DATA-SIZE-DEPENDENT SUPERVISION IN WEAK-TO-STRONG LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training high-capacity models to generalize from limited data remains challenging. Teacher-generated supervision can help, but typically relies on a highly capable expert teacher. We study data-efficient weak-to-strong (W2S) generalization, asking whether pretrained weak teachers can effectively supervise stronger students with limited training data. Our experiments show that students can outperform their weak teachers at small data budgets, although their performance can saturate early as more data are added under a fixed teacher. Comparing supervision sources reveals a data-size-dependent crossover: among the sources evaluated, weaker teachers yield the best student performance at the smallest data budgets, while stronger teachers and eventually hard labels become preferable as the data budget increases. To examine this behavior, we characterize asymptotic risks in a two-stage linear regression model. Numerical evaluation of these risks reveals how the balance between bias and variance shifts with student data size, explaining changes in the preferred teacher within this model. Together, these findings highlight the potential of weak teachers for data-efficient learning and show that effective supervision depends jointly on teacher capability and the student's data budget.

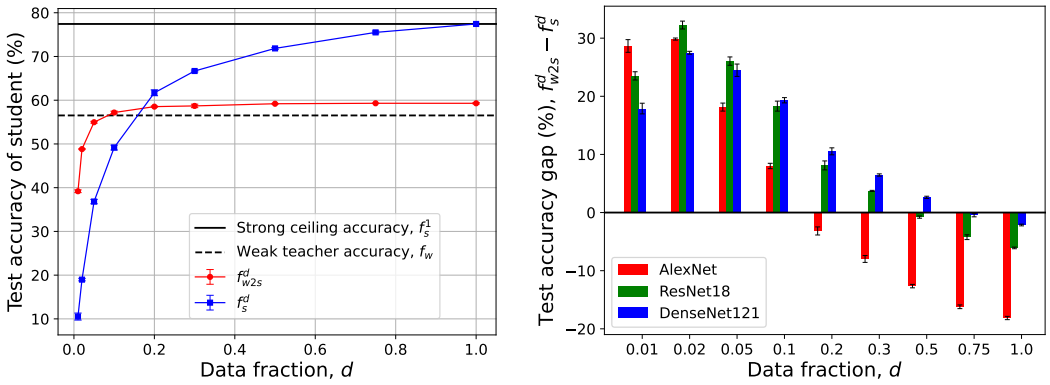
1 INTRODUCTION

Data-efficient learning seeks to train models effectively when the available training data or computational budget is limited. While many approaches focus on selecting or constructing compact sets of training inputs, their effectiveness also depends on the supervision attached to those inputs. This is particularly important in low-data regimes, where soft predictions generated by an expert teacher can encode similarities and uncertainty that are absent from one-hot labels, providing an effective regularization signal when only a small number of training examples is available (Qin et al., 2024; Baruch et al., 2025). These observations suggest that, in data-scarce regimes, the form and structure of supervision can be as consequential as the choice of training inputs.

However, these data-efficient learning methods that leverage model-generated supervision commonly rely on predictions from a highly capable expert teacher. Access to such a teacher may be limited, and generating its predictions can incur substantial computational cost. These considerations motivate exploring weaker teachers as an alternative source of supervision, leading to an initial question: *Can a strong student trained with limited data and supervision from a weak teacher still outperform its supervisor?* This question connects to weak-to-strong generalization(W2SG) (Burns et al., 2024). We investigate it under limited student data budgets, comparing weak supervision with direct hard-label training on the same examples and examining how their relative benefits change as more student data become available.

To explore this question, we first conduct controlled experiments on ImageNet-1K. We compare three models: a weak teacher trained on the full dataset, a strong student trained on a limited subset using soft labels from the weak teacher, and the same strong student trained on ground-truth hard labels using the identical subset. We observe two notable findings (Figure 1a). First, the weakly supervised student can outperform its teacher even when trained on 10% of the dataset. Second, in highly data-scarce regimes, weak supervision can also outperform ground-truth hard-label supervision on the same training examples. These results demonstrate that W2SG is achievable under a limited student data budget. However, the student's accuracy under this fixed weak supervisor saturates

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.



(a) Low-data W2SG

(b) Supervision-strength crossover

Figure 1: **Weak-to-strong generalization across student data scales on ImageNet-1K.** We use ResNet-50 as the student model. **(a)** Test accuracy of students trained using either hard labels or predictions from a fixed weak teacher (AlexNet) as the student data fraction varies. Horizontal lines indicate the weak-teacher accuracy and the full-data ResNet-50 accuracy. **(b)** Test-accuracy gain of weak supervision over hard-label training for AlexNet, ResNet-18, and DenseNet-121 teachers. Positive values indicate an advantage over direct hard-label supervision. As the student data budget increases, the preferred supervision shifts toward stronger weak teachers and eventually to hard labels.

early, while hard-label training continues to improve as more data become available. This raises the question of whether the preferred supervision source changes with the student data budget.

To address this question, we systematically vary both the student data size and the strength of the teacher, using a suite of supervisors with different predictive capabilities. Across the evaluated supervision sources, we observe a *data-size-dependent supervision crossover* in Figure 1b. When student data are extremely limited, supervision from weaker teachers yields the best performance. As the data budget increases, the preferred supervisor shifts toward progressively stronger teachers, and ground-truth hard labels eventually become optimal. Thus, the strongest available supervision is not uniformly the most effective; rather, the optimal supervision strength depends on the amount of data available to the student.

To investigate these empirical findings, we characterize the asymptotic risks of weak, strong, and weak-to-strong estimators in an over-parameterized two-stage linear regression model. Numerical evaluation reveals three main findings: (1) Weak-to-strong generalization can emerge even in low-data regimes, where the student surpasses its weak teacher using limited student training data. (2) The risk-minimizing teacher dimension depends on the student sample size: smaller teacher dimensions are preferable when data are scarce, while larger dimensions become preferable as more data are available. (3) There exists a regime in which the weak-to-strong model achieves lower risk than both the weak teacher and the strong model trained directly on ground-truth labels using the same student sample size. A bias–variance decomposition shows that variance reduction outweighs the additional bias in these regimes, while changes in the bias–variance balance with student data size explain the shift in the preferred teacher.

We next quantify how the observed accuracy gains translate into student-stage resource gains. We measure the hard-label training budget required to match the accuracy of a weakly supervised student, relative to the budget used by that student. We further account for online teacher inference cost by retraining students on reduced subsets under the same total compute budget. Substantial benefits remain in the low-budget regime after accounting for this cost, and the preferred supervision is largely preserved across the evaluated budgets.

Our main contributions are summarized as follows:

- We systematically study a data-size-dependent supervision crossover in weak-to-strong generalization. In our main ImageNet experiments, the best-performing supervision shifts

from weaker teachers to stronger teachers and eventually hard labels as student data increase. Experiments across additional vision and language settings further examine the data dependence of weak-supervision gains.

- We characterize the asymptotic risks of weak, strong, and weak-to-strong estimators in a two-stage linear regression model with a weak teacher estimated from training data. Numerical evaluation shows how the risk-minimizing teacher dimension changes with the student sample size and identifies regimes in which weak-to-strong training outperforms both the teacher and a student-sample-matched hard-label baseline. A bias–variance decomposition explains these gains through variance reduction.

Related work Baruch et al. (2025) study knowledge distillation on pruned datasets, showing that distillation improves student performance and that smaller teachers can be preferable at low retained data fractions. Our work investigates this data dependence in weak-to-strong learning. We compare supervision from teachers of different strengths with ground-truth supervision on the same student examples, examining when students surpass their teachers and how the preferred supervision changes as student data increase.

To analyze this data dependence, we build on the coordinate-restricted surrogate and two-stage regression analyses of Ildiz et al. (2025). We combine these elements by training a teacher that observes only a subset of input coordinates, while its training labels depend on all input coordinates. Under a diagonal Gaussian design, the omitted coordinates act as effective noise during teacher estimation, so the teacher incurs both approximation and estimation errors. Sharing the variance-reduction perspective of Dong et al. (2025), we examine how student sample size changes the balance between bias and variance across teacher dimensions, and how this balance determines the teacher yielding the lowest student risk. A more detailed review of related work is provided in Appendix A.

2 Data-Dependent Supervision in Weak-to-Strong Learning

We first show that weak supervision can outperform both the weak teacher and hard-label training when student data are scarce, but that its benefit saturates as more data become available. This observation motivates us to vary teacher strength. We find that the best-performing supervision shifts systematically toward stronger teachers and eventually to ground-truth labels as the student data size increases.

2.1 Weak-to-Strong Generalization in Data-Scarce Regimes

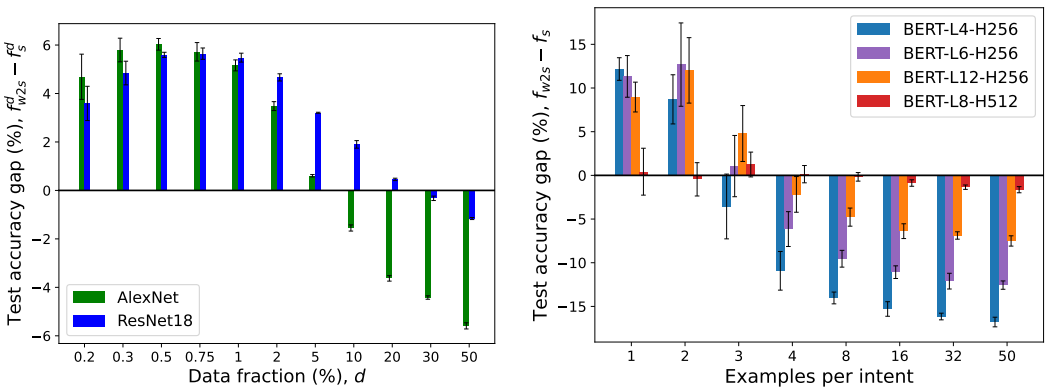
We study how weak-to-strong generalization (W2SG) changes with the amount of student training data. While W2SG—where a student trained on weak supervision surpasses its teacher—is well established (Burns et al., 2024), its behavior across data scales remains less understood. On ImageNet-1K, for each data fraction d, we train a student f^d_{w2s} using predictions from a fixed weak teacher f_w trained on the full dataset, and compare it with the same student f^d_s trained on ground-truth labels from the identical subset.

We make two observations. First, weak supervision improves generalization in the low-data regime: as shown in Figure 1a, f^d_{w2s} outperforms the corresponding hard-label student f^d_s for all evaluated data fractions below 20%. Second, with only 10% of the training data, the student surpasses its weak teacher, demonstrating W2SG under a limited student-data budget. However, performance under this fixed weak supervisor quickly saturates as more student data become available.

This early saturation raises a natural question: if more student data can no longer be exploited by a fixed weak teacher, should the supervision itself become stronger as the data budget increases?

2.2 A Data-Dependent Crossover in Supervision Strength

Motivated by the saturation of a fixed weak teacher, we next ask whether the preferred supervision strength changes with the amount of student data. We keep the student architecture fixed to ResNet-50 and consider three weak teachers of increasing predictive accuracy: AlexNet, ResNet-18, and DenseNet-121. We use each teacher's full-dataset test accuracy as a proxy for supervision strength.



(a) ImageNet linear probing

(b) CLINC150 classification

Figure 2: **Generality of data-dependent supervision across learning settings.** Test-accuracy gain of weak supervision over direct hard-label training as the student data budget varies. (a) Linear probing with a frozen DINO-pretrained ResNet-50 backbone on ImageNet-1K, using AlexNet and ResNet-18 as weak teachers. (b) Full fine-tuning of a pretrained BERT-Large student on CLINC150 using weak BERT teachers of varying capacity. Positive values indicate that weak supervision outperforms direct hard-label training. Across both settings, weak supervision is most beneficial in data-scarce regimes, while stronger supervision becomes increasingly favorable as the student data budget grows.

Figure 1b reveals a clear data-dependent crossover in the preferred supervision. At the lowest evaluated data fraction (1%), AlexNet provides the best student performance. As the data budget increases, the preferred supervisor shifts to ResNet-18 at 2–5% and DenseNet-121 at 10–50%, while direct ground-truth supervision becomes best at 75–100%. Thus, the strongest available supervision is not uniformly optimal: weaker supervision is preferable when student data are scarce, whereas stronger supervision becomes increasingly beneficial as more data become available.

Teacher strength also affects how much student data is required for W2SG itself. As summarized in Table 1, the student first surpasses AlexNet at the 10% evaluated fraction, ResNet-18 at 30%, and DenseNet-121 at 75%. Stronger teachers therefore require more student data to surpass, while weaker teachers reach their performance ceiling earlier. Together, these results show that both the occurrence of W2SG and the preferred supervision strength depend on the student data budget.

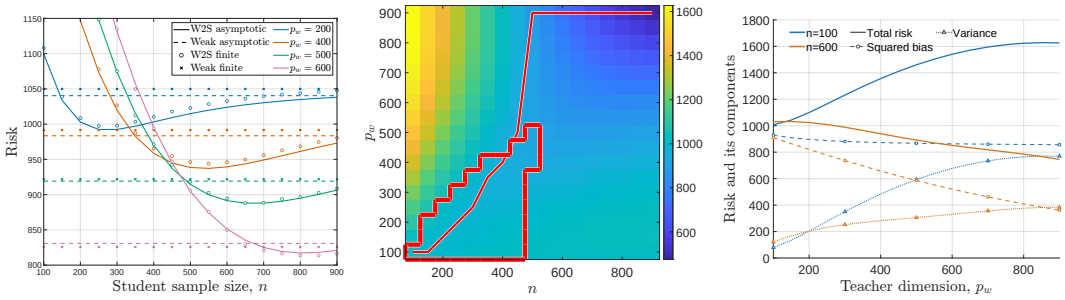
Table 1: ResNet-50 student test accuracy (%) on ImageNet-1K across supervision sources and data fractions. Bold values indicate the smallest evaluated data fraction at which each student surpasses its teacher.

Teacher	**Data Fraction**				**Teacher**
	10%	**30%**	**75%**	**100%**	**Perf.**
AlexNet	**57.08**	58.92	59.31	59.33	56.52
ResNet18	67.32	**70.36**	71.33	71.37	69.76
Densenet121	68.48	73.15	**75.04**	75.29	74.43
Hard	49.54	66.59	75.34	77.38	N/A

2.3 Generality Across Learning Settings and Domains

We next examine whether the data-size-dependent supervision shift observed in Section 2.2 persists beyond our main end-to-end ImageNet setting. We test robustness along two complementary axes: training protocol and task domain. First, we perform linear probing on frozen pretrained vision representations, retaining ImageNet while removing end-to-end feature learning. Second, we study BERT-based intent classification on CLINC150, changing both the model family and task domain.

For vision, we freeze a DINO-pretrained ResNet-50 backbone (Caron et al., 2021) and train only a linear classifier using either hard labels or predictions from weak teachers, following the setup of Burns et al. (2024). As shown in Figure 2a, the preferred supervision again changes with the student data budget: the weaker teacher performs better in the most data-scarce regime, the stronger weak



(a) W2S under various teacher (b) Risk of W2S model (c) Risk decomposition

Figure 3: **Theoretical analysis of the weak-to-strong generalization phenomenon.** (a) Generalization performance under various teacher capacities (p_w). (b) Excess risk comparison. The thin red line indicates the optimal p_w for a given n; the region enclosed by the thick red line highlights where the weak-to-strong model outperforms both the weak and strong baselines. Hyperparameters are set to $p = 1000, \alpha_s = 2, \alpha_w = 4, \boldsymbol{\beta}_{*,i} = i, \kappa_w = 1.25$. (c) Weak-to-strong risk and its squared bias and variance components as functions of p_w at $n = 100$ and $n = 600$, illustrating how the bias–variance tradeoff favors weaker teachers with limited student data and stronger teachers with more data.

teacher becomes preferable with more data, and hard labels eventually dominate. Thus, the crossover does not require end-to-end feature learning.

