STATISTICAL LIMITS OF LEARNING RULES FROM EXCEPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning systems that encounter repeated deviations from an established regularity must decide whether they reflect transient noise, stable but idiosyncratic exceptions, or a reusable secondary rule. We formalize this decision through grouped observations, where each type may be observed through multiple noisy tokens, and study how a fixed observation budget should be divided between repetition and type diversity. For a finite separated rule class, we prove minimax token complexity $\Theta((\log|\mathcal{G}| + \log(1/\delta))/(\alpha(1-2\eta)^2))$. A rate-optimal design heavily repeats one type to establish persistence and uses singleton observations for structural learning, while under balanced acquisition the information-optimal repetition count is $\Theta((1-2\eta)^{-2})$. Controlled simulations support both allocation laws. On CLINC150, persistent unstructured residuals are recoverable on observed intents but do not transfer to unseen intents, whereas structured residuals transfer and benefit from broader supporting-type coverage in all ten domains. On Omniglot, global character count is a poor proxy for relevant diversity, as directly increasing target-alphabet support at a fixed positive-token budget raises structured novel-type AUC from 0.722 to 0.792 (paired gain 0.069, 90% CI $[0.063, 0.076]$), while unstructured controls remain at chance.

1 INTRODUCTION

A learner that repeatedly encounters a violation of an established pattern must decide what representation the evidence supports. The deviation may be a transient observation error, a stable property of one particular type, or the first visible instance of another regularity that should generalize beyond the observations already seen. Modern learning systems can retain both broad structure and rare facts, and recent work has begun to characterize how generalized rules, memorized examples, and shared representations can coexist (Farné et al., 2026; Cheng and Lippl, 2026; Jiang et al., 2021; Park and Choi, 2026). However, before that representational question can be resolved, the data must support a statistical distinction among noise, local persistence, and reusable structure.

This distinction becomes sharp when observations have two natural scales. An entity, lexical item, intent, patient, user, or visual concept can be regarded as a *type*, while repeated measurements, utterances, visits, or images associated with the same entity are *tokens*. Repeated tokens reveal whether a deviation persists through measurement noise, while observations of additional types reveal whether persistent deviations align with a common feature rule. These sources of evidence complement each other, so a fixed observation budget creates an allocation problem between repeated measurements of known types and broader coverage of new types.

We study this problem through a grouped binary observation model. Each type X_i has a latent residual state Z_i, and each token reports that state through independent symmetric noise with rate η. In the transient world N, no stable type state exists. In the persistent-exception world E, the state is stable across repeated tokens but independent of the type features. In the structured world R_g, the state is generated by a secondary rule g from a class $\mathcal{G}$. The balanced formulation matches the one-token marginal across the three worlds, removing anomaly frequency and forcing the learner to use within-type repetition to establish persistence and cross-type regularity to establish structure.

Our first result characterizes the minimax token cost of jointly identifying the data-generating world and recovering the secondary rule when structure is present. For a finite α-separated family of

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.

approximately equidistant rules,

$$B^\star(\delta) = \Theta\left(\frac{\log|\mathcal{G}| + \log(1/\delta)}{\alpha(1-2\eta)^2}\right).$$

The upper bound has a nonuniform form, as one type is observed repeatedly to determine whether deviations persist, while the remaining structural evidence comes from fresh singleton types. Minimax-rate learning does not require denoising every type used to identify the rule.

Our second result studies the constraint in which every acquired type receives the same number r of observations. If $\Gamma_q(r)$ is the worst pairwise Chernoff information per token among the three worlds, with $q = 1 - 2\eta$, then every information-optimal repetition count satisfies

$$r^\star(q) = \Theta(q^{-2}).$$

For $rq^2 \ll 1$, transient and persistent observations differ only through weak within-type dependence, so a type is abandoned before its state can be resolved. For $rq^2 \gg 1$, the latent state is already effectively known, while the information distinguishing a reusable rule from an idiosyncratic exception is bounded at the type level. The optimal balanced allocation lies between these regimes.

We evaluate these predictions in controlled simulations and through residual-label interventions on natural input spaces. The controlled study recovers the predicted joint dependence on rule complexity, separation, and noise, and the nonuniform allocation reduces the finite-sample cost relative to uniformly denoising every observed type. Then, CLINC150 and Omniglot provide semantic and visual hierarchies in which persistence and transfer can be separated. Structured residuals transfer to unseen types, persistent unstructured residuals do not, and transfer improves when additional observations increase *supporting* type diversity rather than merely global type count.

Roadmap.

- Sections 3 and 4 formalize the three worlds and establish the minimax allocation law.
- Sections 5 and 6 analyze balanced replication and an anytime-valid rule-promotion procedure.
- Section 7 tests the resulting allocation predictions in controlled simulations, CLINC150, and Omniglot.
- The appendix contains complete proofs and additional experimental diagnostics.

2 Related Work

Memorization and generalization are not necessarily opposing behaviors. Example-specific memory can be statistically necessary on long-tailed distributions (Feldman, 2020; Feldman and Zhang, 2020), and recent solvable models focus on the coexistence of rules and stored facts. The Rules and Facts model studies simultaneous generalization and memorization across teacher-student, random-feature, and kernel regimes (Farné et al., 2026), while Cheng and Lippl analyze relational generalization when specified exceptions must also be retained (Cheng and Lippl, 2026). Work on structural regularities, lifted representations, and description length studies when collections of examples are compressed into shared structure (Jiang et al., 2021; Park and Choi, 2026; Lan et al., 2024). These results characterize how rules and exceptions can coexist once their roles are expressed in a learner. Our setting instead places uncertainty over those roles in the observation process itself.

Linguistic productivity provides a view that explicitly focuses on the distinction between types and tokens. The Tolerance Principle relates productive rule use to lexical type counts and exceptions (Yang, 2016), recursive variants search for narrower productive regularities when broader ones fail (Belth et al., 2021), and recent language-model experiments test analogous sensitivity to type diversity and exception rates (Friedman et al., 2026). Earlier symbolic induction also distinguished observational noise from stable linguistic exceptions (Déjean, 2002). Grouped and multiple-instance learning formalize settings in which several observations belong to a common latent unit (Sabato and Tishby, 2012; Choi et al., 2021), while model-selection theory studies the statistical price of adapting among candidate structures (Hanneke et al., 2023). Our formulation couples these ingredients by using repeated tokens, which identify whether a stable type-level state exists, as well as diversity across types, which identifies whether those states are predictable from shared features.

A separate empirical literature diagnoses difficult, mislabeled, or influential examples from a fixed training set. Dataset Cartography, AUM, EL2N and GraNd scores, and TracIn use training trajectories, margins, gradients, or influence estimates to characterize individual examples (Swayamdipta et al., 2020; Pleiss et al., 2020; Paul et al., 2021; Pruthi et al., 2020). These methods evaluate examples after data have been collected or during optimization, whereas our objective is to decide how the next observations should be allocated between repeated measurements of known types and new types when the status of the deviations is still unresolved.

