TOPOLOGY AND SHARED ERROR EXPOSURE IN LARGE LANGUAGE MODEL LINEAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

As the population of large language models (LLMs) rapidly grows, understanding how shared training sources shape errors across the LLM ecosystem and how developers' training choices create these exposures becomes an important system level question. We develop a lineage representation that maps directed training relationships into error loadings and decomposes model error covariance into three components arising from common tasks, shared base families, and idiosyncratic errors propagated through the lineage. This decomposition shows when concentrated source exposure sustains aggregate error variance as the model population expands. At fixed individual error distributions, different patterns of shared exposure can yield different aggregate errors. We illustrate the mechanism through three canonical lineage topologies. To examine how individual choices shape concentration, we study a single generation teacher choice model and derive an explicit condition under which every developer prefers the same teacher, while a diversified teacher assignment reduces aggregate error.

1 INTRODUCTION

Large language models (LLMs) are increasingly developed within an interconnected ecosystem, where many models share pretrained foundations (Bommasani et al., 2021). Developers can adapt an existing checkpoint or train a student using a teacher's predictions and generated responses (Hinton et al., 2015; Mukherjee et al., 2023; Gu et al., 2024; Xu et al., 2024). Ecosystem Graphs and HuggingGraph document these dependencies among models, datasets and applications (Bommasani et al., 2024; Rahman et al., 2025). Alongside this structural dependence, empirical studies find substantial overlap in LLM errors and show how model similarity can affect evaluation and supervision (Kim et al., 2025; Goel et al., 2025). As LLMs become widely used in decision making, reliance on shared models and training components raises concerns about algorithmic monoculture (Kleinberg & Raghavan, 2021), where different decision makers can repeatedly produce adverse outcomes for the same individuals (Bommasani et al., 2022; 2026; Peng & Garg, 2024). One mechanism behind such monoculture is error inheritance along model lineages. Errors originating from a shared model can enter the behavior of its descendants (Bommasani et al., 2021; Ojha et al., 2023). Taken together, we ask three questions

- How can lineage topology be represented to track error transmission across models?
- How does lineage topology endogenously emerge from developers' training choices?
- When do inherited errors persist after aggregation when the model population grows?

1.1 OUR CONTRIBUTIONS

We develop a lineage representation which connects directed training relationships to joint model errors. The matrix Φ accumulates exposure along training paths and yields an exact covariance decomposition into common task, shared base family and propagated model specific components. We characterize when propagated variance vanishes as the population grows and when concentrated exposure makes it persist.

Then we investigate how developers' teacher choices shape lineage topology. In a single generation, each developer chooses from a fixed teacher bank to minimize individual error, while aggregate

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.

error also depends on the overlap between students' inherited errors. To isolate the tradeoff between teacher quality and shared exposure, we examine a choice between two teachers and derive an explicit condition under which all developers strictly prefer the same teacher, yet every optimal population assignment uses both. By leveraging the condition, we quantify the trade-off between reducing shared exposure and incurring higher individual error from a suboptimal teacher. Finally, we design synthetic regression experiments illustrate this tradeoff and examine how training transmits errors and how those errors aggregate across lineage topologies.

1.2 Related Work

Our work builds on studies of knowledge distillation that examine knowledge transfer, teacher–student agreement and learning across generations (Phuong & Lampert, 2019; Stanton et al., 2021; Furlanello et al., 2018; Mobahi et al., 2020). Related work investigates model provenance through output similarities, functional representations and tests of independent training (Yax et al., 2025; Wu et al., 2026; Zhu et al., 2025). We take training relationships as given and examine how inherited errors shape population performance.

Our representation draws on graph signal processing (Shuman et al., 2013; Sandryhaila & Moura, 2013). In particular, Seifert et al. (2023) develop a source representation on DAGs through weighted transitive closures, while Sullivant et al. (2010) connect graphical paths to covariance rank constraints. Network models relate information aggregation to the distribution of influence (DeGroot, 1974; Golub & Jackson, 2010) and show how production networks transmit local shocks into aggregate fluctuations (Acemoglu et al., 2012). We apply the accumulated path representation to training errors and characterize when concentrated ancestral exposure sustains aggregate variance.

Classical work relates aggregate error to individual performance and diversity (Krogh & Vedelsby, 1994; Brown et al., 2005), and Wood et al. (2023) develop decompositions for multiple losses and combination rules. Our analysis studies how training relationships generate the error dependencies entering aggregation and determine the gains from adding models.

Finally, endogenous network formation studies how individual linking incentives shape network structure and efficiency (Jackson & Wolinsky, 1996; Bala & Goyal, 2000). Our teacher choice model places this question within model development. Developers choose teachers according to individual accuracy, while aggregate error also depends on shared ancestral exposure. We characterize when these objectives lead to concentrated individual choices and a diversified population optimum.

2 Set up and lineage representation

Consider M models developed sequentially and evaluated on a common distribution of tasks. We use $e_i = f_i(x) - y$ to denote model i's signed scalar prediction error, with mean $b_i = \mathbb{E}[e_i]$ and mean squared error $L_i = \mathbb{E}\left[e_i^2\right]$. For the model population, we define the error vector $e = (e_1, \ldots, e_M)^\top$, mean vector $b = (b_1, \ldots, b_M)^\top$ and covariance matrix $\Sigma = \mathrm{Cov}(e)$.

DAGs. To capture the intricate training relationships, we define a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, M\}$ is the set of models and $(j, i) \in \mathcal{E}$ indicates model i learns from model j. Teachers precede their students, so $\mathcal{G}$ is a directed acyclic graph (DAG). We index its nodes in a topological order with $\mathcal{E} \subseteq \{(j, i) \in \mathcal{V} \times \mathcal{V} : j < i\}$ and let $\mathcal{P}_i = \{j : (j, i) \in \mathcal{E}\}$ denote model i's teacher set. For example, in a roots topology, $\mathcal{P}_i = \varnothing$ for every i. For a star topology rooted at model 1, $\mathcal{P}_1 = \varnothing$ and $\mathcal{P}_i = \{1\}$ for $i \geq 2$. In a chain topology, $\mathcal{P}_1 = \varnothing$ and $\mathcal{P}_i = \{i - 1\}$ for $i \geq 2$ [1].

Ancestry. Since the training lineage dictates how prior errors propagate to each model, we denote $j \preceq i$ if $j = i$ or there is a directed path from j to i by following the DAG and poset representation from Seifert et al. (2023). The relation is reflexive, antisymmetric and transitive, so $(\mathcal{V}, \preceq)$ is a partially ordered set. Let $\mathcal{A}(i) = \{j : j \preceq i, j \neq i\}$ denote the ancestors of model i[2]. For instance, in a chain topology, $\mathcal{A}(i) = \{1, \ldots, i - 1\}$ although each student has only one direct teacher.

[1] More generally, $|\mathcal{P}_i| > 1$ means a student learns from several teachers.

[2] The set $\mathcal{P}_i$ records direct training relationships, while $\mathcal{A}(i)$ includes earlier models connected through intermediate students.

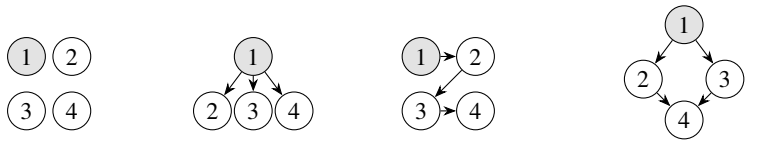

Figure 1: Training paths and accumulated exposure. The first three graphs show the canonical roots, star and chain topologies, while the fourth shows a student with multiple teachers. Each edge has loading a. Beneath each graph, $n_1 = \sum_i \Phi_{i1}$ gives the total exposure to the shaded root.

For each fixed training configuration, we associate every edge $(j, i) \in \mathcal{E}$ with a nonnegative direct error loading A_{ij} and set $A_{ij} = 0$ whenever $j \notin \mathcal{P}_i$. Under the topological ordering, the weighted adjacency matrix is strictly lower triangular,

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ A_{21} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ A_{M1} & \cdots & A_{MM-1} & 0 \end{pmatrix}.$$

We track the evolution of centered errors through the following affine recursion [3]

$$e_i - b_i = \sum_{j \in \mathcal{P}_i} A_{ij} \, (e_j - b_j) + \tilde{u}_i, \quad \mathbb{E}\,[\tilde{u}_i] = 0. \tag{1}$$

The local term $\tilde{u}_i$ denotes the error introduced at node i. Due to shared training sources, these local errors are generally correlated. Letting $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_M)^\top$, we define their covariance matrix as $\Omega = \mathrm{Cov}(\tilde{u}, \tilde{u})$. Applying the recursion along the lineage, following proposition collects these accumulated exposures into the lineage matrix Φ.

Proposition 1 (Lineage representation)**.** *The recursion in Equation* (1) *has the unique solution*

$$e - b = \Phi \tilde{u}, \qquad \Phi = (I - A)^{-1} = I + A + \cdots + A^{M-1}. \tag{2}$$

For $j < i$, Φ_{ij} is the sum of products of direct loadings over all directed paths from j to i. Also, $\Phi_{ii} = 1$ and $\Phi_{ij} = 0$ whenever $j \notin \mathcal{A}(i) \cup \{i\}$. For each power A^ℓ, it collects paths of length ℓ and their sum gives accumulated exposure[4]. Figure 1 illustrates this accumulation with a common edge loading under three canonical lineage structures and a multi-parent setting, highlighting a direct link to an early teacher contributes one path, while intermediate students contribute additional paths.

3 Joint error structure and diversification

Lineage representation describes how errors are inherited through training relationships. We next decompose local errors and derive their contribution to population error so that we can separate the error from the exposure created by the lineage.

