TASK-AWARE ANSWER PRESERVATION UNDER AUDIO COMPRESSION FOR LARGE AUDIO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large audio language models (LALMs) increasingly reason over long audio clips, motivating audio compression to reduce memory use and inference latency. However, audio compression can leave the model's overall answer accuracy acceptable while severely degrading accuracy for particular query families. We introduce a framework for auditing a given audio compression method by measuring the increase in an LALM's answer error relative to uncompressed audio. We formalize an acceptance criterion that limits this increase for every required query family and derive a practical protocol for testing compression settings with statistical confidence. Experiments spanning five multiple-choice audio question-answering benchmarks and four LALMs expose compression-induced damage hidden by overall accuracy.

1 INTRODUCTION

Large audio language models (LALMs) can reason and answer questions about speech, music, and audio events (Chu et al., 2024; Xu et al., 2025). In deployment, long clips, interactive use, and multimodal contexts motivate compressing audio before LALM inference to reduce memory use and response latency (Arora et al., 2025). However, compression can affect answer accuracy differently across query families: questions about lexical content, event timing, or prosody and speaker state depend on different information in the recording. A small overall accuracy drop can therefore conceal severe degradation for a particular query family. An audio compression method should therefore be judged against the answer-accuracy requirements of the tasks the LALM must perform in deployment.

We introduce a task-aware answer-preservation framework for auditing a supplied audio compressor, such as EnCodec (Défossez et al., 2023), with a fixed LALM. The contribution is a criterion and practical protocol for accepting or rejecting the compressor's tested settings, e.g., a target bitrate of 1.5 kbps. The practitioner specifies the query families to evaluate, such as questions about speaker emotion, and the allowed increase in answer error relative to uncompressed audio. The criterion requires that this tolerance be met for every evaluated family. The protocol selects compression settings on development data and tests the selected settings on independent held-out data. It distinguishes three outcomes: evidence that every evaluated family meets the tolerance, evidence that a family exceeds the tolerance, or insufficient evidence for either conclusion. When a setting cannot be approved, the audit identifies the families with excessive degradation or insufficient evidence of answer preservation. The audit can thus change which compression settings are approved for deployment and identify specific failures that subsequent compression methods should address.

The paper has three main contributions. First, it defines answer preservation mathematically by limiting how much compression can reduce the best achievable answer accuracy for each query family. Second, it provides a practical protocol for testing, with statistical confidence, whether a supplied compressor keeps a fixed LALM's increase in answer error within the specified tolerance for every evaluated query family. Third, experiments spanning five English-prompted, multiple-choice audio question-answering benchmarks and four LALMs demonstrate compression-induced damage to particular query families that dataset averages conceal. We also show that grouping queries into different families changes which compression settings pass empirical accuracy checks, and identify when using the query to guide compression reduces the audio budget at the same accuracy tolerance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Related work. Task-oriented source coding studies compression through its effect on task performance, including limitations of the model performing the task (Enttsel & Corlay, 2026). Prompt-compression theory studies the tradeoff between prompt length and a fixed language model's answer quality, including the benefit of giving the compressor access to the question (Nagle et al., 2024). Learn then Test provides statistical procedures for choosing settings that satisfy specified risk limits (Angelopoulos et al., 2025), and CALM uses this approach to reduce language-model computation while limiting the increase in task loss relative to the full model (Schuster et al., 2022). Our focus is auditing audio compression against the answer-accuracy requirements of individual query families. Additional related work is in App. I.

2 Problem Formulation and Setup

Let $\mathcal{X}$ be the space of finite-duration, sampled audio clips and let $\mathcal{Q}$ be the family of queries that may be asked about such clips. Examples include speech-content, event-counting, and speaker or prosody queries. Conditioned on the raw audio input $X \in \mathcal{X}$, for each fixed query $q \in \mathcal{Q}$, let Y_q be the random correct answer taking values in $\mathcal{Y}_q$, let $\mathcal{A}_q$ be the action space of allowed predictions, and let $\ell_q := \ell : \mathcal{A}_q \times \mathcal{Y}_q \to [0, \infty)$ be the answer loss.

We define the set $C = \{C_b(X, q) : b \in \mathcal{B}\}$ as a family of budgeted, q-query-conditioned compressors on X, where $\mathcal{B}$ are retained-interface budgets. A compressed interface is denoted

$$Z_{q,b} = C_b(X, q) \in \mathcal{Z}_b, \tag{2.1}$$

where $\mathcal{Z}_b$ holds the space of admissible compressed interfaces at budget b. Here, admissible means measurable, finite, valid for the downstream LALM, and within budget b. For example, if q asks about voice-activity events and C retains audio fractions, then $C_b(X, q)$ may retain the most energetic 1-second chunks of X, with b controlling how many chunks are retained.

For each q and b, we compare the best answer rule using X, denoted $\mathcal{R}^\star_X(q)$, with the best answer rule using the compressed interface $Z_{q,b}$, denoted $\mathcal{R}^\star_{Z_{q,b}}(q)$:

$$\mathcal{R}^\star_X(q) = \inf_{\psi:\mathcal{X}\to\mathcal{A}_q} \mathbb{E}[\ell(\psi(X), Y_q)]\,, \qquad \mathcal{R}^\star_{Z_{q,b}}(q) = \inf_{\phi:\mathcal{Z}_b\to\mathcal{A}_q} \mathbb{E}[\ell(\phi(Z_{q,b}), Y_q)]\,. \tag{2.2}$$

The expectation is over the distribution of X and the query-specific, conditional answer law $Y_q \mid X$. For a fixed query q, the per-query excess answer-risk at budget b is

$$\Delta_q(b; C) = \mathcal{R}^\star_{Z_{q,b}}(q) - \mathcal{R}^\star_X(q). \tag{2.3}$$

Over the deployment query family $\mathcal{Q}$, the relevant guarantee is the largest such degradation:

$$\Delta_{\mathcal{Q}}(b; C) = \sup_{q\in\mathcal{Q}} \Delta_q(b; C). \tag{2.4}$$

The budgeted interface $C_b(X, q)$ is ε-answer-preserving on $\mathcal{Q}$ if, for $\varepsilon \geq 0$,

$$\Delta_{\mathcal{Q}}(b; C) \leq \varepsilon. \tag{2.5}$$

The supremum is intentional: a compressor should not be approved merely because its average error is small if it substantially damages a deployment-critical query type.

The answer-preservation frontier of C is the smallest retained budget that satisfies

$$b^\star_{\mathcal{Q}}(\varepsilon; C) = \inf\{b \in \mathcal{B} : \Delta_{\mathcal{Q}}(b; C) \leq \varepsilon\}\,, \tag{2.6}$$

with $b^\star_{\mathcal{Q}}(\varepsilon; C) = \infty$ if no budget b yields an ε-answer-preserving interface.

The frontier in (2.6) concerns Bayes-optimal answer rules. The operational audit instead fixes the deployed answerer. For $Z = Z_{q,b}$, write $\mathcal{R}_f(q; U)$ for its expected loss from interface U. Its excess risk satisfies

$$\begin{aligned}\mathcal{R}_f(q; Z) - \mathcal{R}_f(q; X) = \Delta_q(b; C) &+ \big[\mathcal{R}_f(q; Z) - \mathcal{R}^\star_Z(q)\big] \\ &- \big[\mathcal{R}_f(q; X) - \mathcal{R}^\star_X(q)\big].\end{aligned} \tag{2.7}$$

Thus observed degradation combines information loss with changes in the answerer's ability to use the interface. It can be negative and does not identify the Bayes frontier. Section 4 audits the specified answerer and compressor together. Further setup details are in App. J.

3 THEORETICAL FOUNDATIONS: TASK-AWARE ANSWER-PRESERVATION

Notational convention. For a fixed budget b and compressor C, we write $Z = Z_{q,b}$ and use $\Delta_{\mathcal{Q}}(Z; X)$ as shorthand for $\Delta_{\mathcal{Q}}(b; C)$.

The following results specialize classical comparison-of-experiments and task-oriented rate-distortion ideas (Blackwell, 1953; Le Cam, 1964; Witsenhausen, 1980). They provide the audit's foundations.

Assumption 3.1 (Regularity)**.** *For every $q \in \mathcal{Q}$, Bayes-optimal answer rules based on X and on any admissible compressed interface Z exist. The losses are uniformly bounded and all displayed expectations are finite.*

Answer-risk gaps are sufficiency deficits. For $\pi \in \mathcal{P}(\mathcal{Y}_q)$, where $\mathcal{P}(\mathcal{Y}_q)$ denotes probability laws on the answer space, define the Bayes envelope and induced posteriors

$$L_q(\pi) = \inf_{a \in \mathcal{A}_q} \int \ell(a, y)\, \pi(dy), \qquad \Pi_q^X = P_{Y_q|X,q}, \quad \Pi_q^Z = P_{Y_q|Z,q}, \tag{3.1}$$

then $\mathcal{R}_X^\star(q) = \mathbb{E}_X\left[L_q(\Pi_q^X)\right]$ and $\mathcal{R}_Z^\star(q) = \mathbb{E}_Z\left[L_q(\Pi_q^Z)\right]$, where the expectations use the marginal laws induced by the deployment distribution and compressor.

We define the one-sided answer deficiency of Z relative to raw audio X as

$$\delta_{\mathcal{Q}}(Z\|X) = \sup_{q \in \mathcal{Q}} \left\{\mathbb{E}_Z\left[L_q(\Pi_q^Z)\right] - \mathbb{E}_X\left[L_q(\Pi_q^X)\right]\right\}. \tag{3.2}$$

Theorem 3.2 (Restricted sufficiency-risk equivalence)**.** *For every Z,*

$$\delta_{\mathcal{Q}}(Z\|X) = \Delta_{\mathcal{Q}}(Z; X). \tag{3.3}$$

Thus Z is exact task-aware answer sufficient for $\mathcal{Q}$ iff $\Delta_{\mathcal{Q}}(Z; X) = 0$, and is ε-approximately sufficient iff $\Delta_{\mathcal{Q}}(Z; X) \le \varepsilon$.

The proof is in App. K using a Jensen-gap identity for the Bayes envelope. This identity interprets the excess *Bayes* answer risk as the restricted sufficiency deficit for the declared queries and loss. The fixed-answerer measurements in Section 4 retain the same raw-versus-compressed comparison but do not identify this deficit.

Partition refinement and operational budgets. The query-level object in (3.3) is a worst-query guarantee. In experiments, however, we observe finite query-family partitions rather than repeated samples for every individual query. Let μ denote the deployment distribution over queries, and for a partition $\mathcal{P}$ of $\mathcal{Q}$, define the partition-level family gap:

$$\Delta_{\mathcal{P}}^{\mathrm{fam}}(Z; X) = \max_{F \in \mathcal{P}: \mu(F) > 0} \bar{d}_F(Z), \qquad \bar{d}_F(Z) = \mathbb{E}_\mu[R_Z^\star(Q) - \mathcal{R}_X^\star(Q) \mid Q \in F], \tag{3.4}$$

and the deployment-average gap

$$\Delta_\mu^{\mathrm{avg}}(Z; X) = \mathbb{E}_\mu[R_Z^\star(Q) - \mathcal{R}_X^\star(Q)]. \tag{3.5}$$

Thus, $\Delta_{\mathcal{P}}^{\mathrm{fam}}$ is a population-level, lower-resolution version of the worst-query gap in (2.4).

Theorem 3.3 (Monotonicity under partition refinement)**.** *Let $\mathcal{P}' \preceq \mathcal{P}$ denote $\mathcal{P}'$ refines $\mathcal{P}$: for every fine cell $G \in \mathcal{P}'$, there exists a coarse cell $F \in \mathcal{P}$ such that $G \subseteq F$. Then, for all Z,*

$$\Delta_\mu^{\mathrm{avg}}(Z; X) \le \Delta_{\mathcal{P}}^{\mathrm{fam}}(Z; X) \le \Delta_{\mathcal{P}'}^{\mathrm{fam}}(Z; X) \le \Delta_{\mathcal{Q}}(Z; X). \tag{3.6}$$

As in (2.1), let C be a budgeted compressor and Z_b a budget-b interface, and define

$$\mathcal{B}_{\mathcal{P}}(\varepsilon; C) = \left\{b \in \mathcal{B} : \Delta_{\mathcal{P}}^{\mathrm{fam}}(Z_b; X) \le \varepsilon\right\}, \quad b_{\mathcal{P}}^\star(\varepsilon; C) = \inf\left\{b \in \mathcal{B} : \Delta_{\mathcal{P}}^{\mathrm{fam}}(Z_b; X) \le \varepsilon\right\}, \tag{3.7}$$

with $b_{\mathcal{P}}^\star(\varepsilon; C) = \infty$ if the feasible set is empty. If $\mathcal{P}' \preceq \mathcal{P}$, then

$$\mathcal{B}_{\mathcal{P}'}(\varepsilon; C) \subseteq \mathcal{B}_{\mathcal{P}}(\varepsilon; C), \qquad b_{\mathcal{P}}^\star(\varepsilon; C) \le b_{\mathcal{P}'}^\star(\varepsilon; C). \tag{3.8}$$

A budget that passes a refined partition also passes the coarse partition, but a budget that passes the coarse partition may fail once the same examples are split into more specific query families. The proof, including the budget consequence, is in App. K.3.

This theorem is the formal reason for using familywise sign-off rather than a dataset average. Refinement does not create new evidence, it removes averaging across heterogeneous subfamilies. A coarse speech family can pass at tolerance ε because low-damage lexical queries are averaged together with high-damage prosody queries. After refinement, the prosody cell can become the bottleneck, forcing a larger budget or a family-specific fallback.

Answer-risk frontiers and query conditioning. Keep the deployment query distribution μ from the partition definition, and let Q denote the runtime query with support $\mathrm{supp}(\mu) = \mathcal{Q}$. The prior μ weights average rate or length only, whereas the answer-preservation constraint remains worst-case over $\mathcal{Q}$. We write $I(\cdot;\cdot)$ and $H(\cdot)$ for mutual information and entropy (Cover & Thomas, 1991), respectively, and τ for the length or cost of the compressed interface. The frontiers below translate an answer-preservation tolerance into the smallest interface size that can meet it: $R^\star_{\mathrm{Bayes}}$ measures information rate, not Bayes answer-risk, while $L^\star$ measures expected length or cost. For query-agnostic compressors, i.e., when $C_b(X, q)$ does not depend on q:

$$\begin{aligned} R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) &= \inf_{\substack{C,\ b\in\mathcal{B} \\ C_b(X,q)=C_b(X,q')\ \forall q,q'\in\mathcal{Q}}} \{I(X; C_b(X,Q)) : \Delta_{\mathcal{Q}}(b; C) \le \varepsilon\}, \\ L^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) &= \inf_{\substack{C,\ b\in\mathcal{B} \\ C_b(X,q)=C_b(X,q')\ \forall q,q'\in\mathcal{Q}}} \{\mathbb{E}_{X,\,Q\sim\mu}[\tau(C_b(X,Q))] : \Delta_{\mathcal{Q}}(b; C) \le \varepsilon\}. \end{aligned} \tag{3.9}$$

For query-conditioned compressors, when $C_b(X, q)$ may depend on the realized query, then:

$$\begin{aligned} R^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) &= \inf_{C,\ b\in\mathcal{B}} \{I(X; C_b(X,Q) \mid Q) : \Delta_{\mathcal{Q}}(b; C) \le \varepsilon\}, \\ L^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) &= \inf_{C,\ b\in\mathcal{B}} \{\mathbb{E}_{X,\,Q\sim\mu}[\tau(C_b(X,Q))] : \Delta_{\mathcal{Q}}(b; C) \le \varepsilon\}. \end{aligned} \tag{3.10}$$

Theorem 3.4 (Conditioned compression advantage)**.** *For any design prior $Q \sim \mu$ with support $\mathcal{Q}$, independent of X,*

$$R^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \le R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}), \qquad L^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \le L^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}). \tag{3.11}$$

The rate inequality can be strict in an idealized representation class permitting the answer-factor encodings below. Let $\mathcal{Q} = \{q_1, q_2\}$ and $P(Q = q_1) = \lambda \in (0, 1)$, and let the audio contain two independent answer factors, $X = (V_1, V_2, W)$, with $0 < H(V_1), H(V_2) < \infty$, where W is irrelevant side information. Query q_i asks for factor V_i, so the corresponding correct-answer variable is $Y_{q_i} = V_i$, and 0-1 loss is evaluated on the label alphabet of V_i. Then

$$R^\star_{\mathrm{Bayes}}(0, \{q_1, q_2\}) = H(V_1, V_2), \quad R^\star_{\mathrm{Bayes,cond}}(0, \{q_1, q_2\}) = \lambda H(V_1) + (1-\lambda) H(V_2). \tag{3.12}$$

Thus the strict gap is $(1-\lambda)H(V_1) + \lambda H(V_2) > 0$: the agnostic interface must retain both answer factors to answer either possible query, whereas the conditioned interface observes Q first and retains only the realized answer factor.

Conditioning is thus a frontier-level possibility result, not an operational guarantee: it helps when different queries require different answer-relevant audio factors. At the ideal frontier the gain cannot be negative, but in a fixed learned system the measured gain can vanish or reverse when factors are shared, conditioning barely changes the retained interface, or the downstream LALM cannot use the retained information. The proof is in App. K.4.

4 Practical Methodology and Performance Metrics

Consider N dataset examples $\{(x_i, q_i, y_i)\}_{i=1}^N$, where example i consists of an audio clip, query, and correct answer. Let $\mathcal{P}$ be a query-family partition of $\mathcal{Q}$, and let $F_i \in \mathcal{P}$ be the family of query q_i. A candidate compression method m is a rule that, for each budget b, produces compressed interfaces, where $z^{(m)}_{i,b} = \mathcal{C}^{(m)}_b(x_i, q_i)$, following notation from (2.1). We fix one method m and suppress it by writing $z_{i,b} = z^{(m)}_{i,b}$, unless we compare frontiers across methods. Under a frozen answerer f, the raw-audio and compressed interface losses are

$$\ell^x_i = \ell_{q_i}(f(x_i, q_i), y_i), \qquad \ell^z_{i,b} = \ell_{q_i}(f(z_{i,b}, q_i), y_i). \tag{4.1}$$

For a query family $F \in \mathcal{P}$, let $N_F = |\{i : F_i = F\}|$. We estimate three quantities:

$$\widehat{\Delta}_F(b) = \frac{1}{N_F} \sum_{i:F_i=F} \big(\ell^z_{i,b} - \ell^x_i\big), \qquad \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b) = \max_{F\in\mathcal{P}:\, N_F>0} \widehat{\Delta}_F(b), \tag{4.2}$$

$$\widehat{\Delta}_{\mathrm{avg}}(b) = \frac{1}{N} \sum_{i=1}^{N} \big(\ell^z_{i,b} - \ell^x_i\big), \tag{4.3}$$

where $\widehat{\Delta}_{\mathrm{avg}}(b)$ asks how much damage appears after pooling all queries, while $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ asks how much damage appears in the worst observed deployment family, using per-family $\widehat{\Delta}_F(b)$ quantities. The family partition is fixed before observing losses, and unsupported families remain inconclusive rather than being removed from the guarantee. We define the hidden-damage margin:

$$\widehat{H}_{\mathcal{P}}(b) = \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b) - \widehat{\Delta}_{\mathrm{avg}}(b), \tag{4.4}$$

which is the degradation hidden by replacing the partition $\mathcal{P}$ with a dataset mean. Large margins on budget-degenerate controls can also reflect baseline heterogeneity or cancellation rather than budget-sensitive compression damage.

For prediction a and answer y, we use the 0-1 multiple choice loss, defined as $\ell(a, y) = \mathbf{1}\{a \neq y\}$. This matches our empirical claim: the benchmarks ask for one discrete answer, and compression is judged by whether it preserves answer correctness under the same LALM. The risk criterion extends to bounded losses, but the exact binomial endpoints below specifically require binary correctness losses. We use two tolerances: the average tolerance $\varepsilon_{\mathrm{avg}}$ limits overall regression, while the family tolerance ε_F limits damage to the worst deployment family. They may be set equal: when we report a single $\varepsilon = 0.05$, both tolerances are 0.05 in 0-1 loss units, i.e., at most five percentage points of additional error.

For $r \in \{\mathrm{avg}, \mathrm{fam}\}$, define

$$\widehat{\Delta}^{\mathcal{P}}_r(b) = \begin{cases} \widehat{\Delta}_{\mathrm{avg}}(b), & r = \mathrm{avg}, \\ \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b), & r = \mathrm{fam}, \end{cases} \qquad \varepsilon_r = \begin{cases} \varepsilon_{\mathrm{avg}}, & r = \mathrm{avg}, \\ \varepsilon_F, & r = \mathrm{fam}. \end{cases} \tag{4.5}$$

For a budget grid B, the empirical point frontier for a fixed method is

$$\widehat{b}^{\star}_{\mathcal{P}}(\varepsilon_r, \mathcal{Q}; r) = \min_{b \in B} \left\{ b : \widehat{\Delta}^{\mathcal{P}}_r(b) \leq \varepsilon_r \right\}, \tag{4.6}$$

with value ∞ for an empty feasible set. Pointwise confidence intervals alone do not certify a budget selected by searching this grid. For a method m, write $\widehat{b}^{\star}_{\mathcal{P}}(\varepsilon_r, \mathcal{Q}; r \mid m)$ for the same frontier computed after generating compressed interfaces with method m. For the agnostic and conditioned selectors, the operational conditioning gain is

$$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon_r, \mathcal{Q}; r) = \widehat{b}^{\star}_{\mathcal{P}}(\varepsilon_r, \mathcal{Q}; r \mid \mathrm{agn}) - \widehat{b}^{\star}_{\mathcal{P}}(\varepsilon_r, \mathcal{Q}; r \mid \mathrm{cond}). \tag{4.7}$$

Positive gain means that the conditioned selector reaches the same tolerance with less retained-audio budget, not that it has higher accuracy at a fixed budget. Unless stated otherwise, the headline conditioning claim uses $r = \mathrm{fam}$, since sign-off is family-wise.

The original curve intervals use paired example bootstraps (Efron & Tibshirani, 1993). Across-seed Student-t intervals summarize selector-training variability. These are descriptive rather than post-selection certificates. Estimator and bootstrap details are in App. L.

Independent confirmation. Choosing a compression setting and evaluating it on the same recordings can make approval overly optimistic. We therefore separate selection from confirmation: sample independent recordings, choose one question uniformly per recording, and split the recordings into 20% development and 80% confirmation. This gives recordings equal weight, whereas the selector sweeps average question rows. The model, compressor, preprocessing, prompts, scoring, candidate grid, and query families are fixed before confirmation.

On development data, we select the lowest-cost setting whose estimated error increase meets the five-point tolerance overall, c_{avg}, and the lowest-cost setting that meets it for every family, c_{fam}. Raw audio is available as a zero-degradation fallback. If any family lacks development examples, its preservation cannot be assessed, so c_{fam} defaults to raw audio. We also identify the supported family most damaged by c_{avg}, denoted $F^{\dagger}$. The two settings and this witness family are then locked before confirmation outcomes are inspected.

For a fixed candidate and population or family, let n_+ count raw-correct/compressed-wrong pairs and n_- count raw-wrong/compressed-correct pairs among n independent confirmation recordings. Write $u_t(k, n)$ and $l_t(k, n)$ for the one-sided Clopper-Pearson binomial endpoints with tail error t. Set $\eta = 0.05/3$ and $t = \eta/2$ (exact formulas in App. B):

$$U = u_t(n_+, n) - l_t(n_-, n), \qquad L = l_t(n_+, n) - u_t(n_-, n). \tag{4.8}$$

Each endpoint has one-sided coverage at least $1 - \eta$. On the confirmation data, test three assertions:

$$\mathrm{A} : U_{\mathrm{avg}}(c_{\mathrm{avg}}) \le \varepsilon, \qquad \mathrm{B} : L_{F^\dagger}(c_{\mathrm{avg}}) > \varepsilon, \qquad \mathrm{C} : \max_{F \in \mathcal{P}} U_F(c_{\mathrm{fam}}) \le \varepsilon. \tag{4.9}$$

A asks whether the average-selected setting meets the overall tolerance, B asks whether that same setting nevertheless harms the preselected family beyond tolerance, and C asks whether the family-selected alternative meets every family's requirement. Together, they test whether considering families changes a defensible deployment decision.

Bonferroni across A-C limits the probability of any false assertion to 5% within one dataset-model-compressor evaluation. Because C requires every family to pass, a false approval entails an erroneous upper bound for at least one unsafe family. No additional family correction is needed for this decision (App. B). This does not give joint coverage of every displayed family interval. Missing family support or incomplete confirmation prevents approval of a compressed setting. Failure to establish an assertion means insufficient evidence, not proof of its opposite. Raw-only approval means retaining uncompressed audio. A revised setting requires fresh confirmation data. These guarantees concern the evaluated population, not distribution shift or all experiments jointly.

Audit procedure. Algorithm 1 summarizes the audit for a candidate audio compressor. An evidence map is in App. G.

Algorithm 1: sign-off protocol for a candidate compressor.

1. **Specify:** Fix the LALM, compressor, candidate settings and costs, query-family partition, and tolerance $\varepsilon = 0.05$ (Section 5). Sample and split recordings into development and confirmation data as described in "Independent confirmation" in Section 4.
2. **Evaluate paired paths:** Run the same LALM on raw and compressed audio for each sampled question, and record the paired binary losses in (4.1). Keep confirmation outcomes sealed until the choices in Step 4 are locked.
3. **Expose hidden damage:** On development data, compute per-family and pooled error increases using (4.2)-(4.3). Identify the worst-affected family and compute the hidden-damage margin using (4.4).
4. **Select and lock:** Follow the development-data selection rules in "Independent confirmation" in Section 4: choose the lowest-cost overall-passing setting c_{avg}, the lowest-cost all-family-passing setting c_{fam}, and the supported family $F^\dagger$ most damaged by c_{avg}. Apply the stated raw-audio fallback and missing-family rules. Lock these choices before inspecting confirmation outcomes.
5. **Confirm:** On confirmation data, count harmful and beneficial answer flips and compute the bounds in (4.8) using the endpoints in App. B. Test A-C in (4.9). Report the number of evaluated recordings in each family.
6. **Accept, reject, or withhold approval:** Using (4.9), **accept** c_{fam} if $\max_{F \in \mathcal{P}} U_F(c_{\mathrm{fam}}) \le \varepsilon$ (C): every evaluated family meets the tolerance. **Reject** c_{avg} if $L_{F^\dagger}(c_{\mathrm{avg}}) > \varepsilon$ (B): the preselected family exceeds the tolerance, even if the overall check A passes. A setting with neither established family-wise acceptance nor established above-tolerance harm is **inconclusive**; withhold deployment approval. Raw-only acceptance means no compression. A revised setting requires fresh confirmation data.

5 Data and Experimental Setup

We evaluate with five English-prompted multiple-choice audio-question benchmarks: DCASE 2026 dev, a single-family control under the keyword partition (Yang et al., 2026); AudioMCQ-StrongAC, a multi-family benchmark and a selector-training source (He et al., 2026); MMSU, a speech and prosody taxonomy stress test (Wang et al., 2026); MMAR, a mixed speech, audio, and music reasoning check (Ma et al., 2025); and BigBench Audio, a text-dominated control for budget-degenerate behavior (Artificial Analysis, 2024; Suzgun et al., 2023). We evaluate four LALMs: Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2.5-Omni-7B (Xu et al., 2025), MiniCPM-o-4.5 (OpenBMB, 2026), and MERaLiON-2-10B (Institute for Infocomm Research, A*STAR, 2025). All four receive identical compressed waveforms in the comparison in Table 2. Learned-selector sweeps use Qwen2-Audio and Qwen2.5-Omni and independent confirmation uses Qwen2-Audio and MiniCPM.

We evaluate five compressions methodologies as chunk-retention policies: uniform, random, energy-based, learned query-agnostic, and learned query-conditioned. Audio is first divided into N one-second chunks. At budget b, a selector retains $k(b) = \max(1, \lfloor bN \rfloor)$ chunks in temporal order. Learned selectors use MLP scoring heads over 768-dimensional log-mel chunk features, with a 128-dimensional query embedding added in the conditioned case. The objective matches soft chunk-relevance targets by KL loss with count and entropy terms. Targets are normalized nonnegative

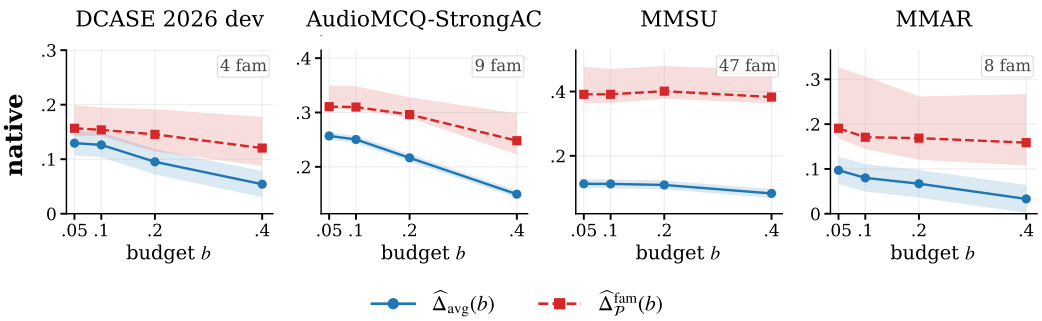


Figure 1: Compression damage across native query families with the learned query-conditioned selector and Qwen2-Audio. Red dashed curves show worst-family excess error (4.2) and blue curves show overall excess error (4.3). Bands are paired-bootstrap intervals. The full partition is in App. E.

increases in correct-answer negative log-likelihood when a chunk is removed from a frozen reference model's audio input. At inference, selectors use only audio and, when applicable, the question. The original AudioMCQ selector sweep is in-domain. The other datasets test zero-shot transfer. Implementation details are in App. N.

Confirmation experiments. We evaluate two distinct compression methods: uniform retention at $b \in \{0.1, 0.2, 0.4, 0.6, 0.8, 0.95\}$ and the EnCodec neural audio codec (Défossez et al., 2023) at nominal rates $\{1.5, 3, 6, 12\}$ kbps, selected on development data before confirmation. AudioMCQ uses 6,400 confirmation sources, with query families defined by source dataset and question type. A separate four-family evaluation uses 5,580 fresh confirmation sources. MMSU uses 3,812 confirmation waveform groups. Each group contributes one question, and recordings are at most 30 seconds. EnCodec reduces bitrate while preserving audio duration, so bitrate savings do not imply faster LALM inference. Protocol details are in App. C.

Query-family partitions. We report three partitions: The keyword (kw) partition maps dataset category strings into shared coarse families, e.g., `speech_content`, `paralinguistic`, `sound_event`, `sound_scene`, `music`, `temporal`, and `general`. The native partition uses benchmark-provided task labels or taxonomies. The semantic partition clusters query text using sentence embeddings and cosine K_{sem}-means clustering. We use one fixed, predeclared semantic partition per dataset, merge rare clusters before reporting, and do not sweep K_{sem} or choose it by the resulting gap. The effective post-merge family counts are in Table 1. The three partitions are not assumed to form one nested chain, and they should be read as alternative operational scopes for family-wise sign-off. The main sweep uses native partitions. Dataset and query details are in App. M.

6 Main Results

Accuracies are reported in percent. Excess-risk and hidden-damage quantities are reported as percentage points (pp) of 0-1 error. Budget frontiers and gains are reported as retained-budget points, i.e., $100\times$ retained-audio fraction. For example, a gain of 4.75 retained-budget points means that the same tolerance is met while retaining 0.0475 less of the audio. Higher accuracy is better and lower excess risk, hidden damage, and required budget are better. Positive conditioning gains mean that query-guided compression requires a smaller audio budget to meet the same accuracy tolerance. Conditioning results are in App. D. Unless stated otherwise, reported 95% intervals use the paired bootstrap described in Section 4.

Averages hide family damage, and the partition determines what is hidden. Figure 1 uses Qwen2-Audio-7B-Instruct with the learned query-conditioned selector. The result is that, in the multi-family budget-sensitive panels, the worst-family excess risk $\widehat{\Delta}^{\text{fam}}_{\mathcal{P}}(b)$ in (4.2) lies above the dataset mean $\widehat{\Delta}_{\text{avg}}(b)$ in (4.3). A budget that looks acceptable on average can therefore still damage the worst deployment family. Table 1 provides a snapshot form: DCASE and BigBench have zero keyword gap

Table 1: For each benchmark with Qwen2-Audio as the fixed model: number of examples N, raw-audio accuracy $\mathrm{Acc}_X = 100(1 - N^{-1}\sum_i \ell_i^x)$ in %, partition family counts (Section 5), and hidden-damage margin $100 \times \widehat{H}_{\mathcal{P}}(0.2)$ (4.4) under the learned query-conditioned selector.

			#families			hidden-damage margin		
Dataset	examples	Acc_X%	kw	native	semantic	kw	native	semantic
DCASE 2026 dev	1,607	44.4	1	4	5	0	5.04	7.68
AudioMCQ-StrongAC	19,480	74.2	4	9	12	6.79	7.96	14.29
MMSU	5,000	55.3	6	47	12	1.56	29.17	22.31
MMAR	1,000	49.6	3	8	6	1.88	10.18	13.34
BigBench Audio[†]	1,000	87.7	1	4	1	0	39.9	n/a

[†]single-family control under the keyword partition. semantic partition collapses to a single cluster.

Table 2: Four LALMs on the same 200 AudioMCQ examples, using identical compressed waveforms. Error increases span uniform and energy-based retention at $b \in \{0.1, 0.2, 0.4\}$. Ranges are not CIs. The last column counts the six paired-bootstrap 95% intervals wholly above zero.

LALM	Raw accuracy (%)	Error increase (pp)	CIs above zero
Qwen2-Audio-7B	70.5	9-17	6/6
Qwen2.5-Omni-7B	80	7.5-13	6/6
MiniCPM-o-4.5	78.5	2.5-8	5/6
MERaLiON-2-10B	60	6-9.5	6/6

because the keyword partition has one effective family, while their native partitions expose 5.04 pp and 39.9 pp hidden-damage margins. Where one partition truly refines another, this is the empirical counterpart of Theorem 3.3, otherwise it is the operational warning that sign-off is partition-indexed. BigBench is the control case: because its compressed predictions are effectively budget-invariant, its large native margin is evidence of baseline cancellation under a degenerate compression frontier. In the original descriptive decision analysis at $\varepsilon = 0.05$, switching between keyword and native families changes 29 of 125 matched pass/inconclusive/fail outcomes. The practitioner takeaway is to report both $\widehat{\Delta}_{\mathrm{avg}}(b)$ and $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ for the declared deployment partition, rather than certifying a compressor from the mean alone.