3 Problem Formulation

3.1 Types, tokens, and residual states

Let $(\mathcal{X}, P_X)$ be an input space, and let $X_1, \dots, X_T$ denote independently sampled types. Type i may be observed r_i times, giving tokens $Y_{i1}, \dots, Y_{ir_i}$, and the total token budget is

$$B = \sum_{i=1}^{T} r_i.$$

The model can be read as operating after a dominant predictor has captured the prevailing regularity, with $Y_{ij} \in \{-1, +1\}$ recording the sign of a remaining binary deviation for token j of type i. For example, a classifier may already capture the dominant semantic decision while residual errors recur for particular intents. The statistical question is whether those errors are token noise, intent-specific effects, or evidence of a shared semantic subrule. Therefore, we want to determine which explanation the residual evidence supports.

Token noise is symmetric. We write

$$q = 1 - 2\eta, \qquad P(\xi_{ij} = +1) = 1 - \eta = \frac{1+q}{2}, \qquad P(\xi_{ij} = -1) = \eta,$$

where $0 < q < 1$. We use q as a channel quality parameter, with smaller values corresponding to noisier token observations.

The main theorems use balanced residual states because balance removes marginal prevalence as an auxiliary signal and isolates the distinction between persistence and structure. The experiments also examine natural hierarchies whose prevalences need not exactly match this idealized setting.

3.2 Three explanations for a deviation

The transient world contains no stable type state. We set

$$\mathsf{N} \qquad Y_{ij} \overset{\text{iid}}{\sim} \mathrm{Rad}(1/2),$$

independently of the type features.

The persistent exception world introduces a stable latent state but makes that state unpredictable from X_i. We draw

$$Z_i \overset{\text{iid}}{\sim} \mathrm{Rad}(1/2), \qquad Z_i \perp X_i,$$

and observe

$$\mathsf{E} \qquad Y_{ij} = Z_i \xi_{ij}.$$

The structured world uses a secondary rule g from a class $\mathcal{G}$ and sets

$$\mathsf{R}_g \qquad Z_i = g(X_i), \qquad Y_{ij} = g(X_i)\xi_{ij}.$$

We assume each g is balanced under P_X in the main analysis.

A single token has the same marginal distribution under N and E, and the same remains true under every balanced R_g. Repeated observations are necessary to establish persistence in this symmetric setting, while correlation between the persistent state and X is necessary to establish reusable structure. If g is unbalanced, singleton observations can additionally reveal R_g through a shifted marginal, but N and E remain exactly indistinguishable when every type is observed once.

3.3 RULE SEPARATION AND MINIMAX RISK

For $g, h \in \mathcal{G}$, define

$$d(g,h) = P_X(g(X) \neq h(X)).$$

Our finite class result assumes that

$$\alpha \leq d(g,h) \leq \kappa\alpha$$

for all distinct g, h, where α controls the minimum separation and κ prevents the family from mixing widely different separation scales.

An estimator outputs N, E, or $(\mathsf{R}, \widehat{g})$. Under the structured world, success requires exact recovery of the generating rule. For a design with total budget B, define its worst-case error

$$R_B(\widehat{H}) = \max\left\{P_{\mathsf{N}}(\widehat{H} \neq \mathsf{N}), P_{\mathsf{E}}(\widehat{H} \neq \mathsf{E}), \max_{g\in\mathcal{G}} P_{\mathsf{R}_g}(\widehat{H} \neq (\mathsf{R}, g))\right\},$$

and let $B^\star(\delta)$ be the minimum budget over grouped designs and estimators whose worst-case error is at most δ.

4 MINIMAX EVIDENCE ALLOCATION

The first theorem gives the minimax token complexity and identifies a rate-optimal acquisition strategy. Exact constants appear in Appendix B.

Theorem 1 (Minimax complexity with sparse replication)**.** *Fix $q_0 < 1$, $\kappa \geq 1$, $M = |\mathcal{G}| \geq 4$, and $\delta \leq 1/4$. Suppose $0 < q \leq q_0$, every $g \in \mathcal{G}$ is balanced, and every distinct pair satisfies*

$$\alpha \leq d(g,h) \leq \kappa\alpha.$$

There exist constants $c = c(q_0, \kappa) > 0$ and $C > 0$ such that

$$c \cdot \frac{\log M + \log(1/\delta)}{\alpha q^2} \leq B^\star(\delta) \leq C \cdot \frac{\log M + \log(1/\delta)}{\alpha q^2}.$$

The upper bound is attained by a design in which one type receives

$$O\big(q^{-2}\log(1/\delta)\big)$$

repetitions and every remaining acquired type receives one token.

The proof separates persistence from structural identification. Repeating one type distinguishes an independent token process from a stable latent type state because its signed average concentrates near zero under N and near magnitude q under either persistent world. Once persistence has been established, singleton types suffice for structural discovery because E makes the persistent state independent of X, whereas R_g makes the state predictable from X. Maximum likelihood over E and the finite rule family succeeds after order $(\log M + \log(1/\delta))/(\alpha q^2)$ singleton observations, while the converse combines a pairwise testing lower bound with a packing argument over $\mathcal{G}$.

Because $\alpha \leq 1$, the repeated-type cost $O(q^{-2}\log(1/\delta))$ is absorbed by the structural term in the displayed minimax rate. Repetition is needed to verify that deviations persist, but that verification does not have to be repeated for every type used to learn structure. The rate can be attained by concentrating repetition on one persistence probe and using the remaining budget for broad singleton coverage.

Corollary 1 (Pure diversity cannot establish persistence)**.** *If every acquired type is observed exactly once, then the complete observed data have the same distribution under N and E for every budget B. Under equal prior probability on these two worlds, every test has average error $1/2$.*

Corollary 1 is an exact identifiability statement rather than a finite-sample concentration effect. New types can provide arbitrarily much evidence about feature-aligned structure, but they cannot reveal whether a feature-independent deviation is stable at the type level when every type is seen only once.

5 Balanced Types and Tokens

The sparse construction in Theorem 1 uses highly unequal replication, while many collection protocols require the same number of observations for each acquired type. We now study the balanced design in which every sampled type receives r observations.

For this section the candidate rule g is known, which isolates the allocation problem from rule search. Let $A = g(X)$ be a balanced sign, and let $P_{\mathsf{N}}^{(r)}$, $P_{\mathsf{E}}^{(r)}$, and $P_{\mathsf{R}}^{(r)}$ denote the distribution of $(A, Y_1, \dots, Y_r)$ under the three worlds. For distributions P and Q with densities p and u, write

$$C(P, Q) = -\log \inf_{0 \le s \le 1} \int p^s u^{1-s}$$

for Chernoff information (Chernoff, 1952). Define the worst pairwise information per token

$$\Gamma_q(r) = \frac{1}{r} \min \left\{ C(P_{\mathsf{N}}^{(r)}, P_{\mathsf{E}}^{(r)}), C(P_{\mathsf{N}}^{(r)}, P_{\mathsf{R}}^{(r)}), C(P_{\mathsf{E}}^{(r)}, P_{\mathsf{R}}^{(r)}) \right\}.$$

Theorem 2 (Optimal balanced replication scale)**.** *There exist universal constants $0 < c < C < \infty$ such that for every $0 < q \le 1/2$, every maximizer*

$$r^\star(q) \in \arg\max_{r \in \mathbb{N}} \Gamma_q(r)$$

satisfies

$$\frac{c}{q^2} < r^\star(q) < \frac{C}{q^2}.$$

Equivalently,

$$r^\star(q) = \Theta\big((1 - 2\eta)^{-2}\big).$$

The proof identifies a different obstruction on each side of the optimal scale. When $rq^2 \le 1$, the equality of the one-token marginals under N and E forces persistence information to arise through dependence among tokens from the same type, and an exact χ^2 calculation gives

$$C(P_{\mathsf{N}}^{(r)}, P_{\mathsf{E}}^{(r)}) \lesssim r^2 q^4.$$

The resulting information per token is at most order rq^4, so replication far below q^{-2} abandons each type before persistence can be resolved reliably.