3.1 Error decomposition and aggregation

Let $k_i > 0$ be the effective compute devoted to local training at node i. Motivated by power law learning curves (Lin et al., 2024), we specify the variance of its model specific local component as

$$s_i^2(k_i) = \sigma_i^2 + a_i k_i^{-\eta}, \qquad \sigma_i^2 \geq 0, \quad a_i > 0, \quad \eta \in (0, 1). \tag{3}$$

[3] Appendix B.1 derives the error recursion from squared loss training with whitened features and ridge regularization, linking the transmission loadings and local error terms to the training objective.

[4] In Appendix B.2, we further examine deviations from this recursion and bounds their effect on aggregate error through the accumulated exposure matrix Φ.

Here σ_i^2 is the persistent local variance, a_i sets the scale of the component reduced by compute, and η determines its rate of decline. For notation simplicity, we use $s_i = s_i(k_i)$ for analysis. Models with shared provenance inherit similar error patterns (Zang et al., 2026; Bommasani et al., 2021), so we represent task, family and model specific local components by

$$\tilde{u}_i = \sigma_c q_i z + \tau_i \zeta_{g(i)} + s_i \xi_i, \qquad \sigma_c, \tau_i \geq 0. \tag{4}$$

For parameters, σ_c measures common task variation, $q_i \in \mathbb{R}$ is its local loading and τ_i is model i's local family loading. For the error sources, z represents variation across evaluation items shared by the population, $\zeta_{g(i)}$ is the error direction associated with model i's base family $g(i)$ and ξ_i shows its own local error direction. Each factor has zero mean and unit variance. Model specific factors ξ_i are mutually uncorrelated and uncorrelated with z and all family factors. Moreover, the task factor is uncorrelated with the family factors. Across families, let $\rho_{gh} = \mathrm{Cov}(\zeta_g, \zeta_h)$ denote the entries of a positive semidefinite correlation matrix with $\rho_{gg} = 1$. Hence,

$$\Omega_{ij} = \sigma_c^2 q_i q_j + \tau_i \tau_j \rho_{g(i)g(j)} + s_i^2 \mathbf{1}_{\{i=j\}}.$$

Shared provenance enters Ω, while direct teacher relationships enter A. Propagation turns these local loadings into final model exposures. Denote $q = (q_1, \ldots, q_M)^\top$ and define

$$c = \Phi q, \qquad (\phi_g)_i = \sum_{j, g(j)=g} \Phi_{ij} \tau_j, \qquad \Lambda = \mathrm{diag}(s_1^2, \ldots, s_M^2). \tag{5}$$

Here the entry c_i is model i's accumulated task loading. The vector ϕ_g records each model's accumulated exposure to family g, including exposure inherited through intermediate teachers. The diagonal matrix Λ collects the variances of the model specific local components. When models are combined into an ensemble, their exposures to each shared source add together, linking the lineage structure to aggregate error. Given aggregation weights $w = (w_1, \ldots, w_M)^\top$ with $w_i \geq 0$ and $\sum_i w_i = 1$, we define the aggregate error $\bar{e}(w) = w^\top e$ and its mean squared error as $L(w) = \mathbb{E}[\bar{e}(w)^2]$. The following theorem reveals how local errors influence the covariance and ensemble performance via aggregation weights.

Theorem 2 (Covariance decomposition and aggregate error)**.** *Under Equation* (4)*,*

$$\Sigma = \Phi \Omega \Phi^\top = \sigma_c^2 c c^\top + \sum_{g,h} \rho_{gh} \phi_g \phi_h^\top + \Phi \Lambda \Phi^\top. \tag{6}$$

The aggregate mean squared error is

$$L(w) = (w^\top b)^2 + (\Phi^\top w)^\top \Omega (\Phi^\top w), \tag{7}$$

and its variance is

$$\mathrm{Var}(\bar{e}(w)) = \sigma_c^2 (w^\top c)^2 + \sum_{g,h} \rho_{gh} (w^\top \phi_g)(w^\top \phi_h) + \sum_j s_j^2 (\Phi^\top w)_j^2. \tag{8}$$

The three covariance terms correspond to shared task variation, family exposure and the transmission of model specific errors. A teacher passes its family component and its own local component to descendants, so both can appear in several models. Family contributions depend on ρ_{gh} because errors associated with different families can move together. When distinct family factors are uncorrelated, the middle covariance term becomes $\sum_g \phi_g \phi_g^\top$. In the aggregate, source j receives loading $(\Phi^\top w)_j = \sum_i w_i \Phi_{ij}$.

Remark (Common task loading). Equal final task exposure means $c = \mathbf{1}$, which corresponds to $q = (I - A)\mathbf{1}$. At node i, the inherited task loading is $\sum_j A_{ij}$ and the local loading is $q_i = 1 - \sum_j A_{ij}$, which means they give unit final loading. The vector q thus keeps track of the part introduced locally.

3.2 System risk

To study whether errors diversify as the model population grows, we consider the equally weighted aggregate $\bar{e}_M = M^{-1} \sum_{i=1}^M e_i$, with mean squared error $L_M = \mathbb{E}[\bar{e}_M^2]$. Each local source j enters this average with loading $M^{-1} \sum_i \Phi_{ij}$. Therefore, the behavior of aggregate variance depends on how these source loadings change with population size. For the equally weighted population, define

$$n_j = \sum_i \Phi_{ij}, \qquad N_g = \sum_{j, g(j)=g} n_j, \qquad H = \sum_j \left(\frac{n_j}{M}\right)^2, \qquad H_{\mathcal{F}} = \sum_g \left(\frac{N_g}{M}\right)^2. \tag{9}$$

The quantity n_j is the total population loading on local source j, and N_g combines these loadings within family g. Their squared exposure indices H and $H_{\mathcal{F}}$ record how strongly the population average depends on individual sources and families. Under homogeneous scales $s_j = s$ and $\tau_j = \tau$, Theorem 2 gives

$$\mathrm{Var}(\bar{e}_M) = \sigma_c^2 \bar{c}_M^2 + \frac{\tau^2}{M^2} \sum_{g,h} \rho_{gh} N_g N_h + s^2 H, \qquad \bar{c}_M = \frac{1}{M} \sum_i c_i. \tag{10}$$

Here $\bar{c}_M$ is the population's average task loading. For uncorrelated family factors, the family term reduces to $\tau^2 H_{\mathcal{F}}$. Positive correlations add to the covariance between family exposures, while negative correlations provide offsetting contributions. The model specific term is governed by H in both cases. A natural question here is when the contribution of model specific errors vanishes as the population grows. The following proposition shows this contribution vanishes exactly when the exposure index H tends to zero under uniformly bounded positive local variances.

Proposition 3 (Diversification of propagated errors)**.** *Let $V_M = M^{-2} \mathbf{1}^\top \Phi \Lambda \Phi^\top \mathbf{1}$ denote the aggregate variance contributed by model specific local errors. Suppose variances satisfy $0 < \underline{s}^2 \le s_j^2 \le \overline{s}^2 < \infty$ uniformly across sources and population sizes, where $\underline{s}^2$ and $\overline{s}^2$ are fixed bounds. Then*

$$\underline{s}^2 H \le V_M = \sum_j s_j^2 (\frac{n_j}{M})^2 \le \overline{s}^2 H, \tag{11}$$

then $V_M \to 0$ exactly when $H \to 0$. Furthermore,

$$\Big(\frac{\max_j n_j}{M}\Big)^2 \le H \le \Big(\frac{\max_j n_j}{M}\Big) \sum_j \frac{n_j}{M}. \tag{12}$$

Under the condition $\sum_j n_j = O(M)$, then $V_M \to 0$ exactly when $\max_j n_j = o(M)$.

The proposition shows when population growth diversifies propagated errors. Under the stated variance bounds and bounded average source exposure, the variance contributed by propagated model specific errors vanishes exactly when the largest total loading on any source grows more slowly than M. If a source's total loading remains proportional to M, its contribution to aggregate variance persists. We call this persistence lineage induced **systemic risk** when $\liminf_{M\to\infty} V_M > 0$.

Furthermore, since $V_M = \sum_j s_j^2(k_j)(n_j/M)^2$, so this contribution depends on both the variance introduced at each source and its exposure in the population average. Additional compute can reduce the local variance through the scaling relation, and changes in lineage structure alter the exposure weights. So far, we capture both effects when the population grows by Equation (11).

To further isolate the role of paths, here we consider centered errors with independent local innovations of common variance $s^2 > 0$ and direct edge loading $0 < a < 1$. In this benchmark, $b = 0$ and $\Omega = s^2 I$. Across the canonical topologies, the error specialize as: independent roots satisfy $e_i = \tilde{u}_i$; the star topology satisfies $e_1 = \tilde{u}_1$ and $e_i = a e_1 + \tilde{u}_i$ for $i \ge 2$; the chain topology meets $e_1 = \tilde{u}_1$ and $e_i = a e_{i-1} + \tilde{u}_i$ for $i \ge 2$.

Proposition 4 (Topology and aggregation limits)**.** *The aggregate variances in this benchmark are*

$$V_M^{\mathrm{Roots}} = \frac{s^2}{M}, \tag{13}$$

$$V_M^{\mathrm{Star}} = \frac{s^2}{M^2}\big([1 + (M-1)a]^2 + M - 1\big) \longrightarrow s^2 a^2, \tag{14}$$

$$V_M^{\mathrm{Chain}} = \frac{s^2}{M^2} \sum_{k=1}^{M} \left(\frac{1-a^k}{1-a}\right)^2 \sim \frac{s^2}{M(1-a)^2}. \tag{15}$$

The proposition explains why the three topologies have different diversification limits. Roots contribute separate local errors, while a star repeatedly transmits the same root error, making its total loading proportional to M. Along a chain, transmission attenuates with distance, keeping each source's total loading bounded. Appendix B.3 extends this comparison to convex distillation, where stronger teacher exposure also reduces the student's local error loading.