Answer preservation across four LALMs. Table 2 compares all four LALMs on the same 200 AudioMCQ examples, balanced over four query families (50 examples each). Each model receives identical compressed waveforms from uniform and energy-based retention at $b \in \{0.1, 0.2, 0.4\}$. All 24 estimated error increases are positive and 23 paired-bootstrap 95% intervals exclude zero (10,000 family-stratified draws). Thus degradation recurs across answerers and both selection methods. This shared-waveform comparison does not identify model-independent information loss.

Compression methods damage different query families. Figure 2 reports all six retention settings and four EnCodec rates on the same 6,400 AudioMCQ sources. Increasing nominal retention from 0.1 to 0.95 reduces pooled excess error from 24.12 to 1.81 pp for Qwen and from 9.98 to 1.02 pp for MiniCPM. At 0.1 retention, speech has the largest coarse-family point estimate in both models (30.12 and 14.13 pp). At 1.5 kbps, the largest EnCodec estimates instead occur for music with Qwen (10.54 pp) and sound with MiniCPM (6.18 pp). At 12 kbps, pooled losses are 0.59 and 0.50 pp, with pointwise intervals including zero. The audit thus reveals how family-level damage changes with the compression method and setting. Details are in App. C.1.

Independent confirmation and neural audio coding. The following bounds use (4.8). On the original 6,400-source AudioMCQ holdout, the primary Qwen retention setting $b = 0.8$ increases overall error by 4.50 $[3.24, 5.75]$ pp. Neither pooled approval nor above-tolerance harm in its preselected family is established. Qwen with EnCodec at 3 kbps passes the pooled check (2.58 $[1.50, 3.66]$ pp), whereas MiniCPM at 1.5 kbps remains inconclusive (4.11 $[2.90, 5.31]$ pp). Missing development support for one fine family forces the original AudioMCQ family-selected policies to raw audio. This fallback reflects missing evidence, not demonstrated harm from every compressed setting.

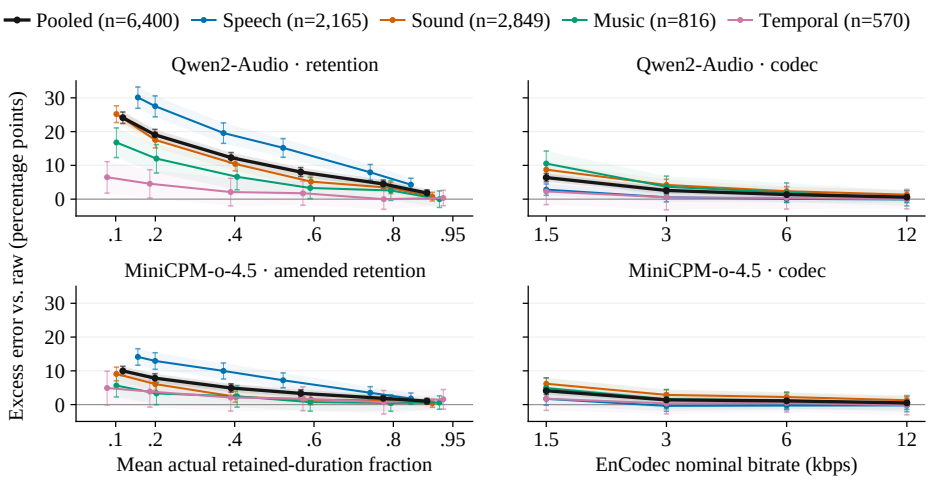


Figure 2: Complete AudioMCQ curves on the original 6,400-source cohort. Black shows pooled excess error and colors show four prespecified question types. Bars and bands are pointwise 95% paired Clopper-Pearson intervals. Retention uses group-specific measured duration fractions, while EnCodec uses nominal bitrate.

Table 3: Fresh AudioMCQ: excess answer error (pp) for Qwen2-Audio on 5,580 sources. Settings were selected on development data. Brackets use (4.8), without joint coverage of all displayed intervals. Speech and sound have 1,991 and 2,511 sources. Full decisions are in Appendix C.

Compression setting	Overall	Speech	Sound
Retention $b = 0.8$	$5.23\ [3.87, 6.59]$	$8.59\ [6.03, 11.11]$	$4.18\ [2.22, 6.13]$
Retention $b = 0.95$	$2.46\ [1.37, 3.53]$	$4.67\ [2.64, 6.67]$	$1.51\ [-0.02, 3.04]$
EnCodec 3 kbps	$2.63\ [1.50, 3.77]$	$0.45\ [-0.94, 1.84]$	$4.98\ [3.02, 6.91]$

On the fresh holdout, average-based selection chooses $b = 0.8$, while family-based selection chooses $b = 0.95$. The preselected speech family at 0.8 has an error increase above the five-point tolerance even at its lower bound (Table 3). Moving to 0.95 improves observed accuracy by 2.78 pp overall and 3.92 pp on speech, while actual retained duration rises from 77.4% to 88.4%. All-family approval remains unresolved: speech's upper bound exceeds five points. Thus the audit identifies a useful budget change without asserting that the higher budget is certified.

Retention at 0.95 and EnCodec at 3 kbps have similar overall point estimates, but different family profiles: retention causes more observed speech damage, while EnCodec causes more sound damage. These are descriptive comparisons, not an efficiency ranking across different resource measures. EnCodec passes the pooled check, but music and sound prevent all-family approval.

On MMSU, EnCodec at 1.5 kbps gives another concrete contrast: Qwen loses 2.99 pp overall versus 25.00 pp on emotion recognition. MiniCPM loses 3.02 versus 20.21 pp on non-verbal sound detection. Across these confirmation experiments, no selected compressed setting is confirmed to meet the error tolerance for every required query family.

7 Conclusion

We provide an acceptance criterion and practical protocol for auditing whether a supplied audio compressor preserves a fixed LALM's answer accuracy across specified query families. Across five benchmarks and four LALMs, the experiments show compression-induced degradation, damage hidden by dataset averages, and dependence on the evaluated families. We do not introduce an improved compressor, but the audit identifies failures and compression decisions that can guide subsequent improvements and enhance reliability of deployed audio compression techniques. Limitations are discussed in App. A.

AI USE STATEMENT

Generative AI tools assisted with aid or polish writing, and were reviewed by the human authors. The authors are responsible for the scientific claims, experimental evidence, and final text.

ETHICAL STATEMENT

No ethical concerns to disclose.

REPRODUCIBILITY STATEMENT

A .zip file has been attached as a supplementary material to this submission. It reproduces the empirical results in the main body and the appendix.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=BvxaP3sVbA.

Artificial Analysis. Big Bench Audio. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/ArtificialAnalysis/big_bench_audio.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, 1993.

Andriy Enttsel and Vincent Corlay. Model-aware rate-distortion limits for task-oriented source coding. *arXiv preprint arXiv:2602.12866*, 2026.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, Chunyat Wu, Qianhua He, Tan Lee, Xie Chen, Wei-Long Zheng, Weiqiang Wang, Mark Plumbley, Jian Liu, and Qiuqiang Kong. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Charles Herrmann, Richard Strong Bowen, and Ramin Zabih. Channel selection using Gumbel softmax. In *European conference on computer vision*, pp. 241–257. Springer, 2020.

Institute for Infocomm Research, A*STAR. MERaLiON-2-10B. Hugging Face model card, 2025. URL https://huggingface.co/MERaLiON/MERaLiON-2-10B. Accessed September 25, 2026.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-AQA: A crowdsourced dataset for audio question answering. In *European Signal Processing Conference*. IEEE, 2022. doi: 10.23919/EUSIPCO55093.2022.9909680.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74837–74868, 2025. doi: 10.52202/085713-2252. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/610a7d6507d55be70c6d13d0b663227d-Abstract-Datasets_and_Benchmarks_Track.html.

Alliot Nagle, Adway Girish, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. Fundamental limits of prompt compression: A rate-distortion framework for black-box language models. *Advances in Neural Information Processing Systems*, 37:94934–94970, 2024.

OpenBMB. MiniCPM-o 4.5. Hugging Face model card, 2026. URL https://huggingface.co/openbmb/MiniCPM-o-4_5. Accessed September 25, 2026.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472. Curran Associates, Inc., 2022. doi: 10.52202/068431-1269. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6fac9e316a4ae75ea244ddcef1982c71-Paper-Conference.pdf.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yHzCDP1tXw.

Hans S. Witsenhausen. Indirect rate distortion problems. *IEEE Transactions on Information Theory*, 26(5):518–521, 1980.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/2503.20215.

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. Multi-domain audio question answering benchmark toward acoustic content reasoning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11464798. URL https://arxiv.org/abs/2505.07365.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

A LIMITATIONS

Query-family coarsening. The empirical audit averages fixed-answerer excess loss within declared query families before taking the family maximum. Keyword rules, dataset-native labels, and semantic clusters can hide differences within a family. The reported $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}()$ values are therefore partition-indexed operational estimates; they do not estimate the worst-query Bayes risk in (2.4).

Fixed-answerer evaluation. Raw and compressed risks are measured from the predictions of a fixed LALM under declared prompts, decoding, answer normalization, and loss. The resulting frontier applies to that evaluation setting. Bootstrap intervals describe sampling uncertainty, not uncertainty from model choice, prompt choice, or distribution shift. Multiple queries sharing one source recording require uncertainty estimation that accounts for their dependence.

Model-class and interface gap. Observed degradation can reflect removed information, retained information the answerer cannot use, or interface mismatch. Comparisons across answerers and interfaces provide diagnostics but do not identify a causal decomposition of these sources. The audit assesses whether the tested configuration preserves answer quality regardless of the internal source of its errors.

Query-conditioned compression assumptions. Conditioned compressors require the query at encoding time. The independent-query assumption is specific to the idealized comparison in Theorem 3.4; the operational audit evaluates the declared joint audio-query population. This matches interactive question answering but not offline archival compression. It also requires care to avoid leakage: the query may guide what to retain, but the compressor must not use answer labels or evaluation metadata unavailable at deployment.

Dataset, model, and seed coverage. The original selector evidence covers five benchmarks and two LALM backbones, with a separate four-answerer control; the evaluated prompts, chunk budgets, and interfaces limit generalization. Some appendix results are single-seed or partial-coverage: V2.1 α-sweep wins are seed-42 only; several Qwen2-Audio V2 per-seed values were reported only as aggregates; V1 Qwen2.5-Omni coverage is incomplete for DCASE, AudioMCQ, and MMSU; and some native-family analyses rely on cross-backbone rather than same-backbone estimators.

Confirmation and codec scope. The completed neural-codec experiments add a bitrate-controlled interface alongside chunk retention. Their nominal and quantizer-payload rates do not establish complete serialized bitrate or LALM latency savings. MMSU waveform deduplication does not resolve original-source independence. Exception-only and minimum-length amendments are separately reported in Appendix C.2. The supplementary free-response scores measure lexical reference agreement; human validation was not performed.

Deployment scope. The paper gives a framework, estimators, and empirical guidance. It does not prove that any particular compressor is universally safe, nor that a family partition chosen on a benchmark will remain correct under distribution shift, longer audio, different languages, new answerers, or different cost models.

B EXACT PAIRED CONFIRMATION BOUNDS

For $n > 0$ and $t = 1/120$, define the one-sided Clopper-Pearson endpoints

$$\begin{aligned} u_t(k,n) &= \begin{cases} 1, & k = n, \\ \mathrm{Beta}^{-1}(1-t; k+1, n-k), & k < n, \end{cases} \\ l_t(k,n) &= \begin{cases} 0, & k = 0, \\ \mathrm{Beta}^{-1}(t; k, n-k+1), & k > 0. \end{cases} \end{aligned} \tag{B.1}$$

Here Beta^{-1} denotes a beta-distribution quantile. Each of n_+ and n_- has a binomial marginal under independent sampling; their mutual dependence is allowed. The union bound gives one-sided error at most $2t = \eta$ for each difference endpoint in (4.8). Conditional on the development choices, A and B each have false-assertion probability at most η. If the family-aware candidate violates any family tolerance, choose one such population family: C implies an erroneous upper-bound decision for that family, also with probability at most η. A union bound over A, B, and C gives error at most

$3\eta = 0.05$, also after averaging over development data. This is decision validity within one declared instance, not simultaneous interval coverage across families or across the study. A missing family ($n = 0$) is unsupported; raw identity is exempt because its paired difference is identically zero.

C Additional experiments with frozen deployment decisions

Models, inputs, and compression. The additional experiments use frozen Qwen2-Audio-7B-Instruct and MiniCPM-o-4.5 answerers, with bfloat16 weights and deterministic decoding. The checkpoint revisions are `0a095220c30b7b31434169c3086508ef3ea5bf0a` and `1f761131fa83f5ed3cd6f2f22b225c4501d154fa`, respectively. Qwen uses constrained choice-letter decoding with at most five new tokens; MiniCPM uses its native greedy decoder with at most 16. Labels are used only for scoring. A refusal or unparsable answer is incorrect; an execution failure is missing evidence.

All conditions start from the same full mono 24 kHz recording, limited to 30 seconds, with the same subsequent resampling to 16 kHz for each answerer. Uniform retention uses nonoverlapping one-second chunks, preserves the final partial chunk when selected, and concatenates selected chunks chronologically. The nominal fractions are $\{0.10, 0.20, 0.40, 0.60, 0.80, 0.95\}$; with N chunks, $k = \max(1, \lfloor bN \rfloor)$ and selected indices are $\lfloor (j + 0.5)N/k \rfloor$, $j = 0, \ldots, k - 1$. EnCodec (Défossez et al., 2023) uses its frozen 24 kHz mono checkpoint, revision `c1dbe2ae3f1de713481a3b3e7c47f357092ee040`, at $\{1.5, 3, 6, 12\}$ kbps. Decoded audio retains the original duration. Raw identity is a separate candidate.

Cohorts and frozen choices. AudioMCQ contains 8,000 original-source groups, with one randomly selected question per group, split into 1,600 development and 6,400 confirmation groups. Sampling and splitting use seed 20260925. Its 16 fine families are the metadata pairs `source_dataset::question_type`. MMSU contains 4,764 deduplicated waveform groups, split into 952 development and 3,812 confirmation groups, with 47 native tasks. Distinct MMSU waveforms could originate from the same recording: original-source independence was not established. Its numerical bounds therefore do not establish unconditional population certificates.

The primary instance is AudioMCQ/Qwen/uniform retention. For each instance, development data select the cheapest candidate with pooled mean excess error at most $\varepsilon = 0.05$, c_{avg}, and the cheapest satisfying every family mean, c_{fam}. Raw is included. The witness family is the largest development-supported mean excess under c_{avg}. Choices and witnesses are frozen before confirmation. In the original AudioMCQ cohort, `MusicCaps::sound` has no development observation, forcing c_{fam} to raw; this is insufficient family support, not evidence that every compressed setting is harmful.

Interpretation of the bounds. For a fixed candidate and population or family, let n_+ count raw-correct/compressed-wrong pairs and n_- count raw-wrong/compressed-correct pairs. The estimated excess error is $(n_+ - n_-)/n$. With one-sided Clopper–Pearson limits U_t, L_t and $t = 1/120$, the endpoints for the selected-comparison A/B/C results are

$$\begin{aligned} U_\Delta &= U_t(n_+, n) - L_t(n_-, n), \\ L_\Delta &= L_t(n_+, n) - U_t(n_-, n). \end{aligned} \tag{C.1}$$

Assertion A is pooled approval of c_{avg} ($U_\Delta \leq 0.05$); B is above-tolerance harm in its frozen witness ($L_\Delta > 0.05$); C is approval of c_{fam} in every required family. Under independent sampling and the frozen-selection procedure, the allocation supports at least 95% simultaneous validity for A/B/C within one instance. C is an intersection–union decision, not simultaneous coverage of every displayed family interval. No study-wide or distribution-shift guarantee is asserted. For raw identity, all differences and bounds are zero by construction.

Original results and limitations. Five of the eight original instances have complete selected comparisons (O1–O5). The primary O1 has 452 harmful and 164 beneficial flips among 6,400 pairs: observed excess error is 4.50 points, but pooled approval and witness harm remain unestablished. O2 establishes pooled approval of EnCodec at 3 kbps. The MMSU codec comparisons exhibit substantial witness damage (Table 5), while retaining the source-independence limitation. No result establishes a complete source-independence-supported A/B/C decision reversal, and no selected compressed

Table 4: The 11 selected-comparison records. Q = Qwen2-Audio; M = MiniCPM; R = retention; E = EnCodec. Settings are nominal retention fractions or codec kbps. Pooled excess and bounds are percentage points. In A/B/C, + denotes an established assertion under the row's stated scope, ? denotes unestablished, and I denotes raw identity only. MMSU rows (†) have unresolved source independence; amended rows (∗) are exploratory and their entries are numerical outcomes, not original confirmatory claims. Row X4 reuses O5 exactly.

ID	Cohort	Model	Interface	n	$c_{\text{avg}}/c_{\text{fam}}$	Pooled excess [bounds]	A/B/C
O1	AudioMCQ	Q	R	6,400	.80/raw	$4.50\ [3.24, 5.75]$	?/?/I
O2	AudioMCQ	Q	E	6,400	3/raw	$2.58\ [1.50, 3.66]$	+/?/I
O3	AudioMCQ	M	E	6,400	1.5/raw	$4.11\ [2.90, 5.31]$	?/?/I
O4	MMSU†	Q	E	3,812	1.5/raw	$2.99\ [1.41, 4.56]$	+/+/I
O5	MMSU†	M	E	3,812	1.5/raw	$3.02\ [1.31, 4.72]$	+/+/I
F1	Fresh AudioMCQ	Q	R	5,580	.80/.95	$5.23\ [3.87, 6.59]$	?/+/?
F2	Fresh AudioMCQ	Q	E	5,580	3/3	$2.63\ [1.50, 3.77]$	+/?/?
X1	AudioMCQ*	M	R	6,400	.40/raw	$4.92\ [3.62, 6.22]$	?/+/I
X2	MMSU†*	M	R	3,812	.95/raw	$3.73\ [2.09, 5.36]$	?/?/I
X3	MMSU†*	Q	R	3,812	.80/raw	$6.79\ [4.95, 8.63]$	?/?/I
X4	MMSU†*, reused	M	E	3,812	1.5/raw	$3.02\ [1.31, 4.72]$	+/+/I

Table 5: Selected development-locked witnesses; row IDs and statistical scopes follow Table 4. These families were not chosen from confirmation outcomes. Error increases and bounds are percentage points.

ID	Witness	n	Harmful/beneficial	Excess [bounds]
O1	MusicCaps::music	70	4/1	$4.29\ [-8.25, 16.01]$
O4	Emotion recognition	88	23/1	$25.00\ [8.21, 38.88]$
O5	Non-verbal sound detection	94	20/1	$20.21\ [5.06, 33.07]$
F1	Speech	1,991	218/47	$8.59\ [6.03, 11.11]$
X1	SpeechCraft::speech	1,743	249/46	$11.65\ [8.63, 14.60]$

family policy receives all-family approval. Raw-only C is not an empirically discovered compression solution.

Fresh coarse-family evaluation. A separately registered supplement retains the original 1,600 development groups and evaluates 5,580 previously unused, source-disjoint AudioMCQ groups. Both policy locks precede access to this holdout. The four native question types contain 629 music, 2,511 sound, 1,991 speech, and 449 temporal questions. These coarser requirements do not establish protection for the original fine families. For retention, $c_{\text{avg}} = 0.80$ and $c_{\text{fam}} = 0.95$. The frozen speech witness exceeds the tolerance at 0.80 (F1). At 0.95, pooled excess is $2.4552\ [1.3728, 3.5332]$ points, while speech excess is $4.6710\ [2.6432, 6.6742]$: its upper endpoint alone prevents C. Pooled accuracy rises descriptively from 69.44% to 72.22%, against raw 74.68%; this improvement is not an additional registered assertion. For EnCodec, both policies select 3 kbps; A holds, but music and sound upper endpoints (5.49 and 6.91 points) prevent C. Identical policies cannot demonstrate a policy-selection reversal.

Measured resources. On the fresh cohort, actual retained-duration fractions are 0.7740 and 0.8844 at nominal 0.80 and 0.95, with mean Qwen audio-token counts 223.87 and 258.67, respectively, against raw 286.56. EnCodec 3 kbps preserves duration and that raw token count. Its mean quantizer-index payload rate is 3,000.52 bits/s, excluding framing and entropy coding; official serialized bitrate was unavailable. Codec results therefore report nominal bandwidth and explicitly identified payload bits, not measured bitstream size. These measurements do not establish end-to-end latency or memory savings.

C.1 Complete descriptive compression grids

All eight dataset–model–interface grids are complete under their stated protocols or amendments: two datasets, two models, and uniform retention or EnCodec. Each of the six retention fractions and four

codec rates has 6,400 paired AudioMCQ sources or 3,812 paired MMSU waveform groups. Figure 2 uses the original AudioMCQ cohort, not the separate 5,580-source evaluation. Its four question types contain 2,165 speech, 2,849 sound, 816 music, and 570 temporal questions. These summaries retain the original question per source and do not replace the 16 fine-family requirements or remedy missing development support.

For these descriptive curves, (C.1) uses tail $t = 0.0125$: a union bound over the four marginal Clopper–Pearson tails gives pointwise 95% interval coverage. This differs from the registered A/B/C allocation $t = 1/120$. The intervals do not give simultaneous coverage across settings or families and do not justify selecting a newly certified budget. Retention points use each displayed group's mean actual retained signal fraction. At nominal $b = 0.10$, this fraction is 0.156 for speech and 0.102 for sound and music; equal nominal budgets need not retain equal duration fractions. EnCodec uses a separate nominal-kbps axis. Neither axis establishes inference latency or memory savings.

AudioMCQ/MiniCPM retention uses the exception-only amendment: one recording is repaired at each of six settings, while original failures remain recorded. MMSU/Qwen retention uses its exception-only amendment, and MMSU/MiniCPM retention completes the existing minimum-length amendment on the exposed cohort. Figure 3 reports the complete pooled MMSU curves. Their unresolved original-source independence prevents unconditional population certification. The completed grids leave the primary designation, original policy locks, tolerance, and registered A/B/C outcomes unchanged.

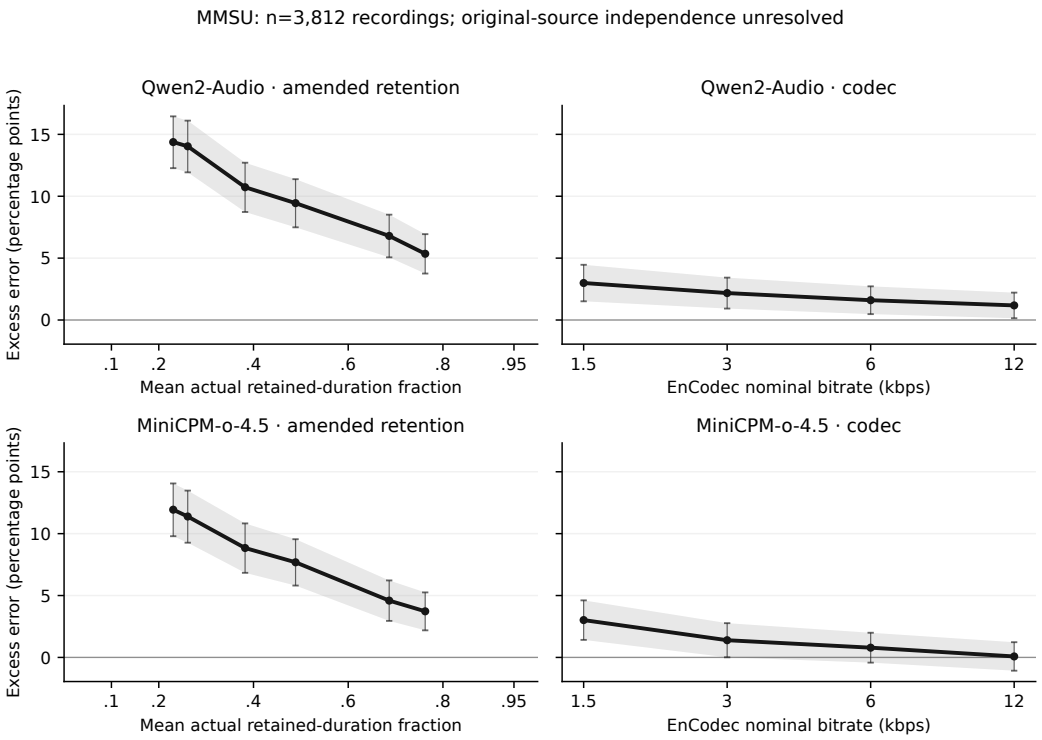


Figure 3: Complete descriptive MMSU curves, with 3,812 paired waveform groups per setting. Retention uses the disclosed amendments, and codec runs use the original protocol. Intervals are pointwise 95% calculations conditional on the unresolved source-independence assumption. Duration and bitrate are separate resource axes. These results do not change the registered decisions.

C.2 Incomplete original comparisons and disclosed amendments

The original AudioMCQ/MiniCPM retention comparison retains 6,399 of 6,400 selected pairs. Original MMSU/Qwen retention retains 3,807 of 3,812. Original MMSU/MiniCPM retention has 6,634 of 6,664 development-condition outputs and no original policy lock. Missing outputs were not dropped, scored as model errors, or overwritten by amendments.

Table 6: Complete native free-response results ($n = 473$ per condition). Exact agreement is a percentage and token-F1 lies in $[0, 1]$; loss changes and descriptive interval endpoints are signed differences on that scale. Positive loss change means worse lexical agreement.

Condition	Exact (%)	Token-F1	Loss increase	Descriptive 95% interval
Raw	31.50	0.3485	0	$[0, 0]$
Retention 0.20	27.91	0.2953	0.0532	$[0.0192, 0.0877]$
Retention 0.80	30.02	0.3304	0.0181	$[-0.0039, 0.0391]$
EnCodec 1.5 kbps	29.81	0.3183	0.0302	$[-0.0030, 0.0633]$
EnCodec 12 kbps	30.66	0.3403	0.0082	$[-0.0123, 0.0287]$

Three separate retention amendments complete those evaluations. MiniCPM's exception-only extension calls the original path first and pads only identified failing subthreshold inputs to 1,281 samples at 16 kHz; X1 uses 6,399 unchanged successes and one repaired answer. Qwen's corresponding minimum is 321 samples; X3 uses 3,807 unchanged successes and five repaired answers. Their original development choices and witnesses remain unchanged. The separate MMSU/MiniCPM padding evaluation X2 applies the 1,281-sample minimum and selects nominal retention 0.95. These modifications followed exposure to original confirmation results and remain exploratory. MMSU source lineage also remains unresolved. No codec or raw input required MiniCPM padding, so X4 reuses O5 exactly and is not a new replication.

As a disclosed sensitivity analysis, all compatible binary completions of the missing original answers leave pooled approval unestablished. AudioMCQ/MiniCPM pooled excess ranges from 4.9219 to 4.9375 points; its fully observed witness remains above tolerance. MMSU/Qwen ranges from 6.6632 to 6.7943 points, with both A and B unestablished in every completion. These are completion ranges, not confidence intervals or manufactured predictions. The amendments do not alter those A/B classifications or complete the original pipelines retroactively.

C.3 Native free-response short-answer evaluation

We additionally evaluate Qwen2-Audio on 473 official Clotho-AQA (Lipping et al., 2022) test recordings, selecting one native non-yes/no short-answer question per recording with seed 20260925. No multiple-choice options or answer vocabulary are supplied. The prompt requests a short answer, preferably one word; greedy generation allows at most 32 new tokens. The five fixed conditions yield 2,365 completed answers. Three released recordings lack original Freesound URLs; their released filenames and decoded-waveform duplicate checks provide the available provenance.

Scoring applies Unicode NFKC normalization, lowercasing, punctuation-to-space conversion, and whitespace collapse. Exact agreement accepts any normalized official reference. Bounded token-F1 loss is $1 - \max_{r \in \mathcal{R}_i} \mathrm{F1}(\hat{y}_i, r)$, where $\mathcal{R}_i$ contains the official answers. Empty answers and refusals receive loss one. Table 6 reports paired loss changes against raw audio; intervals use 2,000 paired recording bootstrap resamples, seed 20260925, and percentile 2.5–97.5% endpoints.

Aggressive retention at 0.20 increases bounded F1 loss, with a descriptive interval excluding zero. The other compressed conditions have positive point changes with intervals containing zero; these results establish neither degradation nor preservation in those conditions. The scores measure lexical reference agreement, not human-rated semantic correctness, and valid alternative wording may be penalized. **Human validation was not performed.** This supplementary evaluation provides no A/B/C certificate, familywise semantic guarantee, or study-wide inference.

D Additional Model, Conditioning, and Feature Controls

Four-model control. Table 2 in the main body reports the model-specific results on the 200-example AudioMCQ cohort. The six conditions combine uniform and energy-based retention at $b \in \{0.1, 0.2, 0.4\}$. Raw and compressed waveforms are held fixed across the four answerers. Intervals are paired, family-stratified bootstrap intervals.

Conditioning gains. The full three-seed comparison supplements the numerical examples in the main text.

Table 7: Operational conditioning gain $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon_F, \mathcal{Q}; \mathrm{fam})$ at $\varepsilon_F = 0.05$ on the retained-budget frontier. Entries are means over seeds $\{42, 123, 456\}$, reported in retained-budget points, where one point equals 0.01 retained-audio fraction. Positive values mean that the conditioned selector reaches the same familywise tolerance with less retained audio than the agnostic selector. Brackets are symmetric across-seed Student-t 95% CIs over selector-training stochasticity, not paired example-bootstrap CIs. Sign labels indicate per-seed sign agreement.

	Qwen2-Audio			Qwen2.5-Omni		
Dataset	$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$	[95% CI] (pts)	sign	$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$	[95% CI] (pts)	sign
DCASE 2026 dev	-3.00	[-15.2, 9.2]	mixed	-3.10	[-12 , 5.8]	mixed
AudioMCQ-StrongAC	4.75	[2.3, 7.2]	consistent+	2.15	[0.2, 4.1]	consistent+
MMSU	-3.36	[-6.3, -0.4]	consistent−	-7.34	[-18.2, 3.5]	consistent−
MMAR	-5.40	[-27.5, 16.7]	mixed	0.80	[-10.1, 11.7]	mixed
BigBench Audio	0		zero	0		zero

Capacity-matched feature comparison. On a content-disjoint set of 2,000 AudioMCQ examples, separate PCA transforms reduce log-mel and native audio features to 512 dimensions for identical 19,740,930-parameter selectors. Native-minus-mel mean accuracy improves by 1.58, 1.63, 1.80, and 1.38 percentage points at retained budgets 0.05, 0.1, 0.2, and 0.4. Across the full-width and PCA-controlled comparisons together, worst-family 0/1 loss is higher with native features in 20 of 24 seed-by-budget contrasts, while mean loss is conclusively lower in 9 of 24. These descriptive comparisons show that the feature representation can improve average accuracy without resolving worst-family preservation. They do not identify a causal decomposition of information loss and model error.

E Complete Partition-Curve Comparison

F Appendix Roadmap

Appendix C reports the confirmation experiments, amendments, and supplementary free-response evaluation. Appendix D provides further context for the four-model comparison, conditioning results, and feature comparison. The remaining material supplements the original analysis. Section I adds related-work context. Section J records supplementary problem-setup and operational-budget details. Section K gives proofs and auxiliary derivations for the theory stated in the main text. Section L defines the practical estimators, uncertainty aggregation, and axis conventions. Section M and Section N document datasets, query families, models, compression methods, budgets, and implementation details. Section O contains all secondary figures, tables, ablations, and family-level analyses. Section P records incomplete or suspicious results and caveats.

G Supplementary Algorithm Evidence Map

H Supplementary Selector-Query Audits

Decoupled selector-query audit. For query-conditioned selectors, let s_i denote the query text supplied to the selector. The anchor run uses $s_i = q_i$. A permuted run uses $s_i = \sigma(q_i)$, where σ is drawn from a finite set Σ of valid query permutations, either globally or within family, while the downstream LALM still receives the true query q_i. We average over $|\Sigma| \leq 10$ valid permutations per training seed. Thus the audit changes only the selector-side query signal and does not ask the LALM the wrong question. For $\varepsilon_F = 0.05$, we report

$$\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}} = \widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(0.05, \mathcal{Q}; \mathrm{fam} \mid s_i = q_i) - \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(0.05, \mathcal{Q}; \mathrm{fam} \mid s_i = \sigma(q_i)), \quad \text{(H.1)}$$

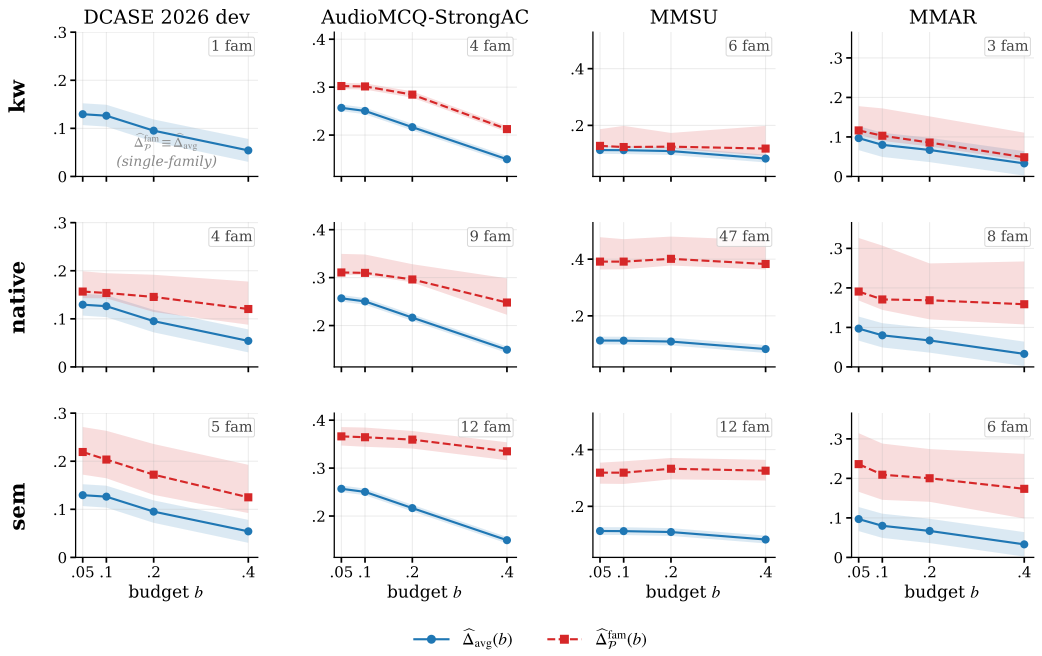


Figure 4: Complete version of Figure 1: keyword, native, and semantic partitions (rows), with Qwen2-Audio and the learned query-conditioned selector. Red dashed curves show worst-family excess error and blue curves show overall excess error. Bands are paired-bootstrap intervals. The main figure retains the native-partition row.

