Critical Subspaces for Emergent Abilities Under Compression: Certified Mixed-Precision Quantization via Margin Sensitivity

Liz Lemma Future Detective

January 17, 2026

Abstract

Quantization is essential for deploying 2026-era large language and reasoning models under tight latency and memory budgets, yet prior work shows that extreme low-bit quantization (e.g., 2-bit) can collapse emergent abilities to near-random performance, while 4-bit often preserves them. We propose a formal framework that explains and exploits this threshold behavior: emergent task success is controlled by a margin that crosses a baseline threshold, and quantization reduces this margin through structured noise. We define parameter/activation groups and estimate their task-aggregated margin sensitivities near emergence tipping points, identifying a small critical subspace whose precision determines whether emergent abilities survive. We then formulate mixed-precision quantization as a budgeted selection problem, prove NP-hardness, and provide approximation algorithms with explicit guarantees (and an FPTAS in a linearized regime). Our main certified guarantee shows that if the unquantized model has sufficient verified margin, then protecting the critical subspace at higher precision while aggressively quantizing the remainder preserves above-baseline performance for emergent tasks. We outline experiments to validate the bound's tightness, reproduce known 2-bit collapse vs 4-bit preservation, and demonstrate near-2-bit average mixed precision that retains both reasoning capability and key safety behaviors, linking emergent abilities to localized circuit fragility emphasized in recent emergence surveys.

Table of Contents

1. 1. Introduction: emergence thresholds, quantization brittleness, and the 'critical subspace' hypothesis; positioning relative to emergence-as-loss and emergence-as-metric debates.

2. 2. Preliminaries and Task Model: verifiers, margins, emergent-vs-nonemergent tasks, baseline performance definitions; discussion of se-

mantic vs syntactic metrics.

3. 3. Quantization as Structured Noise: groupwise quantization model (weights, activations, KV cache), variance functions v_i(b), and independence/weak-dependence assumptions.

4. 4. Sensitivity Measures and Certified Margin Degradation: defining per-group sensitivity s_{t,i} via gradients/Jacobians; deriving expected and high-probability bounds on margin drop.

5. 5. Optimization Problem: Budgeted Mixed-Precision Allocation: formal problem statement; connections to knapsack/submodular maximization; definition of 'critical subspace'.

6. 6. Algorithms: (a) Greedy approximation for submodular benefit under knapsack; (b) DP/FPTAS for linearized special case; (c) Practical estimation and batching tricks for large models.

7. 7. Theoretical Results: correctness and approximation guarantees; NP-hardness; lower bounds on information required to select the critical subspace.

8. 8. Experimental Protocol (flagged as strengthening): datasets/tasks (emergent reasoning, arithmetic, MMLU slices), baselines (uniform 2/4/8-bit), evaluation of certified bounds vs observed drops; safety-behavior retention tests.

9. 9. Discussion: when assumptions fail (correlated noise, post-quant finetuning), extensions (LoRA recovery, activation/KV quantization), and implications for deploy-time policy.

10. 10. Limitations and Open Problems: tightness of bounds, mechanistic alignment with circuits, and adaptive quantization under distribution shift.

1 Introduction

Empirical "emergence" in large language models is often operationalized as the appearance of above-baseline performance on a task once the model (or its training compute) crosses a regime boundary. Although the phenomenology is familiar, two aspects are in tension when one attempts to make emergence actionable for deployment. First, emergence is frequently observed through discrete evaluation metrics whose dependence on underlying model scores is highly nonlinear. Second, deployment constraints (memory, bandwidth, latency) induce interventions—in particular low-bit quantization—that act as structured perturbations to the pretrained parameters and can interact sharply with such nonlinearities. Our guiding premise is that emergence should be studied at the level of *margins* (logit gaps or verifier-score gaps), where "above-baseline" behavior corresponds to a margin exceeding a task-dependent threshold. In this view, quantization brittleness is not mysterious: if the unquantized model sits near an emergence threshold, then even moderate perturbations can flip the sign of the margin on a non-negligible fraction of inputs.

The practical problem is immediate. Modern post-training quantization and mixed-precision deployment can reduce memory footprints by an order of magnitude, yet it is well documented that capability regressions are not uniform across behaviors: some tasks degrade smoothly with bitwidth, while others collapse abruptly. The abrupt failures are particularly salient for tasks that are close to the regime boundary in the full-precision model—including multi-step reasoning, compositional generalization, and certain safety-relevant behaviors whose evaluation is mediated by a learned or programmatic verifier. A uniform-bit quantizer treats all parameters (or activation/KV-cache components) as equally responsible for these fragile margins, which is incompatible with observed heterogeneity across layers, heads, and modules.

We therefore study mixed-precision quantization through the lens of a *critical subspace* hypothesis: for a fixed set of tasks of interest, there exists a comparatively small subset of parameter (or activation) groups whose precision dominates the post-quantization margins. Informally, we posit that many groups can be aggressively quantized with little effect on the relevant margins, provided that a critical subset is preserved at higher precision. Our goal is not merely to state this hypothesis but to make it quantitative and algorithmic under an explicit noise model. Concretely, we model quantization as inducing an additive perturbation in each group, with controlled second moment that decreases with bitwidth, and we ask for a bit-allocation that maximizes a certified lower bound on task margins subject to a total bitrate budget. This yields a natural notion of "criticality": a group is critical to the extent that its quantization noise, when filtered through the local sensitivity of the margin functional, is capable of consuming the available margin slack.

Positioning this approach relative to existing interpretations of emergence clarifies what we do and do not claim. In the "emergence-as-loss" line of thought, a task appears emergent because small improvements in a smooth surrogate (e.g. cross-entropy) produce large changes in a downstream metric once the model crosses a decision boundary. In the "emergence-as-metric" view, the apparent discontinuity is largely an artifact of thresholded evaluation, finite sample size, or metric design. We remain agnostic to the extent to which either explanation accounts for any particular empirical curve; instead, we exploit a common structure shared by both. Regardless of whether the underlying learning dynamics are smooth, deployment decisions are ultimately made using discrete task scores; and these scores are mediated by some real-valued statistic—a logit gap, a calibrated verifier score, or another margin-like quantity—whose sign or magnitude determines correctness. Thus, even if emergence is a metric artifact, it is still the relevant artifact: preserving above-baseline performance under quantization requires controlling how that underlying statistic shifts.

The margin-centric view also suggests why mixed precision should be formulated as a structured optimization problem rather than a layerwise heuristic. If we denote by $m_t(x; \theta)$ the task-specific margin at input x, then the full-precision model is safely above baseline on task t when m_t typically exceeds a threshold η_t calibrated to the desired "emergent" regime (e.g. a correctness rate exceeding random-guessing or a verifier-induced baseline). Under quantization, the parameters become $\theta + \delta\theta$, and the relevant question is whether the degradation

$$m_t(x; \theta + \delta\theta) - m_t(x; \theta)$$

can be controlled uniformly over x in the evaluation distribution or, at minimum, in expectation with sufficiently small variance. This immediately reduces capability preservation to bounding the interaction between (i) the *noise magnitude* induced by assigning a given bitwidth to a group and (ii) the *sensitivity* of the margin to perturbations in that group. It is precisely this interaction that motivates a groupwise sensitivity score and a budgeted allocation procedure: when the cost of increasing precision is constrained, we should spend bits where the sensitivity-weighted noise reduction is largest.

Two further considerations motivate our emphasis on verifier-based semantic margins. First, for many tasks of interest—especially those involving reasoning, alignment, or safety—syntactic string matching is inadequate. Verifiers provide a principled way to define correctness via semantic equivalence classes or property checks, at the cost of introducing an additional model or program into the evaluation pipeline. Second, verifiers naturally induce continuous scores (or at least score differences) that can be used as margins; this is crucial for any certification argument, because discrete success indicators are too coarse to admit useful perturbation bounds without further structure.

From the deployment standpoint, our aim is to formalize the following implication: if the unquantized model has a certified margin slack above an emergence threshold, then there exists a mixed-precision assignment within the budget that preserves above-baseline success on all tasks in a designated emergent set. The certification requirement is not an aesthetic choice. Without an explicit bound relating the assigned bitwidths to margin degradation, mixed-precision selection becomes a combinatorial search guided only by empirical trial-and-error on finite calibration sets, which is both expensive and unreliable near thresholds. Conversely, once we can upper bound the margin loss as a function of per-group sensitivities and per-bit noise proxies, the selection problem becomes a knapsack-like optimization with transparent tradeoffs between memory and guaranteed retained margin.

In summary, we treat quantization as structured noise, emergence as thresholded margin behavior, and mixed precision as an optimization over bit allocations. The central object is a task-aggregated sensitivity profile over groups, which induces a notion of a critical subspace to preserve at higher precision. The remainder of the paper makes these statements precise: we specify the task and verifier model and the margin definitions, derive a degradation bound under the groupwise noise assumptions, and then reduce mixed-precision selection to a budgeted combinatorial problem with corresponding approximation algorithms.