3.3 Error directions and individual evaluation

Persistent aggregate variance raises a related question about the joint error structure as the population grows. Many models can remain exposed to a few dominant error directions. To describe these directions, denote $\Sigma = \sum_{k=1}^{M} \lambda_k v_k v_k^\top$, where the orthonormal eigenvectors v_k identify directions of variation across model errors and the eigenvalues $\lambda_k \geq 0$ measure their variances. Following Bunea & Xiao (2015), we measure the concentration of this variation by the effective rank

$$r(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)}, \qquad \mathrm{tr}(\Sigma) > 0,$$

where $\lambda_{\max}(\Sigma)$ is the largest covariance eigenvalue. This ratio compares total error variance with the variance along the strongest direction. It equals M when all directions have equal positive variance and approaches one when a single direction accounts for most of the variation. The following proposition connects these error directions to aggregate variance and bounds effective rank in terms of individual and aggregate variances.

Proposition 5 (Joint error directions)**.** *Aggregate variance can be expressed as* $\mathrm{Var}(\bar{e}(w)) = \sum_k \lambda_k (w^\top v_k)^2$*. Since* $\Sigma = (\Phi\Omega^{1/2})(\Phi\Omega^{1/2})^\top$*, its eigenvalues are the squared singular values of* $\Phi\Omega^{1/2}$*, with* $\Omega^{1/2}$ *denoting the positive semidefinite square root of* Ω*. For* $\mathrm{Var}(\bar{e}_M) > 0$*,*

$$r(\Sigma) \leq \frac{M^{-1} \sum_i \mathrm{Var}(e_i)}{\mathrm{Var}(\bar{e}_M)}. \tag{16}$$

Proposition 5 links the persistence of aggregate error to the concentration of variation across models. Each error direction contributes according to its variance λ_k and its projection onto the aggregation weights measured by $(w^\top v_k)^2$. For the equally weighted population, Equation (16) shows that effective rank remains bounded whenever average individual variance stays bounded and aggregate variance stays above a positive constant. Thus a growing population can contain more models while its errors remain concentrated in a limited effective dimension.

The preceding analysis suggests that individual performance alone can leave differences in aggregate error unexplained. To isolate the role of shared exposure, consider two population configurations with the same individual error distributions. Their individual variances and aggregate mean coincide, so any difference in aggregate mean squared error comes entirely from cross model covariances

$$\Delta L_M = \Delta \,\mathrm{Var}(\bar{e}_M) = \frac{1}{M^2} \sum_{i \neq j} \Delta \Sigma_{ij}, \tag{17}$$

where Δ denotes the difference between the two configurations. Even with the same individual performance, a population has higher aggregate error when its models' errors are more strongly correlated. Teacher choices affect both individual errors and these correlations. Developers choose teachers to improve their own models, while the population also benefits from diversifying shared error exposure. We next examine when this difference leads developers to concentrate on a common teacher even though a diversified assignment achieves lower aggregate error.

4 Endogenous lineage topology

The preceding analysis takes the training graph as given, a natural following up question is how developers' teacher choices determine its edges. We consider a single generation of M evaluated students with a common feasible teacher set $\mathcal{P}$ consisting of earlier models. Student i chooses $p_i \in \mathcal{P}$, giving its direct teacher set $\mathcal{P}_i = \{p_i\}$. The earlier training graph remains fixed throughout the comparison.

4.1 Training choices and error exposure

For a candidate teacher p, the direct recursion gives

$$e_i(p) - b_i(p) = A_{ip}(e_p - b_p) + \tilde{u}_i(p), \tag{18}$$

where $b_i(p) = \mathbb{E}[e_i(p)]$ and $\tilde{u}_i(p)$ is the centered local error under that training choice. Its task, family and model specific components follow Section 3, with effective compute and local loadings allowed to depend on the selected teacher. The student's mean squared error is

$$L_i(p) = b_i(p)^2 + A_{ip}^2 \Sigma_{pp} + 2A_{ip} \operatorname{Cov}(e_p, \tilde{u}_i(p)) + \operatorname{Var}(\tilde{u}_i(p)), \tag{19}$$

where Σ_{pp} is the variance of teacher p's error. The target of each developer is to choose a teacher minimizing $L_i(p)$. Meanwhile, the population objective includes covariances across students and follows Equation (7). To obtain an explicit comparison, we start from specializing a common transmission coefficient $A_{ip} = a$, with $0 < a \leq 1$, and common local variance $s_i^2(k_i) = s^2$. Suppose teachers and students have the same final task loading one and family loading τ on a common base family g_0. Then the resulting student error is

$$e_i(p) = b(p) + \sigma_c z + \tau \zeta_{g_0} + a \sum_j \Phi_{pj} s_j \xi_j + s\xi_i. \tag{20}$$

$b(p)$ is common across students choosing teacher p and the sum runs over earlier models, where Φ_{pj} records teacher p's exposure to local source j. This specification holds common task and family exposure fixed while teacher choice changes the inherited model specific errors. Students choosing the same teacher share its entire inherited component, and choosing different teachers can still share errors through overlapping ancestry. Therefore, their assignments determine both the new training edges and exposure to earlier error sources. Appendix B.4 embeds the student cohort and its earlier teachers in an augmented DAG, which connects the choice model to the general lineage representation.

To compare individual teacher choices with population assignments, we need both the error associated with each teacher and the overlap between errors associated with different teachers. In Equation (20), the component that varies with teacher choice is $b(p) + a \sum_j \Phi_{pj} s_j \xi_j$. Taking the cross second moments of these components gives

$$Q_{pq} = b(p)b(q) + a^2 \sum_j \Phi_{pj} \Phi_{qj} s_j^2, \qquad p, q \in \mathcal{P}. \tag{21}$$

The diagonal entry Q_{pp} gives the mean squared error contributed by choosing teacher p, while Q_{pq} records the error overlap between students choosing teachers p and q. These cross moments enter the population objective according to how many students follow each teacher. Let N_p count students choosing teacher p and define their share $\alpha_p = N_p/M$. In the population average, the error component associated with teacher p receives weight α_p, so its cross moment with the component associated with teacher q receives weight $\alpha_p \alpha_q$. Summing over teacher pairs gives $\sum_{p,q} \alpha_p \alpha_q Q_{pq} = \alpha^\top Q \alpha$, where $\alpha = (\alpha_p)_{p \in \mathcal{P}}$ and $Q = (Q_{pq})_{p,q \in \mathcal{P}}$. Adding the common task and family components and the students' local errors yields the following objectives.

Proposition 6 (Individual and population objectives)**.** *Under Equation* (20)*,*

$$L_i(p) = \sigma_c^2 + \tau^2 + Q_{pp} + s^2, \qquad L_M = \sigma_c^2 + \tau^2 + \alpha^\top Q \alpha + \frac{s^2}{M}. \tag{22}$$

Developers choose teachers with the smallest Q_{pp}, while a planner chooses assignment shares minimizing $\alpha^\top Q \alpha$. The planner also accounts for the cross moments Q_{pq}. Hence the population value of a teacher depends on both its individual quality and its error overlap with the other selected teachers. The different objectives can lead all developers to choose one teacher even when the population benefits from using several. The following theorem gives the exact condition for a two teacher set $\mathcal{P} = \{p, q\}$.

Theorem 7 (Excess teacher concentration)**.** *Let $M \geq 2$ and suppose $Q_{pp} < Q_{qq}$ under Equation* (20)*. Every developer strictly chooses teacher p. Every population optimum among assignments using these two teachers uses both teachers exactly when*

$$0 < Q_{qq} - Q_{pp} < 2(M-1)(Q_{pp} - Q_{pq}). \tag{23}$$

Starting from the assignment in which all students choose p, moving one student to q changes aggregate error by

$$\Delta L_M = \frac{Q_{qq} - Q_{pp} - 2(M-1)(Q_{pp} - Q_{pq})}{M^2}. \tag{24}$$

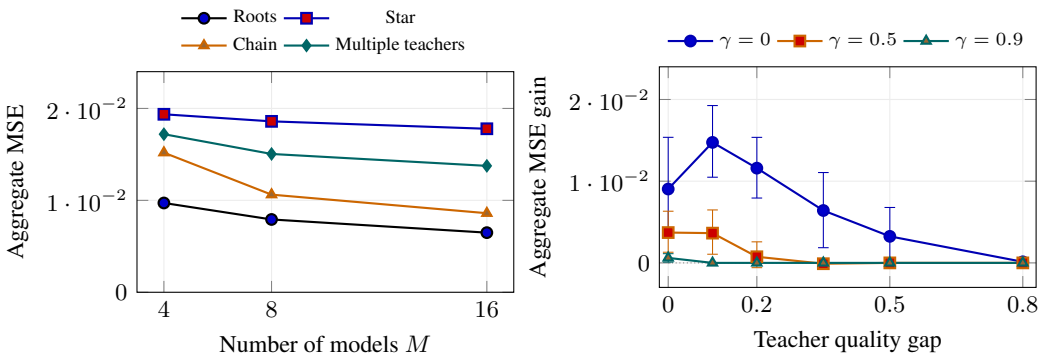

Figure 2: Population growth and teacher choice with neural learners across eight seeds. Left panel shows mean test aggregate MSE at $\beta = 0.7$. Right panel shows the paired gain from population choice relative to individual choice. Assignments are selected on validation data and evaluated on test data. Bars give descriptive pointwise 95% seed bootstrap intervals.