Table 8: Supplementary evidence map for Algorithm 1. This table supplements the main audit; it records how the main empirical objects map to practitioner actions.

Algorithm step	Empirical evidence	Meaning	Practitioner consequence
Specify partition and tolerances	Table 1: native partitions expose margins hidden by keyword partitions.	The partition is part of the claim.	Report the sign-off partition and tolerance before approving a budget.
Run paired raw/compressed interfaces	Figure 1: worst-family excess exceeds the dataset mean.	Compression damage is answer-level paired excess risk.	Evaluate raw and compressed answers under the same frozen LALM and query.
Expose hidden damage	Per-family excess losses identify the families blocking approval.	Remediation depends on the failing-family structure.	Use bypass/retraining for local harm; raise budget/change interface for diffuse harm.
Evaluate conditioning	Table 7: conditioning is positive, negative, mixed, or zero depending on dataset/backbone.	Conditioning is regime-dependent.	Report signed, seeded, backbone-specific conditioned gain.
Audit query use	Table 9: downstream query-use is substantial only in one completed cell.	Selector sensitivity need not change final answers.	Use the decoupled audit for conditioned compressors.
Check contamination	Table 10: naive shadow queries inflate the signal.	Prompt mismatch is not selector query-use.	Treat shadow-query runs as controls, not headline evidence.

which means that (4.7) is recomputed with the indicated selector-side query stream. Positive values mean the correct selector query changes the downstream answer-preservation frontier.

Selector query use must survive downstream. Selector-internal query sensitivity is necessary but insufficient. Table 9 recomputes $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$ after permuting only the selector-side query, as the LALM still receives the true query. Reporting bands use PHI1 for $|\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}| \leq 0.01$, PHI2 for $0.01 < |\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}| < 0.05$, and PHI3 for $|\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}| \geq 0.05$. Both AudioMCQ-StrongAC cells are PHI3 substantial and exclude zero, MMSU is null-to-small with CIs including zero, and BigBench has a zero observed query-use gain. Thus selected-chunk changes count as deployment evidence only when the decoupled audit changes the frontier seen by the frozen answerer.

Table 9: Decoupled selector-query audit $\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$ at $\varepsilon_F = 0.05$. Only the selector query is permuted; the LALM receives the true query. Entries are retained-budget points, where one point equals 0.01 retained-audio fraction. Positive values mean that the correct selector-side query improves the downstream retained-budget frontier. PHI bands use the magnitude thresholds defined in the text.

Backbone	Dataset, regime	$\Delta\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$ [95% CI] (pts)	band
Qwen2-Audio	AudioMCQ-StrongAC, global	7.87 [5.1, 10.6]	**PHI3**
Qwen2-Audio	MMSU, global	1.11 [−4.6, 6.9]	PHI2
Qwen2-Audio	MMSU, within-family	0.92 [−5.3, 7.1]	PHI1
Qwen2-Audio	BigBench Audio, global	0	PHI1†
Qwen2.5-Omni	AudioMCQ-StrongAC, global	7.17 [4.5, 9.9]	**PHI3**
Qwen2.5-Omni	MMSU, global	−0.43 [−4.5, 3.7]	PHI1
Qwen2.5-Omni	MMSU, within-family	−0.05 [−2.8, 2.7]	PHI1
Qwen2.5-Omni	BigBench Audio, global	0	PHI1†

†Observed zero query-use gain; a single family alone does not imply identical query text.

Naive shadow-query evaluation contaminates the audit. Table 10 isolates an audit failure mode: if both selector and LALM receive the permuted query, the effect mixes selector query-use with asking the answerer the wrong question. On AudioMCQ-StrongAC with Qwen2-Audio, the naive protocol is 1.22-$1.25\times$ the decoupled signal. The $\varepsilon_F = 0.05$ decoupled entry intentionally matches Table 9. Hence the decoupled protocol is the headline query-use audit and shadow-query runs are contamination controls.

Table 10: Naive shadow-query contamination versus the decoupled selector-only audit on AudioMCQ-StrongAC with Qwen2-Audio at $\varepsilon_F \in \{0.01, 0.02, 0.05\}$. The decoupled protocol permutes only the selector query, whereas the naive protocol permutes both the selector and LALM queries. Entries are retained-budget points, where one point equals 0.01 retained-audio fraction. The naive-minus-decoupled column uses paired runs; positive values quantify shadow-query contamination.

ε_F	decoupled [95% CI] (pts)	naive [95% CI] (pts)	naive−decoupled [95% CI] (pts)	ratio
0.01	1.9 [0, 3.8]	2.37 [0.3, 4.4]	0.47 [0, 1]	×1.25
0.02	3.8 [0, 7.6]	4.74 [0.6, 8.8]	0.94 [−0.1, 2]	×1.25
0.05	7.87 [5.1, 10.6]	9.57 [6.7, 12.5]	1.7 [0.4, 3]	×1.22

I Additional Related Work

The main paper cites four threads: audio-language modeling, audio tokenization/compression, rate-distortion/information bottleneck theory, and efficient multimodal inference. Qwen2-Audio and Qwen2.5-Omni provide the frozen answerers used here (Chu et al., 2024; Xu et al., 2025). Neural audio codecs and tokenizers such as SoundStream, EnCodec, and AudioLM demonstrate that audio can be mapped to compact discrete or continuous interfaces (Zeghidour et al., 2022; Défossez et al., 2023; Borsos et al., 2023). Those methods optimize reconstruction, generation, or perceptual fidelity; TAAC asks whether the interface preserves the answer to a query family. Classical experiment comparison and approximate sufficiency provide the decision-theoretic language (Le Cam, 1964), while indirect rate-distortion and information bottleneck work provide the rate-constrained view (Tishby et al., 2000; Cover & Thomas, 1991). Recent prompt-compression rate-distortion work is closest in spirit, but focuses on text prompts and black-box language models rather than audio-query families (Nagle et al., 2024). Our empirical design is also related to task-family evaluation: the key methodological claim is that benchmark averages are not adequate when compression damage is heterogeneous across query families.

J Extended Problem Setup and Operational Budgets

We study *task-aware answer-preserving compression* for large audio language models. Let $(\Omega, \mathcal{F}, \mathbb{P})$ carry a raw audio observation $X \in \mathcal{X}$, a random query $Q \in \mathcal{Q}$, and a family of correct-answer random variables $\{Y_q : q \in \mathcal{Q}\}$, where $Y_q \in \mathcal{Y}_q$. For example, X may be a 10-second audio clip, Q may be a query such as "what word was spoken?" or "how many times did the bell ring?", and Y_q denotes the correct answer associated with query q. At inference time only the realized query Q is asked, so only Y_Q is observed, while the full family $\{Y_q\}_{q\in\mathcal{Q}}$ is a bookkeeping device that lets us formalize a task family. For each query q, let $\mathcal{A}_q$ denote the action space of predicted answers, typically $\mathcal{A}_q = \mathcal{Y}_q$, and let

$$\ell : \mathcal{A}_q \times \mathcal{Y}_q \to [0, \infty) \tag{J.1}$$

be the task loss. A compressor C outputs a compressed representation Z, either from X alone or from (X, q) in the query-conditioned setting. Intuitively, for each query q we are specifying a prediction problem. The set $\mathcal{A}_q$ is the set of answers the downstream system is allowed to output, while $\mathcal{Y}_q$ is the set of possible correct answers. In most settings these coincide, so $\mathcal{A}_q = \mathcal{Y}_q$. For example, if q asks "how many times does the bell ring?", then $\mathcal{Y}_q$ may be the set of nonnegative integers; if q asks "which instrument is playing?", then $\mathcal{Y}_q$ may be a finite label set such as $\{\text{piano}, \text{violin}, \text{drums}, \ldots\}$. The loss $\ell(a, y)$ measures the penalty for predicting a when the correct answer is y, for instance the $0/1$ loss $\mathbf{1}\{a \neq y\}$. The compressor then produces a representation Z of the audio: in the query-agnostic setting $Z = C(X)$ depends only on the audio, whereas in the query-conditioned setting $Z = C(X, q)$ may adapt to the realized query and preserve only the information most relevant to answering that query.

Assumption J.1 (Bayes solvability and integrability)**.** *For every $q \in \mathcal{Q}$, Bayes-optimal decision rules based on X and on any admissible Z exist, and all displayed expectations are finite. For the familywise worst-case results we additionally assume bounded losses, $0 \leq \ell(a, y) \leq L_{\max} < \infty$, except in the Gaussian subsection where boundedness is replaced by finite second moments. Here and throughout, "familywise" means that the relevant quantity is taken uniformly over the entire query family $\mathcal{Q}$, typically through a worst-case operation such as $\sup_{q\in\mathcal{Q}}$, rather than for a single fixed query q.*

Notation summary. For a probability law π on $\mathcal{Y}_q$, define the Bayes envelope

$$\underline{L}_q(\pi) := \inf_{a\in\mathcal{A}_q} \int_{\mathcal{Y}_q} \ell(a, y)\, \pi(dy). \tag{J.2}$$

The map $\underline{L}_q$ is concave because it is the infimum of linear functionals of π. We write

$$\Pi_q^X := P_{Y_q|X,q}, \qquad \Pi_q^Z := P_{Y_q|Z,q}, \tag{J.3}$$

for the posterior laws induced by raw and compressed observations. All information quantities are measured in nats. Division by $\log 2$ converts them to bits.

For a realized query q, the Bayes-optimal risk when the decision rule sees raw audio X is

$$\mathcal{R}_X^\star(q) := \inf_{\psi:\mathcal{X}\to\mathcal{A}_q} \mathbb{E}\big[\ell(\psi(X), Y_q)\big] = \mathbb{E}\big[\underline{L}_q(\Pi_q^X)\big]. \tag{J.4}$$

Similarly, when only Z is available,

$$\mathcal{R}_Z^\star(q) := \inf_{\phi:\mathcal{Z}\to\mathcal{A}_q} \mathbb{E}\big[\ell(\phi(Z), Y_q)\big] = \mathbb{E}\big[\underline{L}_q(\Pi_q^Z)\big]. \tag{J.5}$$

The familywise excess Bayes risk is

$$\Delta_{\mathcal{Q}}(Z;X) := \sup_{q\in\mathcal{Q}} \big(\mathcal{R}_Z^\star(q) - \mathcal{R}_X^\star(q)\big). \tag{J.6}$$

This is the answer-preservation criterion of interest: Z is ε-answer-preserving over $\mathcal{Q}$ when $\Delta_{\mathcal{Q}}(Z;X) \le \varepsilon$.

We distinguish three compressor regimes. *Query-agnostic compressors* map $X \mapsto Z$ at run time. Among them, the frontiers below are *family-aware*: the compressor may depend on the fixed family $\mathcal{Q}$ and losses $\{\ell\}$ at design time, but not on the realized query q. Concretely, these include policies that compress the audio in the same way regardless of which question will later be asked, such as uniform chunk selection, energy-based selection, or a learned selector that scores chunks from audio alone. By contrast, *query-conditioned compressors* map $(X, q) \mapsto Z$ and are allowed to adapt to the realized query. For example, if the query asks what word was spoken, the compressor may preferentially retain speech-bearing regions, whereas if the query asks how many times a bell rang, it may preferentially retain temporally informative event regions. We instantiate both regimes empirically in §N.2.

We also distinguish three interface budgets. The information-theoretic rates are

$$\mathrm{Rate}_{\mathrm{info}}(Z) := I(X;Z), \qquad \mathrm{Rate}_{\mathrm{info}}^{\mathrm{cond}}(Z) := I(X;Z|Q). \tag{J.7}$$

If Z is transmitted through a token interface, let $\tau(Z)$ be its sequence length. If the interface vocabulary has cardinality B, then the operational budget associated with a B-ary code is

$$\mathrm{Rate}_{\mathrm{op}}(Z) := \log B \cdot \mathbb{E}[\tau(Z)]. \tag{J.8}$$

The information-theoretic Bayes frontier for family-aware query-agnostic compression is

$$R_{\mathrm{Bayes}}^\star(\varepsilon, \mathcal{Q}) := \inf_{C_{\mathcal{Q}}:X\mapsto Z} \Big\{ I(X;Z) : \Delta_{\mathcal{Q}}(Z;X) \le \varepsilon \Big\}, \tag{J.9}$$

and the corresponding sequence-length frontier is

$$L_{\mathrm{Bayes}}^\star(\varepsilon, \mathcal{Q}) := \inf_{C_{\mathcal{Q}}:X\mapsto Z} \Big\{ \mathbb{E}[\tau(Z)] : \Delta_{\mathcal{Q}}(Z;X) \le \varepsilon \Big\}. \tag{J.10}$$

The two budget notions play different roles. The information-theoretic rate $I(X;Z)$ measures how much information about the raw audio X is preserved in the compressed representation Z: it is zero when Z is independent of X, and increases as Z becomes more informative about X. By contrast, if Z is transmitted through a discrete token interface, then $\tau(Z)$ is the number of transmitted tokens and $\log B \cdot \mathbb{E}[\tau(Z)]$ is the corresponding operational communication budget, where B is the interface vocabulary size. Thus $I(X;Z)$ is the idealized information cost, whereas $\mathbb{E}[\tau(Z)]$ and $\log B \cdot \mathbb{E}[\tau(Z)]$ are concrete length-based surrogates for that cost.

Query-conditioned risks. When the compressor is allowed to depend on the realized query, $P_{Z|X,Q}$, the relevant Bayes risk for query q is computed under the conditional law induced by $Q = q$:

$$\mathcal{R}_{Z|q}^\star(q) := \inf_{\phi:\mathcal{Z}\to\mathcal{A}_q} \mathbb{E}\big[\ell(\phi(Z), Y_q) \mid Q = q\big]. \tag{J.11}$$

Under our standing interpretation that the runtime query variable is exogenous to the audio-generation mechanism and only selects the task, the raw-audio benchmark remains $\mathcal{R}_X^\star(q)$. We therefore define the conditioned familywise excess Bayes risk by

$$\Delta_{\mathcal{Q}}^{\mathrm{cond}}(Z;X) := \sup_{q\in\mathcal{Q}} \big(\mathcal{R}_{Z|q}^\star(q) - \mathcal{R}_X^\star(q)\big). \tag{J.12}$$

The corresponding query-conditioned information-theoretic and sequence-length frontiers are

$$R^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) := \inf_{C:(X,Q)\mapsto Z} \Big\{ I(X; Z \mid Q) : \Delta^{\mathrm{cond}}_{\mathcal{Q}}(Z; X) \le \varepsilon \Big\}, \tag{J.13}$$

and

$$L^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) := \inf_{C:(X,Q)\mapsto Z} \Big\{ \mathbb{E}[\tau(Z)] : \Delta^{\mathrm{cond}}_{\mathcal{Q}}(Z; X) \le \varepsilon \Big\}. \tag{J.14}$$

All later appearances of $R^\star_{\mathrm{Bayes,cond}}$ and $L^\star_{\mathrm{Bayes,cond}}$, including Theorem 3.4, refer to this conditioned excess-risk definition.

To model architectural restrictions of downstream large audio language models, let $\mathcal{F}_{\mathrm{LALM}}$ be a model class of answerers f with $f(z, q) \in \mathcal{A}_q$. Define the model-class Bayes risk

$$\mathcal{R}^\star_{\mathcal{F}}(q; Z) := \inf_{f \in \mathcal{F}_{\mathrm{LALM}}} \mathbb{E}\big[\ell(f(Z, q), Y_q)\big]. \tag{J.15}$$

The model-class frontiers are then

$$R^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}) := \inf_{C_{\mathcal{Q}}:X\mapsto Z} \Big\{ I(X; Z) : \sup_{q\in\mathcal{Q}} \big(\mathcal{R}^\star_{\mathcal{F}}(q; Z) - \mathcal{R}^\star_X(q)\big) \le \varepsilon \Big\}, \tag{J.16}$$

and

$$L^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}) := \inf_{C_{\mathcal{Q}}:X\mapsto Z} \Big\{ \mathbb{E}[\tau(Z)] : \sup_{q\in\mathcal{Q}} \big(\mathcal{R}^\star_{\mathcal{F}}(q; Z) - \mathcal{R}^\star_X(q)\big) \le \varepsilon \Big\}. \tag{J.17}$$

These are the core objects of the paper: $R^\star_{\mathrm{Bayes}}$ and $L^\star_{\mathrm{Bayes}}$ are information-theoretic limits for answer preservation over a family of audio-question tasks, while $R^\star_{\mathcal{F}}$ and $L^\star_{\mathcal{F}}$ are the architecture-restricted frontiers relevant to actual large audio language models.

In this paper we instantiate the theory stated in Section 3 on five real audio-question benchmarks (DCASE 2026 dev, AudioMCQ-StrongAC, MMSU, MMAR, and BigBench Audio) spanning sound-event detection, speech-centric reasoning, multi-modal audio-textual reasoning, and text-dominated control tasks. We make the following empirical contributions. First, we verify the bit-level prediction of Theorem 3.4's strict-separation construction at 66 synthetic finite-alphabet cells, matching the closed-form conditioned gain to within 10^{-9} bits across a sweep of query priors and factor entropies. Second, we demonstrate on three multi-family audio-QA datasets that the theorem-level family-wise excess risk is consistently larger than the dataset-mean metric existing work reports, by 2.5-6.8 percentage points depending on budget. Third, we verify Theorem K.4's nested-family monotonicity along cumulative chains and use chain increments to identify per-dataset bottleneck families. Fourth, we quantify factor-overlap structure in natural audio query taxonomies and show that no observed family-pair summary approaches the factor-disjoint prediction of Corollary K.12. Fifth, we extend the V1 operational conditioned-gain test with V2 three-seed replication on both Qwen2-Audio and Qwen2.5-Omni: AudioMCQ-StrongAC changes from a V1 near-null to a reproducible positive effect on both backbones, DCASE changes from a single-seed positive signal to a mixed/negative three-seed result, and MMSU on Qwen2.5-Omni reveals a large temporal-family failure. Sixth, we run a V2.1 scope-B α-sweep showing that training-target choice is itself a backbone-dependent variable, with improvements over the V2 baseline on all five Qwen2.5-Omni datasets. Finally, we report cross-answerer and interface contrasts as operational diagnostics; they do not identify the theoretical model-class gap $\Gamma_{\mathcal{F}}$.

K Proofs and Auxiliary Theory

This appendix proves the results stated in Section 3. The construction restricts comparison of experiments Le Cam (1964) to the audio-task family $\mathcal{Q}$, and connects that restricted comparison to indirect rate-distortion. The resulting theory is neither a generic sufficiency theory nor a single-task rate-distortion analysis: the central object is the familywise answer-risk gap $\Delta_{\mathcal{Q}}(Z; X)$.

K.1 Task-aware answer sufficiency and restricted experiment comparison

For two observation variables U and V, both available jointly with query q, write

$$U \succeq^{\mathrm{ans}}_{\mathcal{Q}} V \quad \Longleftrightarrow \quad \mathcal{R}^\star_U(q) \le \mathcal{R}^\star_V(q) \text{ for all } q \in \mathcal{Q}. \tag{K.1}$$

This is a $\mathcal{Q}$-restricted comparison preorder. Relative to raw audio X, define the corresponding one-sided answer deficiency

$$\delta^{\mathrm{ans}}_{\mathcal{Q}}(Z\|X) := \sup_{q\in\mathcal{Q}} \mathbb{E}\big[\underline{L}_q(\Pi^Z_q) - \mathbb{E}\big[\underline{L}_q(\Pi^X_q) \mid Z, q\big]\big]\,. \tag{K.2}$$

When $\mathcal{Q}$ is enlarged to the full class of bounded decision problems on a common answer space, the zero-deficiency relation reduces to classical comparison of experiments and approximate sufficiency in the sense of Le Cam Le Cam (1964). Here the restriction to $\mathcal{Q}$ is essential: we deliberately ignore losses that are irrelevant to the target audio-question family.

Intuitively, this construction compares two observations-typically the raw audio X and a compressed representation Z-only through the family of downstream audio tasks we care about. The relation $U \succeq^{\mathrm{ans}}_{\mathcal{Q}} V$ means that, for every query $q \in \mathcal{Q}$, the best possible predictor using U incurs no larger Bayes risk than the best possible predictor using V. Thus, our comparison is deliberately restricted to the target audio-question family $\mathcal{Q}$. The deficiency $\delta^{\mathrm{ans}}_{\mathcal{Q}}(Z\|X)$ then measures how much answer quality is lost by replacing X with Z, in the worst query in the family. For example, if $\mathcal{Q}$ contains both a speech-content query such as "what word was spoken?" and a temporal query such as "how many times did the bell ring?", then a compressed representation that preserves speech regions but blurs event timing may be nearly sufficient for the first query and clearly insufficient for the second; the familywise deficiency records exactly that worst-served task.

Definition K.1 (Exact Task-Aware Answer Sufficiency)**.** *A compressed representation Z is* exactly task-aware answer sufficient *for the query family $\mathcal{Q}$ if for every $q \in \mathcal{Q}$,*

$$\underline{L}_q(\Pi^Z_q) = \mathbb{E}\big[\underline{L}_q(\Pi^X_q) \mid Z, q\big] \qquad \textit{a.s.} \tag{K.3}$$

Definition K.2 (Approximate Task-Aware Answer Sufficiency)**.** *A compressed representation Z is* ε-approximately task-aware answer sufficient *for $\mathcal{Q}$ if*

$$\delta^{\mathrm{ans}}_{\mathcal{Q}}(Z\|X) \le \varepsilon. \tag{K.4}$$

Theorem K.3 (Restricted Sufficiency-Risk Equivalence)**.** *For every admissible compressed representation Z,*

$$\delta^{\mathrm{ans}}_{\mathcal{Q}}(Z\|X) = \Delta_{\mathcal{Q}}(Z; X). \tag{K.5}$$

Consequently, Z is exactly task-aware answer sufficient for $\mathcal{Q}$ if and only if $\Delta_{\mathcal{Q}}(Z; X) = 0$, and Z is ε-approximately task-aware answer sufficient if and only if $\Delta_{\mathcal{Q}}(Z; X) \le \varepsilon$.

Proof. For each fixed q,

$$\mathcal{R}^\star_Z(q) - \mathcal{R}^\star_X(q) = \mathbb{E}\big[\underline{L}_q(\Pi^Z_q)\big] - \mathbb{E}\big[\underline{L}_q(\Pi^X_q)\big]. \tag{K.6}$$

Applying the tower property to the second term gives

$$\mathcal{R}^\star_Z(q) - \mathcal{R}^\star_X(q) = \mathbb{E}\big[\underline{L}_q(\Pi^Z_q) - \mathbb{E}\big[\underline{L}_q(\Pi^X_q) \mid Z, q\big]\big]\,. \tag{K.7}$$

Taking the supremum over $q \in \mathcal{Q}$ proves the identity. Nonnegativity follows from concavity of $\underline{L}_q$ and

$$\Pi^Z_q = \mathbb{E}[\Pi^X_q \mid Z, q]. \tag{K.8}$$

Indeed, $\underline{L}_q(\Pi^Z_q) \ge \mathbb{E}[\underline{L}_q(\Pi^X_q) \mid Z, q]$ almost surely. Therefore zero deficiency is equivalent to zero Jensen gap for every query, which is exactly the almost-sure equality in Definition K.1; the ε-approximate statement follows from the same identity. □

Interpretation. The restricted deficiency is a Jensen gap of the Bayes envelope generated by coarsening X into Z. Exact answer sufficiency is therefore a *local* notion: it does not merely require equality of global risks, but equality of the conditional Bayes envelope after observing Z.

Experimental implication. Paired raw and compressed predictions estimate fixed-answerer excess risk. They do not estimate Bayes risk or certify information preservation; held-out operational certification requires the selection and testing procedure in Section 4.

Theorem K.4 (Monotonicity under Query-Family Refinement)**.** *If $\mathcal{Q}_1 \subseteq \mathcal{Q}_2$, then for every Z,*

$$\delta^{\mathrm{ans}}_{\mathcal{Q}_1}(Z\|X) \le \delta^{\mathrm{ans}}_{\mathcal{Q}_2}(Z\|X), \qquad \Delta_{\mathcal{Q}_1}(Z; X) \le \Delta_{\mathcal{Q}_2}(Z; X). \tag{K.9}$$

Hence, for every $\varepsilon \ge 0$,

$$R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}_1) \le R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}_2), \qquad L^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}_1) \le L^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}_2), \tag{K.10}$$

and likewise

$$R^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}_1) \le R^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}_2), \qquad L^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}_1) \le L^\star_{\mathcal{F}}(\varepsilon, \mathcal{Q}_2). \tag{K.11}$$

Proof. All quantities are defined by a supremum over $q \in \mathcal{Q}$ or an infimum over compressors satisfying such supremum constraints. Enlarging $\mathcal{Q}$ can only tighten the preservation constraints, hence it can only shrink the feasible compressor set and weakly increase the required budget. □

Interpretation. Adding queries can only make answer preservation harder. This creates a partial order over task families: nested audio task families must yield nested rate and length frontiers.

Experimental implication. Later experiments should report nested families $\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \cdots$ and verify that the empirical rate-risk and token-risk curves are ordered accordingly. Violations indicate either estimation noise or optimization failure.

K.2 Task-aware indirect rate-distortion bounds

For a fixed query q, define the reduced distortion

$$\rho_q(x, a) := \mathbb{E}[\ell(a, Y_q) \mid X = x, q] - \underline{L}_q(\Pi_q^X(x)). \tag{K.12}$$

By construction, $\rho_q(x, a) \geq 0$, and $\rho_q(x, a) = 0$ exactly when action a is Bayes-optimal given $X = x$.

Now define the single-query indirect answer-preserving function

$$R_q^{\mathrm{ind}}(\varepsilon) := \inf_{P_{A_q|X}} \Big\{ I(X; A_q) : \mathbb{E}\big[\rho_q(X, A_q)\big] \leq \varepsilon \Big\}. \tag{K.13}$$

By the reduction principle for indirect rate-distortion, R_q^{ind} is the ordinary rate-distortion function of X under the reduced distortion ρ_q.

Proposition K.5 (Task-Aware Indirect Rate-Distortion Sandwich)**.** *For every $\varepsilon \geq 0$, the query-agnostic frontier satisfies*

$$\sup_{q \in \mathcal{Q}} R_q^{\mathrm{ind}}(\varepsilon) \leq R^{\star}_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) \leq \inf_{\substack{P_{U|X},\, \{g_q\} \\ \mathbb{E}[\rho_q(X, g_q(U))] \leq \varepsilon,\ \forall q}} I(X; U). \tag{K.14}$$

For the query-conditioned frontier,

$$\mathbb{E}_Q\big[R_Q^{\mathrm{ind}}(\varepsilon)\big] \leq R^{\star}_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \leq \inf_{\substack{P_{U|X,Q},\, \{g_q\} \\ \mathbb{E}[\rho_q(X, g_q(U)) | Q = q] \leq \varepsilon,\ \forall q}} I(X; U \mid Q). \tag{K.15}$$

The analogous inequalities hold for the sequence-length frontiers after replacing the mutual-information objectives by $\mathbb{E}[\tau(U)]$.

Proof. For the query-agnostic converse, let $X \mapsto Z$ be any feasible universal compressor, and let $A_q = a_q^{\star}(Z)$ be a Bayes-optimal query-q action based on Z. Since A_q is a measurable function of Z,

$$I(X; A_q) \leq I(X; Z) \tag{K.16}$$

by data processing. Moreover,

$$\mathbb{E}[\rho_q(X, A_q)] = \mathcal{R}_Z^{\star}(q) - \mathcal{R}_X^{\star}(q) \leq \Delta_{\mathcal{Q}}(Z; X) \leq \varepsilon. \tag{K.17}$$

Hence $I(X; Z) \geq R_q^{\mathrm{ind}}(\varepsilon)$ for every q, and therefore

$$I(X; Z) \geq \sup_{q \in \mathcal{Q}} R_q^{\mathrm{ind}}(\varepsilon). \tag{K.18}$$

Taking the infimum over feasible Z proves the lower bound.

For the query-agnostic achievability bound, choose any auxiliary U and query-specific decoders $\{g_q\}$ satisfying $\mathbb{E}[\rho_q(X, g_q(U))] \leq \varepsilon$ for all q, and set $Z := U$. Then

$$\mathcal{R}_Z^{\star}(q) - \mathcal{R}_X^{\star}(q) \leq \mathbb{E}[\rho_q(X, g_q(U))] \leq \varepsilon \tag{K.19}$$

for every q, so Z is feasible with rate $I(X; U)$.

For the query-conditioned converse, let $P_{Z|X,Q}$ be any feasible compressor. For each q, let $A_q = a_q^\star(Z)$ be a Bayes-optimal action under the conditional law $Q = q$. Then

$$I(X; A_q \mid Q = q) \le I(X; Z \mid Q = q), \tag{K.20}$$

and

$$\mathbb{E}[\rho_q(X, A_q) \mid Q = q] = \mathcal{R}^\star_{Z|q}(q) - \mathcal{R}^\star_X(q) \le \Delta^{\mathrm{cond}}_{\mathcal{Q}}(Z; X) \le \varepsilon. \tag{K.21}$$

Thus $I(X; Z \mid Q = q) \ge R^{\mathrm{ind}}_q(\varepsilon)$ for every q. Averaging over Q yields

$$I(X; Z \mid Q) = \mathbb{E}_Q\big[I(X; Z \mid Q = q)\big] \ge \mathbb{E}_Q\big[R^{\mathrm{ind}}_Q(\varepsilon)\big]. \tag{K.22}$$

For the query-conditioned achievability bound, choose any auxiliary U generated by a conditional kernel $P_{U|X,Q}$ and query-specific decoders $\{g_q\}$ such that

$$\mathbb{E}[\rho_q(X, g_q(U)) \mid Q = q] \le \varepsilon \qquad \text{for all } q \in \mathcal{Q}. \tag{K.23}$$

Setting $Z := U$, we obtain for every q,

$$\mathcal{R}^\star_{Z|q}(q) - \mathcal{R}^\star_X(q) \le \mathbb{E}[\rho_q(X, g_q(U)) \mid Q = q] \le \varepsilon, \tag{K.24}$$

so Z is feasible with conditional rate $I(X; U \mid Q)$. □

Interpretation. The query-agnostic frontier is controlled from below by the hardest single query, because one universal interface must support every query at once. By contrast, the query-conditioned frontier averages the per-query costs, because the compressor can adapt its representation to the realized query before transmitting anything.

Experimental implication. Later experiments should compare a universal compressor against a query-conditioned compressor on the same task family. The conditioned gain should track the gap between a hardest-query requirement and the average per-query requirement, especially when different audio queries depend on different latent factors.

Remark K.6 (Reduction to classical objects). If $\mathcal{Q} = \{q\}$ and the action a is itself a predictive distribution on $\mathcal{Y}_q$ under logarithmic loss, $\ell(a, y) = -\log a(y)$, then $\underline{L}_q(\pi) = H(\pi)$ and

$$\mathcal{R}^\star_Z(q) - \mathcal{R}^\star_X(q) = H(Y_q|Z, q) - H(Y_q|X, q). \tag{K.25}$$

Hence $R^\star_{\mathrm{Bayes}}(\varepsilon, \{q\})$ is equivalent to an information-bottleneck trade-off Tishby et al. (2000). For discrete prompts and hard compressed prompts, this is related to black-box prompt-compression rate-distortion formulations Nagle et al. (2024); the present Bayes-optimal answer rule differs from a fixed language-model decoder. By contrast, for general audio task families, $\delta^{\mathrm{ans}}_{\mathcal{Q}}$ is a supremum of Bayes-envelope gaps, so no single mutual-information distortion summarizes the entire frontier. In the single-task task-oriented source-coding regime, identifiability can collapse the remote problem to ordinary rate-distortion Enttsel & Corlay (2026); our familywise objective generally prevents that collapse.

K.3 Proof of Theorem 3.3

Proof. Write

$$d_q(Z) = \mathcal{R}^\star_Z(q) - \mathcal{R}^\star_X(q), \qquad d(q) = d_q(Z). \tag{K.26}$$

For any coarse cell $F \in \mathcal{P}$ with $\mu(F) > 0$, refinement gives

$$\bar{d}_F(Z) = \sum_{\substack{G \in \mathcal{P}' : G \subseteq F \\ \mu(G) > 0}} \frac{\mu(G)}{\mu(F)}\, \bar{d}_G(Z), \tag{K.27}$$

so $\bar{d}_F(Z)$ is a convex combination of the fine-cell means inside F. Hence

$$\bar{d}_F(Z) \le \max_{\substack{G \in \mathcal{P}' : G \subseteq F \\ \mu(G) > 0}} \bar{d}_G(Z) \le \Delta^{\mathrm{fam}}_{\mathcal{P}'}(Z; X). \tag{K.28}$$

Taking the maximum over $F \in \mathcal{P}$ gives

$$\Delta^{\mathrm{fam}}_{\mathcal{P}}(Z;X) \le \Delta^{\mathrm{fam}}_{\mathcal{P}'}(Z;X). \tag{K.29}$$

Also,

$$\Delta^{\mathrm{avg}}_{\mu}(Z;X) = \sum_{F\in\mathcal{P}:\mu(F)>0} \mu(F)\bar{d}_F(Z) \le \max_{F\in\mathcal{P}:\mu(F)>0} \bar{d}_F(Z) = \Delta^{\mathrm{fam}}_{\mathcal{P}}(Z;X), \tag{K.30}$$

and for every fine cell G,

$$\bar{d}_G(Z) = \mathbb{E}_\mu[d(Q) \mid Q \in G] \le \sup_{q\in\mathcal{Q}} d_q(Z) = \Delta_{\mathcal{Q}}(Z;X). \tag{K.31}$$

Therefore

$$\Delta^{\mathrm{avg}}_{\mu}(Z;X) \le \Delta^{\mathrm{fam}}_{\mathcal{P}}(Z;X) \le \Delta^{\mathrm{fam}}_{\mathcal{P}'}(Z;X) \le \Delta_{\mathcal{Q}}(Z;X). \tag{K.32}$$

For the budget consequence, suppose $b \in \mathcal{B}_{\mathcal{P}'}(\varepsilon;C)$. Then

$$\Delta^{\mathrm{fam}}_{\mathcal{P}'}(Z_b;X) \le \varepsilon. \tag{K.33}$$

By the inequality just proved,

$$\Delta^{\mathrm{fam}}_{\mathcal{P}}(Z_b;X) \le \Delta^{\mathrm{fam}}_{\mathcal{P}'}(Z_b;X) \le \varepsilon, \tag{K.34}$$

so $b \in \mathcal{B}_{\mathcal{P}}(\varepsilon;C)$. Thus

$$\mathcal{B}_{\mathcal{P}'}(\varepsilon;C) \subseteq \mathcal{B}_{\mathcal{P}}(\varepsilon;C). \tag{K.35}$$

Taking infima, with the convention $\inf\emptyset = +\infty$, gives

$$b^\star_{\mathcal{P}}(\varepsilon;C) = \inf \mathcal{B}_{\mathcal{P}}(\varepsilon;C) \le \inf \mathcal{B}_{\mathcal{P}'}(\varepsilon;C) = b^\star_{\mathcal{P}'}(\varepsilon;C). \tag{K.36}$$

□

K.4 Model-class restriction and interface length

Proposition K.7 (Model-Class Gap)**.** *Define the architecture gap of a compressed representation Z relative to $\mathcal{F}_{\mathrm{LALM}}$ by*

$$\Gamma_{\mathcal{F}}(Z;\mathcal{Q}) := \sup_{q\in\mathcal{Q}} \big(\mathcal{R}^\star_{\mathcal{F}}(q;Z) - \mathcal{R}^\star_Z(q)\big). \tag{K.37}$$

Then

$$\sup_{q\in\mathcal{Q}} \big(\mathcal{R}^\star_{\mathcal{F}}(q;Z) - \mathcal{R}^\star_X(q)\big) \le \Delta_{\mathcal{Q}}(Z;X) + \Gamma_{\mathcal{F}}(Z;\mathcal{Q}). \tag{K.38}$$

Consequently,

$$R^\star_{\mathcal{F}}(\varepsilon,\mathcal{Q}) \ge R^\star_{\mathrm{Bayes}}(\varepsilon,\mathcal{Q}), \qquad L^\star_{\mathcal{F}}(\varepsilon,\mathcal{Q}) \ge L^\star_{\mathrm{Bayes}}(\varepsilon,\mathcal{Q}). \tag{K.39}$$

More quantitatively, any compressor satisfying $\Delta_{\mathcal{Q}}(Z;X) \le \varepsilon$ and $\Gamma_{\mathcal{F}}(Z;\mathcal{Q}) \le \gamma$ is feasible for the model-class frontier at tolerance $\varepsilon+\gamma$.

