BEYOND LABELS: A COMPOSABLE ORDINAL INTERFACE FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

An ordinal label can be correct yet insufficient for subsequent reasoning. We study this gap for language-model predictions that admit an underlying scalar quantity, using nested proportional composition as a concrete setting. We introduce the Differentiable Fuzzy Inference Layer (DFIL), a dual-path head that retains a standard classifier while exposing a supervised scalar and Gaussian memberships with learnable ordered centers. Rather than composing intermediate labels, DFIL performs proportional operations in scalar space and discretizes only the final result. For nearest-center decoding, we characterize exactly when intermediate labels suffice for further composition, construct minimal exact refinements for fixed finite proportional multiplier sets, and derive optimal state budgets for margin-separated threshold queries. For a fixed composition structure and a supplied Cartesian uncertainty box that preserves center ordering, we also derive the exact set of attainable nearest-center labels while accounting for parameters shared across composition and decoding. For the analyzed ordered-center families, we further reduce Gaussian label reachability to a two-variable feasibility problem without requiring globally monotone decoding. Experiments on quantifier reasoning and related ordinal tasks, including human-judgment data and matched controls, separately assess predictive accuracy, structural consistency, and the contribution of the learned membership bank. This work frames ordinal prediction as the design of an inspectable, composable interface rather than a terminal category decision.

1 INTRODUCTION

A correct ordinal label can discard information needed by the next calculation (Figure 1). Two subgroup proportions may share a label, yet selecting the same fraction of each subgroup can place the results in different categories. A procedure receiving only the original label cannot recover both answers. The output has merged values that the next operation must distinguish, even when its original predictions were correct.

We study proportional composition: a base proportion is multiplied by specified positive factors no greater than one, then decoded into an ordinal category. Our approach merges scalars only when every allowed continuation gives the same label, following computation-preserving abstraction (Giacobazzi et al., 2000; Aydin Gol et al., 2012). Fix positive ordered centers below one; nearest-center decoding assigns each scalar to its closest center. The original labels suffice exactly when no allowed multiplication splits an existing category into different output categories. Otherwise, tracing decision thresholds backward through allowed products constructs the smallest exact state refinement. For every fixed finite multiplier set, this refinement is finite and has explicit transitions preserving every decision after every finite sequence.

Allowing every positive multiplier up to one changes the answer: no finite encoding supports all exact queries. Requiring each threshold comparison to be correct only when the product is at least a prescribed margin from that threshold restores a finite solution. We derive the exact minimum state count and an encoding that attains it, with the scalar stored once before the multiplier is supplied. At a fixed single threshold, this count grows inversely with the margin as the margin shrinks. The result determines the information required for the specified decisions, rather than the resolution needed to reconstruct the scalar or minimize average distortion.

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.

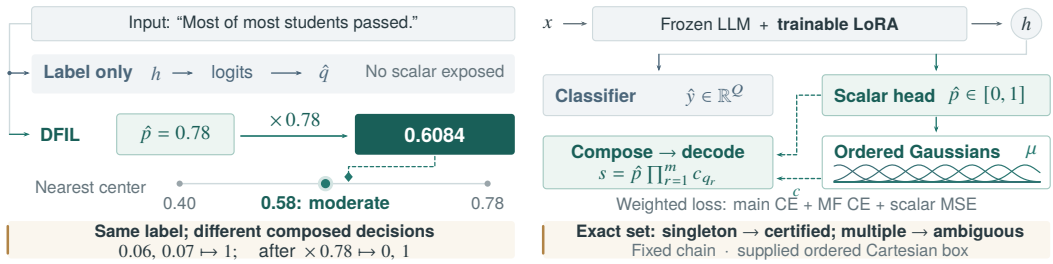


Figure 1: A correct category and a reusable operand preserve different information. Left: under the displayed numerical convention, proportions sharing a label require different labels after the same update. Right: DFIL retains its classifier and exposes a scalar; category centers enter both the product and its decoding reference. The illustrated exact set assumes nearest-center decoding, a specified chain and a supplied ordered Cartesian uncertainty box. The shared center must take the same value in both roles. Classifier and probability-mixture outputs are not covered by this certificate.

The Differentiable Fuzzy Inference Layer (DFIL) provides a learned realization. It retains a language model's full classifier and adds a supervised scalar with Gaussian membership functions centered on learned ordered category representatives. The scalar path composes before decoding; the same centers serve as proportional operands and decoding references. Numerical supervision supplies the scalar's meaning, while substituting a representative for a context-dependent fraction remains an approximation. The structural results also apply to compatible fixed banks.

This shared use of centers creates a second problem when the representatives are uncertain. A center's variation changes both the product and the rule that classifies it. Bounding those occurrences separately admits combinations that no single assignment can realize. For a specified chain, we obtain the exact nearest-center label set over independent base and center intervals, with positive center intervals strictly separated in order. Every returned label is attainable, distinguishing actual ambiguity from extra labels introduced by separate propagation.

Selecting the largest Gaussian membership needs a different analysis: unequal widths can reverse the order of predicted labels despite ordered centers. We characterize global order but do not require it for reachability. With the same center intervals and arbitrary independent positive width intervals, each candidate label's reachability reduces to finitely many feasibility problems in two real variables: the composed scalar and that label's center. Fixing these quantities restricts rival centers to intervals; the remaining centers must both satisfy their decoding restrictions and produce the retained scalar through the shared product. We eliminate the other variables without relaxing this equality. The extension handles DFIL's strict center-spacing and anchor constraints, preserving which endpoints are actually attainable. Across a feasible parameter family, the label set can have gaps or be a singleton even when some decoders are globally nonmonotone.

These results use exact arithmetic and specified numerical operations and uncertainty families. They do not certify classifier or mixture outputs, or supply semantic grounding and calibration. They distinguish two questions about reuse: which information future decisions require, and which decisions remain determined by shared uncertain inputs.

2 Related Work

Preserving future decisions. Complete abstract interpretations admit constructive refinements under suitable continuity conditions (Giacobazzi et al., 2000). Aydin Gol et al. (2012) construct finite bisimulations for switched linear systems on bounded working regions with an absorbing target set. Their predecessor refinement uses a common polyhedral Lyapunov function certifying contraction. We inherit the idea that states may be merged only when future observations agree. Our contribution is the explicit backward threshold set, minimum state count and transitions for proportional updates. Contraction explains finiteness; the threshold set identifies the coarsest exact partition, not just a sufficient discretization.

Quantization for computation. Functional quantization analyzes expected functional distortion at high resolution, with extensions to discontinuous functions (Misra et al., 2011). Hardware-limited task-based quantization also treats finite resolution, optimizing estimation distortion under an analog–digital architecture (Shlezinger et al., 2019). Our margin result instead requires worst-case correctness of every qualifying threshold query. Matching distinguishability and encoding constructions give an exact minimum without a source distribution, even for noninterval encoders. The budget concerns one-time scalar storage, excluding the shared codebook and query; repeated rounding and correctness only of whole labels outside all margin bands are different requirements.

Uncertain decision geometry. Gaussian membership maximization is a one-dimensional multiplicatively weighted nearest-site rule (Gudmundsson et al., 2024). UV-diagrams consider uncertain sites and externally supplied query points, with regions defined by positive probability of nearest-neighbor membership (Xie et al., 2013). Dependent geometric uncertainty has also been studied (Gitik & Joskowicz, 2021). Here the query is computed from the uncertain sites themselves. Our reduction retains their shared product equality while eliminating rival centers and widths. Reachability means existence of a feasible assignment, including isolated ties, with ties resolved toward the lower category. Strict-gap constraints require distinguishing attainable endpoints from limiting values. The result concerns this product and ordered geometry, not arbitrary parameter dependence or a universal solving speedup.

Ordinal outputs and numerical meaning. CORAL and CORN organize predictions through consistent rank probabilities (Cao et al., 2020; Shi et al., 2023); their full real-valued outputs are not subject to our label-only restriction. PRESQUE studies numerical quantifier meanings through pragmatic reasoning (Li et al., 2023). Our analysis takes supplied numerical meanings as inputs; category order alone does not specify them.

3 The Prediction Interface and Its Learned Realization

3.1 Learned Outputs and Training

For text x, let $y \in \{0, \dots, Q-1\}$ be an ordinal label and $p \in [0,1]$ its supervised scalar target. A LoRA-adapted frozen backbone (Hu et al., 2022) yields the last non-padding hidden state $h = f_\theta(x)$. DFIL retains the full classifier $\hat{y} = f_{\text{cls}}(h)$ and exposes a sigmoid scalar $\hat{p} = f_p(h)$ with memberships $\mu_q(z) = \exp\{-(z - c_q)^2/(2\sigma_q^2)\}$. Thus the interface is $h \mapsto (\hat{y}, \hat{p}, \mu(\hat{p}))$: the classifier keeps distinctions the scalar may discard. Supervised numerical targets supply the scalar's interpretation; ordered categories alone do not specify it. The centers are learned in order via a cumulative-softmax parameterization (Appendix A.1; exact feasible set in Section 5.2). Training uses

$$\mathcal{L} = \text{CE}(\hat{y}, y) + \lambda_{\text{DFIL}}\,\text{CE}(\mu(g(\hat{p})), y) + \lambda_p(\hat{p} - p)^2. \tag{1}$$

Here g is the forward-pass identity: the human-probability protocol keeps its input gradient, the few-shot and counting controls stop it, and scalar MSE remains active. Memberships enter the membership-function (MF) cross-entropy as logits. Appendices D and A.6.3 give loss weights, initialization and parameter counts.

3.2 Composition and Decoding

For a specified chain of outer categories $(q_1, \dots, q_m)$, compose before decoding:

$$s = \hat{p} \prod_{r=1}^{m} c_{q_r} = \hat{p} \prod_{j} c_j^{k_j}, \qquad k_j = \#\{r : q_r = j\}. \tag{2}$$

The operation multiplies scalar operands, not membership functions evaluated at a common input. The chain and decoder are specified numerical operations whose operands and category representatives are learned; substituting a center for a context-dependent fraction is an approximation, and .98 for "all" is not the identity 1. We distinguish nearest-center $D_c(s) = \arg\min_q |s - c_q|$, Gaussian $G_{c,\sigma}(s) = \arg\min_q |s - c_q|/\sigma_q$, the main classifier, and validation-selected probability mixtures. Both scalar decoders break ties toward the smaller index; equal widths give $G = D$. Subsequent guarantees concern these specified scalar decoders in exact arithmetic. Numerical certification requires exact or validated comparisons; it does not transfer to classifier or mixture outputs.

4 How Much Information Must an Ordinal Output Retain?

Fix $0 < c_0 < \cdots < c_{Q-1} < 1$ and boundaries $b_i = (c_i + c_{i+1})/2$. Then $D_c(p) = \sum_i 1\{p > b_i\}$ is monotone, and multipliers in $(0, 1]$ keep compositions in $[0, 1]$ (Appendix A.2). An *exact reusable state* consists of an encoding E, transitions F_a for allowed multipliers, and a readout d, satisfying

$$E(ap) = F_a(E(p)), \qquad d(E(p)) = D_c(p), \qquad p \in [0, 1]. \tag{3}$$

The state must support both the present label and all future updates. Its size concerns the exposed numerical interface, not compression of the backbone's hidden representation. Two inputs can share a state only if every allowed continuation has the same outcome. For an allowed set $\mathcal{A} \subset (0, 1]$, with a_w the product of a finite word of its multipliers, this is the future-equivalence relation

$$p \sim_{\mathcal{A}} p' \quad \iff \quad D_c(a_w p) = D_c(a_w p') \text{ for all } w \in \mathcal{A}^*. \tag{4}$$

The empty continuation checks the current label; nonempty ones check sufficiency after reuse.

When labels suffice (Theorem A.8). For a fixed $a \in (0, 1]$, an exact label update exists if and only if $\{b_i/a\}_i \cap (0, 1) \subseteq \{b_i\}_i$. A backward boundary inside an old cell separates inputs with the same present label but different next labels; if none exists, the next label is constant on each old cell. Checking each generator decides whether the original labels support every allowed finite chain.

The smallest exact refinement (Theorem A.10). For finite $\mathcal{A}$, with $a_\varnothing = 1$, define

$$\mathcal{T} = \{b_i/a_w : 0 \le i < Q - 1,\ w \in \mathcal{A}^*\} \cap (0, 1), \qquad E_*(p) = \sum_{t \in \mathcal{T}} 1\{p > t\}. \tag{5}$$

The set $\mathcal{T}$ is finite, and $1 + |\mathcal{T}|$ is the minimum number of exact states. Sorting its boundaries gives cells $[0, t_1], (t_1, t_2], \ldots, (t_N, 1]$. With upper-endpoint representative r_z for cell z, the optimal readout and transition are $d(z) = D_c(r_z)$ and $F_a(z) = E_*(ar_z)$.

Construction and minimality. Start with the original boundaries and repeatedly add $t/a < 1$. Two inputs in one resulting cell have identical comparisons to every backward boundary, so every allowed continuation gives the same label. Multiplying them preserves this equivalence, making the transition independent of the representative. Conversely, if $p \le t < p'$ for $t = b_i/a_w$, then $a_w p \le b_i < a_w p'$. That continuation distinguishes the inputs, forcing every exact encoding to separate them. All cells are nonempty, so the lower bound counts actual distinguishable states and is attained by E_*. This yields an exact update interface, not merely a finer classifier: repeated transitions stay in one system, and the readout agrees with composing the original scalar before decoding.

The fixed finite alphabet ensures termination: if it contains a nonidentity generator, let $\bar{a} < 1$ be the largest and choose H with $\bar{a}^H \le b_0$. Any word with at least H such factors places every backward boundary at or above 1; only finitely many shorter words contribute, and identity factors add nothing. A boundary at 1 never changes the label on the domain under the stated tie rule, hence its exclusion from $\mathcal{T}$. Finite exact reuse thus holds at every finite depth, without depth-dependent rounding.

For $c = (.1, .3, .9)$ and $\mathcal{A} = \{.5\}$, the cuts $.2, .4, .6, .8$ require five states for three labels, with transition $(0, 0, 1, 1, 2)$ and readout $(0, 1, 1, 2, 2)$; in general the table can be large (Appendix A.4).

Optimal storage for arbitrary queries (Theorem A.11). Allowing every multiplier in $(0, 1]$ changes the contract: for any boundary b, every pair $b < x < y \le 1$ is separated by some $a \in (b/y, b/x)$. Thus no finite encoding answers all exact queries (Proposition A.9). Instead, encode p *once* and require bit i to equal $1\{ap > b_i\}$ whenever $|ap - b_i| \ge \gamma$, where $0 < \gamma < \min(b_0, 1 - b_{Q-2})$. Set $L_i = b_i - \gamma$ and $U_i = b_i + \gamma$. Exactly M states are necessary and sufficient, where

$$t_0 = 0, \qquad t_{r+1} = \min_i \max\{U_i, (U_i/L_i)t_r\}, \qquad M = \#\{r \ge 0 : t_r \le 1\}. \tag{6}$$

Cells $[t_r, t_{r+1}) \cap [0, 1]$ with readout $H_i(r, a) = 1\{at_r > L_i\}$ attain this budget. For one boundary b and $\kappa = (b + \gamma)/(b - \gamma)$, $M = 2 + \lfloor \log(1/(b + \gamma))/\log \kappa \rfloor$, so $M = b \log(1/b)/(2\gamma) + O(1)$ as $\gamma \downarrow 0$ for fixed $b \in (0, 1)$: the exact leading constant of the $\Omega(1/\gamma)$ order in Proposition A.9.

Why the bounds match. Two values $x < y$ require different bits precisely when some i satisfies $y \ge \max\{U_i, (U_i/L_i)x\}$. The query $a = U_i/y$ witnesses this distinction by giving $ax \le L_i$ and

$ay = U_i$. The recurrence picks the first such distinguishable value. Its M left endpoints must therefore have distinct encodings, even for noninterval encoders. Conversely, if p lies in cell r, then $ap \le L_i$ forces the readout to zero. If $ap \ge U_i$ but the readout were zero, these inequalities would imply $p \ge t_{r+1}$, contradicting cell membership. The sequence grows by a factor greater than one after $t_1 > 0$, so M is finite. Equality already requires distinction, explaining the left-closed cells. For $b = .5$ and $\gamma = .05$, four states suffice. Summing the returned bits gives the exact ordinal label whenever the composed scalar lies outside every threshold margin. Inside a margin band only that threshold bit is unconstrained. The result is threshold-wise and read-once; whole-label contracts, shared codebook storage and repeated quantization are different questions (Appendix A.5).

5 Exact Decisions Under Shared Parameter Uncertainty

5.1 Exact Nearest-Center Decisions with Shared Operands

Exact label set (Theorem B.5). For a fixed chain, supply $p \in [p^-, p^+] \subseteq [0,1]$ and independent center intervals $[\ell_j, u_j]$, with $0 < \ell_j \le u_j \le 1$ and $u_j < \ell_{j+1}$. Each center appearing in the product and decoder is one shared variable. Put

$$S^- = p^- \prod_j \ell_j^{k_j}, \quad S^+ = p^+ \prod_j u_j^{k_j}, \qquad (d_j^-, d_j^+) = \begin{cases} (\ell_j, u_j), & k_j > 0, \\ (u_j, \ell_j), & k_j = 0. \end{cases} \tag{7}$$

The exact ratio bounds and complete attainable set are

$$L_i = \frac{2S^-}{d_i^- + d_{i+1}^-}, \quad U_i = \frac{2S^+}{d_i^+ + d_{i+1}^+}, \quad \mathcal{R}_D = \left\{ \sum_i 1\{L_i > 1\}, \dots, \sum_i 1\{U_i > 1\} \right\}. \tag{8}$$

Every returned label is attainable. A singleton is therefore necessary and sufficient for a common decision over this box. Evaluation takes $O(Q + m)$ arithmetic operations (Algorithm 1), with bit costs depending on precision and depth.

Why it is exact. For $R_i = 2p \prod_j c_j^{k_j}/(c_i + c_{i+1})$ and $p > 0$, a boundary center has logarithmic derivative $k_j/c_j - 1/(c_i + c_{i+1})$: positive when used in the product, negative when unused. Thus the displayed endpoints attain the ratio extrema without separating shared occurrences. With q_- and q_+ the two sums, if $q_- < Q - 1$, a witness for $L_{q_-} \le 1$ attains the lower label; if $q_+ > 0$, a witness for $U_{q_+ - 1} > 1$ attains the upper one; $p = 0$ and the remaining endpoint cases are immediate. Along a segment joining the witnesses inside the box, the product and boundaries vary continuously and only one boundary can meet the product at a time, so no intermediate label is skipped (Appendix B.2.1).

For $p \in [.84, .86]$, $c_0 = .1$, $c_1 \in [.7, .9]$ and $s = pc_1^2$, the minimum ratio is $1.029 > 1$, giving $\{1\}$. Separate product and boundary intervals give $\{0, 1\}$ by admitting incompatible parameter choices. In general they give only the ratio bounds $\widetilde{L}_i = 2S^-/(u_i + u_{i+1}) \le L_i$ and $\widetilde{U}_i = 2S^+/(\ell_i + \ell_{i+1}) \ge U_i$ (as $d_j^- \le u_j$, $d_j^+ \ge \ell_j$), so the exact set always lies within theirs and certifies every singleton they certify; the example shows the converse fails (Appendix B.2.1).

5.2 Nonmonotone Gaussian Banks: Eliminate Variables, Not Dependence

Ordered centers need not make G monotone. Define $\beta_i = (\sigma_{i+1} c_i + \sigma_i c_{i+1})/(\sigma_i + \sigma_{i+1})$. Theorem B.9 gives a complete global-order test: $G(0) = 0$, $G(1) = Q - 1$ and $G(\beta_i) = i$ for every i. The three learned 7B banks of Table 26 fail it. Whole-box tests and ordered shortcuts are in Appendices B.5–B.6; the following result requires no global order.

Fixed centers, arbitrary widths (Theorem B.12). For fixed centers $c_0 < \cdots < c_{Q-1}$, s in a closed interval I, and independent width intervals $[v_j, w_j]$, $v_j > 0$, with no ordering condition, $G_{c,\sigma}(s) = q$ is attainable if and only if some $s \in I$ has $|s - c_q|/w_q < |s - c_j|/v_j$ for $j < q$ and $\le$ for $j > q$. Splitting I at the at most $2(Q-1)$ equality roots per class and testing each root and piece decides every class exactly in $O(Q^3 + Q^2 \log Q)$ operations; the set may be disconnected. Uncertain shared centers need the reduction below.

What the extension decides. Let centers vary in $[.19, .21] \times [.49, .51] \times [.79, .81]$, widths in $[.39, .41] \times [.019, .021] \times [.049, .051]$, and $s = pc_2$. For $p \in [.85, 1]$, the exact set is $\{0, 2\}$; tightening only p to $[.98, 1]$ gives $\{2\}$ (Figure 2). Thus a globally nonmonotone family can admit a singleton for a specified query. These are algebraic witnesses, not model measurements.

Two-variable reachability (Theorem B.13). Keep the ordered center box of Section 5.1, an independent base interval, and arbitrary independent widths $\sigma_j \in [v_j, w_j]$ with $0 < v_j \le w_j$. For candidate q, retain $s \in [0, 1]$ and $t = c_q \in [\ell_q, u_q]$. Set $d = |s - t|$ and $\rho_j = v_j / w_q$. For a cut $h \in \{0, \dots, Q\}$, require $t \le s$ if $q < h$, and $t \ge s$ otherwise.

For this cut, the rival windows are

$$J_j = \begin{cases} [\ell_j, u_j] \cap (-\infty, s - \rho_j d\,], & j < h, \\ [\ell_j, u_j] \cap [\,s + \rho_j d, \infty), & j \ge h, \end{cases} \qquad J_q = \{t\}. \tag{9}$$

Make each new cutoff endpoint *open for* $j < q$, since ties favor the smaller index. Require nonempty J_j and set $A_j = \inf J_j$, $B_j = \sup J_j$. Then q is attainable *if and only if* some s, t, h satisfy

$$s \in \left[p^- \prod_j A_j^{k_j},\ p^+ \prod_j B_j^{k_j} \right], \tag{10}$$

whose endpoints are included per the flags below, with $a_j = 1\{A_j \in J_j\}$ and $b_j = 1\{B_j \in J_j\}$:

$$e_- = 1\{p^- = 0\} \vee \bigwedge_{j : k_j > 0} a_j, \qquad e_+ = 1\{p^+ = 0\} \vee \bigwedge_{j : k_j > 0} b_j. \tag{11}$$

These are two *real* variables with $Q + 1$ discrete cuts; the feasible label set may be disconnected.

Why the reduction preserves dependence. First, independent widths allow the candidate to take w_q and every rival v_j; widening the candidate improves its normalized distance, while narrowing a rival worsens that rival. Every feasible weak or strict win persists, and these endpoints are feasible. Second, center order places each rival left or right of s, turning its distance comparison into exactly J_j. Different candidate labels may use different favorable width vectors, because reachability asks separately whether each label has a witness. The candidate center remains fixed to t both in decoding and wherever it occurs in the product. Third, the remaining parameter set is connected. Its continuous product image is an interval whose extrema are approached by moving coordinates jointly toward their respective endpoints. With positive base, an excluded endpoint of any used center prevents attainment of the corresponding product extremum; unused centers do not affect attainment. With zero base, zero is attained at any feasible bank. Every interior product value is attained by connectedness. Requiring the retained s to belong to that image restores the *same product equality* using centers that also satisfy the decoder comparisons. A feasible reduced point reconstructs an original witness, and every original witness survives the eliminations. For rational inputs, resolving the finite branches leaves polynomial comparisons in s, t with degree at most $\max(1, m)$; branches, degree and precision affect solving cost, so the two-variable statement is not a constant-cost algorithm (Appendix B.8).

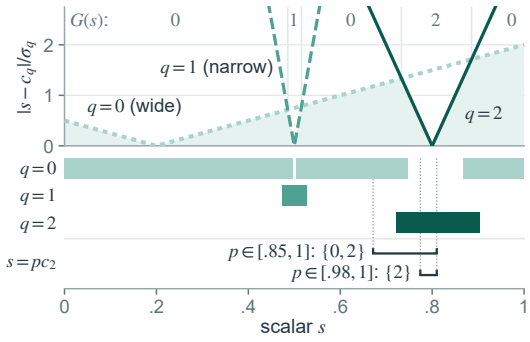


Figure 2: Exact reachability for a nonmonotone Gaussian decoder. *Top:* $|s - c_q|/\sigma_q$ for the midpoint bank of Section 5.2; the smallest wins, so $G(s)$ runs $0, 1, 0, 2, 0$ and fails the global-order test. *Bottom:* classes attainable at each fixed s by some bank in the box, and the query ranges $s = pc_2$. These rows ignore the coupling and only enclose the exact sets of Theorem B.13; here they coincide.

Exactness on the deployed family (Theorem B.15). The parameterization in Appendix A.1 has the exact finite-logit image $c_0 > .02 + g$, $c_j - c_{j-1} > g$, $c_{Q-1} = .98$, with $g = .96\epsilon/(1 + Q\epsilon)$ and

$\epsilon = 10^{-4}$. Intersect it with supplied, possibly overlapping, center intervals; base and width intervals remain independent, with widths $\geq 10^{-4}$. With the same J_j, strict anchor and fixed last center, propagate $L_0 = A_0$, $L_j = \max(A_j, L_{j-1}+g)$ and $U_{Q-1} = B_{Q-1}$, $U_j = \min(B_j, U_{j+1}-g)$. Every active propagated gap excludes its endpoint; a tie retains either exclusion. Feasibility requires $L_j < B_j$, or equality with both the propagated lower and original upper endpoints included. With L_j, U_j (and their endpoint flags) for A_j, B_j, (10) is again necessary and sufficient. The recurrences give simultaneous coordinate infima and suprema: feasible vectors approach all lower bounds together, and can attain all included endpoints together. The constrained set remains convex, so its product image is still an interval; endpoint attainment completes the exact test (Appendix B.8.5). Strictness changes decisions. With $c_0 \in [.4, .5]$, $c_1 = .6$, equal widths and $s = (11/12)c_1 = .55$, the closed box permits a class-zero tie at $c_0 = .5$. Requiring $c_1 - c_0 > .1$ excludes that point and leaves only class one. Exactness thus concerns the actual center family, not a larger enclosing box.

From attainable sets to coverage (Corollary B.6). If the true scalar and centers lie in the supplied ordered box and the true chain in a finite candidate set with joint probability at least $1 - \alpha$, the union of nearest-center sets covers the reference label with at least that probability. Issuing only singletons has marginal wrong-output probability at most α, but conditional error is bounded only by $\alpha / \Pr(\text{accept})$. On the joint containment event, the true chain and its numerical parameters furnish a feasible witness, so a wrong singleton can occur only outside that event. Membership widths do not supply this coverage, and semantic center and parse uncertainty need separate justification.

6 Verification and Empirical Evaluation

Human validation of complete nested quantifier expressions remains open (Appendix D.3), so the subsections test separate claims: exact procedures (direct checks), reuse of model predictions (calibration), the value of learning (matched controls) and the dual-path design (path comparisons).

6.1 Do the Exact Procedures Agree with the Original Problems?

Table 1 compares our constructions with direct decisions or nonlinear-real formulations. The information checks exercise state transitions and demanded threshold bits. For decision sets, the direct formulation retains the base scalar, centers, widths and shared product equality, asking feasibility per label. The 7,740 ordered-box queries (720 instances) include 1,740 shared-center ones. Separate strict-gap checks leave three reduced queries unknown, not unreachable. Both formulations use the same solver; finite checks complement the proofs (Appendices B.11, B.10, B.8.5).

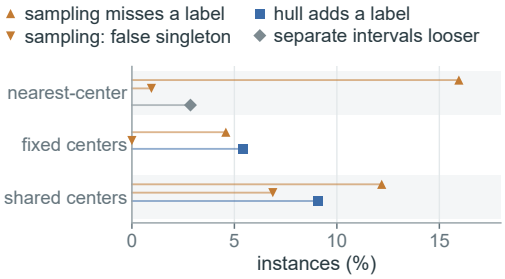


Figure 3: Share of random synthetic boxes (Appendix B.10; not model evaluations) where shortcuts fail. The hull spans extreme exact labels; nearest-center sets are intervals (Theorem B.5), and separate propagation applies only to them.

For $Q = 4$–32, nearest-center closed forms are 1,300–2,000$\times$ and fixed-center Gaussian root enumeration 60–190$\times$ faster than direct solving. The shared-center reduction is slower: direct solving is 1.3–2.7$\times$ faster. Its contribution is exact elimination. The shortcuts fail on ordinary instances (Figure 3): sampling returns singletons on 6.9% of shared-center instances despite additional attainable labels.

6.2 Can the Certificates Reuse Actual Model Predictions?

Two trained Qwen2.5-7B models (Yang et al., 2024) predict scalars on 2,044 FRoG-Hard rows. Split-conformal calibration of $|\hat{p} - p|$ uses 200 random calibration/evaluation half-splits without retraining. Chains are specified; reference labels use each deployed bank. Table 2 reports three nominal levels (center-tolerance analyses: Appendix B.3.1). Calibration covers only the base proportion. Hard shares source problems with Easy training, limiting generalization (Li et al., 2024).

Table 1: Direct verification of the information and decision results.

Construction / family	Checks	Outcome
Finite-state refinement	90 configurations; 18,490 comparisons	All comparisons agree
Margin-state encoding	110 configurations; 62,232 bit checks	All required bits agree
Ordered-box procedures	7,740 label queries (1,740 shared-center)	All complete and agree
Strict-gap reduction	61 direct/reduced pairs	58 agree; 3 reduced unknown

Table 2: Nearest-center outputs (%), deployed-bank reference; means over two models and 200 splits. Coverage is simultaneous over 7 one-step or 28 two-step chains; certified means a singleton. Corollary B.6 bounds only the marginal rate of wrong certified outputs, not these conditional errors.