2 Preliminaries and Task Model

We fix a pretrained model f_θ with parameters θ, and we study its behavior on a finite collection of tasks $\mathcal{T}$. Each task $t \in \mathcal{T}$ comes with an evaluation distribution (or a finite evaluation set) $\mathcal{D}_t$ over inputs x (prompts, questions, contexts), and a notion of semantic correctness implemented by a *verifier*. Our goal in this section is to isolate a real-valued *margin functional* $m_t(x;\theta)$ whose sign (or excess above a threshold) controls the task success indicator and which is sufficiently regular in θ to admit perturbation analysis in later sections.

Verifiers and semantic correctness. For each task t we assume a verifier V_t that, given an input x and a candidate output y, returns either a Boolean decision $V_t(x,y) \in \{0,1\}$ or, more generally, a score $V_t(x,y) \in \mathbb{R}$ that is monotonically related to correctness. The verifier may be programmatic (unit tests, theorem checker, constraint solver), symbolic (exact match up to normalization), or learned (a classifier or reward model). We emphasize that V_t is part of the task definition: it specifies which variations in surface form preserve correctness, and it allows us to treat correctness as a property of the pair (x,y) rather than as equality to a single reference string.

In particular, for open-ended generation tasks, syntactic metrics such as exact match or character-level overlap are often misaligned with semantic correctness; conversely, verifier-based semantics can declare multiple distinct strings correct when they are equivalent under a problem-specific relation. This is essential for our analysis: if correctness is defined only by syntactic coincidence, then the induced decision boundary can be arbitrarily brittle with respect to small perturbations of the model, while a semantic verifier typically induces a smoother separation in an appropriate score space (even if the verifier itself is discrete, it often arises from an underlying scoring procedure).

Model scoring and decoding. We assume the model induces a scoring function $\ell_\theta(x, y) \in \mathbb{R}$ over candidate outputs y given input x. Concretely, ℓ_θ may be the log-probability of y under an autoregressive model, a classifier logit for a label, or a reranker score. We also fix a decoding or selection rule $\hat{y}_\theta(x)$, typically

$$\hat{y}_\theta(x) \in \arg\max_{y \in \mathcal{Y}_t(x)} \ell_\theta(x, y),$$

where $\mathcal{Y}_t(x)$ is the relevant candidate set (e.g. multiple-choice options, a beam-search list, or an implicit large set). Our subsequent bounds are phrased in terms of margins defined from ℓ_θ and V_t; this choice abstracts away the particular decoding algorithm, provided the deployed procedure is consistent with maximizing (or approximately maximizing) ℓ_θ over a candidate set.

Margin functionals. Given V_t and ℓ_θ, we define a margin that compares the model's best verified-correct candidate to its best verified-incorrect candidate. Let

$$\mathcal{Y}_t^+(x) := \{y \in \mathcal{Y}_t(x) : V_t(x, y) = 1\}, \qquad \mathcal{Y}_t^-(x) := \mathcal{Y}_t(x) \setminus \mathcal{Y}_t^+(x).$$

When V_t is Boolean and both sets are nonempty, we set

$$m_t(x; \theta) := \max_{y \in \mathcal{Y}_t^+(x)} \ell_\theta(x, y) \; - \max_{y \in \mathcal{Y}_t^-(x)} \ell_\theta(x, y). \tag{1}$$

When V_t is real-valued, we may instead work with a calibrated score-gap margin, for instance

$$m_t(x; \theta) := V_t\big(x, \hat{y}_\theta(x)\big) \; - \max_{y \in \mathcal{Y}_t(x) :\, y \neq \hat{y}_\theta(x)} V_t(x, y), \tag{2}$$

or any task-specific variant that is (i) larger when the model is more confidently correct and (ii) differentiable (or almost everywhere differentiable) in θ through ℓ_θ and the selection rule used to define the competing outputs. In either case, the fundamental property is that positive margin implies semantic correctness under the induced decision rule in a suitably idealized setting. For (1), if the model selects the overall maximizer of ℓ_θ over $\mathcal{Y}_t(x)$, then $m_t(x; \theta) > 0$ implies $\hat{y}_\theta(x) \in \mathcal{Y}_t^+(x)$, hence $V_t(x, \hat{y}_\theta(x)) = 1$.

Success probabilities and baselines. We define the task success indicator under the deployed decoding rule by

$$\mathrm{Succ}_t(x;\theta) := \mathbf{1}\big\{V_t\big(x, \hat{y}_\theta(x)\big) = 1\big\}\,, \qquad p_t(\theta) := \mathbb{E}_{x\sim\mathcal{D}_t}\big[\mathrm{Succ}_t(x;\theta)\big].$$

We also fix a baseline success probability p_t^{base}, representing performance of a trivial strategy (random guessing for multiple choice, a null heuristic, or a verifier-implied prior). In our setting, "above-baseline" means $p_t(\theta) \geq p_t^{\mathrm{base}} + \epsilon$ for a target slack $\epsilon > 0$.

To connect these discrete success probabilities to margins, we introduce an emergence threshold $\eta_t \in \mathbb{R}$ such that exceeding η_t (in expectation, or with high probability over x) implies above-baseline success. In the simplest logit-gap setting, $\eta_t = 0$ is natural; for verifier-score margins (2), η_t may be calibrated on held-out data so that $\mathbb{P}[m_t(x;\theta) > \eta_t]$ tracks the desired success regime. Our later certification statements are stated in terms of η_t and the *slack* by which the unquantized model exceeds it.

Emergent versus non-emergent tasks. We partition $\mathcal{T}$ into an "emergent" subset $\mathcal{T}_{\mathrm{em}}$ and the remainder $\mathcal{T}_{\mathrm{non}}$. This partition is not meant to encode a claim about training dynamics; it is an operational designation of tasks whose performance is near threshold and hence susceptible to perturbations. Formally, one convenient criterion is that $t \in \mathcal{T}_{\mathrm{em}}$ if (on a calibration set) the empirical distribution of $m_t(x;\theta)$ places nontrivial mass near η_t, so that small negative shifts in margin can cause a large drop in Succ_t. Tasks in $\mathcal{T}_{\mathrm{non}}$ may still be important, but they typically exhibit larger slack and therefore admit more aggressive compression for the same certified guarantee. In the mixed-precision optimization we will aggregate task requirements across $\mathcal{T}_{\mathrm{em}}$ either by worst-case control (via a maximum over t) or by weighted objectives with weights w_t reflecting deployment priorities.

Calibration sets. For each task t we assume access to a finite calibration set $\{x_j^{(t)}\}_{j=1}^{n_t}$ drawn from (or representative of) $\mathcal{D}_t$. We use it to estimate baseline margins $\hat{\mu}_t := \frac{1}{n_t}\sum_j m_t(x_j^{(t)};\theta)$ and, later, to estimate sensitivity quantities derived from gradients of m_t with respect to subsets of θ. The role of calibration is purely algorithmic: it supplies the empirical quantities needed to compute a conservative degradation bound and to allocate precision under a deployment budget. The validation of any resulting allocation is performed on a separate evaluation set, but the certification logic is expressed in terms of the margin functionals introduced above.

3 Quantization as Structured Noise

We model mixed-precision quantization as a structured perturbation acting on a fixed set of disjoint groups. Concretely, we fix an index set $G =$

$\{1, \dots, g\}$ and a decomposition of the deployable state into groupwise blocks. In the simplest case the state is the parameter vector θ and we write

$$\theta \;=\; (\theta_1, \dots, \theta_g), \qquad \theta_i \in \mathbb{R}^{d_i},$$

where groups correspond to layers, submodules (e.g. attention projections, MLPs), or finer partitions such as per-head blocks. In deployments where activation and KV-cache quantization is relevant, we extend the state to include additional tensors that are produced and stored during inference. We will write this abstractly as

$$\zeta \;:=\; (\theta, a, k, v),$$

where a denotes quantized intermediate activations (or pre-activations) and (k, v) denotes the KV cache. The grouping G is then understood as a partition of all quantized quantities that contribute to memory/latency; we still denote the ith block by ζ_i and use the same notation $\delta\zeta_i$ for its quantization-induced perturbation.

Groupwise quantization operator. For each group i and allowed bitwidth $b \in \mathcal{B}$ we fix a quantizer $Q_i(\cdot; b)$ mapping ζ_i to a low-precision representation (uniform affine, per-channel, blockwise, etc.). A mixed-precision assignment $b = (b_1, \dots, b_g)$ induces the quantized state

$$Q(\zeta, b) \;:=\; \big(Q_1(\zeta_1; b_1), \dots, Q_g(\zeta_g; b_g)\big),$$

and we define the perturbation by the identity

$$Q(\zeta, b) \;=\; \zeta + \delta\zeta, \qquad \delta\zeta := (\delta\zeta_1, \dots, \delta\zeta_g). \tag{3}$$

When only weights are quantized, $\zeta = \theta$ and we write $\delta\theta$ accordingly. When activations and KV cache are quantized, (3) should be read as an analysis device: the quantized inference trace is equivalent to injecting additive perturbations into the corresponding tensors at the points where quantization occurs.