We observe a similar pattern beyond vision on CLINC150 (Larson et al., 2019). We use a pretrained BERT-Large model as the student and fine-tune all parameters using either ground-truth labels or predictions from weak BERT teachers of varying capacity. Figure 2b shows large gains from weak supervision with only a few examples per intent, but these gains diminish and eventually become negative as the student data increases. Together, the vision and NLP results provide evidence that the preferred supervision depends on the student data budget across different learning settings and domains.

3 Theoretical Analysis

We analyze the data-size-dependent supervision crossover observed in Section 2 using an overparameterized two-stage linear regression model based on Ildiz et al. (2025). The teacher is trained using only a subset of input coordinates, while its labels depend on all coordinates; under our diagonal Gaussian design, the omitted signal acts as effective noise during teacher estimation. By varying student sample size and teacher dimension, we numerically examine how the preferred teacher changes and when W2S outperforms both the weak and strong baselines under a common target risk.

3.1 Model and Asymptotic Framework

Problem setup We study weak-to-strong training in an over-parameterized two-stage linear regression. Let $\boldsymbol{\beta}_* \in \mathbb{R}^p$ be the ground-truth model. We consider three types of estimated models: (1) the strong model $\hat{\boldsymbol{\beta}}_s \in \mathbb{R}^p$, (2) the weak model $\hat{\boldsymbol{\beta}}_w \in \mathbb{R}^{p_w}$, and (3) the weak-to-strong model $\hat{\boldsymbol{\beta}}_{w2s} \in \mathbb{R}^p$. We define the strong model as a typical regression problem where the training data distribution is $(\boldsymbol{x}, y) \sim \mathcal{D}_s(\boldsymbol{\beta}_*)$ with $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_s)$ where $\boldsymbol{\Sigma}_s \in \mathbb{R}^{p\times p}$ and $y = \boldsymbol{x}^\top \boldsymbol{\beta}_*$. Its estimator is the minimum-ℓ_2 norm interpolator, $\hat{\boldsymbol{\beta}}_s := \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{\beta}\|_2 : \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}\} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ for sample matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with n independent samples and $\boldsymbol{y} \in \mathbb{R}^n$.

We assume that the weak teacher has a lower dimensionality $p_w < p$, meaning it can only access the first p_w coordinates of the ground-truth parameters. The data distribution for the weak model is $(\tilde{\boldsymbol{x}}, \tilde{y}) \sim \mathcal{D}_w(\boldsymbol{\beta}_*)$ with $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_w)$ where $\boldsymbol{\Sigma}_w \in \mathbb{R}^{p\times p}$ and $\tilde{y} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_*$. From this distribution, we draw m independent samples and collect their inputs as the rows of $\tilde{\boldsymbol{X}} \in \mathbb{R}^{m\times p}$, with responses $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta}_*$. The teacher uses only the first p_w input coordinates. We denote the corresponding first p_w columns of $\tilde{\boldsymbol{X}}$ by $\tilde{\boldsymbol{X}}_{p_w} \in \mathbb{R}^{m\times p_w}$ and define the weak estimator as $\hat{\boldsymbol{\beta}}_w = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{\beta}\|_2 : \tilde{\boldsymbol{X}}_{p_w}\boldsymbol{\beta} = \tilde{\boldsymbol{y}}\} = \tilde{\boldsymbol{X}}_{p_w}^\top(\tilde{\boldsymbol{X}}_{p_w}\tilde{\boldsymbol{X}}_{p_w}^\top)^{-1}\tilde{\boldsymbol{y}}$, where $\tilde{\boldsymbol{X}}_{p_w} \in \mathbb{R}^{m\times p_w}$ and $\tilde{\boldsymbol{y}} \in \mathbb{R}^m$.

Finally, the data distribution for the weak-to-strong model is $(\boldsymbol{x}, y^w) \sim \mathcal{D}_s(\hat{\boldsymbol{\beta}}_w^{\text{pad}})$ with $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_s)$ and $y^w = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_w^{\text{pad}}$, where the labels are generated by the zero-padded weak model, $\hat{\boldsymbol{\beta}}_w^{\text{pad}} := \left(\hat{\boldsymbol{\beta}}_w^\top, \mathbf{0}_{p-p_w}^\top\right)^\top \in \mathbb{R}^p$. The estimator is $\hat{\boldsymbol{\beta}}_{w2s} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{\beta}\|_2 : \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}^w\} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}^w$ with $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y}^w \in \mathbb{R}^n$. We evaluate the performance of these estimators using the excess risk, defined as $\mathcal{R}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|_2^2$.

Assumptions and theoretical framework To facilitate a tractable analysis of the excess risk, we adopt the following conditions: (i) *Diagonal Covariance*: The feature covariances are diagonal, $\boldsymbol{\Sigma}_w = \text{diag}(\lambda_1^w, \ldots, \lambda_p^w)$ and $\boldsymbol{\Sigma}_s = \text{diag}(\lambda_1^s, \ldots, \lambda_p^s)$, where eigenvalues follow a power-law decay ($\lambda_i^s = i^{-\alpha_s}, \lambda_i^w = i^{-\alpha_w}$ with $\alpha_s > 1, \alpha_w > 1$). (ii) *Proportional Limit*: We consider the regime $n, p, m, p_w \to \infty$ such that $p/n = \kappa_s > 1$ and $p_w/m = \kappa_w > 1$. Under this high-dimensional regime, we can characterize the distribution of each finite-sample estimator $\hat{\boldsymbol{\beta}}_\alpha$ using a corresponding equivalent Gaussian sequence model denoted as $\tilde{\boldsymbol{\beta}}_\alpha$ for $\alpha \in \{s, w, w2s\}$ (Han & Xu, 2026). For the weak estimator, we use its zero-padded representation when evaluating the risk. Under the appropriate conditions, it has been established that the finite-sample excess risk $\mathcal{R}(\hat{\boldsymbol{\beta}}_\alpha)$ converges in probability to a deterministic asymptotic risk $\bar{\mathcal{R}}_\alpha = \mathbb{E}_{\tilde{\boldsymbol{\beta}}_\alpha}[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*)\|_2^2]$ (Ildiz et al., 2025). Let β_i denote the i-th component of $\boldsymbol{\beta}_*$. For example, the asymptotic risk of the strong estimator is

$$\bar{\mathcal{R}}_s = B_s + \frac{\Omega_s}{1-\Omega_s} B_s, \qquad B_s := \sum_{i=1}^{p} \frac{\tau_s^2 \lambda_i^s}{(\lambda_i^s + \tau_s)^2} \beta_i^2 \tag{1}$$

where B_s means bias term, $\tau_s > 0$ is the unique solution to the first equation below, and Ω_s is defined by the second:

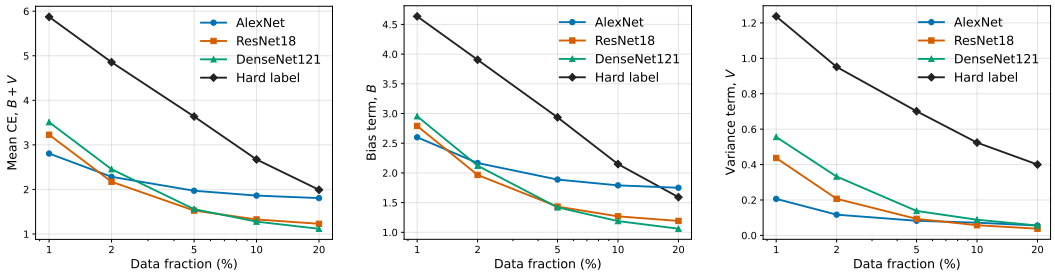
$$n = \sum_{i=1}^{p} \frac{\lambda_i^s}{\lambda_i^s + \tau_s}, \qquad \Omega_s := \frac{1}{n}\sum_{i=1}^{p} \left(\frac{\lambda_i^s}{\lambda_i^s + \tau_s}\right)^2 \tag{2}$$

The corresponding expressions for the weak and weak-to-strong estimators, together with the full distributional framework, are provided in Appendix D, Eqs. 67 and 70. The theoretical learning curves below are obtained by numerically evaluating $\bar{\mathcal{R}}_s$, $\bar{\mathcal{R}}_w$, and $\bar{\mathcal{R}}_{w2s}$.

3.2 Data-Dependent Weak-to-Strong Generalization in Linear Regression

Weak-to-strong generalization occurs in the low-data regime. In Figure 3a, we examine weak-to-strong generalization across weak teachers parameterized by their accessible dimension p_w. With $\kappa_w = p_w/m$ held fixed, increasing p_w also increases the teacher training sample size m and yields a lower-risk teacher in the setting examined. We highlight two main observations. First, the weak-to-strong model can outperform its weak teacher even with limited student data, although its performance eventually saturates as the sample size increases. Second, the amount of student data required to surpass the weak teacher increases as the teacher becomes stronger. For the finite-dimensional simulations, we independently sample 100 teacher training sets for each p_w and 20 student training sets for each n, training one weak or strong model on each corresponding set. For each (n, p_w) configuration, we additionally train 2,000 weak-to-strong models using every pairing of the 100 teachers and 20 student training sets. We report the average risk for each model type. The finite-dimensional simulations closely follow the asymptotic predictions, and the resulting trend is consistent with the empirical observations in Section 2.2.

The optimal weak supervision depends on the data size. Across a grid of (n, p_w) values, Figure 3b plots the weak-to-strong risk, where the thin red line indicates the risk-minimizing teacher dimension for each n. Across the evaluated grid, the risk-minimizing teacher dimension $p_w^\star(n)$ is small when n is small, and it shifts to larger values as n increases. This demonstrates that in the data-scarce regime, training the student with labels from a weaker teacher actually achieves a lower risk, thereby validating the empirical findings in Section 2.2. Furthermore, the region bounded by the thick red line indicates where the weak-to-strong model achieves lower risk than both its weak teacher and the strong student trained with the same sample size n ($\bar{\mathcal{R}}_{w2s} < \bar{\mathcal{R}}_w$ and $\bar{\mathcal{R}}_{w2s} < \bar{\mathcal{R}}_s$). These results highlight how student data size and teacher strength jointly shape weak-to-strong performance: weaker teachers are preferable when student data are scarce, while stronger teachers become more effective as the student data size increases.



(a) CE loss (b) Bias (c) Variance

Figure 4: **Empirical bias–variance analysis on ImageNet-1K.** (a) Mean evaluation CE exhibits a data-dependent crossover among teachers. (b) The weakest teacher yields the lowest bias with the least data, while stronger teachers yield lower bias as data increase. (c) Teacher supervision yields lower variance than hard-label training. Weaker teachers have lower variance when data are scarce, whereas all teachers yield similar variance levels at larger data fractions.

3.3 Data-Dependent Bias–Variance Tradeoff

To understand how the preferred teacher depends on the student data size, we decompose the asymptotic risk into squared bias and variance for $\alpha \in \{s, w, w2s\}$:

$$\bar{\mathcal{R}}_\alpha = \mathbb{E}_{\tilde{\boldsymbol{\beta}}_\alpha}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*)\|_2^2\right] \tag{3}$$

$$= \underbrace{\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\mathbb{E}_{\tilde{\boldsymbol{\beta}}_\alpha}[\tilde{\boldsymbol{\beta}}_\alpha] - \boldsymbol{\beta}_*)\|_2^2}_{\text{Squared bias}} + \underbrace{\mathbb{E}_{\tilde{\boldsymbol{\beta}}_\alpha}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_\alpha - \mathbb{E}_{\tilde{\boldsymbol{\beta}}_\alpha}[\tilde{\boldsymbol{\beta}}_\alpha])\|_2^2\right]}_{\text{Variance}}. \tag{4}$$

Figure 3c shows how the weak-to-strong model's squared bias, variance, and total risk vary with the teacher dimension p_w at two student sample sizes, $n = 100$ and $n = 600$. When student data are scarce ($n = 100$), increasing the teacher dimension yields only a modest bias reduction but substantially increases the variance over much of the evaluated range. The variance increase outweighs the bias improvement, favoring weaker teachers. With more student data ($n = 600$), increasing the teacher dimension produces a larger bias reduction and a smaller variance increase, making stronger teachers preferable. These results illustrate a data-dependent bias–variance tradeoff: weaker teachers can be preferable when student data are scarce because of their lower student variance, whereas additional student data allow the bias advantage of stronger teachers to outweigh the variance cost. We next examine whether weak supervision also reduces variance in deep neural networks.

3.4 Connection to Deep Networks

Our linear analysis highlights the role of variance reduction in the low-data benefit of weaker supervision. We next examine how bias and variance vary with supervision source and student data size in deep networks.

Let s denote the supervision source, either hard labels or predictions from a given teacher. For each student data fraction d, let $\mathcal{T}^{(d,s)}$ denote the randomly sampled student training set with targets provided by s, and let $f(\boldsymbol{x}; \mathcal{T}^{(d,s)})$ denote the class-probability vector predicted for input $\boldsymbol{x}$ by the student trained on $\mathcal{T}^{(d,s)}$. We evaluate all supervision conditions against the same ground-truth labels, with $\boldsymbol{y}$ denoting the one-hot encoding of class y. Let $\mathcal{D}$ denote the evaluation distribution over inputs and ground-truth labels. The expected cross-entropy loss is defined as

$$\mathcal{E}_{\mathrm{CE}}(d, s) := \mathbb{E}_{\mathcal{T}^{(d,s)}}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[-\sum_{c=1}^{C} y_c \log f_c(\boldsymbol{x}; \mathcal{T}^{(d,s)})\right] \tag{5}$$

The expectation over training runs also includes randomness in model initialization and optimization, which is suppressed in the notation. Following Zhou et al. (2021), we decompose the expected

cross-entropy loss as follows (see Appendix D.4 for the derivation):

$$\mathcal{E}_{\mathrm{CE}}(d,s) = \underbrace{\mathbb{E}_{\boldsymbol{x},y}\left[D_{\mathrm{KL}}\left(\boldsymbol{y} \,\|\, \bar{f}^{(d,s)}(\boldsymbol{x})\right)\right]}_{\mathrm{Bias}_{\mathrm{CE}}} + \underbrace{\mathbb{E}_{\mathcal{T}^{(d,s)},\boldsymbol{x}}\left[D_{\mathrm{KL}}\left(\bar{f}^{(d,s)}(\boldsymbol{x}) \,\|\, f(\boldsymbol{x};\mathcal{T}^{(d,s)})\right)\right]}_{\mathrm{Var}_{\mathrm{CE}}} \quad (6)$$

where $\bar{f}^{(d,s)}(\boldsymbol{x}) = \mathrm{softmax}\big(\mathbb{E}_{\mathcal{T}^{(d,s)}}[\log f(\boldsymbol{x};\mathcal{T}^{(d,s)})]\big)$ is the normalized geometric mean of predictions across training runs, with the logarithm applied elementwise. Following the main ImageNet-1K training configuration, we train one student on each of five disjoint subsets per supervision condition at data fractions $d \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ and estimate the decomposition over these runs and the ImageNet-1K validation set.

Figure 4 shows a data-dependent crossover in CE loss among teachers, qualitatively similar to the accuracy trends in Section 2. With very limited data, the weakest teacher yields the lowest loss, whereas stronger teachers become preferable as the data fraction increases.

Students trained with weak-teacher supervision exhibit lower predictive variance than those trained with hard labels across all evaluated data fractions. Among teachers, weaker supervision tends to yield lower student variance when data are scarce, whereas the teachers yield similar levels of student variance at larger data fractions.

The bias behavior is more nuanced. Unlike in the linear-model settings examined above, a stronger teacher does not necessarily yield lower bias when student data are scarce; the weakest teacher also produces the lowest bias at the smallest data fraction. As more data become available, the bias under stronger-teacher supervision decreases more substantially and falls below that under weaker supervision. Thus, the low-data advantage of weaker supervision involves both lower bias and lower variance, while the advantage of stronger supervision at larger data fractions is primarily associated with lower bias, as variance differences become small.