Proof. For each q,

$$\mathcal{R}^\star_{\mathcal{F}}(q;Z) - \mathcal{R}^\star_X(q) = \big(\mathcal{R}^\star_Z(q) - \mathcal{R}^\star_X(q)\big) + \big(\mathcal{R}^\star_{\mathcal{F}}(q;Z) - \mathcal{R}^\star_Z(q)\big). \tag{K.40}$$

Taking the supremum over $q \in \mathcal{Q}$ yields the first inequality. The frontier inequalities follow immediately. □

Interpretation. The proposition bounds model-class excess risk relative to the raw-audio Bayes benchmark by two theoretical terms. Neither term is identified by comparing raw and compressed predictions of a fixed answerer.

Experimental implication. Comparing stronger or specialized answerers on the same compressed interface Z can reveal usable retained information. Such comparisons do not, without a Bayes-risk reference, identify $\Gamma_{\mathcal{F}}$ or fully separate information loss from answerer limitations.

Proposition K.8 (Rate-to-Token Translation)**.** *Let Z be transmitted through a uniquely decodable B-ary token interface, and let $\tau(Z)$ denote the number of interface symbols. Then the lossless source-coding converse implies*

$$H(Z) \le \log B \cdot \mathbb{E}[\tau(Z)]. \tag{K.41}$$

Consequently,

$$I(X;Z) \le H(Z) \le \log B \cdot \mathbb{E}[\tau(Z)]. \tag{K.42}$$

Now let U be a discrete compressed symbol and suppose the interface code may be chosen to match the law of U. Then there exists a uniquely decodable B-ary code c and an induced interface random variable $Z := c(U)$ such that

$$\mathbb{E}[\tau(Z)] < \frac{H(U)}{\log B} + 1. \tag{K.43}$$

Therefore,

$$\frac{H(U)}{\log B} \le \inf_c \mathbb{E}[\tau(c(U))] < \frac{H(U)}{\log B} + 1, \tag{K.44}$$

where the infimum is over uniquely decodable B-ary codes for U. If the compressor is deterministic, $U = g(X)$, then $H(U) = I(X;U)$, and hence

$$\frac{I(X;U)}{\log B} \le \inf_c \mathbb{E}[\tau(c(U))] < \frac{I(X;U)}{\log B} + 1. \tag{K.45}$$

If each interface token is instead a quantized latent carrying b_{tok} bits, replace $\log B$ by $b_{\mathrm{tok}} \log 2$.

Proof. For any uniquely decodable B-ary interface code, the Kraft inequality together with the converse of lossless source coding gives

$$H(Z) \le \log B \cdot \mathbb{E}[\tau(Z)]. \tag{K.46}$$

Since $I(X;Z) \le H(Z)$, the first claim follows immediately.

Now let U be any discrete compressed symbol. A Shannon B-ary code for U yields a uniquely decodable encoding c satisfying

$$\mathbb{E}[\tau(c(U))] < \frac{H(U)}{\log B} + 1 \tag{K.47}$$

by the standard one-symbol source-coding bound Cover & Thomas (1991). The converse lower bound

$$\frac{H(U)}{\log B} \le \inf_c \mathbb{E}[\tau(c(U))] \tag{K.48}$$

follows from the first part applied to $Z = c(U)$. If $U = g(X)$ is deterministic, then $H(U \mid X) = 0$, so $H(U) = I(X;U)$, yielding the final display. The quantized-latent case is identical after interpreting one token as b_{tok} bits. □

Interpretation. Information rate and interface length are linked but not identical. The lower bound $H(Z) \le \log B\, \mathbb{E}[\tau(Z)]$ is universal for any uniquely decodable interface. Near-equality, however, requires an entropy-efficient coding layer or tokenizer matched to the compressed representation. For stochastic compressors the gap between $H(U)$ and $I(X;U)$ is exactly the randomized-encoding overhead $H(U \mid X)$.

Experimental implication. Later experiments should report both rate-risk and sequence-length-risk curves. When the interface is close to entropy-efficient, the horizontal axes should differ mainly by the factor $\log B$, up to the one-token Shannon slack. Persistent deviations beyond that level diagnose tokenizer inefficiency or stochastic-encoding overhead.

Theorem K.9 (Conditioned Compression Advantage)**.** *For any design prior $Q \sim \mu$ with support $\mathcal{Q}$, independent of X,*

$$R^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \le R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}), \qquad L^\star_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \le L^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}). \tag{K.49}$$

The inequalities can be strict. In particular, let $\mathcal{Q} = \{q_1, q_2\}$ with $P(Q = q_1) = \lambda \in (0, 1)$, let the audio contain two independent answer factors and irrelevant side information, $X = (V_1, V_2, W)$, with $H(V_1), H(V_2) > 0$, let $Y_{q_i} = V_i$, and let $\ell_{q_i}(a, Y_{q_i}) = \mathbf{1}\{a \neq Y_{q_i}\}$. Then

$$R^\star_{\mathrm{Bayes}}(0, \{q_1, q_2\}) = H(V_1, V_2), \tag{K.50}$$

while

$$R^\star_{\mathrm{Bayes,cond}}(0, \{q_1, q_2\}) = \lambda H(V_1) + (1 - \lambda) H(V_2). \tag{K.51}$$

Thus the strict rate gap is $(1 - \lambda)H(V_1) + \lambda H(V_2) > 0$.

Proof. The non-strict inequalities are immediate: a query-conditioned compressor can always ignore q and emulate any query-agnostic one.

For the strict-separation construction, query-agnostic zero-risk preservation requires measurable decoders g_1, g_2 such that $g_i(Z) = V_i$ almost surely. Therefore (V_1, V_2) is a function of Z, so

$$I(X; Z) \ge I(V_1, V_2; Z) = H(V_1, V_2). \tag{K.52}$$

Equality is achieved by choosing $Z = (V_1, V_2)$, which discards the irrelevant side information W. In the query-conditioned setting, exact preservation only requires transmitting V_Q. Choosing $Z = V_Q$ gives conditional rate

$$I(X; Z \mid Q) = H(V_Q \mid Q) = \lambda H(V_1) + (1 - \lambda) H(V_2). \tag{K.53}$$

No lower conditional rate can suffice: conditioned on $Q = q_i$, exact recovery of V_i implies $H(V_i \mid Z, Q = q_i) = 0$, and hence $I(X; Z \mid Q = q_i) \ge I(V_i; Z \mid Q = q_i) = H(V_i)$. Averaging over Q proves the lower bound and the claimed strict gap. □

Interpretation. The theorem formalizes the intuitive advantage of query conditioning: a universal query-agnostic interface must preserve the union of all answer-relevant factors, whereas a query-conditioned interface needs only preserve the factors relevant to the realized query.

Experimental implication. A primary empirical quantity is the conditioned gain. We distinguish two estimators that share the theoretical anchor of Theorem 3.4 but live in different operational units:

$$\widehat{G}^{\mathrm{info}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q}) := \widehat{R}^{\star,\mathrm{info}}_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) - \widehat{R}^{\star,\mathrm{info}}_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}), \tag{K.54}$$

the rate-theoretic conditioned gain computed from $I(X; Z)$ or $I(X; Z \mid Q)$ estimates, and

$$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q}) := \widehat{R}^\star_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \mathrm{agnostic}) - \widehat{R}^\star_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \mathrm{conditioned}), \tag{K.55}$$

the operational conditioned gain computed from nominal budget fractions $b \in [0, 1]$. The two are linked by Proposition K.8's rate-to-token translation, modulo the coding slack and interface mismatch described there. In the two-factor construction, the conditioned gain is linear in the query prior; a balanced prior gives a symmetric example, not generally the maximum; the strict-separation construction of Theorem 3.4 predicts $\widehat{G}^{\mathrm{info}}_{\mathrm{cond}}(0) = (1 - \lambda)H(V_1) + \lambda H(V_2)$ for the two-factor finite-alphabet example with $P(Q = q_1) = \lambda$.

K.5 Sharp finite-alphabet and Gaussian special cases

Proposition K.10 (Finite-Alphabet Exact Frontier)**.** *Assume that $\mathcal{Q} = \{q_1, \ldots, q_m\}$ is finite, $\mathcal{X}$ is finite, and each action space $\mathcal{A}_{q_i}$ is finite. Let*

$$\mathcal{U} := \prod_{i=1}^{m} \mathcal{A}_{q_i}, \qquad u = (u_1, \ldots, u_m) \in \mathcal{U}. \tag{K.56}$$

Then the Bayes frontier admits the exact convex characterization

$$R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) = \min_{P_{U \mid X}} I(X; U) \tag{K.57}$$

subject to

$$\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\sum_{y\in\mathcal{Y}_{q_i}} P_X(x)P_{U|X}(u|x)P_{Y_{q_i}|X,q_i}(y|x,q_i)\,\ell_{q_i}(u_i,y) - \mathcal{R}^\star_X(q_i) \le \varepsilon, \quad i=1,\dots,m. \tag{K.58}$$

If a finite interface dictionary assigns a fixed token length $\tau(u)$ to each $u\in\mathcal{U}$, the corresponding sequence-length frontier is the linear program obtained by replacing the objective with

$$\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} P_X(x)P_{U|X}(u|x)\tau(u). \tag{K.59}$$

Proof. Take any query-agnostic compressor $X\mapsto Z$ with query-specific decoders $g_i:\mathcal{Z}\to\mathcal{A}_{q_i}$. Define

$$U := \big(g_1(Z),\dots,g_m(Z)\big)\in\mathcal{U}. \tag{K.60}$$

Then U is a function of Z, so $I(X;U)\le I(X;Z)$, and the risk on query q_i is exactly the risk of using action U_i. Hence no optimal solution is lost by optimizing directly over $P_{U|X}$. Conversely, any $U\in\mathcal{U}$ can itself serve as the compressed representation. The displayed constraints are therefore exact, and they are linear in $P_{U|X}$. The mutual-information objective gives a finite convex program. □

Interpretation. For a finite task family, universal answer-preserving compression can be reduced, without loss, to compressing a *tuple of query-contingent actions*. The exponential growth of $|\mathcal{U}|$ with $|\mathcal{Q}|$ is not an artifact of the proof; it is the exact combinatorial price of simultaneously supporting many queries.

Experimental implication. This proposition yields an exact $\widehat{R}^\star_{\mathrm{Bayes}}$ benchmark on synthetic finite-query tasks. It parallels the exact finite optimization viewpoint recently used in prompt compression Nagle et al. (2024), except that the reproduction alphabet here is the vector of all query-contingent answers rather than a shortened prompt.

Common latent-answer convention in the Gaussian subsection. In the Gaussian special case, all queries refer to the same latent audio-factor vector $S\in\mathbb{R}^m$. Equivalently, $Y_q := S$, $\mathcal{Y}_q=\mathcal{A}_q=\mathbb{R}^m$ for every $q\in\mathcal{Q}$, and the query dependence enters only through the loss ℓ. This models the situation in which lexical, speaker, prosodic, event, or temporal queries emphasize different coordinates of a common underlying factorization.

Proposition K.11 (Gaussian Latent Allocation)**.** *Let*

$$S=(S_1,\dots,S_m), \qquad S_j \overset{\text{ind}}{\sim} \mathcal{N}(0,\sigma_j^2), \tag{K.61}$$

and assume the raw observation is $X=S$. For every query $q\in\mathcal{Q}$, set $Y_q:=S$, $\mathcal{Y}_q=\mathcal{A}_q=\mathbb{R}^m$, and define the weighted quadratic loss

$$\ell(a,S)=\sum_{j=1}^m \alpha_{qj}(a_j-S_j)^2, \qquad \alpha_{qj}\ge 0. \tag{K.62}$$

Then $\mathcal{R}^\star_X(q)=0$ for all q, and, writing $\log^+ t := \max\{0,\log t\}$,

$$R^\star_{\mathrm{Bayes}}(\varepsilon,\mathcal{Q}) = \min_{\substack{0\le d_j\le\sigma_j^2\\ \sum_{j=1}^m \alpha_{qj}d_j\le\varepsilon,\ \forall q\in\mathcal{Q}}} \frac{1}{2}\sum_{j=1}^m \log^+\frac{\sigma_j^2}{d_j}. \tag{K.63}$$

For the query-conditioned frontier,

$$R^\star_{\mathrm{Bayes,cond}}(\varepsilon,\mathcal{Q}) = \mathbb{E}_Q\left[\min_{\substack{0\le d_j(Q)\le\sigma_j^2\\ \sum_{j=1}^m \alpha_{Qj}d_j(Q)\le\varepsilon}} \frac{1}{2}\sum_{j=1}^m \log^+\frac{\sigma_j^2}{d_j(Q)}\right]. \tag{K.64}$$

For a fixed query q, the optimizer is weighted reverse water-filling:

$$d_j^\star(q)=\begin{cases}\sigma_j^2, & \alpha_{qj}=0,\\ \min\{\sigma_j^2,\nu_q/\alpha_{qj}\}, & \alpha_{qj}>0,\end{cases} \tag{K.65}$$

where $\nu_q\ge 0$ is chosen so that $\sum_{j=1}^m \alpha_{qj}d_j^\star(q)=\varepsilon$ whenever the constraint is active.

Proof. Because $X = S$, the raw observation reveals the common latent answer object exactly, hence $\mathcal{R}^\star_X(q) = 0$ for every q. Any compressed representation of S with per-coordinate mean-squared errors $\{d_j\}_{j=1}^m$ incurs query-q Bayes risk

$$\mathcal{R}^\star_Z(q) = \sum_{j=1}^m \alpha_{qj} d_j. \tag{K.66}$$

Thus the familywise constraint $\Delta_{\mathcal{Q}}(Z; X) \le \varepsilon$ is equivalent to

$$\sum_{j=1}^m \alpha_{qj} d_j \le \varepsilon \qquad \text{for all } q \in \mathcal{Q}. \tag{K.67}$$

For independent Gaussian coordinates and additive separable quadratic distortion, the rate-distortion theorem gives the minimal rate

$$\frac{1}{2} \sum_{j=1}^m \log^+ \frac{\sigma_j^2}{d_j} \tag{K.68}$$

for any admissible distortion allocation $\{d_j\}$ Cover & Thomas (1991). Minimizing over all allocations satisfying the familywise constraints proves the first display.

For the query-conditioned frontier, the compressor may choose a separate distortion allocation after observing the realized query Q. Under the conditioned excess-risk definition, the admissible allocations are exactly those satisfying

$$\sum_{j=1}^m \alpha_{Qj} d_j(Q) \le \varepsilon \tag{K.69}$$

for the realized Q. Averaging the per-query optimal Gaussian rates over Q yields the second display. The reverse-water-filling form follows from the KKT conditions of the per-query constrained minimization. □

Interpretation. The proposition treats the latent audio factors themselves as the common answer object, while different queries weight these factors differently. Transcript questions may heavily weight lexical coordinates, speaker questions identity coordinates, and event questions background-sound coordinates. The optimal compressor therefore allocates rate to factors, not to waveform regions per se.

Experimental implication. Later experiments should fit probe factors S_j from raw audio, estimate their variances σ_j^2, and infer query-specific weights α_{qj}. The observed budget allocations should then be compared against the predicted reverse-water-filling solutions $d_j^\star(q)$.

Corollary K.12 (Factor Relevance Decomposition)**.** *Under Proposition K.11, suppose the factor indices partition as*

$$\{1, \dots, m\} = B_1 \cup \cdots \cup B_r, \tag{K.70}$$

and the query family partitions as

$$\mathcal{Q} = \mathcal{Q}_1 \cup \cdots \cup \mathcal{Q}_r \tag{K.71}$$

with the property that $\alpha_{qj} = 0$ whenever $q \in \mathcal{Q}_\ell$ and $j \notin B_\ell$. Then

$$R^\star_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) = \sum_{\ell=1}^r \min_{\substack{0 \le d_j \le \sigma_j^2,\ j \in B_\ell \\ \sum_{j \in B_\ell} \alpha_{qj} d_j \le \varepsilon,\ \forall q \in \mathcal{Q}_\ell}} \frac{1}{2} \sum_{j \in B_\ell} \log^+ \frac{\sigma_j^2}{d_j}. \tag{K.72}$$

Thus disjoint query subfamilies acting on disjoint latent-factor blocks add in rate.

Proof. Under the stated block-separation condition, each constraint in Proposition K.11 involves only the distortions $\{d_j : j \in B_\ell\}$ for the matching block. The Gaussian rate objective is additive across independent coordinates. The feasible set is therefore a Cartesian product of blockwise feasible sets, and minimizing the additive objective over that product decomposes into the displayed sum. □

Interpretation. When one query block cares only about lexical content and another only about speaker identity, the universal answer-preserving rate splits into independent additive contributions. This is a precise version of the idea that unrelated audio capabilities consume separate budgets.

Experimental implication. Later experiments should report factor-group budget tables by clustering queries according to approximately disjoint relevance patterns. The corollary predicts additive rate contributions across such clusters.

Corollary K.13 (Synergy Penalty for Joint-Factor Queries)**.** *Under Proposition K.11, take $m = 2$ with $S_1, S_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Let q_1 emphasize factor 1, q_2 emphasize factor 2, and q_{12} emphasize both jointly, via*

$$\alpha_{q_1} = (1, 0), \qquad \alpha_{q_2} = (0, 1), \qquad \alpha_{q_{12}} = (1, 1). \tag{K.73}$$

Then for $0 < \varepsilon < \sigma^2$,

$$R^\star_{\text{Bayes}}(\varepsilon, \{q_1, q_2\}) = \log \frac{\sigma^2}{\varepsilon}, \tag{K.74}$$

whereas

$$R^\star_{\text{Bayes}}(\varepsilon, \{q_1, q_2, q_{12}\}) = \log \frac{2\sigma^2}{\varepsilon}. \tag{K.75}$$

Hence the joint-factor query incurs the nonnegative synergy penalty

$$\text{Syn}_{12}(\varepsilon) := R^\star_{\text{Bayes}}(\varepsilon, \{q_1, q_2, q_{12}\}) - R^\star_{\text{Bayes}}(\varepsilon, \{q_1, q_2\}) = \log 2, \tag{K.76}$$

namely one bit.

Proof. For $\{q_1, q_2\}$, the constraints are $d_1 \le \varepsilon$ and $d_2 \le \varepsilon$, so the optimal allocation is $d_1 = d_2 = \varepsilon$, giving $\log(\sigma^2/\varepsilon)$. Adding q_{12} imposes $d_1 + d_2 \le \varepsilon$. By symmetry and convexity of the Gaussian rate objective, the optimum is $d_1 = d_2 = \varepsilon/2$, giving $\log(2\sigma^2/\varepsilon)$. Subtracting the two rates yields $\log 2$. □

Interpretation. Supporting two single-factor queries does not automatically support their joint query at the same tolerance. In audio terms, preserving "what was said" and preserving "who said it" separately is easier than preserving the joint answer to "who said that exact phrase?"

Experimental implication. Later experiments should compare separate task families, such as transcript-only and speaker-only, against joint families that require simultaneous access to both factors. A systematic horizontal shift between these curves is the empirical synergy penalty predicted above.

K.6 Experiment-facing consequences

The theory above prescribes concrete empirical quantities. First, one should estimate a family-level surrogate for the theorem object $\Delta_{\mathcal{Q}}(Z; X)$. Let $\mathcal{F}(\mathcal{Q})$ denote the partition of the query family $\mathcal{Q}$ into labeled subfamilies, and let h be the frozen answerer. Define

$$\begin{aligned}\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b) := \max_{F \in \mathcal{F}(\mathcal{Q})} \Bigg[\frac{1}{N_F} \sum_{(x,q,y):\, \text{family}(q) = F} \ell\big(h(Z_b(x), q), y\big) \\ - \frac{1}{N_F} \sum_{(x,q,y):\, \text{family}(q) = F} \ell\big(h(X, q), y\big) \Bigg]\end{aligned} \tag{K.77}$$

as the worst-family empirical excess risk at operational budget b, and

$$\widehat{\Delta}_{\text{avg}}(b) := \frac{1}{N} \sum_{(x,q,y) \in \text{eval}} \Big[\ell\big(h(Z_b(x), q), y\big) - \ell\big(h(X, q), y\big) \Big] \tag{K.78}$$

as the dataset-mean empirical excess risk. The corresponding population family-level quantity is a coarsening of the worst-query fixed-answerer excess risk, not an estimate of the Bayes quantity $\Delta_{\mathcal{Q}}(Z; X)$, and the empirical family maximum is an upper bound on the empirical dataset mean

by $\max \geq \text{mean}$ on the declared partition. The gap $\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b) - \widehat{\Delta}_{\text{avg}}(b)$ is itself an empirical signal of worst-family damage that dataset-mean metrics do not capture. On multi-family datasets we report both quantities with bootstrap CIs; on single-family datasets, the two estimators coincide by construction.

Second, on synthetic finite-query tasks, Proposition K.10 yields an exact plug-in benchmark $\widehat{R}^{\star}_{\text{Bayes}}$; on Gaussian factor tasks, Proposition K.11 yields a fitted analytic proxy. Third, one should estimate both $\widehat{G}^{\text{info}}_{\text{cond}}(\varepsilon, \mathcal{Q})$ on exact synthetic constructions and $\widehat{G}^{\text{op}}_{\text{cond}}(\varepsilon, \mathcal{Q})$ on real audio. Fourth, one should estimate the model-class gap

$$\widehat{\Gamma}_{\mathcal{F}}(Z; \mathcal{Q}) = \max_{q \in \mathcal{Q}} \big(\widehat{\mathcal{R}}^{\star}_{\mathcal{F}}(q; Z) - \widehat{\mathcal{R}}^{\star}_{Z}(q)\big) \tag{K.79}$$

only with justified estimates of both the model-class and unrestricted Bayes risks on the same interface. The learned-native and heuristic-baseline summaries in §P.5 instead report operational contrasts and do not estimate this gap. V1 reports operational frontiers $\widehat{R}^{\star}_{\mathcal{F}}$ against a single backbone and treats the observed frontiers as architecture-specific rather than attempting a Bayes-versus-architecture decomposition. Fifth, one should report the monotonicity increments across nested query families and the factor allocations $\widehat{d}^{\star}_{j}(q)$ or water levels $\widehat{\nu}_q$ when latent-factor probes are available. These are exactly the quantities later needed for empirical rate-risk curves, empirical sequence-length-risk curves, plug-in proxies for $\widehat{R}^{\star}_{\text{Bayes}}$, the gap between query-conditioned and query-agnostic compression, monotonicity across nested query families, and factor-specific budget differences.

L Estimator Details

This section specifies the empirical estimators used throughout the paper. Two design choices are global. First, excess risk is measured at the level of labeled query families rather than individual queries, because every query in V1 appears only once. Second, operational frontiers are reported on two axes: a token axis for single-cell comparisons and a nominal-budget axis for cross-subset analyses. Both choices are forced by the evaluation setup and are justified below.

L.1 Partition granularity and the family-level estimator

The theorem object of central interest is the familywise excess Bayes risk

$$\Delta_{\mathcal{Q}}(Z; X) := \sup_{q \in \mathcal{Q}} \big(\mathcal{R}^{\star}_{Z}(q) - \mathcal{R}^{\star}_{X}(q)\big), \tag{L.1}$$

whose supremum runs over individual queries. Direct empirical estimation of this quantity would require per-query risk estimates $\widehat{\mathcal{R}}^{\star}_{Z}(q)$ and $\widehat{\mathcal{R}}^{\star}_{X}(q)$. That is impossible in V1 because every query appears exactly once in the evaluation data. The finest partition at which per-partition risks are estimable is therefore the labeled query family.

Let $\mathcal{F}(\mathcal{Q})$ denote the partition of the dataset's queries into labeled families, and let h denote the frozen answerer used for the paired raw/compressed evaluation. For family $F \in \mathcal{F}(\mathcal{Q})$ with N_F samples, the family-level mean excess risk at budget b is

$$\begin{aligned}\widehat{\Delta}^{\text{fam}}_{\mathcal{P}}(b) = \max_{F \in \mathcal{F}(\mathcal{Q})} \Bigg[&\frac{1}{N_F} \sum_{(x,q,y):\, \text{family}(q)=F} \ell\big(h(Z_b(x), q), y\big) \\ &- \frac{1}{N_F} \sum_{(x,q,y):\, \text{family}(q)=F} \ell\big(h(X, q), y\big) \Bigg].\end{aligned} \tag{L.2}$$

The corresponding dataset-mean estimator is

$$\widehat{\Delta}_{\text{avg}}(b) = \frac{1}{N} \sum_{(x,q,y) \in \text{eval}} \Big[\ell\big(h(Z_b(x), q), y\big) - \ell\big(h(X, q), y\big)\Big]. \tag{L.3}$$

At the population level, replacing individual queries by a labeled-family partition gives the ordering proved in Theorem 3.3:

$$\Delta^{\text{avg}}_{\mu}(Z; X) \leq \Delta^{\text{fam}}_{\mathcal{F}(\mathcal{Q})}(Z; X) \leq \Delta_{\mathcal{Q}}(Z; X). \tag{L.4}$$

Empirically, $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b) - \widehat{\Delta}_{\mathrm{avg}}(b)$ is the observed hidden-damage margin for the declared partition. On single-family datasets, the two estimators coincide by construction.

L.2 Operational frontier estimator

For compression method m, tolerance ε, and excess-risk variant $\widehat{\Delta} \in \{\widehat{\Delta}_{\mathrm{avg}}(b), \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)\}$, the nominal-axis frontier is

$$\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; m, \widehat{\Delta}) := \min\{b \in \mathcal{B} : \widehat{\Delta}(b) \le \varepsilon\}, \tag{L.5}$$

where $\mathcal{B}$ is the evaluated budget grid. The token-axis frontier is

$$\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q}; m, \widehat{\Delta}) := \widehat{\tau}(\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; m, \widehat{\Delta})), \tag{L.6}$$

where $\widehat{\tau}(b)$ is the empirical mean number of audio tokens passed to the backbone at nominal budget b.

We interpret $\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ as an operational proxy for the rate-theoretic frontier rather than as a calibrated estimator of $R^{\star}_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q})$: our interface is a hard-selected audio-token subset without an entropy-efficient coding layer, so the Shannon-slack bound from Proposition K.8 is not expected to be tight.

L.3 Axis convention

Single-cell analyses use the token axis, because token count corresponds directly to what the backbone processes. Cross-subset analyses use the nominal-budget axis, because the token axis is subset-dependent: restricting to a family subset changes the per-budget distribution of clip lengths and therefore changes $\widehat{\tau}(b)$. In particular, the cumulative-chain analysis of §O.4, the factor-overlap analysis of §O.5, and the conditioned-gain comparison of §O.6 are all reported on the nominal axis.

L.4 Worst-family-constrained frontier

For a query-family subset $\mathcal{Q}_k \subseteq \mathcal{Q}$, the worst-family-constrained nominal frontier is

$$\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_k) := \min\left\{ b \in \mathcal{B} : \max_{F \in \mathcal{F}(\mathcal{Q}_k)} \left[\frac{1}{N_{F,k}} \sum_{\mathrm{family}(q)=F} \ell\big(h(Z_b(x), q), y\big) - \frac{1}{N_{F,k}} \sum_{\mathrm{family}(q)=F} \ell\big(h(X, q), y\big) \right] \le \varepsilon \right\}. \tag{L.7}$$

By construction this frontier is monotone non-decreasing in the subset argument: if $\mathcal{Q}_1 \subseteq \mathcal{Q}_2$, then $\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_1) \le \widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_2)$. Along a cumulative chain, the increment

$$\Delta\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_k) = \widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_k) - \widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_{k-1}) \tag{L.8}$$

measures the additional budget required when the new family is added to the requirement set.

L.5 Operational conditioned gain

The real-audio operational counterpart of Theorem 3.4 is defined for either the average estimator or the familywise estimator:

$$\begin{aligned} \widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q}; \widehat{\Delta}) := \widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \texttt{learned_agnostic}, \widehat{\Delta}) \\ - \widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \texttt{learned_conditioned}, \widehat{\Delta}), \\ \widehat{\Delta} \in \{\widehat{\Delta}_{\mathrm{avg}}(\cdot), \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(\cdot)\}. \end{aligned} \tag{L.9}$$

A positive value means that the conditioned selector requires less nominal budget to achieve the same tolerance and estimator than the agnostic selector. The familywise version $\widehat{\Delta} = \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(\cdot)$ is the

deployment sign-off quantity; the average version is retained for V1 historical comparisons. The synthetic exact quantity

$$\widehat{G}^{\mathrm{info}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q}) = R^{\star}_{\mathrm{Bayes}}(\varepsilon, \mathcal{Q}) - R^{\star}_{\mathrm{Bayes,cond}}(\varepsilon, \mathcal{Q}) \tag{L.10}$$

is reserved for the finite-alphabet constructions of §O.1.

The operational conditioned-gain estimator $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q})$ defined above compares two selector methods evaluated on the same downstream backbone; we refer to this as the *same-backbone* version. V2 further reports a *cross-backbone* variant

$$\begin{aligned}\widehat{G}^{\mathrm{op,cross}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q}; B) := \widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \texttt{learned_agnostic} \text{ on } \texttt{qwen2audio}) \\ - \widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q}; \texttt{learned_conditioned} \text{ on } B).\end{aligned} \tag{L.11}$$

which fixes the reference method to Qwen2-Audio's `learned_agnostic` selector. The same-backbone version asks "does conditioning help, holding everything else constant." The cross-backbone version asks "how much does a backbone plus its conditioned selector lose relative to a shared reference backbone with its agnostic selector." Tables 24 and 25 in §O.7 report the same-backbone version; Table 32 and Table 26 report the cross-backbone version. The two versions differ only for Qwen2.5-Omni because the Qwen2-Audio cross-backbone case is degenerate (A and B use the same backbone). The numerical separation can be large: for MMSU on Qwen2.5-Omni at $\varepsilon = 0.05$, the same-backbone three-seed mean on the keyword `temporal` family is $+0.0057$, whereas the cross-backbone three-seed mean is -0.5738 (see §O.11).

L.6 Bootstrap protocol

All CIs in V1 are computed with non-parametric bootstrap resampling at the sample-ID level with $n_{\mathrm{boot}} = 10{,}000$, confidence level 0.95, and seed 42 Efron & Tibshirani (1993). For single-estimator quantities such as $\widehat{\Delta}_{\mathrm{avg}}(b)$, $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$, and $\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q})$, each bootstrap iteration resamples sample IDs with replacement and recomputes the statistic on the resampled set. The 95% interval is the percentile interval.

For $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q})$ we use a paired bootstrap: each resampling draws a multiset of sample IDs once, then computes both the agnostic and conditioned frontiers on that same resampled set before taking the difference. The paired structure preserves sample-level correlation between the two methods. Two-sided p-values are computed from the bootstrap replicate distribution as

$$p = 2\min\big(\Pr[\widehat{G}^{*}_{\mathrm{cond}} \le 0],\, \Pr[\widehat{G}^{*}_{\mathrm{cond}} \ge 0]\big). \tag{L.12}$$

The factor-overlap diagnostic uses a paired ratio bootstrap, and the cumulative-chain analysis uses a shared-resampling bootstrap across all chain steps.

The bootstrap endpoints below describe the historical analyses; formal confirmation uses the independent protocol in Section 4. Let $U_{95}(\widehat{\theta})$ and $L_{95}(\widehat{\theta})$ be the upper and lower percentile endpoints. In these descriptive analyses, interval positions relative to a tolerance summarize a fixed cell; they do not authorize budget selection or post-selection certification. For conditioned gain, the sign is declared positive only when the lower endpoint is above zero, negative only when the upper endpoint is below zero, and inconclusive otherwise. These intervals quantify sampling uncertainty; cross-seed standard deviations quantify selector/training stochasticity and are reported separately. If several evaluation rows are derived from the same audio source, the same formulas apply with bootstrap resampling at the audio-source ID level.

L.7 Reference conventions

The canonical raw-audio reference for family F is

$$\widehat{R}_X(F) = \frac{1}{N_F} \sum_{(x,q,y):\, \mathrm{family}(q)=F} \ell\big(h(X,q), y\big). \tag{L.13}$$

When the canonical uncompressed parquet is missing for qwen2.5-Omni (AudioMCQ and MMSU), we use the $b = 1.00$ endpoint of the same method as a `self_full` fallback. On qwen2audio cells where both references are available, the difference is bounded by 10^{-4}.