Moment bounds and variance proxies. Our certification arguments will be expressed through conservative second-moment bounds on $\delta\zeta_i$. For each group i and bitwidth b we assume an a priori bound

$$\mathbb{E}\big[\delta\zeta_i\big] = 0, \qquad \mathbb{E}\big\|\delta\zeta_i\big\|_2^2 \;\le\; v_i(b), \tag{4}$$

where $v_i(b)$ is a nonincreasing function of b that summarizes the quantization noise magnitude at that precision. The expectation in (4) is taken over the quantization randomness (e.g. stochastic rounding, dithering), and, when activations are quantized, may also include randomness induced by drawing an input x from the task distribution, since the quantized tensors depend on x.

In many practical schemes $v_i(b)$ admits a simple calibration model. For instance, for a uniform affine quantizer applied elementwise with step size $\Delta_i(b)$ and stochastic rounding, the per-coordinate error is approximately mean-zero with variance $\Delta_i(b)^2/12$, yielding the proxy

$$v_i(b) \;\approx\; d_i \cdot \frac{\Delta_i(b)^2}{12}. \tag{5}$$

More refined versions track per-channel scales and nonuniform ranges, and can be estimated empirically by repeatedly quantizing a representative snapshot of ζ_i (for weights) or of (a, k, v) collected on a calibration set (for activations/KV), then averaging $\|\delta\zeta_i\|_2^2$.

We also allow deterministic quantizers (round-to-nearest) by interpreting (4) as a worst-case-to-moment relaxation: if $\|\delta\zeta_i\|_2 \le r_i(b)$ deterministically, then (4) holds with $v_i(b) := r_i(b)^2$ and an arbitrary choice of mean-zero centering can be enforced by randomized tie-breaking or subtracting the empirical mean over a calibration set. This makes the analysis conservative but keeps the optimization problem and the resulting guarantees in a uniform form.

Weights versus activations versus KV cache. When quantizing weights, $\delta\theta$ is input-independent at deployment time, hence the only source of randomness in (4) is the quantization procedure itself (or an abstracted distribution over weight perturbations capturing the effect of a deterministic rounding rule). When quantizing activations, the perturbation is injected into the forward computation and depends on the realized activation values; a convenient modeling choice is to treat quantization as adding a random perturbation with conditional mean zero given the pre-quantized activation. For KV cache quantization, the same key/value vectors are reused across autoregressive steps; thus δk and δv can have structured temporal reuse. Our grouping permits this explicitly: we may take each KV group to correspond to a layer-head block aggregated over a fixed window of positions, and absorb the reuse into a larger effective d_i and an empirically calibrated $v_i(b)$. The certification bound will depend on $v_i(b)$ but not on how it arose, provided (4) holds for the chosen groups.

Independence and weak dependence. To obtain additive degradation bounds, we impose an (approximate) independence assumption across groups. The cleanest statement is the following.

Assumption (independent group perturbations). The random vectors $\{\delta\zeta_i\}_{i=1}^g$ are mutually independent.

Independence is exact if each group is quantized with independent dithering or independent stochastic rounding conditioned on the unquantized values. It is not exact when quantization shares scale parameters across groups, when activation quantization couples tensors through shared clipping statistics, or when KV-cache quantization introduces repeated use of a fixed perturbation across time. For these cases we note that the role of independence is to justify decompositions of aggregate second moments into sums. A standard relaxation is to allow weak correlations controlled by a correlation matrix. Namely, it suffices for many steps that cross-covariances are bounded as

$$\left|\mathbb{E}\langle \delta\zeta_i, \delta\zeta_j \rangle\right| \;\le\; \rho_{ij}\sqrt{v_i(b_i)v_j(b_j)} \qquad (i \neq j), \tag{6}$$

for some $\rho_{ij} \in [0, 1)$ estimated or upper bounded. Under (6), any subsequent bound that would involve $\sum_i \sqrt{v_i(b_i)}$ under independence typically acquires an additional inflation factor depending on ρ, or an additive term involving $\sum_{i \neq j} \rho_{ij}\sqrt{v_i(b_i)v_j(b_j)}$. For simplicity and because randomized quantization is available in many deployments, we present the main results under independence and treat weak dependence by conservative rescaling of $v_i(\cdot)$ when needed.

Budget coupling via costs. Finally, each bitwidth choice b_i incurs a deployment cost $c_i(b_i)$ (memory footprint for weights and KV, bandwidth/latency for activations), and the mixed-precision assignment must satisfy the budget constraint $\sum_i c_i(b_i) \le B$. The key point for what follows is that (4) supplies a per-group noise magnitude as a function of b_i, while the budget constraint couples the b_i across groups. This sets up the subsequent sensitivity analysis, where we bound the induced change in task margins in terms of the sensitivities of m_t to each group and the corresponding noise proxies $\sqrt{v_i(b_i)}$.

4 Sensitivity Measures and Certified Margin Degradation

We now relate the structured perturbation $\delta\zeta$ induced by mixed-precision quantization to changes in task margins. Fix a task $t \in \mathcal{T}$ and recall that $m_t(x; \zeta)$ denotes a (logit- or verifier-based) margin functional evaluated at input x. Our objective in this section is twofold: (i) to define groupwise sensitivity quantities that can be estimated from calibration data, and (ii) to express certified (i.e., provable under the moment bounds) degradation of the margin in terms of these sensitivities and the variance proxies $\{v_i(b_i)\}$.

Per-group sensitivities. Let $\zeta = (\zeta_1, \dots, \zeta_g)$ be grouped as in Section 3. Assuming $m_t(x; \cdot)$ is differentiable in a neighborhood of the unquantized state, we define the task–group sensitivity at x by the block gradient norm

$$\kappa_{t,i}(x) \ := \ \big\|\nabla_{\zeta_i} m_t(x;\zeta)\big\|_2.$$

Since we ultimately require distributional (or calibration-set) guarantees, we aggregate over $x \sim \mathcal{D}_t$ (or a finite calibration set $\{x_j\}_{j=1}^{n_t}$) by the second moment

$$s_{t,i} \ := \ \Big(\mathbb{E}_{x\sim\mathcal{D}_t}\, \kappa_{t,i}(x)^2\Big)^{1/2} \ \approx \ \Big(\frac{1}{n_t}\sum_{j=1}^{n_t}\big\|\nabla_{\zeta_i} m_t(x_j;\zeta)\big\|_2^2\Big)^{1/2}. \qquad (7)$$

The square-root-of-second-moment form is convenient because it interacts cleanly with Cauchy–Schwarz when combined with the moment bound (4). When only weights are quantized, $\zeta = \theta$ and the gradients in (7) are ordinary parameter gradients of the margin. When activations/KV-cache are included in ζ, the gradients are Jacobians with respect to the stored tensors; these can be obtained by backpropagating through the inference trace on the calibration inputs, treating the quantization injection points as additive perturbations as in (3).

Expected degradation: first-order bound with remainder. Write $\delta\zeta = (\delta\zeta_1, \ldots, \delta\zeta_g)$ for the quantization-induced perturbation. For fixed x, a Taylor expansion yields

$$m_t(x;\zeta+\delta\zeta) = m_t(x;\zeta) + \sum_{i=1}^{g}\big\langle \nabla_{\zeta_i} m_t(x;\zeta),\, \delta\zeta_i\big\rangle + r_t(x;\delta\zeta), \qquad (8)$$

where $r_t(x;\delta\zeta)$ is a second-order remainder. Taking expectation over the quantization randomness and using $\mathbb{E}[\delta\zeta_i] = 0$ from (4), the linear term vanishes in expectation at each x. However, to obtain a usable lower bound that depends only on the second moments, we bound the *magnitude* of the linear term and keep it as a conservative penalty (this is the step that remains valid even if mean-zero holds only approximately after calibration centering). Conditioning on x and applying Cauchy–Schwarz gives

$$\mathbb{E}\Big[\big|\langle \nabla_{\zeta_i} m_t(x;\zeta), \delta\zeta_i\rangle\big| \,\Big|\, x\Big] \ \le \ \big\|\nabla_{\zeta_i} m_t(x;\zeta)\big\|_2 \cdot \Big(\mathbb{E}\big\|\delta\zeta_i\big\|_2^2\Big)^{1/2} \ \le \ \kappa_{t,i}(x)\sqrt{v_i(b_i)}.$$

Averaging over x and using Jensen's inequality in the form $\mathbb{E}\,\kappa_{t,i}(x) \le (\mathbb{E}\,\kappa_{t,i}(x)^2)^{1/2} = s_{t,i}$, we obtain the first-order degradation penalty $\sum_i s_{t,i}\sqrt{v_i(b_i)}$.