4 Quantifying Resource Gains from Weak Supervision

Sections 2 and 3 investigate how the preferred supervision depends on the student data budget. We now express the magnitude of this low-data advantage in units of student-training budget: how much hard-label training budget is required to match the accuracy achieved with weak supervision, and how do these gains change when teacher inference cost is included? We condition on access to pretrained teachers and pre-existing ground-truth labels, and account only for resources consumed during student training. Accordingly, our analysis measures student-stage resource efficiency rather than the end-to-end cost of teacher training or label acquisition.

4.1 Hard-Label-Equivalent Budget Multipliers

We normalize the compute required to train the student on the full dataset with hard labels to a budget of 100%. We first exclude teacher inference cost. Under our fixed training recipe, student training compute is proportional to the number of training examples. For a teacher t and a budget b, let $A_t(b)$ denote the accuracy of the teacher-supervised student, and let $\hat{b}_h(a)$ denote the estimated hard-label training budget required to attain accuracy a. We define the hard-label-equivalent budget multiplier as $M_t(b) := \frac{\hat{b}_h(A_t(b))}{b}$. A multiplier greater than one indicates that hard-label training requires more budget to attain the same accuracy. For each seed, we estimate the matching hard-label budget $\hat{b}_h(a)$ by linearly interpolating log-budget between the two adjacent hard-label measurements that bracket the target accuracy a. We then compute the budget multipliers and report their mean and sample standard deviation across two seeds.

The solid curves in Figure 5 show substantial budget gains in the low-data regime. For example, at a 1% budget, AlexNet supervision achieves an accuracy that would require an estimated 5.6% budget with hard-label training, corresponding to a multiplier of approximately $5.6\times$. The gains depend on both the teacher and the available budget. Students trained with weaker teachers exhibit early accuracy saturation, causing their multipliers to decline as the budget increases. Stronger teachers retain an advantage over a broader budget range. At sufficiently large budgets, the multipliers fall below one, indicating that hard-label training becomes more resource-efficient.

4.2 Accounting for Teacher Inference Cost

We next include the cost of querying the teacher during student training. Let F_s and F_t denote the forward-pass FLOPs of the student and teacher, respectively. We approximate the student backward pass as costing $2F_s$ and query the teacher once per training example at each training step. The resulting cost per example is therefore $3F_s + F_t$, compared with $3F_s$ for hard-label training. To keep the total compute budget fixed while retaining the same training recipe, we reduce the student training set by a factor of $r_t := \frac{3F_s}{3F_s+F_t}$. This retains 94.53%, 87.15%, and 81.15% of the corresponding cost-excluded training set for AlexNet, ResNet18, and DenseNet121, respectively. We retrain the students on these reduced subsets. Let $A_t^{\text{cost}}(b)$ denote the resulting accuracy at total budget b. We compute the cost-aware multiplier as $M_t^{\text{cost}}(b) := \frac{\hat{b}_h(A_t^{\text{cost}}(b))}{b}$ where the denominator includes both student training and teacher inference costs.

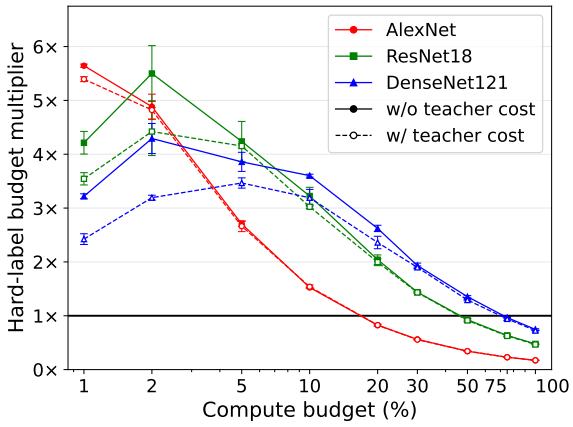


Figure 5: **Hard-label-equivalent budget gains.** Results for a ResNet50 student on ImageNet-1K. Values above $1\times$ indicate greater budget efficiency than hard-label training at matched accuracy. Solid and dashed curves exclude and include teacher **inference** cost, respectively.

The dashed curves in Figure 5 show that teacher inference cost generally reduces the multipliers, particularly for the more expensive teachers at small budgets. Nevertheless, substantial gains remain in the low-budget regime. Among the seven evaluated budgets at which weak supervision achieves the highest mean accuracy when teacher inference cost is excluded, six retain the same preferred teacher after accounting for this cost. At the 2% budget, the preferred teacher changes from ResNet18 to AlexNet, whose lower inference cost permits training on more examples. These results show that, under the evaluated training recipe and cost model, weak supervision can provide student-stage resource gains even after accounting for the cost of generating its labels.

5 Conclusion

Our study shows that the preferred supervision source in weak-to-strong learning depends on the student data budget. In our main experiments, weaker teachers are preferable when student data are scarce, and students can surpass their teachers even with limited training data. As student data increase, performance under a fixed weak teacher saturates, and stronger teachers and eventually hard labels become preferable. Numerical analysis of an overparameterized two-stage linear regression model illustrates how this shift arises from a data-dependent bias–variance tradeoff: lower variance favors weaker teachers at small sample sizes, whereas the bias reduction offered by stronger teachers becomes more beneficial with additional data. Finally, we quantify these benefits through hard-label-equivalent budget multipliers and find that substantial student-stage resource gains remain after accounting for teacher inference cost.

Limitation and Future work Weak supervision does not necessarily improve student performance in all low-data settings. Identifying when weak supervision helps or hurts remains an important direction for future work. Next, our linear analysis also captures only part of the behavior observed in deep networks, where weaker supervision can reduce both bias and variance in the lowest-data regime. Explaining this bias behavior in models that incorporate representation learning remains an open question. Finally, identifying the best supervision source for a given student data budget currently requires comparing multiple student training runs in our experiments. Developing a low-cost method for selecting among teachers and hard-label supervision remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with providing feedback on experimental methodology, implementing methods and running experiments, cleaning and reformatting datasets, interpreting results, and checking mathematical claims and writing derivations. We also used Qwen2.5-Instruct models to generate synthetic training responses for the mathematical reasoning experiments, as described in the experimental details. We have not used generative AI tools to propose or refine hypotheses or to provide critical ingredients for proving mathematical claims. Qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft and revise manuscript text, review related literature, write numerical analysis and visualization code, and refine the paper's structure, title, and keywords. We have reviewed the AI-assisted text, code, and mathematical derivations and take responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

REFERENCES

Emanuel Ben Baruch, Adam Botach, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. Distilling the knowledge in data pruning. In *Forty-second International Conference on Machine Learning*, 2025.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5977–6045. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/busbridge25a.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126474–126499. Curran Associates, Inc., 2024. doi: 10.52202/079017-4017. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4a0d8aef3567f742b0794844d9b5847-Paper-Conference.pdf.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14079–14113. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/dong25g.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *Journal of Machine Learning Research*, 27(23):1–94, 2026.

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

Muhammed Ildiz, Halil Gozeten, Ege Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations*, volume 2025, pp. 27181–27220, 2025.

Nohyun Ki, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XIzO8zr-WbM.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HOSh0SKklE.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, 2019.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43519–43556. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/medvedev25a.html.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pp. 7632–7642. PMLR, 2021.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong generalization through feature learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 169216–169285. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f7977daf15f87e19b56e4d2d363b9002-Paper-Conference.pdf.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf.

Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 131946–131971. Curran Associates, Inc., 2024. doi: 10.52202/079017-4194. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee45939f4403e8abd63a15a29a9c055b-Paper-Conference.pdf.

Qihan Ren, Wang Peng, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://openreview.net/forum?id=gsLVyP22QV.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

David Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *International Conference on Learning Representations*, volume 2025, pp. 88280–88316, 2025.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70120–70159. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/xue25d.html.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3902–3910, 2020. doi: 10.1109/CVPR42600.2020.00396.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gIHd-5X324.

A Extended related Work

Weak-to-strong generalization Weak-to-strong generalization (W2SG) studies whether a strong model trained from weak supervision can outperform the weak supervisor itself (Burns et al., 2024). A growing theoretical literature has explained this phenomenon through pseudolabel correction, model misfit, benign overfitting, regularization, representation geometry, and feature learning (Lang et al., 2024; Charikar et al., 2024; Wu & Sahai, 2025; Medvedev et al., 2025; Xue et al., 2025; Oh et al., 2025).

Among the works most closely related to our theoretical setup, Ildiz et al. (2025) analyze knowledge distillation in high-dimensional regression, characterizing risks under surrogate supervision. They study both data-dependent optimal surrogates and two-stage learning with teachers estimated from data. Building on their coordinate-restricted surrogate and two-stage regression analyses, we study teachers fitted to the full response using only a subset of input coordinates. Under a diagonal Gaussian design, the omitted coordinates act as effective noise during teacher estimation. We use this setting to examine how student data size affects the preferred teacher dimension and the associated bias–variance tradeoff.

Dong et al. (2025) explain W2SG through variance reduction, showing that discrepancies between weak and strong feature subspaces allow the student to attenuate part of the weak teacher's variance. Their strong SFT baseline uses the labeled data employed to train the weak teacher, whereas ours uses the same number of training examples as the weak-to-strong student, allowing us to compare weak and ground-truth supervision at a matched student data budget. Sharing their variance-reduction perspective, we focus on how student sample size changes the bias–variance tradeoff across teacher dimensions and, consequently, the preferred teacher.

Knowledge distillation and soft supervision Knowledge distillation was originally introduced as a method for transferring knowledge from a strong teacher to a weak student (Hinton et al., 2015). Subsequent work showed that distillation can improve a model even when the teacher and student share the same architecture (Furlanello et al., 2018). Later analyses connect such gains to the regularization induced by soft targets: KD can act as learned label smoothing, while imperfect teacher probabilities can reduce variance at the cost of introducing bias (Yuan et al., 2020; Menon et al., 2021; Zhou et al., 2021). This regularization perspective provides a way to understand why supervision from a weaker teacher can benefit a stronger student, despite the errors in the teacher's predictions.

Beyond explaining why distillation can help, prior work has also examined which teachers provide effective supervision. Studies of teacher–student capacity gaps show that more accurate or higher-capacity teachers need not produce better students (Cho & Hariharan, 2019; Mirzadeh et al., 2020). These findings motivate considering teacher effectiveness in relation to the student's training conditions. We focus on student data size in weak-to-strong learning: holding the student architecture fixed, we examine how the preferred supervision changes among teachers of different strengths and ground-truth labels as more student data become available. All supervision sources are compared using the same student training examples.

Data-efficient and resource-constrained learning Data selection and dataset pruning improve efficiency by retaining informative subsets of a larger dataset (Toneva et al., 2019; Paul et al., 2021; Ki et al., 2023; Sener & Savarese, 2018; Mirzasoleiman et al., 2020), whereas dataset distillation constructs compact training sets using synthetic examples and optimized supervision. Recent work highlights that the labels attached to selected examples can be as important as the examples themselves. Qin et al. (2024) show that randomly selected real images paired with expert-generated soft labels can rival substantially more sophisticated dataset-distillation pipelines. Baruch et al. (2025) show that knowledge distillation from a teacher trained on the full dataset improves student performance on pruned subsets. They further find that smaller teachers can be preferable at low retained data fractions and that the preferred distillation weight depends on the retained fraction. Our work investigates this data dependence in weak-to-strong learning. We compare supervision from teachers of different strengths with ground-truth supervision on the same student examples, examining when students surpass their teachers and how the preferred supervision changes as student data increase.

Resource-aware distillation has also been studied through scaling laws. Busbridge et al. (2025) analyze compute-optimal distillation under different teacher training and inference cost assumptions. In our fixed-recipe experiments, we account for teacher inference cost to evaluate whether the student-stage resource gains from weak supervision persist under a fixed compute budget.

B Experimental Details

B.1 Main ImageNet Experiments

Dataset and sampling We use the ImageNet-1K training set and construct student training subsets at data fractions $d \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0\}$. We repeat the main ImageNet experiments using three random seeds. For each seed, we independently shuffle the examples within each class and retain the first $\lfloor dN_c \rfloor$ examples from class c, where N_c is the number of training examples in that class. Consequently, the subsets are nested across data fractions within each seed. For a given seed and data fraction, all supervision conditions use the same student training subset.

Teacher and student models We use AlexNet, ResNet-18, and DenseNet-121 initialized with their ImageNet-1K pretrained weights provided by torchvision as teachers. The teacher models are kept frozen and in evaluation mode throughout student training. The student is a randomly initialized ResNet-50 trained from scratch.

Training objectives We use cross-entropy loss for hard-label training and KL-divergence loss for teacher-supervised training. In both settings, we apply random resized cropping to 224×224, random horizontal flipping, and ImageNet normalization. We then apply mixup to the resulting inputs. For hard-label training, the targets are mixed using the same mixup weights as the inputs. For teacher-supervised training, the teacher is queried online on each mixed input to generate the target probability distribution. Teacher logits are converted to probabilities using a softmax with temperature $T = 1$, unless otherwise specified.

Training configuration and evaluation We train the student for 200 epochs using AdamW with an initial learning rate of 10^{-3}, a cosine annealing learning-rate schedule, a weight decay of 10^{-2}, and a batch size of 256. For mixup, we sample the mixing weight from $\mathrm{Beta}(0.8, 0.8)$. We report top-1 accuracy on the ImageNet-1K validation set using the final-epoch checkpoint. For evaluation, images are resized to 256 pixels and center-cropped to 224×224, followed by ImageNet normalization.

B.2 Image Classification on CIFAR-100

Dataset and sampling We use CIFAR-100, which contains 50,000 training images and 10,000 test images from 100 classes. All teachers are trained on the full training set. For student training, we construct class-balanced subsets with data fractions $d \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0\}$, corresponding to $\{5, 10, 25, 50, 100, 150, 250, 375, 500\}$ training images per class. We conduct three runs using paired data-sampling and training seeds. Within each run, we generate a fixed random permutation of the examples in each class and construct the subsets from prefixes of these permutations; the subsets are therefore nested across data fractions. At each data fraction and seed, all supervision conditions use the same student training subset and training seed. Since the teachers are trained on the full training set, their training data are not disjoint from the student subsets.

Teacher and student models We use three weak-teacher architectures: LeNet, ConvNet-D3 (a three-block, width-128 convolutional network without normalization), and VGG-13 with batch normalization. Each teacher is randomly initialized and trained from scratch on the full CIFAR-100 training set. For the single-teacher experiments, we train one teacher per architecture and reuse its final-epoch checkpoint across all student data fractions and runs. The student is a ResNet-18 adapted to 32×32 CIFAR images, with a 3×3 stride-one input convolution and no initial max-pooling layer. It is randomly initialized and trained from scratch.

Training objectives We use cross-entropy loss for hard-label training and KL-divergence loss for teacher-supervised training. In both settings, we apply random cropping to 32×32 with four-pixel

padding, random horizontal flipping, and CIFAR-100 normalization. We then apply mixup to the resulting inputs. For hard-label training, the targets are mixed using the same mixup weights as the inputs. For teacher-supervised training, the teacher is queried online on each mixed input to generate the target probability distribution. Teacher logits are converted to probabilities using a softmax with temperature $T = 1$, unless otherwise specified.

Training configuration and evaluation We train each teacher for 200 epochs using hard-label cross-entropy without MixUp. We use SGD with an initial learning rate of 0.1, momentum 0.9, a weight decay of 5×10^{-4}, a batch size of 128, and cosine annealing without warmup. We use the checkpoint after the final epoch rather than selecting a checkpoint based on test performance. We train students for 400 epochs using AdamW with an initial learning rate of 10^{-3}, a weight decay of 10^{-2}, a batch size of 128, and cosine annealing without warmup. For mixup, we sample the mixing weight from $\mathrm{Beta}(0.8, 0.8)$. We evaluate the final-epoch student on the official CIFAR-100 test set and report the mean and sample standard deviation of top-1 accuracy over the three paired runs.