The switching student incurs an individual cost $Q_{qq} - Q_{pp}$. Its error overlap with each of the other $M - 1$ students falls by $Q_{pp} - Q_{pq}$. The condition states that the combined reduction in these cross terms outweighs the individual cost. Private choices then form a star, while the population optimum forms two teacher groups.

The cross moment Q_{pq} links this tradeoff to lineage structure. Shared ancestral exposure limits diversification and sufficiently distinct inherited errors can make a less accurate teacher valuable to the population. Appendix B.5 extends the assignment analysis to a larger teacher bank, deriving optimal shares and the effect of reallocating one student. Furthermore, Appendix B.6 incorporates autonomous training and compute costs to examine how access to a teacher changes both inherited error and local learning.

5 Empirical results

We examine proposed theory through synthetic regression experiments with ridge learners and neural networks. We first test how errors propagate through a fixed DAG, then study how aggregate error changes as the population grows and finally compare teacher assignments chosen under individual and population objectives. Each comparison uses the same simulated data and training seeds across conditions, with eight independent repetitions. Appendix C reports the design, metrics and full parameter grid.

Error transmission. We examine how a perturbation to a root's training target reaches its descendants along a fixed DAG. Loadings estimated from one perturbation direction are used to predict responses to another direction and to the reversed perturbation. Ridge training provides an exact benchmark. The training identity holds to a maximum absolute discrepancy of 5.7×10^{-15}, confirming the numerical implementation of the lineage propagation mechanism. Appendix C.2 reports the perturbation design and detailed results.

Aggregate error. We next examine how aggregate error changes as different lineage topologies expand. The left panel of Figure 2 compares populations of 4, 8 and 16 neural models at distillation weight $\beta = 0.7$. Increasing population size from 4 to 16 lowers chain aggregate MSE from 0.01519 to 0.00860, a reduction of 43.4%, while star aggregate MSE falls from 0.01936 to 0.01779, a reduction of 8.1%. The larger improvement along the chain is consistent with the diversification mechanism in Proposition 4. Comparing individual and aggregate performance further reveals the cost of common exposure. At $M = 16$, the star has 39.6% lower average individual MSE than independent roots, yet 2.74 times their aggregate MSE.

Individual choice. Finally, we test whether individual teacher choices lead to assignments with higher aggregate error than those selected for the population. Each developer chooses a teacher by individual validation MSE, while the population objective selects the assignment with the lowest aggregate validation MSE among all allocations of four students to two teachers. Then we fix both

assignments and evaluate them on the test set. The right panel in Figure 2 reports the resulting reduction in aggregate MSE from population choice across teacher quality gaps and prescribed overlaps γ between teacher error directions, where $\gamma = 0$ corresponds to orthogonal directions.

At $\gamma = 0$ and quality gap 0.10, all eight neural runs produce a concentrated individual assignment and a diversified population assignment with lower test error. Mean aggregate MSE falls from 0.04800 under individual choice to 0.03325 under population choice, a 30.7% reduction calculated from the two means. The mean paired gain is 0.01475, with a descriptive 95% seed bootstrap interval of $[0.01049, 0.01926]$. The same joint pattern appears in all eight seeds at gap 0.20. These results illustrate the divergence between individual and population objectives in Theorem 7.

The gains diminish as the prescribed teacher error directions become more similar. At $\gamma = 0.5$ and gap 0.20, the mean gain is 0.00075, with an interval of $[-0.00059, 0.00258]$. At $\gamma = 0.9$, individual and population choices select identical assignments throughout the positive gap grid. Together, these comparisons show how the benefit of diversification depends on both teacher quality differences and error overlap. Appendix C reports the complete grid and paired results across seeds.

6 Discussion

Training relationships turn local errors into shared exposure, which leads the diversification gained from adding models depends on their lineage. Our representation connects this exposure to aggregate error and identifies when propagated variance persists as the population grows. The teacher choice model further shows how individual accuracy incentives can favor concentration even when a diversified assignment improves population performance. The last provided synthetic experiments illustrate these mechanisms under controlled training conditions.

While our controlled experiments illustrate the mechanisms, their relevance to real LLM populations remains an open empirical question. Interventions in teacher selection and distillation could further clarify how training relationships shape shared errors. Extending the choice model across generations with heterogeneous compute budgets and multiple teachers, could bring the analysis closer to actual model development. Positioned as an initial theoretical foundation, we provide a starting point for this research agenda. The inspired future work could inform model evaluation and selection in LLMs research and help explain correlated decision errors in practical applied domains where decision makers rely on LLM.

AI use statement

We use AI to help us search literature, polish sentences, fix bugs, double check proof, clean code and provide critical feedback.

Reproducibility statement

Appendix A provides the proofs in the main text. Appendix B develops the training foundation and theoretical extensions. Appendix C records the experimental design and supplementary results. An executable notebook is available in our anonymous repository.

References

Daron Acemoglu, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012.

Venkatesh Bala and Sanjeev Goyal. A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229, 2000.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in neural information processing systems*, 35:3663–3678, 2022.

Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. Ecosystem graphs: Documenting the foundation model supply chain. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 196–209, 2024.

Rishi Bommasani, Sarah H Bana, Kathleen A Creel, Dan Jurafsky, and Percy Liang. Algorithmic monocultures in hiring. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 6351–6382, 2026.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

Florentina Bunea and Luo Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. 2015.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121, 1974.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313*, 2025.

Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Matthew O Jackson and Asher Wolinsky. A strategic model of social and economic networks. *Journal of economic theory*, 71(1):44–74, 1996.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *arXiv preprint arXiv:2506.07962*, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? *Advances in Neural Information Processing Systems*, 36:11037–11048, 2023.

Kenny Peng and Nikhil Garg. Monoculture in matching markets. *Advances in Neural Information Processing Systems*, 37:81959–81991, 2024.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR, 2019.

Mohammad Shahedur Rahman, Peng Gao, and Yuede Ji. Hugginggraph: Understanding the supply chain of llm ecosystem. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5997–6005, 2025.

Aliaksei Sandryhaila and José MF Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013.

Bastian Seifert, Chris Wendler, and Markus Püschel. Causal fourier analysis on directed acyclic graphs and posets. *IEEE Transactions on Signal Processing*, 71:3805–3820, 2023.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919, 2021.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. 2010.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. Llm dna: Tracing model evolution via functional representations. In *International Conference on Learning Representations*, volume 2026, pp. 7252–7279, 2026.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. Phylolm: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In *International Conference on Learning Representations*, volume 2025, pp. 64792–64844, 2025.

Chuanchao Zang, Xiangtao Meng, Wenyu Chen, Tianshuo Cong, Zha Yaxing, Dong Qi, Zheng Li, and Shanqing Guo. Errortrace: A black-box traceability mechanism based on model family error space. *Advances in Neural Information Processing Systems*, 38:68550–68578, 2026.

Sally Zhu, Ahmed Ahmed, Rohith Kuditipudi, and Percy Liang. Independence tests for language models. *arXiv preprint arXiv:2502.12292*, 2025.

A PROOFS

Proof of Proposition 1. The topological ordering makes A strictly lower triangular, so $A^M = 0$. Hence

$$(I - A)\left(I + A + \cdots + A^{M-1}\right) = I - A^M = I.$$

The centered recursion can be written as

$$(I - A)(e - b) = \tilde{u}.$$

Multiplying by the inverse of $I - A$ gives the unique solution

$$e - b = \Phi\tilde{u}, \qquad \Phi = (I - A)^{-1} = I + A + \cdots + A^{M-1},$$

which proves Equation (2).

For $k \geq 1$, the entry $(A^k)_{ij}$ sums the products of direct loadings along all directed paths of length k from j to i. In particular,

$$(A^k)_{ij} = \sum_{j=i_0<i_1<\cdots<i_k=i} \prod_{\ell=1}^{k} A_{i_\ell i_{\ell-1}}.$$

Entries corresponding to absent edges contribute zero. Summing over path lengths gives the accumulated loading Φ_{ij}. The identity matrix supplies $\Phi_{ii} = 1$, and the path formula gives $\Phi_{ij} = 0$ whenever $j \notin \mathcal{A}(i) \cup \{i\}$. □

Proof of Theorem 2. For a fixed training configuration, Φ is deterministic. Taking the covariance of $e - b = \Phi\tilde{u}$ therefore gives

$$\Sigma = \Phi\Omega\Phi^\top.$$

Substituting the local factor representation in Equation (4) yields

$$e_i - b_i = \sigma_c c_i z + \sum_g (\phi_g)_i \zeta_g + \sum_j \Phi_{ij} s_j \xi_j.$$

Since the task factor has unit variance, its contribution to $\mathrm{Cov}(e_i, e_\ell)$ is $\sigma_c^2 c_i c_\ell$. The family contribution is

$$\mathrm{Cov}\left(\sum_g (\phi_g)_i \zeta_g, \sum_h (\phi_h)_\ell \zeta_h\right) = \sum_{g,h} \rho_{gh} (\phi_g)_i (\phi_h)_\ell.$$

The mutually uncorrelated model specific factors contribute

$$\mathrm{Cov}\left(\sum_j \Phi_{ij} s_j \xi_j, \sum_j \Phi_{\ell j} s_j \xi_j\right) = \sum_j \Phi_{ij}\Phi_{\ell j} s_j^2.$$

Combining these terms gives

$$\Sigma = \sigma_c^2 cc^\top + \sum_{g,h} \rho_{gh} \phi_g \phi_h^\top + \Phi\Lambda\Phi^\top,$$

establishing Equation (6).