At the other extreme, the complete rule versus exception information carried by one type is bounded by the entropy of the latent balanced state,

$$D_{\mathrm{KL}}(P_{\mathsf{R}}^{(r)} \| P_{\mathsf{E}}^{(r)}) = I(A; Y_{1:r}) \le \log 2.$$

The information per token is at most $(\log 2)/r$, so replication far above q^{-2} spends additional tokens after the type state is already effectively known. A construction with r proportional to q^{-2} separates all three worlds by a constant total variation gap within one block, which yields order q^2 Chernoff information per token and establishes the claimed scale.

Theorem 2 quantifies the same evidence tradeoff under a different constraint. Lower noise makes persistence visible after fewer repeated measurements, while higher noise requires more evidence from each type before additional type diversity becomes the more efficient use of a balanced budget.

6 A Sequential Interpretation

Once persistence has been established, structural promotion can be made anytime-valid. Let π be a prior or coding distribution over $\mathcal{G}$, write $K(g) = -\log \pi(g)$, define $L_n(g) = \prod_{i=1}^n (1 + qY_i g(X_i))$, and promote g when $\log L_n(g)$ first crosses $K(g) + \log(1/\delta)$.

Proposition 1 (Anytime complexity-calibrated promotion)**.** *If τ_g is the promotion time above, then*

$$P_{\mathsf{E}}(\exists g : \tau_g < \infty) \le \delta, \qquad \frac{\tau_g}{K(g) + \log(1/\delta)} \to \frac{1}{I(q)} \quad \textit{a.s. under } \mathsf{R}_g,$$

where

$$I(q) = \frac{1+q}{2} \log(1+q) + \frac{1-q}{2} \log(1-q) = \frac{q^2}{2} + O(q^4).$$

The familywise Type I bound follows from nonnegative likelihood-ratio martingales weighted by π, while $I(q)$ is the evidence rate under the true rule. Exact knowledge of the noise level is not needed for validity, as replacing q in the likelihood ratio by any fixed $\widetilde{q} \in (0, 1)$ preserves the same false-promotion bound under E. Under R_g the drift becomes

$$I(q, \widetilde{q}) = \frac{1+q}{2} \log(1+\widetilde{q}) + \frac{1-q}{2} \log(1-\widetilde{q}),$$

which is maximized at $\widetilde{q} = q$, so misspecification affects power rather than Type I control. Appendix D gives the proof and a fixed-rule converse.

7 Experiments

The empirical evaluation separates two goals. Controlled simulations test the finite-sample consequences of the proved allocation laws under matched assumptions, while residual-label interventions on CLINC150 and Omniglot test whether the distinction between persistence and reusable structure remains visible on natural semantic and visual inputs. In the natural-data experiments the inputs and their native hierarchies are unchanged, and only the residual mechanism is controlled, which lets persistence on observed types and transfer to unseen types be measured separately.

7.1 Controlled validation of the minimax scaling

We first evaluate Theorem 1 in a matched finite model with balanced residual states, finite rule families of controlled separation, and exact likelihood inference over N, E, and the structured hypotheses. We vary rule-class size, realized separation, token noise, and observation budget, with structured success requiring both identification of the structured world and recovery of the generating rule.

Figure 1 compares the empirical 80% recovery boundary with the joint complexity predicted by Theorem 1. After multiplying the required budget by q^2, the boundary is approximately linear in $(\log M + \log(1/\delta))/\alpha_{\min}$, with fitted slope 1.085 and 90% confidence interval $[0.991, 1.178]$ for rule families through $M = 256$. Because the experimental grid varies rule-class size and realized separation jointly, this calculation supports the joint complexity law rather than separately estimating the α^{-1} dependence. The structured component averages exact recovery over sampled generating rules, so the empirical boundary is an average-rule analogue of the theorem's worst-case criterion.

The allocation prediction is visible at finite sample sizes as well. Across configurations where both 80% boundaries are estimable, the sparse-replication design requires a median 0.480 times the token budget of uniformly denoising every observed type, with 90% bootstrap interval $[0.465, 0.496]$. Exact rule recovery is substantially harder than deciding which world generated the data, with a median budget ratio of 11.11 in the subset where both boundaries are interior. Pure singleton acquisition reproduces Corollary 1, meaning N and E remain indistinguishable regardless of the number of observed types.

7.2 Balanced replication under exact information

The balanced allocation result can be evaluated without Monte Carlo error because Theorem 2 defines $r^\star$ through an exact Chernoff objective. We maximize this objective at thirteen channel qualities between $q = 0.05$ and $q = 0.90$, where the optimizer ranges from $r^\star = 1997$ to $r^\star = 4$. On log-log axes, the exact optimizer follows the q^{-2} reference scaling across the tested range (Figure 2). A finite-range fit has slope -2.094 with 90% confidence interval $[-2.143, -2.044]$; the deviation of that fitted slope from -2 reflects curvature over the finite grid and is not used as an estimate of a sharper asymptotic law.

7.3 CLINC150: supporting semantic diversity

CLINC150 contains 150 in-scope intents grouped into ten semantic domains of 15 intents each (Larson et al., 2019). We treat an intent as a type, an utterance as a token, and membership in a designated domain as the candidate reusable structure. Because our evaluation constructs new intent-disjoint episode splits rather than using CLINC150's original intent-classification split, we pool the official in-scope train, validation, and test utterances while excluding out-of-scope examples. Each

episode then withholds 52 intents before fitting, using 13 for threshold selection and 39 for novel-type evaluation, and samples training types from the remaining 98 at the target-domain prevalence $\rho = 0.1$. The predictor is logistic regression with $C = 2.0$ and the `liblinear` solver, fit separately with word 1-2 gram TF-IDF features, character 3-5 gram TF-IDF features, and their concatenation; domain and intent identifiers are never features.

Under R, the persistent residual sign is determined by membership in the designated domain. Under E, each intent receives an independent persistent sign with the same positive prevalence, and under N each token is drawn independently from the corresponding one-token marginal. Training labels are corrupted by symmetric noise at rate η. Novel-type ROC-AUC is evaluated on the 39 held-out intents, while persistence AUC evaluates type-state prediction on an utterance panel from observed intents. Reported confidence intervals resample the ten domains after averaging episodes, allocations, and the three TF-IDF representations within each domain.