B.3 Linear probing on Frozen Visual Representations

Dataset and sampling We use the ImageNet-1K training set and construct student training subsets at data fractions $d \in \{0.002, 0.003, 0.005, 0.0075, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$. We repeat the experiments using two random seeds. For each seed, we independently shuffle the examples within each class and retain the first $\lfloor dN_c \rfloor$ examples from class c, where N_c is the number of training examples in that class. Consequently, the subsets are nested across data fractions within each seed. For a given seed and data fraction, all supervision conditions use the same student training subset.

Teacher and student models We use ImageNet-1K-pretrained AlexNet and ResNet-18 models from torchvision as weak teachers. The student consists of a frozen DINO-pretrained backbone (ResNet-50 or ViT-B/8) and a randomly initialized 1000-way linear classification head. The teacher and DINO backbone are kept frozen and in evaluation mode throughout training; only the linear classification head is optimized.

Training objectives We use cross-entropy loss for hard-label training and soft-target cross-entropy loss for teacher-supervised training. For teacher supervision, the frozen teacher is queried online on each augmented input to generate the target probability distribution using a softmax with temperature $T = 1$. No mixup or other target-mixing augmentation is used.

Training configuration and evaluation We train only the randomly initialized linear classification head while keeping the DINO-pretrained backbone frozen. The linear head is trained for 20 epochs with the ResNet-50 backbone and for 5 epochs with the ViT-B/8 backbone. In both cases, we use Adam with an initial learning rate of 10^{-3}, zero weight decay, a batch size of 128, and a cosine annealing learning-rate schedule without warmup. We report top-1 accuracy on the ImageNet-1K validation set using the final-epoch checkpoint.

B.4 Intent Classification on CLINC150

Dataset and sampling We use the official full version of CLINC150 and restrict the task to its 150 in-scope intents, excluding all out-of-scope examples. The official splits contain 100 training, 20 validation, and 30 test utterances per intent, for a total of 15,000 training, 3,000 validation, and 4,500 test examples. We randomly partition the 100 training utterances of each intent into disjoint teacher and student pools containing 50 examples each, keeping this partition fixed across all runs. All weak teachers are trained on the same class-balanced subset of 32 examples per intent drawn from the teacher pool; the remaining 18 teacher-pool examples per intent are not used for student training. From the disjoint student pool, we construct class-balanced subsets with $K \in \{1, 2, 3, 4, 8, 16, 32, 50\}$ examples per intent. We conduct five runs, each with a different pair of data-sampling and training seeds. Within each run, subsets are constructed from prefixes of a fixed per-intent permutation and are therefore nested across K. All supervision conditions use the same student examples and training seed within each run.

Teacher and student models We use four pretrained uncased BERT variants as weak teachers: BERT-L4-H256-A4, BERT-L6-H256-A4, BERT-L12-H256-A4, and BERT-L8-H512-A8. Each teacher is equipped with a newly initialized 150-way classification head and fine-tuned end-to-end on the same teacher subset. For each teacher architecture, we train a single teacher using a fixed teacher–student partition, teacher subset, and training seed, and reuse its validation-selected checkpoint across all values of K and all five student runs. The student is a pretrained BERT-Large-Uncased model with a newly initialized 150-way classification head and is also fine-tuned end-to-end.

Inputs and training objectives The model input consists only of the English utterance; intent names, intent descriptions, and prompts are not provided. We tokenize inputs using the uncased BERT tokenizer, truncate sequences to at most 64 tokens, and pad each batch to the length of its longest sequence. We do not apply token masking or other explicit data augmentation. We use cross-entropy loss for hard-label training and soft-target cross-entropy loss for teacher-supervised training. For teacher supervision, the frozen teacher is queried online on each input to generate the full 150-way target probability distribution using a softmax with temperature $T = 1$.

Training configuration and evaluation We fine-tune each teacher for 20 epochs using AdamW with a peak learning rate of 5×10^{-5}, a weight decay of 10^{-2}, a batch size of 64, and gradient-norm clipping at 1.0. The learning rate increases linearly during the first 10% of training steps and then decays linearly to zero. Teachers are trained using bfloat16 mixed precision and evaluated with a batch size of 128. We evaluate on the official validation split after every epoch and select the checkpoint with the lowest validation negative log-likelihood. We fine-tune students for 50 epochs using AdamW with a peak learning rate of 2×10^{-5}, a weight decay of 10^{-2}, a batch size of 32, and the same gradient-clipping threshold and learning-rate schedule. Students are also trained using bfloat16 mixed precision and evaluated with a batch size of 128. We evaluate students on the validation split at initialization and every five epochs, and select the checkpoint with the lowest validation negative log-likelihood. We report top-1 accuracy on the official test split as the mean and sample standard deviation over the five paired runs.

B.5 Binary classification on SCiQ

Dataset construction and sampling We use the official SciQ training and test splits, containing 11,679 and 1,000 questions, respectively; the official validation split is not used. We convert each question into one binary classification example. With probability $1/2$, the candidate answer is the correct answer and the label is positive; otherwise, one of the three distractors is sampled uniformly and the label is negative. The input is formatted as `Q: <question> A: <candidate answer>`; the support passage and the remaining answer choices are not provided. The binary examples for both splits are generated once and kept fixed across all experiments.

We randomly divide the transformed training set into disjoint teacher and student pools containing 5,839 and 5,840 examples, respectively. All weak teachers are trained on the complete teacher pool. From the student pool, we construct ground-truth-stratified subsets with $K \in \{32, 48, 64, 128, 256, 512, 1024, 2048\}$ total examples, approximately preserving the binary class proportions of the student pool. All supervision conditions use the same selected examples and their ordering for a given K and data-sampling seed. We use 60 data-sampling seeds for $K \in \{32, 48, 64\}$ and 15 for the remaining values of K. For each seed, the subsets are nested across the values of K evaluated with that seed.

Teacher and student models We use pretrained GPT-2, GPT-2-large, and Qwen-1.8B models as weak teachers and pretrained Qwen-7B as the student. Each model is equipped with a bias-free two-way linear classification head applied to the hidden state of the final non-padding token. The classification head is initialized to zero, and the complete model, including the pretrained backbone, is fine-tuned end-to-end. For each teacher architecture, a single teacher is trained once on the fixed teacher pool, and its final-epoch model is reused for all student subset sizes and data-sampling seeds.

Training objectives After training, each teacher generates a two-class probability distribution for every example in the student pool using a softmax with temperature $T = 1$. These distributions are stored and reused during student training. For weak supervision, we minimize soft-target cross-

entropy,

$$\mathcal{L}_{\text{weak}} = -\sum_{c=1}^{2} q_{t,c}(x) \log p_{s,c}(x) \tag{7}$$

where q_t is the stored teacher distribution and p_s is the student's predicted probability distribution. The ground-truth baseline uses cross-entropy with one-hot targets. We do not apply explicit data augmentation.

Training configuration and evaluation We fine-tune the teachers for two epochs using Adam with an effective batch size of 32, zero weight decay, and a cosine learning-rate schedule without warmup. We use an initial learning rate of 5×10^{-5} for GPT-2 and 10^{-5} for GPT-2-large and Qwen-1.8B. We fine-tune the Qwen-7B student for 10 epochs using Adam with a learning rate of 10^{-5}, an effective batch size of 32, zero weight decay, and the same cosine schedule. Dropout is disabled throughout training.

We do not use a validation split or validation-based checkpoint selection. We evaluate the final-epoch student on the fixed binary transformation of the official SciQ test split and report top-1 accuracy. Results are reported as the mean and sample standard deviation across data-sampling seeds, which vary the selected student subset. We do not perform additional repetitions over model-training seeds.

B.6 Mathematical reasoning

Dataset and sampling We use the training split of Math-CoT-20k(Ren et al., 2026), which contains 20,480 problem–solution pairs. We retain examples whose `response_length` values fall between 2,048 and 4,096, inclusive, leaving 4,906 examples. We construct training subsets containing $n \in \{60, 120, 240, 480, 720, 1200\}$ problems by taking the first n examples in the filtered dataset order. The resulting subsets are deterministic and nested. All supervision conditions use the same ordered problems at each n.

Teacher and student models The student is the pretrained base Qwen2.5-7B model and is fine-tuned end-to-end. We use Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, and Qwen2.5-3B-Instruct as fixed weak teachers. For each problem and teacher, we generate 16 candidate solutions using temperature 0.6, top-$p = 0.95$, top-$k = 20$, and at most 8,192 generated tokens.

We extract the contents of the final `\boxed{}` expression in each candidate and identify the majority answer by exact string matching. Among the candidates producing this answer, we retain the solution having the highest average log probability over its generated tokens. This produces one fixed weak solution per problem and teacher, which is reused across all student training-set sizes. Weak solutions are not filtered based on their correctness. As a reference-supervision condition, we instead use the original Math-CoT solution.

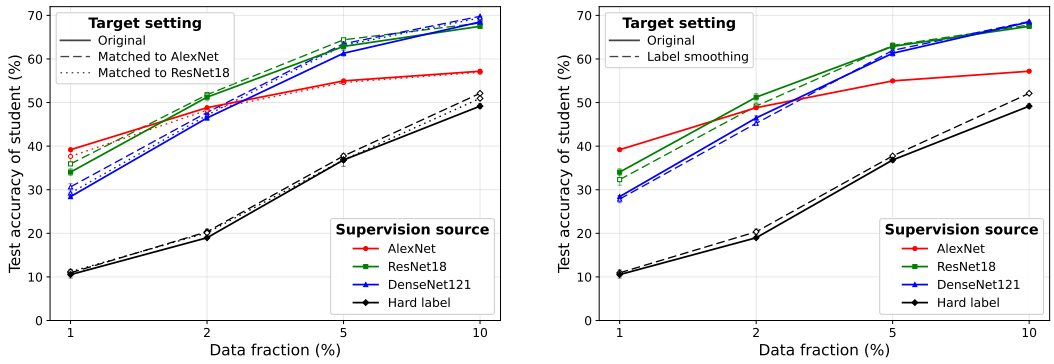
Inputs and training objective The user prompt consists of the mathematical problem followed by "Please reason step by step, and put your final answer within `\boxed{}`." We apply the Qwen chat format and exclude prompt and padding tokens from the loss. Sequences are dynamically right-padded within each batch and are not explicitly truncated.

For both weak and reference supervision, we minimize standard autoregressive cross-entropy over assistant-response tokens:

$$\mathcal{L} = -\sum_{t \in \mathcal{R}} \log p_\theta(y_t \mid x, y_{<t}) \tag{8}$$

where $\mathcal{R}$ denotes the response-token positions. In the weak condition, y is the selected teacher-generated reasoning trace; in the reference condition, it is the original Math-CoT response. Thus, the weak condition performs offline sequence-level imitation rather than online logit distillation: the teacher is not queried during student training, and no KL-divergence loss is used.

Training configuration and evaluation We fine-tune the student for eight epochs using AdamW with a peak learning rate of 10^{-5}, a weight decay of 10^{-2}, and a cosine learning-rate schedule with a



(a) Teacher temperature scaling

(b) Teacher label smoothing

Figure 6: **Student accuracy under entropy-matched supervision on ImageNet-1K.** Teacher targets are matched to the mean entropy of (a) AlexNet or ResNet-18 using temperature scaling, and (b) AlexNet using uniform label smoothing. In both panels, ground-truth targets are entropy-matched using uniform label smoothing. Original supervision results are included for comparison. The broad data-size-dependent shift in teacher preference persists after entropy matching.

10% warmup. The effective batch size is 30, obtained using six devices, a per-device batch size of one, and five gradient-accumulation steps. Training uses bfloat16, gradient checkpointing, dynamic padding, and DeepSpeed ZeRO-3. We do not use a validation split or intermediate checkpoint selection and evaluate the final-epoch model.

We evaluate on all 500 examples of MATH-500 using the same step-by-step and boxed-answer instruction. We generate one response per problem using greedy decoding, with a maximum of 32,768 generated tokens. We use `math_verify` to parse predictions and reference answers and check mathematical equivalence. We report the percentage of problems answered correctly. Each condition is trained once; the reported curves are not averages over multiple training seeds.

C Additional Experiments

C.1 Robustness to Supervision Softness

Entropy-matched supervision We observe the data-size-dependent supervision crossover in Section 2.2. However, teachers differ not only in predictive accuracy but also in the softness of their output distributions. A possible explanation for the low-data advantage of weaker teachers is that they provide softer supervision. To examine this possibility, we compare supervision sources after matching their average target entropy using either temperature scaling or uniform label smoothing.

We construct fixed calibration streams from the ImageNet-1K training set, selecting 50 images per class for fitting and another 50 images per class for a disjoint check stream. Neither stream is constrained to be disjoint from the downstream student training subsets. Images undergo random resized cropping to 224×224, random horizontal flipping, and ImageNet normalization. We use batches of 256 and discard the incomplete batch, yielding 49920 examples per stream.

We apply MixUp to each calibration stream and use the same mixed inputs across all supervision conditions and candidate parameter values. For each batch b, we sample a single mixing weight $\lambda_b \sim \mathrm{Beta}(0.8, 0.8)$ and a random within-batch permutation π_b. Given augmented and normalized images $\boldsymbol{x}_i$, we construct

$$\tilde{\boldsymbol{x}}_i = \lambda_b \boldsymbol{x}_i + (1 - \lambda_b)\boldsymbol{x}_{\pi_b(i)}. \tag{9}$$

We construct the corresponding targets in three ways.

Teacher temperature scaling. Each frozen teacher is evaluated directly on the mixed input. Given its logits $z_t(\tilde{\boldsymbol{x}}_i)$, we define

$$\boldsymbol{q}_i^{\mathrm{temp}} = \mathrm{softmax}\left(z_t(\tilde{\boldsymbol{x}}_i)/T\right) \tag{10}$$

Thus, the target is computed from the mixed image rather than by mixing predictions for the two original images.

Teacher label smoothing. Let $C = 1000$ denote the number of classes. We first obtain the teacher prediction at $T = 1$, $q_t(\tilde{\boldsymbol{x}}_i) = \mathrm{softmax}(z_t(\tilde{\boldsymbol{x}}_i))$, and then mix it with the uniform distribution:

$$\boldsymbol{q}_i^{\text{teacher-smooth}} = (1 - \epsilon_t)\,\mathrm{softmax}\left(z_t(\tilde{\boldsymbol{x}}_i)\right) + \epsilon_t \frac{\mathbf{1}_C}{C} \tag{11}$$

Ground-truth label smoothing. We mix the one-hot ground-truth labels using the same weight and pairing as the inputs, and then apply uniform label smoothing:

$$\boldsymbol{q}_i^{\text{hard}} = (1 - \epsilon_h)\left[\lambda_b \boldsymbol{e}_{y_i} + (1 - \lambda_b)\boldsymbol{e}_{y_{\pi_b(i)}}\right] + \epsilon_h \frac{\mathbf{1}_C}{C} \tag{12}$$

where e_y denotes the one-hot encoding of class y. Ground-truth targets are therefore already soft due to mixup, even when $\epsilon_h = 0$.

For each condition, let $\boldsymbol{q}_i$ denote the resulting target distribution for the i-th mixed input. We compute

$$H(\boldsymbol{q}_i) = -\sum_{c=1}^{C} q_{i,c} \log q_{i,c}, \qquad \bar{H} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{q}_i) \tag{13}$$

where $N = 49920$.

We use natural logarithms and report entropy in nats. The reference entropy $H^\star$ is computed from the anchor teacher's predictions at $T = 1$ on the same mixed inputs. We calibrate one global parameter per supervision source and anchor so that $\bar{H}$ matches $H^\star$. We solve for T by bisection on a logarithmic temperature scale over $T \in [0.02, 100]$, and solve for ϵ_t and ϵ_h by standard bisection over $[0, 1]$. In each case, fitting stops when $|\bar{H} - H^\star| \le 10^{-6}$ nats.