Table 11: Dataset summary for the five audio-QA benchmarks used in V1 evaluation. The **Kw fam.** column is the number of distinct `query_family` labels assigned by the shared keyword-heuristic classifier of §M.6; the **Native** column is the number of families under each dataset's intrinsic task-label metadata (see §M and Table 17). The Acc. column is qwen2audio uncompressed accuracy at budget fraction 1.00, read from the per-cell aggregate JSONs. Partial qwen2.5-Omni coverage reflects the DCASE flash-attention CPU-dispatch failure and the missing canonical uncompressed parquets on AudioMCQ and MMSU in V1.

Dataset	n	Kw fam.	Native	Acc.	Role in paper	qwen2audio	qwen2.5-Omni	E1?
DCASE 2026 dev	1,607	1	6	0.4443	Primary operational target; single-family conditioned-gain test under the keyword partition	Full suite	Heuristics only; DCASE pipeline failed	No
AudioMCQ-StrongAC	19,480	4	10	0.7424	Primary multi-family headline dataset and selector-training source	Full suite	Heuristics only; no canonical uncompressed reference	Yes
MMSU	5,000	6	47	0.5530	Most taxonomically rich multi-family benchmark	Full suite	Heuristics only; no canonical uncompressed reference	Yes
MMAR	1,000	3	9	0.4960	Supplementary multi-family reasoning benchmark	Full suite	Heuristics plus uncompressed reference	Yes
BigBench Audio	1,000	1	4	0.8770	Text-dominated single-family control under the keyword partition (four BBH task types under native partition)	Full suite	Heuristics plus uncompressed reference	No

L.8 Loss-type conventions

The primary loss is the 0/1 multiple-choice loss

$$\ell_{0/1}(q, \hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}, \tag{L.14}$$

which matches the benchmark decision rule and underlies all main-text tables. We also compute the correct-answer negative log-likelihood,

$$\ell_{\mathrm{NLL}}(q, \hat{y}, y) = -\log P_{\mathrm{backbone}}(y \mid Z, q), \tag{L.15}$$

and use it as a secondary measurement in supporting analyses. The 0/1 and NLL variants yield the same qualitative conclusions throughout V1.

L.9 Estimator catalogue

For reference, the empirical estimators used in §O are: $\widehat{\Delta}_{\mathrm{avg}}(b)$ (dataset-mean excess loss), $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ (family-level excess loss), $\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ (token-axis operational frontier), $\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q})$ (nominal-axis operational frontier), $\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_k)$ (worst-family-constrained frontier along cumulative chains), $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q})$ (operational conditioned gain), $\widehat{G}^{\mathrm{info}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q})$ (exact synthetic conditioned gain), and $\widehat{\tau}(b)$ (mean audio-token count at nominal budget b).

M Dataset and Query-Family Details

We evaluate TAAC on five audio-question datasets that together span the principal capabilities a large audio language model is expected to exercise: sound-event understanding on environmental audio, multi-capability multiple-choice reasoning on general audio, speech-centric multi-task understanding, music-and-speech-aware reasoning, and logical reasoning under audio-textual prompts. The datasets are chosen to satisfy two conditions simultaneously: each must expose a nontrivial query family to which Theorem 3.2's family-level estimator can be applied, and the union must include at least one single-family dataset against which family-level effects can be tested for artifacts. Three of the five datasets (AudioMCQ-StrongAC, MMSU, and MMAR) carry multiple labeled query families under the keyword classifier of §M.6 and support the full theorem-level analysis; two (DCASE 2026 dev and BigBench Audio) carry a single labeled family under that classifier and serve as single-family controls at the keyword level. Both DCASE and BigBench Audio are in fact multi-family under their intrinsic dataset-native task metadata (§O.2 Table 17); the single-family status is therefore a property of the keyword classifier, not of the datasets themselves.

M.1 DCASE 2026 dev

DCASE 2026 dev is the development subset of DCASE 2026 Challenge Task 5, Audio-Dependent Question Answering (ADQA) Yang et al. (2026). Each sample consists of an audio clip and a short

Table 12: Duplicate source-copy dataset summary retained from `figures/fig7_dataset_summary.tex`. Dataset summary for the five audio-QA benchmarks used in V1 evaluation. Family count is the number of distinct `query_family` labels assigned by the shared keyword-heuristic classifier. The Acc. column is qwen2audio uncompressed accuracy at budget fraction 1.00, read from the per-cell aggregate JSONs. Partial qwen2.5-Omni coverage reflects the DCASE flash-attention CPU-dispatch failure and the missing canonical uncompressed parquets on AudioMCQ and MMSU in V1.

Dataset	n	Families	Acc.	Role in paper	qwen2audio	qwen2.5-Omni	E1?
DCASE 2026 dev	1,607	1	0.4443	Primary operational target; single-family conditioned-gain test	Full suite	Heuristics only; DCASE pipeline failed	No
AudioMCQ-StrongAC	19,480	4	0.7424	Primary multi-family headline dataset and selector-training source	Full suite	Heuristics only; no canonical uncompressed reference	Yes
MMSU	5,000	6	0.5530	Most taxonomically rich multi-family benchmark	Full suite	Heuristics only; no canonical uncompressed reference	Yes
MMAR	1,000	3	0.4960	Supplementary multi-family reasoning benchmark	Full suite	Heuristics plus uncompressed reference	Yes
BigBench Audio	1,000	1	0.8770	Text-dominated validity control	Full suite	Heuristics plus uncompressed reference	No

natural-language multiple-choice question whose answer must depend on the audio. We use the development split because the official evaluation set is held out by the challenge.

Role in the paper. DCASE 2026 is TAAC's primary *operational* target: the dataset is dominated by non-speech audio content, so learned selectors cannot succeed by defaulting to speech-content heuristics, and the ground-truth signal is sensitive to which audio regions are retained under compression. It is also the dataset on which the operational conditioned gain reaches its largest point estimate, $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(0.05, \mathcal{Q}) = +0.0993$ budget fraction with paired-bootstrap $p = 0.057$; see §O.6.

Query family structure. Under the keyword classifier of §M.6, DCASE 2026's category strings do not match any of the predefined dispatch groups, so the classifier assigns `query_family=general` to every sample. Under this partition, DCASE functions as a single-family evaluation: $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ coincides by construction with $\widehat{\Delta}_{\mathrm{avg}}(b)$, and the keyword-level nested-monotonicity analysis collapses to the trivial chain $\{\texttt{general}\}$. However, DCASE does expose a native content taxonomy: the challenge's own `content_taxonomy.csv` groups samples into six post-merge clusters (e.g., `meaning/stress/intonation`, `speaker/clip/express`, and four smaller clusters; see §O.11 and the `other_rare` caveat noted there). Under this native partition, the family-level excess-risk gap is $+5.04$ pp at $b = 0.20$ with Qwen2-Audio learned-conditioned (Table 17), rather than the zero imposed by the keyword classifier. DCASE is therefore a single-family control only with respect to the keyword classifier, not with respect to the dataset's intrinsic structure.

Sample count and audio characteristics. The DCASE 2026 dev split contributes $n = 1{,}607$ audio-query pairs. Audio durations are variable, with most clips between 5 and 30 seconds; all clips are resampled to 16 kHz mono before chunking. Qwen2-Audio evaluations are complete for all compression methods and budgets. The qwen2.5-Omni DCASE pipeline failed under a flash-attention CPU-dispatch error and is dropped from V1.

M.2 AudioMCQ-StrongAC

AudioMCQ-StrongAC is a four-option multiple-choice audio-question dataset derived from AudioMCQ, with the "StrongAC" qualifier denoting the strongly controlled subset in which the correct answer is supported by specific audio content rather than by inferable background knowledge He et al. (2026). Each sample consists of a natural audio clip, a natural-language query, and four candidate answers, exactly one of which is correct.

Role in the paper. AudioMCQ-StrongAC is the primary dataset for V1's empirical theorem testing. Its native question-type taxonomy (four families, 19,480 evaluation samples) provides the largest multi-family evaluation set in V1 and the one in which the family-level excess-risk gap is most pronounced. The gap $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b) - \widehat{\Delta}_{\mathrm{avg}}(b)$ peaks at $+6.79$ percentage points at $b = 0.20$, with a near-peak value of $+6.29$ percentage points at $b = 0.40$; this is the headline empirical claim of the paper. AudioMCQ is also the dataset on which the selector networks were trained.

Query family structure. AudioMCQ-StrongAC exposes a `question_type` field that our shared classifier maps onto four canonical labels: `general`, `music`, `speech_content`, and `temporal`. The semantic meaning of each family is as follows.

- **`speech_content`** ($n = 10{,}505$, 53.9%): queries about linguistic or lexical content of speech-what was said, transcription, phrase-level understanding, or the language being spoken.
- **`general`** ($n = 6{,}085$, 31.2%): fallback bucket for samples whose native category string matches none of the specific keyword groups. On AudioMCQ this family is the bottleneck family under the nested-monotonicity analysis: adding `general` to the cumulative chain yields the largest increase in $\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(0.05, \mathcal{Q}_k)$, from 0.470 to 0.697 ($+0.227$).
- **`music`** ($n = 1{,}724$, 8.9%): queries about musical content such as instruments, genre, melody, rhythm, or song-level attributes.
- **`temporal`** ($n = 1{,}165$, 6.0%): queries about time, order, counting, duration, or timing relationships between events in the audio.

Sample count and audio characteristics. The AudioMCQ-StrongAC training split contains 19,480 samples with precomputed per-chunk LOO-NLL oracle relevance targets; the evaluation split used here also contains 19,480 samples, with the same per-family counts listed above. Clips are variable-duration (typically 3-30 s), sampled at 16 kHz mono, padded to one chunk when shorter than 1 s and truncated to 120 chunks when longer than 120 s.

M.3 MMSU

MMSU is a multi-task benchmark for spoken-language understanding and reasoning, designed to evaluate fine-grained speech perception and complex reasoning in natural speech Wang et al. (2026). Each sample is a four-option multiple-choice question paired with an audio clip. The benchmark spans linguistic content, paralinguistic inference, speaker attributes, and speech-conditioned scene and event reasoning.

Role in the paper. MMSU is the most taxonomically rich dataset we evaluate, exposing six distinct keyword families and 47 native `task_name` values. Under the keyword partition it is used to show that the family-level excess-risk gap is not an AudioMCQ-specific artifact and to provide the richest keyword-level cumulative-chain analysis in §O.4. Under the 47-task native partition, MMSU is also the dataset on which the keyword partition most severely under-resolves the intrinsic structure - a factor of roughly $8\times$ - and correspondingly the dataset on which the native-partition family-level gap is largest in absolute terms ($+29.17$ pp at $b = 0.20$; Table 17). The per-task analysis in §O.11 identifies `intonation_perception` as the single most-impacted native task on Qwen2.5-Omni under the cross-backbone conditioned-gain estimator (Figure 16).

Query family structure. Our classifier operates on the concatenation of MMSU's native task name and category string. Six canonical labels appear:

- **`general`** ($n = 2{,}886$, 57.7%): the largest fallback bucket.
- **`paralinguistic`** ($n = 803$, 16.1%): emotion, speaker identity, accent, gender, prosody, and pitch.
- **`speech_content`** ($n = 648$, 13.0%): linguistic content of speech.
- **`sound_event`** ($n = 454$, 9.1%): discrete acoustic events and sources. Despite the name, this keyword family is not semantically pure: approximately 115 of 454 samples route to a user-intent-classification semantic cluster under e5-large-v2 embedding analysis (supporting file: `partitions/semantic/mmsu_semantic_partition.csv`; see §P.6).
- **`sound_scene`** ($n = 110$, 2.2%): ambient scene and environment.
- **`temporal`** ($n = 99$, 2.0%): temporal structure of the audio.

The two smallest families, `sound_scene` and `temporal`, have sample counts below 200 and therefore produce wide bootstrap CIs in family-wise frontier estimates.

While the keyword partition groups MMSU queries into six families, the MMSU native metadata carries a `task_name` field with 47 distinct values in the release we evalu-

ate (e.g., `word_identification`, `age_prediction`, `intonation_perception`, `pitch_comparison`, `disfluency_detection`, `volume_comparison`, `couplet_matching`, `dialogue_turn_counting`, among others). The keyword classifier's 6-family partition is therefore approximately $8\times$ under-resolution of the dataset's intrinsic task structure. §O.2 Table 17 shows that this under-resolution materially suppresses the measured family-level gap: under the 47-task native partition, the MMSU worst-family gap at $b = 0.20$ is $+29.17$ pp, compared with the $+1.56$ pp measured under the keyword partition on the same Qwen2-Audio learned-conditioned cells (supporting CSV: `tables/partition_delta_q_summary.csv`, rows with `dataset='mmsu', backbone='qwen2audio', method='learned_conditioned'`).

Sample count and audio characteristics. MMSU contributes $n = 5{,}000$ evaluation samples. The audio is speech-dominated, variable-duration, and resampled to 16 kHz mono. The speech dominance is load-bearing for one observation in §O.4: speech-related families remain relatively well served under aggressive compression because speech dominates the V1 selector's mel-spectrogram feature energy.

M.4 MMAR

MMAR is a benchmark for deep reasoning in speech, audio, music, and their mixtures Ma et al. (2025). We use MMAR's four-option multiple-choice audio-question pairs, treating the textual query as the standard natural-language question and evaluating only the audio-conditioned answer.

Role in the paper. MMAR is a supplementary multi-family dataset that demonstrates the generality of the family-level gap and the nested-monotonicity behavior on a third independently constructed benchmark. It is also the dataset on which the only two numerical monotonicity violations occur, both at the smallest subset and the tightest tolerance, with point-estimate drops of only 0.001-0.003 budget fraction.

Query family structure. Our classifier operates on the concatenation of MMAR's modality and category fields. Three labels appear:

- **`speech_content`** ($n = 618$, 61.8%): linguistic content of speech.
- **`music`** ($n = 217$, 21.7%): music-oriented reasoning.
- **`general`** ($n = 165$, 16.5%): fallback bucket.

Sample count and audio characteristics. MMAR contributes $n = 1{,}000$ evaluation samples. Audio is variable-duration, 16 kHz mono after resampling, with a mixture of speech, music, and mixed-content clips broadly consistent with the three-family taxonomy above.

M.5 BigBench Audio

BigBench Audio is an audio version of a subset of BIG-Bench Hard questions Artificial Analysis (2024); Suzgun et al. (2023). Each item pairs an audio rendering of a reasoning prompt with a multiple-choice answer. In practice, many questions can be answered largely from the textual prompt alone, with the audio providing supporting rather than load-bearing information.

Role in the paper. BigBench Audio is a text-dominated dataset where single-method per-backbone operational effects of compression are small: under the same-backbone conditioned-gain estimator (§L), both Qwen2-Audio and Qwen2.5-Omni show near-null behavior on every axis. However, the dataset is not a universal null control. Under the cross-backbone estimator (§O.7), Qwen2.5-Omni's `formal_fallacies` task exhibits a large negative conditioned gain (three-seed mean -0.8527, std 0.025; see §O.11 and Table 23), which we discuss further in §P.2.

Query family structure. BigBench Audio's native metadata carries a `task_name` field identifying four BIG-Bench Hard task types: `formal_fallacies`, `navigate`, `object_counting`, and `web_of_lies`, with 250 samples each. The keyword classifier of §M.6 does not match any of these to its predefined dispatch, so all samples are routed to `general` under the keyword partition. Under a native-task partition evaluated at $b = 0.20$ with Qwen2-Audio learned-conditioned and 0/1 loss,

Table 13: BigBench Audio per-family per-sample mean excess at $b = 0.20$ with Qwen2-Audio `learned_conditioned` and 0/1 loss. The four native families partition BigBench Audio's 1000 samples into equal-sized 250-sample BIG-Bench Hard task types. Three families exhibit negative mean excess (compression improves predictions); one exhibits strongly positive mean excess. The dataset-aggregate near-null reported in Table 22 is the weighted mean of these four opposing effects. Supporting CSV: `metrics/partition_delta_q_long.csv` rows with `dataset='bigbench_audio'`, `partition='native_fine'`, `backbone='qwen2audio'`, `method='learned_conditioned'`, `budget=0.20`.

Native task	n	per-sample mean excess
`formal_fallacies`	250	$+0.409$
`navigate`	250	-0.123
`object_counting`	250	-0.123
`web_of_lies`	250	-0.123
dataset aggregate	1000	$+0.010$

three of four families exhibit per-sample mean excess of -0.123 (compression improves predictions), while `formal_fallacies` exhibits $+0.409$ (compression harms predictions); see Table 13. The near-null aggregate in Table 22 is thus a cancellation of four opposing effects rather than four neutral effects. Correspondingly, §O.2 reports a family-level excess-risk gap of $+39.90$ percentage points at $b = 0.20$ under the native partition, substantially larger than the keyword-aggregate of $+0$. BigBench Audio is therefore a "single-family control" only with respect to the keyword classifier, not with respect to the dataset's intrinsic structure.

Sample count and audio characteristics. BigBench Audio contributes $n = 1{,}000$ evaluation samples. Short, low-information audio clips are common, and the dataset's reasoning burden is dominated by the textual prompt.

M.6 Query family taxonomy

All query-family labels in this paper are assigned by a single classifier function `infer_family(raw_cat)` defined in `src/taac/datasets/mcq_parser.py`. The function lowercases its input and dispatches in a fixed order through seven keyword groups; the first group that matches wins, and inputs matching no group fall through to `general`. The dispatch order is:

1. **`speech_content`**: {speech, asr, transcript, spoken, language, word, phrase}
2. **`paralinguistic`**: {emotion, speaker, accent, gender, paralinguis, prosody, pitch}
3. **`sound_event`**: {event, detection, sound_event, source}
4. **`sound_scene`**: {scene, environment, ambient, location}
5. **`music`**: {music, instrument, genre, melody, rhythm, song}
6. **`temporal`**: {temporal, count, timing, order, when, duration, time}
7. **`general`**: fallback bucket.

Because the classifier is shared, each canonical label has the same definition on every dataset; what differs is the source field on which the classifier operates.

The taxonomy is a keyword-heuristic partition of each dataset's native category strings, not a semantic partition of query content. A query whose category string contains the word "when" is routed to `temporal` regardless of whether the question actually requires temporal reasoning, and a query whose category string contains "speech" is routed to `speech_content` regardless of the specific speech property at issue. Partition monotonicity applies only when one partition actually refines another; the native and semantic alternatives need not refine the keyword partition. The reported gaps and frontier-overlap ratios are therefore partition-specific, not lower bounds on unidentified semantic quantities. The observed partition sensitivity is illustrated in §O.2: under dataset-native task partitions (e.g., MMSU's 47 `task_name` values, BigBench's 4 BBH task types), measured family-level gaps are $1.2\times$ to $20\times$ larger than under the keyword partition at every tested budget (Table 17).

The family-label robustness of the family-level findings has been quantified directly. Under 20% random label flips with 100 Monte Carlo replicates, the bottleneck-family identity under the semantic partition (§P.6) is preserved in 84%-100% of replicates on MMSU, MMAR, and AudioMCQ-StrongAC, compared with 39%-57% under the keyword partition (supporting CSV: `metrics/label_noise_sensitivity_long.csv`). Label noise therefore affects the observed family structure most under the keyword partition and least under the semantic partition, indicating that finer semantic partitioning both increases the measured gap and reduces its sensitivity to label noise. An empirical instance of the keyword classifier's conservatism: on AudioMCQ-StrongAC, a single semantic cluster of queries about "sequence" or "timing" of sounds ($n = 1619$ per e5-large-v2 + k-means clustering of query text) splits across keyword families as {`general`: 500, `music`: 54, `speech_content`: 64, `temporal`: 1020}, so approximately 500 semantically-temporal samples are routed to `general` under the keyword dispatch.

N Experimental Protocols

This section specifies the backbones, selector architectures, training recipe, evaluation protocol, and inference-time compressor configuration used in V1. Under the no-compromise policy, we surface three methodology caveats explicitly: truncated selector training, the large apparent parameter asymmetry between agnostic and conditioned selectors being dominated by the query embedding table, and the V1 selector's use of 768-dimensional mel-spectrogram features rather than backbone audio-tower features. These caveats interact only with the operational conditioned-gain result in §O.6; the family-level gap, nested-monotonicity, and factor-overlap findings are about theorem-aligned empirical quantities and are invariant to the particular selector instantiation.

N.1 Backbones

V1 uses two open-source audio-language model backbones.

Qwen2-Audio-7B-Instruct. Qwen2-Audio-7B-Instruct is the primary backbone for V1 Chu et al. (2024). The model combines an audio compressor with a Qwen2 language decoder and supports direct textual responses conditioned on audio. All compression methods-`uncompressed`, `uniform`, `random`, `energy_vad`, `learned_agnostic`, and `learned_conditioned`-are evaluated against qwen2audio on all five datasets. Decoder generation is capped at `max_new_tokens=5` because every benchmark is four-option multiple choice.

Qwen2.5-Omni-7B. Qwen2.5-Omni-7B is the secondary backbone Xu et al. (2025). It provides a cross-backbone validity check on MMAR, MMSU, AudioMCQ, and BigBench Audio, but the DCASE pipeline failed due to a flash-attention CPU-dispatch error that returned `predicted_label=ERROR` on all processed rows. V1 therefore treats qwen25omni only as a validity-control backbone; the learned-selector analyses are qwen2audio-only.

Architecture gap. Proposition K.7 defines a theoretical model-class gap $\Gamma_{\mathcal{F}}(Z; \mathcal{Q})$. The reported cross-answerer/interface summaries are operational diagnostics rather than estimates of that gap; see §P.5.

N.2 Selector compression methods

V1 evaluates two learned selectors against three heuristic baselines. All five methods operate on the same interface: segment the audio into 1-second chunks, assign a scalar relevance score to each chunk (or a deterministic selection rule for the heuristics), and retain the top-k chunks where $k = \max(1, \lfloor bN \rfloor)$ for budget fraction b and total chunk count N.

- **`uncompressed`** passes the full audio to the backbone and defines the canonical raw-audio reference.
- **`uniform`** retains evenly spaced chunks at fixed density.
- **`random`** selects chunks uniformly at random without replacement, using a fixed per-sample seed.
- **`energy_vad`** ranks chunks by RMS energy and keeps the top-k.

- **`learned_agnostic`** and **`learned_conditioned`** are small MLP selectors trained on precomputed LOO-NLL oracle relevance targets. The only architectural difference is whether the scoring network sees the query embedding at inference time.

N.3 V1 SELECTOR ARCHITECTURE AND TRAINING

Both learned selectors are implemented as instances of the same `LearnedSelector` class and consume precomputed per-chunk features of dimensionality 768. Those 768-dimensional features are pooled log-mel-spectrogram representations rather than the 1280-dimensional embeddings produced by qwen2audio's own audio tower. This detail is load-bearing: V1 selectors operate on a weaker acoustic representation than the backbone itself uses at inference, so V1 tests a conservative version of the answer-preserving compression problem.

Chunk compressor and scoring heads. The chunk compressor is a two-layer MLP,

$$\mathrm{Linear}(768, 256) \to \mathrm{GELU} \to \mathrm{LayerNorm}(256) \to \mathrm{Linear}(256, 128), \tag{N.1}$$

with 229,760 parameters excluding the LayerNorm affine terms. The agnostic scoring head applies

$$\mathrm{Linear}(128, 256) \to \mathrm{GELU} \to \mathrm{LayerNorm}(256) \to \mathrm{Linear}(256, 1), \tag{N.2}$$

plus a learned scalar `pos_scale` multiplying a normalized temporal-position bias. The conditioned scoring head uses the same structure after concatenating the 128-dimensional chunk embedding with a 128-dimensional query embedding.

Query compressor and parameter asymmetry. The conditioned selector adds a query compressor comprising a token embedding `Embedding(151644,128)`, a positional embedding `Embedding(256,128)`, and a two-layer MLP applied to the mask-aware mean-pooled token-plus-position embedding. Direct checkpoint inspection gives the following counts: `learned_agnostic` has 263,066 parameters across 13 tensors; `learned_conditioned` has 19,806,446 parameters across 21 tensors, of which 19,410,432 live in the query token embedding table. The apparent $75\times$ parameter asymmetry is therefore almost entirely an embedding-table effect; the non-embedding "scoring core" is only $1.51\times$ larger for the conditioned selector.

Training data and target. Both selectors are trained exclusively on the 19,480-sample AudioMCQ-StrongAC training split. The training manifest stores a per-sample `chunk_relevance` vector whose provenance field is `chunk_relevance_source=loo_nll` in every row. For each chunk, the target relevance is the increase in qwen2audio's NLL on the correct answer when that chunk is removed from the audio. At training time the raw relevance vector is normalized to a probability distribution over chunks and used as the target for the selector's soft selection distribution.

Loss and optimization. The training loss is

$$\mathcal{L} = \mathrm{KL}(\text{relevance} \,\|\, \text{selection}) + 0.1\,\mathrm{MSE}(\text{actual_count}, \text{target_count}) - 0.01\,H(\text{selection}), \tag{N.3}$$

where the entropy term prevents early collapse. The selectors use AdamW with learning rate 10^{-4}, weight decay 0.01, batch size 8, gradient accumulation 4, cosine decay with warmup ratio 0.05, and gradient clipping at norm 1.0. The differentiable top-k path uses a Gumbel-softmax perturbation of the chunk scores Herrmann et al. (2020); inference uses exact hard top-k.

V1 training truncation. The nominal V1 schedule is ten epochs, but seed-42 training terminated early under patience-3 stopping on a Gumbel-noisy validation signal. The agnostic selector's final checkpoint was saved at step 2,184 (35.9% of nominal schedule), and the conditioned selector's at step 1,638 (26.9%). The conditioned selector is therefore more undertrained than the agnostic selector, which biases V1 against finding a conditioned-compression advantage. V2 corrects this with patience 8 and a longer schedule.

N.4 EVALUATION PROTOCOL

V1 evaluation is run as a per-backbone, per-dataset, per-method, per-budget sweep that produces one parquet file per cell. Each row stores `sample_id`, `query_family`, the 0/1 multiple-choice loss,

the NLL on the correct answer, the realized `num_audio_tokens`, and the predicted answer label. Every downstream analysis in §L and §O is computed from these parquets.

The canonical raw-audio reference is the mean backbone loss on the uncompressed audio, computed per family. This reference is available for qwen2audio on all five datasets and for qwen2.5-Omni on MMAR and BigBench Audio. On the two qwen25omni cells where the canonical uncompressed parquet is absent (AudioMCQ and MMSU), we use the method's own $b = 1.00$ endpoint as a `self_full` fallback. Empirically, the absolute difference between the canonical reference and the fallback is bounded by 10^{-4} on qwen2audio cells where both are available.

N.5 Chunking and top-k mapping

Audio is segmented into 1-second non-overlapping chunks at 16 kHz. Although the stored configuration block lists `hop_duration_s=0.5`, direct inspection of the training manifest confirms non-overlapping chunking: a 10-second sample contains exactly 10 chunks, and evaluation parquets match the non-overlapping rule $N = \lceil D \rceil$ after padding short clips and truncating clips longer than 120 seconds.

For budget fraction $b \in [0, 1]$ and audio with N chunks, the interface retains

$$k = \max(1, \lfloor bN \rfloor) \tag{N.4}$$

chunks. The floor-of-one is important for very short clips: at $b = 0.05$ on a 10-chunk clip, the selector still retains one chunk; at $b = 1.00$, all chunks are retained and the selection becomes a no-op.

N.6 Budget and tolerance grids

The main-body budget grid is

$$b \in \{0.05, 0.10, 0.20, 0.40, 1.00\}, \tag{N.5}$$

spanning two orders of magnitude of compression while reserving $b = 1.00$ as the uncompressed endpoint. An expanded appendix grid

$$b \in \{0.01, 0.025, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 1.00\} \tag{N.6}$$

provides additional resolution at both extremes. Operational frontiers are reported at tolerances

$$\varepsilon \in \{0.01, 0.02, 0.05\}. \tag{N.7}$$

The $\varepsilon = 0.05$ setting is the primary one in the main text because it corresponds to roughly 5 percentage points of excess 0/1 loss.

N.7 Summary of V1 methodology caveats

We close §N by consolidating the three load-bearing V1 caveats.

1. **Truncated selector training.** Both selectors terminate well short of the nominal schedule, and the conditioned selector terminates earlier than the agnostic selector.
2. **Parameter asymmetry dominated by the query embedding.** The conditioned selector's large parameter count is almost entirely the token embedding table; its scoring core is only modestly larger than the agnostic selector's.
3. **Mel-spectrogram input features.** V1 selectors operate on 768-dimensional mel features rather than the backbone's native 1280-dimensional audio-tower embeddings.

None of these caveats affect the structural findings on the family-level excess-risk gap, nested monotonicity, or factor overlap; they interact with the operational conditioned-gain estimate in §O.6.

O Additional Results and Ablations

This section presents the empirical instantiation of the theory developed in §3. We begin with synthetic tasks where the theorem predictions are verifiable at machine precision, then move to the headline real-audio result on the family-level excess-risk gap, then to the nested-monotonicity and factor-overlap diagnostics. We next report the V1 single-seed conditioned-gain result and then extend it with V2 three-seed, cross-backbone, and training-recipe ablations. The interpretive chain §O.1 → §O.5 → §O.6 → §O.7 is the tightest theorem-experiment bridge in the paper.

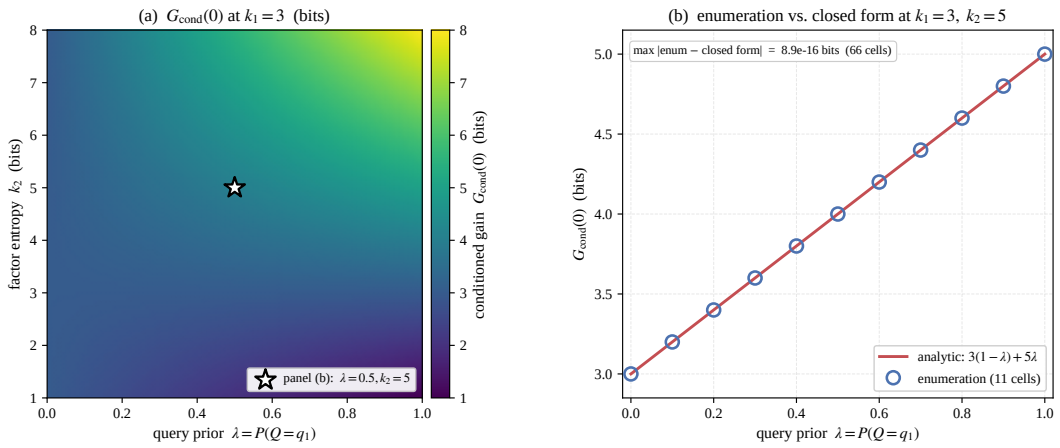


Figure 5: Synthetic verification of Theorem 3.4's strict-separation construction. **Left:** analytic conditioned gain $G_{\text{cond}}(0) = (1-\lambda)k_1 + \lambda k_2$ bits across the query prior λ and the second factor's entropy k_2, at fixed $k_1 = 3$. **Right:** exact enumeration points overlaid on the closed-form line $3(1-\lambda) + 5\lambda$ at $(k_1, k_2) = (3, 5)$. Across 66 test cells the enumerated gains match the closed form to within 10^{-9} bits.

O.1 Synthetic validations of finite, Gaussian, and conditioned-advantage results

We first validate three theoretical constructions on synthetic tasks where the Bayes frontier is analytically computable.

Finite-alphabet frontier. On finite-query synthetic tasks with controlled alphabet sizes and loss structure, the exact Bayes frontier obtained by solving the convex program of Proposition K.10 matches the closed-form optimum to machine precision.

Gaussian latent allocation. On Gaussian factor tasks with controlled variances σ_j^2 and query weights α_{qj}, numerical optimization reproduces the weighted reverse-water-filling solution predicted by Proposition K.11, including the water levels and per-coordinate distortions.

Conditioned-compression separation. For the strict-separation construction with $X = (V_1, V_2, W)$, $Q \in \{q_1, q_2\}$, $P(Q = q_1) = \lambda$, and $Y_{q_i} = V_i$ under zero-one loss, the theorem predicts

$$R^\star_{\text{Bayes}}(0, \{q_1, q_2\}) = H(V_1, V_2), \qquad R^\star_{\text{Bayes,cond}}(0, \{q_1, q_2\}) = \lambda H(V_1) + (1-\lambda) H(V_2), \tag{O.1}$$

and therefore

$$G_{\text{cond}}(0) = (1-\lambda) H(V_1) + \lambda H(V_2). \tag{O.2}$$

We verify this prediction by exact enumeration across six (k_1, k_2) configurations and eleven values of λ, for a total of 66 cells. Every enumerated cell matches the analytic form to better than 10^{-9} bits.

The synthetic verification is important for the remainder of the paper: it establishes that the conditioned gain is large in exactly the factor-disjoint regime described by Theorem 3.4, and therefore motivates reading the real-audio operational results through the measured factor structure of the datasets.

O.2 Family-level excess risk gap

The headline empirical finding of V1 is that the family-level estimator $\widehat{\Delta}^{\text{fam}}_{\mathcal{P}}(b)$ is consistently and substantially larger than the dataset-mean estimator $\widehat{\Delta}_{\text{avg}}(b)$ on every multi-family dataset and every selector family. The gap is therefore a property of the theorem quantity on real data, not a property of any specific compressor.

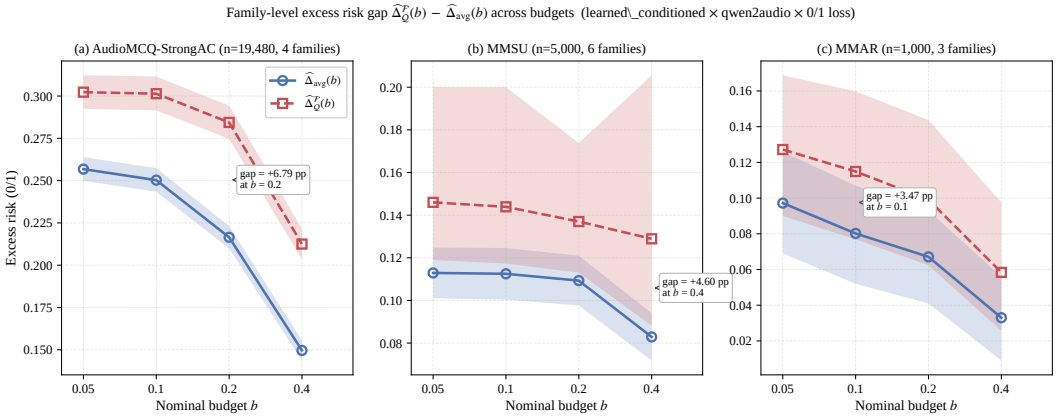


Figure 6: Empirical family-level excess-risk gap across the three multi-family datasets. Each panel shows the dataset-mean excess risk $\widehat{\Delta}_{\mathrm{avg}}(b)$ (blue, solid) and the family-level worst-family excess risk $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ (red, dashed) as a function of nominal budget b, with 95% bootstrap confidence ribbons. The gap is positive at every budget, reaching $+6.79$ percentage points on AudioMCQ-StrongAC at $b = 0.20$. Learned conditioned selector, qwen2audio backbone, and 0/1 loss. On single-family datasets (DCASE and BigBench Audio, not shown), $\widehat{\Delta}_{\mathrm{avg}}(b) \equiv \widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ by construction.