It remains to control the remainder r_t. One convenient sufficient condition is blockwise Hessian smoothness: suppose that for each task t and almost every x the Hessian satisfies $\|\nabla^2_{\zeta} m_t(x;\zeta')\|_{\mathrm{op}} \le H_t$ for all ζ' in a neighborhood of ζ. Then the standard Taylor remainder bound implies

$$|r_t(x;\delta\zeta)| \ \le \ \frac{H_t}{2}\|\delta\zeta\|_2^2 \ \le \ \frac{H_t}{2}\sum_{i=1}^{g}\|\delta\zeta_i\|_2^2,$$

and therefore, using (4),

$$\mathbb{E}_{x,\delta\zeta}\big[m_t(x;\zeta+\delta\zeta)\big] \;\geq\; \mathbb{E}_x\big[m_t(x;\zeta)\big] - \sum_{i=1}^{g} s_{t,i}\sqrt{v_i(b_i)} - \frac{H_t}{2}\sum_{i=1}^{g} v_i(b_i). \quad (9)$$

We will denote the final term by $R_t(b) := \frac{H_t}{2}\sum_i v_i(b_i)$ (or any other valid upper bound derived from a task-specific smoothness estimate). In many regimes of interest the first-order term dominates because $v_i(b_i)$ is small at moderate precision and because the margin gradients concentrate on a small subset of groups.

High-probability bounds and success certification. Expected margins suffice for some objectives, but emergent capability preservation is often formulated as a lower bound on the probability that the post-quantization margin remains positive. A general-purpose route is to bound the variance of the random margin perturbation. Linearizing (8) and using independence across groups, we obtain the proxy

$$\mathrm{Var}\big(m_t(x;\zeta+\delta\zeta)\,\big|\,x\big) \;\lesssim\; \sum_{i=1}^{g}\big\|\nabla_{\zeta_i} m_t(x;\zeta)\big\|_2^2\, v_i(b_i), \quad (10)$$

where the $\lesssim$ hides higher-order contributions controlled by the same smoothness conditions used for $R_t(b)$. Averaging (10) over x yields an unconditional variance proxy

$$\sigma_t^2(b) \;:=\; \sum_{i=1}^{g} s_{t,i}^2\, v_i(b_i),$$

using the definition (7). Writing $\mu_t(b)$ for the certified mean lower bound from (9), Cantelli's inequality gives

$$\mathbb{P}\big[m_t(x;\zeta{+}\delta\zeta) \leq 0\big] \;\leq\; \frac{\sigma_t^2(b)}{\sigma_t^2(b) + \mu_t(b)^2}, \qquad \mu_t(b) := \mathbb{E}_x[m_t(x;\zeta)] - \sum_i s_{t,i}\sqrt{v_i(b_i)} - R_t(b). \quad (11)$$

Thus any allocation b that ensures $\mu_t(b)$ is appreciably positive yields a quantitative lower bound on $\mathbb{P}[m_t > 0]$, and hence (by the task construction) on semantic success probability. When a Gaussian approximation is empirically justified for the aggregated noise, we may replace (11) by the sharper surrogate

$$\mathbb{P}\big[m_t(x;\zeta+\delta\zeta) > 0\big] \;\approx\; \Phi\Big(\frac{\mu_t(b)}{\sigma_t(b)}\Big),$$

which directly exhibits the signal-to-noise ratio $\mu_t(b)/\sigma_t(b)$ as the controlling quantity.

Task aggregation. Since the optimization in the next section will allocate bits jointly across tasks, we also define an aggregated per-group score. For a specified emergent subset $\mathcal{T}_{\text{em}} \subseteq \mathcal{T}$ and weights $\{w_t\}$, we will use either

$$s_i := \max_{t \in \mathcal{T}_{\text{em}}} w_t s_{t,i}, \qquad \text{or} \qquad s_i := \sum_{t \in \mathcal{T}_{\text{em}}} w_t s_{t,i},$$

depending on whether we pursue worst-task guarantees or average-case objectives. The role of s_i is purely to convert the family of bounds (9) into a single benefit function per group that is compatible with a budgeted selection problem.

5 Budgeted Mixed-Precision Allocation

Having expressed certified margin degradation in terms of the sensitivities $\{s_{t,i}\}$ and the noise proxies $\{v_i(b)\}$, we now formalize the mixed-precision allocation problem. Throughout, we fix a finite admissible bitwidth set $\mathcal{B}$ (e.g. $\{2, 3, 4, 8, 16\}$), a grouping $G = \{1, \ldots, g\}$, and per-group costs $c_i(b)$ satisfying $c_i(b') \geq c_i(b)$ when $b' \geq b$. Our decision variable is the vector $b = (b_1, \ldots, b_g) \in \mathcal{B}^g$, and feasibility is the budget constraint

$$\sum_{i=1}^{g} c_i(b_i) \;\leq\; B. \tag{12}$$

Bound-induced objective. For each task $t \in \mathcal{T}_{\text{em}}$ we write the certified expected post-quantization margin lower bound in the form

$$\mu_t(b) \;:=\; \widehat{\mu}_t - \Delta_t(b), \qquad \Delta_t(b) \;:=\; \sum_{i=1}^{g} s_{t,i}\sqrt{v_i(b_i)} + R_t(b), \tag{13}$$

where $\widehat{\mu}_t$ denotes an empirical estimate of $\mathbb{E}_x[m_t(x;\zeta)]$ on calibration data, and $R_t(b)$ is any valid remainder upper bound (for instance, a smoothness-based term scaling with $\sum_i v_i(b_i)$). The most direct certified objective is to maximize the worst-task lower bound:

$$\max_{b \in \mathcal{B}^g} \; \min_{t \in \mathcal{T}_{\text{em}}} \; \mu_t(b) \quad \text{s.t.} \quad \sum_{i=1}^{g} c_i(b_i) \leq B. \tag{14}$$

When one prefers an average-case objective, we may maximize $\sum_t w_t \mu_t(b)$, or (using the variance proxy) a success surrogate such as $\sum_t w_t \Phi(\mu_t(b)/\sigma_t(b))$. In either case, the optimization is discrete and structured by groups.

A useful simplification is to choose an a priori low baseline precision $b_{\text{low}} \in \mathcal{B}$ and view all other bitwidths as *upgrades* from this baseline. Let $b^{\text{low}} := (b_{\text{low}}, \ldots, b_{\text{low}})$. Since $\sqrt{v_i(b)}$ is nonincreasing in b, allocating higher precision reduces the penalty term in (13). We therefore define a per-group *benefit* for assigning bitwidth b to group i by

$$\mathrm{ben}_i(b) \;:=\; s_i\Big(\sqrt{v_i(b_{\mathrm{low}})} - \sqrt{v_i(b)}\Big), \tag{15}$$

where s_i is an aggregate sensitivity score across tasks (e.g. $\max_t w_t s_{t,i}$ for worst-task control, or $\sum_t w_t s_{t,i}$ for weighted averaging). Up to task-specific constants (absorbed into $\widehat{\mu}_t$) and remainder handling, maximizing (14) can be approximated by maximizing the total benefit $\sum_i \mathrm{ben}_i(b_i)$ under the same budget.

Knapsack structure and multi-level discretization. In the special case $\mathcal{B} = \{b_{\mathrm{low}}, b_{\mathrm{high}}\}$, the decision reduces to selecting a subset $S \subseteq G$ of groups to upgrade. Writing $b_i = b_{\mathrm{high}}$ iff $i \in S$, we obtain the 0–1 knapsack form

$$\max_{S\subseteq G} \sum_{i\in S} \Delta_i \quad \text{s.t.} \quad \sum_{i\in S} c_i \le B, \qquad \Delta_i := s_i\Big(\sqrt{v_i(b_{\mathrm{low}})} - \sqrt{v_i(b_{\mathrm{high}})}\Big), \tag{16}$$

with item values Δ_i and costs $c_i := c_i(b_{\mathrm{high}}) - c_i(b_{\mathrm{low}})$. Thus, even under a linearized bound, selecting the best upgrade set is NP-hard (cf. Theorem 3), and exact optimization is intractable in general at the scales of interest.

For multiple precision levels, we obtain a *multi-choice* knapsack: each group i must pick one level $b \in \mathcal{B}$, yielding value $\mathrm{ben}_i(b)$ and cost $c_i(b)$. A standard reduction expands each group into $|\mathcal{B}|-1$ incremental upgrades with appropriate costs and marginal values; however, this introduces precedence constraints (one cannot take a higher upgrade without taking intermediate ones), so we will instead treat the multi-level case directly in the algorithmic section.