	Coverage			Certified			Error if certified		Error otherwise	
$1-\alpha$	p	m=1	m=2	m=0	m=1	m=2	m=1	m=2	m=1	m=2
.95	95.0	95.7	95.5	0.0	18.7	51.9	2.0	0.5	43.4	35.6
.90	89.9	92.6	91.7	0.1	22.4	56.5	3.4	1.1	44.9	38.6
.80	80.0	85.0	83.1	1.4	28.0	62.0	6.5	2.2	47.0	42.2

At $1-\alpha = .90$, observed base coverage is 89.9%, but only 0.1% of base queries give a singleton: their intervals span multiple nearby centers. Proportional chains contract the scalar interval, enabling more decisions at the same input coverage.

6.3 What Does Learning the Interface Add?

Matched FT+Compose (Table 3) augments ordinary fine-tuning (FT) with the same scalar supervision and product rule, but freezes the bank and omits MF cross-entropy. The human comparison uses Qwen2.5-1.5B and 184 stimuli from 22 held-out statements (Belém et al., 2024); each arm selects a probability-mixture weight by validation MSE, then maps probabilities through the same fixed representatives $v_q = q/4$. DFIL's MSE is .007690 against .008154 for FT+Compose. The benefit concerns the complete learned treatment, not isolated center learning or a certified mixture. Task choice followed exploratory development (Appendix D.3).

Table 3: DFIL minus matched FT+Compose, with 20 paired training seeds evaluated on fixed data. Intervals describe training-seed variability: 95% Student-t intervals for human judgments and paired bootstrap intervals otherwise. Negative MSE/MAE and positive accuracy favor DFIL. The Exact-Count MAE contrast is descriptive because the marginal-invariance requirement fails.

Data / target source	Output / endpoint	Difference [95% CI]; evidence status
Human judgments	Selected mixture / MSE	$-.000464$ $[-.000756, -.000172]$
Natural-249 (App. E.2) / model-derived product labels	Nearest-center / accuracy	$-.98$ $[-3.73, 1.45]$pp; unresolved
ExactCount / enumeration	Scalar product / MAE	$+.0014$ $[-.0011, .0039]$; both arms fail marginal invariance

Natural-249 shows no primary nearest-center accuracy gain; its favorable secondary common-bin accuracy does not replace that result. ExactCount retains a descriptive MAE contrast, but both arms

miss the prespecified marginal-invariance requirement: their marginal null-control rates are 0.618 and 0.619, outside the accepted range of 0.40–0.60 (Appendix E.2). Thus the intended conditional-versus-marginal test does not validate a learned-bank-specific compositional benefit. Human-Grid likewise establishes bank-dependent operator fit, not a learned-bank advantage (Appendix E.3).

6.4 Why Retain Both Prediction Paths?

Table 4 retains the original three-seed component study, with $\lambda_{\mathrm{DFIL}} = 0.5$ and an 80/20 Easy row split followed by Hard evaluation. Appendix D.2 documents a percentage-parsing error in the initial diagnostics and a changed checkpoint criterion in corrected follow-ups; this table is therefore interpreted as an original-protocol diagnostic. Its dual-path means exceed the scalar-only means, supporting retention of a classifier option. Unordered-membership variants (Appendix E.4) also change the output path and do not isolate ordering.

Table 4: Core FRoG components, original protocol: accuracy (%), mean ± sample SD, three seeds.

Prediction configuration	Qwen2.5-1.5B	Qwen2.5-7B
FT, full classifier	$48.0 \pm .1$	51.3 ± 2.1
DFIL-only, scalar/Gaussian	41.6 ± 1.0	46.2 ± 2.6
DFIL, dual path	$48.2 \pm .4$	49.9 ± 2.6

DFIL does not dominate ordinary FT. Under the corrected protocol with a shared checkpoint criterion, similar means also fail to establish equivalence (Appendix E.5). On a 40-sentence implicit-proportion probe with hand-assigned targets, DFIL's main classifier and FT both reach 15.8% exact accuracy, whereas nearest-center decoding of the assigned scalars gives $73.3 \pm 1.4\%$ (100% off-by-one; three seeds; Table 31). Given the scalar, the interface recovers most stipulated labels; as input and output paths both change, the gap does not isolate extraction error. Full component, distillation and stronger teacher controls are in Appendices E.4–E.5; other diagnostics are in Appendix F.

Structural consistency. Nearest-center decoding is monotone (Theorem A.3); in an implementation check, 100 applicable checkpoints show zero D_c grid violations. The main classifier has no such guarantee: its text-level violation rates (descriptive; ties count as non-violations) are 9.71% for DFIL and 9.70% for FT (Appendix E.7). D_c on $\hat{p}$ guarantees order in the scalar; across texts, order also needs appropriate extraction. Cross-sentence monotonic consistency (CSMC) is 96.40% for DFIL's Gaussian argmax (94.45% before retraining six of 21 seeds) and 90.90% for FT's classifier; it has no nearest-center guarantee, and constant outputs are trivially consistent (Appendix F.5).

7 Discussion and Limitations

We study ordinal predictions as inputs to further computation rather than terminal category decisions. DFIL exposes numerical operands alongside a classifier. For the proportional setting considered here, our results characterize when labels suffice, how much information future decisions require, and which labels remain attainable under shared parameter uncertainty, including nonmonotone Gaussian decoding. This connects the design of reusable outputs with exact analysis of their subsequent use.

The guarantees apply to the specified scalar decoders and uncertainty families in exact arithmetic. This scope separates numerical composition from context-dependent interpretation: meaningful operands and composition structures must still be supplied, and extending coverage to unrestricted language requires justified joint uncertainty sets for them. Classifier and mixture outputs require separate guarantees. Compatible fixed and learned banks share the structural results; matched composition evaluations leave the added benefit of learning the bank unresolved. Context-sensitive grounding and joint calibration remain priorities for broader application. The present analysis establishes what an output must preserve and what its reuse can determine once the numerical setting is specified.

AI USE STATEMENT

Generative AI assisted writing and literature review; theoretical modeling and hypothesis development; formulation of mathematical claims, proof development and checking; experimental design, method implementation and result interpretation; generation of synthetic evaluation data; and data selection, cleaning, reformatting and annotation. Natural-249 item selection and base-proportion annotations were model-generated, and its sensitivity check used four language-model judges (Appendices D.2 and E.2). These annotations are stipulated targets rather than human judgments. The authors reviewed the AI-assisted text, references, proofs, implementations and results, including exact-arithmetic checks and comparisons of direct and reduced satisfiability formulations on finite decision instances. We take responsibility for the final manuscript and all AI-assisted claims and artifacts.

ETHICS STATEMENT

This work studies scalar interfaces for ordinal prediction. Experiments use synthetic tasks, existing public language datasets and previously collected human-judgment data. Model-generated annotations are distinguished from human judgments. The theoretical guarantees concern specified numerical settings and do not establish context-independent semantic validity for natural-language expressions.

REPRODUCIBILITY STATEMENT

Appendices A–C provide the assumptions, constructions and proofs supporting the theoretical results. Appendix B also describes exact-arithmetic verification and finite numerical checks, distinguishing their roles. Appendices D–F document data processing, experimental settings, evaluation protocols, uncertainty estimates and the scope of reported compute, with protocol-specific limitations stated alongside the results. The code will be made publicly available after publication.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Ebru Aydin Gol, Xuchu Ding, Mircea Lazar, and Calin Belta. Finite bisimulations for switched linear systems. *arXiv preprint arXiv:1208.5471*, 2012.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Catarina G. Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. Perceptions of linguistic uncertainty by language models and humans. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8467–8502, 2024.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. GenericsKB: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*, 2020.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. SemEval-2022 task 8: Multilingual news article similarity. In *International Workshop on Semantic Evaluation (SemEval)*, pp. 1094–1106, 2022.

Patrick Cousot and Radhia Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL)*, pp. 238–252, 1977.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16(72):2377–2404, 2015.

Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 337–340, 2008.

Roberto Giacobazzi, Francesco Ranzato, and Francesca Scozzari. Making abstract interpretations complete. *Journal of the ACM*, 47(2):361–416, 2000.

Rivka Gitik and Leo Joskowicz. Voronoi diagram and Delaunay triangulation with independent and dependent geometric uncertainties. *International Journal of Computational Geometry & Applications*, 31(2–3):75–121, 2021.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Joachim Gudmundsson, Martin P. Seybold, and Sampson Wong. Approximating multiplicatively weighted Voronoi diagrams: Efficient construction with linear size. In *International Symposium on Computational Geometry (SoCG)*, pp. 62:1–62:14, 2024.

Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. Not all LLM reasoners are created equal. *arXiv preprint arXiv:2410.01748*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Jyh-Shing Roger Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685, 1993.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Danial Kamali, Elham J. Barezi, and Parisa Kordjamshidi. NeSyCoCo: A neuro-symbolic concept composer for compositional generalization. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4184–4193, 2025.

Yiyuan Li, Rakesh R. Menon, Sayan Ghosh, and Shashank Srivastava. Pragmatic reasoning unlocks quantifier semantics for foundation models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 573–591, 2023.

Yiyuan Li, Shichao Sun, and Pengfei Liu. FRoG: Evaluating fuzzy reasoning of generalized quantifiers in LLMs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7239–7256, 2024.

Vinicius Macuch Silva, Alexandra Lorson, Michael Franke, Chris Cummins, and Bodo Winter. Strategic use of English quantifiers in the reporting of quantitative information. *Discourse Processes*, 61(10):498–523, 2024.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Vinith Misra, Vivek K. Goyal, and Lav R. Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083, 2017.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3):941–955, 2023.

Nir Shlezinger, Yonina C. Eldar, and Miguel R. D. Rodrigues. Hardware-limited task-based quantization. *IEEE Transactions on Signal Processing*, 67(20):5223–5238, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

Xike Xie, Reynold Cheng, Man Lung Yiu, Liwen Sun, and Jinchuan Chen. UV-diagram: A Voronoi diagram for uncertain spatial databases. *The VLDB Journal*, 22(3):319–344, 2013.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A Interface and Information Results

Status and notation. The results in Appendices A and B analyze the numerical interface in exact arithmetic; apart from the calibration experiment on FRoG predictions in Appendix B.3.1, they are not measurements of trained models. They do not alter the learned weights, training loss, or any reported prediction. Let $Q \geq 2$. Use $D_c(p) = \sum_{i=0}^{Q-2} \mathbf{1}\{p > b_i\}$, $b_i = (c_i + c_{i+1})/2$, with $0 < c_0 < \cdots < c_{Q-1} < 1$.

A.1 The Cumulative-Softmax Center Parameterization

For eight quantifiers, the deployed ordered parameterization is

$$\delta = \frac{\mathrm{softmax}(\delta_{\mathrm{raw}}) + \epsilon}{1 + Q\epsilon}, \qquad c_q = .02 + .96 \sum_{j=0}^{q} \delta_j, \qquad \epsilon = 10^{-4}, \tag{12}$$

with $\sigma_q = \max\{\exp(w_q), 10^{-4}\}$. The final center is fixed at .98. The nominal reference bank is author-specified, not measured from humans, and is initialized only approximately (see the initialization and inverse map below). Parameter counts appear in Appendix A.6.3.

The ordered center parameterization (12) uses cumulative softmax to guarantee strict ordering. We analyze its optimization-relevant properties.

Proposition A.1 (Strict ordering guarantee)**.** *For any $\boldsymbol{\delta}_{raw} \in \mathbb{R}^Q$, the centers* $\mathbf{c}$ *produced by* (12) *satisfy $c_0 < c_1 < \cdots < c_{Q-1}$.*

Proof. Write $s_q = \mathrm{softmax}(\boldsymbol{\delta}_{\mathrm{raw}})_q > 0$. After the shift and normalization, $\delta_q = (s_q + \epsilon)/(1 + Q\epsilon) > \epsilon/(1+Q\epsilon)$. Therefore $c_0 > 0.02 + 0.96\epsilon/(1+Q\epsilon)$, $c_{q+1} - c_q > 0.96\epsilon/(1+Q\epsilon)$, and $c_{Q-1} = 0.98$. These strict inequalities are statements in exact arithmetic; finite-precision implementations require numerical checks. □

Proposition A.2 (Jacobian rank and attainable center configurations)**.** *At finite raw logits, the Jacobian $\partial \mathbf{c}/\partial \boldsymbol{\delta}_{\mathrm{raw}}$ has rank $Q - 1$. The attainable normalized increments are exactly the vectors satisfying $\sum_q \delta_q = 1$ and $\delta_q > \epsilon/(1 + Q\epsilon)$. This characterizes representability; it does not guarantee that gradient descent converges to a chosen configuration.*

Proof. For $\mathbf{s} = \mathrm{softmax}(\boldsymbol{\delta}_{\mathrm{raw}})$, the softmax Jacobian is $\mathrm{diag}(\mathbf{s}) - \mathbf{s}\mathbf{s}^\top$. Its quadratic form is the weighted variance of the test vector under strictly positive weights, so its only null direction is the all-ones vector. Normalization after the ϵ shift multiplies this Jacobian by $1/(1+Q\epsilon)$. Cumulative summation is an invertible lower-triangular map, and scaling by 0.96 preserves rank. Conversely, for any stated feasible increment vector, $s_q = (1+Q\epsilon)\delta_q - \epsilon$ is positive and sums to one; raw logits $\log s_q$ produce it. This proves attainability and rank, not a lower bound on nonzero singular values or optimization convergence. □

Initialization and inverse map. For a feasible target center vector $\mathbf{c}^*$, define $\delta_0^* = (c_0^* - 0.02)/0.96$ and $\delta_q^* = (c_q^* - c_{q-1}^*)/0.96$ for $q \geq 1$. Exact inversion requires $c_{Q-1}^* = 0.98$ and all $\delta_q^* > \epsilon/(1+Q\epsilon)$; then set $\delta_{\mathrm{raw},q} = \log((1+Q\epsilon)\delta_q^* - \epsilon) + K$. The author-specified reference vector in Table 26 has endpoints 0.02 and 0.97 and is not exactly attainable. The implementation initializes from clipped log-increments followed by normalization, so reference initialization denotes an approximate initialization from those reference values.

A.2 Decoder Monotonicity and Compositional Closure

We give proofs with the inference rule and parameter domain stated explicitly. Nearest-center guarantees and free-center Gaussian-argmax representation results concern different hypothesis classes.

A.2.1 Proof of Theorem A.3 (Nearest-Center Rule)

Theorem A.3 (Monotonicity Guarantee, nearest-center inference rule)**.** *Let $Q \geq 2$ and $c_0 < c_1 < \cdots < c_{Q-1}$ be ordered centers in $[0,1]$. Then the nearest-center decision rule*

$$\hat{Q}_{\mathrm{MF}}(p) \;=\; \operatorname*{arg\,min}_{q\in\{0,\ldots,Q-1\}} |p - c_q|$$

with ties broken by the smallest index is monotonically non-decreasing in p. Thus exact nearest-center evaluation on any ordered scalar grid has zero pairwise violations.

Proof. The midpoint $b_q = (c_q + c_{q+1})/2$ for $q = 0, \ldots, Q-2$ is the unique value in $\mathbb{R}$ at which $|p - c_q| = |p - c_{q+1}|$. By the ordering of centers, $b_0 < b_1 < \cdots < b_{Q-2}$. For $p \leq b_0$, $\hat{Q}_{\mathrm{MF}}(p) = 0$; for $p \in (b_{q-1}, b_q]$, $\hat{Q}_{\mathrm{MF}}(p) = q$; for $p > b_{Q-2}$, $\hat{Q}_{\mathrm{MF}}(p) = Q-1$. Since these intervals are ordered by class index, for any $p_1 < p_2$ we have $\hat{Q}_{\mathrm{MF}}(p_1) \leq \hat{Q}_{\mathrm{MF}}(p_2)$. □

Remark A.4 (Why Gaussian $\arg\max$ alone does not guarantee monotonicity)**.** *A natural alternative inference rule is $\arg\max_q \mu_q(p) = \arg\max_q \exp(-(p-c_q)^2/(2\sigma_q^2))$. Setting $|(p-c_q)/\sigma_q| = |(p-c_{q+1})/\sigma_{q+1}|$ yields a quadratic in p with two solutions when $\sigma_q \neq \sigma_{q+1}$: an interior root*

$$b_q^{\mathrm{int}} \;=\; \frac{\sigma_{q+1}\, c_q + \sigma_q\, c_{q+1}}{\sigma_q + \sigma_{q+1}} \in (c_q, c_{q+1}), \tag{13}$$

and an exterior root $b_q^{\mathrm{ext}} = (\sigma_{q+1}c_q - \sigma_q c_{q+1})/(\sigma_{q+1} - \sigma_q)$. These are pairwise equality locations; another class may dominate there. For $Q = 2$, take $c = (0.25, 0.75)$ and $\sigma = (0.5, 0.1)$. Gaussian argmax predicts $0, 1, 0$ at $p = 0.5, 0.75, 1$, respectively, so ordered centers alone do not ensure its monotonicity. Nearest-center decoding uses the midpoint boundary and has no such reversal.

A.2.2 Proof of Theorem A.5 (Compositional Closure)

Theorem A.5 (Proportional Composition Closure)**.** *Let $c_0 < c_1 < \cdots < c_{Q-1} \in (0,1)$ be ordered MF centers, and define the n-step proportional-shift composition operator*

$$Q^{(n)}(q_1, \ldots, q_{n-1};\, p) \;:=\; \operatorname*{arg\,min}_{q\in\{0,\ldots,Q-1\}} \big|\, c_{q_1} c_{q_2} \cdots c_{q_{n-1}}\, p \;-\; c_q \big|,$$

with the empty product equal to one when $n = 1$. Multiplication is performed in the continuous proportion space and quantization is applied only to the final product. Then: (i) $Q^{(n)}$ is well-defined for every $n \geq 1$ and $p \in [0,1]$, with ties broken by smallest index; (ii) for fixed p and a fixed sequence extended by one factor at a time, the composed proportion lies in $[0, p]$ and is non-increasing with depth, as is its predicted class; (iii) the prediction depends only on the centers, not on the widths, and is monotone in p at each fixed depth and fixed outer-index sequence.

Proof of Theorem A.5. **(i) Well-definedness.** For any $p \in [0,1]$ and any choice of outer indices $q_1, \ldots, q_{n-1} \in \{0, \ldots, Q-1\}$, the composed proportion

$$p_{\text{comp}} \;=\; c_{q_1} c_{q_2} \cdots c_{q_{n-1}}\, p$$

is a real number in $[0, p] \subseteq [0,1]$ since each $c_{q_i} \in (0,1)$. The map $q \mapsto |p_{\text{comp}} - c_q|$ takes finitely many real values; its minimum is attained, and ties are broken by smallest index, so $Q^{(n)}$ is well-defined for every $n \geq 1$ and $p \in [0,1]$.

(ii) Non-increasing composition depth. For any $n \geq 2$ and $p \in [0,1]$, the chain of inequalities

$$c_{q_1} \cdots c_{q_{n-1}}\, p \;\leq\; c_{q_2} \cdots c_{q_{n-1}}\, p \;\leq\; \cdots \;\leq\; p$$

holds because each $c_{q_i} \in (0,1)$. Equivalently, extending a fixed sequence gives $p^{(n+1)}_{\text{comp}} = c_{q_n} p^{(n)}_{\text{comp}} \leq p^{(n)}_{\text{comp}}$. Theorem A.3 then gives $Q^{(n+1)} \leq Q^{(n)} \leq Q^{(1)}(p)$. If the base decoder labels p as "most", its extension cannot have a larger index; no particular lower label follows.

(iii) Width independence and monotonicity in p. $Q^{(n)}$ is defined entirely in terms of $\{c_q\}$ and the affine map $p \mapsto c_{q_1} \cdots c_{q_{n-1}}\, p$; widths $\{\sigma_q\}$ never appear. Fixing $q_1, \ldots, q_{n-1}$, the map $p \mapsto Q^{(n)}(q_1, \ldots, q_{n-1}; p)$ is the composition of a non-decreasing affine map and the nearest-center rule (non-decreasing by Theorem A.3), hence non-decreasing in p. □

Remark A.6 (Connection to membership-level t-norm composition)**.** *The product $\mu_{q_1}(p)\mu_q(p)$ is a membership-level conjunction at the same scalar input. Since a Gaussian membership is strictly positive in exact arithmetic,*

$$\arg\max_q [\mu_{q_1}(p)\mu_q(p)] = \arg\max_q \mu_q(p).$$

The first factor is common to all candidate classes. It therefore does not implement $p \mapsto c_{q_1} p$, and there is no general decision-boundary equivalence. Theorem A.5 analyzes the explicit proportion-level product followed by nearest-center decoding.

Remark A.7 (Explicit composition interface)**.** *A label-only classifier does not prescribe how to extract a calibrated proportion or choose class anchors. This is an interface distinction, not an information-theoretic impossibility: real-valued logits can encode a proportion, and a classifier can represent the same decision function. DFIL exposes $\hat{p}$ and $\{c_q\}$ directly, so the proportional-shift operator can be specified without an additional decoding convention for the logits.*

A.3 Label Sufficiency for Composition

Write $D_c(p) = \arg\min_q |p - c_q|$, with smallest-index ties and $0 < c_0 < \cdots < c_{Q-1} < 1$. Its boundaries are $B_c = \{(c_i + c_{i+1})/2 : 0 \leq i < Q-1\}$. This appendix gives the full statements, proofs, and verification for the label-sufficiency results of Section 4; Appendix B does so for the exact-decision results of Section 5. The basic monotonicity and closure proofs appear in Appendix A.2. Multiplication models a fixed nested subgroup selection with a supplied conditional fraction; the language-to-fraction mapping is a separate estimation problem.

Theorem A.8 (Exact criterion for composition from labels)**.** *For a fixed multiplier $a \in (0,1]$, a map T_a satisfying $T_a(D_c(p)) = D_c(ap)$ for every $p \in [0,1]$ exists if and only if*

$$\{b/a : b \in B_c\} \cap (0,1) \;\subseteq\; B_c.$$

For a specified set of multipliers, exact label updates at every depth exist if and only if this condition holds for every generator.

A scaled boundary inside an original decision cell separates two proportions with the same input label but different composed labels; no update of that label can recover the distinction. For example, with centers $(.02, .08, .18, .28, .40, .58, .78, .98)$, both .06 and .07 have label 1, but multiplication by .78 yields labels 0 and 1. Retaining the scalar and quantizing only after composition avoids this loss. This is an information constraint on the exposed interface, not an inability of a classifier retaining its hidden state to compute the product. Appendix A.3.1 gives the proofs, including Proposition A.9.

Proposition A.9 (Finite-state precision requirement)**.** *Let $b \in (0,1)$ be a decoder boundary. Suppose an encoding $E : [0,1] \to \mathcal{S}$ and output map H recover $D_c(ap)$ from $(E(p), a)$ for every $a \in (0,1]$. Then E is injective on $(b,1]$, and $\mathcal{S}$ cannot be finite. More quantitatively, fix $0 < \gamma < \min(b, 1-b)$. Even if only the binary decision $\mathbf{1}\{ap > b\}$ is required, and only when $|ap - b| \geq \gamma$, an encoding with M states must satisfy*

$$M \;\geq\; 1 + \left\lfloor \frac{\log(1/(b+\gamma))}{\log((b+\gamma)/(b-\gamma))} \right\rfloor .$$

For fixed b and $\gamma \downarrow 0$, this is $M = \Omega(1/\gamma)$, or $\log_2 M = \Omega(\log(1/\gamma))$ bits.

A.3.1 Information Retained by the Composition Interface

Proof of Theorem A.8. For necessity, suppose $t = b/a \in (0,1)$ is not an original boundary. It lies strictly inside one decision cell. Choose $p_- < t < p_+$ inside that cell. Their input labels coincide, while $ap_- < b < ap_+$ implies distinct output labels. A function of the input label and the fixed multiplier cannot produce both outputs.

For sufficiency, the decisions $D_c(ap)$ change only at the preimages b/a lying in $(0,1)$. By assumption these are original boundaries, so $D_c(ap)$ is constant on each original decision cell, including its right endpoint under smallest-index tie-breaking. A preimage at 1 causes no change within the domain and belongs to the lower output class. Define T_a as the constant output on each original cell. This gives the required equality at every input, including ties. If the criterion holds for each allowed generator, induction gives exact updates after any sequence of generators. Conversely, exact updates at every depth include the one-step case for each generator. □

The criterion admits nontrivial positive cases: centers $(.1, .4, .6)$ have boundaries $(.25, .5)$, for which $a = .5$ gives $T_a = (0, 0, 1)$. Thus the result is not a blanket impossibility claim about discrete composition. For a particular learned bank, the condition can be checked directly. It expresses whether the chosen partition retains enough information for the specified operations, in the sense of abstraction and precision studied by abstract interpretation (Cousot & Cousot, 1977).

Proof of Proposition A.9. Given $b < x < y \leq 1$, choose $a \in (b/y, b/x)$. This nonempty interval lies in $(0,1)$ and gives $ax < b < ay$. Equal encodings would force equal outputs under this same suffix, a contradiction.

For the margin version, set $\kappa = (b+\gamma)/(b-\gamma) > 1$ and take $x_j = (b+\gamma)\kappa^j \leq 1$. There are exactly the stated number of points. For any $x_j < x_k$, the multiplier $a = (b-\gamma)/x_j$ lies in $(0,1)$ and yields $ax_j = b - \gamma$ and $ax_k \geq b + \gamma$. Both are outside the excluded margin band and require different answers, so all these points need different states. The asymptotic bound follows from $\log \kappa \sim 2\gamma/b$. □

This lower bound allows every multiplier in $(0,1]$ and requires worst-case correctness. It does not establish a sample-complexity separation, does not rule out an exact finite abstraction for a particular finite multiplier family (Theorems A.8 and A.10 give the criterion and the minimal abstraction for such a family), and applies to finite-precision scalar storage as well as categorical storage. The practical advantage of the scalar path is postponing quantization, not unlimited precision or exclusive access to a compositional function. A general classifier retaining its input or hidden representation is outside the label-only restriction.

A.4 Minimal Exact Refinement for a Fixed Multiplier Alphabet

Definition. Let $\mathcal{A} \subset (0,1]$ be an allowed multiplier set, and let $\mathcal{A}^*$ denote its finite words, including the empty word. Put $a_w = \prod_{a \in w} a$ and $a_\varnothing = 1$. Two scalars are future-equivalent if

$$p \sim_{\mathcal{A}} p' \quad \Longleftrightarrow \quad D_c(a_w p) = D_c(a_w p') \text{ for every } w \in \mathcal{A}^*.$$

An exact finite-state interface is an encoding $E : [0,1] \to S$, transitions $F_a : S \to S$, and readout $d : S \to \{0, \ldots, Q-1\}$ satisfying $E(ap) = F_a(E(p))$ and $d(E(p)) = D_c(p)$ for all allowed a, p. Its state count concerns only this numerical post-processing, not the learned representation.

Theorem A.10 (minimal exact refinement)**.** *Define the boundary orbit*

$$\mathcal{T} = \{b_i/a_w : 0 \le i < Q-1,\ w \in \mathcal{A}^*\} \cap (0,1).$$

A finite exact interface exists if and only if $\mathcal{T}$ is finite. If $\mathcal{T} = \{t_1 < \cdots < t_N\}$ is finite, the unique coarsest future-equivalence partition is

$$[0,t_1],\ (t_1,t_2],\ldots,(t_N,1], \qquad E_*(p) = \sum_{t\in\mathcal{T}} \mathbf{1}\{p > t\}.$$

It has $N+1$ states, and every exact interface has at least $N+1$ states. The minimum is attained by selecting the upper endpoint r_s of each cell and setting

$$d(s) = D_c(r_s), \qquad F_a(s) = E_*(ar_s).$$

If $\mathcal{A}$ is finite, remove identity multipliers, let K be the number of remaining distinct values, and suppose $K > 0$. With $\bar{a} = \max(\mathcal{A}\setminus\{1\}) < 1$ and

$$H = \min\{h \ge 1 : \bar{a}^h \le b_0\} = \left\lceil \frac{\log b_0}{\log \bar{a}} \right\rceil,$$

only words with fewer than H nonidentity factors contribute thresholds, and

$$|S_*| = 1 + |\mathcal{T}| \ \le\ 1 + (Q-1)\binom{K+H-1}{K}.$$

If no nonidentity multiplier is allowed, $\mathcal{T} = \{b_0,\ldots,b_{Q-2}\}$ and $|S_*| = Q$.

Proof. For any w, the strict threshold comparisons defining $D_c(a_w p)$ are exactly $p > b_i/a_w$. Thresholds above or equal to 1 never fire on $[0,1]$; equality at 1 remains in the lower class under the specified tie rule. All other thresholds belong to $\mathcal{T}$. Hence two points with the same comparisons to every element of $\mathcal{T}$ give the same output under every suffix. Conversely, if $p < p'$ are separated by some $t = b_i/a_w$, meaning $p \le t < p'$, then $a_w p \le b_i < a_w p'$. The decoder is ordered, so its labels on these two inputs differ. This proves that the stated comparison pattern is precisely future equivalence, including its half-open endpoint convention.

Future equivalence is preserved by every allowed multiplier: if $p \sim p'$, then for each word w the word obtained by adding a gives $D_c(a_w a p) = D_c(a_w a p')$. Therefore all points of one cell map into one future-equivalence cell, and $F_a(s) = E_*(ar_s)$ is well-defined independently of the chosen representative. The empty suffix gives the readout identity. Thus this construction is an exact interface.

If another exact interface assigns the same state to two inputs, repeated transitions and the common readout force equality of their outputs under every word. Its cells therefore refine future equivalence. For finite $\mathcal{T}$ all $N+1$ displayed cells are nonempty and pairwise inequivalent, proving minimality. If $\mathcal{T}$ is infinite, any n distinct thresholds determine $n+1$ pairwise distinguishable nonempty regions, so no fixed finite state count suffices.