For fixed aggregation weights w, the aggregate error has mean $w^\top b$ and variance $w^\top \Sigma w$. Consequently,

$$L(w) = (w^\top b)^2 + w^\top \Sigma w = (w^\top b)^2 + (\Phi^\top w)^\top \Omega (\Phi^\top w),$$

which proves Equation (7). Applying the covariance decomposition to $w^\top \Sigma w$ gives

$$\mathrm{Var}(\bar{e}(w)) = \sigma_c^2 (w^\top c)^2 + \sum_{g,h} \rho_{gh} (w^\top \phi_g)(w^\top \phi_h) + \sum_j s_j^2 (\Phi^\top w)_j^2.$$

This is Equation (8). Finally, $c = \Phi q$ and $\Phi^{-1} = I - A$ imply

$$c = \mathbf{1} \quad \Longleftrightarrow \quad q = (I - A)\mathbf{1}.$$

Entrywise, this condition is $q_i + \sum_{j \in \mathcal{P}_i} A_{ij} = 1$, establishing Remark 3.1. □

Proof of Proposition 3. By definition, the jth entry of $\Phi^\top \mathbf{1}$ is n_j. Since $\Lambda = \mathrm{diag}(s_1^2, \ldots, s_M^2)$,

$$V_M = \frac{1}{M^2} \mathbf{1}^\top \Phi\Lambda\Phi^\top \mathbf{1} = \sum_j s_j^2 \left(\frac{n_j}{M}\right)^2.$$

Applying the uniform variance bounds term by term gives

$$\underline{s}^2 H \leq V_M \leq \overline{s}^2 H.$$

The two bounds are positive constants independent of M. Thus $V_M \to 0$ exactly when $H \to 0$, proving Equation (11) and its stated consequence.

Nonnegative direct loadings give nonnegative entries of Φ, so $n_j \geq 0$. Therefore,

$$\left(\max_j n_j\right)^2 \leq \sum_j n_j^2 \leq \left(\max_j n_j\right) \sum_j n_j.$$

Dividing by M^2 proves Equation (12). Under $\sum_j n_j = O(M)$, the upper bound implies $H \to 0$ whenever $\max_j n_j/M \to 0$. Conversely, the lower bound implies $\max_j n_j/M \to 0$ whenever $H \to 0$. Combining these implications with the variance bounds gives

$$V_M \to 0 \quad \Longleftrightarrow \quad \max_j n_j = o(M).$$

The same variance bounds also imply $\liminf_{M\to\infty} V_M > 0$ exactly when $\liminf_{M\to\infty} H > 0$. □

Proof of Proposition 4. Under $\Omega = s^2 I$, aggregate variance is

$$V_M = \frac{s^2}{M^2} \sum_{j=1}^{M} n_j^2.$$

We compute the source loadings n_j for each topology.

For roots, $A = 0$ and $\Phi = I$. Every source has $n_j = 1$, giving

$$V_M^{\text{Roots}} = \frac{s^2}{M}.$$

For a star rooted at model 1, the lineage matrix has $\Phi_{11} = 1$, $\Phi_{i1} = a$ for $i \geq 2$, and $\Phi_{ii} = 1$. Its remaining entries are zero. Hence

$$n_1 = 1 + (M-1)a, \qquad n_j = 1 \quad \text{for } j \geq 2.$$

Substitution gives

$$V_M^{\text{Star}} = \frac{s^2}{M^2}\left([1 + (M-1)a]^2 + M - 1\right).$$

For fixed a, the first squared term divided by M^2 converges to a^2, while $(M-1)/M^2$ converges to zero. Thus $V_M^{\text{Star}} \to s^2 a^2$.

For a chain, the unique path from j to $i \geq j$ contains $i - j$ edges, each with loading a. Therefore,

$$\Phi_{ij} = a^{i-j} \quad \text{for } j \leq i, \qquad n_j = \sum_{i=j}^{M} a^{i-j} = \frac{1 - a^{M-j+1}}{1-a}.$$

Reindexing by $k = M - j + 1$ gives

$$V_M^{\text{Chain}} = \frac{s^2}{M^2} \sum_{k=1}^{M} \left(\frac{1-a^k}{1-a}\right)^2.$$

For fixed $0 < a < 1$,

$$\sum_{k=1}^{M} (1-a^k)^2 = M - \frac{2a(1-a^M)}{1-a} + \frac{a^2(1-a^{2M})}{1-a^2} = M + O(1).$$

It follows that

$$V_M^{\text{Chain}} = \frac{s^2}{M(1-a)^2} + O(M^{-2}) \sim \frac{s^2}{M(1-a)^2}.$$

These calculations establish Equations (13)–(15). □

Proof of Proposition 5. The covariance matrix Σ is symmetric and positive semidefinite. Its spectral decomposition therefore gives

$$\text{Var}(\bar{e}(w)) = w^\top \Sigma w = \sum_k \lambda_k (w^\top v_k)^2.$$

Also,

$$\Sigma = \Phi\Omega\Phi^\top = (\Phi\Omega^{1/2})(\Phi\Omega^{1/2})^\top.$$

The eigenvalues of this matrix are the squared singular values of $\Phi\Omega^{1/2}$.

The vector $\mathbf{1}/\sqrt{M}$ has unit Euclidean norm. Applying the Rayleigh bound yields

$$\lambda_{\max}(\Sigma) \geq \frac{\mathbf{1}^\top \Sigma \mathbf{1}}{M} = M \operatorname{Var}(\bar{e}_M).$$

Under $\operatorname{Var}(\bar{e}_M) > 0$, this lower bound is positive. Using $\operatorname{tr}(\Sigma) = \sum_i \operatorname{Var}(e_i)$, we obtain

$$r(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} \leq \frac{\sum_i \operatorname{Var}(e_i)}{M \operatorname{Var}(\bar{e}_M)} = \frac{M^{-1} \sum_i \operatorname{Var}(e_i)}{\operatorname{Var}(\bar{e}_M)}.$$

This proves Equation (16). □

Proof of Proposition 6. The factors associated with earlier model specific sources have zero mean, unit variance and zero pairwise covariance. Consequently, the cross second moment of the components associated with teachers p and q is

$$\begin{aligned}\mathbb{E}\left[\left(b(p) + a\sum_j \Phi_{pj} s_j \xi_j\right)\left(b(q) + a\sum_j \Phi_{qj} s_j \xi_j\right)\right] \\ = b(p)b(q) + a^2 \sum_j \Phi_{pj}\Phi_{qj} s_j^2 = Q_{pq}.\end{aligned}$$

This also establishes positive semidefiniteness of Q. For any real vector $v = (v_p)_{p\in\mathcal{P}}$,

$$v^\top Q v = \mathbb{E}\left[\left(\sum_{p\in\mathcal{P}} v_p \left(b(p) + a\sum_j \Phi_{pj} s_j \xi_j\right)\right)^2\right] \geq 0.$$

Taking the second moment of Equation (20) thus gives

$$\begin{aligned}L_i(p) &= b(p)^2 + \sigma_c^2 + \tau^2 + a^2 \sum_j \Phi_{pj}^2 s_j^2 + s^2 \\ &= \sigma_c^2 + \tau^2 + Q_{pp} + s^2.\end{aligned}$$

For a fixed assignment, averaging over the M students gives

$$\bar{e}_M = \sum_{p\in\mathcal{P}} \alpha_p b(p) + \sigma_c z + \tau\zeta_{g_0} + a\sum_j \left(\sum_{p\in\mathcal{P}} \alpha_p \Phi_{pj}\right) s_j \xi_j + \frac{s}{M}\sum_{i=1}^M \xi_i.$$

Here the j sum runs over earlier models and the i sum runs over the evaluated students. The student local factors are mutually uncorrelated, so the last term has variance s^2/M. Taking the second moment yields

$$\begin{aligned}L_M &= \sigma_c^2 + \tau^2 + \frac{s^2}{M} + \left(\sum_{p\in\mathcal{P}} \alpha_p b(p)\right)^2 + a^2 \sum_j s_j^2 \left(\sum_{p\in\mathcal{P}} \alpha_p \Phi_{pj}\right)^2 \\ &= \sigma_c^2 + \tau^2 + \frac{s^2}{M} + \sum_{p,q\in\mathcal{P}} \alpha_p\alpha_q \left(b(p)b(q) + a^2 \sum_j \Phi_{pj}\Phi_{qj} s_j^2\right) \\ &= \sigma_c^2 + \tau^2 + \alpha^\top Q \alpha + \frac{s^2}{M}.\end{aligned}$$

This proves both identities in Equation (22). □

Proof of Theorem 7. By Proposition 6,

$$L_i(q) - L_i(p) = Q_{qq} - Q_{pp} > 0.$$

Every developer therefore strictly prefers teacher p.

Let x be the fraction of students assigned to teacher q. The assignment dependent part of aggregate error is

$$f(x) = (1-x)^2 Q_{pp} + 2x(1-x)Q_{pq} + x^2 Q_{qq}.$$

The common task contribution, common family contribution and student local variance are constant across assignments. Thus minimizing aggregate error is equivalent to minimizing f over $x \in \{0, 1/M, \dots, 1\}$.

Denote $D = Q_{pp} + Q_{qq} - 2Q_{pq}$ for the quadratic coefficient of f. By the definition of Q, this coefficient is

$$D = \big(b(p) - b(q)\big)^2 + a^2 \sum_j \big(\Phi_{pj} - \Phi_{qj}\big)^2 s_j^2.$$

Positive semidefiniteness of Q gives $Q_{pq} \le \sqrt{Q_{pp}Q_{qq}}$. Hence

$$D \ge \left(\sqrt{Q_{qq}} - \sqrt{Q_{pp}}\right)^2 > 0,$$

where the strict inequality follows from $Q_{pp} < Q_{qq}$. Therefore f is strictly convex.