Submodularity in multi-task certification. When we optimize a bound that couples tasks, the effective objective may exhibit diminishing returns. To make this explicit, we consider the set view under two levels and define, for each task t, a task-specific retained-margin surrogate

$$F_t(S) \;:=\; \widehat{\mu}_t - \sum_{i\in G\setminus S} s_{t,i}\sqrt{v_i(b_{\mathrm{low}})} - \sum_{i\in S} s_{t,i}\sqrt{v_i(b_{\mathrm{high}})} - R_t(S), \tag{17}$$

where $R_t(S)$ denotes the remainder bound evaluated at the corresponding bit assignment. If we seek to maximize the number (or weighted fraction) of tasks whose certified margin exceeds a threshold η_t, we are led to objectives of the form

$$F(S) \;:=\; \sum_{t\in\mathcal{T}_{\mathrm{em}}} w_t\,\mathbf{1}\{F_t(S) \ge \eta_t\}\,, \tag{18}$$

or smooth relaxations thereof. Under natural overlap assumptions (the same upgraded groups contribute to multiple tasks, and additional upgrades yield diminishing marginal gains once a task is already comfortably above threshold), F is well-modeled as monotone submodular, rendering greedy selection provably near-optimal (cf. Theorem 4). We emphasize that submodularity is not assumed in the degradation bound itself; rather, it is a property of the *task-aggregated* certification objective one chooses to optimize.

Critical subspace. Finally, we formalize the notion of a *critical subspace* as the subset of groups assigned higher-than-baseline precision under an allocation b:

$$S(b) \;:=\; \{\, i \in G \;:\; b_i > b_{\text{low}} \,\}. \tag{19}$$

Operationally, $S(b)$ identifies the small collection of parameter blocks (or activation/KV blocks) that dominate certified margin degradation through large sensitivities $s_{t,i}$ and/or unfavorable quantization noise $v_i(b)$. The aim of the optimization is therefore twofold: to satisfy the deployment budget (12) while (i) keeping $S(b)$ small and interpretable, and (ii) guaranteeing that for every $t \in \mathcal{T}_{\text{em}}$ the certified post-quantization margin $\mu_t(b)$ remains above the task emergence threshold η_t (or yields a quantitative success lower bound via the probabilistic inequalities of the previous section). In the next section we give approximation algorithms that compute such an allocation efficiently at scale.

6 Theoretical Results

We now record the guarantees that justify the mixed-precision allocation procedure implied by the bound and the subsequent approximation algorithms. Conceptually, the analysis splits into three components: (i) a *certification* inequality translating per-group perturbation magnitudes into a margin lower bound (Theorems 1–2), (ii) an *optimization* layer describing what can and cannot be computed under a budget (Theorems 3–5), and (iii) a *statistical* layer quantifying how much information is required to estimate the sensitivities that drive the certificate (Theorem 6).

Certified correctness from sensitivity-weighted noise. Fix a task $t \in \mathcal{T}_{\text{em}}$ and a feasible bit assignment $b \in \mathcal{B}^g$. Under the independent, zero-mean group noise model, Theorem 1 yields an expected margin guarantee of the form

$$\mathbb{E}_{x,\delta\theta}[m_t(x; \theta + \delta\theta)] \;\geq\; \mathbb{E}_x[m_t(x; \theta)] \;-\; \sum_{i=1}^{g} s_{t,i}\sqrt{v_i(b_i)} \;-\; R_t(b). \tag{20}$$

Thus any allocation algorithm that outputs b immediately induces a task-wise certified lower bound $\mu_t(b) = \widehat{\mu}_t - \Delta_t(b)$ as in (13). Theorem 2 then formalizes the implication for emergence preservation: whenever $\mu_t(b) \geq \eta_t$, the quantized model lies in the above-baseline regime for task t (either in expectation, or with an explicit success lower bound after applying a probabilistic inequality to the margin random variable).

A point worth making explicit is that the certification statement is *algorithm-agnostic*: it does not require the allocation to be optimal, only feasible, and it only depends on $(s_{t,i}, v_i(\cdot), R_t)$. In particular, if an approximation algorithm returns an allocation b that is merely near-optimal with respect to a surrogate objective, the certification (20) remains valid for that returned b.

Robustness to sensitivity estimation error. In practice, $s_{t,i}$ is estimated from a finite calibration set via gradient-based statistics. Let $\widehat{s}_{t,i}$ be an estimator satisfying a uniform deviation bound

$$\Pr\big[\forall t, i: \ \big|\widehat{s}_{t,i} - s_{t,i}\big| \leq \alpha\big] \ \geq \ 1 - \delta. \tag{21}$$

Conditioning on the event in (21), we obtain a conservative certified degradation bound by replacing $s_{t,i}$ with $\widehat{s}_{t,i} + \alpha$:

$$\Delta_t(b) \ \leq \ \sum_{i=1}^{g} (\widehat{s}_{t,i} + \alpha)\sqrt{v_i(b_i)} + R_t(b). \tag{22}$$

Hence, with probability at least $1-\delta$, any computed allocation b that satisfies $\widehat{\mu}_t - \sum_i (\widehat{s}_{t,i} + \alpha)\sqrt{v_i(b_i)} - R_t(b) \geq \eta_t$ also satisfies the corresponding claim with the true sensitivities. This is the standard "plug-in certificate with slack" principle: estimation error affects only the *tightness* of the certificate, not its logical validity.

Approximation guarantees under submodular task aggregation. When the chosen optimization objective over upgrade sets S is a monotone submodular function $F(S)$ under a knapsack constraint (as in Theorem 4), we may apply the classical greedy paradigm for budgeted submodular maximization. Concretely, writing $c(S) = \sum_{i \in S} c_i$ for incremental costs and assuming $F(\varnothing) = 0$, the density-greedy rule that iteratively adds the group with maximal marginal gain-per-cost yields a constant-factor approximation; with standard knapsack-feasibility modifications (e.g. partial enumeration of a constant number of high-value items combined with greedy completion), one obtains the familiar

$$F(S_{\text{greedy}}) \ \geq \ \Big(1 - \tfrac{1}{e}\Big)\, F(S^{\star}) \tag{23}$$

up to the customary constant-factor adjustments specific to knapsack constraints (Theorem 4). The relevance here is that many multi-task certification objectives saturate once tasks clear their thresholds, creating the diminishing-returns structure that submodularity captures. Under such objectives, greedy selection is not merely a heuristic but a provably near-optimal mechanism for identifying a small critical subset of groups.

Pseudo-polynomial optimality and an FPTAS in the linearized two-level case. In the two-precision specialization (16), the allocation reduces exactly to 0–1 knapsack with values Δ_i and costs c_i. Consequently, dynamic programming solves the problem optimally in $O(gB)$ time (and $O(B)$ space with standard rolling-array optimization) when B is moderate and integral. When B is large, Theorem 5 supplies an FPTAS: after scaling values to a bounded range, DP on scaled values returns S such that

$$\sum_{i \in S} \Delta_i \;\geq\; (1-\varepsilon) \sum_{i \in S^\star} \Delta_i, \tag{24}$$

with polynomial runtime in g and $1/\varepsilon$. This provides an explicit sense in which the "critical subspace" can be computed near-optimally under a simplified (but often informative) linearized bound model.

Hardness barriers. Theorem 3 shows that even the two-level linearized allocation problem is NP-hard by direct reduction from 0–1 knapsack. This hardness is not an artifact of our certificate; rather, it reflects the intrinsic combinatorics of distributing a limited precision budget across heterogeneous groups. Moreover, when the multi-task objective is posed as maximizing the number of tasks whose certified margin exceeds thresholds (cf. (18)), the resulting problem subsumes budgeted maximum coverage in natural constructions, suggesting that exact optimization is intractable even when the per-group benefits are easily computed. Accordingly, our algorithmic posture—greedy under submodularity, and DP/FPTAS in a restricted linearized regime—is essentially the strongest one can expect in polynomial time without additional structural assumptions.

Information-theoretic lower bounds for identifying the critical subspace. Even if optimization were free, the selection requires sufficiently accurate sensitivity estimates. Theorem 6 gives a worst-case lower bound: to estimate a given $s_{t,i}$ to additive error α with confidence $1-\delta$, one needs $\Omega(\sigma^2 \log(1/\delta)/\alpha^2)$ samples, where σ^2 controls the variance of the underlying gradient-norm statistic. This directly implies a lower bound on the information required to *choose* $S(b)$ reliably: if two groups i and j have nearly equal benefit-to-cost ratios under the bound, then distinguishing which one should be upgraded (to achieve a target certificate) requires resolving differences on the order of that gap, forcing α to be comparably small and the calibration burden correspondingly large. In particular, in near-threshold regimes—precisely those relevant for emergence preservation—the sample complexity necessarily increases, regardless of the subsequent knapsack solver.