For a finite alphabet, a word with $h \ge H$ nonidentity factors has $a_w \le \bar{a}^h \le b_0 \le b_i$. Consequently $b_i/a_w \ge 1$ and it introduces no threshold. Products with fewer than H factors depend only on the multiplicity vector in $\mathbb{N}_0^K$, not on factor order. There are $\binom{K+H-1}{K}$ such vectors, including zero. Each gives at most $Q-1$ thresholds; duplicate thresholds can only reduce the count. This proves the bound. □

Construction and scope. Start with the original boundaries, repeatedly insert $t/a < 1$ for every stored threshold and allowed nonidentity a, and stop when nothing changes. The bound above guarantees termination for a finite alphabet. Equality testing and transition construction here are exact-arithmetic operations; a floating-point table needs a separate rounding analysis. The states preserve labels under the specified suffixes, not the scalar itself or arbitrary new queries. The table can be combinatorially large and changes when the fitted bank changes. No speedup or memory saving is inferred without measuring that table. For $c = (.1, .3, .9)$ and $\mathcal{A} = \{.5\}$, $\mathcal{T} = \{.2, .4, .6, .8\}$, so exactly five states suffice whereas the original decoder has three labels. With upper-endpoint representatives, the transition is $(0,0,1,1,2)$ and the readout is $(0,1,1,2,2)$. This is an analytic illustration, not a model experiment.

Relation to earlier theory. This is a scalar, threshold-explicit instance of future-equivalence and complete-abstraction construction; its relation to finite bisimulation and general minimization is discussed in Section 2. The contribution here is the exact boundary orbit, minimal state count, endpoint conventions, and multiplicity bound for the paper's proportional decoder. Theorem A.8 is recovered as the special case $\mathcal{T} = B_c$. Proposition A.9 is not contradicted: allowing every multiplier in $(0, 1]$ makes the orbit infinite above the lowest boundary.

A.5 Exact Precision Under a Decision Margin

Contract. Fix a common margin $0 < \gamma < \min(b_0, 1 - b_{Q-2})$. An encoder stores $E(p)$ once. A supplied query $a \in (0, 1]$ asks for each bit $\mathbf{1}\{ap > b_i\}$, and bit i must be correct whenever $|ap - b_i| \geq \gamma$. Bits inside their margin band may be arbitrary. This is a threshold-wise contract: summing the bits gives the exact ordinal label whenever the query is at least γ from every boundary. It is stronger than requiring only a label outside the union of margin bands. Storage of the shared codebook or query, and recursive re-quantization, are not covered by the state budget.

Theorem A.11 (optimal margin-state count)**.** *Put $L_i = b_i - \gamma$, $U_i = b_i + \gamma$, and $\kappa_i = U_i/L_i > 1$. Define*

$$t_0 = 0, \qquad t_{r+1} = \min_i \max\{U_i, \kappa_i t_r\}, \qquad M = \#\{r \geq 0 : t_r \leq 1\}.$$

The sequence strictly increases and M is finite. The exact minimum number of states for the stated contract is M, even among arbitrary, not necessarily interval-valued, encoders. An optimal encoding uses $[t_r, t_{r+1}) \cap [0, 1]$, and reads its state r as

$$H_i(r, a) = \mathbf{1}\{at_r > L_i\}.$$

For a single boundary b, writing $\kappa = (b + \gamma)/(b - \gamma)$ gives the closed form

$$M = 2 + \left\lfloor \frac{\log(1/(b + \gamma))}{\log \kappa} \right\rfloor.$$

For fixed $b \in (0, 1)$, as $\gamma \downarrow 0$,

$$M = \frac{b \log(1/b)}{2\gamma} + O(1).$$

Proof. For $0 \leq x < y \leq 1$, bit i can require different outputs under the same query if and only if there exists $a \in (0, 1]$ such that $ax \leq L_i$ and $ay \geq U_i$. Such an a exists exactly when

$$y \geq U_i \quad \text{and} \quad y \geq \kappa_i x.$$

Necessity follows from $a \leq 1$ and the two inequalities. For sufficiency take $a = U_i/y$: this lies in $(0, 1]$ and gives $ax \leq L_i$ (also when $x = 0$). Thus the first value that can be distinguished from x is $\min_i \max(U_i, \kappa_i x)$, with equality already distinguishable because the margin condition is inclusive.

Let $U_{\min} = \min_i U_i > 0$ and $\kappa_{\min} = \min_i \kappa_i > 1$. Then $t_1 = U_{\min}$ and $t_{r+1} \geq \kappa_{\min} t_r > t_r$ for $r \geq 1$. Consequently the sequence eventually exceeds one. Every pair $t_r < t_s \leq 1$ is distinguishable: choose an index attaining the minimum defining t_{r+1} and note that $t_s \geq t_{r+1} \geq \max(U_i, \kappa_i t_r)$. All M such points therefore need distinct states in any encoder.

For achievability, take p in the state with left endpoint t_r and fix a query a. If $ap \leq L_i$, then $at_r \leq L_i$, so $H_i(r, a) = 0$. If $ap \geq U_i$ but the proposed readout is zero, then $at_r \leq L_i$ and $p \geq U_i/a \geq U_i$. For $t_r > 0$ these also imply $p \geq \kappa_i t_r$; for $t_r = 0$ the latter inequality is automatic. Therefore $p \geq \max(U_i, \kappa_i t_r) \geq t_{r+1}$, contradicting state membership. The final state includes 1 because its next endpoint is strictly larger than 1; if some $t_r = 1$, it is a legitimate singleton state. Thus both demanded outputs are correct, including margin equalities, and M states suffice.

For a single boundary, $t_1 = b + \gamma$ and $t_r = (b + \gamma)\kappa^{r-1}$ for $r \geq 1$. Counting these terms at or below one, together with t_0, yields the displayed floor formula. Finally $\log \kappa = 2\gamma/b + O(\gamma^3)$ and $\log(1/(b + \gamma)) = \log(1/b) - \gamma/b + O(\gamma^2)$. Substitution, with the floor changing the count by less than one, gives the asymptotic expression. □

Meaning for DFIL. Theorem A.11 supplies an optimal constructive counterpart to Proposition A.9's lower precision bound. For $b = .5$ and $\gamma = .05$, it gives four states, including the zero/lower-range state. The result concerns read-once scalar storage with the total multiplier supplied as a query. It neither permits arbitrary repeated re-quantization at the same error budget nor proves that any archived model has the required decision margin. It leaves the actual scalar head and inference implementation unchanged.

Why the query contract matters. The threshold-wise requirement cannot silently be replaced by an optimality claim for whole-label queries outside every margin band. With boundaries $(.1, .2)$, margin $.04$, and inputs $x = .125$, $y = .25$, the upper threshold can distinguish the inputs under the stated bit-wise contract. But whenever both ax and ay avoid both margin bands, they have the same ordinal label: the valid ranges are $[0, .06]$, $[.14, .16]$, and $[.24, 1]$; $ax \leq .125$ forces a valid ax into the first range, and $ay = 2ax \leq .12$ then forces a valid ay into that same range. Thus the pairwise distinguishability relations differ. Likewise, repeated left-endpoint re-quantization is not the read-once construction: with $b = .25$, $\gamma = .01$, $p = .33$, and factors $.85, .95$, re-quantizing the intermediate reconstructed scalar can return the zero state although the true product $.266475$ exceeds $b + \gamma = .26$. Neither extension is claimed.

A.6 Capacity and Representation

Capacity of the ordered decoder (Proposition A.14). With the scalar map held fixed, let $\mathcal{H}_Q$ contain the nearest-center decoders whose centers arise from (12). Because every labeling of sorted scalars is non-decreasing, the growth function satisfies $\Pi_{\mathcal{H}_Q}(n) \leq \binom{n+Q-1}{Q-1}$, and the Natarajan dimension is exactly $Q - 1$; composing with a fixed scalar map preserves both upper bounds. This measures the decoder, not the jointly trained backbone, and implies no sample-complexity separation.

Representation versus guarantee (Propositions A.12 and A.13). An unconstrained linear head can realize nearest-center decoding exactly, yet the same class contains rules whose label falls as the proportion rises; the difference DFIL provides is an architectural guarantee, not expressive capacity. Conversely, Gaussian argmax with free ordered centers realizes any monotone partition away from its boundaries, whereas nearest-center decoding and the deployed center family cannot realize every partition: for three classes, boundaries $.1$ and $.9$ admit no nearest-center centers in $[0, 1]$.

A.6.1 Free-Center Gaussian Representation

Proposition A.12 (Realization by free-center Gaussian argmax)**.** *Let $Q \geq 2$ and $0 = b_0^* < b_1^* < \cdots < b_Q^* = 1$. A monotone target that labels the interior of $[b_q^*, b_{q+1}^*]$ by q is realized, away from internal boundaries, by Gaussian argmax with freely chosen ordered centers and positive widths. This statement concerns the free-center Gaussian family, not the constrained parameterization in Eq.* (12) *or the nearest-center decoder.*

Proof. Put $\ell_q = b_{q+1}^* - b_q^* > 0$, $c_q = (b_q^* + b_{q+1}^*)/2$, and $\sigma_q = \ell_q$. The centers are strictly ordered. At any point in the interior of interval q, $|p - c_q|/\sigma_q < 1/2$. For every $j \neq q$, the point lies outside interval j, so $|p - c_j|/\sigma_j > 1/2$. Since $\exp(-u^2/2)$ decreases strictly with $u \geq 0$, class q uniquely maximizes membership. At $p = 0$ and $p = 1$ the respective endpoint class has normalized distance $1/2$ and every other class has larger distance. Adjacent classes tie at internal boundaries, whose convention is excluded from the claim. □

Limits for nearest-center decoding and constrained centers. Nearest-center boundaries must satisfy $b_{q+1} = (c_q + c_{q+1})/2$. Not every target partition has such centers in $[0, 1]$: boundaries $0.1, 0.9$ for three classes would require $c_0 + c_1 = 0.2$ and $c_1 + c_2 = 1.8$, whereas $c_0 \geq 0$ gives $c_1 \leq 0.2$ and $c_2 \leq 1$ gives $c_1 + c_2 \leq 1.2$.

Equation (12) also fixes the last center at 0.98 and imposes a minimum center separation greater than $0.96\epsilon/(1 + Q\epsilon)$. For $Q = 8$ and $\epsilon = 10^{-4}$, the target boundaries $(0.1, 0.2, 0.5, 0.500001, 0.500002, 0.8, 0.9)$ satisfy the interior-envelope conditions but cannot be realized by the constrained Gaussian family. The centers of the two consecutive narrow classes

must lie strictly inside their respective intervals: each Gaussian uniquely attains membership one at its own center and remains the winner in a neighborhood. Their separation would be less than 2×10^{-6}, contradicting the minimum separation greater than 9.59×10^{-5}. Thus an interior-envelope condition alone does not give universality for the deployed family.

A.6.2 Representation Versus an Architectural Guarantee

Proposition A.13 (Unconstrained linear heads admit both monotone and non-monotone rules)**.** *For $Q \geq 2$ and $d \geq 2$, the class of linear heads on unconstrained hidden representations contains rules that violate monotonicity in a designated proportion coordinate. It also contains a head that exactly realizes nearest-center decoding for any strictly ordered centers. Therefore lack of an architectural guarantee does not imply lack of expressive capacity.*

Proof. For a violating example, take $\mathbf{h}_1 = (0.3, 1, 0, \ldots)$ and $\mathbf{h}_2 = (0.7, -1, 0, \ldots)$, with the first coordinate denoting the proportion. Set $f_0(\mathbf{h}) = -h_2$, $f_{Q-1}(\mathbf{h}) = h_2$, and, if present, all intermediate logits to -2. The predictions are $Q-1$ and 0, respectively, despite increasing proportion. This is an existence construction within the unconstrained class; it makes no assertion that a particular encoder produces these states.

For a monotone realization, let $\mathbf{h} = (p, 0, \ldots)$ and set $f_q(\mathbf{h}) = 2c_q p - c_q^2$. Then

$$\arg\max_q f_q(\mathbf{h}) = \arg\min_q (p - c_q)^2 = \arg\min_q |p - c_q|,$$

because p^2 is common to all classes. Use the same smallest-index tie convention. Theorem A.3 proves monotonicity of this linear-head realization. □

A.6.3 Parameter Accounting and Statistical Scope

The cumulative-softmax centers have $Q-1$ effective degrees of freedom, and the Q width parameters contribute at most Q more. The bank therefore has at most $2Q-1$ effective degrees of freedom, of which only the $Q-1$ center degrees affect nearest-center decisions. It stores $2Q$ raw scalars because a common shift of the center logits is redundant. The implemented width floor makes the local width derivative zero below its threshold.

The complete scalar branch also includes the trained numerical head. Like the main classifier, it applies layer normalization, a $d \to 256$ affine layer, GELU, dropout .1, and an output affine layer. With affine layer normalization and biases in both linear layers, its parameter count is $2d + (256d + 256) + (256 + 1) = 258d + 513$, before the MF bank. The deployed main classifier is also an MLP, with $2d + (256d + 256) + (256Q + Q) = 258d + 256 + 257Q$ parameters; the bare affine head in Proposition A.13 is an expressivity construction, not the implemented baseline. DFIL retains the MLP classifier and adds the scalar branch. At $d = 3584$, the numerical head has $925{,}185$ parameters and the raw MF bank adds 16, for $925{,}201$ additional stored parameters. Thus the small MF bank does not imply a smaller complete prediction head.

These counts do not establish a VC-dimension ratio or a sample-complexity separation. Multiclass capacity requires a specified complexity notion and a hypothesis class that includes every trained component, including the numerical head and LoRA. The few-shot results are empirical observations; no learning-rate or generalization guarantee is inferred from parameter counts.

An exact capacity result for the ordered decoder. To make the head-level capacity question precise, hold the scalar representation fixed and analyze only nearest-center decisions. We use Natarajan dimension, a multiclass notion of shattering (Daniely et al., 2015): each of n points has two distinct allowed labels, and all 2^n choices must be realizable. This definition concerns labelings, not the number of stored parameters.

Proposition A.14 (Capacity of the constrained nearest-center decoder)**.** *Fix $Q \geq 2$ and $\epsilon > 0$. Let $\mathcal{H}_Q$ consist of $p \mapsto \arg\min_q |p - c_q|$ on $[0, 1]$, with smallest-index ties and centers generated by Equation* (12) *as the raw logits range over $\mathbb{R}^Q$. Then*

$$\mathrm{Ndim}(\mathcal{H}_Q) = Q - 1, \qquad \Pi_{\mathcal{H}_Q}(n) \leq \binom{n + Q - 1}{Q - 1},$$

where Π *counts the maximum number of label vectors on* n *fixed points. For a fixed scalar map* $s : \mathcal{X} \to [0,1]$*, the composed class* $\{D_c \circ s\}$ *satisfies the same upper bounds.*

Proof. Every labeling on sorted scalar points is non-decreasing by Theorem A.3. Such a vector is determined by the Q nonnegative class counts summing to n, giving the binomial growth bound.

For the dimension upper bound, suppose $p_1 < \cdots < p_n$ are Natarajan-shattered with label pairs whose lower and upper members are $l_i < u_i$. Realizing all choices, in particular the upper choice at i and the lower choice at $i+1$, requires $u_i \le l_{i+1}$. Since labels are integers, $u_i \ge l_i + 1$, and hence $n \le Q - 1$.

For the lower bound, choose the strictly interior feasible bank

$$c_j^* = .02 + .96(j+1)/Q, \qquad j = 0, \ldots, Q-1,$$

obtained from equal raw logits, and set $x_i = (c_i^* + c_{i+1}^*)/2$ for $i = 0, \ldots, Q-2$. Prescribe arbitrary choices between labels i and $i+1$ at each x_i. Set boundary $b_i = x_i + t$ for the lower choice and $b_i = x_i - t$ for the upper choice, with $t > 0$ sufficiently small. Keep $c_{Q-1} = .98$ fixed and solve backwards by $c_i = 2b_i - c_{i+1}$. At $t = 0$ this gives c^*; by continuity, sufficiently small t preserves all strict feasibility inequalities and keeps neighboring boundaries on opposite sides of each x_i.

Writing $c_{-1} = .02$, the normalized increments $\delta_j = (c_j - c_{j-1})/.96$ sum to one because $c_{Q-1} = .98$, and the strict inequalities $\delta_j > \epsilon/(1 + Q\epsilon)$ hold at the equal-increment bank and hence at every sufficiently nearby bank. By Proposition A.2, finite raw logits realize each such bank exactly in Equation (12). Thus every prescribed binary choice is attained within the deployed center family, proving the lower bound. Restricting the input to the image of a fixed s cannot increase either complexity measure. □

This gives a formal capacity characterization with an explicit scope: at $Q = 8$ the fixed-scalar nearest-center family has Natarajan dimension 7, even though it cannot realize every ordered partition (Proposition A.12 and the counterexamples following it). Widths do not affect this decoder. A learned numerical head or LoRA representation changes the hypothesis class; the result supplies neither an end-to-end sample-complexity ratio against FT nor a proof of the observed few-shot advantage. Exact rational-arithmetic checks enumerate all binary boundary choices for $Q = 2, \ldots, 8$.

B Exact Decisions Under Parameter Uncertainty

B.1 Error Propagation and a Sufficient Decision Margin

A sufficient decision margin (Proposition B.3). Let $p, a_i \in [0,1]$ be reference operands of one parsed chain and $\hat{p}, \hat{a}_i \in [0,1]$ their estimates. Telescoping the product gives $|\hat{s} - s| \le B$ with

$$B = |\hat{p} - p| \prod_i \hat{a}_i + p \sum_i |\hat{a}_i - a_i| \prod_{j<i} \hat{a}_j \prod_{j>i} a_j.$$

If the estimated centers lie within η of reference centers c^* and $B + \eta$ is smaller than the distance from s to the nearest reference boundary $(c_q^* + c_{q+1}^*)/2$, nearest-center decoding of $\hat{s}$ returns the reference decision. With exact factors all at most $c_{\max} < 1$, it reduces to $|\hat{s} - s| \le |\hat{p} - p|\, c_{\max}^m$, so longer chains contract absolute base-proportion error (Proposition B.1 and its corollary); this does not imply smaller classification error. The margin is sufficient, not necessary, and supplies no statistical control of the estimation errors; the analysis below replaces it by an exact test for supplied boxes.

We distinguish perturbations of the base proportion from uncertainty in the factors and decoding centers.

Proposition B.1 (Base-proportion error with fixed factors)**.** *Let* $p, \hat{p} \in [0,1]$ *and* $\epsilon_p = |\hat{p} - p|$*. For a fixed outer-index sequence of length* $n-1$ *and fixed centers* $c_{q_i} \in (0,1)$*, define* $A = \prod_{i=1}^{n-1} c_{q_i}$*, with empty product one. The estimate* $\hat{s} = A\hat{p}$ *and reference product* $s = Ap$ *satisfy*

$$|\hat{s} - s| = A\epsilon_p.$$

The reference uses the same centers; it is not assumed to be the true semantic proportion.

Proof. Factor the common nonnegative scalar: $|A\hat{p} - Ap| = A|\hat{p} - p|$. □

Corollary B.2 (Conditional contraction)**.** *If all factors are bounded by $c_{\max} < 1$, then $|\hat{s} - s| \leq \epsilon_p c_{\max}^{n-1}$. At $n = 1$ the factor is one; at larger depths it is less than one.*

This contracts only absolute base-proportion error with the factors fixed. It does not imply decreasing classification error, accurate centers, or improved natural-language understanding.

Proposition B.3 (Factor perturbations and a sufficient decision margin)**.** *Let $m = n-1$, $p, \hat{p} \in [0, 1]$, and let $a_i, \hat{a}_i \in [0, 1]$ be reference and estimated factors for the same parsed chain. Put $s = p \prod_i a_i$, $\hat{s} = \hat{p} \prod_i \hat{a}_i$, and*

$$B = |\hat{p} - p| \prod_{i=1}^{m} \hat{a}_i + p \sum_{i=1}^{m} |\hat{a}_i - a_i| \prod_{j<i} \hat{a}_j \prod_{j>i} a_j.$$

Then $|\hat{s} - s| \leq B$. Let $c_0^* < \cdots < c_{Q-1}^*$ and $c_0 < \cdots < c_{Q-1}$ be reference and estimated decoding centers, set $\eta = \max_q |c_q - c_q^*|$, and let

$$\gamma(s) = \min_{0 \leq q < Q-1} \left| s - \frac{c_q^* + c_{q+1}^*}{2} \right|.$$

If $B + \eta < \gamma(s)$, nearest-center decoding of $\hat{s}$ using c equals nearest-center decoding of s using c^*.

Proof. Use the telescoping identity

$$\prod_i \hat{a}_i - \prod_i a_i = \sum_{i=1}^{m} (\hat{a}_i - a_i) \prod_{j<i} \hat{a}_j \prod_{j>i} a_j$$

and write $\hat{s} - s = (\hat{p} - p) \prod_i \hat{a}_i + p(\prod_i \hat{a}_i - \prod_i a_i)$. The triangle inequality yields B. Each estimated midpoint differs from the corresponding reference midpoint by at most η. Consequently each signed distance to a boundary changes by at most $B + \eta$. Under the strict margin condition all signs are preserved, so both decoders select the same ordered interval. □

The reference factors and centers must be externally meaningful for this certificate to imply semantic correctness. The result assumes the same parsed chain, supplies no statistical bound on the estimation errors, and is not an unconditional off-by-one guarantee.

Proposition B.4 (Dynamic range of a fixed chain)**.** *For fixed factors, the map $p \mapsto Ap$ sends $[0, 1]$ exactly onto $[0, A]$.*

Proof. $A > 0$, so the linear map sends the endpoints to $0, A$ and attains every intermediate value. □

Both the range and fixed-factor perturbations contract. Classification accuracy also depends on where the products fall relative to decision boundaries; range contraction alone does not explain a depth-dependent accuracy gain or loss. A smaller scalar error need not leave the class unchanged near a boundary, so an error such as .3 cannot be declared harmless without its location relative to the boundaries (Proposition B.3).

Endpoint semantics. The product t-norm has identity 1, but the deployed representative center for "all" is 0.98. Repeated "all" factors therefore shrink a proportion ($0.98^{100} \approx 0.1326$ at base one), rather than implementing an exact semantic identity. Exact "all" and "none" semantics would require factors 1 and 0, respectively. The present closure theorem does not assert those endpoint identities for the learned-center approximation.

Represented structure. Multiplying known factors is executable once the interface exposes them, whereas a label alone may lose needed precision (Theorem A.8). Correct parsing and meaningful factors are additional requirements. Neither contraction nor an oracle/text gap proves that linguistic extraction is the only remaining bottleneck. The exact-box extension in Appendix B strengthens the sufficient margin test when the relevant uncertainty set is available.

B.2 Exact Nearest-Center Certification

Point-estimate closure alone does not determine whether estimation errors change a label. Let a fixed parsed chain use center c_j exactly $k_j \in \mathbb{N}_0$ times, with $m = \sum_j k_j$ outer factors. Consider the full box

$$\mathcal{U} = [p_-, p_+] \times \prod_j [\ell_j, u_j], \qquad 0 \le p_- \le p_+ \le 1, \quad 0 < \ell_j \le u_j \le 1, \quad u_j < \ell_{j+1}.$$

The same uncertain center is reused at every occurrence and in the decoder. Define

$$S_- = p_- \prod_j \ell_j^{k_j}, \quad S_+ = p_+ \prod_j u_j^{k_j}, \qquad (d_j^-, d_j^+) = \begin{cases} (\ell_j, u_j), & k_j > 0, \\ (u_j, \ell_j), & k_j = 0. \end{cases}$$

Theorem B.5 (Exact set of attainable composed labels)**.** *For $R_i(p, c) = 2p \prod_j c_j^{k_j}/(c_i + c_{i+1})$, its exact minimum and maximum over $\mathcal{U}$ are*

$$L_i = \frac{2S_-}{d_i^- + d_{i+1}^-}, \qquad U_i = \frac{2S_+}{d_i^+ + d_{i+1}^+}.$$

Set $q_- = \sum_i \mathbf{1}\{L_i > 1\}$ and $q_+ = \sum_i \mathbf{1}\{U_i > 1\}$. Then

$$\left\{ D_c\left(p \prod_j c_j^{k_j} \right) : (p, c) \in \mathcal{U} \right\} = \{q_-, q_- + 1, \ldots, q_+\}.$$

Thus a common label is certified if and only if $q_- = q_+$. The bounds require $O(Q + m)$ arithmetic operations when the chain is given as a list.

The proof compares the product with each boundary through their ratio: boundary centers used in the product increase this ratio and unused ones decrease it, which resolves the shared-variable dependence exactly and can be strictly tighter than separate intervals for the product and the boundary. The certificate extends the sufficient perturbation bound in Proposition B.3 to an exact decision test for the stated box and fixed chain. Its intervals must be justified externally; Corollary B.6 transfers their coverage to the returned label set. The certificate applies equally to FT+Compose supplied with the same interface and intervals. Proofs, the comparison with separate intervals, rational-arithmetic verification results, and stress-test settings appear in Appendix B.2.1.

B.2.1 Proof and Use of the Exact Certificate

Proof of Theorem B.5. First fix a boundary i and $p > 0$. All centers are positive. For a center not in the boundary, $\partial_{c_j} \log R_i = k_j/c_j \ge 0$. For $j \in \{i, i+1\}$,

$$\partial_{c_j} \log R_i = \frac{k_j}{c_j} - \frac{1}{c_i + c_{i+1}}.$$

If $k_j = 0$, this derivative is negative. If $k_j \ge 1$, it is positive because $c_i + c_{i+1} > c_j$. The derivative with respect to p is positive. Thus the minimum uses the lower base endpoint, lower endpoints for every used center, and upper endpoints for unused centers in the denominator; the maximum uses the opposite choices. These choices are jointly feasible because the entire box consists of ordered banks. Centers with exponent zero do not change the numerator. This gives precisely L_i and U_i, with both extrema attained. The case $p = 0$ follows directly from $R_i = 0$, so boxes including zero and point intervals are also covered.

Let $s = p \prod_j c_j^{k_j}$. Smallest-index ties give the identity

$$D_c(s) = \sum_{i=0}^{Q-2} \mathbf{1}\{s > (c_i + c_{i+1})/2\} = \sum_{i=0}^{Q-2} \mathbf{1}\{R_i > 1\}.$$

Because the boundaries are ordered, R_i is non-increasing in i at every point. Its minimum sequence $(L_i)_i$ and maximum sequence $(U_i)_i$ are therefore non-increasing as well. Every feasible decision is at least q_- and at most q_+. If $q_- < Q - 1$, attain $L_{q_-} \le 1$: at that point the decision is at most q_-, so

Algorithm 1 Exact labels for a supplied ordered box and chain

Require: Base interval $[p_-, p_+]$, center intervals $[\ell_j, u_j]$, chain $(q_1, \dots, q_m)$
1: **if** the box fails the domain or ordering conditions of Theorem B.5 **then**
2: **return** $\{0, \dots, Q-1\}$
3: **end if**
4: Count $k_j = \#\{r : q_r = j\}$; compute $S_- = p_- \prod_j \ell_j^{k_j}$ and $S_+ = p_+ \prod_j u_j^{k_j}$
5: Set $(d_j^-, d_j^+) = (\ell_j, u_j)$ if $k_j > 0$, and (u_j, ℓ_j) otherwise
6: Compute $L_i = 2S_-/(d_i^- + d_{i+1}^-)$ and $U_i = 2S_+/(d_i^+ + d_{i+1}^+)$
7: **return** $\{\sum_i \mathbf{1}\{L_i > 1\}, \dots, \sum_i \mathbf{1}\{U_i > 1\}\}$

the lower endpoint is attainable. The case $q_- = Q-1$ is immediate. If $q_+ > 0$, attain $U_{q_+-1} > 1$: the decision is at least q_+, so the upper endpoint is attainable. The case $q_+ = 0$ is immediate.

It remains to show that no label between these endpoints is skipped. Join points attaining the two extreme labels by a straight segment in the box. The product and all boundaries vary continuously along it, and adjacent boundaries remain strictly separated. At any point at most one boundary can equal the product. A neighborhood of such a point therefore contains only two adjacent possible labels; away from boundaries the label is locally constant. A path from the minimum to the maximum cannot skip an intermediate class. This proves exactness of the whole returned set, including singleton necessity.

To implement the bounds, count the m chain entries, accumulate the two products, and form $Q-1$ denominators. Using a list of factors takes $O(Q+m)$ arithmetic operations. This is an arithmetic-operation count, not a constant-bit-cost claim about exact rationals at arbitrarily large depth. □

Comparison with separate interval propagation. If the product and decoding bank are bounded separately, their ratios satisfy only

$$\widetilde{L}_i = \frac{2S_-}{u_i + u_{i+1}}, \qquad \widetilde{U}_i = \frac{2S_+}{\ell_i + \ell_{i+1}}.$$

Our formulas give $\widetilde{L}_i \le L_i \le U_i \le \widetilde{U}_i$. Thus the exact output set is contained in the separate-interval set, and every singleton certified by the latter is certified by ours. In the witness in Section 5.1, separate intervals give $S \in [.4116, .6966]$ and boundary $[.4, .5]$, returning $\{0, 1\}$; the exact lower ratio is $2(.84)(.7)^2/(.1 + .7) = 1.029$, returning $\{1\}$. The improvement follows from retaining dependency between repeated parameters. General interval-bound methods need not be exact (Gowal et al., 2018); our exact result uses this particular product-to-boundary ratio and non-negative integer multiplicities, rather than claiming a general neural-network verifier.

Scope of the uncertainty set. The theorem is exact for the entire specified Cartesian box. Arbitrary correlated sets, intersecting center intervals, fractional exponents, or a different decoder require separate analysis. A box enclosing a smaller feasible set remains sound for that set but need not be tight. In particular, an uncertainty description for the deployed cumulative-softmax family can fix its last center at 0.98; any additional parameterization constraints must be respected when claiming exactness for that family. None of the reported scores is retroactively certified by choosing intervals after observing a test outcome.

If box coverage fails with probability at most α_U and chain coverage with probability at most α_K, the union bound gives $\alpha = \alpha_U + \alpha_K$ without independence. A singleton error rate conditional on acceptance is bounded only by $\alpha / \Pr(\text{accept})$ when acceptance probability is positive. As in standard conformal prediction (Angelopoulos & Bates, 2021), marginal coverage must not be confused with conditional validity after selection.