Moving one student from p to q changes the share from zero to $1/M$. The resulting change is

$$\begin{aligned}\Delta L_M &= f(1/M) - f(0) \\ &= \frac{2(Q_{pq} - Q_{pp})}{M} + \frac{D}{M^2} \\ &= \frac{Q_{qq} - Q_{pp} - 2(M-1)(Q_{pp} - Q_{pq})}{M^2}.\end{aligned}$$

This proves Equation (24).

To characterize all optimal assignments, let k denote the number of students assigned to q. Consecutive feasible assignments satisfy

$$f((k+1)/M) - f(k/M) = \frac{2(Q_{pq} - Q_{pp})}{M} + \frac{(2k+1)D}{M^2}, \qquad k = 0, \dots, M-1.$$

These increments increase strictly with k because $D > 0$.

If the first increment is nonnegative, all subsequent increments are positive. The concentrated assignment $k = 0$ is then a minimizer. When the first increment equals zero, $k = 0$ and $k = 1$ attain the same minimum.

If the first increment is negative, assigning one student to q strictly improves on $k = 0$. The other concentrated assignment satisfies

$$f(1) = Q_{qq} > Q_{pp} = f(0) > f(1/M).$$

Both endpoints therefore have strictly higher error than a feasible diversified assignment. A minimum exists on the finite feasible set, and every minimizer lies in $\{1, \dots, M-1\}$.

It follows that every population optimum uses both teachers exactly when

$$Q_{qq} - Q_{pp} < 2(M-1)(Q_{pp} - Q_{pq}).$$

Together with the assumed strict individual preference $Q_{qq} - Q_{pp} > 0$, this is Equation (23). □

B THEORETICAL EXTENSIONS

B.1 AN EXACT TRAINING FOUNDATION

The direct recursion in Section 2 follows exactly from a squared loss training problem with shared linear features. The following construction identifies both the direct transmission loadings and the local errors generated by training.

Let $X \in \mathbb{R}^{n\times d}$ be the common training design, where n is the number of observations and d is the number of features. Suppose $X^\top X/n = I_d$ and let the evaluation target be $f_*(x) = x^\top\theta_*$. Model i receives local targets $y_i = X(\theta_* + \nu_i)$, where $\nu_i \in \mathbb{R}^d$ is its local target perturbation. For a student, let $W_{ij} \geq 0$ be teacher mixture weights satisfying $\sum_{j\in\mathcal{P}_i} W_{ij} = 1$ and denote $\beta_i \in [0,1]$ as the distillation weight. Setting $W_{ij} = 0$ outside $\mathcal{P}_i$, then a root has $\beta_i = 0$ and a zero row in W.

The mixed target and fitted coefficients are

$$\begin{aligned} t_i &= (1-\beta_i)y_i + \beta_i \sum_{j\in\mathcal{P}_i} W_{ij} X\theta_j, \\ \theta_i &= \arg\min_\theta \left\{ \frac{1}{n}\|X\theta - t_i\|_2^2 + \lambda\|\theta\|_2^2 \right\}, \qquad \lambda \geq 0. \end{aligned} \tag{25}$$

The first order condition gives $(X^\top X/n + \lambda I_d)\theta_i = X^\top t_i/n$. Using the feature normalization yields

$$\theta_i = \frac{(1-\beta_i)(\theta_* + \nu_i) + \beta_i \sum_{j\in\mathcal{P}_i} W_{ij}\theta_j}{1+\lambda}. \tag{26}$$

Subtracting θ_* gives

$$\theta_i - \theta_* = \sum_{j\in\mathcal{P}_i} \frac{\beta_i W_{ij}}{1+\lambda}(\theta_j - \theta_*) + \frac{(1-\beta_i)\nu_i - \lambda\theta_*}{1+\lambda}. \tag{27}$$

The same expression applies to roots because $\beta_i = 0$. For the prediction error $e_i(x) = x^\top(\theta_i - \theta_*)$, we define

$$A_{ij} = \frac{\beta_i W_{ij}}{1+\lambda}, \qquad u_i(x) = x^\top \frac{(1-\beta_i)\nu_i - \lambda\theta_*}{1+\lambda}. \tag{28}$$

Then $e = Ae + u$ holds pointwise. Taking expectations over the evaluation distribution gives $b = Ab + \mathbb{E}[u]$. Subtracting this identity establishes Equation (1) with $\tilde{u} = u - \mathbb{E}[u]$. The local term includes both the model's target perturbation and the shrinkage error introduced by regularization.

We can use the construction to determine local covariance. Conditional on the trained coefficients, let $U \in \mathbb{R}^{d\times M}$ have column i equal to $[(1-\beta_i)\nu_i - \lambda\theta_*]/(1+\lambda)$. Since $u = U^\top x$, then

$$\Omega = U^\top \operatorname{Cov}(x) U. \tag{29}$$

Shared directions in the columns of U therefore generate correlated local errors. In particular, the common shrinkage term can contribute to this dependence. Theorem 2 propagates the resulting Ω through the lineage. The task and family decomposition in Equation (4) further separates these sources when its factor assumptions hold.

B.2 Approximate propagation

To evaluate how training deviations propagate to the evaluated population, we leverage the lineage matrix. Fixing A and the local terms $\tilde{u}$ from calibration, suppose the centered errors satisfy

$$e - b = A(e-b) + \tilde{u} + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0.$$

Here ε collects the remaining local discrepancies. Define the affine prediction by $e^{\text{aff}} = b + \Phi\tilde{u}$. Solving the recursion gives

$$e - e^{\text{aff}} = \Phi\varepsilon. \tag{30}$$

Thus a discrepancy introduced at one node reaches later models through the same paths as a local error source. For a scalar random variable Y, we rewrite $\|Y\|_{L^2} = (\mathbb{E}[Y^2])^{1/2}$, then point-wise Cauchy Schwarz followed by expectation yields

$$\left\| w^\top(e - e^{\text{aff}}) \right\|_{L^2} \leq \|\Phi^\top w\|_2 \left(\mathbb{E}\|\varepsilon\|_2^2\right)^{1/2}. \tag{31}$$

For convenience, we use C to denote the right side. Expanding the difference between the two aggregate squared errors and applying Cauchy Schwarz gives

$$\left|\mathbb{E}[(w^\top e)^2] - \mathbb{E}[(w^\top e^{\text{aff}})^2]\right| \leq 2\|w^\top e^{\text{aff}}\|_{L^2} C + C^2. \tag{32}$$

The bound separates the accuracy of the local approximation from the exposure that propagates its discrepancies. For equal weights, $\|\Phi^\top w\|_2^2 = H$. Therefore, source concentration also governs sensitivity to local approximation errors.

B.3 Local rescaling under distillation

Proposition 4 holds local innovation variance fixed to isolate the effect of transmission paths. The training foundation provides a related comparison in which distillation also rescales the student's local contribution.

Here we consider zero regularization and one teacher per student. Let ε_i be centered local target errors with $\mathrm{Cov}(\varepsilon) = s^2 I$ and suppose

$$e_1 = \varepsilon_1, \qquad e_i = a e_{p_i} + (1-a)\varepsilon_i, \qquad 0 < a < 1.$$

The direct teacher loading remains a and the root has local variance s^2, while each student has local variance $(1-a)^2 s^2$. For independent roots, every model has error ε_i, giving aggregate variance s^2/M.

For the star, the root's total loading is $1 + (M-1)a$ and each student's own source has total loading one. The covariance formula therefore gives

$$V_M^{\mathrm{Star}} = \frac{s^2}{M^2}\left([1+(M-1)a]^2 + (M-1)(1-a)^2\right) \longrightarrow s^2 a^2. \tag{33}$$

For the chain, the total loading on source j is $(1-a^{M-j+1})/(1-a)$. Separating the root from the student sources gives

$$V_M^{\mathrm{Chain}} = \frac{s^2}{M^2}\left[\left(\frac{1-a^M}{1-a}\right)^2 + (1-a)^2 \sum_{k=1}^{M-1}\left(\frac{1-a^k}{1-a}\right)^2\right] \sim \frac{s^2}{M}. \tag{34}$$

B.4 Earlier teachers and the student population

In the single generation choice model, students inherit errors from their chosen teachers and those teachers' ancestors. We include these earlier models in an augmented DAG to trace this inheritance. Let $\mathcal{V}_T$ contain the earlier models, including the candidate teachers $\mathcal{P}$ and their ancestors. Use A_T for their direct loading matrix and $\Phi_T = (I - A_T)^{-1}$ for their lineage matrix. Define the assignment matrix $Z \in \mathbb{R}^{M \times |\mathcal{V}_T|}$ by $Z_{ip} = \mathbf{1}_{\{p_i = p\}}$. Columns corresponding to earlier models outside $\mathcal{P}$ are zero. Under the common student transmission coefficient a, the augmented matrices are

$$A_{\mathrm{aug}} = \begin{pmatrix} A_T & 0 \\ aZ & 0 \end{pmatrix}, \qquad \Phi_{\mathrm{aug}} = \begin{pmatrix} \Phi_T & 0 \\ aZ\Phi_T & I_M \end{pmatrix}. \tag{35}$$

Multiplying the two block matrices verifies $(I - A_{\mathrm{aug}})\Phi_{\mathrm{aug}} = I$. The block $aZ\Phi_T$ contains each student's exposure to earlier sources through its selected teacher. Its entry at student i and source j is $a\Phi_{p_i j}$. Under Equation (20), each candidate teacher has centered error

$$e_p - b_p = \sigma_c z + \tau\zeta_{g_0} + \sum_j \Phi_{pj} s_j \xi_j.$$

The corresponding centered student local term is

$$\tilde{u}_i(p) = (1-a)(\sigma_c z + \tau\zeta_{g_0}) + s\xi_i.$$

Adding this term to $a(e_p - b_p)$ gives

$$e_i(p) - b(p) = \sigma_c z + \tau\zeta_{g_0} + a\sum_j \Phi_{pj} s_j \xi_j + s\xi_i.$$

The inherited and locally introduced task and family components therefore add to the common final loadings specified in the main text.