Finally, these lower bounds clarify the role of practical estimation tricks used at scale (mini-batching, gradient subsampling, or Hutchinson-type estimators): such techniques primarily trade compute for variance. They can reduce wall-clock time, but they cannot circumvent the α^{-2} dependence in the information requirement, and thus they cannot eliminate the need for sufficiently rich calibration evidence when the critical subspace is only weakly identifiable.

7 7. Theoretical Results: correctness and approximation guarantees; NP-hardness; lower bounds on information required to select the critical subspace.

To make the optimization layer completely explicit, we view a mixed-precision policy as a vector $b = (b_1, \ldots, b_g) \in \mathcal{B}^g$, or equivalently as a set of *upgrade decisions* relative to a fixed baseline precision b_{low}. Writing $\Delta_t(b)$ for the certificate-implied degradation term in (20), the most direct certified objective is

$$\max_{b\in\mathcal{B}^g} \min_{t\in\mathcal{T}_{\mathrm{em}}} \Big(\widehat{\mu}_t - \Delta_t(b)\Big) \qquad \text{s.t.} \qquad \sum_{i=1}^{g} c_i(b_i) \le B, \tag{25}$$

or the thresholded feasibility variant $\widehat{\mu}_t - \Delta_t(b) \ge \eta_t$ for all $t \in \mathcal{T}_{\mathrm{em}}$. Even under the first-order bound (dropping R_t), (25) is a combinatorial budgeted allocation, because each coordinate b_i is discrete and the costs need not be uniform.

A standard simplification is to introduce per-group *benefits* as reductions in the bound relative to b_{low}. Fixing an aggregation of sensitivities s_i (e.g. $s_i = \max_{t\in\mathcal{T}_{\mathrm{em}}} w_t s_{t,i}$), define the per-group retained-margin contribution at precision b by

$$\mathrm{gain}_i(b) \;:=\; s_i\Big(\sqrt{v_i(b_{\mathrm{low}})} - \sqrt{v_i(b)}\Big), \qquad \mathrm{gain}_i(b_{\mathrm{low}}) = 0, \tag{26}$$

and let $\mathrm{cost}_i(b) := c_i(b) - c_i(b_{\mathrm{low}})$. Maximizing $\sum_i \mathrm{gain}_i(b_i)$ subject to $\sum_i \mathrm{cost}_i(b_i) \le B' := B - \sum_i c_i(b_{\mathrm{low}})$ yields a *multiple-choice knapsack* instance (each group chooses exactly one level). Theorem 3 corresponds to the two-choice restriction, in which each group either stays at b_{low} or upgrades to b_{high}, with value $\Delta_i = \mathrm{gain}_i(b_{\mathrm{high}})$ and weight $\mathrm{cost}_i(b_{\mathrm{high}})$; the NP-hardness then follows by a direct encoding of items as groups. The same reduction also shows hardness for multi-level precision: by creating dummy bitwidth options whose $(\mathrm{gain}, \mathrm{cost})$ pairs replicate an arbitrary set of items, one obtains NP-hardness of the general discrete allocation even when $v_i(\cdot)$ is monotone and $c_i(\cdot)$ is increasing.

Given this hardness barrier, the correctness statement we rely on is deliberately modular: *any* feasible output b implies (20), hence any post-processing step (greedy, DP, or heuristic) is automatically certified once its output is plugged into $\Delta_t(b)$. Thus the only algorithm-dependent claims concern proximity of the chosen allocation to the optimum of a chosen surrogate objective (and, separately, the statistical reliability of estimated sensitivities).

For approximation, two structural regimes are relevant. First, in the linearized two-level case, the objective $\sum_{i\in S}\Delta_i$ under a budget is exactly 0–1 knapsack, admitting (i) pseudo-polynomial dynamic programming and (ii) the FPTAS of Theorem 5. In this regime, the approximation guarantee is particularly interpretable: if S_ε is the FPTAS solution, then the resulting certified bound under the linearized model is within a factor $(1-\varepsilon)$ of the best achievable improvement over baseline, and consequently the certified retained margin $\widehat{\mu}_t - \Delta_t(b)$ is correspondingly close to the optimum among two-level allocations.

Second, in multi-task settings it is often more faithful to optimize a *saturated* objective that reflects emergence thresholds. For instance, one may define

$$F(S) := \sum_{t\in\mathcal{T}_{\rm em}} w_t \min\Big\{\big(\widehat{\mu}_t - \eta_t\big) - \Delta_t(S),\, M_t\Big\}_+, \tag{27}$$

where $\Delta_t(S)$ denotes the degradation bound induced by upgrading the set S (with all other groups at $b_{\rm low}$), and M_t caps marginal value once a task is safely above threshold. Under the common case that $\Delta_t(S)$ decomposes as a sum of per-group contributions and the only nonlinearity is the outer truncation in (27), F is monotone and submodular: each additional upgrade yields diminishing returns because tasks that already exceed their capped slack stop benefiting. In that case, Theorem 4 applies and a density-greedy procedure gives a constant-factor approximation to the best upgrade set under the knapsack constraint. Importantly, the certificate is compatible with this saturation: when the greedy algorithm increases $F(S)$, it is directly increasing a conservative proxy for the number (or weighted measure) of tasks that remain above threshold under the bound.

We finally connect the statistical layer (Theorem 6) to the *identifiability* of the critical subspace. Suppose we are in a two-level regime for clarity. Let $\rho_i := \Delta_i/c_i$ denote the (unknown) value-to-cost density. Any algorithm that attempts to select an (approximately) optimal set must, implicitly, separate groups with near-tied densities, because swapping one such group for another can change feasibility and objective value while leaving the certificate near the threshold. More formally, if there exist two groups $i \neq j$ with $|\rho_i - \rho_j| \leq \gamma$, then any procedure that outputs, with probability at least $1-\delta$, a set whose total value is within $o(\gamma)$ of the optimum must estimate ρ_i and ρ_j to accuracy $o(\gamma)$. Since Δ_i depends linearly on s_i through (26), Theorem 6 implies a calibration lower bound of order $\Omega(\log(1/\delta)/\gamma^2)$ samples (up to problem-dependent variance factors) to resolve the ordering. This is the same "gap" phenomenon familiar from best-arm identification: the closer the instance is to having multiple nearly optimal critical subspaces, the larger the unavoidable sample requirement for reliably selecting one.

Consequently, the difficulty of selecting S is not purely computational. Even if one could solve the knapsack problem exactly, near-threshold deployment (where $\widehat{\mu}_t - \eta_t$ is small for some t) forces α in (21) to be small enough that the plug-in slack in (22) does not dominate the margin, and this in turn forces calibration sets large enough to overcome the α^{-2} information barrier. This clarifies why, in practice, we treat the critical subspace as an empirically stable object only when it is separated by a nontrivial "benefit gap" from competing allocations, and why the certificate is best interpreted as a conservative sufficient condition rather than a tight characterization in regimes of weak identifiability.

8 Experimental Protocol

We describe an experimental protocol intended to (i) instantiate the calibration objects appearing in the certificate (margins, sensitivities, and noise moments), (ii) compare the resulting certified degradation to observed post-quantization regressions, and (iii) evaluate whether mixed-precision policies that preserve emergent capabilities also preserve safety-relevant behaviors. Throughout, we treat the certificate as a sufficient condition and report both certified and empirical outcomes.

Models and grouping. We consider a pretrained transformer f_θ and form groups $G = \{1, \dots, g\}$ by layerwise blocks, with a default partition into (a) embeddings, (b) attention projection matrices (optionally split into Q, K, V, O), (c) MLP matrices, (d) layer norms and biases, and (e) output head. When activation or KV-cache quantization is studied, we introduce additional groups for per-layer activations and per-layer KV tensors. Costs $c_i(b)$ are taken as realized deployment memory (for weight-only) or peak memory/latency proxies (for activation/KV), measured using the target runtime. We impose hardware constraints when applicable (e.g. embeddings $\geq$ 8-bit).

Task suite. We instantiate $\mathcal{T}_{\text{em}}$ using a mixture of emergent reasoning and arithmetic tasks, and we include a broad general-knowledge slice to test distributional robustness. Concretely, we recommend: (i) multi-step arithmetic (GSM-style word problems) and synthetic long addition/multiplication with controlled length; (ii) symbolic and logical reasoning (e.g. compositional deduction, chain-of-thought-free variants with verifier scoring); (iii) selected MMLU slices emphasizing reasoning (mathematics, formal logic, abstract algebra) alongside non-reasoning controls (history, sociology) to detect uneven regressions. For each task t, we define a margin functional $m_t(x; \theta)$ that is compatible with the evaluation: for multiple-choice tasks, we use the logit gap between the correct option and the best incorrect option; for free-form tasks with a verifier, we use a verifier-score gap between a correct and an incorrect completion or a calibrated score minus a threshold. The threshold η_t is set by a baseline criterion (random-guess or majority-class) plus a fixed slack, or by calibrating the margin-to-success link on the unquantized model.