The three worlds separate in the expected way. Across the full experimental grid, structured novel-type AUC is 0.666 with 90% confidence interval $[0.630, 0.709]$, while both N and E remain at 0.500. Persistence AUC is 0.891 under R and 0.855 under E, compared with 0.500 under N, so persistent unstructured residuals are recoverable on observed types without transferring to unseen intents.

At fixed budget $B = 120$ and $\eta = 0$, broader type coverage improves structured transfer (Figure 3). Novel-type AUC rises from 0.693 with 10 observed intents and 12 utterances per intent to 0.781 with 60 observed intents and two utterances per intent; the paired domain-level gain is 0.088 with 90% confidence interval $[0.068, 0.108]$, and all ten domains have positive effects. Increasing token noise reduces the domain-mean structured AUC from 0.734 at $\eta = 0$ to 0.602 at $\eta = 0.3$, while novel-type performance under N and E remains at chance.

7.4 OMNIGLOT: GLOBAL AND SUPPORTING DIVERSITY

Omniglot contains 1,623 handwritten characters from 50 alphabets, with 20 drawings per character (Lake et al., 2015). We treat a character as a type, a drawing as a token, and membership in a designated alphabet as reusable structure. Pixel, HOG, and 512-dimensional ImageNet-pretrained ResNet18 features are extracted once and held fixed, after which a StandardScaler and `liblinear` logistic regression with $C = 2.0$ are fit within each episode; alphabet identity is never an input feature. Each episode withholds a disjoint character set and samples T training characters and r drawings per character without replacement. Novel-type AUC is evaluated on held-out characters and persistence AUC on held-out drawings of observed characters.

Across the full grid, novel-type AUC is 0.500 under both N and E and 0.580 under R (90% CI $[0.567, 0.593]$), showing that structured residuals transfer across characters while persistent unstructured residuals do not. Nevertheless, a global type-count intervention appears to favor repetition.

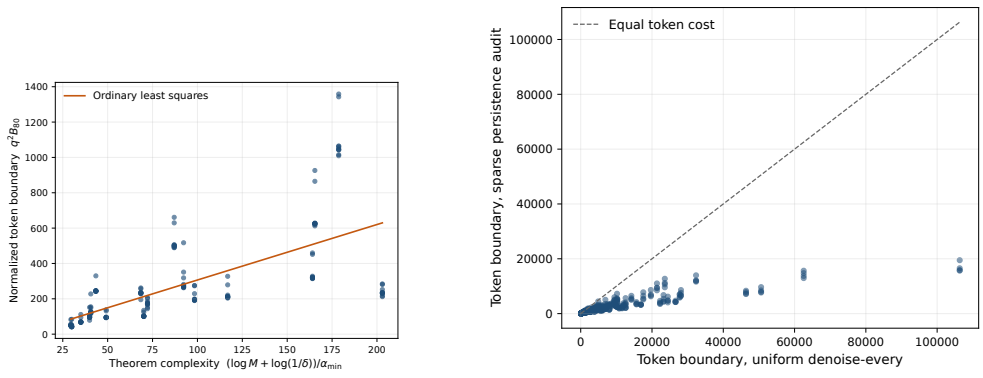


Figure 1: **Finite-sample behavior of the minimax allocation.** Left: after q^2 normalization, the empirical 80% recovery boundary grows approximately linearly with $(\log M + \log(1/\delta))/\alpha_{\min}$ (slope 1.085, 90% CI $[0.991, 1.178]$). Right: matched 80% recovery boundaries for sparse and uniform replication; points below the diagonal favor sparse replication.

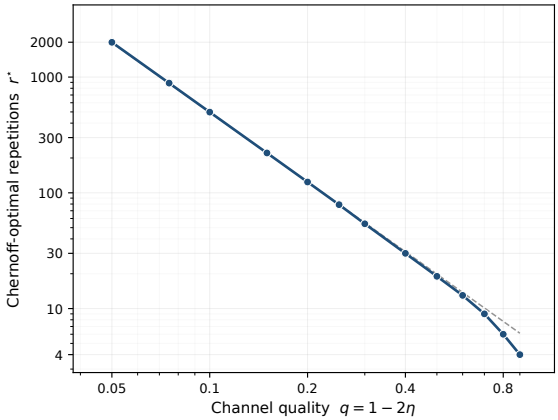


Figure 2: **Chernoff-optimal balanced replication.** Exact repetition count $r^\star(q)$ versus channel quality $q = 1 - 2\eta$ on log-log axes. The dashed line is a reference line of slope -2, normalized through an interior grid point,

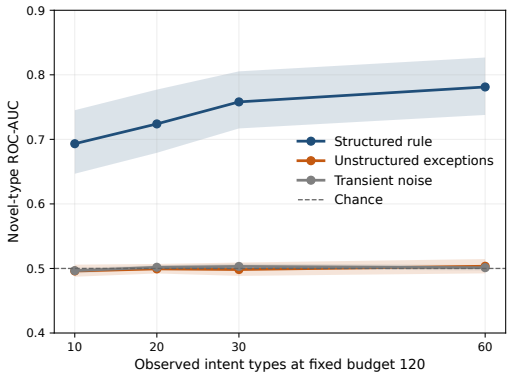


Figure 3: **Supporting semantic diversity improves structured transfer on CLINC150.** At $B = 120$ and $\eta = 0$, structured novel-type ROC-AUC increases as the balanced allocation shifts from repeated utterances toward more observed intents. Bands are 90% bootstrap intervals over the ten domains; N and E remain at chance on unseen intents.

At $B = 120$ and $\eta = 0$, structured AUC falls from 0.665 at $T = 10$ to 0.631 at $T = 120$, a paired change of -0.035 (90% CI $[-0.042, -0.028]$). The intervention changes the relevant support only weakly, because the mean number of observed target-alphabet characters rises from 1.00 to 2.36 while drawings per character fall from 12 to 1. Consistent with this explanation, the global-diversity effect is more favorable for larger alphabets (Spearman $\rho = 0.45$, $p = 0.001$).

We next vary supporting diversity directly. Using the ResNet18 representation, we hold the positive-token budget at 12 drawings and the negative panel fixed, then vary K_+, the number of observed target-alphabet characters, while keeping held-out target characters disjoint from training. Structured novel-type AUC rises monotonically from 0.722 at $K_+ = 1$ to 0.792 at $K_+ = 8$ (Figure 4). The primary paired gain across all 50 alphabets is 0.069 with 90% confidence interval $[0.063, 0.076]$; 49 alphabets have gains above 0.01, one is effectively unchanged, and none decreases by more than 0.01. At $K_+ = 12$, available for 48 alphabets, structured AUC reaches 0.796. Throughout the intervention, N and E remain at chance on novel types. Increasing relevant type diversity improves transfer even though the broader global-T intervention moves in the opposite direction.