A calibration route, with observable targets. For a fixed fitted predictor and positive, pre-fixed scale functions $r_j(X)$, suppose calibration examples provide the full target vector $z = (p, c_0, \dots, c_{Q-1})$. Form simultaneous scores $V_t = \max_j |z_{tj} - \hat{z}_j(X_t)|/r_j(X_t)$. With N calibration cases exchangeable with the test case, take the $\lceil (N+1)(1-\alpha_U) \rceil$-th order statistic of the scores, using $+\infty$ if this index exceeds N. The resulting coordinate intervals have joint marginal

coverage at least $1-\alpha_U$ by the exchangeable-rank argument. Intersect with the known domain $[0,1]$; if an ordered positive box cannot be supplied, return all labels instead of applying the exact formula. This preserves coverage. Calibration data must be separate from fitting and selecting the predictor or the scales. If only base proportions are observed, this argument calibrates only p and requires separately valid center bounds. FRoG observes p but neither centers nor parses, so no dataset used here provides a calibrated joint box and candidate-chain set; Appendix B.3.1 carries out the p-only route with declared centers and specified chains. The corollary is a route to a validated extension, not a new human-semantic accuracy claim.

Computational verification and numerical provenance. An initial implementation checked 600 ordered boxes, 18,000 corners, and 7,200 interior points. A separate reference implementation used for verification checks 256 rational boxes with $Q = 2, \ldots, 5$ and multiplicities $0, \ldots, 4$ against 7,680 complete corner evaluations and 2,048 sampled interior points. It also checks 512 random boundary/multiplier configurations and the explicit positive and negative label-update examples. No mismatch was found in these checks. They complement the mathematical proofs; they do not independently reproduce the initial implementation or any model training result.

Reanalysis using the printed center banks. To make a stress sweep reproducible without checkpoint files, we treat every three-decimal entry of the six banks in Table 26 as an exact decimal input. For $\rho \in \{0, .1, .25\}$, each non-final center receives radius $\rho \min\{c_j, 1-c_j, c_j - c_{j-1}, c_{j+1} - c_j\}$, omitting absent-neighbor terms; the final center stays fixed at .98. Base values are $r/20$, $r = 0, \ldots, 20$, with radius .02 clipped to $[0,1]$. Each of the eight centers is repeated $m \in \{1, 2, 4, 8, 16\}$ times. This gives 3,024 declared boxes per depth. All arithmetic in the primary implementation is rational. A separate implementation also checked 100 selected boxes by enumerating all 512 corners per box and reproduced the sweep counts.

Table 17 reports both the initial-implementation counts and this rounded-input reanalysis. The initial count sequence is not recovered from these explicit inputs. The higher-precision centers used by the initial sweep were not retained, so the source of the discrepancy is unresolved; it is not established that rounding explains it. The printed-input sweep is therefore not a replication of the initial sweep. It is a deterministic illustration of the exact certificate for a fully specified input collection, with no calibration or semantic-accuracy interpretation.

B.3 Coverage Transfer

Corollary B.6 (Coverage transfer and selective output)**.** *Let the true scalar and shared centers lie in a proposed ordered box, and the true chain in a finite proposed set $\mathcal{K}(X)$, with joint probability at least $1-\alpha$. Return the union of the exact label sets in Theorem B.5 over those chains. Then the reference composed label belongs to the returned set with probability at least $1-\alpha$. If a prediction is issued only for a singleton, the probability of issuing a wrong prediction is at most α.*

Proof of Corollary B.6. On the joint containment event, the true chain is one of the candidates and its true parameters are feasible in the box. Theorem B.5 includes its label. Therefore failure of coverage, and hence issuing a wrong singleton, can occur only outside that event. □

The corollary requires observable targets or external bounds that justify the input set. Appendix B.2.1 gives the distinction between marginal and conditional error and a simultaneous split-conformal construction with its data requirements.

Certified set-valued extension. Algorithm 1 implements Theorem B.5 as a post-processing operation. It accepts a supplied uncertainty box and a fixed chain; it does not infer their validity from the MF widths. For several candidate parses, run it on each and take the union. A singleton permits a point decision; otherwise the set exposes the unresolved ambiguity. Invalid or overlapping boxes return all labels. This extension leaves the training objective, the trained weights, and the evaluation protocols unchanged. Appendix B.3.1 calibrates the base interval where the base proportion is observed; learning valid center intervals and a candidate parse set remains an additional calibration task, and the semantic coverage of composed expressions is not established. Main-classifier and mixture outputs remain separate, uncertified choices.

Table 5: Calibrated certification on FRoG-Hard: means over two 7B models and 200 random calibration/evaluation halves, with the deployed bank as reference. Simultaneous coverage requires every chain of an item to be covered. Certified means a singleton output. Error rates are conditional on certification or its absence; the corollary bounds only the marginal rate of wrong certified outputs.

		Simultaneous coverage		Certified (%)		
$1-\alpha$	p coverage	$m=1$	$m=2$	$m=0$	$m=1$	$m=2$
.95	.950	.957	.955	0.0	18.7	51.9
.90	.899	.926	.917	0.1	22.4	56.5
.80	.800	.850	.831	1.4	28.0	62.0

	Error when certified (%)		Error otherwise (%)	
$1-\alpha$	$m=1$	$m=2$	$m=1$	$m=2$
.95	2.0	0.5	43.4	35.6
.90	3.4	1.1	44.9	38.6
.80	6.5	2.2	47.0	42.2

B.3.1 Calibrated Certification on FRoG Predictions

FRoG records the target percentage of every item, so the base-proportion route of Appendix B.2.1 can be carried out on real predictions; centers are not observed and are handled by declaration. We use the numerical heads and learned banks of two independently trained Qwen2.5-7B dual-path models with detached MF input. The 2,044 Hard rows are split uniformly at random into calibration and evaluation halves, 200 times. The calibration score is $|\hat{p}-p|$, that is, $r \equiv 1$ in the route above, and the resulting interval, intersected with $[0,1]$, is applied to the evaluation half. Queries are the base scalar and every chain of one or two of the seven non-null quantifiers (7 and 28 chains); the null center is excluded because it sends every product to the lowest class. The chain is specified, not parsed. The reference label is the nearest-center label of $p\prod_j c_j^{k_j}$ under the deployed bank, and the box is that bank itself. The reference label is then a function of p alone, and Corollary B.6 guarantees coverage of at least $1-\alpha$ for all 35 composed queries of an item simultaneously, over random partitions of this fixed collection. Hard shares source questions with the Easy training rows (Appendix D.4.1); this can make residuals optimistic but does not affect the validity of the random partition.

Table 5 shows base-proportion coverage at the nominal level and simultaneous label coverage above it. Base queries are almost never certified: the calibrated radius, about .18 at $\alpha = .1$, spans several of the closely spaced low centers. A chain multiplies this radius by its factors (Proposition B.1), and 19–28% of one-step and 52–62% of two-step queries receive singletons. Plug-in decisions from $\hat{p}$ err on 35.6% of one-step and 17.4% of two-step queries overall; the certificate separates the certified ones, which err on 0.5–6.5%, from the rest, which err on 36–47%. The observed marginal rate of wrong certified outputs is at most 1.8% for every α. Calibrating separately within quartiles of $\hat{p}$ changes these certified fractions by at most 2.3 points and lowers the certified error, for example from 2.0% to 1.2% at $\alpha = .05$, $m = 1$.

Center uncertainty separates the exact sets from separate interval propagation. With a declared tolerance of $\pm.01$ around every deployed center, capped at 45% of each adjacent gap and with the two extreme centers kept as points, the exact sets certify 14.0, 17.9 and 21.2% of one-step queries at $\alpha = .05, .10, .20$, against 10.7, 15.4 and 18.4% when product and boundary intervals are bounded separately. Composed-query label coverage under the deployed bank and both box corners is at least .95. For the coordinatewise hull of the two trained banks, with each bank as reference, the gain is 0.8 points; for two-step queries it is at most one point in every setting. The experiment thus tests the base-proportion calibration route and the set-valued decision rule on real predictions. It does not calibrate centers or parses, and its reference label follows the deployed bank rather than human judgments of composed expressions.

B.4 Gaussian Decoding and the Scalar Guarantees

If every adjacent width ratio satisfies $(1-c_{i+1})/(1-c_i) < \sigma_{i+1}/\sigma_i < c_{i+1}/c_i$, then G is monotone with boundaries $\beta_i = (\sigma_{i+1}c_i + \sigma_i c_{i+1})/(\sigma_i + \sigma_{i+1})$, and G and D disagree exactly on the disjoint intervals between β_i and the midpoints $(c_i + c_{i+1})/2$, by at most one class; they agree everywhere if and only if all widths are equal (Propositions B.7 and B.8). The condition is sufficient, not necessary, and unequal widths otherwise need separate analysis.

The counterexample in Remark A.4 motivates a precise connection between the Gaussian decoder used in the original experiments and the nearest-center rule. The following results analyze the existing membership family; they introduce no new training procedure or width parameterization.

Proposition B.7 (A sufficient width condition for monotone Gaussian decoding)**.** *Let $0 < c_0 < \cdots < c_{Q-1} < 1$, $\sigma_i > 0$, and let $G_{c,\sigma}(p) = \arg\max_i \exp(-(p-c_i)^2/(2\sigma_i^2))$ use smallest-index ties. If every adjacent pair satisfies*

$$\frac{1-c_{i+1}}{1-c_i} < \frac{\sigma_{i+1}}{\sigma_i} < \frac{c_{i+1}}{c_i}, \qquad 0 \le i < Q-1, \tag{14}$$

then $G_{c,\sigma}$ is non-decreasing on $[0,1]$ and has ordered boundaries

$$\beta_i = \frac{\sigma_{i+1}c_i + \sigma_i c_{i+1}}{\sigma_i + \sigma_{i+1}}, \qquad G_{c,\sigma}(p) = \sum_{i=0}^{Q-2} \mathbf{1}\{p > \beta_i\}.$$

Consequently, Gaussian decoding of a product of fixed positive factors is monotone in the base proportion and non-increasing when the chain is extended by factors in $(0,1)$. Its decisions can still depend on widths.

Proof. For adjacent centers write $d_i(p) = |p - c_i|/\sigma_i$. On $[0, c_i]$, the difference $d_{i+1} - d_i$ is affine and strictly positive at both endpoints, using the upper width-ratio bound at zero. On $[c_{i+1}, 1]$ it is affine and strictly negative at both endpoints, using the lower bound at one. Between the centers, d_i increases and d_{i+1} decreases; equality occurs only at $\beta_i \in (c_i, c_{i+1})$. Thus the lower index wins their comparison for $p \le \beta_i$, with equality only at the boundary, and the higher index wins for $p > \beta_i$. The boundaries are ordered because each lies between successive centers. If $p \in (\beta_{k-1}, \beta_k]$, successive adjacent comparisons show that class k beats every smaller class strictly and every larger class weakly; the tie rule selects k. The first and last intervals follow identically. Composing this monotone decoder with the product yields the final assertions. □

Condition (14) is sufficient, not necessary: another class can mask a pairwise tail crossing. Equal widths satisfy it, but unequal widths may also do so. The strict lower inequality handles endpoint ties. For $c = (1/4, 3/4)$ and $\sigma = (3/10, 1/10)$, equality holds at the lower ratio bound; class 1 wins near $p = 1$, but the tie at $p = 1$ returns class 0. Center ordering alone, or a finite grid with no observed reversal, does not establish this condition.

Proposition B.8 (Decoder agreement and a Gaussian composition margin)**.** *Under the assumptions of Proposition B.7, put $D_c = \hat{Q}_{\mathrm{MF}}$, $b_i = (c_i + c_{i+1})/2$, and*

$$\Delta_\sigma = \max_i |\beta_i - b_i|, \qquad \beta_i - b_i = \frac{(c_{i+1} - c_i)(\sigma_i - \sigma_{i+1})}{2(\sigma_i + \sigma_{i+1})}.$$

The exact disagreement set is

$$\{p : G_{c,\sigma}(p) \ne D_c(p)\} = \bigcup_{i=0}^{Q-2} \big(\min\{b_i, \beta_i\}, \max\{b_i, \beta_i\}\big].$$

These intervals are disjoint, and the two predictions differ by at most one class. They agree everywhere if and only if all widths are equal. In particular, $\min_i |p - b_i| > \Delta_\sigma$ guarantees agreement at p.

For the reference and estimated chains in Proposition B.3, suppose the estimated bank (c, σ) satisfies (14)*. Its Gaussian prediction agrees with the reference nearest-center prediction whenever*

$$B + \eta + \Delta_\sigma < \gamma(s). \tag{15}$$

Proof. Both decoders count strict crossings of their respective boundaries. For boundary i, the counts differ exactly on the stated half-open interval. Each such interval lies in (c_i, c_{i+1}), so different intervals cannot overlap or cancel. The boundary-shift identity follows by subtraction; because each center gap is positive, every shift vanishes exactly when every adjacent width pair is equal. The margin condition for a common scalar follows immediately. For the perturbed chain, writing $b_i^* = (c_i^* + c_{i+1}^*)/2$ gives

$$|(\hat{s} - \beta_i) - (s - b_i^*)| \le |\hat{s} - s| + |\beta_i - b_i| + |b_i - b_i^*| \le B + \Delta_\sigma + \eta.$$

Under (15), every threshold comparison has the same strict sign as its reference comparison, proving equality of the labels. □

This extends the original error analysis to Gaussian predictions through an explicit width-dependent term. It explains precisely when a scalar guarantee transfers between the two decoders, without equating their measured accuracies. The original training objective does not enforce (14); neither this conditional result nor the nearest-center theorem certifies every trained Gaussian bank. Exact rational-arithmetic checks cover boundary ties, the reversal of Remark A.4, and the margin implications. The uncertainty-box certificate in Appendix B.2.1 continues to concern nearest-center boundaries.

B.5 Exact Global-Order Test for a Gaussian Bank

Theorem B.9 (finite exact Gaussian-order test)**.** *Fix strictly ordered centers in* $(0, 1)$ *and arbitrary positive widths. Define normalized distances* $d_j(s) = |s - c_j|/\sigma_j$*, with smallest-index ties, and let* $G(s) = \arg\min_j d_j(s)$*. Define the adjacent interior crossings*

$$\beta_i = \frac{\sigma_{i+1} c_i + \sigma_i c_{i+1}}{\sigma_i + \sigma_{i+1}}.$$

Then G *is nondecreasing on* $[0, 1]$ *if and only if*

$$G(0) = 0, \quad G(1) = Q - 1, \quad G(\beta_i) = i \quad (0 \le i < Q - 1).$$

When these tests hold, $G(s) = \sum_i \mathbf{1}\{s > \beta_i\}$ *everywhere. Checking the conditions requires* $O(Q^2)$ *normalized-distance comparisons and no scalar grid. This is an exact analytic test, not a numerical tolerance rule.*

Proof. At each center c_j, class j is the unique minimizer because its distance is zero and the other distances are positive. If G is monotone, it must consequently be zero at 0 and $Q - 1$ at 1. Between c_i and c_{i+1} only labels i and $i + 1$ can occur. Their distances have exactly one equality in that interval, at β_i, and the tie selects i. This proves necessity and forces the claimed switch location.

For sufficiency, consider the candidate cell $[\beta_{k-1}, \beta_k]$, interpreting the outer endpoints as 0 and 1. It contains c_k in its interior. For a competitor $j \neq k$, c_j is outside the open cell. Hence d_j is affine there while d_k is convex, and $d_j - d_k$ is concave. The tests at the right endpoint make $d_j - d_k \ge 0$ there; at the left endpoint the winning adjacent class $k - 1$ ties k, so the tests also make $d_j - d_k \ge 0$. At c_k the difference is strictly positive. Concavity on each side of c_k then makes the difference strictly positive throughout the open cell. At each internal right endpoint the explicit test selects k; at the outer endpoints the explicit endpoint tests select the correct extreme class. Therefore G has exactly the stated cells and is monotone. Computing $Q - 1$ crossings and comparing Q distances at those crossings and the two endpoints uses $O(Q^2)$ comparisons. □

Relation and limitation. With the weighted-Voronoi view of Section 2, whose diagram computation is established geometry, the contribution is this exact finite monotonicity criterion for the fitted DFIL decoder, not a new general Voronoi algorithm. The adjacent-width inequalities in Proposition B.7 remain an inexpensive sufficient condition. They need not hold for a globally monotone bank because a distant class can mask a pairwise tail crossing. Conversely, passing a finite uniform grid is not this exact test. No archived prediction is retroactively claimed to pass it.

Printed-bank applicability check. This condition is not automatic for the displayed learned banks. Treating the three 7B center/width rows in Table 26 as exact decimals, all have $G(0) = 1$ and $G(c_0) = 0$. In particular $c_0/\sigma_0 = 20$, while c_1/σ_1 is below .52. At c_0, class zero uniquely wins. Even allowing $\pm.0005$ around each of these first two displayed centers and widths leaves $c_0/\sigma_0 \geq 13$ and $c_1/\sigma_1 \leq 49/93 < 1$. This is a check of an algebraic condition on printed parameters, not recovery of checkpoint predictions or a new grid-frequency measurement. It illustrates why Theorem B.11 must keep its order precondition and why Theorem B.12 separately handles nonmonotone fixed-center banks.

Corollary B.10 (exact robust-order test on a full box)**.** *Let $0 < \ell_j \leq c_j \leq u_j < 1$, $u_j < \ell_{j+1}$, and independently $0 < v_j \leq \sigma_j \leq w_j$. Every bank in this Cartesian box is globally monotone if and only if every local corner satisfies all the following inequalities:*

$$\begin{aligned}
A_j &= c_j\sigma_0 - c_0\sigma_j \geq 0 && (j > 0),\\
B_j &= (1 - c_j)\sigma_{Q-1} - (1 - c_{Q-1})\sigma_j > 0 && (j < Q - 1),\\
C_{ij} &= \sigma_{i+1}(c_i - c_j) + \sigma_i(c_{i+1} - c_j) - \sigma_j(c_{i+1} - c_i) > 0 && (j < i),\\
E_{ij} &= \sigma_{i+1}(c_j - c_i) + \sigma_i(c_j - c_{i+1}) - \sigma_j(c_{i+1} - c_i) \geq 0 && (j > i + 1).
\end{aligned}$$

For an inequality, its local corners use only the center and width endpoints of the at most three class indices appearing in it. At most $2^6 = 64$ corners are tested per inequality. Hence the full robust test costs $O(Q^2)$ arithmetic operations and comparisons, rather than enumerating every corner of the $2Q$-dimensional box. Point intervals simply reduce the corner count.

Proof. At zero, $G(0) = 0$ is equivalent to $c_0/\sigma_0 \leq c_j/\sigma_j$ for all $j > 0$, which gives $A_j \geq 0$. At one, the highest index must strictly beat every lower index, giving $B_j > 0$. At the interior crossing β_i, the two adjacent distances are $(c_{i+1} - c_i)/(\sigma_i + \sigma_{i+1})$. For $j < i$, the competitor lies to its left; multiplying the comparison $(\beta_i - c_j)/\sigma_j > (c_{i+1} - c_i)/(\sigma_i + \sigma_{i+1})$ by the positive denominators gives $C_{ij} > 0$. A rival $j > i + 1$ lies to its right, and the analogous weak inequality gives $E_{ij} \geq 0$. The strictness is necessary precisely for smaller-index rivals. Thus these inequalities are equivalent, pointwise, to Theorem B.9's tests.

Each displayed polynomial is affine in every individual coordinate when the others are held fixed. A multiaffine function on a rectangle equals a convex combination of its corner values: write the affine interpolation in the first nondegenerate coordinate, then successively substitute the affine interpolation in each remaining coordinate. The resulting nonnegative weights sum to one, including on the boundary. Therefore nonnegativity everywhere is equivalent to nonnegativity at all corners. Strict positivity everywhere on this closed finite box is likewise equivalent to strict positivity at all corners, since a convex combination of finitely many positive values is positive. Coordinates not appearing in an inequality do not affect its value. Applying this fact to every order inequality proves necessity, sufficiency, and the operation bound. □

Strict gain over adjacent-width sufficiency. The centers $(1/20, 3/40, 3/20, 17/20)$ and widths $(19/40, 13/40, 3/4, 3/4)$ are a monotone bank that fails Proposition B.7's adjacent-width condition. Giving every center and width radius $1/10000$ still passes Corollary B.10 but fails the interval version of that sufficient condition. The local-corner check is exact for these rational inputs. Thus the robust test is not merely a rewriting of the old sufficient test. This witness establishes an analytic difference in domains, not its frequency among trained models.

B.6 Gaussian Decision Sets with Shared Centers and Uncertain Widths

Theorem B.11 (Gaussian shared-box extension)**.** *Use the preceding center and scalar box, with $0 < \ell_j \leq u_j < 1$ and $u_j < \ell_{j+1}$. Independently supply width intervals $0 < v_j \leq \sigma_j \leq w_j$. Assume the full center–width box passes the necessary-and-sufficient robust-order test in Corollary B.10. No additional adjacent-width ratio restriction is needed. Let $k_j \in \mathbb{N}_0$, $s = p\prod_j c_j^{k_j}$, and retain $S^{\pm}, d_j^{\pm}$ from Theorem B.5. Put $r_i^- = v_i/w_{i+1}$ and $r_i^+ = w_i/v_{i+1}$. Over the full scalar–center–width Cartesian box the exact extrema of the ratio to the Gaussian boundary are*

$$L_i^G = \frac{S^-(1 + r_i^+)}{d_i^- + r_i^+ d_{i+1}^-}, \qquad U_i^G = \frac{S^+(1 + r_i^-)}{d_i^+ + r_i^- d_{i+1}^+}.$$

Its exact attainable Gaussian-label set is

$$\{G_{c,\sigma}(s)\} = \left\{ \sum_i \mathbf{1}\{L_i^G > 1\}, \ldots, \sum_i \mathbf{1}\{U_i^G > 1\} \right\}.$$

The notation on the left means the set over all parameters in the supplied box. Thus singleton equality is necessary and sufficient for a constant Gaussian decision on that box. After the $O(Q^2)$ bank-box check, the arithmetic-operation count is $O(Q + m)$ for each chain of m factors.

Proof. By Corollary B.10, every Gaussian decoder in the full box is monotone and has the strictly ordered boundaries $\beta_i = (c_i + r_i c_{i+1})/(1 + r_i)$, where $r_i = \sigma_i/\sigma_{i+1}$. The tie convention is the same for every feasible bank.

For $p > 0$, its ratio is

$$R_i^G = \frac{s(1 + r_i)}{c_i + r_i c_{i+1}}.$$

A nonboundary center has log derivative $k_j/c_j \geq 0$. For c_i the derivative is $k_i/c_i - 1/(c_i + r_i c_{i+1})$; for c_{i+1} it is $k_{i+1}/c_{i+1} - r_i/(c_i + r_i c_{i+1})$. Each is strictly positive when its multiplicity is positive and strictly negative when its multiplicity is zero. The log derivative in p is positive. Finally, at fixed centers,

$$\frac{\partial}{\partial r_i} \frac{1 + r_i}{c_i + r_i c_{i+1}} = \frac{c_i - c_{i+1}}{(c_i + r_i c_{i+1})^2} < 0.$$

Thus the minimum uses the minimum endpoint choices of Theorem B.5 for the centers and base scalar, and $r_i = r_i^+$; the maximum uses the opposite choices and $r_i = r_i^-$. These width-ratio extremes are attained by the relevant two width endpoints. Other widths do not affect this boundary ratio. All chosen coordinates are jointly feasible because the domain is the full Cartesian box. This proves the displayed extrema; when $p = 0$, all ratios are zero and the same formulas apply by direct substitution.

At each feasible point the Gaussian boundaries are strictly ordered, so the ratio sequence is decreasing in boundary index. Its minimum and maximum sequences are therefore ordered. The endpoint-attainment argument of Theorem B.5 applies: attain the first lower ratio not exceeding one to attain the minimum label, and attain the last upper ratio exceeding one to attain the maximum label, with the extreme classes handled directly. Along a segment joining those two feasible points, s and the Gaussian boundaries vary continuously and the boundaries never coincide. A label can change only between adjacent classes, including the specified tie value; hence no intermediate label is omitted. Counting factors, evaluating products, and evaluating the $Q - 1$ ratios establishes the operation count. □

Scope. This extends the shared-center calculation to the Gaussian rule actually used in several existing protocols, without changing that rule. Corollary B.10 exactly decides uniform global ordering over the full box. Failure means at least one bank in that box is not globally ordered; it does not imply that the fitted point bank fails or that a particular queried interval is ambiguous. No claim is made that all archived widths pass the test. The cheaper sufficient inequalities $v_i/w_{i+1} > u_i/\ell_{i+1}$ and $w_i/v_{i+1} < (1 - u_i)/(1 - \ell_{i+1})$ remain an optional shortcut, but are not assumptions of this theorem. As for Theorem B.5, exactness concerns the full box; under extra constraints the image is an outer set (Appendix B.2.1). Fixing all widths to the same positive value recovers Theorem B.5. Independent, nondegenerate intervals that happen to have the same endpoints do not force equal widths and should not be confused with that special case.

A correlated-set counterexample to endpoint-only checking. Cartesian independence is substantive. Let $p = .8$, $c = (x, x + .01, x + .30)$, $x \in [.02, .60]$, and use $s = pc_2^2$ with equal fixed widths. At the two endpoints the nearest-center (and Gaussian) label is 1; at $x = .29$ it is 0, since $s = .27848 < (.29 + .30)/2 = .295$. The ratio to the first boundary has endpoint values $2048/625$ and $648/605$, both above one, but interior value $118/125 < 1$. Checking only the endpoints of this correlated curve would miss an attainable label. An enclosing Cartesian box still provides an outer set, but its tightness on the curve is not asserted.

B.7 Arbitrary Gaussian Widths with Fixed Centers

Fixed centers, arbitrary widths (Theorem B.12). With the centers fixed, class q is attainable if and only if some s in the scalar interval satisfies the Gaussian comparisons at the favorable width endpoints, w_q for q and v_j for each rival, strictly against lower indices. Splitting the interval at the at most $2(Q-1)$ equality roots per class decides every class exactly in $O(Q^3 + Q^2 \log Q)$ operations; the attainable set need not be contiguous.

Theorem B.12 (exact reachability without a monotonicity assumption)**.** *Fix distinct centers $c_0 < \cdots < c_{Q-1}$, a closed scalar interval $I = [s_-, s_+]$, and independent positive width intervals $[v_j, w_j]$. No ordering condition on the widths is imposed. A class q is attainable as the smallest-index Gaussian argmax for some scalar and width vector if and only if there exists $s \in I$ such that*

$$\frac{(s-c_q)^2}{w_q^2} < \frac{(s-c_j)^2}{v_j^2} \quad (j<q), \qquad \frac{(s-c_q)^2}{w_q^2} \le \frac{(s-c_j)^2}{v_j^2} \quad (j>q).$$

All attainable classes can be obtained by finite one-dimensional inequality intersection. For each q, split I at the equality roots against its $Q-1$ competitors, test all split points and one point in each open subinterval with the displayed strictness, and include q precisely when a test succeeds. At most $2(Q-1)$ roots are required for each q; a direct implementation uses $O(Q^3 + Q^2 \log Q)$ arithmetic operations and comparisons. The returned set need not be a contiguous label interval.

Proof. At fixed s, increasing σ_q weakly decreases its normalized squared distance, while decreasing any competitor width weakly increases that competitor's distance. Thus any feasible win by q persists when its width is set to w_q and all competitor widths to v_j, including strict inequalities against lower-index competitors. Conversely that endpoint vector is itself feasible, proving equivalence.

For positive widths, equality factors as equality or negation of normalized signed distances. It has an interior root $(v_j c_q + w_q c_j)/(v_j + w_q)$ and, when $v_j \neq w_q$, a possible exterior root $(v_j c_q - w_q c_j)/(v_j - w_q)$. Distinct centers prevent the comparison polynomial from vanishing identically. Signs of all comparison polynomials are constant between consecutive roots; checking a sample per open subinterval and all endpoints therefore determines feasibility exactly, including isolated tie-feasible points and a singleton input interval. Sorting $O(Q)$ roots and making $O(Q)$ comparisons at $O(Q)$ samples for each q gives the stated direct bound. □

Interpretation. With the fitted centers fixed, a supplied base interval maps exactly to I under the original product, so this analyzes unchanged Gaussian predictions even when their decoder is nonmonotone. For centers $(.2, .5, .8)$, widths $(.4, .02, .05)$, and $I = [.65, .95]$, the exact set is $\{0, 2\}$; filling the gap with class 1 would not be exact. Uncertain centers shared with the product are outside this theorem; Theorem B.11 handles one meaningful such case. A practical implementation must preserve inequality signs, for example by exact rational inputs or validated arithmetic. Ordinary floating-point sampling is not a certificate.

B.8 Exact Reachability with Shared Uncertain Centers

The proof of Theorem B.13 has three steps. Width monotonicity gives the favorable endpoint vector. At fixed s, t, h, the Gaussian comparisons cut each rival center down to its interval J_j. Their product image is an interval by continuity and connectedness, with the stated extremal values and endpoint flags. Requiring the retained s to belong to that image restores the *same* product equality used by decoding. Appendices B.8.1–B.8.5 give the complete statement, proof, rational formulation and a strict-endpoint counterexample.

This appendix analyzes the unchanged proportional product and Gaussian argmax. It does not change DFIL training, report additional model evaluations, or assume that an uncertainty set has semantic coverage. All statements concern exact real arithmetic. The two-variable reduction below is distinct from the shared-box and fixed-center results of Appendices B.6 and B.7, which retain their own statements and scopes.

B.8.1 Problem and Endpoint Conventions

Let $Q \geq 2$, $k_j \in \mathbb{N}_0$, and $m = \sum_j k_j$. Supply a nonempty full Cartesian set

$$0 \leq p^- \leq p \leq p^+ \leq 1, \qquad 0 < \ell_j \leq c_j \leq u_j \leq 1, \quad u_j < \ell_{j+1}, \qquad 0 < v_j \leq \sigma_j \leq w_j.$$

Define $s = p \prod_j c_j^{k_j}$ and $G_{c,\sigma}(s) = \arg\min_j |s - c_j|/\sigma_j$, breaking ties toward the smaller index. Repeated occurrences of a center use the same variable. An attainable label means that a feasible parameter vector realizes that label, including feasible isolated ties; it does not mean that the label has positive probability under a specified distribution.