Table 14: Excess-risk summary on AudioMCQ-StrongAC: dataset-mean $\widehat{\Delta}_{\mathrm{avg}}(b)$ versus worst-family $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ at four budgets.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$	gap (pp)
0.05	0.2568 [0.2499, 0.2639]	0.3023 [0.2924, 0.3123]	+4.55
0.10	0.2503 [0.2434, 0.2574]	0.3014 [0.2915, 0.3116]	+5.11
0.20	0.2165 [0.2099, 0.2232]	0.2844 [0.2747, 0.2944]	+6.79
0.40	0.1496 [0.1436, 0.1555]	0.2125 [0.2035, 0.2216]	+6.29

AudioMCQ-StrongAC. Under `learned_conditioned` on qwen2audio with the canonical uncompressed reference and 0/1 loss, the family-level gap peaks at $b = 0.20$ and remains above six points at $b = 0.40$.

MMSU. MMSU exhibits the same phenomenon, albeit with wider family-level CIs because the `temporal` and `sound_scene` families are small.

MMAR. MMAR is the smallest of the multi-family datasets and therefore yields the noisiest family-level max, but the gap remains positive at every tested budget.

The keyword-partition analyses in Table 14, Table 15, and Table 16 are conservative in a measurable sense. Under dataset-native task partitions (AudioMCQ-StrongAC from source-dataset × question-type metadata; MMSU from 47 `task_name` values; MMAR from category × subcategory; BigBench Audio from its 4 BBH task types; DCASE from its post-merge 6-family content taxonomy), the family-level gaps grow substantially. Under semantic partitions using e5-large-v2 sentence embeddings and cosine k-means clustering on query text (see §P.6), gaps grow further still:

The paper's keyword partition under-measures the family-level gap by a factor of roughly $1.2\times$ (AudioMCQ) to $20\times$ (MMSU), and entirely misses the multi-family structure of BigBench and DCASE. The §O.2 headline of $+6.79$ pp on AudioMCQ-StrongAC should therefore be read as a lower bound; the true worst-family gap under honest partitioning is larger on every dataset. BigBench Audio's semantic cell is marked "n/a" because e5-large-v2 clustering on BBH query text collapses to a single cluster at $k = 12$ (BBH queries share strong semantic structure at the cluster level); its native 4-family partition is the correct fine-grained scope.

Table 15: Family-level excess-risk gap on MMSU under the same protocol as Table 14.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$	gap (pp)
0.05	0.1129 [0.1010, 0.1248]	0.1460 [0.1190, 0.2000]	+3.31
0.10	0.1125 [0.1004, 0.1246]	0.1439 [0.1173, 0.2000]	+3.14
0.20	0.1093 [0.0976, 0.1210]	0.1370 [0.1130, 0.1739]	+2.77
0.40	0.0829 [0.0718, 0.0940]	0.1289 [0.0882, 0.2056]	+4.59

Table 16: Family-level excess-risk gap on MMAR under the same protocol as Table 14.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$	gap (pp)
0.05	0.0972 [0.0690, 0.1260]	0.1272 [0.0900, 0.1688]	+3.01
0.10	0.0802 [0.0520, 0.1070]	0.1149 [0.0769, 0.1598]	+3.47
0.20	0.0671 [0.0410, 0.0930]	0.0991 [0.0624, 0.1437]	+3.19
0.40	0.0330 [0.0090, 0.0560]	0.0584 [0.0256, 0.0976]	+2.54

Across the three multi-family datasets in the keyword partition, the family-level excess-risk gap is consistently positive and ranges from $+2.54$ to $+6.79$ percentage points. On DCASE and BigBench Audio the keyword gap is exactly zero because the keyword classifier assigns a single `general` family; under native-task partitions, however, both datasets are multi-family and show substantial family-level gaps: $+5.04$ pp on DCASE (6 post-merge families) and $+39.90$ pp on BigBench Audio (4 BBH task families), both at $b = 0.20$ with Qwen2-Audio learned conditioned. Table 17 below consolidates the keyword-versus-native/semantic comparison.

O.3 Rate-risk and length-risk summaries

The full per-cell rate-risk and length-risk tables are large and are best generated directly from the released CSVs. The main operational pattern is clear on the nominal axis. At $\varepsilon = 0.05$, DCASE is the most compression-friendly dataset: both learned selectors reach tolerance at budget fractions around 0.5, while MMAR reaches tolerance in the 0.2-0.3 range. AudioMCQ and MMSU are the most compression-hostile datasets, with frontiers near 0.8 at $\varepsilon = 0.05$. Tightening the tolerance to $\varepsilon = 0.02$ moves DCASE to roughly 0.78-0.82, AudioMCQ to roughly 0.92, MMSU to roughly 0.85, and MMAR to roughly 0.63-0.69. At $\varepsilon = 0.01$, all datasets except MMAR require budget fractions near 0.9 or above. These frontiers are architecture-specific operational baselines rather than bit-level rate measurements.

O.4 Nested-family monotonicity and bottleneck identification

Theorem K.4 predicts that adding query families can only increase the required budget. We test this empirically with cumulative chains ordered by increasing family sample count and evaluate the worst-family-constrained frontier $\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(\varepsilon, \mathcal{Q}_k)$ at $\varepsilon = 0.05$.

Under the keyword partition, the AudioMCQ chain is strictly monotone, with increments $+0.1969$, $+0.2272$, and $+0.1826$; adding `general` is therefore the dominant keyword-level step. MMSU is also monotone, but the first two steps have very wide CIs because they are driven by the `temporal` and `sound_scene` families ($n = 99$ and $n = 110$). The scientifically relevant MMSU observation at this granularity is that once the small-family end of the chain is admitted, subsequent additions cost very little. MMAR is monotone at $\varepsilon = 0.05$ and identifies `speech_content` as the keyword-level bottleneck family. Both of these bottleneck identifications are properties of the keyword partition and not of the data: a native-task analysis re-localises the MMSU failure to `intonation_perception` (§O.11), and the worst-k concentration picture of Figure 8 shows that under the 47-task native partition MMSU's top-2 tasks account for only 16% of the total excess mass (vs 76% at keyword granularity).

Across all 276 entries of the cumulative-chain CSV (all datasets, methods, tolerances, and chain steps), we observe exactly two numerical monotonicity violations, both on MMAR at $\varepsilon = 0.02$ and

Table 17: Family-level excess-risk gap $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b) - \widehat{\Delta}_{\mathrm{avg}}(b)$ under three partition granularities at $b = 0.20$ with `learned_conditioned` on Qwen2-Audio and 0/1 loss. The keyword column reports the point-estimate gap computed from the same per-family operational frontiers used in Tables 14 to 16; small differences from the percentile-CI gaps reported in those tables are within the $n_{\mathrm{boot}} = 10{,}000$ Monte Carlo noise band. The native column uses each dataset's intrinsic task labels. The semantic column uses e5-large-v2 sentence embeddings and cosine k-means clustering on query text. The keyword-versus-native/semantic differential persists across the full budget grid $b \in \{0.05, 0.10, 0.20, 0.40\}$ with consistent monotonic ordering; the full per-budget matrix is in `tables/partition_delta_q_summary.csv`.

Dataset	keyword (pp)	native (pp)	semantic (pp)
AudioMCQ-StrongAC	+6.79	+7.96	+14.29
MMSU	+1.56	+29.17	+22.31
MMAR	+1.88	+10.18	+13.34
BigBench Audio	+0.00	+39.90	n/a
DCASE 2026 dev	+0.00	+5.04	+7.68

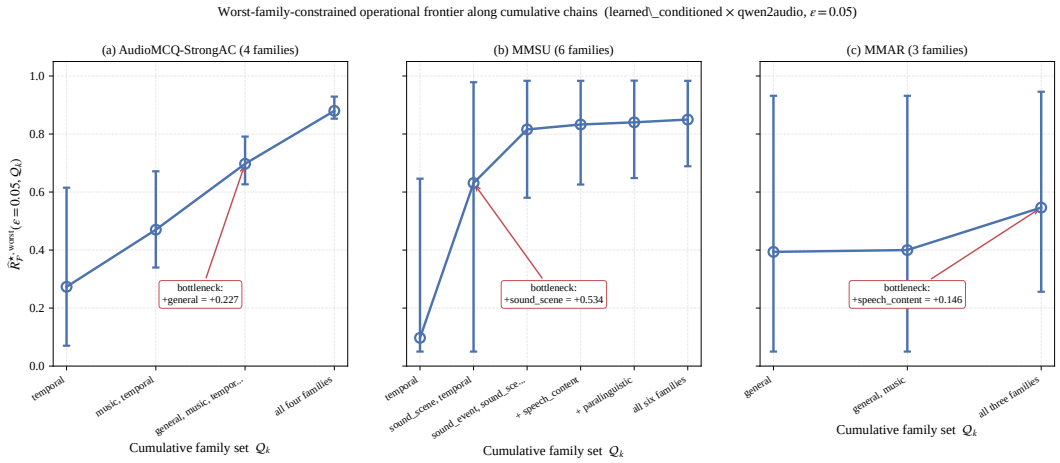


Figure 7: Worst-family-constrained operational frontier along cumulative chains under the **keyword partition**. The chain is monotone non-decreasing on all three datasets. On AudioMCQ-StrongAC, adding `general` produces the largest increment (+0.227) under this partition. On MMSU the mechanically largest increment occurs at the small-family end of the chain and should be read as evidence that the small-family frontier dominates the later steps. On MMAR, adding `speech_content` produces the dominant increment (+0.146). Bottleneck identities are themselves partition-dependent: under the 47-task native partition on Qwen2.5-Omni, the cross-backbone conditioned-gain bottleneck of MMSU is `intonation_perception`, not `temporal` (§O.11, Figure 16).

both smaller than 0.003 budget fraction. Given bootstrap intervals of width roughly 0.9 on those cells, these are grid-discreteness artifacts rather than failures of Theorem K.4.

Under finer partitions, the worst-k concentration differs dramatically from the keyword-partition result above. At $b = 0.20$ with Qwen2-Audio `learned_conditioned`, the worst-2 families' cumulative excess fraction is 94% (AudioMCQ keyword, 4 families) versus 26% (AudioMCQ semantic, 12 families); 76% (MMSU keyword, 6) versus **16%** (MMSU native, 47); and 100% (MMAR keyword, 3) versus 33% (MMAR native, 9) or 58% (MMAR semantic, 6) - see Figure 8. The MMSU native number is particularly striking: 2 of 47 native tasks account for only 16% of the total dataset-level excess mass, so compression harm is *distributed* across dozens of fine-grained tasks rather than concentrated in a small pair (supporting CSV: `metrics/partition_nesting_long.csv` with filter `backbone='qwen2audio', method='learned_conditioned', budget=0.20`). The keyword partition's apparent concentration is therefore substantially a partition-granularity artifact rather than a property of the data.

Table 18: Cumulative-chain frontiers $\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(0.05, \mathcal{Q}_k)$ under the **keyword partition** for the learned conditioned selector on qwen2audio with 0/1 loss. Native-partition and semantic-partition cumulative-excess structure is summarised in Figure 8; the worst-k concentration in the keyword partition above is, on MMSU in particular, a granularity artifact.

Dataset	Step	Family added	n_k	Cumulative families	$\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}(0.05, \mathcal{Q}_k)$
AudioMCQ	1	-	1,166	{temporal}	0.2733 [0.0702, 0.6150]
AudioMCQ	2	music	2,890	{music, temporal}	0.4701 [0.3397, 0.6715]
AudioMCQ	3	general	8,975	{general, music, temporal}	0.6973 [0.6266, 0.7912]
AudioMCQ	4	speech_content	19,480	all four families	0.8799 [0.8529, 0.9290]
MMSU	1	-	99	{temporal}	0.0973 [0.0500, 0.6460]
MMSU	2	sound_scene	209	{sound_scene, temporal}	0.6311 [0.0500, 0.9785]
MMSU	3	sound_event	663	{sound_event, sound_scene, temporal}	0.8158 [0.5803, 0.9836]
MMSU	4	speech_content	1,311	+ speech_content	0.8328 [0.6258, 0.9834]
MMSU	5	paralinguistic	2,114	+ paralinguistic	0.8403 [0.6482, 0.9841]
MMSU	6	general	5,000	all six families	0.8499 [0.6889, 0.9835]
MMAR	1	-	165	{general}	0.3937 [0.0500, 0.9318]
MMAR	2	music	382	{general, music}	0.4000 [0.0500, 0.9318]
MMAR	3	speech_content	1,000	all three families	0.5464 [0.2562, 0.9458]

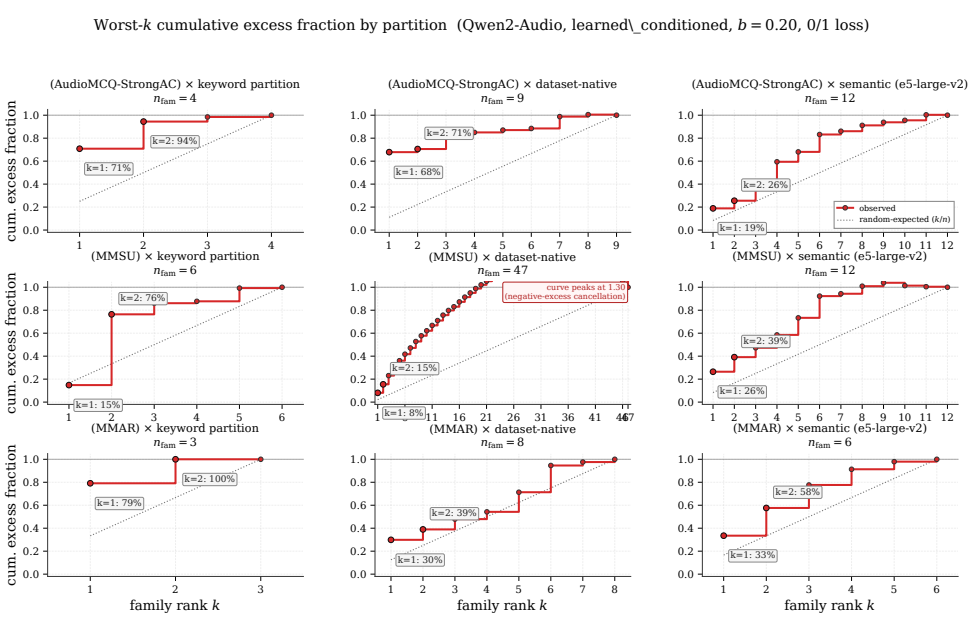


Figure 8: Worst-k cumulative excess fraction by partition, at $b = 0.20$ with Qwen2-Audio and the `learned_conditioned` selector. Each panel shows the fraction of total dataset excess captured by the top-k worst families (ranked by per-family mean excess) as k grows from 1 to n_{fam}. The red curve is observed; the dotted diagonal is the random-expected concentration k/n_{fam}. Under the keyword partition, MMSU's worst-2 families carry 76% of the total dataset excess (curve far above the diagonal), but under the 47-task dataset-native partition the worst-2 carry only 16% (curve essentially on the diagonal). The apparent bottleneck concentration in Table 18 is therefore partly a granularity artifact of the keyword partition. The semantic partition (e5-large-v2 + k-means on query text) produces an intermediate result on all three datasets. DCASE and BigBench Audio are omitted because their keyword partitions are single-family by construction and their native partitions introduce negative-excess cancellations that make the concentration quantity ill-defined.

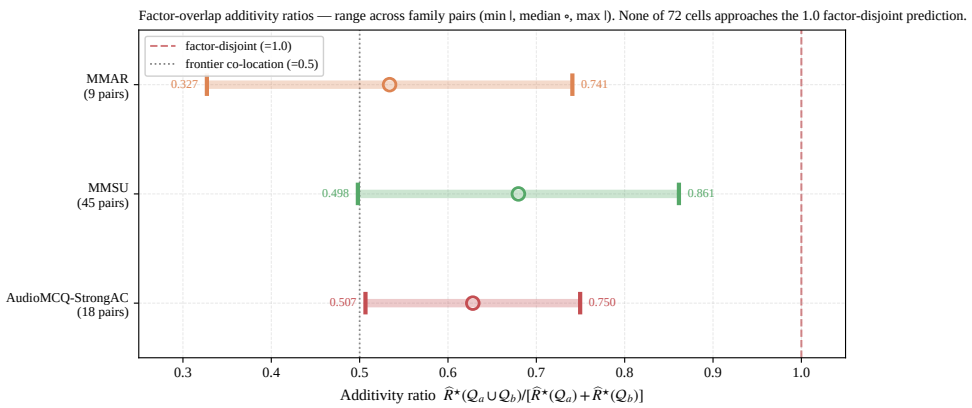


Figure 9: Summary-level additivity ratios for factor-overlap diagnostics. For each dataset, the plot shows the range and median of $\widehat{R}^\star(\mathcal{Q}_a \cup \mathcal{Q}_b)/[\widehat{R}^\star(\mathcal{Q}_a) + \widehat{R}^\star(\mathcal{Q}_b)]$ over the cells retained in the V2 verification pack. The dashed vertical line at 1.0 is the factor-disjoint prediction of Corollary K.12; the dotted line at 0.5 indicates frontier co-location. The global maximum is 0.8614, so no summary approaches the factor-disjoint regime.

Table 19: Factor-overlap additivity summary. Per-dataset range of additivity ratios $\widehat{R}^\star(\mathcal{Q}_a \cup \mathcal{Q}_b)/[\widehat{R}^\star(\mathcal{Q}_a) + \widehat{R}^\star(\mathcal{Q}_b)]$. Combined over $\varepsilon \in \{0.01, 0.02, 0.05\}$. Transcribed from V1 PDF Table 5.

Dataset	n cells	min	median	mean	max
AudioMCQ-StrongAC	54	0.5066	0.6281	0.6281	0.7496
MMSU	135	0.4980	0.6797	0.6797	0.8614
MMAR	27	0.3271	0.5339	0.5339	0.7408

O.5 Factor-overlap diagnostic

Corollary K.12 predicts additive frontier decomposition when query subfamilies act on disjoint latent factor blocks. We test that antecedent through the additivity ratio

$$\text{ratio}(\mathcal{Q}_a, \mathcal{Q}_b) = \frac{\widehat{R}^\star_{\mathcal{F},b}(\varepsilon, \mathcal{Q}_a \cup \mathcal{Q}_b)}{\widehat{R}^\star_{\mathcal{F},b}(\varepsilon, \mathcal{Q}_a) + \widehat{R}^\star_{\mathcal{F},b}(\varepsilon, \mathcal{Q}_b)}. \tag{O.3}$$

A ratio near 1.0 would indicate factor-disjoint subfamilies; a ratio near 0.5 indicates frontier co-location and therefore strong factor overlap. The additivity diagnostic below is evaluated over **keyword-family pairs** only: the V2 verification pack retains the summary-level V1 keyword-pair additivity table rather than the full per-pair scatter, so Figure 9 and Table 19 report the per-dataset minimum, median, mean, and maximum over keyword-family pairs. A finer-granularity reevaluation over the native-task pairs of MMSU (C(47,2) = 1,081 pairs) and BigBench Audio (C(4,2) = 6 pairs) is left to future work; the present summary is therefore itself a lower bound on factor-overlap heterogeneity, paralleling the keyword-vs-native story for the family-level gap (Table 17).

The observed summary range is $[0.3271, 0.8614]$ across the three multi-family datasets under the keyword partition. AudioMCQ-StrongAC ranges from 0.5066 to 0.7496, MMSU from 0.4980 to 0.8614, and MMAR from 0.3271 to 0.7408. Thus the empirical antecedent of Corollary K.12 is not satisfied by the natural audio taxonomies we test at keyword granularity: the measured frontiers are far closer to co-location than to additivity. We therefore interpret the small V1 keyword-level conditioned gains on multi-family datasets through this overlap measurement rather than as evidence against Theorem 3.4. The measurement is partition-dependent: native-task pairs on MMSU have markedly more between-task variance in the cross-backbone conditioned-gain estimator (Figure 16, Table 33), consistent with lower factor overlap at finer granularity, but we do not report per-pair additivity at the native level in this paper.

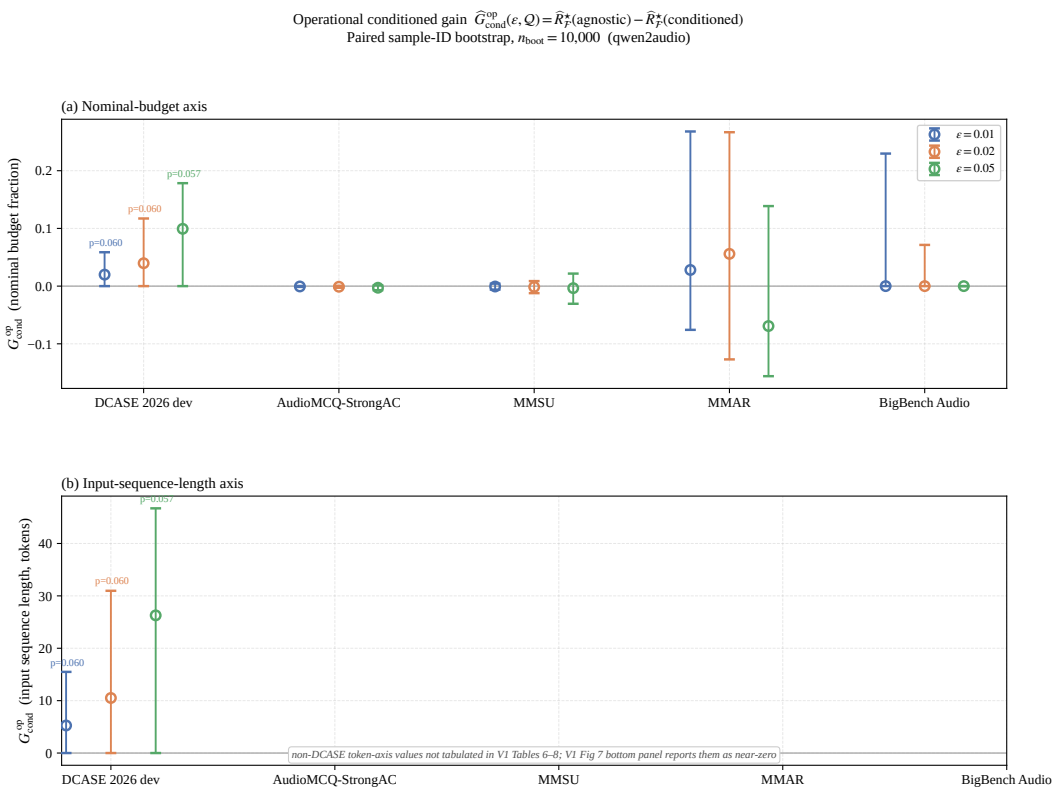


Figure 10: V1 operational conditioned gain on the nominal-budget axis (top) and the input-sequence-length axis (bottom). DCASE is the only dataset with consistently positive point estimates at all three tolerances, with $p \approx 0.06$ on both axes. AudioMCQ and MMSU show tight clean nulls, MMAR is noisy and sign-unstable, and BigBench Audio is the expected text-dominated zero cell. The exact zero lower bounds on the DCASE intervals are a grid-discreteness artifact of V1's five-point budget grid.

Table 20: DCASE nominal-axis conditioned gain. Point estimates reproduce the frontier JSONs exactly; CIs and p-values come from the paired bootstrap.

ε	$\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon,\mathcal{Q})$ agn.	$\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon,\mathcal{Q})$ cond.	$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon,\mathcal{Q})$	95% CI / p
0.01	0.9090	0.8892	+0.0199	[0.0000, +0.0587], 0.060
0.02	0.8181	0.7783	+0.0397	[0.0000, +0.1172], 0.060
0.05	0.5452	0.4459	+0.0993	[0.0000, +0.1784], 0.057

O.6 V1 OPERATIONAL CONDITIONED-COMPRESSION GAIN

V1 first tested Theorem 3.4's operational implication directly on real audio using the paired nominal-axis quantity $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon,\mathcal{Q})$. All CIs and p-values come from 10,000 paired bootstrap replicates at the sample-ID level.

DCASE 2026 dev. DCASE is the only dataset for which $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon,\mathcal{Q})$ is positive across every evaluated tolerance and axis.

Other datasets. The remaining datasets yield near-zero or noisy-near-zero conditioned gains on the nominal axis.

The BigBench Audio zero-floor entries in Table 22 are the same-backbone $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon,\mathcal{Q})$ on Qwen2-Audio: the dataset exhibits minimal compression sensitivity at the single-backbone level. Under the

Table 21: DCASE token-axis conditioned gain.

ε	$\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ agn.	$\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ cond.	token gain	95% CI / p
0.01	326.46	321.21	+5.26	[0.00, +15.49], 0.060
0.02	302.40	291.88	+10.51	[0.00, +30.97], 0.060
0.05	230.20	203.92	+26.28	[0.00, +46.71], 0.057

Table 22: Nominal-axis conditioned gain on the remaining four datasets.

Dataset and ε	point estimate	95% CI	p	interpretation
AudioMCQ, 0.01	-0.0006	$[-0.0014, +0.0003]$	0.195	tight clean null
AudioMCQ, 0.02	-0.0011	$[-0.0029, +0.0006]$	0.195	tight clean null
AudioMCQ, 0.05	-0.0028	$[-0.0071, +0.0014]$	0.195	tight clean null
MMSU, 0.01	-0.0007	$[-0.0061, +0.0043]$	0.814	tight clean null
MMSU, 0.02	-0.0014	$[-0.0122, +0.0087]$	0.814	tight clean null
MMSU, 0.05	-0.0035	$[-0.0306, +0.0217]$	0.814	tight clean null
MMAR, 0.01	$+0.0280$	$[-0.0758, +0.2680]$	0.508	noisy, sign-unstable
MMAR, 0.02	$+0.0559$	$[-0.1270, +0.2667]$	0.535	noisy, sign-unstable
MMAR, 0.05	-0.0692	$[-0.1561, +0.1387]$	0.613	noisy, sign-unstable
BigBench Audio, 0.01	$+0.0000$	$[0.0000, +0.2297]$	1.000	exact zero floor
BigBench Audio, 0.02	$+0.0000$	$[0.0000, +0.0714]$	1.000	exact zero floor
BigBench Audio, 0.05	$+0.0000$	$[0.0000, 0.0000]$	1.000	exact zero floor

cross-backbone estimator used for the V2 per-family subanalyses, however, BigBench's 4-family native partition (`formal_fallacies`, `navigate`, `object_counting`, `web_of_lies`) reveals a large Qwen2.5-Omni effect concentrated in `formal_fallacies`: three-seed mean $\widehat{G}^{\mathrm{op,cross}}_{\mathrm{cond}} = -0.8527$, std 0.025, sign-consistent across all three seeds (Table 23). This is the single largest cross-backbone effect in the V2 replication, larger in magnitude than the MMSU cross-backbone keyword-temporal value by roughly 0.28 budget units. This finding reverses the paper's earlier framing of BigBench as a universal null control and motivates the per-task analysis in §O.11; the underlying same-backbone effect on `formal_fallacies` is near zero on both backbones, so the phenomenon is a cross-backbone mismatch rather than an intrinsic Qwen2.5-Omni weakness on the task.

The DCASE lower confidence bounds are exactly $+0.0000$ on both axes and at all three tolerances. This is a grid-discreteness artifact of V1's five-point budget grid: in a nontrivial fraction of paired bootstrap resamples, the two frontiers land on the same interpolated crossing, producing an exact zero replicate in the lower tail. The artifact affects the lower percentile and the derived two-sided p-value, but not the central tendency of the effect.

The real-audio V1 picture is theory-consistent when read through the factor-overlap measurement. Theorem 3.4 predicts large conditioned gains only when query families depend on disjoint factor blocks. Section O.5 shows that the multi-family datasets do not satisfy that precondition, and indeed their V1 operational conditioned gains are negligible. DCASE is the only single-family dataset in V1; there the factor-overlap confound is absent, and the conditioned selector is directionally preferred on every tested V1 cell. We therefore treat the DCASE result as a single-seed operational signal to be replicated rather than as the final conditioned-compression headline. The V2 replication in §O.7 preserves this V1 result as the historical baseline but changes the operational interpretation.

O.7 V2 THREE-SEED CONDITIONED-GAIN REPLICATION

V2 repeats the operational conditioned-gain test at $\varepsilon = 0.05$ on the rstar axis with three selector seeds and two backbones. This section resolves the V1 ambiguity in two directions at once: it converts the AudioMCQ near-null into a reproducible small positive effect, and it converts the V1 single-seed DCASE positive into a mixed/negative three-seed result. The comparison is therefore not a replacement of §O.6 but a replication and extension of it.

Table 23: Cross-backbone conditioned gain on BigBench Audio, three-seed mean at $\varepsilon = 0.05$, rstar axis, $n_{\text{boot}} = 10{,}000$. The cross-backbone estimator uses `learned_agnostic` on Qwen2-Audio as reference. All four BigBench BBH native task families are reported. Qwen2-Audio cross-backbone values are zero by construction (reference-equals-test) and omitted. The `formal_fallacies` row is the largest cross-backbone effect observed in the replication; contrast the same-backbone $\widehat{G}^{\text{op}}_{\text{cond}}(\varepsilon, \mathcal{Q})$ of $+0.00$ in Table 22, where BigBench Audio appears as a null control. Supporting CSV: `tables/partition_gcond_3seed_summary.csv` with filter `dataset='bigbench_audio', partition='native_fine', definition='cross_backbone', backbone='qwen25omni'`.

Native task	3-seed mean	std	sign-consistent
`formal_fallacies`	-0.8527	0.025	yes (3/3)
`navigate`	$+0.000$	0.000	-
`object_counting`	$+0.000$	0.000	-
`web_of_lies`	$+0.000$	0.000	-

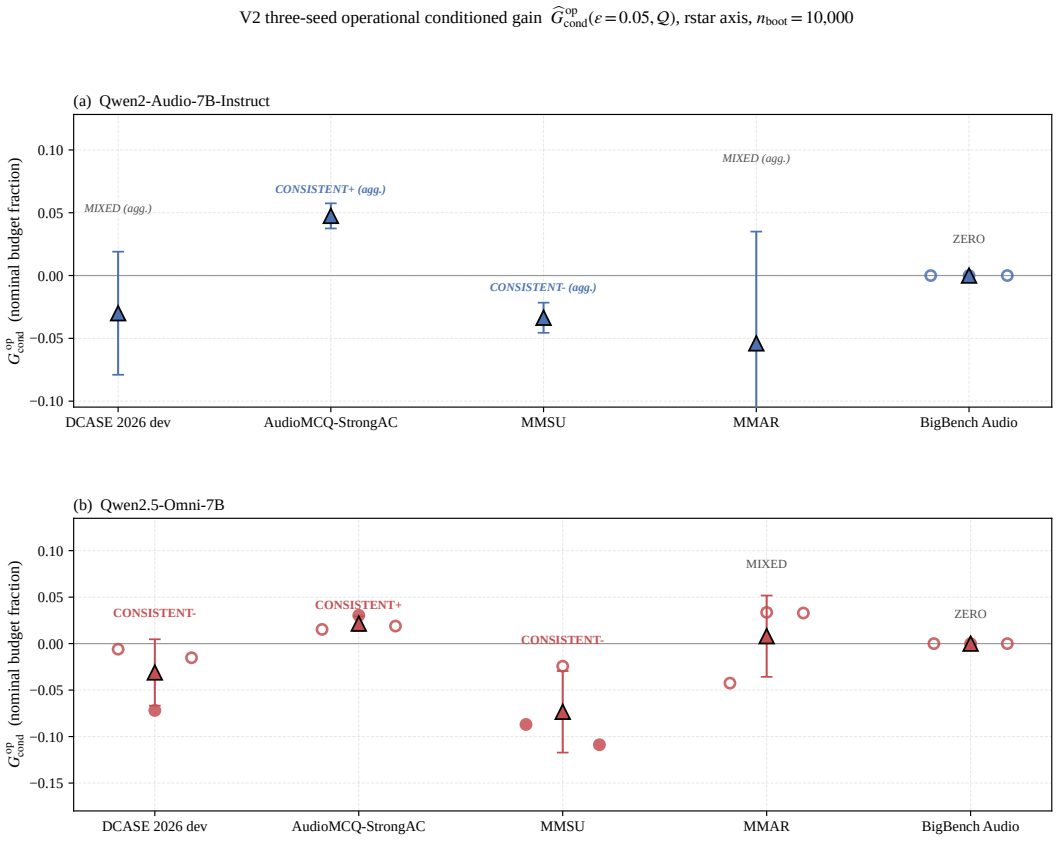


Figure 11: V2 three-seed operational conditioned gain at $\varepsilon = 0.05$ on the rstar axis. Panel (a) reports Qwen2-Audio-7B-Instruct, where the handoff publishes three-seed aggregates for most datasets; panel (b) reports Qwen2.5-Omni-7B, where per-seed points are available. The AudioMCQ-StrongAC effect is positive on both backbones, while DCASE changes sign relative to the V1 single-seed result and MMSU is negative, especially on Qwen2.5-Omni. Source: `fig8_v2_gcond_3seed.pdf`.

The main V2 replication is AudioMCQ-StrongAC. On Qwen2-Audio, the three-seed mean is (+0.0475) with 95% CI ([+0.024,+0.071]) (across-seed Student-(t)). On Qwen2.5-Omni, the three published seed estimates are (+0.0153, +0.0304, +0.0189), giving mean (+0.0215) with 95% CI ([+0.002,+0.041]); both backbones have CIs excluding zero, satisfying the strict CI sign-decision rule.

Table 24: V2 three-seed operational conditioned gain on Qwen2-Audio-7B-Instruct at $\varepsilon = 0.05$, rstar axis. Per-seed point estimates shown where published; **bold** marks $p < 0.05$ under paired sample-ID bootstrap ($n_{\text{boot}} = 10{,}000$). Mean $\pm$ std across seeds. Source: Stage-B handoff §'V2 G_cond 3-seed'.