Temperature scaling uses AlexNet and ResNet-18 as two separate anchors, whereas teacher label smoothing uses the AlexNet anchor only. Ground-truth label smoothing is calibrated for both anchors. The target entropies and calibrated parameters are reported in Table 2.

Without refitting, we apply the fitted parameters to the disjoint check stream. Across all reported matched conditions, the resulting mean target entropy differs from the corresponding anchor entropy measured on the same stream by at most 0.0324 nats.

We follow the main ImageNet training protocol (Section B.1), replacing only the supervision targets with their entropy-calibrated versions. All calibrated parameters are fixed across data fractions, training seeds, and epochs. During student training, temperature scaling is applied only to teacher logits. Student logits remain unscaled, and we minimize KL divergence without a T^2 loss multiplier. We evaluate data fractions of 1%, 2%, 5%, and 10% over three training seeds.

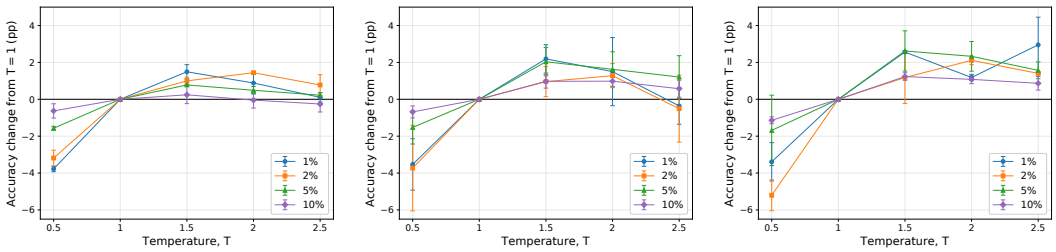
Figure 6 shows that the overall pattern of data-size-dependent supervision crossover is preserved under both entropy-matching methods: AlexNet supervision yields the highest student accuracy at the smallest data fraction, while stronger teachers become preferable as more data are available. However, the effect on accuracy depends on the matching method. Temperature-based softening improves mean student accuracy under ResNet-18 and DenseNet-121 supervision across all evaluated data fractions. In contrast, uniform teacher label smoothing reduces accuracy at the smallest data fractions and yields smaller changes at larger fractions. Temperature scaling can also alter the mean accuracy ordering of closely performing teachers, as observed for ResNet-18 and DenseNet-121 at the 5% data fraction under the ResNet-18 entropy anchor. These results suggest that the superior student performance obtained with weaker teachers in the low-data regime cannot be explained solely by differences in average target entropy.

Temperature scaling The entropy-matching experiments show that temperature-based softening can improve student accuracy. We further examine this effect over a wider range of temperatures to assess whether additional softening yields further gains. We evaluate $T \in \{0.5, 1, 1.5, 2, 2.5\}$ for AlexNet, ResNet-18, and DenseNet-121 at student data fractions of 1%, 2%, 5%, and 10%, using two random seeds.

Figure 7 shows the change in student accuracy relative to $T = 1$. Across all evaluated teachers and data fractions, $T = 1.5$ improves mean accuracy, whereas sharpening the targets with $T = 0.5$

Table 2: Entropy-matching parameters and mean target entropy on the disjoint check stream. Teacher targets are calibrated using temperature scaling (T) or uniform label smoothing (ϵ). Ground-truth MixUp targets use label smoothing (ϵ). All entropy values are in nats.

Matching method	Supervision source	Parameter	Check entropy
AlexNet anchor: target entropy = 2.662813			
Temperature	AlexNet	$T = 1$	2.6207
	ResNet-18	$T = 1.236695$	2.6034
	DenseNet-121	$T = 1.362106$	2.6060
Smoothing	AlexNet	$\epsilon_t = 0$	2.6207
	ResNet-18	$\epsilon_t = 0.086753$	2.6219
	DenseNet-121	$\epsilon_t = 0.127705$	2.6294
Smoothing	Hard labels	$\epsilon_h = 0.252463$	2.6502
ResNet-18 anchor: target entropy = 1.997025			
Temperature	AlexNet	$T = 0.792175$	1.9634
	ResNet-18	$T = 1$	1.9506
	DenseNet-121	$T = 1.123906$	1.9504
Smoothing	Hard labels	$\epsilon_h = 0.166551$	1.9830



(a) AlexNet (b) ResNet18 (c) DenseNet121

Figure 7: **Effect of teacher temperature on student accuracy on ImageNet-1K.** Panels show accuracy changes relative to $T = 1$ in percentage points for a ResNet-50 student supervised by (a) AlexNet, (b) ResNet-18, and (c) DenseNet-121 at different student data fractions. Moderate target softening improves mean accuracy, but further increases in temperature do not consistently yield additional gains. Points and error bars show the mean and sample standard deviation of paired accuracy differences across two seeds, using each seed's $T = 1$ result as its baseline.

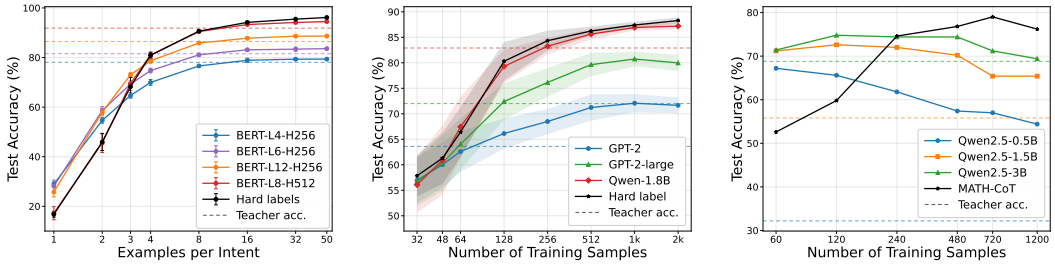
reduces it. Further increasing the temperature does not consistently improve performance, and the best-performing temperature varies across teachers and data fractions. These results suggest that moderate target softening can improve student performance, but increasing softness does not always yield additional gains.

C.2 Additional Results Across Learning Settings

Our main experiments in Section 2 reveal three findings: students can outperform their weak teachers using limited training data; in low-data regimes, weak supervision can outperform hard-label training of the same student on the same examples; and the preferred supervision source changes as the student data size increases.

Here, we examine how these findings extend to additional datasets, architectures, training protocols, and tasks, including CIFAR-100 classification, vision transformers trained from scratch, linear probing with pretrained visual representations, and language classification and generative reasoning. We report which findings are observed in each setting and discuss differences from the main experiments.

CLINC150 intent classification Section 2.3 shows a data-dependent supervision crossover on CLINC150. Here, we additionally compare student performance with the accuracy of each teacher.



(a) CLINC150 (b) SciQ (c) MATH500

Figure 8: **Additional results on language tasks.** (a) CLINC150 classification: students can outperform their respective teachers using limited training data. (b) SciQ binary classification: students can outperform their teachers, although weak supervision does not consistently outperform hard-label training. (c) MATH500 reasoning: supervision from weaker models can yield higher student accuracy than stronger-model supervision in the low-data regime.

Figure 8a shows that students can outperform their respective teachers using only a subset of the available student training data. However, unlike in the main ImageNet experiments, we do not observe a clear trend in which stronger teachers require more student data to be surpassed.

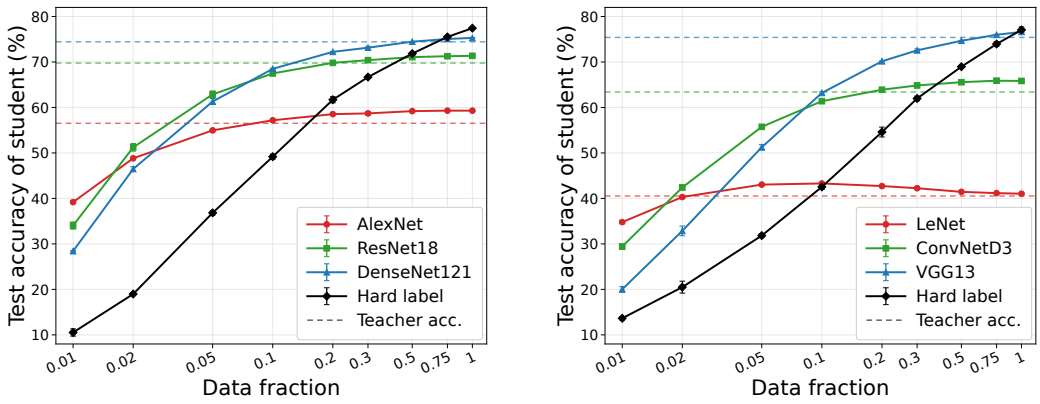
NLP binary classification We evaluate SciQ as a binary classification task following Burns et al. (2024), using Qwen-7B as the student. Figure 8b shows that students can outperform their respective teachers as the training set grows. However, hard-label training generally matches or exceeds teacher-supervised training, and we do not observe a clear data-dependent crossover among teachers. Thus, outperforming the teacher does not necessarily imply an advantage over hard-label training or a change in the preferred teacher.

Mathematical reasoning We evaluate Qwen2.5-7B on MATH500 after fine-tuning on prompts from Math-CoT-20k (Ren et al., 2026). We compare the original Long-CoT responses with responses generated by weaker teacher models. At 60 and 120 training examples, all three weaker-model target conditions yield higher student accuracy than the original MATH-CoT targets (Figure 8c). At 240 examples, MATH-CoT supervision reaches approximately the performance of the best weaker-model supervision, and it performs best at the larger evaluated training sizes. These results show that weaker-model targets can be more effective with limited training data, whereas the original MATH-CoT targets become preferable as more training examples are available.

ImageNet-1K and CIFAR-100. Figure 9 presents the full ImageNet-1K accuracy curves alongside results on CIFAR-100. The ImageNet-1K panel complements the main results in Section 2 by displaying all supervision conditions together with the individual teacher accuracies. CIFAR-100 exhibits the same three patterns: students can outperform their teachers using limited training data, teacher supervision improves over hard-label training in the low-data regime, and the preferred supervision changes with student data size. In terms of mean accuracy, LeNet supervision performs best at 1%, ConvNetD3 at 2–10%, VGG13 at 20–75%, and hard labels at 100%. Students first surpass their respective teachers at the evaluated fractions of 2%, 20%, and 75% for LeNet, ConvNetD3, and VGG13, respectively.

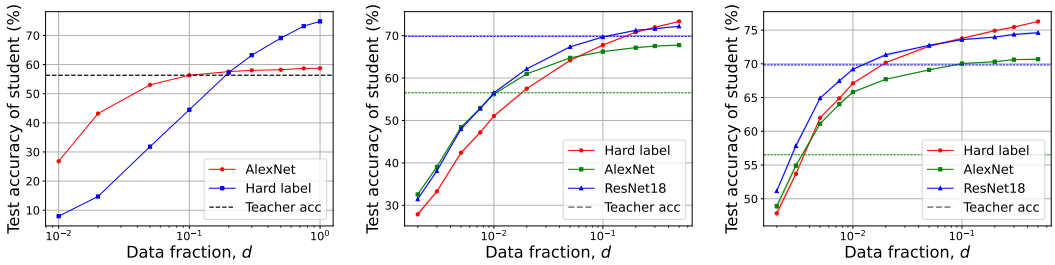
Larger architecture trained from scratch Figure 10a reports results for a ViT-B/16 student (Dosovitskiy et al., 2021) trained from scratch with AlexNet as a weak teacher. Weak supervision outperforms hard-label training in the low-data regime, and the student can also surpass its teacher. These results show that the benefits observed with CNN students extend to this transformer-based setting.

Linear probing on frozen visual representations Following Burns et al. (2024), we freeze a DINO backbone (Caron et al., 2021) and train only a linear classifier. We use ResNet-50 and ViT-B/8 as student backbones, with AlexNet or ResNet-18 as weak teachers. For ResNet-50 (Figure 10b), the preferred supervision shifts from AlexNet to ResNet-18 and eventually to hard labels as student data increase. For ViT-B/8 (Figure 10c), teacher supervision also outperforms hard-label training at low



(a) ImageNet1K (b) CIFAR-100

Figure 9: **Student performance across data fractions on ImageNet-1K and CIFAR-100.** Solid curves show student test accuracy under teacher supervision or hard-label training, and horizontal dashed lines indicate the corresponding teachers' accuracies. Across both datasets, students can surpass their teachers with limited data, and teacher supervision outperforms hard-label training in the low-data regime. The best-performing supervision shifts toward stronger teachers and eventually hard labels as student data increase.



(a) ViT-B/16, scratch (b) ResNet50, linear probing (c) ViT-B/8, linear probing

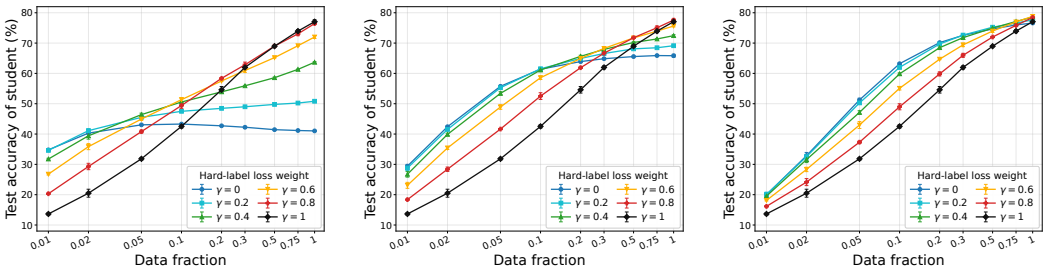
Figure 10: **Additional results on vision tasks.** (a) ViT-B/16 trained from scratch with AlexNet supervision outperforms hard-label training at low data and can surpass its teacher. (b) Linear probing on frozen DINO ResNet-50 exhibits a supervision crossover from AlexNet to ResNet-18 and eventually hard labels as student data increase. (c) With frozen DINO ViT-B/8, ResNet-18 consistently outperforms AlexNet as a supervisor, while hard-label training overtakes both teacher-supervised conditions as data increase.

data, but ResNet-18 consistently outperforms AlexNet. With more data, hard-label training surpasses both teacher-supervised conditions. Thus, the ViT-B/8 setting exhibits a crossover between teacher and hard-label supervision, rather than between the two teachers. Students can also outperform their respective teachers in both backbone settings.

C.3 Combining Supervision Sources

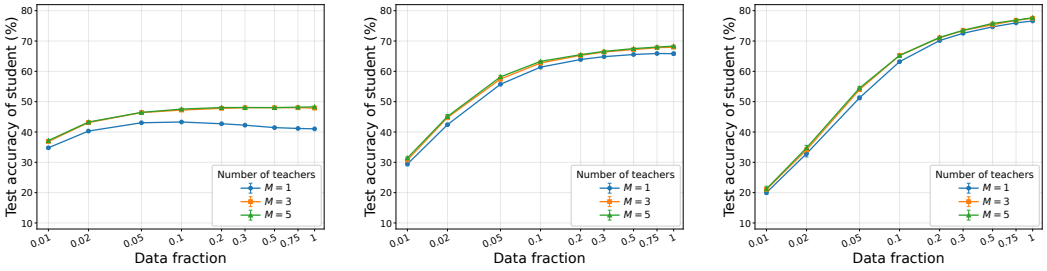
The relative benefits of teacher supervision and hard labels vary with student data size. This motivates examining whether combining their training losses can improve performance over either form of supervision alone. We also investigate teacher ensembling as an alternative to increasing the strength of individual teachers, asking whether averaging predictions from multiple teachers of the same architecture improves student performance and alleviates the saturation observed under supervision from a single weak teacher. We evaluate both approaches on CIFAR-100 across student data fractions

Combining ground-truth and teacher supervision We train students using a weighted combination of the hard-label and teacher-supervision losses,



(a) LeNet (b) ConvNetD3 (c) VGG13

Figure 11: **Combining ground-truth and teacher supervision on CIFAR-100.** Student accuracy is shown for different hard-label loss weights γ, where $\gamma = 0$ denotes teacher-only training and $\gamma = 1$ denotes hard-label-only training. As student data increase, larger hard-label weights become preferable, and intermediate weights can outperform both forms of supervision alone.