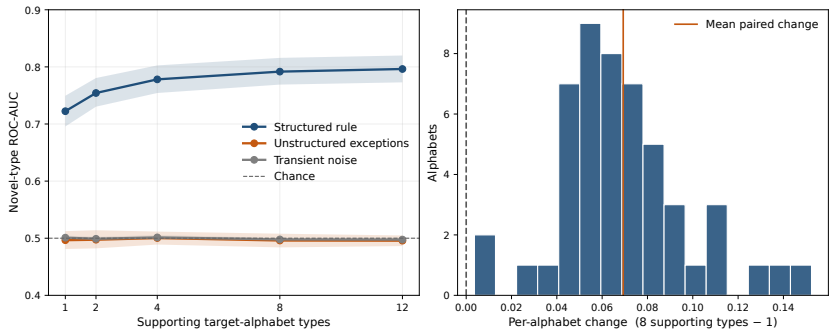


Figure 4: **Directly increasing supporting visual diversity improves structured transfer on Omniglot.** The number of observed target-alphabet characters K_+ varies at a fixed 12-drawing positive-token budget and fixed negative panel with $\eta = 0$. Left: novel-type ROC-AUC with 90% alphabet-level intervals. Right: per-alphabet change from $K_+ = 1$ to $K_+ = 8$, the primary paired endpoint.

8 Discussion

The theory separates evidence that a deviation persists from evidence that persistent deviations share reusable structure. Repetition resolves the first uncertainty, while diversity across informative types resolves the second. Under flexible acquisition these tasks can be separated by concentrating repeated observations on one persistence probe and using fresh types for structural learning. Balanced acquisition cannot separate the tasks in this way, and the optimal compromise occurs at the noise-dependent scale $r^\star = \Theta(q^{-2})$.

This perspective also clarifies the role of memorization. Under E, the exceptional state is stable but unpredictable from the observed features, so retaining the state of an observed type without transferring it is statistically appropriate rather than a failure to generalize. Under R, the same local persistence is accompanied by cross-type predictability, so the evidence supports a representation that transfers to new types. The distinction is not between memorization and generalization as competing objectives, but between local information that should remain local and repeated structure that supports compression into a rule.

CLINC150 adds supporting semantic types directly, and structured transfer improves in every domain. In Omniglot, increasing global character count adds little target-alphabet support and simultaneously removes repeated drawings, so the global intervention favors repetition. Directly increasing target-alphabet support under a fixed positive-token budget reverses that pattern. Type count alone is not the relevant quantity; the value of diversity depends on whether the added types carry evidence about the candidate structure.

The framework is intentionally stylized. The main theorems assume binary residual states, symmetric token noise, balanced prevalence, and finite approximately equidistant rule families. The controlled simulation averages structured recovery over sampled generating rules rather than estimating the theorem's literal worst case, and its design does not separately identify the α^{-1} factor. CLINC150 and Omniglot retain natural inputs and native hierarchies but impose controlled residual mechanisms, so the experiments test the proposed evidence distinction under known interventions rather than asserting that organic exceptions in either dataset follow the three-world model. Extensions to richer rule classes and adaptive acquisition require corresponding notions of local complexity and evidence-dependent sampling.

9 Conclusion

Repeated deviations require repetition to establish persistence and supporting-type diversity to justify reusable structure. We prove the corresponding allocation laws for flexible and balanced acquisition, and observe the same distinction in controlled simulations, CLINC150, and Omniglot.

Reproducibility Statement

Complete assumptions and proofs appear in the appendix. The controlled simulations use exact likelihood inference under fixed configuration grids. For CLINC150 and Omniglot, the supplementary material records the episode construction, representations, classifiers, residual mechanisms, evaluation metrics, and source data for every reported figure. The CLINC analysis uses the pinned in-scope CLINC150 data and domain mapping, while Omniglot features and character splits are generated from the stated preprocessing pipeline. Code and figure source data are provided to reproduce the reported analyses.

AI Use Statement

Large language models were used to assist with experiment design and implementation, to review mathematical derivations and proof development, and to edit portions of the manuscript. Synthetic experimental data were generated programmatically from the probabilistic models specified in the paper, where AI tools assisted with the associated code and workflow. The authors independently verified the mathematical arguments, experimental methodology and outputs, citations, and manuscript claims, and take full responsibility for the contents of the paper.

References

Caleb Belth, Sarah Payne, Deniz Beser, Jordan Kodner, and Charles Yang. The greedy and recursive search for morphological productivity. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pages 2869–2875, 2021.

Luke Cheng and Samuel Lippl. A mathematical theory of balancing relational generalization and memorization. arXiv preprint arXiv:2605.22972, 2026.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Youngwon Choi, Sungdong Lee, and Joong Ho Won. Learning from nested data with Ornstein auto encoders. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of PMLR, 2021.

Hervé Déjean. Learning rules and their exceptions. *Journal of Machine Learning Research*, 2:669–693, 2002.

Gabriele Farné, Fabrizio Boncoraglio, and Lenka Zdeborová. The Rules and Facts model for simultaneous generalization and memorization in neural networks. arXiv preprint arXiv:2603.25579, 2026.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why. Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891, 2020.

Adam E. Friedman, Stevan Harnad, and Rushen Shi. Tolerance Principle and small language model learning. arXiv preprint arXiv:2601.12179, 2026.

Steve Hanneke, Samory Kpotufe, and Yasaman Mahdaviyeh. Limits of model selection under transfer learning. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of PMLR, pages 5781–5812, 2023.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. Characterizing structural regularities of labeled data in overparameterized models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of PMLR, pages 5034–5044, 2021.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Nur Lan, Emmanuel Chemla, and Roni Katzir. Bridging the empirical theoretical gap in neural network formal language learning using minimum description length. arXiv preprint arXiv:2402.10013, 2024.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out of scope prediction. In *Proceedings of EMNLP IJCNLP*, pages 1311–1316, 2019.

Bumjin Park and Jaesik Choi. Lifted representation hypothesis in language models. arXiv preprint arXiv:2607.19360, 2026.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, 2020.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, 2020.

Sivan Sabato and Naftali Tishby. Multi instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13:2999–3039, 2012.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*, pages 9275–9293, 2020.

Charles Yang. *The Price of Linguistic Productivity. How Children Learn to Break the Rules of Language*. MIT Press, 2016.

A Proof Preliminaries

All logarithms in the proofs are natural. We repeatedly use the Rademacher concentration bound below.

Lemma 1 (Rademacher tail bound)**.** *Let $V_1, \dots, V_r \in \{-1, +1\}$ be independent and let $E[V_j] = m$. Then for every $t > 0$,*

$$P\left(\sum_{j=1}^{r}(V_j - m) \le -t\right) \le \exp\left(-\frac{t^2}{2r}\right).$$

If $m = 0$, then

$$P\left(\left|\sum_{j=1}^{r} V_j\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2r}\right).$$

Proof. For any $\lambda > 0$, exponential Markov inequality gives

$$P\left(\sum(V_j - m) \le -t\right) \le e^{-\lambda t}\prod_{j=1}^{r} E[e^{-\lambda(V_j - m)}].$$

For $V \in \{-1, +1\}$ with mean m, define

$$\phi(\lambda) = \log E[e^{\lambda(V-m)}].$$

We have $\phi(0) = \phi'(0) = 0$, while $\phi''(\lambda)$ is the variance of a random sign under an exponentially tilted law and is at most one. Taylor's theorem gives

$$\phi(\lambda) \le \frac{\lambda^2}{2}.$$

Consequently,

$$P\left(\sum (V_j - m) \le -t\right) \le \exp\left(-\lambda t + \frac{r\lambda^2}{2}\right).$$

The choice $\lambda = t/r$ proves the first inequality, and applying the same argument to the upper and lower tails proves the second. □

B Proof of Theorem 1

We prove explicit upper and lower bounds, from which the asymptotic statement in the main text follows.