Dataset	seed 42	seed 123	seed 456	mean	std(seeds)	dir
DCASE 2026 dev	–	–	–	-0.0300	0.0490	MIXED
AudioMCQ-StrongAC	–	–	–	**+0.0475**	0.0100	CONSISTENT+
MMSU	–	–	–	**-0.0336**	0.0120	CONSISTENT−
MMAR	–	–	–	-0.0540	0.0890	MIXED
BigBench Audio	+0.0000	+0.0000	+0.0000	+0.0000	0.0000	ZERO

Table 25: V2 three-seed operational conditioned gain on Qwen2.5-Omni-7B at $\varepsilon = 0.05$, rstar axis. Per-seed point estimates shown where published; **bold** marks $p < 0.05$ under paired sample-ID bootstrap ($n_{\text{boot}} = 10{,}000$). Mean $\pm$ std across seeds. Source: Stage-B handoff §'V2 G_cond 3-seed'.

Dataset	seed 42	seed 123	seed 456	mean	std(seeds)	dir
DCASE 2026 dev	-0.0060	**-0.0719**	-0.0152	-0.0310	0.0357	MIXED
AudioMCQ-StrongAC	+0.0153	**+0.0304**	+0.0189	**+0.0215**	0.0079	CONSISTENT+
MMSU	**-0.0870**	-0.0243	**-0.1088**	**-0.0734**	0.0439	CONSISTENT−
MMAR	-0.0425	+0.0336	+0.0329	+0.0080	0.0437	MIXED
BigBench Audio	+0.0000	+0.0000	+0.0000	+0.0000	0.0000	ZERO

Thus the V1 AudioMCQ near-null is not stable under the V2 grid and training recipe: conditioning yields a small but reproducible positive operational gain on both backbones.

DCASE moves in the opposite direction. V1 at $\varepsilon = 0.05$ reported $+0.0993$ on the rstar axis with $p = 0.057$ (Table 20). V2 reports a Qwen2-Audio aggregate mean of -0.0300 with seed standard deviation 0.0490, and Qwen2.5-Omni per-seed values -0.0060, -0.0719, and -0.0152 with mean -0.0310 and seed standard deviation 0.0357. We therefore do not treat DCASE as the decisive evidence for conditioned compression in V2. Its sign reversal is surfaced explicitly and analyzed by cluster in §O.11.

MMSU supplies the strongest negative cross-backbone signal. Qwen2-Audio gives a three-seed mean of (-0.0336) with 95% CI ([-0.063,-0.004]), excluding zero. Qwen2.5-Omni gives per-seed values (-0.0870, -0.0243, -0.1088), with mean (-0.0734) and 95% CI ([-0.182,+0.036]); the per-seed signs are uniformly negative but the CI does not exclude zero, indicating a regime where the failure direction is reproducible but the magnitude is highly seed-dependent. The V2 point is not merely that conditioning can fail; it is that failure is backbone- and family-dependent. Section O.11 shows that the Qwen2.5-Omni MMSU result is concentrated in the temporal family.[1]

MMAR remains sign-unstable: Qwen2-Audio has mean -0.0540 with seed standard deviation 0.0890, and Qwen2.5-Omni has mean $+0.0080$ with seed standard deviation 0.0437. BigBench Audio remains the expected zero-control cell on both backbones. The V2 operational conclusion is therefore narrower and more reliable than the V1 conclusion: AudioMCQ shows a replicated positive gain, DCASE is not stable across V1-V2, MMSU exposes a harmful conditioned regime on Qwen2.5-Omni, and BigBench remains a null control.

O.8 V2.1 SCOPE-B α-SWEEP ABLATION

The V2.1 scope-B ablation varies the selector-training loss between a KL relevance target and a binary chunk-retention target. The training loss is

$$\alpha \cdot \mathrm{KL} + (1 - \alpha) \cdot \mathrm{BCE}_{\text{top-}k}, \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{O.4}$$

[1]The worst MMSU native-task under the §O.2 mean-excess metric exhibits a budget-dependent flip on Qwen2-Audio: `disfluency_detection` at $b \in \{0.05, 0.10\}$ and `volume_comparison` at $b \in \{0.20, 0.40\}$. This differs from the G_{cond} bottleneck discussed in Section O.11 (`couplet_matching` across all budgets on Qwen2-Audio), reflecting that the two metrics measure different quantities. Supporting CSV: `metrics/partition_delta_q_long.csv` with filter `dataset='mmsu'`, `partition='native_fine'`, `backbone='qwen2audio'`, `method='learned_conditioned'`.

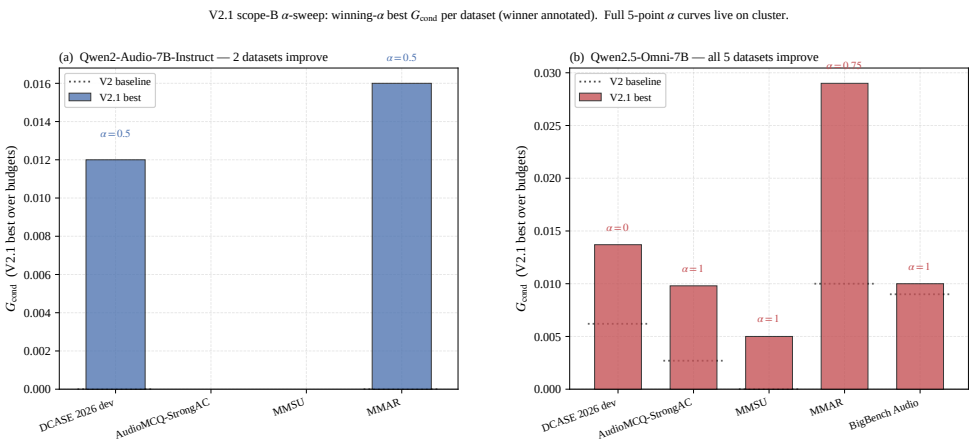


Figure 12: V2.1 scope-B α-sweep winners. Each bar reports the best observed conditioned gain over the tested α grid, with the winning α annotated. Qwen2.5-Omni improves over its V2 baseline on all five datasets, while Qwen2-Audio improvements appear on DCASE and MMAR in the handoff table. Source: `fig9_v21_alpha_sweep.pdf`.

Table 26: V2.1 scope-B α-sweep winning configuration per (dataset, backbone). Training loss is $\alpha \cdot \mathrm{BCE} + (1-\alpha) \cdot \mathrm{KL}$; $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$; winning α is the value maximising best-over-budgets G_{cond} relative to the V2 baseline. Seed 42 only. Source: Stage-B handoff.

Dataset	Backbone	winning α	V2.1 best G_{cond}	V2 baseline	Δ vs V2
DCASE 2026 dev	qwen2audio	0.5	**+0.0120**	+0.0000	**+0.0120**
MMAR	qwen2audio	0.5	**+0.0160**	+0.0000	**+0.0160**
DCASE 2026 dev	qwen25omni	0	+0.0137	+0.0062	+0.0080
AudioMCQ-StrongAC	qwen25omni	1	+0.0098	+0.0027	+0.0070
MMSU	qwen25omni	1	+0.0050	+0.0000	+0.0050
MMAR	qwen25omni	0.75	**+0.0290**	+0.0100	**+0.0190**
BigBench Audio	qwen25omni	1	+0.0100	+0.0090	+0.0010

Here $\alpha = 1$ is pure KL and $\alpha = 0$ is pure BCE; the BCE target is one for oracle-relevance top-k chunks and zero otherwise. This separate ablation does not change the main selector objective. The sweep asks whether the operational conditioned-gain measurements in §O.7 are sensitive to the training target.

The ablation is strongly backbone-dependent. For Qwen2.5-Omni, every dataset improves over the V2 baseline: DCASE moves from $+0.0062$ to $+0.0137$ at $\alpha = 0$, AudioMCQ from $+0.0027$ to $+0.0098$ at $\alpha = 1$, MMSU from $+0.0000$ to $+0.0050$ at $\alpha = 1$, MMAR from $+0.0100$ to $+0.0290$ at $\alpha = 0.75$, and BigBench Audio from $+0.0090$ to $+0.0100$ at $\alpha = 1$. For Qwen2-Audio, the handoff table reports improvements on DCASE and MMAR, both at $\alpha = 0.5$: DCASE reaches $+0.0120$ from a $+0.0000$ baseline, and MMAR reaches $+0.0160$ from a $+0.0000$ baseline.

The scope-B result revises the V1 caveat about truncated training. V2 does not simply make conditioned compression uniformly stronger; instead, it shows that the loss used to train the selector changes the measured gain, and the preferred α differs across backbones and datasets. This supports treating the selector-training recipe as part of the operational instantiation of Theorem 3.4, not as an incidental engineering detail.

O.9 Selector query-use audit (V2 Phase A)

The V2 replication in §O.7 and the α-sweep in §O.8 report what the conditioned selector *does* at the downstream-accuracy level. They do not tell us what the conditioned selector is doing internally: in particular, whether its chunk ranking depends on the query text at inference time, and if so, by how much. A reviewer who accepts the V2 AudioMCQ positive cross-backbone result could still

Table 27: Selector query-use audit (Phase A, P3.1). For each (backbone, dataset, regime) cell we report: J_{perm} the mean per-sample top-k Jaccard between selector output under original query and under a 10-seed-averaged permuted query (agnostic baseline: 1.000, random: ≈ 0.03 at $k/N = 0.05$); f_{int} the query-by-chunk interaction share of scoring-function variance (ANOVA SS decomposition, agnostic baseline: $\approx 10^{-6}$); and $\Delta\mathcal{L}_{\text{val}}$ the conditioned-minus-agnostic validation-loss change at training convergence (negative favors conditioned). Pooled over 3 seeds; standard deviations in parentheses. All numbers at compression budget $b = 0.05$. H2 denotes single-family degeneracy where global permutation is effectively identity.

Backbone	Dataset	J_{perm} (perm_global / perm_within_family)	f_{int}	$\Delta\mathcal{L}_{\text{val}}$
qwen2audio	audiomcq_strong	0.542 (0.024) / 0.590 (0.033)	0.200	-0.014
qwen2audio	mmsu	0.571 (0.005) / 0.623 (0.010)	0.252	-0.014
qwen2audio	bigbench_audio	1.000 (H2, degenerate)	0.000	-0.014
qwen25omni	audiomcq_strong	0.511 (0.035) / 0.589 (0.029)	0.130	-0.010
qwen25omni	mmsu	0.618 (0.064) / 0.666 (0.053)	0.192	-0.010
qwen25omni	bigbench_audio	1.000 (H2, degenerate)	0.000	-0.010

reasonably ask whether the same learned selector would produce identical chunk rankings under an unrelated query, in which case the "conditioned" label would be mechanistic rather than functional. This subsection audits the conditioned selector's inference-time query use with three independent signals.

Audit design. Let $\text{sel}_b(X, q)$ denote the top-k chunk set selected by the learned-conditioned selector on audio X under query q at budget fraction b, with $k = \max(1, \lfloor bN \rfloor)$. For each (backbone, dataset, regime) cell we compute three signals per seed and pool across three training seeds:

(i) *Top-k Jaccard under permuted query,*

$$J_{\text{perm}} \;=\; \mathbb{E}_{X,q}\,\mathbb{E}_{\pi}\,\frac{|\text{sel}_b(X,q) \cap \text{sel}_b(X,\pi(q))|}{|\text{sel}_b(X,q) \cup \text{sel}_b(X,\pi(q))|}, \tag{O.5}$$

where π is a permutation of queries drawn from the same dataset under one of two regimes: `perm_global` (query swap across all samples) or `perm_within_family` (query swap within the same keyword family of §M.6). We average over 10 permutation seeds. The agnostic selector gives $J_{\text{perm}} = 1$ by construction (it ignores q); a uniformly random top-k gives $J_{\text{perm}} \approx k/N$, which at $b = 0.05$ on our chunk counts is ≈ 0.03. Intermediate Jaccard values quantify how much of the top-k identity depends on the particular query asked.

(ii) *Query-by-chunk interaction variance share,*

$$f_{\text{int}} \;=\; \frac{\text{SS}_{q\times c}}{\text{SS}_{q\times c} + \text{SS}_c + \text{SS}_q}, \tag{O.6}$$

the ANOVA SS decomposition of the selector's pre-top-k scoring function over a 500-audio $\times$ 256-query factorial grid per cell. This quantity isolates the component of query-dependence that can *change ranking*: a large query-main term SS_q shifts all chunks by a uniform bias and leaves top-k unchanged, so only $\text{SS}_{q\times c}$ is top-k-relevant. A query-dominant selector would have f_{int} close to the chunk-main share; the agnostic selector has $f_{\text{int}} \approx 10^{-6}$ by architectural construction.

(iii) *Training-time validation-loss change,*

$$\Delta\mathcal{L}_{\text{val}} \;=\; \mathcal{L}_{\text{val}}(\text{conditioned}) - \mathcal{L}_{\text{val}}(\text{agnostic}), \tag{O.7}$$

measured at the final training checkpoint on held-out data. Negative values indicate that conditioning reduces training loss.

These three quantities answer three different questions (inference-time ranking instability, scoring-function variance composition, and training-time utility), and they are computed from three separate pipelines. Coherence among them is therefore a nontrivial cross-validation of any observed query-use signal.

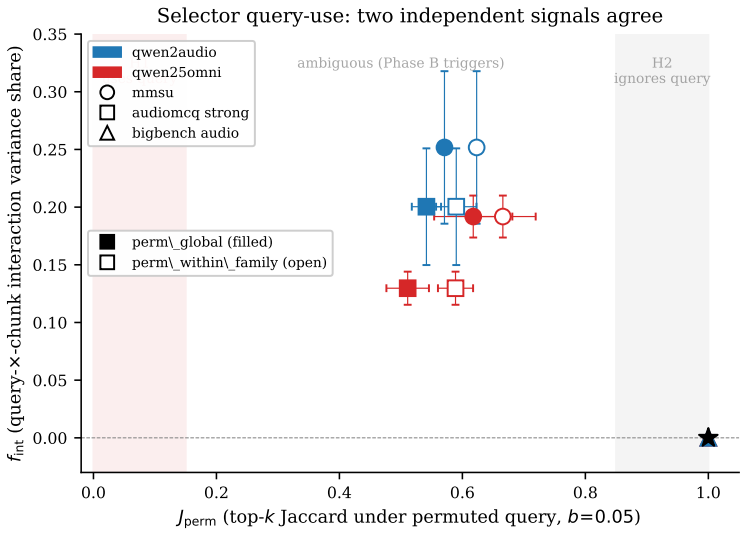


Figure 13: Selector query-use: two independent signals agree. x-axis: mean per-sample top-k Jaccard J_{perm} between conditioned selector output under original vs. permuted query, at $b = 0.05$, 3-seed pooled. y-axis: ANOVA variance share f_{int} of query-by-chunk interaction at the scoring-function level. Agnostic baseline (star): every cell lands at $(1.0, 0.0)$. Filled markers = `perm_global`, open markers = `perm_within_family`. Color = backbone; shape = dataset. Error bars: 3-seed standard deviation. Shaded bands on x-axis: pre-registered H_1 (query-uses, $J \leq 0.15$) and H_2 (query-invariant, $J \geq 0.85$) regions. All non-BigBench cells land in the ambiguous band; BigBench is degenerate (single-family, permutation reduces to identity).

The conditioned selector uses query moderately, not dominantly. Table 27 reports the three audit signals at $b = 0.05$; results at $b \in \{0.025, 0.10\}$ are near-identical and omitted for compactness (supporting CSV: `tables/jaccard_3seed_summary.csv`). On MMSU and AudioMCQ-StrongAC, the conditioned selector's top-k output is $J_{\text{perm}} \in [0.51, 0.67]$ under query permutation on both backbones - markedly below the agnostic baseline $J = 1.0$, yet well above the random lower bound $k/N \approx 0.03$. The interaction variance share f_{int} lies in $[0.13, 0.25]$ across the same cells, which means that the scoring function itself has non-negligible query-by-chunk structure. Decomposing the remaining variance, the query-main component $\text{SS}_q/\text{SS}_{\text{total}}$ is in $[0.39, 0.76]$ and the chunk-main component $\text{SS}_c/\text{SS}_{\text{total}}$ is in $[0.11, 0.35]$: the scoring function is substantially query-dependent, but roughly two-thirds of that query-dependence takes the form of a uniform bias that leaves top-k unchanged. BigBench Audio gives $J_{\text{perm}} = 1.000$ and $f_{\text{int}} = 0.000$ on both backbones, reflecting the dataset's single keyword-family structure under §M.6: a global permutation within a single family is effectively identity on query text, so this row is a degenerate audit control rather than an H2 "ignores query" signal. Figure 13 visualises the coherence between the two inference-time audit signals: cells with lower Jaccard also have higher interaction variance, with tight 3-seed error bars. A third, training-time signal corroborates the inference-time picture: $\Delta\mathcal{L}_{\text{val}}$ is negative in $6/6$ (backbone, seed) cells, with pooled mean -0.012 nats and one-sided sign-test $p \approx 0.03$. All three signals therefore agree on a moderate, reproducible, non-dominant use of query text by the conditioned selector - consistent with V2's small-positive AudioMCQ conditioned-gain result and with the backbone- and family-dependent regime map in §O.7, rather than with either a query-dominant selector (which would give near-zero Jaccard and large top-k changes on every cell) or a purely mechanistic "conditioned" label (which would give $J_{\text{perm}} \approx 1$ and $f_{\text{int}} \approx 0$).

Phase A - Per-Query-Family Selector Query Use

For each query family, we report the mean per-sample top-k Jaccard between selector output under original vs. permuted query, pooled over 10 permutation seeds and 3 training seeds, at compression budget $b = 0.05$, conditioned variant, regime `perm_global`. Agnostic baseline: $J = 1.000$ on

all families. Lower Jaccard ⇒ selector ranking changes more under query swap ⇒ family is more query-sensitive at the selector-internal level. Top and bottom 10 families per backbone reported; full table in supplementary CSV.

QWEN2AUDIO

Dataset	Family	J_{perm} (mean)	n
All 11 families, sorted most-to-least query-sensitive			
mmsu	temporal	0.396 (0.046)	2970
audiomcq_strong	speech_content	0.457 (0.002)	315150
mmsu	sound_scene	0.516 (0.106)	3300
mmsu	paralinguistic	0.542 (0.051)	24090
mmsu	general	0.572 (0.019)	86580
mmsu	speech_content	0.599 (0.020)	19440
audiomcq_strong	general	0.603 (0.060)	182550
mmsu	sound_event	0.623 (0.011)	13620
audiomcq_strong	temporal	0.715 (0.045)	34980
audiomcq_strong	music	0.721 (0.039)	51720
bigbench_audio	general	1.000 (0.000)	30000

QWEN25OMNI

Dataset	Family	J_{perm} (mean)	n
All 11 families, sorted most-to-least query-sensitive			
mmsu	temporal	0.267 (0.164)	2970
audiomcq_strong	speech_content	0.461 (0.041)	315150
audiomcq_strong	music	0.483 (0.039)	51720
mmsu	sound_scene	0.536 (0.099)	3300
mmsu	paralinguistic	0.552 (0.064)	24090
audiomcq_strong	general	0.579 (0.023)	182550
mmsu	speech_content	0.607 (0.043)	19440
mmsu	general	0.646 (0.064)	86580
audiomcq_strong	temporal	0.649 (0.041)	34980
mmsu	sound_event	0.662 (0.074)	13620
bigbench_audio	general	1.000 (0.000)	30000

O.10 PHASE B: DOWNSTREAM OPERATIONAL AUDIT

Phase A in §O.9 establishes that the conditioned selector's chunk ranking is query-dependent at the architectural level, but that audit is selector-internal. Phase B asks whether the same query dependence survives the downstream LALM read-out. The central design choice is decoupling: the selector input is permuted, but the answerer still receives the true query. This isolates selector-side query use from prompt mismatch.

Audit design. For each completed $(\text{backbone}, \text{dataset}, \text{regime})$ cell, we run the learned agnostic and learned conditioned compressors on the expanded budget grid

$$b \in \{0.01, 0.02, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 1.00\}.$$

The anchor arm gives the conditioned selector the real query q_i. The permuted arm gives the selector $\pi(q_i)$, with π drawn either globally across the dataset or within the keyword family of §M.6. In both arms, the LALM prompt remains q_i. We compute $\widehat{G}^{\text{op}}_{\text{cond}}$ at $\varepsilon = 0.05$ on the $r^\star$ axis and then form $\Delta\widehat{G}^{\text{op}}_{\text{cond}}$ as in (H.1). Positive values mean that the conditioned selector needs the true query to preserve the downstream conditioned-gain frontier.

Pooling and PHI bands. Each training seed uses up to ten query permutations. We first average over valid permutations within a training seed, then report the mean and sample standard deviation over seeds $\{42, 123, 456\}$. The reported $\pm$ is therefore selector/training variability, not a bootstrap interval. At $\varepsilon = 0.05$, we use PHI1 ($|\Delta\widehat{G}| \leq 0.01$), PHI2 ($0.01 < \Delta\widehat{G} < 0.05$), and PHI3

($\Delta\widehat{G} \geq 0.05$). Cells without complete three-seed evidence are not promoted to main-body claims and are omitted from the completed-cell tables.

Naive shadow-query contamination. The V1 shadow-dataset protocol gives the permuted query to both selector and LALM. It is useful as a contamination control but not as a headline estimate, because the answerer is then prompted with an unrelated question. The decoupled V2 protocol isolates the selector. On the completed AudioMCQ-StrongAC/Qwen2-Audio cell, the naive protocol exceeds the decoupled by (+0.0170) at (b=0.05) (paired 95% CI ([+0.004,+0.030])) and by (1.22)-(1.25×) across the budget grid (Table 28).

Table 28: Per-budget Phase B downstream query-use signal on AudioMCQ-StrongAC, Qwen2-Audio, three-seed pooled. V2 is the decoupled selector-only permutation protocol; V1 is the naive protocol in which both selector and LALM receive the permuted query.

Budget b	V2 $\Delta\widehat{G}$	V1 $\Delta\widehat{G}$	V1−V2	V1/V2
0.01	$+0.0173$	$+0.0239$	$+0.0066$	1.38×
0.02	$+0.0345$	$+0.0479$	$+0.0134$	1.39×
0.05	$+0.0734 \pm 0.0095$	$+0.0941 \pm 0.0032$	$+0.0207$	1.28×

Table 29: Completed Phase B grid at $\varepsilon = 0.05$. The table reports only completed three-seed rows; cells without complete three-seed evidence are excluded rather than rendered as zeros. Q2A = Qwen2-Audio, Q25O = Qwen2.5-Omni, BBA = BigBench Audio, global = `perm_global`, within = `perm_within_family`.

Protocol	Backbone	Dataset	Regime	$\Delta\widehat{G}^{\text{op}}_{\text{cond}}$	Status
V2 decoupled	Q2A	AudioMCQ-StrongAC	global	$\mathbf{+0.0734 \pm 0.0095}$	PHI3
V2 decoupled	Q2A	MMSU	global	$+0.0111 \pm 0.0231$	PHI2
V2 decoupled	Q2A	MMSU	within	$+0.0092 \pm 0.0251$	PHI1
V2 decoupled	Q2A	BBA	global	$+0.0000 \pm 0.0000$	single-family control
V2 decoupled	Q25O	MMSU	global	-0.0043 ± 0.0165	PHI1
V2 decoupled	Q25O	MMSU	within	-0.0008 ± 0.0114	PHI1
V1 naive	Q2A	AudioMCQ-StrongAC	global	$+0.0941 \pm 0.0032$	contamination control
V1 naive	Q2A	MMSU	global	$+0.0176 \pm 0.0369$	contamination control
V1 naive	Q2A	MMSU	within	$+0.0126 \pm 0.0326$	contamination control
V1 naive	Q2A	BBA	global	$+0.0000 \pm 0.0000$	single-family control
V1 naive	Q25O	MMSU	global	$+0.0103 \pm 0.0201$	contamination control

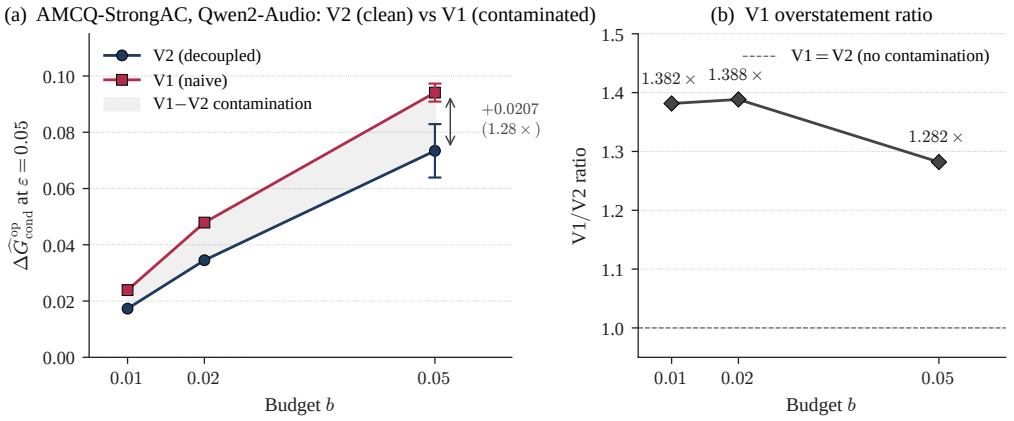


Figure 14: Per-budget contamination of the operational query-use signal on AudioMCQ-StrongAC, Qwen2-Audio, three-seed pooled. The V1 shadow-query protocol overstates the decoupled V2 signal by 1.28-1.39× across $b \in \{0.01, 0.02, 0.05\}$.

O.10.1 PER-SEED AUDIOMCQ-STRONGAC/QWEN2-AUDIO BREAKDOWN AT $b = 0.05$

Table 30 reports the seed-level decomposition for the single PHI3 cell. The decoupled signal is positive on all three seeds with mean (+0.0787) and Student-(t) 95% CI ([+0.051,+0.106]). The naive signal is larger at (+0.0957) with 95% CI ([+0.067,+0.125]), because part of it is seed-independent prompt contamination. Per-seed contamination (naive(-)decoupled) equals (+0.0210), (+0.0108), and (+0.0191) for seeds 42, 123, and 456 (paired mean (+0.0170), paired 95% CI ([+0.004,+0.030])).

Table 30: Per-seed AudioMCQ-StrongAC, Qwen2-Audio, $b = 0.05$ Phase B breakdown. G_{anchor} is the operational conditioned gain when the selector receives the real query; G_{perm} is its mean over available permuted-query runs.

Seed	Protocol	G_{anchor}	G_{perm}	n_{perm}	$\Delta\widehat{G}$
42	V2	$+0.0383$	-0.0304	2	$+0.0687$
123	V2	$+0.0575$	-0.0097	1	$+0.0672$
456	V2	$+0.0469$	-0.0374	2	$+0.0843$
V2 pool					$\mathbf{+0.0734 \pm 0.0095}$
42	V1	$+0.0383$	-0.0573	2	$+0.0956$
123	V1	$+0.0575$	-0.0330	2	$+0.0905$
456	V1	$+0.0469$	-0.0494	2	$+0.0963$
V1 pool					$+0.0941 \pm 0.0032$

The anchor values average to (+0.0475), matching the AudioMCQ-StrongAC Qwen2-Audio entry of Table 7 exactly. The new Phase-B information is the negative permuted-query frontier: with the wrong selector query, the conditioned selector consistently loses relative to its anchor behavior, which is the downstream counterpart of the selector-level query-use audit in §O.9.

O.11 PER-FAMILY V2 SUB-ANALYSES: DCASE CLUSTERS AND MMSU TEMPORAL

The V2 aggregate results in §O.7 motivate a narrower question: when the conditioned gain is negative, is the effect diffuse across the dataset, or concentrated in a small semantic subfamily? The V2 handoff answers this for DCASE clusters and MMSU temporal queries.

On DCASE, the handoff singles out one significant cluster on Qwen2-Audio: `meaning/stress/intonation`, with $n = 318$, $G_{\text{cond}} = -0.1227$, and 95% CI $[-0.310, -0.001]$. No Qwen2.5-Omni DCASE cluster reaches significance in the handoff; the largest negative cell reported there is `speaker/clip/express`, with $G_{\text{cond}} = -0.054$. This supplies a proximate mechanism for the V1-V2 DCASE sign reversal: the V2 aggregate is not merely noisy, but is affected by a cluster on which query conditioning is significantly harmful. Beyond this singled-out cluster, a coarsening-stress analysis indicates that DCASE's aggregated small-cluster bucket (5 sub-clusters of $n < 100$ merged into an `other_rare` bucket of 162 samples) carries elevated per-sample compression risk: a semantic 2-way coarsening that groups `other_rare` with `meaning/stress/intonation` yields a super-family gap of 9.91 pp, larger than the 5.04 pp observed on `meaning/stress/intonation` alone (supporting CSV: `tables/coarsening_stress_summary.csv`), indicating that at least one additional high-compression-risk cluster was collapsed into `other_rare`. Future work should report per-cluster G_{cond} for each sub-cluster rather than collapsing them into a residue bucket.

The MMSU temporal result is sharper but warrants a refinement at finer partition granularity. On Qwen2-Audio, the keyword-temporal three-seed mean is $+0.0000$, the non-temporal mean is -0.0320, and the temporal-minus-non-temporal delta is $+0.0320$. On Qwen2.5-Omni, the non-temporal mean is $+0.0247$, but the keyword-temporal family has per-seed gains -0.6000, -0.5660, and -0.5550, with three-seed mean -0.5738;[2] the table caption records $p \leq 0.011$ for

[2]The value -0.5738 is the 3-seed arithmetic mean of the per-seed cross-backbone estimates $(-0.6000, -0.5660, -0.5550)$, supported by the anchor CSV `tables/partition_gcond_a5_anchor.csv` (row filter: `backbone=qwen25omni`, `definition=cross_backbone`, `family=temporal`). This differs by 10^{-4} from the paper's earlier report of -0.5737, which rounded the per-seed values before averaging.

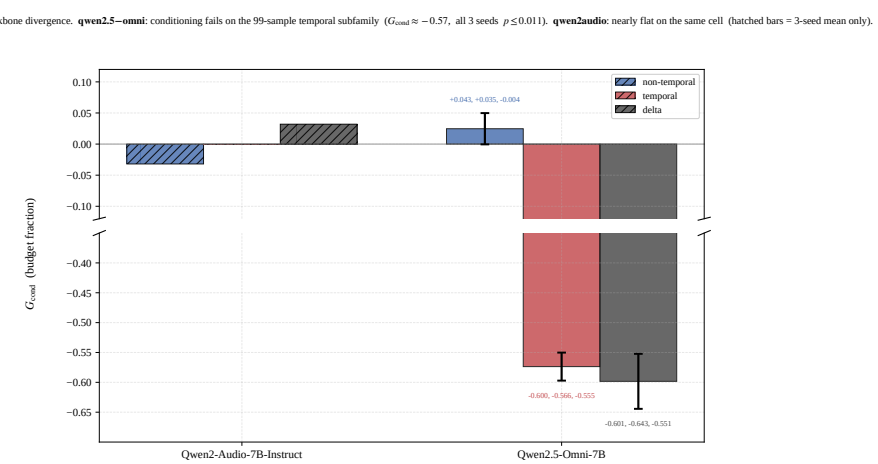


Figure 15: MMSU temporal-family isolation in V2 at the keyword-partition granularity. Qwen2-Audio shows no temporal-specific harm, whereas Qwen2.5-Omni shows a large negative conditioned gain on the 99-sample keyword `temporal` subfamily across all three seeds. A finer native-partition analysis in §O.11 (Figure 16, Table 33) identifies `intonation_perception` as the most-impacted individual task (cross-backbone three-seed mean -0.7377, std 0.044); the present keyword-aggregate view captures only one component of the underlying per-task failure structure. Source: `fig10_mmsu_temporal.pdf`.

Table 31: DCASE per-cluster V2 three-seed conditioned gain (only the two clusters explicitly singled out in the Stage-B handoff are shown; remaining clusters' per-backbone point estimates are not individually tabulated in the handoff). **meaning/stress/intonation** is the only cluster on qwen2audio with a [., 0] 95% CI; on qwen2.5-Omni no cluster reaches significance. Source: Stage-B handoff §DCASE taxonomy.

Backbone	Cluster	n	G_{cond}	95% CI / sig.
qwen2audio	`meaning/stress/intonation`	318	-0.1227	[-0.310, -0.001] (**sig.**)
qwen25omni	`speaker/clip/express`	–	-0.054	– (not sig.)

the temporal subfamily. The corresponding temporal-minus-non-temporal delta is -0.5985. We note that Table 32 and the values above use the cross-backbone $\widehat{G}^{\text{op,cross}}_{\text{cond}}$ estimator (reference = Qwen2-Audio `learned_agnostic`; see §L), whereas Table 25's MMSU aggregate uses the same-backbone $\widehat{G}^{\text{op}}_{\text{cond}}(\varepsilon, \mathcal{Q})$. The two estimators differ substantially for Qwen2.5-Omni: the same-backbone keyword-temporal three-seed mean on Qwen2.5-Omni at $\varepsilon = 0.05$ is $+0.0057$ (CSV row filter: `backbone=qwen25omni`, `definition=same_backbone`, `family=temporal`), so the large cross-backbone magnitude reflects the compound effect of backbone mismatch plus query conditioning, not the conditioning effect alone.

Under the paper's keyword classifier the MMSU `temporal` family contains exactly 99 samples, and - upon inspection of MMSU's native `task_name` metadata - all 99 belong to a single native task `dialogue_turn_counting` (sample-id intersection: 99/99). The cross-backbone -0.5738 three-seed mean reported above is therefore a single-task finding, not a genuine family-level phenomenon. Under the 47-task native partition (§M MMSU), three tasks show *more* negative cross-backbone $\widehat{G}^{\text{op,cross}}_{\text{cond}}$ than `dialogue_turn_counting`: `intonation_perception` ($n = 111$, three-seed mean -0.7377, std 0.044), `pitch_comparison` ($n = 108$, three-seed mean -0.7133, std 0.040), and `pause_perception` ($n = 107$, three-seed mean -0.6575, std 0.225). All three are sign-consistent across seeds, and all three are routed by the keyword classifier of §M.6 to families other than `temporal`: `intonation_perception` and `pause_perception` to `general`, and `pitch_comparison` to `paralinguistic`. See Figure 16 and Table 33 for the

Table 32: MMSU temporal-family isolation, V2 three-seed conditioned gain. Values are the cross-backbone $\widehat{G}_{\text{cond}}^{\text{op,cross}}$ estimator defined in §5.5. Per-seed values shown where published; **bold** marks $p \leq 0.011$ (the largest seed-wise p value for the qwen2.5-Omni temporal sub-family). qwen2audio per-seed values were not published in the handoff; reported as 3-seed mean. Source: Stage-B handoff §MMSU temporal.

