by distortion elbow and routing constraints
Shared spectral atoms R	1, 2, 4, 8, 16
Innovation sparsity ρ	0.1%, 0.5%, 1%, 2%, 5%, 10%
Weight codecs	8, 6, 5, 4, 3, 2 bit; ternary; VQ/AQ; sparse outlier
KV/state bits	8, 4, 3, 2 with anchor/reset intervals
Repair rank	4–256, layer dependent
Capability-pack cache	2–8 GiB within resident budget