For a nonempty bounded interval J, let $A = \inf J$, $B = \sup J$ and write $a = 1\{A \in J\}$, $b = 1\{B \in J\}$. Thus a singleton is nonempty only when both endpoints are included. We preserve these endpoint flags rather than replacing open intervals by closed hulls.

B.8.2 Two-Variable Reduction

Theorem B.13 (exact reachability without global Gaussian order)**.** *Fix a candidate label q. Retain two real variables $s \in [0,1]$ and $t \in [\ell_q, u_q]$, with t denoting the same c_q used in the product and decoder. Set $d = |s - t|$, $\rho_j = v_j/w_q$ for $j \neq q$, and $J_q = \{t\}$. For one cut $h \in \{0, \ldots, Q\}$, require $t \leq s$ when $q < h$ and $t \geq s$ otherwise. Define the rival intervals by*

	$j < q$	$j > q$
$j < h$	$[\ell_j, u_j] \cap (-\infty, s - \rho_j d)$	$[\ell_j, u_j] \cap (-\infty, s - \rho_j d]$
$j \geq h$	$[\ell_j, u_j] \cap (s + \rho_j d, \infty)$	$[\ell_j, u_j] \cap [s + \rho_j d, \infty)$.

Require every J_j to be nonempty. For their endpoints and flags define

$$P_- = p^- \prod_j A_j^{k_j}, \qquad P_+ = p^+ \prod_j B_j^{k_j},$$

$$e_- = 1\{p^- = 0\} \vee \bigwedge_{j:k_j>0} a_j, \qquad e_+ = 1\{p^+ = 0\} \vee \bigwedge_{j:k_j>0} b_j.$$

Then q is attainable if and only if there exist s, t, h as above such that

$$\big(s > P_- \ \vee \ (s = P_- \wedge e_-)\big) \quad \wedge \quad \big(s < P_+ \ \vee \ (s = P_+ \wedge e_+)\big). \tag{16}$$

For rational inputs, this is a finite Boolean combination of polynomial comparisons with only two real unknowns; h contributes $Q + 1$ cases. The result makes no global-order assumption on G. Its attainable label set may be disconnected.

Proof. Eliminating widths. Fix s, c. Label q wins exactly when

$$|s - c_q|/\sigma_q < |s - c_j|/\sigma_j \ (j < q), \qquad |s - c_q|/\sigma_q \leq |s - c_j|/\sigma_j \ (j > q).$$

As in the proof of Theorem B.12, raising σ_q to w_q weakly decreases the left sides and lowering each rival width to v_j weakly increases the right sides, so every weak and every strict win comparison is preserved; any feasible win persists at this width vector, which belongs to the full width box. Different candidate classes may use different width endpoint vectors, as each attainability question is existential. Width feasibility is therefore equivalent to $|s - c_j| > \rho_j |s - c_q|$ for $j < q$ and the weak version for $j > q$. This elimination uses independence of the width intervals, not Gaussian monotonicity.

Eliminating rival positions at fixed s, t. Since the center box is strictly ordered, any feasible vector admits a cut h: $c_j \leq s$ on its left and $c_j \geq s$ on its right. Equality at a center can be assigned to either adjacent cut. On the left, $|s - c_j| = s - c_j$, so the win comparison gives the stated upper cutoff. On the right it gives the stated lower cutoff. The sign condition for t handles the candidate center. Thus for this cut the feasible center vectors are exactly $\prod_j J_j$, with $J_q = \{t\}$. All vectors in this Cartesian product retain the original ordering because the original intervals were separated. No center appearing in the product has been replaced by an independently chosen decoding center.

Restoring the product equality. On the nonempty connected set $[p^-, p^+] \times \prod_j J_j$, the continuous map $F(p, c) = p \prod_j c_j^{k_j}$ has a connected image in $\mathbb{R}$, hence an interval. Every center is strictly

positive. Monotonicity in each coordinate gives infimum P_- and supremum P_+; these values can be approached by simultaneously approaching the respective interval endpoints. If $p^- > 0$, the minimum is attained exactly when all positive-exponent lower endpoints are included: strict monotonicity in any such missing coordinate otherwise makes the product larger. If $p^- = 0$, the value zero is attained for any feasible centers. The same argument applies at the upper endpoint, with $p^+ = 0$ treating the identically-zero product. Coordinates with exponent zero never affect attainment. These are precisely e_-, e_+. Every strict intermediate value is attained by connectedness, even if the parameter domain is open at some endpoints. Thus its image is exactly the interval specified by (16). In particular (16) is equivalent to the existence of one feasible p, c whose product is the retained s.

Combining the three equivalences proves both implications. At rational inputs ρ_j is rational. Within a cut d is affine in s, t, rival endpoints are minima or maxima of affine expressions, and nonemptiness and inclusion flags are affine comparisons. Expanding their finitely many branches yields polynomial comparisons of degree at most $\max(1, m)$ in s, t. A compact implementation may retain the minima and maxima as if–then–else expressions. This eliminates continuous variables; it does not assert a linear-time solver or constant bit cost at arbitrary depth. □

B.8.3 Two Analytic Witnesses and the Role of Ties

Take three centers in $[.19, .21] \times [.49, .51] \times [.79, .81]$, widths in $[.39, .41] \times [.019, .021] \times [.049, .051]$, and product $s = pc_2$. For $p \in [.85, 1]$ the exact set is $\{0, 2\}$. At midpoint parameters $p = .85$ gives $s = .68$ and label zero, whereas $p = 1$ gives $s = c_2$ and label two. Class one is impossible: $s \in [.6715, .81]$, so its normalized distance is at least $(.6715 - .51)/.021 > 7$, while class zero's is at most $(.81 - .19)/.39 < 2$. Filling the gap with class one would not be exact. Tightening only p to $[.98, 1]$ gives the singleton $\{2\}$: distance to class two is at most $.02(.81)/.049 < .331$, while class zero's distance is at least $(.98(.79) - .21)/.41 > 1.37$ and class one's is larger still. These are algebraic examples, not model experiments. The midpoint bank is globally nonmonotone because $G(0) = 0$, $G(.5) = 1$, and $G(.68) = 0$, so the box cannot pass the global-order prerequisite of Theorem B.11.

Endpoint inclusion is essential. Let $q = 1$, $p = .9$, $c_0 \in [.2, .5]$, $c_1 = .8$, widths $(.1, 1)$, and $s = pc_0$. Fix $s = 36/95$ and $t = 4/5$. Winning against index zero requires $c_0 > 8/19$, but the product equality forces $c_0 = 8/19$. The allowed center interval is nonempty but the product infimum is not attained. Replacing the strict comparison or its product-endpoint flag by a weak one incorrectly makes this instance feasible.

B.8.4 Relation, Implementation, and Limits

Relative to the weighted-Voronoi and uncertain nearest-neighbor work discussed in Section 2, the specific step is to retain the endogenous query $s = p \prod_j c_j^{k_j}$, whose operands are also the uncertain sites, while eliminating rival centers without breaking that equality. The reduction extends Theorems B.11–B.12 in a different direction: arbitrary positive width boxes and uncertain shared centers are handled together, without a global Gaussian-order assumption.

The mathematical checks compare a direct rational nonlinear-real formula with the reduced two-variable formula. Finite cross-checks on 177 label-feasibility instances returned matching answers (88 satisfiable, 89 unsatisfiable), with no unknown answers in this collection. A deliberately incorrect weak-tie variant fails the endpoint witness above. The collection includes 52 rational boxes with two to four classes, zero and point base intervals, empty and repeated-factor chains, and 13 fixed-(s, t) queries. Both formulations use the same solver; like the checks in Appendix B.11, they are not empirical model evaluations, a formal proof, or independent expert reviews. The comparisons use exact rational inputs. Solver timeouts must be returned as unknown, never interpreted as infeasibility. Appendix B.10 repeats the comparison on 320 random instances with up to eight classes and ten factors, without disagreement or timeout.

B.8.5 Exactness on the Deployed Center Family

The full-box assumption in Theorem B.13 can be weakened to the chain constraints generated by the actual center parameterization. This extension is useful when supplied center intervals overlap

or their Cartesian product includes centers excluded by the minimum increments. It still does not cover arbitrary correlated uncertainty in network weights or raw logits.

Lemma B.14 (strict interval-chain feasibility and endpoint attainment)**.** *Let J_j be bounded intervals, possibly open at either endpoint, with algebraic lower/upper bounds A_j, B_j and exclusion flags $o_j^-, o_j^+ \in \{0,1\}$. A flag one means that the endpoint is excluded. Let $g \geq 0$, and impose $c_j - c_{j-1} > g$ for $j = 1, \ldots, Q-1$. Define a lower-bound maximum as follows: the value is the larger of two candidate lower bounds; its exclusion flag is the flag of the larger one, or their logical OR when the values tie. Set*

$$(L_0, \lambda_0) = (A_0, o_0^-), \quad (L_j, \lambda_j) = \max_{\text{lower}}\{(A_j, o_j^-), (L_{j-1} + g, 1)\}. \tag{17}$$

The interval-chain set $\mathcal{F}(J, g)$ is nonempty if and only if, for every j,

$$L_j < B_j \quad \text{or} \quad (L_j = B_j \ \wedge \ \lambda_j = 0 \ \wedge \ o_j^+ = 0). \tag{18}$$

When it is nonempty, $L_j = \inf_{c \in \mathcal{F}} c_j$; all coordinates with $\lambda_j = 0$ can simultaneously attain these infima, and a coordinate with $\lambda_j = 1$ cannot attain its infimum. Define the upper-bound minimum with the analogous tie-OR flag, and set

$$(U_{Q-1}, \nu_{Q-1}) = (B_{Q-1}, o_{Q-1}^+), \quad (U_j, \nu_j) = \min_{\text{upper}}\{(B_j, o_j^+), (U_{j+1} - g, 1)\}. \tag{19}$$

Then the symmetric assertions hold for the suprema U_j, with ν_j indicating nonattainment. These are linear forward/backward recurrences once the interval bounds and flags are given.

Proof. Induction in j shows that every feasible vector satisfies $c_j \geq L_j$, strictly when $\lambda_j = 1$. Indeed, an open lower window imposes a strict comparison, and the strict adjacent constraint gives $c_j > c_{j-1} + g \geq L_{j-1} + g$ independently of the previous flag. Taking the larger bound, with a tie retaining either strict requirement, gives exactly (17). Comparing to the upper window proves necessity of (18).

For sufficiency and attainability, for $\eta > 0$ construct

$$c_0(\eta) = A_0 + \eta o_0^-, \qquad c_j(\eta) = \max\{A_j + \eta o_j^-,\ c_{j-1}(\eta) + g + \eta\}. \tag{20}$$

Each coordinate satisfies its lower window and every adjacent gap is strictly greater than g. Continuity of the finite maximum recurrence gives $c_j(\eta) \to L_j$ as $\eta \downarrow 0$. If $\lambda_j = 0$, the recurrence defining it must select its own closed lower bound; for $j > 0$ it must strictly exceed $L_{j-1} + g$, since the competing propagated bound has flag one. Thus for all sufficiently small positive η, simultaneously over the finite list of such indices, $c_j(\eta) = A_j = L_j$. For $j = 0$ the same fact follows directly from $o_0^- = 0$.

Under (18), every upper bound either has positive slack above L_j, which sufficiently small η respects even if the upper endpoint is open, or equals L_j with both flags zero. In the latter case the preceding argument makes $c_j(\eta)$ exactly that allowed endpoint. Taking the minimum of finitely many permissible positive choices of η yields a feasible vector. Consequently all infima are approached together, all zero-flag coordinates attain them together, and one-flag coordinates cannot attain them by the necessity argument. This proves the entire lower assertion. Reverse the index order and replace a coordinate by its negative to obtain the upper statement, or apply the identical argument to $\min\{B_j - \eta o_j^+, c_{j+1}(\eta) - g - \eta\}$. This gives (19) and its simultaneous endpoint property. □

Theorem B.15 (two-variable reduction with strict center-gap constraints)**.** *In Theorem B.13 replace the separated center box by the nonempty set*

$$\mathcal{C}_g = \{c : \ell_j \leq c_j \leq u_j,\ c_j - c_{j-1} > g\ (j \geq 1)\}, \quad 0 < \ell_j \leq u_j \leq 1, \quad g \geq 0. \tag{21}$$

Center intervals need not be disjoint. Optionally impose a fixed lower anchor $c_0 > a + g$ and point intervals such as $c_{Q-1} = C$. Widths and the base scalar retain their independent supplied intervals.

For fixed q, s, t, h, construct J_j by the same Gaussian cutoffs as Theorem B.13 and intersect J_0 with $(a + g, \infty)$ when an anchor is imposed. Apply Lemma B.14. Reject the cut if (18) *fails. Otherwise replace Theorem B.13's product bounds and inclusion flags by*

$$P_- = p^- \prod_j L_j^{k_j}, \qquad P_+ = p^+ \prod_j U_j^{k_j}, \tag{22}$$

$$e_- = 1\{p^- = 0\} \vee \bigwedge_{j:k_j>0} (\lambda_j = 0), \qquad e_+ = 1\{p^+ = 0\} \vee \bigwedge_{j:k_j>0} (\nu_j = 0). \tag{23}$$

Then the same endpoint-aware test (16) *is necessary and sufficient for attainability of q, after existentially choosing s, t and one of the $Q+1$ cuts. The formula still has only two real unknowns. No global Gaussian-order assumption is used.*

Proof. The width elimination in Theorem B.13 does not depend on center independence, since the width variables remain independent of the centers. Positive strict gaps imply ordered centers; consequently every feasible vector still has one of the stated cuts. At fixed s, t, h, the remaining center set is precisely $\mathcal{F}(J, g)$, not the Cartesian product of the windows. The anchor is an additional open lower window, and the candidate center is a point window $J_q = \{t\}$; both fit Lemma B.14.

That lemma decides nonemptiness and identifies simultaneous coordinate infima and suprema. Every center is positive. Thus monotonicity of the product and the simultaneous limiting constructions give exactly the infimum and supremum in (22). If $p^- > 0$, attainment of the lower product requires every positive-exponent coordinate to attain its infimum: any strict increase in such a coordinate increases the positive product. By Lemma B.14 this is possible if and only if every corresponding $\lambda_j = 0$, and all these endpoints can then be attained together. With $p^- = 0$, zero is attained at any feasible center vector. The upper endpoint follows by the same argument, including the identically-zero case $p^+ = 0$. This proves (23).

The center set is an intersection of intervals and strict linear half-spaces, hence convex and connected, though not necessarily closed. Its product with the base interval is connected. The product map is continuous and its real image is therefore an interval; its extremal values and inclusion are now completely specified by (22)–(23). Membership of s in that interval is equivalent to the original shared product equality at a feasible center vector. Combining this equivalence with the unchanged Gaussian comparisons gives both directions.

All recurrences involve minima, maxima, additions of the rational constant g, and Boolean endpoint flags. Expanding their branches again yields polynomial comparisons in only s, t, with degree at most $\max(1, m)$. The recurrences are interval-chain elimination operations, not extra real variables. They can be represented with shared expressions in an SMT formula. There is no claimed linear total solve time; interval overlap can make solving harder despite the reduced dimension. □

Corollary B.16 (the cumulative-softmax center image is covered exactly)**.** *For the deployed parameterization* (12)*, put*

$$g = \frac{.96\epsilon}{1 + Q\epsilon}, \qquad a = .02, \qquad C = .98.$$

Its center image at finite raw logits is exactly

$$c_0 > a + g, \qquad c_j - c_{j-1} > g \ (j \geq 1), \qquad c_{Q-1} = C. \tag{24}$$

Intersecting this image with supplied center intervals, then using Theorem B.15, gives the exact attainable Gaussian-label set for that center family and the specified independent scalar/width intervals. To match the deployed width parameterization as well, the supplied intervals must lie in $[10^{-4}, \infty)$.

Proof. Necessity follows as in the proof of Proposition A.1: every center increment, including $c_0 - a$, is .96 times a normalized increment exceeding $\epsilon/(1 + Q\epsilon)$, hence strictly greater than g, and the increments sum to .96. Conversely, for any center vector satisfying (24), set $c_{-1} = a$; the normalized increments $(c_j - c_{j-1})/.96$ are then positive, exceed $\epsilon/(1+Q\epsilon)$ and sum to one, so Proposition A.2 gives finite raw logits producing exactly these increments and hence the stated centers. Every width at or above 10^{-4} is represented by $\max\{\exp(w_j), 10^{-4}\}$. The center constraints therefore have exactly the form in Theorem B.15, with a strict anchor and fixed final center. □

Why open constraints cannot be replaced silently by their closure. Let $c_0 \in [.4, .5]$, $c_1 = .6$, equal fixed widths, $p = 11/12$ and $s = pc_1 = .55$. The closed Cartesian box attains labels $\{0, 1\}$: at $c_0 = .5$ the midpoint equals .55, and the tie selects zero. Imposing $c_1 - c_0 > .1$ excludes that endpoint; every feasible midpoint is strictly below .55, so the exact set becomes $\{1\}$. The strict gap changes an attainable decision, not just a measure-zero detail that a universal certificate may ignore. Analogous strict boundaries occur in (24).

Checks and remaining scope. A separately written direct specification keeps every center, width and base variable and imposes the gap constraints explicitly. The reduced specification retains only s, t and symbolic endpoint flags. In the final 61-instance nonlinear check collection, 58 direct/reduced pairs both terminated and agreed (36 satisfiable and 22 unsatisfiable); the reduced solver returned unknown at the per-instance limit on three additional instances, while the direct formulation terminated. No answer was inferred from those unknown outcomes. Separately, 240 exact linear interval-chain tests and 288 joint endpoint-attainment tests agreed with the recurrences. They include zero gaps, open and closed windows, point windows, zero base intervals, empty/repeated-factor chains, overlapping center windows, and exact cumulative-softmax constraints for two, three, four and eight classes. These finite checks have the status stated in Appendix B.11.

Corollary B.16 avoids replacing the deployed center family by a larger Cartesian set. These results assume externally supplied intervals and a fixed composition structure; widths are membership shapes, not confidence radii. They do not remove the need for meaningful input uncertainty, certify a learned encoder over unrestricted prose, handle arbitrary dependence among base scalars and parameters, or prove faster numerical solving or an empirical speedup. Main-classifier and probability-mixture outputs remain outside this certificate, and mathematical set exactness does not establish semantic calibration or end-to-end accuracy. For a different constrained set, an enclosing set is an outer approximation unless a new exact elimination is established.

B.9 Scope and Limitations of the Extensions

The external human-target confirmation tests task utility beyond method-generated labels; the matched controls constrain attribution to DFIL's learned bank or fuzzy loss.

Limitations. The one-dimensional representation requires a defensible scalar target and adequate extraction. A point center does not capture all contextual or speaker variation in vague language. Natural-249 uses model-generated, product-derived labels, and independent counting does not reproduce a DFIL-specific composition advantage. The positive human-target result uses one 1.5B backbone and a validation-selected mixture, with seed-level uncertainty on a fixed test set. The main classifier and mixed output have no nearest-center guarantee. In the initial diagnostics, the 32B results show large dispersion in both conditions, and a six-seed replication under the corrected selection protocol does not establish its cause or remove it (Appendix D.2). MF saturation also had to be corrected in some of the initial diagnostic runs. These limitations define where further evaluation is needed.

B.10 Runtime and Agreement of the Exact Procedures

We compare the three exact procedures with two alternatives on the same random instances. The direct formulation keeps every uncertain quantity as a solver variable: the base scalar, all centers, and for Gaussian decoding all widths, with one nonlinear-real satisfiability query per label. Random sampling draws $2{,}000$ (nearest center) or $4{,}000$ (Gaussian) parameter vectors uniformly from the box. Each instance has ordered centers with random spacings, disjoint center intervals whose half-widths are 10–49% of the adjacent gap, a base interval of half-width $.02$–$.3$, and a random chain of length m; Gaussian instances add width intervals with lower endpoints in $[.02, .5]$ and relative spread up to 50%. There are 20 instances per configuration. Nearest-center instances use $Q = 4, 8, \ldots, 256$ and $m \in \{1, 2, 4\}$. Fixed-center instances use the box midpoints as centers, $Q = 4, \ldots, 128$ and $m \in \{1, 2\}$. Shared-center instances use $Q \in \{3, 4, 6, 8\}$ with $m \in \{1, 2, 3\}$ and $Q \in \{4, 8\}$ with $m \in \{6, 10\}$. The closed forms use exact rational arithmetic. The two-variable reduction and every direct formulation, run for $Q \leq 32$, use Z3 (de Moura & Bjørner, 2008) with a 10 s limit per query. Timings are single-process and indicate orders of magnitude only.

On 720 instances with a direct formulation, comprising $7{,}740$ label queries, every solver call terminated and every attainable set equals the exact one. The closed-form certificate takes $37\,\mu$s at $Q = 4$ and 2.7 ms at $Q = 256$. From $Q = 4$ to 32 it is $1{,}300$ to $2{,}000$ times faster than the direct formulation, and fixed-center root enumeration is 60 to 190 times faster, a margin that narrows with its cubic cost (Figure 4, left). For shared centers, the reduction decides all $1{,}740$ label queries, each in under 1.7 s. The direct formula is 1.3 to 2.7 times faster at these sizes, so Theorem B.13 is used for exactness with two unknowns, not for speed.

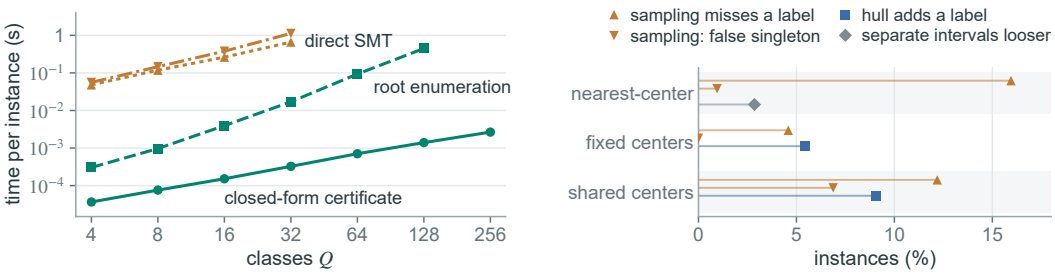


Figure 4: Exact decisions are fast, and the shortcuts fail on ordinary instances. Left: median time per instance against Q, pooled over chain lengths, for the closed-form nearest-center certificate (circles), fixed-center root enumeration (squares) and the direct formulations of both problems (triangles). Right: share of instances in which sampling misses an attainable label or returns a false singleton, the interval between extreme exact labels adds an unattainable one, or separate product and boundary intervals return a larger set. Nearest-center sets are intervals by Theorem B.5; separate propagation applies only to that problem.

The alternatives fail on ordinary instances (Figure 4, right). Sampling misses an attainable label in 16.0% of nearest-center, 4.6% of fixed-center and 12.2% of shared-center instances. In 6.9% of shared-center instances it returns a singleton although another label is attainable, which would be issued as a certified decision. The exact Gaussian sets are disconnected in 5.4% of fixed-center and 9.1% of shared-center instances, where the interval between extreme labels adds an unattainable one. Separate product and boundary intervals are strictly looser in 2.9% of these random nearest-center boxes, and in 2 of 420 they lose a singleton. These are checks of implementations against theorems on synthetic boxes, not model evaluations.

B.11 Mathematical Checks and Their Status

Rational-arithmetic checks cover the partition and decoder tests. They check 90 finite-alphabet configurations with 18,490 transition/suffix comparisons, 110 margin configurations with 62,232 demanded-bit comparisons, and 1,500 Gaussian banks against their complete pairwise-root arrangements, including ties. The initial Gaussian-box checks cover 38 boxes, 2,848 complete endpoint vectors, and 1,140 interior points. The strengthened robust-order tests cover a separate 130 boxes and compare 15,528 local-corner inequalities with 12,832 full endpoint banks; ratio checks include 27,744 evaluations. The arbitrary-width checks cover 160 fixed-center interval cases and 2,240 width-endpoint vectors. No mismatch was found in these finite checks.

A separate floating-point toy check of Proposition C.3 uses 320 bounded-logit examples. The largest observed absolute loss-identity error is 7.55×10^{-15} and the largest finite-difference gradient error is 2.19×10^{-10}. Symbolic differentiation and rational witnesses check the scope counterexamples above. These are checks for proof errors, not statistical validation or new empirical evaluations. The two program formulations are not independent expert reviews, and no formal proof assistant was used. Their success cannot replace the proofs or author verification. No model was trained, loaded, or run for these checks, and no archived prediction, seed, dataset, or empirical endpoint was added or changed.

C Analysis of the Training Objective

C.1 Gradient Geometry of the Dual-Path Loss

Both scalar-path losses act on h through one direction. With $u = \nabla_h z$, residual $r = z - p$, and $a = \partial \ell_{\mathrm{MF}} / \partial z$, their gradients are au and $2ru$. When both gradients are nonzero, they align if and only if $ar > 0$ and oppose if and only if $ar < 0$ (Proposition C.2). If either gradient vanishes, no directional comparison is asserted. Detachment removes the MF gradient through its input, but does not by itself remove optimizer coupling or imply a convergence advantage (Appendix C.3).

The DFIL training objective (1) combines three terms. We analyze the interaction between these terms and their effect on the learned representations.

C.1.1 Gradient Analysis

Consider the gradient of the total loss with respect to the backbone parameters θ_{LoRA}:

$$\nabla_\theta \mathcal{L} = \underbrace{\nabla_\theta \mathcal{L}_{\text{CE}}(\hat{y}, y)}_{\text{classification signal}} + \lambda_{\text{DFIL}} \underbrace{\nabla_\theta \mathcal{L}_{\text{CE}}(\boldsymbol{\mu}, y)}_{\text{MF alignment signal}} + \lambda_p \underbrace{\nabla_\theta \|\hat{p} - p\|_2^2}_{\text{proportion regression signal}} \tag{25}$$

When g is the identity, MF cross-entropy can update the bank, numerical head, and LoRA parameters. When g is stop-gradient, its derivative with respect to the extractor and LoRA parameters is zero; the bank can still receive that loss. The regression term supplies scalar supervision in both cases. These are computational dependencies, not guarantees of semantic alignment.

Remark C.1 (Gradient roles)**.** *The classification, MF, and regression losses act through different computational paths, so their gradients can differ. They can also align, vanish, or conflict. Without assumptions on the parameters and data, the decomposition above implies neither complementary directions nor an optimization benefit.*

The MF loss rewards agreement with the supplied labels through the scalar path. Whether its gradients improve extraction, calibration, or task accuracy is an empirical question.

Proposition C.2 (Scalar-path gradient geometry and conflict)**.** *For one example, fix the MF parameters and a differentiable scalar extractor $z = z(\mathbf{h})$. Let $\mathbf{u} = \nabla_{\mathbf{h}} z$, $r = z - p$, and $a = \partial \ell_{\text{MF}} / \partial z$, where ℓ_{MF} is the implemented cross-entropy using memberships as logits. With an attached MF input,*

$$\nabla_{\mathbf{h}} \ell_{\text{MF}} = a\mathbf{u}, \qquad \nabla_{\mathbf{h}} (z - p)^2 = 2r\mathbf{u}, \qquad \nabla_{\mathbf{h}} \mathcal{L} = \mathbf{g}_{\text{cls}} + (\lambda_{\text{DFIL}} a + 2\lambda_p r)\mathbf{u}.$$

For nonzero scalar-path gradients, their directions agree exactly when $ar > 0$ and oppose each other exactly when $ar < 0$. If the MF input is detached, its computed gradient to $\mathbf{h}$ is zero; the regression gradient remains $2r\mathbf{u}$. The same statements hold for parameters entering these two losses only through z.

Proof. The chain rule gives both scalar multiples of $\mathbf{u}$ and their sum. Their inner product is $2ar\|\mathbf{u}\|^2$, establishing the sign criterion; if a, r, or $\mathbf{u}$ vanishes, at least one gradient is zero. For Gaussian memberships and a positive per-example class weight w_y, put $\pi_q = \exp(\mu_q) / \sum_j \exp(\mu_j)$. The derivative is explicitly

$$a = w_y \sum_q (\pi_q - \mathbf{1}\{q = y\})\, \mu_q(z) \frac{c_q - z}{\sigma_q^2},$$

since $\partial \mu_q / \partial z = \mu_q(z)(c_q - z)/\sigma_q^2$. A batch reduction can add a common positive normalization without changing the per-example sign. Stop-gradient sets the computed derivative through its input to zero, while leaving derivatives with respect to the bank parameters available. □

A conflict can occur even inside the correct decision cell. Take $Q = 2$, $c = (.25, .98)$, equal widths .1, target $p = .8$ with label $y = 1$, and estimate $z = .9$. Both p and z have nearest-center label 1. Here $\mu_0'(z) < 0$ and $\mu_1'(z) > 0$, so the formula gives $a < 0$, whereas $r = .1 > 0$. Gradient descent on the MF loss moves z upward, toward its class center; regression moves it downward, toward the supplied proportion. Thus classification alignment and scalar fidelity can conflict without a misclassified example. This bank is feasible in Equation (12); the witness is an analytic construction, not a measured training outcome.

The main classifier can supply directions outside the scalar path, but no non-parallelism is guaranteed. The proposition is a per-example decomposition; gradients summed over examples can have additional interactions. It explains precisely what detachment removes and what remains to be evaluated empirically, without assuming an optimization or generalization benefit. Numerical checks compare the displayed derivative and chain rule with finite differences, including the conflict example.

C.1.2 ROLE OF λ_{DFIL} AND λ_p

The small-seed ablations and weight sweep do not establish an optimal balance or complementary gradient directions.

Three regimes of the dual-path objective. **(i) No scalar auxiliary losses.** Setting $\lambda_{\text{DFIL}} = \lambda_p = 0$ leaves the main-classifier cross-entropy as the training signal. This identifies an objective-level relation to fine-tuning while retaining the architecture; it does not train a useful scalar head by itself. The implemented MF-only ablation instead trains $\mathcal{L}_{\text{CE}}(\boldsymbol{\mu}, y) + \lambda_p(z-p)^2$, so it is a different objective and inference path.