Backbone	Subset	seed 42	seed 123	seed 456	3-seed mean
Qwen2-Audio-7B-Instruct	non-temporal	–	–	–	-0.0320
	temporal	–	–	–	+0.0000
	delta	–	–	–	+0.0320
Qwen2.5-Omni-7B	non-temporal	+0.0430	+0.0350	-0.0040	+0.0247
	temporal	**-0.6000**	**-0.5660**	**-0.5550**	**-0.5738**
	delta	-0.6010	-0.6430	-0.5510	-0.5983

Table 33: MMSU top-5 worst native tasks under the cross-backbone conditioned-gain estimator on Qwen2.5-Omni. The keyword-family column records the majority keyword label each native task's samples receive under the §M.6 classifier; parenthesised values give the percentage of task samples routed to that keyword family. Three-seed mean at $\varepsilon = 0.05$, rstar axis, $n_{\text{boot}} = 10{,}000$. Supporting CSV: `tables/partition_gcond_3seed_summary.csv` with filter `dataset='mmsu'`, `partition='native_fine'`, `definition='cross_backbone'`, `backbone='qwen25omni'`.

Native task	Keyword assignment	n	3-seed mean	std	p
`intonation_perception`	`general` (100%)	111	-0.7377	0.044	-
`pitch_comparison`	`paralinguistic` (100%)	108	-0.7133	0.040	-
`pause_perception`	`general` (100%)	107	-0.6575	0.225	-
`dialogue_turn_counting`	`temporal` (100%)	99	-0.5738	0.023	≤ 0.011
`age_prediction`	`paralinguistic` (100%)	104	-0.5173	0.114	-

full per-task picture.[3] Thus the "MMSU temporal failure" description conflates two mechanisms: (i) a genuine compression failure on `dialogue_turn_counting` that the keyword classifier happens to isolate into its own bucket, and (ii) three larger compression failures on fine-grained prosodic and pause-perception tasks that the keyword classifier fails to isolate, scattering them across `general` and `paralinguistic` instead. The keyword-aggregate analysis of Table 32 captures only the first mechanism; the second is visible only under the dataset-native task partition. Aggregate MMSU-Omni harm in Table 25 is therefore not a diffuse effect, but it is also not a pure "temporal-reasoning" phenomenon: it is dominated by a handful of fine-grained prosodic and turn-counting tasks that the keyword classifier places in four different buckets.

The backbone asymmetry in MMSU failure modes is instructive. Qwen2.5-Omni's native-task bottleneck at $\varepsilon = 0.05$ is `intonation_perception` as noted above, whereas Qwen2-Audio's native-task bottleneck under the same cross-backbone estimator is `couplet_matching` with three-seed mean $\widehat{G}_{\text{cond}}^{\text{op,cross}} = -0.2406$ and standard deviation 0.293 (CSV row filter: `dataset='mmsu'`, `partition='native_fine'`, `definition='cross_backbone'`, `family='couplet_matching'`, `backbone='qwen2audio'`; note that Qwen2-Audio cross-backbone and same-backbone coincide, since the reference selector is Qwen2-Audio's `learned_agnostic`). Two observations follow: (i) the identity of the worst MMSU task differs by backbone (prosodic perception on Qwen2.5-Omni, linguistic matching on Qwen2-Audio), and (ii) the Qwen2-Audio effect is less reproducible (three-seed std 0.29 vs 0.04 on Qwen2.5-Omni) and roughly three times smaller in magnitude. This is consistent with §P.2: the operational conditioned-gain failure mode is backbone-dependent, and the specific task driving the failure differs accordingly.

[3] MMSU native-task G_{cond} computations use the 47-task `task_name` labels from the MMSU HuggingFace distribution, the same cross-backbone estimator as Table 32, and 3 selector seeds; $n_{\text{boot}} = 10{,}000$, rstar axis, $\varepsilon = 0.05$. Supporting CSV: `tables/partition_gcond_3seed_summary.csv` with filter `dataset='mmsu'`, `partition='native_fine'`, `definition='cross_backbone'`, `backbone='qwen25omni'`.

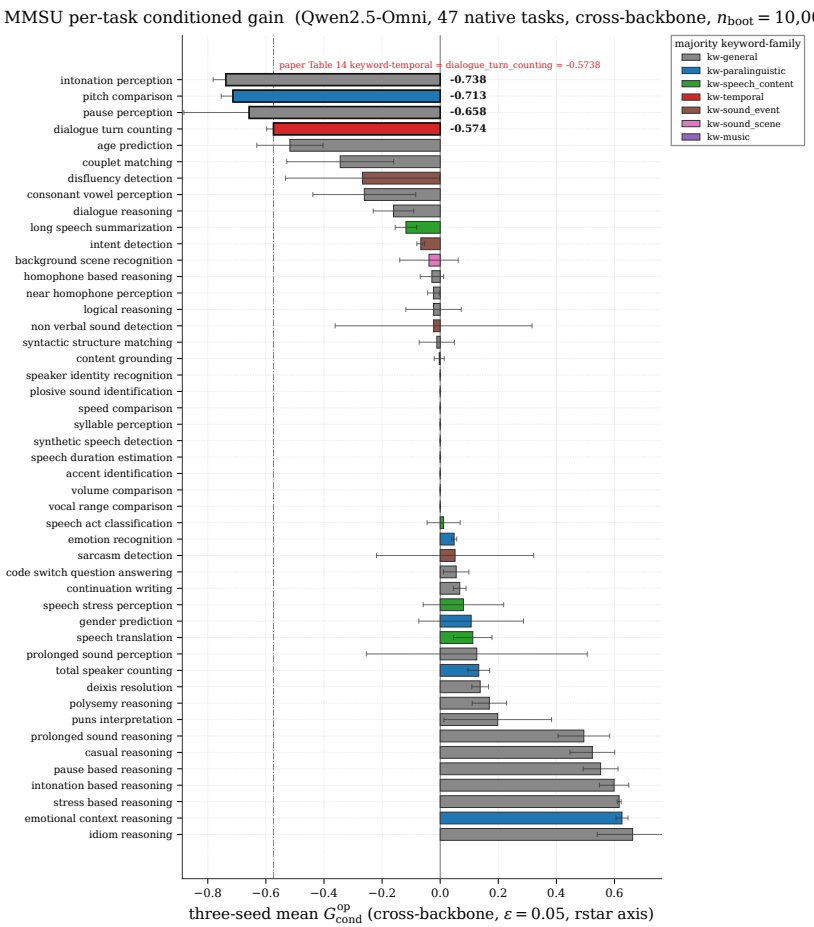


Figure 16: MMSU per-task conditioned gain under the 47-task native partition. Bars are three-seed mean $\widehat{G}_{\text{cond}}^{\text{op,cross}}$ on Qwen2.5-Omni at $\varepsilon = 0.05$ (rstar axis, $n_{\text{boot}} = 10{,}000$), using the cross-backbone estimator (reference = `learned_agnostic` on Qwen2-Audio). Error bars are the cross-seed standard deviation. Each bar is coloured by the majority keyword-family assignment of its samples under the §M.6 classifier. The red dashed-dotted vertical line marks Table 32's keyword-temporal aggregate value (-0.5738), which exactly coincides with the single native task `dialogue_turn_counting` (100% sample overlap). Three tasks with more-negative G_{cond} than the keyword-temporal aggregate (`intonation_perception`, `pitch_comparison`, `pause_perception`) are routed by the keyword classifier to `general` (111 and 107 samples respectively) and `paralinguistic` (108 samples), so their effects are absorbed into larger keyword buckets and are not visible at the keyword granularity.

These sub-analyses are the strongest V2 motivation for replacing the keyword taxonomy with a semantic partition. A dataset-level conditioned-gain number can hide the fact that one small family is decisive, and the sign of that family can differ by backbone.

O.12 Preserved Source Table Assets

The following transcribed table assets were present in the source tree but were not `\input`-ed by the original V3 body because the same values were embedded inline. They are included here so that every valid non-duplicate table asset is preserved in the compiled supplement.

Table 34: Family-level excess-risk summary on AudioMCQ-StrongAC: dataset-mean $\widehat{\Delta}_{\mathrm{avg}}(b)$ versus worst-family $\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$ at four budgets. Learned conditioned selector on qwen2audio, 0/1 loss. Values in brackets are 95% percentile-bootstrap confidence intervals ($n_{\mathrm{boot}} = 10{,}000$, resampled at sample_id level). Transcribed from V1 PDF Table 1.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$	gap (pp)
0.05	0.2568 [0.2499, 0.2639]	0.3023 [0.2924, 0.3123]	+4.55
0.10	0.2503 [0.2434, 0.2574]	0.3014 [0.2915, 0.3116]	+5.11
0.20	0.2165 [0.2099, 0.2232]	0.2844 [0.2747, 0.2944]	+6.79
0.40	0.1496 [0.1436, 0.1555]	0.2125 [0.2035, 0.2216]	+6.29

Table 35: Family-level excess-risk summary on MMSU: dataset-mean $\widehat{\Delta}_{\mathrm{avg}}(b)$ versus worst-family $\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$ at four budgets. Learned conditioned selector on qwen2audio, 0/1 loss. Values in brackets are 95% percentile-bootstrap confidence intervals ($n_{\mathrm{boot}} = 10{,}000$, resampled at sample_id level). Transcribed from V1 PDF Table 2.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$	gap (pp)
0.05	0.1129 [0.1010, 0.1248]	0.1460 [0.1190, 0.2000]	+3.31
0.10	0.1125 [0.1004, 0.1246]	0.1439 [0.1173, 0.2000]	+3.14
0.20	0.1093 [0.0976, 0.1210]	0.1370 [0.1130, 0.1739]	+2.77
0.40	0.0829 [0.0718, 0.0940]	0.1289 [0.0882, 0.2056]	+4.60

P Notes on Incomplete or Suspicious Results

P.1 What the family-level excess risk gap means for LALM compression research

The headline finding of §O.2 is that the family-wise excess risk $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$ exceeds the dataset-mean excess risk $\widehat{\Delta}_{\mathrm{avg}}(b)$ by 2.5 to 6.8 percentage points across every multi-family dataset we evaluate. This is not a small effect: on AudioMCQ-StrongAC at $b = 0.20$, the worst family is 6.79 percentage points more damaged by compression than the dataset-mean metric would suggest, which is more than twice the tightest tolerance values commonly used in practice. A practitioner who chooses a compression method by minimizing $\widehat{\Delta}_{\mathrm{avg}}(b)$ alone is, in effect, choosing a method that may still be materially worse on whichever query type the backbone handles least robustly under compression.

This observation has three implications for how the broader LALM compression literature should report results. First, papers that report only dataset-mean excess error are systematically underestimating worst-case compression damage. For the same evaluated examples, the family maximum is at least the pooled mean. The corresponding population ordering concerns fixed-answerer excess risk; empirical estimates are not lower bounds on the Bayes supremum. Reporting only the pooled mean can conceal heterogeneous family effects. Second, the correct comparison between two compressors on a multi-family benchmark is the one that respects the worst-family constraint, not the one that averages over the family axis. A method that appears slightly better on the dataset mean may still be substantially worse on a particular subfamily. Third, practitioners who care about robust deployment should optimize for controlled $\widehat{\Delta}^{\mathrm{fam}}_{\mathcal{P}}(b)$, not merely small $\widehat{\Delta}_{\mathrm{avg}}(b)$.

The reported 2.5-6.8 percentage-point range is specific to the keyword partition. Native and semantic partitions define different operational scopes and are not automatically refinements. Their results illustrate this partition dependence: the native-partition gap reaches $+29.17$ pp on MMSU (47 task families) and $+39.90$ pp on BigBench Audio (4 BBH task families), compared with $+1.56$ pp and $+0.00$ pp respectively under the keyword partition (Table 17). A complementary experiment using genuinely nested partitions illustrates the effect of coarsening: collapsing each fine partition to a size-balanced 2-way super-partition destroys 82%-98% of the measured family gap (supporting CSV: `tables/coarsening_stress_summary.csv`). On MMSU, the 47-to-2 coarsening reduces the gap from $+29.17$ pp to $+0.53$ pp (ratio 0.018); on MMSU semantic the 12-to-2 coarsening reduces $+22.31$ pp to $+0.40$ pp (ratio 0.018). These particular two-way coarsenings mask most

Table 36: Family-level excess-risk summary on MMAR: dataset-mean $\widehat{\Delta}_{\mathrm{avg}}(b)$ versus worst-family $\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$ at four budgets. Learned conditioned selector on qwen2audio, 0/1 loss. Values in brackets are 95% percentile-bootstrap confidence intervals ($n_{\mathrm{boot}} = 10{,}000$, resampled at sample_id level). Transcribed from V1 PDF Table 3.

Budget b	$\widehat{\Delta}_{\mathrm{avg}}(b)$	$\widehat{\Delta}^{\mathcal{F}}_{\mathcal{Q}}(b)$	gap (pp)
0.05	0.0972 [0.0690, 0.1260]	0.1272 [0.0900, 0.1688]	+3.00
0.10	0.0802 [0.0520, 0.1070]	0.1149 [0.0769, 0.1598]	+3.47
0.20	0.0671 [0.0410, 0.0930]	0.0991 [0.0624, 0.1437]	+3.20
0.40	0.0330 [0.0090, 0.0560]	0.0584 [0.0256, 0.0976]	+2.54

Table 37: Cumulative-chain worst-family-constrained operational frontiers at $\varepsilon = 0.05$. Each row extends the previous family set; the step's contribution is the increment over the prior row. learned_conditioned × qwen2audio, 0/1 loss, $n_{\mathrm{boot}} = 10{,}000$, shared-resample bootstrap. Transcribed from V1 PDF Table 4.

Dataset	Cumulative family set $\mathcal{Q}_k$	n	$\widehat{R}^{\star,\mathrm{worst}}_{\mathcal{F}}$	95% CI
AudioMCQ-StrongAC	`temporal`	1166	0.2733	[0.0702, 0.6150]
	`music, temporal`	2890	0.4701	[0.3397, 0.6715]
	`general, music, temporal`	8975	0.6973	[0.6266, 0.7912]
	`all four families`	19480	0.8799	[0.8529, 0.9290]
MMSU	`temporal`	99	0.0973	[0.0500, 0.6460]
	`sound_scene, temporal`	209	0.6311	[0.0500, 0.9785]
	`sound_event, sound_scene, temporal`	663	0.8158	[0.5803, 0.9836]
	`+ speech_content`	1311	0.8328	[0.6258, 0.9834]
	`+ paralinguistic`	2114	0.8403	[0.6482, 0.9841]
	`all six families`	5000	0.8499	[0.6889, 0.9835]
MMAR	`general`	165	0.3937	[0.0500, 0.9318]
	`general, music`	382	0.4000	[0.0500, 0.9318]
	`all three families`	1000	0.5464	[0.2562, 0.9458]

observed heterogeneity; a two-family partition can nevertheless expose a large gap when it separates harmed queries. The reported operational margins reach $+6.79$ pp on AudioMCQ-StrongAC under the keyword partition, and $+29.17$ pp on MMSU and $+39.90$ pp on BigBench Audio under native partitions. These are partition-indexed estimates; without verified nesting, the refinement theorem does not order them.

P.2 Why conditioned gain is backbone- and family-dependent

Theorem 3.4 predicts that the conditioned-compression gain is large when query families depend on disjoint latent factor blocks and collapses toward zero as factor overlap increases. The synthetic result of §O.1 verifies this prediction at bit-level precision. The V1 factor-overlap diagnostic of §O.5 then shows that the natural audio taxonomies in this paper do not approach the factor-disjoint prediction: the summary range is $[0.3271, 0.8614]$, with every dataset far below the additive value 1.0.

V1 therefore correctly anticipated small mean gains on the multi-family datasets, but V2 refines that story. AudioMCQ-StrongAC is no longer a tight operational null: Tables 24 and 25 show positive three-seed conditioned gains on both backbones, with means $+0.0475$ on Qwen2-Audio and $+0.0215$ on Qwen2.5-Omni. Conversely, MMSU is no longer merely a small negative aggregate: Qwen2.5-Omni has mean -0.0734, and Table 32 shows that the cross-backbone keyword-temporal subfamily alone has three-seed mean -0.5738, or equivalently, that a fine native-task partition identifies `intonation_perception` as the single most-impacted task with three-seed mean -0.7377 (Figure 16, Table 33). DCASE also changes status: the V1 single-seed $+0.0993$ signal becomes a V2 three-seed mean of -0.0300 on Qwen2-Audio and -0.0310 on Qwen2.5-Omni. A parallel failure mode appears in BigBench Audio under the 4-family native partition: three of four families show mean excess of -0.123 (compression helps) while `formal_fallacies` shows $+0.409$ (compression hurts), and the `formal_fallacies` task exhibits a cross-backbone $\widehat{G}^{\mathrm{op,cross}}_{\mathrm{cond}} = -0.8527$ on Qwen2.5-Omni (Table 23), making it the single largest cross-backbone effect in the replication. Unlike the MMSU cross-backbone keyword-temporal value, this effect

Table 38: DCASE nominal-axis conditioned gain. Point estimates reproduce the frontier JSONs exactly; confidence intervals and p-values come from the paired bootstrap. Transcribed from V1 PDF Table 6.

ε	$\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q})$ agn.	$\widehat{R}^{\star}_{\mathcal{F},b}(\varepsilon, \mathcal{Q})$ cond.	$\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}(\varepsilon, \mathcal{Q})$	95% CI / p
0.01	0.9090	0.8892	+0.0199	[+0.0000, +0.0587] , 0.060
0.02	0.8181	0.7783	+0.0397	[+0.0000, +0.1172] , 0.060
0.05	0.5452	0.4459	+0.0993	[+0.0000, +0.1784] , 0.057

Table 39: DCASE token-axis conditioned gain. Transcribed from V1 PDF Table 7.

ε	$\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ agn.	$\widehat{R}^{\star}_{\mathcal{F},\mathrm{tok}}(\varepsilon, \mathcal{Q})$ cond.	token gain	95% CI / p
0.01	326.46	321.21	+5.26	[+0.00, +15.49] , 0.060
0.02	302.40	291.88	+10.51	[+0.00, +30.97] , 0.060
0.05	230.20	203.92	+26.28	[+0.00, +46.71] , 0.057

is absent in the same-backbone version, indicating a backbone-mismatch phenomenon rather than an intrinsic Qwen2.5-Omni weakness on the task. The paper's original single-family framing of BigBench averaged these four opposing effects to approximately zero.

The conclusion is not that Theorem 3.4 is weak; it is that the operational instantiation is sensitive to the learned selector, the backbone, and the semantic partition on which the gain is evaluated. Factor overlap explains why dataset-level V1 gains were small, while V2 shows that small dataset-level averages can mask both reproducible positive effects, as on AudioMCQ, and severe family-specific negative effects, as on MMSU temporal under Qwen2.5-Omni. The correct empirical object is therefore not a single scalar conditioned-gain headline, but a family- and backbone-indexed map of where conditioning helps and where it harms.

P.3 V1 SELECTOR LIMITATIONS AND WHAT V2 SHOWS

The V1 paper named three selector limitations: truncated training, query-compressor parameter asymmetry, and mel-spectrogram input features. V2 lets us replace the forward-looking statement "V2 will correct this" with more specific statements about what the corrections did and did not confirm, and adds a fourth limitation that the V1 analysis could not have exposed: the partition at which V1 trained and evaluated its selectors was the keyword partition, and that partition under-resolves the native task structure on every multi-family dataset we evaluate (Table 17).

Truncated selector training. V2 confirms that the operational conditioned-gain estimate is training-recipe sensitive, but disconfirms the simple expectation that longer or corrected training would uniformly strengthen the V1 DCASE result. The V1 DCASE cell was $+0.0993$ at $\varepsilon = 0.05$ with $p = 0.057$; V2 moves DCASE to -0.0300 on Qwen2-Audio and -0.0310 on Qwen2.5-Omni at the same tolerance. At the same time, V2 converts AudioMCQ-StrongAC from a V1 near-null into a positive cross-backbone result: $+0.0475$ on Qwen2-Audio and $+0.0215$ on Qwen2.5-Omni. The V2.1 scope-B sweep further confirms training-target dependence: Qwen2.5-Omni improves over the V2 baseline on all five datasets, while Qwen2-Audio improves on DCASE and MMAR in the handoff table. Thus the V1 truncated-training caveat was real, but its correction changes the locus of the effect rather than simply amplifying the V1 DCASE point estimate.

Parameter asymmetry dominated by the query embedding. V2 partially disconfirms the concern that any conditioned gain would be only a parameter-count artifact. The AudioMCQ result replicates across two backbones with different downstream architectures, and the Qwen2.5-Omni seed-123 AudioMCQ cell is significant at $p = 0.001$. However, V2 also shows that the query side remains a delicate part of the system: the V2.1 winning α values differ by backbone and dataset, with Qwen2.5-Omni favouring $\alpha = 1$ on AudioMCQ, MMSU, and BigBench Audio, $\alpha = 0.75$ on MMAR, and $\alpha = 0$ on DCASE, while Qwen2-Audio's reported wins use $\alpha = 0.5$. A more parameter-efficient or pretrained query compressor remains an important next step, but V2 makes the caveat empirical rather than speculative.

Table 40: Nominal-axis conditioned gain on the remaining four datasets. Transcribed from V1 PDF Table 8.

Dataset and ε	point estimate	95% CI	p	interpretation
AudioMCQ, 0.01	-0.0006	[-0.0014, +0.0003]	0.195	tight clean null
AudioMCQ, 0.02	-0.0011	[-0.0029, +0.0006]	0.195	tight clean null
AudioMCQ, 0.05	-0.0028	[-0.0071, +0.0014]	0.195	tight clean null
MMSU, 0.01	-0.0007	[-0.0061, +0.0043]	0.814	tight clean null
MMSU, 0.02	-0.0014	[-0.0122, +0.0087]	0.814	tight clean null
MMSU, 0.05	-0.0035	[-0.0306, +0.0217]	0.814	tight clean null
MMAR, 0.01	+0.0280	[-0.0758, +0.2680]	0.508	noisy, sign-unstable
MMAR, 0.02	+0.0559	[-0.1270, +0.2667]	0.535	noisy, sign-unstable
MMAR, 0.05	-0.0692	[-0.1561, +0.1387]	0.613	noisy, sign-unstable
BigBench Audio, 0.01	+0.0000	[+0.0000, +0.2297]	1.000	exact zero floor
BigBench Audio, 0.02	+0.0000	[+0.0000, +0.0714]	1.000	exact zero floor
BigBench Audio, 0.05	+0.0000	[+0.0000, +0.0000]	1.000	exact zero floor

Mel-spectrogram input features. V2 confirms that the selector feature representation is load-bearing, but not in the monotone sense anticipated by V1. Moving to the V2 pipeline does not uniformly lower every frontier or uniformly increase conditioned gain. Instead, it exposes representation-backbone interactions: AudioMCQ becomes positive on both backbones, DCASE reverses sign, and MMSU temporal is harmless on Qwen2-Audio but severely harmful on Qwen2.5-Omni. The mel-versus-audio-tower caveat should therefore be read as a warning that selector features determine which task families are served, not as a guarantee that native features monotonically improve every operational cell.

Partition under-resolution. The V1 selector training target is defined per sample and does not reference any query-family label, so in principle V1 training is partition-agnostic. The V1 family-level *evaluation*, however, is computed over the keyword partition $\mathcal{F}_{\mathrm{kw}}(\mathcal{Q})$, and the keyword classifier systematically under-resolves the native structure by factors of $1.2\times$ (AudioMCQ) to $8\times$ (MMSU), collapsing four distinct BigBench BBH task types into a single `general` bucket and six distinct DCASE content clusters into another `general` bucket. Under the 47-task MMSU native partition, the family-level gap is $+29.17$ pp at $b = 0.20$, versus the $+1.56$ pp reported under the keyword partition; the cross-backbone conditioned-gain bottleneck on Qwen2.5-Omni is `intonation_perception` (-0.7377), not the keyword-level `temporal` family; and the worst-k concentration on MMSU is 16% of excess on the worst 2 of 47 native tasks, versus 76% of excess on the worst 2 of 6 keyword families. These native-partition observations change the operational interpretation of the keyword results; they do not establish a general ordering of every reported quantity across nonnested partitions.

These observations leave the structural V1 findings intact. The family-level excess-risk gap, nested monotonicity, factor-overlap summaries, and synthetic theorem verifications do not depend on the V1 selector's exact training recipe. What V2 changes is the operational interpretation of $\widehat{G}^{\mathrm{op}}_{\mathrm{cond}}$: it is reproducibly positive in some regimes, reproducibly harmful in others, and cannot be summarized by the V1 DCASE cell alone.

P.4 What "uses query" means for the V2 conditioned selector

§O.9 establishes that the conditioned selector's ranking is query-dependent in the architectural sense - non-trivial f_{int}, Jaccard bounded away from 1 - but that the top-k overlap under query permutation remains above 0.5 on MMSU and AudioMCQ-StrongAC across both backbones. This nuances rather than contradicts the V2 operational picture. A reviewer asking "does the conditioned selector actually use the query" is answered *yes, at the* 13-25% *interaction-variance level and the* 0.5-0.7 *top-k Jaccard level, with* 3%*-consistent training-loss reduction*. A reviewer asking "is the conditioned selector query-dominant" is answered *no, the majority of top-k identity is preserved under query permutation*. Both answers are consistent with §O.7's small-positive AudioMCQ cross-backbone effect and with the backbone- and family-dependent regimes we observe: a moderately query-using selector will, in aggregate, deliver a small-positive-or-near-null operational gain rather than a uniformly large one.

Table 41: Descriptive cross-answerer/interface contrasts across 50 cells (5 datasets $\times$ 5 budgets $\times$ 2 selectors, seed 42). The historical $\widehat{\Gamma}^{\text{native}}$ and $\widehat{\Gamma}^{\text{heur}}$ labels denote these contrasts, not estimates of the theoretical model-class gap. Reported values are retained.

Quantity	n cells	mean	std	min	max
$\widehat{\Gamma}^{\text{native}}_{\mathcal{F}}$ (both selectors)	50	+0.1052	0.063	-0.009	+0.217
agnostic selector	25	+0.1047	–	–	–
conditioned selector	25	+0.1058	–	–	–
$\widehat{\Gamma}^{\text{heur}}_{\mathcal{F}}$ (three baselines)	50	+0.1260	–	–	–
selector-induced drift (learned − heur.)	50	-0.0025	0.010	–	–

Per-family heterogeneity (appendix §O.9) identifies MMSU/`temporal` as the most query-sensitive keyword family ($J_{\text{perm}} = 0.40$ on Qwen2-Audio and 0.27 on Qwen2.5-Omni), consistent with the temporal-family isolation observed in Table 32. AudioMCQ-StrongAC/`speech_content` is the most query-sensitive family on AudioMCQ ($J_{\text{perm}} = 0.46$ on both backbones), and BigBench Audio/`general` is the degenerate H2 control. The family-level ranking of query sensitivity is positively correlated across the two backbones over the 10 non-degenerate families (Spearman $\rho = 0.60$, Pearson $r = 0.66$, $p = 0.04$), indicating that the conditioning mechanism partially captures task-dependent structure rather than purely backbone-idiosyncratic artifacts. The largest cross-backbone disagreement in the per-family table is AudioMCQ-StrongAC/`music`, where $J_{\text{perm}} = 0.72$ on Qwen2-Audio versus 0.48 on Qwen2.5-Omni: Qwen2.5-Omni's smaller audio compressor produces chunk representations for music queries whose selector rankings are more strongly shaped by query content than Qwen2-Audio's.

The audit is strictly about the selector's *internal* behaviour - it observes the top-k chunk set and the pre-top-k scores, not downstream LALM predictions. Whether the 0.13-0.25 interaction share and the 0.5-0.7 Jaccard translate into a downstream accuracy effect that survives the LALM's own capacity to answer from context is a separate question, one that requires a decoupled-query experiment at the end-to-end level (we sketch its design in §P.9 under the heading of operational scope). The present audit therefore establishes the selector-level precondition for the V2 operational claims without conflating selector-side query use with downstream-side query use.

P.5 Cross-answerer and interface diagnostics

Proposition K.7 gives a theoretical bound involving Bayes information loss and model-class excess risk. The following learned-native and heuristic-baseline summaries compare deployed configurations; they do not identify either theoretical term.

The learned-native estimate averages $+0.1052$ over 50 cells, with standard deviation 0.063 and range $[-0.009, +0.217]$. The agnostic and conditioned selector means are nearly identical: $+0.1047$ and $+0.1058$. The heuristic-baseline estimate is $+0.1260$, and the learned-minus-heuristic selector-induced drift has mean -0.0025 with standard deviation 0.010. These descriptive contrasts vary more across configurations than between the two selector averages; the latter differ by 0.0011. This comparison does not identify the magnitude of the theoretical architecture gap.

The learned-native contrast changes both the answerer and its available interface. Shared heuristic interfaces remove that particular difference, but comparisons among a finite set of answerers still do not supply a Bayes-risk reference. Even evaluating stronger answerers on a common Z would remain a diagnostic unless the required optimal risks were identified.

P.6 Toward a semantic query-family partition

Three V1 findings-the family-level excess-risk gap, the cumulative-chain bottleneck analysis, and the factor-overlap diagnostic-depend on the query-family partition $\mathcal{F}(\mathcal{Q})$. V2 makes this dependence operational rather than merely methodological, and reports first results under two concrete non-keyword partitions: the dataset-native task partition of each benchmark (summarised in Table 17 and analysed in §O.11) and a semantic partition obtained from e5-large-v2 sentence embeddings

and cosine k-means clustering on query text (see Table 17, the semantic column). Both partitions produce substantially larger family-level gaps than the keyword partition on every multi-family dataset, and both preserve the bottleneck-family identity under 20% random label noise more robustly than the keyword partition does (§M.6, 84%-100% under semantic vs 39%-57% under keyword). The MMSU temporal isolation in Table 32 shows that a 99-sample family can dominate the sign of the conditioned-gain result on one backbone: Qwen2.5-Omni has non-temporal mean $+0.0247$ but cross-backbone keyword-temporal mean -0.5738, whereas Qwen2-Audio has keyword-temporal mean $+0.0000$ and non-temporal mean -0.0320. A dataset-level average alone would hide this divergence, and a finer native-task partition points to `intonation_perception` rather than temporal-reasoning as the precise failure mechanism (Figure 16).

DCASE gives the complementary lesson. The V1 taxonomy treated DCASE as single-family, but the V2 cluster analysis identifies `meaning/stress/intonation` as a Qwen2-Audio cluster with $n = 318$, $G_{\text{cond}} = -0.1227$, and 95% CI $[-0.310, -0.001]$. Thus even a dataset that is single-family under the V1 classifier contains subfamilies on which conditioning can behave differently. This directly motivates a semantic partition of query content rather than a partition inherited only from dataset metadata or keyword strings.

Three concrete next steps follow. First, the per-pair additivity diagnostic of §O.5 should be recomputed over native-task pairs on MMSU and BigBench Audio, not only over keyword pairs, to test whether finer granularity recovers any factor-disjoint regime that the keyword aggregation obscures. Second, the operational conditioned-gain replication of §O.7 should be reported by native task as well as by dataset aggregate: the aggregate `CONSISTENT−` label for Qwen2.5-Omni MMSU in Table 25 already hides the fact that the effect is concentrated on three fine-grained native tasks (`intonation_perception`, `pitch_comparison`, `pause_perception`). Third, a backbone-indexed, native-partition cumulative-chain analysis would allow nested-monotonicity bottleneck identification at the same granularity as Figure 16's per-task conditioned-gain picture. None of these steps requires new data collection; all three require rerunning existing estimators over the partitions already introduced in Table 17.

P.7 From operational frontiers to rate-theoretic frontiers

V1 reports operational frontiers on the nominal-budget and token axes, interpreting them as monotone proxies for the rate-theoretic frontiers of §2. That proxy is appropriate for the present paper because the compressed interface is a hard-selected token subset without an entropy-efficient coding layer. But it leaves open an important next step: moving from proxy frontiers to measured rate frontiers.

Two concrete directions are especially promising. First, an entropy-efficient coding layer on the selector output would bring operational token counts closer to the mutual-information quantities that define the theory. Second, stochastic interfaces would make the randomized-encoding overhead in Proposition K.8 measurable rather than purely conceptual. Neither extension is required for the core V1 claims, but both would sharpen the empirical interpretation of Theorem 3.4 on real audio by allowing a rate-axis analogue of $\widehat{G}^{\text{info}}_{\text{cond}}(\varepsilon, \mathcal{Q})$.

P.8 End-to-end selector training

V1 trains selectors against precomputed LOO-NLL oracle relevance targets rather than through the downstream answer likelihood itself. This decouples selector training from the backbone and makes V1 computationally manageable, but it also means the selector cannot adapt if the oracle target is systematically misspecified. A natural V3 direction is end-to-end selector training, in which the selector's output is connected differentiably to the backbone's final answer likelihood through a soft top-k or straight-through estimator. The main challenges are computational cost and gradient estimation through discrete selection, both of which are standard engineering rather than conceptual obstacles. Prompt-compression work in text provides a natural template for this transition (Nagle et al., 2024).

P.9 Scope, limitations, and when the theory applies

The paper makes three types of claim. First, it makes exact theorem claims in settings where the relevant rate quantities are analytically tractable; these are the synthetic validations of §O.1. Second,

it makes structural empirical claims about theorem-aligned quantities on real data: the family-level excess risk gap, nested monotonicity, and factor overlap. Third, it makes staged operational claims about conditioned compression on real audio. V1 provided a suggestive single-seed DCASE signal; V2 changes the operational picture to a backbone- and family-dependent map, with cross-backbone positive AudioMCQ, negative DCASE means, and a large Qwen2.5-Omni temporal failure on MMSU.

The paper does *not* claim that V1 or V2 directly measures the true rate-theoretic frontier on real audio, that the learned-native diagnostic table estimates the theoretical architecture gap, that V1's keyword partition is the final semantic story, or that query conditioning is uniformly beneficial across datasets and backbones. Those absences are deliberate and explicitly documented. They are not weaknesses hidden in the appendix; they define the honest scope of what the present evidence establishes.

The underlying theory is broader than the present instantiation. It applies to any compression interface that produces a random variable Z and any downstream answerer that incurs per-query Bayes risks: hard token selection, audio re-encoding, prompt compression, or hybrid interfaces. The family-level excess-risk refinement and the monotonicity structure are therefore general phenomena, not artifacts of the specific selector architecture used here. Theorem 3.4 is more regime-specific: it predicts strict separation when query families act on disjoint latent factor blocks, and it collapses smoothly toward zero as factor overlap increases. The present paper tests those statements conservatively, and V2 shows why the operational version must be reported by backbone and by semantic subfamily rather than as a single dataset-level scalar.