Calibration sets and data hygiene. For each t we form a calibration set $\{x_j\}_{j=1}^{n_t}$ disjoint from the test set. We recommend n_t in the range 256–2048, with larger n_t for tasks known to be near emergence thresholds. All sensitivity and noise-moment estimates are computed exclusively on calibration data; evaluation uses held-out test splits. We repeat the entire pipeline over several random seeds for subsampling and quantization stochasticity (when present).

Estimating sensitivities. We estimate $s_{t,i} = \sqrt{\mathbb{E}_x \|\nabla_{\theta_i} m_t(x;\theta)\|_2^2}$ by Monte Carlo averaging over the calibration set. To reduce cost, we optionally use (a) microbatching across tasks, (b) per-layer gradient checkpointing, and (c) randomized estimators (e.g. Hutchinson-style sketches) for large groups. We report both point estimates and confidence intervals obtained by empirical Bernstein or bootstrap, which are then propagated into a conservative sensitivity $\overline{s}_{t,i}$ used by the certificate. When aggregating across tasks we use $s_i = \max_{t\in\mathcal{T}_{\mathrm{em}}} w_t \overline{s}_{t,i}$ with weights w_t either uniform or chosen to emphasize tasks near threshold.

Calibrating quantization noise moments. For each group i and bitwidth $b \in \mathcal{B}$, we estimate $v_i(b)$ by directly quantizing the group in isolation (keeping the rest at a high-precision reference), computing $\delta\theta_i$, and measuring $\|\delta\theta_i\|_2^2$ under the chosen quantizer (uniform, per-channel, or learned scale). We take $v_i(b)$ as an upper confidence bound (e.g. mean plus two standard errors) over several calibration batches and, if applicable, over several quantization parameter initializations. This step makes the independence and zero-mean assumptions operational: we enforce mean-zero by centering stochastic rounding noise or by subtracting the empirical mean perturbation when a deterministic quantizer is used.

Policies and baselines. We evaluate: (i) uniform b-bit baselines for $b \in \{2,4,8\}$ (and 16-bit reference), (ii) layerwise heuristics (attention at higher precision than MLP, embeddings high), and (iii) our certificate-driven mixed-precision policy obtained by optimizing the surrogate objective induced by $\mathrm{gain}_i(b)$ under the budget. In the two-level setting we solve the induced knapsack either exactly (pseudo-polynomial DP when feasible) or via an FPTAS; for multi-level precision we use a multiple-choice knapsack solver or a greedy density heuristic, and we always record the achieved budget and the resulting critical set $S = \{i : b_i > b_{\text{low}}\}$. Budgets are reported as average bits per parameter and as realized memory/latency.

Evaluating certificate tightness. For each task t we compute: (a) the certified degradation $\Delta_t(b)$ and the certified retained margin $\widehat{\mu}_t - \Delta_t(b)$; (b) the observed margin drop on test data, $\Delta_t^{\text{obs}} := \widehat{\mu}_t - \widehat{\mu}_t^Q$, where $\widehat{\mu}_t^Q$ is the empirical mean margin under the quantized model; and (c) the observed success drop. We summarize tightness by the ratio $\Delta_t^{\text{obs}}/\Delta_t(b)$ and by the fraction of instances where the sign of the margin is preserved. We also measure correlation between per-group predicted importance (e.g. $\rho_i = \text{gain}_i(b_{\text{high}})/\text{cost}_i(b_{\text{high}})$) and empirical ablations that upgrade a single group at a time from b_{low}.

Safety-behavior retention. We evaluate safety-relevant behaviors under the same quantized policies, treating them as additional tasks with margins defined by a safety verifier. We recommend two classes of tests: (i) harmlessness and refusal compliance under adversarial prompts (jailbreak-style and benign but sensitive topics), and (ii) toxicity and bias metrics on standard prompt sets. For each, we define a margin as the difference between a refusal/compliance score and a threshold, or the negative of a toxicity score relative to an acceptable bound. We report whether mixed precision increases unsafe completion rates, and whether the certificate (computed on a safety calibration set) predicts when safety margins are at risk. To avoid conflating calibration with evaluation, we maintain disjoint safety calibration and safety test sets, and we report worst-case degradation over the safety tasks under the same budget that targets $\mathcal{T}_{\text{em}}$.

Reporting. We report: (i) the chosen bitwidth vector b and the size and composition of S; (ii) per-task accuracy and margin statistics pre/post quantization; (iii) certified versus observed degradation; and (iv) safety metrics. All results are stratified by budget and by whether activation/KV quantization is enabled, and we include ablations that vary group granularity to assess how sensitive the learned critical subspace is to the partitioning choice.

9 Discussion: Assumption Failures, Extensions, and Deploy-Time Policy

Our certificate is deliberately modular: it isolates (i) a perturbation model for quantization, (ii) a sensitivity object derived from gradients of margins, and (iii) a budgeted optimization of bitwidths. This modularity makes clear where the argument may become loose or invalid, and it suggests several extensions that preserve the overall structure while altering the technical inputs.

Correlated or biased quantization noise. Theorem 1 uses two structural assumptions that may fail in practice: independence across groups and mean-zero perturbations. Independence is often violated when a shared quantization scale couples multiple tensors, when per-layer clipping is tuned jointly, or when activation quantization introduces correlated errors along the computation graph. In this case, the first-order Taylor term no longer decouples by groups in expectation, and one obtains cross-covariance contributions. Concretely, letting $g_{t,i}(x) := \nabla_{\theta_i} m_t(x;\theta)$, the leading term becomes

$$\mathbb{E}\langle \nabla_\theta m_t(x;\theta), \delta\theta \rangle = \sum_i \mathbb{E}\langle g_{t,i}(x), \delta\theta_i \rangle + \sum_{i \neq j} \mathbb{E}\langle g_{t,i}(x), \delta\theta_j \rangle,$$

and a conservative bound can be written in terms of a covariance operator for $\delta\theta$. One simple repair is to aggregate groups into larger blocks so that the remaining inter-block correlations are reduced, at the cost of a coarser allocation. A more quantitative repair is to replace the diagonal moment bounds $v_i(b_i)$ by a block covariance bound: for a partition into blocks $P_1, \ldots, P_k$, assume $\mathbb{E}\|\delta\theta_{P_\ell}\|_2^2 \leq v_{P_\ell}(b_{P_\ell})$ and proceed as before. This preserves the knapsack structure but may reduce granularity.

Mean-zero can also fail for deterministic round-to-nearest quantizers, especially when scales are estimated from a finite calibration sample and then held fixed. In that setting, the perturbation decomposes as $\delta\theta_i = \mu_i + \xi_i$ with $\mathbb{E}[\xi_i] = 0$ and $\mu_i \neq 0$. The certificate then acquires an additional deterministic bias term $\sum_i \|g_{t,i}\|_2 \|\mu_i\|_2$, which can dominate at low bitwidth. Operationally, we can estimate μ_i empirically and either (a) subtract it by recentering (when stochastic rounding is available), or (b) include it explicitly as a separate penalty in $\Delta_t(b)$. The latter typically changes the optimization only through modified per-group "values" and thus remains compatible with our selection algorithms.

Second-order effects and near-threshold tasks. Even when first-order assumptions hold, the remainder R_t can be non-negligible in regimes where quantization is aggressive (e.g. $b = 2$) or where m_t is highly curved in relevant directions. This is most salient for tasks near emergence thresholds η_t, where the relevant slack is small. A practical implication is that the certificate should be treated as a sufficient condition with an explicit "margin buffer": we should demand $\widehat{\mu}_t - \Delta_t(b) \geq \eta_t + \tau$ for a tunable $\tau > 0$ chosen to absorb unmodeled curvature. Methodologically, one can tighten R_t by estimating a local smoothness constant along the quantization directions (e.g. by finite differences on $\delta\theta$ restricted to candidate groups), but we emphasize that such tightening is model- and quantizer-dependent.

Post-quantization fine-tuning and the meaning of a certificate. If we fine-tune after quantization, we are no longer analyzing $f_{\theta+\delta\theta}$ but rather

$f_{\theta+\delta\theta+\Delta\theta_{\mathrm{ft}}}$, where $\Delta\theta_{\mathrm{ft}}$ depends on optimization dynamics and data. Two interpretations are then possible. The first is conservative: we certify the initial quantized model and treat fine-tuning as an empirical improvement step that may (but need not) recover margins beyond the certified lower bound. The second is analytic: we incorporate fine-tuning into the perturbation model by bounding $\|\Delta\theta_{\mathrm{ft},i}\|_2$ as a function of learning rate, gradient norms, and step count, and then add $\sum_i s_{t,i}\|\Delta\theta_{\mathrm{ft},i}\|_2$ to $\Delta_t(b)$. This yields a joint certificate for "quantize-then-train" pipelines, but it is only meaningful if we can upper bound the optimization trajectory in a way that is not vacuous. In practice, we view fine-tuning as a mechanism for *reallocating* error: it can reduce effective sensitivity in some groups while increasing it in others, suggesting an alternating scheme (estimate $s_{t,i}$, choose b, fine-tune briefly, re-estimate) with the understanding that each iteration certifies only its current iterate.