(ii) Strong MF pressure. With attached MF inputs, increasing λ_{DFIL} weights the MF gradient $a\mathbf{u}$ more heavily relative to the other terms in Proposition C.2. Depending on the example, this can reinforce or oppose proportion regression. It does not algebraically turn the dual-path model into the MF-only variant. Under detachment, changing this weight affects bank training directly but does not add that MF gradient to the numerical head or LoRA. The observed MF saturation and the corresponding reruns therefore need to be read together with the actual computational path.

(iii) Joint training with finite weights. The original component study uses $\lambda_{\text{DFIL}} = \lambda_p = .5$; the original scaling configuration uses $\lambda_{\text{DFIL}} = .1$. Equal coefficients do not imply equal gradient norms or complementary directions. The single-seed sweep reports 49.9–51.2% accuracy over its tested weights (Table 8); it is sensitivity evidence for that run, not a proof of a globally optimal balance. These regimes connect the loss-design choices to an explicit gradient calculation.

C.2 WHAT THE MEMBERSHIP TEACHER SUPPLIES

Proposition C.3 makes the term concrete: with memberships in $[0,1]$, the detached teacher $t = \text{softmax}(\mu/\tau)$ satisfies $\max_q t_q \le e^{1/\tau}/(e^{1/\tau}+Q-1)$ and $\text{KL}(t\|u) \le 1/(8\tau^2)$ against the uniform u, so at $\tau = 2$ and $Q = 8$ no class exceeds .191 and $\text{KL}(t\|u) \le 1/32$. The distillation loss is thus exactly a uniform-target temperature loss plus a bounded linear correction on the student logits. This rewriting is consistent with the absence of an ordering-specific gain but does not attribute the measured effect.

Proposition C.3 (exact loss decomposition; supporting analysis)**.** *Let $\mu \in [0,1]^Q$ and $\tau > 0$. Set $t = \text{softmax}(\mu/\tau)$ and $s = \text{softmax}(z/\tau)$, and let u be uniform on the Q classes. Treat t as detached, as in the evaluated teacher extension. Then*

$$\tau^2 \,\text{KL}(t\|s) = \tau^2 \,\text{CE}(u,s) - \tau \sum_q (t_q - u_q) z_q + \tau^2 \sum_q t_q \log t_q,$$

$$\nabla_z[\tau^2 \,\text{KL}(t\|s)] = \tau(s-t),$$

$$\max_q t_q \le \frac{e^{1/\tau}}{e^{1/\tau}+Q-1}, \qquad \text{KL}(t\|u) \le \frac{1}{8\tau^2}.$$

Here $\text{CE}(u,s) = -\sum_q u_q \log s_q$. Thus the stopped teacher gives a uniform-target temperature loss plus an input-dependent linear correction with bounded coefficients on the student logits. This is an equivalent expression for the existing loss, not a proposed replacement or a generalization guarantee.

Proof. Writing $Z = \sum_j e^{z_j/\tau}$ gives $\log s_q = z_q/\tau - \log Z$. Substitute this into the KL and separate $t = u + (t-u)$; the last vector sums to zero, canceling its $\log Z$ term and giving the identity. Differentiating with t fixed gives $\tau(s-t)$. For the maximum probability, every logit difference $\mu_j - \mu_q$ is at least -1, so

$$t_q = \frac{1}{1 + \sum_{j \ne q} e^{(\mu_j - \mu_q)/\tau}} \le \frac{1}{1 + (Q-1)e^{-1/\tau}}.$$

For the entropy bound put $\beta = 1/\tau$ and $\psi(\beta) = \log(Q^{-1}\sum_j e^{\beta\mu_j})$. Direct differentiation gives $\psi'(\beta) = \sum_j t_j \mu_j$ and $\psi''(\beta) = \text{Var}_t(\mu_j)$. Since $0 \le \mu_j \le 1$, the variance is at most $m - m^2 \le 1/4$, where $m = \sum_j t_j \mu_j$ (use $\mu_j^2 \le \mu_j$). Also $\text{KL}(t\|u) = \beta\psi'(\beta) - \psi(\beta)$, zero at $\beta = 0$. Integrating its derivative $\beta\psi''(\beta)$ from zero to β proves the bound. □

Interpretation and non-attribution. The evaluated eight-class, $\tau = 2$ teacher has largest class probability below .191 and KL to uniform at most $1/32$ nats. These are analytic bounds, not measurements of the saved teacher outputs. They apply to ordered and unordered Gaussian membership banks alike. The decomposition explains what information can enter the distillation term; it does not show that the uniform component or its correction caused the observed gains. This is an application of elementary softmax calculus and bounded variance, included to interpret the existing stronger teacher controls rather than as a headline novelty claim.

C.3 Gradient Paths Versus Optimizer Trajectories

Stop-gradient makes the direct MF-loss derivative with respect to the extractor zero (Proposition C.2). Equality of complete training trajectories is a stronger assertion. A sufficient condition for equality, proved by induction, is that the extractor/main update is block-separable from bank gradients and other bank-dependent update decisions, with the same initialization, optimizer state, minibatches, randomness, and scalar/main losses. A common global clipping factor can violate block separability: for scalar gradient $g = 1$, bank gradients $h = 0$ or $h = 2$, and norm cap one, the clipped scalar gradients are respectively 1 and $1/\sqrt{5}$. Even a first Adam step with a positive denominator epsilon need not be invariant to this rescaling. This is an analytic counterexample to an unqualified trajectory claim, not evidence that a particular archived runner used such a coupling. We therefore describe detachment as removing a direct gradient path, not as proving identical extractors across experimental arms. No optimization algorithm or empirical result is changed.

C.4 MF-to-Main Distillation

This section documents the separately evaluated MF-to-main distillation variant. Its added training loss and measured main-classifier output remain distinct from default DFIL and from the nearest-center interface guarantees.

C.4.1 Training objective and existing evaluations

The distillation follow-up adds a temperature-scaled teacher-to-student loss to Equation (1):

$$\mathcal{L}_{\text{distill}} = \tau^2 \,\mathrm{KL}(\mathrm{softmax}(\boldsymbol{\mu}/\tau) \parallel \mathrm{softmax}(\hat{y}/\tau)) \,, \tag{26}$$

with $\tau = 2$ and loss weight 0.3. The MF teacher is detached in this term; the term updates the student and its backbone path, not the teacher directly. The KL is taken in the direction displayed. The original three-seed result is given in Table 6; Table 15 reports the larger follow-up experiments, which use a checkpoint criterion shared across arms and stronger teacher controls.

Table 6: Original three-seed distillation experiment on Qwen2.5-7B (%). These measurements use only three seeds, fewer than the larger follow-ups, and do not establish non-inferiority of the structural diagnostics.

Configuration	FRoG Acc.	CSMC Consistency	Pairwise Viol.
DFIL baseline	51.0 ± 1.3	92.3 ± 2.4	10.8
+ Distillation	54.4 ± 1.3	91.8 ± 0.8	9.9 ± 0.1
Δ	$+3.4$	-0.5	-0.9

In a separate pre-specified follow-up on Qwen2.5-7B with 20 paired seeds and the shared checkpoint criterion (Appendix D.2), distillation improves over default DFIL by 8.75pp (95% CI $[6.08, 11.70]$, $20/20$ favorable seeds); this experiment is distinct from the 30-seed arms in Table 15. Main-path violation rate changes by -0.35pp ($p = .672$) and CSMC consistency by -1.28pp ($p = .296$). These null comparisons do not establish equivalence or absence of harm. The stronger teacher controls in Appendix E.5 fail to isolate an ordering-specific accuracy benefit, and no distillation loss imposes the nearest-center decision rule on the main classifier. We therefore retain distillation as an optional accuracy-oriented variant rather than as a transfer of the structural certificate.

D EXPERIMENTAL SETUP

Quantifier protocols supply proportions; other ordinal tasks use the scalar encoding stated in their protocol.

Memberships enter cross-entropy as logits. Each protocol fixes class weights and g: the human-probability task leaves MF inputs attached, whereas few-shot and counting controls detach them; scalar MSE remains active. Component experiments use $\lambda_{\mathrm{DFIL}} = \lambda_p = .5$ and original scaling $\lambda_{\mathrm{DFIL}} = .1$, distinct regimes rather than one optimized configuration (Appendix D.1).

D.1 TRAINING DETAILS

Original experiment hyperparameters. The original FRoG recipe uses a 20-epoch budget, batch size 8 except identified early ablations, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), weight decay 10^{-2}, warmup–cosine scheduling, and gradient clipping at 1. LoRA uses $r = 16$, $\alpha = 32$, dropout 0.05, and query/value projections; cross-task projections are specified with their protocol. The original class-weight vector is $(6.59, 1.93, 0.39, 0.51, 1.01, 0.42, 1.03, 2.04)$, fixed in the training configuration. Scalar-loss weight is $\lambda_p = 0.5$; original scaling uses $\lambda_{\mathrm{DFIL}} = 0.1$ and component ablations use 0.5. Nominal Gaussian widths are $(0.03, 0.05, 0.08, 0.08, 0.10, 0.12, 0.10, 0.04)$, rather than a shared 0.1; uniform-initialization controls use their specified common widths. The single-seed weight sweep does not establish multi-seed robustness. The 7B scaling means use six seeds at batch size 8; the three-seed per-class diagnostics and the batch-64 results come from separate experiments.

Compute scope. The original experiments use approximately 200 GPU-hours, mainly on H200 GPUs, with indicative per-seed training times of 15, 25, 45, and 90 minutes for the reported 7B, 14B, 32B, and quantized 72B configurations. The human-target confirmation experiments run on a single consumer GPU. This figure excludes the follow-up controls and the external-data work, and is therefore not a compute total for the paper as a whole; we do not report such a total.

Dataset. We use the FRoG quantifier-prediction task (Li et al., 2024), distributed through Hugging Face. Easy is randomly divided 80/20 for training/validation, followed by Hard evaluation; this is not a source-problem-disjoint partition (Appendix D.4.1). The compositional test set is synthetically generated with 900 examples (600 two-step, 300 three-step compositions).

D.2 PROTOCOLS OF THE MATCHED COMPARISONS

The main-text follow-ups are separate experiments and are not pooled with the original results. The protocols below state how each of them was run; none introduces an additional benchmark.

Evaluation protocol and initialization. In the initial diagnostics, a percentage target was divided by 100 only when its magnitude exceeded 1, which misreads targets such as 1%. The corrected protocol interprets the percent sign directly, including values at or below 1%. The follow-ups also select checkpoints with an arm-independent objective computed over the full validation set. Where the decoding rule or checkpoint criterion behind an original result cannot be verified, that result is reported as a diagnostic rather than as verification of the exact nearest-center theorem. Numerical differences between the two protocols therefore cannot be attributed solely to a method component.

The eight nominal reference centers are author-specified, not estimates copied from QuRe. For comparison, QuRe Table 3 reports mean strengths 0.004 for *none*, 0.225 for *some*, 0.687 for *most*, and 0.885 for *all* under its $g = 0.01, w = 1$ configuration (Li et al., 2023). It also lists *tiny*, *small*, and *moderate*; equating them with FRoG's corresponding "amount" labels requires a stated vocabulary mapping. Our bank is not evidence of human grounding.

Human probability confirmation. We use a fixed version of the public human data of Belém et al. (2024). Targets are medians of at least three included human ratings for a statement–expression stimulus. Statement-disjoint partitions contain 70 training stimuli from 8 statements, 69 selection stimuli from 8 statements, a separate 8-statement development partition, and the 184 confirmation stimuli from the remaining 22 statements. The confirmation protocol follows earlier development on the visible partitions; no confirmation row is used for selecting checkpoints, mixture weights, seeds, or subsets.

Both arms use Qwen2.5-1.5B, five ordered bins, 20 new paired seeds (2001–2020), a 20-epoch budget, batch size 8, and maximum length 192. They share scalar supervision with $\lambda_p = 0.5$; DFIL adds MF cross-entropy with weight 0.5, while FT+Compose freezes the bank and sets that weight to zero. MF inputs are not detached. For class probabilities π^{main} and π^{MF}, the scored output is

$$\hat{r}_\alpha = \sum_{q=0}^{4} v_q \big[\alpha \pi_q^{\text{main}} + (1-\alpha)\pi_q^{\text{MF}}\big], \qquad v_q = q/4.$$

Here $\pi^{\text{MF}} = \text{softmax}(\boldsymbol{\mu})$; this is not nearest-center inference. Each arm chooses $\alpha \in \{0, 0.1, \ldots, 1\}$ by selection-set MSE, breaking ties toward smaller α. Both arms receive the same selection opportunity.

The primary inference unit is the paired training seed on this fixed confirmation set. The frozen success criteria require a negative mean MSE difference, a Student-t 95% interval below zero, exact two-sided sign-flip $p < .05$, at least $15/20$ favorable seeds, a negative mean MAE difference, and both arms beating the training-mean constant. All hold. The exact test enumerates 2^{20} signed assignments (1,884 as or more extreme, $p = .0018$). The effect concerns the selected mixture; a main-only advantage is not established.

Independent counting (ExactCount). Gold is the exact enumerated intersection proportion. Three fixed test sets each contain 1,200 worlds in 600 matched-marginal pairs (1,800 pairs in total). The 20 seed-level summaries average over these sets. Calibration uses a separate split. Both arms estimate the two scoped conditional-chain factors, and their calibrated estimates are multiplied outside the model. They share scalar supervision, uniform initialization, and a checkpoint objective excluding MF cross-entropy. Detachment removes its direct extractor gradient (Appendix C.3). This limits the treatment contrast. The primary marginal-pull contrast is null ($p = .873$), and both arms fail the marginal-invariance validity criterion; composed MAE is an additional arm comparison, not a substitute primary endpoint. Its paired difference is $+0.001422$, 95% CI $[-0.0011, +0.0039]$, exact $p = .289$. Repeating the same procedure reproduces identical scalar predictions; this is a consistency check, not an independent replicate.

Natural-249 uniform-initialization control. Both arms use uniform initialization and share the corrected evaluation protocol and checkpoint criterion over 20 paired seeds. The frozen primary is native nearest-center accuracy. Secondary common-bin accuracy differs by $+5.22$pp (95% CI $[2.19, 7.99]$, $p = .004$); composed MAE differs by -0.00428 ($[-0.01780, 0.01099]$, $p = .579$), and base-proportion MAE by $+0.01394$ ($[-0.00339, 0.03479]$, $p = .243$). Positive MAE differences are worse. No secondary replaces the null primary. The final dataset and its base targets were produced by a single Qwen model; agreement statistics from a separate 595-candidate pool judged by four language models do not establish consensus selection of these 249 items.

Accuracy, distillation, and teacher controls. The Qwen comparisons use 30 paired seeds and the Mistral comparison 20. The two additional Qwen teacher controls reuse the original FT and DFIL-distillation arms and add 30 runs for each control. Checkpoints minimize weighted main-classifier validation cross-entropy, excluding the arm-specific MF term; the aggregation uses complete-dataset sufficient statistics. The two teacher-control intervals are paired bootstraps, and their sign-flip tests are Monte Carlo estimates using 20,000 draws, not exact enumeration. The self-distillation teacher is a frozen same-seed FT checkpoint, adding a separate training stage. The unordered teacher uses independently sigmoid-bounded centers in $[0.02, 0.98]$, initialized to the ordered bank's realized values and widths. These controls test teacher source and ordering, with the stated compute asymmetry.

These three comparisons respectively contribute 30, 30, and 40 applicable dual-path checkpoints, 100 in total, each reporting zero nearest-center grid violations, with no missing applicable entries. The main-path violation metric treats predicted ties as non-violations; its rate is descriptive and does not establish a text-level guarantee.

32B dispersion. The six original pairs give $55.42 \pm 5.61\%$ for DFIL and $53.68 \pm 6.12\%$ for FT, with a paired gap of $+1.74$pp (95% CI $[-7.92, 11.41]$, $p = .66$, $3/6$ favorable seeds). A separate rerun under the shared checkpoint criterion gives $49.50 \pm 7.01\%$ and $47.22 \pm 8.46\%$ ($+2.28$pp, exact $p = .47$, $3/6$ favorable). The dispersion does not fall below the original SDs, so this rerun does not

support checkpoint selection as its explanation. Both result sets remain distinct; bimodality and a scale-specific causal mechanism are unestablished.

D.3 External Screening and Generalization Limits

The human-probability confirmation was selected after exploratory external-data work. We retain the development/confirmation distinction and report relevant adverse follow-ups rather than treating the confirmed dataset as representative of all ordinal tasks. On a deduplicated ten-level drug-review task (20 seeds), DFIL's ordinal MAE exceeds FT's by 0.0167 (95% CI $[0.0056, 0.0281]$, exact $p = .0093$). A binary monotonicity-inference candidate identified during development did not meet the stricter confirmation criteria when retested on new statement partitions with 100 seeds: MSE difference -0.00207, 95% CI $[-0.00452, 0.00026]$, Monte Carlo sign-flip $p = .092$. These tasks and outputs differ from the human-probability confirmation. They restrict generalization claims; they neither replace its held-out endpoint nor support universal predictive improvement. External human validation of complete nested quantifier expressions remains open.

D.4 Datasets and Evaluation Protocol

D.4.1 FRoG Dataset Statistics

Source overlap and protocol scope. FRoG-Easy and FRoG-Hard differ in distractor selection; masking conditions are a separate design axis (Li et al., 2024). In the quantifier-prediction data we use, both contain $2{,}044$ rows derived from the same $1{,}276$ unique source–original-question pairs. Every aligned row has the same underlying question without options, target percentage mention, and quantifier. There are $2{,}042$ distinct option-free questions shared by the two splits, but only 4 exact full-question strings in common before tokenization. Different option strings therefore do not imply independent underlying problems.

Both the original and the follow-up procedures split Easy randomly by row and evaluate on Hard; they do not separate source-question groups. The documented 20% validation split rounds up to 409 rows, leaving $1{,}635$ training rows. Table 7 reports the seed-42 split obtained from this row order; other seeds change class counts. The source class totals agree with the data used here.

These statistics describe the data and preprocessing we use; they do not certify the partition seen by every earlier training run, and the exact dataset revision and tokenized inputs behind some earlier results cannot be confirmed. Where follow-up summaries use different evaluation counts, we report those counts rather than assuming the ones above. All measurements are reported under the protocol that produced them, but we make no unseen-problem generalization claim from FRoG scaling, few-shot, component, or teacher comparisons. More seeds and a corrected checkpoint criterion do not remove source overlap. The statement-disjoint human confirmation uses a separate dataset and partitioning procedure.

Table 7: FRoG dataset statistics. Training/validation columns give the seed-42 row split; Hard shares its underlying problems. The final column retains illustrative proportion ranges, not empirical bounds on the dataset or universal quantifier meanings.

Quantifier	**Train (42)**	**Val (42)**	**Hard**	**Illustrative range**
none	23	8	31	$[0.00, 0.00]$
tiny amount	83	23	106	$[0.01, 0.10]$
few	414	104	518	$[0.05, 0.30]$
small amount	322	80	402	$[0.10, 0.35]$
some	164	38	202	$[0.25, 0.50]$
moderate amount	391	95	486	$[0.35, 0.65]$
most	154	45	199	$[0.60, 0.90]$
all	84	16	100	$[1.00, 1.00]$
Total	1,635	409	2,044	

The class distribution is highly imbalanced ("few" has 16× more examples than "none"), motivating our use of inverse-frequency weighted cross-entropy loss. The illustrative ranges overlap between adjacent labels (e.g., "few" and "small amount" share $[0.10, 0.30]$); these intervals were not estimated as empirical class supports and are not used as certificate inputs.

D.4.2 QuRe Adaptation

We evaluate FRoG-trained checkpoints without retraining on a 401-item adaptation of QuRe (Li et al., 2023), whose original collection has 744 sentences. Four gold classes are represented; the mapped output space has five labels, as specified below. On Qwen-7B, the MF branch reaches $31.3 \pm 2.7\%$, compared with FT at $16.6 \pm 4.4\%$ and DFIL main at $14.9 \pm 2.5\%$. On Mistral-7B, DFIL main reaches $19.0 \pm 2.3\%$, compared with FT at $16.5 \pm 0.9\%$ and MF at $14.2 \pm 5.4\%$ (Table 23). The contrasts describe relative behavior under this adaptation; choosing the better test path retrospectively would require a separate selection evaluation. The reference initialization is author-specified, not estimated from QuRe annotations.

The original QuRe collection contains 744 Wikipedia-derived sentences (Li et al., 2023). Our evaluation uses 401 items after filtering the distributed annotations to *no*, *few*, *some*, *most*, and *all*. The evaluated items carry 76 *few*, 196 *some*, 115 *most*, and 14 *all* gold labels, with no *no* gold items. Thus four represented classes does not imply a four-label decoder. Predictions map *none* to *no*; *tiny amount*, *few*, and *small amount* to *few*; *some* to itself; *moderate amount* and *most* to *most*; and *all* to itself.

Each evaluation input prepends a masked quantifier clause and the available percentage annotation to the original sentence, falling back to the quantifier sentence when needed; prediction then proceeds through the trained head and the mapping above. This input adaptation and label filtering differ from the original QuRe evaluation; Table 23 reports transfer under this protocol. The item and gold-class counts above are established, but we cannot identify a single immutable source revision behind every evaluated input. Both prediction paths are reported separately, without selecting a winner on the transfer test set.

Metric and constant reference. We report ordinary item accuracy, $N^{-1} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}$, without class weighting. Always predicting *some* would give $196/401 = 48.88\%$, exceeding every reported model mean (maximum 31.3%). This is a descriptive reference computed from the evaluation-label distribution, not a validation-selected or prospective baseline. It leaves the DFIL–FT contrasts intact but does not support practically strong transfer performance. We therefore treat QuRe as an adaptation diagnostic, with its percentage-augmented inputs and class imbalance.

D.4.3 Evaluation Metrics

Accuracy. We report classification accuracy on the hard split. A uniform random predictor has expected accuracy 12.5% on eight classes. A predictor sampling labels from the reported test-class frequencies has expected accuracy $\sum_q (n_q/N)^2 \approx 18.4\%$; this is a descriptive prior baseline, not a training-only estimate.

Per-class accuracy. Accuracy computed separately for each quantifier class, revealing whether a method systematically fails on specific classes.

Monotonicity violation rate. The fraction of ordered test pairs (x_i, x_j) with $p_i < p_j$ where the predicted class index for x_i exceeds that for x_j (see Appendix E.8).

D.4.4 Input Formatting

Preprocessing tokenizes each full question directly, including its option text, with padding and truncation to at most 512 tokens. The prediction head uses the fixed eight-label vocabulary rather than the item-specific option letters. The illustration below lists that vocabulary; it is not the literal template applied to every FRoG example:

```
Choose the most appropriate quantifier for the
blank in the following sentence:
```

```
"___ of the 20 students passed the exam,
 specifically 15 out of 20."

Options: (A) none (B) tiny amount (C) few
(D) small amount (E) some (F) moderate amount
(G) most (H) all

Answer:
```

The prediction head is trained on the corresponding class index. The main path decodes classifier logits; the scalar path uses the specified numerical head and decoder. This head-based evaluation should not be confused with free-form generation of an option letter.

D.5 Hyperparameter Sensitivity: λ_{DFIL}

We report an exploratory sweep over $\lambda_{\text{DFIL}} \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$ on Qwen2.5-7B (seed 42, $\lambda_p = 0.5$ fixed). Table 8 gives FRoG-Hard test accuracy; it is not a validation-based hyperparameter selection.

Table 8: λ_{DFIL} single-seed sensitivity probe on FRoG-Hard (Qwen2.5-7B, seed 42); reports the magnitude of variation across λ values, not a multi-seed sensitivity analysis.

λ_{DFIL}	0.1	0.5 (default)	1.0	2.0	5.0
Accuracy (%)	50.4	**51.2**	49.9	51.2	50.4

The reported main-path accuracy ranges from 49.9% to 51.2%, a spread of 1.3pp over a 50-fold weight range. The values 0.5 and 2.0 tie in this seed. These observations do not establish robustness across seeds, an optimal default, or a causal balance between the loss terms.

E Details of the Main Experimental Evidence

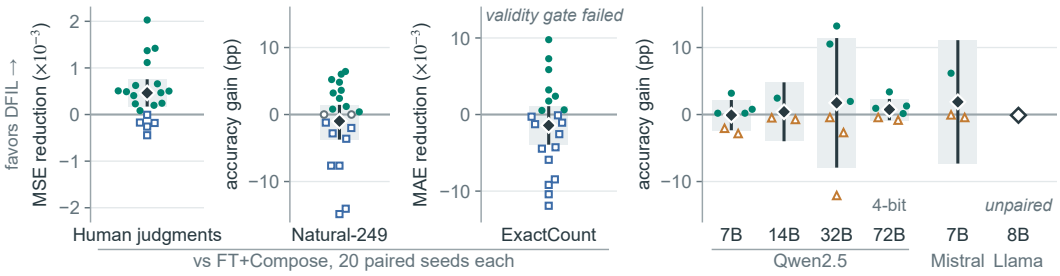


Figure 5: Only the human-judgment interval excludes zero. Differences favor DFIL upward. Left three panels: FT+Compose controls, 20 paired seeds; both ExactCount arms fail the null-control requirement. Right: FRoG-Hard, original protocol, 3–6 pairs per backbone; Llama-3.1-8B is unpaired and mean-only. Filled circles favor DFIL, open squares FT+Compose, open triangles FT, grey circles tie; diamonds, bars and shading show means and 95% intervals (paired bootstrap for Natural-249 and ExactCount, Student-t otherwise). Exact values: Table 9 and Appendix E.6.1.

FRoG (Li et al., 2024) supplies eight-class single-step training and Hard evaluation, with $2{,}044$ examples in each source split. Easy and Hard share base problems; these are option-variation diagnostics rather than problem-disjoint tests (Appendix D.4.1). The original experiments span Qwen2.5, Mistral, Llama, and Phi backbones (Yang et al., 2024; Jiang et al., 2023; Grattafiori et al., 2024; Abdin et al., 2024); their per-backbone breakdowns appear in the appendix. We additionally report paired controls that use a corrected interpretation of percentage targets and a checkpoint criterion

Table 9: Learned-treatment utility and composition controls. DFIL minus matched FT+Compose; 20 paired seeds per row on fixed data. The human-judgment row uses a Student-t interval; the other intervals are paired bootstraps. It evaluates a selected probability mixture. Natural-249 evaluates native nearest-center decisions against model-annotated, product-derived targets. ExactCount's additional MAE contrast is inconclusive because its validity gate fails.

Endpoint	DFIL	FT+Compose	Difference [95% CI]
Human-judgment mixture MSE	.007690	.008154	$-.000464$ $[-.000756, -.000172]$
Natural-249 accuracy (%)	31.31	32.29	$-.98$ $[-3.73, +1.45]$
ExactCount composed MAE	.0357	.0343	$+.0014$ $[-.0011, +.0039]$

shared across arms. These follow-ups are separate experiments: because both the checkpoint criterion and the target interpretation differ, differences from the original results cannot be read as single-factor effects. Appendix D.2 states their protocols. Unless otherwise stated, uncertainty intervals compare training seeds on fixed evaluation data; they do not measure variation across new datasets.

Section 6 reports direct verification, calibrated reuse, matched human and composition comparisons, and the core dual-path contrast. This appendix retains the complete Human-Grid bank sensitivity, component variants, distillation and stronger teacher controls, with additional endpoints and protocol qualifications. Results from distinct training and selection procedures are not pooled.

E.1 Human Probability Judgments

We use the public human-judgment data of Belém et al. (2024), which asks how uncertainty expressions modify the perceived probability of natural-language statements. Human-median targets are independent of DFIL's centers and composition operator. The confirmation set contains all 184 eligible stimuli from 22 statements held out from training, validation, and earlier development, spanning five ordered bins. The protocol was fixed before constructing or scoring this set, following development on other statements; it is not a study-wide blind choice of task.

We compare DFIL with matched FT+Compose on Qwen2.5-1.5B using 20 new paired seeds. Both arms receive identical human scalar targets, initialization, training budget, and selection rules. The primary output is a mixture of the main and MF class probabilities, converted to a scalar using fixed class representatives. Each arm selects its mixture weight on the validation split only. Table 9 reports the pre-specified primary endpoint, and Figure 5 shows the seed-level differences for all three controls and FRoG-Hard.

The paired MSE difference is $-.000464$ with 95% interval $[-.000756, -.000172]$. The exact sign-flip $p = .0018$ and $15/20$ favorable seeds accompany the absolute MSE values in Table 9. The relative MSE reduction is 5.69%; the secondary mean MAE difference is -0.001612. This supports utility on this external human-target task. It does not test nested composition, and the selected mixture does not inherit the nearest-center theorem. Other external screens included negative or unreplicated findings (Appendix D.3); the result does not establish cross-task, cross-scale, or fuzzy-specific superiority.

E.2 Composition Controls

Operator-consistency diagnostics. The original 900-example composition set uses the same proportional-shift family as inference. Reference-bank targets use fixed author-specified centers; self-consistent targets use each model's own centers. The latter gives a reported $+5.9$pp DFIL–FT+Compose gap (20 seeds), increasing to $+23.2$pp under uniform initialization (3 seeds; Tables 27 and 28). These are agreement measurements against protocol-defined targets. Learning the centers also changes self-consistent targets, so uniform initialization does not remove label-generation coupling or isolate a causal advantage. The 100% self-consistent oracle score follows from identical target and prediction operations.

Independent counting labels. ExactCount constructs matched pairs of finite worlds with fixed marginal counts and changing intersection counts. Gold is computed by enumeration, without learned centers or proportional-shift labels. Models trained on single-edge counting estimate $P(A)$ and $P(B \mid A)$ separately; the two estimates are then multiplied outside the model by a fixed rule. Across 20 paired seeds and three fixed sets of 600 pairs each, both arms recover conditional-edge direction at approximately 99%. However, the marginal null-control rates are 0.618 and 0.619, outside the pre-specified 0.5 ± 0.10 range, so the full validity criterion fails. The comparison uses detached MF inputs and hence does not test a fuzzy-loss effect on scalar extraction. Its MAE contrast shows no detected DFIL advantage (Table 9).