(a) LeNet (b) ConvNetD3 (c) VGG13

Figure 12: **Teacher ensembling on CIFAR-100.** Student accuracy under supervision from $M = 1$, 3, or 5 teachers of the same architecture. Ensembling improves mean accuracy over single-teacher supervision, but increasing M from three to five yields limited additional gains.

$$\mathcal{L}_\gamma = \gamma\,\mathrm{CE}\big(\tilde{\boldsymbol{y}}, p_s(\tilde{\boldsymbol{x}})\big) + (1-\gamma)\,D_{\mathrm{KL}}\big(q_t(\tilde{\boldsymbol{x}}) \,\|\, p_s(\tilde{\boldsymbol{x}})\big) \tag{14}$$

where $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ denote the MixUp input and the corresponding mixed ground-truth target, respectively. The distributions $q_t(\tilde{\boldsymbol{x}})$ and $p_s(\tilde{\boldsymbol{x}})$ are the teacher and student class-probability predictions on the mixed input. We vary the hard-label loss weight $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, where $\gamma = 0$ and $\gamma = 1$ correspond to teacher-only and hard-label-only training, respectively. Figure 11 shows that the benefit of combining the losses depends on both student data size and teacher strength. At the smallest data fraction, teacher-only supervision yields the highest mean accuracy for all three teachers. As more data become available, the best-performing weight among the evaluated candidates shifts toward a larger hard-label contribution, while stronger teachers remain useful at larger teacher-loss weights. Intermediate weights can outperform both endpoints. For example, at the full data fraction, VGG13 supervision with $\gamma = 0.6$ achieves 79.06% mean accuracy, compared with 76.71% for teacher-only training and 77.38% for hard-label training. Thus, teacher supervision can remain beneficial as part of a combined objective even when hard-label training outperforms teacher-only training.

Ensembling teachers For each architecture, we train five teachers independently using the same training set and optimization configuration but different random seeds, which vary model initialization, data ordering, and stochastic data augmentation. We construct ensembles of size $M \in \{1, 3, 5\}$ using fixed prefixes of the ordered teacher bank. All members within an ensemble share the same architecture. For each mixed input, we query all frozen ensemble members and average their class-probability predictions:

$$\boldsymbol{q}_{\mathrm{ens}}(\tilde{\boldsymbol{x}}) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{softmax}\left(z_m(\tilde{\boldsymbol{x}})\right). \tag{15}$$

We minimize $D_{\mathrm{KL}}(\boldsymbol{q}_{\mathrm{ens}} \| \boldsymbol{p}_s)$ without an additional hard-label loss. All other student-training settings follow Section B.2.

Figure 12 shows that both ensemble sizes improve mean student accuracy over single-teacher supervision across all evaluated architectures and data fractions. Increasing the ensemble size from three to five provides smaller and less consistent additional gains. Ensembling also does not eliminate the performance saturation observed with weaker teachers. For example, LeNet ensembles reach a higher accuracy level but still plateau as student data increase. These results show that ensembling improves the effectiveness of weak supervision while retaining its dependence on teacher architecture and student data size.

D Theoretical Details

We provide the details of the two-stage linear regression analysis in the main text, building on Ildiz et al. (2025). The weak teacher is fitted to responses generated from the full input while using only a subset of input coordinates. We derive the corresponding Gaussian-sequence risk expressions and use them to examine how student sample size and teacher dimension affect weak-to-strong performance and the bias–variance tradeoff.

D.1 Full Problem Setup and Notation

Weak model Let the ground-truth model be $\boldsymbol{\beta}_* \in \mathbb{R}^p$. The weak model $\boldsymbol{\beta}_w \in \mathbb{R}^{p_w}$ only accesses a subset of the features ($p_w < p$). It is estimated from the source dataset $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)_{i=1}^m$ drawn from the distribution $(\tilde{\boldsymbol{x}}, \tilde{y}) \sim \mathcal{D}_w(\boldsymbol{\beta}_*)$, where the linear model is

$$\tilde{y} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_* + \tilde{z}, \quad \tilde{\boldsymbol{x}}, \boldsymbol{\beta}_* \in \mathbb{R}^p, \quad \tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w), \quad \tilde{z} \sim \mathcal{N}(0, \sigma_w^2). \tag{16}$$

We assume that the weak model uses the first p_w coordinates of $\tilde{\boldsymbol{x}}_i$, i.e., $\tilde{\boldsymbol{x}}_{i,p_w}$ is the truncated p_w-dimensional vector of $\tilde{\boldsymbol{x}}_i$. In the over-parameterized regime ($m < p_w$), the weak estimator is

$$\hat{\boldsymbol{\beta}}_w = \arg\min_{\boldsymbol{\beta}} \{ \|\boldsymbol{\beta}\|_2 : \tilde{\boldsymbol{X}}_{p_w} \boldsymbol{\beta} = \tilde{\boldsymbol{y}} \} \tag{17}$$

where $\tilde{\boldsymbol{X}}_{p_w} \in \mathbb{R}^{m \times p_w}$ and $\tilde{\boldsymbol{y}} \in \mathbb{R}^m$ are the data matrix and label vector, respectively.

Weak-to-strong model Given the weak teacher model $\hat{\boldsymbol{\beta}}_w \in \mathbb{R}^{p_w}$, the weak-to-strong model $\boldsymbol{\beta}_{w2s} \in \mathbb{R}^p$ is estimated from the target dataset $(\boldsymbol{x}_i, y_i^w)_{i=1}^n$ drawn from $(\boldsymbol{x}, y^w) \sim \mathcal{D}_s(\boldsymbol{\beta}'_w)$, with the linear model

$$y^w = \boldsymbol{x}^\top \boldsymbol{\beta}'_w + z, \quad \boldsymbol{x} \in \mathbb{R}^p, \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_s), \quad z \sim \mathcal{N}(0, \sigma_s^2), \tag{18}$$

where $\boldsymbol{\beta}'_w = \begin{bmatrix} \hat{\boldsymbol{\beta}}_w \\ \boldsymbol{0}_{p-p_w} \end{bmatrix}$. In the over-parameterized regime ($n < p$), the estimator is

$$\hat{\boldsymbol{\beta}}_{w2s} = \arg\min_{\boldsymbol{\beta}} \{ \|\boldsymbol{\beta}\|_2 : \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{y}^w \} \tag{19}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y}^w \in \mathbb{R}^n$ are the data matrix and label vector.

Strong model We denote the original regression model as the strong model. The estimator $\boldsymbol{\beta}_s \in \mathbb{R}^p$ is obtained from the target dataset $(\boldsymbol{x}_i, y_i)_{i=1}^n$ drawn from $(\boldsymbol{x}, y) \sim \mathcal{D}_s(\boldsymbol{\beta}_*)$, with the linear model

$$y = \boldsymbol{x}^\top \boldsymbol{\beta}_* + z, \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_s), \quad z \sim \mathcal{N}(0, \sigma_s^2). \tag{20}$$

In the over-parameterized regime ($n < p$), the estimator is

$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \{ \|\boldsymbol{\beta}\|_2 : \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{y} \} \tag{21}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$.

Excess risk For any estimator $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, the excess risk on the target distribution is given by

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_s(\boldsymbol{\beta}_*)}\left[(y - \boldsymbol{x}^\top\hat{\boldsymbol{\beta}})^2\right] - \sigma_s^2 = \left\|\Sigma_s^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\right\|_2^2. \tag{22}$$

D.2 Risk Characterizations

Following Han & Xu (2026), we analyze the asymptotic risk of the estimators by applying distributional characterization results for the minimum-ℓ_2-norm interpolators under the equivalent Gaussian sequence model. Before presenting the exact risk expressions, we introduce the fundamental quantities that dictate the high-dimensional asymptotic limits.

Aspect ratios (κ_s, κ_w): We define the over-parameterization ratios for the strong and weak models as $\kappa_s = p/n$ and $\kappa_w = p_w/m$, respectively. In our proportional asymptotic regime, these ratios are fixed constants strictly greater than 1 ($\kappa_s, \kappa_w > 1$).

Implicit regularizers (τ_s, τ_w): The high-dimensional nature of the ridgeless interpolation induces an implicit ridge penalty. We denote these implicit regularizers as $\tau_s > 0$ and $\tau_w > 0$, which are defined as the unique solutions to the following fixed-point equations:

$$n = \operatorname{tr}\left((\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\right), \tag{23}$$

$$m = \operatorname{tr}\left((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}\right). \tag{24}$$

Auxiliary quantities (Ω_s, Ω_w): To capture the variance components of the asymptotic risk, we define the following auxiliary quantities based on the data covariance and the implicit regularizers:

$$\Omega_s = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_s^2(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-2}\right), \quad \Omega_w = \frac{1}{m}\operatorname{tr}\left(\boldsymbol{\Sigma}_{w,p_w}^2(\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-2}\right). \tag{25}$$

Asymptotic risk ($\bar{\mathcal{R}}$): We denote $\bar{\mathcal{R}}$ as the deterministic limit of the finite-sample excess risk evaluated under the high-dimensional proportional regime, where the sample sizes and dimensions grow to infinity (e.g., $n, p \to \infty$) while maintaining their respective fixed aspect ratios.

Asymptotic risk of strong model We first define two quantities: the distribution of the strong model estimator and the noise-amplitude functional $\gamma_s : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$, which are characterized by the following fixed-point equation:

$$\tilde{\boldsymbol{\beta}}_s(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_*, \boldsymbol{g}_s) := (\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\beta}_* + \gamma_s(\boldsymbol{\beta}_*)(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\frac{\boldsymbol{g}_s}{\sqrt{p}} \tag{26}$$

$$\gamma_s^2(\boldsymbol{\beta}_*) := \kappa_s(\sigma_s^2 + \bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_s(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_*, \boldsymbol{g}_s))) \tag{27}$$

where $\boldsymbol{g}_s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Under the assumptions of Theorem 1 in Ildiz et al. (2025), the strong estimator's risk concentrates around the expected risk of its Gaussian-sequence counterpart. These assumptions include a fixed positive lower bound on the noise variance, uniformly bounded covariance and inverse-covariance operator norms, and bounded signal norms as the dimension grows. Note that the finite-sample estimator depends on the realized training samples, whereas its Gaussian-sequence counterpart represents this randomness through $\boldsymbol{g}_s$. In our noiseless setting, the distribution of this counterpart is determined by $\boldsymbol{\Sigma}_s$, $\boldsymbol{\beta}_*$, and the sample size and dimension.

The asymptotic risk of strong model is defined as

$$\begin{aligned}
\bar{\mathcal{R}}_s &:= \bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_s) := \mathbb{E}_{\boldsymbol{g}_s}\left[\left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_*)\right\|_2^2\right] \\
&= \left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}\left((\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\beta}_* - \boldsymbol{\beta}_*\right)\right\|_2^2 + \gamma_s^2(\boldsymbol{\beta}_*)\,\mathbb{E}_{\boldsymbol{g}_s}\left[\left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\frac{\boldsymbol{g}_s}{\sqrt{p}}\right\|_2^2\right] \qquad (28) \\
&= \left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}\left(\boldsymbol{I}_p - (\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\right)\boldsymbol{\beta}_*\right\|_2^2 + \frac{\gamma_s^2(\boldsymbol{\beta}_*)}{p}\operatorname{tr}\left(\boldsymbol{\Sigma}_s^2(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-2}\right) \\
&= \tau_s^2\left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_*\right\|_2^2 + \frac{n\Omega_s}{p}\kappa_s\left(\sigma_s^2 + \bar{\mathcal{R}}_s\right). \qquad (29)
\end{aligned}$$

So,

$$\bar{\mathcal{R}}_s = \frac{\sigma_s^2\Omega_s + \tau_s^2\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_*\|_2^2}{1 - \Omega_s} \tag{30}$$

Asymptotic risk of weak model Similarly, the distribution of weak model estimator and the noise-amplitude functional $\gamma_w : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is characterized by the following fixed-point equation:

$$\tilde{\boldsymbol{\beta}}_w(\boldsymbol{\Sigma}_w, \boldsymbol{\beta}_*, \boldsymbol{g}_w) := (\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w} \boldsymbol{\beta}_{*,p_w} + \gamma_w(\boldsymbol{\beta}_*)(\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}} \frac{\boldsymbol{g}_w}{\sqrt{p_w}} \tag{31}$$

$$\gamma_w^2(\boldsymbol{\beta}_*) := \kappa_w \Big(\sigma_w^2 + \|\boldsymbol{\Sigma}_{w,(p_w+1:p)}^{\frac{1}{2}} \boldsymbol{\beta}_{*,(p_w+1:p)}\|_2^2 + \bar{\mathcal{R}}^{(w)}(\tilde{\boldsymbol{\beta}}_w(\boldsymbol{\Sigma}_w, \boldsymbol{\beta}_*, \boldsymbol{g}_w)) \Big) \tag{32}$$

where $\boldsymbol{g}_w \sim \mathcal{N}(0, \boldsymbol{I}_{p_w})$, and $\bar{\mathcal{R}}^{(w)}(\tilde{\boldsymbol{\beta}}_w) := \mathbb{E}_{\boldsymbol{g}_w} \left[\left\| \boldsymbol{\Sigma}_{w,p_w}^{1/2} (\tilde{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_{*,p_w}) \right\|_2^2 \right]$ denotes the expected squared estimation error in the first p_w coordinates, weighted by $\boldsymbol{\Sigma}_{w,p_w}$. To simplify the analysis, we assume that the features are partitioned into used and unused dimensions such that they are uncorrelated, represented by the block-diagonal covariance matrix $\boldsymbol{\Sigma}_w = \begin{bmatrix} \boldsymbol{\Sigma}_{w,p_w} & \boldsymbol{0}_{p_w \times (p-p_w)} \\ \boldsymbol{0}_{(p-p_w)\times p_w} & \boldsymbol{\Sigma}_{w,(p_w+1:p)} \end{bmatrix}$. The weak estimator in equation 17 has access to only the first p_w dimensions of the data. Consequently, the remaining dimensions effectively function as additional label noise, as the label generation process for the weak model can be decomposed as follows:

$$\tilde{y} = \tilde{\boldsymbol{x}}_{p_w}^\top \boldsymbol{\beta}_{*,p_w} + \tilde{\boldsymbol{x}}_{(p_w+1):p}^\top \boldsymbol{\beta}_{*,(p_w+1:p)} + \tilde{z} = \tilde{\boldsymbol{x}}_{p_w}^\top \boldsymbol{\beta}_{*,p_w} + \tilde{z}', \quad \text{where } \tilde{z}' \sim \mathcal{N}\left(0, \sigma_w^2 + \|\boldsymbol{\Sigma}_{w,(p_w+1:p)}^{\frac{1}{2}} \boldsymbol{\beta}_{*,(p_w+1:p)}\|_2^2\right) \tag{33}$$

Also, we introduce zero padding function embedding function, $Z : \mathbb{R}^{p_w} \to \mathbb{R}^p$ and $\boldsymbol{Z} : \mathbb{R}^{p_w \times p_w} \to \mathbb{R}^{p \times p}$, which are $Z(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0}_{p-p_w} \end{bmatrix}$ and $\boldsymbol{Z}(\boldsymbol{\Sigma}) = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{p_w \times (p-p_w)} \\ \boldsymbol{0}_{(p-p_w)\times p_w} & \boldsymbol{0}_{(p-p_w)\times(p-p_w)} \end{bmatrix}$.