LoRA-style recovery as a certified extension. A more structured recovery mechanism is to add a small set of trainable high-precision parameters (e.g. LoRA adapters) while quantizing the base weights aggressively. In our framework this corresponds to augmenting the parameter vector by additional groups θ_k^{A} with bitwidth fixed at 16 (or 8) and cost accounted for in the budget. The certificate then applies to the *composite* parameterization, and the optimizer may rationally spend a small fraction of B on adapters if they yield large effective margin gains. The technical point is that adapters can change both $\widehat{\mu}_t$ (baseline margins of the adapted model) and the sensitivities $s_{t,i}$ with respect to quantized groups, potentially shrinking the needed critical set S. This suggests a deploy-time design principle: when the budget is extremely tight, it may be preferable to maintain a tiny high-precision subspace that is explicitly trained to be robust to quantization, rather than attempting to protect many original groups by raising their bitwidth.

Activation and KV-cache quantization. Weight-only quantization treats $\delta\theta$ as the sole perturbation. Activation and KV quantization introduce *state* perturbations that depend on inputs and on intermediate representations. One can still reuse the margin-sensitivity template by defining groups over activation/KV tensors and replacing $\nabla_{\theta_i} m_t$ with the appropriate Jacobian of m_t with respect to the quantized state. Formally, for an activation group a_i with perturbation δa_i satisfying $\mathbb{E}\|\delta a_i\|_2^2 \le v_i(b_i)$, we obtain a bound of the same shape with $s_{t,i} := \sqrt{\mathbb{E}\|\nabla_{a_i} m_t\|_2^2}$. The principal complication is that independence and stationarity are less plausible for activations, so one should expect to fall back to blockwise grouping or empirically calibrated worst-case variance bounds.

Implications for deploy-time mixed-precision policy. The preceding points imply that mixed precision should be treated as a *policy* rather than a one-shot decision. In deployment, we recommend enforcing three guardrails derived from the certificate: (i) maintain explicit slack in certified margins on emergent and safety tasks, rather than operating exactly at threshold; (ii) prefer allocations that are stable under plausible model drift (e.g. minor prompt distribution shift), which can be operationalized by using $s_i = \max_{t \in \mathcal{T}} s_{t,i}$ over a broader task set than the one optimized for; and (iii) reserve a small "precision contingency" (unused budget) that can be spent to raise bitwidth for a small number of groups if monitoring indicates margin erosion. In this sense, the critical subspace S is not merely descriptive; it is an actionable handle for safe adaptation when assumptions or operating conditions change.

Limitations and open problems: tightness, mechanisms, and adaptation. Our guarantees are only as useful as they are tight. The degradation terms in Theorem 1 (and hence the slack conditions in Theorem 2) are driven by three upper bounds: (i) the second-moment proxy $v_i(b)$ for quantization error, (ii) the Cauchy–Schwarz step that produces $\sum_i s_{t,i}\sqrt{v_i(b_i)}$, and (iii) the treatment of curvature through the remainder R_t. Each of these steps can introduce orders-of-magnitude looseness, particularly when the effective quantization noise is highly anisotropic within a group. In practice, quantization error often concentrates on a low-dimensional subspace (e.g. a few principal directions induced by scale/clipping), whereas our bound treats it as if it could align adversarially with $\nabla_{\theta_i} m_t$. A natural technical goal is therefore to replace the scalar variance proxy by a covariance-sensitive quantity. One concrete direction is to model $\mathbb{E}[\delta\theta_i \delta\theta_i^\top] \preceq \Sigma_i(b_i)$ and to bound the first-order term by

$$\mathbb{E}\langle \nabla_{\theta_i} m_t, \delta\theta_i \rangle \geq -\sqrt{\mathbb{E}\big[\nabla_{\theta_i} m_t^\top \Sigma_i(b_i) \nabla_{\theta_i} m_t\big]},$$

which interpolates between isotropic and highly structured noise. The open difficulty is that estimating $\Sigma_i(b)$ reliably is expensive and quantizer-dependent, yet without it we cannot expect certificates to predict the empirically observed "sharp transition" behavior of extreme low-bit quantization.

A related limitation is that our use of groupwise sensitivities implicitly assumes that the natural grouping reflects the geometry of the margin. When groups are chosen by layer or tensor type, $s_{t,i}$ conflates directions that have very different functional roles. This can lead to overly conservative allocations in which we spend budget to protect parameters that are large in norm but functionally redundant for the tasks of interest. Conversely, if we choose groups that are too fine (e.g. per-channel or per-head), the optimization improves but sensitivity estimation becomes sample-inefficient, and the noise model becomes harder to justify. This exposes a basic open problem:

determine, from limited calibration data, a grouping that is simultaneously (a) stable under quantization backends, (b) aligned with the local curvature of margins, and (c) tractable for knapsack-style selection. Even partial progress—for instance, grouping by low-rank directions discovered via Hessian sketches or Fisher information—would directly strengthen the certificate without changing its overall logic.

A second set of open questions concerns mechanistic alignment: does the critical set S correspond to identifiable computational circuits? Our method outputs a subset of groups whose precision is predicted to matter for certified margins, but it does not explain *why* those groups matter, nor whether they constitute a coherent mechanism across tasks. We view this as an opportunity rather than a defect. If emergent capabilities are mediated by sparse circuits (e.g. specific attention heads, MLP neurons, or compositionally interacting modules), then a tight mixed-precision policy ought to recover those circuits as the "high-precision" subspace, and should do so robustly across prompt paraphrases and task variants. Formally, this suggests studying whether the sensitivity map $i \mapsto s_{t,i}$ concentrates on groups that coincide with independently discovered mechanistic features (via activation patching, path attribution, or linear probes). An affirmative result would provide evidence that our bound is not merely a worst-case inequality but is tracking real causal pathways. A negative result would indicate either that the circuit picture is incomplete at the scale relevant for quantization, or that our sensitivity surrogate is missing key second-order interactions (e.g. products of perturbations across layers). Making this precise appears to require a joint theory of (a) margin geometry and (b) circuit identifiability under weight perturbations, which is currently unavailable.

Third, distribution shift creates a fundamental tension for deploy-time quantization. Our optimization uses calibration distributions $\{\mathcal{D}_t\}$ both to estimate baseline margins and to define $s_{t,i}$; under shift, the relevant gradients and margins can change, so the chosen S may cease to be critical in the intended sense. This motivates adaptive quantization policies: instead of a single $b \in \mathcal{B}^g$, we may wish to output a mapping π from a lightweight prompt statistic (or a task classifier) to a bitwidth assignment $\pi(z)$. The challenge is that the budget constraint is now either an average constraint (expected cost under the prompt stream) or a worst-case constraint (cost must never exceed B), and the certificate must account for selection bias: if we raise precision on "hard" prompts, we are implicitly conditioning on prompts with smaller margins, where the Taylor approximation is least stable. A principled approach would couple a shift-robust sensitivity definition, e.g. $s_i = \sup_{t \in \mathcal{T}} \sqrt{\mathbb{E}_{x \sim \mathcal{D}_t} \|\nabla_{\theta_i} m_t(x; \theta)\|_2^2}$, with an online monitor that estimates margin proxies and triggers contingency precision increases. Providing a non-vacuous guarantee for such a closed-loop policy remains open: one needs to control both estimation error (how well we detect margin erosion) and actuation error (how quickly increasing b_i restores margins).

Finally, even in the static setting, we do not yet have a characterization of when the knapsack relaxation matches the true “minimal precision” structure. Theorems 3–5 explain computational hardness and approximation, but they do not address statistical or geometric conditions under which the optimal solution is *stable* (small changes in calibration data do not change S) or *sparse* (a small fraction of groups dominate). Establishing such conditions would matter operationally: stable sparsity is what would make the notion of a critical subspace actionable. One plausible hypothesis is a separation condition on sensitivities, e.g. a gap between a small set of high-s_i groups and the rest, together with diminishing returns in $v_i(b)$; under such a condition, greedy selection should be near-optimal and insensitive to estimation noise. Proving this would require combining approximation analysis with concentration bounds for $\widehat{s}_{t,i}$, thereby linking Theorem 6-style sample complexity to stability of the selected set.

In summary, the certificate provides a coherent scaffold, but three research directions appear decisive for turning it into a sharp and mechanistically informative tool: covariance-aware noise models to improve tightness, circuit-aligned groupings to improve interpretability, and adaptive policies to remain valid under distribution shift.