Theorem 3 (Explicit minimax bounds)**.** *Under the assumptions of Theorem 1,*

$$B^\star(\delta) \ge \frac{1 - q_0^2}{16\kappa\alpha q^2}\left(\log M + \log\frac{1}{2\delta}\right),$$

while

$$B^\star(\delta) \le \left\lceil \frac{8}{q^2}\log\frac{8}{\delta}\right\rceil + \left\lceil \frac{8}{\alpha q^2}\log\frac{4(M+1)}{\delta}\right\rceil.$$

Proof. We first construct the upper bound.

Choose one probe type and observe it

$$r_0 = \left\lceil \frac{8}{q^2}\log\frac{8}{\delta}\right\rceil$$

times. Let

$$S_0 = \sum_{j=1}^{r_0} Y_{0j}.$$

Declare the data transient if

$$|S_0| < \frac{r_0 q}{2},$$

and otherwise proceed to a rule versus exception stage.

Under N, $E[Y_{0j}] = 0$, so Lemma 1 gives

$$P_{\mathsf{N}}\left(|S_0| \ge \frac{r_0 q}{2}\right) \le 2\exp\left(-\frac{r_0 q^2}{8}\right) \le \frac{\delta}{4}.$$

Under either persistent world, conditional on the latent sign Z_0 we have

$$Z_0 Y_{0j} = \xi_{0j}, \qquad E[Z_0 Y_{0j}] = q.$$

Therefore,

$$P\left(|S_0| < \frac{r_0 q}{2}\right) \le \exp\left(-\frac{r_0 q^2}{8}\right) \le \frac{\delta}{8}.$$

Now collect

$$n = \left\lceil \frac{8}{\alpha q^2}\log\frac{4(M+1)}{\delta}\right\rceil$$

fresh types, each observed once. Under E,

$$p_E(x, y) = p_X(x)\frac{1}{2},$$

while under R_g,

$$p_g(x, y) = p_X(x)\frac{1 + qyg(x)}{2}.$$

Run maximum likelihood over

$$\{\mathsf{E}, \mathsf{R}_{g_1}, \ldots, \mathsf{R}_{g_M}\}.$$

For two distributions P, Q, define

$$\mathrm{BC}(P, Q) = \int \sqrt{pq}.$$

If $X_1, \ldots, X_n$ are drawn from P, then

$$P^n\left(q^n(X_{1:n}) \geq p^n(X_{1:n})\right) \leq \mathrm{BC}(P, Q)^n,$$

because $p^n \leq \sqrt{p^n q^n}$ on the event that $q^n \geq p^n$.

For E and one rule,

$$\mathrm{BC}(P_E, P_g) = \frac{\sqrt{1+q} + \sqrt{1-q}}{2} =: b_0(q).$$

Since

$$b_0(q)^2 = \frac{1 + \sqrt{1-q^2}}{2} \leq 1 - \frac{q^2}{4},$$

we obtain

$$-\log b_0(q) \geq \frac{q^2}{8}.$$

For two distinct rules with disagreement $d = d(g, h)$,

$$\mathrm{BC}(P_g, P_h) = 1 - d + d\sqrt{1-q^2}.$$

Using $d \geq \alpha$ and

$$1 - \sqrt{1-q^2} \geq \frac{q^2}{2},$$

we have

$$\mathrm{BC}(P_g, P_h) \leq 1 - \frac{\alpha q^2}{2},$$

and this gives

$$-\log \mathrm{BC}(P_g, P_h) \geq \frac{\alpha q^2}{2}.$$

Since $\alpha \leq 1$, every distinct pair in the maximum likelihood family satisfies the common weaker bound

$$-\log \mathrm{BC}(P_a, P_b) \geq \frac{\alpha q^2}{8}.$$

A union bound over at most M competitors gives

$$P(\text{rule stage error}) \leq M \exp\left(-\frac{n\alpha q^2}{8}\right) < \frac{\delta}{4}.$$

Combining the repeated-type and structural stages yields worst-case error below δ and proves the upper bound.

We now prove the lower bound. Choose any distinct $g, h \in \mathcal{G}$. If a type receives r_i repetitions, the conditional distributions under R_g and R_h coincide when $g(X_i) = h(X_i)$ and differ by a binary symmetric channel when the rules disagree. The one-token KL divergence between opposite channel inputs is

$$D_{\mathrm{BSC}}(q) = q \log \frac{1+q}{1-q} = 2q \operatorname{arctanh} q.$$

For $q \leq q_0$,

$$\operatorname{arctanh} q = \int_0^q \frac{dt}{1-t^2} \leq \frac{q}{1-q_0^2},$$

so

$$D_{\mathrm{BSC}}(q) \le \frac{2q^2}{1-q_0^2}.$$

Summing over the grouped design gives

$$D_{\mathrm{KL}}(P_g \| P_h) = B\, d(g,h) D_{\mathrm{BSC}}(q) \le \frac{2\kappa\alpha B q^2}{1-q_0^2}.$$

Suppose the worst-case error is at most δ. Let A be the event that the estimator returns (R, g). Then

$$P_g(A) \ge 1-\delta, \qquad P_h(A) \le \delta.$$

Data processing through the binary event A gives

$$D_{\mathrm{KL}}(P_g \| P_h) \ge d_{\mathrm{bin}}(1-\delta \| \delta) = (1-2\delta)\log\frac{1-\delta}{\delta}.$$

For $\delta \le 1/4$,

$$d_{\mathrm{bin}}(1-\delta\|\delta) \ge \frac{1}{2}\log\frac{1}{2\delta}.$$

Therefore,

$$B \ge \frac{1-q_0^2}{4\kappa\alpha q^2}\log\frac{1}{2\delta}.$$

For the complexity lower bound, choose G uniformly from $\{1,\dots,M\}$ and generate data from R_{g_G}. If the estimator has worst-case error at most δ, then its average rule recovery error is at most δ. Fano's entropy argument gives

$$I(G;D) \ge \log M - \log 2 - \delta\log(M-1) \ge \frac{1}{4}\log M,$$

where the final inequality uses $M \ge 4$ and $\delta \le 1/4$.