Then, asymptotic risk of weak model can be characterized as

$$\bar{\mathcal{R}}_w := \bar{\mathcal{R}}(Z(\tilde{\boldsymbol{\beta}}_w)) = \mathbb{E}_{\boldsymbol{g}_w} \left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}} (Z(\tilde{\boldsymbol{\beta}}_w) - \boldsymbol{\beta}_*)\|_2^2 \right] \tag{34}$$

$$= \left\| \boldsymbol{\Sigma}_s^{\frac{1}{2}} (Z((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w} \boldsymbol{\beta}_{*,p_w}) - \boldsymbol{\beta}_*) \right\|_2^2 + \gamma_w^2(\boldsymbol{\beta}_*) \mathbb{E}_{\boldsymbol{g}_w} \left[\left\| \boldsymbol{\Sigma}_s^{\frac{1}{2}} Z \left((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}} \frac{\boldsymbol{g}_w}{\sqrt{p_w}} \right) \right\|_2^2 \right] \tag{35}$$

$$= \left\| \boldsymbol{\Sigma}_s^{\frac{1}{2}} (Z((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w} \boldsymbol{\beta}_{*,p_w}) - \boldsymbol{\beta}_*) \right\|_2^2 + \frac{\gamma_w^2(\boldsymbol{\beta}_*)}{p_w} \mathrm{tr}(\boldsymbol{\Sigma}_s \boldsymbol{Z}((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1} \boldsymbol{\Sigma}_{w,p_w} (\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1})) \tag{36}$$

where

$$\gamma_w^2(\boldsymbol{\beta}_*) = \kappa_w \frac{\sigma_w^2 + \|\boldsymbol{\Sigma}_{w,(p_w+1:p)}^{\frac{1}{2}} \boldsymbol{\beta}_{*,(p_w+1:p)}\|_2^2 + \tau_w^2 \|(\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}} \boldsymbol{\beta}_{*,p_w}\|_2^2}{1 - \Omega_w} \tag{37}$$

Asymptotic risk of weak-to-strong model The distribution of the weak-to-strong model estimator and the noise-amplitude functional $\gamma_s : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is characterized by the following fixed-point equation:

$$\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s) := (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1} \boldsymbol{\Sigma}_s Z(\tilde{\boldsymbol{\beta}}_w) + \gamma_s(Z(\tilde{\boldsymbol{\beta}}_w))(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1} \boldsymbol{\Sigma}_s^{\frac{1}{2}} \frac{\boldsymbol{g}_s}{\sqrt{p}} \tag{38}$$

$$\gamma_s^2\big(Z(\tilde{\boldsymbol{\beta}}_w)\big) := \kappa_s \left(\sigma_s^2 + \mathbb{E}_{\boldsymbol{g}_s} \left[\left\| \boldsymbol{\Sigma}_s^{1/2} \big(\tilde{\boldsymbol{\beta}}_{w2s} - Z(\tilde{\boldsymbol{\beta}}_w) \big) \right\|_2^2 \right] \right) \tag{39}$$

Asymptotic risk of weak-to-strong model can be characterized as

$$\bar{\mathcal{R}}_{w2s} := \bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s)) \tag{40}$$

$$= \mathbb{E}_{\boldsymbol{g}_w, \boldsymbol{g}_s}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s) - \boldsymbol{\beta}_*)\|_2^2\right] \tag{41}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}((\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s Z(\tilde{\boldsymbol{\beta}}_w) - \boldsymbol{\beta}_*)\|_2^2\right] \tag{42}$$

$$+ \mathbb{E}_{\boldsymbol{g}_w}\left[\gamma_s^2(Z(\tilde{\boldsymbol{\beta}}_w))\mathbb{E}_{\boldsymbol{g}_s}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\frac{\boldsymbol{g}_s}{\sqrt{p}}\|_2^2\right]\right] \tag{43}$$

The first term, equation 42, is

$$\mathbb{E}_{\boldsymbol{g}_w}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}((\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s Z(\tilde{\boldsymbol{\beta}}_w) - \boldsymbol{\beta}_*)\|_2^2\right] \tag{44}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}\left\{(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s Z\left((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}\boldsymbol{\beta}_{*,p_w}\right.\right.\right.\right.$$

$$\left.\left.\left.\left. + \gamma_w(\boldsymbol{\beta}_*)(\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\frac{\boldsymbol{g}_w}{\sqrt{p_w}}\right) - \boldsymbol{\beta}_*\right\}\right\|_2^2\right] \tag{45}$$

$$= \left\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}\left\{\boldsymbol{\beta}_* - (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s \boldsymbol{Z}\left((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\right)\boldsymbol{Z}(\boldsymbol{\Sigma}_{w,p_w})Z(\boldsymbol{\beta}_{*,p_w})\right\}\right\|_2^2 \tag{46}$$

$$+ \frac{\gamma_w^2(\boldsymbol{\beta}_*)}{p_w}\mathrm{tr}\Big(\boldsymbol{Z}\left(\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}(\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\right)\boldsymbol{\Sigma}_s(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s$$

$$\times (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s \boldsymbol{Z}\left((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\right)\Big) \tag{47}$$

The second term, equation 43, is

$$\mathbb{E}_{\boldsymbol{g}_w}\left[\gamma_s^2(Z(\tilde{\boldsymbol{\beta}}_w))\mathbb{E}_{\boldsymbol{g}_s}\left[\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\frac{\boldsymbol{g}_s}{\sqrt{p}}\|_2^2\right]\right] \tag{48}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\gamma_s^2(Z(\tilde{\boldsymbol{\beta}}_w))\right]\frac{1}{p}\mathrm{tr}(\boldsymbol{\Sigma}_s^2(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I})^{-2}) \tag{49}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\gamma_s^2(Z(\tilde{\boldsymbol{\beta}}_w))\right]\frac{n}{p}\Omega_s \tag{50}$$

Also, similar with Lemma 1. in Ildiz et al. (2025),

$$\mathbb{E}_{\boldsymbol{g}_w}\left[\gamma_s^2(Z(\tilde{\boldsymbol{\beta}}_w))\right] \tag{51}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\kappa_s\left(\sigma_s^2 + \mathbb{E}_{\boldsymbol{g}_s}\left[\left\|\boldsymbol{\Sigma}_s^{1/2}\big(\tilde{\boldsymbol{\beta}}_{w2s} - Z(\tilde{\boldsymbol{\beta}}_w)\big)\right\|_2^2\right]\right)\right] \tag{52}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\kappa_s\left(\sigma_s^2 + \frac{\sigma_s^2\Omega_s + \tau_s^2\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}Z(\tilde{\boldsymbol{\beta}}_w)\|_2^2}{1-\Omega_s}\right)\right] \tag{53}$$

$$= \mathbb{E}_{\boldsymbol{g}_w}\left[\frac{\kappa_s}{1-\Omega_s}\left(\sigma_s^2 + \tau_s^2\Big\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\right.\right.$$
$$\left.\left.\times Z\Big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}\boldsymbol{\beta}_{*,p_w} + \gamma_w(\boldsymbol{\beta}_*)(\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\frac{\boldsymbol{g}_w}{\sqrt{p_w}}\Big)\Big\|_2^2\right)\right] \tag{54}$$

$$= \frac{\kappa_s}{1-\Omega_s}\Big(\sigma_s^2 + \tau_s^2\big\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{Z}((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w})Z(\boldsymbol{\beta}_{*,p_w})\big\|_2^2\Big)$$
$$+ \frac{\kappa_s\tau_s^2\gamma_w^2(\boldsymbol{\beta}_*)}{p_w(1-\Omega_s)}\mathrm{tr}\Big(\boldsymbol{Z}\big(\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}(\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\big)\boldsymbol{\Sigma}_s^{\frac{1}{2}}$$
$$\times(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{Z}\big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\big)\Big) \tag{55}$$

So, the asymptotic risk of weak-to-strong model is

$$\bar{\mathcal{R}}_{w2s} := \bar{\mathcal{R}}\big(\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s)\big) \tag{56}$$

$$= \Big\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}\Big\{\boldsymbol{\beta}_* - (\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\boldsymbol{Z}\big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}\big)Z(\boldsymbol{\beta}_{*,p_w})\Big\}\Big\|_2^2 \tag{57}$$
$$+ \frac{\Omega_s}{1-\Omega_s}\Big(\sigma_s^2 + \tau_s^2\big\|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}$$
$$\times\boldsymbol{Z}\big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}\big)Z(\boldsymbol{\beta}_{*,p_w})\big\|_2^2\Big) \tag{58}$$
$$+ \frac{\gamma_w^2(\boldsymbol{\beta}_*)}{p_w}\Big\{\frac{\Omega_s\tau_s^2}{1-\Omega_s}\mathrm{tr}\Big(\boldsymbol{Z}\big(\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}(\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\big)\boldsymbol{\Sigma}_s^{\frac{1}{2}}$$
$$\times(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{Z}\big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\big)\Big) \tag{59}$$
$$+ \mathrm{tr}\Big(\boldsymbol{Z}\big(\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}(\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\big)\boldsymbol{\Sigma}_s(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s$$
$$\times(\boldsymbol{\Sigma}_s + \tau_s\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s\boldsymbol{Z}\big((\boldsymbol{\Sigma}_{w,p_w} + \tau_w\boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w}^{\frac{1}{2}}\big)\Big)\Big\} \tag{60}$$

D.3 Bias-Variance Decomposition

In this section, we analyze how the dimension of the weak model (p_w) and the amount of data used to train the student model (n) affect the excess risk of the weak-to-strong model through bias–variance decomposition.

In the risk of an estimator, the variance component arises from the noise-amplitude functional γ, also referred to as the effective noise (Han & Xu, 2026), which is influenced by both the label noise σ and the error (risk). For example, the bias–variance decomposition of the risk of the strong model

estimator equation 30 is

$$\bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_s(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_*, \boldsymbol{g}_s)) = \underbrace{\tau_s^2 \|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_*)\|_2^2}_{\text{Bias term}} \tag{61}$$

$$+ \underbrace{\frac{\Omega_s}{1-\Omega_s}\left(\sigma_s^2 + \tau_s^2 \|\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_*)\|_2^2\right)}_{\text{Variance term}} \tag{62}$$

because the mean of the estimator is

$$\mathbb{E}_{\boldsymbol{g}_s}[\tilde{\boldsymbol{\beta}}_s(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_*, \boldsymbol{g}_s)] = (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s \boldsymbol{\beta}_*. \tag{63}$$

Similarly, the bias–variance decomposition of the weak-to-strong model risk equation 56 has its bias component given by equation 57, since the mean of the estimator is

$$\mathbb{E}_{\boldsymbol{g}_w, \boldsymbol{g}_s}[\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s)] = (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_s \boldsymbol{Z}((\boldsymbol{\Sigma}_{w,p_w} + \tau_w \boldsymbol{I}_{p_w})^{-1}\boldsymbol{\Sigma}_{w,p_w})\boldsymbol{\beta}_* \tag{64}$$

The remaining contributions which are equation 58, 59 and 60 correspond to the variance terms.

Assumption. For simplicity, we assume that both covariances are diagonal:

$$\boldsymbol{\Sigma}_w = \operatorname{diag}\big(\lambda_1^w, \ldots, \lambda_p^w\big), \qquad \boldsymbol{\Sigma}_s = \operatorname{diag}\big(\lambda_1^s, \ldots, \lambda_p^s\big), \tag{65}$$

with each sequence ordered in descending magnitude. In this setup, the weak teacher uses only the coordinates with the largest covariance eigenvalues, i.e., the top-p_w coordinates. When varying p_w, we keep the aspect ratio $\kappa_w = p_w/m$ fixed.

We can express risk through eigenvalue of covariance matrix. The risk of strong model is

$$\bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_s(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_*, \boldsymbol{g}_s)) = \underbrace{\sum_{i=1}^{p} \lambda_i^s \left(1 - \frac{\lambda_i^s}{\lambda_i^s + \tau_s}\right)^2 \beta_i^2}_{\text{Bias term}} + \frac{\Omega_s}{1-\Omega_s}\left(\sigma_s^2 + \sum_{i=1}^{p} \frac{\tau_s^2 \lambda_i^s}{(\lambda_i^s + \tau_s)^2}\beta_i^2\right) \tag{66}$$

The risk of weak model can be characterized as

$$\bar{\mathcal{R}}(Z(\tilde{\boldsymbol{\beta}}_w(\boldsymbol{\Sigma}_w, \boldsymbol{\beta}_*, \boldsymbol{g}_w))) \tag{67}$$

$$= \underbrace{\sum_{i=1}^{p_w} \lambda_i^s \left(1 - \frac{\lambda_i^w}{\lambda_i^w + \tau_w}\right)^2 \beta_i^2 + \sum_{i=p_w+1}^{p} \lambda_i^s \beta_i^2}_{\text{Bias term}} \tag{68}$$

$$+ \frac{1}{m(1-\Omega_w)}\left(\sigma_w^2 + \sum_{i=p_w+1}^{p} \lambda_i^w \beta_i^2 + \sum_{i=1}^{p_w} \frac{\tau_w^2 \lambda_i^w}{(\lambda_i^w + \tau_w)^2}\beta_i^2\right)\left(\sum_{i=1}^{p_w} \frac{\lambda_i^s \lambda_i^w}{(\lambda_i^w + \tau_w)^2}\right) \tag{69}$$

Also, risk of weak-to-strong model is

$$\bar{\mathcal{R}}(\tilde{\boldsymbol{\beta}}_{w2s}(\boldsymbol{\Sigma}_s, Z(\tilde{\boldsymbol{\beta}}_w), \boldsymbol{g}_s)) \tag{70}$$

$$= \underbrace{\sum_{i=1}^{p_w} \lambda_i^s \left(1 - \frac{\lambda_i^s}{\lambda_i^s + \tau_s} \frac{\lambda_i^w}{\lambda_i^w + \tau_w}\right)^2 \beta_i^2 + \sum_{i=p_w+1}^{p} \lambda_i^s \beta_i^2}_{\text{Bias term}} \tag{71}$$

$$+ \frac{\Omega_s}{1-\Omega_s} \left(\sigma_s^2 + \sum_{i=1}^{p_w} \frac{\tau_s^2 \lambda_i^s}{(\lambda_i^s + \tau_s)^2} \frac{(\lambda_i^w)^2}{(\lambda_i^w + \tau_w)^2} \beta_i^2\right) \tag{72}$$

$$+ \frac{1}{m(1-\Omega_w)} \left(\sigma_w^2 + \sum_{i=p_w+1}^{p} \lambda_i^w \beta_i^2 + \sum_{i=1}^{p_w} \frac{\tau_w^2 \lambda_i^w}{(\lambda_i^w + \tau_w)^2} \beta_i^2\right) \times \left(\frac{\Omega_s}{1-\Omega_s} \sum_{i=1}^{p_w} \frac{\tau_s^2 \lambda_i^s \lambda_i^w}{(\lambda_i^w + \tau_w)^2 (\lambda_i^s + \tau_s)^2}\right) \tag{73}$$

$$+ \frac{1}{m(1-\Omega_w)} \left(\sigma_w^2 + \sum_{i=p_w+1}^{p} \lambda_i^w \beta_i^2 + \sum_{i=1}^{p_w} \frac{\tau_w^2 \lambda_i^w}{(\lambda_i^w + \tau_w)^2} \beta_i^2\right) \times \left(\sum_{i=1}^{p_w} \frac{\lambda_i^w (\lambda_i^s)^3}{(\lambda_i^w + \tau_w)^2 (\lambda_i^s + \tau_s)^2}\right) \tag{74}$$