Natural prose with matched initialization. Natural-249 contains mined sentences from GenericsKB, PubMedQA, and CNN/DailyMail (Bhakthavatsalam et al., 2020; Jin et al., 2019; See et al., 2017). The final item selection and the proportion annotations come from a single Qwen model; $191/249$ items additionally meet the stricter sensitivity rule requiring agreement of three of four language-model judges. All 249 items are kept fixed. Its product-derived labels are not independent human composition judgments. The 20-seed matched-uniform follow-up on Qwen2.5-7B gives 31.31% accuracy for DFIL and 32.29% for FT+Compose. The primary interval crosses zero (exact sign-flip $p = .492$); ten seeds favor DFIL, eight favor FT+Compose, and two tie, with all seeds retained. A secondary common-bin metric favors DFIL by 5.22pp, but neither composed nor base-proportion MAE shows a detected improvement (Appendix D.2). This metric-dependent outcome does not establish overall superiority or explain the original 3.6pp FT+Compose lead (Table 10) as an initialization effect.

Template diagnostic. Table 11 reports an original three-seed diagnostic on Qwen2.5-7B with 320 author-constructed template examples from eight domains, separate from Natural-249 and scored against reference-bank targets. It compares DFIL's composed prediction with direct text-only fine-tuning, not with the matched FT+Compose control, and therefore does not measure the added benefit of learning the bank.

Table 10: Original Natural-249 results (mean $\pm$ SD over 3 seeds). Natural sentences were mined from GenericsKB, PubMedQA, and CNN/DailyMail. The set was constructed with Qwen-based filtering and Qwen base-proportion annotations rather than by consensus of four language-model judges. A post hoc check finds $191/249$ items satisfy a strict criterion requiring three of these four judges to agree. Targets apply $c_{Q_1} p_{\text{base}}$ and fixed author-specified centers, so these are model-annotated, operator-derived labels. The corrected matched-uniform follow-up appears in Table 9.

Stratum	n	**DFIL from-text**	**Text-only**	Δ
Headline (mean over Q_1/sources, 3 seeds)				
DFIL (dual-path)	249	32.8 ± 3.7	12.6 ± 0.6	$+20.2$
FT+Compose	249	$\mathbf{36.4 \pm 1.9}$	12.9 ± 0.4	$\mathbf{+23.6}$
Per outer quantifier Q_1 (FT+Compose, mean over seeds)				
`most`	80	28.7	9.6	$+19.2$
`some`	80	31.2	14.2	$+17.1$
`few`	64	**56.8**	17.7	$\mathbf{+39.1}$
`all`	23	18.8	7.2	$+11.6$
`none`*	2	100	0	$+100$
Per source corpus (FT+Compose, mean over seeds)				
GenericsKB-Best	106	43.4	17.3	$+26.1$
PubMedQA	67	37.8	14.9	$+22.9$
CNN/DailyMail	76	25.4	4.8	$+20.6$

* `none` bucket has n=2, statistical noise; reported for completeness.

E.3 Human-Elicited Quantifiers and Exact Matrix Targets

The Human-Grid analysis reuses Experiment 1 of Macuch Silva et al. (2024), in which participants selected nested quantifiers to describe school-exam outcomes. Its complete-case set contains 1,192 responses from 60 participants across 20 matrices of 5 students by 12 questions. The quantifiers

Table 11: Naturalistic compositional reasoning accuracy (%). 320 examples from 8 domain templates (clinical trials, education, manufacturing, elections, environmental science, corporate governance, sports, urban planning), with reference-bank target labels (composed proportion bucketed via fixed reference centers, independent of either model's learned MF). Mean ± sample std over 3 seeds (Qwen2.5-7B).

Method	**DFIL-from-text**	**Text-only**
DFIL	$\mathbf{39.5 \pm 3.3}$	12.8 ± 2.4
Fine-tune	—	11.1 ± 0.9
Δ (DFIL-from-text − Fine-tune text-only)	$+28.3$	

Table 12: The representative bank changes which operator fits human-elicited expressions. MAE against counted matrix density, all 1,192 complete Human-Grid responses; bold denotes a row minimum, not significance. This exploratory analysis runs no text model.

Center bank	Product	Repeat inner	Repeat outer	Minimum	Łuk.
Reference initialization	**.1464**	.2688	.2162	.1673	.1754
PRESQUE, $g = .01, w = 1$	.2243	.2283	.2188	**.1917**	.2974
PRESQUE, $g = .01, w = 4$	.2237	.2308	.2217	**.1936**	.2923
Uniform initialization	.1669	.3050	.2565	.2106	**.1548**
Learned, seed 42	**.1630**	.2835	.2341	.1813	.1878
Learned, seed 123	**.1559**	.2671	.2192	.1690	.1884
Learned, seed 456	**.1577**	.2579	.2107	.1631	.2044

are human choices; the target is the fraction of matrix cells satisfying the selected predicate ("right" or "wrong"), obtained by counting. Thus neither the target nor the original quantifier choices are generated by DFIL's proportional-shift rule. Matrix density is an aggregate target, however, not a unique truth-conditional interpretation of the chosen sentence.

For each fixed center bank, the chosen outer and inner quantifiers yield scalars a and b. We compare product ab, repeat-inner b, repeat-outer a, minimum $\min(a, b)$, and Łukasiewicz $\max(0, a + b - 1)$. The sensitivity analysis uses all seven center banks: the author-specified reference initialization, two published PRESQUE strength estimates from Table 3 of Li et al. (2023), the original uniform initialization, and three learned banks. All bank comparisons are post-result descriptive analyses. No text model is run in this operator comparison.

Table 12 reports every bank and operator from this analysis. The participant-level intervals and the positive-scope sensitivity results below provide additional detail; they do not replace the complete-response comparison or turn the exploratory analysis into a confirmatory model evaluation.

Under the reference initialization, the participant-mean product-minus-repeat-inner MAE difference is $-.1223$, with descriptive participant-bootstrap 95% interval $[-.1414, -.1025]$. Product also has lower MAE than repeat-outer, minimum, and Łukasiewicz for that bank. Changing the grounding changes the result: minimum has the lowest MAE under both published PRESQUE banks, whose product-minus-repeat-inner intervals are $[-.0197, .0121]$ and $[-.0229, .0088]$; Łukasiewicz is best under uniform initialization. There the product-minus-repeat-inner interval, $[-.1548, -.1207]$, also excludes zero while Łukasiewicz still has the lower MAE: repeat-inner ignores the outer quantifier and is a degenerate control, so an interval excluding zero does not make product the best operator for that bank. The 1,116-response positive-scope subset gives the same winning operator for each bank. This supports compatibility of the submitted center-plus-product configuration with an external elicitation task, while exposing dependence on the centers. It does not establish center-invariant product superiority, a DFIL–FT+Compose text-model advantage, or replication of the original $+5.9$pp difference.

Scope of this analysis. The source study makes its materials and data available at `https://osf.io/g6bep/`. Our analysis uses its complete-case responses together with the seven center banks listed above. It was carried out after an earlier result had been examined, and is therefore exploratory rather than confirmatory. The intervals resample participant means with 100,000 bootstrap draws; the table reports response-mean absolute errors. The analysis uses fixed data and center-bank versions. All 70 bank–operator–subset means are recomputed from these inputs; the analysis involves no new annotations, model inference, or training.

E.4 Component Ablation

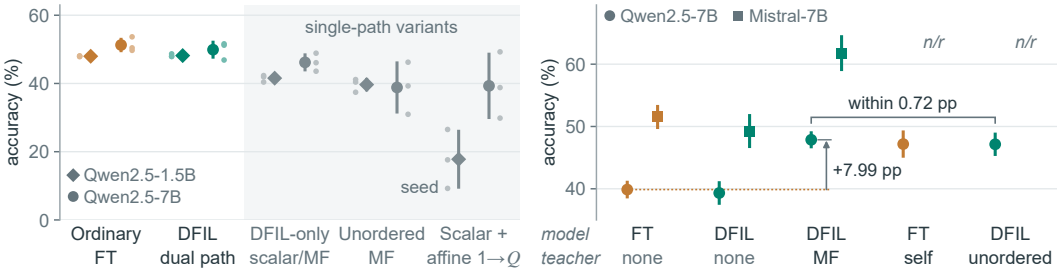


Figure 6: Why keep both paths, and what distillation adds. FRoG-Hard. Left: original protocol, $\lambda_{\mathrm{DFIL}} = .5$, mean $\pm$ 1 sample SD over three seeds (dots); grey variants are single-path. Right: corrected protocol, main-classifier accuracy, 30 Qwen/20 Mistral seeds, unpaired 95% t-intervals, axis from 35; n/r, not run. Exact values: Tables 13 and 15.

Table 13 reports the original five-variant component comparison at two scales, plotted in the left panel of Figure 6.

Table 13: The full classifier and the reusable scalar serve different roles. Existing FRoG component study, accuracy (%), mean $\pm$ sample SD over three seeds; $\lambda_{\mathrm{DFIL}} = .5$. This is the original shared-problem protocol, not the corrected follow-up or the $\lambda_{\mathrm{DFIL}} = .1$ scaling recipe. All five original variants are shown; Table 14 distinguishes reused measurements and the batch-8 single-path follow-up. The unordered and affine variants are single-path systems; differences to dual-path DFIL do not isolate ordering or membership learning. Scalar-path accuracy uses the evaluated Gaussian rule, not a new nearest-center measurement.

Training/inference configuration	Qwen2.5-1.5B	Qwen2.5-7B
Fine-tuning, full classifier	$48.0 \pm .1$	51.3 ± 2.1
DFIL-only, scalar/MF path	41.6 ± 1.0	46.2 ± 2.6
Unordered MF, single path	39.7 ± 2.0	38.8 ± 7.6
Scalar with affine $1 \to Q$ decoder	17.8 ± 8.6	39.3 ± 9.7
DFIL, dual path	$48.2 \pm .4$	49.9 ± 2.6

The additional table below identifies the batch-8 scalar-only measurement separately from the batch-64 scalar-only entry in that main comparison. Reused rows are not independent replications or a controlled batch-size study.

Implemented ablations. In the single-path DFIL variant, training is $\mathrm{CE}(\boldsymbol{\mu}, y) + \lambda_p \mathrm{MSE}(\hat{p}, p)$: setting the dual-path auxiliary coefficient to zero does not remove its primary MF cross-entropy. The scalar linear-classifier variant retains the numerical head and replaces the Gaussian bank with an affine $1 \to Q$ classifier. It therefore differs from the full-hidden-state fine-tuning MLP. The no-ordering variant replaces ordered centers by free parameters in the single-path model; it is not a matched removal of ordering from the dual-path system. These comparisons describe different pathways and losses. They do not isolate ordering as the cause of an accuracy gap, show that MF supervision is absent, or establish that a classifier without MF cannot train successfully.

Table 14: Extended ablation on FRoG-Hard (Qwen2.5-1.5B), mean ± sample SD over three seeds. The DFIL-only row is the batch-8 measurement (39.2%); Table 13 retains the batch-64 value (41.6%). Several other rows repeat the same measurements across the two settings. They are not independent replications or a controlled batch-size experiment. Parameter counts include LoRA and all trained heads.

Variant	Accuracy (%)	Trainable params	Description
DFIL (full)	48.2 ± 0.4	2,974,489	Full dual-path model
Fine-tune only	48.0 ± 0.1	2,577,672	Standard classifier, no DFIL
DFIL-only	39.2 ± 1.3	2,575,889	Single scalar/MF path; CE + MSE
No MF branch	17.8 ± 8.6	2,575,889	Scalar head + affine $1 \to Q$ decoder
No ordering	39.7 ± 2.0	2,575,889	MF without cumulative softmax

Component-by-component interpretation. Full classifier versus the MF-only path. The observed dual-path gains are 6.6pp at 1.5B and 3.7pp at 7B in Table 13. A scalar bottleneck restricts the information passed to the output; retaining the classifier lets the system also use the full representation. These data support the design in this setting without proving that every task needs both paths.

Ordered versus unordered memberships. The original unordered variant obtains $39.7 \pm 2.0\%$ and $38.8 \pm 7.6\%$. The 8.5/11.1pp contrasts to dual-path DFIL change the output architecture as well as the bank constraint, so they are not isolated ordering effects. The ordering theorem remains exact for nearest-center decoding regardless of that empirical attribution. Table 15 supplies a stronger matched unordered-teacher comparison for the distillation extension.

Membership mapping versus a linear scalar classifier. The "No MF" variant keeps the numerical head and replaces the membership bank by a learned $1 \to Q$ affine classifier. Its $17.8 \pm 8.6\%$ and $39.3 \pm 9.7\%$ means describe that one-dimensional classifier, not removal of proportion supervision or the full scalar branch. A linear head can implement nearest-center decisions (Proposition A.13); these training outcomes do not establish an expressive impossibility.

Auxiliary weight and retained accuracy. At 7B the component study with $\lambda_{\mathrm{DFIL}} = .5$ gives 49.9% for dual-path DFIL and 51.3% for FT; the scaling configuration with $\lambda_{\mathrm{DFIL}} = .1$ gives 50.4% and 50.5%. These are separate runs rather than a controlled estimate of the weight's effect. Proposition C.2 and Appendix C.1 explain the possible loss interactions; small-seed accuracy differences do not establish equivalence or an optimal coefficient.

E.5 Single-Step Accuracy and Teacher Controls

Table 15: MF-to-main teaching improves this accuracy endpoint; the Qwen controls do not isolate an ordering-specific gain. Existing corrected-protocol FRoG-Hard results, main-classifier accuracy (%). All conditions use the shared main-loss checkpoint criterion. Qwen uses 30 paired seeds; Mistral uses 20. Self-distillation includes an additional teacher-training stage. Dashes are unmeasured controls, not zero gains. Source-problem overlap limits this to option-variation adaptation.

Training condition	Qwen2.5-7B	Mistral-7B
Ordinary fine-tuning	39.87	51.52
DFIL, no distillation	39.32	49.27
DFIL + MF distillation	47.86	61.77
Fine-tuning + self-distillation	47.18	—
DFIL + unordered teacher	47.14	—

The original cross-backbone results have similar observed DFIL and fine-tuning means (Table 16); small seed counts do not establish equivalence. A follow-up that selects checkpoints on the shared main loss gives $39.32 \pm 5.07\%$ for default DFIL and $39.87 \pm 3.80\%$ for ordinary fine-

tuning on Qwen2.5-7B (30 paired seeds). The paired difference is -0.55pp with a 90% CI of $[-2.42, +1.31]$pp and a 95% CI of $[-2.80, +1.69]$pp; the pre-specified $\pm$2pp equivalence test fails (TOST $p = .099$). These means are obtained under the corrected evaluation protocol; they do not replace the estimates obtained under the original protocol.

MF-to-main distillation. The extension in Appendix C.4.1 uses the MF branch as a training-time teacher; the measured inference output is the main classifier. It improves over ordinary fine-tuning on Qwen2.5-7B by 7.99pp (95% CI $[6.07, 9.87]$, $29/30$ seeds) and on Mistral-7B by 10.25pp ($[6.93, 13.29]$, $17/20$). To examine attribution, we add a frozen same-seed FT teacher and a trainable Gaussian teacher with independently bounded, unordered centers. The ordered and unordered banks start from the same realized centers and widths. Both stronger controls obtain similar Qwen means (Table 15).

The Qwen paired gaps are $+0.69$pp versus self-distillation (95% CI $[-1.35, +2.97]$, Monte Carlo sign-flip $p = .552$) and $+0.72$pp versus the unordered teacher ($[-1.18, +2.68]$, $p = .470$). Per-seed differences in the two contrasts range from -10.9 to $+20.1$pp; $18/30$ seeds favor self-distillation over DFIL with MF distillation, while $19/30$ favor DFIL with MF distillation over the unordered teacher. Thus the gain over ordinary FT does not isolate a benefit from ordering. The self-distillation control uses a separately trained teacher and more total training compute; the unordered-bank control retains the single-stage arrangement. Neither result transfers the MF-path guarantee to the main classifier.

E.6 Accuracy Across Model Scales

E.6.1 Full Single-Step Scaling Table

The complete original scaling table below retains all six backbone configurations and every alternative baseline. These measurements describe the shared-source-problem adaptation protocol and are separate from the corrected-selector teacher experiments in Table 15. A note to Table 16 gives their paired seed-level contrasts where pairing is available.

Interpretation of the original comparisons. Table 16 retains the original per-backbone and alternative-baseline measurements. CORAL and CORN encode ordinal output structure, but neither their text-level ordering nor DFIL's is determined solely by a scalar-decoder theorem. Zero-shot quantifier scoring and trained heads use different supervision and inference regimes. Similar observed means across a few seeds do not establish equivalent accuracy or lower variance. The 32B dispersion and the follow-up under a checkpoint criterion shared across arms are reported separately in Appendix D.2; the cause of the spread is unresolved.

Comparison by baseline and backbone. Fine-tuning and FT+Compose. The original six-backbone comparison shows that adding DFIL is compatible with similar observed full-data means in these runs, while exposing an additional scalar output. FT+Compose reaches $50.2 \pm 1.9\%$ on Qwen-7B and shares the product interface and nearest-center guarantee; it is therefore the closest control for the incremental learned-bank contribution.

Ordinal heads. On Qwen-7B, CORAL reaches $43.2 \pm 2.6\%$ and CORN $52.1 \pm 1.8\%$, compared with DFIL's $50.4 \pm 1.8\%$. CORN's higher mean prevents claiming that DFIL uniformly outperforms ordinal alternatives. CORAL at 14B reaches $51.7 \pm 2.4\%$; its per-class results in Table 25 retain the original analysis of extreme and intermediate categories.

Quantifier scoring without fine-tuning. The shared-backbone, length-normalized conditional token scores give 24.4% on Qwen-7B and 23.9% on Mistral-7B, versus trained DFIL at 50.4% and 64.1%. This baseline does not implement the NLI and pragmatic-listener inference of PRESQUE (Li et al., 2023). The comparison records different supervision and inference regimes and cannot identify a fuzzy-head effect independently of training. Related compositional stress tests (Hosseini et al., 2024; Mirzadeh et al., 2025) motivate evaluating nested operations explicitly; they do not show that an unconstrained classifier cannot learn those operations.

Scale and family variation. Qwen DFIL means are 50.4, 57.3, 55.4, and 77.2% at 7B, 14B, 32B, and 72B. This observed sequence is not monotone, and the 72B experiment also changes to a quantized backbone. Mistral and Llama differences are $+1.9$ and $-.1$pp. At 32B the paired six-seed

mean difference is $+1.74$pp with a 95% CI of $[-7.92, 11.41]$pp and $3/6$ positive pairs; those records establish dispersion, not a bimodal mechanism or an ordering-specific instability.

Table 16: Original FRoG-Hard scaling measurements (mean $\pm$ SD); DFIL and Fine-tune entries are accuracies in %. These results come from the original evaluation pipeline and are distinct from the follow-ups in Table 15, which use a checkpoint criterion shared across arms. MF diagnostic rates are descriptive and must not be read as exact nearest-center certificates.

Backbone	**DFIL**	**Fine-tune**	Δ^*	**MF mono.**‡
Qwen2.5-7B	50.4 ± 1.8	50.5 ± 1.8	-0.1	0.0%
Qwen2.5-14B	57.3 ± 1.7	56.9 ± 2.8	+0.4	0.1%
Qwen2.5-32B¶	55.4 ± 5.6	53.7 ± 6.1	+1.7	0.1%
Qwen2.5-72B†	$\mathbf{77.2 \pm 1.5}$	76.5 ± 1.6	$+0.7$	0.1%
Additional backbones:				
Mistral-7B-v0.3	64.1 ± 0.9	62.2 ± 4.5	$+1.9$	—
Llama-3.1-8B	61.5 ± 4.3	61.6 ± 1.1	-0.1	—
Alternative baselines (Qwen2.5-7B):				
FT+Compose (frozen MF)§	—	50.2 ± 1.9	—	0.0%
CORAL (Cao et al., 2020)	—	43.2 ± 2.6	—	—
CORN (Shi et al., 2023)	—	52.1 ± 1.8	—	—
Quantifier scoring (zero-shot, Qwen-7B)$^\#$	—	24.4	—	—
Quantifier scoring (zero-shot, Mistral-7B)$^\#$	—	23.9	—	—
CORAL on Qwen2.5-14B	—	51.7 ± 2.4	—	—

† 72B uses QLoRA 4-bit. ‡ MF diagnostic from the original evaluation pipeline; scalar-grid and text-level ordering are distinguished in Section 6.4.

§ Classifier + numerical head + frozen MF; trained with $\lambda_{\text{DFIL}} = 0$ (no fuzzy loss).

* Paired per-seed DFIL$-$FT difference, pp, mean [95% paired-t CI] (seed pairs): Qwen2.5-7B -0.09 $[-2.34, +2.16]$ (6); 14B $+0.41$ $[-3.99, +4.80]$ (3); 32B $+1.74$ $[-7.92, +11.41]$ (6); 72B 4-bit $+0.74$ $[-0.85, +2.33]$ (6); Mistral-7B $+1.89$ $[-7.31, +11.10]$ (3). Every interval contains zero. Llama-3.1-8B (-0.11) is unpaired because its arms come from separate runs ($3 + 3$), so it has no interval.

¶ Six 32B seeds show large dispersion in both arms. The paired gap is $+1.74$pp, 95% CI $[-7.92, +11.41]$, with DFIL higher in $3/6$ seeds. Neither bimodality nor a causal explanation is established.

$^\#$ Length-normalized conditional log-probability of quantifier tokens on the shared backbone; no training or NLI/RSA inference. This is not a reproduction of PRESQUE (Li et al., 2023).

E.6.2 Observed Differences Across Model Scales

The sign and dispersion of the original DFIL–fine-tuning contrasts vary across configurations. The paired intervals and the separate 32B follow-up are reported in Table 16 and Appendix D.2; these observations do not establish equivalence or identify a capacity-based cause.

Observations:

- **Dispersion across seeds.** The reported standard-deviation ratios are close to one at 7B and 72B and lower at 14B and 32B. They are empirical dispersion summaries, not variance ratios or consequences of a head-level complexity bound. No causal explanation for the scale dependence is established here.
- **Statistical scope.** The original small-seed Welch comparisons do not detect per-backbone accuracy differences. This is not evidence of equivalence. The 30-pair equivalence test in Appendix E.5 also fails its pre-specified $\pm$2pp criterion, and it concerns a separate experiment under the corrected evaluation protocol.

Table 17: Certificate sweeps, 3,024 cases per depth. "Initial implementation" gives the counts of the first sweep, computed before this rounded-input reanalysis. "Printed-input" uses exactly the displayed center values and the fully specified rational sweep in Appendix B.2.1. The source of their numerical difference is unresolved. No row is a calibrated-coverage estimate.

	Initial implementation			Printed-input reanalysis		
m	Exact	Separate	Extra	Exact	Separate	Extra
1	2224	2027	197	2299	2118	181
2	2320	2316	4	2526	2506	20
4	2549	2549	0	2622	2622	0
8	2709	2709	0	2773	2773	0
16	2829	2829	0	2850	2850	0

- **Scope of compositional evidence.** The reported compositional experiments assess the configurations specified in Appendix F.4. Theorems A.3 and A.5 do not imply that a compositional accuracy advantage persists or grows with model scale.

E.6.3 Comparison with Zero-Shot Performance

FRoG (Li et al., 2024) studies zero-shot quantifier reasoning, whereas Table 16 uses task-specific supervised training. The latter reports higher means at larger Qwen scales, but the regimes differ in supervision and, at 72B, precision. This comparison does not identify task alignment or head design as the cause of a change in scaling behavior.

E.7 Structural Checks and Scalar Availability

Exact-certificate verification. Appendix B.2.1 separates the initial implementation's checks from the reference implementation and printed-input stress sweep. Table 17 includes every tested depth and reports the unresolved discrepancy with the initial-implementation counts. The mathematical containment result is independent of these frequencies: exact propagation contains every feasible label and can be tighter than bounding product and boundary separately. Neither sweep provides calibrated semantic coverage.

Implementation and text-level ordering. Across the follow-up experiments, $30 + 30 + 40 = 100$ applicable dual-path checkpoints report zero nearest-center grid violations, with no missing grid entries. This finite check corroborates implementation of the rule; the universal scalar statement comes from Theorem A.3. In the corresponding text-level diagnostic, main-path violation rates are 9.71% for DFIL and 9.70% for FT. That metric credits prediction ties as non-violations and is descriptive. Joint training therefore supplies no observed main-path certificate.

Scalar targets and diagnostic paths. On the author-built implicit-proportion probe (IPR), DFIL's main classifier and FT each obtain 15.8% accuracy; injecting author-assigned scalars into a separate nearest-center MF path gives 73.3% (Appendix F.6). Because this comparison changes the input and output path, it does not isolate extraction error. CSMC is summarized in Section 6.4; sentiment, similarity, QuRe transfer and few-shot results appear in Appendix F, which also gives the full CSMC diagnostics. Gaussian-argmax scores carry no nearest-center guarantee. CSMC also includes six seeds that were retrained, preventing an isolated estimate of the fuzzy-loss effect.

Table 18: Quantifier-order diagnostic accuracy (%). MF-based uses learned center ordering ($c_{Q_i} \leq c_{Q_j}$); classifier-based uses argmax labels on sentence pairs. Mean $\pm$ sample std over 3 seeds, Qwen2.5-7B.

Method	MF-based	Classifier-based
DFIL	$\mathbf{100.0 \pm 0.0}$	56.0 ± 2.8
Fine-tune	—	52.6 ± 2.7

E.8 Monotonicity Analysis

E.8.1 Evaluation Protocol

We evaluate monotonicity violations on the FRoG-Hard test set ($N = 2044$ examples). For each ordered pair (x_i, x_j) where the ground-truth proportion $p_i < p_j$, we check whether the predicted class index satisfies $\hat{q}_i \leq \hat{q}_j$. We report two metrics:

- **Pairwise violation rate:** $\frac{|\{(i,j):p_i<p_j \wedge \hat{q}_i>\hat{q}_j\}|}{|\{(i,j):p_i<p_j\}|}$. Each unordered pair with unequal reference proportions contributes one orientation; equal-proportion pairs are excluded, so the denominator need not equal $\binom{N}{2}$.
- **Mean violation magnitude:** $\frac{1}{|\text{violations}|}\sum_{(i,j)\in\text{violations}}(\hat{q}_i - \hat{q}_j)$. When violations exist, this captures their severity; otherwise we report it as undefined: a violation by 1 class (e.g., predicting "few" instead of "some") is less severe than a violation by 4 classes.

E.8.2 Results Across Methods

Table 19: Monotonicity diagnostics, mean over three seeds. DFIL/FT text-level rows come from the batch-64 runs, whereas the scaling table uses batch 8. The MF row uses Gaussian argmax on 1,000 scalar grid points (999 adjacent comparisons); other rows use text-level classifier pairs.

Method / Path	Scale	Pairwise viol. (%)	Mean magnitude	Total pairs
DFIL (MF grid)	7B	0.03	—	999
DFIL (main classifier)	7B	10.8	1.69	1,959,395
Fine-tune (classifier)	7B	10.6	1.69	1,959,395
CORAL (classifier)	7B	8.2	2.03	1,959,395
CORAL (classifier)	14B	6.6	2.13	1,959,395
CORN (Shi et al., 2023) (classifier)	7B	9.0	1.62	1,959,395

Analysis. The mean 0.03% MF grid rate under the original evaluation protocol (zero, zero, and one reversal over 999 comparisons per seed) measures adjacent Gaussian-argmax reversals on a scalar grid and is not the nearest-center theorem's zero-violation statement. Text-level rates concern a different denominator and include extraction and classifier behavior. CORAL has lower reported pairwise rates but larger conditional violation magnitudes than DFIL/FT in these runs; the observations do not establish the cause or a scale-dependent improvement. A guaranteed ordering in the supplied scalar requires the nearest-center rule, and ordering across texts additionally requires appropriate extraction.

F Additional Diagnostics

On the 40-sentence implicit-proportion probe, both main classifiers reach 15.8%, whereas nearest-center decoding of author-assigned scalars reaches $73.3 \pm 1.4\%$ exact and 100% off-by-one accuracy (Table 31); the input and output paths both change, precluding an isolated extraction diagnosis. Across 21 seeds, CSMC consistency is 96.40% for the DFIL membership output versus 90.90% for the FT main output (Table 30); different outputs, six retrained DFIL seeds and Gaussian argmax make this a combined-protocol comparison without a nearest-center guarantee. Few-shot, cross-task, QuRe, CSMC and IPR diagnostics, with their decoder, selection, retraining and weak-reference limitations, are in Appendices D.4.2 and F.1–F.6; with the validation-selected ensemble, few-shot FRoG gains over FT are $+16.4$, $+11.4$ and $+3.7$pp at $N = 16, 32, 64$ but -0.8 and $+0.8$pp at $N = 128, 256$, and SST-5 gains stay between $+2.4$ and $+3.2$pp; these are three-seed diagnostics on a shared-problem protocol, not sample-complexity results. On quantifier-order pairs, comparing ordered centers is correct by construction, whereas the classifiers reach only $56.0 \pm 2.8\%$ (DFIL) and $52.6 \pm 2.7\%$ (FT; Table 18). Across 100 applicable follow-up checkpoints the nearest-center

grid has zero violations, whereas main-path text-level rates are 9.71% DFIL and 9.70% FT, so joint training supplies no observed main-path certificate.

These supplementary experiments examine composition, ordering, and transfer; the matched controls appear in Appendix E. Synthetic and model-specific targets measure agreement with the stipulated operator. MF accuracy, cross-task, and CSMC results use Gaussian argmax, while the IPR oracle uses nearest-center decoding. The scalar theorem therefore does not certify the Gaussian or main-classifier scores. Results from distinct training and evaluation protocols are reported separately; unresolved configuration and checkpoint information limits some comparisons to descriptive diagnostics.

F.1 Few-Shot Adaptation

We evaluate sample efficiency on both FRoG and SST-5 by subsampling the training set (stratified by class), with Qwen2.5-7B and 3 seeds per configuration.