Let

$$\overline{P} = \frac{1}{M}\sum_h P_h.$$

Convexity of negative logarithm yields

$$D_{\mathrm{KL}}(P_g \| \overline{P}) \le \frac{1}{M}\sum_h D_{\mathrm{KL}}(P_g \| P_h).$$

Averaging and using the pairwise upper bound gives

$$I(G;D) \le \frac{2\kappa\alpha B q^2}{1-q_0^2}.$$

Consequently,

$$B \ge \frac{1-q_0^2}{8\kappa\alpha q^2}\log M.$$

Combining the confidence and complexity lower bounds gives

$$B \ge \frac{1-q_0^2}{16\kappa\alpha q^2}\left(\log M + \log\frac{1}{2\delta}\right),$$

which completes the proof. □

Proof of Corollary 1. If every type is observed once, then under N each Y_i is an independent uniform sign. Under E,

$$Y_i = Z_i \xi_i,$$

where Z_i is an independent uniform sign. Multiplication by any independent noise sign leaves Y_i uniform, and independence across types is preserved. Therefore, the complete observed data have exactly the same law under N and E for every budget. □

C Proof of Theorem 2

We begin with two standard information inequalities that are included for completeness.

Lemma 2. *For any distributions P, Q,*

$$C(P, Q) \le D_{\mathrm{KL}}(P\|Q).$$

Proof. For $0 \le s < 1$, set $t = 1 - s$. Jensen's inequality gives

$$\log E_P\left[\left(\frac{q}{p}\right)^t\right] \ge tE_P \log \frac{q}{p} = -tD_{\mathrm{KL}}(P\|Q).$$

Therefore,

$$-\log \int p^s q^{1-s} \le tD_{\mathrm{KL}}(P\|Q) \le D_{\mathrm{KL}}(P\|Q).$$

Taking the supremum over s proves the claim. □

Lemma 3. *For any distributions P, Q with $P \ll Q$,*

$$D_{\mathrm{KL}}(P\|Q) \le \log\left(1 + \chi^2(P\|Q)\right) \le \chi^2(P\|Q).$$

Proof. Let $L = p/q$. Jensen's inequality under P gives

$$D_{\mathrm{KL}}(P\|Q) = E_P[\log L] \le \log E_P[L] = \log E_Q[L^2] = \log\left(1 + \chi^2(P\|Q)\right).$$

The second inequality follows from $\log(1 + x) \le x$. □

Proof of Theorem 2. We first control the small repetition regime. Under N,

$$P_N(Y_1 = y_1, \ldots, Y_r = y_r) = 2^{-r}.$$

After marginalizing the latent uniform sign under E,

$$P_E(y_{1:r}) = 2^{-r-1}\left[\prod_{j=1}^{r}(1 + qy_j) + \prod_{j=1}^{r}(1 - qy_j)\right].$$

Therefore, the likelihood ratio $L = dP_E/dP_N$ is

$$L(y_{1:r}) = \frac{1}{2}\left[\prod_{j=1}^{r}(1 + qy_j) + \prod_{j=1}^{r}(1 - qy_j)\right].$$

Under P_N, let

$$A_+ = \prod_{j=1}^{r}(1 + qY_j), \qquad A_- = \prod_{j=1}^{r}(1 - qY_j).$$

Direct calculation gives

$$E_N[A_+^2] = E_N[A_-^2] = (1 + q^2)^r$$

and

$$E_N[A_+A_-] = (1 - q^2)^r.$$

Therefore,

$$\chi^2(P_E^{(r)}\|P_N^{(r)}) = \frac{1}{2}\left[(1 + q^2)^r + (1 - q^2)^r\right] - 1.$$

Writing $x = rq^2$ and using

$$(1 + q^2)^r \le e^x, \qquad (1 - q^2)^r \le e^{-x},$$

we obtain

$$\chi^2(P_E^{(r)}\|P_N^{(r)}) \le \cosh x - 1.$$

If $x \le 1$, then

$$\cosh x - 1 \le \frac{\cosh 1}{2} x^2.$$

Therefore, the two preceding lemmas imply

$$C(P_N^{(r)}, P_E^{(r)}) \le \frac{\cosh 1}{2} r^2 q^4,$$

and consequently

$$\Gamma_q(r) \le \frac{\cosh 1}{2} r q^4 \qquad \text{whenever } rq^2 \le 1.$$

We next construct a benchmark at the scale q^{-2}. Set

$$r_0 = \left\lceil \frac{64}{q^2} \right\rceil$$

and let

$$S = \sum_{j=1}^{r_0} Y_j, \qquad t = \frac{r_0 q}{2}.$$

Under N, Lemma 1 gives

$$P_N(|S| \ge t) \le 2e^{-r_0 q^2/8} \le 2e^{-8}.$$

Under E, conditional on the latent sign Z,

$$ZS = \sum_{j=1}^{r_0} \xi_j,$$

so

$$P_E(|S| \ge t) \ge 1 - e^{-8}.$$

It follows that

$$\mathrm{TV}(P_N^{(r_0)}, P_E^{(r_0)}) \ge 1 - 3e^{-8}.$$

Marginally over the balanced rule sign A, the vector $Y_{1:r_0}$ has the same law under R as under E, so the same event gives

$$\mathrm{TV}(P_N^{(r_0)}, P_R^{(r_0)}) \ge 1 - 3e^{-8}.$$

For E versus R, use

$$B = \{AS \ge t\}.$$

Under R,

$$AY_j = \xi_j,$$

and this gives

$$P_R(B) \ge 1 - e^{-8}.$$

Under E, A is independent of S and is uniform, so

$$P_E(B) = \frac{1}{2} P_E(S \ge t) + \frac{1}{2} P_E(S \le -t) \le \frac{1}{2}.$$

Therefore,

$$\mathrm{TV}(P_E^{(r_0)}, P_R^{(r_0)}) \ge \frac{1}{2} - e^{-8}.$$

Let

$$v = \frac{1}{2} - e^{-8} > 0.$$

Every pairwise total variation distance is at least v.

The Bhattacharyya coefficient satisfies

$$\mathrm{TV}(P, Q) \le \sqrt{1 - \mathrm{BC}(P, Q)^2}.$$

Therefore,

$$\mathrm{BC}(P, Q) \le \sqrt{1 - v^2}.$$

Since $s = 1/2$ is admissible in the Chernoff coefficient,

$$C(P, Q) \geq -\log \mathrm{BC}(P, Q) \geq c_0, \qquad c_0 = -\frac{1}{2}\log(1 - v^2) > 0.$$

It follows that

$$\Gamma_q(r_0) \geq \frac{c_0}{r_0}.$$

For $q \leq 1/2$,

$$r_0 \leq \frac{257}{4q^2},$$

so

$$\Gamma_q(r_0) \geq aq^2, \qquad a = \frac{4c_0}{257} > 0.$$

We can now exclude repetition counts that are too small. Let

$$K = \frac{\cosh 1}{2}$$

and choose

$$c = \min\left\{1, \frac{a}{2K}\right\}.$$

If

$$r \leq \frac{c}{q^2},$$

then $rq^2 \leq 1$ and

$$\Gamma_q(r) \leq Krq^4 \leq Kcq^2 \leq \frac{a}{2}q^2 < \Gamma_q(r_0).$$

Therefore, no maximizer can satisfy $r \leq c/q^2$.