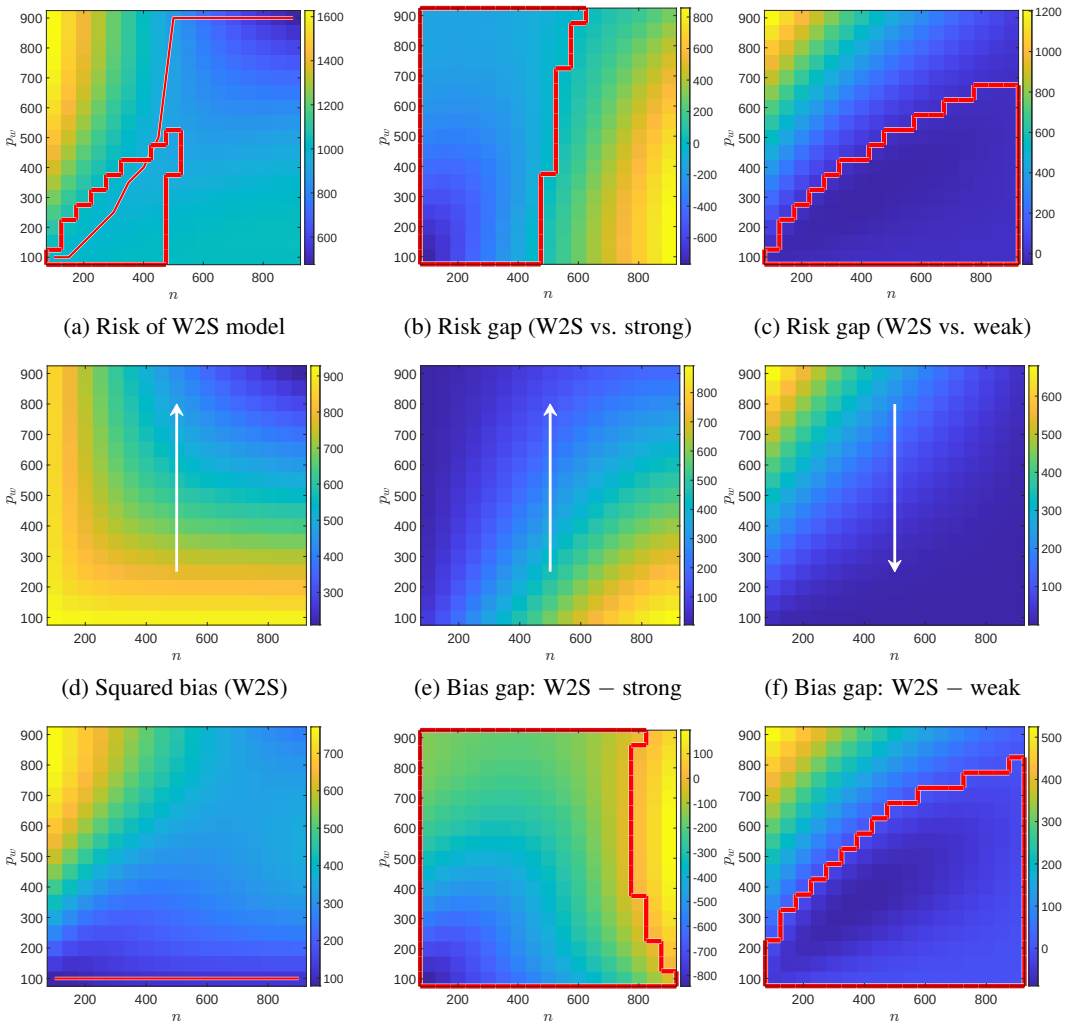
For the numerical results, we evaluate these risk expressions in the noiseless power-law setting specified in the main text. This setting does not satisfy all the assumptions of the cited convergence theorem, so the theorem does not directly guarantee convergence for these numerical configurations.

Our main results show how the weak-teacher dimension p_w and the target data size n affect the risk achieved by the weak-to-strong model. Across a grid of (n, p_w) values, Figure 13a plots the weak-to-strong risk, Figure 13b shows the risk gap between the weak-to-strong and strong models, and Figure 13c shows the risk gap between the weak-to-strong and weak models. As observed in Figure 13a, for each fixed n, the risk-minimizing weak-teacher dimension $p_w^\star(n)$ is small when n is small and shifts to larger values as n increases. This indicates that, in the data-scarce regime, training the student on labels provided by a weaker teacher can achieve lower risk.

Moreover, Figure 13a reveals a regime in which the weak-to-strong model attains lower risk than both the strong and weak models; this corresponds to the overlapping red region in Figures 13b and 13c. The red region in Figure 13b indicates where the risk of the weak-to-strong model is lower than that of the strong model, mainly in the low-data regime. As p_w increases, the range of n for which the weak-to-strong model outperforms the strong model also expands. This implies that, by using a better weak model, the weak-to-strong model can become more advantageous as the amount of data increases. The red region in Figure 13c indicates where the weak-to-strong model achieves lower risk than the weak model. As p_w increases—i.e., as we use a better weak teacher—a larger sample size n is required for the weak-to-strong model to outperform the weak model. These trends mirror the empirical observations in Section 2.2 and highlight how the target data size n and the label quality—governed by the weak-teacher dimension p_w—jointly affect weak-to-strong performance.

We next explain why a weaker teacher can be beneficial in the low-data regime by examining the bias–variance decomposition. Across the same (n, p_w) grid, Figures 13d and 13g plot the squared bias and variance components, respectively. In the evaluated setting, squared bias decreases as p_w increases. A larger teacher dimension allows additional signal coordinates to contribute to the student predictions, reducing the contribution from omitted coordinates in equation 71. In contrast, the variance tends to be smaller for smaller p_w. When student data are scarce, the variance increase outweighs the bias reduction over much of the evaluated range, favoring smaller teacher dimensions. With more student data, the bias reduction becomes more substantial relative to the variance cost, making larger teacher dimensions preferable.

Figures 13e and 13f show that the weak-to-strong model has higher squared bias than both baselines in the evaluated setting. It can nevertheless achieve lower total risk when its variance reduction



(a) Risk of W2S model (b) Risk gap (W2S vs. strong) (c) Risk gap (W2S vs. weak)

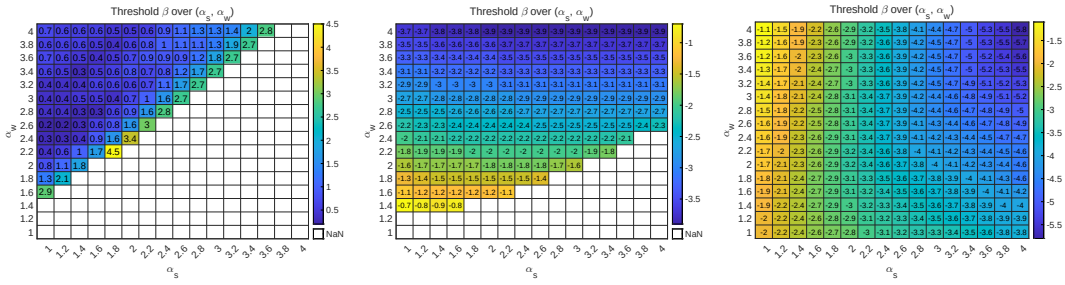
(d) Squared bias (W2S) (e) Bias gap: W2S − strong (f) Bias gap: W2S − weak

(g) Variance (W2S) (h) Variance gap: W2S − strong (i) Variance gap: W2S − weak

Figure 13: **Risk and bias–variance decomposition over the (n, p_w) grid. (a–c) Risk.** (a) Risk of the weak-to-strong estimator. The red curve connects the risk-minimizing teacher dimensions $p_w^\star(n)$ over the evaluated grid; the preferred dimension increases with student sample size. Red shading indicates where W2S outperforms both baselines. (b,c) Risk differences between W2S and the strong and weak baselines, respectively; red regions indicate lower W2S risk. **(d–f) Squared bias.** (d) W2S squared bias decreases as n and p_w increase in the evaluated setting. (e,f) Squared-bias differences relative to the strong and weak baselines, respectively; W2S has no squared-bias advantage over either baseline in this setting. **(g–i) Variance.** (g) W2S variance tends to be smaller at smaller p_w. (h,i) Variance-component differences relative to the strong and weak baselines, respectively; red regions indicate lower W2S variance. All differences are computed as W2S minus the corresponding baseline. **Parameter settings:** $\kappa_w = 1.25$, $\lambda_i^w = i^{-4}$, $\lambda_i^s = i^{-2}$, $\beta_i = i$, and $\sigma_w = \sigma_s = 0$.

outweighs this additional bias. We next examine the variance contributions relative to the strong and weak baselines.

First, the fundamental reason why the variance of the weak-to-strong model can be lower than that of the strong model is that the weak teacher uses only p_w input coordinates, restricting the signal used to train the student. As shown in equation 66, the variance term for the strong model is the sum of the contributions over all p dimensions. In the evaluated setting, the strong model's variance tends to increase as the student sample size n decreases, although this trend need not hold outside the examined range. For the weak-to-strong model, the variance contribution in equation 72, associated



(a) Both baselines (b) Strong baseline (c) Weak baseline

Figure 14: **Parameter analysis.** Threshold β over (α_s, α_w). For each (α_s, α_w), the heatmaps show the minimum β such that there exists a region in (n, p_w) where the weak-to-strong model attains lower risk than (a) both models, (b) the strong model, or (c) the weak models. For each (α_s, α_w), we fix $p = 1000$ and search over $\beta \in \{-10, -9.9, \ldots, 5\}$ and (n, p_w) on the same grid $n, p_w \in \{0.1p, 0.15p, \ldots, 0.9p\}$, and plot the smallest β for which there exists (n, p_w) such that the weak-to-strong model achieves lower risk than the strong model (b), the weak model (c), or both models (a), respectively.

with the teacher's expected prediction, includes only the first p_w signal coordinates rather than all p coordinates, and the retained contributions are further reduced by teacher shrinkage. Although teacher estimation error contributes additional variance, the net effect is a reduction relative to the strong baseline over a broad region of the evaluated grid, as shown in Figure 13h. However, when the teacher uses only the first p_w coordinates, the omitted coordinates contribute $\sum_{i=p_w+1}^{p} \lambda_i^s \beta_i^2$ to the squared bias. Consequently, when considering the bias gap, the region where the total risk is smaller is more concentrated in the low-data regime, where the bias gap is relatively small, as shown in Figure 13e.

Second, the reason why the variance of the weak-to-strong model can be lower than that of the weak model is noise reduction. For the weak model, the dimensions $p_w + 1$ through p, which the weak model has not seen, act as effective noise during weak-teacher estimation. This additional noise, represented by the term $\sum_{i=p_w+1}^{p} \lambda_i^w \beta_i^2$ in the variance of the weak model, is shown in equation 69. The contribution of this noise to the variance is weighted by the term $\sum_{i=1}^{p_w} \frac{\lambda_i^s \lambda_i^w}{(\lambda_i^w + \tau_w)^2}$, which can be reduced to $\sum_{i=1}^{p_w} \frac{\lambda_i^s \lambda_i^w}{(\lambda_i^w + \tau_w)^2} \left(\frac{\lambda_i^s}{\lambda_i^s + \tau_s} \right)^2$ in the propagated teacher variance term through the two-stage training process. Weak-to-strong training helps to regularize the learning process by reducing the variance inherited from the weak teacher, acting like a noise reduction mechanism.

However, the above phenomena—where the weak-to-strong model outperforms both the weak and strong models—do not hold universally across all parameter settings. We therefore empirically analyze how the eigenvalue decay of the data covariance, α_w and α_s, and the signal magnitude $\beta_i^2 = i^\beta$ affect the weak-to-strong model. In Figure 14, for each pair (α_s, α_w) we mark whether there exists a nontrivial region in (n, p_w) where the weak-to-strong model outperforms both the weak and strong models. The color indicates the minimum tested value of β for which such a region exists, and for all larger tested values of β, there is some regime in which the weak-to-strong model is superior to both baselines.

Within the evaluated parameter grid, we observe two main trends associated with the weak-to-strong model outperforming both baselines. First, we observe that it is favorable when α_w is larger than α_s. This trend is closely related to when the weak-to-strong model can outperform the strong model. Intuitively, a larger α_w means that the covariance eigenvalues decay more rapidly, reducing the effective noise from omitted coordinates. This can help the weak model estimate the retained signal components more accurately, making it easier for the weak-to-strong model to outperform the strong model.

Second, within the evaluated grid, the parameter β must reach a positive threshold for the weak-to-strong model to outperform both baselines. The parameter β controls the signal strength through $\beta_i^2 = i^\beta$, and $\beta > 0$ means that the squared signal coefficients increase toward the tail directions,

corresponding to smaller eigenvalues. Recall that the weak-to-strong model can outperform both the weak and strong models when variance reduction outweighs the additional bias. We also observe that the region where the weak-to-strong model outperforms both baselines expands as β increases.

D.4 Derivation of the cross-entropy bias–variance decomposition

Following Zhou et al. (2021), we derive the decomposition using the notation in Section 3.4. Since the ground-truth evaluation labels $\boldsymbol{y}$ are one-hot, their entropy is zero, so cross-entropy equals KL divergence. Therefore,

$$\mathcal{E}_{\mathrm{CE}}(d,s) = \mathbb{E}_{\mathcal{T}^{(d,s)}}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[-\sum_{c=1}^{C} y_c \log f_c(\boldsymbol{x};\mathcal{T}^{(d,s)})\right] \tag{75}$$

$$= \mathbb{E}_{\mathcal{T}^{(d,s)}}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[-\sum_{c=1}^{C} y_c \log \bar{f}_c^{(d,s)}(\boldsymbol{x}) + \sum_{c=1}^{C} y_c \log \frac{\bar{f}_c^{(d,s)}(\boldsymbol{x})}{f_c(\boldsymbol{x};\mathcal{T}^{(d,s)})}\right] \tag{76}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[D_{\mathrm{KL}}\big(\boldsymbol{y}\,\|\,\bar{f}^{(d,s)}(\boldsymbol{x})\big)\right] + \mathbb{E}_{\mathcal{T}^{(d,s)}}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\sum_{c=1}^{C} y_c \log \frac{\bar{f}_c^{(d,s)}(\boldsymbol{x})}{f_c(\boldsymbol{x};\mathcal{T}^{(d,s)})}\right] \tag{77}$$

$$= \underbrace{\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[D_{\mathrm{KL}}\big(\boldsymbol{y}\,\|\,\bar{f}^{(d,s)}(\boldsymbol{x})\big)\right]}_{\mathrm{Bias}_{\mathrm{CE}}} + \underbrace{\mathbb{E}_{\mathcal{T}^{(d,s)}}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[D_{\mathrm{KL}}\big(\bar{f}^{(d,s)}(\boldsymbol{x})\,\|\,f(\boldsymbol{x};\mathcal{T}^{(d,s)})\big)\right]}_{\mathrm{Var}_{\mathrm{CE}}} \tag{78}$$

To explain the last equality, fix $\boldsymbol{x}$, d, and s, and write $f_c = f_c(\boldsymbol{x};\mathcal{T}^{(d,s)})$, $\bar{f}_c = \bar{f}_c^{(d,s)}(\boldsymbol{x})$, and $\mathbb{E}_{\mathcal{T}} = \mathbb{E}_{\mathcal{T}^{(d,s)}}$. By the normalized geometric-mean definition,

$$\bar{f}_c = \frac{\exp(\mathbb{E}_{\mathcal{T}}[\log f_c])}{Z}, \qquad Z = \sum_{j=1}^{C}\exp(\mathbb{E}_{\mathcal{T}}[\log f_j]). \tag{79}$$

Taking logarithms gives

$$\log \bar{f}_c = \mathbb{E}_{\mathcal{T}}[\log f_c] - \log Z \tag{80}$$

and hence

$$\mathbb{E}_{\mathcal{T}}\left[\log \frac{\bar{f}_c}{f_c}\right] = \log \bar{f}_c - \mathbb{E}_{\mathcal{T}}[\log f_c] = -\log Z. \tag{81}$$

This expected log ratio is the same for every class. Since $\sum_c y_c = \sum_c \bar{f}_c = 1$, we obtain

$$\mathbb{E}_{\mathcal{T}}\left[\sum_{c=1}^{C} y_c \log \frac{\bar{f}_c}{f_c}\right] = -\log Z \tag{82}$$

$$= \mathbb{E}_{\mathcal{T}}\left[\sum_{c=1}^{C} \bar{f}_c \log \frac{\bar{f}_c}{f_c}\right] \tag{83}$$

$$= \mathbb{E}_{\mathcal{T}}\left[D_{\mathrm{KL}}(\bar{f}\|f)\right]. \tag{84}$$

Taking the expectation over the evaluation distribution yields the variance term.