Assigning zero evaluation weight to earlier models and weight $1/M$ to each student gives the student aggregate, while Φ_{aug} retains their exposure to teachers and ancestors. This construction embeds the single generation choice model in the general DAG framework, so the covariance and aggregation results can apply directly to the evaluated student population.

B.5 STUDENTS ALLOCATION

Theorem 7 establishes when using both teachers improves on the individually preferred assignment. We now examine how to allocate students across teachers, first deriving the change in aggregate error when one student switches teachers, then characterizing optimal allocations for two teachers and for a general teacher bank. Let $H_{\mathcal{P}} = \sum_{p\in\mathcal{P}} \alpha_p^2$ measure concentration in teacher assignments and suppose one student moves from teacher q to a distinct teacher p. Using counts and shares before the move, we have

$$\Delta H_{\mathcal{P}} = \frac{2(N_p - N_q + 1)}{M^2}, \qquad \Delta L_i = Q_{pp} - Q_{qq}. \tag{36}$$

The first identity follows by expanding $(N_p+1)^2+(N_q-1)^2-N_p^2-N_q^2$. Expanding the population quadratic at the new shares gives

$$\Delta L_M = \frac{2}{M}\left[(Q\alpha)_p - (Q\alpha)_q\right] + \frac{Q_{pp} + Q_{qq} - 2Q_{pq}}{M^2}. \tag{37}$$

For two teachers satisfying $Q_{pp} < Q_{qq}$, let x be the fraction assigned to q. The assignment dependent objective is

$$f(x) = Q_{pp} + 2x(Q_{pq} - Q_{pp}) + x^2(Q_{pp} + Q_{qq} - 2Q_{pq}).$$

Its quadratic coefficient is positive under the strict quality ordering. Completing the square gives

$$\alpha_q^* = \left[\frac{Q_{pp} - Q_{pq}}{Q_{pp} + Q_{qq} - 2Q_{pq}}\right]_{[0,1]}, \qquad [x]_{[0,1]} = \min\{1, \max\{0, x\}\}. \tag{38}$$

For a finite cohort, the optimal number assigned to q is a nearest feasible integer to $M\alpha_q^*$. Both nearest integers are optimal when their objective values coincide.

Denote $C_{pq} = Q_{pq} - b(p)b(q) = a^2 \sum_j \Phi_{pj}\Phi_{qj}s_j^2$ for the covariance between the inherited model specific components under teachers p and q. A useful specialization holds inherited variances equal and inherited components uncorrelated,

$$C_{pp} = C_{qq} = \omega^2, \qquad C_{pq} = 0, \qquad 0 \le b(p) < b(q).$$

Since $d = b(q) - b(p) > 0$, the condition in Theorem 7 becomes

$$2Mb(p)d + d^2 < 2\omega^2(M-1). \tag{39}$$

After division by M^2, the left side is the increase in squared aggregate bias from moving one student to q, and the right side is the reduction in inherited variance. The continuous optimal share reduces to

$$\alpha_q^* = \max\left\{0, \frac{\omega^2 - b(p)d}{2\omega^2 + d^2}\right\}.$$

For an arbitrary finite teacher set $\mathcal{P}$, the exact allocation problem is

$$\min_{\alpha}\ \alpha^\top Q\alpha, \qquad \alpha_p \in \{0, 1/M, \dots, 1\}, \qquad \sum_{p\in\mathcal{P}} \alpha_p = 1. \tag{40}$$

The common contribution $\sigma_c^2 + \tau^2 + s^2/M$ is fixed across assignments. Allowing continuous shares gives a convex problem on the simplex and a lower bound for the finite cohort optimum. A minimizer exists because the simplex is compact. Positive definiteness of Q makes it unique.

The continuous optimum can be characterized by its occupied teacher set $\mathcal{S} \subseteq \mathcal{P}$. Under positive definiteness, let $Q_{\mathcal{S}\mathcal{S}}$ be the corresponding principal submatrix and $\mathbf{1}_{\mathcal{S}}$ a vector of ones. The candidate active shares and their objective value are

$$\alpha_{\mathcal{S}}^* = \frac{Q_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{\mathcal{S}}}{\mathbf{1}_{\mathcal{S}}^\top Q_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{\mathcal{S}}}, \qquad v^* = \frac{1}{\mathbf{1}_{\mathcal{S}}^\top Q_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{1}_{\mathcal{S}}}. \tag{41}$$

This candidate is optimal exactly when its active shares are strictly positive and

$$Q_{q\mathcal{S}}\alpha_{\mathcal{S}}^* \ge v^*, \qquad q \in \mathcal{P}\setminus\mathcal{S}. \tag{42}$$

To obtain these conditions, consider an infinitesimal transfer between two occupied teachers. Optimality requires equal marginal contributions, so $(Q\alpha^*)_p = v^*$ on $\mathcal{S}$. Multiplying by the shares and summing identifies $v^* = (\alpha^*)^\top Q\alpha^*$. Solving the active linear system and imposing unit total share gives Equation (41). A transfer toward an unused teacher has a nonnegative directional derivative exactly when Equation (42) holds. Convexity makes these conditions sufficient as well as necessary.

For positive semidefinite Q, the same marginal conditions characterize optimality, with the active equations written as $Q_{\mathcal{S}\mathcal{S}}\alpha_{\mathcal{S}} = v\mathbf{1}_{\mathcal{S}}$ and $\mathbf{1}_{\mathcal{S}}^\top \alpha_{\mathcal{S}} = 1$. The inverse formula applies whenever the active submatrix is invertible.

B.6 Autonomous training and local compute

The main choice model fixes local training variance across teacher assignments. Access costs and compute budgets introduce an additional tradeoff between inherited accuracy and local learning.

Given a compute budget $B_i > 0$, teacher p is feasible for student i if its access cost satisfies $0 \le K_p < B_i$. Let

$$k_{ip} = \lambda_p(B_i - K_p), \qquad 0 < \lambda_p \le 1, \qquad k_{i0} = B_i. \tag{43}$$

Here λ_p is the fraction of the remaining budget converted into effective local compute, and option 0 denotes autonomous training. The local variance follows the main specification

$$s_i^2(k) = \sigma_i^2 + a_i k^{-\eta}.$$

Retaining the common final task and family loadings from Section 4, we allow the teacher transmission coefficient A_{ip} and student mean $b_i(p)$ to depend on the choice. Then

$$e_i(p) = b_i(p) + \sigma_c z + \tau\zeta_{g_0} + A_{ip}\sum_j \Phi_{pj} s_j \xi_j + s_i(k_{ip})\xi_i,$$

for a feasible teacher, while autonomous training gives

$$e_i(0) = b_i(0) + \sigma_c z + \tau\zeta_{g_0} + s_i(B_i)\xi_i.$$

The earlier source factors and student local factors satisfy the orthogonality assumptions of the main choice model. The individual objectives are therefore

$$\begin{aligned} L_i(p) &= \sigma_c^2 + \tau^2 + b_i(p)^2 + A_{ip}^2 \sum_j \Phi_{pj}^2 s_j^2 + s_i^2(k_{ip}), \\ L_i(0) &= \sigma_c^2 + \tau^2 + b_i(0)^2 + s_i^2(B_i). \end{aligned} \tag{44}$$

Teacher p strictly improves individual performance exactly when

$$b_i(0)^2 - b_i(p)^2 - A_{ip}^2 \sum_j \Phi_{pj}^2 s_j^2 > a_i\left(k_{ip}^{-\eta} - B_i^{-\eta}\right). \tag{45}$$

The left side measures the reduction in squared bias after accounting for inherited variance. The right side is the increase in local variance from using less effective local compute. Autonomous training minimizes individual error when its objective is weakly below every feasible teacher objective.

For an assignment $(p_1, \ldots, p_M)$, the population objective is

$$\begin{aligned} L_M = &\left(\frac{1}{M}\sum_i b_i(p_i)\right)^2 + \sigma_c^2 + \tau^2 \\ &+ \sum_j s_j^2 \left(\frac{1}{M}\sum_{i:p_i>0} A_{ip_i}\Phi_{p_i j}\right)^2 + \frac{1}{M^2}\sum_i s_i^2(k_{ip_i}). \end{aligned} \tag{46}$$

Teacher choices now affect mean error, exposure to earlier sources and local training variance. Their population value depends on all three contributions.

A general switching identity makes the difference between individual and population incentives explicit. We hold every other student's trained error fixed and let $\delta e_i = e_i(p) - e_i(q)$, where either option may be autonomous training. If $\bar{e}_M$ is the aggregate before the move, direct expansion gives

$$\begin{aligned} \Delta L_M &= \frac{2}{M}\mathbb{E}[\bar{e}_M \delta e_i] + \frac{1}{M^2}\mathbb{E}[(\delta e_i)^2], \\ \Delta L_i &= 2\mathbb{E}[e_i(q)\delta e_i] + \mathbb{E}[(\delta e_i)^2]. \end{aligned} \tag{47}$$

Combining the two identities yields

$$\Delta L_M = \frac{\Delta L_i}{M^2} + \frac{2}{M^2}\sum_{\ell \neq i} \mathbb{E}[e_\ell \delta e_i].$$

The additional cross terms measure how the switching student's error change interacts with the rest of the population. For interventions that also retrain descendants, the change in their errors enters the aggregate through the corresponding lineage paths.