It remains to exclude excessive repetition. Under R, the joint law is $P_R(A, Y_{1:r})$. Under E, A is independent of $Y_{1:r}$, while the marginal law of $Y_{1:r}$ is the same in the two worlds because the latent sign is uniform in both cases. Therefore,

$$D_{\mathrm{KL}}(P_R^{(r)} \| P_E^{(r)}) = I_R(A; Y_{1:r}) \leq H(A) = \log 2.$$

Chernoff information is bounded by KL, so

$$\Gamma_q(r) \leq \frac{\log 2}{r}.$$

Choose

$$C = \frac{2\log 2}{a}.$$

If

$$r \geq \frac{C}{q^2},$$

then

$$\Gamma_q(r) \leq \frac{\log 2}{C}q^2 = \frac{a}{2}q^2 < \Gamma_q(r_0).$$

Therefore, no maximizer can satisfy $r \geq C/q^2$.

Combining the two exclusions proves

$$\frac{c}{q^2} < r^\star(q) < \frac{C}{q^2}.$$

□

D SEQUENTIAL PROMOTION DETAILS

For completeness, we prove Proposition 1 and record its fixed-sample converse. For any promotion event A_n satisfying

$$P_{\mathsf{E}}(A_n) \le \delta, \qquad P_{\mathsf{R}_g}(A_n) \ge 1 - \beta,$$

data processing implies

$$n \ge \frac{d_{\mathrm{bin}}(1-\beta\|\delta)}{I(q)}.$$

Proof of Proposition 1. Under E,

$$E_{\mathsf{E}}\left[1 + qY_n g(X_n) \mid \mathcal{F}_{n-1}, X_n\right] = 1,$$

so $L_n(g)$ is a nonnegative mean one martingale. For

$$b_g = \frac{1}{\delta\pi(g)},$$

optional stopping at the first crossing of b_g, truncated at any finite horizon, gives

$$P_{\mathsf{E}}\left(\sup_n L_n(g) \ge b_g\right) \le \delta\pi(g).$$

Summing over g proves the familywise bound. The same argument applies verbatim with any fixed $\widetilde{q} \in (-1, 1)$ because under E, $E[Y_n g(X_n) \mid \mathcal{F}_{n-1}, X_n] = 0$, so each factor $1 + \widetilde{q}Y_n g(X_n)$ has conditional mean one. Under R_g, its expected log increment is $I(q, \widetilde{q})$; differentiating in $\widetilde{q}$ shows that this drift is maximized at $\widetilde{q} = q$.

Under the true world R_g,

$$Y_i g(X_i) = \xi_i,$$

so

$$\log L_n(g) = \sum_{i=1}^{n} W_i, \qquad W_i = \log(1 + q\xi_i).$$

These variables are iid and have mean

$$E[W_i] = I(q) > 0.$$

The strong law gives

$$\frac{1}{n}\log L_n(g) \to I(q) \qquad \text{almost surely.}$$

The standard positive drift hitting time argument then yields

$$\frac{\tau_g}{K(g) + \log(1/\delta)} \to \frac{1}{I(q)} \qquad \text{almost surely.}$$

Differentiating I gives

$$I'(q) = \frac{1}{2}\log\frac{1+q}{1-q} = \operatorname{arctanh} q,$$

and integrating the power series of $\operatorname{arctanh} q$ gives

$$I(q) = \frac{q^2}{2} + \frac{q^4}{12} + O(q^6).$$

For the fixed-sample converse stated above, the one-sample KL divergence between R_g and E is $I(q)$, so the n sample divergence is $nI(q)$. Data processing through the promotion event A_n gives

$$nI(q) \ge d_{\mathrm{bin}}\left(P_{\mathsf{R}_g}(A_n)\|P_{\mathsf{E}}(A_n)\right) \ge d_{\mathrm{bin}}(1-\beta\|\delta),$$

which proves the claim. □

E Additional Experimental Results

The controlled simulation study evaluates rule families through $M \leq 256$; larger families with $M = 512$ and 1024 are not included. Under uniform replication, the median finite-range slope of B_{80}^{minimax} against q on logarithmic axes is -1.89, while the sparse allocation produces shallower slopes over the tested range. We use the simulation to assess the joint complexity law rather than to estimate an exact finite-range noise exponent. Normalizing the budget by $z = B\alpha_{\min}q^2/(\log M + \log(1/\delta))$ reduces cross-configuration prediction error modestly, with normalized-to-raw RMSE ratio 0.906. In the singleton regime, equal-prior mean N/E correctness is 0.491, consistent with exact indistinguishability.

Figure 5 shows the domain-level CLINC contrast underlying the aggregate result in Section 7. Every domain improves when the allocation moves from $T = 10$ to $T = 60$ at fixed $B = 120$ and $\eta = 0$.

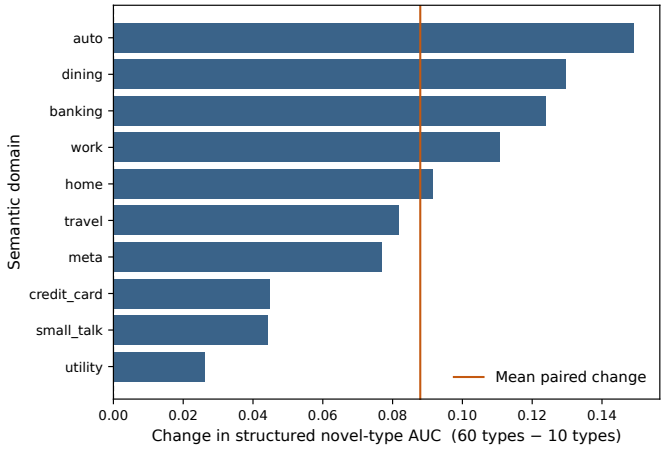


Figure 5: **CLINC150 domain-level diversity effects.** Change in structured novel-type AUC from 10 to 60 observed intents at $B = 120$ and $\eta = 0$. All ten semantic domains have positive effects.

Figure 6 records the global type-count intervention that motivates the targeted K_+ experiment. At $B = 120$ and $\eta = 0$, structured novel-type AUC changes from 0.665 at $T = 10$ to 0.631 at $T = 120$, a paired effect of -0.035 (90% CI $[-0.042, -0.028]$), with 2/6/42 positive/null/negative alphabets under a ± 0.01 threshold. Mean target-alphabet support changes only from 1.00 to 2.36, while drawings per observed character fall from 12 to 1.

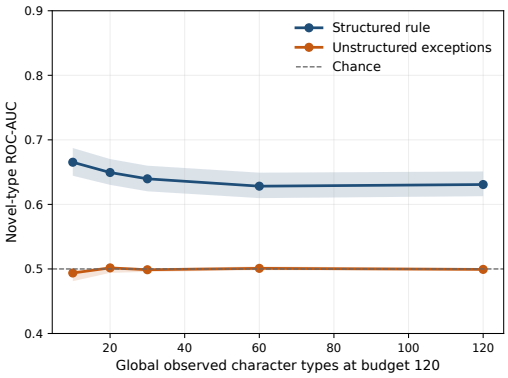


Figure 6: **Global type count is a weak intervention on target-alphabet support.** Increasing total observed characters at fixed budget reduces structured transfer because relevant support grows slowly while repetition per character falls sharply.