We report the main classifier, Gaussian-argmax MF branch, and a convex probability ensemble against fine-tuning. For each seed and budget, $\alpha \in \{0, 0.1, \ldots, 1\}$ maximizes validation accuracy, with ties resolved toward smaller α; the corresponding test score is then read out. Unlike this budget-specific selection, the cross-task table uses fixed $\alpha = 0.5$. Neither the main output nor the probability ensemble has a nearest-center certificate.

FRoG-Hard (budgets $\{16, 32, 64, 128, 256\}$): Table 20.

Table 20: Few-shot learning on FRoG-Hard (Qwen2.5-7B). Accuracy (%), mean $\pm$ std over 3 seeds. Ensemble coefficient α is selected on validation. Bold marks validation-selected ensemble means above fine-tune; this does not denote statistical significance.

Budget	16	32	64	128	256
Fine-tune	4.6 ± 1.3	9.6 ± 3.6	19.4 ± 5.7	27.5 ± 0.4	33.0 ± 1.7
DFIL (main)	5.1 ± 1.2	8.6 ± 0.4	23.2 ± 2.6	26.4 ± 1.8	34.0 ± 1.4
DFIL (MF)	21.1 ± 1.3	21.3 ± 6.0	8.5 ± 5.7	22.7 ± 4.2	25.7 ± 1.6
DFIL (ens., val-α)	$\mathbf{21.1 \pm 1.3}$	$\mathbf{21.0 \pm 6.3}$	$\mathbf{23.1 \pm 2.7}$	26.7 ± 1.7	$\mathbf{33.8 \pm 1.5}$
Δ (ensemble $-$ FT)	$\mathbf{+16.4}$	$\mathbf{+11.4}$	$\mathbf{+3.7}$	-0.8	$\mathbf{+0.8}$

SST-5 (budgets $\{32, 64, 128, 256, 512\}$): Table 21.

Table 21: Few-shot learning on SST-5 (Qwen2.5-7B). Accuracy (%), mean $\pm$ std over 3 seeds, α selected on validation.

Budget	32	64	128	256	512
Fine-tune	34.9 ± 0.5	36.6 ± 1.4	40.2 ± 3.1	45.5 ± 2.3	47.6 ± 2.0
DFIL (main)	34.9 ± 0.1	35.1 ± 1.9	39.8 ± 2.5	46.0 ± 2.4	48.8 ± 0.9
DFIL (MF)	38.1 ± 1.6	39.0 ± 1.7	40.3 ± 7.5	46.8 ± 3.9	50.9 ± 1.2
DFIL (ens., val-α)	$\mathbf{37.9 \pm 1.4}$	$\mathbf{39.0 \pm 1.6}$	$\mathbf{43.4 \pm 2.3}$	$\mathbf{48.1 \pm 2.9}$	$\mathbf{50.8 \pm 1.1}$
Δ (ensemble $-$ FT)	$\mathbf{+3.1}$	$\mathbf{+2.4}$	$\mathbf{+3.2}$	$\mathbf{+2.6}$	$\mathbf{+3.2}$

Observations. On FRoG-Hard, MF means exceed FT by 16.4 and 11.7pp at $N = 16$ and 32; validation-selected ensemble gaps are $16.4, 11.4, 3.7, -0.8, 0.8$pp across the five budgets. On SST-5, ensemble gaps range from 2.4 to 3.2pp. Differences are computed before rounding. At FRoG-Hard $N = 16$ the selected α is 0 for all three seeds, so the ensemble coincides with the MF path; the advantage is carried by the MF and ensemble paths, not uniformly by the main classifier. No intervals are reported because three-seed inference would require strong distributional assumptions; the paired per-seed ensemble-minus-FT differences straddle zero at FRoG-Hard $N = 64, 128, 256$ and at SST-5 $N = 128$, so those means should not be read as reliable effects. The three-seed

observations concern the stated paths and training regimes, not a sample-complexity guarantee or a proof of predictive superiority.

Stop-gradient protocol. The reported few-shot DFIL runs use $\mu_q = \exp[-(\hat{p} - c_q)^2/(2\sigma_q^2)]$ at every listed budget, including SST-5 at $N = 512$. The scalar predictor still receives MSE supervision, while MF cross-entropy updates the bank without passing through $\hat{p}$ to the extractor. This configuration was motivated by observed saturation in earlier runs. Recovery under the modified training rule does not by itself identify class imbalance as the cause or establish monotonic text predictions.

Regime dependence. Detachment is protocol-specific. It is also used for the six CSMC runs retrained after the saturation described above and for the independent-counting control, whereas the human-judgment confirmation (Appendix D.2) leaves the MF input attached. Thus it cannot be described as absent from all full-data experiments. No uniform, pre-specified criterion governs the detachment setting or the decision to retrain, so the affected results are diagnostic. A complete claim about detachment would require a matched comparison with selection fixed in advance.

F.2 Cross-Task and Transfer Diagnostics

Table 22 reports SST-5 and five-bucket STS22 accuracy for four backbones, with three seeds per cell. LoRA is applied to the four logical attention projections (q, k, v, o), with Phi-3.5-mini using its fused QKV layout. Main, Gaussian-argmax MF, and fixed-$\alpha = 0.5$ probability-ensemble outputs are shown separately. The MF column reports the Gaussian decoder as used in these runs, not a re-evaluation under nearest-center decoding. "Mono" is its finite synthetic-grid diagnostic.

Data processing and protocol limitations. SST-5 preprocessing uses the SetFit distribution of the Stanford Sentiment Treebank (Socher et al., 2013), with labels $y \in \{0, 1, 2, 3, 4\}$ and scalar targets $y/4$. It uses the supplied training, validation, and test partitions; when no validation partition is available, the test partition can take its place. Inputs are truncated to 128 tokens. The STS22 preprocessing uses the English-filtered cross-lingual collection of the SemEval-2022 multilingual news-article similarity task (Chen et al., 2022), splits its supplied training partition $85/15$ using the run seed, and retains the supplied test partition. After clipping the similarity score s to $[0, 5]$, labels are $y = \min(4, \lfloor s \rfloor)$ and scalar targets are again $y/4$. Sentence pairs are concatenated with a separator and truncated to 256 tokens. Thus the scalar target represents the ordinal bucket, not the original continuous similarity score.

Both methods use seeds $42, 123, 456$. The evaluation sets contain 2,210 SST-5 items with class counts $(279, 633, 389, 510, 399)$ and 197 STS22 items with counts $(0, 58, 36, 30, 73)$. Table 22 reports means and sample standard deviations; the mixture fixes $\alpha = 0.5$. On STS22, always predicting the highest class gives $73/197 = 37.06\%$, above all model means. This test-frequency reference describes class imbalance rather than a validation-selected baseline, and limits the absolute-utility interpretation of relative gains.

These comparisons use method-specific validation objectives and the final training state rather than restoring the minimum-loss checkpoint. Per-run data versions, partitions, and MF-detachment settings are not fully established, including whether the SST-5 validation fallback was active. We therefore treat the results as supplementary diagnostics, without claiming verified partition independence or the shared checkpoint selection used by the human-judgment confirmation and matched follow-ups.

Path-per-task summary. Selecting the MF path for SST-5 and the main path for STS22 gives larger reported DFIL means in $7/8$ backbone–task cells. This descriptive tally uses task-level routing and correlated comparisons; it is not a test of general superiority. The MF result uses Gaussian argmax, while main-path results use the unconstrained classifier; neither carries the nearest-center guarantee. These experiments do not establish why an individual backbone benefits or loses from a particular path.

F.3 Per-Class Accuracy, Learned Centers, and CORAL Comparison

F.3.1 Per-Class Accuracy Across All Scales

Table 24 presents per-class accuracy for DFIL and fine-tuning across all four model scales.

Table 22: Cross-task accuracy (%) on SST-5 and STS22 across four backbone families, grouped by backbone. Mean ± std over 3 seeds. Bold marks the largest reported mean per row, not a significance test. "Mono" is the synthetic-grid violation rate of the MF branch.

Task	DFIL main	DFIL MF	DFIL ens$_{\alpha=0.5}$	Fine-tune	Mono
Qwen2.5-7B-Instruct					
SST-5	57.78 ± 1.33	$\mathbf{57.98 \pm 0.82}$	57.86 ± 1.41	57.47 ± 0.64	0.0%
STS22	34.18 ± 0.59	28.93 ± 1.52	$\mathbf{34.86 \pm 1.06}$	32.99 ± 0.88	0.1%
Mistral-7B-Instruct-v0.3					
SST-5	60.47 ± 0.98	$\mathbf{60.71 \pm 0.35}$	60.39 ± 0.96	59.22 ± 0.11	0.0%
STS22	$\mathbf{36.21 \pm 1.47}$	29.78 ± 1.17	35.87 ± 1.06	35.03 ± 3.05	0.1%
Phi-3.5-mini-3.8B					
SST-5	53.80 ± 0.30	$\mathbf{54.99 \pm 0.35}$	53.97 ± 0.15	54.31 ± 1.15	0.0%
STS22	32.49 ± 1.83	29.95 ± 0.88	32.83 ± 1.47	$\mathbf{33.50 \pm 1.34}$	0.1%
Llama-3.1-8B-Instruct					
SST-5	58.51 ± 0.68	$\mathbf{59.17 \pm 0.69}$	58.55 ± 0.71	58.60 ± 0.21	0.0%
STS22	$\mathbf{35.03 \pm 2.21}$	27.92 ± 1.34	34.69 ± 3.67	32.49 ± 3.33	0.1%

Table 23: Transfer diagnostic on the 401-item QuRe adaptation (Appendix D.4.2); four represented gold classes and five mapped output labels. Ordinary accuracy (%), mean ± sample std over 3 seeds, no retraining. The descriptive constant-*some* reference is 48.88%, above every model mean.

Backbone	DFIL (main cls)	DFIL (MF branch)	Fine-tune
Mistral-7B-v0.3	$\mathbf{19.0 \pm 2.3}$	14.2 ± 5.4	16.5 ± 0.9
Qwen2.5-7B	14.9 ± 2.5	$\mathbf{31.3 \pm 2.7}$	16.6 ± 4.4

Observations. The weakest class varies by scale and method; "some" is not uniformly the hardest. For "most", DFIL − FT is −1.7pp at 7B and +9.4pp at 14B. At 72B, DFIL has higher means on "some" (59.2 vs. 53.3) and "most" (78.2 vs. 74.4), but lower accuracy on "all" (62.7 vs. 75.7). A 13pp gap on the latter cannot be dismissed as seed noise without an appropriate uncertainty analysis, but its per-seed values overlap (DFIL 51–73%, FT 70–79%), so the three seeds do not separate the arms. Class sizes range from 31 to 518, and differences on the small classes should not be read as reliable. These main-classifier outcomes do not identify a causal role for MF center placement. The three-seed per-class numbers are not the same aggregation as the six-seed scaling rows.

Table 24: Per-class accuracy (%) on FRoG-Hard across scales. Mean over 3 seeds. Sample count per class in parentheses in header row.

Scale	Method	none (31)	tiny (106)	few (518)	small (402)	some (202)	mod. (486)	most (199)	all (100)
7B	DFIL	43.0	30.5	66.2	45.7	30.9	60.4	24.0	44.7
	FT	36.6	31.1	72.8	45.1	36.1	60.2	25.6	28.3
14B	DFIL	48.4	58.5	72.7	56.9	33.8	62.9	38.0	39.7
	FT	43.0	50.3	78.9	52.1	32.2	65.3	28.6	39.3
32B	DFIL	47.3	56.6	80.1	56.7	42.9	65.2	41.5	56.3
	FT	39.8	50.3	75.4	52.3	37.1	61.7	41.0	56.7
72B	DFIL	64.5	86.5	85.8	74.2	59.2	85.4	78.2	62.7
	FT	69.9	80.2	84.7	78.2	53.3	83.3	74.4	75.7

F.3.2 CORAL Ordinal Baseline: Detailed Analysis

Table 25: CORAL ordinal regression per-class accuracy (%) on FRoG-Hard. Mean ± std over 3 seeds. DFIL and FT 7B per-class numbers, for comparison, are reported in Table 24.

	Scale	**none**	**tiny**	**few**	**small**	**some**	**mod.**	**most**	**all**
CORAL	7B	83.9	31.4	57.6	30.6	7.3	48.8	34.0	83.3
		±8.5	±1.4	±9.8	±7.3	±7.7	±4.0	±1.0	±1.5
CORAL	14B	92.5	28.0	68.9	38.8	10.9	59.9	44.9	83.7
		±4.9	±10.6	±8.8	±5.9	±10.7	±4.3	±0.3	±4.5

CORAL exhibits a distinctive bimodal accuracy pattern at both scales: very high accuracy on extreme classes ("none" 83.9–92.5%, "all" 83.3–83.7%) but poor accuracy on middle classes ("some" 7.3–10.9%, "small amount" 30.6–38.8%). Scaling from 7B to 14B improves CORAL's overall accuracy (+8.5pp), but the bimodal pattern persists: "some" improves only marginally (7.3% → 10.9%). The pattern may reflect the learned score, threshold placement, and data distribution; it does not establish that a shared scalar cannot represent the middle classes. DFIL and fine-tuning, with their full Q-way classifiers, achieve more balanced performance across all classes (Table 24).

F.3.3 Learned MF Centers and Widths Across Seeds

Table 26 reports learned centers and widths alongside the nominal author-specified initialization. The ordering parameterization maps these nominal values into its feasible set, including the fixed top center 0.98.

Table 26: Learned MF centers and widths (Qwen2.5-7B, each seed), with the reference initialization shown for comparison.

	none	tiny	few	small	some	mod.	most	all
Reference initialization:								
c_{init}	.020	.080	.180	.280	.400	.580	.780	.970
Learned centers (7B):								
seed 42	.020	.024	.089	.243	.353	.478	.832	.980
seed 123	.020	.024	.104	.239	.352	.490	.774	.980
seed 456	.020	.024	.082	.232	.347	.470	.736	.980
Learned widths (7B):								
seed 42	.001	.047	.085	.078	.066	.054	.102	.004
seed 123	.001	.051	.073	.075	.064	.049	.119	.089
seed 456	.001	.053	.078	.073	.057	.047	.115	.090
Learned centers (72B, QLoRA 4-bit per Table 16 footnote):†								
seed 42	.020	.032	.038	.051	.067	.103	.565	.980
seed 123	.020	.034	.044	.071	.128	.192	.655	.980
seed 456	.020	.029	.032	.044	.064	.093	.437	.980

† The 72B centers are compressed toward the lower anchor. The MF parameters are represented in 32-bit floating-point precision after loading the quantized backbone, so direct 4-bit rounding of the center logits is not the implemented mechanism. No matched precision ablation establishes the cause of this pattern. Ordered centers retain the nearest-center scalar guarantee, but neither semantic calibration nor the main classifier's accuracy follows from their spread.

Interpretation. The top center is fixed at 0.98 by parameterization; its stability is not learned evidence about the meaning of "all". The lower centers can approach the lower anchor, subject to the positive-increment floor. The three 7B runs show similar middle centers, but this is not a human-calibration result. Widths describe the fitted membership shape: the "all" widths $0.004, 0.089, 0.090$

are not uniformly near-degenerate. Nearest-center inference does not use widths, while Gaussian-argmax predictions and training gradients can depend strongly on them.

F.4 Synthetic Composition Diagnostics

The composition test contains 600 two-step and 300 three-step examples; training contains only single-step examples. From-text inference estimates the base proportion before composition, while oracle inference supplies it. We retain both original target protocols: a shared, author-specified reference bank and each model's own bank (self-consistent targets). Table 27 reports both, because they answer different questions.

Against reference-bank targets, DFIL obtains $31.5 \pm 0.4\%$, compared with $27.8 \pm 3.6\%$ for FT+Compose and $16.4 \pm 7.3\%$ for ordinary text-only fine-tuning. Against self-consistent targets, DFIL obtains $31.2 \pm 9.2\%$ versus $25.4 \pm 3.5\%$ over 20 paired seeds (+5.9pp from unrounded estimates; paired t $p = .011$, $d = .63$). The 100% self-consistent oracle follows from identical target and prediction operations. These measurements support execution of the specified operator from single-step training; independently correct linguistic composition requires independently justified targets. The reference-bank oracle difference (50.5% DFIL versus 94.3% FT+Compose) shows that learning centers can also reduce agreement with the stipulated reference. Independent counting controls give MAE .0357 for DFIL versus .0343 for FT+Compose (20 paired seeds), but the validity gate fails, leaving the intended comparison inconclusive (Table 9). Depth and per-seed analyses follow below; Appendix E.2 details the controls.

Table 27: Original operator-consistency diagnostic on 900 examples (Qwen2.5-7B, accuracy %). Self-consistent targets depend on each model's own bank; reference-bank targets use the fixed author-specified bank. Oracle predictions inject the reference base scalar. The self-consistent oracle is 100% by construction. The ordinary FT value was evaluated only against fixed reference-bank targets and is therefore reported only in that block. Differences are computed from unrounded estimates.

Method	**From-text**	**Oracle**[†]
Self-consistent GT (model-specific labels, n=20 seeds for FT+Compose / DFIL)		
FT+Compose	25.4 ± 3.5	$100.0 \pm 0.0^{\flat}$
DFIL (ours)	$\mathbf{31.2 \pm 9.2}$	$100.0 \pm 0.0^{\flat}$
Δ vs. FT+Compose	$\mathbf{+5.9}^{\ddagger}$	0.0
Reference-bank GT (shared, product-derived targets)		
Fine-tune (text-only)	16.4 ± 7.3	—
FT+Compose	27.8 ± 3.6	94.3 ± 0.0
DFIL	31.5 ± 0.4	50.5 ± 13.0
Δ DFIL − FT+Compose	$+3.7$	—

[‡] +5.9pp significant at n=20 (paired t p=0.011, Cohen's d=0.63).
[♭] 100% on the self-consistent oracle is by construction (each model vs. its own MF).

Table 28: Original self-consistent composition diagnostic (accuracy %, mean ± sample SD). Each model defines its own targets. Uniform initialization uses 3 seeds; the author-specified reference initialization uses 20. Center learning and target changes are coupled.

Init	**Method**	**From-text**	**Text-only**
Uniform	DFIL	$\mathbf{41.6 \pm 7.3}$	13.4 ± 4.5
	FT+Compose	18.4 ± 6.2	19.5 ± 4.1
Reference	DFIL	31.2 ± 9.2	18.5 ± 9.2
	FT+Compose	25.4 ± 3.5	17.6 ± 3.6

F.4.1 PER-SEED VARIANCE ANALYSIS

Two cohorts must be distinguished. The six-seed reference-initialized comparison uses seeds $42, 123, 456, 789, 1000, 2024$. Over these six seeds, DFIL gives $30.2037 \pm 9.3369\%$ and FT+Compose $23.3704 \pm 0.7492\%$ (sample SD). The 20-seed extension in Table 27 instead gives $31.2 \pm 9.2\%$ and $25.4 \pm 3.5\%$; its mean and uncertainty must be reported together.

In the six-seed exploratory regularization comparison, adding $\lambda_{\text{MF-reg}} = 0.1$ to penalize center drift gives $27.8889 \pm 8.7113\%$: the mean moves from 30.2 to 27.9 and the SD from 9.3 to 8.7, with 3 seeds improving and 3 degrading. This probe did not improve the mean and is not part of default DFIL. Since learned centers define both predictions and self-consistent targets, between-seed variability combines extraction, optimization, and target-partition changes. Freezing centers fixes the target partition but does not remove other training variability. These experiments do not identify a unique variance mechanism. Learned-but-shared centers and wider variance budgets remain possible future interventions.

F.4.2 COMPOSITIONAL TEST SET CONSTRUCTION

The compositional test set is generated synthetically to evaluate execution of the stipulated composition rule without compositional training examples. No compositional examples appear in the training data.

Level 1 (2-step, 600 examples): We use 12 quantifier pairs Q_1 of Q_2 (for example most of most, few of all and moderate amount of most), with $Q_1 \in \{\text{few}, \text{some}, \text{moderate amount}, \text{most}, \text{all}\}$ and $Q_2 \in \{\text{few}, \text{some}, \text{moderate amount}, \text{most}, \text{all}\}$. For each pair, 50 base proportions are drawn from a normal distribution with standard deviation 0.1 centered at the center of Q_2, clipped to $[0.01, 0.99]$. The composed proportion is $p_{\text{comp}} = c_{Q_1} \cdot p_{\text{base}}$. Under the reference-bank protocol, centers are the reference values and the target thresholds p_{comp} at the fixed cut points $(.05, .13, .23, .34, .49, .68, .88)$, which round the midpoints of adjacent reference centers; under the self-consistent protocol, the model's own centers are used and the target is its own Gaussian argmax (Appendix F.4.3).

Level 2 (3-step, 300 examples): We use 6 triples $Q_1 \circ Q_2 \circ Q_3$ with each $Q_i \in \{\text{few}, \text{some}, \text{most}, \text{all}\}$ (for example most of most of most and some of most of all). For each triple, 50 base proportions are drawn around the center of Q_3 in the same way, and $p_{\text{comp}} = c_{Q_1} \cdot c_{Q_2} \cdot p_{\text{base}}$.

Text templates: Each compositional example is rendered as a natural-language scenario. For a two-quantifier chain, "most of most" with estimated base proportion $p_{\text{base}} = 0.8$ uses one outer representative: $p_{\text{comp}} = 0.78 \times 0.8 = 0.624$. Nearest-center decoding with the illustrative reference bank gives "moderate amount." A three-step numerical chain uses two outer factors, $0.78 \times 0.78 \times 0.8 = 0.48672$; the same bank gives "some" under nearest-center decoding but "moderate amount" under the nominal unequal-width Gaussian argmax. These are specified numerical illustrations, not independently established semantic labels; the reference-bank protocol uses the fixed cut points above.

F.4.3 PER-SEED COMPOSITIONAL RESULTS

Self-Consistent Compositional Evaluation To eliminate any external ground-truth dependency, we evaluate each model against its own MF: ground-truth labels are defined as $\arg\max_q \mu_q(p_{\text{comp}})$ using the model's own membership functions. This measures internal compositional consistency.

The reported self-consistent gaps are $+23.2$pp with uniform initialization and $+5.9$pp with reference initialization. Because learning centers changes the targets themselves, these contrasts measure internal operator consistency rather than superiority on a common semantic target. The lower uniform text-only value (13.4 vs. 19.5%) documents sensitivity to the input regime; it does not identify the extractor's internal failure mechanism.

Reference-bank results. Table 29 uses shared, product-derived targets and three seeds. From-text means are 31.5% for DFIL and 27.8% for FT+Compose; the DFIL sample SD is 0.4pp, versus 7.3pp for FT text-only and 3.6pp for FT+Compose. The DFIL oracle has sample SD 13.0pp. Its centers differ across seeds, but these associations do not identify the cause of oracle variability. With p

Table 29: Compositional accuracy (%) per seed (Qwen2.5-7B). Top: overall accuracy; bottom: Level 1 / Level 2 accuracy. SD denotes sample standard deviation. Compose-text values come from the later batch-64 re-run, because results from the initial evaluation pipeline are not available; per-seed results are reported below.

	Overall			
Seed	FT-text	Compose-text	DFIL-text	DFIL-oracle
42	9.2	32.0	31.9	54.3
123	23.9	25.9	31.6	36.0
456	16.1	25.6	31.1	61.2
Mean	16.4	27.8	31.5	50.5
SD	7.3	3.6	0.4	13.0

	Level 1 / Level 2			
Seed	FT-text	Compose-text	DFIL-text	DFIL-oracle
42	7.0/13.7	25.0/46.0	30.8/34.0	65.7/31.7
123	34.0/3.7	19.0/39.7	30.5/33.7	37.0/34.0
456	9.5/29.3	22.0/32.7	28.7/36.0	69.3/45.0
Mean	16.8/15.6	22.0/39.4	30.0/34.6	57.3/36.9
SD	14.9/12.9	3.0/6.7	1.2/1.3	17.7/7.1

fixed, $|\hat{c}p - cp| \leq |\hat{c} - c|$; multiplication does not amplify absolute single-factor error, although quantization near a boundary can change labels.

F.4.4 Proportion Extraction on Compositional Text

Mean absolute base-proportion errors are substantial (approximately 0.27–0.35 for DFIL and 0.26–0.31 for FT+Compose). This is a diagnostic on synthetic composed text. Proposition B.1 bounds only the absolute error with fixed factors; it does not explain an observed accuracy gap or imply that an error of 0.30 is harmless. Label stability additionally depends on the factor and boundary errors in Proposition B.3. These measurements do not independently validate semantic correctness.

F.5 CSMC Per-Domain Consistency

Table 30 uses Gaussian membership argmax for DFIL and the main classifier for FT on 200 sentence pairs across ten domains. The original 21 DFIL runs give 94.45% consistency; replacing six runs with saturated memberships by their stop-gradient reruns gives the reported 96.40%, versus 90.90% for FT. The paired difference is 5.50pp (t-test $p = 5.56 \times 10^{-5}$; Wilcoxon $p \approx 0.0003$), with 18 positive, two zero, and one negative seed difference. These numbers therefore describe a mixed cohort in which six of the 21 DFIL seeds were retrained with the MF input detached, rather than a single uniformly trained cohort.

The CSMC evaluation decodes with $\arg\max \boldsymbol{\mu}$ rather than nearest-center decoding, so these results carry no nearest-center guarantee. The six replaced runs also prevent attributing the gap solely to the decoder or fuzzy loss. A constant prediction satisfies non-strict consistency, so these rates must be read alongside accuracy and the predicted-class distribution; high consistency alone does not establish language understanding.

F.6 IPR Per-Category Results

The Implicit Proportion Reasoning (IPR) benchmark consists of 40 sentences across four categories where the proportion is described in natural language without explicit fractions. Examples per category:

- **Linguistic** (n=12): "The vast majority of voters supported the measure" (≈ 0.85).

Table 30: Original CSMC consistency by domain (%). The comparison uses DFIL's Gaussian-argmax MF output and FT's main output over 21 seeds; six DFIL seeds were retrained after membership saturation. It describes this combined protocol rather than an isolated fuzzy-loss effect.

Domain	DFIL	FT	Δ	Domain	DFIL	FT	Δ
Public health	94.5	79.0	**+15.5**	Clinical trials	99.5	97.4	+2.1
Education	94.8	80.5	**+14.3**	Manufacturing	93.1	91.2	+1.9
Environment	98.6	88.6	**+10.0**	Research	97.9	96.2	+1.7
Governance	91.2	86.2	+5.0	Finance	95.7	94.8	+1.0
Election	99.8	96.4	+3.3	Sports	99.0	98.8	+0.2
Overall ($n=21$): DFIL MF 96.40 vs FT main 90.90**, Δ = +5.50pp**							

- **Approximate** (n=12): "Approximately one in five participants experienced the side effect" (≈ 0.20).
- **Comparative** (n=8): "Three times as many supporters as opponents" (≈ 0.75).
- **Emphatic** (n=8): "An overwhelming majority endorsed the resolution" (≈ 0.92).

The 40 proportions and class labels are hand-assigned probe targets. They are not uniformly obtained by nearest-center decoding of the nominal reference bank: $15/40$ labels disagree with that mapping. Comparative examples additionally stipulate the relevant population and category exhaustiveness. The text results use three Qwen2.5-7B checkpoints, and the oracle uses their learned centers; this is not an independently annotated semantic benchmark.

Table 31: IPR accuracy (%), mean ± sample SD over three seeds. DFIL (text) uses the main classifier logits; it does not evaluate the numerical head. DFIL (oracle) injects hand-assigned proportions into nearest-center decoding with the learned centers. The comparison changes both available input and output path and therefore does not isolate extraction error. Off-by-one allows one ordinal class of error.

Category	DFIL (text)	DFIL (oracle)	Fine-tune
Linguistic (n=12)	22.2 ± 12.7	$\mathbf{91.7 \pm 0.0}$	30.6 ± 4.8
Approximate (n=12)	16.7 ± 14.4	$\mathbf{77.8 \pm 4.8}$	13.9 ± 4.8
Comparative (n=8)	4.2 ± 7.2	$\mathbf{50.0 \pm 0.0}$	12.5 ± 21.7
Emphatic (n=8)	16.7 ± 19.1	$\mathbf{62.5 \pm 0.0}$	0.0 ± 0.0
Overall (exact)	15.8 ± 7.6	$\mathbf{73.3 \pm 1.4}$	15.8 ± 5.8
Overall (off-by-one)	48.3 ± 6.3	$\mathbf{100.0 \pm 0.0}$	45.0 ± 2.5

Interpretation. Recomputing all 120 oracle predictions from the stipulated proportions and the learned centers gives 73.3% exact and 100% off-by-one accuracy. This measures agreement with the probe's stipulated labels. The 15.8% text scores come from the main classifiers, so their difference from the oracle cannot be attributed specifically to the numerical head or backbone extraction. The oracle's remaining errors likewise do not establish a particular cause such as center drift. A direct extraction diagnosis would require comparing predicted and supplied proportions through the same fixed decoder on defensible targets. Neither the finite off-by-one result nor the scalar monotonicity theorem guarantees semantic accuracy on new text.

G Comparison with Related Approaches

Table 32: Interfaces and assumptions of related approaches. These are design comparisons, not claims that alternative architectures cannot realize DFIL's decoder.

Approach	**Structured object**	**Relation to this work**
CORAL / CORN (Cao et al., 2020; Shi et al., 2023)	Ordered output probabilities	Ordinal output constraints; a proportion-composition interface requires additional specification
ANFIS (Jang, 1993)	Trainable fuzzy rule system	Established membership learning and rule combination; input domain alone is not a novelty distinction
Logic Tensor Networks (Badreddine et al., 2022)	Differentiable logical satisfaction	Broader logical constraints; the present certificate concerns a specified scalar decoder
NeSyCoCo (Kamali et al., 2025)	Symbolic predicate composition	Compositional reasoning through predicates rather than the particular scalar interface studied here
FT+Compose	Scalar, frozen ordered bank, product	Shares nearest-center and product guarantees; controls for learning the bank and its auxiliary loss
DFIL	Scalar, learned ordered bank, product	Jointly trained dual-path implementation with a conditional scalar guarantee