C Experimental details and supplementary results

C.1 Protocol and evaluation

The completed synthetic experiment uses seeds 2701 through 2708, input dimension 8, 192 training observations, 256 validation observations and 1024 test observations. Each seed is shared across the paired topology and training comparisons. The protocol uses linear and nonlinear target tasks, ridge learners and neural learners with hidden width 16. Neural training budgets are 200 and 600 updates with learning rate 0.01. The source perturbation scale is 0.15, the configured local noise scale is 0.18, and the propagation graphs contain eight nodes. The growth experiment evaluates populations of 4, 8 and 16 nodes at distillation weights 0.4 and 0.7, with ridge penalties 0 and 0.05.

For test errors $e_i(t)$ on T observations indexed by t, individual MSE and aggregate MSE are

$$\widehat{L}_i = T^{-1}\sum_t e_i(t)^2, \qquad \widehat{L}_M = T^{-1}\sum_t \left(M^{-1}\sum_i e_i(t)\right)^2.$$

All reported MSE values use the original task target. The empirical mean and covariance give the decomposition into squared aggregate bias and aggregate variance. Paired comparisons use the same evaluation observations. All reported averages give each independent seed equal weight.

The numerical summaries and frozen assignment records are stored under protocol `dag_training_full_v1`. The configuration records NumPy 2.1.3 and Python 3.13.15. Training records contain the final objective and the relative decrease over the last saved interval. At 600 updates the median relative decrease is 16.65%. The neural results therefore describe fixed training budgets.

C.2 Transmission and response prediction

The perturbation experiment compares paired trainings with a root target perturbation. Loadings are fitted from validation responses in direction h_1 and evaluated on test responses in direction h_2 and a negative direction. The measured root response initializes the descendant prediction. For a descendant response δe_i and its prediction $\widehat{\delta e_i}$, the recorded relative RMSE is

$$RMSE = \frac{\sqrt{T^{-1}\sum_t (\delta e_i(t) - \widehat{\delta e_i}(t))^2}}{\sqrt{T^{-1}\sum_t \delta e_i(t)^2}}.$$

Positive response energy defines this statistic. Descendants of separate roots have zero transmitted response and are summarized separately. The reported means first average eligible descendant and probe results within each seed and then average seeds. The stored skill score is one minus the squared relative RMSE, using zero response as its reference prediction.

The ridge identity check has maximum absolute discrepancy 5.7×10^{-15}. Ridge response linearity holds on both target tasks because the fitted prediction is linear in its training targets. Neural response

Table 1: Relative response RMSE averaged within seeds and then across seeds.

Learner	Target	Updates	Mean relative RMSE
Neural	Linear	200	0.409
Neural	Linear	600	0.420
Neural	Nonlinear	600	0.714
Ridge	Linear	600	$< 10^{-10}$
Ridge	Nonlinear	600	$< 10^{-10}$

error is larger on the nonlinear task. At 600 updates on the linear task, mean relative RMSE is 0.549 for chains, 0.345 for stars and 0.365 for the multiple teacher configuration. The corresponding values for the nonlinear task are 0.847, 0.642 and 0.652. This identifies successive transmission as the more demanding response prediction in the present grid.

The experiment also records superposition and sign diagnostics. Neural superposition relative RMSE averages 0.188 at 200 updates and 0.210 at 600 updates on the linear task, and 0.426 at 600 updates on the nonlinear task. These diagnostics quantify departures from the exact affine response alongside the prediction results.

C.3 Growth and regularization

Table 2: Mean aggregate MSE across eight seeds. Distillation weight is $\beta = 0.7$ and ridge penalty is zero. Neural models use 600 updates.

Learner	M	Roots	Star	Chain	Multiple teachers
Ridge	4	0.00611	0.01549	0.01082	0.01313
Ridge	8	0.00394	0.01459	0.00529	0.01060
Ridge	16	0.00265	0.01404	0.00285	0.00938
Neural	4	0.00971	0.01936	0.01519	0.01720
Neural	8	0.00792	0.01859	0.01062	0.01505
Neural	16	0.00649	0.01779	0.00860	0.01375

The growth comparisons use the full empirical source covariance. In the ridge records, source based aggregate variance and directly measured aggregate variance agree within 3.1×10^{-16}. At $M = 16$, $\beta = 0.7$ and zero penalty, neural average individual MSE is 0.03796 for roots and 0.02293 for a star. The corresponding aggregate MSE values are 0.00649 and 0.01779. Both the individual improvement and aggregate increase occur in all eight seeds.

Table 3: Mean test aggregate MSE at $M = 16$ and $\beta = 0.7$. Each entry averages eight seeds.

Learner	λ	Roots	Star	Chain	Multiple teachers
Ridge	0.00	0.00265	0.01404	0.00285	0.00938
Ridge	0.05	0.00549	0.01820	0.02018	0.01670
Neural	0.00	0.00649	0.01779	0.00860	0.01375

The penalty contributes the local shrinkage term $-\lambda\theta_*/(1+\lambda)$ in Equation (27). This shared direction is repeatedly inherited. At penalty 0.05, the ridge chain has higher aggregate MSE than the star. The comparison illustrates why the general formula uses both propagation and the associated local covariance. The independent innovation benchmark isolates the effect of paths under its specified local moments.

C.4 Teacher quality and correlation

The choice experiment trains four student slots for each of two teachers. Using δ for the quality gap in this grid, designed teacher errors have scale 0.3 for the first teacher and $0.3+\delta$ for the second, with $\delta \in \{0, 0.10, 0.20, 0.35, 0.50, 0.80\}$. The prescribed overlaps between teacher error directions are $\gamma \in \{0, 0.5, 0.9\}$. These parameters describe teacher construction, while Q in the theory describes

the error inherited by a student. Training determines the relation between them. Empirical teacher correlations are retained in the result records.

For each seed and grid cell, each developer chooses the teacher minimizing its own validation MSE. The population rule evaluates all 2^4 assignments and selects the smallest validation aggregate MSE. The test gain is $\widehat{L}_M^{\text{private}} - \widehat{L}_M^{\text{population}}$. The reported joint pattern requires a concentrated private assignment, a diversified population assignment, positive test margins for all individual teacher preferences and a positive test aggregate gain. Autonomous training is outside this experimental choice set.

Tables 4 and 5 report the complete grid. Intervals use 2000 paired bootstrap resamples of the eight seeds, with nominal 95% coverage for each cell considered separately. They are descriptive pointwise intervals. The grid is exploratory, and simultaneous coverage across cells requires an additional multiplicity adjustment. Relative reductions in the main text are computed from the ratio of the two mean MSE values.

Table 4: Neural teacher choice across the complete grid. Gain is private aggregate MSE minus population aggregate MSE on the test set. The final columns count concentrated private assignments and the joint pattern out of eight seeds.

γ	Gap	Mean gain	Lower	Upper	Private	Joint
0.0	0.00	0.009040	0.003492	0.015368	3/8	1/8
0.0	0.10	0.014747	0.010485	0.019256	8/8	8/8
0.0	0.20	0.011596	0.007937	0.015364	8/8	8/8
0.0	0.35	0.006412	0.001858	0.011059	8/8	7/8
0.0	0.50	0.003249	-0.000028	0.006782	8/8	3/8
0.0	0.80	0.000130	0.000000	0.000391	8/8	1/8
0.5	0.00	0.003711	0.001304	0.006323	3/8	2/8
0.5	0.10	0.003641	0.001058	0.006481	8/8	7/8
0.5	0.20	0.000749	-0.000585	0.002577	8/8	2/8
0.5	0.35	-0.000101	-0.000303	0.000000	8/8	0/8
0.5	0.50	0.000000	0.000000	0.000000	8/8	0/8
0.5	0.80	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.00	0.000610	0.000093	0.001128	2/8	0/8
0.9	0.10	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.20	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.35	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.50	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.80	0.000000	0.000000	0.000000	8/8	0/8

Table 5: Ridge teacher choice across the complete grid. Gain is private aggregate MSE minus population aggregate MSE on the test set. The final columns count concentrated private assignments and the joint pattern out of eight seeds.

γ	Gap	Mean gain	Lower	Upper	Private	Joint
0.0	0.00	0.009022	0.004477	0.013307	5/8	5/8
0.0	0.10	0.010454	0.005242	0.016556	6/8	6/8
0.0	0.20	0.010350	0.005672	0.015001	8/8	8/8
0.0	0.35	0.006204	0.001140	0.011247	8/8	4/8
0.0	0.50	0.004087	0.000642	0.007743	8/8	4/8
0.0	0.80	-0.000017	-0.000052	0.000000	8/8	0/8
0.5	0.00	0.003457	0.001649	0.004951	6/8	6/8
0.5	0.10	0.002175	-0.000035	0.004681	7/8	3/8
0.5	0.20	0.001248	-0.000178	0.002854	8/8	3/8
0.5	0.35	-0.000124	-0.000371	0.000000	8/8	0/8
0.5	0.50	0.000000	0.000000	0.000000	8/8	0/8
0.5	0.80	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.00	0.000212	-0.000069	0.000578	7/8	3/8
0.9	0.10	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.20	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.35	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.50	0.000000	0.000000	0.000000	8/8	0/8
0.9	0.80	0.000000	0.000000	0.000000	8/8	0/8